QBIO490: Multi-omic Data Analysis
Fall 2022 Clinical Data Partner Activity

Due: Meeting 6.1 (Monday, 9/26)

Deliverables:

1. Submit your answers to the written activity by adding a pdf document to your GitHub (and sharing a google doc with Nicole + TAs).
2. Submit your coding responses by creating an R script (not a Notebook!) called `partner_activity_yourname.R` and adding it to your GitHub.
    a. Because this is a script, you will need to charge your working directory via `setwd("/PATH/TO/ANALYSIS/DATA")`
3. Though this is a partner activity, each student needs to push both their written answers and their code to GitHub. It is okay if your answers are the same as you are working together!

A Few Steps Before Starting

1. Ensure that your data is in your `analysis_data` folder.
2. Read in `brca_clinical_data.csv` file from your local machine. This is the file that you should have saved at the end of the clinical tutorial. Saving and reading in this file allows you to skip the `GDCquery` step every time you want to read in your data.

```
clinical <- read.csv("PATH/TO/brca_clinical_data.csv")
```

3. We will also be working with drug and radiation data. The data were already queried and downloaded, just prepare them with the following two commands:

```
clinical.drug <- GDCprepare_clinic(query = clinical_query,
                                   clinical.info = "drug")

clinical.rad <- GDCprepare_clinic(query = clinical_query,
                                  clinical.info = "radiation")
```

Written Activity

1. Define the following: *categorical variable, discrete variable, continuous variable.* Provide examples of each.

   A categorical variable can only take a limited range of possible values based on some qualitative characteristic. An example is sex.

   A discrete variable is a type of quantitative variable that takes on finite numerical values that cannot be subdivided. An example is number of children.

   A continuous variable is another type of quantitative variable that can take on an infinite number of dividable values. An example is height.

2. Look at the different column names of the `clinical` dataframe. Choose one that is interesting to you and your partner. Ensure that there are not too many NAs in this column by using `is.na(clinical$COLUMN_NAME)`. Remember that in coding, TRUE is equal to 1 and FALSE is equal to 0. You can then use the `sum()` function to find how many TRUEs exist. Which variable have you chosen?

   Histological type, which has 0 NA values.

3. Google your chosen variable. How is your variable measured or collected? Is your variable categorical, discrete, or continuous?

   Histological type is a categorical variable that is most commonly categorized by examination of the primary tumor, which can be done through tissue biopsy, endoscopy, and various other procedures.

4. Find two research articles that mention your clinical variable. Provide the links and a brief description of the findings.

   https://www.sciencedirect.com/science/article/pii/S1092913420300319

   This first paper finds that histological subtype in the highly aggressive triple negative breast cancer (ER-, PR-, HER2-) is highly associated with the histological subtype of the primary tumor. For instance, the invasive lobular carcinoma subtype was found to have the worst recurrence outcomes, with a recurrence rate of about

40%.

https://www.frontiersin.org/articles/10.3389/fonc.2021.635237/full

This second paper discusses the validity of anti-programmed death ligand 1 (PD-L1) therapy in treating metaplastic breast carcinoma, which traditionally is associated with poor prognosis and limited treatment options. Here, they find that in a cohort study of 5 patients with metaplastic breast carcinoma, 3 responded favorably to anti-PD-L1 treatment, suggesting a possible avenue of research moving forward.

5. Look at the different column names of the `clinical.drug`, `clinical.rad`, and `clinical` dataframes. Choose a variable from one of these data frames. Ensure there are not too many NAs (there will likely be more NAs in the drug and radiation dfs than in the patient data, don't worry about it too much). Which variable have you chosen? Provide a brief description of the variable.

   From the clinical dataframe, I chose the lymph_nodes_examined_count variable, which has 139 NA occurrences.

   This variable encapsulates the number of lymph nodes examined within a patient. Lymph node status is correlated with cancer progression, so examining lymph nodes, typically through tissue biopsy, can potentially guide clinical decision-making.

6. Scientists generate hypotheses before experimenting or exploring data. Generate three hypotheses: (1) Relate your variables to each other, (2) Relate your first variable to survival in breast cancer, (3) Relate your second variable to survival in breast cancer.

   (1) With the knowledge that some histological subtypes present as more aggressive/with worse outcomes than others, it is likely that histological subtype and lymph nodes examined is linked. More aggressively presenting subtypes will probably be examined more thoroughly.
   (2) More aggressive subtypes will present with poorer survival outcomes.
   (3) Number of lymph nodes examined will have little to no association with survival. An examination during a diagnostic phase should not have a significant impact on survival.

7. Summarize what you learned from your graphs! What is the significance of these findings? (Answer this question after you finish your analyses)

   From my boxplot, I see that the distribution of the number of lymph nodes examined is not statistically significant for each histological subtype, as all error bars overlap. This disproves hypothesis (1), instead suggesting that the number of patient lymph nodes examined does not depend on the histological subtype. This makes sense, because the subtype might not be known at the time of diagnosis.

   From the histological subtype KM plot, I can conclude that histological subtype does influence the survival of breast cancer patients, although maybe not to a statistically significant extent (p=0.29). This validates hypothesis (2). From the graph, we see that those of subtype infiltrating lobular carcinoma seem to have the best survival outcomes, while those of subtype mucinous carcinoma seem to have the worst outcomes.

   Results cannot be accurately extrapolated from the lymph nodes examined KM plot because of the high p-value of 0.46. The high p-value demonstrates that the number of lymph nodes examined likely has no association with survival outcomes. This validates hypothesis (3).

Coding Activity:

1. Perform an analysis looking at the two variables that you chose. First brainstorm and sketch out a plot that contains both variables. Feel free to get creative, if you are struggling, feel free to ask for ideas! (Helpful functions/packages: plot(), hist(), boxplot(), pairs(), ggplot2 package + associated functions)
   - TIP: Sometimes it can be hard to plot a continuous variable with another variable. You can convert a continuous variable to a categorical one. For example, we previously defined age < 50yrs old as "Young" and age ≥ 50 yrs old as "Old." Here we have converted age, a continuous variable, to young and old, a categorical variable.

2. Perform a survival analysis, following the steps of the clinical data tutorial with the first variable.
   - As with the previous tip, the survival analysis needs a categorical variable. If you have a continuous variable, use an `ifelse()` statement to create a new column

with a categorical version of the variable.

3. Repeat with the second variable. Note that for drug and radiation data, there might be many categories in one column. Try to keep the KM plot simple by limiting the data to ~5 stratification categories.

4. For an extra challenge (optional) perform a survival analysis where survival is stratified by *both* variables.

5. Save your plots and write any data frames you used to your local machine.

Check before submitting:

You **must** include informative comments throughout your code.

```
str(clinical) # view structure of clinical data frame
head(clinical) # view first few rows of clinical data frame
```

You **must** install and load all necessary packages at the top of your coding fall.

```
if (!require(package)){
install.packages("package")
}

library(package)
```

You **must** change your working directory at the top of your coding file.

```
setwd("/Users/nicoleblack/Desktop/QBIO/qbio_nicole/analysis_data")
```

You **must** be able to run your script from top to bottom (with a clean environment) without any issues.

- Before turning it in, hit the broom in the top right corner of Environment to clear all values and data. Then run the entire script by hitting the run button in the top right of your source panel. Your code should run all the way through with no errors.