**Homework Question:**

Our goal is to compare 58 patients who have bladder mucosae surrounding their cancer (pre-cancerous samples) vs. 10 patients with normal bladder mucosae (normal control samples)

# Analysis Plan of Action

## **Step 0**

### In clinical data file

Open clinical data file in excel to get familiar with the data. Run the following checks:

* Rows are patients
* Columns are clinical attributes
* There is one row per patient
* Check/Notice the Number of rows and number of columns
* All data must be de-identified

In Excel, all of the above checks were performed. The number of rows

(patients) was 233, and the number of columns (features) was 21.

### In Gene Expression Data File

Open gene expression data file in excel (if possible) to get familiar with the data. Run the following checks:

* Columns are patients
* Rows are features (genes in this case)
* There is one column per patient
* Check/Notice the Number of rows and number of columns
* All data must be de-identified
* It was given to us that gene expression data was normalized in log2 scale and includes gene annotation

In Excel, all of the above checks were performed. The number of columns (patients) is 233, and the number of rows (genes) is 43148. The gene expression data did include gene annotation.

## **Step 1 - Read in data into R. Check if data was imported correctly**

### **Read clinical data into R**

* General Checks
  + Rows are patients
  + Columns are clinical attributes
  + There is one row per patient
  + How many patients are there (dimensions of the data frame)
* Find the column which is your outcome of interest
* What are the groups to compare?

### **Read gene expression data into R**

* Read in cleaned processed molecular data into R
  + Check the dimensions of the data frame
  + Rows are molecular features/variables (genes in our example)
  + Columns are genes
  + Rows are patients

## **Step 2 - Clean/Filter data**

**Filter clinical data**

* Make sure one row per patient
* Filter out values from `PrimaryBladderCancerType` that we’re not comparing

**Don’t need to Filter gene exp data**

## **Step 3 - Identify the groups to be compared**

Identify the groups to be compared (Baseline and Comparison Grps)

* “Normal bladder mucosae” is the baseline group
* “Bladder mucosae surrounding cancer” is the comparison group

Subset groups into

* clinBase
* clinComp
* geneBase
* geneComp

What is the identifier in each sample to subset by:

* Clinical data sample identifiers: The GSMid column contains the Patient ID’s.
* Gene expression data sample identifiers: The column headers represent the Patient ID’s.

## **Step 4: Sanity check**

* See if filtering of clinical data in R matches filtering of clinical data in excel
* See if sample ids in clinical data match sample ids in gene exp data (if they don’t match it means your step 1 and/or 2 is wrong)
* Verify seeing correct number of samples in baseline and comp groups
* Export the column names from gene expression data and checked to see if it contained only probe/gene names and no other garbage

## **Step 5: Prep data for T-test**

* Make sure Gene expression data has features (genes in this case) as rows, and samples as columns.
* Make sure objects must be data frame
* Make sure it is numeric data only

## **Step 6- Call function for T-test**

* Use t-test function from “fnTTest” file to run the t-test