NetID: au198
Aizhan Uteubayeva

Steps 1
Final Project: Differential Gene Expression (t-test)
HIDS-7003-01

Goal: Team 1 goal is to compare Pre-Cancer (normal looking bladder mucosae surrounding cancer) vs Normal patients in order to identify differentially expressed genes

Workflow:

1) Reading in the data
    a) Clinical data - 233 patients; confirmed de-identification
    b) Gene expression data in log2 scale: includes gene annotation - 43,148 rows and 233 columns
2) Clean/filter the data
    a) Clinical data contains - "PrimaryBladderCancerType" column has values "Normal bladder mucosae" (baseline group) and "Bladder muscosae surrounding cancer" (comparison group)
    b) Gene expression data - genes are in rows, patients are in columns;
        i) NB: the names are going to cleaned from “|” in Step 2
3) Identify the groups to be compared
    a) Baseline - "Normal bladder mucosae" (baseline group)
    b) Comparison group - "Bladder muscosae surrounding cancer"
4) Sanity check
    a) (Yes) See if filtering of clinical data in R matches filtering of clinical data in excel
    b) (Yes) 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)
    c) (Yes) Verify you see correct number of samples in baseline and comp groups
    d) (Yes) Export the column names from gene expression data to see if it contains only probe/gene names and no other garbage
5) Preparing the data for t-test
    a) Checking to make sure data is a numeric data frame
6) Calling the function
7) Creating two csv files with ordered p-value>0.001 and top 20 genes