**A Brief Guide for Constructing Survival Plots in R**

**Data Configuration**

The following guide will show the basic steps to format TCGA gene expression data mapped to TCGA clinical data so that survival analysis can be performed using the survival and survminer R packages.

For the R libraries to read the survival data correctly, there needs to be at least three columns:

Expression group:

* For every patient, categorize as either high expression or low expression. We bifurcate these groups based on the median expression data in the group we are analyzing.
* Example: In a group of 41 patients for Colon Adenocarcinoma from TCGA, we examined differential expression of the gene CCL2. In this group the median log2fold change is -0.477.
* In a new column named “group”, all patients that are below the median TPM value are given the label “low” while all patients that are above the median log2fold change are given the label “high”.
* The excel formula for creating the “group” column is:
  + “=IF(D2=MEDIAN($D$2:$D$1079),"median", IF(D2<MEDIAN($D$2:$D$1079),"low","high"))”
  + Set the ranges specified in the above formula to the first and last sample row

Outcome:

* For every patient, there is a category for “alive” or “dead”.
* Make a new column that has a “0” if the patient has “alive” listed and a “1” if the patient has “dead’ listed. We named this column “outcome\_log”.
* The excel formula for creating the “outcome” column is:
  + =IF(L2="Dead",1,0)
  + In the above formula L2 represents the vital\_status column

Survival time:

* For every patient, there needs to be a time listed for how many days they spent in the study. This number is listed in TCGA as “days\_to\_death” for patients listed as “dead” and “days\_to\_last\_follow\_up” for patients listed as “alive”.
* Make a column that includes these numbers for each patient. We labeled this column “survival”
* The excel formula for the ‘survival” column is:
  + =IF(M2=1,N2,O2)
  + In the above formula M2 is the outcome column, N2 is the survival days column, and O2 is the days to days\_to\_last\_follow\_up column

Finally, we need to remove entries that should not be included in the survival analysis.

First, remove all entries in the survival column that are not > 0, including entries that are 0, negative numbers, and non-number characters.

Second, remove the median value (if there is one) as indicated by the group column.

Here is an R markdown file for running the survival analysis and generating the plot:

Kaplan Meier Survival Analysis

Ned Cauley

1/20/2021

##Load libraries to make survival plots and comparisons.

library(survival)  
library (survminer)

## Loading required package: ggplot2

## Loading required package: ggpubr

##Indicate the TCGA study and gene to be used for survival analysis.

study = 'BRCA'  
  
gene = 'PYGM'

##Load the csv file with mapped gene expression data, formatted correctly for survival analysis.

data\_file <- read.csv(file = paste('TCGA-',study,'/',gene,'.csv',sep = ''), header = TRUE, stringsAsFactors = FALSE)

##Determine the same size for the current analysis.

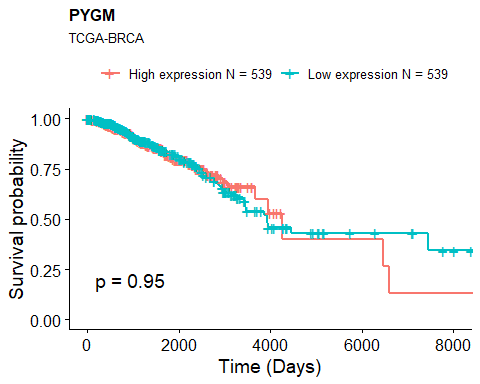
sample\_size = nrow(data\_file)/2

##Generate Kaplan Meier curves for the two expression groups.

surv\_object <- Surv(time = data\_file$survival, event = data\_file$outcome)  
  
fit1 <- survfit(surv\_object ~ group, data = data\_file)

##Plot the Kaplan Meier Analysis and perform a log-rank test to determine the p-value.

ggsurvplot(fit1, data = data\_file,  
 pval = TRUE,  
 title = gene,  
 font.title = c(12, "bold", "black"),  
 subtitle = paste('TCGA-',study,sep = ''),  
 font.subtitle = c(10, "plain", "black"),  
 legend.title =c(""),  
 legend.labs = c(paste('High expression N = ',sample\_size,sep = ''),paste('Low expression N = ',sample\_size,sep = '')),  
 xlab = "Time (Days)"  
 )



Further information on Kaplan Meier Curves  
  
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959/