Homework 6 - Spring 2021 STAT115/215 BIO/BST282

Jenna Landy

Data for this HW is stored at /n/stat115/2021/HW6/data

Part I: Data exploration on TCGA

The Cancer Genome Atlas (TCGA) is an NCI project to comprehensively profile over 10K tumors in 33 cancer types. In this homework, we are going to explore TCGA data analysis.

Q1. Go to TCGA GDC website (https://portal.gdc.cancer.gov/ (https://portal.gdc.cancer.gov/)) and explore the GDC data portal. How many glioblastoma (GBM) cases in TCGA meet ALL of the following requirements?

1. Male;
2. Diagnosed at the age above 45;
3. Still alive.

```
I went to the website and clicked "Exploration", then under the "Cases" tab, filtered cases by Program = TCGA and
Project = TCGA-GBM. Next, under the "Clinical" tab, filtered Gender = Male, Vital Status = Alive, and Age at Diag
nosis from 46 Years (above 45).
```

Answer link (https://portal.gdc.cancer.gov/exploration?
facetTab=clinical&filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22
GBM%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22genes.is_cancer_gene_census%22%2C%2

```
This resulted in 42 cases glioblastoma cases in TCGA that are male, still alive, and were diagnosed at the age ab
ove 45.
```

Q2. TCGA GDC (https://portal.gdc.cancer.gov/ (https://portal.gdc.cancer.gov/)) and Broad Firehose (http://firebrowse.org/ (http://firebrowse.org/)) both provide processed TCGA data for downloading and downstream analysis. Download clinical data of GBM. What's the average diagnosed age of all GBM patients?

```
From the previous search, I removed the specifications for gender, vital status, and age. Then I downloaded clini
cal data in TSV format.
```

```
clinical <- read.csv(
  'clinical.cases_selection.2021-04-18/clinical.tsv',
  sep = '\t'
)
clinical$age_at_diagnosis %>%
  as.numeric() %>%
  sapply(function(x) {floor(x / 365.25)}) %>%
  mean(na.rm = T)
```

```
## [1] 58.01531
```

```
clinical$age_at_index %>% as.numeric() %>% mean(na.rm = T)
```

```
## [1] 58.01531
```

Answer

```
The average age of diagnosis of all TCGA GBM patients on the GDC Data Portal is 58.02 years old.
```

Part II – Tumor Subtypes

Q1. GBM is one of the earliest cancer types to be processed by TCGA, and the expression profiling was initially done with Affymetrix microarray. Also, with brain cancer, it is hard to get sufficient number of normal samples. We provide the pre-processed expression matrix in (GBM_expr.txt) where samples are columns and genes are rows. Do a K-means (k=3) clustering from all the genes and the most variable 2000 genes. Do tumor and normal samples separate in different clusters? Do the tumors samples consistently separate into 2 clusters, regardless of whether you use all the genes or most variable genes?

```
Data: /n/stat115/2021/HW6/data/GBM_expr.txt
```

```
GBM_expr <- read.csv(
  'GBM_expr.txt', sep = '\t', row.names = 1
)
GBM_expr_scaled <- scale(GBM_expr)

tumor <- !startsWith(colnames(GBM_expr),'normal')
```

K-means (k=3) clustering from all the genes

```
set.seed(123)
k3_all <- kmeans(t(GBM_expr_scaled), centers = 3)
```

```
table(k3_all$cluster, tumor)
```

```
##    tumor
##     FALSE TRUE
##   1     0   30
##   2    10    0
##   3     0   30
```

K-means (k=3) clustering from the most variable 2000 genes

```
geneVars <- apply(GBM_expr_scaled, 1, function(x) {var(x)})
topVarGenes <- sort(geneVars, decreasing = T)[1:2000] %>%
  names()
GBM_expr_scaled_top2000 <- GBM_expr_scaled[
  rownames(GBM_expr_scaled) %in% topVarGenes,
]

set.seed(123)
k3_top_2000 <- kmeans(t(GBM_expr_scaled_top2000), centers = 3)
```

```
table(k3_top_2000$cluster, tumor)
```

```
##    tumor
##     FALSE TRUE
##   1     0   30
##   2    10    0
##   3     0   30
```

Do tumor and normal samples separate in different clusters? Do the tumors samples consistently separate into 2 clusters, regardless of whether you use all the genes or most variable genes?

Answer

```
Regardless of whether I used all the genes or the most variable, genes, all normal samples are in cluster 2, and
the tumor samples separate into 2 clusters (clusters 1 and 3).
```

Q2. LIMMA is a BioConductor package that does differential expression between microarrays, RNA-seq, and can remove batch effects (especially if you have experimental design with cmplex batches). Use LIMMA to see how many genes are differentially expressed between the two GBM subtypes (with FDR < 0.05% and logFC > 1.5)?

Note 0.05% cutoff = 0.0005.

```
cluster <- unname(k3_top_2000$cluster)
GBM_expr_scaled_g13 <- GBM_expr_scaled[,cluster != 2]

sample_names <- colnames(GBM_expr_scaled_g13)
group1 <- as.logical(cluster[cluster!=2] == 1)

mm <- model.matrix(~0 + group1)

# fit LIMMA
fit <- lmFit(GBM_expr_scaled_g13, mm)

contr <- makeContrasts(
  group1FALSE - group1TRUE,
  levels = colnames(coef(fit))
)
tmp <- contrasts.fit(fit, contr)

# Summary statistics
tmp <- eBayes(tmp)

# Summary statistics table
# adj.p.value cutoff at 0.05% = 0.0005
top.table_allexprs <- topTable(
  tmp, sort.by = "logFC", n = Inf
)
top.table <- topTable(
  tmp, sort.by = "logFC", n = Inf,
  p.value = 0.0005, lfc = 1.5 #cutoffs!
)
nrow(top.table)
```

```
## [1] 24
```

Answer:

```
24 genes are differentially expressed between the two GBM subtypes.
```

Q3. For Graduate Students: From the DNA methylation profiles (GBM_meth.txt), how many genes are significantly differentially methylated between the two subtypes? Are DNA methylation associated with higher or lower expression of these genes? How many differentially expressed genes have an epigenetic (DNA methylation) cause?

How many genes are significantly differentially methylated between the two subtypes?

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

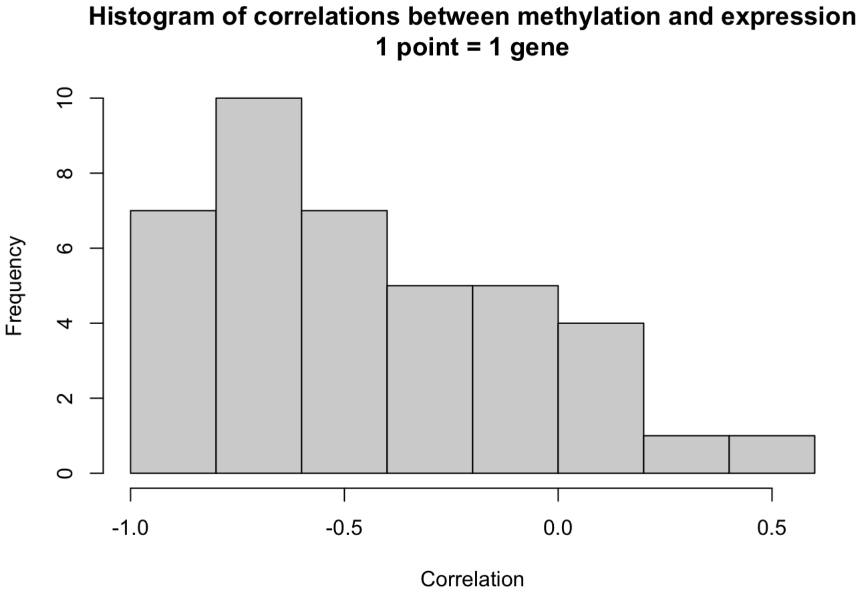
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
GBM_meth <- read.csv(
  'GBM_meth.txt',
  sep = '\t',
  row.names = 1
)
GBM_meth_logit <- logit(GBM_meth)
GBM_meth_scaled <- scale(GBM_meth_logit)

meth_clusters <- k3_top_2000$cluster[colnames(GBM_meth)] %>%
  unname()
group1 <- as.logical(meth_clusters[meth_clusters!=2] == 1)

mm_meth <- model.matrix(~0 + group1)

GBM_meth_scaled_g13 <- GBM_meth_scaled[,meth_clusters != 2]

# fit LIMMA
fit_meth <- lmFit(GBM_meth_scaled_g13, mm_meth)

contr_meth <- makeContrasts(
  group1FALSE - group1TRUE,
  levels = colnames(coef(fit_meth))
)
tmp_meth <- contrasts.fit(fit_meth, contr_meth)

# Summary statistics
tmp_meth <- eBayes(tmp_meth)

# Summary statistics table
# adj.p.value cutoff at 0.05% = 0.0005
top.table_allmeth <- topTable(
  tmp_meth, sort.by = "logFC", n = Inf
)
top.table_meth <- topTable(
  tmp_meth, sort.by = "logFC", n = Inf,
  p.value = 0.0005, lfc = 1.5 #cutoffs!
)
nrow(top.table_meth)
```

```
## [1] 63
```

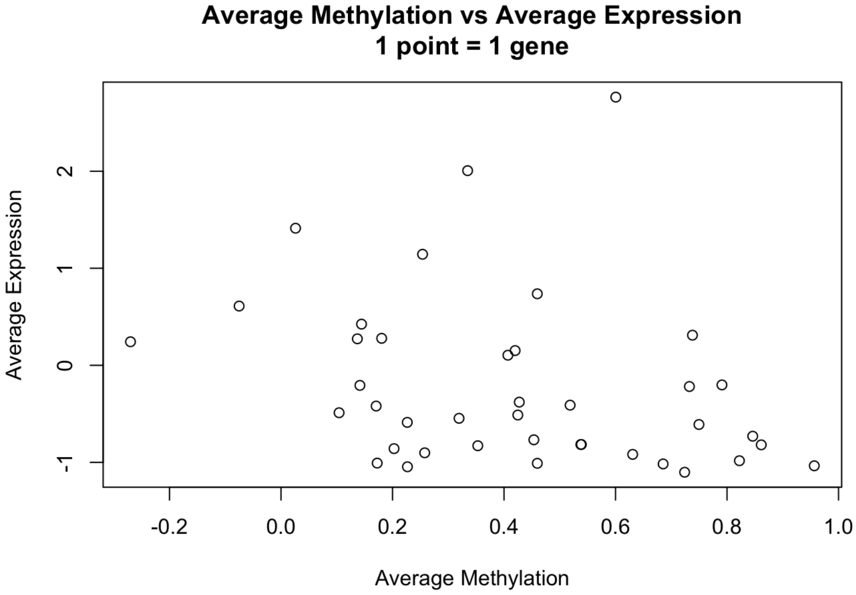
Are DNA methylation associated with higher or lower expression of these genes?

```
common_rows <- intersect(rownames(GBM_expr_scaled_g13), rownames(GBM_meth_scaled))
common_samples <- intersect(colnames(GBM_expr_scaled_g13), colnames(GBM_meth_scaled))

subset_expr <- GBM_expr_scaled_g13[
  rownames(GBM_expr_scaled_g13) %in% rownames(top.table_meth) & rownames(GBM_expr_scaled_g13) %in% common_rows,
  colnames(GBM_expr_scaled_g13) %in% common_samples
]

subset_meth <- GBM_meth_scaled[
  rownames(GBM_meth_scaled) %in% rownames(top.table_meth) & rownames(GBM_meth_scaled) %in% common_rows,
  colnames(GBM_meth_scaled) %in% common_samples
]

# make sure = 0, same order
# sum(colnames(subset_meth) != colnames(subset_expr))
# sum(rownames(subset_meth) != rownames(subset_expr))

corrs <- c()
for (row in rownames(subset_meth)) {
  corrs <- c(
    corrs,
    cor(subset_expr[row,], subset_meth[row,])
  )
}

hist(corrs, main = "Histogram of correlations between methylation and expression\n1 point = 1 gene",
     xlab = 'Correlation')
```


Histogram of correlations between methylation and expression
1 point = 1 gene

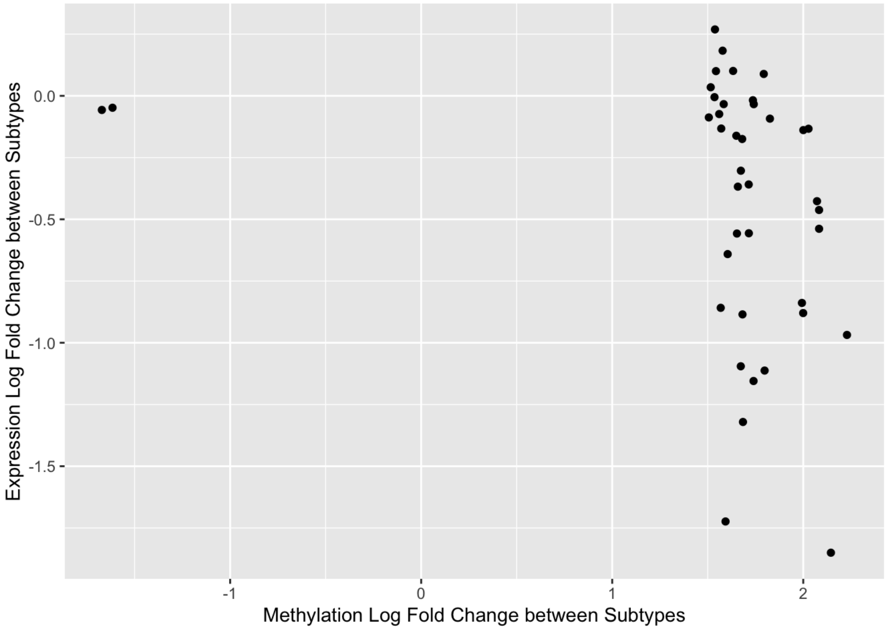
```
The majority of correlation values are below 0, so higher DNA methylation seems to be associated with lower expre
ssion of these genes.
```

```
avg_expr <- GBM_expr_scaled_g13[
  rownames(GBM_expr_scaled_g13) %in% rownames(top.table_meth) &
  rownames(GBM_expr_scaled_g13) %in% common_rows,
] %>% rowMeans()

avg_meth <- GBM_meth_scaled[
  rownames(GBM_meth_scaled) %in% rownames(top.table_meth) &
  rownames(GBM_meth_scaled) %in% common_rows,
] %>% rowMeans()

# make sure = 0, same order
# sum(names(avg_meth) != names(avg_expr))

plot(avg_meth, avg_expr, main = 'Average Methylation vs Average Expression\n1 point = 1 gene',
     xlab = 'Average Methylation', ylab = 'Average Expression')
```

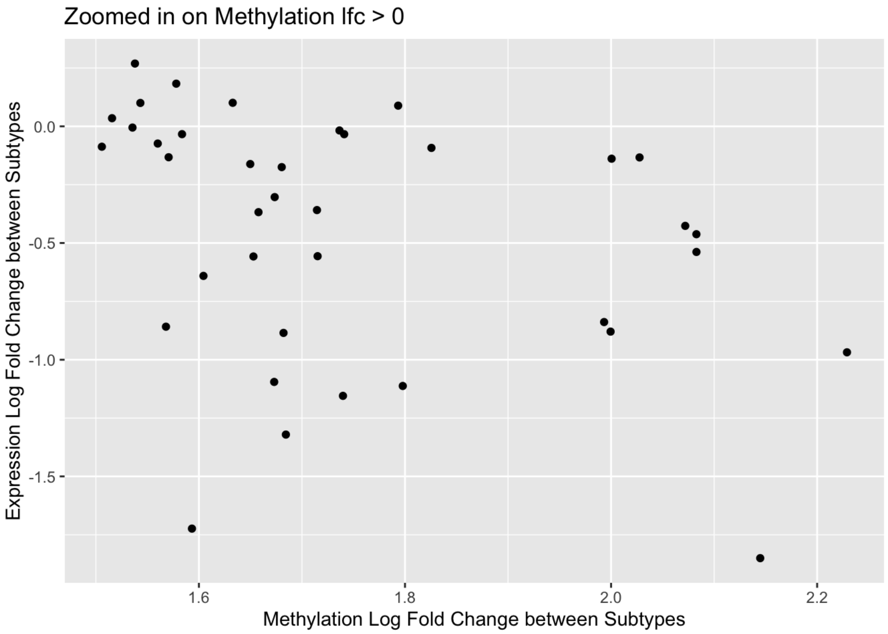

Average Methylation vs Average Expression
1 point = 1 gene

```
There is a slight negative association between average methylation and expression, so higher DNA methylation seem
s to be associated with lower expression of these genes.
```

```
top_common <- intersect(rownames(top.table_meth), rownames(top.table_allexprs))

data.frame(
  meth_lfc = top.table_meth[top_common, 'logFC'],
  expr_lfc = top.table_allexprs[top_common, 'logFC']
) %>%
  ggplot(aes(x = meth_lfc, y = expr_lfc)) +
  geom_point() +
  xlab('Methylation Log Fold Change between Subtypes') +
  ylab('Expression Log Fold Change between Subtypes')
```



```
data.frame(
  meth_lfc = top.table_meth[top_common, 'logFC'],
  expr_lfc = top.table_allexprs[top_common, 'logFC']
) %>%
  dplyr::filter(meth_lfc > 0) %>%
  ggplot(aes(x = meth_lfc, y = expr_lfc)) +
  geom_point() +
  xlab('Methylation Log Fold Change between Subtypes') +
  ylab('Expression Log Fold Change between Subtypes') +
  ggtitle('Zoomed in on Methylation lfc > 0')
```



```
There is a slight negative association between log fold change for methylation and log fold change for expressio
n, so higher DNA methylation seems to be associated with lower expression of these genes.
```

How many differentially expressed genes have an epigenetic (DNA methylation) cause?

```
sum(!is.na(top.table_allmeth[rownames(top.table), 'logFC']))
```

```
## [1] 16
```

```
data.frame(
  expr_lfc = top.table[, 'logFC'],
  meth_lfc = top.table_allmeth[rownames(top.table), 'logFC']
) %>%
  dplyr::filter(
    expr_lfc < 0 & meth_lfc > 0 |
    expr_lfc > 0 & meth_lfc < 0
  ) %>%
  nrow()
```

```
## [1] 12
```

```
top_both <- intersect(rownames(top.table), rownames(top.table_meth))
top_both
```

```
## [1] "LGALS3" "PDPN"
```

```
data.frame(
  name = top_both,
  expr_lfc = top.table[top_both, 'logFC'],
  meth_lfc = top.table_allmeth[top_both, 'logFC']
)
```

```
##     name  expr_lfc meth_lfc
## 1 LGALS3 -1.850144 2.144759
## 2   PDPN -1.723604 1.593239
```

Answer

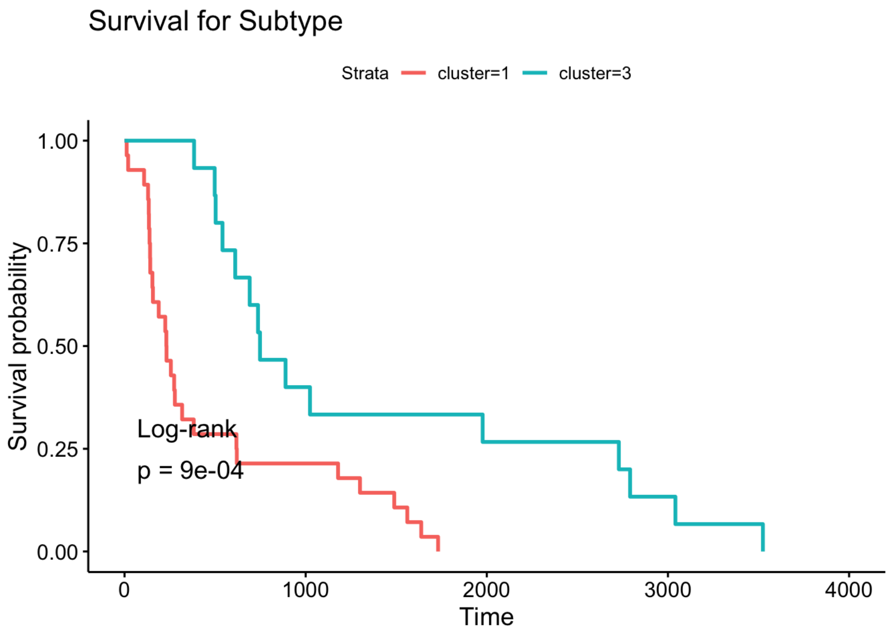
```
63 genes are significantly differentially methylated between the two subtypes.

Based on the three plots and descriptions above, higher DNA methylation seems to be associated with lower express
ion of these genes.

Among all 16 differentially expressed genes which have methylation data, 12 have a log fold change in methylation
in the opposite direction of their log fold change in expression, so may have an epigenetic cause. If we look onl
y at genes that are differentially expressed and significantly differentially methylated, there are only two such
genes and both have opposite direction log fold changes--so both have a significant epigenetic cause. These two g
enes are LGALS3 and PDPN.
```

Q4. With the survival data of the GBM tumors (GBM_clin.txt), make a Kaplan-Meier Curve to compare the two subtypes of GBM patients. Is there a significant difference in patient outcome between the two subtypes?

```
GBM_clin <- read.csv('GBM_clin.txt', sep = '\t', row.names = 1)

gsub('\\.', '-', names(k3_top_2000$cluster))[1:60] == rownames(GBM_clin) # same order
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
GBM_clin$cluster <- k3_top_2000$cluster[1:60]

km_fit <- survfit(Surv(days.to.death, vital.status) ~ cluster, data = GBM_clin)
ggsurvplot(km_fit, pval = TRUE, pval.method = TRUE) +
    ggtitle("Survival for Subtype")
```



Answer

```
There is a significant (p-value = 0.0009 < 0.01 significance level) difference in patient outcome between the two
subtypes.
```

Q5. For Graduate Students: Use the differential genes (say this is Y number of genes) between the two GBM subtypes as a gene signature to do a Cox regression of the tumor samples. Does it give significant predictive power of patient outcome?

```
diff_genes <- rownames(top.table)
diff_genes_expr <- data.frame(t(data.frame(GBM_expr[diff_genes,])))
rownames(diff_genes_expr) <- gsub('\\.', '-', rownames(diff_genes_expr))

diff_genes_expr$sample <- rownames(diff_genes_expr)
GBM_clin$sample <- rownames(GBM_clin)

GBM_clin_cox <- GBM_clin %>%
  merge(diff_genes_expr, by = 'sample', all.x = TRUE) %>%
  mutate(surv = Surv(days.to.death, vital.status))

diff_genes[diff_genes =="RP11-35N6.1"] = "RP11.35N6.1"
colnames(GBM_clin_cox)[colnames(GBM_clin_cox) =="RP11-35N6.1"] = "RP11.35N6.1"

form <- as.formula(paste(
  "surv ~ ",
  paste(diff_genes, collapse = ' + '),
  sep = ''
))
cox_mod <- coxph(form, data = GBM_clin_cox)
summary(cox_mod)
```

```
## Call:
## coxph(formula = form, data = GBM_clin_cox)
## 
##   n= 43, number of events= 43 
##    (17 observations deleted due to missingness)
## 
##                 coef exp(coef) se(coef)      z Pr(>|z|)    
## POSTN        0.63904   1.89466  0.24393  2.620 0.008798 ** 
## CHI3L1      -0.74330   0.47554  0.30301 -2.453 0.014166 *  
## NNMT        -1.09068   0.33599  0.34036 -3.204 0.001353 ** 
## CXCL14      -0.08943   0.91445  0.27918 -0.320 0.748710    
## MOXD1        1.23616   3.44237  0.37817  3.269 0.001080 ** 
## FABP5       -0.70555   0.49384  0.33669 -2.096 0.036122 *  
## LGALS3       0.12213   1.12990  0.36411  0.335 0.737316    
## DLL3        -2.59233   0.07485  0.61830 -4.193 2.76e-05 ***
## DCX          0.38278   1.46636  0.31746  1.206 0.227903    
## EMP3         0.07151   1.07413  0.37697  0.190 0.849542    
## PTX3         0.27880   1.32155  0.36201  0.770 0.441213    
## RBP1        -1.24722   0.28730  0.44964 -2.774 0.005540 ** 
## SCN3A       -1.87484   0.15338  0.48070 -3.900 9.61e-05 ***
## PDPN         2.31137  10.08828  0.59734  3.869 0.000109 ***
## TIMP1       -0.56142   0.57040  0.77248 -0.727 0.467362    
## TOX3        -0.93354   0.39316  0.56435 -1.654 0.098090 .  
## RP11.35N6.1 -0.28505   0.75198  0.32780 -0.870 0.384531    
## LTF         -0.48623   0.61494  0.19566 -2.485 0.012952 *  
## C20orf42     0.74417   2.10469  0.35083  2.121 0.033908 *  
## PLA2G5       0.32711   1.38695  0.60382  0.542 0.588001    
## ADM          1.05459   2.87078  0.38359  2.749 0.005973 ** 
## IGFBP2      -3.06758   0.04653  0.71087 -4.315 1.59e-05 ***
## MAOB         1.83668   6.27566  0.51389  3.574 0.000351 ***
## KCND2        3.89929  49.36747  0.74425  5.239 1.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##             exp(coef) exp(-coef) lower .95 upper .95
## POSTN         1.89466    0.52780   1.17463    3.0561
## CHI3L1        0.47554    2.10285   0.26258    0.8612
## NNMT          0.33599    2.97630   0.17243    0.6547
## CXCL14        0.91445    1.09355   0.52908    1.5805
## MOXD1         3.44237    0.29050   1.64043    7.2237
## FABP5         0.49384    2.02495   0.25527    0.9554
## LGALS3        1.12990    0.88504   0.55349    2.3066
## DLL3          0.07485   13.36086   0.02228    0.2515
## DCX           1.46636    0.68196   0.78708    2.7319
## EMP3          1.07413    0.93099   0.51308    2.2487
## PTX3          1.32155    0.75669   0.65003    2.6868
## RBP1          0.28730    3.48065   0.11902    0.6935
## SCN3A         0.15338    6.51975   0.05979    0.3935
## PDPN         10.08828    0.09912   3.12870   32.5290
## TIMP1         0.57040    1.75316   0.12550    2.5925
## TOX3          0.39316    2.54349   0.13008    1.1883
## RP11.35N6.1   0.75198    1.32983   0.39553    1.4296
## LTF           0.61494    1.62618   0.41907    0.9024
## C20orf42      2.10469    0.47513   1.05818    4.1862
## PLA2G5        1.38695    0.72100   0.42471    4.5293
## ADM           2.87078    0.34834   1.35359    6.0885
## IGFBP2        0.04653   21.48975   0.01155    0.1874
## MAOB          6.27566    0.15935   2.29212   17.1823
## KCND2        49.36747    0.02026  11.47977  212.2993
## 
## Concordance= 0.87  (se = 0.037 )
## Likelihood ratio test= 76.86  on 24 df,   p=2e-07
## Wald test            = 37.1  on 24 df,   p=0.04
## Score (logrank) test = 64.3  on 24 df,   p=2e-05
```

```
library(glmnet)
set.seed(321)

GBM_clin_cox_nona <- na.omit(GBM_clin_cox) #Omit NAs

x <- as.matrix(GBM_clin_cox_nona[,diff_genes]) #Subset the covariates and assign it to x
survobj <- Surv(GBM_clin_cox_nona$days.to.death, GBM_clin_cox_nona$vital.status) #Create the Survival object, whi
ch is the outcome we are interested in

cvfit <- cv.glmnet(x, survobj, family = "cox", alpha = 1) #Cross-validation

cvfit
```

```
## 
## Call:  cv.glmnet(x = x, y = survobj, family = "cox", alpha = 1) 
## 
## Measure: Partial Likelihood Deviance 
## 
##     Lambda Index Measure      SE Nonzero
## min 0.2400    10   7.290 0.13640       3
## 1se 0.5052     2   7.425 0.01417       2
```

```
coef(cvfit, s = "lambda.min") #The set of coefficients using lambda.min, choose those that are non-zero to run co
x regression
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## POSTN        .         
## CHI3L1       .         
## NNMT         .         
## CXCL14       0.06353838
## MOXD1        .         
## FABP5        .         
## LGALS3       .         
## DLL3        -0.02925364
## DCX          .         
## EMP3         .         
## PTX3         .         
## RBP1         .         
## SCN3A        .         
## PDPN         0.09074997
## TIMP1        .         
## TOX3         .         
## RP11.35N6.1  .         
## LTF          .         
## C20orf42     .         
## PLA2G5       .         
## ADM          .         
## IGFBP2       .         
## MAOB         .         
## KCND2        .
```

```
cox_mod2 <- coxph(surv~CXCL14 + DLL3 + PDPN, data = GBM_clin_cox)
summary(cox_mod2)
```

```
## Call:
## coxph(formula = surv ~ CXCL14 + DLL3 + PDPN, data = GBM_clin_cox)
## 
##   n= 43, number of events= 43 
##    (17 observations deleted due to missingness)
## 
##            coef exp(coef) se(coef)      z Pr(>|z|)
## CXCL14  0.09579   1.10052  0.10081  0.950    0.342
## DLL3   -0.10901   0.89672  0.13626 -0.800    0.424
## PDPN    0.15623   1.16910  0.11409  1.369    0.171
## 
##        exp(coef) exp(-coef) lower .95 upper .95
## CXCL14    1.1005     0.9087    0.9032     1.341
## DLL3      0.8967     1.1152    0.6866     1.171
## PDPN      1.1691     0.8554    0.9348     1.462
## 
## Concordance= 0.742  (se = 0.039 )
## Likelihood ratio test= 18.86  on 3 df,   p=3e-04
## Wald test            = 16.75  on 3 df,   p=8e-04
## Score (logrank) test = 18.87  on 3 df,   p=3e-04
```

```
extractAIC(cox_mod2)
```

```
## [1]   3.0000 230.2051
```

```
extractAIC(cox_mod)
```

```
## [1]  24.0000 214.2073
```

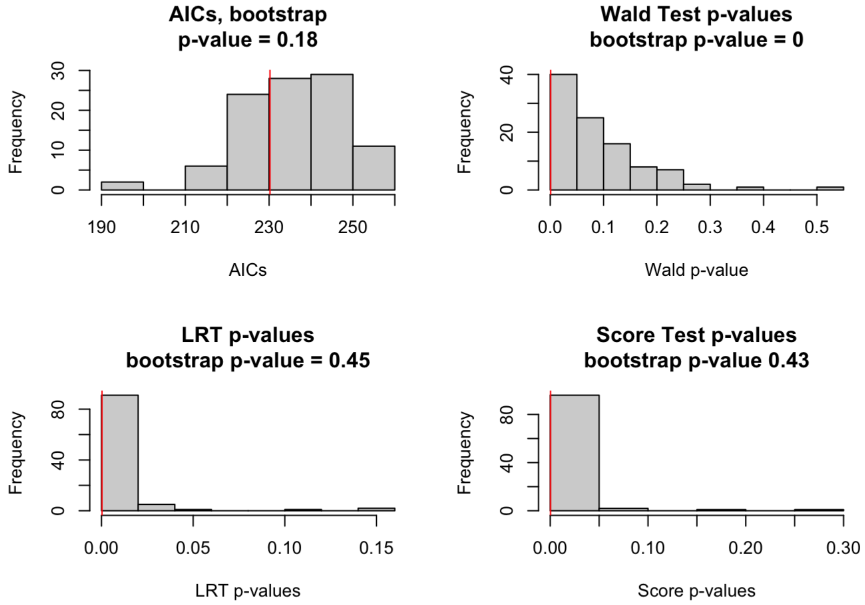
Answer:

```
The likelihood ratio test and score test result in very significant p-values (<0.0001) and the wald test results
in a moderately significant p-value (<0.05), so yes, the basic cox proportional hazard model gives significant pr
edictive power of patient outcome.

When LASSO is applied, only three genes have non-zero coefficients: CXCL14, DLL3, and PDPN. These are the most im
portant genes in predicting patient outcome. The cox proportional hazard model with only these three genes as cov
ariates results in very significant p-values for all three tests (p-value < 0.001). The likelihood test and score
test have slightly worse p-values than before, but now all three tests agree. However, the AIC of the lasso model
is slightly worse than the AIC of the original model.
```

Q6. For Graduate Students: Many studies use gene signatures to predict prognosis of patients. Take a look at this paper: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002240 (http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002240). It turns out that most published gene signatures are not significantly more associated with outcome than random predictors. Write a script to randomly sample Y genes in this expression data as a gene signature and do Cox regression on the sampled signature to predict patient outcome. Automate the script and random sample followed by Cox regression 100 times. How does your signature in Q5 compared to random signatures in predicting outcome?

```r
Y <- length(diff_genes)
rownames(GBM_expr_scaled) <- gsub('@','.',rownames(GBM_expr_scaled))
rownames(GBM_expr_scaled) <- gsub('-','.',rownames(GBM_expr_scaled))

all_genes <- rownames(GBM_expr_scaled)

rownames(GBM_clin) <- gsub('-','\\.',rownames(GBM_clin))
GBM_clin <- GBM_clin %>%
  mutate(
    surv = Surv(days.to.death, vital.status),
    sample = rownames(GBM_clin)
  )

all_genes_expr <- data.frame(t(data.frame(GBM_expr_scaled)))
all_genes_expr$sample <- rownames(all_genes_expr)

GBM_clin_cox <- GBM_clin %>%
  merge(all_genes_expr, by = 'sample', all.x = TRUE)


AICs <- c()
wald_ps <- c()
log_ps <- c()
score_ps <- c()

for (i in 1:100) {
  genes <- sample(all_genes, size = Y, replace = F)

  genes_expr <- GBM_clin_cox[,c(genes, 'surv')]

  cox_mod <- coxph(surv ~ ., data = genes_expr)

  summ <- summary(cox_mod)
  AICs <- c(AICs, extractAIC(cox_mod)[2])
  wald_ps <- c(wald_ps, summ$waldtest['pvalue'] %>% unname())
  log_ps <- c(log_ps, summ$logtest['pvalue'] %>% unname())
  score_ps <- c(score_ps, summ$sctest['pvalue'] %>% unname())

  GBM_expr_scaled
}
```

```r
par(mfrow = c(2,2))

bootstrap_p <- mean(AICs < extractAIC(cox_mod2))
hist(AICs, main = paste('AICs, bootstrap\np-value =', bootstrap_p))
abline(v = extractAIC(cox_mod2), col = 'red', xlab = 'AIC')

bootstrap_p <- mean(wald_ps < summary(cox_mod2)$waldtest['pvalue'])
hist(wald_ps, main = paste('Wald Test p-values\nbootstrap p-value =', bootstrap_p), xlab = 'Wald p-value')
abline(v = summary(cox_mod2)$waldtest['pvalue'],
       col = 'red')

bootstrap_p <- mean(log_ps < summary(cox_mod2)$logtest['pvalue'])
hist(log_ps, main = paste('LRT p-values\nbootstrap p-value =', bootstrap_p), xlab = 'LRT p-values')
abline(v = summary(cox_mod2)$logtest['pvalue'],
       col = 'red')

bootstrap_p <- mean(score_ps < summary(cox_mod2)$sctest['pvalue'])
hist(score_ps, main = paste('Score Test p-values\nbootstrap p-value',bootstrap_p), xlab = 'Score p-values')
abline(v = summary(cox_mod2)$sctest['pvalue'],
       col = 'red')
```



Answer:

```
The signature in Q5 (post-lasso metrics marked in red on the plots above) doesn't perform better than random sign
atures in predicting outcome by any metric besides the Wald p-test. I conclude that it doesn't perform any better
than random signatures.
```

Part III – Tumor mutation analyses and precision medicine

Q1. The MAF files contain the mutations of each tumor compared to the normal DNA in the patient blood. Write a script to parse out the mutations present in each tumor sample and write out a table. The table should rank the genes by how many times mutation happens in the tumor samples provided. Submit the table with the top 20 genes.

```
mutations <- c()
files <- list.files('MAF')
#length(files)

first = TRUE
for (file in files) {
  this_dat <- read.csv(paste('MAF/', file, sep = ''), sep = '\t')
  this_dat$file <- file
  if (first) {
    all_dat <- this_dat
    first = FALSE
  } else {
    all_dat <- rbind(all_dat, this_dat)
  }
}

top20 <- all_dat %>%
  group_by(Hugo_Symbol) %>%
  tally() %>%
  arrange(-n) %>%
  filter(Hugo_Symbol != 'Unknown') %>%
  head(20)

top20
```

```
## # A tibble: 20 x 2
##    Hugo_Symbol     n
##    <chr>       <int>
##  1 TP53           16
##  2 IDH1           12
##  3 ATRX            9
##  4 NF1             7
##  5 EGFR            6
##  6 TTN             6
##  7 NBPF10          5
##  8 APOB            4
##  9 MUC17           4
## 10 PIK3CA          4
## 11 PIK3R1          4
## 12 PTEN            4
## 13 ABCB4           3
## 14 AHNAK2          3
## 15 CACNA1S         3
## 16 CCDC129         3
## 17 DLG5            3
## 18 DMBT1           3
## 19 GFRAL           3
## 20 HMCN1           3
```

Answer:

```
The most commonly mutated gene is TP53 with 16 mutations across the 23 tumor samples. The next most commonly muta
ted gene is IDH1, which is mutated 12 times across the 23 samples. The table above shows the gene and number of m
utations for the top 20 most frequently mutated genes. Because there were many mutations with 3 occurences (the s
mallest # in top 20), ties were broken by alphabetical order of the gene symbol.
```

Q2. Existing clinical genetic testing laboratories use information about the frequency of a mutation in cohorts, like from the GBM cohort in TCGA, to assess a mutation's clinical significance (guidelines: https://www.ncbi.nlm.nih.gov/pubmed/27993330 (https://www.ncbi.nlm.nih.gov/pubmed/27993330)). Of the top 20 mutated genes in Q1, what is the most frequent protein change (hint: count mutations with the exact same amino acid change as the same)? In what gene does the most frequent protein change occur? Do you think this gene and protein mutation pair forms a genetic subtype of GBM (i.e. Is this gene-protein mutation pair seen more in one GBM subtype than the other)?

```
# Of the top 20 mutated genes in Q1, what is the most frequent protein change
top20genes <- as.character(top20$Hugo_Symbol)

all_dat %>%
  filter(
    Hugo_Symbol %in% top20genes,
    Protein_Change != ''
  ) %>%
  group_by(Protein_Change) %>%
  tally() %>%
  arrange(-n) %>%
  head()
```

```
## # A tibble: 6 x 2
##   Protein_Change     n
##   <chr>          <int>
## 1 p.R132H           12
## 2 p.E545K            2
## 3 p.R175H            2
## 4 p.R248Q            2
## 5 p.Y220C            2
## 6 p.461_462RE>Q      1
```

```
# In what gene does the most frequent protein change occur
all_dat %>%
  filter(
    Hugo_Symbol %in% top20genes,
    Protein_Change != ''
  ) %>%
  group_by(Hugo_Symbol, Protein_Change) %>%
  tally() %>%
  arrange(-n) %>%
  head()
```

```
## # A tibble: 6 x 3
## # Groups:   Hugo_Symbol [4]
##   Hugo_Symbol Protein_Change     n
##   <chr>       <chr>          <int>
## 1 IDH1        p.R132H           12
## 2 PIK3CA      p.E545K            2
## 3 TP53        p.R175H            2
## 4 TP53        p.R248Q            2
## 5 TP53        p.Y220C            2
## 6 ABCB4       p.A6T              1
```

```
# Is this gene-protein mutation pair seen more in one GBM subtype than the other
idh1.pr132h_dat <- all_dat %>%
  group_by(file) %>%
  mutate(
    idh1.pr132h = Hugo_Symbol == 'IDH1' & Protein_Change == 'p.R132H'
  ) %>%
  dplyr::summarize(idh1.pr132h = max(idh1.pr132h)) %>%
  mutate(
    sample = file %>%
      str_replace(".maf.txt", "") %>%
      str_replace_all('-', '.'),
    cluster = k3_top_2000$cluster[sample]
  )

idh1.pr132h_dat %>%
  group_by(cluster, idh1.pr132h) %>%
  tally()
```

```
## # A tibble: 3 x 3
## # Groups:   cluster [2]
##   cluster idh1.pr132h     n
##     <int>       <int> <int>
## 1       1           0    10
## 2       3           0     1
## 3       3           1    12
```

```
idh1.pr132h_dat %>%
  filter(cluster == 3, idh1.pr132h == 0)
```

```
## # A tibble: 1 x 4
##   file                 idh1.pr132h sample       cluster
##   <chr>                      <int> <chr>          <int>
## 1 TCGA-26-5133.maf.txt           0 TCGA.26.5133       3
```

Answer:

```
The most common protein change is p.R132H, which occurs 12 times among the top 20 mutated genes in the 23 sample
s. Every instance of this protein change is in a mutation of the IDH1 gene. Each sample where this occurs (IDH1 g
ene with p.R132H protein change) is in cluster 3, so this gene and protein mutation pair is related to a genetic
subtype of GBM. However, there is one sample in cluster three where the gene and protein mutation pair does not o
ccur: sample TCGA.26.5133.
```

Q3. CBioPortal has a comprehensive list of tumor profiling results for interactive visualization. Go to cBioPortal (http://www.cbioportal.org (http://www.cbioportal.org)), and select either “Glioblastoma” under “CNS/Brian” (left) or select “TCGA PanCancer Atlas Studies” under “Quick Select” (middle). Input each gene in Q1 and click Submit. From the OncoPrint tab, you can see how often each gene is mutated in GBM or all TCGA cancer types. Based on this, which of the genes in Q1 is likely to be a cancer driver gene?

I selected “Glioblastoma” under “CNS/Brian”, and there were 6 studies selected. I then clicked “Query by Gene” and pasted the list “TP53 IDH1 ATRX NF1 EGFR TTN NBPF10 APOB MUC17 PIK3CA PIK3R1 PTEN ABCB4 AHNAK2 CACNA1S CCDC129 DLG5 DMBT1 GFRAL HMCN1” into the query box. Gene CCDC129 was marked as invalid and was recommended to change to ITPRID1. According to the GeneCards website (https://www.genecards.org/cgi-bin/carddisp.pl?gene=ITPRID1), these are both aliases for the same gene, so I accepted the recommended change, and clicked “Submit Query”.

Answer:

```
From the OncoPrint tab, I can see how often each gene is mutated in GBM cancers. The most often mutated is EGFR,
with an altered/profiled ratio of 0.46. Next is PTEN, at 0.26. Then is TP53, which was the most frequently mutate
d in our data, with a ratio of 0.17. In comparison, gene IDH1, which was the focus of question 2, is only altered
in 3/100 profilings (0.03).

Based on this, EGFR is most likely to be a driver gene of glioblastomas, as it is the most frequently mutated acr
oss the 6 studies selected.
```

Q4. From the Mutation tab on the cBioPortal result page, is this mutation a gain or loss of function mutation on the gene you identified from Q2?

Answer:

```
In question 2, I identified the gene IDH1 as related to a genetic subtype of GBM. The Mutations tab on the cBioPo
rtal shows that this is most often a missense mutation. The mutation is almost always at the same position, R132
H/G/C, so is a gain of function mutation.
```

Q5. From cBioPortal, select Glioblastoma (TCGA provisional, which has the largest number of samples) and enter the driver mutation gene in Q2. From the Survival tab, do GBM patients with this mutation have better outcome in terms of progression free survival and overall survival?

I selected all Glioblastoma studies that included TCGA in the title, clicked “Query by Gene”, and entered “IDH1” in the query box. Under the “Comparison/Survival tab, I selected the”Survival" tab.

Answer:

```
The Kaplan-Meier Curve shows that overall survival is, on average, longer within the altered group (or the indivi
duals with the IDH1 mutation). The logrank test p-value of 2.11e-7 = 0.000000211 < 0.01 is significant at any rea
sonable p-value. This reflects the Kaplan-Meier Curve in Part II question 4 for survival by cluster 1 vs cluster
3.

The trend is maintained when looking at progression-free survival, with p-value of 1.191e-3 = 0.001191 < 0.01.

Thus, GBM patients with the IDH1 mutation have better outcome in terms of progression free survival and overall s
urvival
```

Q6. You are working with an oncologist collaborator to decide the treatment option for a GBM patient. From exome-seq of the tumor, you identified the top mutation in Q2. To find out whether there are drugs that can target this mutation to treat the cancer, go to https://www.clinicaltrials.gov (https://www.clinicaltrials.gov) to find clinical trials that target the gene in Q2. How many trials are related to glioblastoma? How many of these are actively recruiting patients which this patient could potentially join? Hint: Search by the disease name and gene name.

In the advanced search, I put Glioblastoma as “Condition or disease” and IDH1 under “other terms”, since there was not a gene name option.

Answer:

```
My search returned 18 studies related to glioblastoma and the gene IDH1. 16 of these are interventional, so would
potentially have a drug involved (as opposed to simple observation).

9 of the total studies are actively recruiting, 8 of which are interventional.
```

Part IV- CRISPR screens

We will learn to analyze CRISPR screen data from this paper: https://www.ncbi.nlm.nih.gov/pubmed/?term=26673326 (https://www.ncbi.nlm.nih.gov/pubmed/?term=26673326). To identify therapeutic targets for glioblastoma (GBM), the author performed genome-wide CRISPR-Cas9 knockout (KO) screens in patient-derived GBM stem-like cell line (GSCs0131).

MAGeCK tutorial: https://sourceforge.net/p/mageck/wiki/Home/ (https://sourceforge.net/p/mageck/wiki/Home/)
https://sourceforge.net/projects/mageck/ (https://sourceforge.net/projects/mageck/)

Q1. Use MAGeCK to do a basic QC of the CRISPR screen data (e.g. read mapping, ribosomal gene selection, replicate consistency, etc). Comment on the quality of the data based on your results.

Directory: /n/stat115/2021/HW6/data/crispr_data

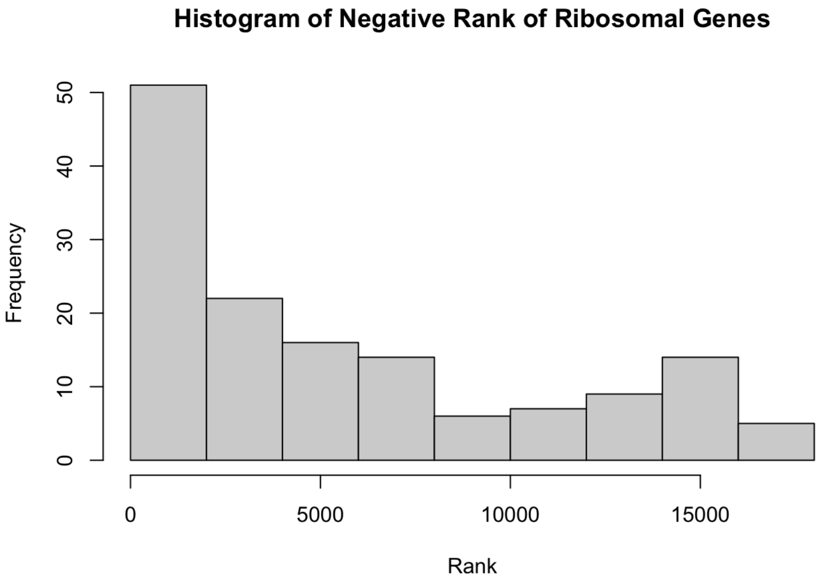
```
module load Anaconda
source activate /n/stat115/2021/HW6_env2
```

```
mageck count -l /n/stat115/2021/HW6/data/crispr_data/library.csv -n OUT --sample-label Day0,Day23 --fastq /n/stat
115/2021/HW6/data/crispr_data/GSC_0131_Day0_Rep1.fastq.gz,/n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day0_Rep
2.fastq.gz /n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day23_Rep1.fastq.gz,/n/stat115/2021/HW6/data/crispr_dat
a/GSC_0131_Day23_Rep2.fastq.gz

mageck test -k OUT.count.txt -t Day23 -c Day0 -n OUT

# ----

mageck count -l /n/stat115/2021/HW6/data/crispr_data/library.csv -n sep_OUT --sample-label Day0_Rep1,Day0_Rep2,Da
y23_Rep1,Day23_Rep2 --fastq /n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day0_Rep1.fastq.gz /n/stat115/2021/HW6/
data/crispr_data/GSC_0131_Day0_Rep2.fastq.gz /n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day23_Rep1.fastq.gz /
n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day23_Rep2.fastq.gz

mageck test -k sep_OUT.count.txt -t Day23 -c Day0 -n sep_OUT
```

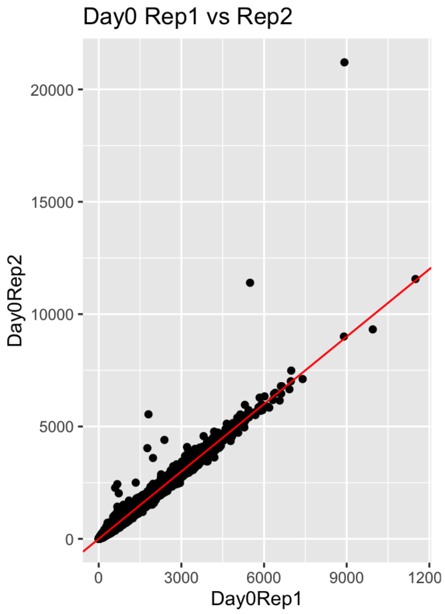
```
countsummary <- read.csv('OUT.countsummary.txt', sep = '\t')
countsummary
```

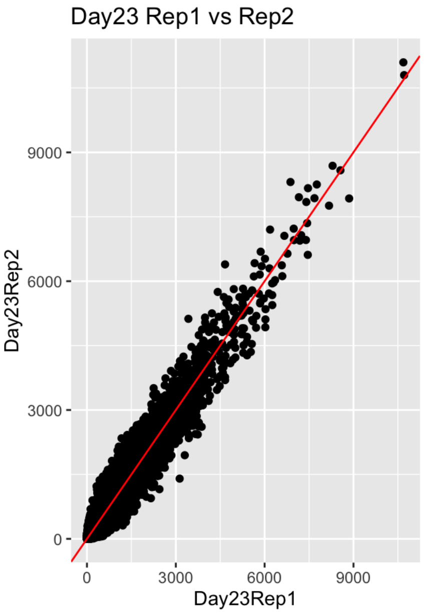
```
##                                                                 File Label
## 1 /n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day23_Rep1.fastq.gz Day23
## 2 /n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day23_Rep2.fastq.gz Day23
## 3  /n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day0_Rep2.fastq.gz  Day0
## 4  /n/stat115/2021/HW6/data/crispr_data/GSC_0131_Day0_Rep1.fastq.gz  Day0
##      Reads   Mapped Percentage TotalsgRNAs Zerocounts GiniIndex NegSelQC
## 1 62818064 39992777     0.6366       64076         57   0.08510        0
## 2 58686580 37836392     0.6447       64076         51   0.08587        0
## 3 47289074 31709075     0.6705       64076         17   0.07496        0
## 4 51190401 34729858     0.6784       64076         14   0.07335        0
##   NegSelQCPval NegSelQCPvalPermutation NegSelQCPvalPermutationFDR NegSelQCGene
## 1            1                       1                          1            0
## 2            1                       1                          1            0
## 3            1                       1                          1            0
## 4            1                       1                          1            0
```

```
genesummary <- read.csv('OUT.gene_summary.txt', sep = '\t')
genesummary %>%
  filter(grepl('^RP', id)) %>%
  pull(neg.rank) %>%
  hist(main = 'Histogram of Negative Rank of Ribosomal Genes',
       xlab = 'Rank')
```



```
count_normalized <- read.csv('OUTQC.count_normalized.txt', sep = '\t')

p1 <- ggplot(count_normalized, aes(x = Day0Rep1, y = Day0Rep2)) +
  geom_point() +
  ggtitle('Day0 Rep1 vs Rep2') +
  geom_abline(slope = 1, intercept = 0, color = 'red')

p2 <- ggplot(count_normalized, aes(x = Day23Rep1, y = Day23Rep2)) +
  geom_point() +
  ggtitle('Day23 Rep1 vs Rep2') +
  geom_abline(slope = 1, intercept = 0, color = 'red')

library(patchwork)
p1 + p2
```





Answer:

```
The mapped read percentages are all over 0.6, the zero counts are all under 0.1, and all gini indices are less th
an 0.1, so the data is good!

The rank of ribosomal genes tend to be ranked lower than non-ribosomal genes, which is good! This is clear from t
he histogram of ribosomal gene ranks, which shows the density of the ranks near 0, and fewer observations with hi
gh ranks.

In the plots of replicates against eachother show that for both days, replicates are fairly consistent.
```

Q2. Analyze CRISPR screen data with MAGeCK to identify positive and negative selection genes. How many genes are selected as positive or negative selection genes, respectively, and what are their respective enriched pathways?

```
neg_genes <- genesummary %>%
  filter(neg.fdr < 0.05)

length(neg_genes %>%
  pull(id))
```

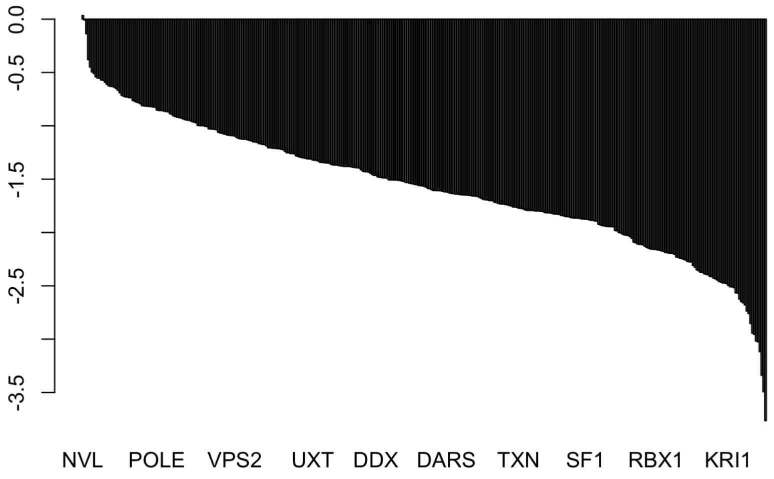
```
## [1] 369
```

```
# cat(neg_genes%>% pull(id), sep = ', ')

pos_genes <- genesummary %>%
  filter(pos.fdr < 0.05)

length(pos_genes%>%
  pull(id))
```

```
## [1] 8
```

```
# cat(pos_genes%>% pull(id), sep = ', ')
```

```
library(fgsea)

neg_ranks <- neg_genes$neg.lfc
names(neg_ranks) <- neg_genes$id
neg_ranks <- sort(neg_ranks, decreasing = T)
barplot(neg_ranks)
```



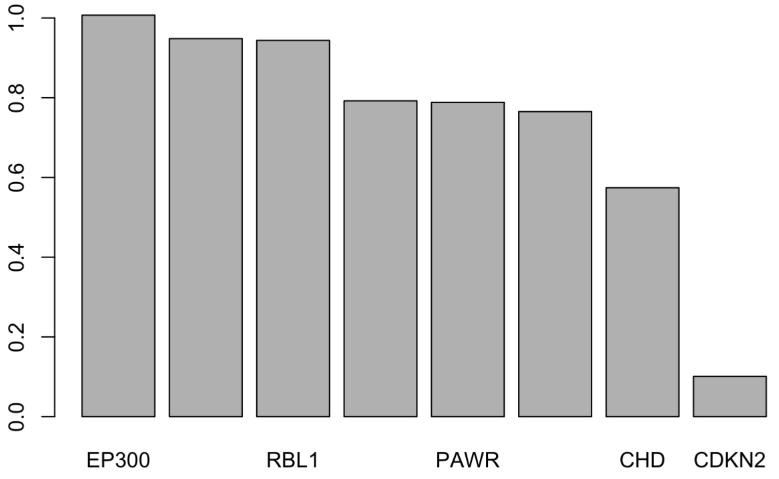
```
pathways <- gmtPathways('../HW2/c2.cp.kegg.v7.1.symbols.gmt')
fgseaRes <- fgsea(pathways, neg_ranks, minSize=10, maxSize = 500)
```

```
## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize, gseaParam, : There are ties in the prera
nked stats (1.08% of the list).
## The order of those tied genes will be arbitrary, which may produce unexpected results.
```

```
fgseaRes
```

```
##                            pathway       pval      padj    log2err         ES
## 1:                 KEGG_CELL_CYCLE 0.84931507 0.8493151 0.12154328  0.1657303
## 2: KEGG_OXIDATIVE_PHOSPHORYLATION 0.05728667 0.2099057 0.32177592  0.3983287
## 3:                   KEGG_RIBOSOME 0.80542453 0.8493151 0.03113351 -0.2318436
## 4:                KEGG_SPLICEOSOME 0.10495283 0.2099057 0.14641624 -0.4368551
##           NES size                                  leadingEdge
## 1:  0.7423026   13       ANAPC1,MCM7,MCM6,CDC25B,BUB3,RAD21,...
## 2:  1.5207484   10 PPA2,COX6B1,NDUFB9,NDUFA2,NDUFS2,NDUFB10,...
## 3: -0.7285148   11        RPS11,RPS5,RPS18,RPS8,RPL14,RPS20,...
## 4: -1.3727160   11              XAB2,PRPF38B,LSM2,EFTUD2,HNRNPU
```

```
pos_ranks <- pos_genes$pos.lfc
names(pos_ranks) <- pos_genes$id
pos_ranks <- sort(pos_ranks, decreasing = T)
barplot(pos_ranks)
```



```
fgseaRes <- fgsea(pathways, pos_ranks, minSize=3, maxSize = 500)
```

```
## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize,
## gseaParam, : All values in the stats vector are greater than zero and scoreType
## is "std", maybe you should switch to scoreType = "pos".
```

```
fgseaRes
```

```
##                             pathway       pval       padj   log2err ES      NES
## 1:                  KEGG_CELL_CYCLE 0.04562064 0.04562064 0.3217759  1 1.784247
## 2: KEGG_TGF_BETA_SIGNALING_PATHWAY 0.04562064 0.04562064 0.3217759  1 1.784247
##    size       leadingEdge
## 1:    3 EP300,CREBBP,RBL1
## 2:    3 EP300,CREBBP,RBL1
```

Answer:

```
There were 369 negative selection genes and 8 positive selection genes. Negative selection genes fall under the c
ell cycle, oxidative phosporylation, ribosome, and splicosome kegg pathways. The positive selection genes fall un
der the cell cycle and tgf beta signaling pathways.
```

Q3. For Graduate Students: Genes negatively selected in this CRISPR screen could be potential drug targets. However, if they are always negatively selected in many cells, targeting such genes might create too much toxicity to the normal cells. Go to depmap (DepMap.org) which has CRISPR / RNAi screens of over 500 human cell lines, Click “Tools” Data Explorer. Pick the top 3 negatively selected genes to explore. Select Gene Dependency from CRISPR (Avana) on the X axis and Omics from Expression on the Y axis, to see the relationship between the expression level of the gene and dependency (CRISPR screen selection) of the gene across ~500 cell lines. Are the top 3 genes good drug targets?

```
neg_genes %>% pull(id) %>% head(3)
```

```
## [1] "NF2"    "SRSF10" "EIF2B4"
```

Answer:

```
For NF2, expression is very weakly associated with gene dependency (pearson correlation around 0.11). For SRSF10,
expression is even less correlated with gene dependency (pearson 0.014). For EIF2B4, it is also a weak correlatio
n (pearson correlation 0.116).

Since we want uncorrelated expression and gene dependency, I would investigate SRSF10 first. However, since all t
hree genes show weak correlations, any would probably be fine.
```

Q4. For Graduate Students: Let’s filter out pan essential genes (PanEssential.txt) from the negatively selected genes in Q2. Take the remaining top 10 genes, and check whether those genes have drugs or are druggable from this website: http://www.oasis-genomics.org/ (http://www.oasis-genomics.org/). Go to Analysis -> Pan Cancer Report, enter the top 10 genes and check the table for druggability (more druggable for higher number on Dr). Which of these genes are druggable?

Note: If the above website cannot be opened, try https://www.dgidb.org/ (https://www.dgidb.org/). Go to Search Potential Druggability or Search Drug-Gene Interactions ann enter the top 10 genes.

```
PanEssential <- read.csv('PanEssential.txt', sep = '\t')

druggable <- neg_genes %>%
  filter(!(id %in% PanEssential$GeneSymbol)) %>%
  arrange(neg.fdr) %>%
  head(10)
druggable$id %>% cat(sep = ', ')
```

```
## NF2, SRSF10, ATP6V0, TCOF1, RPL, PPP1R8, PMVK, MRPL9, HMB, XRN1
```

Answer:

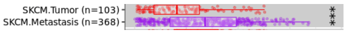
```
I pasted this list into the drug gene interaction database and found 3 drugable genes: NF2, TCOF1, and PMVK.
```

Part V. Cancer immunology and immunotherapy

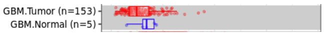
Immune checkpoint inhibitors, which primarily activate CD8 T cells, have shown remarkable efficacy in melanoma (SKCM), but haven't worked as well in GBM patients. Let's explore the tumor immune microenvironment from TCGA data. Although the cancer patients in TCGA were not treated with immunotherapy, their response to other drugs and clinical outcome might be influenced by pre-treatment tumor immune microenvironment

Q1. TIMER (http://timer.cistrome.org/ (http://timer.cistrome.org/)) estimated the infiltration level of different immune cells of TCGA tumors using different immune deconvolution methods. CD8A and CD8B are two gene markers on CD8 T cells. On the Diff Exp tab, compare the expression level of either CD8A or CD8B between GBM and SKCM (Metastatic Melanoma). Based on this, which cancer type have more CD8 T cells?

```
library(knitr)
knitr::include_graphics('pic1.png')
```



```
knitr::include_graphics('pic2.png')
```



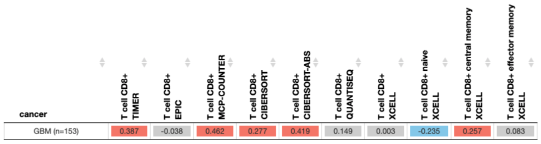
Answer:

```
I looked at the expression of CD8A. The boxplots for expression levels of GBM and SKCM are above. It seems that S
KCM tumors tend to have more CD8 T cells than GBM tumors, though this difference may not be signifiacnt.
```

Q2. On the Gene tab, select both GBM and SKCM (Metastatic Melanoma), include CD8 T cells as the cell infiltrate. Check the following genes, PDCD1(PD1), CD274(PDL1), CTLA4 which are the targets of immune checkpoint inhibitors, to see whether their expression level is associated with immune cell infiltration in the GBM and SKCM tumors. Their higher expression usually indicate that T cells are in a dysfunctional state, which immune checkpoint inhibitors aim to revive.

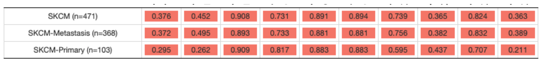
PDCD1: expression strongly associated with immune cell infiltration in SKCM, less often significant in GBM.

```
knitr::include_graphics('pic4.png')
```



cancer	T cell CD8+ TIMER	T cell CD8+ EPIC	T cell CD8+ MCP-COUNTER	T cell CD8+ CIBERSORT	T cell CD8+ CIBERSORT-ABS	T cell CD8+ QUANTISEQ	T cell CD8+ XCELL	T cell CD8+ naive XCELL	T cell CD8+ central memory XCELL	T cell CD8+ effector memory XCELL
GBM (n=153)	0.387	-0.038	0.462	0.277	0.419	0.149	0.003	-0.235	0.257	0.083

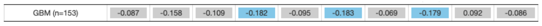
```
knitr::include_graphics('pic3.png')
```



SKCM (n=471)	0.376	0.452	0.908	0.731	0.891	0.894	0.739	0.365	0.824	0.363
SKCM-Metastasis (n=368)	0.372	0.495	0.893	0.733	0.881	0.881	0.756	0.382	0.832	0.389
SKCM-Primary (n=103)	0.295	0.262	0.909	0.817	0.883	0.883	0.595	0.437	0.707	0.211

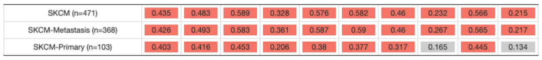
CD274: expression strongly positively associated with immune cell infiltration in SKCM, either not-associated or negatively associated in GBM.

```
knitr::include_graphics('pic6.png')
```



GBM (n=153)	-0.087	-0.158	-0.109	-0.182	-0.095	-0.183	-0.069	-0.179	0.092	-0.086

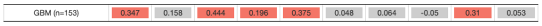
```
knitr::include_graphics('pic5.png')
```



SKCM (n=471)	0.435	0.483	0.589	0.328	0.576	0.582	0.46	0.232	0.566	0.215
SKCM-Metastasis (n=368)	0.426	0.493	0.583	0.361	0.587	0.59	0.46	0.267	0.565	0.217
SKCM-Primary (n=103)	0.403	0.416	0.453	0.206	0.38	0.377	0.317	0.165	0.445	0.134

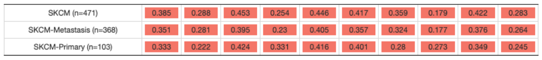
CTLA4: expression strongly positively associated with immune cell infiltration in SKCM, and less frequently associated in GBM.

```
knitr::include_graphics('pic7.png')
```



GBM (n=153)	0.347	0.158	0.444	0.196	0.375	0.048	0.064	-0.05	0.31	0.053

```
knitr::include_graphics('pic8.png')
```



SKCM (n=471)	0.385	0.288	0.453	0.254	0.446	0.417	0.359	0.179	0.422	0.283
SKCM-Metastasis (n=368)	0.351	0.281	0.395	0.23	0.405	0.357	0.324	0.177	0.376	0.264
SKCM-Primary (n=103)	0.333	0.222	0.424	0.331	0.416	0.401	0.28	0.273	0.349	0.245

Answer:

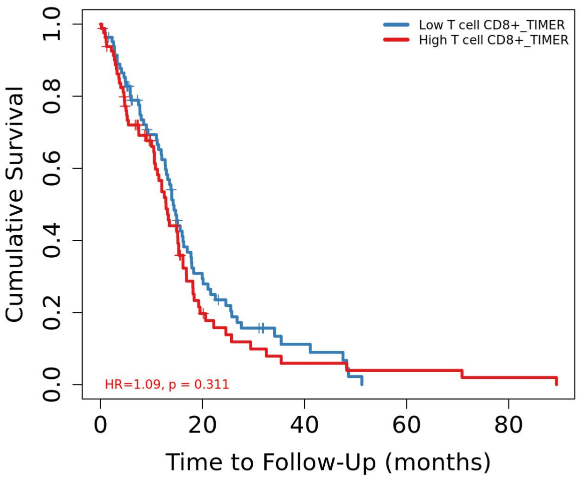
```
For all three genes, expression is strongly positively associated with immune cell infiltration in SKCM, and less
frequently associated, and sometimes negatively associated, in GBM. This helps to explain why it is harder to use
this type of treatment with GBMs.
```

Q3. On the Survival tab, select both GBM and SKCM, include CD8 T cell as the cell infiltrate, add tumor stage and patient age as the clinical variables to conduct survival analyses. Based on the Cox PH model, what factors are the most significantly associated with patient survival in each cancer type? Plot the Kaplan-Meier curve to evaluate how the immune cell infiltrate is associated with survival. Is CD8 T cell associated with patient survival in each cancer type?

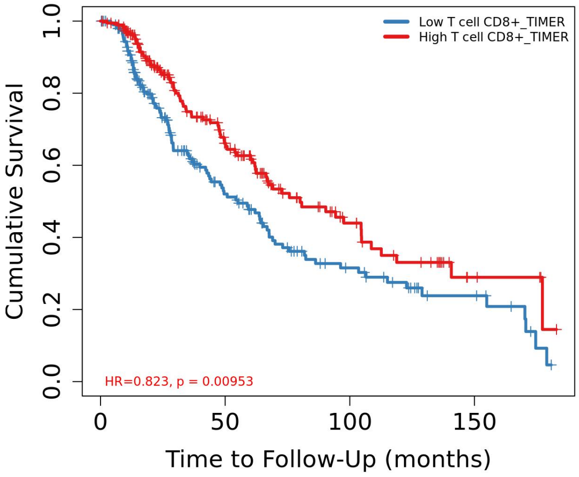
Model details:

```
T cell CD8+_TIMER in GBM (n=153):
Model: Surv(OS, EVENT) ~ `T cell CD8+_TIMER` + Age
150 patients with 120 dying ( 3 missing obs. )
                     coef     HR se(coef) 95%CI_l 95%CI_u     z     p signif
`T cell CD8+_TIMER` 2.637 13.975    1.004   1.955  99.894 2.628 0.009     **
Age                 0.028  1.028    0.008   1.013   1.044 3.606 0.000    ***
Rsquare = 0.132 (max possible = 9.98e-01 )
Likelihood ratio test p = 2.42e-05
Wald test p = 5.12e-05
Score (logrank) test p = 5.98e-05

T cell CD8+_TIMER in SKCM (n=471):
Model: Surv(OS, EVENT) ~ `T cell CD8+_TIMER` + Age + Stage
407 patients with 192 dying ( 64 missing obs. )
                      coef    HR se(coef) 95%CI_l 95%CI_u      z     p signif
`T cell CD8+_TIMER` -1.216 0.296    0.401   0.135   0.650 -3.032 0.002     **
Age                  0.019 1.019    0.005   1.009   1.029  3.636 0.000    ***
Stage2               0.270 1.310    0.215   0.859   1.998  1.254 0.210
Stage3               0.636 1.890    0.200   1.278   2.795  3.187 0.001     **
Stage4               1.234 3.436    0.349   1.735   6.802  3.541 0.000    ***
Rsquare = 0.105 (max possible = 9.92e-01 )
Likelihood ratio test p = 1.39e-08
Wald test p = 5.95e-08
Score (logrank) test p = 3.31e-08
```

```
# GBM KM Curve
knitr::include_graphics('outcome_plot.jpg')
```



```
# SKCM KM Curve
knitr::include_graphics('outcome_plot_skcm.jpg')
```



Answer:

```
The factor most associated with patient survival in both cancer types. Stage is also significantly associated wit
h survival in SKCM.

The Kaplan Meier curves show that increased CD8 T cells is significantly associated with better outcomes for SKCM
(p-value = 00953), but there is no significant difference in GBM (p-value = 0.311).
```

Q4. For Graduate Students: Based on the above observations, can you hypothesize why immune checkpoint inhibitors don't work well for some GBM patients?

Answer:

```
In (1), we found that SKCM tumors tend to have more CD8 T cells than GBM tumors. Further, in (2), we found that t
he expression of genes that are targets of immune checkpoint inhibitors were much more associated with immune cel
l infiltration in SKCM tumors than in GBM tumors. Finally, in (3) we found that increased CD8 T cells is associat
ed with improved survival outcomes only for SKCM tumors, and there was no significant difference for GBM tumors.

All of these things together show that CD8 T cells are not associated in GBM, so it's not a realistic target. Sin
ce T cells are not associated with survival outcomes, aren't associated with immune infiltration, a change in the
amount or function of T cells would not be expected to change the cancer.
```

Rules for submitting the homework:

Please submit your solution directly on the canvas website. Please provide both your code in this Rmd document and an html file for your final write-up. Please pay attention to the clarity and cleanness of your homework.

The teaching fellows will grade your homework and give the grades with feedback through canvas. Some of the questions might not have a unique or optimal solution. TFs will grade those according to your creativity and effort on exploration, especially in the graduate-level questions.