BIOS 855 Final Project

Olivia Rippee

2024-07-25

Introduction

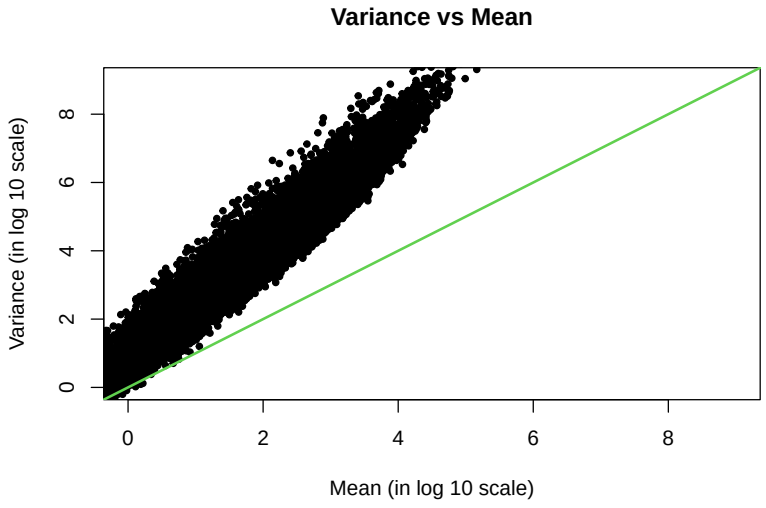
Head and neck squamous cell carcinoma (HNSCC) is the cancer of the mouth, nasal cavity, throat, and/or larynx. Common clinical symptoms include sinus congestion, sore throat, earaches, and swollen lymph nodes. More serious cases also have ulcers, bleeding, and tumors. HNSCC can metastasize to the lymph nodes and lungs. About 50% of patients survive at least 5 years after diagnosis. Understanding the molecular underpinnings of HNSCC may lead to targeted treatments and better prognosis for patients.

The data for this project comes from the The Cancer Genome Atlas Network and their manuscript entitled "Comprehensive genomic characterization of head and neck squamous cell carcinomas" (doi.org/10.1038/nature14129).

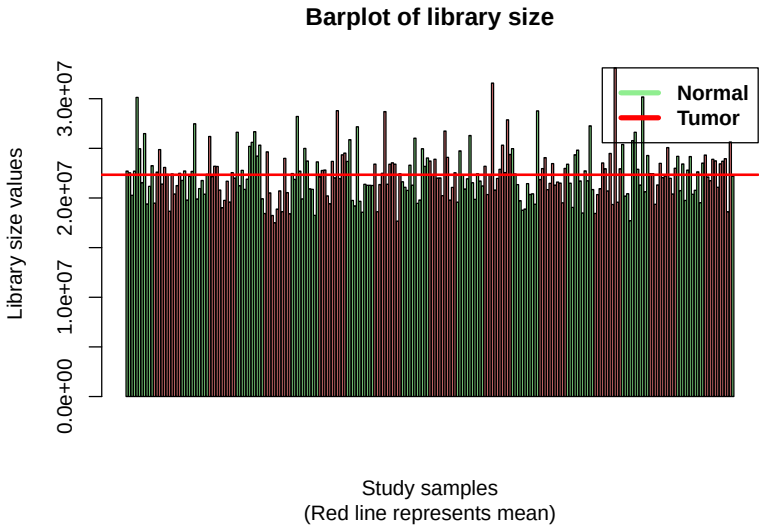
Our goal is to identify differentially expressed features between HNSCC patients with tumors and without tumors, and to determine the pathways affected by this differential expression.

Preliminary Checks

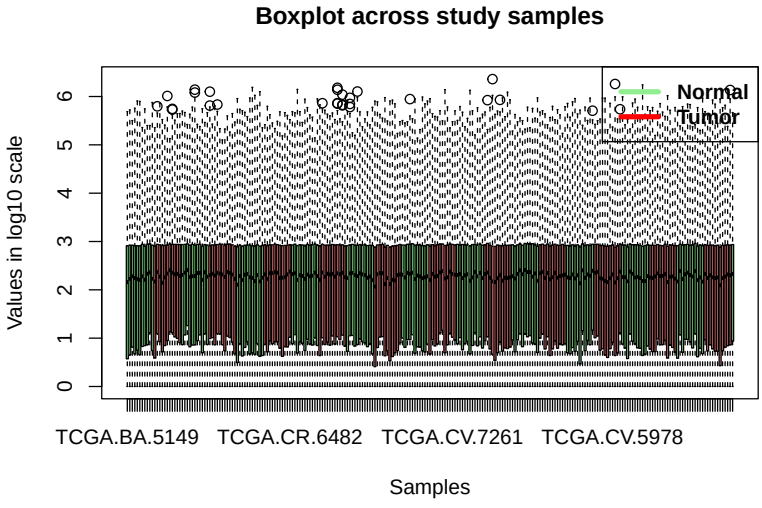
For count-based data such as RNAseq, we would fit either a Poisson or negative binomial model. High-throughput data tends to have higher variance of counts than the Poisson distribution can handle, so we graphed mean vs variance to narrow down a model. The graph shows that the variance of the data (given as black dots) is higher than the mean, and this is beyond what the Poisson distribution allows (represented by the green line). Therefore, we concluded that a negative binomial distribution should be used for these data.



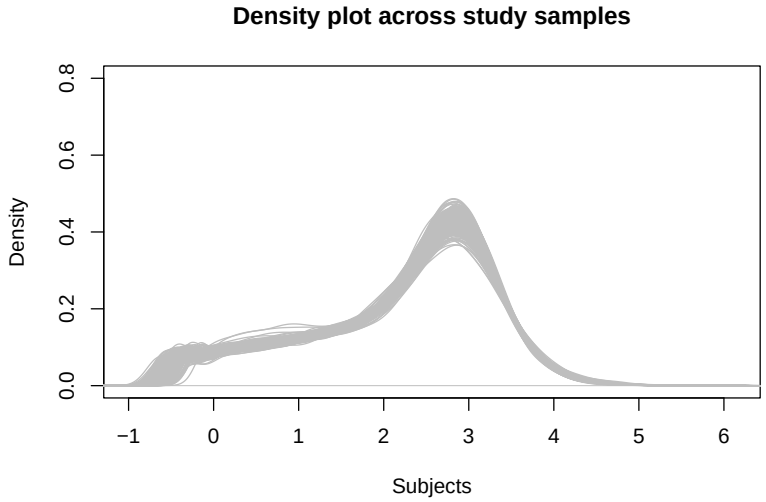
To determine if any genes should be removed from the data before analysis, we plotted the number of reads for each sample (library size) to visualize potential imbalances of coverage. Normalization can account for some imbalance, however if counts are too small then the samples have underlying problems and need to be removed. We did not note any issues in this case.



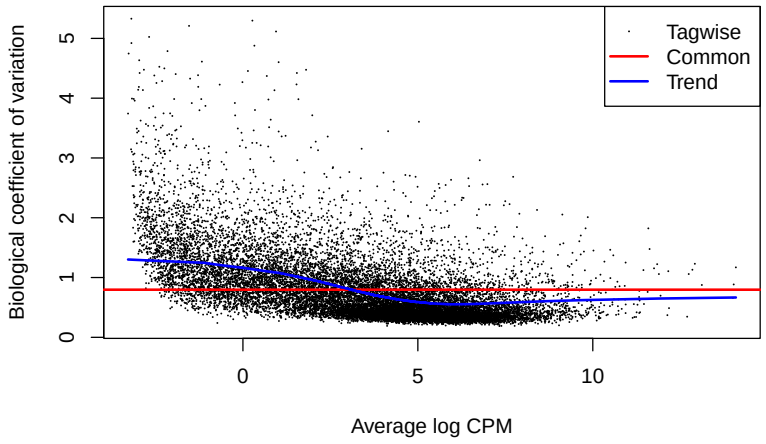
Next we viewed the count boxplots for each sample. Since count data is not normally distributed, we used the log10 transformation. Since any 0 values in the dataset would generate an error for log10, we added 1 to all values in the dataset. We can see that the centers are relatively constant and the distributions are not significantly different. As such, we did not remove any genes from the analysis.



To evaluate whether any subjects need to be removed before the analysis, we plotted the density for each subject. Any stark deviations from the overall pattern would indicate an outlier in need of removal. We noted that there are no outlying density distributions, and thus did not remove any subjects.



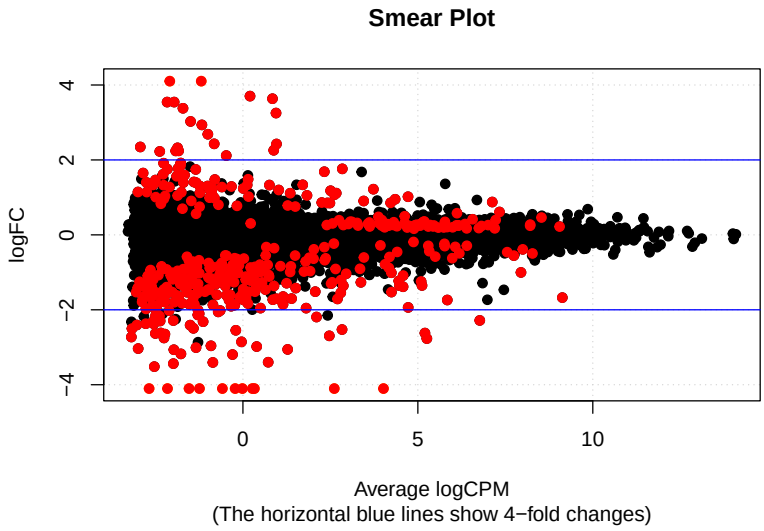
Then, we filtered and normalized the data. We filtered out genes such that each remaining gene has at least 2 samples with at least 1 count per million. We then calculated the dispersion of the data and plot the abundance (log counts per million) vs dispersion (biological coefficient of variation). The trendline is relatively close to zero, so we concluded that dispersion of the genes is (relatively) constant.

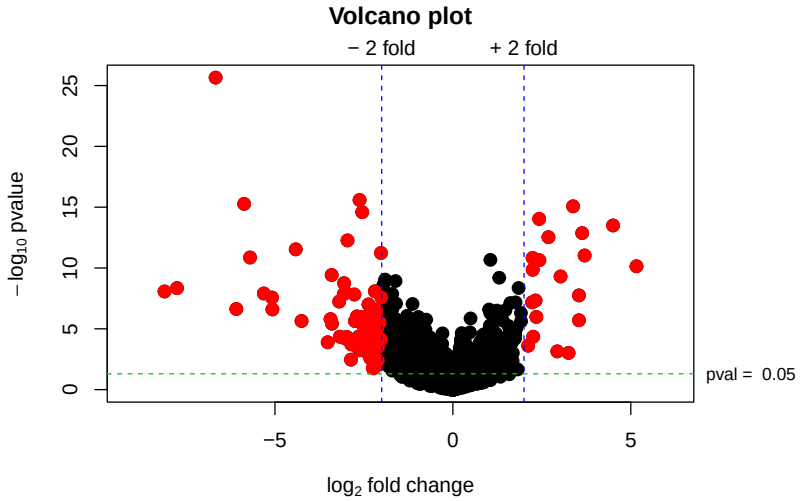


Next, we conducted differential expression analysis. Our contrast of interest is patients with tumor vs patients without tumor. There are a total of 427 significantly differentially expressed genes (see table below).

```
##        WITH TUMOR-TUMOR FREE
## Down                     274
## NotSig                 16907
## Up                       153
```

We visualized this information using a smear plot or volcano plot. Significantly DE genes are found outside of the logFC = 2 (4-fold change) lines and the p-value=0.05 line.





The top 10 differentially expressed genes are given below.

```
##                logFC     logCPM       PValue          FDR
## MUC6       -6.667729  2.6112912 2.225593e-26 3.857842e-22
## AMICA1     -2.621876  5.2026294 2.570734e-16 2.228055e-12
## TKTL1      -5.867111  4.0174139 5.342037e-16 3.086629e-12
## LOC645323   3.378178 -1.7124679 8.489856e-16 3.679079e-12
## ALDH8A1    -2.547559 -0.2033563 2.565122e-15 8.892763e-12
## FXYD2       2.425359  0.9598868 9.176970e-15 2.651227e-11
## RPL10L      4.501115 -2.0903812 3.184549e-14 7.885853e-11
## PCSK2       3.634420  0.8449425 1.345021e-13 2.914324e-10
## KLK2        2.685679 -1.0025817 2.937506e-13 5.657636e-10
## SCG3       -2.960947 -0.9126571 5.400180e-13 9.360671e-10
```

DE Genes Literature Analysis

Several of the top 10 genes have already been implicated in HNSCC. The top hit, *MUC6*, is featured in the manuscript "Analysis of MUC6 polymorphisms on the clinicopathologic characteristics of Asian patients with oral squamous cell carcinoma" by Hua et al. *AMICA1* is described in Frontiers in Genetics article "Network-Based Predictors of Progression in Head and Neck Squamous Cell Carcinoma" by Sanati et al. The *TKTL1* gene has been elucidated in a functional study by Sun et al. entitled "TKTL1 is activated by promoter hypomethylation and contributes to head and neck squamous cell carcinoma carcinogenesis through increased aerobic glycolysis and HIF1alpha stabilization." Another functional study related to HNSCC has been conducted on the aldehyde dehydrogenases, specifically *ALDH8A1*, by Kim et al. in their manuscript "Targeting aldehyde dehydrogenase activity in head and neck squamous cell carcinoma with a novel small molecule inhibitor."

Others of these top 10 genes have been implicated in other types of cancers. For example, *FXYD2* was found to be a potential target in ovarian clear cell carcinoma by Hsu et al. in their manuscript "Targeting FXYD2 by cardiac glycosides potently blocks tumor growth in ovarian clear cell carcinoma." *LOC645323* was found to be associated with colorectal cancer by Yang et al. in "Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns". *PCSK2* was identified as a factor in pheochromocytoma-paraganglioma by Remes et al. in "PCSK2 expression in neuroendocrine tumors points to a midgut, pulmonary, or pheochromocytoma-paraganglioma origin."

We can use our DE gene list to corroborate the existing evidence of roles in HNSCC and to identify new potential targets (that currently have no known associations or are only known to be associated with other cancers) to better understand the molecular effects of HNSCC and develop therapeutics to treat HNSCC.

Survival Analysis

We then merged the RNAseq and clinical datasets to fit the Coxph model for survival for each significantly DE gene. We converted the categorical variable **vital status** using a numerical code (living=0, deceased=1) to put vital status into the model. The Coxph model is semi-parametric, as it doesn't have a required distribution but does assume sample independence, a linear association between log(risk) and risk factors, and proportionality of risk effects over time.

```
##                       coef exp(coef)    se(coef)         z   Pr(>|z|)
## d.merge[, i] -0.01600748   0.98412 0.008486922 -1.886134 0.05927685
```

The Coxph model is given by: $h_t = h_0(t) \cdot exp(b_i x_i)$, where

- t is survival time
- h_t is a random variable representing the risk of dying at time t
- h_0 is the baseline risk, when the impact of the gene is zero
- X_i is the covariate, in this case the ith gene, and
- b_i are the coefficients representing the impact of the ith gene.

The $exp(b_i x_i)$'s are the hazard ratios (HR). A HR = 1 indicates no effect, HR > 1 indicates an increase in hazard, and HR < 0 indicates reduced hazard ("Cox Proportional-Hazards Model", STHDA).

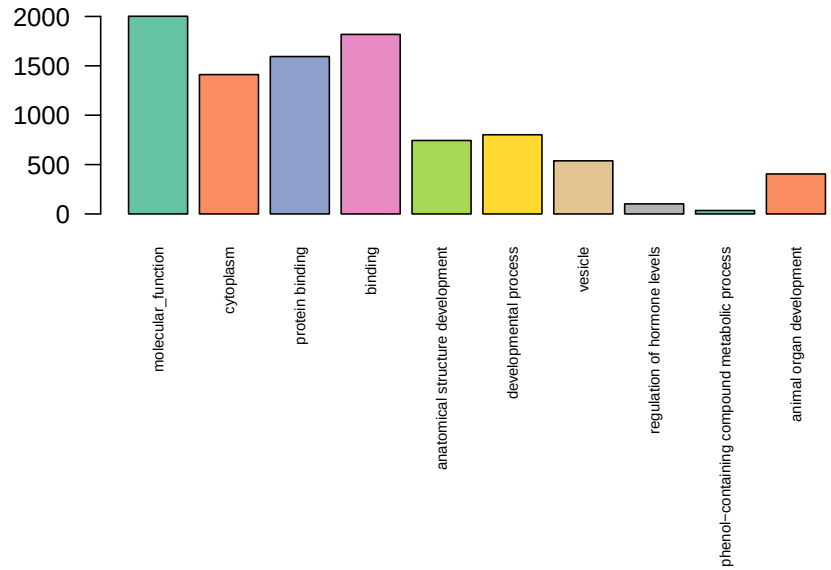
We then filtered the results to only include genes with p-value < 0.05, resulting in 447 genes with a significant impact on survival. We did not implement multiple testing correction here.

```
##     gene.names p.vals
## 7         TTC36 0.0135
## 17     SLC7A14 0.0056
## 20       TMED4 0.0026
## 31        ACP1 0.0116
## 32     SCARNA6 0.0108
## 40    B4GALNT1 0.0017
## 62         RAN 0.0022
## 70      DIAPH3 0.0087
## 72    KIAA0319 0.0357
## 73       ABHD8 0.0242
## 75     FAM162B 0.0017
## 77       PDPK1 0.0173
## 86        ENC1 0.0023
## 87        ING4 0.0106
## 91        RGL4 0.0018
## 95        SELL 0.0102
## 103   LYPLA2P1 0.0435
## 124      PREX1 0.0011
## 128      FITM2 0.0253
## 140    N4BP2L1 0.0208
## 149      6-Sep 0.0229
## 150  LST-3TM12 0.0122
## 154      RIBC2 0.0078
## 159      AMPD3 0.0325
## 160      RBM11  0.036
## 172      DLEU1 0.0404
## 183     TXNDC5 0.0143
## 190     ARID1B 0.0103
## 194       CD55 0.0116
## 197   AASDHPPT 0.0325
```

Pathway Analysis

We performed Gene Ontology (GO) analysis on the 81 significant DE genes. We found that several of the top 10 affected pathways are essential to cell function and development (see table below). The majority of affected pathways are molecular functions, binding, and cytoplasm functions. This may suggest a major affect on cell signaling, which could explain the disregulation that leads to cancer.

```
##                                                  Term Ont     N   DE
## GO:0003674                          molecular_function  MF 18522 2002
## GO:0005737                                  cytoplasm  CC 12174 1411
## GO:0005515                             protein binding  MF 14090 1594
## GO:0005488                                    binding  MF 16737 1818
## GO:0048856          anatomical structure development  BP  6016  744
## GO:0032502                   developmental process  BP  6593  802
## GO:0031982                                 vesicle  CC  4180  538
## GO:0010817              regulation of hormone levels  BP   539  102
## GO:0018958 phenol-containing compound metabolic process  BP   113   35
## GO:0048513                  animal organ development  BP  3052  405
##                    P.DE
## GO:0003674 1.994468e-21
## GO:0005737 2.699357e-16
## GO:0005515 4.692491e-16
## GO:0005488 2.578971e-12
## GO:0048856 2.993873e-11
## GO:0032502 9.942121e-11
## GO:0031982 1.879783e-10
## GO:0010817 4.154728e-10
## GO:0018958 9.250175e-10
## GO:0048513 1.806135e-09
```



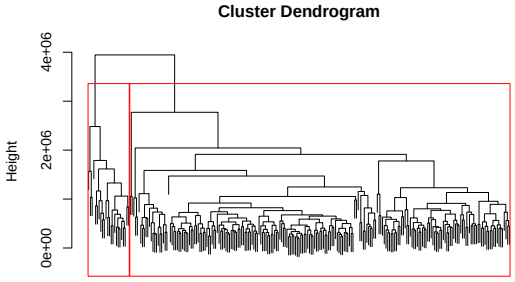
Clustering for Disease Sub-Types

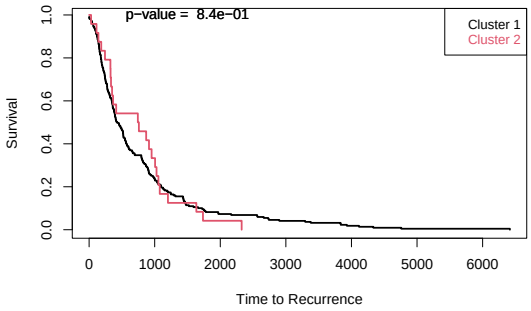
Finally, we conducted cluster analysis to identify subtypes of HNSCC. We created hierarchical clustering plots to visualize the relatedness between the individuals and determine how many clusters to use for analysis. Creating two clusters splits the data into groups of size 219 and 24, however the difference is not statistically significant according to a log rank test (p=0.842). Thus we concluded that there are no observable disease subtypes.

```
## cluster.mem
##   1   2
## 219  24
```

```
## Call:
## survdiff(formula = Surv(days_to_last_followup) ~ as.factor(cluster.mem),
##     data = D, na.action = na.exclude)
##
##                             N Observed Expected (O-E)^2/E (O-E)^2/V
## as.factor(cluster.mem)=1 219      219    218.1   0.00402    0.0396
## as.factor(cluster.mem)=2  24       24     24.9   0.03515    0.0396
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.8
```

```
## [1] 0.8421797
```





Discussion

Head and neck squamous cell carcinoma patients have differentially expressed genes depending on their tumor status. Patients with tumors show differential expression of 427 genes such as *MUC6*, *AMICA1*, *TKTL1*, *ALDH8A1*, *FXYD2*, *RPL10L*, *PCSK2*, *KLK2*, *SCG3* as compared to their counterparts without tumors. There are 447 total genes with a significant impact on survival. The pathways most significantly affected in HNSCC patients with tumors are those with significant biological function, for example protein binding, hormone regulation, development, and metabolism.

These identified genes and their de-regulation patterns may serve as molecular markers for HNSCC. Future studies should be conducted on each of these genes and their protein products to develop therapeutics for HNSCC patients.

References

Brooks, Susan S et al. “A novel ribosomopathy caused by dysfunction of RPL10 disrupts neurodevelopment and causes X-linked microcephaly in humans.” Genetics vol. 198,2 (2014): 723-33. doi:10.1534/genetics.114.168211.

The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517, 576–582 (2015). 28 January 2015, https://doi.org/10.1038/nature14129.

“Cox Proportional-Hazards Model.” Cox Proportional-Hazards Model , Statistical tools for high-throughput data analysis, www.sthda.com/english/wiki/cox-proportional-hazards-model. Accessed 25 July 2024.

Hannu, Koistinen et al. “KLK-targeted Therapies for Prostate Cancer.” EJIFCC vol. 25,2 207-18. 4 Sep. 2014.

Hsu, I-Ling et al. “Targeting FXYD2 by cardiac glycosides potently blocks tumor growth in ovarian clear cell carcinoma.” Oncotarget vol. 7,39 (2016): 62925-62938. doi:10.18632/oncotarget.7497.

Hua, Chun-Hung et al. “Analysis of MUC6 polymorphisms on the clinicopathologic characteristics of Asian patients with oral squamous cell carcinoma.” Journal of cellular and molecular medicine vol. 27,17 (2023): 2594-2602. doi:10.1111/jcmm.17886.

Kim, Jeewon et al. “Targeting aldehyde dehydrogenase activity in head and neck squamous cell carcinoma with a novel small molecule inhibitor.” Oncotarget vol. 8,32 52345-52356. 10 Apr. 2017, doi:10.18632/oncotarget.17017.

Remes, Satu Maria et al. “PCSK2 expression in neuroendocrine tumors points to a midgut, pulmonary, or pheochromocytoma-paraganglioma origin.” APMIS : acta pathologica, microbiologica, et immunologica Scandinavica vol. 128,11 (2020): 563-572. doi:10.1111/apm.13071.

Sanati, Nasim et al. “Network-Based Predictors of Progression in Head and Neck Squamous Cell Carcinoma.” Frontiers in genetics vol. 9 183. 29 May. 2018, doi:10.3389/fgene.2018.00183.

Sun, Wenyue et al. “TKTL1 is activated by promoter hypomethylation and contributes to head and neck squamous cell carcinoma carcinogenesis through increased aerobic glycolysis and HIF1alpha stabilization.” Clinical cancer research : an official journal of the American Association for Cancer Research vol. 16,3 (2010): 857-66. doi:10.1158/1078-0432.CCR-09-2604.

Wang, Yi et al. “SCG3 Protein Expression in Glioma Associates With less Malignancy and Favorable Clinical Outcomes.” Pathology oncology research : POR vol. 27 594931. 26 Feb. 2021, doi:10.3389/pore.2021.594931.

Yang, Xiaofei et al. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns, Briefings in Bioinformatics, Volume 18, Issue 5, September 2017, Pages 761–773, https://doi.org/10.1093/bib/bbw063.

Appendix: R Code

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE, out.width = "65%", fig.align = "c
    tidy = TRUE, tidy.opts = list(width.cutoff = 100))
library(readxl)
library(readr)
library(tidyverse)
library(formatR)
library(ggplot2)
library(ggrepel)
library(RColorBrewer)
library(BiocManager)
library(Biobase)
library(edgeR)
library(limma)
library(fgsea)
library(data.table)
library(org.Hs.eg.db)
library(GO.db)
library(clusterProfiler)
library(biomaRt)
library(survival)

load("C:/Users/orippee/OneDrive - University of Kansas Medical Center/Graduate/BIOS 855 - Genomics/Final

d.clin <- read.csv("C:/Users/orippee/OneDrive - University of Kansas Medical Center/Graduate/BIOS 855 -
    header = T)


rownames(d.clin) <- d.clin$bcr_patient_barcode
rownames(d.clin) <- gsub(pattern = "\\-", ".", rownames(d.clin))

all(rownames(d.mRNA) == rownames(d.clin))
common.subjects <- intersect(rownames(d.mRNA), rownames(d.clin))
length(common.subjects)
d.mRNA <- d.mRNA[common.subjects, ]
d.clin <- d.clin[common.subjects, ]
all(rownames(d.mRNA) == rownames(d.clin))
d.mRNA[1:6, 1:5]
head(d.clin)
table(d.clin$cancer_status)

# TUMOR FREE WITH TUMOR 154 89 154 + 89 = 243 279-243 = 36 missing cancer status

d.clin <- subset(d.clin, !is.na(cancer_status))
dim(d.clin)
d.clin <- d.clin[order(d.clin$cancer_status), ]
d.mRNA <- d.mRNA[rownames(d.clin), ]  # Subsets d.mRNA and puts in the same order as d.clin
all(rownames(d.mRNA) == rownames(d.clin))
table(d.clin$cancer_status)


### Genes filltering based on counts
```

```r
dat <- t(d.mRNA)
dat <- dat[rowSums(dat) != 0, ]  ## Remove if gene counts is zero for all samples

### Mean Variance plot by gene
###-----------------------------------------------
mean.x <- apply(dat, 1, mean)
var.x <- apply(dat, 1, var)
plot(log10(mean.x), log10(var.x), pch = 20, xlab = "Mean (in log 10 scale)", ylab = "Variance (in log 1
    xlim = c(0, 9), ylim = c(0, 9), main = "Variance vs Mean")
abline(0, 1, col = 3, lwd = 2)

### Barplot
###-----------------------------------------------
lib.size <- colSums(dat)
barplot(lib.size, xaxt = "n", xlab = "Study samples", ylab = "Library size values", main = "Barplot of
    col = c(rep("lightgreen", 11), rep("lightcoral", 11)), sub = "(Red line represents mean)")
abline(h = mean(lib.size), lwd = 2, col = "red")
legend("topright", c("Normal", "Tumor"), lty = c(1, 1), lwd = c(4, 4), col = c("lightgreen", "red"),
    bty = "o", cex = 1, text.font = 2)

### Boxplot (In log10 scale)
###-----------------------------------------------
boxplot(x = as.list(as.data.frame(log10(dat + 1))), xlab = "Samples", ylab = "Values in log10 scale",
    main = "Boxplot across study samples", col = c(rep("lightgreen", 11), rep("lightcoral", 11)))
legend("topright", c("Normal", "Tumor"), lty = c(1, 1), lwd = c(4, 4), col = c("lightgreen", "red"),
    bty = "o", cex = 1, text.font = 2)


for (i in 1:ncol(dat)) {
    if (i == 1) {
        plot(density(log10(dat[, 1])), main = "Density plot across study samples", xlab = "Subjects",
            col = "gray", ylim = c(0, 0.8))
    } else {
        den <- density(log10(dat[, i]))
        lines(den$x, den$y, col = "gray")
    }
}


## Filter and normalize data
##-------------------------------------------------------------------------------------------
keep <- rowSums(cpm(as.matrix(dat)) > 1) >= 2
# At least 2 samples have to have cpm > 1.
dat.filtered <- dat[keep, ]
# dim(dat.filtered)
rm(keep)
d <- DGEList(counts = as.matrix(dat.filtered), lib.size = colSums(dat.filtered), group = c(rep("TUMOR F
    154), rep("WITH TUMOR", 89)))
# dim(d)
d <- calcNormFactors(d, method = "TMM")  ## Calculates normalization factors
d <- estimateDisp(d)  ## Calculates genewise dispersion parameter adjusted using bayesian empirical met

## BCV
```

```r
###-------------------------------------------------------------
plotBCV(d)


de.test <- exactTest(d, pair = c("TUMOR FREE", "WITH TUMOR"))  ## First value is baseline
de.test.FDR <- topTags(de.test, n = Inf, adjust.method = "BH", sort.by = "PValue")
# head(de.test.FDR$table)
summary(de <- decideTestsDGE(de.test, p = 0.05, adjust = "BH"))  ## Counts up- and down-regulated genes


### plotSmear
###-------------------------------------------------------------
detags <- rownames(d)[as.logical(de)]
plotSmear(de.test, de.tags = detags, main = "Smear Plot", sub = "(The horizontal blue lines show 4-fold
    cex = 1)
abline(h = c(-2, 2), col = "blue")


### Volcano plot
###-------------------------------------------------------------
d.volcano <- de.test.FDR$table[, c("logFC", "PValue")]
par(mar = c(5, 4, 4, 5))
plot(d.volcano$logFC, -log(d.volcano$PValue, 10), main = "", pch = 20, cex = 2, xlab = expression(log[2]
    fold ~ change), ylab = expression(-log[10] ~ pvalue), xlim = c(min(d.volcano$logFC) - 1, max(d.volca
    1))
title("Volcano plot")

lfc <- 2
pval <- 0.05

# Selecting interesting genes
sigGenes <- (abs(d.volcano$logFC) > lfc & -log(d.volcano$PValue, 10) > -log10(pval))

# Identifying the selected genes
points(d.volcano[sigGenes, ]$logFC, -log(d.volcano[sigGenes, ]$PValue, 10), pch = 20, col = "red", cex =
abline(h = -log10(pval), col = "green3", lty = 2)
abline(v = c(-lfc, lfc), col = "blue", lty = 2)
mtext(paste("pval = ", round(pval, 2)), side = 4, at = -log10(pval), cex = 0.8, line = 0.5, las = 1)
mtext(c(paste("-", lfc, "fold"), paste("+", lfc, "fold")), side = 3, at = c(-lfc, lfc), cex = 1, line =
d.edgeR.result <- de.test.FDR$table
head(d.edgeR.result, n = 10)
d.merge <- cbind(d.clin, d.mRNA)

d.merge$death.ind <- ifelse(d.merge$vital_status == "DECEASED", 1, 0)

gene.names <- NULL
p.vals <- NULL

# 8 = first column in merge dataset where gene expression starts
for (i in 8:247) {
    fit <- coxph(Surv(days_to_last_followup, death.ind) ~ d.merge[, i], data = d.merge)
    gene.names <- c(gene.names, colnames(d.merge)[i])  #append gene name to list
    p.vals <- c(p.vals, summary(fit)$coef[5])
```

```
}

summary(fit)$coef
p.vals <- round(p.vals, digits = 4)
surv.results <- data.frame(cbind(gene.names, p.vals))

surv.sig.genes <- surv.results[surv.results$p.vals < 0.05, ]
head(surv.sig.genes, n = 30)

d.go <- d.edgeR.result
d.go.DE <- subset(d.go, PValue < 0.05)
d.entrez.id <- mapIds(org.Hs.eg.db, keys = rownames(d.go.DE), column = "ENTREZID", keytype = "SYMBOL")

# length(d.entrez.id) head(d.entrez.id) all(d.go.DE$hgnc_symbol==names(d.entrez.id))
go.test <- goana(d.entrez.id, species = "Hs")
go.results <- topGO(go.test, sort = "DE", number = Inf)
head(go.results, 10)
# sum(go.results$P.DE<10^(-5))


term <- c("molecular_function", "cytoplasm", "protein binding", "binding", "anatomical structure develop
    "developmental process", "vesicle", "regulation of hormone levels", "phenol-containing compound meta
    "animal organ development")
DE <- c(2002, 1411, 1594, 1818, 744, 802, 538, 102, 35, 405)

par(mar = c(12, 4, 4, 4))
barplot(height = DE, names.arg = term, las = 2, cex.names = 0.6, col = brewer.pal(8, "Set2"))

mm <- as.matrix(d.mRNA)

require(graphics)
d <- dist(mm, method = "euclidean")  # distance matrix
h.clust <- hclust(d, method = "complete")
# str(h.clust)
plot(h.clust, labels = F, xlab = "", sub = "")  # display dendogram


# Cluster dendogram with 2 clusters ----------------------------------------
cluster.mem <- cutree(h.clust, k = 2)  # cut tree into clusters

# draw dendogram with red borders around the clusters
rect.hclust(h.clust, k = 2, border = "red")
table(cluster.mem)


### Assess the clusters using CoxPH analysis among the clusters
d.merge$bcr_patient_barcode <- gsub(pattern = "\\-", ".", d.merge$bcr_patient_barcode)
# all(names(cluster.mem)==d.merge$bcr_patient_barcode)

D <- cbind(cluster.mem, d.merge[, 1:7])  # Combine the cluster.id with clinical part of data
# D[1:5,1:5]

### Log rank test:
```

```r
###-------------------------------------------------------
lr.test <- survdiff(Surv(days_to_last_followup) ~ as.factor(cluster.mem), data = D, na.action = na.exclu
lr.test
pval <- pchisq(lr.test$chisq, 1, lower.tail = F)
pval

### Kaplan Meier plot of the two clusters
###-------------------------------------------------------
for (i in 1:length(unique(sort(cluster.mem)))) {
    ii <- unique(sort(cluster.mem))[i]
    ddd <- D[D$cluster.mem == ii, ]
    f <- survfit(Surv(days_to_last_followup) ~ 1, data = ddd)
    if (i == 1) {
        plot(f, col = "black", conf.int = FALSE, xlab = "Time to Recurrence", ylab = "Survival", main =
            lwd = 2)
        clust.name <- "Cluster 1"
        clr <- i
    } else {
        lines(f, col = i, conf.int = F, lwd = 2)
        clust.name <- c(clust.name, paste("Cluster", i))
        clr <- c(clr, i)
    }
    text(1500, 1, paste("p-value = ", format(pval, digits = 2, scientific = T)))
}
legend("topright", clust.name, text.col = clr, cex = 0.9)
```