Insights from TCGA and GEO Analyses of Mesothelioma

Brad Kaptur

Overview

The goal of this project is to use publicly available genetic data corresponding to mesothelioma patient isolated tumor samples to analyze key genes involved in mesothelioma and evaluate those key genes as they relate to clinical outcomes.

Introduction

Mesothelioma is a form of cancer that develops from the cells of the protective lining covering many of the body's internal organs. This malignancy is often caused by exposure to asbestos, and most people who develop the disease have worked in careers related to mining. Epidemiological studies have shown that malignant mesothelioma is primarily caused by exposure to asbestos fibers; even with asbestos abatement efforts, the incidence of mesothelioma may continue to increase until 2020 due to the extended latency period of the condition [1].The disease presents as shortness of breath as a result of pleural effusion, as well as pain in the chest wall. Diagnosis is difficult because the symptoms are similar to many other conditions, and a biopsy is usually needed [2]. Additionally, while there have been modest improvements in the outlook for patients afflicted with mesothelioma, the prognosis for malignant mesothelioma remains disappointing [2]. Chemotherapy is the only treatment that has been proven to improve survival. In 2004, the FDA approved Pemetrexed in combination with Cisplatin for treatment of malignant pleural mesothelioma. Although the Cis/Pem regimen has been a major advance in the treatment of this condition, the response rate is only about 40%. Thus, the majority of patients who receive the drug will be exposed to its toxic side effects, without any benefit over a placebo [3]. Thus, there is a pressing need to find predictive biomarkers to identify patients from the mesothelioma population who would be candidates for Cis/Pem therapy [3]. Diseases are often the result of aberrances on several levels-and mesothelioma is no exception. Insight can be gained from disciplines such as cellular engineering, epidemiology, and even bioinformatics and machine learning. Taking a multidisciplinary approach to disease allows for understanding and treatment from several levels from the cell to the population.

Methods

Previous datasets used in the study of mesothelioma are publicly available on NCBI's Gene Expression Omnibus (GEO). One specific dataset of interest was GSE51024, which was a gene expression study performed with the goal of discovering critical pathways and networks or therapeutic targets in pleural mesothelioma. The dataset contains gene expression data for malignant pleural mesothelioma tumor and paired normal lung tissue for 41 tumors using Affymetrix U133 plus 2.0 chips. The overall design of this study used RNA isolated from frozen resected tumor and normal tissues, and the dataset was made public on August 31, 2014. In the analysis performed in this work, GSE51024 dataset was loaded into RStudio from the GEO database, and each sample was classified as normal or tumor using the sample information provided on the GEO website. There were a total of 14 unpaired tumor samples in addition to the 41 tumor and 41 normal paired samples supplied in the dataset. For subsequent analyses, these unpaired samples were discarded. From the raw dataset, no significant outliers were detected by visualization of the raw probe intensities on a boxplot, so no further samples were discarded. Normalization was necessary for the paired samples, and thus Robust Multi-array Averaging (RMA) was performed. Through this procedure, incorporating the median polish algorithm and quantile normalization, a cleaned, normalized dataset of the 41 pairs was attained, as again visualized through boxplot comparison. A model matrix was created which incorporated the paired study design using two factors: sibship (the pairing of samples taken from the same patient) and status (a differentiator of tumor and normal samples). The lmFit linear model for series of arrays function from the limma package was used to fit a linear model for each gene for the series of arrays. The output of this function was then used as the input for the ebayes empirical bayes statistics for differential expression function-also of the limma package. This output ranked the genes in order of their evidence for differential expression using the empirical Bayes method to shrink the gene-wise sample variances toward a common value. From there, a Benjamini-Hochberg correction was applied to diminish the effect of the false discovery rate on the statistical significance of the output. Gene name symbols were attained for the significant probes using the hgu133plus2.db package file to reference probe names. From here, a top expression set was generated using the output of the ranked top gene list. This expression set was then used as the input for a heatmap using hierarchical clustering to group similar samples together. Subsequently, gene specific comparisons were performed for the top genes that were significantly more highly or more lowly expressed in tumor samples relative to normal, and graphical output was obtained for these genes of interest. Additionally, the list of top genes was refined to remove NAs and duplicate genes and exported as a text document for subsequent analysis as a network model.

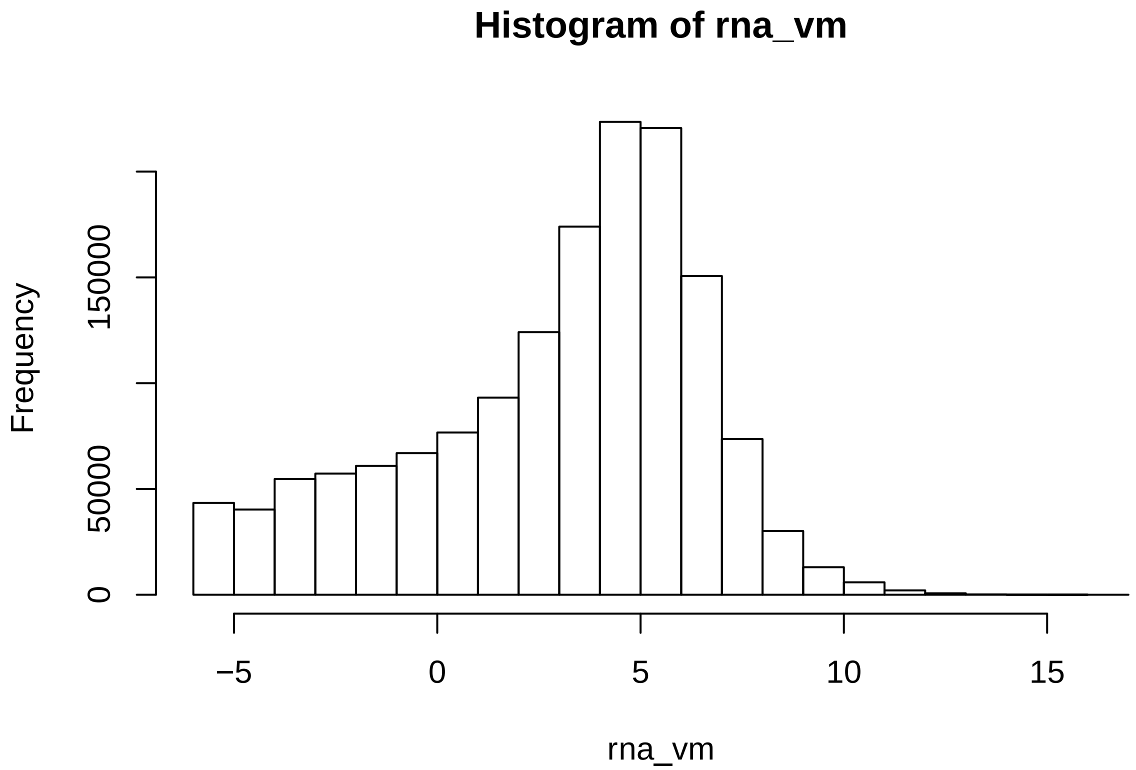
Network analysis was performed in Cytoscape starting with a blank network. The application Agilent Literature Search was used to generate networks in this work. The list of top non-redundant genes was inputted into the Terms box of the Agilent Literature Search application using the default Context parameters of “Homo sapiens” and “human” as well as the default search controls of 10 max engine matches and using context and concept lexicon to restrict search. The application was run using these parameters, which outputted a network based on these top genes as well as genes of interest. The outputted literature-based network consisted of one main highly connected branch as well as several branches that only contained one of the search genes and a small number of literature-only genes (that were not search terms). These small gene branches were discarded and not part of subsequent analyses. The remaining network was checked for nodes that were not actual gene names or textual errors. From this highly connected network of search genes and result genes, it was of interest to simplify the network down to smaller, highly-connected branches. To do so, the result genes were sorted by number of connections, and the top 5 result genes were used to create a network of their connected search genes.

While the GEO dataset GSE51024 offered interesting insights into differential gene expression, one area in which it was lacking was relevant accompanying clinical data associated with the tumor specimens. Currently, The Cancer Genome Atlas of the National Cancer Institute and the National Human Genome Research Institute hosts Mesothelioma data consisting of 87 cases with accompanying SNP, methylation, mRNA, miRNA, and clinical data accompanying. The FireBrowse pipeline interface hosted by the Broad Institute of MIT & Harvard performs automated, accelerated analysis of these TCGA samples and stores analysis-ready patient samples in their stddata archieves. As part of the automated statistical testing, the association between 18203 genes and 11 clinical features across 86 of the 87 mesothelioma samples hosted on TCGA was performed. Statistical significance was defined by P value < 0.05 and Q value < 0.3. Under this analysis, 5 of the 7 clinical features were significantly associated with at least one gene. Of particular interest was the parameter termed “DAYS_TO_DEATH_OR_LAST_FUP”, which is reflective of the duration from a patient’s initial diagnosis until they either become diseased or their last return visit. In this metric, the assumption that their last return visit is substitutable for death is justified by the low long-term survival rate of malignant mesothelioma after initial diagnosis, though it is noted that separation of this category into the two distinct variables may have improved the predictive validity of subsequent statistical tests by decreasing noise. FireBrowse’s analysis revealed 30 genes correlated with this variable, and the top five were of interest for visualization and further study. Of particular interest were correlation of gene expression data against clinical outcome, and thus the Illumina RNA-seq data file (illuminahiseq_rnaseqv2-RESM_genes_normailzed) and the Meso clinical data files (Merge_Clinical) were downloaded from FireBrowse and loaded into R. In the cleaning step, genes whose expression was 0 in greater than 50% of the samples were removed. The limma package’s voom function was used to normalize the data, which were then converted to z-score values. For each gene of interest, a survival analysis was performed. Survival curves were plotted corresponding to normal and low expression of each gene (defined by a t-score of -1). Additionally, the point at which 50% of patients were deceased was found for each gene, and the corresponding date was compared for both expression populations.

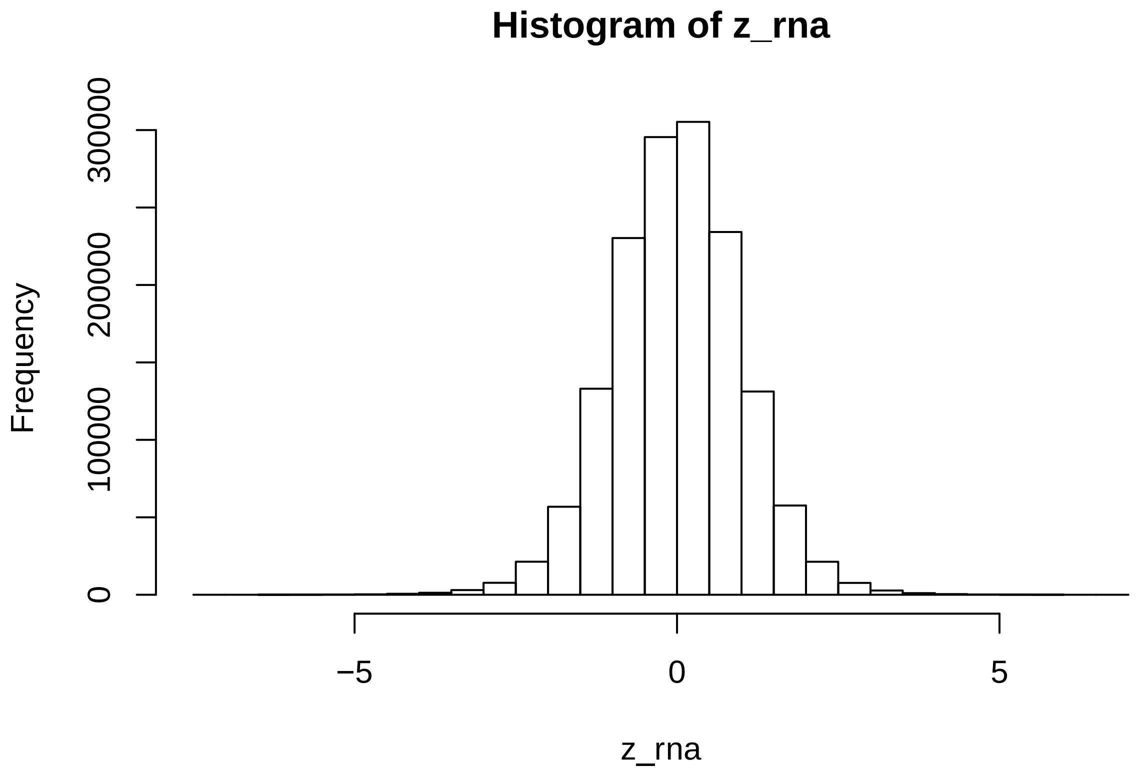
```r
library(survival)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(limma)
#code adapted from BioStars tutorial here: https://www.biostars.org/p/153013/

#load data from saved files
rna <- read.table("Meso_RNA/MESO.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normali
rna <- rna[-1,]
clinical <- data.frame(t(read.table("Meso_Clinical/MESO.merged_only_clinical_clin_format.txt",header=T,

#clean RNA file
rem <- function(x){
  x <- as.matrix(x)
  x <- t(apply(x,1,as.numeric))
  r <- as.numeric(apply(x,1,function(i) sum(i == 0)))
  remove <- which(r > dim(x)[2]*0.5)
  return(remove)
}
remove <- rem(rna)
rna <- rna[-remove,]
vm <- function(x){
  x <- t(apply(x,1,as.numeric))
  ex <- voom(x)
  return(ex$E)
}
rna_vm  <- vm(rna)
colnames(rna_vm) <- gsub("\\.","-",substr(colnames(rna),1,12))

#look at rna distribution pre-normalization
hist(rna_vm)
```



```
#normalize RNA
scal <- function(x){
  mean_n <- rowMeans(x)
  sd_n <- apply(x,1,sd)
  res <- matrix(nrow=nrow(x), ncol=ncol(x))
  colnames(res) <- colnames(x)
  rownames(res) <- rownames(x)
  for(i in 1:dim(x)[1]){
    for(j in 1:dim(x)[2]){
      res[i,j] <- (x[i,j]-mean_n[i])/sd_n[i]
    }
  }
  return(res)
}
z_rna <- scal(rna_vm)

#look at RNA distribution post-normalization
hist(z_rna)
```



```
#set row names to gene names
rownames(z_rna) <- sapply(rownames(z_rna), function(x) unlist(strsplit(x,"\\|"))[[1]])

#match patient info, output number of useable patients
clinical$IDs <- toupper(clinical$patient.bcr_patient_barcode)
sum(clinical$IDs %in% colnames(z_rna))
```

```
## [1] 86
```

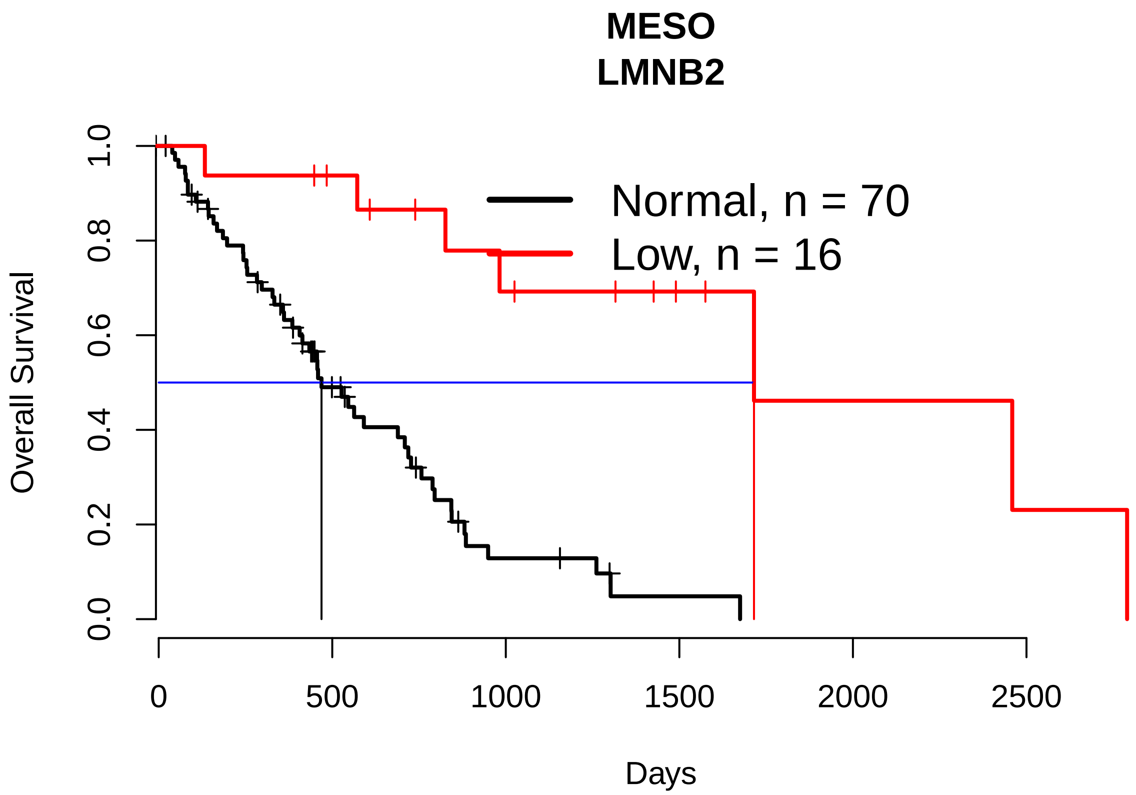
```
#retain useful columns
ind_keep <- grep("days_to_new_tumor_event_after_initial_treatment",colnames(clinical))

#condense followups
new_tum <- as.matrix(clinical[,ind_keep])
new_tum_collapsed <- c()
for (i in 1:dim(new_tum)[1]){
  if(sum(is.na(new_tum[i,])) < dim(new_tum)[2]){
    m <- max(new_tum[i,],na.rm=T)
    new_tum_collapsed <- c(new_tum_collapsed,m)
  } else {
    new_tum_collapsed <- c(new_tum_collapsed,"NA")
  }
}
ind_keep <- grep("days_to_death",colnames(clinical))
death <- as.matrix(clinical[,ind_keep])
```

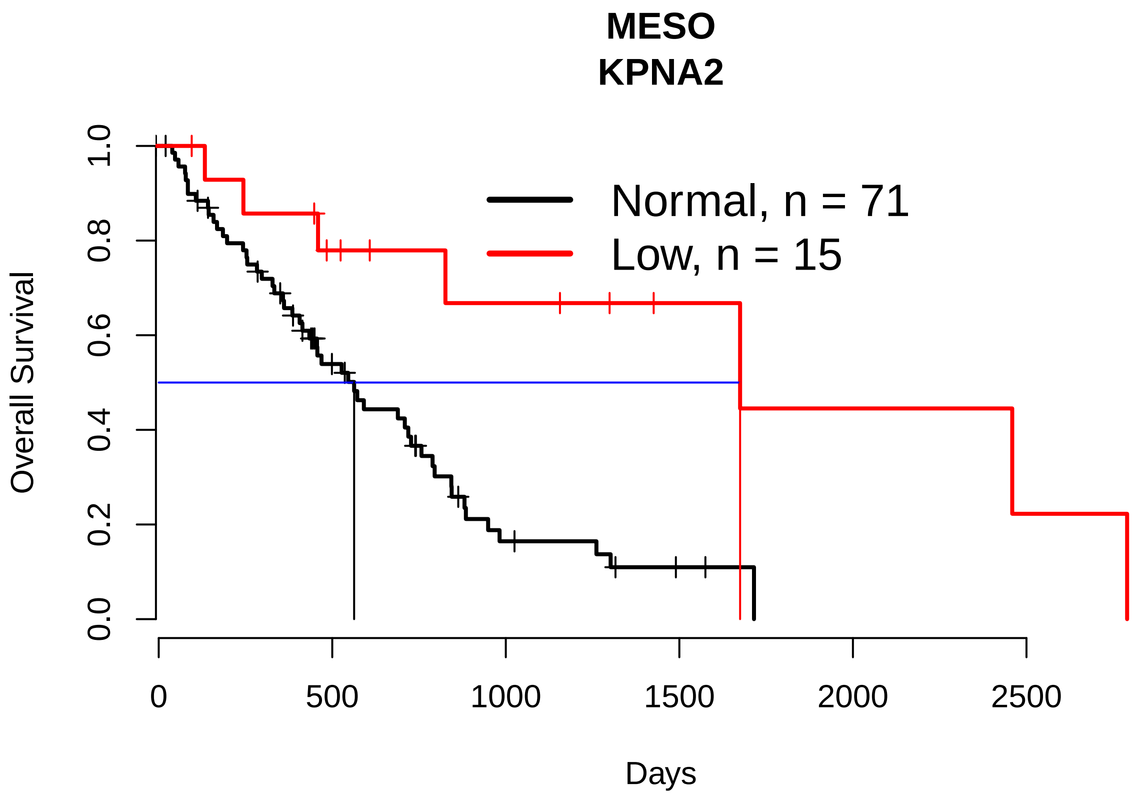
```r
death_collapsed <- c()
for (i in 1:dim(death)[1]){
  if(sum(is.na(death[i,])) < dim(death)[2]){
    m <- max(death[i,],na.rm=T)
    death_collapsed <- c(death_collapsed,m)
  } else {
    death_collapsed <- c(death_collapsed,"NA")
  }
}
ind_keep <- grep("days_to_last_followup",colnames(clinical))
fl <- as.matrix(clinical[,ind_keep])
fl_collapsed <- c()
for (i in 1:dim(fl)[1]){
  if(sum(is.na(fl[i,])) < dim(fl)[2]){
    m <- min(fl[i,],na.rm=T)
    fl_collapsed <- c(fl_collapsed,m)
  } else {
    fl_collapsed <- c(fl_collapsed,"NA")
  }
}
all_clin <- data.frame(new_tum_collapsed,death_collapsed,fl_collapsed)
colnames(all_clin) <- c("new_tumor_days", "death_days", "followUp_days")

#create vectors for clinical parameters
all_clin$new_time <- c()
for (i in 1:length(as.numeric(as.character(all_clin$new_tumor_days)))){
  all_clin$new_time[i] <- ifelse(is.na(as.numeric(as.character(all_clin$new_tumor_days))[i]),
                                 as.numeric(as.character(all_clin$followUp_days))[i],as.numeric(as.chara
}
all_clin$new_death <- c()
for (i in 1:length(as.numeric(as.character(all_clin$death_days)))){
  all_clin$new_death[i] <- ifelse(is.na(as.numeric(as.character(all_clin$death_days))[i]),
                                  as.numeric(as.character(all_clin$followUp_days))[i],as.numeric(as.cha
}

#output number alive and dead patients
table(clinical$patient.vital_status)
```

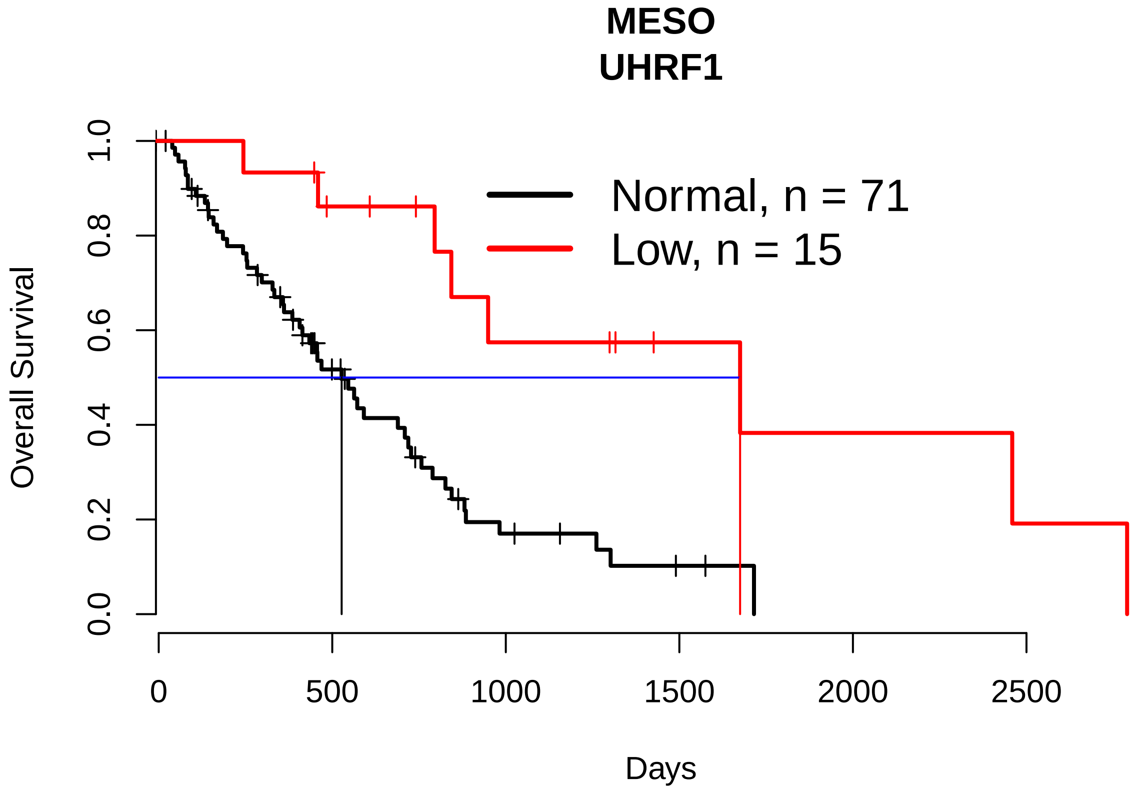
```
##
## alive  dead
##    29    58
```

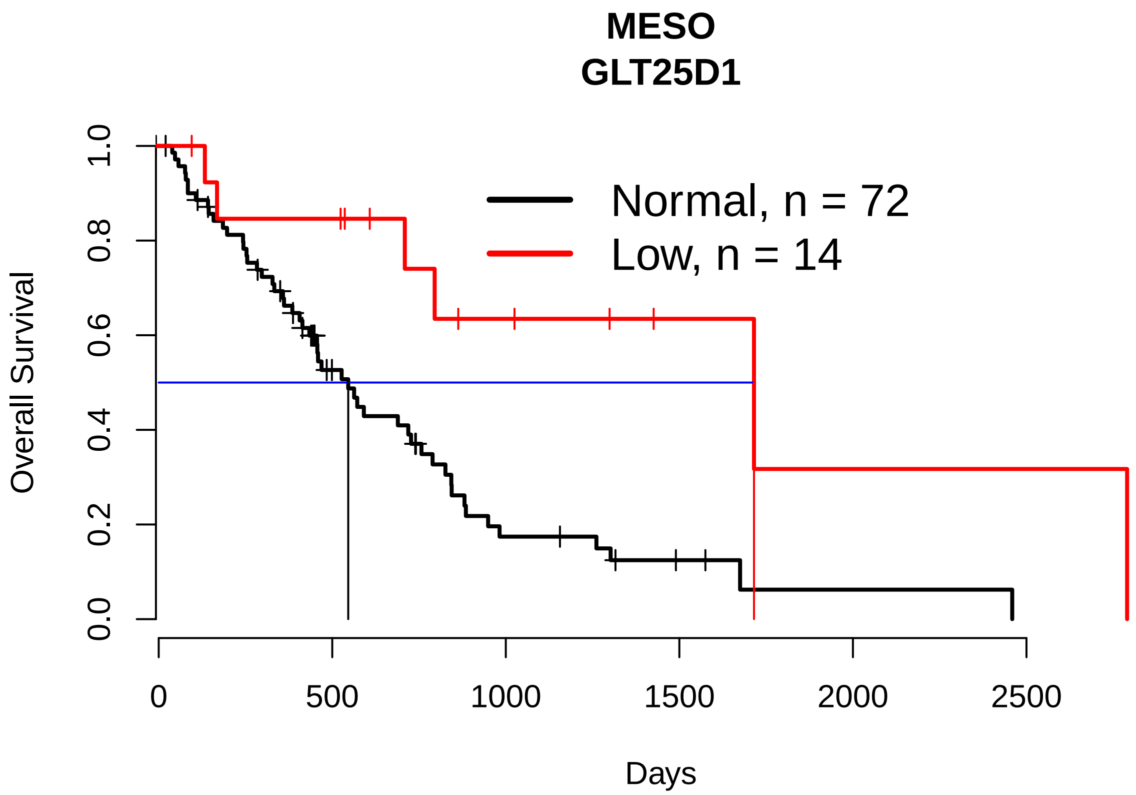
```r
all_clin$death_event <- ifelse(clinical$patient.vital_status == "alive", 0,1)

#add clinical IDs as row names
rownames(all_clin) <- clinical$IDs

#use t-score of 1 to define low expression levels
event_rna <- t(apply(z_rna, 1, function(x) ifelse(x < -1,1,0)))
ind_tum <- which(unique(colnames(z_rna)) %in% rownames(all_clin))
ind_clin <- which(rownames(all_clin) %in% colnames(z_rna))
```

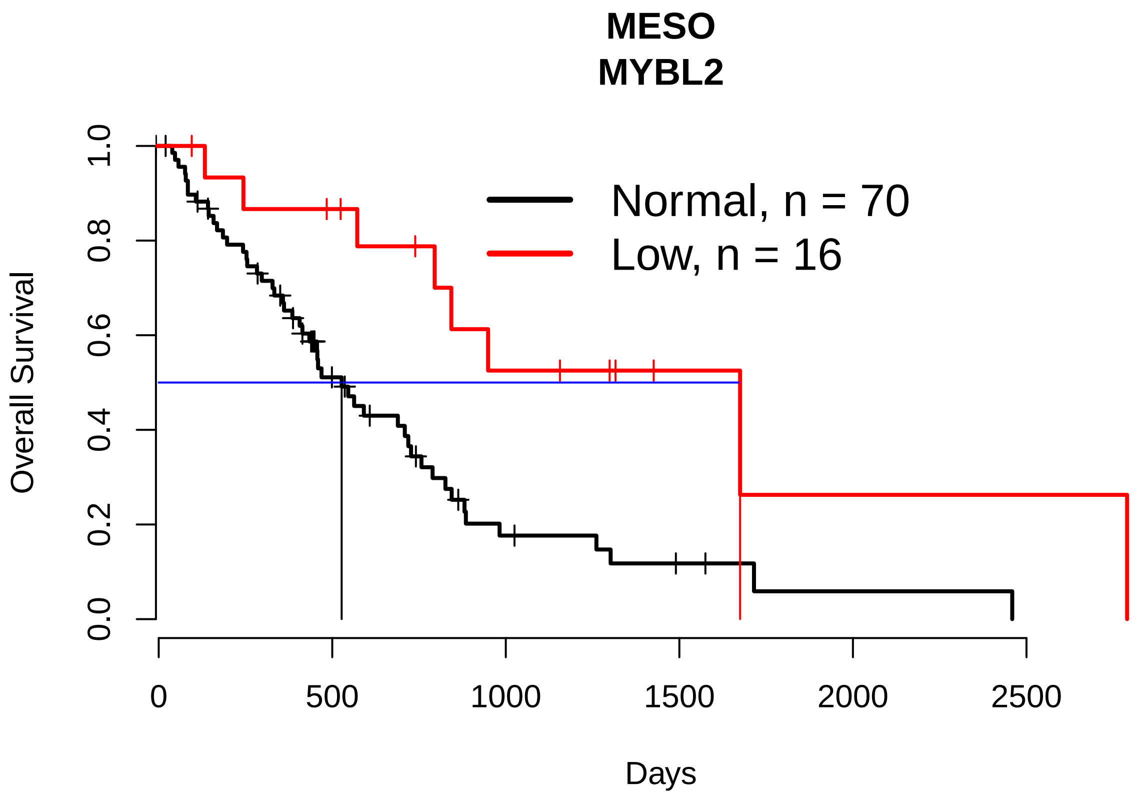
```
#create list of genes of interest
gene_list <- c("LMNB2", "KPNA2", "UHRF1", "GLT25D1","MYBL2")
gene_list_length <- length(gene_list)

#use loop to iterate through genes of interest
for (i in 1:gene_list_length){
  #define gene of interest for graph
  ind_gene <- which(rownames(z_rna) == gene_list[i])

  #check how many samples are altered
  sample_dist <- table(event_rna[ind_gene,])

  #perform survival analysis
  s <- survfit(Surv(as.numeric(as.character(all_clin$new_death))[ind_clin],all_clin$death_event[ind_cli
  s1 <- tryCatch(survdiff(Surv(as.numeric(as.character(all_clin$new_death))[ind_clin],all_clin$death_eve
  #find p-value
  p_val <- ifelse(is.na(s1),next,(round(1 - pchisq(s1$chisq, length(s1$n) - 1),3)))[[1]]
  p_val

  #graph survival curves
  plot(survfit(Surv(as.numeric(as.character(all_clin$new_death))[ind_clin],all_clin$death_event[ind_cli
       col=c(1:3), frame=F, lwd=2,main=paste("MESO",rownames(z_rna)[ind_gene],sep="\n"))
  x1 <- ifelse(is.na(as.numeric(summary(s)$table[,'median'][1])),"NA",as.numeric(summary(s)$table[,'medi
  x2 <- as.numeric(summary(s)$table[,'median'][2])
  if(x1 != "NA" & x2 != "NA"){
    lines(c(0,max(x1,x2)),c(0.5,0.5),col="blue")
    lines(c(x1,x1),c(0,0.5),col="black")
    lines(c(x2,x2),c(0,0.5),col="red")
  }

  #add legend to plot
  legend(max(as.numeric(as.character(all_clin$death_days)[ind_clin]),na.rm = T)*0.3,1,
         legend=c(paste("Normal, n =", sample_dist[1]),paste("Low, n =", sample_dist[2])),bty="n",cex=1
  title(xlab = "Days", ylab = "Overall Survival")
}
```


MESO
LMNB2
Overall Survival
0.0
0.2
0.4
0.6
0.8
1.0
Normal, n = 70
Low, n = 16
0
500
1000
1500
2000
2500
Days


MESO
KPNA2
Normal, n = 71
Low, n = 15
Overall Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Days


MESO
UHRF1
Normal, n = 71
Low, n = 15
Overall Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Days


MESO
GLT25D1
Normal, n = 72
Low, n = 14
Overall Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Days



```r
#load GEO file for study
library(GEOquery)
getGEOSuppFiles("GSE51024")
untar("./GSE51024/GSE51024_RAW.tar", exdir="./GSE51024/data")
```

```r
rm(list=ls(all=TRUE))

library(affy)
```

```
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
##
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
##
## The following object is masked from 'package:limma':
##
##     plotMA
##
## The following objects are masked from 'package:dplyr':
```

```
##
##     combine, intersect, setdiff, union
##
## The following object is masked from 'package:stats':
##
##     xtabs
##
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, as.vector, cbind,
##     colnames, do.call, duplicated, eval, evalq, Filter, Find, get,
##     intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rep.int, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unlist, unsplit
##
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
library(limma)

#load and label raw data
raw.data <- read.affybatch(list.celfiles("./GSE51024/data/", full.names=TRUE), compress=TRUE)
```

```
## Warning in read.affybatch(list.celfiles("./GSE51024/data/", full.names = TRUE), : Incompatible pheno
```

```
pData(raw.data)
```

```
##                              sample
## GSM1235016_Normal1.CEL.gz         1
## GSM1235017_Tumor1.CEL.gz          2
## GSM1235018_Normal2.CEL.gz         3
## GSM1235019_Tumor2.CEL.gz          4
## GSM1235020_Normal3.CEL.gz         5
## GSM1235021_Tumor3.CEL.gz          6
## GSM1235022_Normal4.CEL.gz         7
## GSM1235023_Tumor4.CEL.gz          8
## GSM1235024_Normal5.CEL.gz         9
## GSM1235025_Tumor5.CEL.gz         10
## GSM1235026_Normal6.CEL.gz        11
## GSM1235027_Tumor6.CEL.gz         12
## GSM1235028_Normal7.CEL.gz        13
## GSM1235029_Tumor7.CEL.gz         14
## GSM1235030_Normal8.CEL.gz        15
## GSM1235031_Tumor8.CEL.gz         16
## GSM1235032_Normal9.CEL.gz        17
## GSM1235033_Tumor9.CEL.gz         18
## GSM1235034_Normal10.CEL.gz       19
## GSM1235035_Tumor10.CEL.gz        20
## GSM1235036_Normal11.CEL.gz     21
## GSM1235037_Tumor11.CEL.gz      22
## GSM1235038_Normal12.CEL.gz     23
## GSM1235039_Tumor12.CEL.gz      24
## GSM1235040_Normal13.CEL.gz     25
## GSM1235041_Tumor13.CEL.gz      26
## GSM1235042_Normal14.CEL.gz     27
## GSM1235043_Tumor14.CEL.gz      28
## GSM1235044_Normal15.CEL.gz     29
## GSM1235045_Tumor15.CEL.gz      30
## GSM1235046_Normal16.CEL.gz     31
## GSM1235047_Tumor16.CEL.gz      32
## GSM1235048_Normal17.CEL.gz     33
## GSM1235049_Tumor17.CEL.gz      34
## GSM1235050_Normal18.CEL.gz     35
## GSM1235051_Tumor18.CEL.gz      36
## GSM1235052_Normal19.CEL.gz     37
## GSM1235053_Tumor19.CEL.gz      38
## GSM1235054_Normal20.CEL.gz     39
## GSM1235055_Tumor20.CEL.gz      40
## GSM1235056_Normal21.CEL.gz     41
## GSM1235057_Tumor21.CEL.gz      42
## GSM1235058_Normal22.CEL.gz     43
## GSM1235059_Tumor22.CEL.gz      44
## GSM1235060_Normal23.CEL.gz     45
## GSM1235061_Tumor23.CEL.gz      46
## GSM1235062_Tumor52.CEL.gz      47
## GSM1235063_Tumor53.CEL.gz      48
## GSM1235064_Normal24.CEL.gz     49
## GSM1235065_Tumor24.CEL.gz      50
## GSM1235066_Normal25.CEL.gz     51
## GSM1235067_Tumor25.CEL.gz      52
## GSM1235068_Normal26.CEL.gz     53
## GSM1235069_Tumor26.CEL.gz      54
## GSM1235070_Normal27.CEL.gz     55
## GSM1235071_Tumor27.CEL.gz      56
## GSM1235072_Normal28.CEL.gz     57
## GSM1235073_Tumor28.CEL.gz      58
## GSM1235074_Tumor54.CEL.gz      59
## GSM1235075_Tumor55.CEL.gz      60
## GSM1235076_Normal29.CEL.gz     61
## GSM1235077_Tumor29.CEL.gz      62
## GSM1235078_Normal30.CEL.gz     63
## GSM1235079_Tumor30.CEL.gz      64
## GSM1235080_Normal31.CEL.gz     65
## GSM1235081_Tumor31.CEL.gz      66
## GSM1235082_Normal32.CEL.gz     67
## GSM1235083_Tumor32.CEL.gz      68
## GSM1235084_Tumor42.CEL.gz      69
## GSM1235085_Tumor43.CEL.gz      70
## GSM1235086_Normal33.CEL.gz     71
## GSM1235087_Tumor33.CEL.gz      72
## GSM1235088_Normal34.CEL.gz     73
## GSM1235089_Tumor34.CEL.gz      74
## GSM1235090_Normal35.CEL.gz     75
## GSM1235091_Tumor35.CEL.gz      76
## GSM1235092_Tumor44.CEL.gz      77
## GSM1235093_Tumor45.CEL.gz      78
## GSM1235094_Tumor46.CEL.gz      79
## GSM1235095_Tumor47.CEL.gz      80
## GSM1235096_Tumor48.CEL.gz      81
## GSM1235097_Tumor49.CEL.gz      82
## GSM1235098_Tumor50.CEL.gz      83
## GSM1235099_Tumor51.CEL.gz      84
## GSM1235100_Normal36.CEL.gz     85
## GSM1235101_Tumor36.CEL.gz      86
## GSM1235102_Normal37.CEL.gz     87
## GSM1235103_Tumor37.CEL.gz      88
## GSM1235104_Normal38.CEL.gz     89
## GSM1235105_Tumor38.CEL.gz      90
## GSM1235106_Normal39.CEL.gz     91
## GSM1235107_Tumor39.CEL.gz      92
## GSM1235108_Normal40.CEL.gz     93
## GSM1235109_Tumor40.CEL.gz      94
## GSM1235110_Normal41.CEL.gz     95
## GSM1235111_Tumor41.CEL.gz      96
```

```
pData(raw.data)$status <- c(rep(c("Normal","Tumor"),23),"Tumor","Tumor",rep(c("Normal","Tumor"), 5),"Tu

pData(raw.data)
```

```
##                            sample status
## GSM1235016_Normal1.CEL.gz       1 Normal
## GSM1235017_Tumor1.CEL.gz        2  Tumor
## GSM1235018_Normal2.CEL.gz       3 Normal
## GSM1235019_Tumor2.CEL.gz        4  Tumor
## GSM1235020_Normal3.CEL.gz       5 Normal
## GSM1235021_Tumor3.CEL.gz        6  Tumor
## GSM1235022_Normal4.CEL.gz       7 Normal
## GSM1235023_Tumor4.CEL.gz        8  Tumor
## GSM1235024_Normal5.CEL.gz       9 Normal
## GSM1235025_Tumor5.CEL.gz       10  Tumor
## GSM1235026_Normal6.CEL.gz      11 Normal
## GSM1235027_Tumor6.CEL.gz       12  Tumor
## GSM1235028_Normal7.CEL.gz      13 Normal
## GSM1235029_Tumor7.CEL.gz       14  Tumor
## GSM1235030_Normal8.CEL.gz      15 Normal
## GSM1235031_Tumor8.CEL.gz       16  Tumor
## GSM1235032_Normal9.CEL.gz      17 Normal
## GSM1235033_Tumor9.CEL.gz       18  Tumor
## GSM1235034_Normal10.CEL.gz     19 Normal
## GSM1235035_Tumor10.CEL.gz      20  Tumor
## GSM1235036_Normal11.CEL.gz     21 Normal
## GSM1235037_Tumor11.CEL.gz      22  Tumor
## GSM1235038_Normal12.CEL.gz     23 Normal
## GSM1235039_Tumor12.CEL.gz      24  Tumor
## GSM1235040_Normal13.CEL.gz     25 Normal
## GSM1235041_Tumor13.CEL.gz      26  Tumor
## GSM1235042_Normal14.CEL.gz     27 Normal
## GSM1235043_Tumor14.CEL.gz      28  Tumor
## GSM1235044_Normal15.CEL.gz     29 Normal
## GSM1235045_Tumor15.CEL.gz      30  Tumor
## GSM1235046_Normal16.CEL.gz     31 Normal
## GSM1235047_Tumor16.CEL.gz      32  Tumor
## GSM1235048_Normal17.CEL.gz     33 Normal
## GSM1235049_Tumor17.CEL.gz      34  Tumor
## GSM1235050_Normal18.CEL.gz     35 Normal
## GSM1235051_Tumor18.CEL.gz      36  Tumor
## GSM1235052_Normal19.CEL.gz     37 Normal
## GSM1235053_Tumor19.CEL.gz      38  Tumor
## GSM1235054_Normal20.CEL.gz     39 Normal
## GSM1235055_Tumor20.CEL.gz      40  Tumor
## GSM1235056_Normal21.CEL.gz     41 Normal
## GSM1235057_Tumor21.CEL.gz      42  Tumor
## GSM1235058_Normal22.CEL.gz     43 Normal
## GSM1235059_Tumor22.CEL.gz      44  Tumor
## GSM1235060_Normal23.CEL.gz     45 Normal
## GSM1235061_Tumor23.CEL.gz      46  Tumor
## GSM1235062_Tumor52.CEL.gz      47  Tumor
## GSM1235063_Tumor53.CEL.gz      48  Tumor
## GSM1235064_Normal24.CEL.gz     49 Normal
## GSM1235065_Tumor24.CEL.gz      50  Tumor
## GSM1235066_Normal25.CEL.gz     51 Normal
## GSM1235067_Tumor25.CEL.gz      52  Tumor
## GSM1235068_Normal26.CEL.gz     53 Normal
## GSM1235069_Tumor26.CEL.gz      54  Tumor
## GSM1235070_Normal27.CEL.gz     55 Normal
## GSM1235071_Tumor27.CEL.gz      56  Tumor
## GSM1235072_Normal28.CEL.gz     57 Normal
## GSM1235073_Tumor28.CEL.gz      58  Tumor
## GSM1235074_Tumor54.CEL.gz      59  Tumor
## GSM1235075_Tumor55.CEL.gz      60  Tumor
## GSM1235076_Normal29.CEL.gz     61 Normal
## GSM1235077_Tumor29.CEL.gz      62  Tumor
## GSM1235078_Normal30.CEL.gz     63 Normal
## GSM1235079_Tumor30.CEL.gz      64  Tumor
## GSM1235080_Normal31.CEL.gz     65 Normal
## GSM1235081_Tumor31.CEL.gz      66  Tumor
## GSM1235082_Normal32.CEL.gz     67 Normal
## GSM1235083_Tumor32.CEL.gz      68  Tumor
## GSM1235084_Tumor42.CEL.gz      69  Tumor
## GSM1235085_Tumor43.CEL.gz      70  Tumor
## GSM1235086_Normal33.CEL.gz     71 Normal
## GSM1235087_Tumor33.CEL.gz      72  Tumor
## GSM1235088_Normal34.CEL.gz     73 Normal
## GSM1235089_Tumor34.CEL.gz      74  Tumor
## GSM1235090_Normal35.CEL.gz     75 Normal
## GSM1235091_Tumor35.CEL.gz      76  Tumor
## GSM1235092_Tumor44.CEL.gz      77  Tumor
## GSM1235093_Tumor45.CEL.gz      78  Tumor
## GSM1235094_Tumor46.CEL.gz      79  Tumor
## GSM1235095_Tumor47.CEL.gz      80  Tumor
## GSM1235096_Tumor48.CEL.gz      81  Tumor
## GSM1235097_Tumor49.CEL.gz      82  Tumor
## GSM1235098_Tumor50.CEL.gz      83  Tumor
## GSM1235099_Tumor51.CEL.gz      84  Tumor
## GSM1235100_Normal36.CEL.gz     85 Normal
## GSM1235101_Tumor36.CEL.gz      86  Tumor
## GSM1235102_Normal37.CEL.gz     87 Normal
## GSM1235103_Tumor37.CEL.gz      88  Tumor
## GSM1235104_Normal38.CEL.gz     89 Normal
## GSM1235105_Tumor38.CEL.gz      90  Tumor
## GSM1235106_Normal39.CEL.gz     91 Normal
## GSM1235107_Tumor39.CEL.gz      92  Tumor
## GSM1235108_Normal40.CEL.gz     93 Normal
## GSM1235109_Tumor40.CEL.gz      94  Tumor
## GSM1235110_Normal41.CEL.gz     95 Normal
## GSM1235111_Tumor41.CEL.gz      96  Tumor
```

```
#boxplot(exprs(raw.data), col="red",main="Raw Probe Intensities")
#boxplot(raw.data, col="red",main="Raw Probe Intensities")

#quality control
GSE51024.rma <- rma(raw.data)
```

```
## Creating a generic function for 'nchar' from package 'base' in package 'S4Vectors'

## Background correcting
## Normalizing
## Calculating Expression
```

```
#boxplot(exprs(GSE51024.rma), col="blue", main="RMA Expression Values")

GSE51024.qc <- raw.data[, !sampleNames(raw.data) %in% c("GSM1235084_Tumor42.CEL.gz", "GSM1235085_Tumor43
                                                        "GSM1235092_Tumor44.CEL.gz", "GSM1235093_Tumor45
                                                        "GSM1235094_Tumor46.CEL.gz", "GSM1235095_Tumor47
                                                        "GSM1235096_Tumor48.CEL.gz", "GSM1235097_Tumor49
                                                        "GSM1235098_Tumor50.CEL.gz", "GSM1235099_Tumor51
                                                        "GSM1235062_Tumor52.CEL.gz", "GSM1235063_Tumor53
                                                        "GSM1235074_Tumor54.CEL.gz", "GSM1235075_Tumor55



#use sibship pairs for paired study design
pData(GSE51024.qc)$SibShip <- ceiling(c(1:82)/2)
GSE51024.qc.rma <- rma(GSE51024.qc)
```

```
## Background correcting
## Normalizing
## Calculating Expression
```

```
SibShip <- factor(pData(GSE51024.qc.rma)$SibShip)
Status <- factor(pData(GSE51024.qc.rma)$status, levels = c("Normal","Tumor"))

#design incorporates sipship and status variables
```

```
design <- model.matrix(~-1+SibShip+Status)

fit <- lmFit(GSE51024.qc.rma, design)

#adjust fit coefficients using empirical Bayes moderation of standard errors
fit2 <- eBayes(fit)

#output hypothesis test results
Tumor_results <- topTable(fit2, coef ="StatusTumor" , adjust="BH", num=100, p.value=0.05)
head(Tumor_results)
```

```
##                  logFC   AveExpr         t      P.Value    adj.P.Val
## 212814_at   -2.209248 10.514953 -23.94835 2.795674e-26 1.528535e-21
## 202893_at   -2.040677  9.226271 -23.41406 6.819620e-26 1.864313e-21
## 206069_s_at -2.178871  7.301401 -21.72628 1.284909e-24 2.341747e-20
## 204388_s_at -3.257243  9.155738 -21.02183 4.633603e-24 6.333556e-20
## 212741_at   -2.944609 10.090108 -20.23096 2.041929e-23 2.232850e-19
## 226228_at   -6.153408  8.820715 -20.02041 3.054937e-23 2.783811e-19
##                    B
## 212814_at   49.17071
## 202893_at   48.33661
## 206069_s_at 45.57075
## 204388_s_at 44.35371
## 212741_at   42.94032
## 226228_at   42.55533
```

```
#load reference gene IDs
library(hgu133plus2.db)
```

```
## Loading required package: AnnotationDbi
## Loading required package: stats4
## Loading required package: GenomeInfoDb
## Loading required package: S4Vectors
##
## Attaching package: 'S4Vectors'
##
## The following object is masked from 'package:dplyr':
##
##     rename
##
## Loading required package: IRanges
##
## Attaching package: 'IRanges'
##
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
##
##
## Attaching package: 'AnnotationDbi'
##
## The following object is masked from 'package:dplyr':
##
```

```
##     select
##
## Loading required package: org.Hs.eg.db
## Loading required package: DBI
```

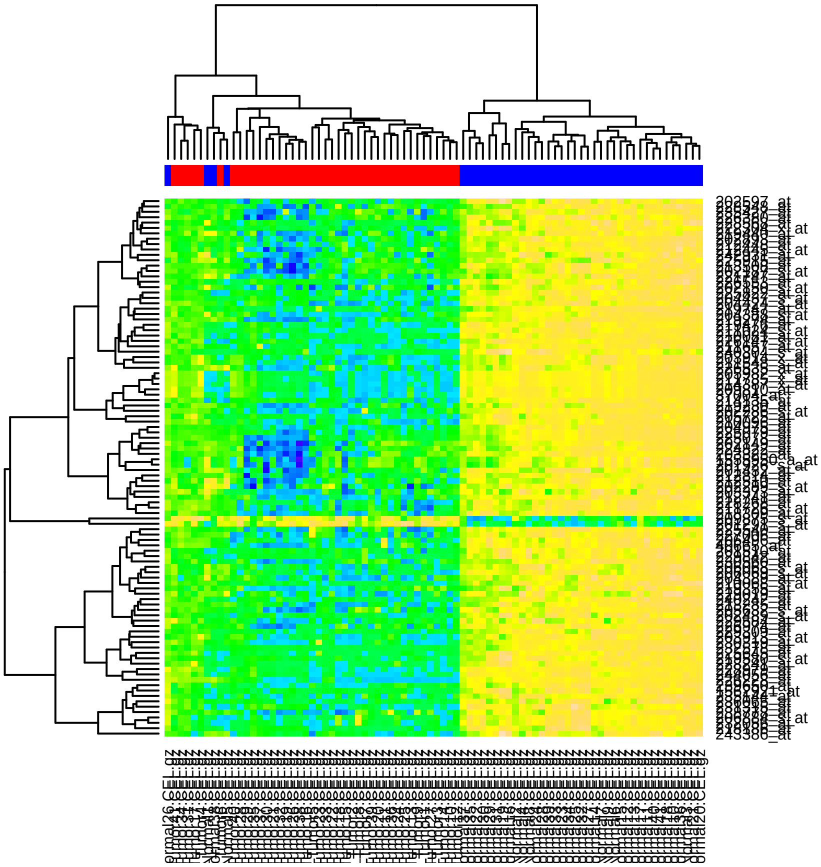
```
Tumor_results$ID = row.names(Tumor_results)
Tumor_results$SYMBOL <- lapply(Tumor_results$ID, function(x) mget(x, env=hgu133plus2SYMBOL, ifnotfound=
head(Tumor_results)
```

```
##                 logFC   AveExpr         t      P.Value    adj.P.Val
## 212814_at   -2.209248 10.514953 -23.94835 2.795674e-26 1.528535e-21
## 202893_at   -2.040677  9.226271 -23.41406 6.819620e-26 1.864313e-21
## 206069_s_at -2.178871  7.301401 -21.72628 1.284909e-24 2.341747e-20
## 204388_s_at -3.257243  9.155738 -21.02183 4.633603e-24 6.333556e-20
## 212741_at   -2.944609 10.090108 -20.23096 2.041929e-23 2.232850e-19
## 226228_at   -6.153408  8.820715 -20.02041 3.054937e-23 2.783811e-19
##                    B          ID SYMBOL
## 212814_at   49.17071   212814_at AHCYL2
## 202893_at   48.33661   202893_at UNC13B
## 206069_s_at 45.57075 206069_s_at  ACADL
## 204388_s_at 44.35371 204388_s_at   MAOA
## 212741_at   42.94032   212741_at   MAOA
## 226228_at   42.55533   226228_at   AQP4
```

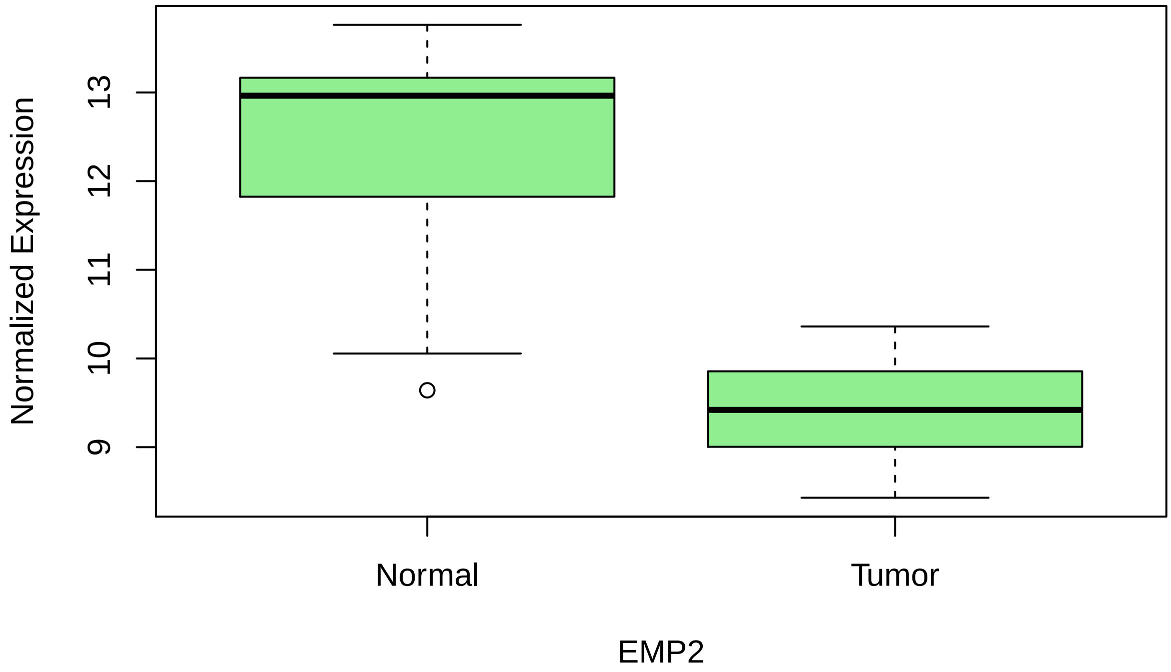
```
cat('There are', nrow(Tumor_results), 'significant probes.')
```

```
## There are 100 significant probes.
```

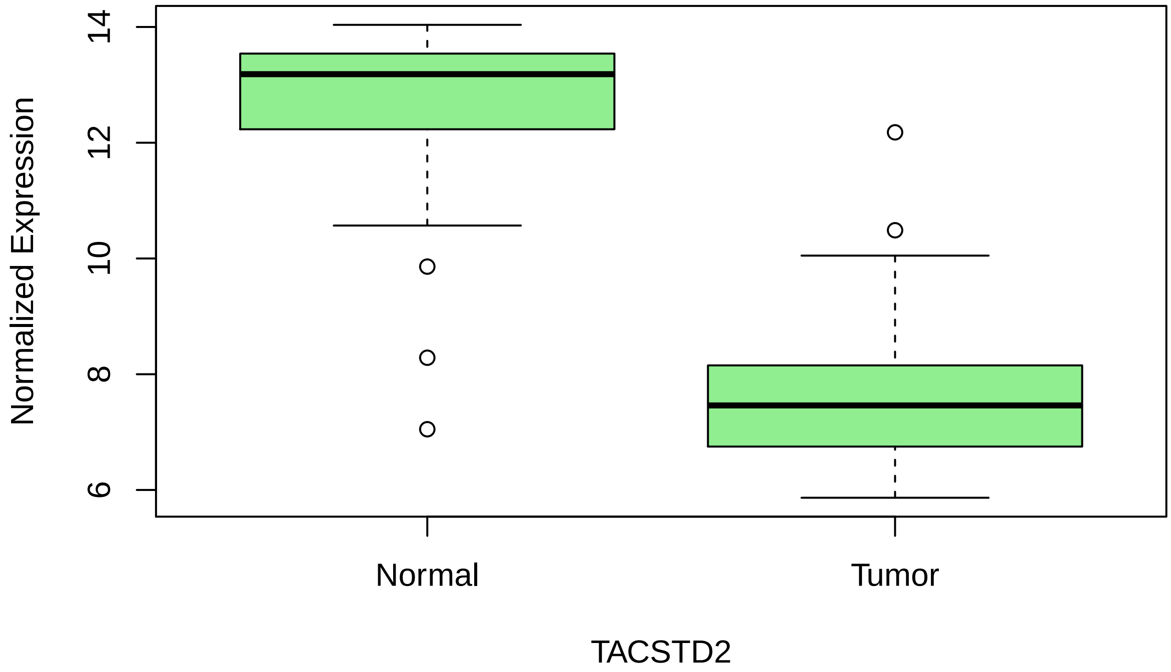
```
#output heatmap with status label at top of plot
top.eset <- GSE51024.qc.rma[row.names(exprs(GSE51024.qc.rma)) %in% row.names(Tumor_results)]
treatment.colors <- unlist(lapply(GSE51024.qc.rma$status, function(x){if (x=="Tumor") "red" else "blue"
heatmap(exprs(top.eset), col=topo.colors(100), ColSideColors=treatment.colors)
```



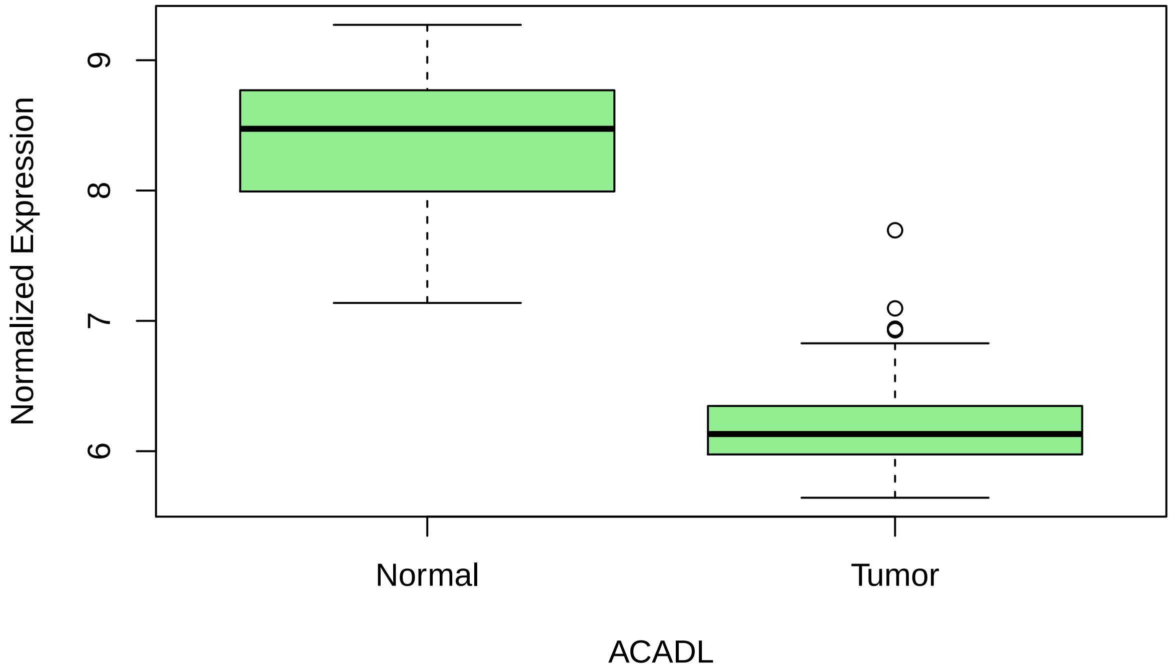
```
#output boxplots of expression values
boxplot(exprs(GSE51024.qc.rma)["204975_at" , ]~ GSE51024.qc.rma$status, col="lightgreen", xlab='EMP2', 
```



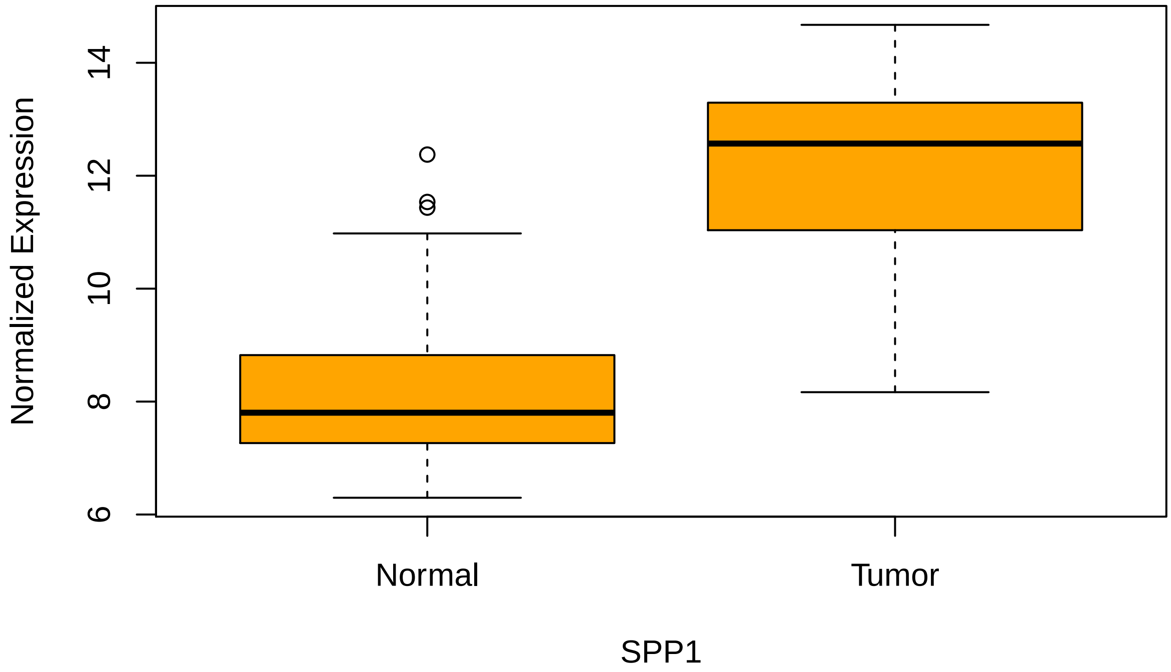
```
boxplot(exprs(GSE51024.qc.rma)["202286_s_at" , ]~ GSE51024.qc.rma$status, col="lightgreen", xlab='TACSTI
```



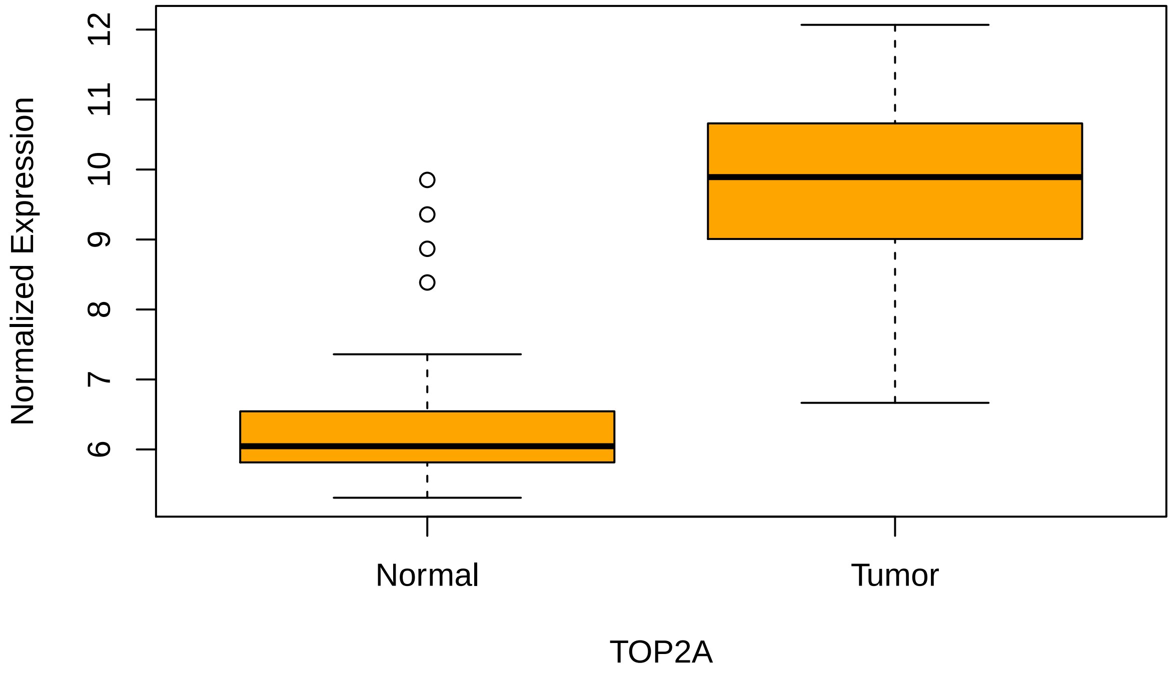
```
boxplot(exprs(GSE51024.qc.rma)["206069_s_at" , ]~ GSE51024.qc.rma$status, col="lightgreen", xlab='ACADL
```



```
boxplot(exprs(GSE51024.qc.rma)["209875_s_at" , ]~ GSE51024.qc.rma$status, col="orange", xlab='SPP1', yl
```



```
boxplot(exprs(GSE51024.qc.rma)["201291_s_at" , ]~ GSE51024.qc.rma$status, col="orange", xlab='TOP2A', yl
```



```
#for obtaining gene names for probe
Tumor_results$SYMBOL[Tumor_results$ID=='201291_s_at']
head(Tumor_results)


#for outputting list for network analysis of significant genes and processing
l_results <- (Tumor_results$SYMBOL)
l_results <- l_results[!is.na(l_results)]
l_results <- unique(l_results)
length(l_results)
lapply(l_results, write, "GEO_genes.txt", append=TRUE, ncolumns=100)
```

Results

The GEO dataset GSE51024 was used in this study to observe and classify gene expression differences between normal cell lines and malignant mesothelioma cells. After cleaning, this dataset contained 41 pairs of samples that corresponded to normal and tumor tissue for a given patient. The subset of top differentially expressed genes was obtained as described above. For preliminary analysis, this subset was limited to the top 100 differentially expressed genes. An output of this resultant heatmap demonstrating hierarchical clustering is shown in Figure 1.

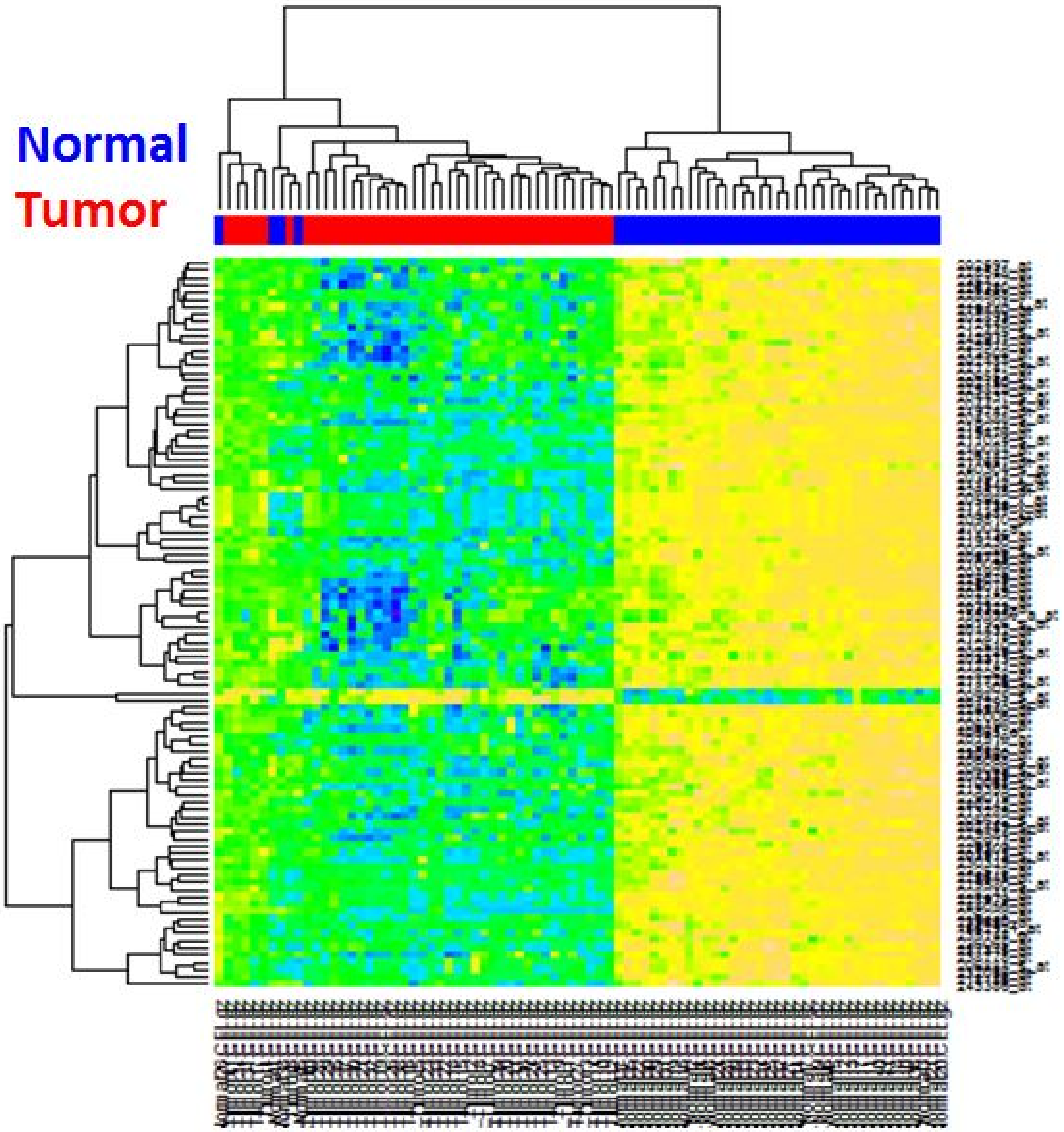


Figure 1: A heatmap demonstrating differential gene expression and clustering of normal and tumor samples.

In this figure, the tissue samples that are designated normal are annotated in blue at the top of the heatmap, and those that are tumor samples are shown in red. The samples fall roughly into two clusters, with a small subset of normal samples being interspersed within the tumor cluster, which suggests that using these top genes for these clustering purposes was not sufficient information to perfectly cluster the samples or, perhaps, there are some normal samples that display tumor-like characteristics in their gene expression profile. It is also interesting to note that for the vast majority of differentially expressed genes, most were more highly expressed in the normal samples than the tumor samples. Of the top 100 probes, 98 of them demonstrated higher expression levels in normal than tumor, and only 2 of them demonstrated higher levels in tumor than normal. This result is somewhat surprising-though not completely unexpected. Given that there are two main types of genes that one may expect differentially expressed in cancer cells (oncogenes and tumor-suppressor genes), and that these gene types behave differently, this result is not implausible. For an oncogene (cancer-promoting), it would be expected that levels of expression would be higher in tumor than normal; for a tumor suppressor gene the opposite would be expected (higher levels of expression in normal than tumor). The boxplots of these top two differentially expressed genes with higher levels in tumor samples and a subset of three of the top differentially expressed genes with higher levels in normal than tumor samples are displayed in Figure 2 to check for outliers as well as overlap between the two distributions. This figure suggests that of these genes of interest, the overlap between the interquartile range is negligible; however, there are a small number of outliers in each case.

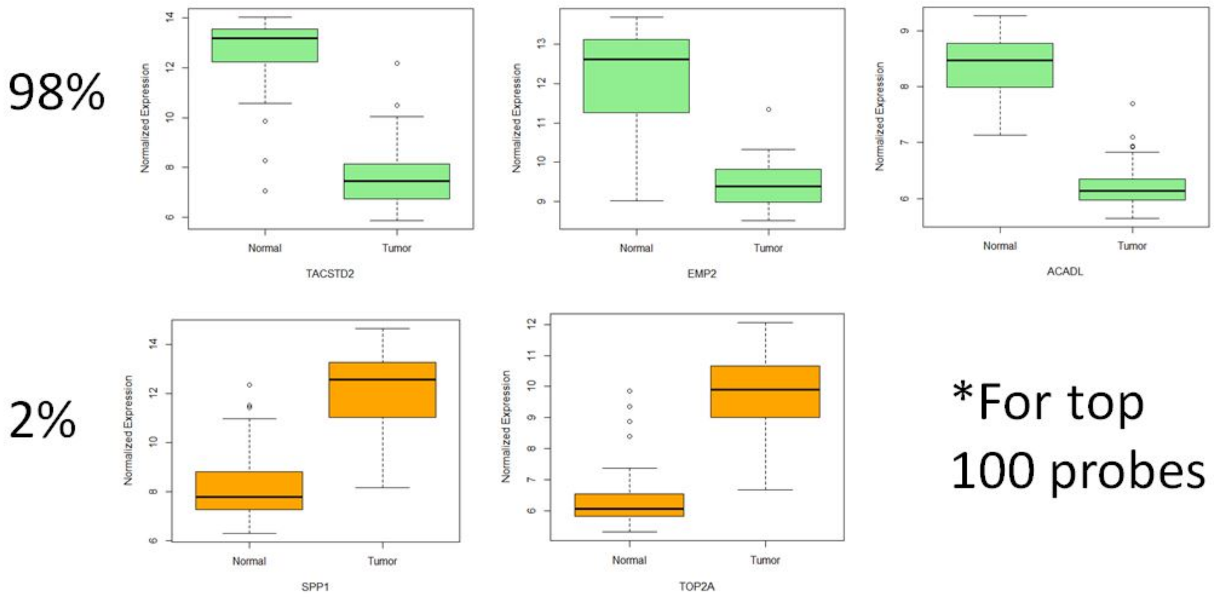


Figure 2: A selected subset of differentially expressed genes analyzed via boxplot.

Additionally, it was of interest to reference these top genes within the recent cancer literature to make sense of these differential expression patterns in the context of correlation and mechanism within other cancers or related diseases. TACSTD2, one of the top ten differentially expressed genes, was found to have contributed to squamous cell carcinoma progression when a comparative analysis was performed in both normal and squamous cell carcinoma tissues. Accompanying data demonstrated that Tacstd2 expression and membrane localization were tightly associated with stratified with stratified epithelial homeostasis. In samples in which there was demonstrable loss of TACSTD2, it was found that these samples were often poorly differentiated SCC tissues collected from the cervix, esophagus, head, and neck. This study supported that loss of TACSTD2 could potentially promote SCC progression and treatment resistance [4]. One of the two top differentially expressed genes with oncogene-type expression, SPP1, also has a body of work motivating its importance in the study of cancer. In one such study, it was found that SPP1 (secreted phosphoprotein-1) promoted cancer cell survival and regulated tumor-associated angiogenesis and inflammation, which are both inherent mechanisms to the pathogenesis of malignant pleural effusion, a condition in which cancer causes an abnormal amount of fluid to collect within the pleura [5]. Again, these results are consistent with the findings of this study, in that one would expect increased levels of SPP1 in tumor-cells due to its nature as an oncogene. While these two results are supported by the body of literature in other cancer types, there are other cases in which the literature is in apparent disagreement with these results. In the case of EMP2-again one of the top ten differentially expressed genes identified by this approach-there are clearly lowered levels of EMP2 in the tumor sample when compared to normal. This result would suggest that one may expect to find literature implicating EMP2 as a tumor suppressor gene or verifying lowered expression levels in tumorous tissue. However, the vast majority of the current literature for other cancers suggests the opposite. In one such work, it was found that EMP2 was significantly increased in expression levels in bladder tumor tissue compared to normal adjacent tissue, and identified it as a potential target for bladder cancer immunological therapy [6]. In an unrelated study in glioblastoma multiforme (GBM), it was found that EMP2 expression was significantly associated with activated Src kinase in patient samples and promoted tumor cell invasion in intracranial mouse models [7]. These sources provide several instances in which EMP2 levels were either more highly expressed in tumorous tissue or were implicated in a pro-metastatic outcome that contributes to cancer progression. Given these stark differences in findings, it is clear that further study is necessary regarding the specific actions of the EMP2 gene in mesothelioma. Since neither of the referenced studies dealt with malignant mesothelioma specifically, it is possible that the EMP2 gene behaves differently in mesothelioma than in glioblastoma or bladder cancer. Given that EMP2 has been associated with a variety of functions including endocytosis, cell signaling, cell proliferation, cell, migration, cell adhesion, cell death, and others, it is possible that the gene be significantly differentially expressed in either direction to yield oncogenic outcomes. Regardless, further study is justified on the role of EMP2 in mesothelioma, and studies that answer the outcomes associated with both high and low expression of EMP2 would be of interest. One potential criticism here that has not gone unnoticed is arbitrarily restricting the top genes dataset to the top 100 genes and using that for clustering. For this reason, additional clustering analyses using heatmaps were performed for the top 10 and top 1000 probes, which represent a decrease and an increase of one order of magnitude respectively. In both cases, the impact on the characteristics of the clusters relative to the case of the top 100 genes was negligible. These additional heatmaps are shown in Figure 3.

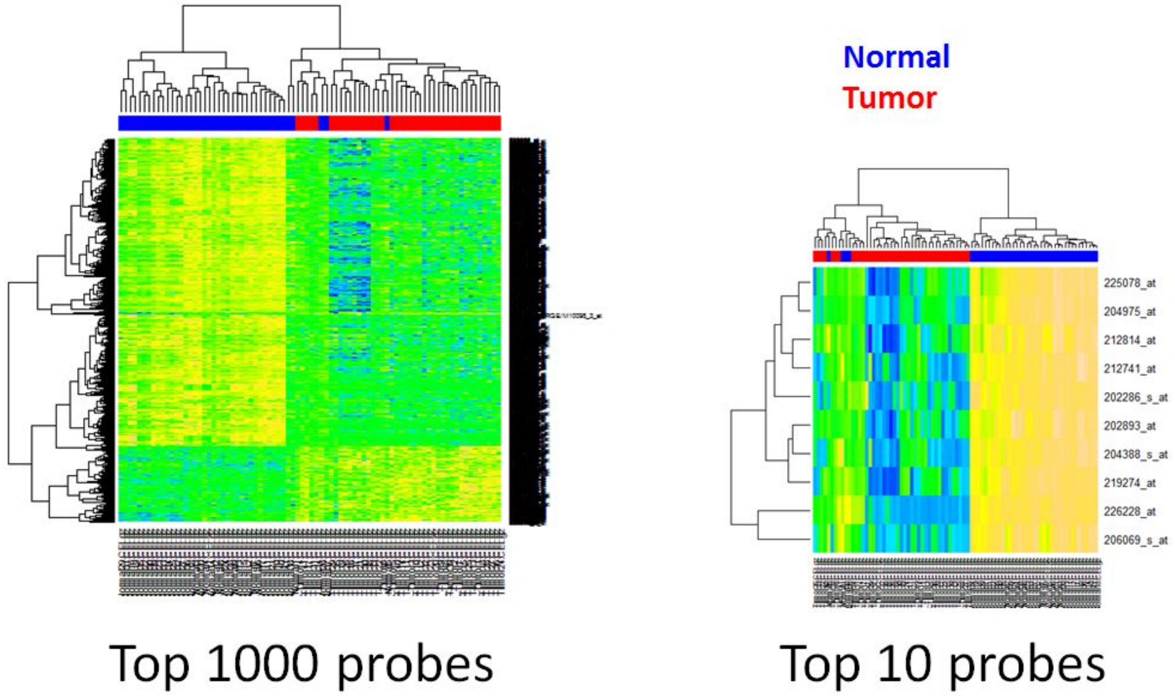


Figure 3: Heatmaps representing clustering using 10 and 1000 probes.

While searching for individual genes that were outputted as the top significantly expressed genes in the GEO dataset yields interesting information about these genes and may suggest potential therapeutic approaches, this method is rather time-consuming and approaches the problem without using all of the available tools that have been previously developed. Additionally, recent approaches to molecular biology and the study of cancer have shifted to more of a network view of interactions and complexes. Network analysis has been used previously to elucidate common pathways and genes of interest based on their connectivity to other genes. In this work, network analysis was used to accelerate this aspect of literature study and make sense of these complex gene interactions in a way that may not be easily visualized by looking up genes one-by-one. Building off of the previous GEO analysis, the top 100 probes were cross-referenced to their target genes, cleaned of non-matches and duplicates, and narrowed down to a list of 72 genes. Agilent Literature Search was used to generate a network that ignored unconnected genes and was cleaned of extraneous connections. This resultant network is shown in Figure 4.

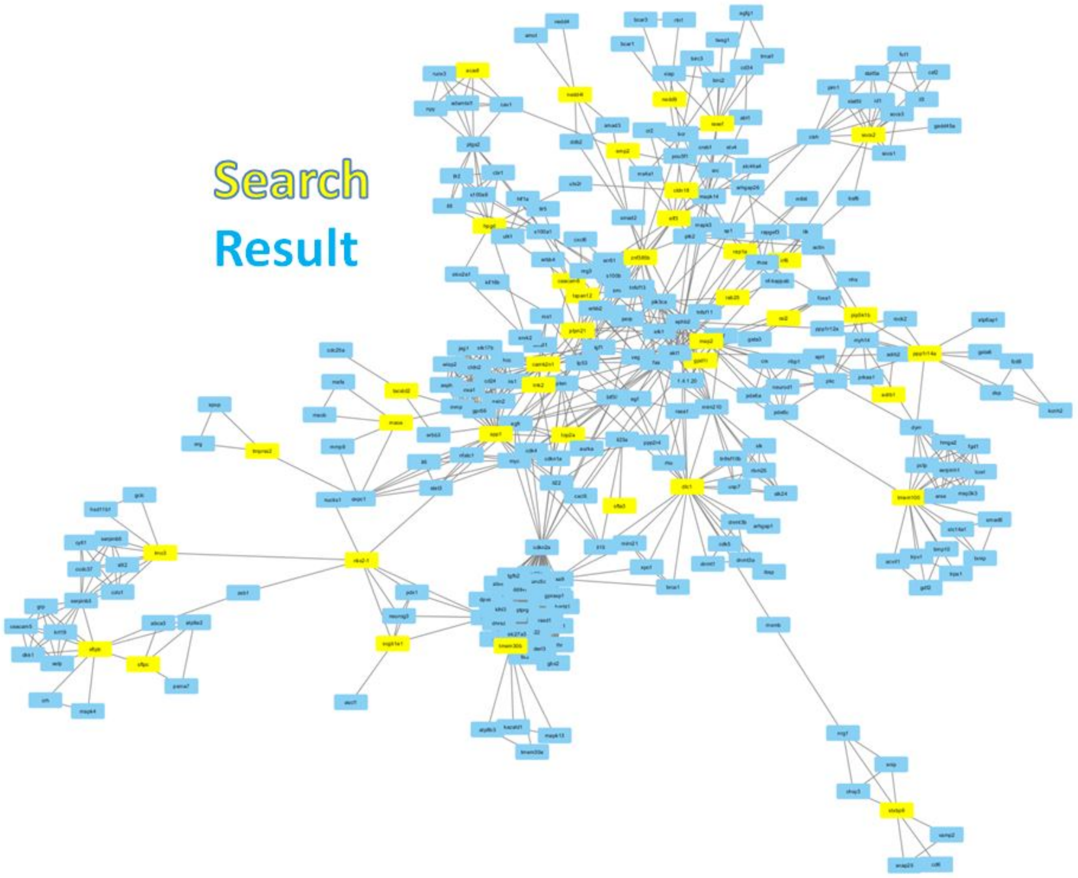


Figure 4: Preliminary network analysis generated from Agilent Literature Search.

This procedure differentiates between “search” genes-those that were outputted by the GEO analysis and used in the search-and “result” genes that were generated by the search. It is noteworthy that approximately 40 of the 72 input genes were part of this main branched network, which is an interesting result, though not entirely surprising due to the multitude of interactions any one of those complexes could be involved in and the interconnectivity of current models of cell signaling. It would be difficult, however, to generate further insights from the network in Figure 4 without further refinement due to the complexity of the situation. While there are valid concerns about losing information as components of the network are discarded and separated, it is somewhat of a necessity in this case due to its unwieldly nature. In this subsequent analysis, the network was broken down into highly connected hubs of result genes as described above. There are two characteristics of importance in this resultant smaller network as displayed in Figure 5.

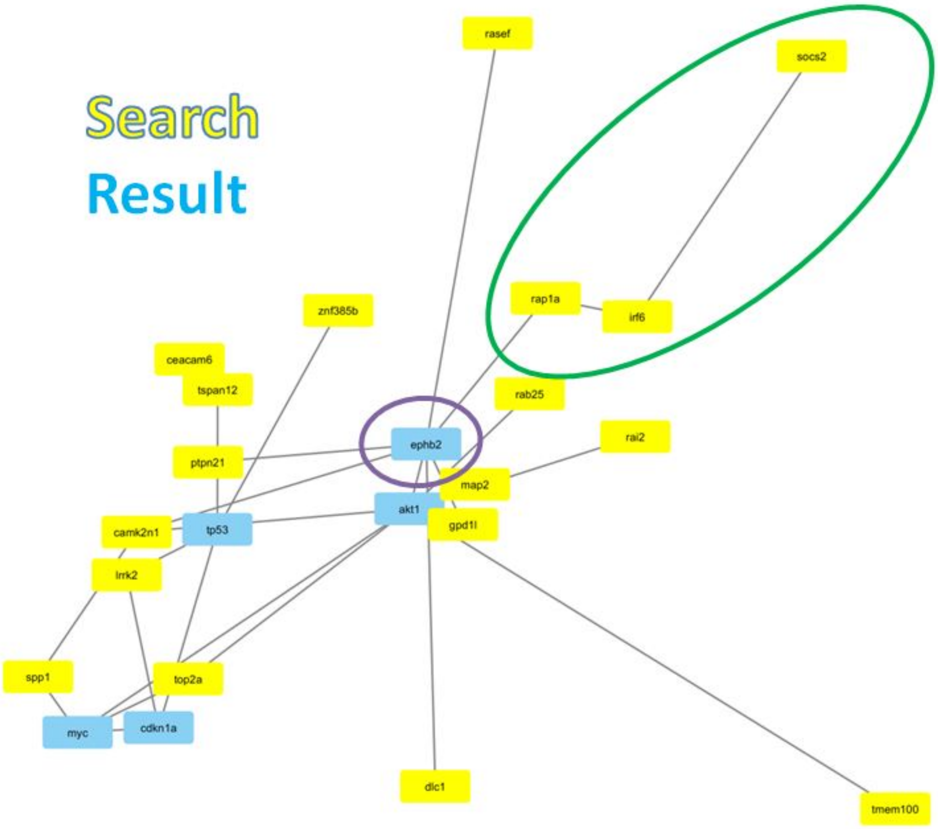


Figure 5: Refined network analysis generated from Agilent Literature Search.

The first is those highly connected hub genes that are blue, representing results. These genes were not part of the top 100 significant DE genes according to the GEO analysis, yet are implicated in several of these interactions with genes that are DE in some sort of human context (either patient or tissue sample). The second is those lines of connected search genes in which they form a line with 3 or more search genes. This is suggestive of a signaling cascade in which one gene may regulate another gene, which regulates a third gene (and thus aberrant expression of the first gene leads to aberrant expression of the other two). While such signaling cascades that demonstrate DE levels may contribute to cancerous processes, it is important to note the functions of the genes involved. In this refined network, the centrally located hub gene EphB2 leads to several branches, some of which demonstrate signaling cascade-like appearances. While EphB2 did not appear in the list of top 100 DE genes, it has been heavily implicated in other cancers in the literature. In one such study, it was shown that EphB2 promoted cervical cancer progression by inducing epithelial-mesenchymal transition (EMT) [8]. In another study, EphB2 was targeted using a conjugated antibody-drug target approach for the treatment of colorectal cancer [9]. While both of these studies focused on EphB2's role in other cancers, its high degree of connectivity to differentially expressed tumor genes in conjunction with the previous literature showing that it promotes cervical cancer progression warrant further study of this gene in the context of mesothelioma specifically.

While the previous GEO and network analyses have supported the idea that there are a subset of genes that are significantly differentially expressed in the disease mesothelioma and that these genes have substantial relevance in previous related cancer literature, one weakness of the dataset is that it does not supply relevant clinical information associated with the gene expression profiles. In previous works, a number of microarray-based gene expression models have been proposed to predict patient clinical outcomes. Given the substantial differences in the gene expression profiles between mesothelioma and normal patients, it is clear that these markers can be used to differentiate patient and normal with some degree of accuracy. What is less obvious is whether these markers can be used to divide mesothelioma tumor samples into grades and whether gene expression profiles correspond to clinical outcomes. To answer these questions, another dataset was needed that incorporated clinical outcome elements for a given patient with some form of genomic data. While TCGA hosts such data directly, many of the preliminary statistical analyses and cleaning had already been automated using MIT's Broad Institute FireBrowse pipeline. Thus, cleaned data was downloaded directly from FireBrowse, with the statistically significant findings used as a starting point for further analysis and visualization in R. Of interest was the variable that the Broad Institute referred to as "DAYS_TO_DEATH_OR_LAST_FUP." This variable represents the number of days past initial diagnosis that the patient lived until either they became deceased or no longer returned for follow-up appointments in the system. While grouping the two variables together makes the dataset more complete by providing information for all patients, it also introduces the potential complication of added noise due to additional clinical information that likely has nothing to do with the gene of interest. As long as the reasonable assumption is made that genetic profile of the chosen genes does not influence a patient to no longer return for follow-up appointments, there is no reason to believe its addition would bias the results. FireBrowse was utilized in order to find the top genes associated with the "DAYS_TO_DEATH_OR_LAST_FUP" variable, and a total of 30 genes were revealed. The top 5 genes were drawn from FireBrowse and subjected to this subsequent analysis. This analysis is shown in Figure 6.

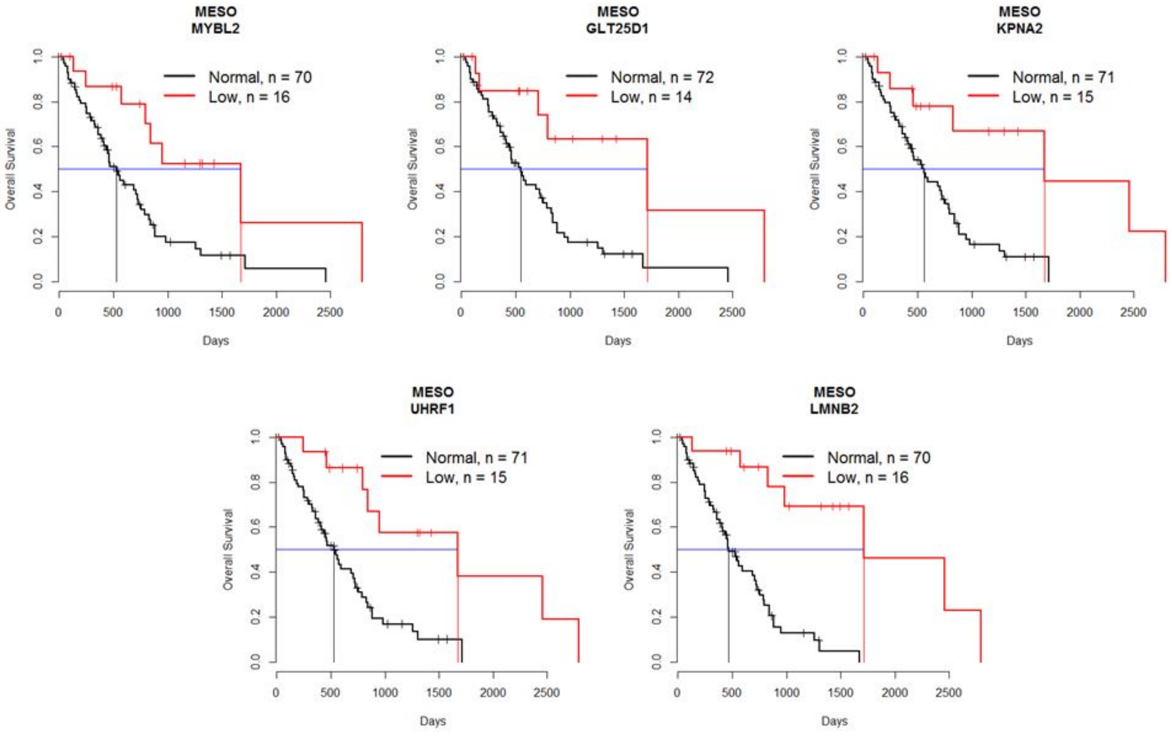


Figure 6: Survival analysis using 5 of the top DE genes from FireBrowse analysis.

This analysis consisted of differentiation between normal and abnormal gene expression, where abnormal corresponds to a normalized intensity value with a t-value of less than -1. Thus no clinically relevant threshold was defined. However, in all cases, a clear trend was demonstrated that patients with normalized lower expressions of these gene values tended to live longer on average. This is particularly evident when looking at the points at which 50% of patients were deceased, and observing that on average patients with lower gene expression values tended to live, on average, nearly three times as long. Literature research regarding these genes may shed some light on these findings. For example, KPNA2 is a gene that is noteworthy for promotion of cell proliferation and tumorigenicity in epithelial ovarian carcinoma [10]. Additionally, its expression is significantly associated with higher tumor stage, positive lymph node status, among other markers for metastatic breast cancer. Further experimentation in lung cancer has shown that depletion of KPNA2 expression using siRNAs inhibited proliferation in lung cancer cell lines [11]. Given these prior findings, it is not surprising that this gene would appear in this list, and it is also not surprising that high expression of this gene would be correlated with a more rapid disease progression and death. Most notably, one study showed that KPNA2 is a nuclear export protein that contributed to aberrant localization of key proteins and poor prognosis of breast cancer [12].

In this preliminary study of publicly available datasets relating to mesothelioma, several tools and analyses were performed to identify important genes and pathways. Analysis using a GEO dataset containing microarray data on mesothelioma yielded significantly DE genes that had previously been implicated in other cancers. In this analysis, there was an interesting trend of decreased expression of these genes in tumor cells amongst the top hits. While this yielded a large amount of information, Cytoscape's Agilent Literature Search application allowed for accelerated analysis of the literature to generate a highly connected and encompassing network. Analysis of this network linked key genes as well as suggesting pathways. One gene in particular, EphB2, had previously been the subject of research in other cancers and was highly linked to genes of interest in this study despite not being one of the top DE genes. This analysis provides justification for further study of this gene in the context of mesothelioma specifically. While many of these insights focused on differentiation of mesothelioma from normal cells, they were lacking in the sense that they did not have accompanying clinical data. TCGA provides a wealth of genomic datasets that are paired with clinical outcomes, and MIT's FireBrowse pipeline draws from that for accelerated analysis. While much of the statistical analysis for significance and cleaning had already been performed, the task then switches to visualization and understanding of complex, high volumes of information. Of some of these top genes identified by MIT's FireBrowse pipeline, it was apparent that they had a significant clinical effect on patient outcomes through visualization of their survival curves. Again, many of these genes that were identified had already been extensively studied in the literature in other cancers, though not typically in mesothelioma specifically. Overall, this preliminary work narrowed down a wealth of information into a subset of genes that would be of specific interest to labs that experiment with mesothelioma cell lines to try to optimize therapies for improved clinical outcomes.

Works Cited

[1] Yong He et al., "The Role of PKR/eIF2a Signaling Pathway in Prognosis of Non-Small Cell Lung Cancer," PLOS One, 2011.

[2] Steven M Albelda and Daniel H Sterman, "Advances in the Diagnosis, Evaluation, and Management of Malignant Pleural Mesothelioma," Respirology, vol. 10, no. 3, pp. 266-283, 2005.

[3] Deborah A Altomare et al., "Human and Mouse Mesotheliomas Exhibit Elevated AKT/PKB Activity, Which Can Be Targeted Pharmacologically to Inhibit Tumor Cell Growth," Oncogene, vol. 24, pp. 6080-6089, 2005.

[4] F Wang et al., "Loss of TACSTD2 contributed to squamous cell carcinoma progression through attenuating TAp63-dependent apoptosis," Cell Death and Disease, pp. 1-8, 2014.

[5] I Psallidas et al., "Secreted phosphoprotein-1 directly provokes vascular leakage to foster malignant pleural effusion," Oncogene, vol. 32, pp. 528-535, 2013.

[6] Wujiang Liu, Jin Jie, Liqun Zhou, and Yinglu Guo, "EMP2 as a novel target of bladder cancer immunotherapy," Journal of Urology, 2015.

[7] Yu Qin et al., "Epithelial Membrane Protein-2 (EMP2) activates Src protein and is a novel therapeutic target for glioblastoma," J Biol Chem, pp. 13974-13985, 2014.

[8] Qing Gao et al., “EphB2 promotes cervical cancer progression by inducing epithelial-mesenchymal transition,” Human Pathology, vol. 45, no. 2, pp. 372-381, 2014.

[9] Weiguang Mao et al., “EphB2 as a Therapeutic Antibody Drug Target for the Treatment of Colorectal Cancer,” Cancer Research, pp. 781-790, 2004.

[10] L Huang et al., “KPNA2 promotes cell proliferation and tumorigenicity in epithelial ovarian carcinoma through upregulation of c-Myc and downregulation of FOXO3a,” Cell Death Dis, vol. 4, 2013.

[11] S Ma and X Zhou, “KPNA2 is a promising biomarker candidate for esophageal squamous cell carcinoma and correlates with cell proliferation,” Oncol Rep, 2014.

[12] A Alshareeda et al., “KPNA2 is a nuclear export protein that contributes to aberrant localisation of key proteins and poor prognosis of breast cancer,” Br J Cancer, pp. 1929-37, 2015.