MDS plots for potential covariates

Annika Jorgensen

2023-06-07

R Markdown

Title: "2023 06/07 Version 2 New MDS Plots"
Author: Annika Jorgensen
Date: 06/07/2023

Purpose The purpose of this file is twofold. This script is meant to remove the samples that were found to be low quality and then remake the MDS plot to illustrate the low quality samples were removed. This script also conducts MDS analysis on a number of variables in the metadata thought to be significantly different between males and females to see if the variables are potentially driving variation in the tumor samples.

Results ### **Libraries** The first chunk of code is dedicated to installing the libraries. These libraries are to help execute the differential analysis and helps visualize the data. The code was not included for concision.

Defining Colors

This chunk defines color palette variables that are going to be used in plots later on the script. These variables are defined by converting BrewerCode palettes into palettes that can be used in R.

```
viralPalette <- brewer.pal(8, "Set1")
hbvColor <- viralPalette[1]
hcvColor <- viralPalette[2]
bothColor <- viralPalette[3]
neitherColor <- viralPalette[4]

sexTissuePalette <- brewer.pal(12, "Paired")
maleTumorColor <- sexTissuePalette[4]
maleAdjacentColor <- sexTissuePalette[3]
femaleTumorColor <- sexTissuePalette[6]
femaleAdjacentColor <- sexTissuePalette[5]
```

Read in data

This code is where you read in all the data files that are going to be used in the script. The data is also converted into a variety of variables that makes the data easier to handle. The data is also cleaned up to make sure the analysis done later is accurate and precise.

```
metadata <- read.table("~/Desktop/ResearchProjects/LiverCancer/Metadata/metadata_for_de.csv", row.names=
tumorAdjacentExp <- read.table("~/Desktop/ResearchProjects/LiverCancer/Metadata/japan_all_samples_salmo
colnames(tumorAdjacentExp) <- gsub("\\.", "-", colnames(tumorAdjacentExp)) #changing the column names

# Importing gene annotations
```

```r
#genes <- read.table("gencode.v25.chr_patch_hapl_scaff.annotation.bed", header=FALSE, sep="\t")
genes <- read.table("~/Desktop/ResearchProjects/LiverCancer/Metadata/gencodeTranscripts.txt", header=TRU
genes <- data.frame(genes)
tumorAdjacentExp <- tumorAdjacentExp[rownames(tumorAdjacentExp) %in% genes$GENEID ,]
genes <- genes[match(rownames(tumorAdjacentExp), genes$GENEID),]
# Calculating gene length, this is needed for calculating the FPKM values
genes$length <- with(genes, end - start)

# Removing Samples due to low quality
metadata <- metadata[!(metadata$ID == "RK023") , ]
metadata <- metadata[!(metadata$ID == "RK106") , ]
metadata <- metadata[!(metadata$ID == "RK113") , ]
metadata <- metadata[!(metadata$ID == "RK135") , ]
metadata <- metadata[!(metadata$ID == "RK105") , ]
metadata <- metadata[!(metadata$ID == "RK116") , ]
metadata <- metadata[!(metadata$ID == "RK066") , ]
metadata <- metadata[!(metadata$ID == "RK096") , ]

#Removing both and NBNC samples
metadata <- metadata[!(metadata$Virus_infection == "NBNC"), ]
metadata <- metadata[!(metadata$Virus_infection == "both"), ]

# Subsetting and ordering metadata to match the count matrix
tumorAdjacentExpSubset <- tumorAdjacentExp[,colnames(tumorAdjacentExp) %in% metadata$sampleid]
metadataSubset <- metadata[metadata$sampleid %in% colnames(tumorAdjacentExpSubset),]
metadataSubset <- metadataSubset[match(colnames(tumorAdjacentExpSubset), metadataSubset$sampleid),]
rownames(metadataSubset) <- metadataSubset$sampleid

# Adding tissue type, converting categorical variables to factors
metadataSubset$tumor <- as.numeric(grepl('tumor', metadataSubset$sampleid, ignore.case=T))

#Swapping lesion type for sample RK169
metadataSubset["RK169-tumor-XY","tumor"] <- 0
metadataSubset["RK169-adjacent-XY","tumor"] <- 1

#Changing rownames to match swapped lesion type
rownames(metadataSubset)[rownames(metadataSubset)=="RK169-tumor-XY"] <- "RK169_adjacent-XY"
rownames(metadataSubset)[rownames(metadataSubset)=="RK169-adjacent-XY"] <- "RK169_tumor-XY"
rownames(metadataSubset)[rownames(metadataSubset)=="RK169_adjacent-XY"] <- "RK169-adjacent-XY"

rownames(tumorAdjacentExpSubset)[rownames(tumorAdjacentExpSubset)=="RK169-tumor-XY"] <- "RK169_adjacent
rownames(tumorAdjacentExpSubset)[rownames(tumorAdjacentExpSubset)=="RK169-adjacent-XY"] <- "RK169_tumor
rownames(tumorAdjacentExpSubset)[rownames(tumorAdjacentExpSubset)=="RK169_adjacent-XY"] <- "RK169-adjac

#Swapping lesion type for sample RK065
metadataSubset["RK065-tumor-XX","tumor"] <- 0
metadataSubset["RK065-adjacent-XX","tumor"] <- 1

#Changing rownames in metadata to match swapped lesion type
rownames(metadataSubset)[rownames(metadataSubset)=="RK065-tumor-XY"] <- "RK065_adjacent-XY"
rownames(metadataSubset)[rownames(metadataSubset)=="RK065-adjacent-XY"] <- "RK065_tumor-XY"
rownames(metadataSubset)[rownames(metadataSubset)=="RK065_adjacent-XY"] <- "RK065-adjacent-XY"
```

```
rownames(tumorAdjacentExpSubset)[rownames(tumorAdjacentExpSubset)=="RK065-tumor-XY"] <- "RK065_adjacent
rownames(tumorAdjacentExpSubset)[rownames(tumorAdjacentExpSubset)=="RK065-adjacent-XY"] <- "RK065_tumor
rownames(tumorAdjacentExpSubset)[rownames(tumorAdjacentExpSubset)=="RK065_adjacent-XY"] <- "RK065-adjac

metadataSubset$gender_tissue <- paste(metadataSubset$Gender, metadataSubset$tumor, sep="_")
metadataSubset$gender_tissue_viral <- paste(metadataSubset$gender_tissue, metadataSubset$Virus_infectio
metadataSubset$library_type <- metadataSubset$strandedness
metadataSubset$library_type <- factor(metadataSubset$library_type)
metadataSubset$tumor <- factor(metadataSubset$tumor)
metadataSubset$Ta <- factor(metadataSubset$Ta)
metadataSubset$Portal_vein_invasion <- factor(metadataSubset$Portal_vein_invasion)
metadataSubset$Hepatic_vein_invasion <- factor(metadataSubset$Hepatic_vein_invasion)
metadataSubset$Bile_duct_invasion <- factor(metadataSubset$Bile_duct_invasion)
metadataSubset$Liver_fibrosisc <- factor(metadataSubset$Liver_fibrosisc)
metadataSubset$Prognosis <- factor(metadataSubset$Prognosis)
```

```
# Creating the DGEList object
dge <- DGEList(counts=tumorAdjacentExpSubset, genes=genes)
colnames(dge) <- colnames(tumorAdjacentExpSubset)
dge$samples$sex <- metadataSubset$Gender
dge$samples$viral <- factor(metadataSubset$Virus_infection)
dge$samples$ID <- metadataSubset$ID
dge$samples$tumor <- metadataSubset$tumor
dge$samples$gender_tissue <- metadataSubset$gender_tissue
dge$samples$gender_tissue_viral <- metadataSubset$gender_tissue_viral
dge$samples$library_type <- metadataSubset$library_type
dge$samples$edmonson_grade <- metadataSubset$Edmondson_grade
dge$samples$Ta <- metadataSubset$Ta
dge$samples$survival <- metadataSubset$Overall_survival_month
dge$samples$smoking <- factor(metadataSubset$Smoking)
dge$samples$alcohol <- factor(metadataSubset$Alcohol_intake)
dge$samples$fibrosis <- factor(metadataSubset$Liver_fibrosisc)

# Inspecting the N of samples in each group
table(dge$samples$gender_tissue_viral)
```

```
##
## F_0_HBV F_0_HCV F_1_HBV F_1_HCV M_0_HBV M_0_HCV M_1_HBV M_1_HCV
##       7      32       8      33      31      59      37      71
```

```
# ======================================
# Filtering expression data
# ======================================

# Keeping genes that have a mean FPKM of at least 0.5 in at least one of the
# groups under investigation and at least 6 reads in at least 10 samples
fpkm <- rpkm(dge, gene.length=dge$genes$length)

M_1_HBV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="M_1_HBV")],1,mean,na.
M_0_HBV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="M_0_HBV")],1,mean,na.
M_1_HCV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="M_1_HCV")],1,mean,na.
M_0_HCV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="M_0_HCV")],1,mean,na.
```

```
F_1_HBV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="F_1_HBV")],1,mean,na.
F_0_HBV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="F_0_HBV")],1,mean,na.
F_1_HCV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="F_1_HCV")],1,mean,na.
F_0_HCV_mean_fpkm <- apply(as.data.frame(fpkm)[(dge$samples$gender_tissue_viral=="F_0_HCV")],1,mean,na.

keep <- (M_1_HBV_mean_fpkm > 0.5 | M_0_HBV_mean_fpkm > 0.5 |
           M_1_HCV_mean_fpkm > 0.5 | M_0_HCV_mean_fpkm > 0.5 |
           F_1_HBV_mean_fpkm > 0.5 | F_0_HBV_mean_fpkm > 0.5 |
           F_1_HCV_mean_fpkm > 0.5 | F_0_HCV_mean_fpkm > 0.5 )

dge <- dge[keep,,keep.lib.sizes=FALSE]
dge <- calcNormFactors(dge, method="TMM")
keep <- rowSums(dge$counts > 6) >= 10
dge <- dge[keep,,keep.lib.size=FALSE]
dge <- calcNormFactors(dge, method="TMM")

# N of genes retained after filtering
dim(dge$genes)
```

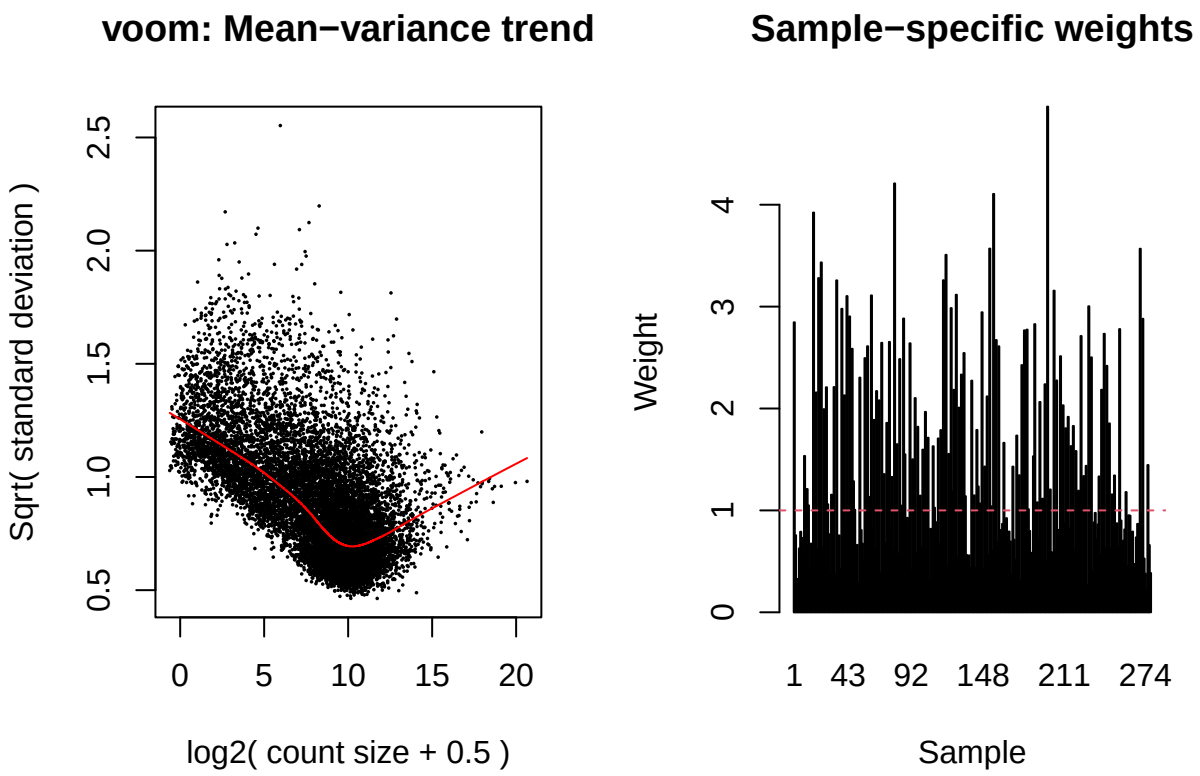
```
## [1] 12465     7
```

```
# ================================== ====================
# ============================================================
# Analysis of all tumor vs. tumor-adjacent regardless of sex
# ============================================================
# ============================================================

# Creating a new design model matrix with the variable of interest and the
# library type
design <- model.matrix(~0+dge$samples$tumor+dge$samples$library_type+dge$samples$Ta+dge$samples$smoking
colnames(design) <- gsub("dge\\$samples\\$tumor", "tumor", colnames(design))
colnames(design) <- gsub("dge\\$samples\\$library_typeunstranded", "library_type", colnames(design))
colnames(design) <- gsub("dge\\$samples\\$Ta2", "Ta2", colnames(design))
colnames(design) <- gsub("dge\\$samples\\$Ta3", "Ta3", colnames(design))
colnames(design) <- gsub("dge\\$samples\\$Ta4", "Ta4", colnames(design))
colnames(design) <- gsub("dge\\$samples\\$smoking", "smoking", colnames(design))
colnames(design) <- gsub("dge\\$samples\\$alcohol", "alcohol", colnames(design))
colnames(design) <- gsub("dge\\$samples\\$fibrosis", "fibrosis", colnames(design))
#colnames(design) <- gsub("dge\\$samples\\$Ta4", "Ta4", colnames(design))

head(design)
```

```
##   tumor0 tumor1 library_type Ta2 Ta3 Ta4 smoking1 alcohol1 alcohol2 alcohol3
## 1      1      0            1   0   0   1        1        1        0        0
## 2      0      1            1   0   0   1        1        1        0        0
## 3      1      0            1   1   0   0        0        0        0        0
## 4      0      1            1   1   0   0        0        0        0        0
## 5      1      0            1   0   1   0        1        0        1        0
## 6      0      1            1   0   1   0        1        0        1        0
##   fibrosis1 fibrosis2 fibrosis3 fibrosis4
## 1         0         0         0         1
## 2         0         0         0         1
```

```
## 3          0          0          1          0
## 4          0          0          1          0
## 5          0          0          0          1
## 6          0          0          0          1
```

```
# Running voom again with the new design matrix.
v <- voomWithQualityWeights(dge, design, plot=TRUE)
```



Top 25 gene MDS Plot

Recreating the MDS plot we have seen previously with the removed samples to see if all low quality samples have been removed.

```
# Removing batch effects
vCorrectLibtype <- removeBatchEffect(v, batch=v$targets$library_type)
```

the variable "identifiers" is a dataframe that houses all of the metadata information of the samples. The information is pull from the voom output and in order to ensure the sample order matches the expression data we are plotting I use the identical function. The data frame includes the following information (note that many of the number levels seem to be off by one that is intentional as variables were zero indexed and R changed them because it one indexed)

– The sample ID. Column name is id

– The lesion type where 1 means tumor - adjacent and 2 means tumor (identifiers is quickly subsetted to just include tumor samples so all values should say 2). Column name is tumor.

– Tumor stage where 1 means stage 1, 2 means stage, 3 means stage 3, 4 means stage 4. Column name is stage.

– Sex of the sample where M means male and F means female. Column name is sex.

– Edmondson grade of the tumor sample where 1 means grade 1, 1~2 means grade 1.5, 2 means grade 2, 2~3 means grade 2.5, 3 means grade 3 and 4 means grade 4. NA values were replaced with zeros because we are assuming the lack of a value means that the tumor was not graded. Column name is grade

– Viral etiology of the tumor sample where 1 means HBV and 2 means HCV. Column name is viral.

– Whether or not patient smokes where 1 means sample does not smoke and two means sample does smoke. Column name is smoking

– Level of liver fibrosis in the sample. Where 1 means liver fibrosis severity of zero, 2 means liver fibrosis severity of one, 3 means liver fibrosis severity of two, 4 means liver fibrosis severity of three, and five means liver fibrosis severity of four. Column name is fibrosis.

– Alcohol intake of the sample where where 1 means alcohol intake level zero, 2 means alcohol intake level one, 3 means alcohol intake level two, 4 means alcohol intake level three. Column name is alcohol

```
#creating data frame with metadata variables
identifiers<- cbind(id = v[["targets"]][["ID"]], tumor = v[["targets"]][["tumor"]], stage = v[["targets"

identifiers<- as.data.frame(identifiers)

#removing NAs fram edmondson grade and replacing with zero
identifiers$grade <- ifelse(is.na(identifiers$grade), 0, identifiers$grade)

#subsetting by tumor samples
subset_condition <- which(identifiers$tumor == "2")
identifiers <- identifiers[subset_condition, ]

#subsetting expression data to tumor samples
Tumor<- vCorrectLibtype[, grepl("tumor",colnames(vCorrectLibtype),fixed=TRUE)]
Tumor<- as.data.frame(Tumor)

#extracting id from sample name since the column name includes additional information
first_five_chars <- substr(colnames(Tumor), 1, 5)

#checking if sample orders match
result <- identical(identifiers$id, first_five_chars)
print(result)
```

```
## [1] TRUE
```

MDS plot was created to see if tumor stage was driving signal. Principal Components 1-4 were checked on the top 25,50,100,1000 genes were checked. To limit the risk of making a mistake the arguments dim.plot was changed from c(1,2) to c(3,4) to check Principal Components 3 and 4 and then top was changed from 25 to 50, 100, and finally 1000 to check the different gene amounts. No clustering was observed.

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/stage_P
#creating MDS plot top 25 genes PC 1 and 2 coloring by tumor stage
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(1,2), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$stage=="1","#FFC20A",
```

```
                           ifelse(identifiers$stage=="2", "#E66100",
                                  ifelse(identifiers$stage=="3", "#40B0A6","#006CD1"))), gene.selection =
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("bottomright", pch=c(15),
      col=c( "#FFC20A", "#E66100","#40B0A6", "#006CD1"),
      legend=c("Stage 1","Stage 2", "Stage 3","Stage 4"))
legend("topleft", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
dev.off()
```

```
## pdf
##   2
```

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/stage_PC
#creating MDS plot top 25 genes PC 3 and 4 coloring by tumor stage
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(3,4), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$stage=="1","#FFC20A",
                           ifelse(identifiers$stage=="2", "#E66100",
                                  ifelse(identifiers$stage=="3", "#40B0A6","#006CD1"))), gene.selection =
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("bottomright", pch=c(15),
      col=c( "#FFC20A", "#E66100","#40B0A6", "#006CD1"),
      legend=c("Stage 1","Stage 2", "Stage 3","Stage 4"))
legend("topleft", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
dev.off()
```

```
## pdf
##   2
```

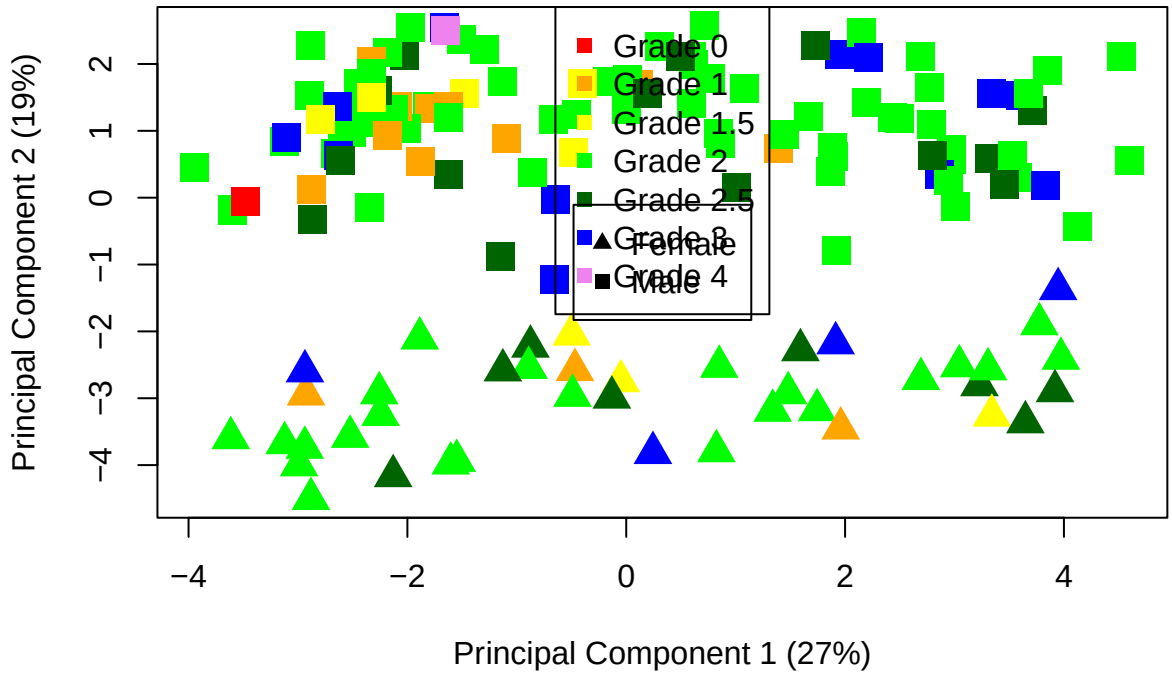
MDS plot was created to see if viral etiology was driving signal.Principal Components 1-4 were checked on the top 25,50,100,1000 genes were checked. To limit the risk of making a mistake the arguments dim.plot was changed from c(1,2) to c(3,4) to check Principal Components 3 and 4 and then top was changed from 25 to 50, 100, and finally 1000 to check the different gene amounts. No clustering was observed.

```
#creating MDS plot top 25 genes principal component 1 and 2 coloring by viral etiology

pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/viral_e
plotMDS(Tumor, top = 25, dim.plot = c(1,2), plot=TRUE, cex=2,
               pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col=ifelse(identifiers$viral %in% c("1"), "#1A85FF", #1 means HBV
                          ifelse(identifiers$viral %in% c("2"), "#D41159", "black")), #2 means HCV
               gene.selection = "common")
legend("top", pch=c(15),
       col=c("#D41159", "#1A85FF") ,
       legend=c("HCV", "HBV"))
legend("center", pch=c(17, 15),
       col=c("black"),
       legend=c("Female", "Male"))
```

```
#creating MDS plot top 25 genes principal component 3 and 4 coloring by viral etiology

pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/viral_e
plotMDS(Tumor, top = 25, dim.plot = c(3,4), plot=TRUE, cex=2,
               pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col=ifelse(identifiers$viral %in% c("1"), "#1A85FF", #1 means HBV
                          ifelse(identifiers$viral %in% c("2"), "#D41159", "black")), #2 means HCV
               gene.selection = "common")
legend("top", pch=c(15),
       col=c("#D41159", "#1A85FF") ,
       legend=c("HCV", "HBV"))
legend("center", pch=c(17, 15),
       col=c("black"),
       legend=c("Female", "Male"))
```

MDS plot was created to see if Edmondson Grade was driving signal.Principal Components 1-4 were checked on the top 25,50,100,1000 genes were checked. To limit the risk of making a mistake the arguments dim.plot was changed from c(1,2) to c(3,4) to check Principal Components 3 and 4 and then top was changed from 25 to 50, 100, and finally 1000 to check the different gene amounts. No clustering was observed.

```
plotMDS(Tumor, top = 25, dim.plot = c(1,2), plot=TRUE, cex=2,
                pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$grade=="0","red",
                           ifelse(identifiers$grade=="1", "orange",
                                  ifelse(identifiers$grade=="1~2", "yellow",
                                         ifelse(identifiers$grade=="2",  "green",
                                                ifelse(identifiers$grade=="2~3", "darkgreen",
                                                       ifelse(identifiers$grade=="3", "blue", "violet"))))),
                gene.selection = "common")
legend("top", pch=c(15),
       col=c( "red", "orange","yellow", "green", "darkgreen", "blue", "violet"),
      legend=c("Grade 0","Grade 1", "Grade 1.5","Grade 2", "Grade 2.5", "Grade 3", "Grade 4"))
legend("center", pch=c(17, 15),
       col=c("black"),
       legend=c("Female", "Male"))
```



MDS plot was created to see if Smoking was driving signal.Principal Components 1-4 were checked on the top 25,50,100,1000 genes were checked. To limit the risk of making a mistake the arguments dim.plot was changed from c(1,2) to c(3,4) to check Principal Components 3 and 4 and then top was changed from 25 to 50, 100, and finally 1000 to check the different gene amounts. No clustering was observed.

```
#MDS plot top 25 genes principal components 1 & 2 colored by patients who smoke and patients who do not

pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/smoking_
plotMDS(Tumor, top = 25, dim.plot = c(1,2), plot=TRUE, cex=2,
                pch=ifelse(identifiers$sex %in% c("F"),17, 15),
```

```
               col=ifelse(identifiers$smoking %in% c("1"), "#D41159", "#1A85FF"),
               gene.selection = "common")
legend("bottomright", pch=c(15),
       col=c("#1A85FF", "#D41159") ,
       legend=c("Smoking", "Not smoking"))
legend("center", pch=c(17, 15),
       col=c("black"),
       legend=c("Female", "Male"))
dev.off()
```

```
## pdf
##   2
```

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/smoking_
plotMDS(Tumor, top = 25, dim.plot = c(3,4), plot=TRUE, cex=2,
               pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col=ifelse(identifiers$smoking %in% c("1"), "#D41159","#1A85FF"),
               gene.selection = "common")
legend("topleft", pch=c(15),
       col=c("#1A85FF", "#D41159") ,
       legend=c("Smoking", "Not Smoking"))
legend("center", pch=c(17, 15),
       col=c("black"),
       legend=c("Female", "Male"))
dev.off()
```

```
## pdf
##   2
```

MDS plot was created to see if Alcohol Intake was driving signal.Principal Components 1-4 were checked on the top 25,50,100,1000 genes were checked. To limit the risk of making a mistake the arguments dim.plot was changed from c(1,2) to c(3,4) to check Principal Components 3 and 4 and then top was changed from 25 to 50, 100, and finally 1000 to check the different gene amounts. No clustering was observed.

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/alcohol_
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(1,2), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$alcohol=="1","#FFC20A", #one means intake level zero
                           ifelse(identifiers$alcohol=="2", "#E66100", #two means intake level one
                                  ifelse(identifiers$alcohol=="3", "#40B0A6","#006CD1"))), gene.selectio
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter
```

```
## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("bottomleft", pch=c(15),
      col=c( "#FFC20A", "#E66100","#40B0A6", "#006CD1"),
      legend=c("No alcohol","Severity 1", "Severity 2","Severity 3"))

legend("topleft", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
dev.off()
```

```
## pdf
##   2
```

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/alcohol_
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(1,2), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$alcohol=="1","#FFC20A",
                           ifelse(identifiers$alcohol=="2", "#E66100",
                                  ifelse(identifiers$alcohol=="3", "#40B0A6","#006CD1"))), gene.selectio
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("bottom", pch=c(15),
      col=c( "#FFC20A", "#E66100","#40B0A6", "#006CD1"),
      legend=c("No alcohol","Severity 1", "Severity 2","Severity 3"))

legend("top", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
```

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/alcohol_
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(1,2), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$alcohol=="1","#FFC20A","#40B0A6"), gene.selection = "common")
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("bottom", pch=c(15),
      col=c( "#FFC20A", "#40B0A6"),
      legend=c("No alcohol","Alcohol"))

legend("top", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
dev.off()
```

```
## pdf
##   2
```

These MDS plots compare the variation amongst patients who drink alcohol versus patients who do not. Same process was followed as above.

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/alcohol_
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(1,2), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$alcohol=="1","#FFC20A","#40B0A6"), gene.selection = "common")
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("bottom", pch=c(15),
      col=c( "#FFC20A", "#40B0A6"),
      legend=c("No alcohol","Alcohol"))

legend("top", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
dev.off()
```

```
## pdf
##   2
```

```
pdf("~/Desktop/ResearchProjects/SexChromosomeGithubUpload/MDSplots_potential_covariates/figures/alcohol_
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(3,4), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$alcohol=="1","#FFC20A","#40B0A6"), gene.selection = "common")
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("bottom", pch=c(15),
      col=c( "#FFC20A", "#40B0A6"),
      legend=c("No alcohol","Alcohol"))

legend("top", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
dev.off()
```

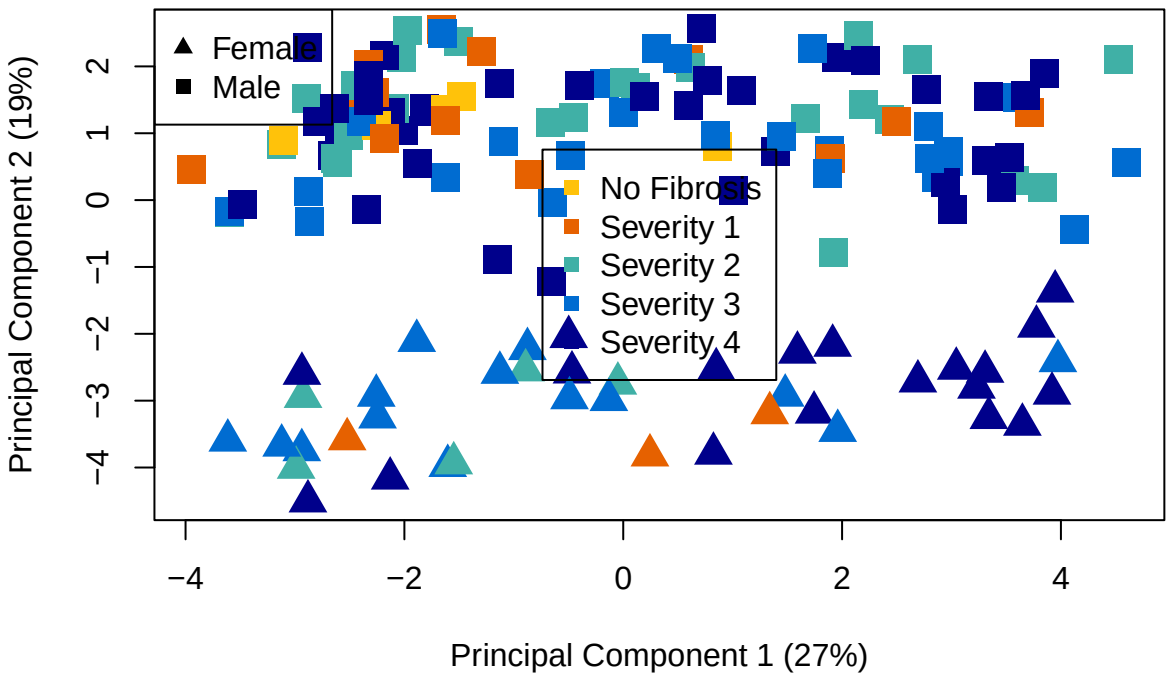
```
## pdf
##   2
```

MDS plot was created to see if Liver Fibrosis was driving signal.Principal Components 1-4 were checked on the top 25,50,100,1000 genes were checked. To limit the risk of making a mistake the arguments dim.plot was changed from c(1,2) to c(3,4) to check Principal Components 3 and 4 and then top was changed from 25 to 50, 100, and finally 1000 to check the different gene amounts. No clustering was observed.

```
mds <- plotMDS(Tumor, top = 25, ndim = 10, dim.plot = c(1,2), plot=TRUE, cex=2,
            pch=ifelse(identifiers$sex %in% c("F"),17, 15),
               col= ifelse(identifiers$fibrosis=="1","#FFC20A", #one = fibrosis level zero
                           ifelse(identifiers$fibrosis=="2", "#E66100", #two = fibrosis level one
                                  ifelse(identifiers$fibrosis=="3", "#40B0A6",#three = fibrosis level tw
                                         ifelse(identifiers$fibrosis== "4", "#006CD1", "darkblue")))), g
```

```
## Warning in plot.window(...): "ndim" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "ndim" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "ndim" is not a
## graphical parameter

## Warning in box(...): "ndim" is not a graphical parameter

## Warning in title(...): "ndim" is not a graphical parameter
```

```
legend("center", pch=c(15),
      col=c( "#FFC20A", "#E66100","#40B0A6", "#006CD1", "darkblue"),
      legend=c("No Fibrosis","Severity 1", "Severity 2","Severity 3", "Severity 4"))

legend("topleft", pch=c(17, 15),
     col=c("black"),
legend=c("Female", "Male"))
```


Female
Male
No Fibrosis
Severity 1
Severity 2
Severity 3
Severity 4
Principal Component 2 (19%)
Principal Component 1 (27%)
−4
−2
0
2
4
−4
−3
−2
−1
0
1
2