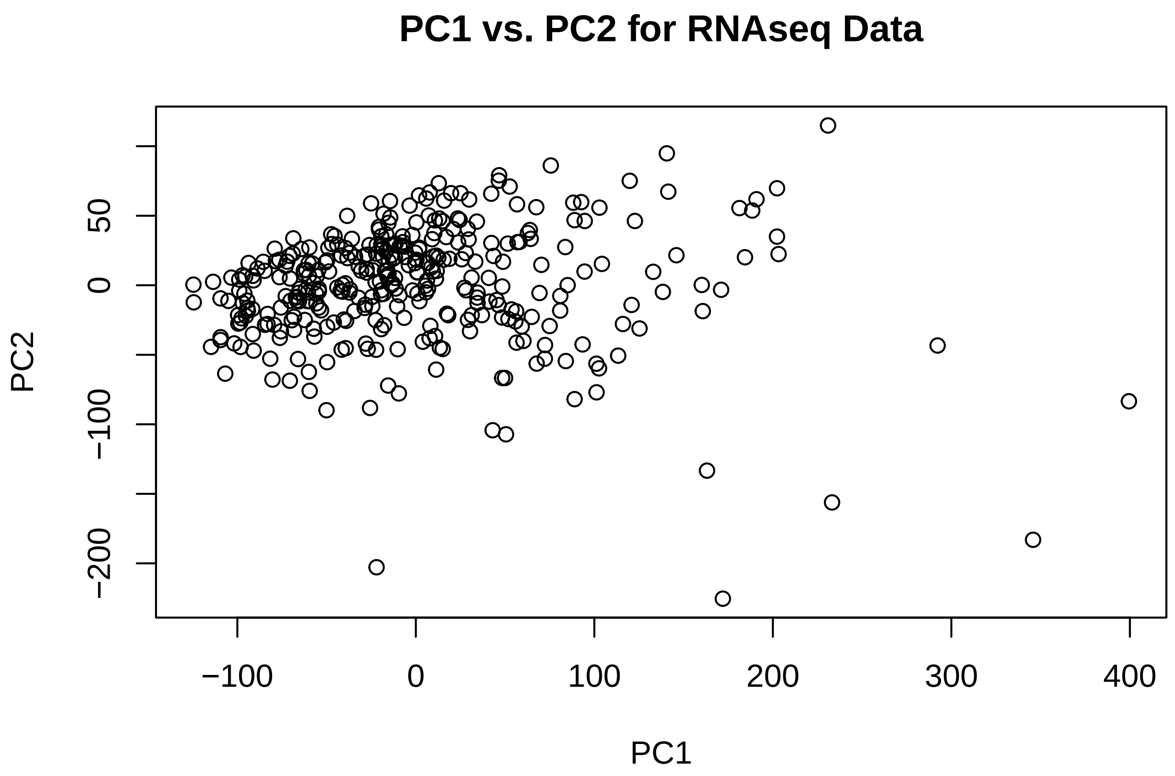
Differential Expression Analysis

Group 20

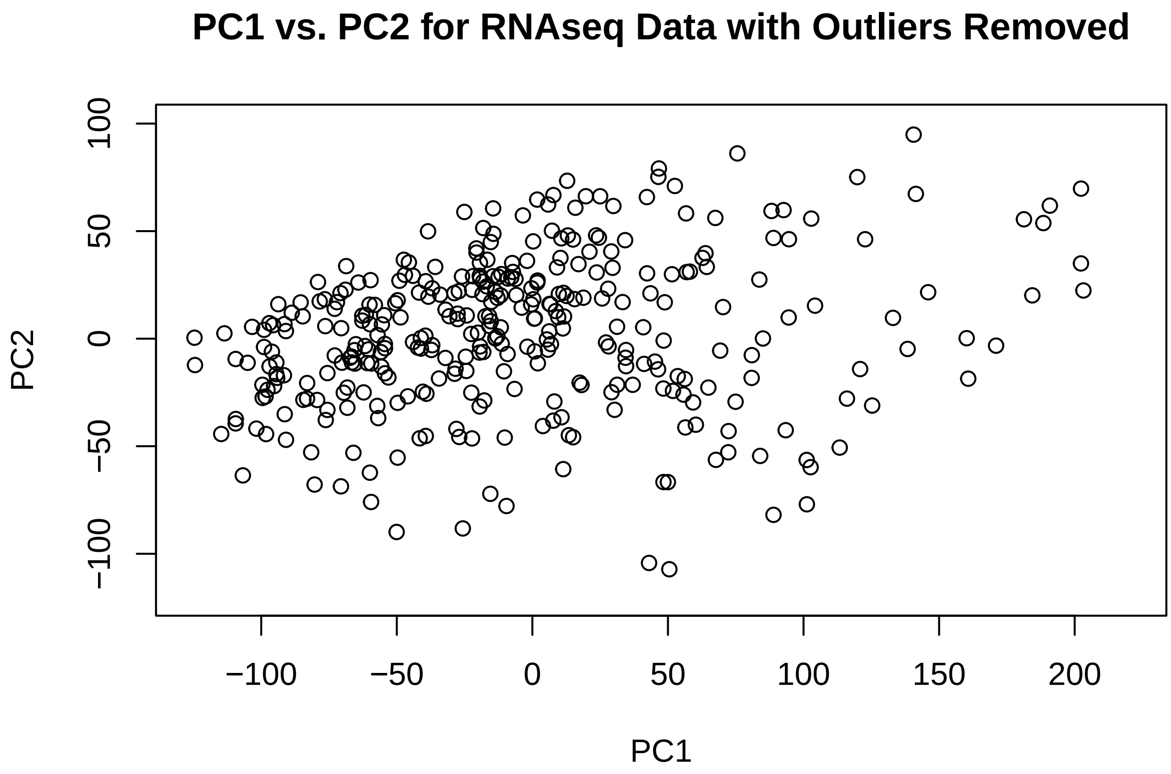
2023-11-12

PCA to remove outliers

```r
liver_data_cst_removed <- read.csv("rnaseq_data_shared.csv", row.names = 1)

liver_pca <- prcomp(liver_data_cst_removed, center = TRUE, scale = TRUE)

pc_1_2 <- liver_pca$x

# cutoff is < 220 for PC1, > -120 for PC2
plot(pc_1_2[, 1], pc_1_2[, 2], main = "PC1 vs. PC2 for RNAseq Data", xlab = "PC1",
    ylab = "PC2")
```



```
plot(pc_1_2[, 1], pc_1_2[, 2], main = "PC1 vs. PC2 for RNAseq Data with Outliers Removed",
    xlab = "PC1", ylab = "PC2", xlim = c(-125, 220), ylim = c(-120, 100))
```



```
# subset into new dataframe with no outliers
rownames_to_filter <- row.names(pc_1_2[(pc_1_2[, 1] > 220 | pc_1_2[, 2] < -120),
    ])
liver_data_filtered <- liver_data_cst_removed[!(rownames(liver_data_cst_removed) %in%
    rownames_to_filter), ]
```

```
library(DESeq2)
```

```
## Warning: package 'DESeq2' was built under R version 4.3.1

## Warning: package 'S4Vectors' was built under R version 4.3.1

## Warning: package 'IRanges' was built under R version 4.3.1

## Warning: package 'GenomicRanges' was built under R version 4.3.1

## Warning: package 'GenomeInfoDb' was built under R version 4.3.1

## Warning: package 'MatrixGenerics' was built under R version 4.3.1
## Warning: package 'matrixStats' was built under R version 4.3.1
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```r
library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.3.2
```

DESeq Pipeline

```r
# cluster data for each patient
clinical_clusters <- read.csv("clinical_with_groups.csv")
head(clinical_clusters)
```

```
##     PATIENT_ID                 DSS_STATUS DSS_MONTHS DAYS_LAST_FOLLOWUP
## 1 TCGA-2V-A95S 0:ALIVE OR DEAD TUMOR FREE         NA                 NA
## 2 TCGA-2Y-A9GS          1:DEAD WITH TUMOR   723.9921                 NA
## 3 TCGA-2Y-A9GT          1:DEAD WITH TUMOR  1623.9822                 NA
## 4 TCGA-2Y-A9GU 0:ALIVE OR DEAD TUMOR FREE  1938.9788               1939
## 5 TCGA-2Y-A9GV          1:DEAD WITH TUMOR  2531.9723                 NA
## 6 TCGA-2Y-A9GW          1:DEAD WITH TUMOR  1270.9861                 NA
##   subgroup deceased overall_survival
## 1        1    FALSE               NA
## 2        2     TRUE         723.9921
## 3        1     TRUE        1623.9822
## 4        1    FALSE        1939.0000
## 5        1     TRUE        2531.9723
## 6        1     TRUE        1270.9861
```

```r
colData <- clinical_clusters[, c("PATIENT_ID", "subgroup")]

colData$subgroup <- as.factor((colData$subgroup))
colData <- colData[!(colData$PATIENT_ID %in% rownames_to_filter), ]  #removing the outlier patients

countData <- as.data.frame(t(as.matrix(liver_data_filtered)))  #format for DESeq
countData <- countData[, order(colnames(countData))]  #order the patient names to match colData
```

```r
dds = DESeqDataSetFromMatrix(countData = countData, colData = colData, design = ~subgroup)
```

```
## converting counts to integer mode
```

```r
# countData is count numbers colData is the conditions for each sample design
# is the formula used (i.e. what variable we want to compare between the
# samples)
```

```
dds = DESeq(dds)
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```
## -- replacing outliers and refitting for 5457 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)
```

```
## estimating dispersions
```

```
## fitting model and testing
```

```
dds
```

```
## class: DESeqDataSet
## dim: 54563 344
## metadata(1): version
## assays(6): counts mu ... replaceCounts replaceCooks
## rownames(54563): ENSG00000000003.15 ENSG00000000005.6 ...
##   ENSG00000288674.1 ENSG00000288675.1
## rowData names(23): baseMean baseVar ... maxCooks replace
## colnames(344): TCGA-2V-A95S TCGA-2Y-A9GS ... TCGA-ZS-A9CF TCGA-ZS-A9CG
## colData names(4): PATIENT_ID subgroup sizeFactor replaceable
```

```
res = results(dds, contrast = c("subgroup", "1", "2"))
summary(res)
```

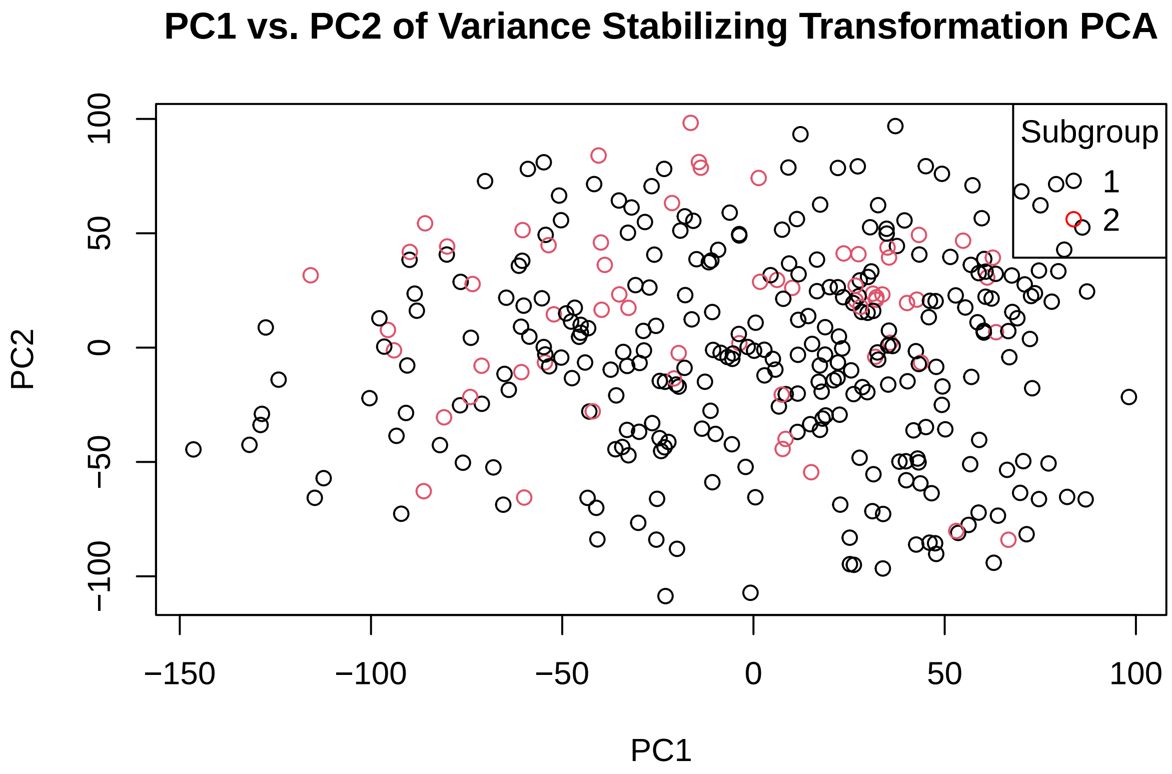
```
##
## out of 54515 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 2599, 4.8%
## LFC < 0 (down)     : 2818, 5.2%
## outliers [1]       : 0, 0%
## low counts [2]     : 21140, 39%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
res <- results(dds)
res.table <- table(res$padj < 0.01 & abs(res$log2FoldChange) > 1.5)
res.table
```

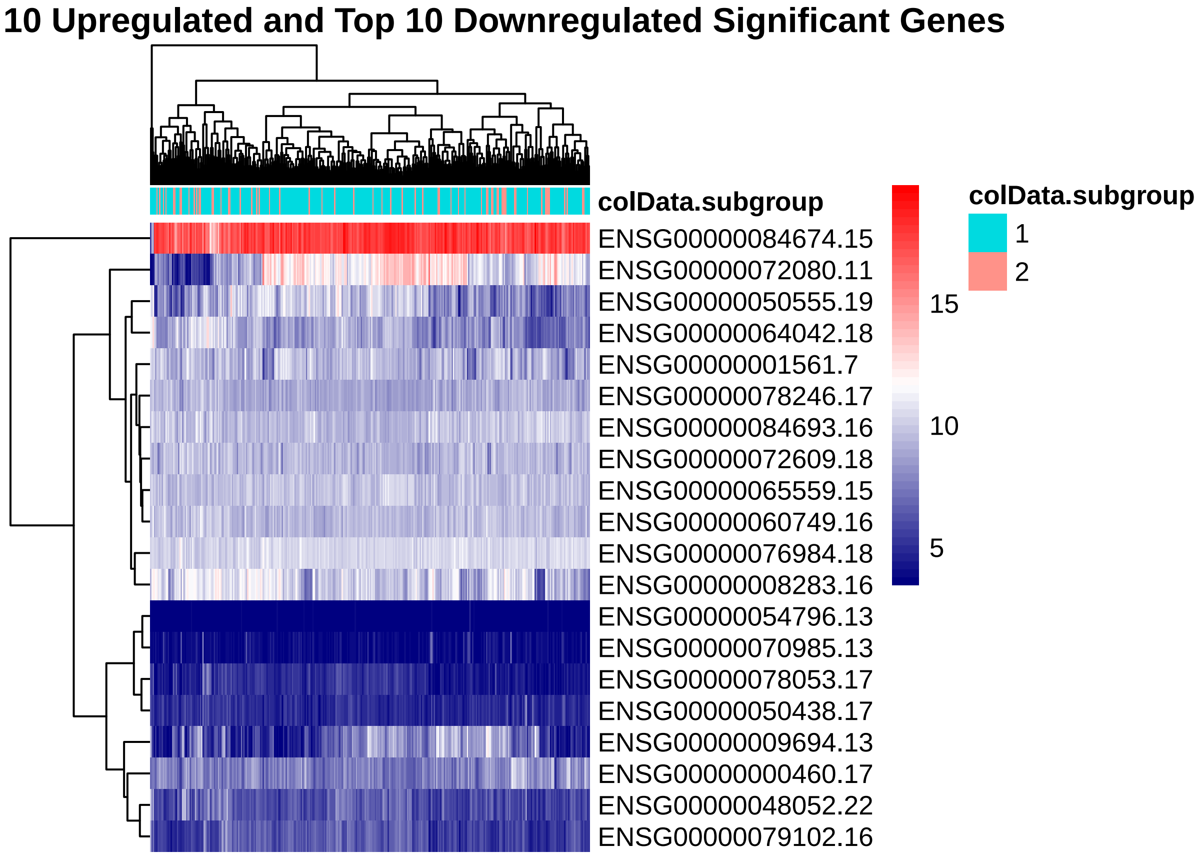
```
##
## FALSE  TRUE
## 54205   284
```

```
# filter results until 200-300 genes are obtained as 'TRUE'
```

```
# Variance stabilizing transformation
vsd <- vst(dds)

vsd.results <- t(assay(vsd))
vsd_pca <- prcomp(vsd.results, center = TRUE)
vsd.pca.res <- as.data.frame(vsd_pca$x)

# demonstration of clusters on the PCs
vsd.pca.res$subgroup <- colData$subgroup[match(row.names(vsd.pca.res), colData$PATIENT_ID)]

plot(vsd.pca.res$PC1, vsd.pca.res$PC2, col = vsd.pca.res$subgroup, main = "PC1 vs. PC2 of Variance Stabi
    xlab = "PC1", ylab = "PC2")
legend("topright", legend = levels(vsd.pca.res$subgroup), col = c("black", "red"),
    pch = 1, title = "Subgroup")
```



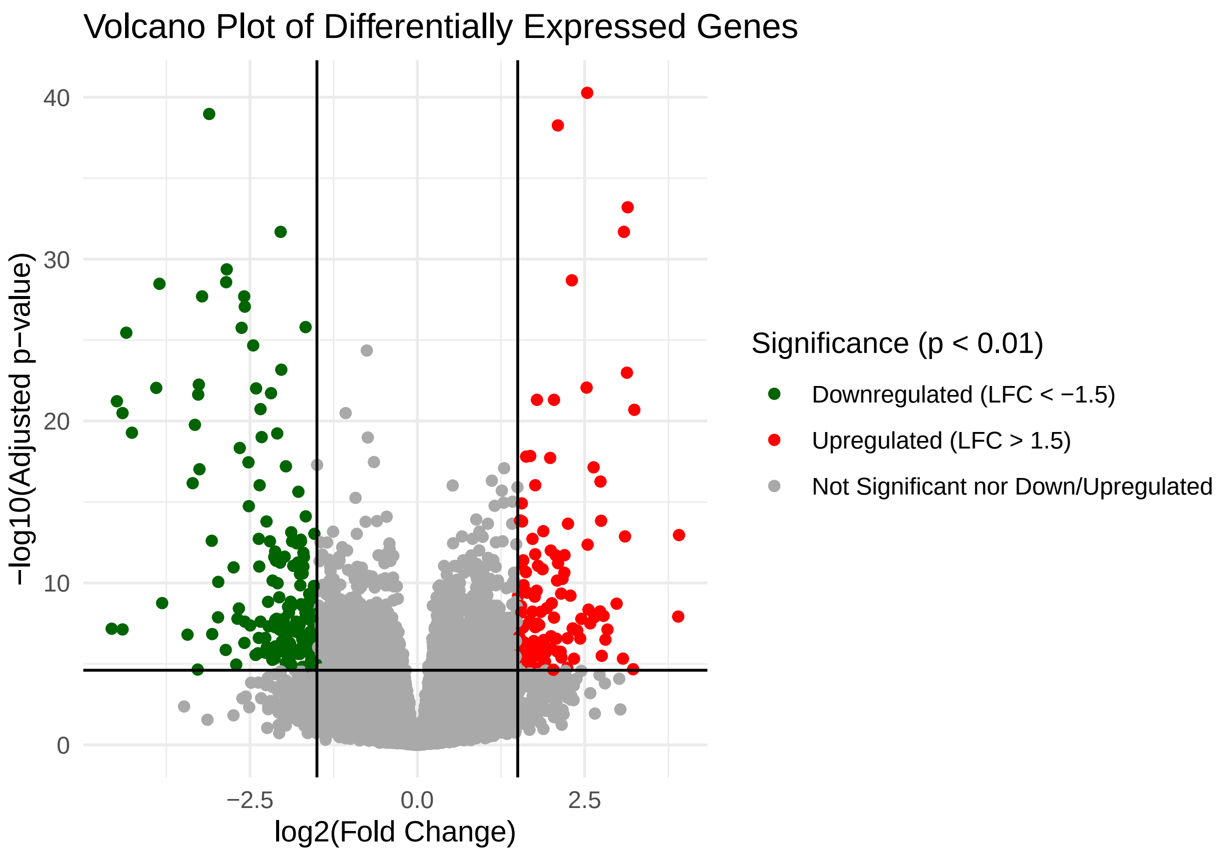
```r
# subset the most significant genes, then obtain the top 10 most upregulated
# and top 10 most downregulated genes

resSig <- subset(res, padj < 0.01)
up <- order(resSig$log2FoldChange, decreasing = TRUE)[1:10]
down <- order(resSig$log2FoldChange, decreasing = FALSE)[1:10]
genes3 <- c(up, down)

# generate heatmap of the 20 genes
annot_col = data.frame(colData$subgroup)
row.names(annot_col) <- colData$PATIENT_ID

sampleMatrix <- assay(vsd)[genes3, ]

rownames(sampleMatrix) = rownames(countData[genes3, ])
colnames(sampleMatrix) = colnames(countData)

pheatmap(sampleMatrix, cluster_rows = TRUE, show_rownames = TRUE, cluster_cols = TRUE,
    show_colnames = FALSE, annotation_col = annot_col, main = "Heatmap of Top 10 Upregulated and Top 10
    color = colorRampPalette(c("navy", "white", "red"))(50))
```



```r
#generating volcano plot of genes
genes_volcano <- data.frame(res$log2FoldChange, res$padj)
rownames(genes_volcano) <- rownames(res)
```

```
vec <- rownames(genes_volcano)[abs(genes_volcano$res.log2FoldChange) > 1.5 & genes_volcano$res.padj < 0

volcano_label <- vec[!is.na(vec)]

library(ggplot2)
ggplot(data=genes_volcano, aes(x=res.log2FoldChange, y=-log(res.padj))) +
  geom_point(
    aes(
      color =
        ifelse(
        res.padj < 0.01,
        ifelse(res.log2FoldChange > 1.5, "red", ifelse(res.log2FoldChange < -1.5, "darkgreen", "darkgre
        "darkgrey"
      ) #subsetting gene colors based on
      #significance, upregulation, and downregulation
    )
  ) +
  scale_color_manual(values = c("darkgreen", "red", "darkgrey"),
                     breaks = c("darkgreen", "red", "darkgrey"),
    labels = c("Downregulated (LFC < -1.5)", "Upregulated (LFC > 1.5)", "Not Significant nor Down/Upreg
  labs(title = "Volcano Plot of Differentially Expressed Genes",
       x = "log2(Fold Change)",
       y = "-log10(Adjusted p-value)",
       color = "Significance (p < 0.01)"
       ) +
  geom_hline(yintercept=-log(0.01)) +
  geom_vline(xintercept = 1.5) +
  geom_vline(xintercept = -1.5) +
  theme_minimal()
```

```
## Warning: Removed 21186 rows containing missing values (`geom_point()`).
```


Volcano Plot of Differentially Expressed Genes
−log10(Adjusted p−value)
40
30
20
10
0
−2.5
0.0
2.5
log2(Fold Change)
Significance (p < 0.01)
Downregulated (LFC < −1.5)
Upregulated (LFC > 1.5)
Not Significant nor Down/Upregulated