PGC1α Loss Promotes Lung Cancer Metastasis through Epithelial-Mesenchymal Transition

By: Cody Watson and Uzair Qadir

Abstract

PGC1α has been observed to downregulate cancer metastasis in melanoma, breast, and pancreatic cancer, but it has also been associated with worse outcomes in other tumor types (Torrano et al., 2016; LeBleu et al., 2014; Luo et al., 2016; Jiang et al., 2003; LaGory et al., 2015; Shoag et al., 2013; Vazquez et al., 2013). Little is currently known about its role in mediating lung cancer metastasis. A better understanding of its role in lung cancer could lead to the emergence of a useful therapeutic target. This study examines A549 human lung cancer cells and the impact of PGC1α knockdown on EMT related markers to better understand this pathway. Analysis of gene expression reveals the suppression of PGC1α significantly downregulated the expression of epithelial marker CDH1, while also upregulating mesenchymal markers such as CDH2, VIM, ITGA5, SNAI1, and SNAI2. This supports the idea that PGC1α is an opposing regulator of lung cancer metastasis and that this pathway is a potential target for future therapies.

Background and Experimental Design Rationale

The epithelial–mesenchymal transition (EMT) is a process in which epithelial cells lose cell polarity and cell adhesion. They can become migratory and turn into multipotent stromal cells which can then further differentiate into a number of cell types such as fibroblasts which can lead to dangerous growth elsewhere in the body. It is a key process implicated in lung cancer metastasis (Chaffer et al., 2016).

In order to evaluate the role of PGC1α on this EMT pathway in lung cancer specifically, two groups of cells were differentially treated. A549 lung adenocarcinoma cells were selected due to their particularly high levels of PGC1α expression. Then, 3 treatment groups were selected and treated with siRNA to knockdown PGC1α expression. Further, the 3 control groups were treated with random siRNA to ensure the treatment process itself did not have an impact on the results. The dataset we analyzed was produced by isolation of RNA via TRIzol. RNA-Seq was prepared via TruSeq Stranded mRNA LT Sample Prep Kit and sequencing with a NovaSeq6000 system.

Analyses Results

Obtaining RNA-seq dataset

We utilized sratoolkit to download the fastq files from the NIH GEO database.

```
#!/bin/bash

#SBATCH -A e30836

#SBATCH -p normal

#SBATCH -N 1

#SBATCH -n 2

#SBATCH -t 48:00:00
```

```
#SBATCH --mem=50G

module load sratoolkit/3.0.0

fastq-dump -I --split-files SRR12514551 -O /projects/e30836/project/group10_2022/fastq_data/

fastq-dump -I --split-files SRR12514552 -O /projects/e30836/project/group10_2022/fastq_data/

fastq-dump -I --split-files SRR12514553 -O /projects/e30836/project/group10_2022/fastq_data/

fastq-dump -I --split-files SRR12514554 -O /projects/e30836/project/group10_2022/fastq_data/

fastq-dump -I --split-files SRR12514555 -O /projects/e30836/project/group10_2022/fastq_data/

fastq-dump -I --split-files SRR12514556 -O /projects/e30836/project/group10_2022/fastq_data/
```

This step resulted in the generation of the fastq files for each SRR file in the NIH GEO database.

Performing Read Alignment

In order to perform further analysis, we need to map all the reads we obtained from the fastq file to the genome. We used STAR
Read Alignment code:

```
#!/bin/bash
#SBATCH -A e30836
#SBATCH -p normal
#SBATCH -N 1
#SBATCH -n 2
#SBATCH -t 48:00:00
#SBATCH --mem=80G
# load modules you need to use
module load STAR/2.5.2

cd /projects/e30836/project/group10_2022/fastq_data/SRR12514551_1

STAR --runThreadN 10 --quantMode TranscriptomeSAM --genomeDir
/projects/e30836/hw1/hg38.index/STAR --readFilesIn
/projects/e30836/project/group10_2022/fastq_data/SRR12514551_1/SRR12514551_1.fastq
```

Track file generation code:

```
#!/bin/bash
#SBATCH -A e30836
#SBATCH -p normal
#SBATCH -N 1
#SBATCH -n 2
#SBATCH -t 48:00:00
#SBATCH --mem=80G

# load modules you need to use
module load samtools/1.6
module load deeptools/3.1.1



samtools view -b ../Aligned.out.sam > ../Aligned.out.bam
samtools sort ../Aligned.out.bam > Aligned.out.sort.bam
samtools index Aligned.out.sort.bam
bamCoverage --bam Aligned.out.sort.bam --normalizeUsing CPM --outFileName
SRR12514551_1.-.bigWig --filterRNAstrand reverse --binSize 1 --numberOfProcessors 60
bamCoverage --bam Aligned.out.sort.bam --normalizeUsing CPM --outFileName
SRR12514551_1.+.bigWig --filterRNAstrand forward --binSize 1 --numberOfProcessors 60
```

This code was run for each fastq file generated in step 1 and resulted in the bam and BigWig files that were necessary for future gene expression analysis and for IGV analysis.

Calculating Gene Expression Levels

We used RSEM on the Quest server to generate a matrix of gene counts for comparing our different samples.

RSEM code:

```
#!/bin/bash
#SBATCH -A e30836
#SBATCH -p normal
#SBATCH -N 1
```

```
#SBATCH -n 2
#SBATCH -t 48:00:00
#SBATCH --mem=80G
# load modules you need to use
module load rsem/1.3.0
# Set your working directory
cd /projects/e30836/project/group10_2022/RSEM/SRR12514551_1
# A command you actually want to execute:
rsem-calculate-expression -p 10 --bam
/projects/e30836/project/group10_2022/fastq_data/SRR12514551_1/Aligned.toTranscriptome.ou
t.bam /projects/e30836/hw1/hg38.index/STAR/gencode.v28 SRR12514551_1
```

This code was run for each fastq file generated in step 1 and generated the files: Generated (SRR filename).genes.results, (SRR filename).isoforms.results, (SRR filename).stat, and (SRR filename).transcript.bam files. The genes.results files were used to generate the count matrix for DESeq analysis and TPM values for the differentially expressed gene heatmaps.

Generating Count Matrix and Annotation Files

Some of the reads were split into multiple files and needed to be concatenated. We performed data cleaning and generation of the annotation files for further analysis in python.

Count Matrix, TPM matrix, Annotation file for DESeq and Gene Ontology:

Using python,

```
import pandas as pd
SRR12514551_1 = pd.read_csv("SRR12514551_1.genes.results", delimiter='\t')
SRR12514552_1 = pd.read_csv("SRR12514552_1.genes.results", delimiter='\t')
SRR12514553_1 = pd.read_csv("SRR12514553_1.genes.results", delimiter='\t')
SRR12514554_1 = pd.read_csv("SRR12514554_1.genes.results", delimiter='\t')
SRR12514555_1 = pd.read_csv("SRR12514555_1.genes.results", delimiter='\t')
SRR12514556_1 = pd.read_csv("SRR12514556_1.genes.results", delimiter='\t')
SRR12514551_1 = SRR12514551_1.loc[:,["gene_id","expected_count"]]
SRR12514552_1 = SRR12514552_1.loc[:,"expected_count"]
SRR12514553_1 = SRR12514553_1.loc[:,"expected_count"]
SRR12514554_1 = SRR12514554_1.loc[:,"expected_count"]
SRR12514555_1 = SRR12514555_1.loc[:,"expected_count"]
SRR12514556_1 = SRR12514556_1.loc[:,"expected_count"]
new_SRR12514551_1 = pd.read_csv("SRR12514551_1.genes.results", delimiter='\t')
```

```
new_SRR12514552_1 = pd.read_csv("SRR12514552_1.genes.results", delimiter='\t')
new_SRR12514553_1 = pd.read_csv("SRR12514553_1.genes.results", delimiter='\t')
new_SRR12514554_1 = pd.read_csv("SRR12514554_1.genes.results", delimiter='\t')
new_SRR12514555_1 = pd.read_csv("SRR12514555_1.genes.results", delimiter='\t')
new_SRR12514556_1 = pd.read_csv("SRR12514556_1.genes.results", delimiter='\t')
new_SRR12514551_1 = new_SRR12514551_1.loc[:,["gene_id","TPM"]]
new_SRR12514552_1 = new_SRR12514552_1.loc[:,"TPM"]
new_SRR12514553_1 = new_SRR12514553_1.loc[:,"TPM"]
new_SRR12514554_1 = new_SRR12514554_1.loc[:,"TPM"]
new_SRR12514555_1 = new_SRR12514555_1.loc[:,"TPM"]
new_SRR12514556_1 = new_SRR12514556_1.loc[:,"TPM"]
```

```
new_1 = pd.concat((SRR12514551_1,SRR12514552_1),axis =1)
new_2 = pd.concat((new_1,SRR12514553_1),axis = 1)
new_3 = pd.concat((new_2,SRR12514554_1), axis = 1)
new_4 = pd.concat((new_3,SRR12514555_1), axis = 1)
new_5 = pd.concat((new_4,SRR12514556_1), axis = 1)
```

	gene_id	control	control	control	treated	treated	treated
0	TSPAN6	1601.00	1283.00	1441.00	1595.00	1724.00	2082.00
1	TNMD	0.00	0.00	0.00	0.00	0.00	0.00
2	DPM1	1381.00	1250.00	1349.00	1434.00	1490.00	1690.00
3	SCYL3	350.92	288.34	291.49	237.95	296.50	375.10
4	C1orf112	843.08	712.66	812.51	832.05	855.50	1026.90
...	...	...	...	...	...	...	...
58376	CTD-2643I7.6	0.00	0.00	0.00	0.00	0.00	0.00
58377	CTD-2575K13.8	0.00	0.00	0.00	21.84	0.00	0.00
58378	RP5-931K24.3	839.35	739.77	801.42	1060.51	1144.26	1298.84
58379	CMB9-75A1.1	8.43	11.99	7.78	4.42	8.34	3.97
58380	RP11-87O6.1	0.00	0.00	0.00	0.00	0.00	0.00

Generated count matrix

```
new_1 = pd.concat((SRR12514551_1,SRR12514552_1),axis =1)
new_2 = pd.concat((new_1,SRR12514553_1),axis = 1)
new_3 = pd.concat((new_2,SRR12514554_1), axis = 1)
new_4 = pd.concat((new_3,SRR12514555_1), axis = 1)
new_5 = pd.concat((new_4,SRR12514556_1), axis = 1)
new_5.columns = ["gene_id","control","control","control","treated","treated","treated"]
```

	gene_id	control	control	control	treated	treated	treated
0	TSPAN6	39.63	34.40	37.88	45.82	43.53	43.61
1	TNMD	0.00	0.00	0.00	0.00	0.00	0.00
2	DPM1	70.10	69.98	74.28	85.26	78.62	74.39
3	SCYL3	4.65	4.24	4.14	3.91	4.31	4.67
4	C1orf112	17.27	16.66	17.39	18.24	17.55	19.11
...	...	...	...	...	...	...	...
58376	CTD-2643I7.6	0.00	0.00	0.00	0.00	0.00	0.00
58377	CTD-2575K13.8	0.00	0.00	0.00	0.37	0.00	0.00
58378	RP5-931K24.3	53.39	51.38	54.27	78.92	75.36	70.75
58379	CMB9-75A1.1	0.24	0.40	0.15	0.09	0.16	0.06
58380	RP11-87O6.1	0.00	0.00	0.00	0.00	0.00	0.00

Generated TPM values matrix

	condition	type
control_1	control	paired-end
control_2	control	paired-end
control_3	control	paired-end
treated_1	treated	paired-end
treated_2	treated	paired-end
treated_3	treated	paired-end

Annotation File

DESeq Analysis

We input the count matrix and the annotation files into R and used them as the basis for a DESeq analysis workflow. Our objective was to examine upregulated and downregulated genes, perform statistical tests of significance, and generate useful visualizations.

```
library("DESeq2")
cts <-as.matrix(read.csv("New_Merged_RSEM.csv", row.names="gene_id"))
coldata <- read.csv("Annotation_File_1.csv", row.names=1)
cts = subset(cts, select = -c(X) )
row.names(coldata)
colnames(cts)
```

```
dds <- DESeqDataSetFromMatrix(countData = round(cts), colData = coldata, design =
~condition)
dds <- DESeq(dds)
res <- results(dds)
res
```

Our results here illustrate the log2foldchange for our treatment group vs control group as well as the p values indicating significance of these differences.

```
log2 fold change (MLE): condition treated vs control
Wald test p-value: condition treated vs control
DataFrame with 58381 rows and 6 columns
                  baseMean log2FoldChange     lfcSE      stat      pvalue
                 <numeric>      <numeric> <numeric> <numeric>   <numeric>
ENSG00000000003.14  1606.244      0.2759327 0.0671936  4.106533 4.01643e-05
ENSG00000000005.5      0.000             NA       NA        NA         NA
ENSG00000000419.12  1425.464      0.1730254 0.0714528  2.421536 1.54551e-02
ENSG00000000457.13   302.938     -0.0879454 0.1341158 -0.655743 5.11990e-01
ENSG00000000460.16   841.715      0.1539271 0.0849716  1.811511 7.00617e-02
...                      ...            ...       ...       ...         ...
ENSG00000285498.1    0.00000           NA       NA        NA         NA
ENSG00000285505.1    4.15222       5.479463 3.9173639   1.39876 1.61884e-01
ENSG00000285508.1  973.24945       0.515952 0.0791211   6.52104 6.98222e-11
ENSG00000285509.1    7.40550      -0.843180 0.8266472  -1.02000 3.07728e-01
ENSG00000285513.1    0.00000           NA       NA        NA         NA
                       padj
                  <numeric>
ENSG00000000003.14  0.00029576
ENSG00000000005.5         NA
ENSG00000000419.12  0.05689346
ENSG00000000457.13  0.74537923
ENSG00000000460.16  0.19386982
...                      ...
ENSG00000285498.1         NA
ENSG00000285505.1  3.62783e-01
ENSG00000285508.1  1.11292e-09
ENSG00000285509.1  5.54706e-01
ENSG00000285513.1         NA
```

```
res <- results(dds, contrast=c("condition","treated","control"))
```

```
resultsNames(dds)
"Intercept"                 "condition_treated_vs_control"

resLFC <- lfcShrink(dds, coef="condition_treated_vs_control", type="apeglm")
resLFC
```

Here, we use the lfcshrink function to emphasize LFC for which we are more confident due to having a higher number of counts.

```
log2 fold change (MAP): condition treated vs control
Wald test p-value: condition treated vs control
DataFrame with 58381 rows and 5 columns
                  baseMean log2FoldChange     lfcSE      pvalue        padj
                 <numeric>      <numeric> <numeric>   <numeric>   <numeric>
ENSG00000000003.14  1606.244      0.2661399 0.0666766 4.01643e-05  0.00029576
ENSG00000000005.5     0.000             NA        NA         NA          NA
ENSG00000000419.12  1425.464      0.1646875 0.0701785 1.54551e-02  0.05689346
ENSG00000000457.13   302.938     -0.0729566 0.1229300 5.11990e-01  0.74537923
ENSG00000000460.16   841.715      0.1433266 0.0825927 7.00617e-02  0.19386982
...                      ...            ...       ...         ...         ...
ENSG00000285498.1    0.00000           NA        NA         NA          NA
ENSG00000285505.1    4.15222      0.0179839 0.2951056 1.61884e-01 3.62783e-01
ENSG00000285508.1  973.24945      0.5012355 0.0793472 6.98222e-11 1.11292e-09
ENSG00000285509.1    7.40550     -0.0989356 0.2988431 3.07728e-01 5.54706e-01
ENSG00000285513.1    0.00000           NA        NA         NA          NA
```

We order our results by significance, and also filter by a number of p value thresholds to view results of only a certain significance.

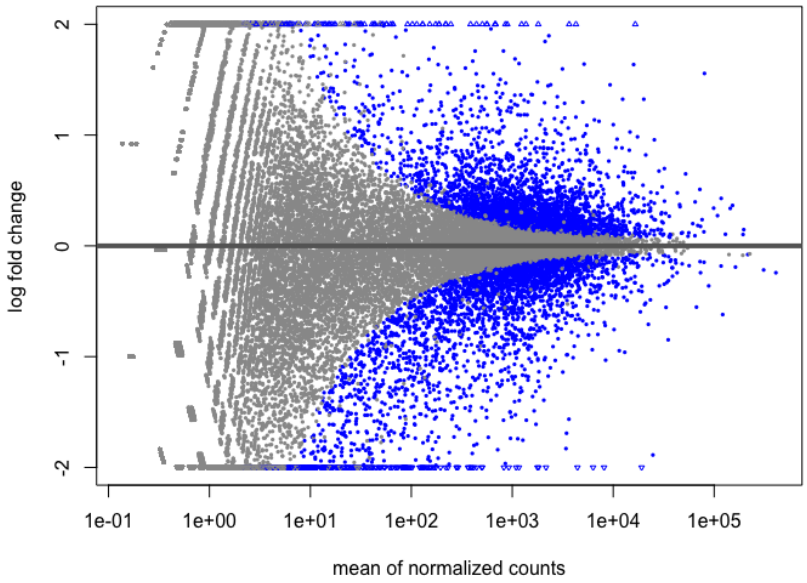
```
resOrdered <- res[order(res$pvalue),]
summary(res)
```

```
out of 30311 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 3230, 11%
LFC < 0 (down)     : 3003, 9.9%
outliers [1]       : 6, 0.02%
low counts [2]     : 10062, 33%
(mean count < 2)
```

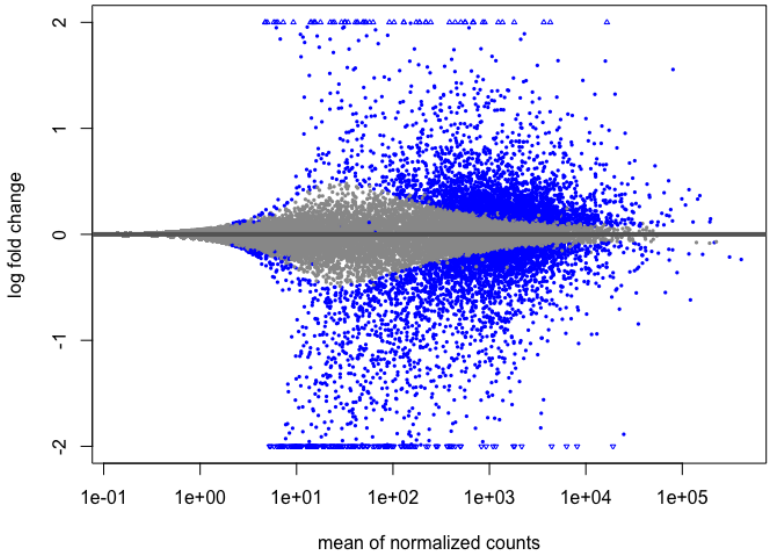
```
sum(res$padj < 0.1, na.rm=TRUE)
6233

res05 <- results(dds, alpha=0.05)
summary(res05)

out of 30311 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 2821, 9.3%
LFC < 0 (down)     : 2616, 8.6%
outliers [1]       : 6, 0.02%
low counts [2]     : 12857, 42%
(mean count < 6)

sum(res05$padj < 0.05, na.rm=TRUE)
5437

library("IHW")
resIHW <- results(dds, filterFun=ihw)
summary(resIHW)

out of 30311 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 3349, 11%
LFC < 0 (down)     : 2973, 9.8%
outliers [1]       : 6, 0.02%

sum(resIHW$padj < 0.1, na.rm=TRUE)
6322

metadata(resIHW)$ihwResult

ihwResult object with 58381 hypothesis tests
Nominal FDR control level: 0.1
Split into 20 bins, based on an ordinal covariate
```

We compare the plots of our results before and after shrinkage, as you can see the shrink function has minimized the log fold change for our noisiest data points.

plotMA(res, ylim=c(-2,2))



plotMA(resLFC, ylim=c(-2,2))



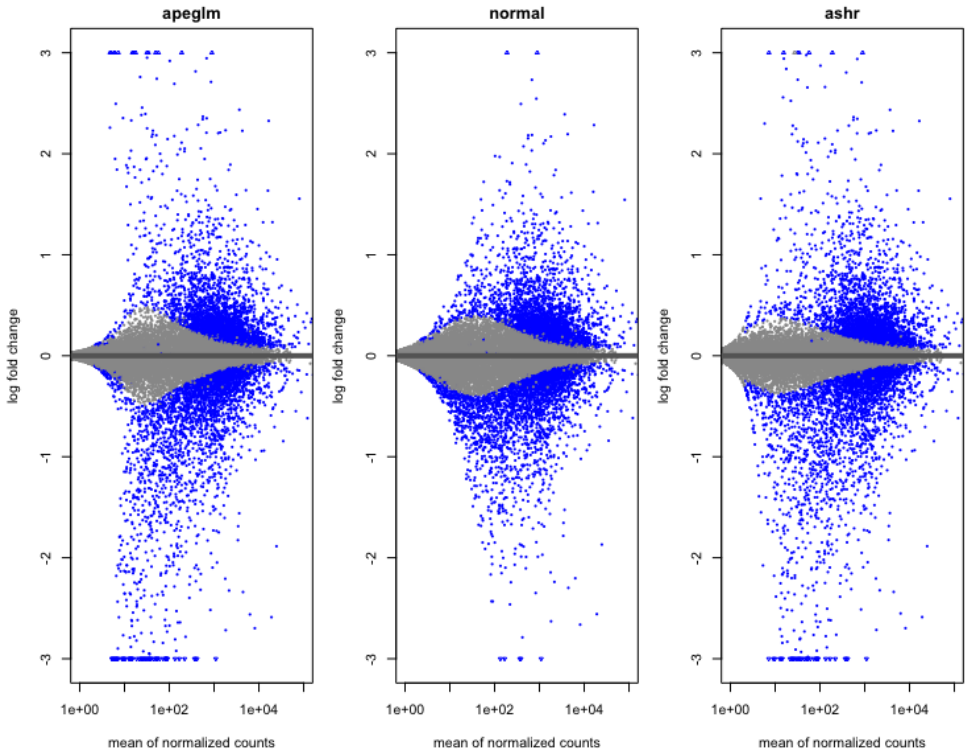
idx <- identify(res$baseMean, res$log2FoldChange)
rownames(res)[idx]

"ENSG00000099864.17" "ENSG00000136811.16" "ENSG00000168301.12"

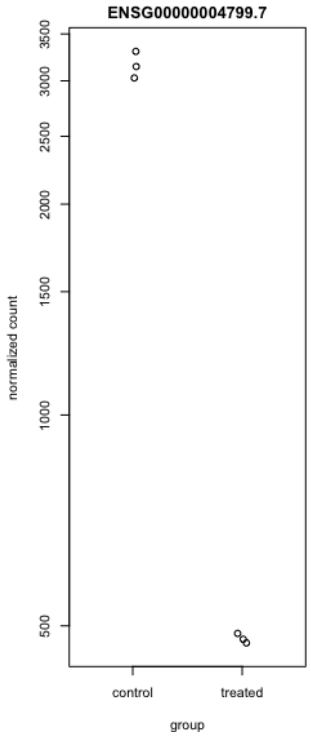
[4] "ENSG00000189057.10" "ENSG00000204196.5"

```
resNorm <- lfcShrink(dds, coef=2, type="normal")
resAsh <- lfcShrink(dds, coef=2, type="ashr")
par(mfrow=c(1,3), mar=c(4,4,2,1))
xlim <- c(1,1e5)
ylim <- c(-3,3)
plotMA(resLFC, xlim=xlim, ylim=ylim, main="apeglm")
plotMA(resNorm, xlim=xlim, ylim=ylim, main="normal")
plotMA(resAsh, xlim=xlim, ylim=ylim, main="ashr")
```

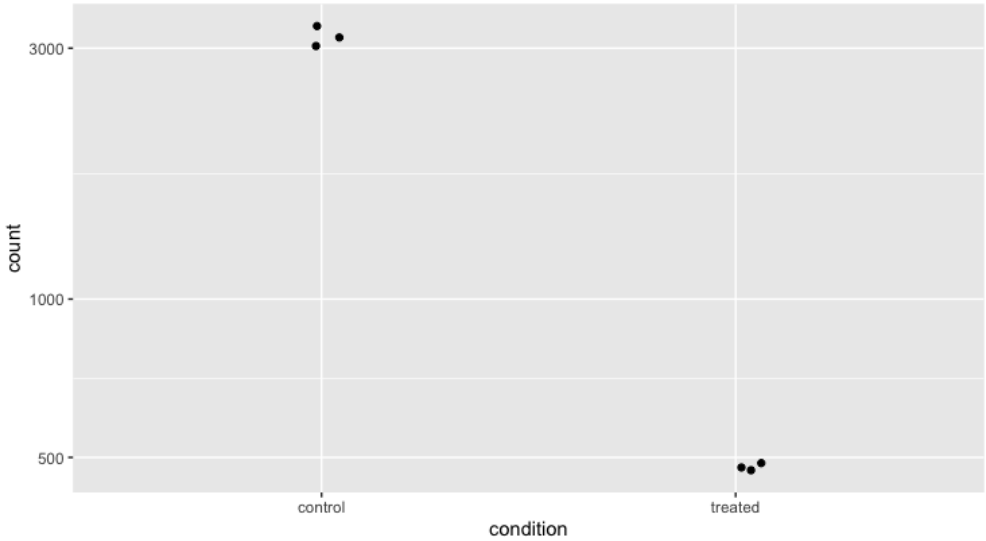
We compare the results of shrinkage using 3 different methods.




```
plotCounts(dds, gene=which.min(res$padj), intgroup="condition")
```




```
d <- plotCounts(dds, gene=which.min(res$padj), intgroup="condition",
                returnData=TRUE)
library("ggplot2")
ggplot(d, aes(x=condition, y=count)) +
  geom_point(position=position_jitter(w=0.1,h=0)) +
  scale_y_log10(breaks=c(500,1000,3000))
```




```
resSig <- subset(resOrdered, padj < 0.05)
resSig
```

```
log2 fold change (MLE): condition treated vs control
Wald test p-value: condition treated vs control
DataFrame with 5344 rows and 6 columns
                   baseMean log2FoldChange     lfcSE      stat    pvalue      padj
                  <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
ENSG00000004799.7    1818.95       -2.72087 0.0690460  -39.4065         0         0
ENSG00000021826.15   8118.57       -2.69985 0.0437869  -61.6588         0         0
ENSG00000035862.12  15931.30        1.63625 0.0331581   49.3469         0         0
ENSG00000095752.6    4274.14        2.22886 0.0466205   47.8085         0         0
ENSG00000107984.9    4423.65       -2.34178 0.0507385  -46.1539         0         0
...                      ...            ...       ...       ...       ...       ...
ENSG00000154723.12  1986.123       0.147989 0.0596848   2.47950 0.0131565 0.0498687
ENSG00000164086.9    327.421       0.316795 0.1277667   2.47948 0.0131576 0.0498687
ENSG00000155329.11   330.989      -0.332712 0.1342146  -2.47896 0.0131767 0.0499253
ENSG00000145241.10   453.352       0.275697 0.1112158   2.47894 0.0131774 0.0499253
ENSG00000179627.9    145.834      -0.490096 0.1977139  -2.47881 0.0131820 0.0499332
```

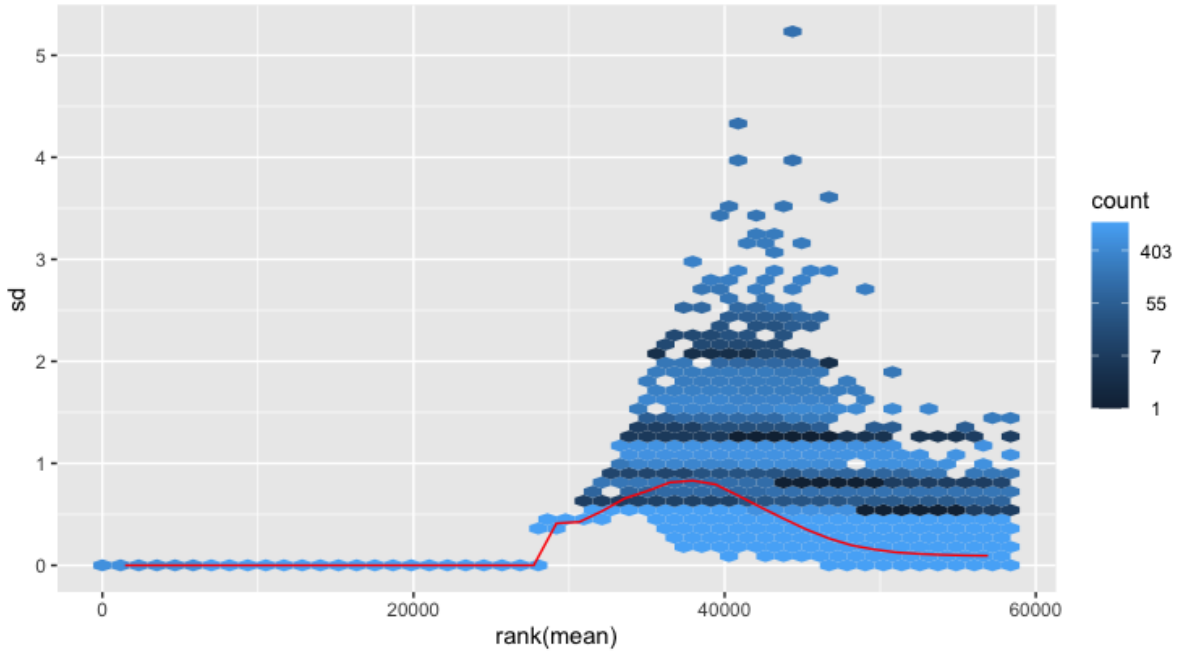
We use a variance stabilizing transformation and regularized log transformation in order to generate data with normalized variance for clustering.

```
write.csv(as.data.frame(resSig),
        file="final_condition_treated_results.csv")
vsd <- vst(dds, blind=FALSE)
rld <- rlog(dds, blind=FALSE)
head(assay(vsd), 3)
```

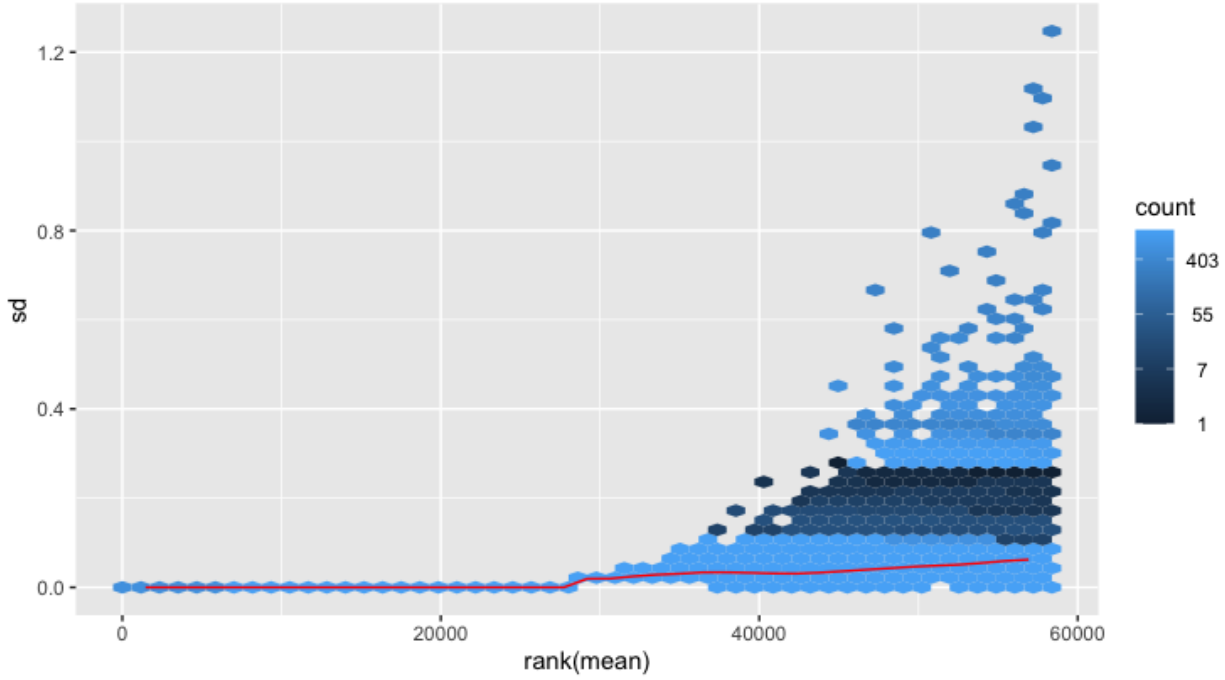
```
 control_1 control_2 control_3 treated_1 treated_2 treated_3
ENSG00000000003.14 11.621855 11.522849 11.580291 11.747813 11.720291 11.711594
ENSG00000000005.5   9.886847  9.886847  9.886847  9.886847  9.886847  9.886847
ENSG00000000419.12 11.510127 11.503651 11.530665 11.662147 11.605012 11.549164
```

Here, we plot mean vs SD rowwise for each of our transformations.

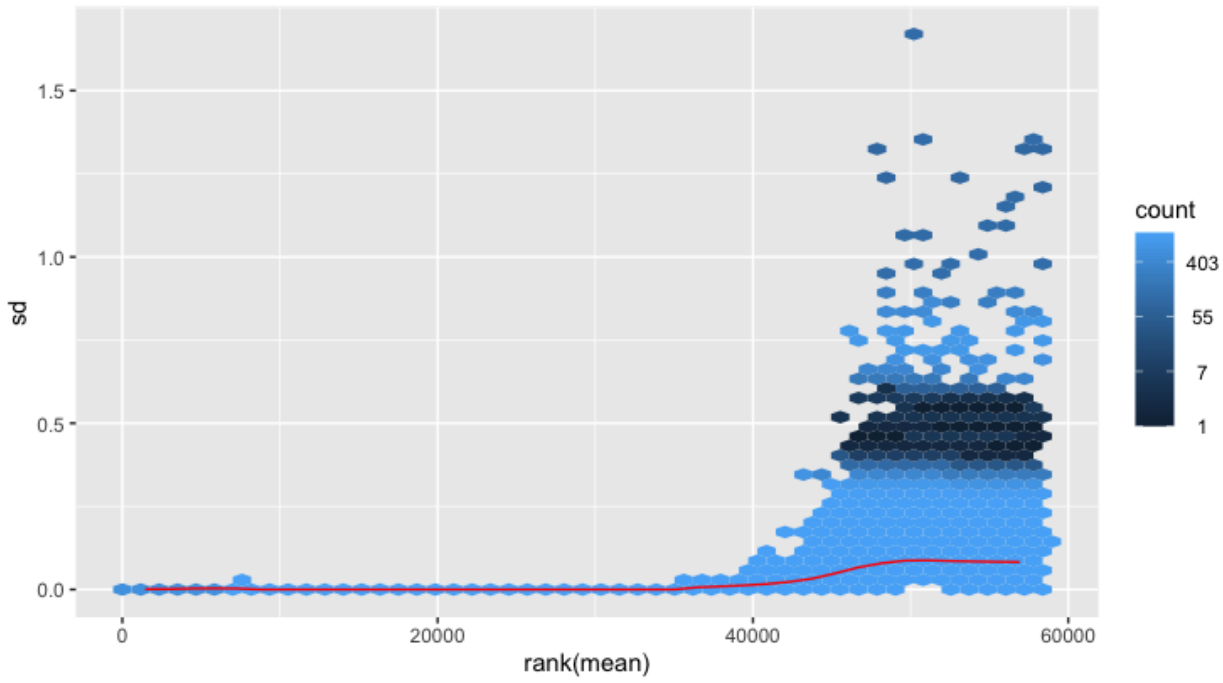
```
ntd <- normTransform(dds)
library("vsn")
meanSdPlot(assay(ntd))
```



meanSdPlot(assay(vsd))

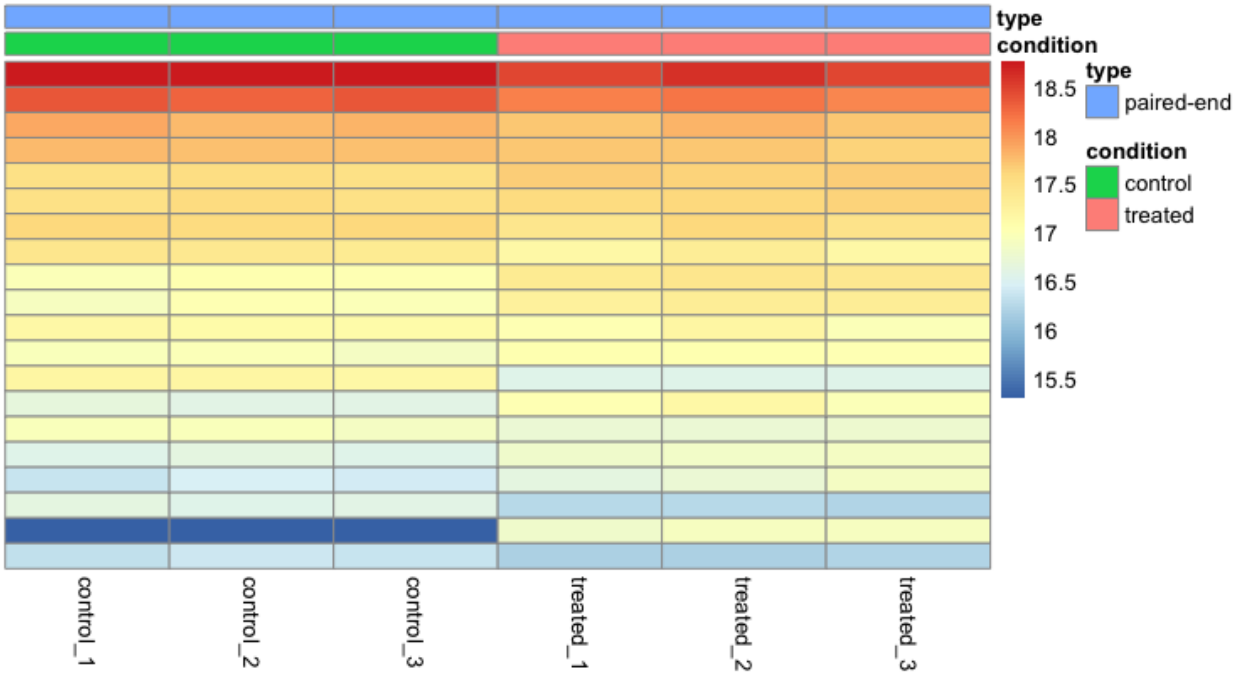


meanSdPlot(assay(rld))

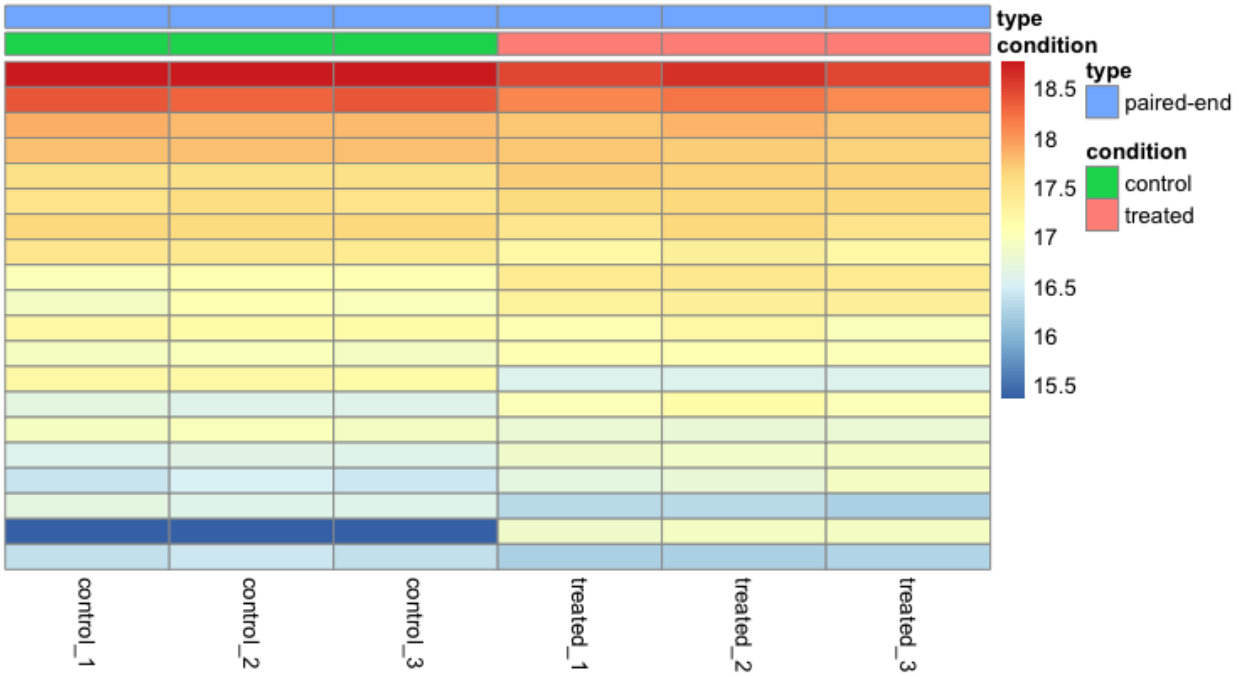


Now, we produce heat maps contrasting our treatment and control groups using all three transformed count matrices.

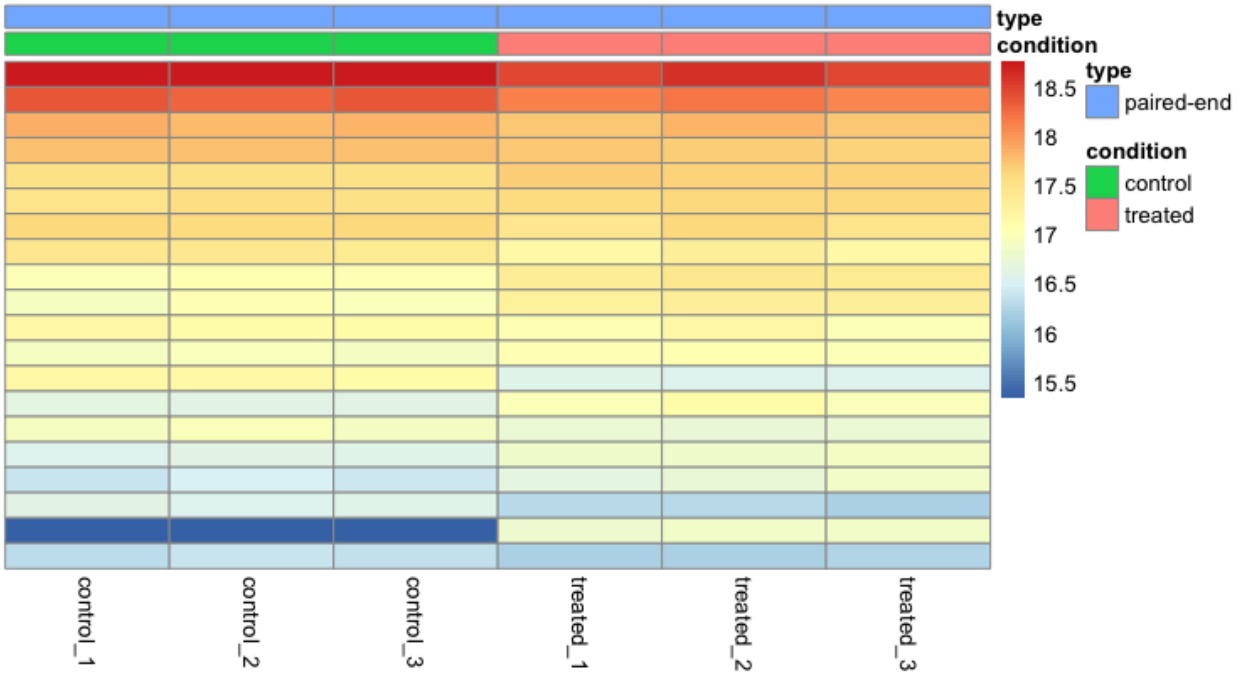
```
library("pheatmap")
select <- order(rowMeans(counts(dds,normalized=TRUE)),
                decreasing=TRUE)[1:20]
df <- as.data.frame(colData(dds)[,c("condition","type")])
pheatmap(assay(ntd)[select,], cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df)
```



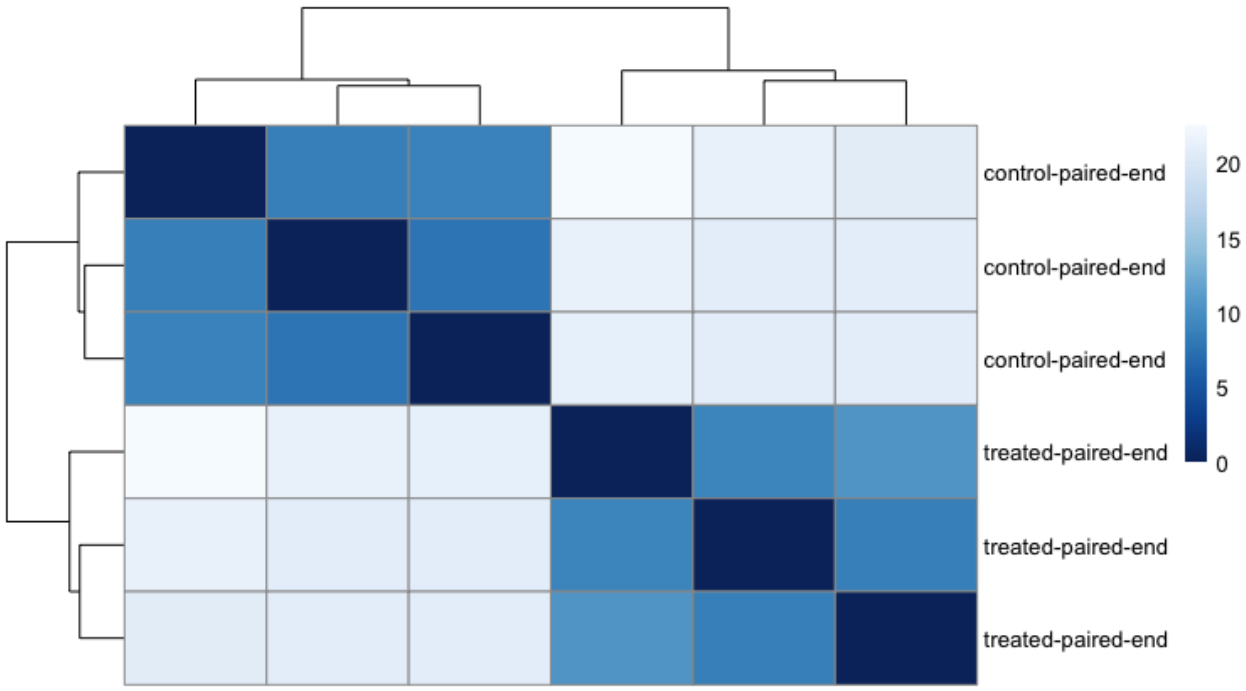
pheatmap(assay(vsd)[select,], cluster_rows=FALSE, show_rownames=FALSE,
cluster_cols=FALSE, annotation_col=df)



pheatmap(assay(rld)[select,], cluster_rows=FALSE, show_rownames=FALSE,
cluster_cols=FALSE, annotation_col=df)

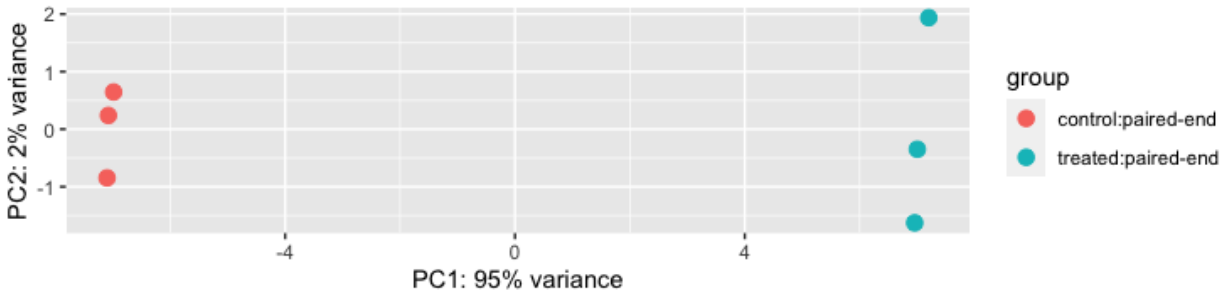



```
sampleDists <- dist(t(assay(vsd)))
library("RColorBrewer")
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(vsd$condition, vsd$type, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
pheatmap(sampleDistMatrix,
        clustering_distance_rows=sampleDists,
        clustering_distance_cols=sampleDists,
        col=colors)
```

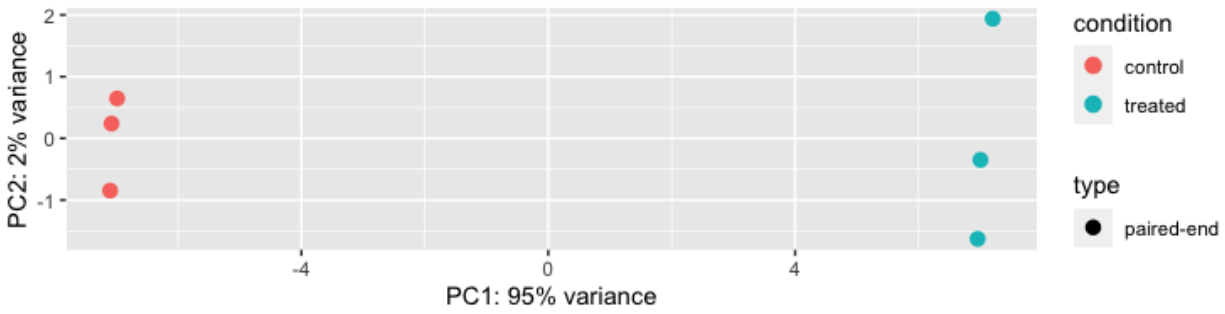


Below, we plot the principal component analysis for our treatment and control groups.

```
plotPCA(vsd, intgroup=c("condition", "type"))
```



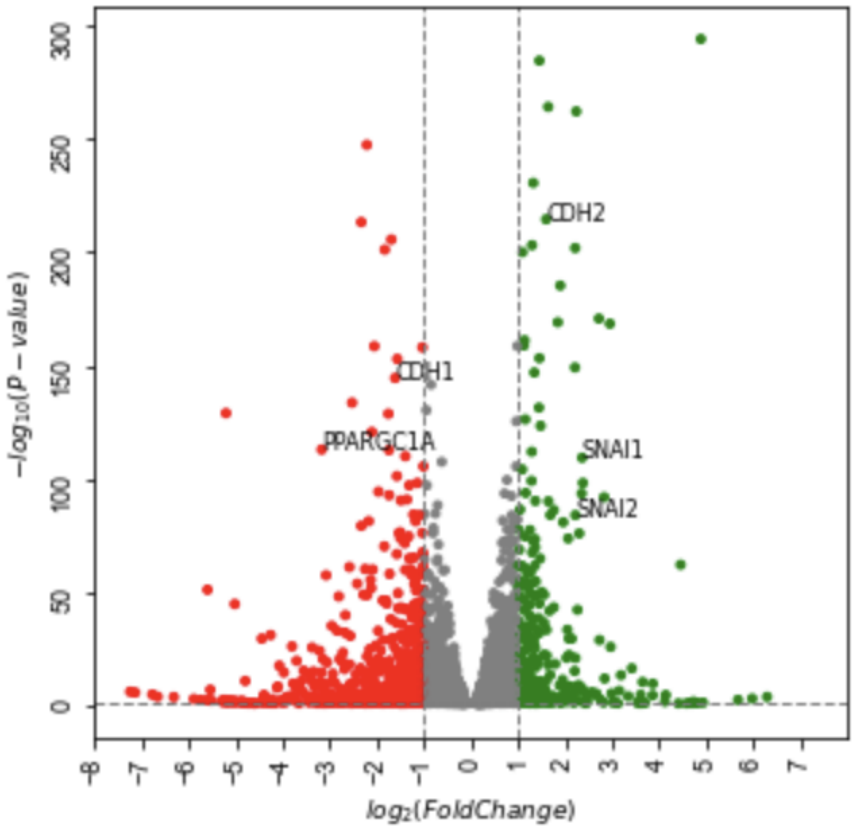

```
pcaData <- plotPCA(vsd, intgroup=c("condition", "type"), returnData=TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"))
ggplot(pcaData, aes(PC1, PC2, color=condition, shape=type)) +
  geom_point(size=3) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  coord_fixed()
```



The important results to pull from the DESeq analysis is that the control groups and treatment groups are very different from one another but within their own respective groups they are very similar to one another. This conclusion can be drawn from the sample to sample heat matrix and PCA analysis graphs. Additionally, from the MA plots and summaries of res and res05, it is clear that there are a lot of upregulated and downregulated genes in this study. To determine which genes are significantly upregulated or downregulated, a volcano plot will need to be created to determine a cutoff for the log2FoldChange values.

Gene Ontology Analysis

Volcano Plot



Up regulated log2foldchange >1 (cutoff used in paper, confirmed in volcano plot)
Down regulated log2foldchange < -1 (cutoff used in paper, confirmed in volcano plot)
P value of 0.05 (value used in paper)
297 upregulated genes
728 down regulated genes

Top 5 enriched pathways for each ontology term show below

Down regulated BP

	Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
0	GOTERM_BP_DIRECT	GO:0007275~multicellular organism development	23	3.668262	5.008784e-07	PDGFRA, NFE2, CSF3, EHF, EPHA7, FLT1, HLF, EPH...	570	217	19308	3.590298	0.001609	0.001610	0.001597
1	GOTERM_BP_DIRECT	GO:0007156~homophilic cell adhesion via plasma...	18	2.870813	1.376498e-05	CADM4, PCDH10, PCDH7, VSTM2L, L1CAM, CDH8, CEA...	570	172	19308	3.544920	0.043277	0.021212	0.021047
2	GOTERM_BP_DIRECT	GO:0001525~angiogenesis	22	3.508772	1.979946e-05	VAV3, ACVRL1, FLT1, GPR15, CXCL8, SYK, PLXND1,...	570	252	19308	2.957226	0.061654	0.021212	0.021047
3	GOTERM_BP_DIRECT	GO:0030593~neutrophil chemotaxis	12	1.913876	2.784119e-05	VAV3, CXCL10, PREX1, CXCL11, CXADR, CXCL8, SYK...	570	82	19308	4.957125	0.085596	0.022370	0.022196
4	GOTERM_BP_DIRECT	GO:1990573~potassium ion import across plasma ...	9	1.435407	4.460691e-05	KCNK5, KCNH2, SLC12A2, KCNJ11, KCNJ15, KCNJ16,...	570	45	19308	6.774737	0.133566	0.028673	0.028450

Down regulated CC

	Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
0	GOTERM_CC_DIRECT	GO:0005886~plasma membrane	229	36.523126	1.746960e-14	GLDC, SLC23A1, AQP3, ICAM1, PREX1, BAIAP2L2, R...	591	5071	20562	1.571158	8.366641e-12	8.385407e-12	7.686623e-12
1	GOTERM_CC_DIRECT	GO:0005887~integral component of plasma membrane	95	15.151515	5.949984e-13	KCNK5, ACVRL1, PTPRU, CHRM3, CD40, TENM2, FLT1...	591	1508	20562	2.191796	2.855849e-10	1.427996e-10	1.308996e-10
2	GOTERM_CC_DIRECT	GO:0005615~extracellular space	109	17.384370	1.008913e-11	FCGBP, CSF3, F13A1, XYLT1, ICAM1, CXCL16, IL13...	591	1938	20562	1.956819	4.842793e-09	1.614261e-09	1.479739e-09
3	GOTERM_CC_DIRECT	GO:0005576~extracellular region	110	17.543860	1.198188e-09	CSF3, HHIP, F13A1, IFI30, CXCL16, CDH1, EVA1C,...	591	2129	20562	1.797608	5.751299e-07	1.437825e-07	1.318006e-07
4	GOTERM_CC_DIRECT	GO:0016323~basolateral plasma membrane	26	4.146730	2.752341e-08	CHRM3, SLC43A3, C5AR1, TGFA, ARRB2, AQP3, ERBB...	591	239	20562	3.784891	1.321115e-05	2.443678e-06	2.240038e-06

Down regulated MF

	Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
0	GOTERM_MF_DIRECT	GO:0005102~receptor binding	30	4.784689	0.000017	TENM2, BEX1, FGL1, ARRB2, C5, IFNL2, IFNL1, WI...	582	397	18869	2.449947	0.013867	0.013964	0.013779
1	GOTERM_MF_DIRECT	GO:0008009~chemokine activity	9	1.435407	0.000132	CXCL10, CXCL11, C5, CXCL8, CCL20, CCL5, CXCL3,...	582	50	18869	5.835773	0.103973	0.054889	0.054162
2	GOTERM_MF_DIRECT	GO:0030506~ankyrin binding	6	0.956938	0.000287	SPTBN4, KCNJ11, CDH1, KCNQ2, CACNA1D, RHCG	582	20	18869	9.726289	0.212382	0.064218	0.063368
3	GOTERM_MF_DIRECT	GO:0005543~phospholipid binding	13	2.073365	0.000332	SPTBN4, UNC13A, SYT16, RASAL1, SNAP91, PREX1, ...	582	120	18869	3.512271	0.241333	0.064218	0.063368
4	GOTERM_MF_DIRECT	GO:0001228~transcriptional activator activity,...	30	4.784689	0.000388	FOXA1, MESP1, CEBPA, EHF, HLF, ONECUT3, HNF4G,...	582	475	18869	2.047640	0.275885	0.064218	0.063368

Upregulated BP

	Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
0	GOTERM_BP_DIRECT	GO:0007155~cell adhesion	22	9.090909	0.000002	LAMA4, NEXN, PTPRK, THY1, LOXL2, CDH6, EFNB2, ...	223	555	19308	3.432117	0.002970	0.002975	0.002932
1	GOTERM_BP_DIRECT	GO:0030198~extracellular matrix organization	11	4.545455	0.000022	CCDC80, COL5A1, CRISPLD2, ABI3BP, MMP2, COL5A3...	223	165	19308	5.772197	0.034160	0.016560	0.016321
2	GOTERM_BP_DIRECT	GO:0007156~homophilic cell adhesion via plasma...	11	4.545455	0.000031	CDH6, CDH4, TENM3, CDH2, CADM1, CDH10, CDH11, ...	223	172	19308	5.537282	0.048466	0.016560	0.016321
3	GOTERM_BP_DIRECT	GO:0030199~collagen fibril organization	7	2.892562	0.000064	GREM1, COL5A1, LUM, COL5A3, COL5A2, LOXL2, DDR2	223	60	19308	10.101345	0.097512	0.025649	0.025279
4	GOTERM_BP_DIRECT	GO:0032331~negative regulation of chondrocyte ...	5	2.066116	0.000151	GREM1, EFEMP1, SNAI2, PTHLH, NKX3-2	223	24	19308	18.038117	0.213648	0.048066	0.047372

Upregulated CC

	Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
)	GOTERM_CC_DIRECT	GO:0005576~extracellular region	56	23.140496	3.596355e-09	SPARC, CGB8, C4BPB, IL1RAP, EFEMP1, BMPER, GLI...	234	2129	20562	2.311330	9.889971e-07	9.889976e-07	9.566304e-07
1	GOTERM_CC_DIRECT	GO:0031012~extracellular matrix	18	7.438017	7.651396e-09	FBN2, CCBE1, LRRC15, LUM, MMP2, WNT5A, EFEMP1,...	234	260	20562	6.083432	2.104132e-06	1.052067e-06	1.017636e-06
2	GOTERM_CC_DIRECT	GO:0005615~extracellular space	47	19.421488	1.052008e-06	MOXD1, LRRC15, SPARC, HTRA1, STC1, C4BPB, FSTL...	234	1938	20562	2.131050	2.892605e-04	9.643405e-05	9.327803e-05
3	GOTERM_CC_DIRECT	GO:0005788~endoplasmic reticulum lumen	16	6.611570	2.343488e-06	IGFBP3, WNT5A, FSTL1, BACE1, CDH2, COL5A1, EVA...	234	306	20562	4.594604	6.442525e-04	1.611148e-04	1.558420e-04
4	GOTERM_CC_DIRECT	GO:0005886~plasma membrane	85	35.123967	6.415200e-05	SNAP25, ITK, VIPR1, SPARC, TENM3, CD82, DYSF, ...	234	5071	20562	1.472905	1.748765e-02	3.528360e-03	3.412887e-03

Upregulated MF

	Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
0	GOTERM_MF_DIRECT	GO:0005201~extracellular matrix structural con...	16	6.611570	6.274666e-11	FBN2, TECTB, SPARC, LUM, LAMA4, EFEMP1, COL5A1...	218	138	18869	10.035368	2.459670e-08	2.459669e-08	2.440845e-08
1	GOTERM_MF_DIRECT	GO:0005518~collagen binding	9	3.719008	1.099217e-06	LRRC15, CCBE1, SPARC, LUM, ABI3BP, COL5A3, ITG...	218	68	18869	11.455815	4.308003e-04	2.154465e-04	2.137976e-04
2	GOTERM_MF_DIRECT	GO:0005509~calcium ion binding	26	10.743802	1.919915e-06	FBN2, ANXA8L1, SPARC, DYSF, RASGRP1, FSTL1, SL...	218	751	18869	2.996579	7.523242e-04	2.508689e-04	2.489490e-04
3	GOTERM_MF_DIRECT	GO:0005178~integrin binding	11	4.545455	1.761700e-05	CXCL12, COL5A1, LGALS12, ITGA11, TIMP2, ITGAV,...	218	161	18869	5.913699	6.882133e-03	1.726466e-03	1.713253e-03
4	GOTERM_MF_DIRECT	GO:0001968~fibronectin binding	5	2.066116	4.752294e-04	LRRC15, CCDC80, MMP2, IGFBP3, ITGAV	218	32	18869	13.524226	1.700039e-01	3.725799e-02	3.697285e-02

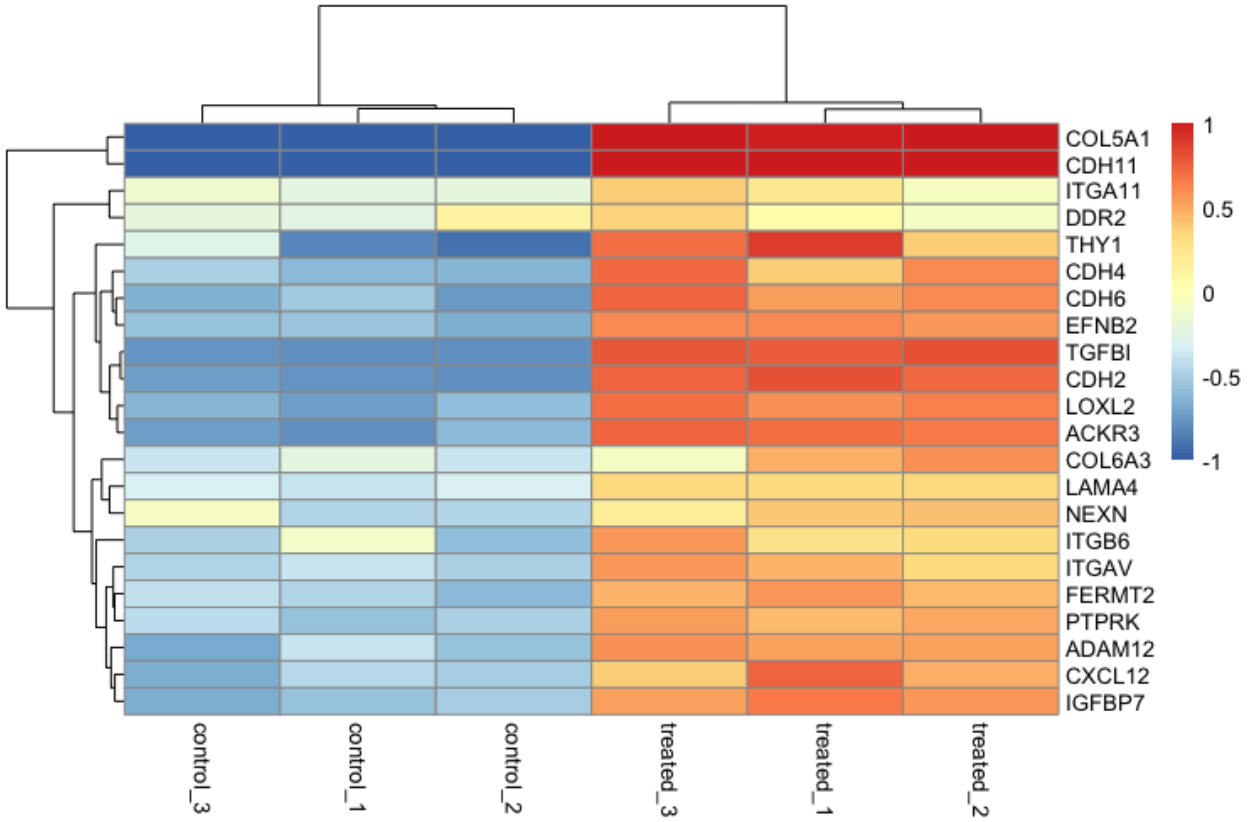
From the volcano plot we see using a cutoff value of 1 and -1 for our gene ontology will give us the significantly expressed genes. After filtering by those cutoffs, gene ontology was performed on DAVID and the top 5 enriched pathways for each gene ontology term for the upregulated and downregulated terms were obtained. Looking at the downregulated terms,

we see a lot of the molecular function terms have binding terms and this aligns with what occurs during EMT as cell to cell adhesion is decreased during EMT and binding most likely plays a role in that process. Additionally, the most enriched down regulated biological process term is multicellular organism development which also makes sense with what occurs during EMT. Oddly enough though, looking at the upregulated biological process terms, cell adhesion appears and in the molecular function terms a few binding terms pop up again which do not align with how EMT works. Further investigation should be conducted to determine the exact function of this term.

Differentially Expressed Gene Heatmap:

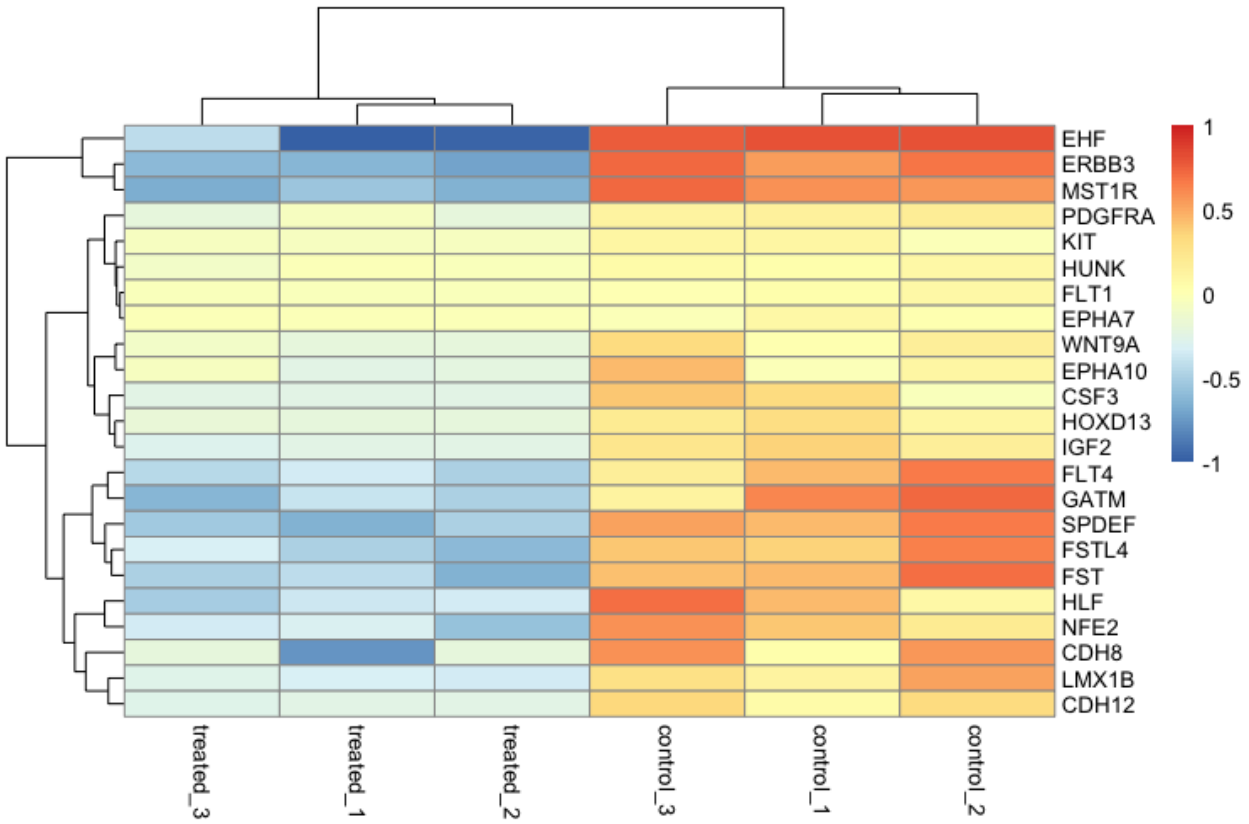
```
install.packages ("pheatmap")
library(tidyverse)
library("pheatmap")
range <- 1
genes_1 <- c("LAMA4", "NEXN", "PTPRK", "THY1", "LOXL2", "CDH6", "EFNB2",
"CDH4", "CXCL12", "CDH2", "COL5A1", "ADAM12", "ITGA11", "CDH11", "COL6A3",
"ACKR3", "IGFBP7","ITGAV", "TGFBI", "ITGB6", "FERMT2", "DDR2")

cols_1 <- c("control_1","control_2","control_3","treated_1","treated_2","treated_3")
A <- read_csv("upregulated_tpm_1")
A <- A[-c(1)]
rownames(A) <- genes_1
colnames(A) <- cols_1
pheatmap(A, breaks=seq(-range, range, length.out = 100))
```

Heatmap of genes in top enriched upregulated BP term (Cell Adhesion)




```
range <- 1
genes_2 <- c("PDGFRA", "NFE2", "CSF3", "EHF", "EPHA7", "FLT1", "HLF", "EPHA10",
"FST",  "FLT4", "IGF2", "HUNK", "HOXD13", "MST1R", "WNT9A", "FSTL4", "CDH8",
"GATM", "ERBB3", "KIT", "CDH12", "LMX1B", "SPDEF")

cols_2 <- c("control_1","control_2","control_3","treated_1","treated_2","treated_3")
B <- read_csv("downregulated_tpm_1")
B <- B[-c(1)]
rownames(B) <- genes_2
colnames(B) <- cols_2
pheatmap(B, breaks=seq(-range, range, length.out = 100))
```

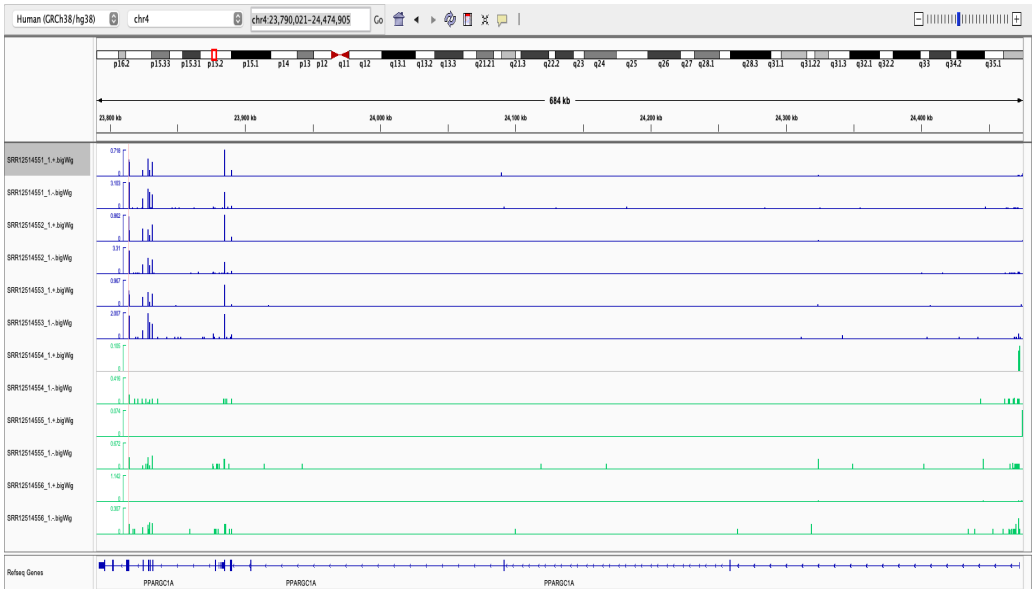
Heatmap of genes in top enriched downregulated BP term (Multicellular organism development)



From the differential gene expression analysis heatmaps of the most enriched upregulated and downregulated biological process term, we see the varying levels of upregulation and downregulation of the genes responsible for these terms. One thing to take note of is that in the paper CDH2 was said to be an upregulated gene and we see that gene appearing here in the upregulated table above.

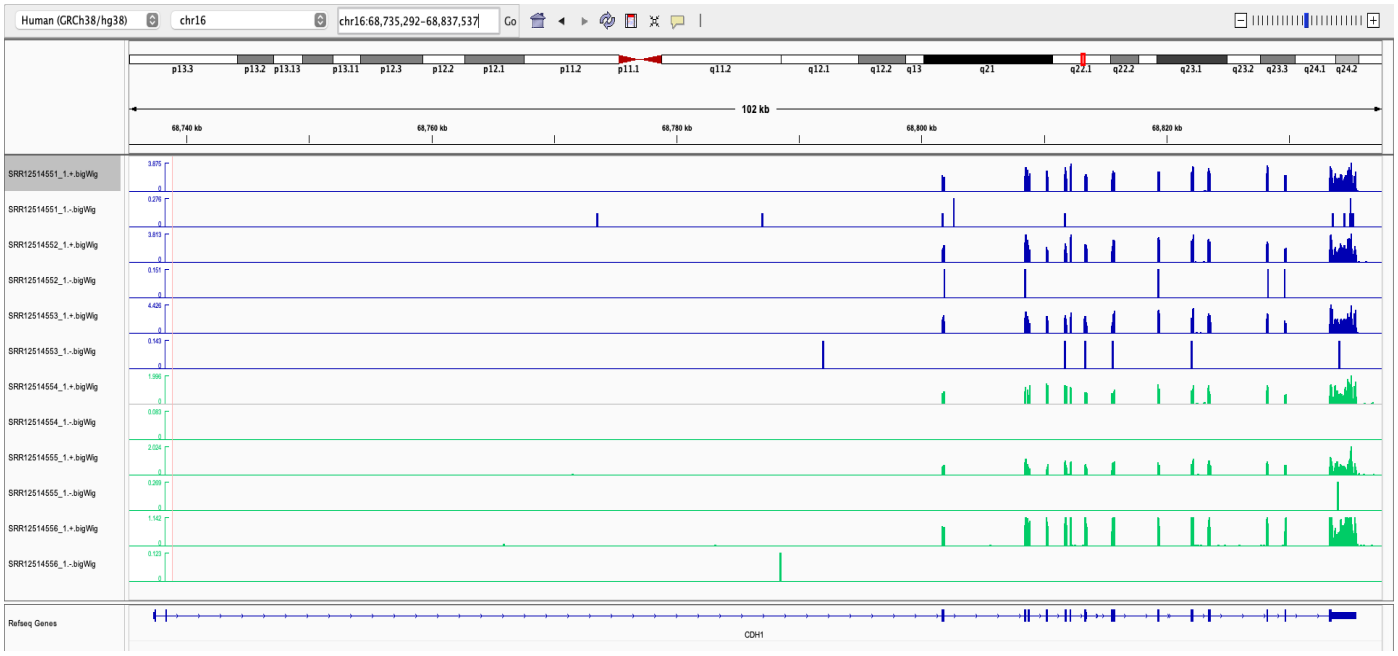
IGV:

PPARGC1A (PGC1α)



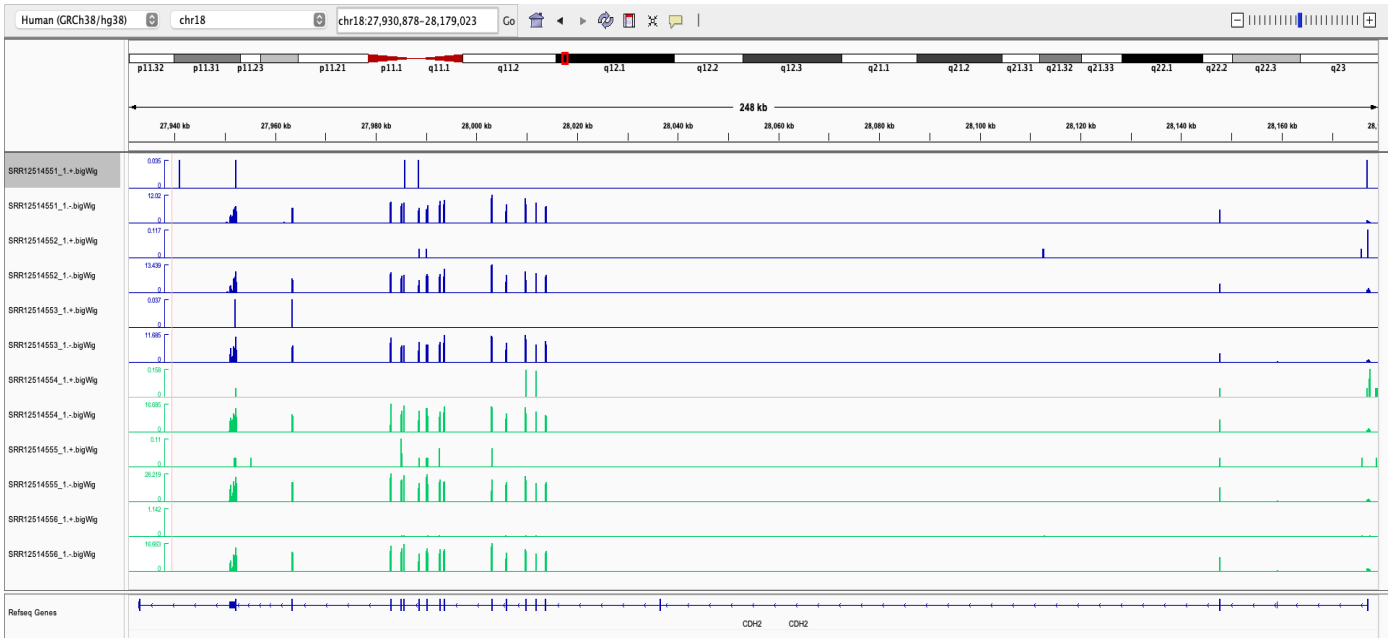
	gene_id	control	control	control	treated	treated	treated
3708	PPARGC1A	12.08	12.61	12.5	0.72	0.94	0.94

CDH1



	gene_id	control	control	control	treated	treated	treated
555	CDH1	41.46	40.87	44.78	12.61	10.94	12.57

CDH2



	gene_id	control	control	control	treated	treated	treated
12860	CDH2	21.26	20.65	21.58	64.64	61.64	62.57

To summarize the IGV results, we see that PPARGC1A, the gene knocked down in the treatment, shows a downregulation based on the IGV screenshot and TPM values as expected. Next, we looked into gene CDH1. In the paper, this gene was mentioned to be an epithelial marker and therefore should be downregulated. This is exactly what we see from the CDH1 IGV screenshot and TPM values. Lastly, we looked into the gene CDH2. This gene was mentioned to be a mesenchymal marker in the paper and therefore we should observe an upregulation of this gene. Based on the IGV screenshot and TPM values we can see that this is exactly the case. Overall, our results matched the results from the paper.

Summary and Discussion

To summarize, our results from DESeq analysis, gene ontology analysis, and IGV visualizations align with those seen in the paper this project is based on. From the DESeq analysis, we observed a large difference between the control and treatment groups and this was expected given the PPARGC1A gene was knocked down in the treatment group. Additionally, a large number of upregulated and downregulated genes were observed from the MA plots and summaries of the res and res05. This meant filtering by a p-value of 0.05 and log2FoldChanges of 1 and -1, as was done in the paper to obtain the significantly expressed genes. Gene ontology results aligned with how EMT was described, showing a downregulation in multicellular

organism development and numerous binding terms. However, an interesting result we observed was an upregulation in cell adhesion and numerous other binding terms. To determine the exact function of these terms, further analysis should be conducted and is something future studies should look into. Still, when looking at specific EMT genes of interest like CDH1, CDH2, VIM, ITGA5, and SNAI1, our results were as expected. Lastly, IGV visualizations were created to check if key genes mentioned in the paper to be upregulated or downregulated matched our results. The genes we chose to study were PPARGC1A, CDH1, and CDH2. The corresponding upregulation or downregulation for each gene based on the IGV screenshots and TPM values matched with what was observed in the paper.

The results of this project supports the idea proposed in the paper that PGC1α is an opposing regulator of lung cancer metastasis and that this pathway is a potential target for future therapies.

References

Chaffer, C.L.; San Juan, B.P.; Lim, E.; Weinberg, R.A. EMT, cell plasticity and metastasis. Cancer Metastasis Rev. **2016**, 35, 645–654.

Torrano, V.; Valcarcel-Jimenez, L.; Cortazar, A.R.; Liu, X.; Urosevic, J.; Castillo-Martin, M.; Fernandez-Ruiz, S.; Morciano, G.; Caro-Maldonado, A.; Guiu, M.; et al. The metabolic co-regulator PGC1alpha suppresses prostate cancer metastasis. Nat. Cell Biol. **2016**, 18, 645–656.

LeBleu, V.S.; O'Connell, J.T.; Gonzalez Herrera, K.N.; Wikman, H.; Pantel, K.; Haigis, M.C.; de Carvalho, F.M.; Damascena, A.; Domingos Chinen, L.T.; Rocha, R.M.; et al. PGC-1alpha mediates mitochondrial biogenesis and oxidative phosphorylation in cancer cells to promote metastasis. Nat. Cell Biol. **2014**, 16, 992–1003, 1001–1015.

Luo, C.; Lim, J.H.; Lee, Y.; Granter, S.R.; Thomas, A.; Vazquez, F.; Widlund, H.R.; Puigserver, P. A PGC1alpha-mediated transcriptional axis suppresses melanoma metastasis. Nature **2016**, 537, 422–426.

Jiang, W.G.; Douglas-Jones, A.; Mansel, R.E. Expression of peroxisome-proliferator activated receptor-gamma (PPARgamma) and the PPARgamma co-activator, PGC-1, in human breast cancer correlates with clinical outcomes. Int. J. Cancer **2003**, 106, 752–757.

LaGory, E.L.; Wu, C.; Taniguchi, C.M.; Ding, C.C.; Chi, J.T.; von Eyben, R.; Scott, D.A.; Richardson, A.D.; Giaccia, A.J. Suppression of PGC-1alpha Is Critical for Reprogramming Oxidative Metabolism in Renal Cell Carcinoma. Cell Rep. **2015**, 12, 116–127.

Shoag, J.; Haq, R.; Zhang, M.; Liu, L.; Rowe, G.C.; Jiang, A.; Koulisis, N.; Farrel, C.; Amos, C.I.; Wei, Q.; et al. PGC-1 coactivators regulate MITF and the tanning response. Mol. Cell **2013**, 49, 145–157.

Vazquez, F.; Lim, J.H.; Chim, H.; Bhalla, K.; Girnun, G.; Pierce, K.; Clish, C.B.; Granter, S.R.; Widlund, H.R.; Spiegelman, B.M.; et al. PGC1alpha expression defines a subset of human melanoma tumors with increased mitochondrial capacity and resistance to oxidative stress. Cancer Cell **2013**, 23, 287–301.

Author Contributions

Cody - 50 %
Uzair - 50 %