INTERACT

Using GEOquery to load in phenodata associated with count data file

```
library(GEOquery)
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Setting options('download.file.method.GEOquery'='auto')
```

```
## Setting options('GEOquery.inmemory.gpl'=FALSE)
```

```
gse = getGEO("GSE124161") #unfortunately the data is not included in this file, so we need to load that in separately.
```

```
## Found 1 file(s)
```

```
## GSE124161_series_matrix.txt.gz
```

Loading in dataset previously downloaded from NCBI-GEO directly from stored computer location.

```
GSE124161_readcount <- read.delim("~/NYU/BIGY-7633 Transcriptomics/project/GSE124161_readcoun
t.txt", row.names=1)
```

readcount file needs to have the metadata associated with the sample ID names

since they re-ordered the count data, it is different than their metadata file, we need to re-create the metadata to fit the revised order that they utilized in the count data file.

Retreiving metadata from series_matrix file, int the order of the count data file

```
pheno_data <-gse[["GSE124161_series_matrix.txt.gz"]]@phenoData@data[["title"]]
D0 <-pheno_data[1:3]
W1T<-pheno_data[c(4,6,8)]
W1C<-pheno_data[c(5,7,9)]
W2T<-pheno_data[c(10,12,14)]
W2C<-pheno_data[c(11,13,15)]
W3T<-pheno_data[c(16,18,20)]
W3C<-pheno_data[c(17,19,21)]
W4T<-pheno_data[c(22,24,26)]
W4C<-pheno_data[c(23,25,27)]
W5T<-pheno_data[c(28,30,32)]
W5C<-pheno_data[c(29,31,33)]
W6T<-pheno_data[c(34,36,38)]
W6C<-pheno_data[c(35,37,39)]
W8T<-pheno_data[c(40,42,44)]
W8C<-pheno_data[c(41,43,45)]
count_pheno <-c(D0,W1T, W1C, W2T, W2C, W3T, W3C, W4T, W4C, W5T, W5C, W6T, W6C, W8T, W8C)
count_pheno
```

```
##  [1] "GT01 D0"                          "GT09 D0"
##  [3] "GT19 D0"                          "GT57 Week 1 Treated [GT57_1_T]"
##  [5] "GT58 Week 1 Treated [GT58_1_T]"   "GT59 Week 1 Treated [GT59_1_T ]"
##  [7] "GT57 Week 1 Control  [GT57_1_C ]" "GT58 Week 1 Control [GT58_1_C ]"
##  [9] "GT59 Week 1 Control [GT59_1_C]"   "GT51 Week 2 Treated [GT51_2_T]"
## [11] "GT52 Week 2 Treated [GT52_2_T]"   "GT53 Week 2 Treated [GT53_2_T]"
## [13] "GT51 Week 2 Control [GT51_2_C]"   "GT52 Week 2 Control [GT52_2_C]"
## [15] "GT53 Week 2 Control [GT53_2_C ]"  "GT45 Week 3 Treated [GT45_3_T]"
## [17] "GT46 Week 3 Treated [GT46_3_T]"   "GT47 Week 3 Treated [GT47_3_T]"
## [19] "GT45 Week 3 Control [GT45_3_C]"   "GT46 Week 3 Control [GT46_3_C]"
## [21] "GT47 Week 3 Control [GT47_3_C]"   "GT17 Week 4 Treated [GT17_4_T]"
## [23] "GT18 Week 4 Treated [GT18_4_T]"   "GT19 Week 4 Treated [GT19_4_T]"
## [25] "GT17 Week 4 Control [GT17_4_C]"   "GT18 Week 4 Control [GT18_4_C]"
## [27] "GT19 Week 4 Control [GT19_4_C]"   "GT39 Week 5 Treated [GT39_5_T]"
## [29] "GT40 Week 5 Treated [GT40_5_T]"   "GT41 Week 5 Treated [GT41_5_T]"
## [31] "GT39 Week 5 Control [GT39_5_C]"   "GT40 Week 5 Control [GT40_5_C]"
## [33] "GT41 Week 5 Control [GT41_5_C]"   "GT34 Week 6 Treated [GT34_6_T]"
## [35] "GT35 Week 6 Treated [GT35_6_T]"   "GT37 Week 6 Treated [GT37_6_T]"
## [37] "GT34 Week 6 Control [GT34_6_C]"   "GT35 Week 6 Control [GT35_6_C]"
## [39] "GT37 Week 6 Control [GT37_6_C]"   "GT27 Week 8 Treated [GT27_8_T]"
## [41] "GT28 Week 8 Treated [GT28_8_T]"   "GT29 Week 8 Treated [GT29_8_T]"
## [43] "GT27 Week 8 Control [GT27_8_C]"   "GT28 Week 8 Control [GT28_8_C]"
## [45] "GT29 Week 8 Control [GT29_8_C]"
```

Capturing count data file column names to match the metadata against sample names and treatment levels to be created in a dataframe below

```
count_cols <- names(GSE124161_readcount)#get the column names from the read count data
count_cols
```

```
##  [1] "GT01_D0"  "GT09_D0"  "GT19_D0"  "GT57_1_T" "GT58_1_T" "GT59_1_T"
##  [7] "GT57_1_C" "GT58_1_C" "GT59_1_C" "GT51_2_T" "GT52_2_T" "GT53_2_T"
## [13] "GT51_2_C" "GT52_2_C" "GT53_2_C" "GT45_3_T" "GT46_3_T" "GT47_3_T"
## [19] "GT45_3_C" "GT46_3_C" "GT47_3_C" "GT17_4_T" "GT18_4_T" "GT19_4_T"
## [25] "GT17_4_C" "GT18_4_C" "GT19_4_C" "GT39_5_T" "GT40_5_T" "GT41_5_T"
## [31] "GT39_5_C" "GT40_5_C" "GT41_5_C" "GT34_6_T" "GT35_6_T" "GT37_6_T"
## [37] "GT34_6_C" "GT35_6_C" "GT37_6_C" "GT27_8_T" "GT28_8_T" "GT29_8_T"
## [43] "GT27_8_C" "GT28_8_C" "GT29_8_C"
```

Constructing a data frame for later use with sample metadata

```
#this is matching the original count column names to the phenotype names I ordered in prior R code
pheno_df<-cbind(count_pheno,count_cols)
pheno_df<-as.data.frame(pheno_df)
```

Continue to build the metadata dataframe object

```
#I may need a factor separating the count data by week, so I am creating the information in an ordered fashion, to integrate into the data frame as a column.
day0 <-rep("D0", each = 3)
wknames <- c("W1", "W2", "W3", "W4", "W5", "W6", "W8")
weeks1_8 <-rep(wknames, each = 6)


all_weeks <-c(day0,weeks1_8)

pheno_df$weeks <-all_weeks
```

Keep building the dataframe, adding columns of metadata as necessary.

```
#adding another column to designate the treatment and control groups associated with the colum ns in the count file
group0 <-c("Day0","Day0","Day0")
expnames <- c("EGCG", "Placebo")
Group1_8 <-rep(expnames, each =3, times =7)

groups<-append(group0, Group1_8)

pheno_df$groups <-groups
```

Keep building the dataframe, adding columns of metadata as necessary.

```
#adding uninjured-injured to separate out the groups, in case we want to use this comparison.
uninjured <-c("uninjured","uninjured", "uninjured")
injured <-rep("injured", each =1, times =42)

treatment <-append(uninjured, injured)
pheno_df$treatment <-treatment #adding column "status" to pheno_df
```

Keep building the dataframe, adding columns of metadata as

necessary.

```
#adding "none", "zonal" and "direct" to pheno table to designate application type
none<- rep("none", each = 1, times =3)
zonal <-rep("zonal", each =1, times = 12)
direct <-rep("direct", each = 1, times = 30)

appl <- append(none, zonal)
application <-append(appl,direct)


pheno_df$application <-application #adding column "application" to pheno_df
```

Keep building the dataframe, adding columns of metadata as necessary

```
#adding "D0", "Zw1", "Zw2" and "Dw1", "Dw2", "Dw3", "Dw4", "Dw6" to pheno table to designate a
pplication type by week
non<- rep("W0", each = 1, times =3)
zw1 <-rep("Zw1", each =1, times = 6)
zw2 <-rep("Zw2", each =1, times = 6)
dw1 <-rep("Dw1", each = 1, times = 6)
dw2 <-rep("Dw2", each = 1, times = 6)
dw3 <-rep("Dw3", each = 1, times = 6)
dw4 <-rep("Dw4", each = 1, times = 6)
dw6 <-rep("Dw6", each = 1, times = 6)

zon1 <- append(non, zw1)
zon2 <- append(zon1, zw2)
dir1 <- append(zon2, dw1)
dir2 <- append(dir1, dw2)
dir3 <- append(dir2, dw3)
dir4 <- append(dir3, dw4)
dir6 <- append(dir4, dw6)

pheno_df$appl_by_wk <-dir6 #adding column "application" to pheno_df
```

Rename columns in pheno_df for clarity

```
#renaming column names in data frame for clarity
colnames(pheno_df) <-c("samples", "count_colnames", "week", "treatments", "status", "applicati
on", "appl_by_wk")
pheno_df
```

samples <chr>	count_colnames <chr>	w…	treatments <chr>	status <chr>	application <chr>	app <ch
GT01 D0	GT01_D0	D0	Day0	uninjured	none	W0
GT09 D0	GT09_D0	D0	Day0	uninjured	none	W0
GT19 D0	GT19_D0	D0	Day0	uninjured	none	W0

samples <chr>	count_colnames <chr>	w…	treatments <chr>	status <chr>	application <chr>	app <ch
GT57 Week 1 Treated [GT57_1_T]	GT57_1_T	W1	EGCG	injured	zonal	Zw1
GT58 Week 1 Treated [GT58_1_T]	GT58_1_T	W1	EGCG	injured	zonal	Zw1
GT59 Week 1 Treated [GT59_1_T]	GT59_1_T	W1	EGCG	injured	zonal	Zw1
GT57 Week 1 Control [GT57_1_C]	GT57_1_C	W1	Placebo	injured	zonal	Zw1
GT58 Week 1 Control [GT58_1_C]	GT58_1_C	W1	Placebo	injured	zonal	Zw1
GT59 Week 1 Control [GT59_1_C]	GT59_1_C	W1	Placebo	injured	zonal	Zw1
GT51 Week 2 Treated [GT51_2_T]	GT51_2_T	W2	EGCG	injured	zonal	Zw2

1-10 of 45 rows Previous **1** 2 3 4 5 Next

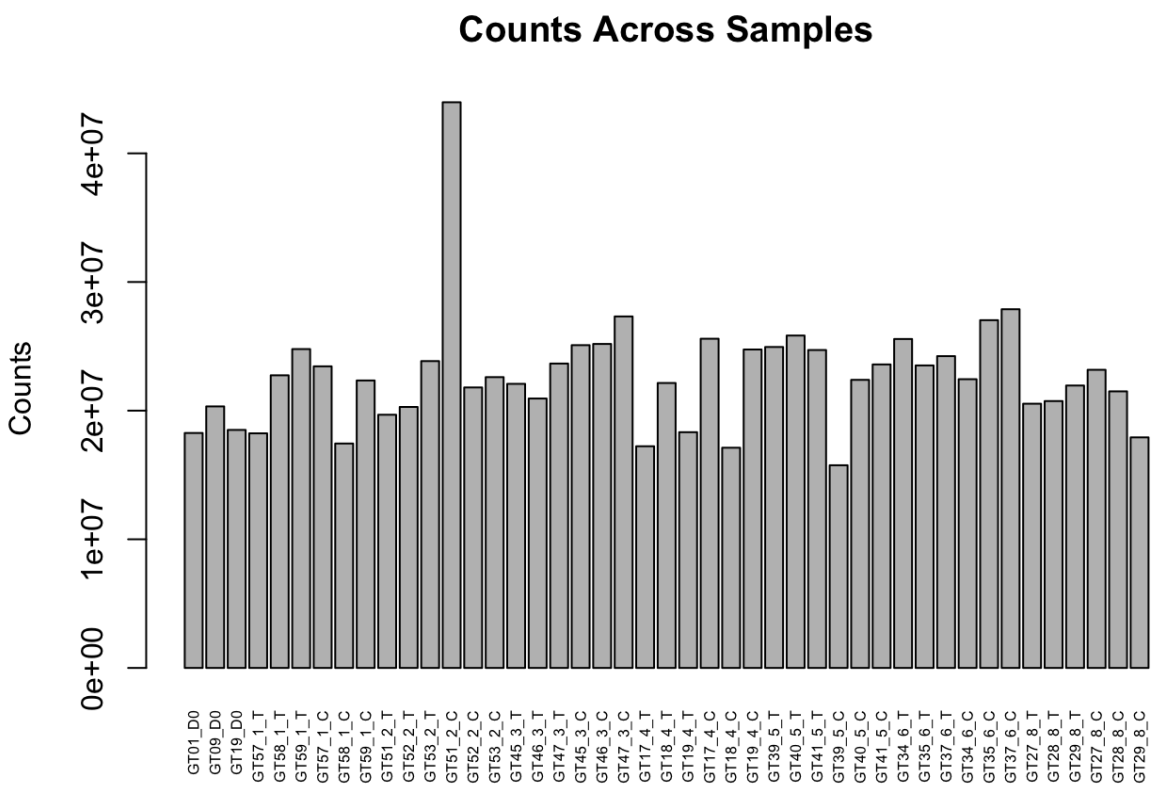
Quality Control

```
sums <- colSums(GSE124161_readcount)

#GT51_2_C, which is one of the week 2 control samples, is double the depth of the entire exper
iment 43,971,422 (divided in half it's = 21,985,711), so yes it is double. I wonder if they fo
und this, and removed it. One way is to check the week 2 heat maps to see if this sample has a
n extreme up regulation. It's there, so they did not exclude it, but nothing shows as extremel
y up-regulated???? What is going on here??
sums
```

```
##  GT01_D0  GT09_D0  GT19_D0 GT57_1_T GT58_1_T GT59_1_T GT57_1_C GT58_1_C
## 18261863 20328995 18501060 18235034 22746777 24785167 23443164 17444604
## GT59_1_C GT51_2_T GT52_2_T GT53_2_T GT51_2_C GT52_2_C GT53_2_C GT45_3_T
## 22345084 19680608 20286078 23851421 43971422 21802025 22604825 22081693
## GT46_3_T GT47_3_T GT45_3_C GT46_3_C GT47_3_C GT17_4_T GT18_4_T GT19_4_T
## 20941586 23660164 25090699 25184243 27321326 17230960 22148122 18323913
## GT17_4_C GT18_4_C GT19_4_C GT39_5_T GT40_5_T GT41_5_T GT39_5_C GT40_5_C
## 25588820 17116435 24752227 24947885 25837140 24712048 15753422 22393291
## GT41_5_C GT34_6_T GT35_6_T GT37_6_T GT34_6_C GT35_6_C GT37_6_C GT27_8_T
## 23591332 25568002 23513634 24235461 22444354 27031627 27882501 20537596
## GT28_8_T GT29_8_T GT27_8_C GT28_8_C GT29_8_C
## 20743466 21951239 23173809 21498675 17921576
```

```
barplot(sums,
        main = "Counts Across Samples",
        ylab = "Counts",
        cex.names = 0.5,
        las = 3)
```



Some Quick Analysis:

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::combine()    masks Biobase::combine(), BiocGenerics::combine()
## ✖ dplyr::filter()     masks stats::filter()
## ✖ dplyr::lag()        masks stats::lag()
## ✖ ggplot2::Position() masks BiocGenerics::Position(), base::Position()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to bec
ome errors
```

```
library(reshape2)
```

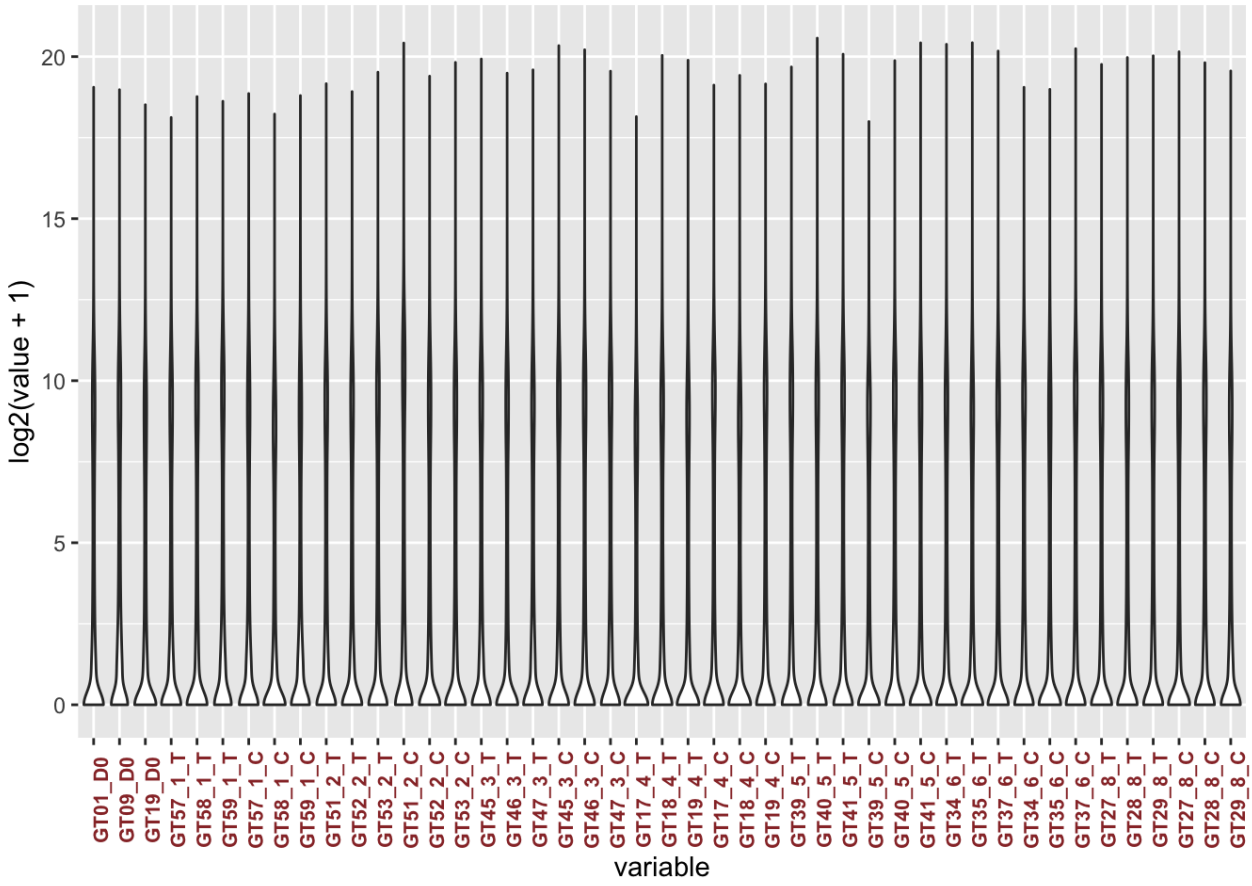
```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
datamat = apply(GSE124161_readcount, 2, as.integer)
data= as.data.frame(datamat)
rownames(data) = rownames(GSE124161_readcount)

data_wnames = data #data with gene names
data_wnames$gene = rownames(data) #creating a column $gene in the data stored in variable data
_wnames, using the rownames from the data variable
data_melt = melt(data_wnames) #melting the data into long form (see in environment) This is hu
ge, for every gene in the dataset 48,162 X 45 samples = 2,167,290 entries
```

```
## Using gene as id variables
```

```
ggplot(data_melt) +
  geom_violin(mapping = aes(x=variable, y=log2(value +1))) + theme(axis.text.x = element_text
(face="bold", color="#993333",
                           size=8, angle=90))
```



```
# this violin plot gives you an idea on how the data is distributed. You would expected all of them to look the same and the samples should not vastly deviate from each other over the entire data set. So we are looking to see that all the samples and replicates do not have big differences

# in this example we want the variable as the x axis (the variable is the names in the melted graph, that will group/condense according to the name, and will be the samples in the graph, we should have 45 samples plotted)

# the (value+1) is added because if you have a value of 0, and you take the log of 0, you have a problem - it's undefined, so if you add 1 to all of the values, then a log of 1 will = 0 and everything will be scaled identically with 1 extra count added across the board, so we can get the log(1) = 0

# and we can see this in the violin plot as the thick base in the violin plot is the genes that actually have 0 counts.
```

Trim the dataset from low count genes

Removing low expressed genes. When you do the differential expression you do not want to run statistics on the background noise.Look at the data in the following histogram to determine how many genes are low expressed. 6,000 of the genes (look at the histogram below) are not going to give you results, they are too low, so you should trim the dataset.

```
genesum = rowSums(data) # we are asking R to calculate the sums of the counts in each row(gene) across all 45 samples, and we will graph the result as a histogram below.
hist(log2(genesum+1), ylim = c(0,10000), xlim = c(0,30), breaks = 25) #this generates the histogram below, and from looking at the histogram, we can see that 6K genes have a near-zero expression level, really low values, across the entire dataset.
```


```
sum(genesum ==0) #here we are asking exactly how many genes are equal to 0, which we get 4467 out of 48,162 total genes, this is not bad, it is expected, not all genes are going to be expressed in an RNAseq experiment.
```

```
## [1] 4467
```

```
sum(genesum <2) # and 6518 genes have a count across the 45 samples when totaled up are less than 2,
```

```
## [1] 6518
```

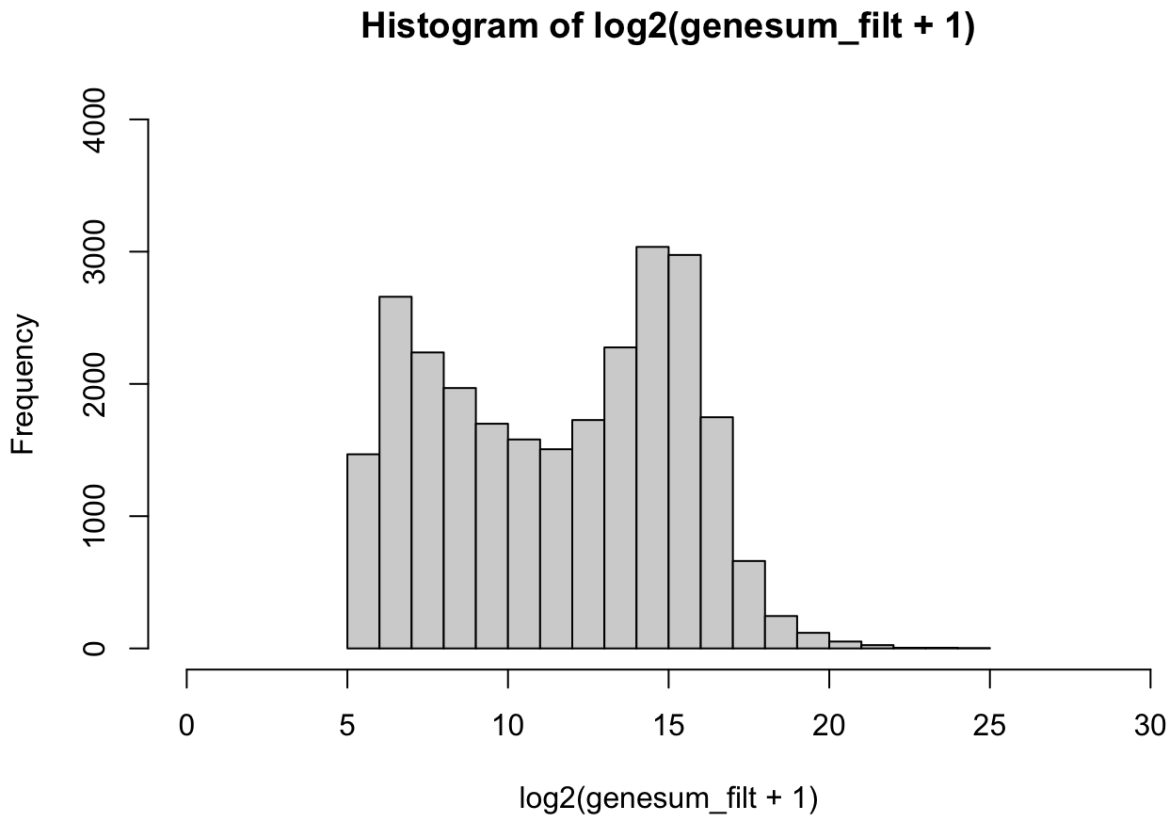
```
#For this filtering we wanted to keep the subset of data where the genesum was over 30
#genesum = 45 + 1 = 46 log2(46) = 5.52 = ~5.5, everything below 5 on the above graph, so we want to keep everything 45 and above

data_filt = subset(data, genesum >45)
genesum = rowSums(data)
```

Take a look at the histogram again.

Is this the presence of 2 means forming in the data?

```
genesum_filt = rowSums(data_filt)
hist(log2(genesum_filt+1), ylim = c(0,4000), xlim = c(0,30), breaks = 25)
```




```
dim(data_filt)
```

```
## [1] 25995    45
```

Load limma and edgeR

```
library(limma)
```

```
##
## Attaching package: 'limma'
```

```
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
library(edgeR)
```

Create a design matrix for lm.

First we create the levels that we are potentially interested in. Some of these we may not use, but we have them in case we want to make a different comparison.

```
week_mm <- factor(pheno_df$week, levels = c("D0", "W1", "W2", "W3", "W4", "W5", "W6", "W8"))
treatments_mm <- factor(pheno_df$treatments, levels = c("Day0", "Placebo", "EGCG"))
status_mm <-factor(pheno_df$status, levels = c("uninjured", "injured"))
application_mm <-factor(pheno_df$application, levels = c("none", "zonal", "direct"))
appl_by_wk_mm <-factor(pheno_df$appl_by_wk, levels = c("W0", "Zw1", "Zw2", "Dw1", "Dw2", "Dw
3", "Dw4", "Dw6"))
```

This model matrix defines all of the things that you are interested in comparing It creates a matrix that defines the various experimental categories of the samples in your experiment that you want to compare. We will use the week_mm and treatments_mm as we are interested in comparing treatments across the weeks.

```
weektreat = factor(paste(week_mm,treatments_mm, sep=""))
weektreat
```

```
##  [1] D0Day0    D0Day0    D0Day0    W1EGCG    W1EGCG    W1EGCG    W1Placebo
##  [8] W1Placebo W1Placebo W2EGCG    W2EGCG    W2EGCG    W2Placebo W2Placebo
## [15] W2Placebo W3EGCG    W3EGCG    W3EGCG    W3Placebo W3Placebo W3Placebo
## [22] W4EGCG    W4EGCG    W4EGCG    W4Placebo W4Placebo W4Placebo W5EGCG
## [29] W5EGCG    W5EGCG    W5Placebo W5Placebo W5Placebo W6EGCG    W6EGCG
## [36] W6EGCG    W6Placebo W6Placebo W6Placebo W8EGCG    W8EGCG    W8EGCG
## [43] W8Placebo W8Placebo W8Placebo
## 15 Levels: D0Day0 W1EGCG W1Placebo W2EGCG W2Placebo W3EGCG W3Placebo ... W8Placebo
```

```
design = model.matrix(~0+weektreat)
design
```

```
##    weektreatD0Day0 weektreatW1EGCG weektreatW1Placebo weektreatW2EGCG
## 1                1                0                   0                0
## 2                1                0                   0                0
## 3                1                0                   0                0
## 4                0                1                   0                0
## 5                0                1                   0                0
## 6                0                1                   0                0
## 7                0                0                   1                0
## 8                0                0                   1                0
## 9                0                0                   1                0
## 10               0                0                   0                1
## 11               0                0                   0                1
## 12               0                0                   0                1
## 13               0                0                   0                0
## 14               0                0                   0                0
## 15               0                0                   0                0
## 16               0                0                   0                0
## 17               0                0                   0                0
## 18               0                0                   0                0
## 19               0                0                   0                0
## 20               0                0                   0                0
## 21               0                0                   0                0
## 22               0                0                   0                0
## 23               0                0                   0                0
## 24               0                0                   0                0
## 25               0                0                   0                0
## 26               0                0                   0                0
## 27               0                0                   0                0
## 28               0                0                   0                0
## 29               0                0                   0                0
## 30               0                0                   0                0
## 31               0                0                   0                0
## 32               0                0                   0                0
## 33               0                0                   0                0
## 34               0                0                   0                0
## 35               0                0                   0                0
## 36               0                0                   0                0
## 37               0                0                   0                0
## 38               0                0                   0                0
## 39               0                0                   0                0
## 40               0                0                   0                0
## 41               0                0                   0                0
## 42               0                0                   0                0
## 43               0                0                   0                0
## 44               0                0                   0                0
## 45               0                0                   0                0
##    weektreatW2Placebo weektreatW3EGCG weektreatW3Placebo weektreatW4EGCG
## 1                   0                0                   0                0
## 2                   0                0                   0                0
## 3                   0                0                   0                0
## 4                   0                0                   0                0
## 5                   0                0                   0                0
## 6                   0                0                   0                0
## 7                   0                0                   0                0
## 8                   0                0                   0                0
```

```
## 9                    0                0                    0                0
## 10                   0                0                    0                0
## 11                   0                0                    0                0
## 12                   0                0                    0                0
## 13                   1                0                    0                0
## 14                   1                0                    0                0
## 15                   1                0                    0                0
## 16                   0                1                    0                0
## 17                   0                1                    0                0
## 18                   0                1                    0                0
## 19                   0                0                    1                0
## 20                   0                0                    1                0
## 21                   0                0                    1                0
## 22                   0                0                    0                1
## 23                   0                0                    0                1
## 24                   0                0                    0                1
## 25                   0                0                    0                0
## 26                   0                0                    0                0
## 27                   0                0                    0                0
## 28                   0                0                    0                0
## 29                   0                0                    0                0
## 30                   0                0                    0                0
## 31                   0                0                    0                0
## 32                   0                0                    0                0
## 33                   0                0                    0                0
## 34                   0                0                    0                0
## 35                   0                0                    0                0
## 36                   0                0                    0                0
## 37                   0                0                    0                0
## 38                   0                0                    0                0
## 39                   0                0                    0                0
## 40                   0                0                    0                0
## 41                   0                0                    0                0
## 42                   0                0                    0                0
## 43                   0                0                    0                0
## 44                   0                0                    0                0
## 45                   0                0                    0                0
##    weektreatW4Placebo weektreatW5EGCG weektreatW5Placebo weektreatW6EGCG
## 1                    0                0                    0                0
## 2                    0                0                    0                0
## 3                    0                0                    0                0
## 4                    0                0                    0                0
## 5                    0                0                    0                0
## 6                    0                0                    0                0
## 7                    0                0                    0                0
## 8                    0                0                    0                0
## 9                    0                0                    0                0
## 10                   0                0                    0                0
## 11                   0                0                    0                0
## 12                   0                0                    0                0
## 13                   0                0                    0                0
## 14                   0                0                    0                0
## 15                   0                0                    0                0
## 16                   0                0                    0                0
## 17                   0                0                    0                0
## 18                   0                0                    0                0
```

```
## 19                  0                0                  0                0
## 20                  0                0                  0                0
## 21                  0                0                  0                0
## 22                  0                0                  0                0
## 23                  0                0                  0                0
## 24                  0                0                  0                0
## 25                  1                0                  0                0
## 26                  1                0                  0                0
## 27                  1                0                  0                0
## 28                  0                1                  0                0
## 29                  0                1                  0                0
## 30                  0                1                  0                0
## 31                  0                0                  1                0
## 32                  0                0                  1                0
## 33                  0                0                  1                0
## 34                  0                0                  0                1
## 35                  0                0                  0                1
## 36                  0                0                  0                1
## 37                  0                0                  0                0
## 38                  0                0                  0                0
## 39                  0                0                  0                0
## 40                  0                0                  0                0
## 41                  0                0                  0                0
## 42                  0                0                  0                0
## 43                  0                0                  0                0
## 44                  0                0                  0                0
## 45                  0                0                  0                0
##    weektreatW6Placebo weektreatW8EGCG weektreatW8Placebo
## 1                   0                0                  0
## 2                   0                0                  0
## 3                   0                0                  0
## 4                   0                0                  0
## 5                   0                0                  0
## 6                   0                0                  0
## 7                   0                0                  0
## 8                   0                0                  0
## 9                   0                0                  0
## 10                  0                0                  0
## 11                  0                0                  0
## 12                  0                0                  0
## 13                  0                0                  0
## 14                  0                0                  0
## 15                  0                0                  0
## 16                  0                0                  0
## 17                  0                0                  0
## 18                  0                0                  0
## 19                  0                0                  0
## 20                  0                0                  0
## 21                  0                0                  0
## 22                  0                0                  0
## 23                  0                0                  0
## 24                  0                0                  0
## 25                  0                0                  0
## 26                  0                0                  0
## 27                  0                0                  0
## 28                  0                0                  0
```

```
## 29                  0                 0                 0
## 30                  0                 0                 0
## 31                  0                 0                 0
## 32                  0                 0                 0
## 33                  0                 0                 0
## 34                  0                 0                 0
## 35                  0                 0                 0
## 36                  0                 0                 0
## 37                  1                 0                 0
## 38                  1                 0                 0
## 39                  1                 0                 0
## 40                  0                 1                 0
## 41                  0                 1                 0
## 42                  0                 1                 0
## 43                  0                 0                 1
## 44                  0                 0                 1
## 45                  0                 0                 1
## attr(,"assign")
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$weektreat
## [1] "contr.treatment"
```

```
colnames(design) = levels(weektreat)
design
```

```
##    D0Day0 W1EGCG W1Placebo W2EGCG W2Placebo W3EGCG W3Placebo W4EGCG W4Placebo
## 1       1       0          0       0          0       0          0       0          0
## 2       1       0          0       0          0       0          0       0          0
## 3       1       0          0       0          0       0          0       0          0
## 4       0       1          0       0          0       0          0       0          0
## 5       0       1          0       0          0       0          0       0          0
## 6       0       1          0       0          0       0          0       0          0
## 7       0       0          1       0          0       0          0       0          0
## 8       0       0          1       0          0       0          0       0          0
## 9       0       0          1       0          0       0          0       0          0
## 10      0       0          0       1          0       0          0       0          0
## 11      0       0          0       1          0       0          0       0          0
## 12      0       0          0       1          0       0          0       0          0
## 13      0       0          0       0          1       0          0       0          0
## 14      0       0          0       0          1       0          0       0          0
## 15      0       0          0       0          1       0          0       0          0
## 16      0       0          0       0          0       1          0       0          0
## 17      0       0          0       0          0       1          0       0          0
## 18      0       0          0       0          0       1          0       0          0
## 19      0       0          0       0          0       0          1       0          0
## 20      0       0          0       0          0       0          1       0          0
## 21      0       0          0       0          0       0          1       0          0
## 22      0       0          0       0          0       0          0       1          0
## 23      0       0          0       0          0       0          0       1          0
## 24      0       0          0       0          0       0          0       1          0
## 25      0       0          0       0          0       0          0       0          1
## 26      0       0          0       0          0       0          0       0          1
## 27      0       0          0       0          0       0          0       0          1
## 28      0       0          0       0          0       0          0       0          0
## 29      0       0          0       0          0       0          0       0          0
## 30      0       0          0       0          0       0          0       0          0
## 31      0       0          0       0          0       0          0       0          0
## 32      0       0          0       0          0       0          0       0          0
## 33      0       0          0       0          0       0          0       0          0
## 34      0       0          0       0          0       0          0       0          0
## 35      0       0          0       0          0       0          0       0          0
## 36      0       0          0       0          0       0          0       0          0
## 37      0       0          0       0          0       0          0       0          0
## 38      0       0          0       0          0       0          0       0          0
## 39      0       0          0       0          0       0          0       0          0
## 40      0       0          0       0          0       0          0       0          0
## 41      0       0          0       0          0       0          0       0          0
## 42      0       0          0       0          0       0          0       0          0
## 43      0       0          0       0          0       0          0       0          0
## 44      0       0          0       0          0       0          0       0          0
## 45      0       0          0       0          0       0          0       0          0
##    W5EGCG W5Placebo W6EGCG W6Placebo W8EGCG W8Placebo
## 1       0          0       0          0       0          0
## 2       0          0       0          0       0          0
## 3       0          0       0          0       0          0
## 4       0          0       0          0       0          0
## 5       0          0       0          0       0          0
## 6       0          0       0          0       0          0
## 7       0          0       0          0       0          0
## 8       0          0       0          0       0          0
```

```
## 9       0          0       0          0       0          0
## 10      0          0       0          0       0          0
## 11      0          0       0          0       0          0
## 12      0          0       0          0       0          0
## 13      0          0       0          0       0          0
## 14      0          0       0          0       0          0
## 15      0          0       0          0       0          0
## 16      0          0       0          0       0          0
## 17      0          0       0          0       0          0
## 18      0          0       0          0       0          0
## 19      0          0       0          0       0          0
## 20      0          0       0          0       0          0
## 21      0          0       0          0       0          0
## 22      0          0       0          0       0          0
## 23      0          0       0          0       0          0
## 24      0          0       0          0       0          0
## 25      0          0       0          0       0          0
## 26      0          0       0          0       0          0
## 27      0          0       0          0       0          0
## 28      1          0       0          0       0          0
## 29      1          0       0          0       0          0
## 30      1          0       0          0       0          0
## 31      0          1       0          0       0          0
## 32      0          1       0          0       0          0
## 33      0          1       0          0       0          0
## 34      0          0       1          0       0          0
## 35      0          0       1          0       0          0
## 36      0          0       1          0       0          0
## 37      0          0       0          1       0          0
## 38      0          0       0          1       0          0
## 39      0          0       0          1       0          0
## 40      0          0       0          0       1          0
## 41      0          0       0          0       1          0
## 42      0          0       0          0       1          0
## 43      0          0       0          0       0          1
## 44      0          0       0          0       0          1
## 45      0          0       0          0       0          1
## attr(,"assign")
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$weektreat
## [1] "contr.treatment"
```

What if we let LIMMA and edgeR select our low expressed genes for us?

how would that be different than the genesum cutoff we chose of 45?

```
dge = DGEList(counts = GSE124161_readcount)
dim(dge$counts) #before filtering
```

```
## [1] 48162    45
```

```
keep = filterByExpr(dge, design)
dge = dge[keep,,keep.lib.sizes=FALSE]

dim(dge$counts) #after filtering there are 20,935 genes, this is a lot stricter than the 25,99
5 genes we kept by filtering using genesum which was arbitrary and selected by judgement.
```

```
## [1] 20935    45
```

Create a PCA plot, after low expressed genes are filtered out

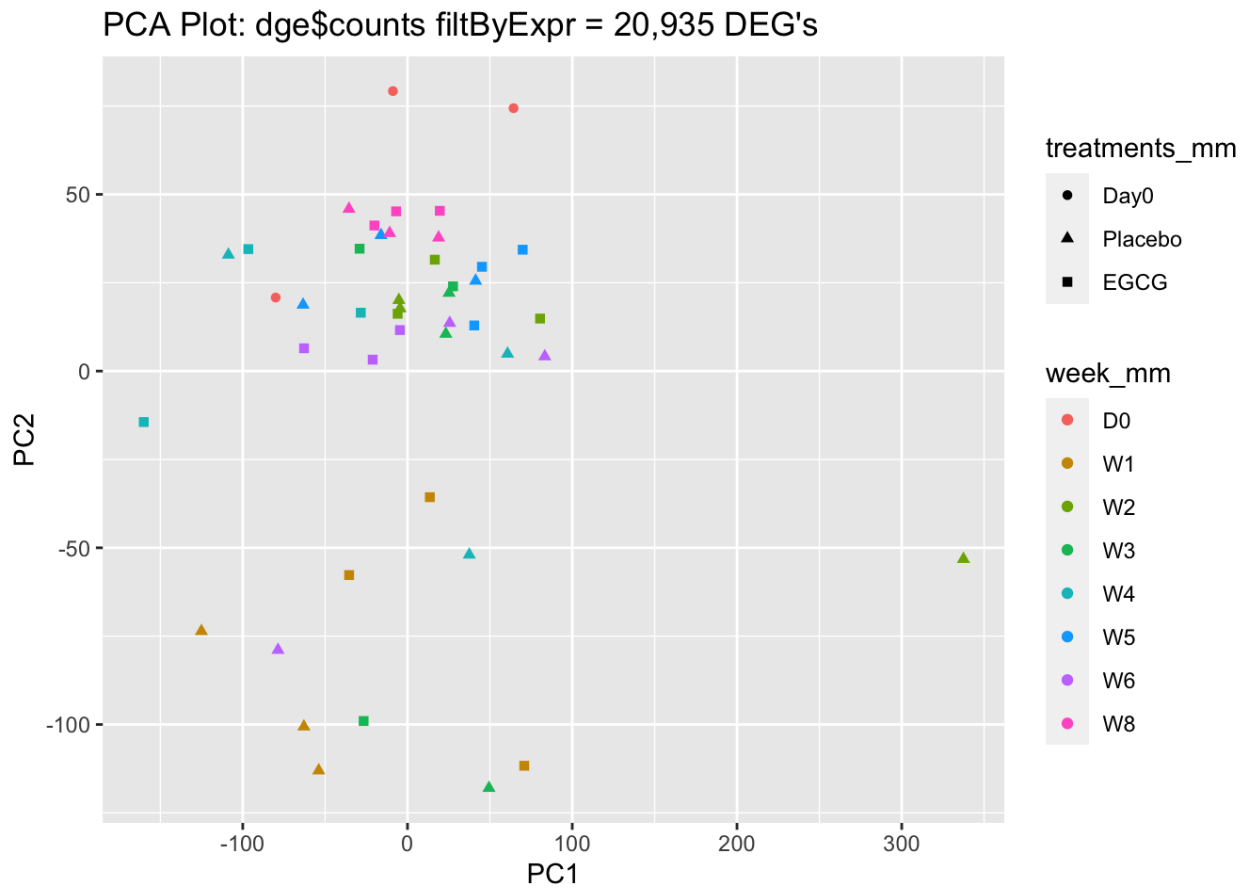
```
library(dplyr)

data_prcomp = prcomp(t(dge$counts), scale=TRUE, center=TRUE)

library(ggplot2)

coords2draw = as.data.frame(data_prcomp$x)

ggplot(coords2draw) +
  geom_point(mapping=aes(x = PC1, y= PC2,
                         col = week_mm, shape = treatments_mm)) +
  labs(title = "PCA Plot: dge$counts filtByExpr = 20,935 DEG's")
```



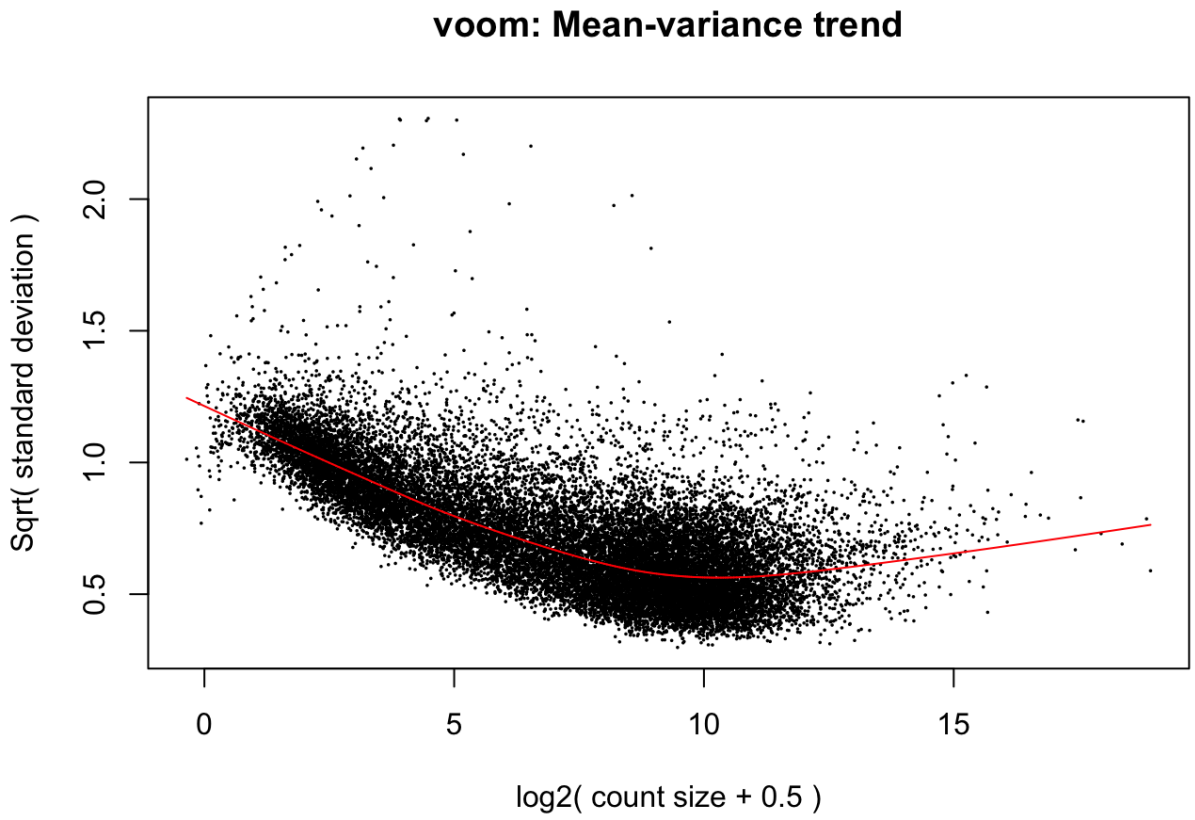
Now that we created a matrix that defines the various experimental categories of the samples, now we want to normalize the data. We need to normalize the data first before making the comparisons, as the normalized data is needed to proceed in next steps.

Voom normalization

Voom provides data in a format that can be used for standard limma methods. In the Limma manual, another normalization process is called “eset”, which is normalization through the AFFY package, however we are normalizing this data with “voom”, so “v” is the object we are storing the normalized data in.

Voom is the normalization method that allows us to use the data in downstream analyses. For this expression, voom is acting on the 20,935 genes captured in “dge”, using the design outlined

```
v = voom(dge, design, plot=TRUE, normalize="quantile")
```



Then create the lmfit (this calculates the “within” variance). This fits a linear model to the data.

```
nfit = lmFit(v,design)
```

Now specifically compare the different coefficients for the comparison

This gives us a lot more control to make the specific comparisons we want. This gives us a lot of control over a complex data set, one with a lot of levels, time-series data.

```
newcontrasts = makeContrasts(Zw1EGCG_vs_Zw1_placebo = W1EGCG - W1Placebo,#these are comparing the Treatment to the control at a single time point
                             Zw2EGCG_vs_Zw2_placebo = W2EGCG - W2Placebo,
                             Dw1_EGCG_vs_Dw1_placebo = W3EGCG - W3Placebo,
                             Dw2_EGCG_vs_Dw2_placebo  = W4EGCG - W4Placebo,
                             Dw3_EGCG_vs_Dw3_placebo  = W5EGCG - W5Placebo,
                             Dw4_EGCG_vs_Dw4_placebo  = W6EGCG - W6Placebo,
                             Dw6_EGCG_vs_Dw6_placebo  = W8EGCG - W8Placebo,
                             interact = (W1EGCG - W1Placebo) - (W2EGCG - W2Placebo), #Change in expression levels from Zonal week1 to wk2 differs between the ECGC-treated group and the placebo-treated group. If statistically significant, it would suggest that the change in expression levels over time (from week1 --> week 2) differs between the EGCG-treated group and the placebo-treated group.
                             interact2 = (W1EGCG - W1Placebo) - (W3EGCG - W3Placebo), #Change in expression levels for Zonal wk 1 to Direct wk 1. If statistically significant, it would suggest that the change in expression levels over time (from week1 --> week 2) differs between the EGCG-treated group and the placebo-treated group.
                             interact3 = (W2EGCG - W2Placebo) - (W4EGCG - W4Placebo), #Change in Expression levels for Zonal wk 2 and Direct wk 2
                             interact4 = W2EGCG - W1EGCG - W2Placebo + W1Placebo, # EGCG vs Placebo, Significant means that EGCG is having a statistically differential response between the two time points W1-W2
                             interact5 = W1EGCG + W1Placebo - W2EGCG - W2Placebo, # Significant means that EGCG is having a statistically differential response between the two time points, relative to the control

                             levels = weektreat)
```

```
newcontrasts
```

```
##            Contrasts
## Levels      Zw1EGCG_vs_Zw1_placebo Zw2EGCG_vs_Zw2_placebo
##   D0Day0                        0                      0
##   W1EGCG                        1                      0
##   W1Placebo                    -1                      0
##   W2EGCG                        0                      1
##   W2Placebo                     0                     -1
##   W3EGCG                        0                      0
##   W3Placebo                     0                      0
##   W4EGCG                        0                      0
##   W4Placebo                     0                      0
##   W5EGCG                        0                      0
##   W5Placebo                     0                      0
##   W6EGCG                        0                      0
##   W6Placebo                     0                      0
##   W8EGCG                        0                      0
##   W8Placebo                     0                      0
##            Contrasts
## Levels      Dw1_EGCG_vs_Dw1_placebo Dw2_EGCG_vs_Dw2_placebo
##   D0Day0                         0                       0
##   W1EGCG                         0                       0
##   W1Placebo                      0                       0
##   W2EGCG                         0                       0
##   W2Placebo                      0                       0
##   W3EGCG                         1                       0
##   W3Placebo                     -1                       0
##   W4EGCG                         0                       1
##   W4Placebo                      0                      -1
##   W5EGCG                         0                       0
##   W5Placebo                      0                       0
##   W6EGCG                         0                       0
##   W6Placebo                      0                       0
##   W8EGCG                         0                       0
##   W8Placebo                      0                       0
##            Contrasts
## Levels      Dw3_EGCG_vs_Dw3_placebo Dw4_EGCG_vs_Dw4_placebo
##   D0Day0                         0                       0
##   W1EGCG                         0                       0
##   W1Placebo                      0                       0
##   W2EGCG                         0                       0
##   W2Placebo                      0                       0
##   W3EGCG                         0                       0
##   W3Placebo                      0                       0
##   W4EGCG                         0                       0
##   W4Placebo                      0                       0
##   W5EGCG                         1                       0
##   W5Placebo                     -1                       0
##   W6EGCG                         0                       1
##   W6Placebo                      0                      -1
##   W8EGCG                         0                       0
##   W8Placebo                      0                       0
##            Contrasts
## Levels      Dw6_EGCG_vs_Dw6_placebo interact interact2 interact3 interact4
##   D0Day0                         0        0         0         0         0
##   W1EGCG                         0        1         1         0        -1
```

```
##   W1Placebo                      0        -1         -1          0          1
##   W2EGCG                         0        -1          0          1          1
##   W2Placebo                      0         1          0         -1         -1
##   W3EGCG                         0         0         -1          0          0
##   W3Placebo                      0         0          1          0          0
##   W4EGCG                         0         0          0         -1          0
##   W4Placebo                      0         0          0          1          0
##   W5EGCG                         0         0          0          0          0
##   W5Placebo                      0         0          0          0          0
##   W6EGCG                         0         0          0          0          0
##   W6Placebo                      0         0          0          0          0
##   W8EGCG                         1         0          0          0          0
##   W8Placebo                     -1         0          0          0          0
##            Contrasts
## Levels      interact5
##   D0Day0             0
##   W1EGCG             1
##   W1Placebo          1
##   W2EGCG            -1
##   W2Placebo         -1
##   W3EGCG             0
##   W3Placebo          0
##   W4EGCG             0
##   W4Placebo          0
##   W5EGCG             0
##   W5Placebo          0
##   W6EGCG             0
##   W6Placebo          0
##   W8EGCG             0
##   W8Placebo          0
```

Fit the data to new contrasts and then calculate the p-value for each gene.

```
nfit2= contrasts.fit(nfit, newcontrasts)
nfit2 = eBayes(nfit2)
topTable(nfit2,  adjust="BH") #BH = one of the multiple hypothesis testing methods we talked a
bout the FDR correction.
```

	Zw1EGCG_vs_Zw1_placebo <dbl>	Zw2EGCG_vs_Zw2_placebo <dbl>	Dw1_EGCG_vs_Dw1_placebo <dbl>
ENSG00000140519	-1.6248092	0.51406004	1.0167165352
ENSG00000021355	-0.2395233	-0.01146786	-0.0999755754
ENSG00000171848	0.4620292	0.08929963	0.0575356387
ENSG00000183696	-0.5946972	0.94157598	-0.0009506325
ENSG00000163209	-0.9096414	0.75378615	-0.9568299708
ENSG00000189410	-0.2724934	0.93794466	0.3123970609

	Zw1EGCG_vs_Zw1_placebo <dbl>	Zw2EGCG_vs_Zw2_placebo <dbl>	Dw1_EGCG_vs_Dw1_placebo <dbl>
ENSG00000128965	-1.8278921	0.72243711	-0.0759112571
ENSG00000115602	-0.6573179	0.18092313	-0.0095954582
ENSG00000074317	-0.9754783	1.15879278	0.2942885039
ENSG00000106819	0.3898805	-0.46008831	-0.0828503291

1-10 of 10 rows | 1-4 of 17 columns

Coeff = INTERACT

get details of specific coeff defined in the contrast. Selected contrast “interact”

topTable() is a function in limma which summarizes the results of the linear model, perform hypothesis tests, and adjust the p-values for multiple testing.Results include (log2) fold changes, standard errors, t-statistics and p-values. A number of summary statistics are presented by topTable() for the top genes and the selected contrast “interact”.

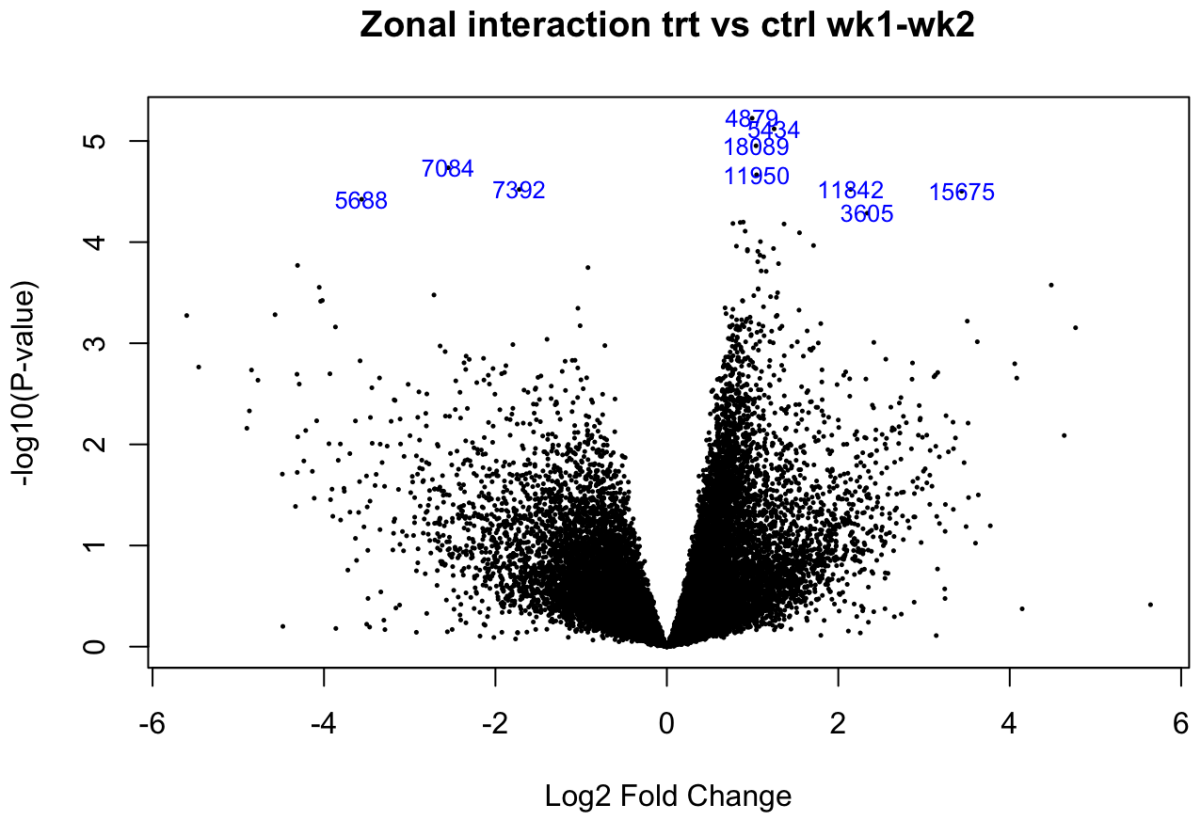
```
topTable(nfit2, coef = "interact",  adjust="BH") #we want to specify a specific coefficient, we can look at the Treatment effect EGCG "vs" the Placebo in the interaction of week 1 and 2 Zonal Treatment
```

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
ENSG00000085491	0.9948163	4.9843135	5.346265	5.956497e-06	0.07781009	3.884279
ENSG00000138071	1.2515705	8.4230263	5.265892	7.584321e-06	0.07781009	3.591048
ENSG00000137710	1.0401063	5.7848368	5.137563	1.115024e-05	0.07781009	3.316062
ENSG00000206560	1.0463782	5.4347981	4.917067	2.158403e-05	0.08221305	2.711787
ENSG00000128965	-2.5503292	0.3528284	-4.970273	1.840855e-05	0.08221305	2.506032
ENSG00000217801	-1.7218036	1.6768825	-4.805082	3.015282e-05	0.08221305	2.221923
ENSG00000154874	3.4397189	0.3176969	4.791309	3.141650e-05	0.08221305	1.836474
ENSG00000008952	0.8913389	6.7788842	4.554955	6.338436e-05	0.09896020	1.729630
ENSG00000065150	0.8557484	6.6350538	4.552109	6.392000e-05	0.09896020	1.722080
ENSG00000167193	0.7705619	6.4417247	4.544263	6.542041e-05	0.09896020	1.701122

1-10 of 10 rows

Make a volcano plot of this data

```
volcanoplot(nfit2, "interact", highlight = 10, main="Zonal interaction trt vs ctrl wk1-wk2")#The higlight=10 highlights the top 10, but gives the rownames... Changing to Ensembl gene name truncates the name, so it is not useful. However, we can capture the row name and get the gene name, but why bother, as the names are already in the top10 gene list above from the toptable() function. So see above ^)
```



Where are the normalized values from the Zoom normalization for all of the comparisons made in the dataset. ### Normalized values are stored in v$E

```
normexpvalues = v$E
```

```
head(normexpvalues)
```

```
##                   GT01_D0    GT09_D0   GT19_D0  GT57_1_T  GT58_1_T  GT59_1_T
## ENSG00000123159  7.271149  7.4122624 8.0319992  6.987033  7.314820  7.098878
## ENSG00000233005 -1.235546 -0.5587764 0.4922798 -2.100866 -2.745624 -2.038246
## ENSG00000131242  5.666664  5.4804867 5.3694893  3.815836  4.336988  3.616872
## ENSG00000139168  5.134622  5.1683002 5.9914752  4.952089  5.118815  4.801069
## ENSG00000115541  2.893089  2.6645463 3.4987717  3.383017  3.165651  3.065886
## ENSG00000105486  4.717573  5.0625303 4.3603528  4.447360  4.814297  4.264415
##                  GT57_1_C  GT58_1_C  GT59_1_C   GT51_2_T  GT52_2_T  GT53_2_T
## ENSG00000123159  7.518985  7.684010  7.597574  6.9607315  7.261658  7.128385
## ENSG00000233005 -1.485629 -1.488546 -2.827593 -0.8579596 -1.345410 -1.195722
## ENSG00000131242  4.264950  2.849192  4.520628  4.5330654  4.342466  4.672421
## ENSG00000139168  4.827302  6.253074  4.654369  4.9533066  5.269447  4.875076
## ENSG00000115541  4.253053  5.025696  3.750854  2.3880567  3.052152  2.288165
## ENSG00000105486  4.273606  3.904416  4.233130  4.3465493  4.168258  4.423441
##                  GT51_2_C   GT52_2_C  GT53_2_C  GT45_3_T  GT46_3_T  GT47_3_T
## ENSG00000123159 7.3061832  7.3284764  6.961145  6.967745  7.288389  6.772299
## ENSG00000233005 0.2076713 -0.2889084 -2.032891 -1.941631 -1.101122 -1.509705
## ENSG00000131242 4.7960280  4.8491830  4.562683  3.794021  4.511481  3.548694
## ENSG00000139168 5.1067209  5.2165167  5.089568  4.833474  4.960957  5.778976
## ENSG00000115541 2.4973033  2.5490538  3.021289  2.534274  2.621711  4.571800
## ENSG00000105486 4.0003567  4.1640820  3.937719  3.937066  3.949419  4.026190
##                  GT45_3_C  GT46_3_C  GT47_3_C  GT17_4_T  GT18_4_T  GT19_4_T
## ENSG00000123159  7.098878  7.014791  6.528873 7.7103974  7.340138  7.341442
## ENSG00000233005 -1.779650 -1.905098 -2.262058 0.6500853 -1.509705 -1.862430
## ENSG00000131242  4.375879  4.206759  3.789523 3.9281970  4.786679  4.504358
## ENSG00000139168  5.038846  4.960489  5.702488 6.5497827  5.216517  4.944889
## ENSG00000115541  2.524548  2.909242  4.306894 4.1797316  2.504025  2.556141
## ENSG00000105486  3.984542  4.044198  4.230057 2.7321736  4.247714  4.249908
##                   GT17_4_C   GT18_4_C  GT19_4_C  GT39_5_T  GT40_5_T  GT41_5_T
## ENSG00000123159  7.6458680  7.5539732  7.442093  6.964603  7.354397  7.247190
## ENSG00000233005 -0.5900602 -0.7800433 -0.853282 -2.838380 -3.840517 -1.674588
## ENSG00000131242  3.9592267  4.7818214  5.126004  4.459279  4.398804  4.546997
## ENSG00000139168  5.7875683  5.4415423  5.021878  4.945479  4.684047  4.905898
## ENSG00000115541  3.3237744  2.4725512  2.682647  2.477482  3.143244  2.368212
## ENSG00000105486  3.3436448  4.3932545  4.612627  4.277105  4.225955  4.354620
##                  GT39_5_C  GT40_5_C   GT41_5_C  GT34_6_T  GT35_6_T  GT37_6_T
## ENSG00000123159  7.005782  7.101623  7.1745103  7.381414  7.048813  7.162279
## ENSG00000233005 -0.853282 -2.341753 -0.6415115 -2.077288 -1.146907 -1.320596
## ENSG00000131242  3.348706  4.118488  4.5110221  4.214178  4.067442  4.869286
## ENSG00000139168  5.569744  4.852133  4.9444328  5.032897  5.375121  5.465900
## ENSG00000115541  3.212673  2.332835  2.3253250  2.640732  2.825147  2.737686
## ENSG00000105486  4.223734  4.181322  4.1387828  4.193323  4.328932  4.112569
##                  GT34_6_C GT35_6_C  GT37_6_C  GT27_8_T  GT28_8_T  GT29_8_T
## ENSG00000123159  7.375818 7.469538  6.956143  6.991880  6.861242  6.612324
## ENSG00000233005 -2.182059 1.275623 -2.032891 -5.481118 -5.481118 -4.304903
## ENSG00000131242  5.073132 4.490853  4.758655  4.710662  4.344061  4.303919
## ENSG00000139168  4.980795 6.445494  5.427946  4.946310  5.131683  5.223881
## ENSG00000115541  2.706947 4.458751  2.419008  2.608850  2.498930  2.192418
## ENSG00000105486  4.786679 2.885908  4.092467  4.161710  4.073175  4.160029
##                  GT27_8_C  GT28_8_C  GT29_8_C
## ENSG00000123159  6.854044  6.726125  6.568899
## ENSG00000233005 -1.228285 -2.077288 -1.281308
## ENSG00000131242  4.690596  4.141276  4.222744
## ENSG00000139168  5.104816  5.132168  5.217227
```

```
## ENSG00000115541  2.354893  2.398867  2.812218
## ENSG00000105486  4.086042  3.890042  4.019271
```

Get the genes that have adjpvalue < 0.2 and absolute log2fc > 1.5

Because the 0.05 p-value produced **no** gene candidates for downstream processing, the p-value was selected for 0.2, providing enough of a gene set for downstream cluster analysis and heat map generation.

This uses the coefficient “interact” of the following interaction of week 1 vs week 2, which is Zonal treatment: interact = (W1EGCG - W1Placebo) - (W2EGCG - W2Placebo)

```
interact_sig = topTable(nfit2,
                coef = "interact",
                adjust="BH", #method used to adjust the p-values for multiple testing. Option
s, in increasing conservatism, include                                    "none", "BH", "BY", "ho
lm"
                p.value=0.2, #cutoff value for adjusted p-values. Only genes with lower p-valu
es are listed
                number=10000, #max number of genes to list
                sort.by = "P", #sort by p-value
                lfc=log2(1.5)) #log fold change cutoff, the minimum absolute log2-fold-change
required
```

Get the voom values for these genes.

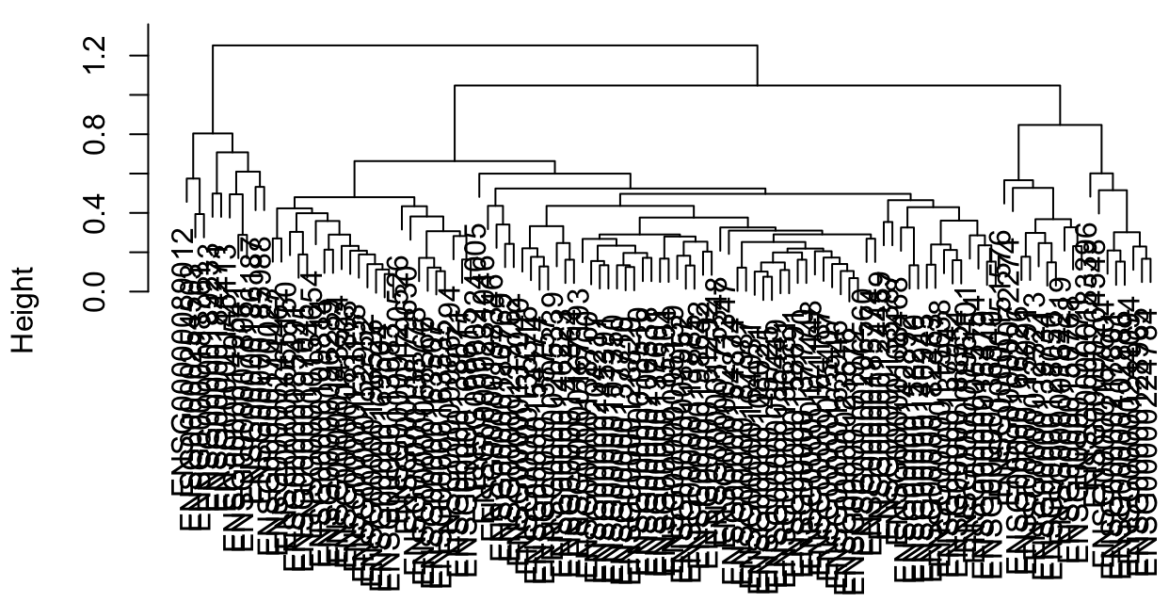
```
interact_sig_normvalues = normexpvalues[rownames(interact_sig),]
```

Calculate the distance using pairwise correlation of genes.

Use hclust to perform the clustering.

This is the interaction of (W1EGCG - W1Placebo) - (W2EGCG - W2Placebo)

```
interact_sig_dist = as.dist(1 - cor(t(interact_sig_normvalues))) #this is correlation, not euc
lidean ()
interact_sig_hclust = hclust(interact_sig_dist,
                             method="average")
plot(interact_sig_hclust, main = "Interaction Zonal (W1EGCG - W1Placebo) - (W2EGCG - W2Placeb
o)")
```



Let's

determing the ideal number of cluster by calculating the avg silhouette width at each cutting.

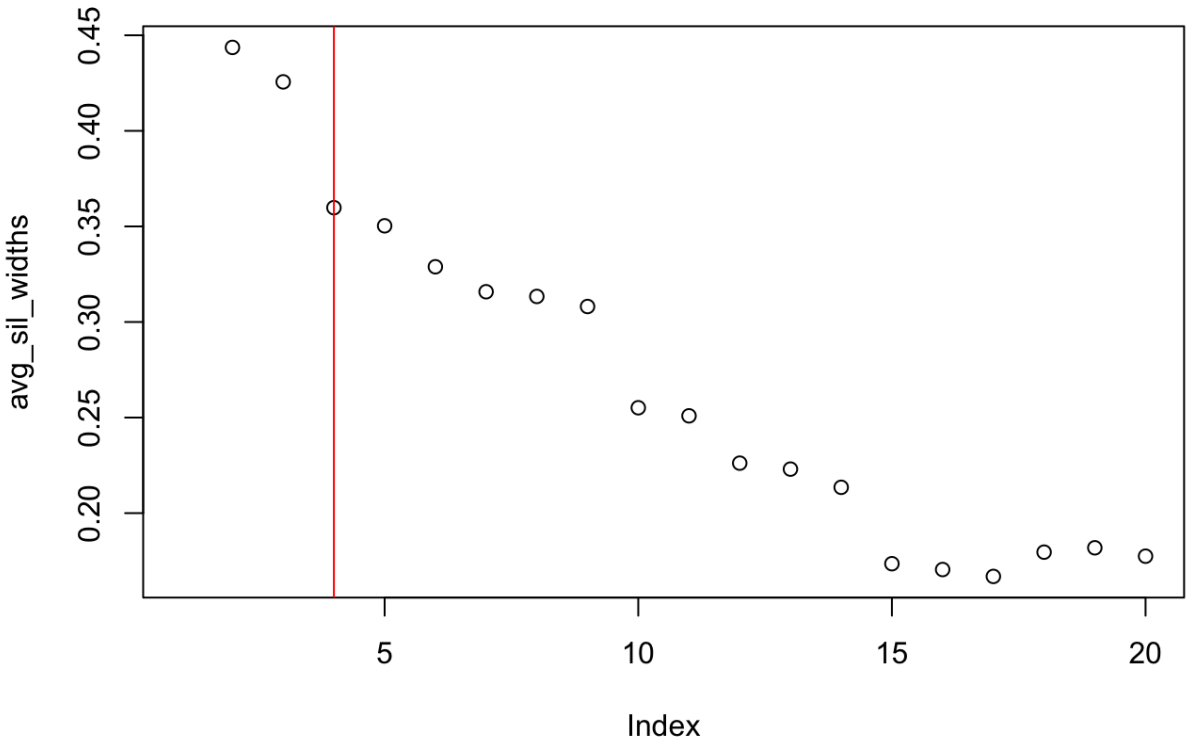
```
library(cluster)

avg_sil_widths = numeric()
for ( i in 2:20) {
  tempclust = cutree(interact_sig_hclust, k = i)
  avg_sil_widths[i] = mean(silhouette(tempclust, interact_sig_dist)[,"sil_width"])

}
```

4 looks promising. Let's go with 4 for now.

```
plot(avg_sil_widths)
abline(v=4, col="red")
```



Create

the groups. Notice the result is actually a vector of number and the gene names are the labels.

```
interact_sig_hclust_4 = cutree(interact_sig_hclust, k=4)
head(interact_sig_hclust_4)
```

```
## ENSG00000085491 ENSG00000138071 ENSG00000137710 ENSG00000128965 ENSG00000206560
##               1               1               1               2               1
## ENSG00000185774
##               1
```

To get the gene names that are in the different groups, use the which command to find out which genes are in the different groups, but then use the names function to get the actual names.

```
interact_sig_hclust_g1= normexpvalues[names(which(interact_sig_hclust_4==1)),]
interact_sig_hclust_g2= normexpvalues[names(which(interact_sig_hclust_4==2)),]
interact_sig_hclust_g3= normexpvalues[names(which(interact_sig_hclust_4==3)),]
interact_sig_hclust_g4= normexpvalues[names(which(interact_sig_hclust_4==4)),]
```

Create heatmap of each cluster group

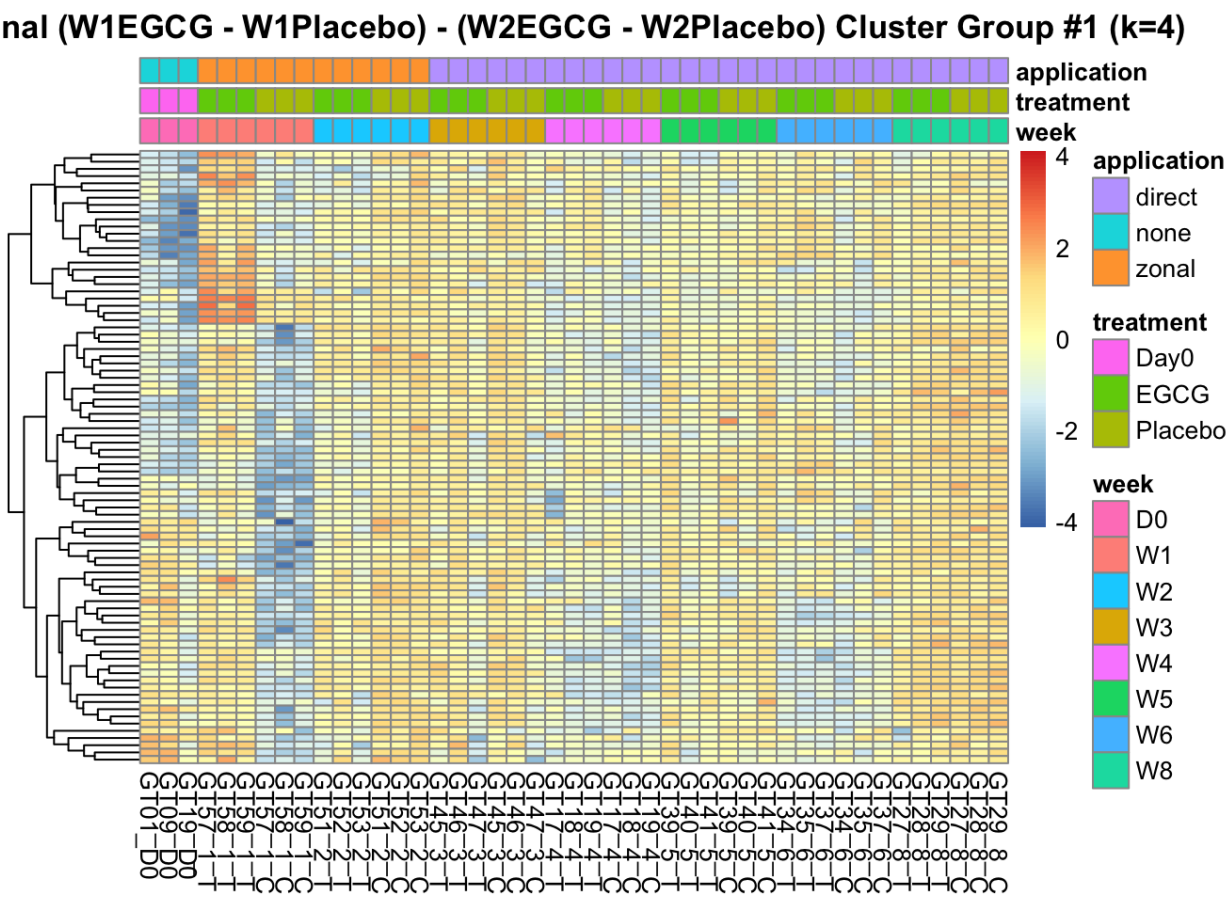
Cluster#1

Use “pheatmap” to draw cluster. “annot_col” defines how to create the legend. “scale” allows us to see the pattern for each gene. To make it easier to compare the different groups, I asked the columns not to be clustered “cluster_cols = F” , and to not show the gene names “show_rownames = F”.

```
library(pheatmap)

annotation <-as.data.frame(cbind(pheno_df$week, pheno_df$treatments, pheno_df$application))
colnames(annotation) <- c('week','treatment', 'application')
rownames(annotation) <-pheno_df$count_colnames

pheatmap(interact_sig_hclust_g1,annotation_col = annotation, scale="row", cluster_cols = F, sh
ow_rownames = F, main = "Zonal (W1EGCG - W1Placebo) - (W2EGCG - W2Placebo) Cluster Group #1 (k
=4)" )
```



Perform Go-Term Enrichment analysis

```
# Load the proper packages
library(GOstats)
```

```
## Loading required package: Category
```

```
## Loading required package: stats4
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: IRanges
```

```
## Loading required package: S4Vectors
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:lubridate':
##
##     second, second<-
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, rename
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:lubridate':
##
##     %within%
```

```
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
```

```
## The following object is masked from 'package:purrr':
##
##     reduce
```

```
##
## Attaching package: 'AnnotationDbi'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:S4Vectors':
##
##     expand
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loading required package: graph
```

```
##
## Attaching package: 'graph'
```

```
## The following object is masked from 'package:stringr':
##
##     boundary
```

```
##
```

```
##
## Attaching package: 'GOstats'
```

```
## The following object is masked from 'package:AnnotationDbi':
##
##     makeGOGraph
```

```
library(GO.db)
library(Category)
library(org.Hs.eg.db)
```

```
##
```

Go-Term Enrichment Part 1

Create HyperGoparpam

Converting the Ensemble to Entrez was achieved with this code: https://www.biostars.org/p/441386/ (https://www.biostars.org/p/441386/)

```
library("AnnotationDbi")

#adding ENTREZ ID's to global gene data file
GSE124161_readcount$entrez = mapIds(org.Hs.eg.db,
                    keys=rownames(GSE124161_readcount), #Column containing Ensembl gene ids
                    column="ENTREZID",
                    keytype="ENSEMBL",
                    multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
#Wrangling the ensemble gene ID's to Entrez in the interact_sig_hclust_g1
diffexpgenes_names_df <-rownames(as.data.frame(interact_sig_hclust_g1))
diffexpgenes_names_df <-as.data.frame(diffexpgenes_names_df)

diffexpgenes_names_df$entrez = mapIds(org.Hs.eg.db,
                    keys= diffexpgenes_names_df$diffexpgenes_names_df, #Column containing Ense
mbl gene ids
                    column="ENTREZID",
                    keytype="ENSEMBL",
                    multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
diffexpgenes_names <-diffexpgenes_names_df$entrez
readcount_names <-GSE124161_readcount$entrez

#Utilized following resource for below code format https://bioconductor.org/packages/release/b
ioc/vignettes/GOstats/inst/doc/GOstatsHyperG.pdf

params <- new("GOHyperGParams",
              geneIds = diffexpgenes_names, #don't use quotes here, it will not work, you will
get an error message. This is the variable name where you stored your differentially expressed
gene names
              universeGeneIds = readcount_names, #don't use quotes here, it will not work, you
will get an error message. This is the variable name where you stored all of the gene names fr
om the whole unfiltered data set. Its the whole list of the "universe" of gene IDs for your ar
ray or reference genome.
              annotation = "org.Hs.eg",
              ontology = "BP",
              pvalueCutoff=0.01, #don't use quotes here, it will not work, you will get an err
or message
              testDirection = "over")
```

```
## Warning in makeValidParams(.Object): removing duplicate IDs in geneIds
```

```
## Warning in makeValidParams(.Object): removing duplicate IDs in universeGeneIds
```

```
hypGO <- hyperGTest(params)
hypGO
```

```
## Gene to GO BP  test for over-representation
## 1989 GO BP ids tested (64 have p < 0.01)
## Selected gene set size: 78
##     Gene universe size: 17259
##     Annotation package: org.Hs.eg
```

The summary function returns a data.frame summarizing the result.

By default, only the results for terms with a p-value less than the cutoff specified in the parameter instance will be shown.You can also set a minimum number of genes for each term using the “categorySize” argument. I chose a grouping of 10.

```
sumGo <- summary(hypGO, categorySize =10)
sumGo
```

GOBPID <chr>	Pvalue <dbl>	OddsRatio <dbl>	ExpCount <dbl>	Count <int>	Size <int>
GO:0051641	0.0002856908	2.424293	14.69702764	28	3252
GO:0008104	0.0007084221	2.447917	10.83747610	22	2398
GO:0070727	0.0007499909	2.436011	10.88266991	22	2408
GO:0031293	0.0010801006	50.210526	0.04971319	2	11
GO:0051649	0.0016436536	2.396818	9.28280897	19	2054
GO:0045184	0.0022066584	2.495853	7.34851382	16	1626
GO:0045144	0.0023223136	32.268797	0.07231010	2	16
GO:0051177	0.0023223136	32.268797	0.07231010	2	16
GO:0007135	0.0026242633	30.115789	0.07682948	2	17
GO:0061983	0.0026242633	30.115789	0.07682948	2	17

1-10 of 35 rows | 1-6 of 7 columns Previous **1** 2 3 4 Next

```
GoPlot <- data.frame(sumGo$GOBPID,sumGo$Pvalue,sumGo$Term)
colnames(GoPlot) <-c("GO_ID_BP", "P-value", "Term")
GoPlot
```

GO_ID_BP <chr>	P-value <dbl>
GO:0051641	0.0002856908
GO:0008104	0.0007084221

GO_ID_BP <chr>	P-value <dbl>
GO:0070727	0.0007499909
GO:0031293	0.0010801006
GO:0051649	0.0016436536
GO:0045184	0.0022066584
GO:0045144	0.0023223136
GO:0051177	0.0023223136
GO:0007135	0.0026242633
GO:0061983	0.0026242633

1-10 of 35 rows | 1-2 of 3 columns Previous **1** 2 3 4 Next

KEGG ENRIGHMENT Part1

```
#install Libraries needed for KEGG Enrifhment Analysis
library(clusterProfiler)
```

```
## clusterProfiler v4.6.2  For help: https://yulab-smu.top/biomedical-knowledge-mining-book/
##
## If you use clusterProfiler in published research, please cite:
## T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo,
and G Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The In
novation. 2021, 2(3):100141
```

```
##
## Attaching package: 'clusterProfiler'
```

```
## The following object is masked from 'package:AnnotationDbi':
##
##     select
```

```
## The following object is masked from 'package:IRanges':
##
##     slice
```

```
## The following object is masked from 'package:S4Vectors':
##
##     rename
```

```
## The following object is masked from 'package:purrr':
##
##     simplify
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(pathview)
```

```
## ##############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
## formally cite the original Pathview paper (not just mention it) in publications
## or products. For details, do citation("pathview") within R.
##
## The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## ##############################################################################
```

```
library(gage)
library(gageData)

#Now perform KEGG ENRICHMENT

keggEnrich <- enrichKEGG(
  diffexpgenes_names_df$entrez,
  organism = "hsa",
  keyType = "kegg",
  pvalueCutoff = 0.05, #adjust this if you are not seeing any results
  pAdjustMethod = "BH",
  )
```

```
## Reading KEGG annotation online: "https://rest.kegg.jp/link/hsa/pathway"...
```

```
## Reading KEGG annotation online: "https://rest.kegg.jp/list/pathway/hsa"...
```

```
#Show results from enrichKEGG
head(keggEnrich)
```

0 rows

```
keggEnrich
```

```
## #
## # over-representation test
## #
## #...@organism     hsa
## #...@ontology     KEGG
## #...@keytype      kegg
## #...@gene     chr [1:83] "29957" "10097" "5962" "23243" "80333" "284047" "4685" "7095" ...
## #...pvalues adjusted by 'BH' with cutoff <0.05
## #...0 enriched terms found
## #...Citation
##  T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo,
and G Yu.
##  clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
##  The Innovation. 2021, 2(3):100141
```

```
#Generate a graph for the two KEGG results
#Edit the pathway id to that which is appropriate based on the ID column from the enrichKEGG o
utput

#These will generate images that will be saved to the working directory or the downloads folde
r
#Repeat for however many results you get from keggEnrich

pv.out_htmp1a <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa00565", s
pecies = "hsa")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa00565.pathview.png
```

```
#Repeat for the second result
pv.out_htmp1b <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa00480", s
pecies = "hsa")
```

```
## Warning: None of the genes or compounds mapped to the pathway!
## Argument gene.idtype or cpd.idtype may be wrong.
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa00480.pathview.png
```

```
#Also show the genes involved in the pathway
#These correspond to the elements included in the image of the KEGG pathway generated earlier
pv.out_htmp1a$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
60	8540	AGPS	8540	gene	396	168	46	17	1
62	8611	PLPP1		gene	613	383	46	17	*NA*
65	10390	CEPT1		gene	180	453	46	17	*NA*
66	10390	CEPT1		gene	613	453	46	17	*NA*
67	10390	CEPT1		gene	314	453	46	17	*NA*
69	5048	PAFAH1B1		gene	159	521	46	17	*NA*
70	5319	PLA2G1B		gene	271	558	46	17	*NA*
71	54947	LPCAT2		gene	223	576	46	17	*NA*
72	54947	LPCAT2		gene	203	541	46	17	*NA*
73	5319	PLA2G1B		gene	541	575	46	17	*NA*

1-10 of 20 rows | 1-10 of 11 columns Previous **1** 2 Next

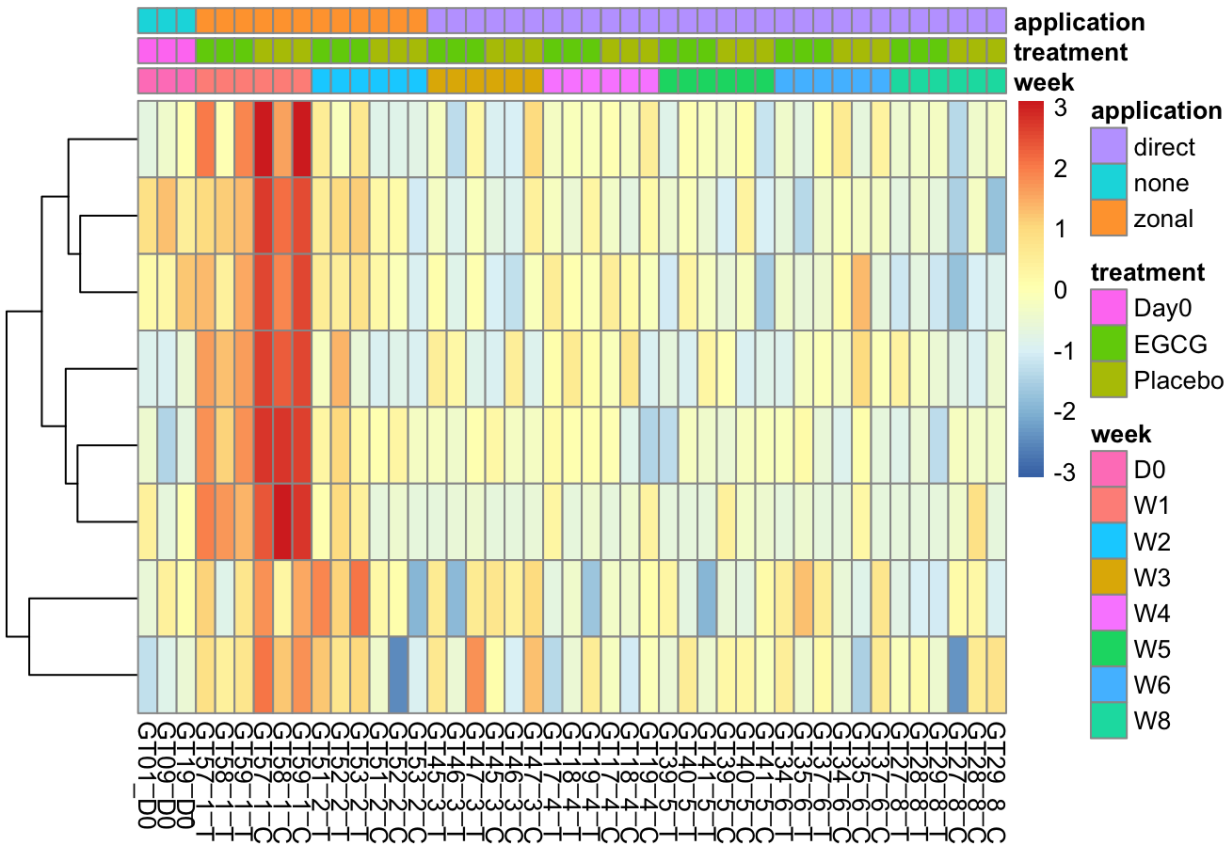
```
pv.out_htmp1b$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
37	290	ANPEP		gene	592	515	46	17	*NA*
38	2678	GGT1		gene	489	526	46	17	*NA*
39	2876	GPX1		gene	194	567	46	17	*NA*
46	2879	GPX4		gene	217	546	46	17	*NA*
47	51060	TXNDC12		gene	169	546	46	17	*NA*
49	2539	G6PD		gene	217	511	46	17	*NA*
51	3417	IDH1		gene	169	490	46	17	*NA*
52	2937	GSS		gene	458	457	46	17	*NA*
53	2729	GCLC		gene	542	275	46	17	*NA*
54	2678	GGT1		gene	332	430	46	17	*NA*

1-10 of 29 rows | 1-10 of 11 columns Previous **1** 2 3 Next

Cluster#2

```
pheatmap(interact_sig_hclust_g2,annotation_col = annotation, scale="row", cluster_cols = F, sh
ow_rownames = F, main = "Zonal (W1EGCG - W1Placebo) - (W2EGCG - W2Placebo) Cluster Group #2 (k
=4)" )
```



Go-Term Enrichment Part 2

Create HyperGoparpam

Converting the Ensemble to Entrez was achived with this code: https://www.biostars.org/p/441386/ (https://www.biostars.org/p/441386/)

```
library("AnnotationDbi")

#adding ENTREZ ID's to global gene data file
GSE124161_readcount$entrez = mapIds(org.Hs.eg.db,
                     keys=rownames(GSE124161_readcount), #Column containing Ensembl gene ids
                     column="ENTREZID",
                     keytype="ENSEMBL",
                     multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
#Wrangling the ensemble gene ID's to Entrez in the interact_sig_hclust_g2
diffexpgenes_names_df <-rownames(as.data.frame(interact_sig_hclust_g2))
diffexpgenes_names_df <-as.data.frame(diffexpgenes_names_df)

diffexpgenes_names_df$entrez = mapIds(org.Hs.eg.db,
                     keys= diffexpgenes_names_df$diffexpgenes_names_df, #Column containing Ense
mbl gene ids
                     column="ENTREZID",
                     keytype="ENSEMBL",
                     multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
diffexpgenes_names <-diffexpgenes_names_df$entrez
readcount_names <-GSE124161_readcount$entrez

#Utilized following resource for below code format https://bioconductor.org/packages/release/b
ioc/vignettes/GOstats/inst/doc/GOstatsHyperG.pdf

params <- new("GOHyperGParams",
              geneIds = diffexpgenes_names, #don't use quotes here, it will not work, you will
get an error message. This is the variable name where you stored your differentially expressed
gene names
              universeGeneIds = readcount_names, #don't use quotes here, it will not work, you
will get an error message. This is the variable name where you stored all of the gene names fr
om the whole unfiltered data set. Its the whole list of the "universe" of gene IDs for your ar
ray or reference genome.
              annotation = "org.Hs.eg",
              ontology = "BP",
              pvalueCutoff=0.01, #don't use quotes here, it will not work, you will get an err
or message
              testDirection = "over")
```

```
## Warning in makeValidParams(.Object): removing duplicate IDs in universeGeneIds
```

```
hypGO <- hyperGTest(params)
hypGO
```

```
## Gene to GO BP  test for over-representation
## 234 GO BP ids tested (23 have p < 0.01)
## Selected gene set size: 4
##     Gene universe size: 17259
##     Annotation package: org.Hs.eg
```

```
sumGo <- summary(hypGO, categorySize =10)
sumGo
```

GOBPID <chr>	Pvalue <dbl>	OddsRatio <dbl>	ExpCount <dbl>	Count <int>	Size <int>
GO:0034638	0.003472219	410.5000	0.003476447	1	15
GO:2000318	0.003472219	410.5000	0.003476447	1	15
GO:0006750	0.003703378	383.1111	0.003708210	1	16
GO:0019184	0.004165576	338.0000	0.004171736	1	18
GO:2000319	0.004627613	302.3860	0.004635263	1	20
GO:0045624	0.005089490	273.5556	0.005098789	1	22
GO:0010955	0.006704791	205.0833	0.006721131	1	29
GO:0072539	0.006704791	205.0833	0.006721131	1	29
GO:1903318	0.006704791	205.0833	0.006721131	1	29
GO:0043171	0.006935388	198.0000	0.006952894	1	30

1-10 of 19 rows | 1-6 of 7 columns Previous **1** 2 Next

```
GoPlot <- data.frame(sumGo$GOBPID,sumGo$Pvalue,sumGo$Term)
colnames(GoPlot) <-c("GO_ID_BP", "P-value", "Term")
GoPlot
```

GO_ID_BP <chr>	P-value <dbl>	Term <chr>
GO:0034638	0.003472219	phosphatidylcholine catabolic process
GO:2000318	0.003472219	positive regulation of T-helper 17 type immune response
GO:0006750	0.003703378	glutathione biosynthetic process
GO:0019184	0.004165576	nonribosomal peptide biosynthetic process
GO:2000319	0.004627613	regulation of T-helper 17 cell differentiation
GO:0045624	0.005089490	positive regulation of T-helper cell differentiation
GO:0010955	0.006704791	negative regulation of protein processing
GO:0072539	0.006704791	T-helper 17 cell differentiation
GO:1903318	0.006704791	negative regulation of protein maturation
GO:0043171	0.006935388	peptide catabolic process

1-10 of 19 rows Previous **1** 2 Next

KEGG ENRIGHMENT Part2

```
#Now perform KEGG ENRICHMENT

keggEnrich <- enrichKEGG(
  diffexpgenes_names_df$entrez,
  organism = "hsa",
  keyType = "kegg",
  pvalueCutoff = 0.05, #adjust this if you are not seeing any results
  pAdjustMethod = "BH",
  )
```

```
#Show results from enrichKEGG
head(keggEnrich)
```

	ID	Description	GeneRatio	BgRa...	pvalue	p.adjust	qvalue	gen
	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<lgl>	<ch
hsa00565	hsa00565	Ether lipid metabolism	1/2	50/8390	0.01188414	0.01354225	*NA*	791
hsa00480	hsa00480	Glutathione metabolism	1/2	57/8390	0.01354225	0.01354225	*NA*	790

2 rows | 1-9 of 10 columns

```
keggEnrich
```

```
## #
## # over-representation test
## #
## #...@organism     hsa
## #...@ontology     KEGG
## #...@keytype      kegg
## #...@gene     chr [1:8] "79094" "100288175" "112488748" "440955" "6705" NA "79153" ...
## #...pvalues adjusted by 'BH' with cutoff <0.05
## #...2 enriched terms found
## 'data.frame':    2 obs. of  9 variables:
##  $ ID         : chr  "hsa00565" "hsa00480"
##  $ Description: chr  "Ether lipid metabolism" "Glutathione metabolism"
##  $ GeneRatio  : chr  "1/2" "1/2"
##  $ BgRatio    : chr  "50/8390" "57/8390"
##  $ pvalue     : num  0.0119 0.0135
##  $ p.adjust   : num  0.0135 0.0135
##  $ qvalue     : logi  NA NA
##  $ geneID     : chr  "79153" "79094"
##  $ Count      : int  1 1
## #...Citation
##  T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo,
and G Yu.
##  clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
##  The Innovation. 2021, 2(3):100141
```

```
#Generate a graph for the two hsa00565 and hsa00480 KEGG results
#Edit the pathway id to that which is appropriate based on the ID column from the enrichKEGG o
utput

#These will generate images that will be saved to the working directory or the downloads folde
r
#Repeat for however many results you get from keggEnrich

pv.out_htmp2a <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa00565", s
pecies = "hsa")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa00565.pathview.png
```

```
#Repeat for the second result
pv.out_htmp2b <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa00480", s
pecies = "hsa")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa00480.pathview.png
```

```
#Also show the genes involved in the pathway
#These correspond to the elements included in the image of the KEGG pathway generated earlier
pv.out_htmp2a$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
60	8540	AGPS		gene	396	168	46	17	*NA*
62	8611	PLPP1		gene	613	383	46	17	*NA*
65	10390	CEPT1		gene	180	453	46	17	*NA*
66	10390	CEPT1		gene	613	453	46	17	*NA*
67	10390	CEPT1		gene	314	453	46	17	*NA*
69	5048	PAFAH1B1		gene	159	521	46	17	*NA*
70	5319	PLA2G1B		gene	271	558	46	17	*NA*
71	54947	LPCAT2		gene	223	576	46	17	*NA*
72	54947	LPCAT2		gene	203	541	46	17	*NA*

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
73	5319	PLA2G1B		gene	541	575	46	17	*NA*

1-10 of 20 rows | 1-10 of 11 columns Previous **1** 2 Next

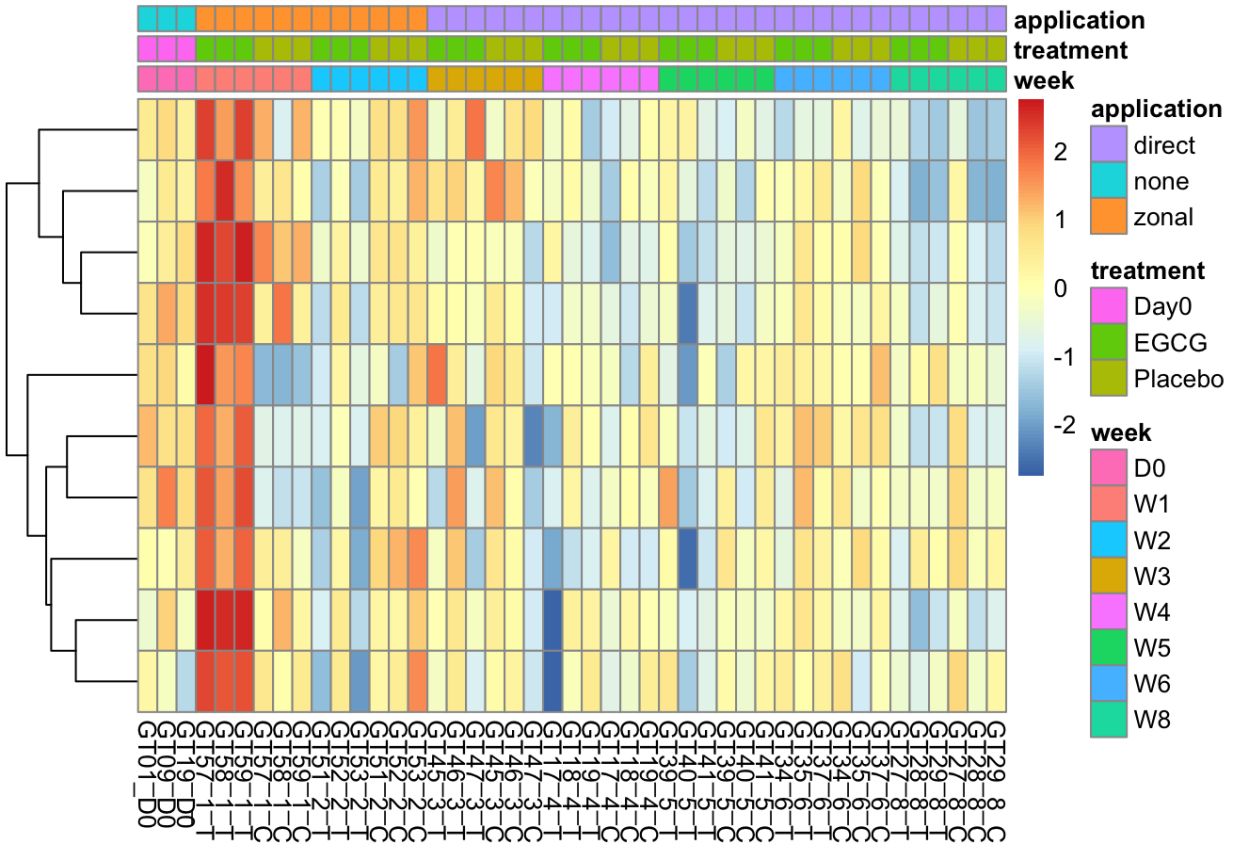
```
pv.out_htmp2b$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
37	290	ANPEP		gene	592	515	46	17	*NA*
38	2678	GGT1		gene	489	526	46	17	*NA*
39	2876	GPX1		gene	194	567	46	17	*NA*
46	2879	GPX4		gene	217	546	46	17	*NA*
47	51060	TXNDC12		gene	169	546	46	17	*NA*
49	2539	G6PD		gene	217	511	46	17	*NA*
51	3417	IDH1		gene	169	490	46	17	*NA*
52	2937	GSS		gene	458	457	46	17	*NA*
53	2729	GCLC		gene	542	275	46	17	*NA*
54	2678	GGT1		gene	332	430	46	17	*NA*

1-10 of 29 rows | 1-10 of 11 columns Previous **1** 2 3 Next

Cluster#3

```
pheatmap(interact_sig_hclust_g3,annotation_col = annotation, scale="row", cluster_cols = F, sh
ow_rownames = F, main = "Zonal (W1EGCG - W1Placebo) - (W2EGCG - W2Placebo) Cluster Group #3 (k
=4)" )
```



Go-Term

Enrichment Part 3

Create HyperGoparpam

Converting the Ensemble to Entrez was achived with this code: https://www.biostars.org/p/441386/ (https://www.biostars.org/p/441386/)

```
library("AnnotationDbi")

#adding ENTREZ ID's to global gene data file
GSE124161_readcount$entrez = mapIds(org.Hs.eg.db,
                     keys=rownames(GSE124161_readcount), #Column containing Ensembl gene ids
                     column="ENTREZID",
                     keytype="ENSEMBL",
                     multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
#Wrangling the ensemble gene ID's to Entrez in the interact_sig_hclust_g3
diffexpgenes_names_df <-rownames(as.data.frame(interact_sig_hclust_g3))
diffexpgenes_names_df <-as.data.frame(diffexpgenes_names_df)

diffexpgenes_names_df$entrez = mapIds(org.Hs.eg.db,
                     keys= diffexpgenes_names_df$diffexpgenes_names_df, #Column containing Ense
mbl gene ids
                     column="ENTREZID",
                     keytype="ENSEMBL",
                     multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
diffexpgenes_names <-diffexpgenes_names_df$entrez
readcount_names <-GSE124161_readcount$entrez

#Utilized following resource for below code format https://bioconductor.org/packages/release/b
ioc/vignettes/GOstats/inst/doc/GOstatsHyperG.pdf

params <- new("GOHyperGParams",
              geneIds = diffexpgenes_names, #don't use quotes here, it will not work, you will
get an error message. This is the variable name where you stored your differentially expressed
gene names
              universeGeneIds = readcount_names, #don't use quotes here, it will not work, you
will get an error message. This is the variable name where you stored all of the gene names fr
om the whole unfiltered data set. Its the whole list of the "universe" of gene IDs for your ar
ray or reference genome.
              annotation = "org.Hs.eg",
              ontology = "BP",
              pvalueCutoff=0.01, #don't use quotes here, it will not work, you will get an err
or message
              testDirection = "over")
```

```
## Warning in makeValidParams(.Object): removing duplicate IDs in universeGeneIds
```

```
hypGO <- hyperGTest(params)
hypGO
```

```
## Gene to GO BP  test for over-representation
## 615 GO BP ids tested (48 have p < 0.01)
## Selected gene set size: 10
##     Gene universe size: 17259
##     Annotation package: org.Hs.eg
```

```
sumGo <- summary(hypGO, categorySize =10)
sumGo
```

GOBPID <chr>	Pvalue <dbl>	OddsRatio <dbl>	ExpCount <dbl>	Count <int>	Size <int>
GO:1900182	0.001101061	50.482353	0.050408483	2	87
GO:1900180	0.002547072	32.667939	0.077061243	2	133
GO:0006006	0.004808767	23.443681	0.106611044	2	184
GO:0009113	0.005780498	212.839506	0.005794078	1	10
GO:0033572	0.005780498	212.839506	0.005794078	1	10
GO:1904851	0.005780498	212.839506	0.005794078	1	10
GO:0006086	0.006356891	191.544444	0.006373486	1	11
GO:0006188	0.006356891	191.544444	0.006373486	1	11
GO:0070203	0.006356891	191.544444	0.006373486	1	11
GO:1904869	0.006356891	191.544444	0.006373486	1	11

1-10 of 32 rows | 1-6 of 7 columns Previous **1** 2 3 4 Next

```
GoPlot <- data.frame(sumGo$GOBPID,sumGo$Pvalue,sumGo$Term)
colnames(GoPlot) <-c("GO_ID_BP", "P-value", "Term")
GoPlot
```

GO_ID_BP <chr>	P-value <dbl>	Term <chr>
GO:1900182	0.001101061	positive regulation of protein localization to nucleus
GO:1900180	0.002547072	regulation of protein localization to nucleus
GO:0006006	0.004808767	glucose metabolic process
GO:0009113	0.005780498	purine nucleobase biosynthetic process
GO:0033572	0.005780498	transferrin transport
GO:1904851	0.005780498	positive regulation of establishment of protein localization to telomere
GO:0006086	0.006356891	acetyl-CoA biosynthetic process from pyruvate
GO:0006188	0.006356891	IMP biosynthetic process
GO:0070203	0.006356891	regulation of establishment of protein localization to telomere
GO:1904869	0.006356891	regulation of protein localization to Cajal body

1-10 of 32 rows Previous **1** 2 3 4 Next

KEGG ENRIGHMENT Part3

```
#Now perform KEGG ENRICHMENT

keggEnrich <- enrichKEGG(
  diffexpgenes_names_df$entrez,
  organism = "hsa",
  keyType = "kegg",
  pvalueCutoff = 0.2, #adjust this if you are not seeing any results
  pAdjustMethod = "BH",
  )
```

```
#Show results from enrichKEGG
head(keggEnrich)
```

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust
	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
hsa04810	hsa04810	Regulation of actin cytoskeleton	2/6	229/8390	0.01034827	0.1010302
hsa04144	hsa04144	Endocytosis	2/6	251/8390	0.01234916	0.1010302
hsa04614	hsa04614	Renin-angiotensin system	1/6	23/8390	0.01634067	0.1010302
hsa00020	hsa00020	Citrate cycle (TCA cycle)	1/6	30/8390	0.02126952	0.1010302
hsa04216	hsa04216	Ferroptosis	1/6	41/8390	0.02897327	0.1049876
hsa00620	hsa00620	Pyruvate metabolism	1/6	47/8390	0.03315397	0.1049876

6 rows | 1-7 of 10 columns

```
keggEnrich
```

```
## #
## # over-representation test
## #
## #...@organism     hsa
## #...@ontology     KEGG
## #...@keytype      kegg
## #...@gene     chr [1:10] "57486" "1737" "81624" "79892" "5229" "10576" "10096" "10606" ...
## #...pvalues adjusted by 'BH' with cutoff <0.2
## #...19 enriched terms found
## 'data.frame':    19 obs. of  9 variables:
##  $ ID         : chr  "hsa04810" "hsa04144" "hsa04614" "hsa00020" ...
##  $ Description: chr  "Regulation of actin cytoskeleton" "Endocytosis" "Renin-angiotensin sy
stem" "Citrate cycle (TCA cycle)" ...
##  $ GeneRatio  : chr  "2/6" "2/6" "1/6" "1/6" ...
##  $ BgRatio    : chr  "229/8390" "251/8390" "23/8390" "30/8390" ...
##  $ pvalue     : num  0.0103 0.0123 0.0163 0.0213 0.029 ...
##  $ p.adjust   : num  0.101 0.101 0.101 0.101 0.105 ...
##  $ qvalue     : num  0.0672 0.0672 0.0672 0.0672 0.0698 ...
##  $ geneID     : chr  "81624/10096" "10096/7037" "57486" "1737" ...
##  $ Count      : int  2 2 1 1 1 1 1 1 1 ...
## #...Citation
##  T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo,
and G Yu.
##  clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
##  The Innovation. 2021, 2(3):100141
```

```
#Generate a graph for the two KEGG results
#Edit the pathway id to that which is appropriate based on the ID column from the enrichKEGG o
utput

#These will generate images that will be saved to the working directory or the downloads folde
r
#Repeat for however many results you get from keggEnrich

pv.out_htmp3a <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa04144", s
pecies = "hsa")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa04144.pathview.png
```

```
## Info: some node width is different from others, and hence adjusted!
```

```
#Repeat for the second result
pv.out_htmp3b <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa04614", s
pecies = "hsa")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa04614.pathview.png
```

```
#Also show the genes involved in the pathway
#These correspond to the elements included in the image of the KEGG pathway generated earlier
pv.out_htmp3a$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
32	8027	STAM		gene	461	753	46	17	*NA*
33	9146	HGS		gene	461	770	46	17	*NA*
44	10617	STAMBP		gene	1006	828	46	17	*NA*
45	9101	USP8		gene	1006	806	46	17	*NA*
89	1759	DNM1		gene	270	181	46	17	*NA*
90	867	CBL		gene	379	271	46	17	*NA*
91	3949	LDLR		gene	533	181	46	17	*NA*
92	1211	CLTA		gene	292	269	46	17	*NA*
93	160	AP2A1		gene	292	286	46	17	*NA*
95	89853	MVB12B		gene	547	753	46	17	*NA*

1-10 of 120 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 … 12 Next

```
pv.out_htmp3b$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
6	59272	ACE2		gene	430	309	46	17	*NA*
7	59272	ACE2		gene	292	238	46	17	*NA*
8	1636	ACE		gene	430	124	46	17	*NA*
9	1636	ACE		gene	392	157	46	17	*NA*
10	1636	ACE		gene	302	437	46	17	*NA*
18	185	AGTR1		gene	603	402	46	17	*NA*
19	186	AGTR2		gene	603	317	46	17	*NA*
20	57486	NLN	57486	gene	368	267	46	17	1
21	7064	THOP1		gene	368	247	46	17	*NA*

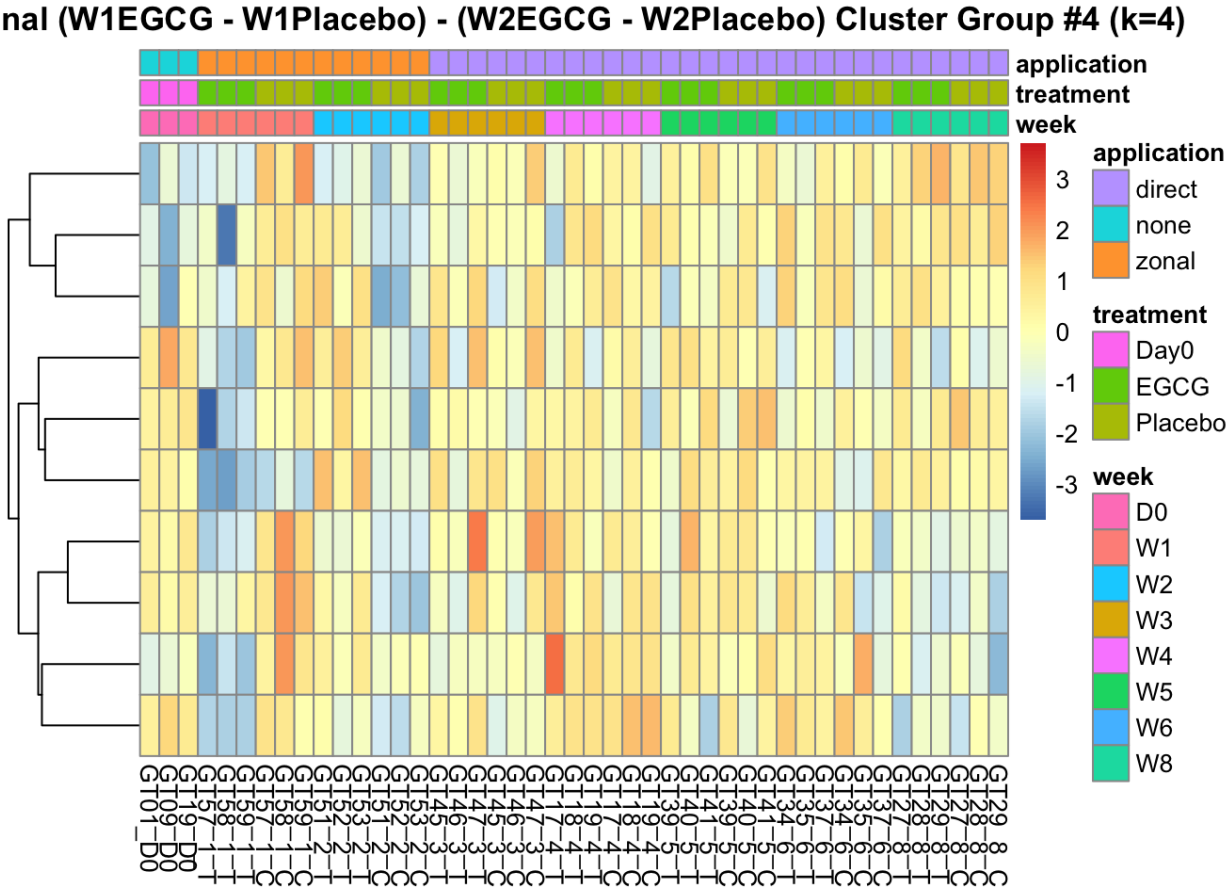
	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
22	4311	MME		gene	368	287	46	17	*NA*

1-10 of 29 rows | 1-10 of 11 columns

Previous **1** 2 3 Next

Cluster#4

```
pheatmap(interact_sig_hclust_g4,annotation_col = annotation, scale="row", cluster_cols = F, sh
ow_rownames = F, main = "Zonal (W1EGCG - W1Placebo) - (W2EGCG - W2Placebo) Cluster Group #4 (k
=4)" )
```



Go-Term Enrichment Part 4

Create HyperGoparpam

Converting the Ensemble to Entrez was achived with this code: https://www.biostars.org/p/441386/ (https://www.biostars.org/p/441386/)

```
library("AnnotationDbi")

#adding ENTREZ ID's to global gene data file
GSE124161_readcount$entrez = mapIds(org.Hs.eg.db,
                    keys=rownames(GSE124161_readcount), #Column containing Ensembl gene ids
                    column="ENTREZID",
                    keytype="ENSEMBL",
                    multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
#Wrangling the ensemble gene ID's to Entrez in the interact_sig_hclust_g4
diffexpgenes_names_df <-rownames(as.data.frame(interact_sig_hclust_g4))
diffexpgenes_names_df <-as.data.frame(diffexpgenes_names_df)

diffexpgenes_names_df$entrez = mapIds(org.Hs.eg.db,
                    keys= diffexpgenes_names_df$diffexpgenes_names_df, #Column containing Ense
mbl gene ids
                    column="ENTREZID",
                    keytype="ENSEMBL",
                    multiVals="first") #This selects the first gene alias, if there are multip
le gene names under the single EntrezID
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
diffexpgenes_names <-diffexpgenes_names_df$entrez
readcount_names <-GSE124161_readcount$entrez

#Utilized following resource for below code format https://bioconductor.org/packages/release/b
ioc/vignettes/GOstats/inst/doc/GOstatsHyperG.pdf

params <- new("GOHyperGParams",
              geneIds = diffexpgenes_names, #don't use quotes here, it will not work, you will
get an error message. This is the variable name where you stored your differentially expressed
gene names
              universeGeneIds = readcount_names, #don't use quotes here, it will not work, you
will get an error message. This is the variable name where you stored all of the gene names fr
om the whole unfiltered data set. Its the whole list of the "universe" of gene IDs for your ar
ray or reference genome.
              annotation = "org.Hs.eg",
              ontology = "BP",
              pvalueCutoff=0.01, #don't use quotes here, it will not work, you will get an err
or message
              testDirection = "over")
```

```
## Warning in makeValidParams(.Object): removing duplicate IDs in universeGeneIds
```

```
hypGO <- hyperGTest(params)
hypGO
```

```
## Gene to GO BP  test for over-representation
## 351 GO BP ids tested (10 have p < 0.01)
## Selected gene set size: 9
##     Gene universe size: 17259
##     Annotation package: org.Hs.eg
```

```
sumGo <- summary(hypGO, categorySize =10)
sumGo
```

GOBPID <chr>	Pvalue <dbl>	OddsRatio <dbl>	ExpCount <dbl>	Count <int>	Size <int>
GO:0042981	0.004051467	8.980298	0.737875891	4	1415
GO:0043067	0.004363210	8.783333	0.752998436	4	1444
GO:0033262	0.006241672	195.897727	0.006257605	1	12
GO:0010941	0.006274914	7.868342	0.832261429	4	1596
GO:0002357	0.006760245	179.562500	0.006779072	1	13
GO:0048308	0.007796669	153.892857	0.007822006	1	15
GO:0048313	0.007796669	153.892857	0.007822006	1	15

7 rows | 1-6 of 7 columns

```
GoPlot <- data.frame(sumGo$GOBPID,sumGo$Pvalue,sumGo$Term)
colnames(GoPlot) <-c("GO_ID_BP", "P-value", "Term")
GoPlot
```

GO_ID_BP <chr>	P-value <dbl>	Term <chr>
GO:0042981	0.004051467	regulation of apoptotic process
GO:0043067	0.004363210	regulation of programmed cell death
GO:0033262	0.006241672	regulation of nuclear cell cycle DNA replication
GO:0010941	0.006274914	regulation of cell death
GO:0002357	0.006760245	defense response to tumor cell
GO:0048308	0.007796669	organelle inheritance
GO:0048313	0.007796669	Golgi inheritance

7 rows

KEGG ENRIGHMENT Part4

```
#Now perform KEGG ENRICHMENT

keggEnrich <- enrichKEGG(
  diffexpgenes_names_df$entrez,
  organism = "hsa",
  keyType = "kegg",
  pvalueCutoff = 0.1, #adjust this if you are not seeing any results
  pAdjustMethod = "BH",
  )
```

```
#Show results from enrichKEGG
head(keggEnrich)
```

	ID <chr>	Description <chr>
hsa00532	hsa00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
hsa00534	hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
hsa05219	hsa05219	Bladder cancer
hsa05213	hsa05213	Endometrial cancer
hsa04730	hsa04730	Long-term depression
hsa04720	hsa04720	Long-term potentiation

6 rows | 1-3 of 10 columns

```
keggEnrich
```

```
## #
## # over-representation test
## #
## #...@organism     hsa
## #...@ontology     KEGG
## #...@keytype      kegg
## #...@gene     chr [1:10] "4585" "84937" "389643" "55423" "100420015" "26229" "126520" ...
## #...pvalues adjusted by 'BH' with cutoff <0.1
## #...35 enriched terms found
## 'data.frame':    35 obs. of  9 variables:
##  $ ID         : chr  "hsa00532" "hsa00534" "hsa05219" "hsa05213" ...
##  $ Description: chr  "Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfa
te" "Glycosaminoglycan biosynthesis - heparan sulfate / heparin" "Bladder cancer" "Endometrial
cancer" ...
##  $ GeneRatio  : chr  "1/3" "1/3" "1/3" "1/3" ...
##  $ BgRatio    : chr  "21/8390" "24/8390" "41/8390" "58/8390" ...
##  $ pvalue     : num  0.00749 0.00856 0.01459 0.0206 0.0213 ...
##  $ p.adjust   : num  0.0704 0.0704 0.0704 0.0704 0.0704 ...
##  $ qvalue     : num  0.0234 0.0234 0.0234 0.0234 0.0234 ...
##  $ geneID     : chr  "26229" "26229" "369" "369" ...
##  $ Count      : int  1 1 1 1 1 1 1 1 1 1 ...
## #...Citation
##  T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo,
and G Yu.
##  clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
##  The Innovation. 2021, 2(3):100141
```

```
#Generate a graph for the two KEGG results
#Edit the pathway id to that which is appropriate based on the ID column from the enrichKEGG o
utput

#These will generate images that will be saved to the working directory or the downloads folde
r
#Repeat for however many results you get from keggEnrich

pv.out_htmp4a <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa05219", s
pecies = "hsa")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa05219.pathview.png
```

```
## Info: some node width is different from others, and hence adjusted!
```

```
#Repeat for the second result
pv.out_htmp4b <- pathview(gene.data = diffexpgenes_names_df$entrez, pathway.id = "hsa05213", s
pecies = "hsa")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Users/dawndestefano/NYU/BIGY-7633 Transcriptomics/project
```

```
## Info: Writing image file hsa05213.pathview.png
```

```
#Also show the genes involved in the pathway
#These correspond to the elements included in the image of the KEGG pathway generated earlier
pv.out_htmp4a$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
5	3265	HRAS		gene	361	124	46	17	*NA*
8	4312	MMP1		gene	796	559	46	17	*NA*
9	1890	TYMP		gene	487	605	46	17	*NA*
10	7422	VEGFA		gene	796	506	46	17	*NA*
11	3576	CXCL8		gene	796	611	46	17	*NA*
12	7057	THBS1		gene	943	460	46	17	*NA*
13	1956	EGFR		gene	255	506	46	17	*NA*
14	999	CDH1		gene	255	575	46	17	*NA*
15	2064	ERBB2		gene	255	528	46	17	*NA*
16	1029	CDKN2A		gene	675	303	46	17	*NA*

1-10 of 32 rows | 1-10 of 11 columns Previous **1** 2 3 4 Next

```
pv.out_htmp4b$plot.data.gene
```

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
6	1499	CTNNB1		gene	641	365	46	17	*NA*
7	1499	CTNNB1		gene	469	403	46	17	*NA*
8	3265	HRAS		gene	466	263	46	17	*NA*
9	324	APC		gene	641	348	46	17	*NA*
10	8312	AXIN1		gene	596	348	46	17	*NA*
11	2932	GSK3B		gene	558	365	46	17	*NA*
12	1495	CTNNA1		gene	311	343	46	17	*NA*
13	1499	CTNNB1		gene	296	360	46	17	*NA*
15	2309	FOXO3		gene	577	216	46	17	*NA*

	kegg.names <chr>	labels <chr>	all.mapped <chr>	type <chr>	x <dbl>	y <dbl>	width <dbl>	height <dbl>	mol.data <dbl>
16	572	BAD		gene	577	185	46	17	*NA*