DGCs expression in the *P. aeruginosa* compendium

Because DGC genes have relatively low dynamic ranges in their expression as measured by nanostring in different carbon sources, perhaps DGC genes in P. aeruginosa also have relatively low dynamic ranges in the P. aeruginosa compendium of PAO1 microarray data. As a group pf comparator genes we will look at the ranges of expression of Pho genes across the compendium thinking that they are genes that likely very wildly in their expression levels.

Aquiring updated *P. aeruginosa* gene expression compendium

First, I got an updated version of the P.a compendium by cloning and following the protocol in Jie's ADAGE repository

link

Briefly, I ran scripts written by Jie to download every file in the EBI ArrayExpress FTP site with 'A-AFFY-30' in the array name. These files were converted from CEL to text files and control probes were removed then concatenated into a file containing gene expression values for each PAO1 gene probe in each sample in the compendium. The compendium was then transformed linearly to zero-one normalize within samples

$M = M - rowMin(M)/(rowMax(M) - rowMin(M))$

where M is a matrix of gene expression values made of rows of samples and columns of PAO1 genes.

The matrix of normalized gene expression values can now be loaded into R to begin the analysis. (Note: I also ran this analysis on the compendium without normalization and normalize by gene and this made the most sense to me and looked like the version without any normalization because if we are looking at expression range from the perspective of genes it is confounding to normalize those ranges to all be the same.)

Loading data

Load the compendium:

```
# load compendium
compendium <- read.csv('../data/Pa_compendium_norm_samp.csv', sep = '\t', row.names=1, stringsAsFactors
```

Note the compendium is made of 5549 genes and 1185 samples.

Also load list of PAO1 DGC gene from Alan:

```
# load DGC genes
dgc_genes <- read.csv('../data/dgc_pao1.csv', stringsAsFactors = FALSE)
```

Node there are 49 DGC genes. Also the list of pho genes from the Haussler regulon (*Bielecki et al 2015, Nuc. Ac. Res.*):

```
# load pho genes
hous_pho_regulon <- read.csv('../data/Hous_pho_genes.csv', stringsAsFactors = FALSE)
# regulon was defined in PAO1 but we only use genes with PAO1 homologs
pho_genes <- hous_pho_regulon[hous_pho_regulon$PAO1.ID != "",]
```

Note this results in 160 pho genes.

Now that I have gene lists I can subset the compendium to just look at DGC and pho genes:

```
# only include DGC and pho genes that are in the compendium
dgc_comp <- dgc_genes$Locus[dgc_genes$Locus %in% rownames(compendium)]
pho_comp <- pho_genes$PAO1.ID[pho_genes$PAO1.ID %in% rownames(compendium)]
dgc_act <- compendium[dgc_comp,]
dgc_pho_act <- compendium[c(dgc_comp, pho_comp),]
```

Note that 0 DGC genes and 2 pho genes are not in the compendium (PA4212, PA1899).

Plotting gene expression

First, look at straightforward gene expression in a heatmap of DGC genes:

```
library(heatmap3)
pdf("plots/dgc_activities.pdf", width = 14, height=14)
heatmap3(as.matrix(dgc_act), scale="none", labCol=NA)
dev.off()
```

```
png("plots/dgc_activities.png", width = 800, height = 800)
heatmap3(as.matrix(dgc_act), scale="none", labCol=NA)
dev.off()
```

Overall, it looks pretty blue, pretty monotone. But this doesn't tell us about the ranges of expression compares to other genes, so lets add the Pho genes as comparators:

```
pdf("plots/dgc_pho_activities.pdf",width = 14, height=14)
heatmap3(as.matrix(dgc_pho_act),  scale="none", labCol=NA,
         Rowv=NA,
         RowSideLabs = "DGC:Pho",
         RowSideColors =c( rep("orangered", length(dgc_comp)), rep("turquoise",length(pho_comp)))
         )
dev.off()
```

```
png("plots/dgc_pho_activities.png",width = 800, height = 800)
heatmap3(as.matrix(dgc_pho_act),  scale="none", labCol=NA,
         Rowv=NA,
         RowSideLabs = "DGC:Pho",
         RowSideColors =c( rep("orangered", length(dgc_comp)), rep("turquoise",length(pho_comp)))
         )
dev.off()
```

I think that it does look like some Pho genes have deeper blues and brighter reds than the DGC genes, but it is hard to say if there is an overall trend or just examples that could be picked out as good examples... I turned off clustering so that the DGC genes (green) and the Pho genes(orange) would be separate, but let's see if the two groups cluster separately:

```
pdf("plots/dgc_pho_activities_clust.pdf",width = 14, height=14)
heatmap3(as.matrix(dgc_pho_act),  scale="none", labCol=NA,
         RowSideLabs = "DGC:Pho",
         RowSideColors =c( rep("orangered", length(dgc_comp)), rep("turquoise",length(pho_comp)))
         )
dev.off()
```

```
png("plots/dgc_pho_activities_clust.png",width = 800, height = 800)
heatmap3(as.matrix(dgc_pho_act),  scale="none", labCol=NA,
         RowSideLabs = "DGC:Pho",
         RowSideColors =c( rep("orangered", length(dgc_comp)), rep("turquoise",length(pho_comp)))
```

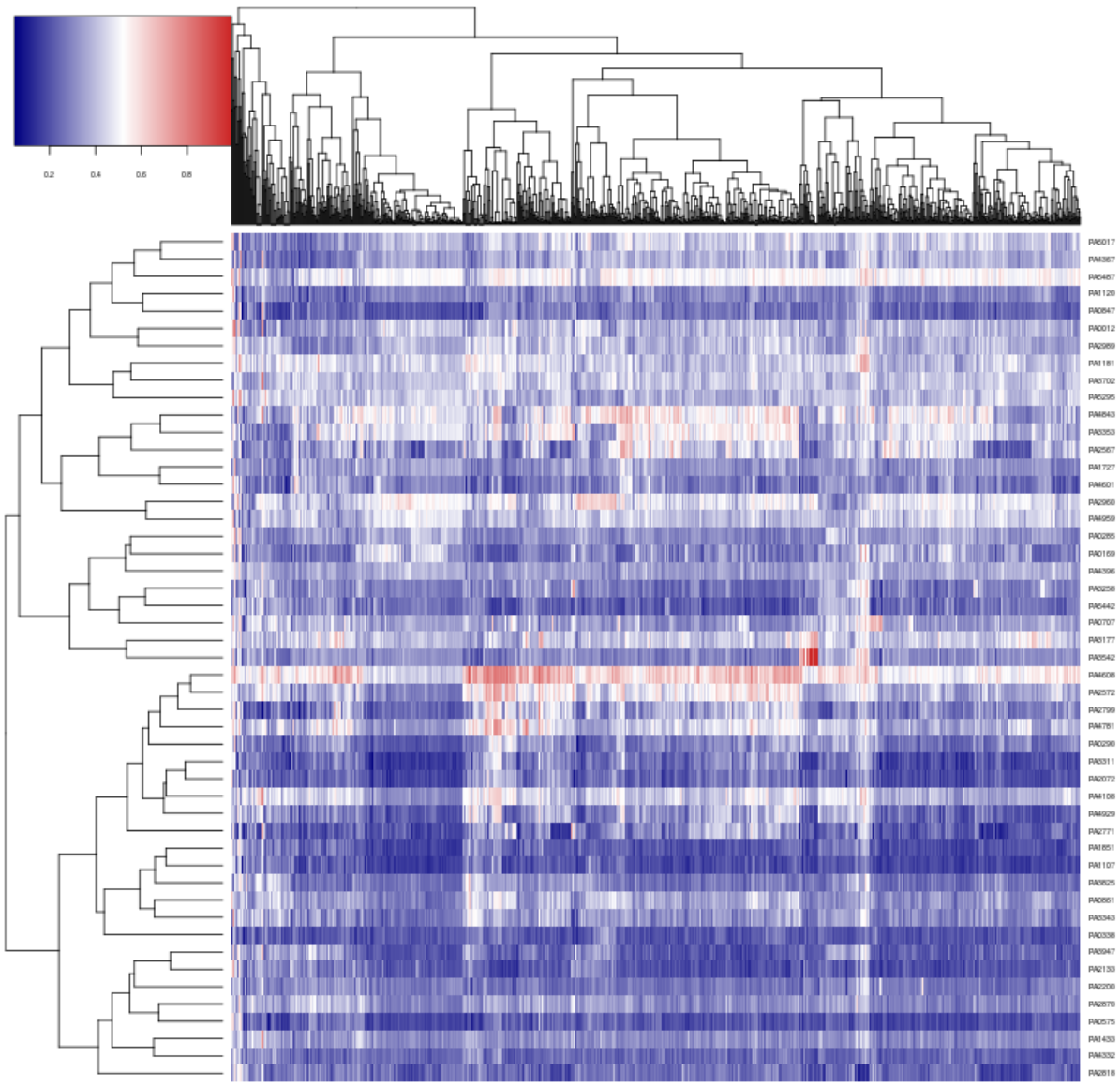


Figure 1:

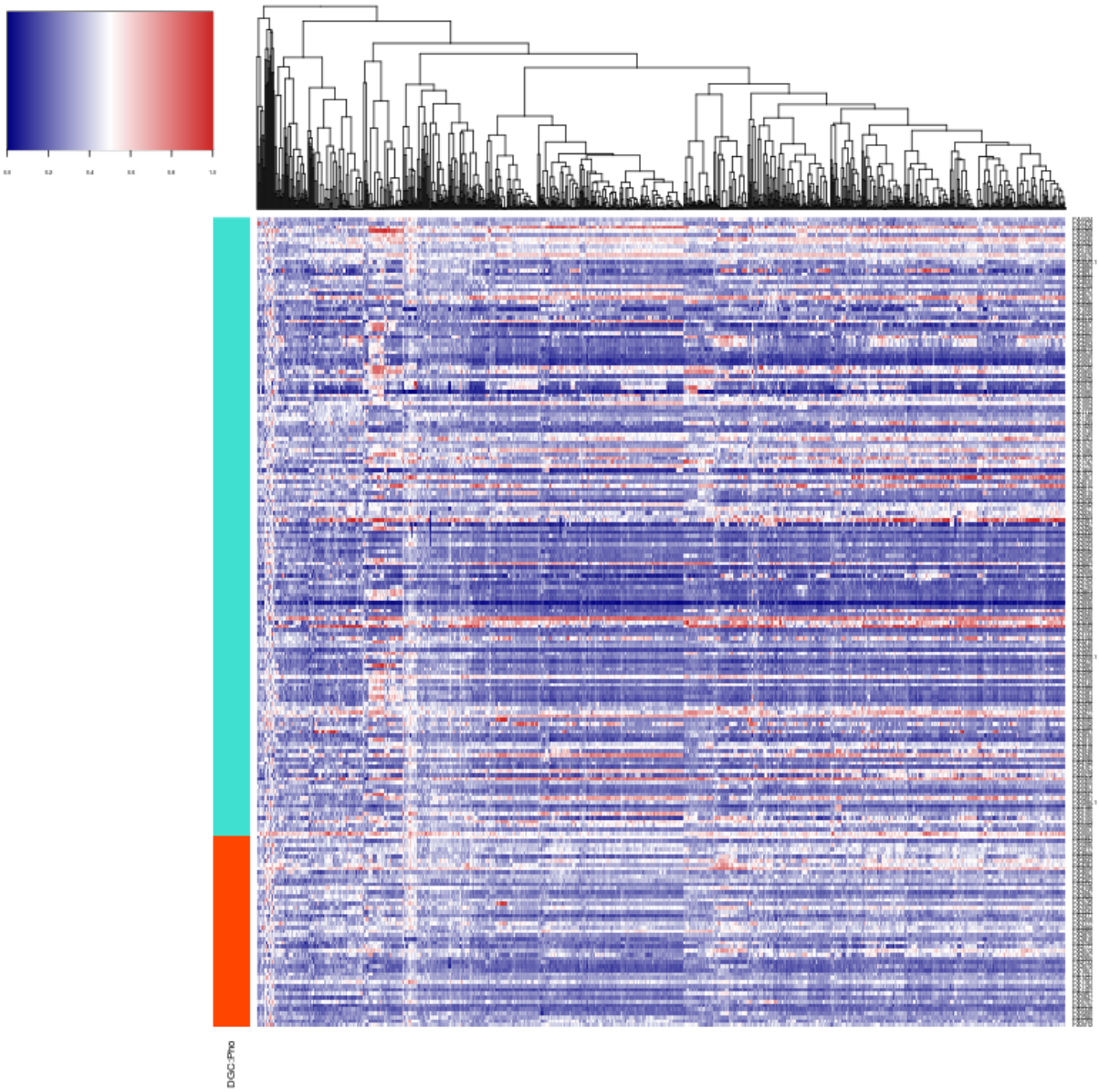


Figure 2:

```
        )
dev.off()
```

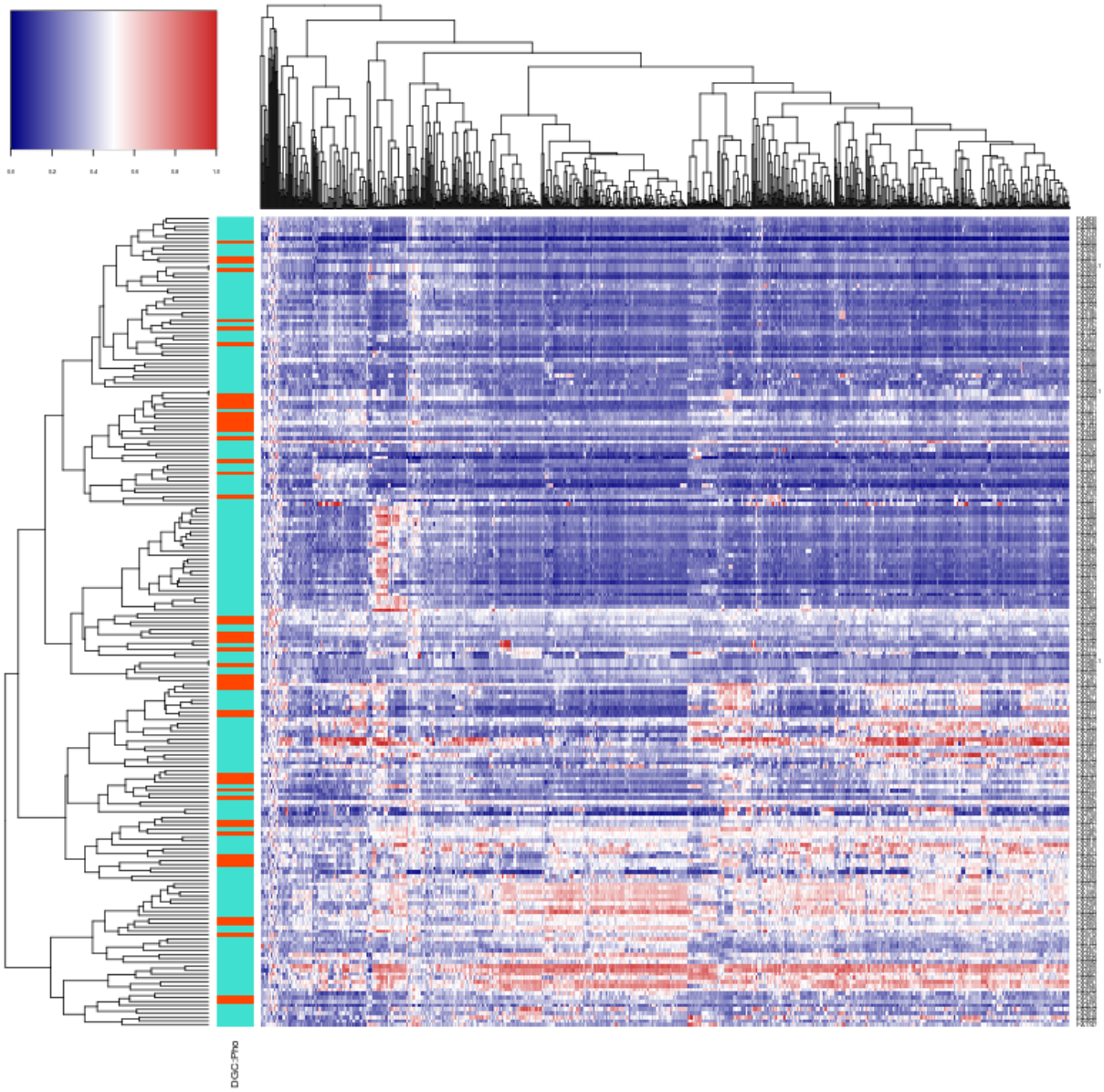


Figure 3:

Because the Pho genes and DGC genes don't cluster in their respective groups, I don't thin k clearly illustrates anything more about their respective ranges. I think I will take another approach that Alan suggested yesterday, looking at each gene and it's distribution of expression values.

Plotting ranges of gene expression

```
library(ggplot2)
library(reshape2)
```

To make the data easier to use in ggplot2, I just have to do a little massaging.

```
dgc_melt <- melt(as.matrix(dgc_pho_act))
colnames(dgc_melt) <- c("Gene", "Sample", "Normalized_Expression")
dgc_melt$Group <-rep(c(rep("DGC", length(dgc_comp)), rep("Pho", length(pho_comp))), ncol(dgc_pho_act))
```

```
g1 <- ggplot(dgc_melt, aes(Gene, Normalized_Expression)) +
  geom_boxplot(aes(color = Group)) +
  theme(axis.text.x = element_text(angle=90, hjust=1, size=7))

#g1
pdf("plots/dgc_pho_activities_boxplots.pdf",width = 20, height=5)
g1
dev.off()
```

```
png("plots/dgc_pho_activities_boxplots.png",width = 1800, height=600)
g1
dev.off()
```

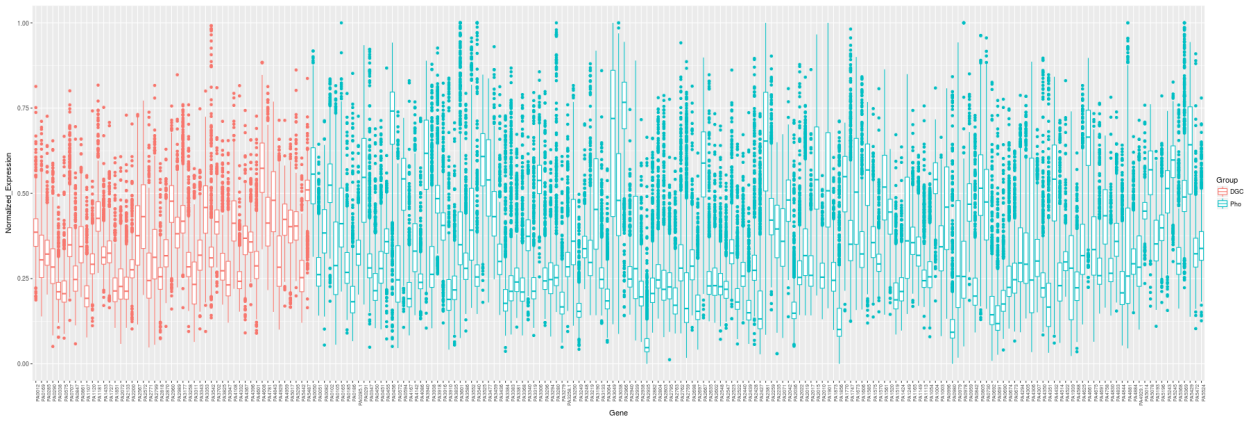


Figure 4:

Here we see the DGC genes in orange and pho genes in blue. It is difficult to compare the ranges of the genes to each other because the ranges are fairly wise with lots of outliers and there are so many genes. Maybe ordering the genes by standard deviation will allow us to see if all DGC genes are on the lower end and pho genes on the upper end:

```
dgc_melt_order <- melt(as.matrix(dgc_pho_act))
colnames(dgc_melt_order) <- c("Gene", "Sample", "Normalized_Expression")
dgc_melt_order$Group <-rep(c(rep("DGC", length(dgc_comp)), rep("Pho", length(pho_comp))), ncol(dgc_pho_a
dgc_melt_order$Gene <- reorder(dgc_melt_order$Gene, dgc_melt_order$Normalized_Expression, function(x) s
```

```
g1.5 <- ggplot(dgc_melt_order, aes(Gene, Normalized_Expression)) +
  geom_boxplot(aes(color = Group)) +
  theme(axis.text.x = element_text(angle=90, hjust=1, size=7))

#g1.5
pdf("plots/dgc_pho_activities_ordered_boxplots.pdf",width = 20, height=5)
g1.5
```

```
dev.off()
```

```
png("plots/dgc_pho_activities_ordered_boxplots.png",width = 1800, height=600)
g1.5
dev.off()
```

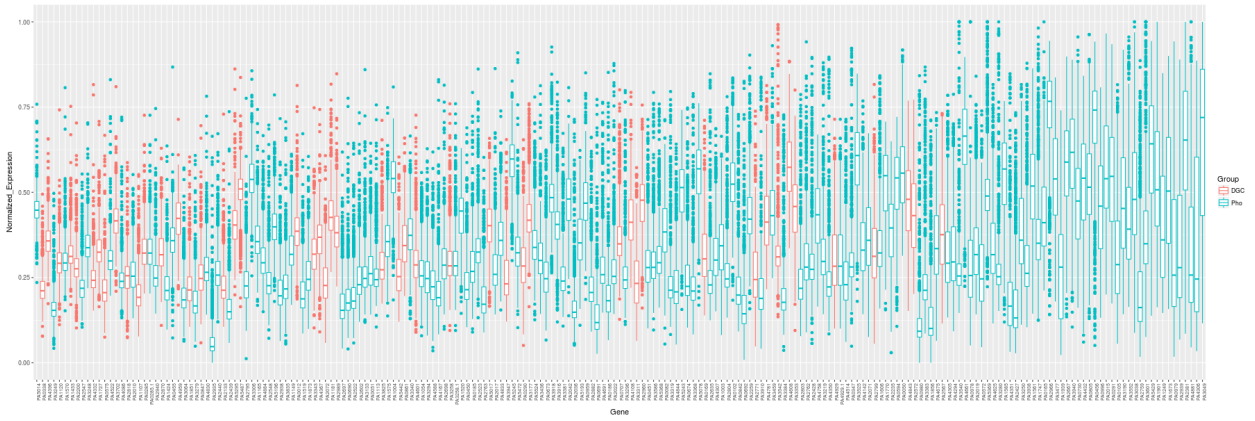


Figure 5:

A little easier to see. The Pho genes make up the right of the plot with larger st devs but it is not a very sharp divide. Also there are over twice as many pho genes as DGC genes here so it is hard to say that the exclusion of DGC genes from the left of the graph isn't because of sampling bias.

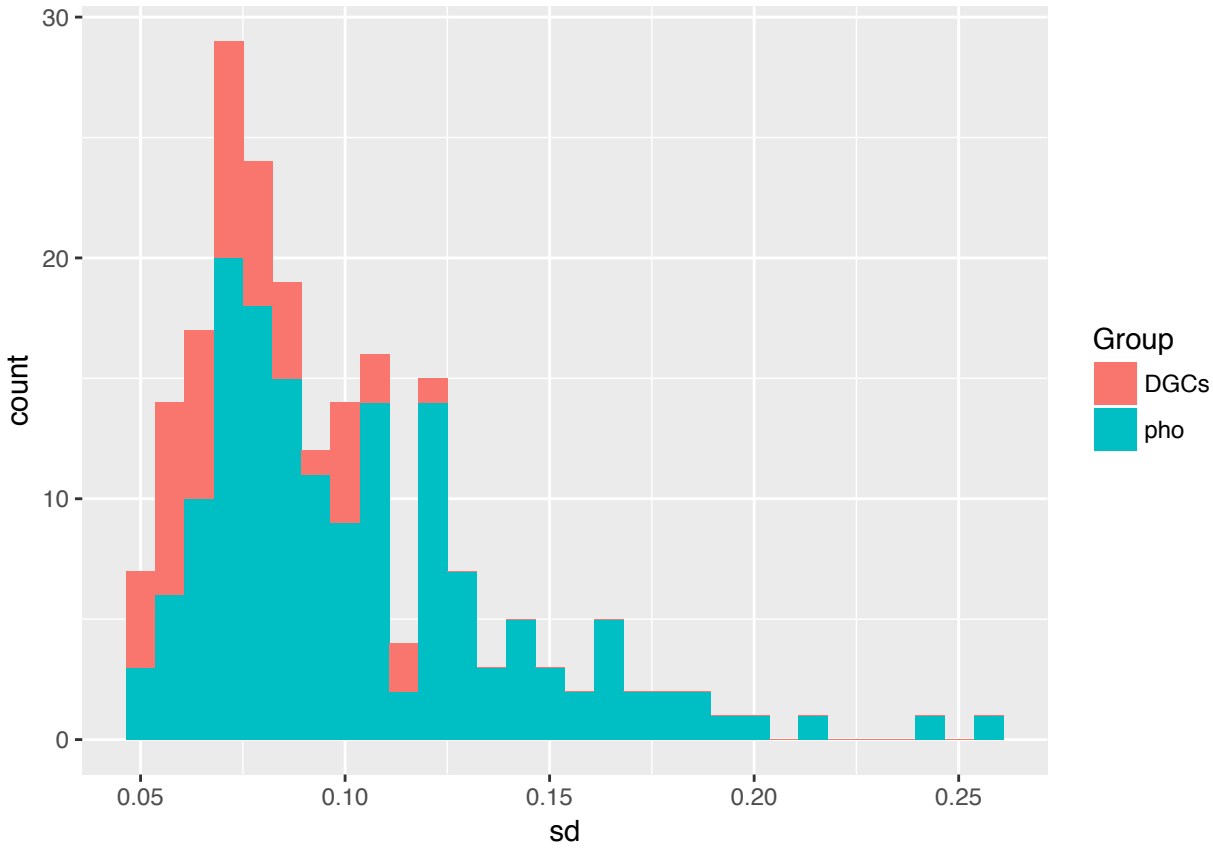
I can ask the question a little more directly by plotting the DGC genes and Pho genes as groups and looking at summary statistics.

Plotting summary stats

First I'll look at st dev as a measure of variability that is a little more robust to outliers than range. I will look at the distributions of st dev in gene expression value for both DGC and pho genes:

```
dgc_pho_summ <- data.frame(rownames(dgc_pho_act))
dgc_pho_summ$sd <- apply(as.matrix(dgc_pho_act), 1,function(x) sd(x))
dgc_pho_summ$range <- apply(as.matrix(dgc_pho_act), 1,function(x) max(x) - min(x))
dgc_pho_summ$mad <- apply(as.matrix(dgc_pho_act), 1,function(x) mad(x))
dgc_pho_summ$Group <- c(rep("DGCs", length(dgc_comp)), rep("pho", length(pho_comp)))
g3 <- ggplot(dgc_pho_summ, aes(sd)) +
  geom_histogram(aes(fill=Group))
g3
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
pdf("plots/dgc_pho_activities_dist.pdf",width = 14, height=10)
g3
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
dev.off()
```

```
png("plots/dgc_pho_activities_dist.png",width = 800, height = 600)
g3
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
dev.off()
```

The Pho genes do appear to have a tail extending toward higher st dev suggesting some pho genes, while not the majority, have st dev above those of DGC genes. This plot also shows that more DGC genes have st devs between 0.05 and 0.1 than pho genes which may indicate DCG genes are more tightly distributed around a relatively low mean.

If we are looking at distributions, we may as well include the entire compendium. I will plot a third group of genes which is all genes from the PAO1 chip that aren't on the pho or DGC lists.

Again, calculate range and st dev:

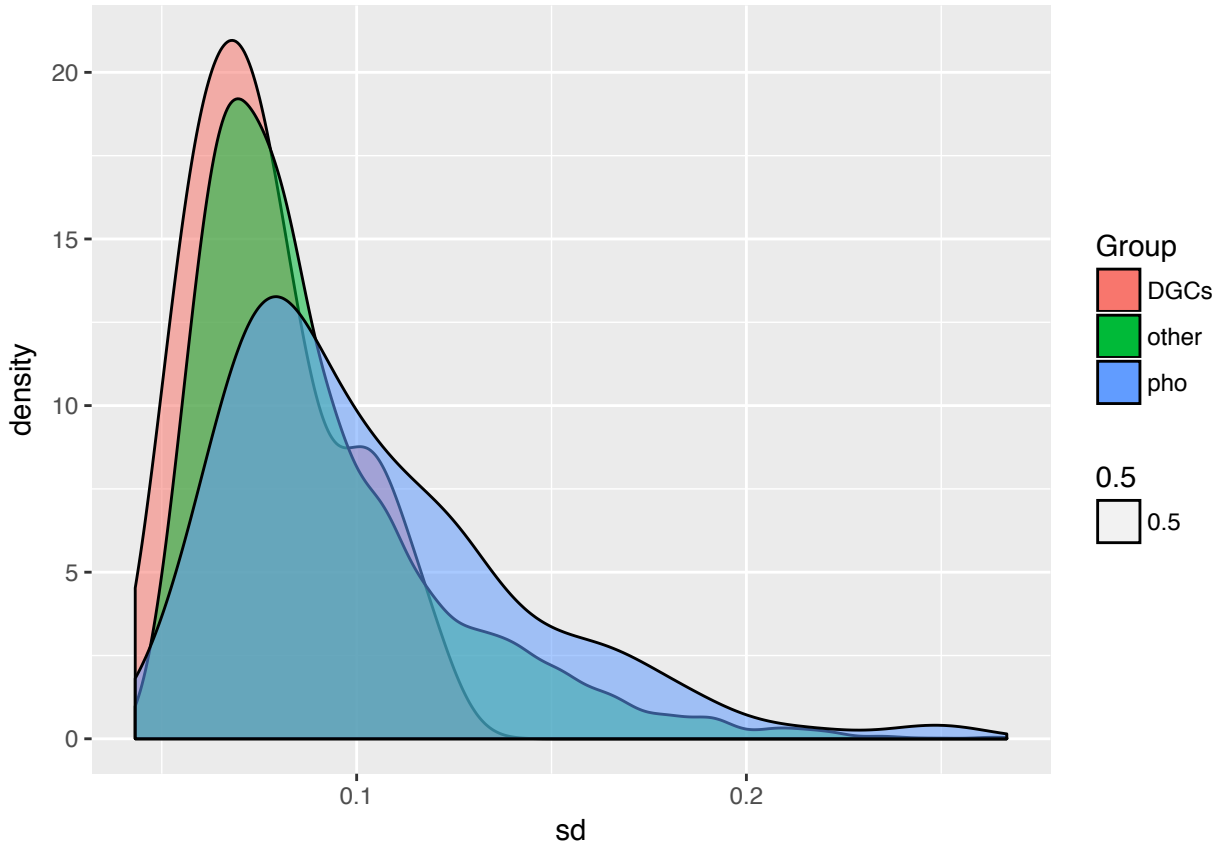
```
other_act <- compendium[!(rownames(compendium) %in% rownames(dgc_pho_act)),]
all_act <- compendium[c(dgc_comp, pho_comp, rownames(other_act)),]

all_act_summ <- data.frame(rownames(all_act))
all_act_summ$sd <- apply(as.matrix(all_act), 1,function(x) sd(x))
all_act_summ$range <- apply(as.matrix(all_act), 1,function(x) max(x) - min(x))
all_act_summ$mad <- apply(as.matrix(all_act), 1,function(x) mad(x))
```

```
all_act_summ$Group <- c(rep("DGCs", length(dgc_comp)), rep("pho", length(pho_comp)), rep("other", nrow(
all_act_summ$range <- apply(as.matrix(all_act), 1,function(x) max(x) - min(x))
```

Standard Deviation distributions

Plot distributions of st dev:

```
g4 <- ggplot(all_act_summ, aes(sd)) +
  geom_density(aes(fill=Group, alpha = .5))
g4
```



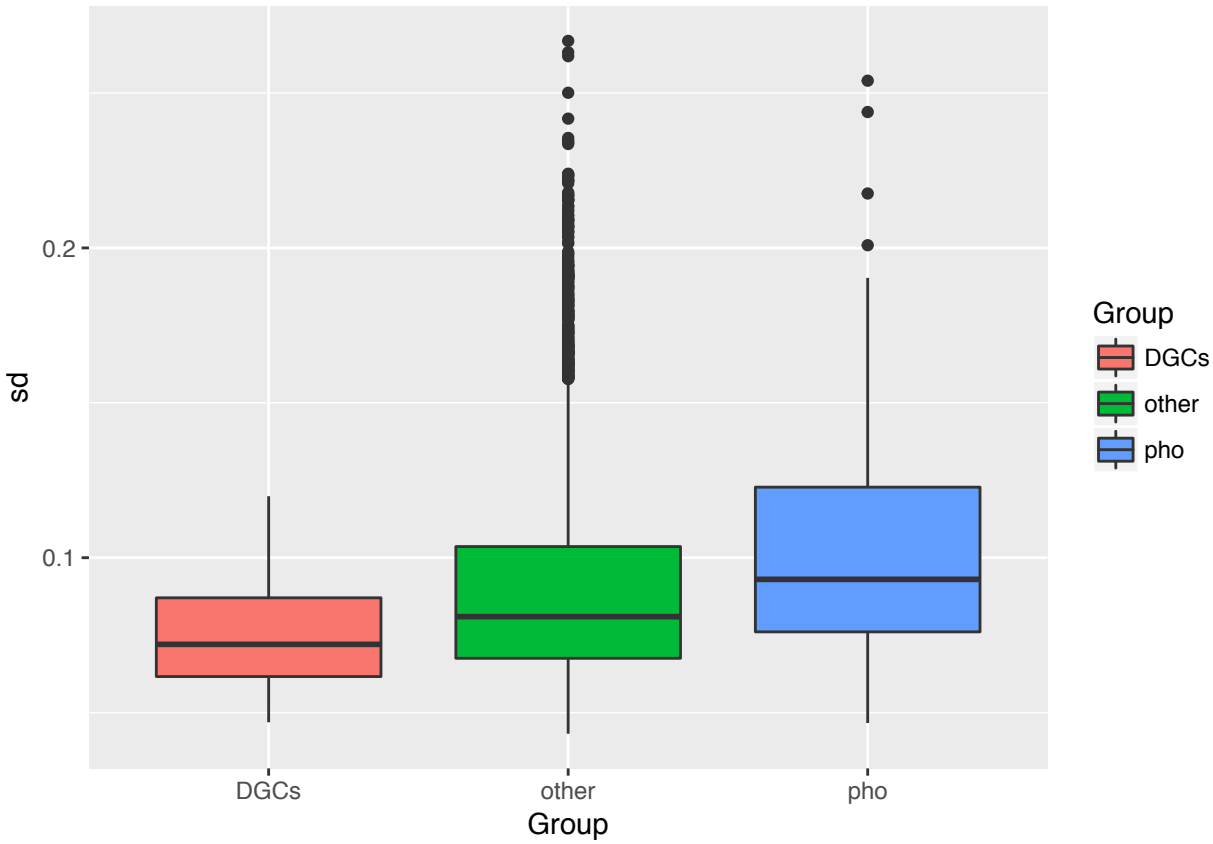

```
pdf("plots/all_activities_dens.pdf",width = 14, height=10)
g4
dev.off()
```

```
png("plots/all_activities_dens.png",width = 800, height = 600)
g4
dev.off()
```

I think this plot shows that the DGC genes have pretty average st dev of expression across the compendium but also represent genes with some of the lowest st devs. Again we see that pho genes are present in the outliers and DGC genes have a tighter distribution, albeit with a small shoulder here.

A boxplot is just another way of representing the same thing:

```
g5 <- ggplot(all_act_summ, aes(Group, sd)) +
  geom_boxplot(aes(fill=Group))
g5
```




```
pdf("plots/all_activities_box.pdf",width = 10, height=14)
g5
dev.off()
```

```
png("plots/all_activities_box.png",width = 600, height = 800)
g5
dev.off()
```

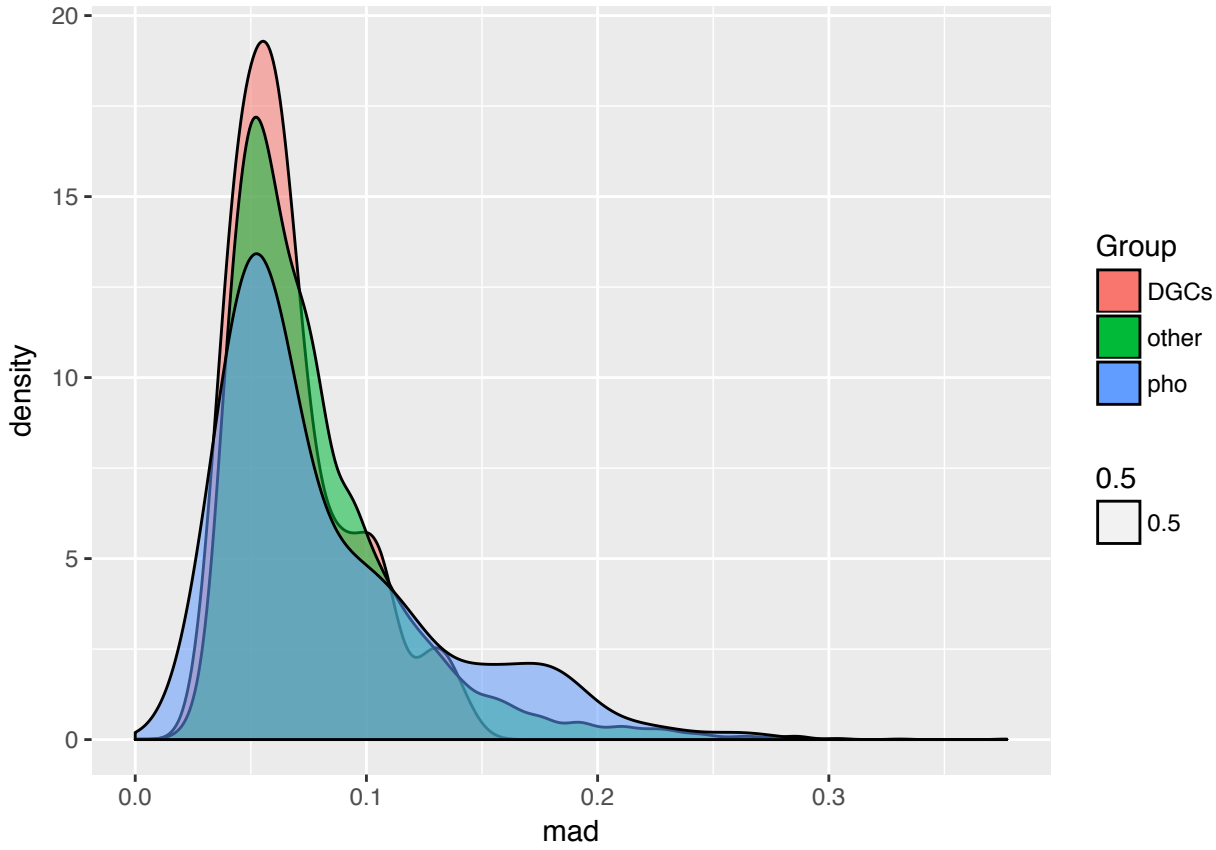
We can see that DGC genes are not outliers in either direction - they don't vary as much as some of the pho genes and some other genes and their average st dev is slightly below all other genes and even further below pho genes but there are some genes with smaller st dev than any of the DGC genes. I'm not quite sure if it appropriate to run a statistic like a t-test here to compare DGC st dev to pho or other genes..

Median Absolute Devaition distributions

And just to look at median absolute deviation as another measure of variation robust to outliers where

$MAD = 1.4826 * median(|X_i - median(X)|)$

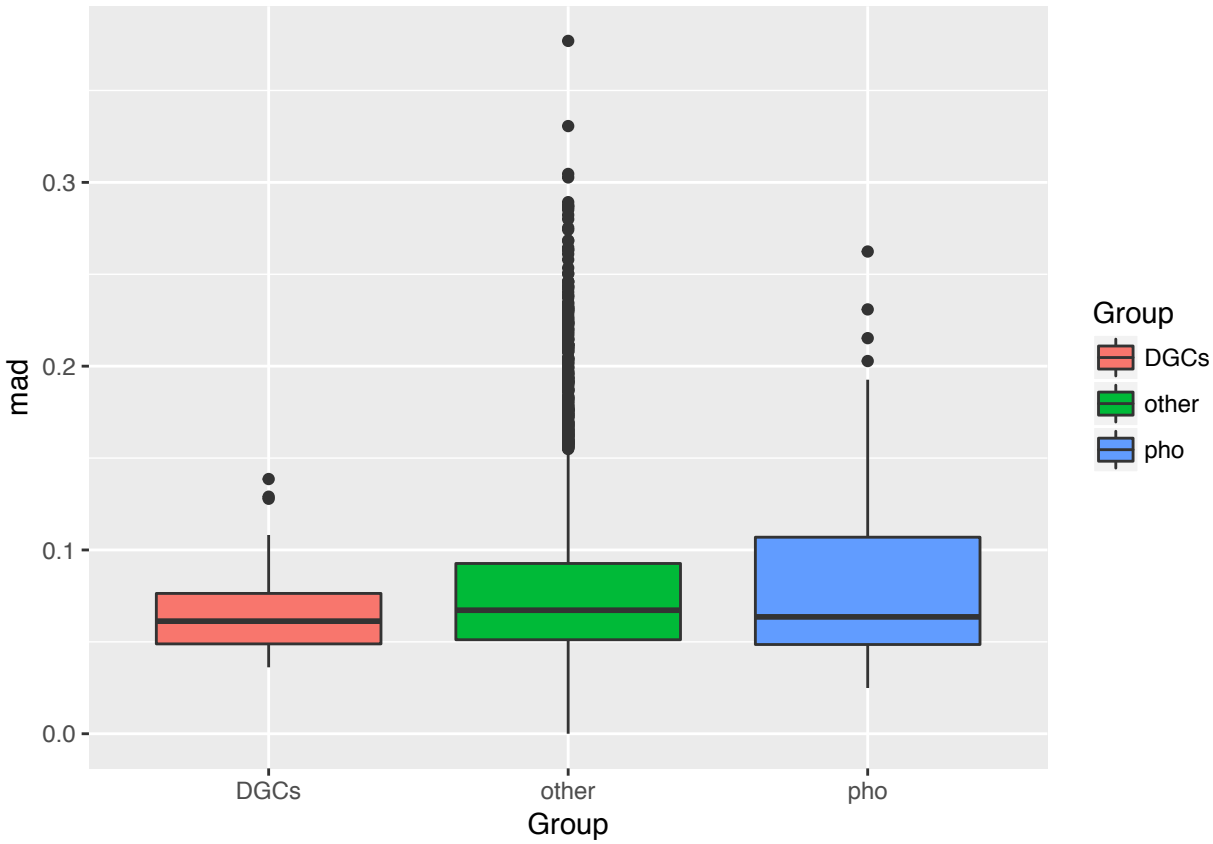
```
g6 <- ggplot(all_act_summ, aes(mad)) +
  geom_density(aes(fill=Group, alpha = .5))
g6
```




```
pdf("plots/all_activities_mad_dens.pdf",width = 14, height=10)
g6
dev.off()
```

```
png("plots/all_activities_mad_dens.png",width = 800, height = 600)
g6
dev.off()
```

```
g7 <- ggplot(all_act_summ, aes(Group, mad)) +
  geom_boxplot(aes(fill=Group))
g7
```




```
pdf("plots/all_activities_mad_box.pdf",width = 10, height=14)
g7
dev.off()
```

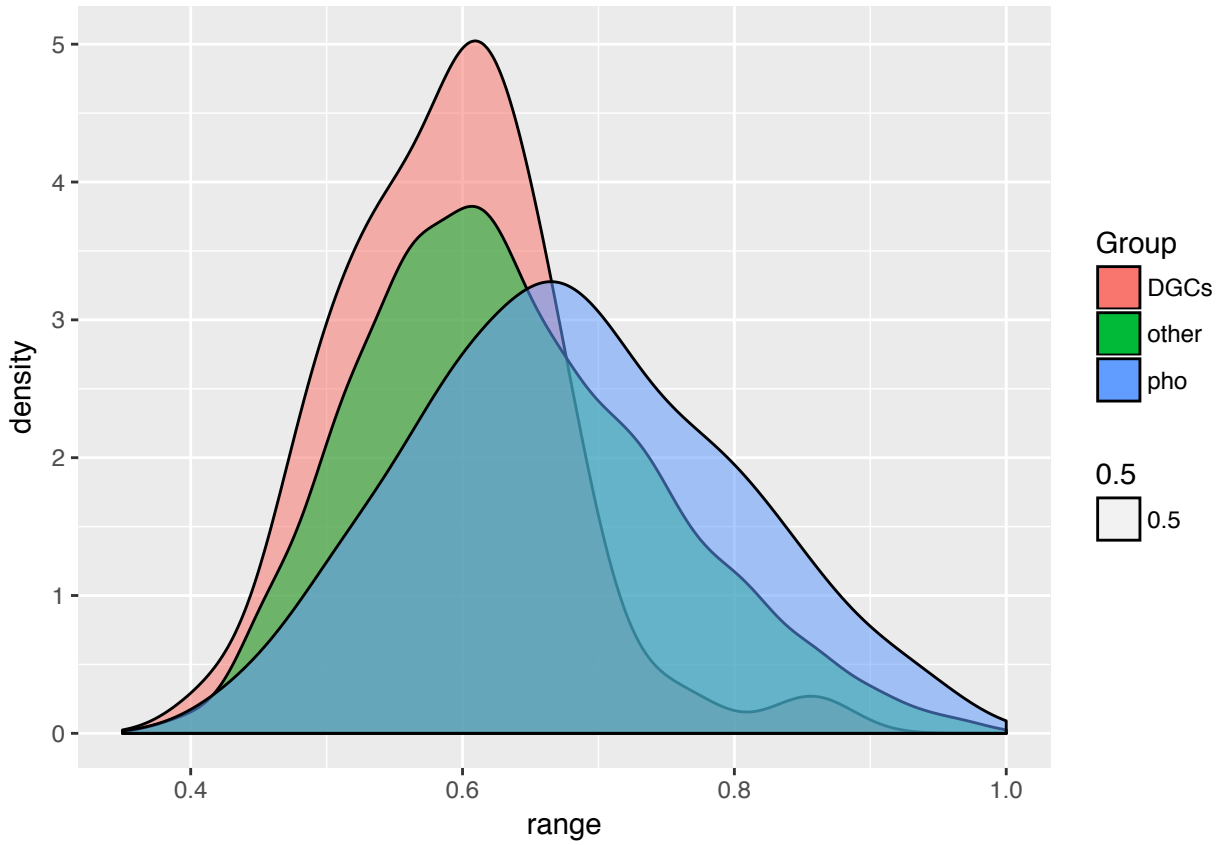
```
png("plots/all_activities_mad_box.png",width = 600, height = 800)
g7
dev.off()
```

The trends here are similar. Perhaps they are less dramatic because the outliers in the pho and other categories are part of what contrasts against the DGC genes.

Range distribtutions

And for completeness, I'll also look at range:

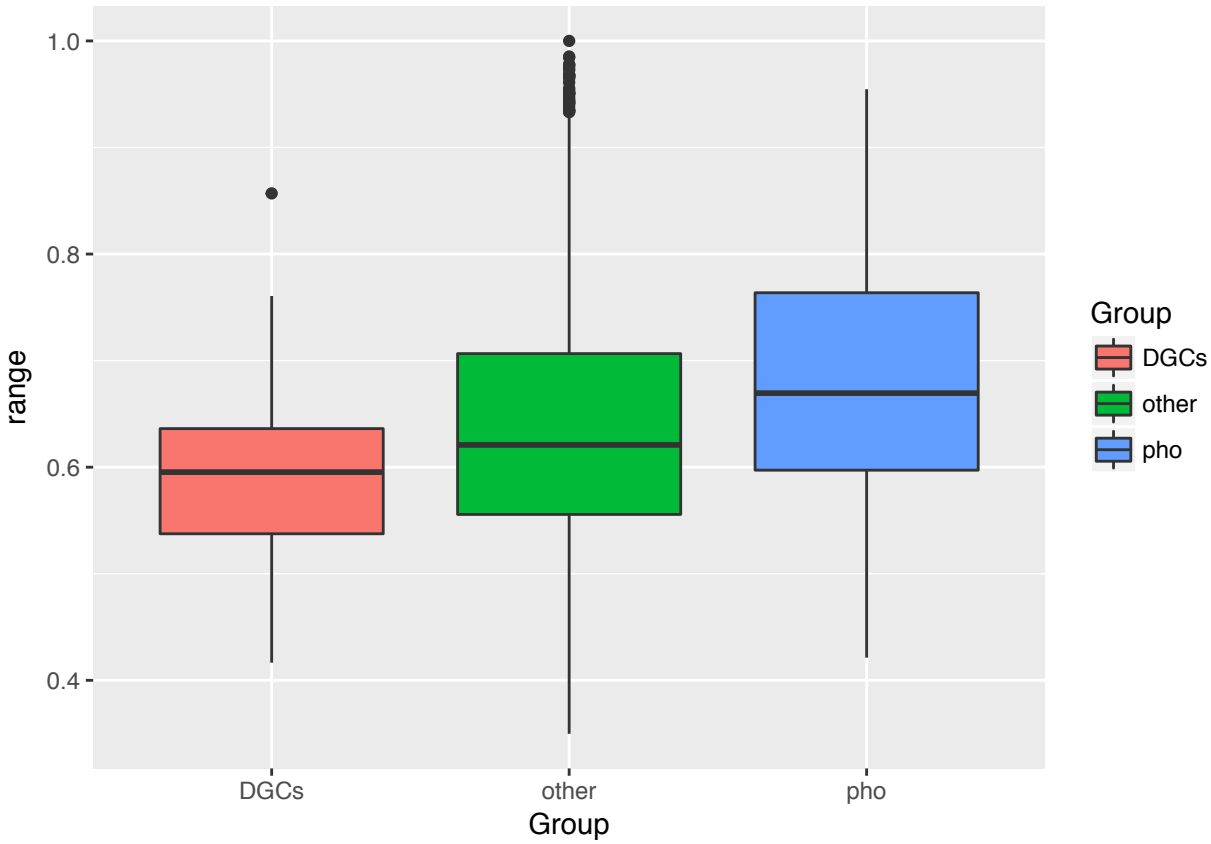
```
g8 <- ggplot(all_act_summ, aes(range)) +
  geom_density(aes(fill=Group, alpha = .5))
g8
```




```
pdf("plots/all_activities_range_dens.pdf",width = 14, height=10)
g8
dev.off()
```

```
png("plots/all_activities_range_dens.png",width = 800, height = 600)
g8
dev.off()
```

```
g9 <- ggplot(all_act_summ, aes(Group, range)) +
  geom_boxplot(aes(fill=Group))
g9
```




```
pdf("plots/all_activities_range_box.pdf",width = 10, height=14)
g9
dev.off()
```

```
png("plots/all_activities_range_box.png",width = 600, height = 800)
g9
dev.off()
```

Interesting that the density plot for range is perhaps the most compelling showing the most separation between DGC and Pho genes but even with the wider distributions, the differences in the modes and (as seen in the boxplot) means may be slightly less compelling. But I guess those are just aesthetic thoughts as the trends are about the same whether I look at st dev, mad or range and that consistency is confirming.

The data directory in the parent dir of this R project contains the compendium of gene expression (normalized by sample, normalized by gene and non-normalized) so it should be readily usable for any further analysis for whatever Alan wants.