#6.1

Linear Regression and Correlation

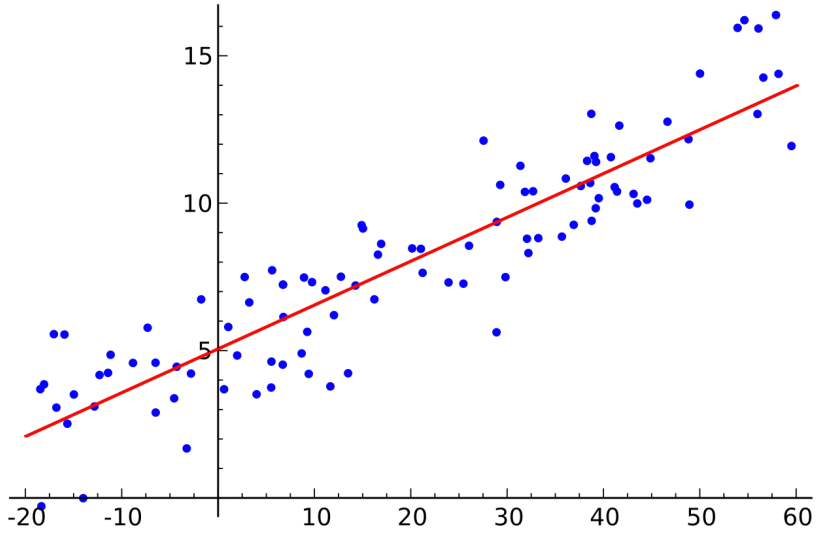
Heatmaps

Boxplots, violin plots, swarm plots

Outputting figures

Linear regression

- Model a scalar relationship between a dependent continuous variable and independent variable(s)
- Attempts to fit a slope to the data to best describe it
- Allows for correction of confounding variables



Using the same data from last class…

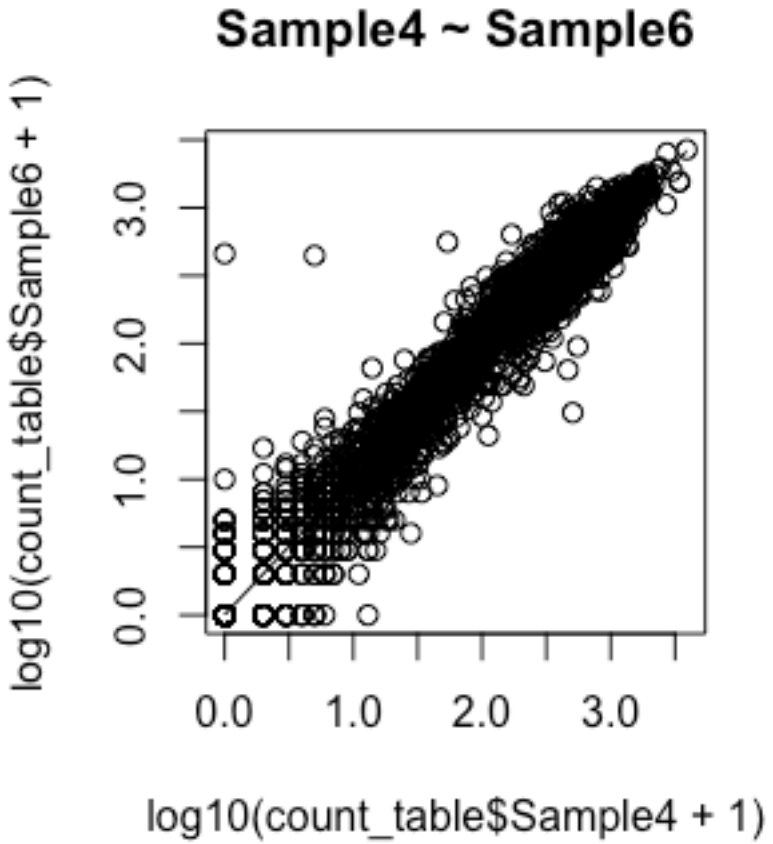
```
count_table <- read.table("~/Desktop/bioinformatics-bootcamp/class_6_1/sce_mct1_03hr_counts.txt", sep = '\t', header = TRUE, row.names = 1, check.names=F)
```

Linear modeling

```
> linearMod <- lm(
      Sample4 ~ Sample6,
      data=count_table)
> print(linearMod)
```

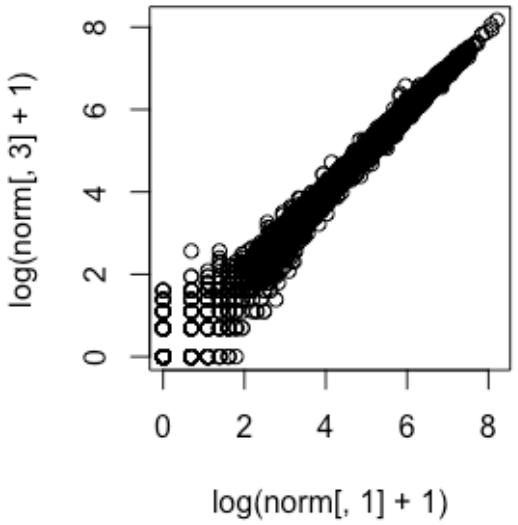
Coefficients:

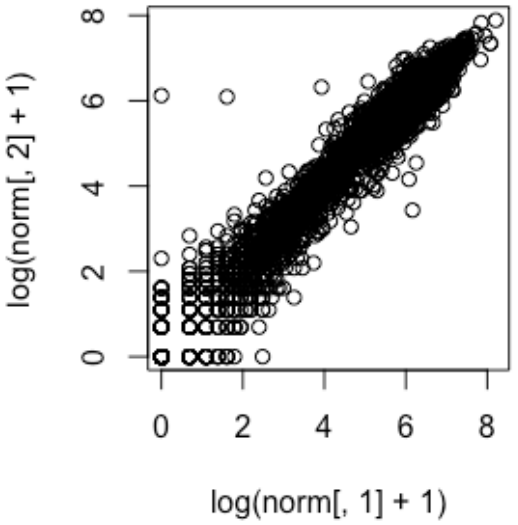
(Intercept)	Sample6	
-0.7282	1.1097	<- a near linear relationship



Correlation

- A common measure in many biological contexts to compare replicates, etc.
- Level of interdependence between two variables





Correlation examples in R

```
> cor(count_table["Sample4"], count_table["Sample6"])
            Sample6
Sample4 0.943669
> cor(count_table["Sample4"], count_table["Sample10"])
             Sample10
Sample4 0.9950528
```

Which samples are biological replicates?

Correlation examples in R

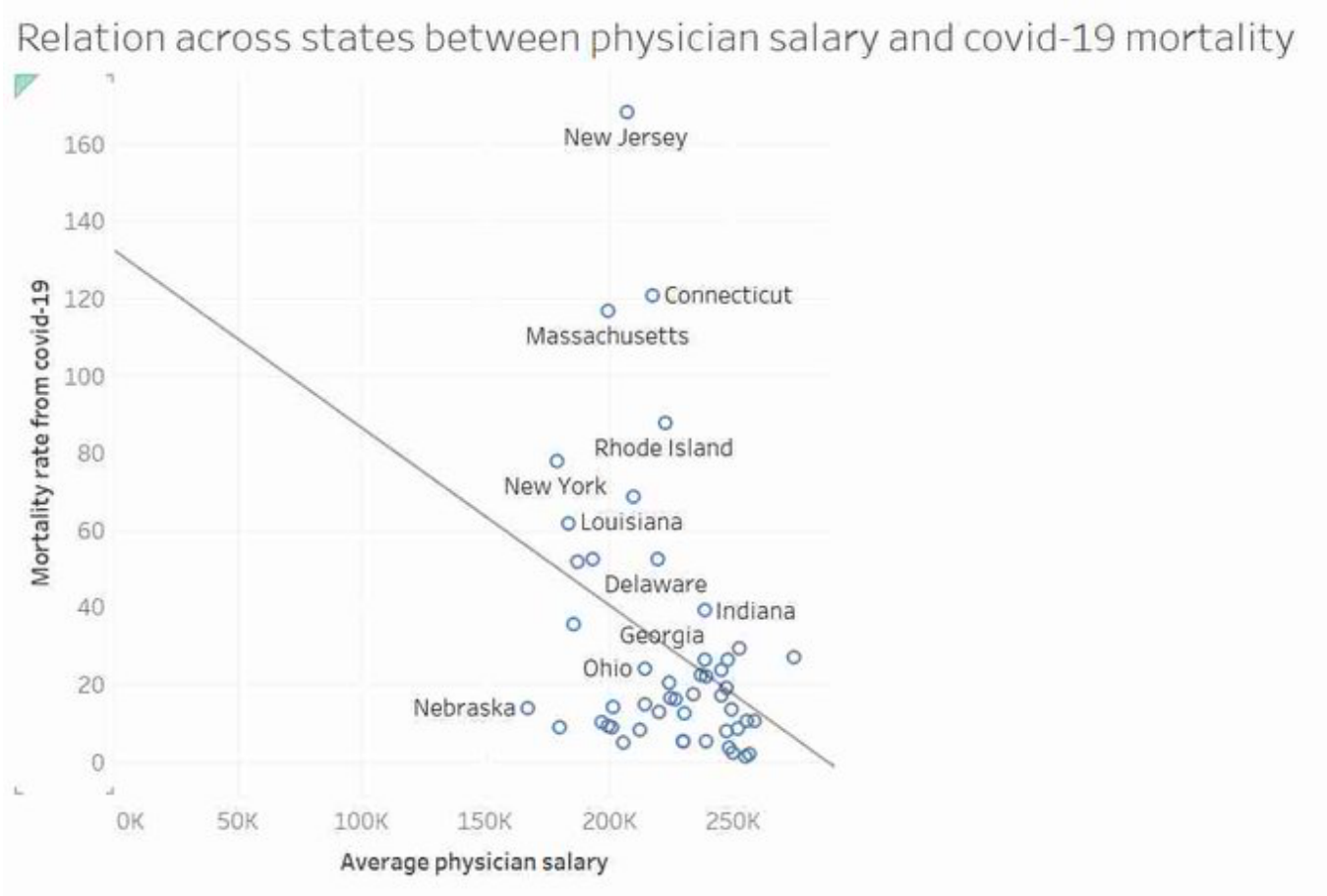
- The cor() function by default uses a Pearson (R) correlation
  - Expects datasets to be normally distributed
- By running cor(…, method="spearman"), a Spearman (ρ) correlation coefficient will be calculated
  - Rank order correlation
  - Does not assume datasets are normally distributed (better for RNA-seq)
- Coefficients are directional (negative or positive correlation)
- The Coefficient of Determination (R^2) measures the proportion of the variance in the dependent variable that is predictable from the independent variable

Correlation coefficients

- The meaning of a correlation coefficient will vary field to field
- Examples:
  - Biological replicates: > 0.99 indicate correlation
  - Human biological samples: something over 0.60 might be a correlation
  - Psychology: something over 0.20 might show a correlation

What isn’t correlation?

- An example of why fitting a line can be misleading
- The math will try to fit a line as best it can, but it’s always important to look at your scatterplot before fitting a line and see if you can naturally see a trend
- My recommendation: Use a scatterplot with no line, provide correlation coefficient

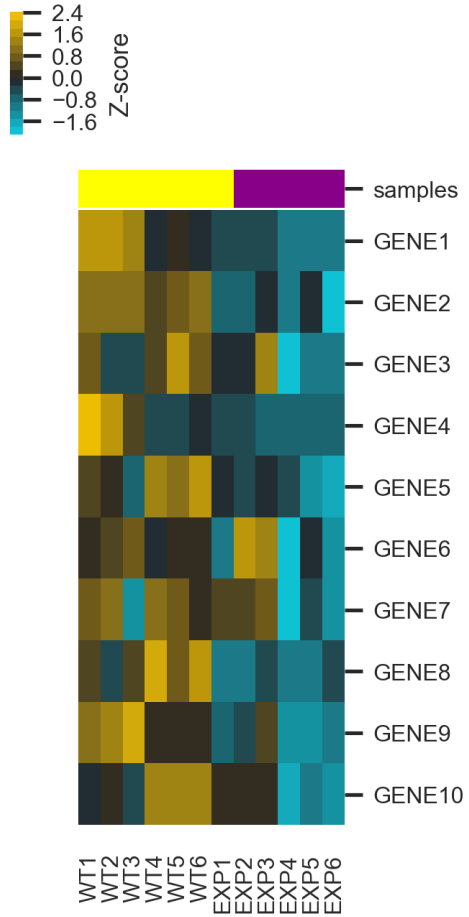


Heatmaps

- Look at broad patterns across several variables between samples
- Use sample to sample normalized data (as appropriate)
- Use z-scored gene values

Why do we use Z-scores?

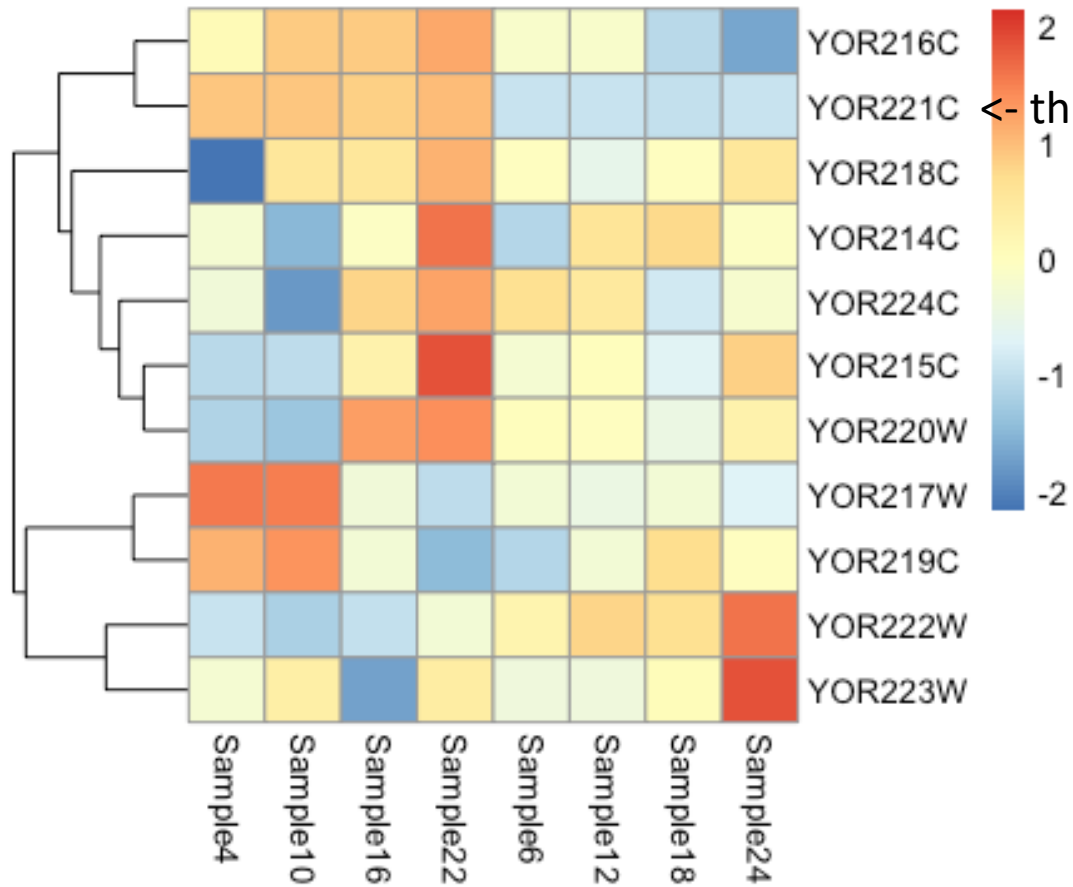
Why do we use Z-scores?



Plotting a heatmap in R

```
> library(tweeDEseq)
> norm <- normalizeCounts(count_table, method="TMM")
> top_data = norm[1000:1030,]

> library(pheatmap) #download if not installed
> pheatmap(as.matrix(top_data), scale = c("row"), cluster_cols=FALSE)
```

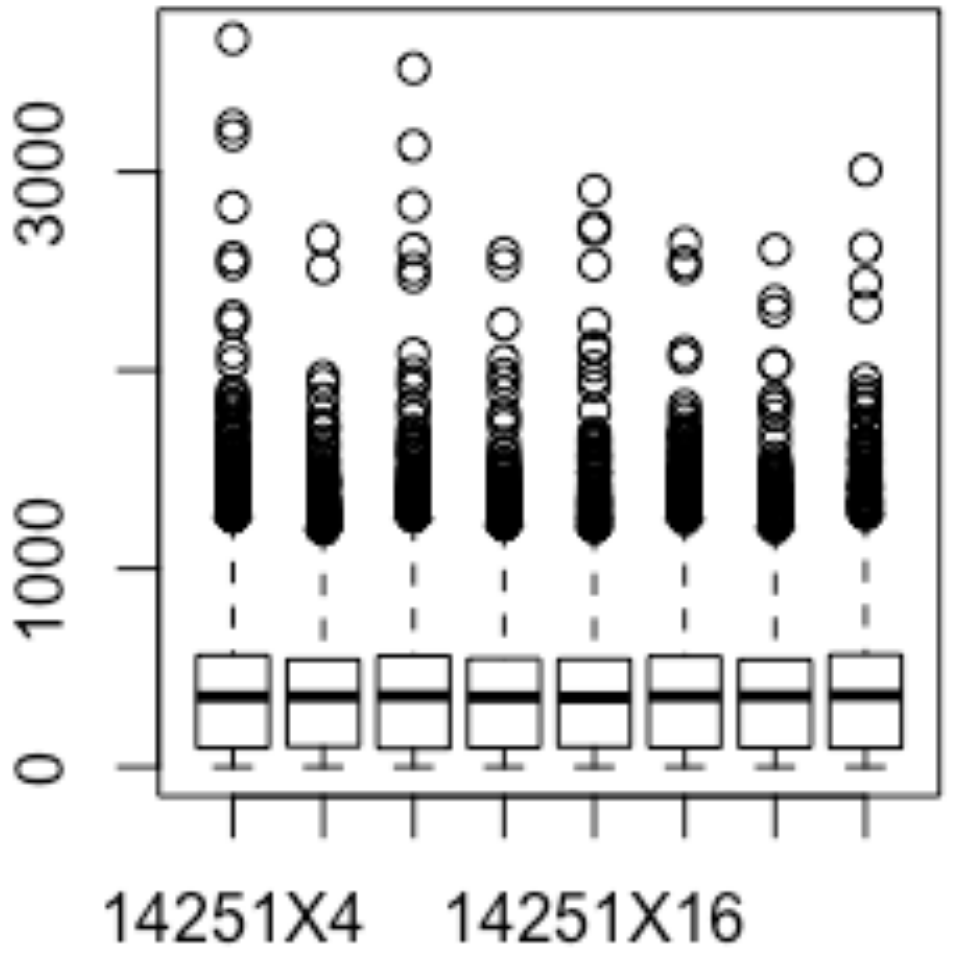


<- this is our deleted gene

Boxplots

- Hide data

> boxplot(norm)

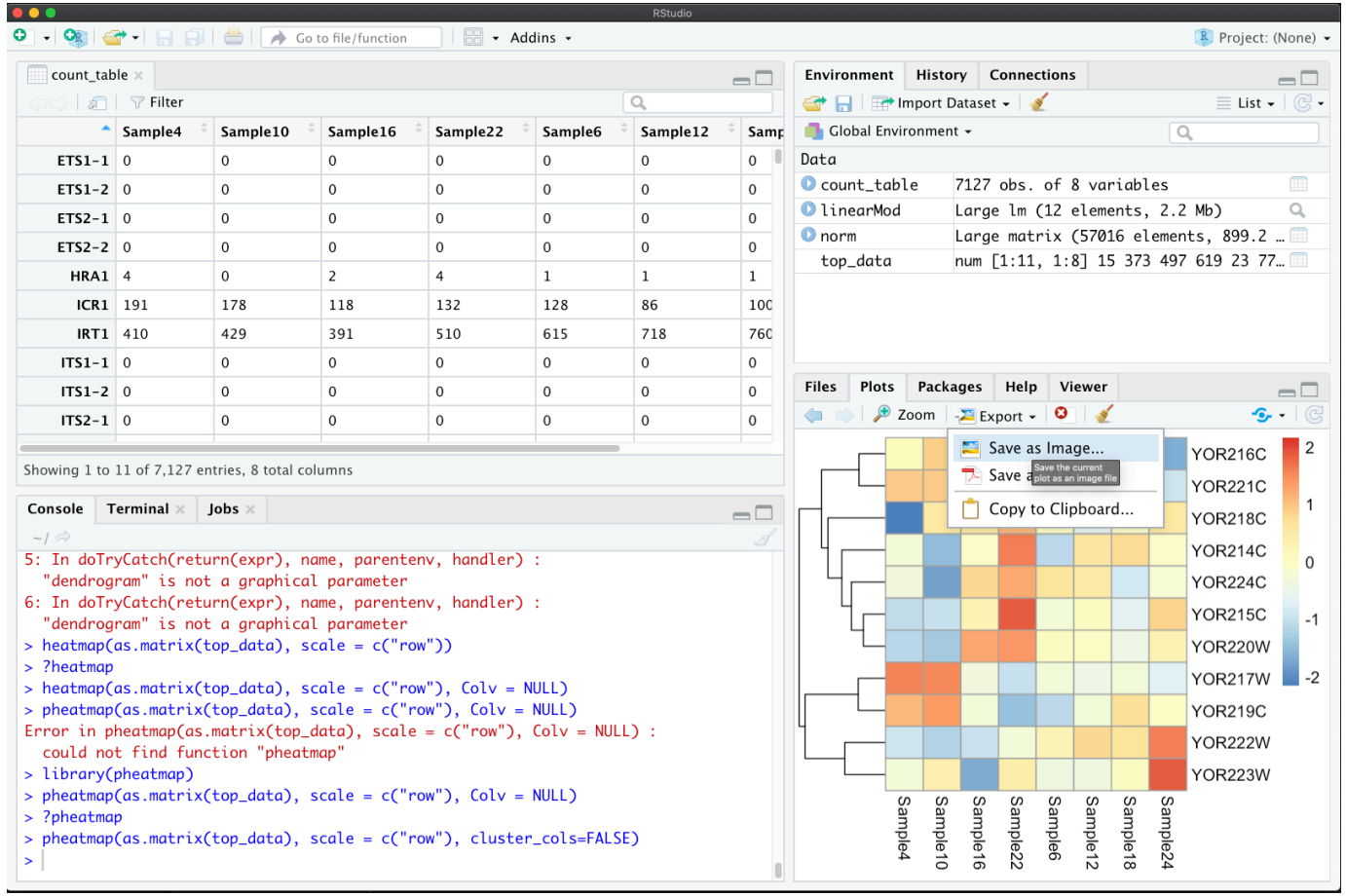


Beeswarm plot

- Similar in concept to a boxplot, but more transparent

```
> library(beeswarm)
> beeswarm(as.matrix(norm[1:1000,]))
```

Outputting a figure



Outputting a figure