# STAT115 Lab 1

*(your name)*

*2017-02-07*

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

We can have several different sizes based on the text structure:

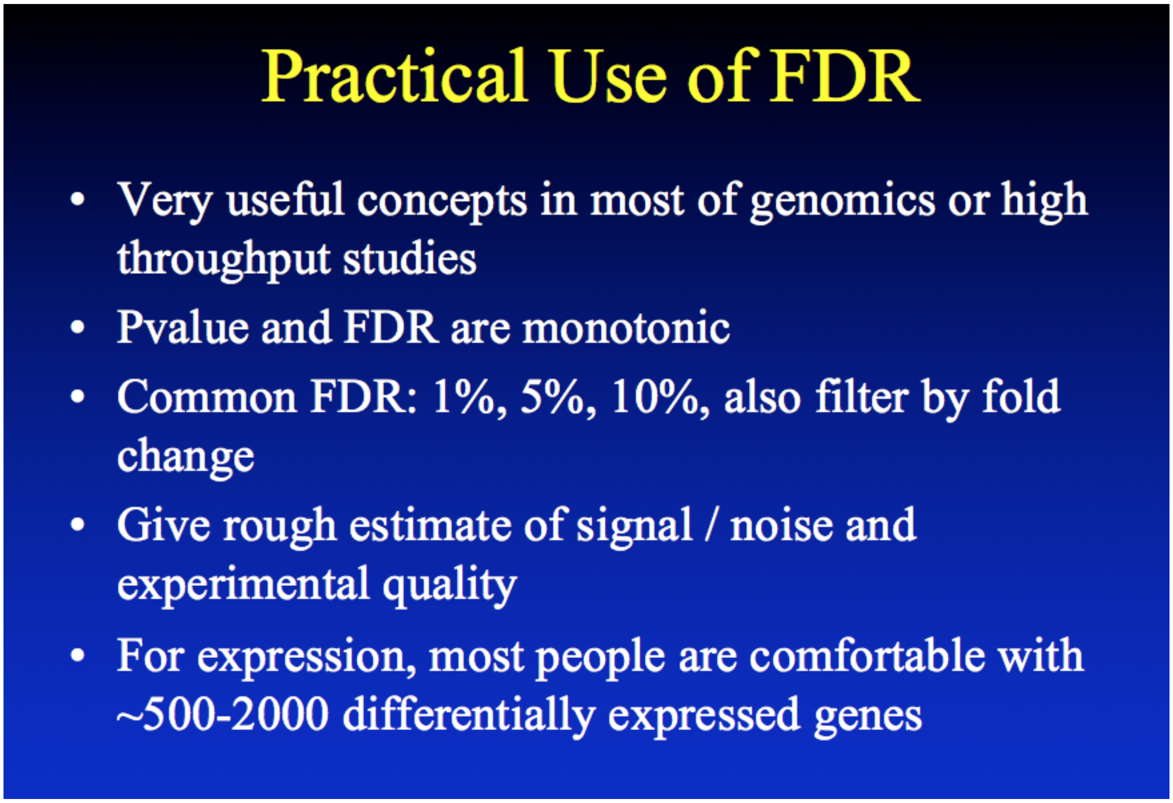
# Big

## Kinda big

### Still biggish

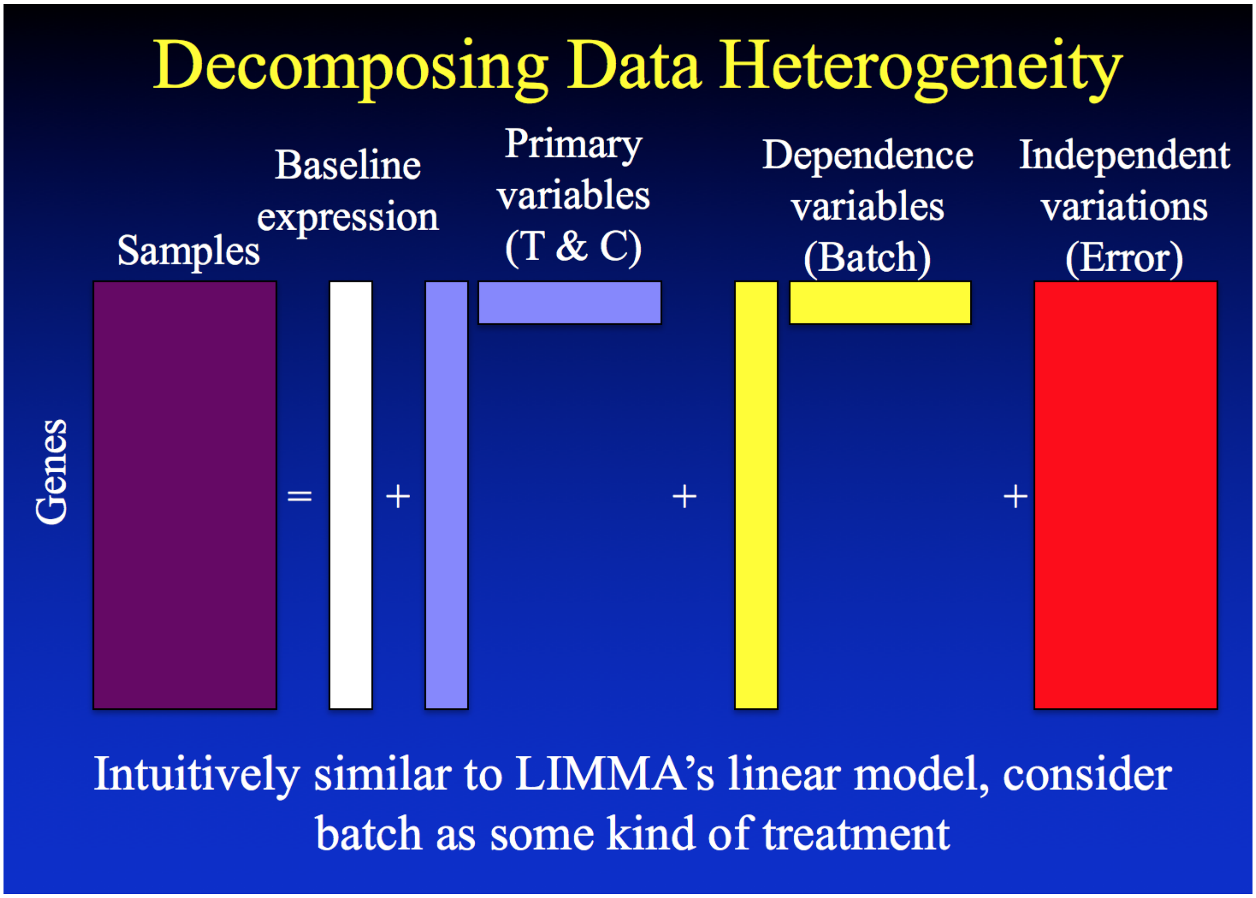
#### Pretty Normal

You can **bold** or *italicize* text very easily. Most significantly, probably, is the ease of embedding images:



Practical Use of FDR

- Very useful concepts in most of genomics or high throughput studies
- Pvalue and FDR are monotonic
- Common FDR: 1%, 5%, 10%, also filter by fold change
- Give rough estimate of signal / noise and experimental quality
- For expression, most people are comfortable with ~500-2000 differentially expressed genes



# Decomposing Data Heterogeneity

Samples

Baseline expression

Primary variables (T & C)

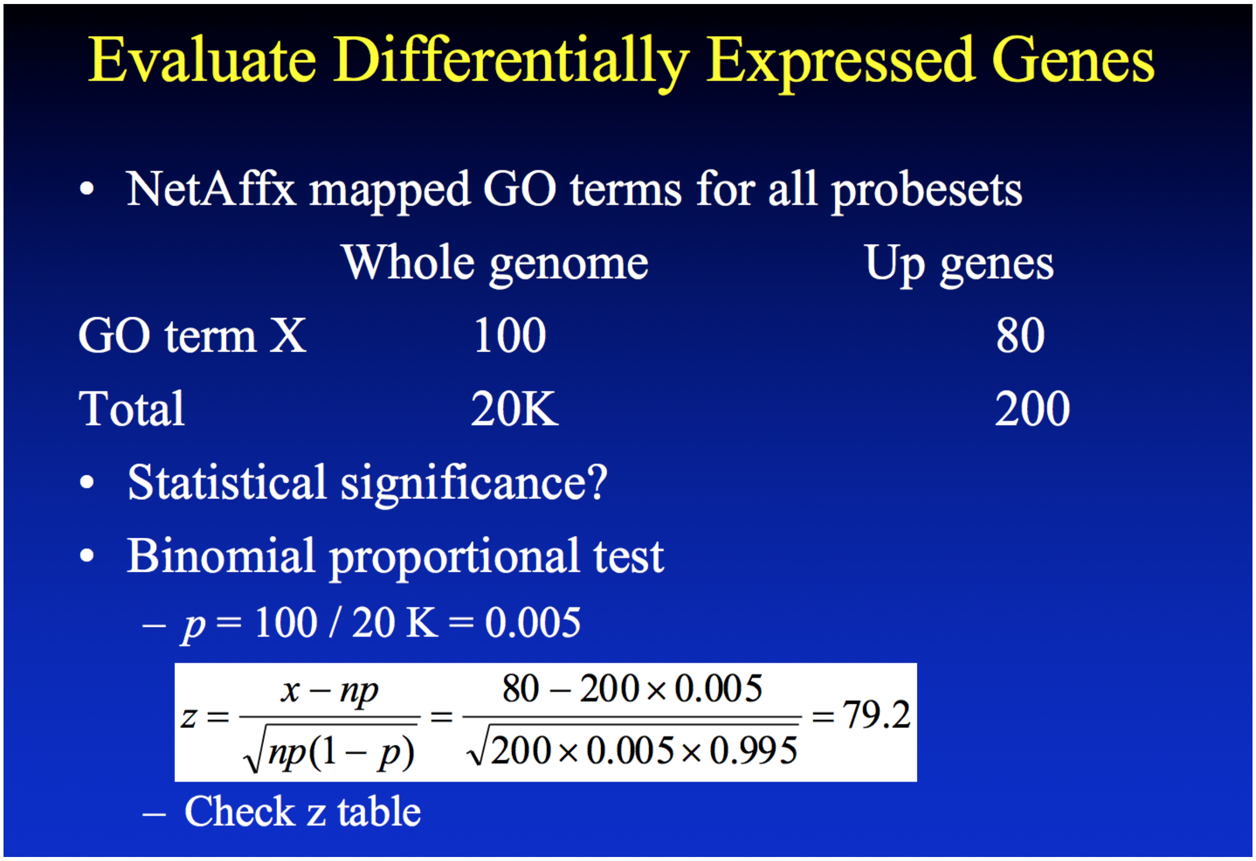
Dependence variables (Batch)

Independent variations (Error)

Genes

= + + +

Intuitively similar to LIMMA’s linear model, consider batch as some kind of treatment



# Evaluate Differentially Expressed Genes

- NetAffx mapped GO terms for all probesets

| | Whole genome | Up genes |
|---|---|---|
| GO term X | 100 | 80 |
| Total | 20K | 200 |

- Statistical significance?
- Binomial proportional test
  - $p = 100 \,/\, 20\text{ K} = 0.005$

$$z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{80 - 200 \times 0.005}{\sqrt{200 \times 0.005 \times 0.995}} = 79.2$$

  - Check z table

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("affy")
biocLite("hgu133plus2.db")
biocLite("drosophila2.db")

# etc.
```

```
library(knitr)
library(affy)
library(hgu133plus2.db)
library(drosophila2.db)
```

```
celFiles <- list.celfiles(path = "data", full.names=TRUE)
data.affy <- ReadAffy(filenames = celFiles)
data.affy
```

Show an MA Plot

```
MAplot(data.affy, pairs = TRUE, which=c(1,2,3,4), plot.method = "smoothScatter")
```

```
# Raw data visualization
# MA plot for only the first 4 samples
image(data.affy)
hist(data.affy)
boxplot(data.affy, col=c(1:4))

# Normalize data
data.rma <- rma(data.affy)
expr.rma <- exprs(data.rma)

#compare normalized data with raw data
library(affyPLM)
MAplot(data.rma, pairs = TRUE, which=c(1,2,3,4), plot.method = "smoothScatter")
boxplot(data.frame(expr.rma), col=c(1:4))

# Look at gene names
head(rownames(expr.rma), 20)

# Convert gene names
Annot <- data.frame(REFSEQ=sapply(contents(drosophila2REFSEQ), paste, collapse=", "),
                    SYMBOL=sapply(contents(drosophila2SYMBOL), paste, collapse=", "),
                    DESC = sapply(contents(drosophila2GENENAME), paste, collapse=", ")  )

# Merge data frames together (like a database table join)
all <- merge(Annot, expr.rma, by.x=0, by.y=0, all = TRUE)

#find significant up- and down-regulated genes
control <- c(1,2)
mutants <- c(3,4)

# Do a basic T test
foldchange <- apply(expr.rma, 1, function(x) mean( x[mutants] ) - mean( x[control] ) )
T.p.value <- apply(expr.rma, 1, function(x) t.test( x[mutants], x[control], var.equal = TRUE)$p.value )
```

```r
# fdr <- p.adjust(T.p.value, method="fdr")
genes.up <- all$SYMBOL[ which( T.p.value < 0.05 & foldchange > 0 ) ]
genes.down <- all$SYMBOL[ which( T.p.value < 0.05 & foldchange < 0 ) ]

# Use limma to call differentially expressed genes
library(limma)
design <- c(0, 0, 1, 1) # assign control and trt groups
design.mat <- model.matrix(~design)
fit <- lmFit(expr.rma, design.mat)
fit <- eBayes(fit)
topTable(fit, coef = "design")
```