Lab 12 | BINF 6310 | Jon Lee | Spring 2020

Code

Lab #12 (the last assignment)

Please e-mail code, graphs and answers to questions to afodor@uncc.edu (mailto:afodor@uncc.edu). Please put lab #12 in the subject line. This lab is a “normal” sized lab and will be graded equally to other labs.

Please have lab submitted (whatever you have) by May 5th.

This week’s dataset is again (same as last week): http://afodor.github.io/classes/stats2015/prePostPhylum.txt (http://afodor.github.io/classes/stats2015/prePostPhylum.txt)

(This datset is described, albeit from a different analysis pipeline, in these papers: http://www.sciencemag.org/content/sci/338/6103/120.full.html (http://www.sciencemag.org/content/sci/338/6103/120.full.html) and http://www.nature.com/ncomms/2014/140903/ncomms5724/full/ncomms5724.html (http://www.nature.com/ncomms/2014/140903/ncomms5724/full/ncomms5724.html))

Note that WT and IL10-/- animals are in different cages. So “Cage1_WT” is a different cage from “Cage1_10-/-“.

For the POST timepoints only:

1. For each phyla, graph the relative abundance of that phyla vs. cage. Does there appear to be a cage effect across different phyla?

Hide

```
#load in data
pFile <- paste("prePostPhylum.txt", sep = "")

pTable <-read.table(pFile,header = TRUE, sep = "\t")
numCols <- ncol(pTable)
pClasses <- c(rep("character", 4), rep("numeric", numCols-4))
pTable <- read.table(pFile, header = TRUE, sep = "\t", colClasses = pClasses)

pData<- pTable[,5:10]

#generate table of POST data only
postTable <- data.frame()

for(i in 1:length(pTable[,3]))
{
    if(pTable[i,3] == "POST")
    {
        postTable <- rbind(postTable, pTable[i,])
    }
}

postData <- postTable[,5:10]
```

Hide

```
#generate cage data table
cage1WTTable <- data.frame()
cage2WTTable <- data.frame()
cage3WTTable <- data.frame()
cage4WTTable <- data.frame()
cage5WTTable <- data.frame()
cage6WTTable <- data.frame()

cage1MutTable <- data.frame()
cage2MutTable <- data.frame()
cage3MutTable <- data.frame()
cage4MutTable <- data.frame()
cage5MutTable <- data.frame()

for(i in 1:length(postTable[,2]))
{
    if(postTable[i,2] == "Cage1_WT")
    {
        cage1WTTable <- rbind(cage1WTTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage2_WT")
    {
        cage2WTTable <- rbind(cage2WTTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage3_WT")
    {
        cage3WTTable <- rbind(cage3WTTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage4_WT")
    {
        cage4WTTable <- rbind(cage4WTTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage5_WT")
    {
        cage5WTTable <- rbind(cage5WTTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage6_WT")
    {
        cage6WTTable <- rbind(cage6WTTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage1_10-/-")
    {
        cage1MutTable <- rbind(cage1MutTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage2_10-/-")
    {
        cage2MutTable <- rbind(cage2MutTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage3_10-/-")
    {
        cage3MutTable <- rbind(cage3MutTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage4_10-/-")
```

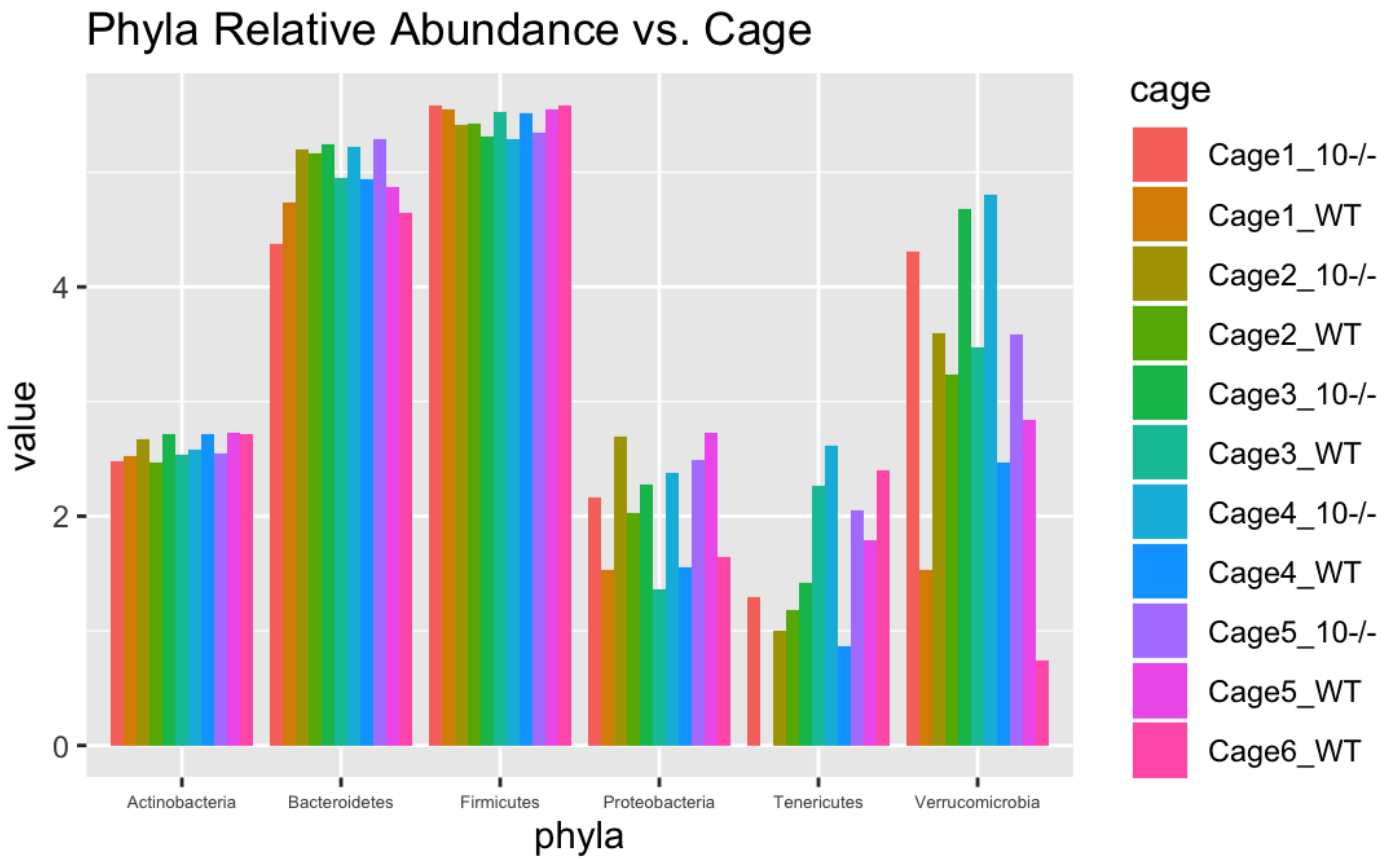
```
    {
        cage4MutTable <- rbind(cage4MutTable, postTable[i,])
    }
    if(postTable[i,2] == "Cage5_10-/-")
    {
        cage5MutTable <- rbind(cage5MutTable, postTable[i,])
    }
}
```

Hide

```
#Tenericutes Data
tenerData <- c(mean(cage1WTTable[,5]), mean(cage2WTTable[,5]), mean(cage3WTTable[,5]), m
ean(cage4WTTable[,5]), mean(cage5WTTable[,5]), mean(cage6WTTable[,5]), mean(cage1MutTabl
e[,5]), mean(cage2MutTable[,5]), mean(cage3MutTable[,5]), mean(cage4MutTable[,5]), mean
(cage5MutTable[,5]))

#Verrucomicrobia Data
verrData <- c(mean(cage1WTTable[,6]), mean(cage2WTTable[,6]), mean(cage3WTTable[,6]), me
an(cage4WTTable[,6]), mean(cage5WTTable[,6]), mean(cage6WTTable[,6]), mean(cage1MutTable
[,6]), mean(cage2MutTable[,6]), mean(cage3MutTable[,6]), mean(cage4MutTable[,6]), mean(c
age5MutTable[,6]))

#Bacteroidetes Data
bactData <- c(mean(cage1WTTable[,7]), mean(cage2WTTable[,7]), mean(cage3WTTable[,7]), me
an(cage4WTTable[,7]), mean(cage5WTTable[,7]), mean(cage6WTTable[,7]), mean(cage1MutTable
[,7]), mean(cage2MutTable[,7]), mean(cage3MutTable[,7]), mean(cage4MutTable[,7]), mean(c
age5MutTable[,7]))

#Actinobacteria Data
actiData <- c(mean(cage1WTTable[,8]), mean(cage2WTTable[,8]), mean(cage3WTTable[,8]), me
an(cage4WTTable[,8]), mean(cage5WTTable[,8]), mean(cage6WTTable[,8]), mean(cage1MutTable
[,8]), mean(cage2MutTable[,8]), mean(cage3MutTable[,8]), mean(cage4MutTable[,8]), mean(c
age5MutTable[,8]))

#Firmicutes Data
firmData <- c(mean(cage1WTTable[,9]), mean(cage2WTTable[,9]), mean(cage3WTTable[,9]), me
an(cage4WTTable[,9]), mean(cage5WTTable[,9]), mean(cage6WTTable[,9]), mean(cage1MutTable
[,9]), mean(cage2MutTable[,9]), mean(cage3MutTable[,9]), mean(cage4MutTable[,9]), mean(c
age5MutTable[,9]))

#Proteobacteria Data
protData <- c(mean(cage1WTTable[,10]), mean(cage2WTTable[,10]), mean(cage3WTTable[,10]),
mean(cage4WTTable[,10]), mean(cage5WTTable[,10]), mean(cage6WTTable[,10]), mean(cage1Mut
Table[,10]), mean(cage2MutTable[,10]), mean(cage3MutTable[,10]), mean(cage4MutTable[,10
]), mean(cage5MutTable[,10]))
```

Hide

```
#Graph Data
phyla <- c(rep(colnames(postTable)[5], 11), rep(colnames(postTable)[6], 11), rep(colname
s(postTable)[7], 11), rep(colnames(postTable)[8], 11), rep(colnames(postTable)[9], 11),
 rep(colnames(postTable)[10], 11))
cage <- c(rep(c("Cage1_WT", "Cage2_WT", "Cage3_WT", "Cage4_WT", "Cage5_WT", "Cage6_WT",
"Cage1_10-/-", "Cage2_10-/-", "Cage3_10-/-", "Cage4_10-/-", "Cage5_10-/-"), 6))
value <- c(tenerData, verrData, bactData, actiData, firmData, protData)
graphData <- data.frame(phyla, cage, value)

library(ggplot2)
require(ggplot2)
ggplot(graphData, aes(fill = cage, y = value, x = phyla)) + geom_bar(position = "dodge",
stat = "identity") + ggtitle("Phyla Relative Abundance vs. Cage") + theme(axis.text.x =
 element_text(size = 5))
```



Answer:

Looking at the relative abundance only, there does not seem to be a cage effect on the first three phyla (Actinobacteria, Bacteroidetes, and Firmicutes), but there does seem to be an effect on the last three phyla (Proteobacteria, Tenericutes, and VErrucomicrobia).

2. For each phyla build a mixed linear model with genotype as the fixed variable and cage as a random variable. Report the intraclass correlation coefficient for each phyla. Are there any phyla that are significantly different for genotype in the mixed model at a 10% false discovery rate? (Note: FDR corrected p-values).

Hide

```
#generate intermidiate table for cage and genotype info & p-value vector
intTable <- cbind(postTable[,2], postTable[,4])
pvals <- vector()

#Tenericutes Model
tenerTable <- data.frame(cbind(intTable, postTable[,5]))
colnames(tenerTable) <- c("Cage", "Genotype", "Tenericutes")
tenerVals <- as.numeric(tenerTable[,3])
tenerCage <- tenerTable[,1]
tenerGen <- tenerTable [,2]
tenerMod <- glm(tenerVals ~ tenerCage + tenerGen, data = tenerTable)
tenerCF <- data.frame(coef(summary(tenerMod)))
pvals <- c(pvals, tenerCF$Pr...t..[1])
tenerCF$Estimate[1]
```

```
[1] 9.666667
```

Hide

```
#Verrucomicrobia Model
verrTable <- data.frame(cbind(intTable, postTable[,6]))
colnames(verrTable) <- c("Cage", "Genotype", "Verrucomicrobia")
verrVals <- as.numeric(verrTable[,3])
verrCage <- verrTable[,1]
verrGen <- verrTable[,2]
verrMod <- glm(verrVals ~ verrCage + verrGen, data = verrTable)
verrCF <- data.frame(coef(summary(verrMod)))
pvals <- c(pvals, verrCF$Pr...t..[1])
verrCF$Estimate[1]
```

```
[1] 31.66667
```

Hide

```
#Bacteroidetes Model
bactTable <- data.frame(cbind(intTable, postTable[,7]))
colnames(bactTable) <- c("Cage", "Genotype", "Bacteroidates")
bactVals <- as.numeric(bactTable[,3])
bactCage <- bactTable[,1]
bactGen <-bactTable[,2]
bactMod <- glm(bactVals ~ bactCage + bactGen, data = bactTable)
bactCF <- data.frame(coef(summary(bactMod)))
pvals <- c(pvals, bactCF$Pr...t..[1])
bactCF$Estimate[1]
```

```
[1] 3.666667
```

Hide

```
#Actinobacteria Model
actiTable <- data.frame(cbind(intTable, postTable[,8]))
colnames(actiTable) <- c("Cage", "Genotype", "Actinobacteria")
actiVals <- as.numeric(actiTable[,3])
actiCage <- actiTable[,1]
actiGen <- actiTable[,2]
actiMod <- glm(actiVals ~ actiCage + actiGen, data = actiTable)
actiCF <- data.frame(coef(summary(actiMod)))
pvals <- c(pvals, actiCF$Pr...t..[1])
actiCF$Estimate[1]
```

```
[1] 15
```

Hide

```
#Firmicutes Model
firmTable <- data.frame(cbind(intTable, postTable[,9]))
colnames(firmTable) <- c("Cage", "Genotype", "Firmicutes")
firmVals <- as.numeric(firmTable[,3])
firmCage <- firmTable[,1]
firmGen <- firmTable[,2]
firmMod <- glm(firmVals ~ firmCage + firmGen, data = firmTable)
firmCF <- data.frame(coef(summary(firmMod)))
pvals <- c(pvals, firmCF$Pr...t..[1])
firmCF$Estimate[1]
```

```
[1] 34.33333
```

Hide

```
#Proteobacteria Model
protTable <- data.frame(cbind(intTable, postTable[,10]))
colnames(protTable) <- c("Cage", "Genotype", "Proteobacteria")
protVals <- as.numeric(protTable[,3])
protCage <- protTable[,1]
protGen <- protTable[,2]
protMod <- glm(protVals ~ protCage + protGen, data = protTable)
protCF <- data.frame(coef(summary(protMod)))
pvals <- c(pvals, protCF$Pr...t..[1])
protCF$Estimate[1]
```

```
[1] 22.66667
```

Hide

```
#adjust p-values
pvalsFDR <- p.adjust(pvals, method = "BH")
pvalsFDR
```

```
[1] 4.316231e-02 4.133871e-10 4.138698e-01 4.138698e-01 4.316231e-02
[6] 9.568548e-09 3.439647e-05
```

Asnwer:

The p-values suggest that all 6 phyla are significantly different between cage and genotype.

Hints:

1. If you use par(mfrow=c(3,2)) you can fit all 6 plots for phyla vs. cage on one graph. You can put the p-values and intraclass correlation coefficient in the “main” text above each graph to make a nice summary figure.
2. It can be useful to make a dataframe with just the data you want before building your model. So if you are looping through columns in a “myT” that you’ve read with read.table and i is your column index..

   myT <- myT[myT$time == “POST”,] bug <- myT[,i] cage <- myT$cage genotype <- myT$genotype myFrame <- data.frame(bug, cage, genotype)

(and then build your models with data=myFrame…)

3. Getting a p-value out of the mixed linear model could be done with something like:

   unclass(summary(M.mixed))$tTable[2,5]

Getting the rho(intraclass correlation coefficient) out of a GLS model can be done with:

```
coef(M.gls$modelStruct[1]$corStruct,unconstrained=FALSE)[[1]]
```

4. You can have both points and boxplots on a scatter graph with something like:

   boxplot(myFrame$bug ~ myFrame$cage) stripchart(bug ~ cage, data = myFrame,vertical = TRUE, pch = 21, add=TRUE