Exercise in quantitative metagenomics

In this exercise, we conceptually continue from the previous exercise, where you *de novo* assembled contigs from metagenomics data and counted reads that mapped to these. In this exercise we have done the assembly and counting across a cohort of hundreds of human fecal samples in advance and in addition provide the gene-wise taxonomy and the BMI of the human donors.

From this data we shall estimate the species richness, diversity and look at the effect of downsizing. Furthermore we shall see if we can identify any differences between the microbiome of lean and obese.

First let's install the "vegan" pacakge and thereafter load the read count data from a series of stool samples.

```
install.packages("vegan")
library("vegan")
load(url("http://teaching.healthtech.dtu.dk/material/22126/Counts_NGS.RData"))
head(Counts)
str(Counts)
```

Q1. How many samples do we have and how many genes?

The different samples may have been sequenced to different depths. Try to count the reads per sample

```
sampleDepth<-(colSums(Counts))
hist(sampleDepth, breaks=100, ylab="Number of samples", xlab="Number of reads",
main="Sample depth")
range(sampleDepth)
dev.print(file="sample.depth.pdf", device=pdf)
```

Q2. Whats the sample depth range?

Species

Lets get the genes associated to species. Here is the gene-wise species taxonomy

```
load(url("http://teaching.healthtech.dtu.dk/material/22126/taxonomy_species.RData"))
head(taxonomy_species)
```

We then combine (by summing) the read counts pr. gene to read counts per species.

```
taxCounts<-apply(Counts, 2, tapply, INDEX=taxonomy_species, sum)
# wait ~1 min
```

Try looking at the taxCounts matrix:

```
str(taxCounts)
head(taxCounts)
```

Q3. How many species are there in total?

Richness and Diversity

What is the species richness and diversity (Shannon) for the different samples.

Q4. What does a high Shannon diversity index mean?

OK, lets see it for our samples

```
species_richness<-(colSums(taxCounts>0))
names(species_richness)<-NULL
require(vegan)
speciesDiversity<-diversity(t(taxCounts), index = "shannon")
names(speciesDiversity)<-NULL

par(mfrow=c(2,2), pch=20)
barplot(sort(species_richness), las=3, main="Species richness", xlab="Samples",
ylab="Richness")
barplot(sort(speciesDiversity), xlab="Samples", las=3, main="Diversity (Shannon)")
plot(species_richness,speciesDiversity,xlab="Richness",
ylab="Shannon diversity index")
dev.print(file="richness_diversity_full.pdf", device=pdf)
par(mfrow=c(1,1))
```

Downsizing

But this was on the raw count data with different sampling depth (number of counts) per sample. We should downsize so that we get fair comparisons.

First suggest the number of reads we should sample per sample for the downsizing [target]. If we chose a low target we will loose abundance resolution and detection sensitivity. If we chose it higher we will loose samples.

```
plot(sampleDepth, pch=20, log="y", xlab="Samples", ylab="Number of reads")
```

NB. there is no right answer (but there are less good suggestions). Insert the number you want to downsize to below and plot it again - the samples above the horizontal line we will keep and the samples below the line we will throw out.

```
downsizeTarget <- INSERT_NUMBER_HERE

plot(sampleDepth, pch=20, log="y", xlab="Samples", ylab="Number of reads");
abline(h=downsizeTarget)
dev.print(file="sample_depth_downsize_target.pdf", device=pdf)
```

Q5. Which threshold did you chose and why? How many samples did you loose?

OK lets downsize

```
dz_Counts<-round(t(t(Counts)*downsizeTarget/sampleDepth))
weak_samples<-sampleDepth<downsizeTarget
```

```
dz_Counts[,weak_samples]<-NA # samples that did not make the cut
```

This is a quick and dirty downsizing (ideally one resampled the reads to a given depth, but that will take days)
We re-analyze on the downsized data

```
dz_taxCounts<-apply(dz_Counts, 2, tapply, INDEX=taxonomy_species, sum);
gc() # drop some memory
```

And the richness and diversity again, now on downsized data

```
dz_species_richness<-(colSums(dz_taxCounts>0))
names(dz_species_richness)<-NULL
require(vegan)
dz_speciesDiversity<-diversity(t(dz_taxCounts), index = "shannon")
dz_speciesDiversity[weak_samples]<-NA
names(dz_speciesDiversity)<-NULL
```

Now plot the richness and diversity with downsized data

```
par(mfrow=c(2,2), pch=20)
barplot(sort(dz_species_richness), las=3, main="Species richness (Downsized)",
xlab="Species", ylab="Richness")
barplot(sort(dz_speciesDiversity), las=3,main="Shannon's diversity index (downsized)",
xlab="Species", ylab="Shannon diversity")
```

And compare to the raw data

```
plot(dz_species_richness,species_richness, xlab="downsized richness", ylab="raw
richness", main="Richness")
plot(dz_speciesDiversity,speciesDiversity,xlab="downsized species diversity", ylab="raw
species diversity",main="Diversity (Shannon)")
dev.print(file="richness_diversity_down.pdf", device=pdf)
par(mfrow=c(1,1), pch=1)
```

Q6. What is the effect on the downsizing on richness
Q7. What is the effect on the downsizing on diversity (shannon)

Lets plot the abundance of each species in a sample with low diversity and a sample with high diversity. You should be able to see a clear difference between the two samples!

```
par(mfrow=c(1,2))
barplot(taxCounts[,4], main="Person 4, SD > 3", xaxt="n", xlab="Species", ylab="Normalized
abundance")
barplot(taxCounts[,240], main="Person 240, SD < 0.5", xaxt="n", xlab="Species",
ylab="Normalized abundance")
dev.print(file="diversity_differences.pdf", device=pdf)
par(mfrow=c(1,1))
```

Comparisons

Now lets see if there is a difference between the microbiome of lean and obese humans. But first load some sample more information: BMI and Class

```
load(url("http://teaching.healthtech.dtu.dk/material/22126/BMI.RData"))
boxplot(BMI$BMI.kg.m2 ~ BMI$Class, col=c("red", "gray","blue"), ylab="BMI")
dev.print(file="bmi_differences.pdf", device=pdf)
```

First let us see if the abundance of *E. coli* differs between obese and lean individuals using a Wilcoxon rank sum test (look for the p-value in the output):

```
wilcox.test(x=dz_taxCounts["Escherichia coli",BMI$Classification=="ob"],
y=dz_taxCounts["Escherichia coli",BMI$Classification=="le"] )
```

Also lets get the mean abundance of E. coli in the tree groups

```
tapply(dz_taxCounts["Escherichia coli",], BMI$Classification, mean, na.rm=TRUE)
```

Q8. Is there any significant difference in abundance of *E. coli* between the different BMI groups?

Let's test all species:

```
pval<-apply(dz_taxCounts, 1,
function(V){wilcox.test(x=V[BMI$Classification=="ob"],y=V[BMI$Classification=="le"])$p.val
ue})

Abundance_ratio<-log2(apply(dz_taxCounts,
1,function(V){mean(x=V[BMI$Classification=="ob"],
na.rm=TRUE)/mean(V[BMI$Classification=="le"], na.rm=TRUE)}))
```

Also lets correct for multiple testing using Benjamini-Hochberg (False Discovery Rate) (we are testing 120 species) and plot them:

```
pval.adjust = p.adjust(pval, method="BH")
plot(sort(pval.adjust), log="y", pch=16, xlab="Species", ylab="p-values")
abline(h=0.05, col="grey", lty=2)
dev.print(file="BMI_pvals.pdf", device=pdf)
```

Q9. How many species are significant with an FDR < 0.05?

Here one could write to file and open in Microsoft Excel:

```
o<-order(pval)
BMIstat<-data.frame(pval,pval.adjust, Abundance_ratio)[o,]
write.csv(BMIstat, file="BMI_microbiome_stat.csv")
```

or look at the top 10 in R

```
BMIstat[1:10,]
par(mar=c(5,18,5,5))
barplot(BMIstat[1:10,3], names.arg=rownames(BMIstat)[1:10], las=1,xlab="log fold
difference between lean and obese", horiz=TRUE)
```

```
dev.print(file="obese_lean_diff.pdf", device=pdf)
```

Q10. Can you see any differences in the abundances - which species have large differences, what are their p-values?
Q11. What type of bacteria is the most significant one? [try google]

Beta-diversity

Plot the Bray-curtis distance between samples

```
library(RColorBrewer)
library(gplots)

vdist = as.matrix(vegdist(t(taxCounts)))
rownames(vdist) = colnames(vdist)
hmcol = colorRampPalette(brewer.pal(9, "GnBu"))(100)

heatmap.2(vdist, trace='none', col=rev(hmcol))
dev.print(file="bray-curtis.pdf", device=pdf)
```