**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://www.cbs.dtu.dk/courses/27626/misc/qmetagenomics/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://www.cbs.dtu.dk/courses/27626/misc/qmetagenomics/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

**Q3. Try to explain what we are doing when we are summing the read counts per gene to read counts per species?**

Try looking at the taxCounts matrix:

str(taxCounts)

head(taxCounts)

**Q4. How many species are there in total?**

**Richness and Diversity**

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

**Q5. 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 /Rarefy**

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 lose abundance resolution and detection sensitivity. If we chose it higher we will lose 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)

**Q6. Which threshold did you chose and why? How many samples did you lose?**

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)

Count the species again, now 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="Samples", ylab="Richness")

barplot(sort(dz\_speciesDiversity), las=3,main="Shannon's diversity index (downsized)", xlab="Samples", 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)

**Q7A. What is the effect on the downsizing on richness**

**Q7B. What is the effect on the downsizing on diversity (shannon)**

Let’s 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="Individual 4, SD > 3", xaxt="n", xlab="Species", ylab="Normalized abundance")

barplot(taxCounts[,240], main="Individual 240, SD < 0.5", xaxt="n", xlab="Species", ylab="Normalized abundance")

dev.print(file="diversity\_differences.pdf", device=pdf)

par(mfrow=c(1,1))

**Beta-diversity and PCA**

Lets try to plot the Bray-curtis disimilarity between samples

library(RColorBrewer)

library(gplots)

Calculate the Bray-curtis dissimilarity and plot it as a heatmap

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)

**Q8. Can you see some clusters of samples?**

Finally for the PCA

dz\_taxCounts\_noNA <- dz\_taxCounts[,!is.na(colSums(dz\_taxCounts))]

my.rda <- rda(t(dz\_taxCounts\_noNA))

biplot(my.rda, display = c("sites", "species"), type = c("text", "points"))

dev.print(file="pca.pdf", device=pdf)

**Q9. Can you see which species that seems to be driving the differences between the samples?**

**Statistical comparisons**

Now let’s 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://www.cbs.dtu.dk/courses/27626/misc/qmetagenomics/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)

**Q10A. Is there any significant difference in abundance of *E. coli* between the different BMI groups?**

**Q10B. Try to explain what a “wilcox.test” is. Use the R help, but also try google.**

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.value})

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)

**Q11. 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)

**Q12. Can you see any differences in the abundances - which species have large differences, what are their p-values?**

**Q13. What type of bacteria is the most significant one? [try google]**

**Statistically modelling the variance using DESeq2**

Statistically modelling the variance instead of downsizing using DESeq2. This program has been use for analysis of transcriptomics data and also widely used for

To install this package, start R (version "3.6") and enter:

if (!requireNamespace("BiocManager", quietly = TRUE))

 install.packages("BiocManager")

BiocManager::install("DESeq2")

library(DESeq2)

Lets use the non-downsized data

cts <- taxCounts

coldata = BMI[,1]

coldata = matrix(NA, nrow=nrow(BMI), ncol=1)

coldata[,1] = as.vector(BMI[,1])

rownames(coldata) = rownames(BMI)

colnames(coldata) = "BMI"

Take a look at coldata, it contains information for each individuals

coldata

Make sure that all individuals are in our coldata (information) and also in the data

all(rownames(coldata) == colnames(cts))

Load data into DESeq format and take a look at the data (dds)

dds <- DESeqDataSetFromMatrix(countData = cts, colData = coldata, design = ~ BMI)

dds

Perform the statistical analysis

dds <- DESeq(dds)

dds

Get the results

res <- results(dds)

res

Order the results according to adjusted p-value

resOrdered <- res[order(res$pvalue),]

head(res)

**Q14**. which are the most significant species? Is there an overlap between these and using downsizing+wilcoxon test (what you did above)?

You can try to plot the different species and their estimated abundances in the three categories

**Q15**. Lets try to plot Ruminococcus gnauvus,

plotCounts(dds, gene="Ruminococcus gnavus", intgroup="BMI")

**Compositional analysis using ALDEx2**

Lets try to do the analysis using a compositional approach. Here we will use ALDEx2 to do this – you can find a vignette (eg manual) [here](https://bioconductor.org/packages/release/bioc/vignettes/ALDEx2/inst/doc/ALDEx2_vignette.pdf):

To install this package, start R (version "3.6") and enter:

if (!requireNamespace("BiocManager", quietly = TRUE))

 install.packages("BiocManager")

BiocManager::install("ALDEx2")

library(ALDEx2)

Let us try to do a t-test (Welch’s t and Wilcocon rank test) between the lean and the obese individuals in the dataset. First we need to make a matrix with only the lean and the obese individuals:

taxCounts\_lean = taxCounts[, coldata[,1] == "le"]

taxCounts\_obese = taxCounts[, coldata[,1] == "ob"]

taxCounts\_comb = cbind(taxCounts\_lean, taxCounts\_obese)

Then lets do the statistical analysis. First we set the conditions (conds) and then we run the aldex analysis:

conds <- c(rep("le", ncol(taxCounts\_lean)), rep("ob", ncol(taxCounts\_obese)))

x <- aldex(taxCounts\_comb, conds, mc.samples=128, test="t", effect=TRUE, include.sample.summary=FALSE, denom="iqlr", verbose=FALSE)

Since we performed two tests we can try to see if there are two species that are found significant in both examples (after Benjamini-Hochberg correction):

found.by.all <- which(x$we.eBH < 0.05 & x$wi.eBH < 0.05)

found.by.one <- which(x$we.eBH < 0.05 | x$wi.eBH < 0.05)

plot(x$diff.win, x$diff.btw, pch=19, cex=0.3, col=rgb(0,0,0,0.3), xlab="Difference within", ylab="Difference between")

points(x$diff.win[found.by.one], x$diff.btw[found.by.one], pch=19, cex=0.5, col=rgb(0,0,1,0.5))

points(x$diff.win[found.by.all], x$diff.btw[found.by.all], pch=19, cex=0.5, col=rgb(1,0,0,1))

abline(0,1,lty=2)

abline(0,-1,lty=2)

dev.print(file="lean\_obese.diff.aldex.pdf", device=pdf)

Species identified by both tests are shown in red. Species identified by only one test are shown in blue dots. Non-significant features represent rare features if black and abundant features if grey dots.

**Q16. Do you find any significant species, and if so which did you find (hint see below)?**

x[which(x$we.eBH < 0.05 & x$wi.eBH < 0.05),]

**Q17. Did you find the same species to be overrepresented in all analyses?**

Optional: Using the manual for ALDEx2 try to see if you can do an ANOVA including also the overweight invdividuals.