Microbial community data analysis in R

Antti Karkman

Wednesday 18.5. 2016

After quality trimming, chimera removal, OTU picking and taxonomic assignment of OTUs we have transformed the data to a biom -file that contains the OTU table and taxonomic assignment for each OTU.

In this tutorial we will use the Phyloseq package for further analysis. We will only show some features, more information and many good tutorials can be found from the website. You should have all the files on your own laptop. If not, then get them from taito and then we can continue by reading it all to R with the sample metadata (tab delimited file with sample names, sampling year and sampling month).

So let's start with R

Importing data and preprocessing

Open R or Rstudio and follow the instructions. **BUT**, don't just copy and paste, you need to change the object names and paths.

If you haven't downloaded `phyloseq`, `ggplot2`or `vegan` yet, do it now. This part you can copy and paste.

```r
#phyloseq
source('http://bioconductor.org/biocLite.R')
biocLite('phyloseq')
#ggplot2
install.packages("ggplot2")
#vegan
install.packages("vegan")
```

When you have all the packages installed, load them in R.

```r
library(phyloseq)
library(ggplot2)
library(vegan)
```

```
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.2-1
```

Now we have phyloseq hopefully working and we can move on reading in the data. Take care that the paths to the files and all filenames are correct. The biom fle can be read in to R with the function `import_biom()`.

```r
my_data <- import_biom("~/Desktop/microbial_community_course/coursedata.biom")
```

Modify the taxonomy a bit and check the naming of the different taxonomic levels in your taxonomy table

If they don't make any sense, change them to something that makes sense, hopefully.

```r
colnames(tax_table(my_data)) <- c("Domain", "Phylum", "Class", "Subclass", "Order",
                                  "Suborder", "Family", "Genus", "Species")
colnames(tax_table(my_data))
```

```
## [1] "Domain"   "Phylum"   "Class"    "Subclass" "Order"    "Suborder"
## [7] "Family"   "Genus"    "Species"
```

And the last thing is to read in the metadata.

```
metadata <- read.table("~/Desktop/microbial_community_course/metadata.txt", sep="\t",
                       header=TRUE, row.names=1)
sample_data(my_data) <- sample_data(metadata)
```

The year is written in number, but we should have it as a factor.

```
sample_data(my_data)$year <- factor(sample_data(my_data)$year)
```

Now we have a phyloseq object with all the data we need,we can have a look at it and how the different parts can be extracted from the object. Just by giving the name of the object you see the different objects wrapped inside the phyloseq object. It also tells you the the number of samples, taxa and metadata variables. Also you can see the functions to call the different objects. For example the first rows of your taxonomy table.

```
my_data
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10513 taxa and 10 samples ]
## sample_data() Sample Data:       [ 10 samples by 2 sample variables ]
## tax_table()   Taxonomy Table:    [ 10513 taxa by 9 taxonomic ranks ]
```

```
head(tax_table(my_data))
```

```
## Taxonomy Table:     [6 taxa by 9 taxonomic ranks]:
##           Domain     Phylum           Class            Subclass
## Otu00001 "Bacteria" "Actinobacteria" "Actinobacteria" "Acidimicrobidae"
## Otu00002 "Bacteria" "unclassified"   "unclassified"   "unclassified"
## Otu00003 "Bacteria" "Actinobacteria" "Actinobacteria" "Actinobacteridae"
## Otu00004 "Bacteria" "Actinobacteria" "Actinobacteria" "Acidimicrobidae"
## Otu00005 "Bacteria" "Actinobacteria" "Actinobacteria" "Actinobacteridae"
## Otu00006 "Bacteria" "Actinobacteria" "Actinobacteria" "Actinobacteridae"
##           Order              Suborder           Family
## Otu00001 "Acidimicrobiales" "Acidimicrobineae" "Candidatus_Microthrix"
## Otu00002 "unclassified"     "unclassified"     "unclassified"
## Otu00003 "Actinomycetales"  "Micrococcineae"   "Intrasporangiaceae"
## Otu00004 "Acidimicrobiales" "Acidimicrobineae" "Candidatus_Microthrix"
## Otu00005 "Actinomycetales"  "Micrococcineae"   "Intrasporangiaceae"
## Otu00006 "Actinomycetales"  "Micrococcineae"   "Intrasporangiaceae"
##           Genus          Species
## Otu00001 "unclassified" "unclassified"
## Otu00002 "unclassified" "unclassified"
## Otu00003 "Tetrasphaera" "unclassified"
## Otu00004 "unclassified" "unclassified"
## Otu00005 "Tetrasphaera" "unclassified"
## Otu00006 "Tetrasphaera" "unclassified"
```

```r
sample_data(my_data)
```

```
## Sample Data:        [10 samples by 2 sample variables]:
##        year month
## ER1089 2001     a
## ER1094 2001     a
## ER1100 2002     a
## ER1168 2002     a
## ER538  2002     a
## ER625  2003     b
## ER626  2003     b
## ER716  2003     b
## ER745  2003     b
## ER746  2005     b
```

We already removed some unwanted sequences with `mothur` and that can also be done with `phyloseq`. There might be still some *Chloroplast* sequences in the data, so we should remove them.

```r
#check if you have any
subset_taxa(my_data, Class=="Chloroplast")
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 2 taxa and 10 samples ]
## sample_data() Sample Data:       [ 10 samples by 2 sample variables ]
## tax_table()   Taxonomy Table:    [ 2 taxa by 9 taxonomic ranks ]
```

```r
#and then remove them
my_data <- subset_taxa(my_data, Class!="Chloroplast")
```

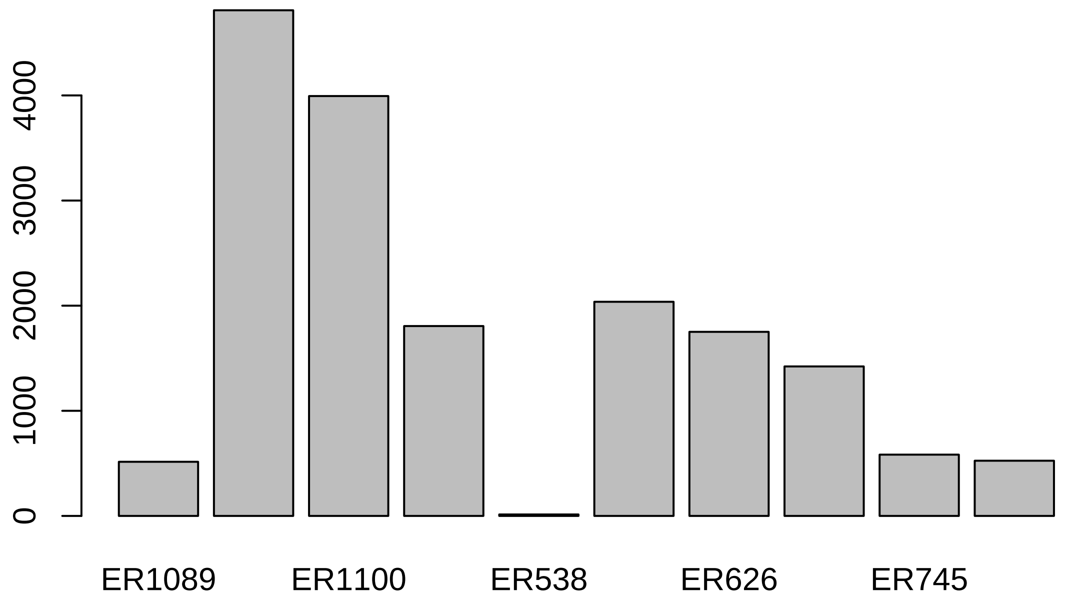
Library sizes and normalisation

Not all samples have the same amount of sequences or in another words have different library sizes and that might influence our results. There are several ways to normalise the data and we will use the most simple one, *proportions*. Calculate the library sizes of your samples with `sample_sums()`function. Use `barplot()` to visualize the library size differences.

```r
lib_sizes <- sample_sums(my_data)
lib_sizes
```

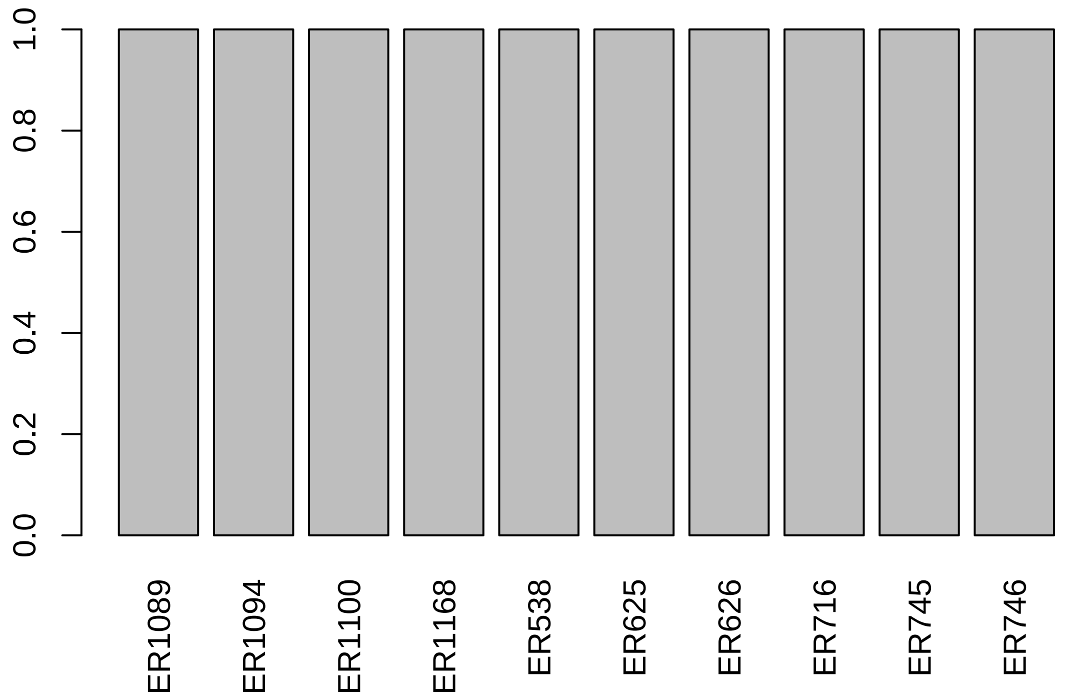
```
## ER1089 ER1094 ER1100 ER1168  ER538  ER625  ER626  ER716  ER745  ER746
##    515   4810   3994   1806     14   2037   1751   1422    583    525
```

```r
barplot(lib_sizes)
```



After seeing the library size differences we will transform the OTU counts to proportions with function `transform_sample_counts()`.

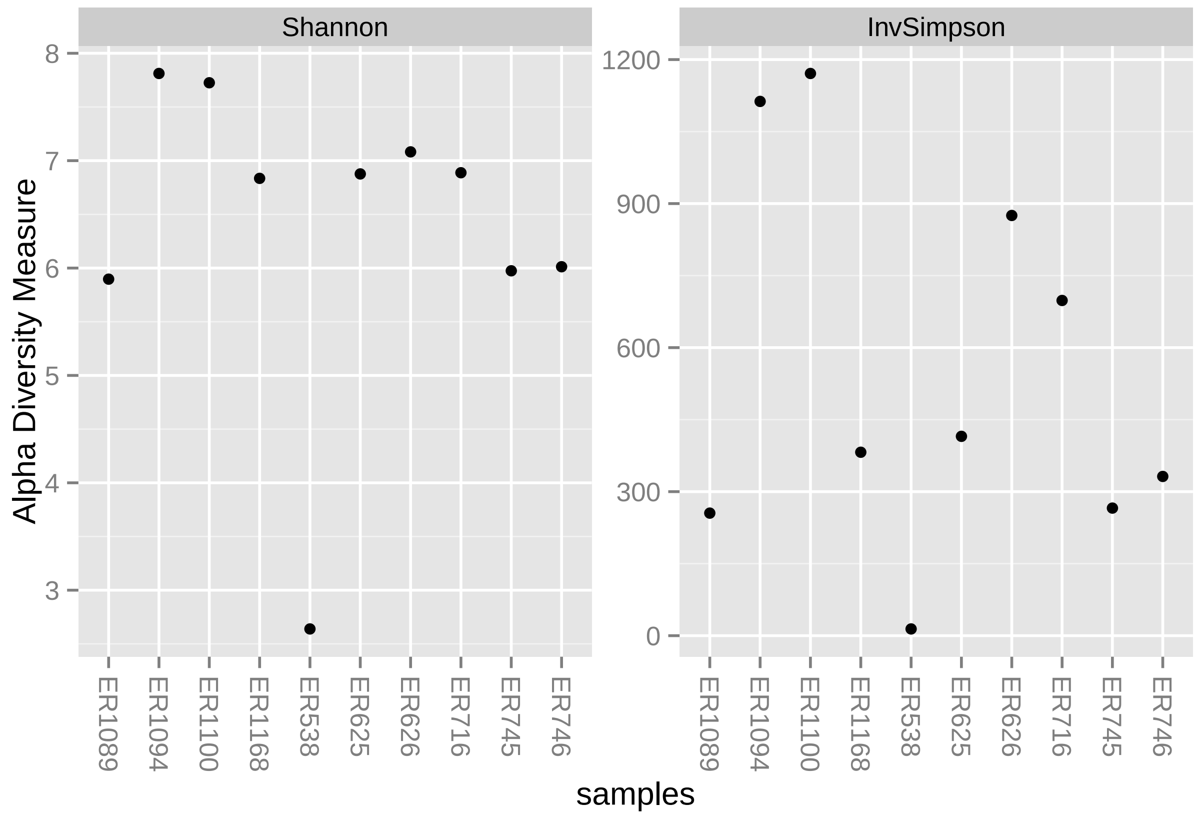
```
my_data_prop <- transform_sample_counts(my_data, function(x) x/sum(x))
barplot(sample_sums(my_data_prop), las=3)
```



Few figures from the data

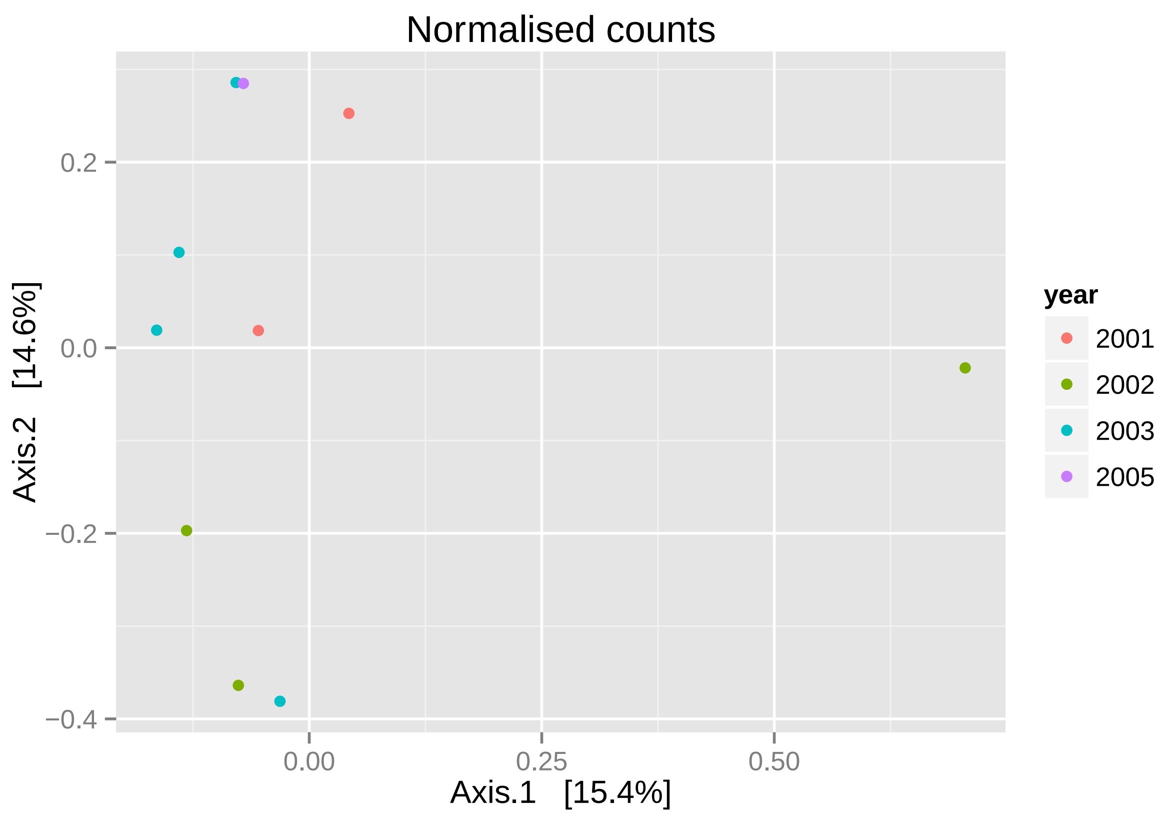
We can also use `phyloseq` to calculate and visualise some α- and β- diversity measures. Calculate Shannon and inverse Simpson indeces using the raw count data.

```
plot_richness(my_data, measures=c("shannon", "invsimpson"))
```

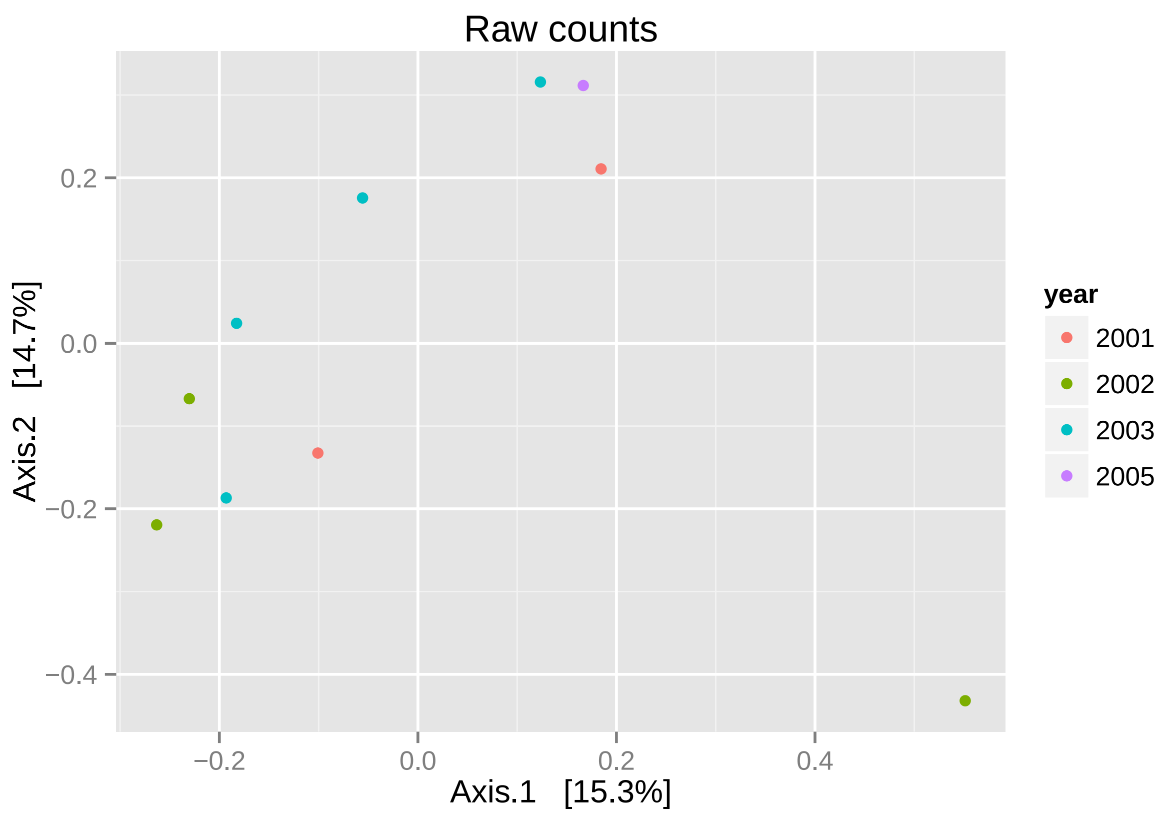


Then β-diversity with Bray-Curtis dissimilarity index and visualise it with PCoA ordination. In here use the normalised data. You can also do another one with raw counts and see if there is any difference. Use the sampling year for the color and the sampling month for the shape of the points.

```
my_data_ord <- ordinate(my_data_prop, method="PCoA", distance="bray")
plot_ordination(my_data_prop, my_data_ord, col="year", title="Normalised counts")
```

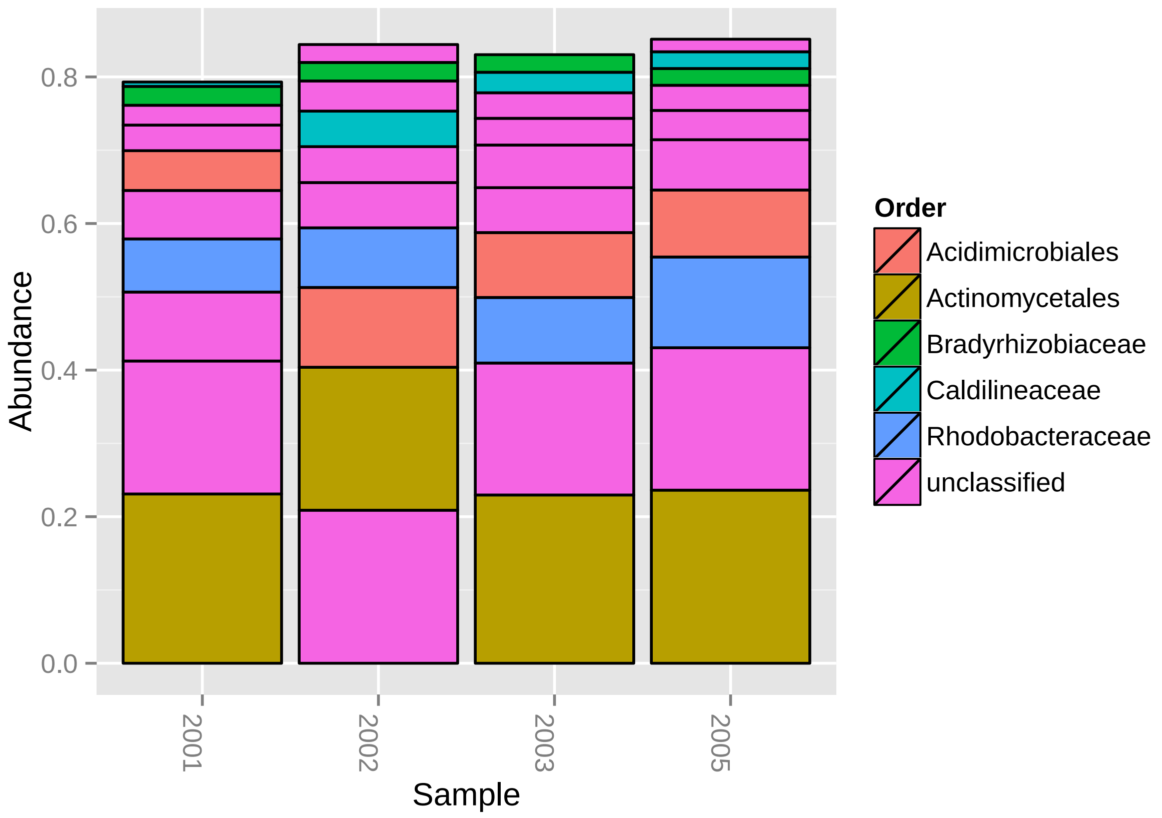


```r
my_data_ord <- ordinate(my_data, method="PCoA", distance="bray")
plot_ordination(my_data, my_data_ord, col="year", title="Raw counts")
```



As the last exercise we try to replicate the figure 1B from the original article. `phyloseq` has two funtions that are handy for this, one that merges samples based on metadata and one that calculates counts for different taxonomic levels. So we'll first merge the samples by year, calculate counts for order level and then make the barplot.

```r
my_data_merged <- merge_samples(my_data, "year")
my_data_merged <- transform_sample_counts(my_data_merged, function(x) x/sum(x))
my_data_merged_order <- tax_glom(my_data_merged, taxrank="Order")
my_data_abund <- prune_taxa(names(sort(taxa_sums(my_data_merged_order),TRUE)[1:10]),
                            my_data_merged_order)
plot_bar(my_data_abund, fill="Order")
```



Close enough.