

Multiple samples and sample groups



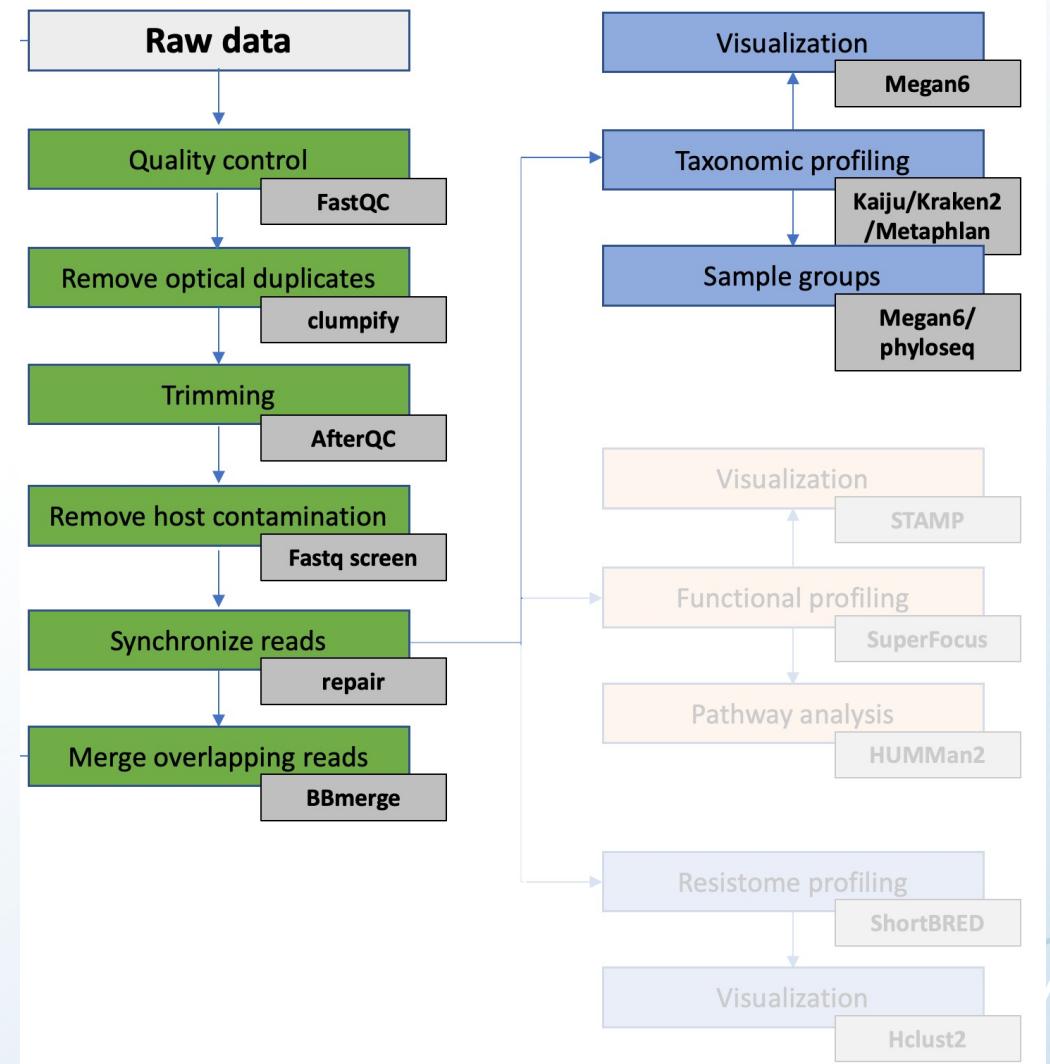
***experimental
group***



***control
group***

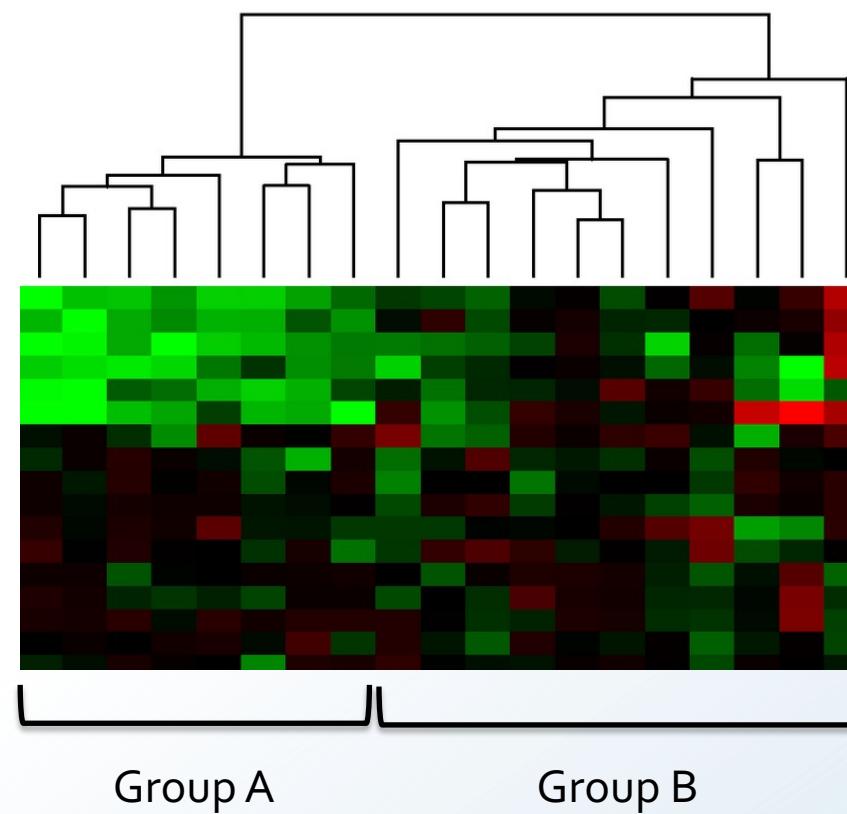
Overview

Comparison of groups
Alpha and beta diversity
Differential abundance
Introduction to R and RStudio
Introduction to phyloseq



Why do we want to perform comparison of sample groups?

Large within group variation - Need multiple samples in order to build evidence



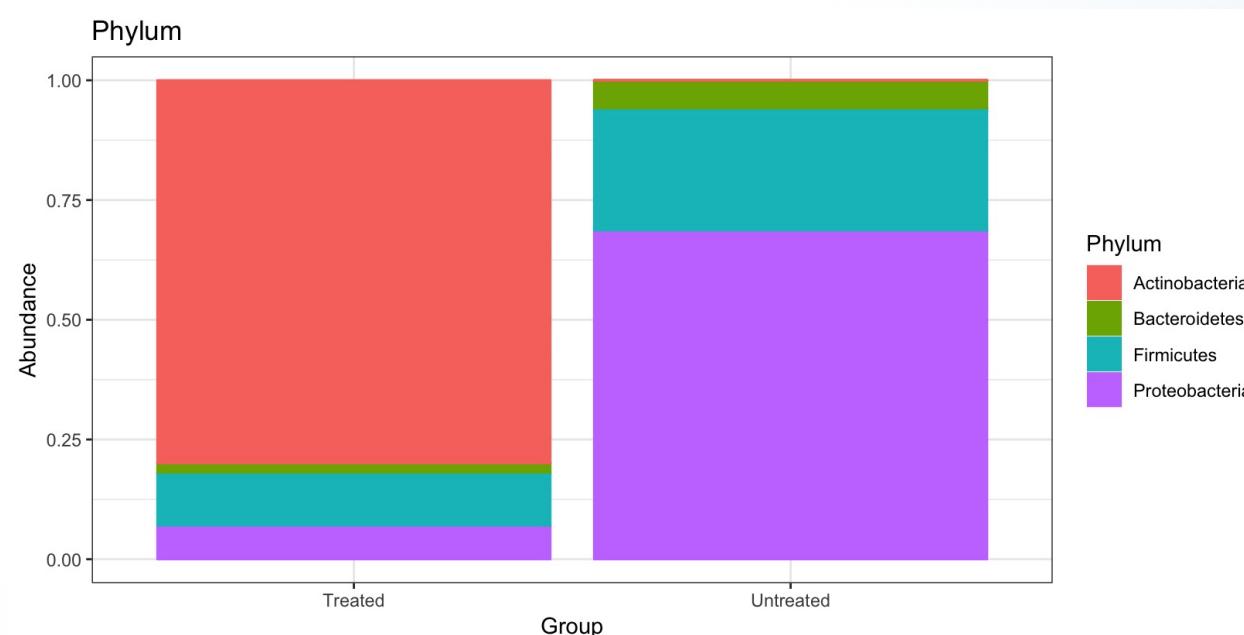
Compositional comparison of samples or groups

Compare the microbial composition between samples or groups of samples, eg:

Microbiome of patients treated with medication against untreated patients

Microbiome of healthy people against sick people

Microbiome of individuals before and after treatment



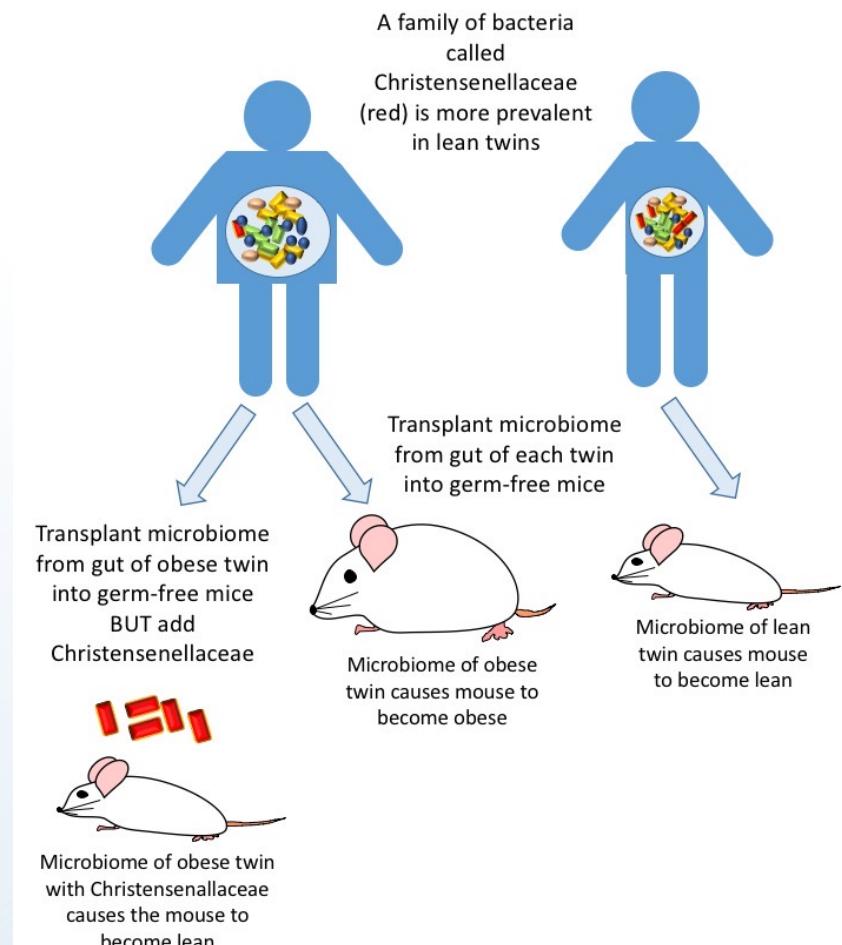
Gut microbiome comparison of lean versus obese twins

Obese individuals have lower diversity

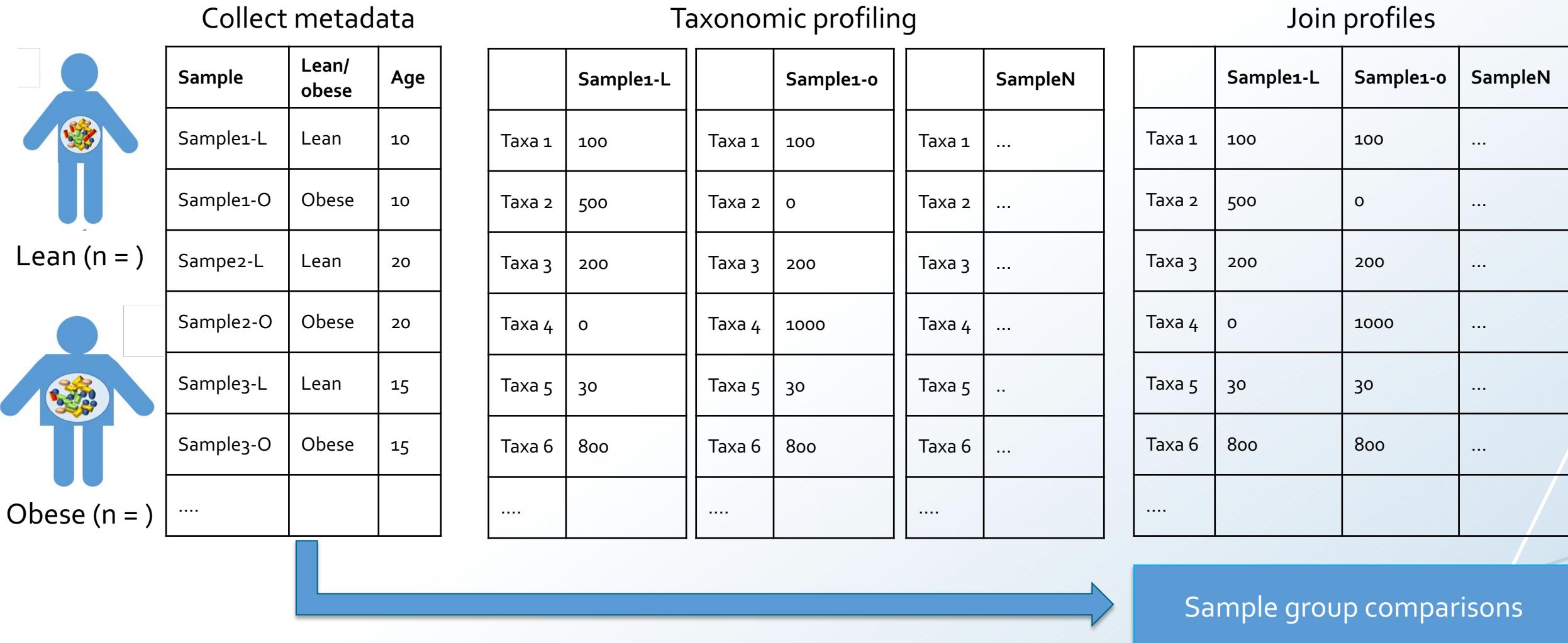
Lean individuals have higher abundance of
Christensenellaceae

Faecal transplantation to germ free mice:

- From obese donors = obese mice
- From lean donors = lean mice
- From obese donors + *Christensenellaceae* = lean mice

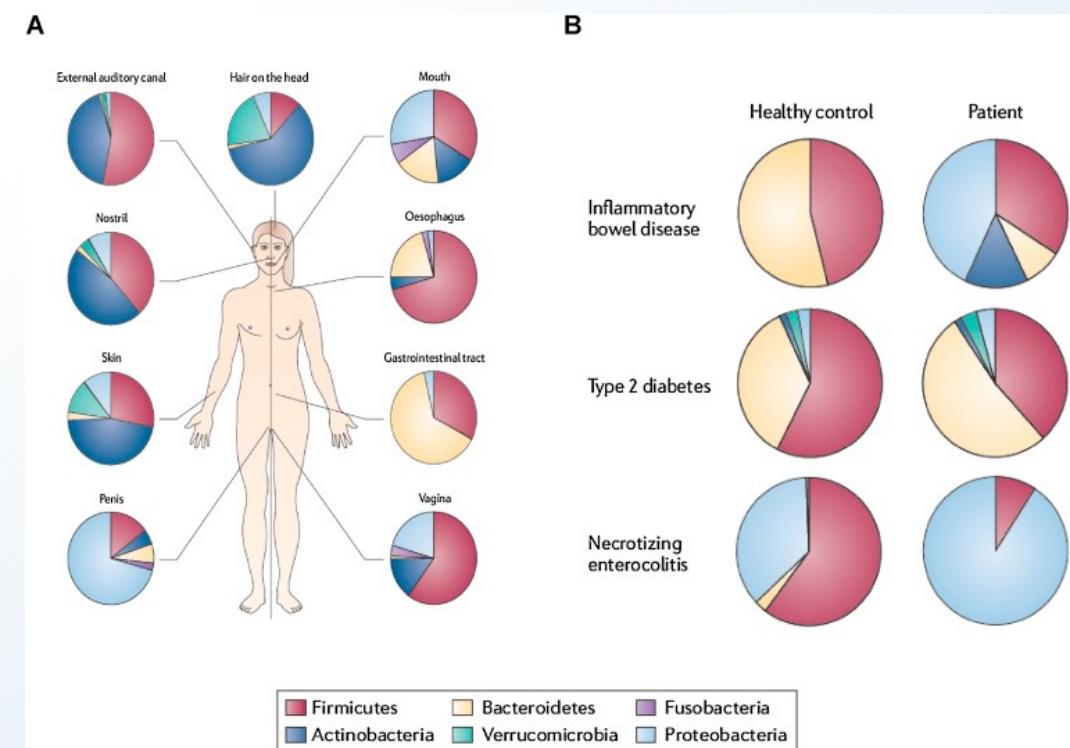
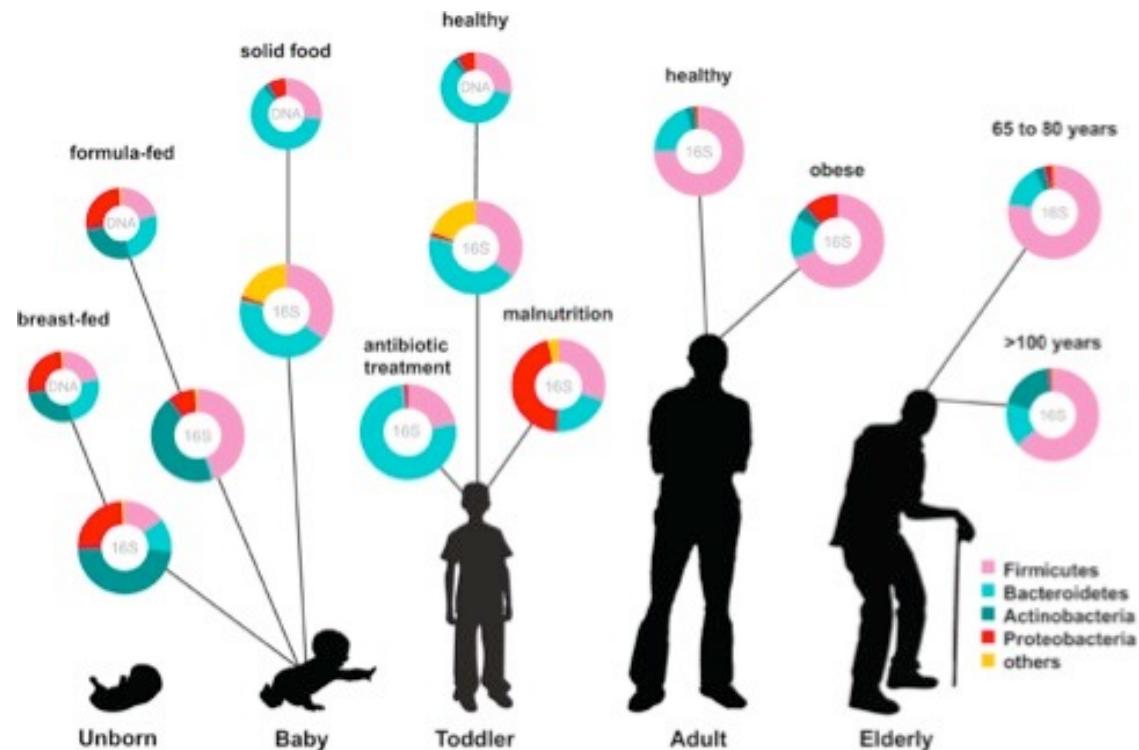


How can we perform comparison of microbial communities?



The human microbiome project – characterisation of the human microbiome

Characterised microbial communities at several different sites on the human body
Investigated the role of these microbes in human health and disease



Microbial community profiling - terminology

Community

“Group or association of populations of two or more different species”

Richness

The species richness is how many species there are in a sample

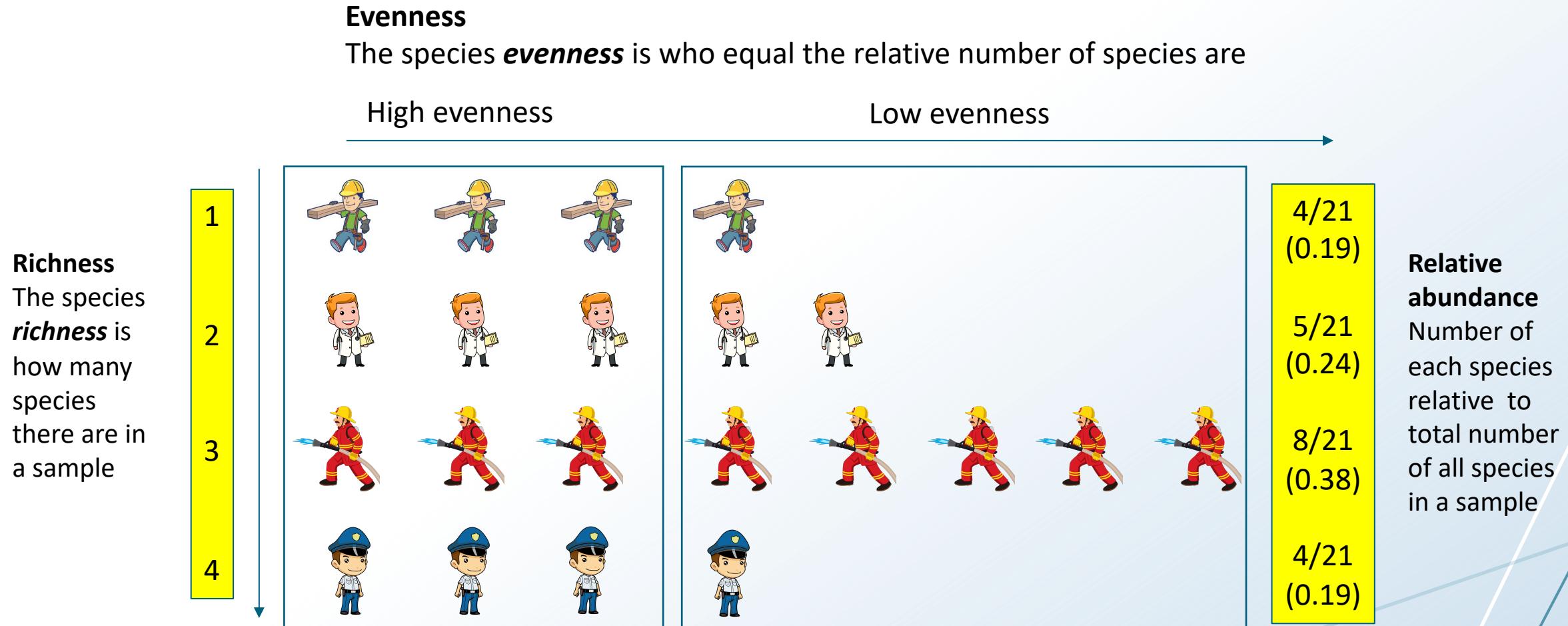
Evenness

The species evenness is how equal the relative number of species are in a sample

Diversity = Richness + Evenness



Alpha diversity – Within sample diversity



Alpha diversity – How many different species are in each sample and how evenly they are distributed

Community 1



Richness = 4

Evenness (high) >

Diversity

Community 2



Richness = 4

Evenness (low)

Diversity

Alpha diversity – How many different species are in each sample and how evenly they are distributed

Community 1



Taxa table

Taxa ID	Community 1	Community 2	Taxon
Taxa 1	4	8	
Taxa 2	4	4	
Taxa 3	4	2	
Taxa 4	3	1	

Community 2



How do we measure Alpha diversity - Diversity indices

A diversity index is a mathematical measure of species diversity in a community

Richness estimators: “How many?”

Basically count the number of different species in the sample

OTU richness (amplicon) – count of different OTUs

Observed Species – count of unique species

Chao1 richness – estimate richness by adding a correction factor to the observed number of species

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{n_1^2}{2n_2}$$

S_{obs} is the number of observed species

n_1 is the number of singletons (species captured once)

n_2 is the number of doubletons (species captured twice).

How do we measure Alpha diversity - Diversity indices

A diversity index is a mathematical measure of species diversity in a community

Richness estimators: “How many?”

Basically count the number of different species in the sample

OTU richness (amplicon) – count of different OTUs

Observed Species – count of unique species

Chao1 richness – estimate richness by adding a correction factor to the observed number of species

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{n_1^2}{2n_2} = 4 + \frac{1}{1} = 5$$

Taxa ID	Community 2	Taxon
Taxa 1	8	
Taxa 2	4	
Taxa 3	2	
Taxa 4	1	

How do we measure Alpha diversity - Diversity indices

A diversity index is a mathematical measure of species diversity in a community

Richness and evenness estimators: “How different?”

The calculated value of diversity increases both when the number of species increases and when evenness increases.

Simpson and Fisher index

Shannon diversity: accounts for both abundance and evenness of the species present

$$H' = \frac{N \ln N - \sum(n_i \ln n_i)}{N}$$

N is the total number of species counts

n_i is the number of individuals in species i

How do we measure Alpha diversity - Diversity indices

A diversity index is a mathematical measure of species diversity in a community

Richness and evenness estimators: “How different?”

The calculated value of diversity increases both when the number of species increases and when evenness increases.

Shannon diversity: accounts for both abundance and evenness of the species present

$$\begin{aligned} H' &= \frac{N \ln N - \sum(n_i \ln n_i)}{N} = \frac{15 \times \ln 15 - ((8 \times \ln 8) + (4 \times \ln 4) + (2 \times \ln 2) + (1 \times \ln 1))}{15} \\ &= \frac{40,620 - (16,636 + 5,545 + 1,386 + 0)}{15} \\ &= \underline{\underline{1,137}} \end{aligned}$$

Taxa ID	Community 2	Taxon
Taxa 1	8	
Taxa 2	4	
Taxa 3	2	
Taxa 4	1	

N = 15

How do we measure Alpha diversity - Diversity indices

A diversity index is a mathematical measure of species diversity in a community

Richness and evenness estimators: “How different?”

The calculated value of diversity increases both when the number of species increases and when evenness increases.

Shannon diversity: accounts for both abundance and evenness of the species present

$$\begin{aligned} H' &= \frac{N \ln N - \sum(n_i \ln n_i)}{N} = \frac{15 \times \ln 15 - ((4 \times \ln 4) + (4 \times \ln 4) + (4 \times \ln 4) + (3 \times \ln 3))}{15} \\ &= \frac{40,620 - (5,545 + 5,545 + 5,545 + 3,295)}{15} \\ &= \underline{\underline{1,379}} \end{aligned}$$

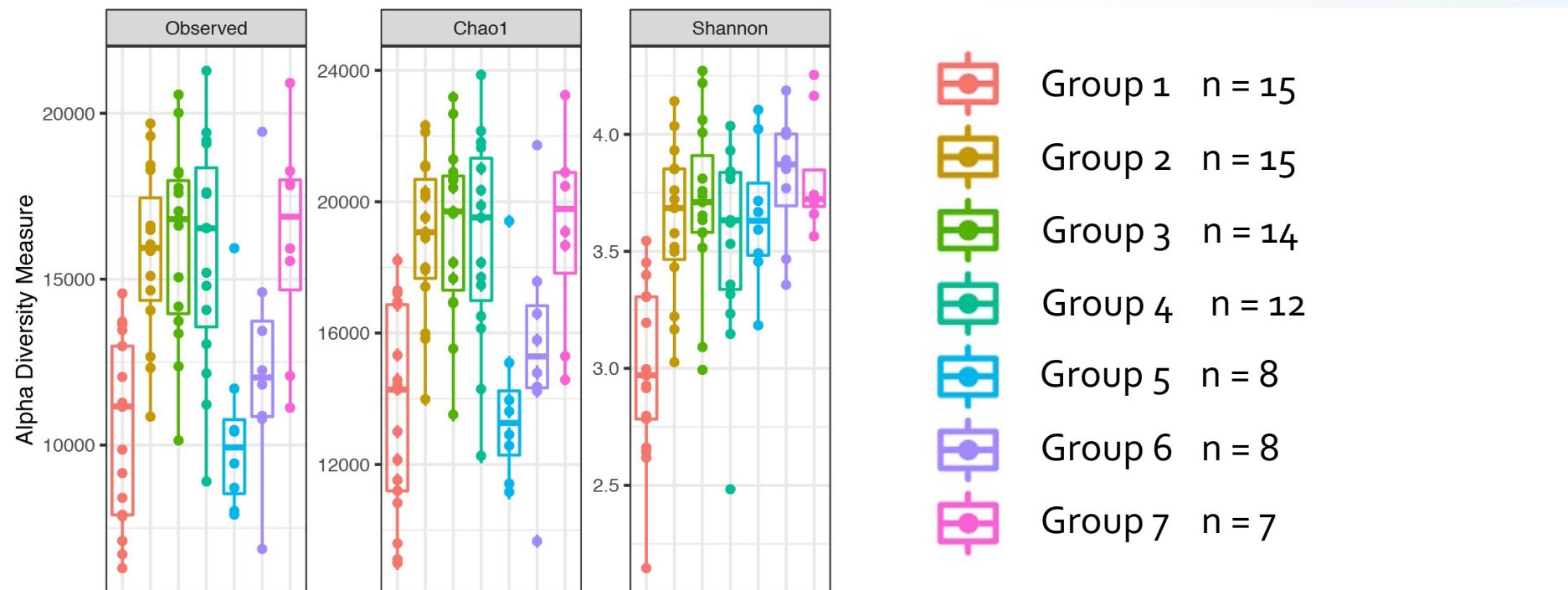
Taxa ID	Community 1	Taxon
Taxa 1	4	
Taxa 2	4	
Taxa 3	4	
Taxa 4	3	

N = 15

Visualisation of Alpha diversity

Alpha diversity is normally visualised through box plots – using multiple diversity indices

Analysis can be performed on different taxonomical levels (eg. species or genus)



Determine if alpha diversity is statistical significant between samples or groups that are compared

We are testing if the samples or groups we are comparing are equal (null hypothesis)

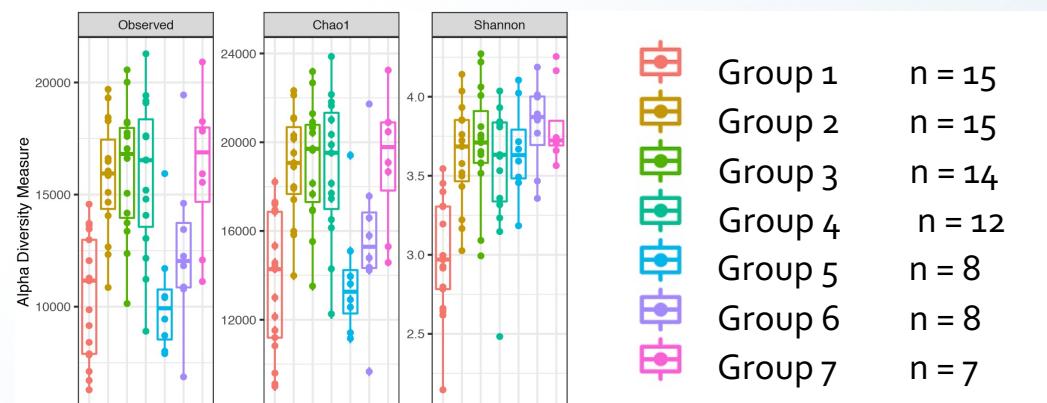
Output is normally a p-value, and $p < 0.05$ is regarded as “statistically significant”

This means that it is highly unlikely that the samples or groups are equal

The statistical tests normally performed with statistical packages

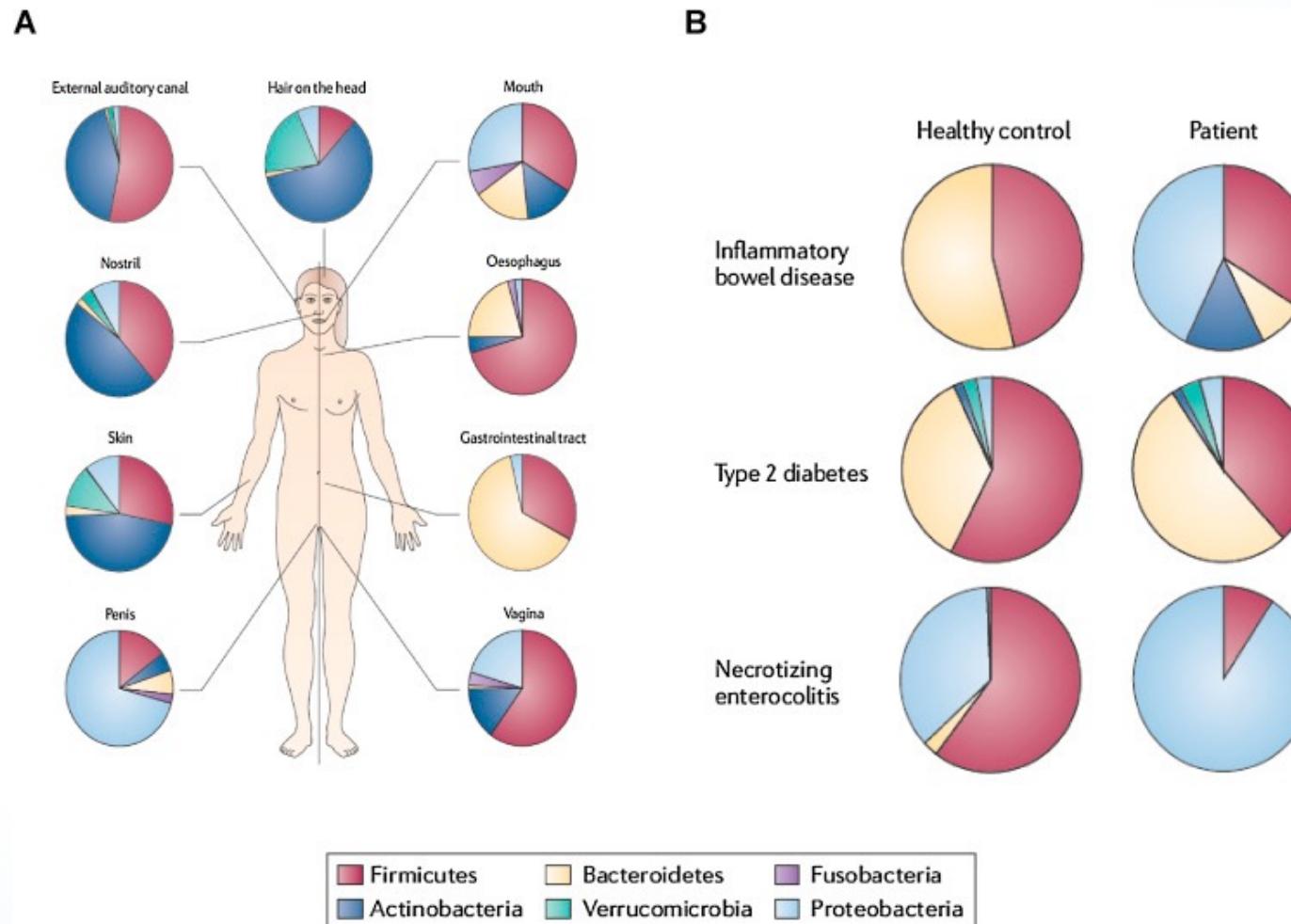
ANOVA - when the data is roughly normally distributed

Wilcoxon rank-sum test (Mann-Whitney) - when the data is not normally distributed



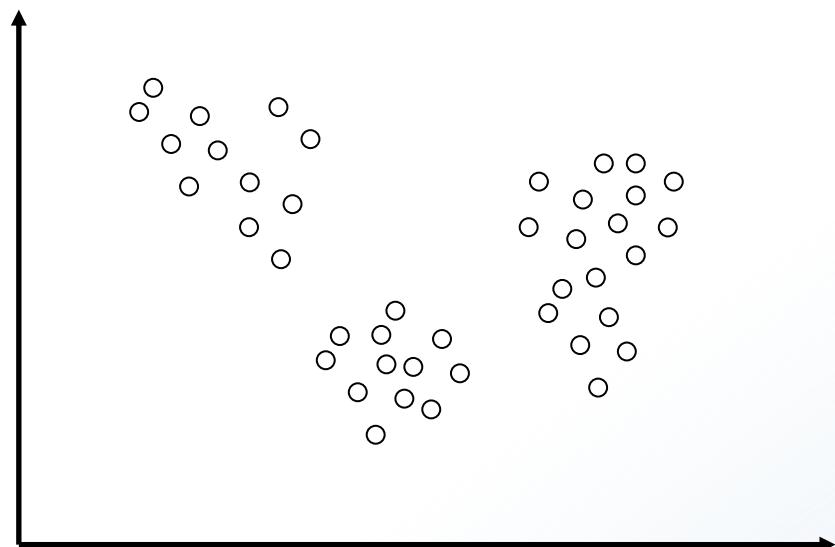
Beta diversity – Between sample diversity, or the number of species that are not the same in two different environments

How different is the microbial composition in one environment compared to another?



Beta diversity – differences between environments

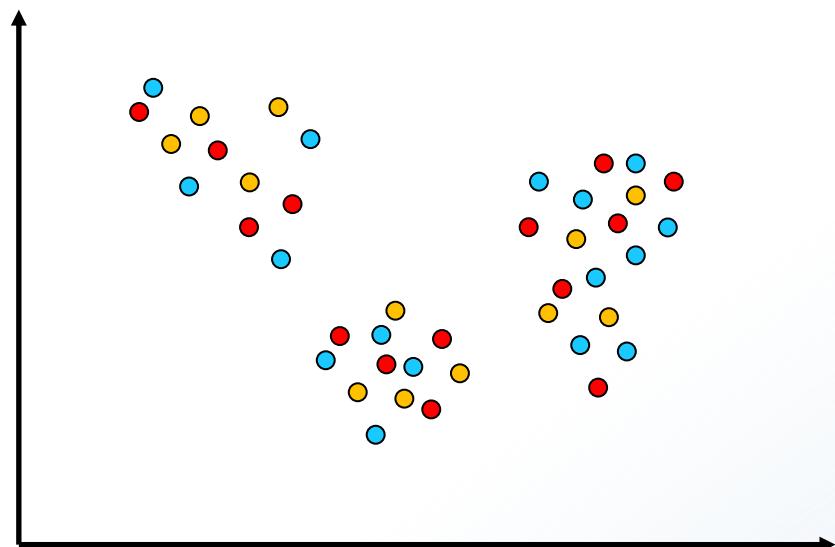
Beta diversity analysis is often performed to identify the influence of a treatment versus a control, or to explore the differences between two sample sites



	Age	Gender	Body site
Sample 1	10	Male	Skin
Sample 2	10	Female	Skin
Sample 3	20	Male	Skin
Sample 4	20	Female	Tounge
Sample 5	30	Male	Tounge
Sample 6	10	Male	Stool
Sample 7	30	Female	Stool
...			
Sample N	30	Female	Skin

Beta diversity – differences between environments

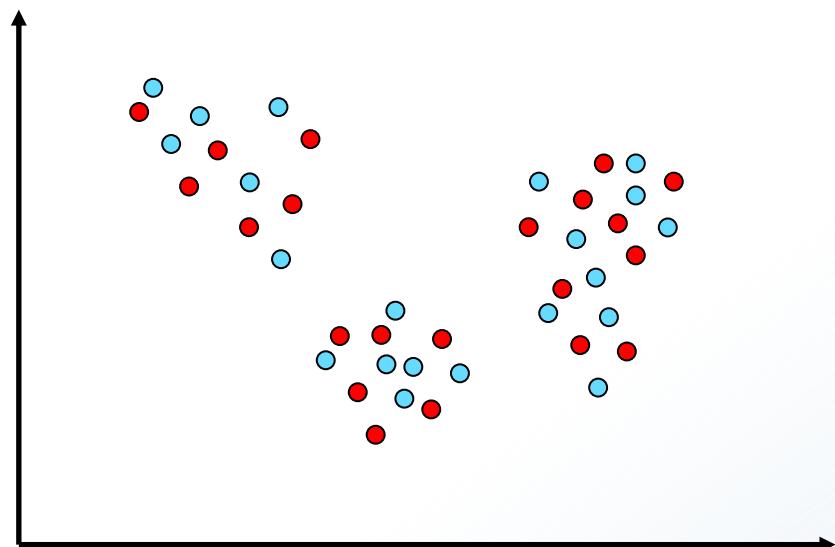
Beta diversity analysis is often performed to identify the influence of a treatment versus a control, or to explore the differences between two sample sites



	Age	Gender	Body site
Sample 1	10	Male	Skin
Sample 2	30	Female	Skin
Sample 3	20	Male	Skin
Sample 4	20	Female	Tounge
Sample 5	30	Male	Tounge
Sample 6	10	Male	Stool
Sample 7	30	Female	Stool
...			
Sample N	30	Female	Skin

Beta diversity – differences between environments

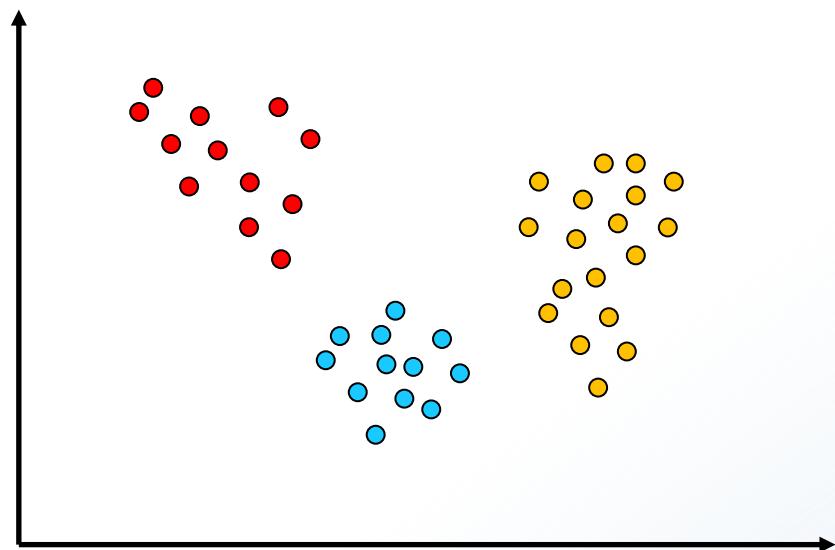
Beta diversity analysis is often performed to identify the influence of a treatment versus a control, or to explore the differences between two sample sites



	Age	Gender	Body site
Sample 1	10	Male	Skin
Sample 2	10	Female	Skin
Sample 3	20	Male	Skin
Sample 4	20	Female	Tounge
Sample 5	30	Male	Tounge
Sample 6	10	Male	Stool
Sample 7	30	Female	Stool
...			
Sample N	30	Female	Skin

Beta diversity – differences between environments

Beta diversity analysis is often performed to identify the influence of a treatment versus a control, or to explore the differences between two sample sites



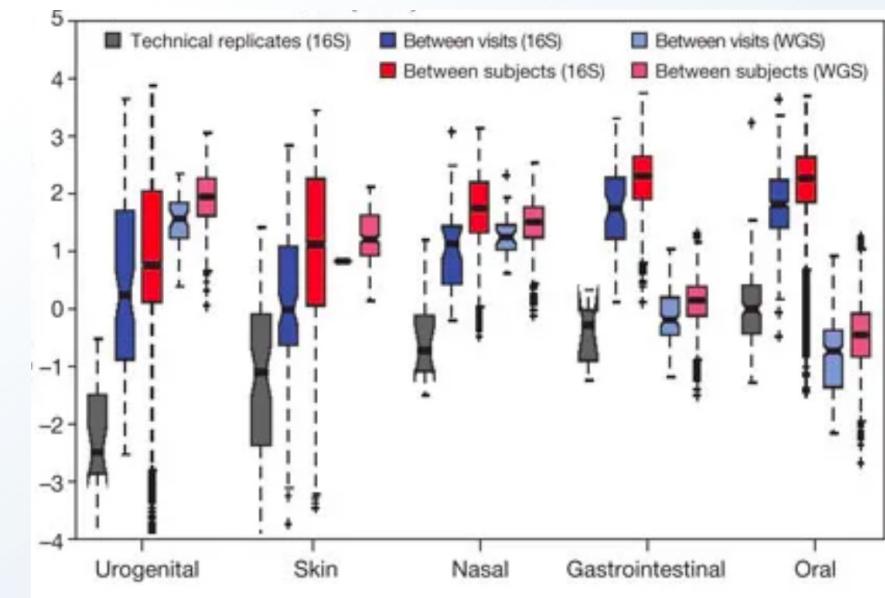
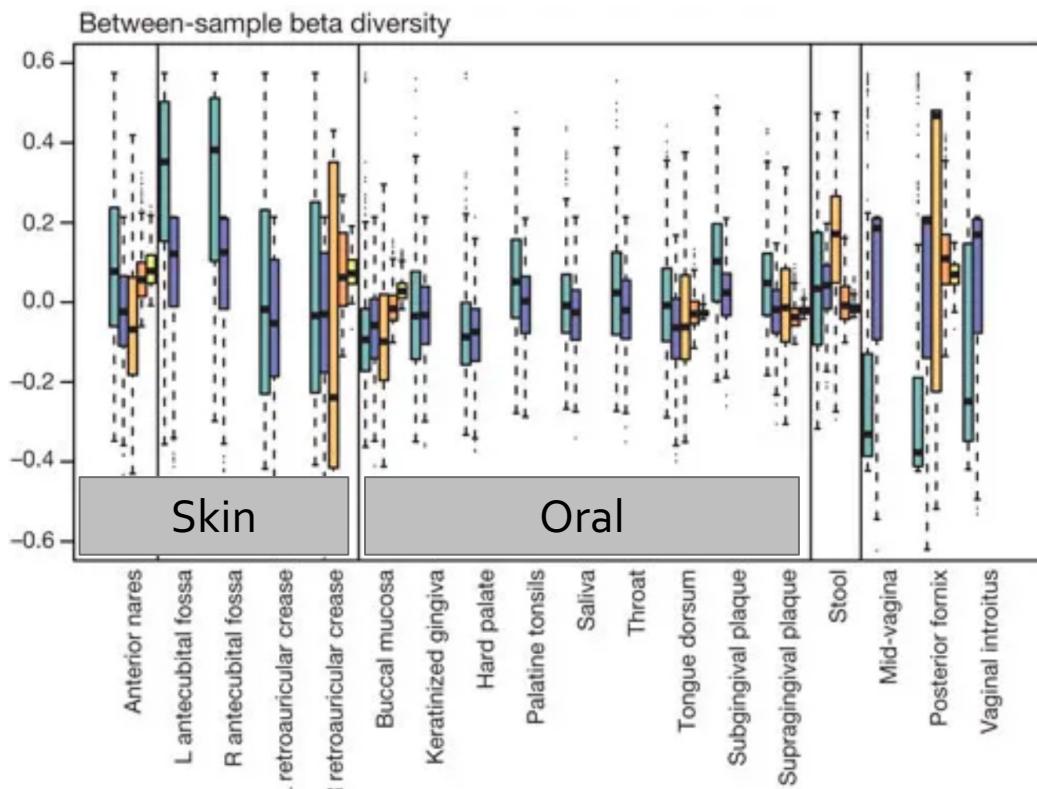
	Age	Gender	Body site
Sample 1	10	Male	Skin
Sample 2	10	Female	Skin
Sample 3	20	Male	Skin
Sample 4	20	Female	Tounge
Sample 5	30	Male	Tounge
Sample 6	10	Male	Stool
Sample 7	30	Female	Stool
...			
Sample N	30	Female	Skin

Beta diversity – differences between environments

Beta diversity between samples within body sites

More variability between the samples on the skin than in the oral cavity

More variability between individuals than between visits



How do we measure Beta diversity – Diversity metrics

Beta diversity describes how different every sample is from every other sample. Thus, each sample has more than one value.

Many different distance measures:

Some metrics take abundance into account (*i.e.* diversity: Bray-Curtis, weighted UniFrac)

Some only calculate based on presence-absence (*i.e.* richness: Jaccard, unweighted UniFrac)

Calculation of beta-diversity appears like this (made up numbers)

	Sample 1	Sample 2	Sample 3
Sample 1	0	0,444	0,888
Sample 2	0,444	0	0,666
Sample 3	0,888	0,666	0

How do we measure Beta diversity – Diversity metrics

Bray–Curtis dissimilarity is based on abundance or read count data

Measure differences in microbial abundances between two samples (e.g., at species level)

values are from 0 to 1

0 means both samples share the same species at exact the same abundances

1 means both samples have complete different species abundances

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

i & j are the two sites

S_i is the total number of specimens counted on site i

S_j is the total number of specimens counted on site j

C_{ij} is the sum of only the lesser counts for each species found in both sites

How do we measure Beta diversity – Diversity metrics

Bray–Curtis dissimilarity is based on abundance or read count data

Measure differences in microbial abundances between two samples (e.g., at species level)

values are from 0 to 1

0 means both samples share the same species at exact the same abundances

1 means both samples have complete different species abundances

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j} = 1 - \frac{2 \times 9}{15 + 15} = 0,4$$

i & j are the two sites

S_i is the total number of specimens counted on site i

S_j is the total number of specimens counted on site j

C_{ij} is the sum of only the lesser counts for each species found in both sites

$$\begin{matrix} 4 \\ 4 \\ 2 \\ 1 \end{matrix} \quad C_{ij} = 9$$

Taxa ID	Community 1	Community 2	Taxon
Taxa 1	4	8	
Taxa 2	4	4	
Taxa 3	4	2	
Taxa 4	3	1	

$S_i = 4+4+4+3 = 15$ $S_j = 8+4+4+1 = 15$

Ordination - summarise multivariate data in fewer dimensions than the original data set

True biological samples can consist of > 1000 species – impossible to visualise in plot

Solution: reduce to a few key trends that are shared = possible to derive a smaller set of axes (e.g., two) that could be plotted to summarize most of the variation in the data set

Pair wise distances of all samples = huge table.

Need to project the whole beta diversity table down to 2-3 dimensions in order to visualise

	Sample 1	Sample 2	Sample 3
Sample 1	0	0,444	0,888
Sample 2	0,444	0	0,666
Sample 3	0,888	0,666	0

Ordination - summarise multivariate data in fewer dimensions than the original data set

PCoA (Principal Coordinates Analyses) also called MDS (Metric Dimensional Scaling)

NMDS (Non Metric Dimensional Scaling)

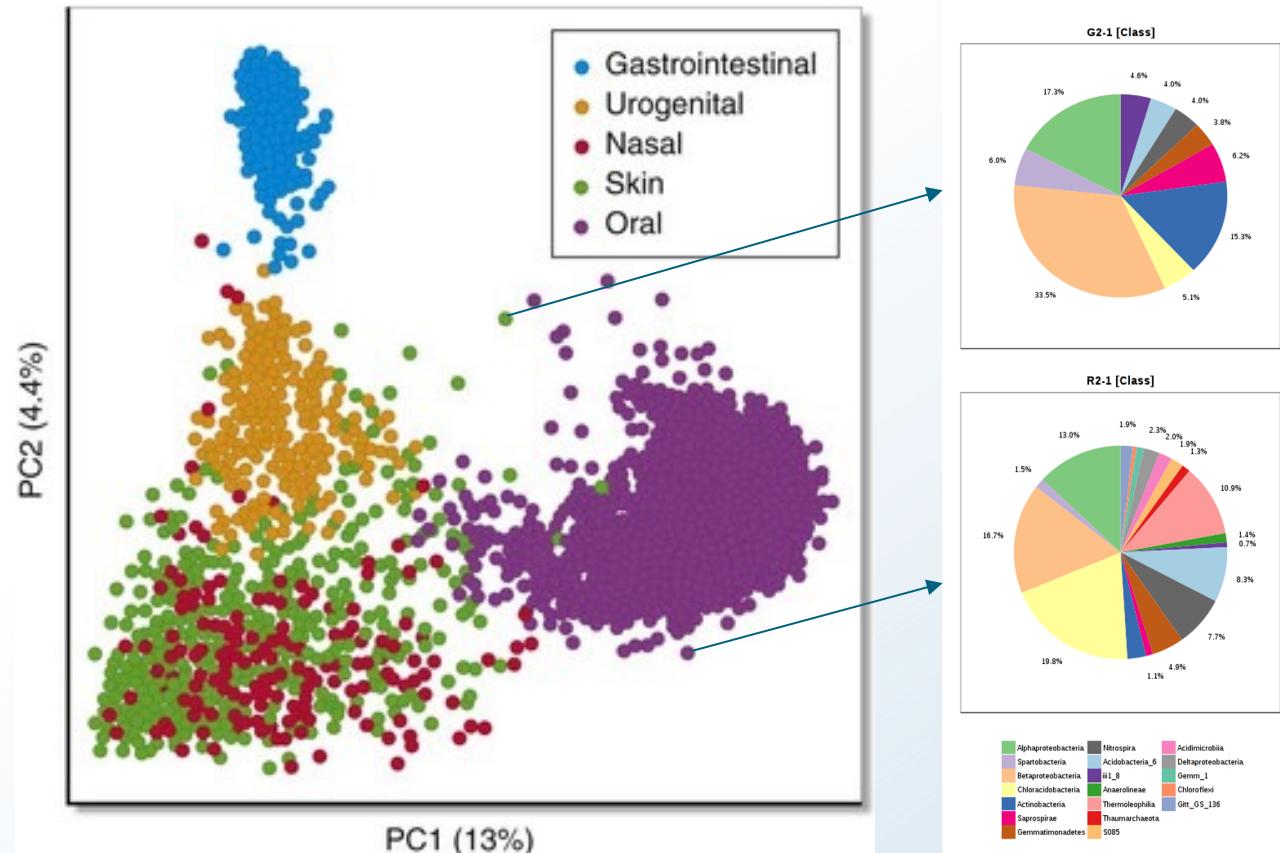
PCA (Principal Components Analysis)

Visualisation of Beta diversity

Beta diversity is normally visualised through PCA, PCoA or NMDS plots

Each symbol in the plot represents the total microbial community of that sample

Symbols closer together have more similar microbiotas while those farther apart have less similar



Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207-214.

Determine if beta diversity is statistical significant between samples or groups that are compared

Common to test statistically whether there is a significant difference between groups

We test whether the overall microbial community differs by your variable of interest

We are testing if the samples or groups we are comparing are equal (null hypothesis)

Output is normally a p-value, and $p < 0.05$ is regarded as “statistically significant”

This means that it is highly unlikely that the samples or groups are equal

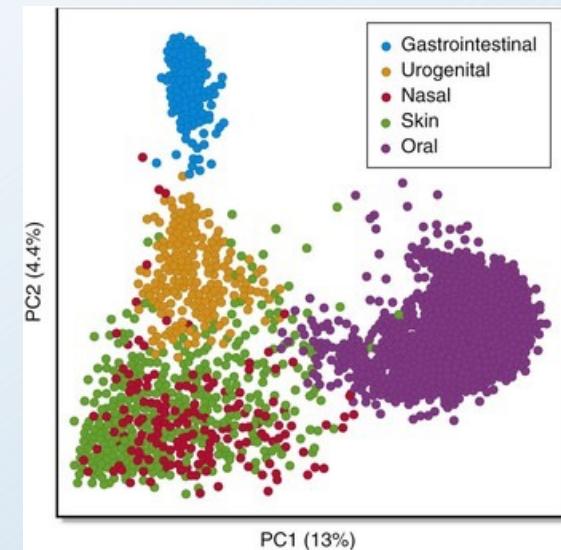
The statistical tests normally performed with statistical packages

ANOISM (Analysis of similarities)

tests whether distances between groups are greater than within groups

PERMANOVA (Multivariate ANOVA with permutations)

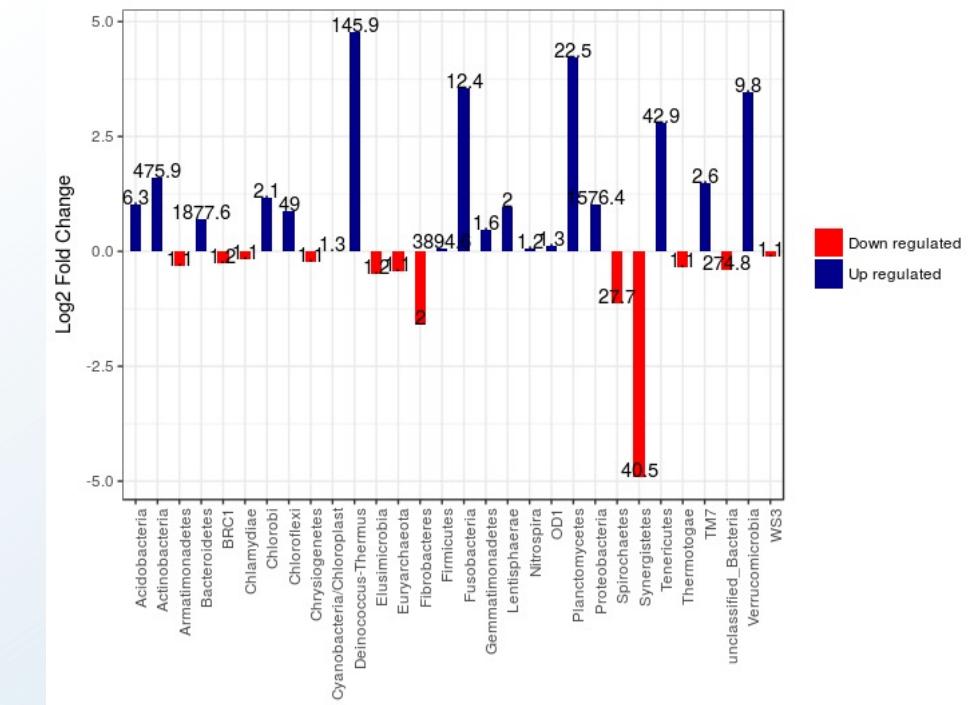
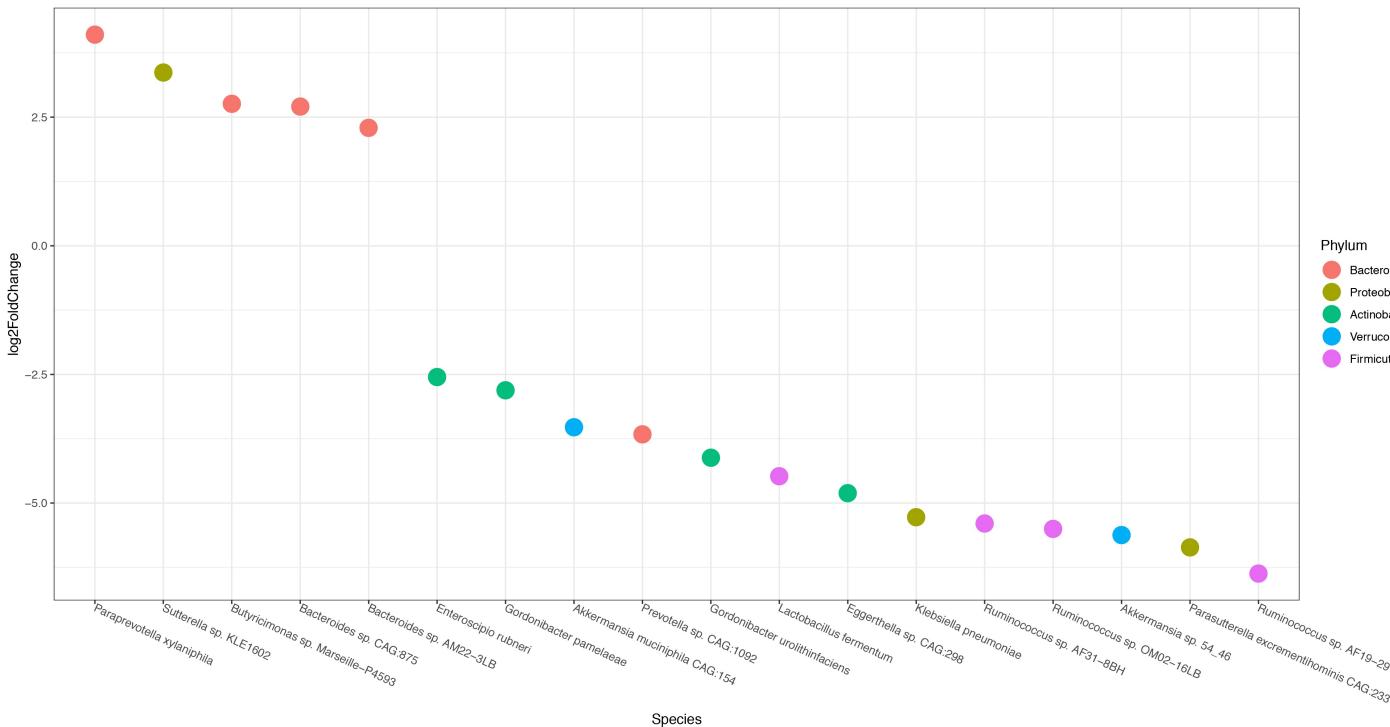
tests whether distance differ between groups



Differential abundance analysis - identify taxa that are significantly different between two groups

Compare the relative proportions of each bacterial species across microbiome samples and sample groups

DESeq2 estimate variance-mean dependence in count data and test for differential expression



Time for a break



R is a programming language for statistical computing and graphics

Widely used by professional statisticians and in bioinformatics

R is a dynamically typed interpreted language, and is typically used interactively

This means that you type commands and run them interactively to produce results – all in the console

It has many built-in functions and libraries, and is extensible, allowing users to define their own functions

R has lots of great functions for producing publication quality plots

Reproducible analysis

Document what you have done with your data in code

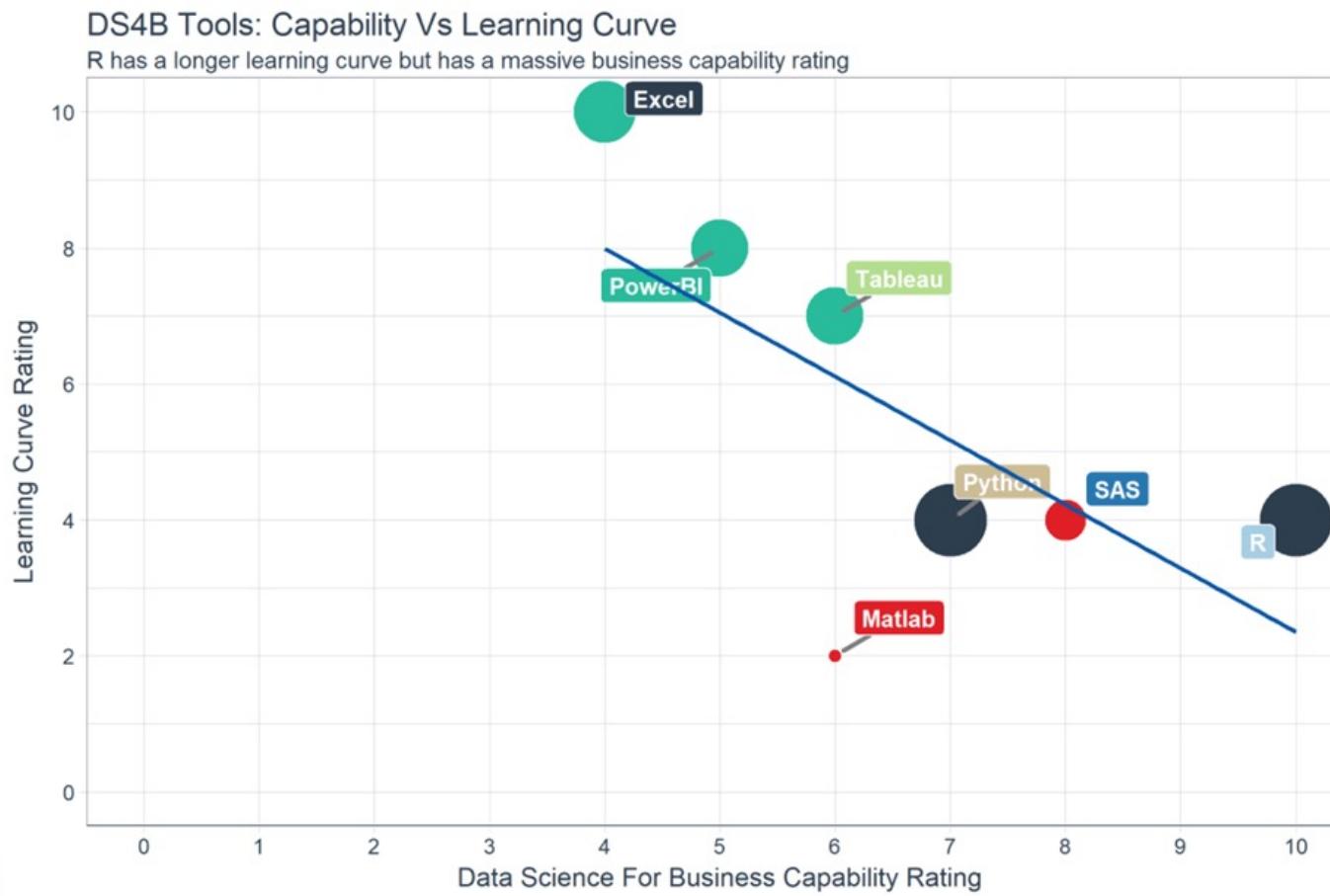
Collaborative

Share your data and analysis



R is a programming language for statistical computing and graphics

R is harder to learn than excel, but the capabilities is much greater



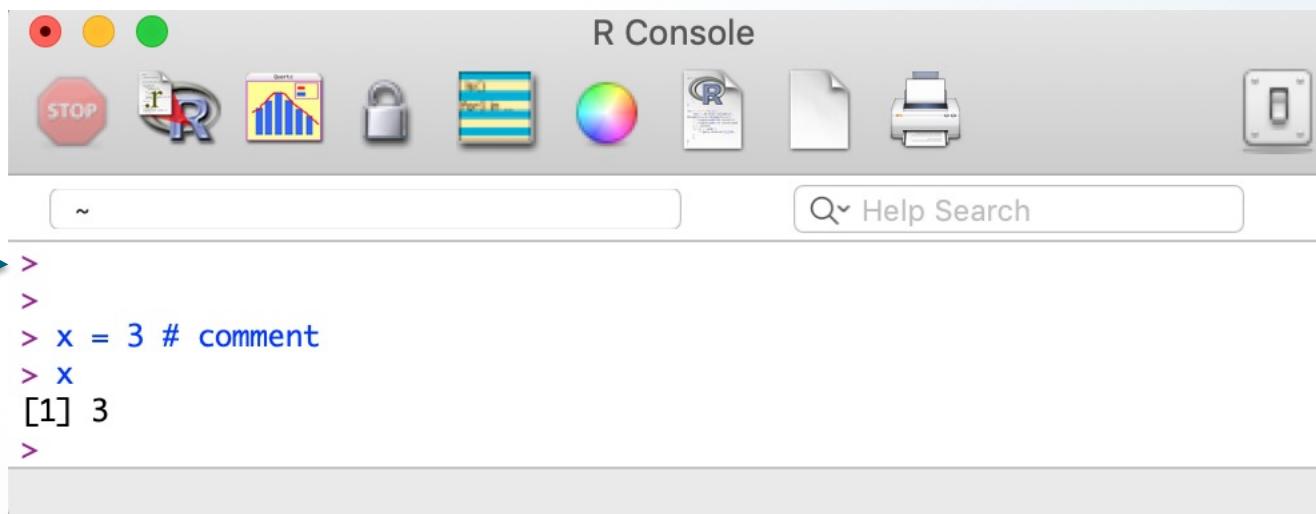
Some R basics

Typical usage is to read data into R from other sources (eg. taxonomy read count tables from Kaiju)

What you do with the data is then written in code (R language)

Assignment operator <- or = (It assigns values on the right to objects on the left)

All text after the pound sign "#" within the same line is considered a comment



The image shows a screenshot of the R Console window. At the top, there is a toolbar with various icons: a red STOP sign, a yellow R logo, a green file, a lock, a bar chart, a color palette, and a printer. To the right of the toolbar, it says "R Console". Below the toolbar is a search bar with a magnifying glass icon and the text "Help Search". The main area of the window is a text editor showing R code. A blue arrow points to the first line of code, which starts with a command prompt ">". The code is as follows:

```
>
>
> x = 3 # comment
> x
[1] 3
>
```

R packages - additional functionality beyond those offered by the core R library

An R package is a collection of functions, data, and documentation that extends the capabilities of base R.

Packages are the fundamental units of reproducible R code

Packages includes documentation that describes how to use them, and sample data

Once you have installed a package, you can load it with the 'library()' function

RStudio – a graphical interface for R

RStudio allows the user to run R in a more user-friendly environment

Install R packages

Visualise tables and plots

Import/export functionalities

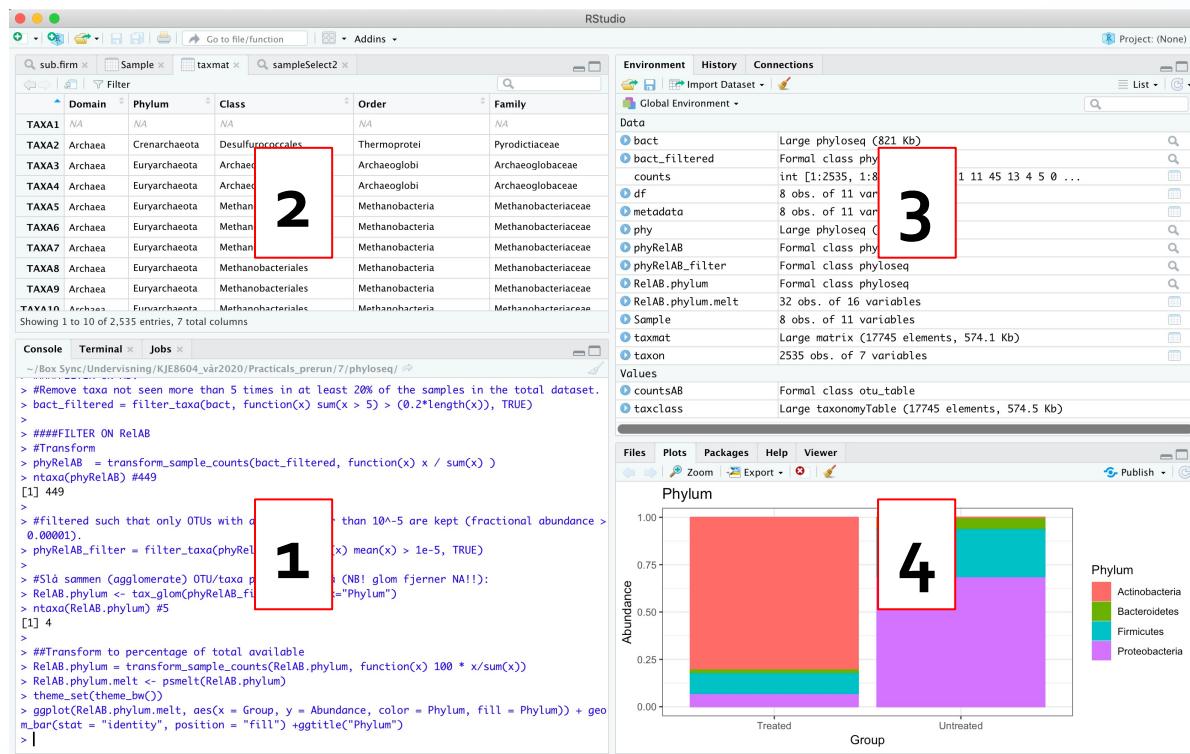


RStudio - screen

RStudio screen is divided into four parts:

The upper left part of RStudio displays information about objects and show tables (will be empty when you start RStudio). If you click on a table in the Environment list, you can see the data on a screen to the left.

The console is where you can type commands and see output



The **Environment** tab shows all the active objects.

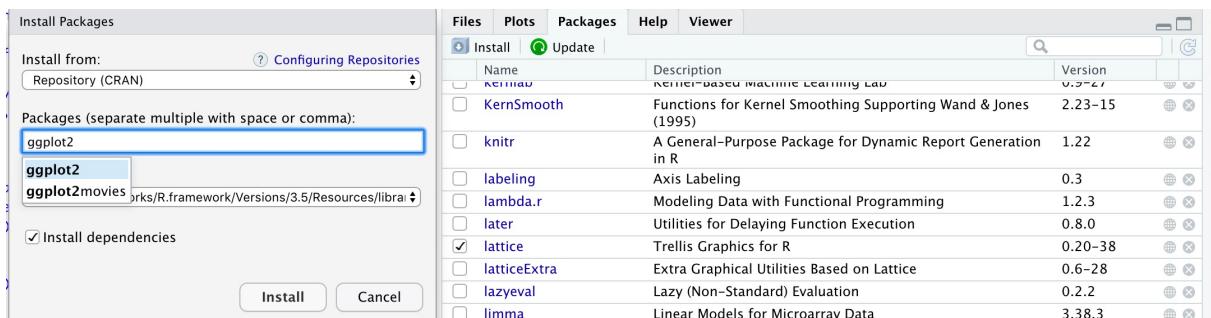
The **History** tab shows a list of commands used so far

The **Files** tab shows all the files and folders in your default workspace as if you were on a PC/Mac window. The **Plots** tab will show all your graphs.

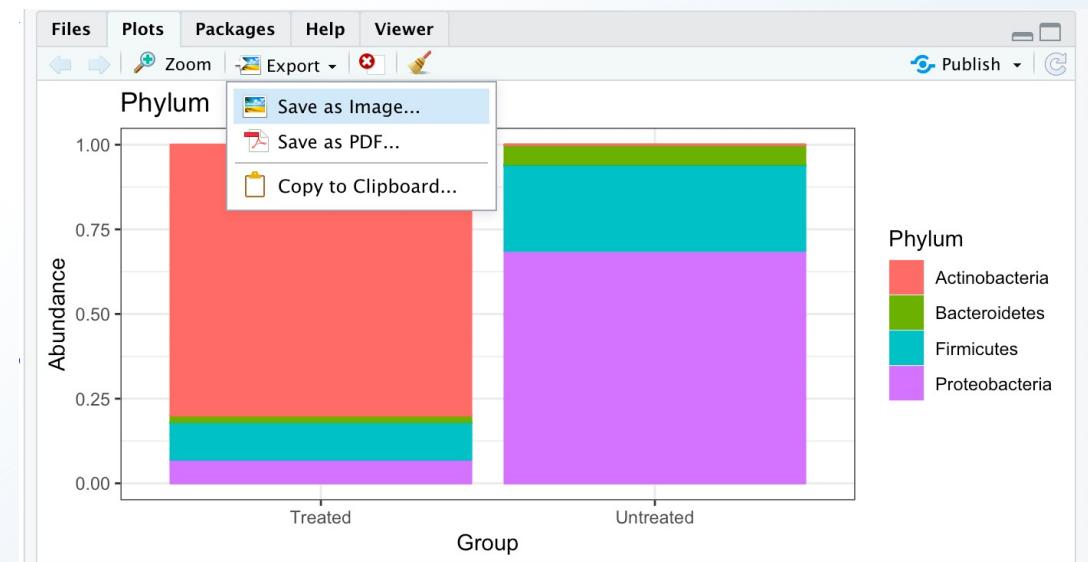
The **Packages** tab will list a series of packages or add-ons needed to run certain processes. For additional info see the **Help** tab

RStudio – Some nice features

Allow you to change general appearance (eg. set a default working directory)



Install R packages



Export plots (png, pdf, eps, svg, etc)

Bioconductor – source for most bioinformatics libraries and tools for R

[Bioconductor](#) project provides many additional R packages for statistical data analysis in different life science areas

Bioconductor is needed to install and run other R packages such as phyloseq

The screenshot shows the Bioconductor website homepage. At the top, there is a navigation bar with links for Home, Install, Help, Developers, and About. A search bar is also present. Below the navigation bar, there are several sections: 'About Bioconductor' which provides an overview of the project; 'News' which lists recent news items; 'Install' which links to software packages and installation instructions; 'Learn' which provides resources for learning Bioconductor tools; 'Use' which shows how to create bioinformatic solutions; and 'Develop' which provides developer resources. The overall design is clean and organized, with a focus on providing practical tools and resources for bioinformatics research.

phyloseq - R package to analyse community composition data in a phylogenetic framework



A tool to import, store, analyse, and graphically display complex phylogenetic sequencing data

Provides an object-oriented programming infrastructure

Simplifies many of the common data management and preprocessing tasks

Provide powerful analysis functions, building upon related packages available in R

Eg. calculating ecological distances

Use advanced/flexible graphic systems (ggplot2)

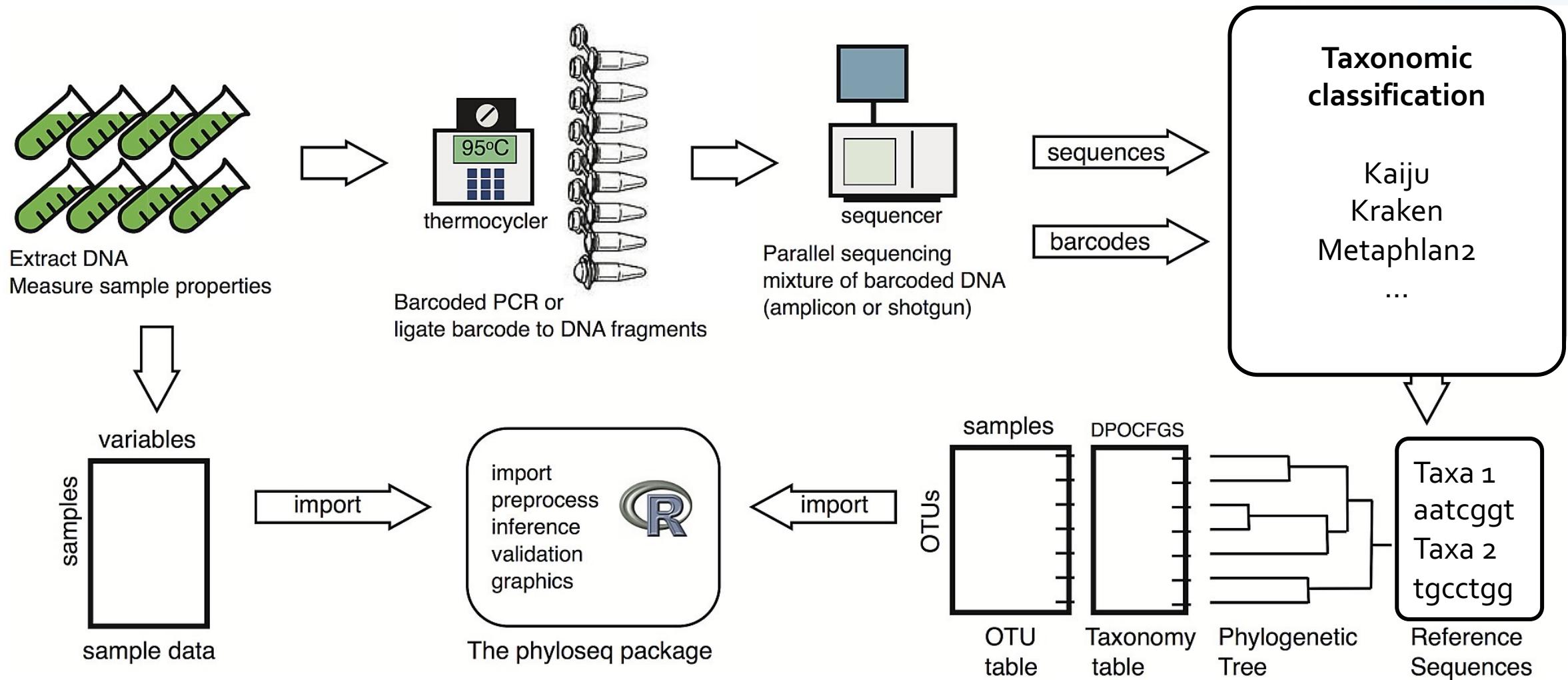
Produce publication-quality graphics of complex phylogenetic data

Well documented functions and active user community

Easy to share data and reproduce analyses



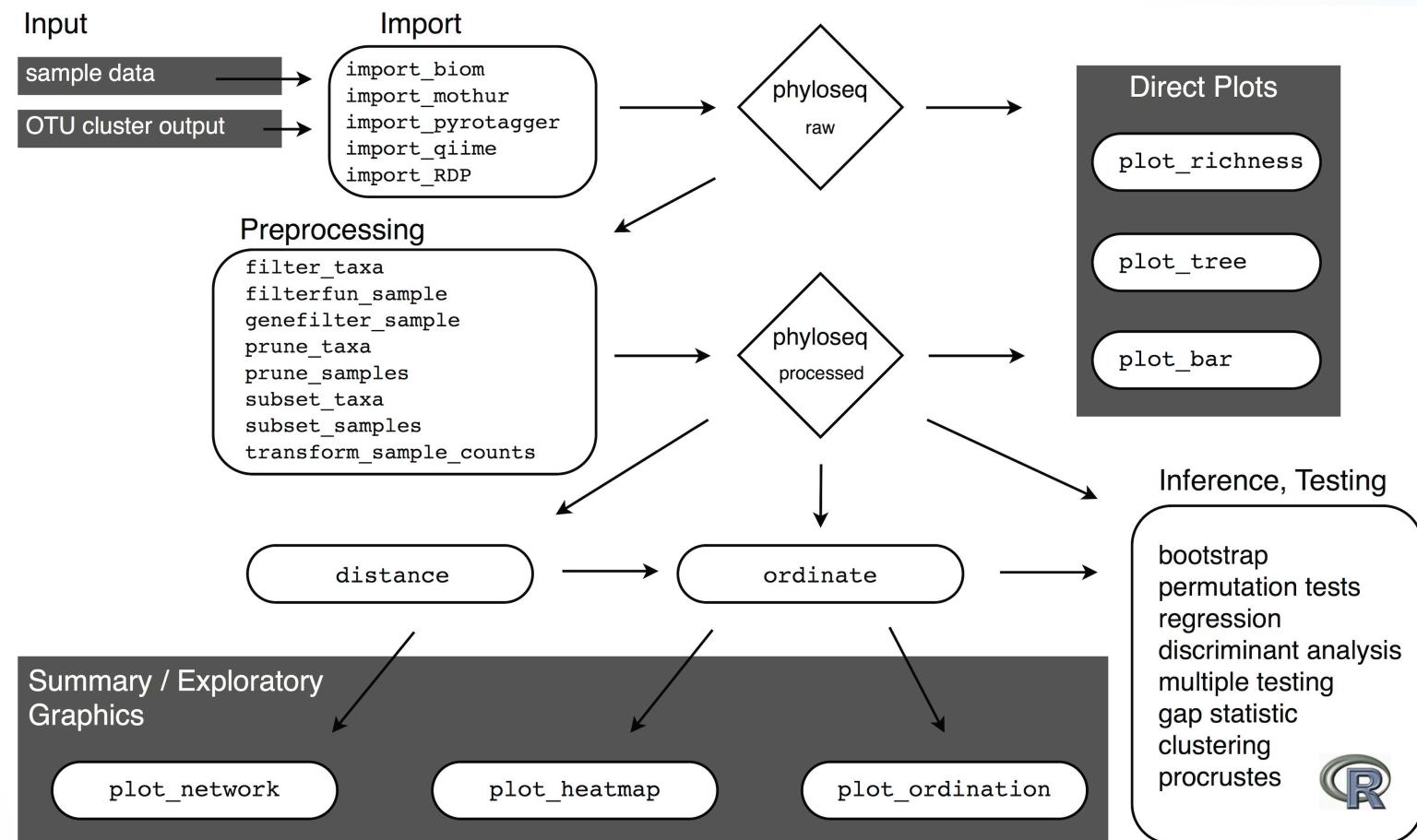
phyloseq in the experimental and analysis workflow





phyloseq analysis workflow

Need input data that has already been clustered into OTUs for amplicon data, or that has been taxonomically classified for metagenomic data



phyloseq - access and query the data in the phyloseq object

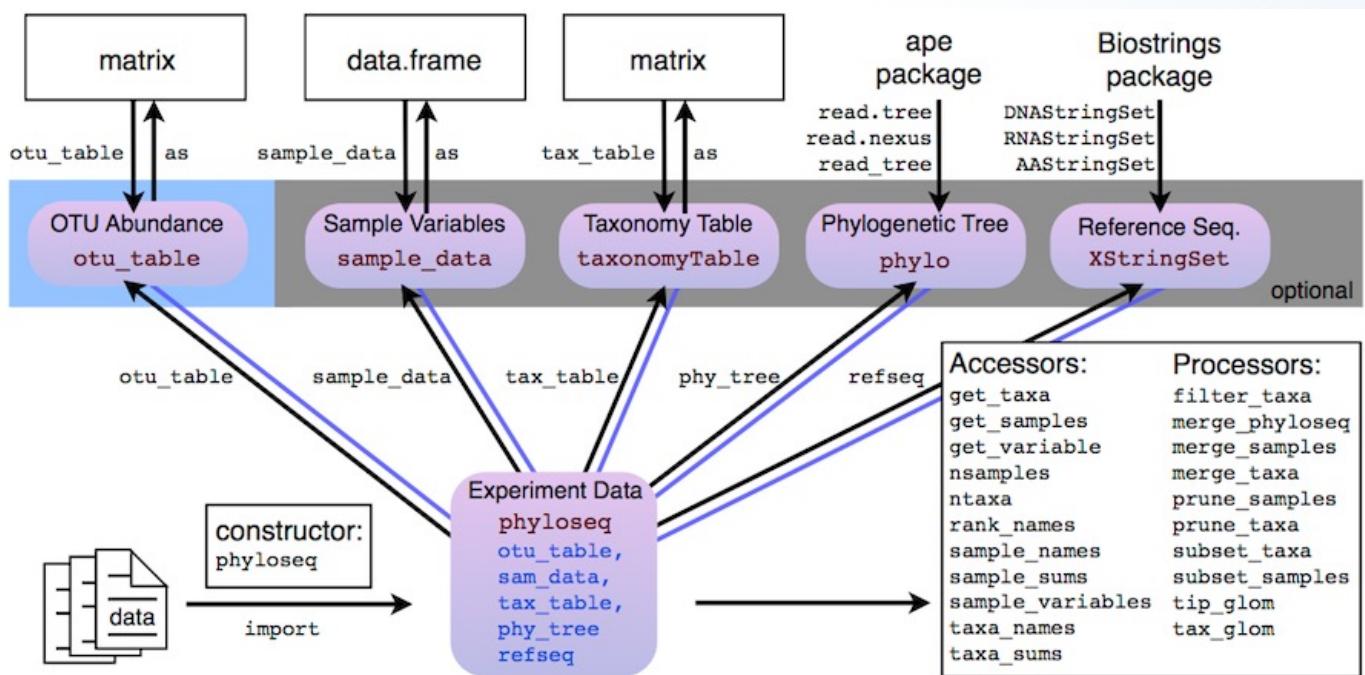


Accessors are functions to access components in a phyloseq object

Processors are functions to extracts parts of a phyloseq object

In addition are functions that performs analysis and/or graphics task

A matrix is a collection of data elements arranged in a two-dimensional rectangular layout. A matrix can only contain characters or only logical values.



Data frames are tabular data objects. Unlike a matrix in data frame each column can contain different modes of data. The first column can be numeric while the second column can be character



phyloseq – Some basic example after importing data

The main phyloseq object in this example is named phy

Accessors: How many taxa are there in my data (in phy)?

```
> ntaxa(phy)  
> 1000
```

Processors: Collaps all taxa to phylum level

```
> phy.phylum <- tax_glom(phy, taxrank="Phylum")  
> ntaxa(phy)  
> 10
```

Analysis functions: Calculate alpha diversity for all samples

```
<alpha.diversity <- estimate_richness(phy)  
> head (alpha.diversity)
```

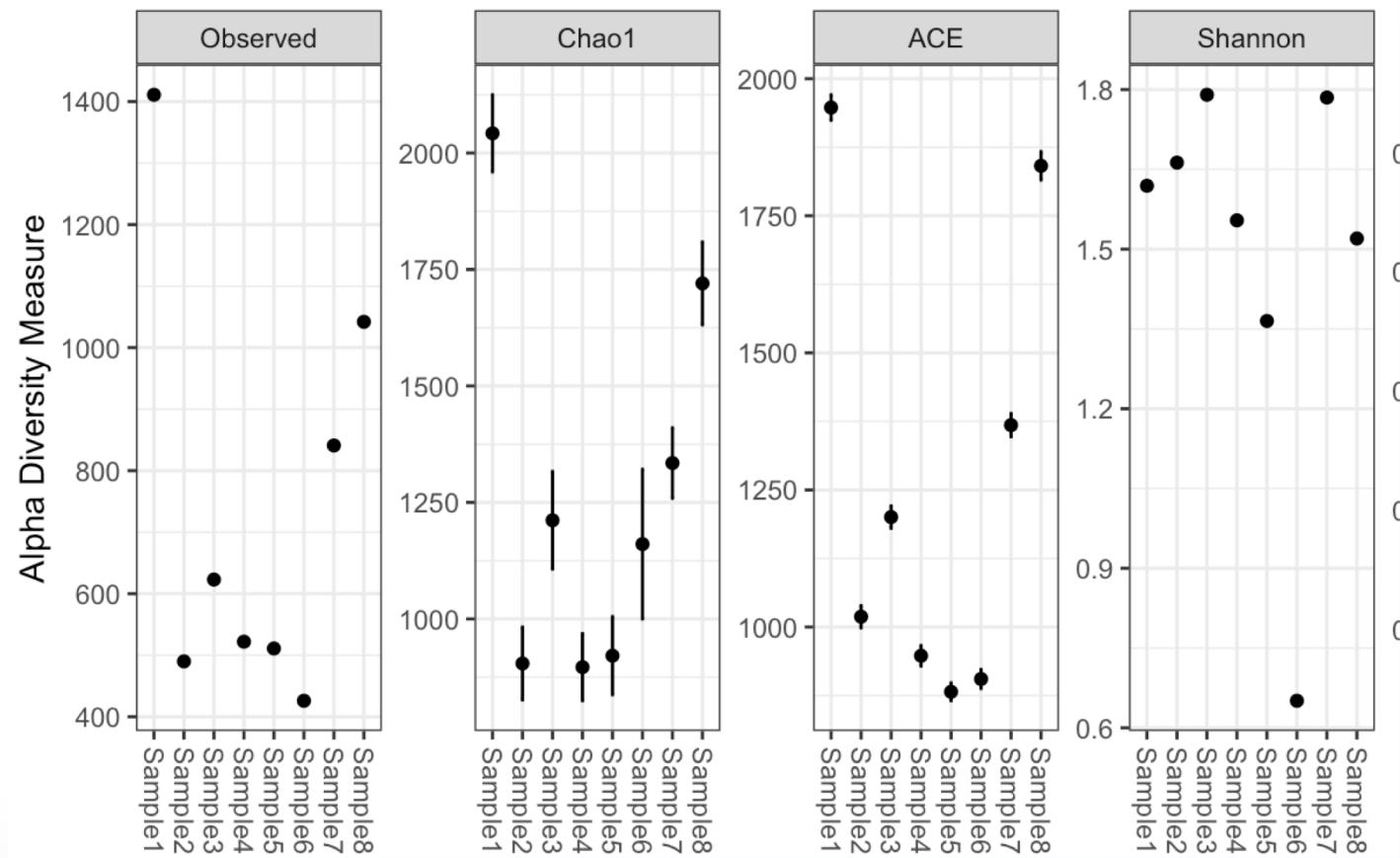
	Observed	Chao1	se.chao1	ACE	se.ACE	Shannon
Sample1	1411	2042.2418	82.76011	1947.4397	23.28101	1.6190928
Sample2	490	904.3810	78.16085	1018.8332	20.28168	1.6626745
Sample3	623	1211.6719	104.77347	1200.5020	20.27128	1.7901943
Sample4	522	896.5161	72.17895	947.5645	18.72522	1.5540733
Sample5	511	921.0600	83.87716	881.8277	16.36221	1.3649780
Sample6	426	1160.5882	160.67905	905.1891	17.48230	0.6509008



phyloseq – example of basic plot

Plot functions: Plot alpha diversity for all samples

```
> plot_richness(phy.prune)
```

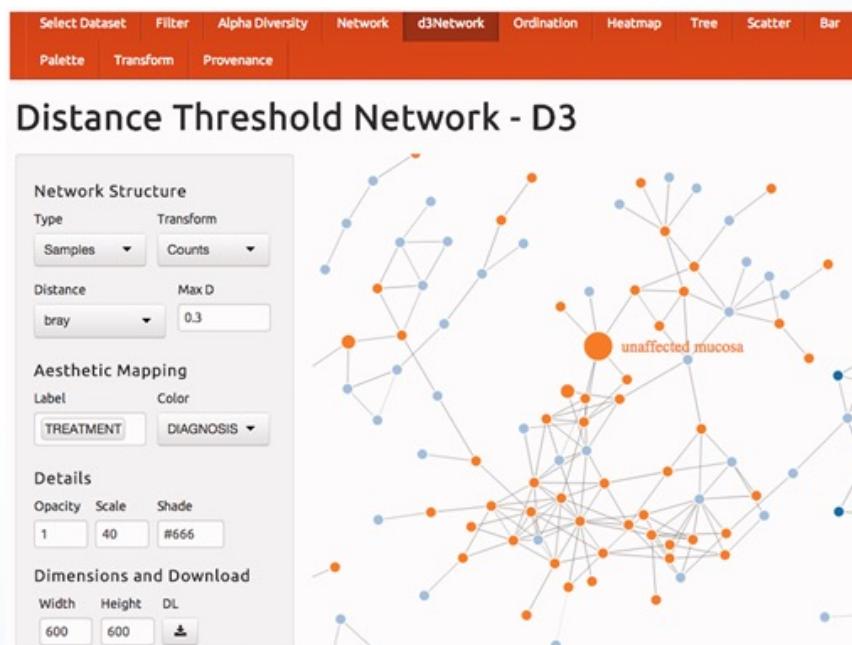


Shiny-phyloseq

Dynamic interaction with microbiome data

Runs on any modern Web browser

Requires no programming, increasing the accessibility and decreasing the entrance requirement to using phyloseq and related R tools





Phyloseq - Highlights

Import abundance and related data

Convenience analysis wrappers for common analysis tasks

44 supported distance methods (UniFrac, Jensen-Shannon, etc)

Ordination → many supported methods, including constrained methods

Microbiome plot functions using ggplot2 for powerful, flexible exploratory analysis

Modular, customisable preprocessing functions supporting fully reproducible work.

Functions for merging data based on taxa/sample variables, and for supporting manually-imported data.

Native R/C, parallelised implementation of UniFrac distance calculations.

Multiple testing methods specific to high-throughput amplicon sequencing data.

Examples for analysis and graphics using real published data.