Introduction to the Microbiome Package for R

By: Moaaz Moazzam and Tenzin Jordan Shawa

Installing the Microbiome Package

1. First install R.
2. Install an IDE such as RStudio
3. Furthermore, install RTools.
4. Install Bioconductor.
5. Then the Microbiome library can be installed.
6. Lastly, install the Phyloseq, vegan, dplyr, tidyverse, and MicrobeDS libraries.

You can do this with a text editor, or you can even do it from R like so:

```
writeLines('PATH="${RTOOLS40_HOME}\\usr\\bin;${PATH}"', con = "~/.Renviron")
```

Now restart R, and verify that `make` can be found, which should show the path to your Rtools installation.

```
Sys.which("make")
## "C:\\rtools40\\usr\\bin\\make.exe"
```

If this works, you can try to install an R package from source:

```
install.packages("jsonlite", type = "source")
```

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.12")
```

Installing microbiome R package

Open R and install the package. If the installation fails, ensure from the RStudio tools panel that you have access to the Bioconductor repository.

```
library(BiocManager)
BiocManager::install("microbiome")
```

Alternatively, to install the bleeding edge (potentially unstable) development version, run in R:

```
library(devtools) # Load the devtools package
install_github("microbiome/microbiome") # Install the package
```

Using the tools

Once the package has been installed, load it in R

```
library(microbiome)
```

Bioconductor Installer (Recommended)

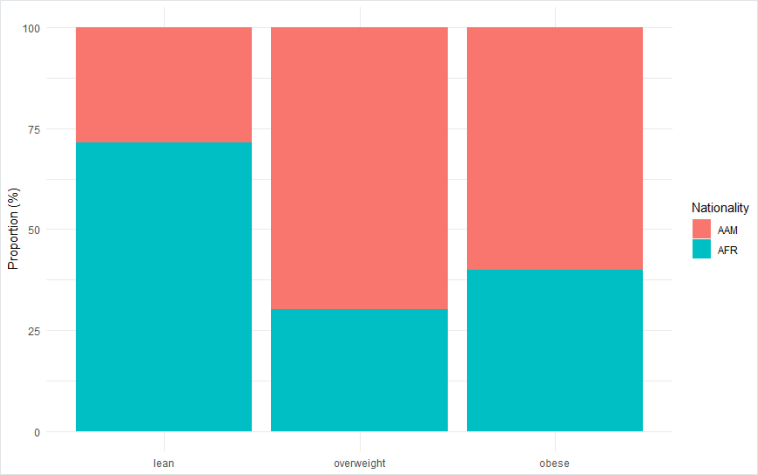
Recommended two lines for installing phyloseq (execute from within a fresh R session).

```
source('http://bioconductor.org/biocLite.R')
biocLite('phyloseq')
```

Installation	Install the latest version of this package by entering the following in R: `install.packages("remotes")` `remotes::install_github("twbattaglia/MicrobeDS")`

Brief Intro to the Microbiome Package

- Package provides tools and functions for the statistical analysis of metagenomic and microbial data.
  - •Alpha Diversity
  - •Beta Diversity
- Visualization of data is also a component of the package.

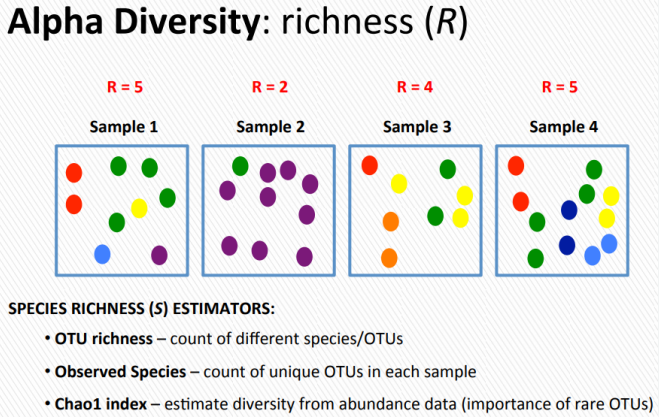


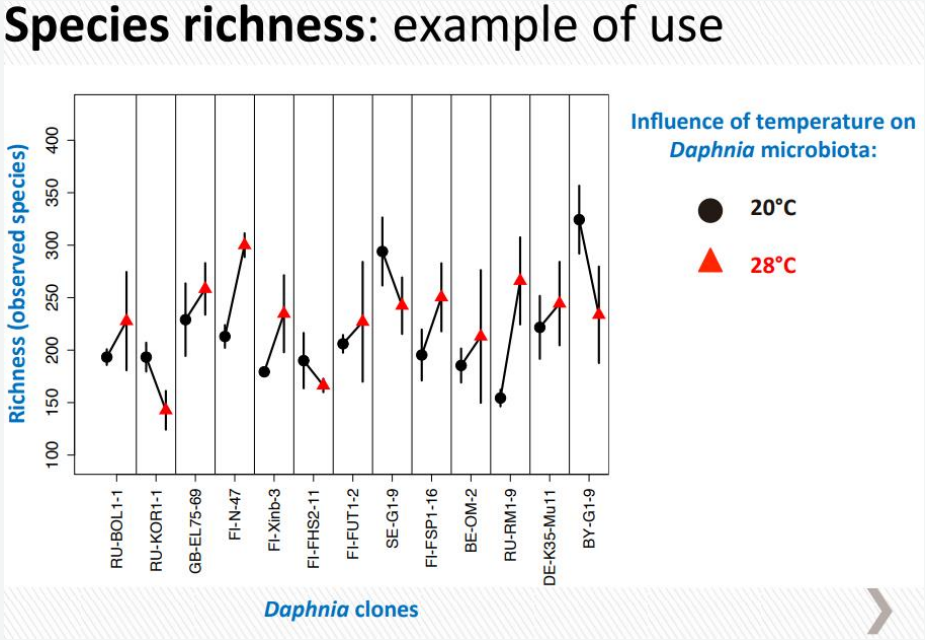
	diversity_shannon
Sample-1	3.187815
Sample-2	3.394462
Sample-3	2.864855
Sample-4	3.056922
Sample-5	3.073742
Sample-6	2.941993

Alpha Diversity

- In terms of microbiomes, Alpha Diversity can be seen as the measure of variation between microbes in a single sample set.
- Variation can be measured in several ways: Richness, Evenness, and Dominance
- Iterations of the Alpha value affect the diversity index, along with the variations between the species.
- "alpha()" function or "diversities()" function.

$$D_\alpha = \left(\sum_{i=1}^{s} p_i^\alpha \right)^{\frac{1}{1-\alpha}}.$$





Richness

- Measure of the different number of species within a sample.
- Measured in a variety of methods: Shannon-Weiner Index, Simpson Index, Chao1 Index, ACE.
- "estimate_richness()" function.

Richness Example

```
# import packages along with some example data sets

library(microbiome)
library(knitr)
data("atlas1006")
data("dietswap")
data("peerj32")

# ----------------------------------------
#assign variables to the data sets in phyloseq format
pseq <- atlas1006
pseq2 <- dietswap
pseq3 <- peerj32$phyloseq

#however the above function is pretty encompassing and confusing the handle with
#functions below are more specific ----------------------------------------------


estimate_richness(pseq, split = TRUE, measures = NULL) #gives estimator functions such as chao1 and ace
```

```
> estimate_richness(pseq, split = TRUE, measures = NULL) #gives estimator functions such as chao1 and ace
           Observed    Chao1  se.chao1      ACE   se.ACE  Shannon   Simpson InvSimpson   Fisher
Sample.1         99 107.0769  5.599727 110.9867 5.238297 3.187815 0.9229817  12.983930 16.07126
Sample.2         98 106.6667  6.413703 107.7171 5.012227 3.394462 0.9397430  16.595578 15.04043
Sample.3         99 108.5455  6.566400 108.9817 5.098539 2.864855 0.8851036   8.703493 16.26890
Sample.4        100 109.5455  6.566415 113.0700 5.139539 3.056922 0.9066208  10.709023 15.21763
Sample.5         98 109.6667  8.000223 110.5109 5.297317 3.073742 0.9183541  12.248008 14.59865
Sample.6         99 110.6667  8.000246 110.7823 5.262665 2.941993 0.8965359   9.665190 15.94374
Sample.7         99 107.0769  5.599727 110.0732 5.257411 2.594319 0.8286794   5.837009 16.88443
Sample.8         97 105.0769  5.599705 106.8910 5.082473 2.626788 0.8673823   7.540473 15.28406
Sample.9         96 100.0909  3.595256 100.8046 4.887528 2.689003 0.8516698   6.741716 15.76293
```

```
> sample_data(atlas1006)
Sample Data:        [1151 samples by 10 sample variables]:
         age    sex   nationality DNA_extraction_method project diversity   bmi_group subject time    sample
Sample-1   28   male            US                  <NA>       1      5.76 severeobese       1    0  Sample-1
Sample-2   24 female            US                  <NA>       1      6.06       obese       2    0  Sample-2
Sample-3   52   male            US                  <NA>       1      5.50        lean       3    0  Sample-3
Sample-4   22 female            US                  <NA>       1      5.87 underweight       4    0  Sample-4
Sample-5   25 female            US                  <NA>       1      5.89        lean       5    0  Sample-5
Sample-6   42   male            US                  <NA>       1      5.53        lean       6    0  Sample-6
Sample-7   25 female            US                  <NA>       1      5.49 underweight       7    0  Sample-7
Sample-8   27 female            US                  <NA>       1      5.38        lean       8    0  Sample-8
Sample-9   21 female            US                  <NA>       1      5.34        lean       9    0  Sample-9
Sample-10  25 female            US                  <NA>       1      5.64        lean      10    0 Sample-10
```

Chao1 Measure

- Chao1 is an estimator that calculates the estimate of richness of a given set of samples.
- Can use the "estimate_richness" function.
- Higher Chao1 values = higher estimated richness.

$$S_{chao1} = S_{obs} + \frac{n_1 (n_1 - 1)}{2 (n_2 + 1)}$$

where,

S_{chao1} = the estimated richness

S_{obs} = the observed number of species

n_1 = the number of OTUs with only one sequence (i.e. “singletons”)

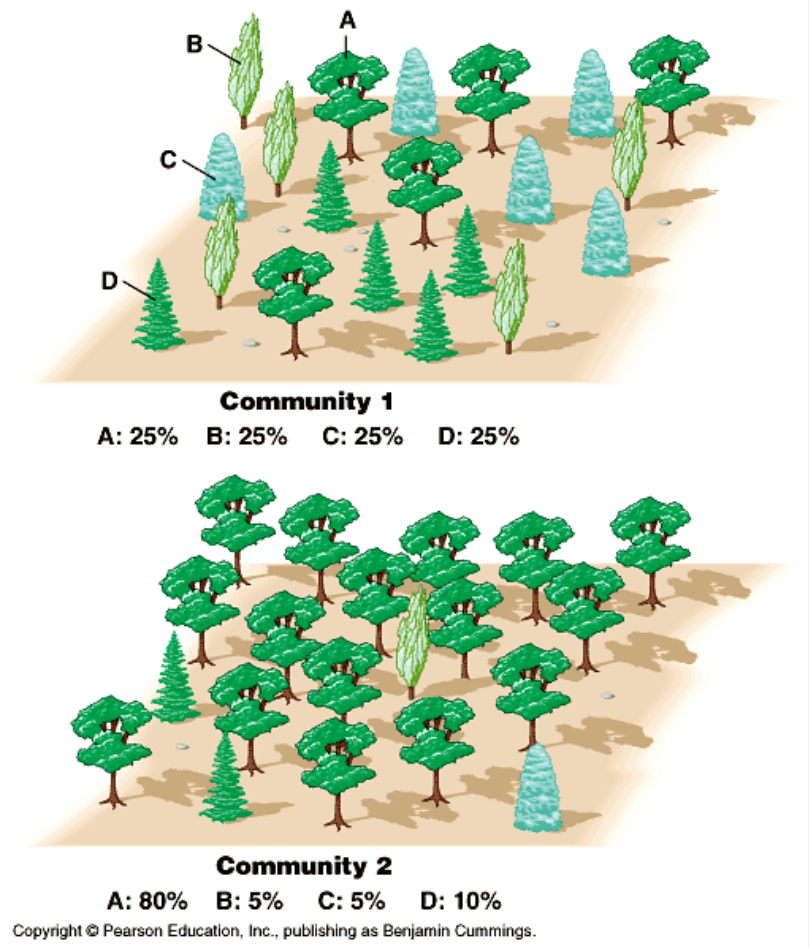
n_2 = the number of OTUs with only two sequences (i.e. “doubletons”)

Evenness

- Measure of equivalence in amount of each species in a given sample.
- Evenness can be calculated and measured using the Pielou's Evenness Index.
- "Evenness()" Function.

$$H'_{\max} = -\sum_{i=1}^{S} \frac{1}{S} \ln \frac{1}{S} = \ln S.$$

$$J' = \frac{H'}{H'_{\max}}$$



Evenness Example

```
> evenness(pseq2,index = "all", zeroes = TRUE, detection = 0)
              camargo    pielou    simpson      evar     bulla
Sample-1   0.20140360 0.6331719 0.07270887 0.1709714 0.3298916
Sample-2   0.22619920 0.6004646 0.07366312 0.1334372 0.2755652
Sample-3   0.21091757 0.5195476 0.04166102 0.1373098 0.2541325
Sample-4   0.34223085 0.6429969 0.07553142 0.1711334 0.3176017
Sample-5   0.14078174 0.4544002 0.03078386 0.1661081 0.2310307
Sample-6   0.17092214 0.4450266 0.02813098 0.1625409 0.2308364
Sample-7   0.33542227 0.6482503 0.11202315 0.1499344 0.2924694
Sample-8   0.14102222 0.4678982 0.03185407 0.1501392 0.2384971
Sample-9   0.38893289 0.7222761 0.17155753 0.1841683 0.3491324
Sample-10  0.22828737 0.6091818 0.07883334 0.1484012 0.2806217
```

```
Sample Data:        [222 samples by 8 sample variables]:
           subject    sex nationality group     sample timepoint timepoint.within.group  bmi_group
Sample-1       byn   male         AAM    DI   Sample-1         4                     1      obese
Sample-2       nms   male         AFR    HE   Sample-2         2                     1       lean
Sample-3       olt   male         AFR    HE   Sample-3         2                     1 overweight
Sample-4       pku female         AFR    HE   Sample-4         2                     1      obese
Sample-5       qjy female         AFR    HE   Sample-5         2                     1 overweight
Sample-6       riv female         AFR    HE   Sample-6         2                     1      obese
Sample-7       shj female         AFR    HE   Sample-7         2                     1      obese
Sample-8       tgx   male         AFR    HE   Sample-8         2                     1 overweight
Sample-9       ufm   male         AFR    HE   Sample-9         2                     1       lean
Sample-10      nms   male         AFR    HE  Sample-10         3                     2       lean
```

```
                                      Sample-158
Actinomycetaceae                               0
Aerococcus                                     0
Aeromonas                                      0
Akkermansia                                   90
Alcaligenes faecalis et rel.                  23
Allistipes et rel.                           327
Anaerobiospirillum                             0
Anaerofustis                                   0
Anaerostipes caccae et rel.                   74
Anaerotruncus colihominis et rel.             68
Anaerovorax odorimutans et rel.               20
Aneurinibacillus                               0
Aquabacterium                                  0
Asteroleplasma et rel.                         0
Atopobium                                      1
Bacillus                                       1
Bacteroides fragilis et rel.                 546
Bacteroides intestinalis et rel.              37
Bacteroides ovatus et rel.                   443
Bacteroides plebeius et rel.                  66
Bacteroides splachnicus et rel.               44
Bacteroides stercoris et rel.                 38
Bacteroides uniformis et rel.                294
Bacteroides vulgatus et rel.                1157
Bifidobacterium                              283
Bilophila et rel.                              2
Brachyspira                                    0
Bryantella formatexigens et rel.             152
Bulleidia moorei et rel.                       2
Burkholderia                                   0
Butyrivibrio crossotus et rel.               139
Campylobacter                                  4
Catenibacterium mitsuokai et rel.              0
Clostridium (sensu stricto)                   17
Clostridium cellulosi et rel.                400
Clostridium colinum et rel.                   26
Clostridium difficile et rel.                 12
Clostridium felsineum et rel.                  0
```

Dominance

- Measure of how much a specific species dominates in count versus others in a sample.
- Can be measured using the Simpson and Shannon Indexes.
- Negatively, affects diversity.
- "dominance()" or "dominant()" function.

Dominance Example

```
> otu_table(pseq3)
OTU Table:          [130 taxa and 44 samples]
                     taxa are rows
```

	sample-1	sample-2	sample-3	sample-4	sample-5	sample-6	sample-7	sample-8	sample-9	sample-10	sample-11	sample-12
Actinomycetaceae	0	2	9	4	0	3	10	24	9	19	0	0
Aerococcus	6	6	2	0	13	0	13	21	9	0	6	0
Aeromonas	0	16	11	18	0	0	7	0	10	11	7	8
Akkermansia	224	186	224	179	241	416	357	276	804	334	714	650
Alcaligenes faecalis et rel.	0	0	12	0	0	0	0	0	0	0	0	7
Allistipes et rel.	169	107	194	216	277	312	99	123	44	86	189	151
Anaerobiospirillum	0	3	0	8	0	0	2	0	0	1	0	20
Anaerofustis	20	30	20	24	20	24	30	32	46	39	28	37
Anaerostipes caccae et rel.	360	619	385	312	529	505	490	400	802	904	596	337
Anaerotruncus colihominis et rel.	10	54	64	88	42	49	154	152	35	29	91	115
Anaerovorax odorimutans et rel.	64	69	74	107	82	166	101	113	174	69	159	266
Aneurinibacillus	8	7	8	7	4	4	15	17	35	25	26	15
Aquabacterium	45	77	89	79	17	32	70	54	88	295	17	53
Asteroleplasma et rel.	0	0	0	0	0	0	0	5	0	9	0	0
Atopobium	6	6	66	51	1	3	11	8	0	0	16	24
Bacillus	7	14	7	12	14	9	10	20	35	23	11	14
Bacteroides fragilis et rel.	655	131	180	133	215	113	29	64	339	710	126	226
Bacteroides intestinalis et rel.	948	515	280	375	293	123	462	1263	174	185	146	208
Bacteroides ovatus et rel.	494	141	126	95	156	47	24	46	208	246	122	149
Bacteroides plebeius et rel.	184	91	144	144	147	107	89	104	122	134	121	142
Bacteroides splachnicus et rel.	186	100	69	66	94	69	30	64	58	98	51	74
Bacteroides stercoris et rel.	116	33	47	38	140	48	12	33	105	101	55	61
Bacteroides uniformis et rel.	2459	571	549	801	301	192	93	135	705	959	301	463
Bacteroides vulgatus et rel.	4682	1423	1895	1562	1850	737	563	3105	260	250	890	822

```
> dominant(pseq3,level = NULL)
 [1] "Bacteroides vulgatus et rel."         "Faecalibacterium prausnitzii et rel." "Faecalibacterium prausnitzii et rel."
 [4] "Faecalibacterium prausnitzii et rel." "Faecalibacterium prausnitzii et rel." "Ruminococcus obeum et rel."
 [7] "Faecalibacterium prausnitzii et rel." "Bacteroides vulgatus et rel."         "Ruminococcus obeum et rel."
[10] "Ruminococcus obeum et rel."           "Faecalibacterium prausnitzii et rel." "Ruminococcus bromii et rel."
```

Species	Number (n)	n(n-1)
Sea holly	2	2
Sand couch	8	56
Sea bindweed	1	0
Sporobolus pungens	1	0
Echinophora spinosa	3	6
Total	**15**	**64**
	N = 15	**$\sum$ n(n-1) = 64**

Putting the figures into the formula for Simpson's Index:

$$D = 1 - \left(\frac{\Sigma\, n(n-1)}{N(N-1)} \right)$$

$$D = 1 - \left(\frac{64}{15(14)} \right)$$

Simpson's Index of Diversity = 0.7

Simpson Index

- An index that, considering richness and evenness, measures the diversity of the sample set by being given the number of species along with count of each species.
- Often used in its inverse or Gini-Simpson form.
- The larger (near 1) the measured value = the more diverse the sample (Gini-Simpson).

$$D = 1 - \left(\frac{\Sigma\, n(n-1)}{N(N-1)} \right)$$

Simpson Index Example

	Sample-158	Sa
Actinomycetaceae	0	
Aerococcus	0	
Aeromonas	0	
Akkermansia	90	
Alcaligenes faecalis et rel.	23	
Allistipes et rel.	327	
Anaerobiospirillum	0	
Anaerofustis	0	
Anaerostipes caccae et rel.	74	
Anaerotruncus colihominis et rel.	68	
Anaerovorax odorimutans et rel.	20	
Aneurinibacillus	0	
Aquabacterium	0	
Asteroleplasma et rel.	0	
Atopobium	1	
Bacillus	1	
Bacteroides fragilis et rel.	546	
Bacteroides intestinalis et rel.	37	
Bacteroides ovatus et rel.	443	
Bacteroides plebeius et rel.	66	
Bacteroides splachnicus et rel.	44	
Bacteroides stercoris et rel.	38	
Bacteroides uniformis et rel.	294	
Bacteroides vulgatus et rel.	1157	
Bifidobacterium	283	
Bilophila et rel.	2	
Brachyspira	0	
Bryantella formatexigens et rel.	152	
Bulleidia moorei et rel.	2	
Burkholderia	0	
Butyrivibrio crossotus et rel.	139	
Campylobacter	4	
Catenibacterium mitsuokai et rel.	0	
Clostridium (sensu stricto)	17	
Clostridium cellulosi et rel.	400	
Clostridium colinum et rel.	26	
Clostridium difficile et rel.	12	
Clostridium felsineum et rel.	0	
Clostridium leptum et rel.	181	
Clostridium nexile et rel.	33	
Clostridium orbiscindens et rel.	855	
Clostridium ramosum et rel.	4	
Clostridium sphenoides et rel.	83	
Clostridium stercorarium et rel.	68	
Clostridium symbiosum et rel.	1560	
Clostridium thermocellum et rel.	0	
Collinsella	35	
Coprobacillus catenaformis et rel.	2	
Coprococcus eutactus et rel.	107	
Corynebacterium	1	
Desulfovibrio et rel.	11	

```
> alpha(pseq2, index = "inverse_simpson", zeroes = TRUE)
Observed richness
Other forms of richness
Diversity
Evenness
Dominance
Rarity
           diversity_inverse_simpson
Sample-1                    7.561722
Sample-2                    8.102943
Sample-3                    4.291085
Sample-4                    7.930799
Sample-5                    3.170738
Sample-6                    2.953753
Sample-7                   11.650407
Sample-8                    3.280969
Sample-9                   17.327310
Sample-10                   8.198667
```

```
> alpha(pseq2, index = "gini_simpson", zeroes = TRUE)
Observed richness
Other forms of richness
Diversity
Evenness
Dominance
Rarity
           diversity_gini_simpson
Sample-1                0.8677550
Sample-2                0.8765881
Sample-3                0.7669587
Sample-4                0.8739093
Sample-5                0.6846160
Sample-6                0.6614476
Sample-7                0.9141661
Sample-8                0.6952120
Sample-9                0.9422876
Sample-10               0.8780290
```

Shannon-Weiner Index

- An index that, similar to Simpson, measures the diversity of a given sample set by using the number of individual species along with the counts of each species.
- Higher value implies a greater diversity of a given sample set.

$$H' = -\sum_{i=1}^{R} p_i \ln p_i = -\sum_{i=1}^{R} \ln p_i^{p_i}$$

Step 1:

First, let us calculate the sum of the given values.

sum	= (60+10+25+1+4)
	= 100

Step 2:

No of sample	pi=sample/sum	ln(pi)	pi*ln(pi)
60	0.60	-0.51	-0.31
10	0.10	-2.30	-0.23
25	0.25	-1.39	-0.35
1	0.01	-4.61	-0.05
4	0.04	-3.22	-0.13
sum=100			SUM = -1.07

H=1.07

Step 3:

H_{max}=ln(N)=ln(5)=1.61

Evenness=H/H_{max}=1.07/1.61=0.66

Shannon Index Example

```
> alpha(pseq3, index = "shannon", zeroes = TRUE)
Observed richness
Other forms of richness
Diversity
Evenness
Dominance
Rarity
          diversity_shannon
sample-1           3.478264
sample-2           3.672656
sample-3           3.610991
sample-4           3.659335
sample-5           3.577881
sample-6           3.641771
sample-7           3.679855
sample-8           3.689226
sample-9           3.680276
sample-10          3.772017
```

	sample-1	sample-2	sample-3	sample-4	sample-5	sample-6	sample-7
Actinomycetaceae					0	2	9
Aerococcus					6	6	2
Aeromonas					0	16	11
Akkermansia					224	186	224
Alcaligenes faecalis et rel.					0	0	12
Allistipes et rel.					169	107	194
Anaerobiospirillum					0	3	0
Anaerofustis					20	30	20
Anaerostipes caccae et rel.					360	619	385
Anaerotruncus colihominis et rel.					10	54	64
Anaerovorax odorimutans et rel.					64	69	74
Aneurinibacillus					8	7	8
Aquabacterium					45	77	89
Asteroleplasma et rel.					0	0	0
Atopobium					6	6	66
Bacillus					7	14	7
Bacteroides fragilis et rel.					655	131	180
Bacteroides intestinalis et rel.					948	515	280
Bacteroides ovatus et rel.					494	141	126
Bacteroides plebeius et rel.					184	91	144
Bacteroides splachnicus et rel.					186	100	69
Bacteroides stercoris et rel.					116	33	47
Bacteroides uniformis et rel.					2459	571	549
Bacteroides vulgatus et rel.					4682	1423	1895
Bifidobacterium					274	377	2229
Bilophila et rel.					4	19	20
Brachyspira					0	0	0
Bryantella formatexigens et rel.					347	570	355
Bulleidia moorei et rel.					8	17	4
Burkholderia					62	58	43
Butyrivibrio crossotus et rel.					235	191	151
Campylobacter					0	0	0
Catenibacterium mitsuokai et rel.					19	10	11
Clostridium (sensu stricto)					0	2	42
Clostridium cellulosi et rel.					140	148	182
Clostridium colinum et rel.					133	167	20
Clostridium difficile et rel.					25	53	27
Clostridium felsineum et rel.					0	2	30
Clostridium leptum et rel.					231	248	318
Clostridium nexile et rel.					501	548	745
Clostridium orbiscindens et rel.					255	203	121
Clostridium ramosum et rel.					75	72	35
Clostridium sphenoides et rel.					613	687	421
Clostridium stercorarium et rel.					41	75	69
Clostridium symbiosum et rel.					323	458	405
Clostridium thermocellum et rel.					0	2	8
Collinsella					89	157	217
Coprobacillus catenaformis et rel.					99	118	46
Coprococcus eutactus et rel.					498	750	1013
Corynebacterium					0	1	7
Desulfovibrio et rel.					0	12	27
Dialister					17	48	99

Beta Diversity

- Beta Diversity attempts to showcase the differences of species amongst different sample sets (environments).
- Measures dissimilarity/divergence which can be measured in several methods.
- Bray-Curtis, Jaccard Distance, UniFrac

$$\beta = \gamma/\alpha$$

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

Where:

- 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.

Simple Example

For a simple example, consider two aquariums;

- **Tank one**: 6 goldfish, 7 guppies and 4 rainbow fish,
- **Tank two:** 10 goldfish and 6 rainbow fish.

To calculate Bray-Curtis, let's first calculate C_{ij} (the sum of only the lesser counts for each species found in both sites). Goldfish are found on both sites; the lesser count is 6. Guppies are only on one site, so they can't be added in here. Rainbow fish, though, are on both, and the lesser count is 4.
So C_{ij} = 6 + 4 = 10.

S_i (total number of specimens counted on site i) = 6 + 7 + 4 = 17, and
S_j (total number of specimens counted on site j) = 10 + 6 = 16.

So our BC_{ij} = 1 – (2 * 10) / (17 + 16), or 0.39.

Bray-Curtis Method

- A dissimilarity measuring method that compares two sample sites by comparing the differences in abundances between the two.
- 0 = being the most similar, while 1 = being the least similar.

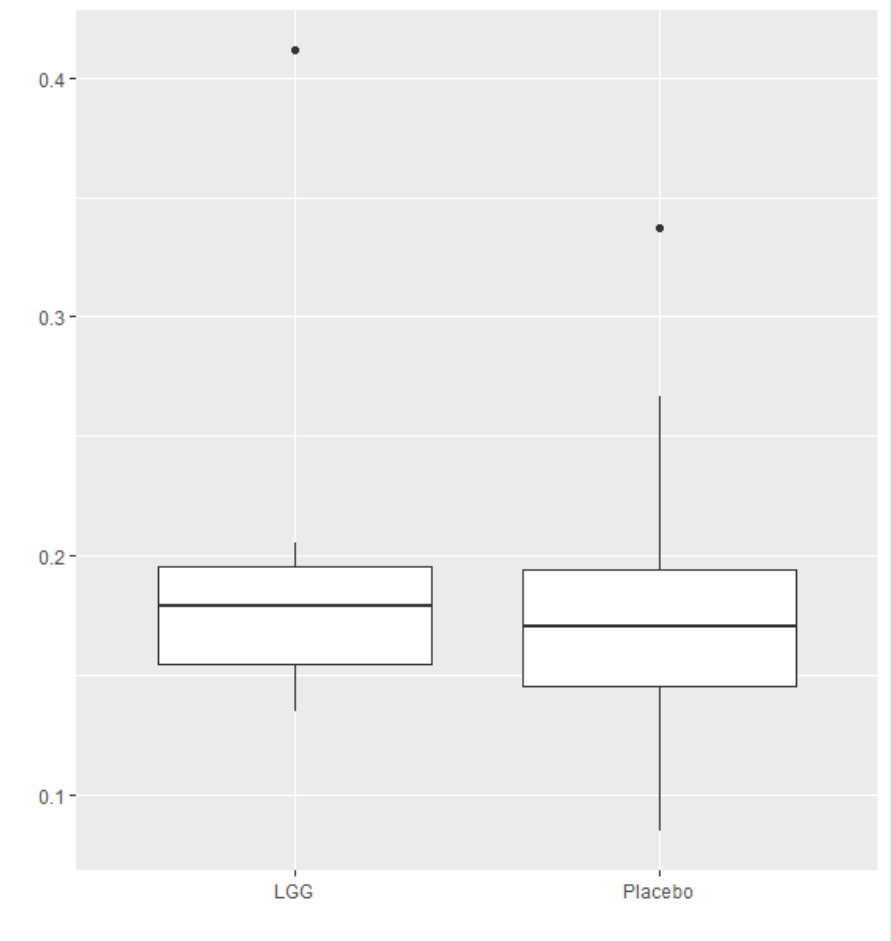
Bray-Curtis Example

```
> sample_data(pseq)
Sample Data:        [1151 samples by 10 sample variables]:
          age    sex   nationality DNA_extraction_method project diversity   bmi_group subject time    sample
Sample-1   28   male            US                   <NA>       1      5.76 severeobese       1    0  Sample-1
Sample-2   24 female            US                   <NA>       1      6.06       obese       2    0  Sample-2
Sample-3   52   male            US                   <NA>       1      5.50        lean       3    0  Sample-3
Sample-4   22 female            US                   <NA>       1      5.87 underweight       4    0  Sample-4
Sample-5   25 female            US                   <NA>       1      5.89        lean       5    0  Sample-5
Sample-6   42   male            US                   <NA>       1      5.53        lean       6    0  Sample-6
Sample-7   25 female            US                   <NA>       1      5.49 underweight       7    0  Sample-7
Sample-8   27 female            US                   <NA>       1      5.38        lean       8    0  Sample-8
Sample-9   21 female            US                   <NA>       1      5.34        lean       9    0  Sample-9
Sample-10  25 female            US                   <NA>       1      5.64        lean      10    0 Sample-10
```

```
> #Simple Divergence Function
> #compares a subset of samples within the atlas1006 study. primarily adults that are lean are compared to the
> #whole sample set for dissimilarity based on different species abundances (0 = 100% similar, 1 = 0% similar)
>
> pseqsub <- subset_samples(atlas1006, bmi_group == "lean")
> ref <- apply(abundances(pseqsub), 1, median)
> divSUB <- divergence(pseq, ref, method = "bray")
> print("lean group divergence")
[1] "lean group divergence"
> print(divSUB)
  Sample-1   Sample-2   Sample-3   Sample-4   Sample-5   Sample-6   Sample-7   Sample-8   Sample-9  Sample-10
 0.3571828  0.2743132  0.5320902  0.3945900  0.3869019  0.5015391  0.5342835  0.6168473  0.5472196  0.4281898
```

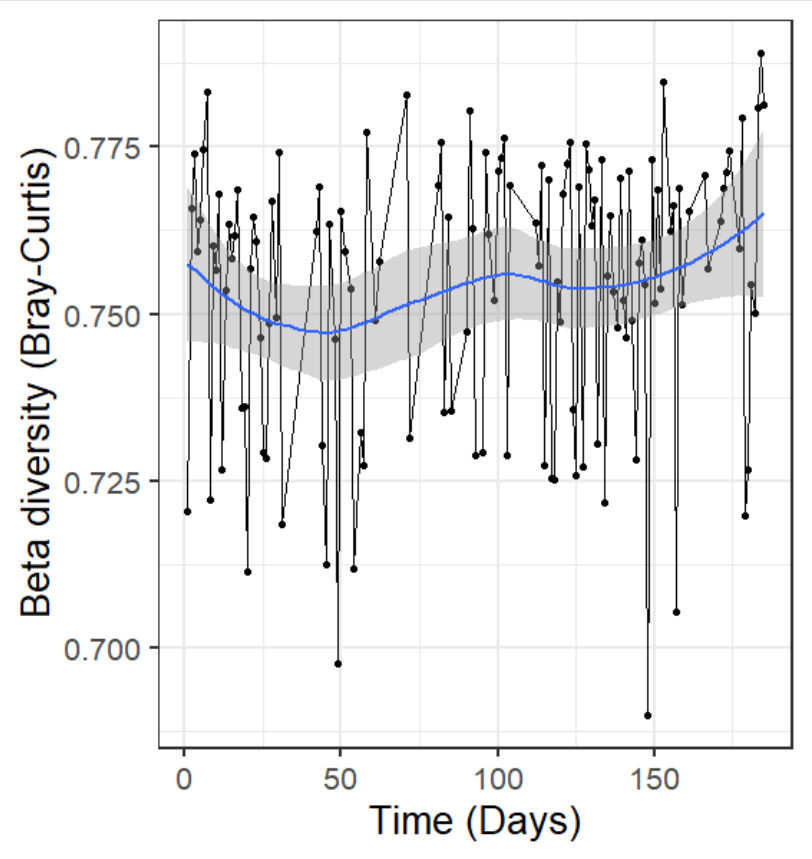
More Advanced Bray Curtis Example

```r
#-------------------------------------------
#more complex use cases for divergence.

betas <- list()
groups <- as.character(unique(meta(pseq3)$group))
for (g in groups) {

  df <- subset(meta(pseq3), group == g)
  beta <- c()

  for (subj in df$subject) {
    # Pick the samples for this subject
    dfs <- subset(df, subject == subj)
    # Check that the subject has two time points
    if (nrow(dfs) == 2) {
      s <- as.character(dfs$sample)
      # Here with just two samples we can calculate the
      # beta diversity directly
      beta[[subj]] <- divergence(abundances(pseq3)[, s[[1]]],abundances(pseq3)[, s[[2]]],method = "bray")
    }
  }
  betas[[g]] <- beta
}
# boxplot
df <- as.data.frame(unlist(betas))
s<- rownames(df)
si<- as.data.frame(s)
si<- separate(si, s, into = c('names','s'))
df1<- bind_cols(df, si)
rownames(df1)<- df1$s ; df1$s<- NULL

p<- ggplot(df1, aes(x = names, y = `unlist(betas)`))+ geom_boxplot() + ylab('') + xlab('')

plot(p)
```



Change in Divergence over Time

```
## Divergence experienced to a single subject may chance over time (increase). The blue best fit
# curve demonstrates the change in divergence as the days progress. Some days have high similarity = 49
# others not so much like day ~180

s <- "F4" # Selected subject
b <- "UBERON:feces" # Selected body site

# Let us pick a subset
pseq4 <- subset_samples(MovingPictures, host_subject_id == s & body_site == b)

# Rename variables
sample_data(pseq4)$subject <- sample_data(pseq4)$host_subject_id
sample_data(pseq4)$sample <- sample_data(pseq4)$X.SampleID

# Tidy up the time point information (convert from dates to days)
sample_data(pseq4)$time <- as.numeric(as.Date(gsub(" 0:00", "", as.character(sample_data(pseq4)$collection_timestamp)), "%m/%d/%Y") - as.Date("10/21/08", "%m/%d/%Y"))

# Order the entries by time
df <- meta(pseq4) %>% arrange(time)

# Calculate the beta diversity between each time point and
# the baseline (first) time point
beta <- c() # Baseline similarity
s0 <- subset(df, time == 0)$sample
# Let us transform to relative abundance for Bray-Curtis calculations
a <-microbiome::abundances(microbiome::transform(pseq4, "compositional"))
for (tp in df$time[-1]) {
  # Pick the samples for this subject
  # If the same time point has more than one sample,
  # pick one at random
  st <- sample(subset(df, time == tp)$sample, 1)
  # Beta diversity between the current time point and baseline
  b <- vegdist(rbind(a[, s0], a[, st]), method = "bray")
  # Add to the list
  beta <- rbind(beta, c(tp, b))
}
colnames(beta) <- c("time", "beta")
beta <- as.data.frame(beta)

theme_set(theme_bw(20))
library(ggplot2)
p <- ggplot(beta, aes(x = time, y = beta)) +
  geom_point() +
  geom_line() +
  geom_smooth() +
  labs(x = "Time (Days)", y = "Beta diversity (Bray-Curtis)")
print(p)
```



Jaccard Distance Example

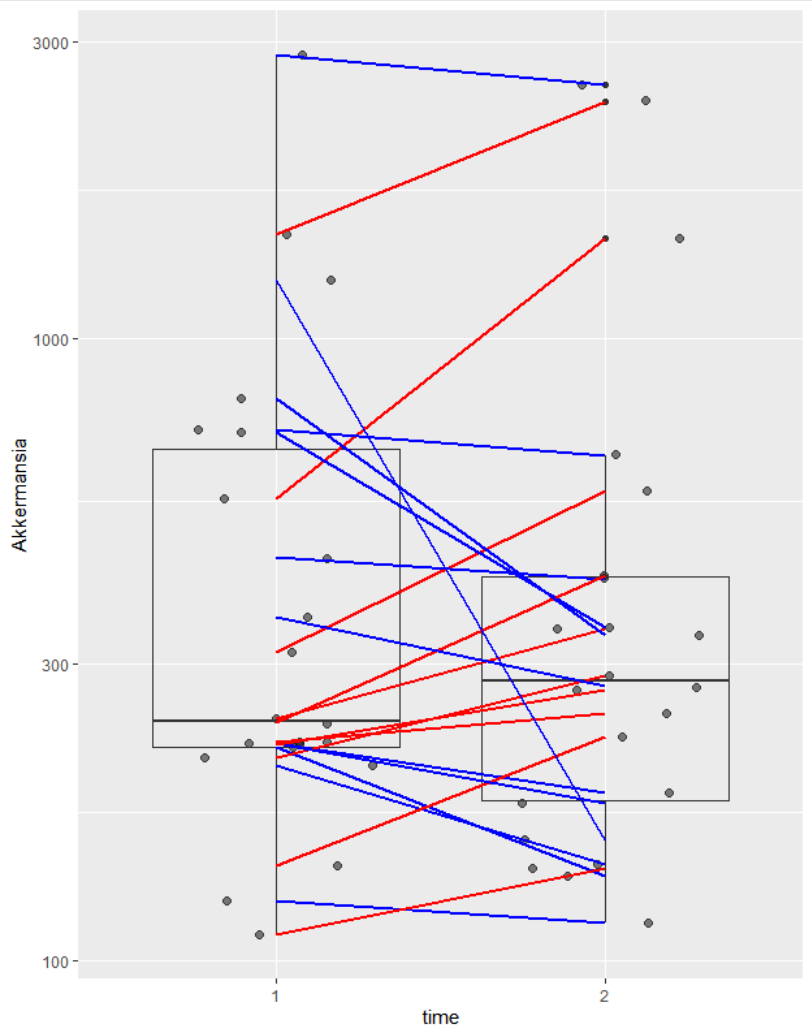
- Jaccard Method measures the dissimilarity of all the species contain in two sample sets (environments).
- 0 = both sets have the same set of species, 1 = both sets have no species in common.

```
> distcalc <- distance(pseq3, "jaccard", binary = TRUE)
> print("jaccard distance function")
[1] "jaccard distance function"
> print(distcalc)
           sample-1   sample-2   sample-3   sample-4   sample-5   sample-6   sample-7   sample-8   sample-9  sample-10  sample-11  sample-12
sample-2  0.19642857
sample-3  0.22689076 0.09322034
sample-4  0.20535714 0.09734513 0.16393443
sample-5  0.10101010 0.21238938 0.24166667 0.22123894
sample-6  0.19444444 0.18260870 0.18333333 0.17543860 0.14285714
sample-7  0.14018692 0.16379310 0.18032787 0.18803419 0.19090909 0.20869565
sample-8  0.18750000 0.09649123 0.10084034 0.15384615 0.20353982 0.17391304 0.15517241
sample-9  0.21428571 0.13913043 0.14166667 0.16379310 0.24561404 0.21551724 0.14912281 0.11403509
sample-10 0.26666667 0.13445378 0.13709677 0.11111111 0.28099174 0.20833333 0.20491803 0.11016949 0.10344828
sample-11 0.16346154 0.18584071 0.20168067 0.21052632 0.14563107 0.16666667 0.18018018 0.17699115 0.15454545 0.21186441
sample-12 0.24166667 0.14049587 0.12800000 0.14876033 0.22689076 0.18333333 0.19512195 0.14754098 0.14166667 0.10655738 0.18644068
```

Univariate Comparisons

Univariate data involves statistics that only have observations of a single type of data

In Microbiome R there are functions that allows one to make boxplots for the abundance measure as well as testing the paired comparison for a single taxonomic group with a random subject effect



Abundance boxplot

Using the command boxplot_abundance one can get a simple boxplot of the species abundance with time vs akkermansia

LME4

Linear Mixed-Effects Models using 'Eigen' and S4, or LME4 for short, can be used for linear model comparison to test an individual taxonomic group with a random subject effect

```
out <- lmer(signal ~ group + (1|subject), data = dfs)

out0 <- lmer(signal ~ (1|subject), data = dfs)

comp <- anova(out0, out)

pv <- comp[["Pr(>Chisq)"]][[2]]

print(pv)

[1] 0.4556962
```

Multivariate Comparisons

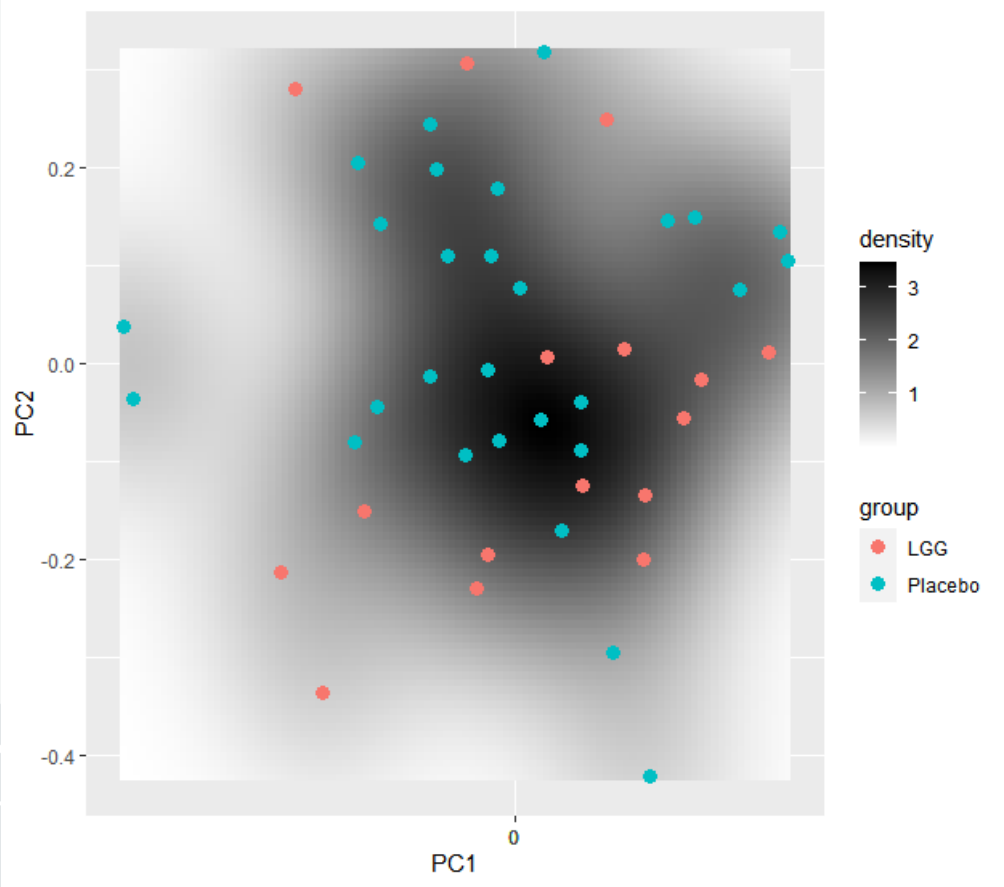
Multivariate comparisons involve analysis of two or more variables

The Microbiome package for R also has a function to do Permutational Multivariate Analysis of Variance, PERMANOVA for short

PERMANOVA is used to do geometric partitioning of multivariate variation within a given space with a dissimarily measure, such as counts of abundance

Visualizing Microbiome Variation

Using the plot landscape function, the population density of both a probiotic treament LGC and a placebo can be compared



PERMANOVA

The permanova and adonis commands PERMANOVA can be conducted on a data set

```
permanova <- adonis(t(otu) ~ group, data = meta,
permutations=99, method = "bray")
print(as.data.frame(permanova$aov.tab)["group", "Pr(>F)"])
[1] 0.31
```

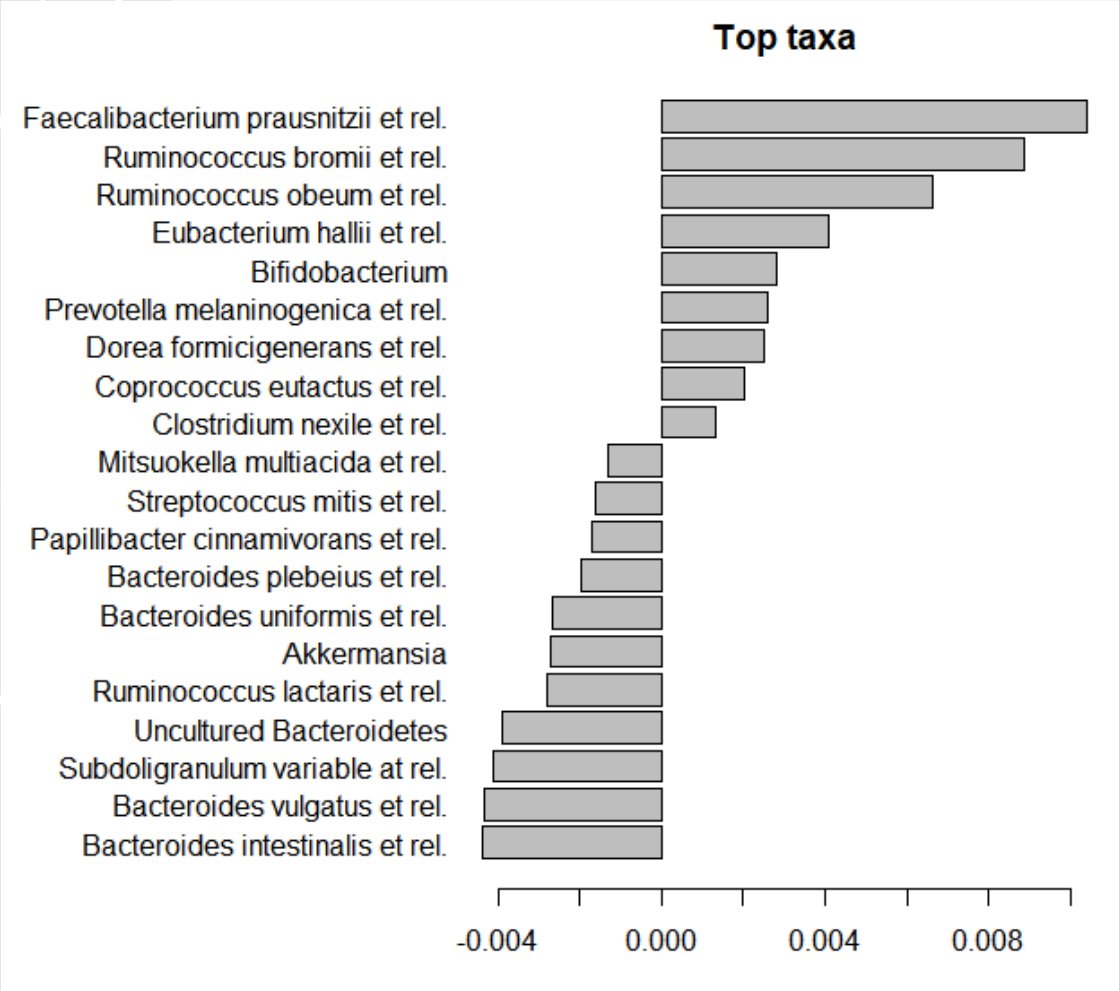
Homogeneity Condition

Analysis of Variance, aka ANOVA, may also be conducted on a dataset to check the variance homogeneity

```
dist <- vegdist(t(otu))
anova(betadisper(dist, meta$group))
Analysis of Variance Table

Response: Distances
        Df   Sum Sq   Mean Sq F value Pr(>F)
Groups     1 0.000016 0.0000160  0.0043 0.9483
Residuals 42 0.157881 0.0037591
```

Top Taxa Factors

Using PERMANOVA the coefficients for the top taxa can be computed to show the difference between various groups



References (Diversity)

1. https://socratic.org/questions/how-is-biodiversity-measured
2. http://www.evolution.unibas.ch/walser/bacteria_community_analysis/2015-02-10_MBM_tutorial_combined.pdf
3. https://slidetodoc.com/chapter-4-species-diversity-tables-figures-and-equations/
4. https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12462
5. http://www.metagenomics.wiki/pdf/definition/alpha-beta-diversity
6. https://www.easycalculation.com/statistics/learn-shannon-wiener-diversity.php
7. http://rewhc.org/biomeasures.shtml
8. https://geographyfieldwork.com/Simpson'sDiversityIndex.htm#:~:text=Simpson's Diversity Index is a,evenness increase, so diversity increases.
9. https://bioconductor.org/packages/devel/bioc/manuals/microbiome/man/microbiome.pdf
10. https://mothur.org/wiki/chao/
11. http://www.countrysideinfo.co.uk/simpsons.htm
12. https://microbiome.github.io/tutorials/Alphadiversity.html
13. https://microbiome.github.io/tutorials/Betadiversity.html
14. https://www.webpages.uidaho.edu/~beth/rem357_lectures_files/17dominance%20diversity.pdf
15. https://www.statisticshowto.com/bray-curtis-dissimilarity/

References (Variate Analysis)

1. https://sciencing.com/similarities-of-univariate-multivariate-statistical-analysis-12549543.html
2. https://microbiome.github.io/tutorials/Mixedmodels.html
3. https://microbiome.github.io/tutorials/PERMANOVA.html
4. https://cran.r-project.org/web/packages/lme4/index.html
5. https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat07841