Leiden University
Medical Center

Exploratory microbiome analysis

**Dr. Quinten Ducarmon, Dr. Morgan Essex
& Prof. Georg Zeller**

Quinten Ducarmon

IAI Course

LEIDEN, OCTOBER 27TH 2025

Learning objectives of today's lecture and practical

- Understand a sequence count table from a real microbiome study

- Understand and calculate sparsity and library size

- Understand and apply relative abundance transformation to a count table

- Ability to join metadata and count data by Sample IDs

- Apply data frame manipulation and perform summary statistics to isolate specific bacterial taxa

A teaser on typical microbiome plots to make biological sense out of the complex data

Schirmer et al. Nat Microbiol 2018
Milani et al., Sci Rep 2016
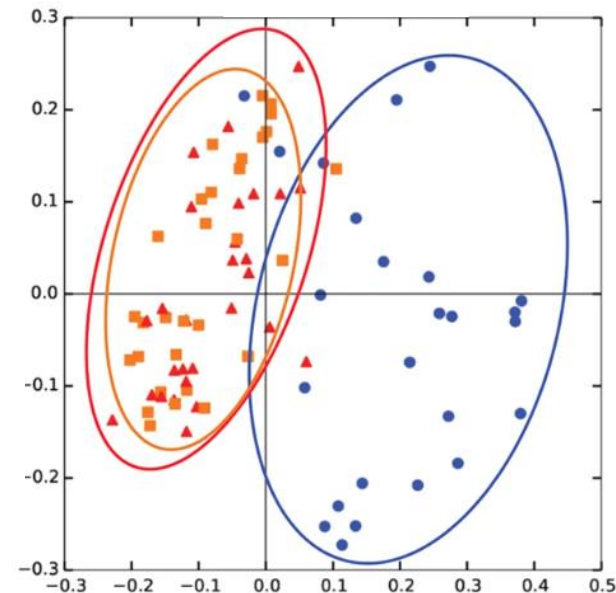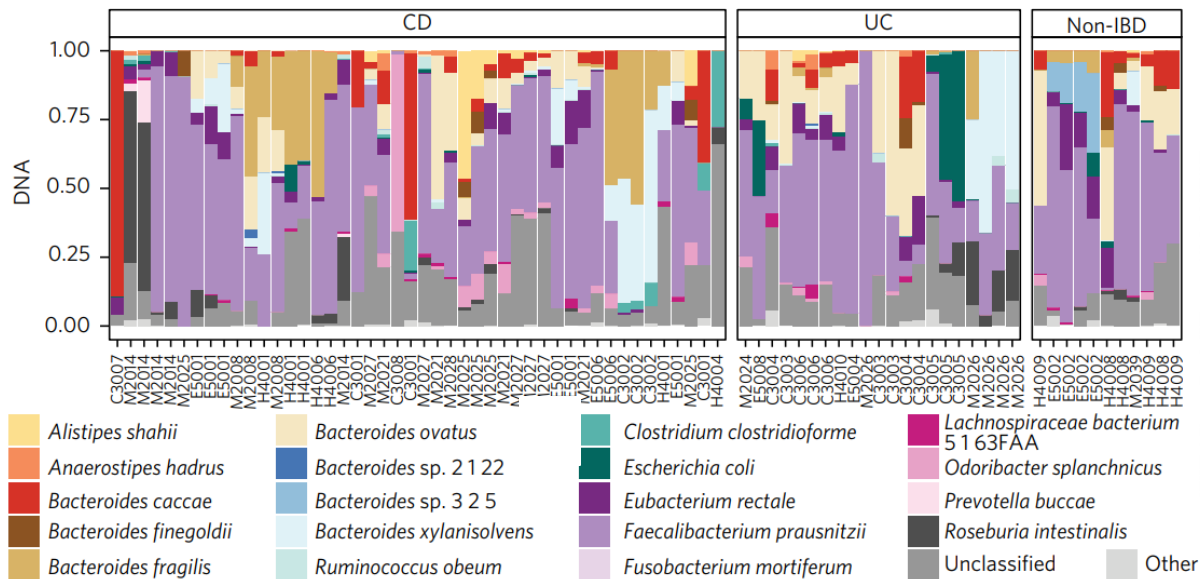
What does microbiome data look like?

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
Roseburia	80	10	70	40	50	55	35
Bacteroides	120	170	130	30	130	150	120
Enterococcus	30	20	30	35	40	20	80
Eubacterium_hallii	40	15	50	45	45	30	10
Faecalibacterium	70	80	70	65	65	40	30
Prevotella	50	50	55	80	45	25	40
Paraprevotella	40	10	50	45	50	20	40
Akkermansia	40	15	65	50	55	40	70
Collinsella	45	45	30	45	35	45	50
Clostridioides	0	0	0	0	0	0	30
Ruminococcus_2	50	15	35	40	45	35	40
Subdoligranulum	50	0	55	40	45	45	25
Lactobacillus	30	80	25	35	35	40	70
Alistipes	45	25	45	30	30	35	20
Escherichia-Shigella	40	35	20	30	45	40	60
Bifidobacterium	40	60	35	40	0	45	25
Veillonella	10	45	15	30	50	20	90
Fusobacterium	20	70	30	30	20	0	25
Anaerostipes	30	20	40	55	35	60	20
Clostridium	25	90	40	50	40	55	20
Blautia	30	65	35	55	35	40	20
Dialister	30	20	15	30	25	45	25
Christensenellaceae_R-7_group	35	25	20	40	30	50	15
Fusicatenibacter	30	20	25	35	35	55	10
Rumminococcaceae_UCG-002	20	15	15	25	15	10	30

Characteristics of gut microbiome data

- Highly individualistic, with high inter-individual variation and low intra-individual variation (i.e. a microbial fingerprint)
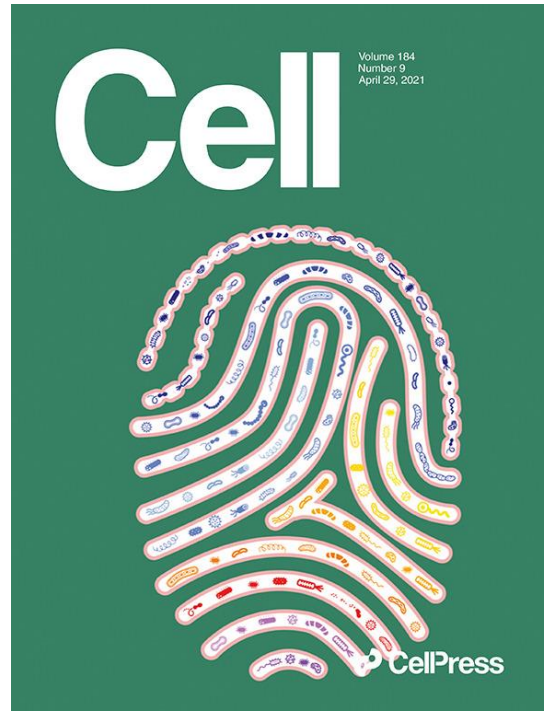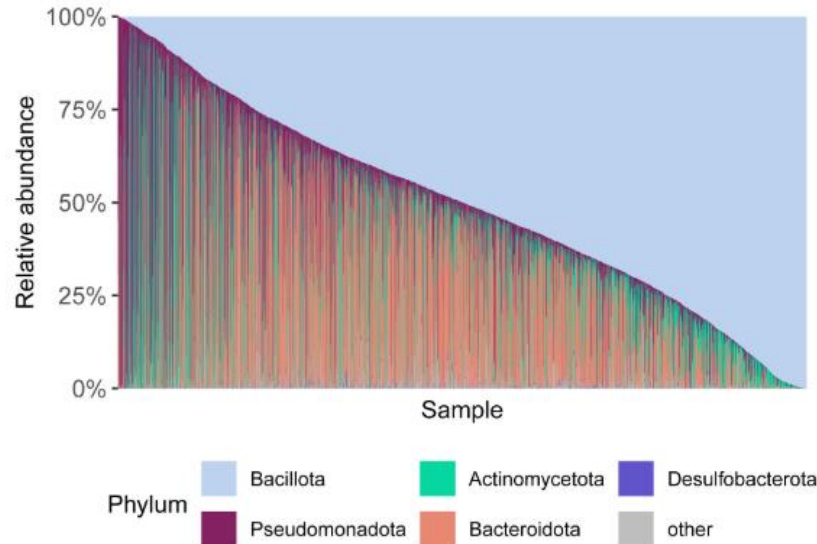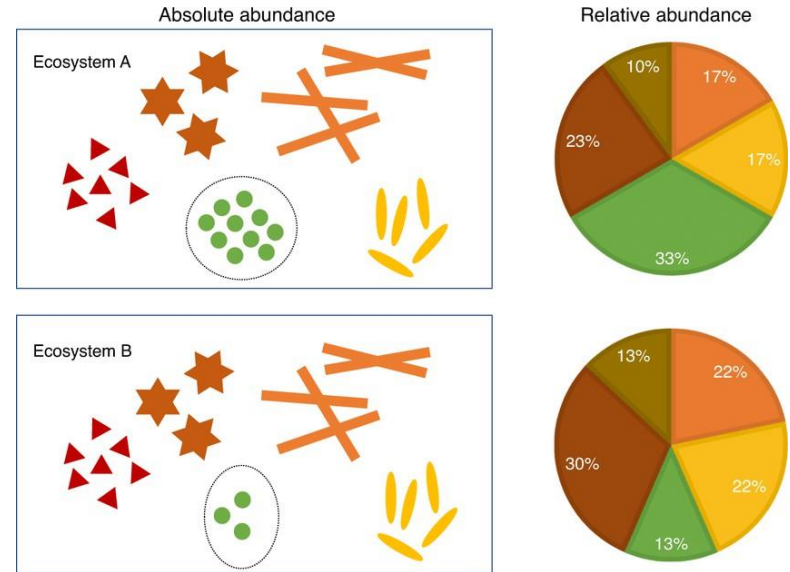

Chen et al., Cell, 2021

Characteristics of gut microbiome data

- Highly individualistic with high inter-individual variation and low intra-individual variation
- Many zeroes in the data (sparsity)

Chen et al., Cell, 2021
Abdill et al., Cell 2025

Characteristics of gut microbiome data

- Highly individualistic with high inter-individual variation and low intra-individual variation
- Many zeroes in the data (sparsity)
- Sequencing data is <u>always</u> relative

Chen et al., Cell, 2021
Abdill et al., Cell 2025
https://www.cd-genomics.com/resource-absolute-abundance-relative-abundance.html

Why do we transform microbiome data prior to testing?

- Ensure comparability between samples

- Account for sequencing depth (transform count to relative abundances), as raw counts are not comparable between samples
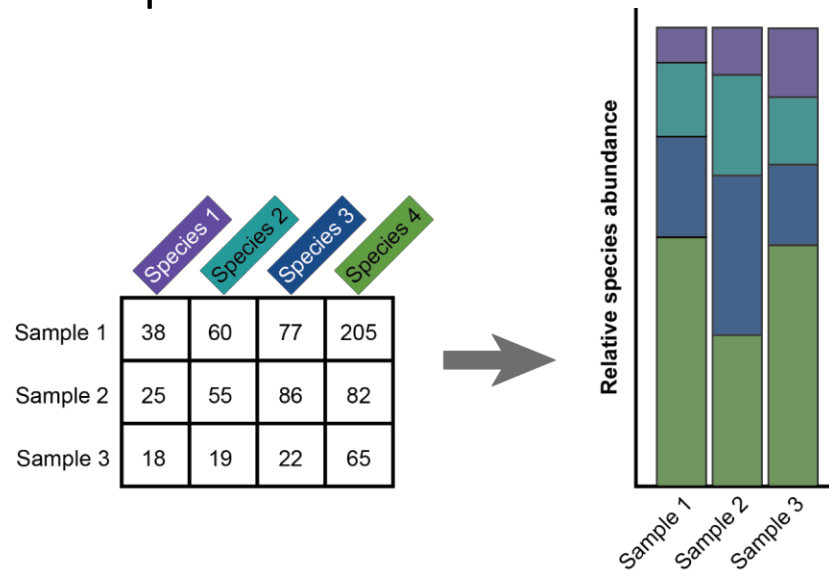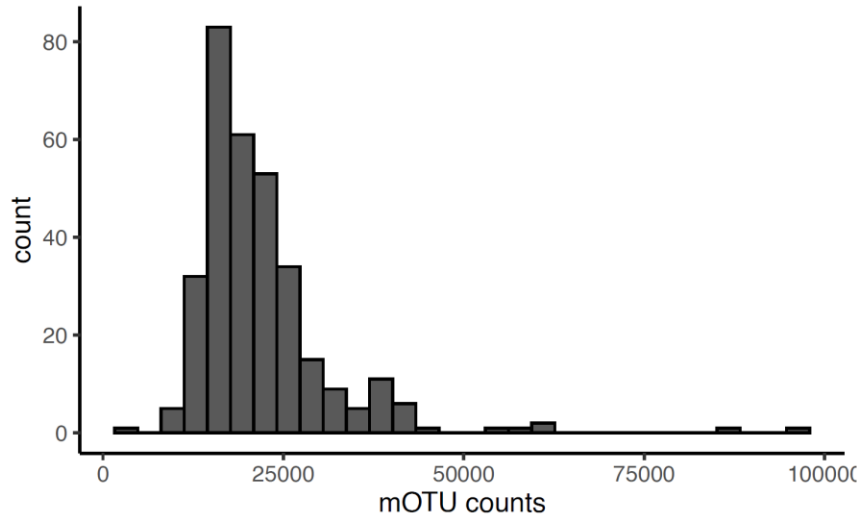
# Which dataset will you be working with today?

RESEARCH PAPER

Establishing high-accuracy biomarkers for colorectal cancer by comparing fecal microbiomes in patients with healthy families

Jian Yang[a], Dongfang Li[b], Zhenyu Yang[c,d], Wenkui Dai[e,f], Xin Feng[e], Yanhong Liu[e], Yiqi Jiang[g], Pingang Li[a], Yinhu Li[e], Bo Tang[a], Qian Zhou[g], Chuangzhao Qiu[e], Chao Zhang[a], Ximing Xu[c,d], Su Feng[c,d], Daxi Wang[e], Heping Wang[f,h], Wenjian Wang[f,h], Yuejie Zheng[f,h], Lin Zhang[i], Wenjie Wang[j,k], Ke Zhou[b], Shuaicheng Li[g], and Peiwu Yu[a]

Visualizing microbial abundances
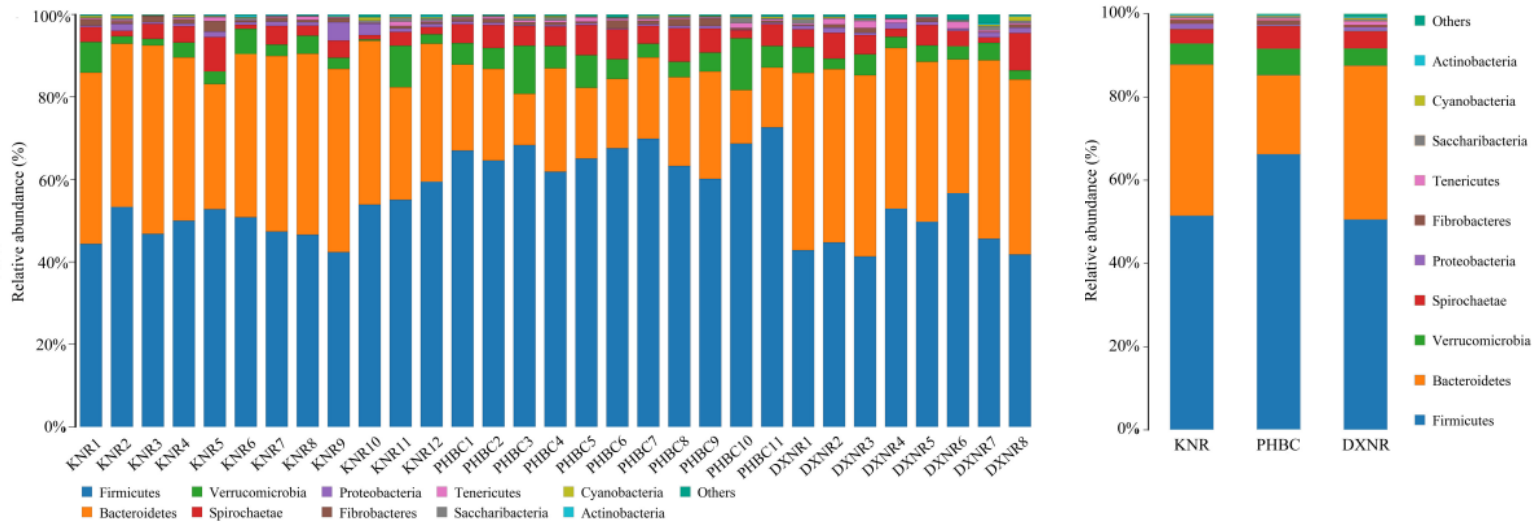
- What is a potential disadvantage of only visualizing relative abundances at group level (the right plot) ?

Visualizing microbial abundances
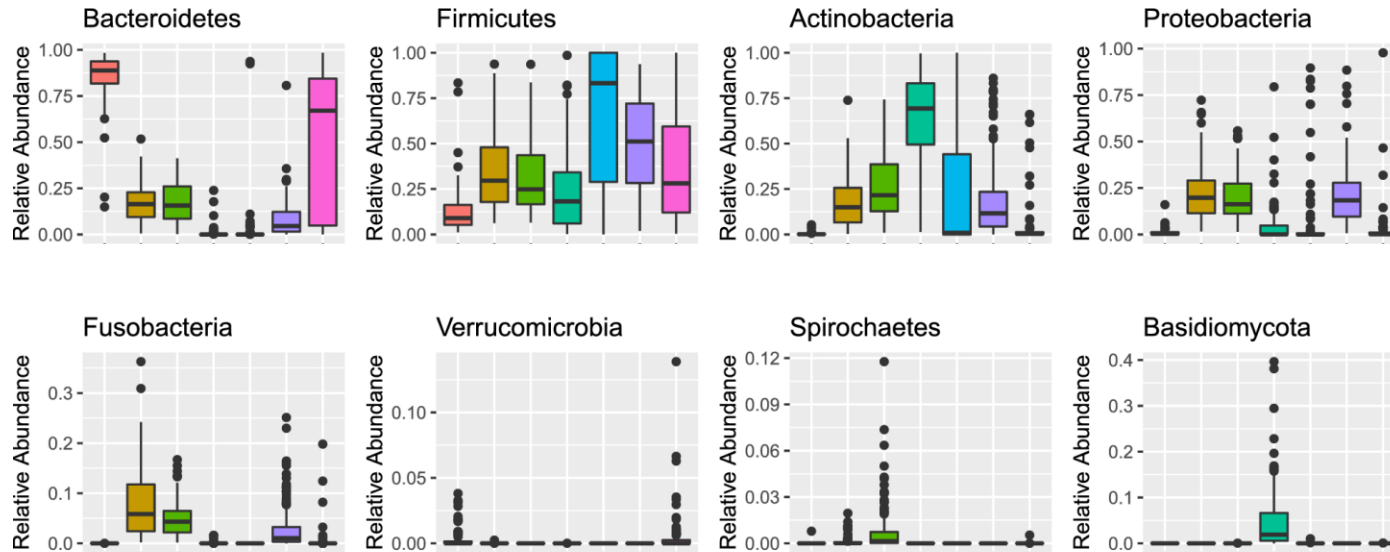
- How insightful do you think all these boxplots are
  and what are (dis)advantages of this visualization?

Many factors influence the microbiome

- Many factors influence the gut microbiome.

- These cannot always be disentangled.

- Therefore, always be careful when attributing specific microbiome changes to a disease (e.g. CRC)

Schmidt et al., Cell 2018
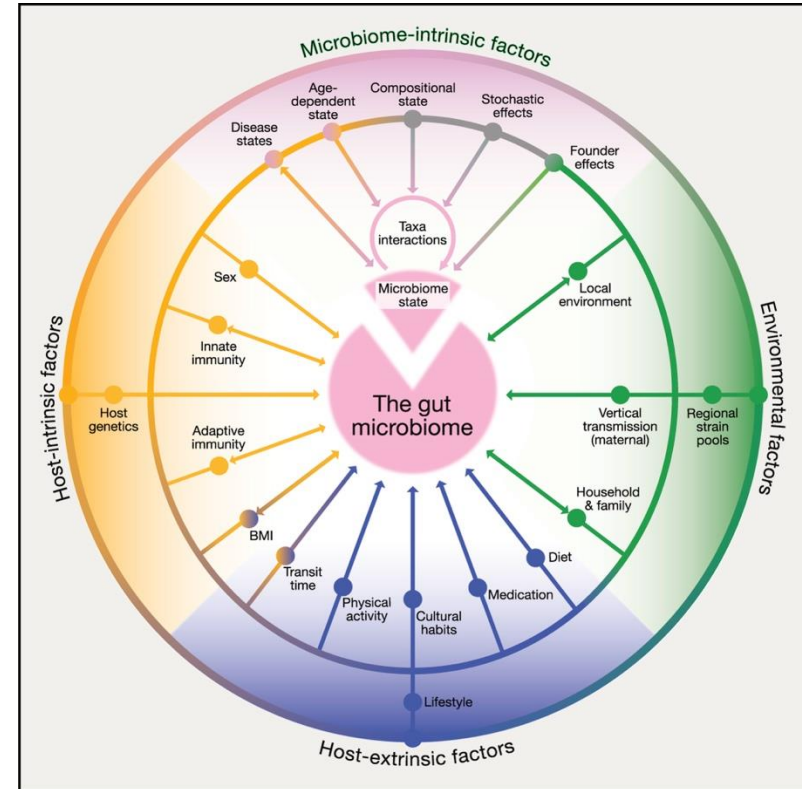
Recap on computer science 101

- *Syntax*: rules that govern a language (+ most common type of errors!)
- *Function*: a unit of code designed to perform a specific (defined) task
- *Library/Package:* related sets of functions that are stored + work together

```
1   # comments start with a hashtag
2   everything after the # is ignored
3   # (and line 2 would cause an error!)

4   # loading functions into a script
5   library(package_name)
6   source("helper_script.R")

7   getwd() # where am I, computer?
8   here::here() # preferred
```
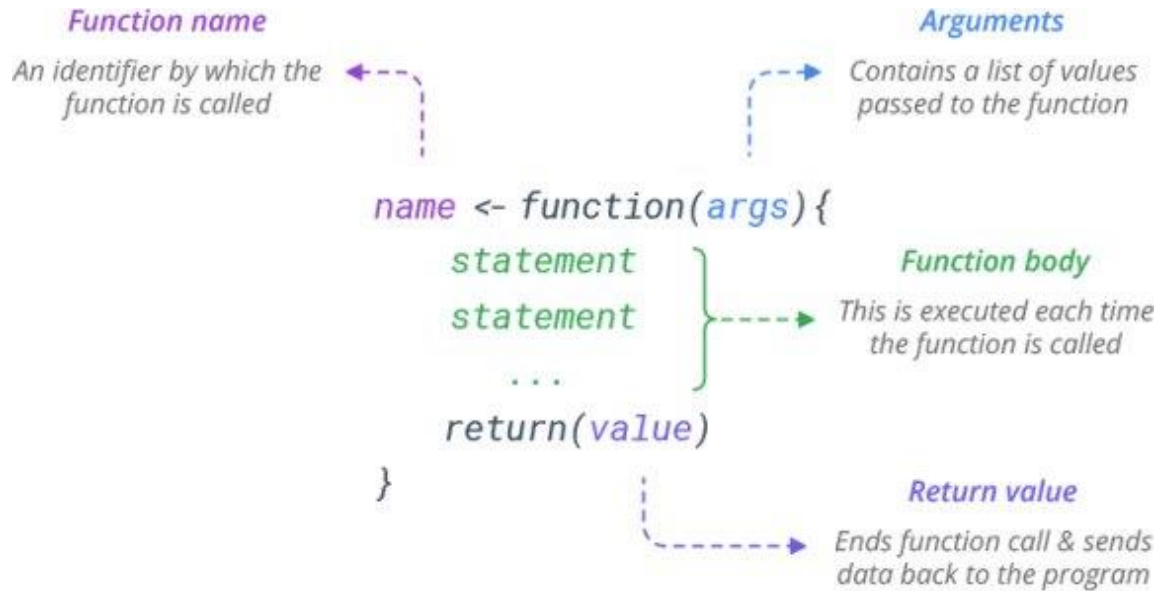
When a calculation is needed many times

```r
```{r}
my_function <- function(x) {
  return (5 * x)
}

print(my_function(3))
```
```

```
[1] 15
```

Function name
An identifier by which the function is called

Arguments
Contains a list of values passed to the function

```r
name <- function(args){
    statement
    statement
    ...
    return(value)
}
```

Function body
This is executed each time the function is called

Return value
Ends function call & sends data back to the program

Take-home messages

- Gut microbiome / sequencing data has very specific characteristics such as high sparsity

- Be mindful of your data characteristics when visualizing

- Sequencing data is complex