CODA 4 MICROBIOME

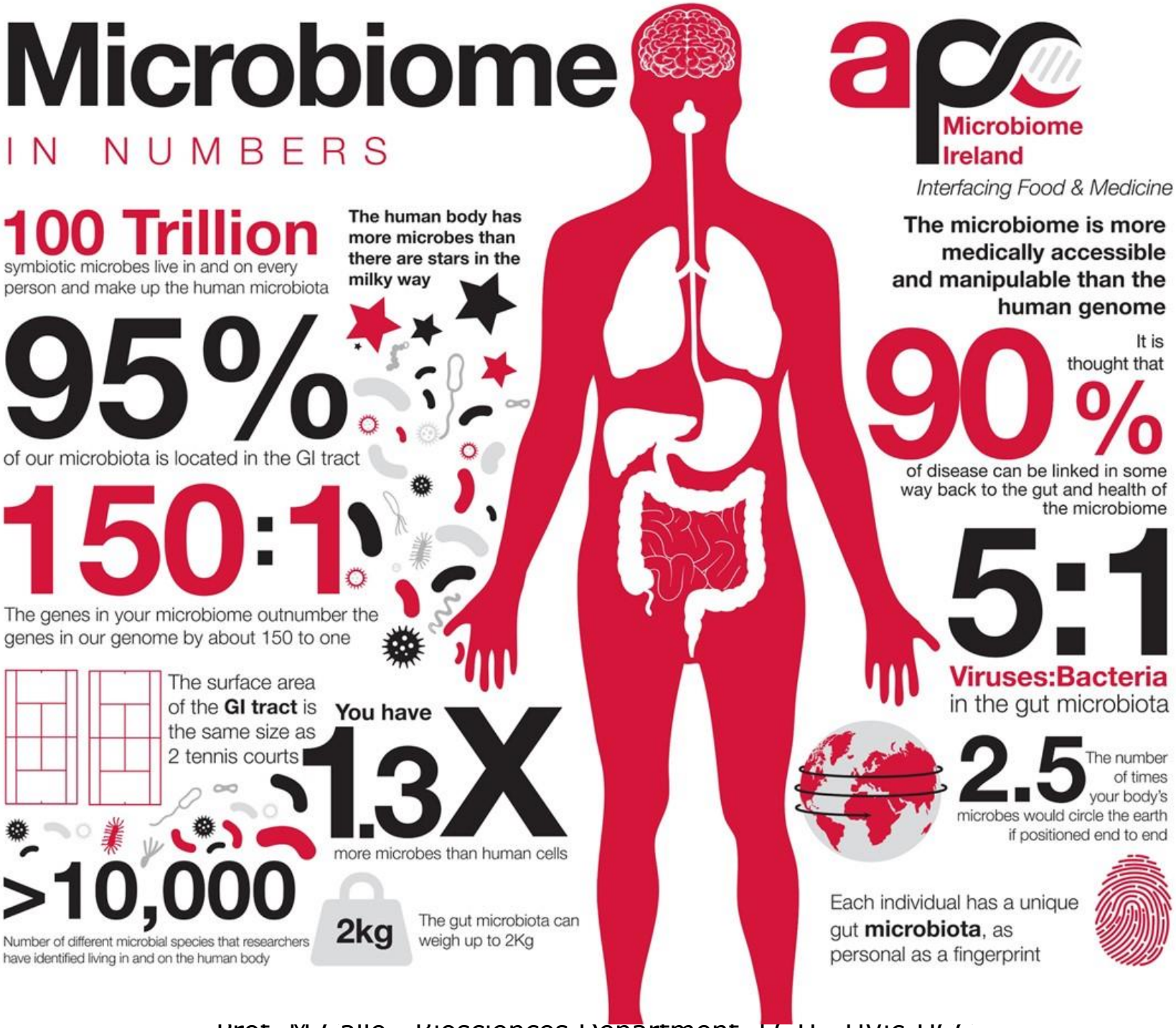
Statistical methods for microbiome analysis

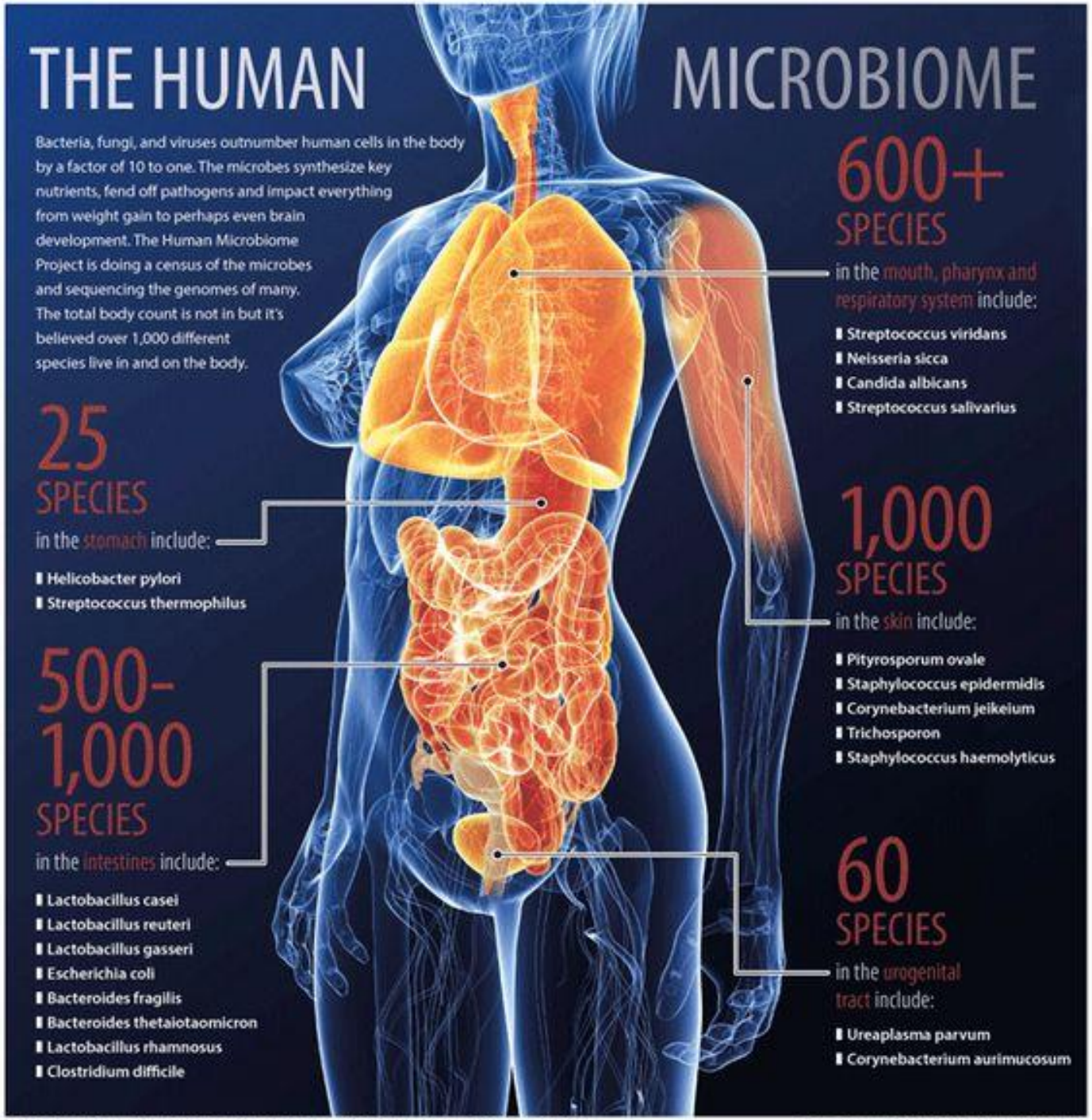
Malu Calle
Universitat de Vic - UCC

UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA

Program

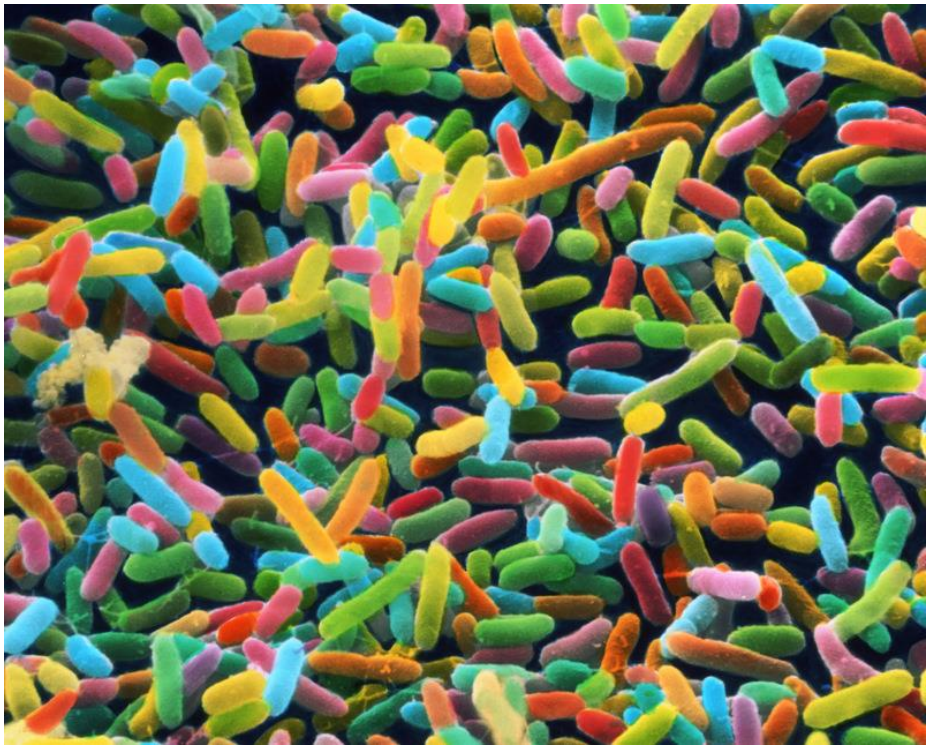
1. The human microbiome
2. NGS microbiome studies
   - 2.1. Microbial DNA extraction and sequencing
   - 2.2. 16S ribosomal RNA gene
   - 2.3. 16S Bioinformatics pipeline
3. Microbiome statistical analysis
   - 3.1. Main goals and challenges of microbiome statistical analysis
   - 3.2. Exploratory analysis - Abundance plots
   - 3.3. Ecological measures of richness and diversity
   - 3.4. Ordination: Visualization of beta diversity
   - 3.5. Standard methods for Microbiome differential abundance testing
4. Compositional data analysis of microbiome cross-sectional studies
   - 4.1. Exploratory analysis of log-ratios
   - 4.2. Identification of microbial signatures
5. Compositional data analysis of microbiome longitudinal studies
   - 5.1. Log-ratio analysis of microbiome longitudinal data
   - 5.2. Identification of dynamic microbial signatures


Microbiome
IN NUMBERS
Microbiome Ireland
Interfacing Food & Medicine
100 Trillion
symbiotic microbes live in and on every person and make up the human microbiota
The human body has more microbes than there are stars in the milky way
95%
of our microbiota is located in the GI tract
150:1
The genes in your microbiome outnumber the genes in our genome by about 150 to one
The surface area of the GI tract is the same size as 2 tennis courts
You have
1.3X
more microbes than human cells
>10,000
Number of different microbial species that researchers have identified living in and on the human body
2kg
The gut microbiota can weigh up to 2Kg
The microbiome is more medically accessible and manipulable than the human genome
90%
It is thought that
of disease can be linked in some way back to the gut and health of the microbiome
5:1
Viruses:Bacteria
in the gut microbiota
2.5
The number of times your body's microbes would circle the earth if positioned end to end
Each individual has a unique gut microbiota, as personal as a fingerprint


THE HUMAN MICROBIOME
Bacteria, fungi, and viruses outnumber human cells in the body by a factor of 10 to one. The microbes synthesize key nutrients, fend off pathogens and impact everything from weight gain to perhaps even brain development. The Human Microbiome Project is doing a census of the microbes and sequencing the genomes of many. The total body count is not in but it's believed over 1,000 different species live in and on the body.
600+ SPECIES in the mouth, pharynx and respiratory system include:
Streptococcus viridans
Neisseria sicca
Candida albicans
Streptococcus salivarius
25 SPECIES in the stomach include:
Helicobacter pylori
Streptococcus thermophilus
1,000 SPECIES in the skin include:
Pityrosporum ovale
Staphylococcus epidermidis
Corynebacterium jeikeium
Trichosporon
Staphylococcus haemolyticus
500–1,000 SPECIES in the intestines include:
Lactobacillus casei
Lactobacillus reuteri
Lactobacillus gasseri
Escherichia coli
Bacteroides fragilis
Bacteroides thetaiotaomicron
Lactobacillus rhamnosus
Clostridium difficile
60 SPECIES in the urogenital tract include:
Ureaplasma parvum
Corynebacterium aurimucosum
SOURCES: NATIONAL INSTITUTES OF HEALTH, SCIENTIFIC AMERICAN; HUMAN MICROBIOME PROJECT
Dean Tweed • POSTMEDIA NEWS / IMAGE: Fotolia

The human microbiome

The human microbiome *-> important role in human health*



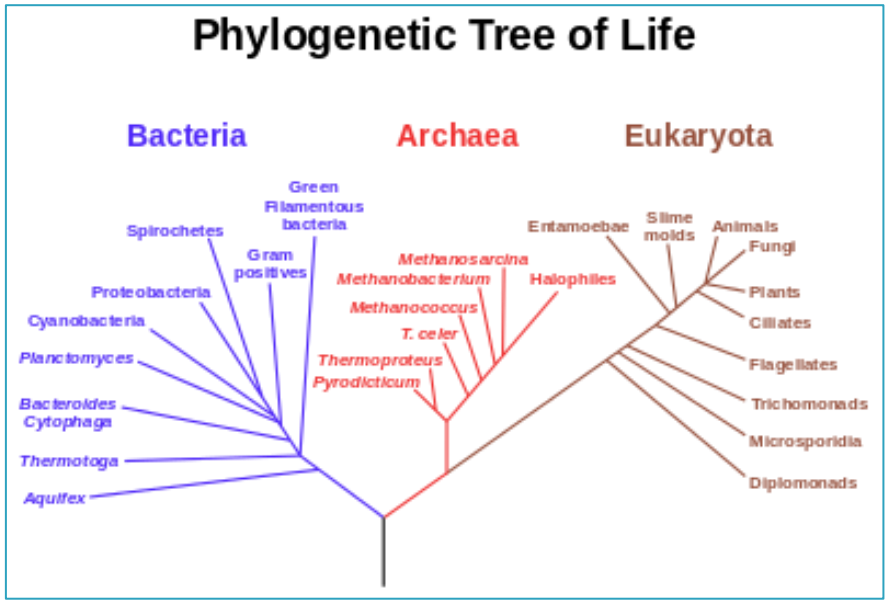
Microbiome or microbiota *= Community of microscopic organisms that live in a given environment*

Metagenome *= The collective genomes of a given community of micro-organisms. It is a measure of the functional potential of a given microbiota.*

Metagenomics *is the study of metagenomes*

The human microbiome

Microbiome *--> **bacteria**, archaea, fungi, viruses*

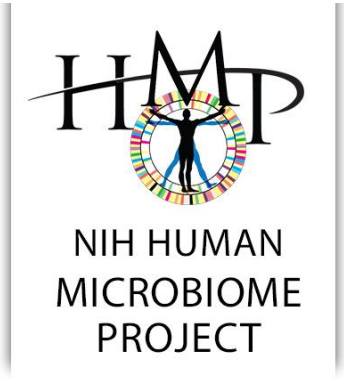


- *The number of bacteria associated with a human body is about the same as human body cells --> 40 trillion*
- *~ 2 Kg of body weight*

- *> 1,000 types of bacteria living in the guts*

Major international microbiome projects

*Large research initiatives in this field are the **Human Microbiome Project** (USA) and **MetaHIT**, Metagenomics of the Intestinal tract (EU) both integrated now in the **International Human Microbiome Consortium** (IHMC).*



What is a "healthy human microbiome"?

*The human microbiome has **extensive diversity** between anatomical sites, but also per individual and over time*

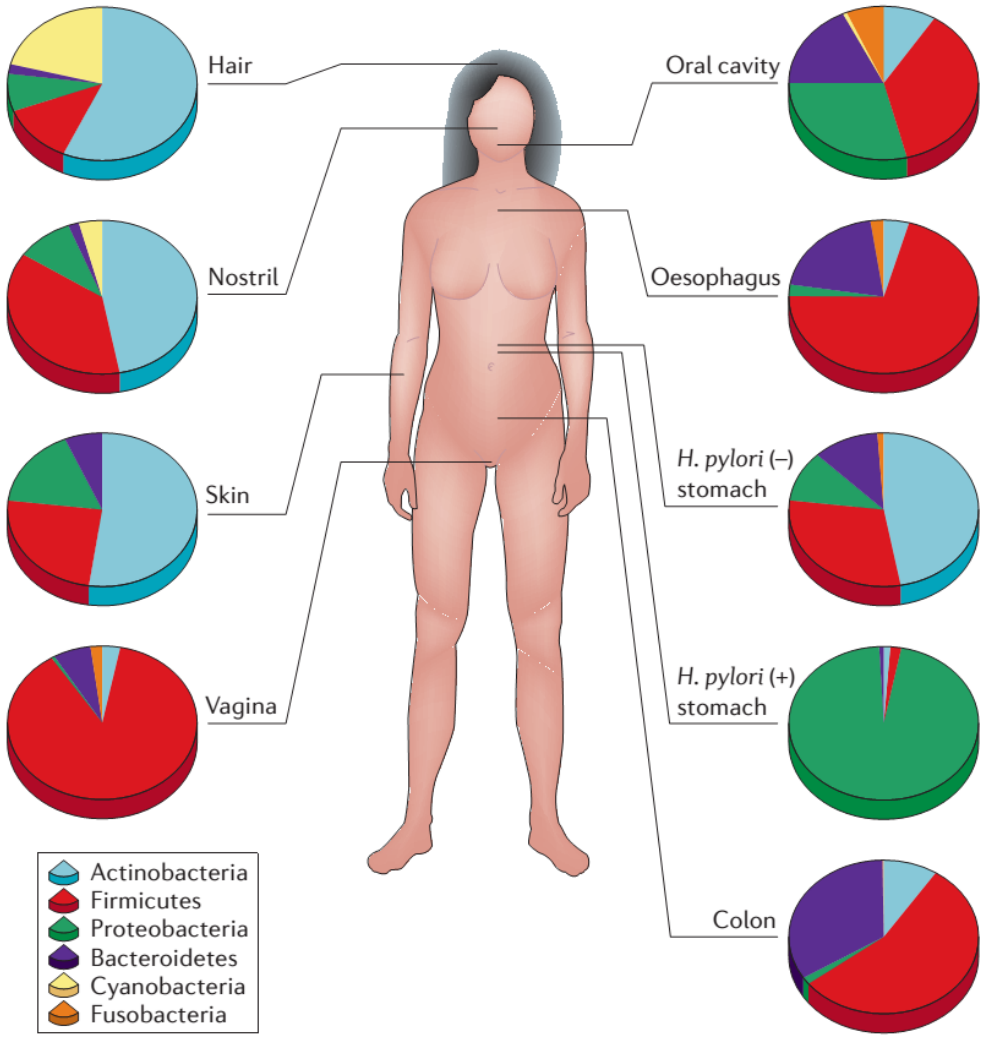
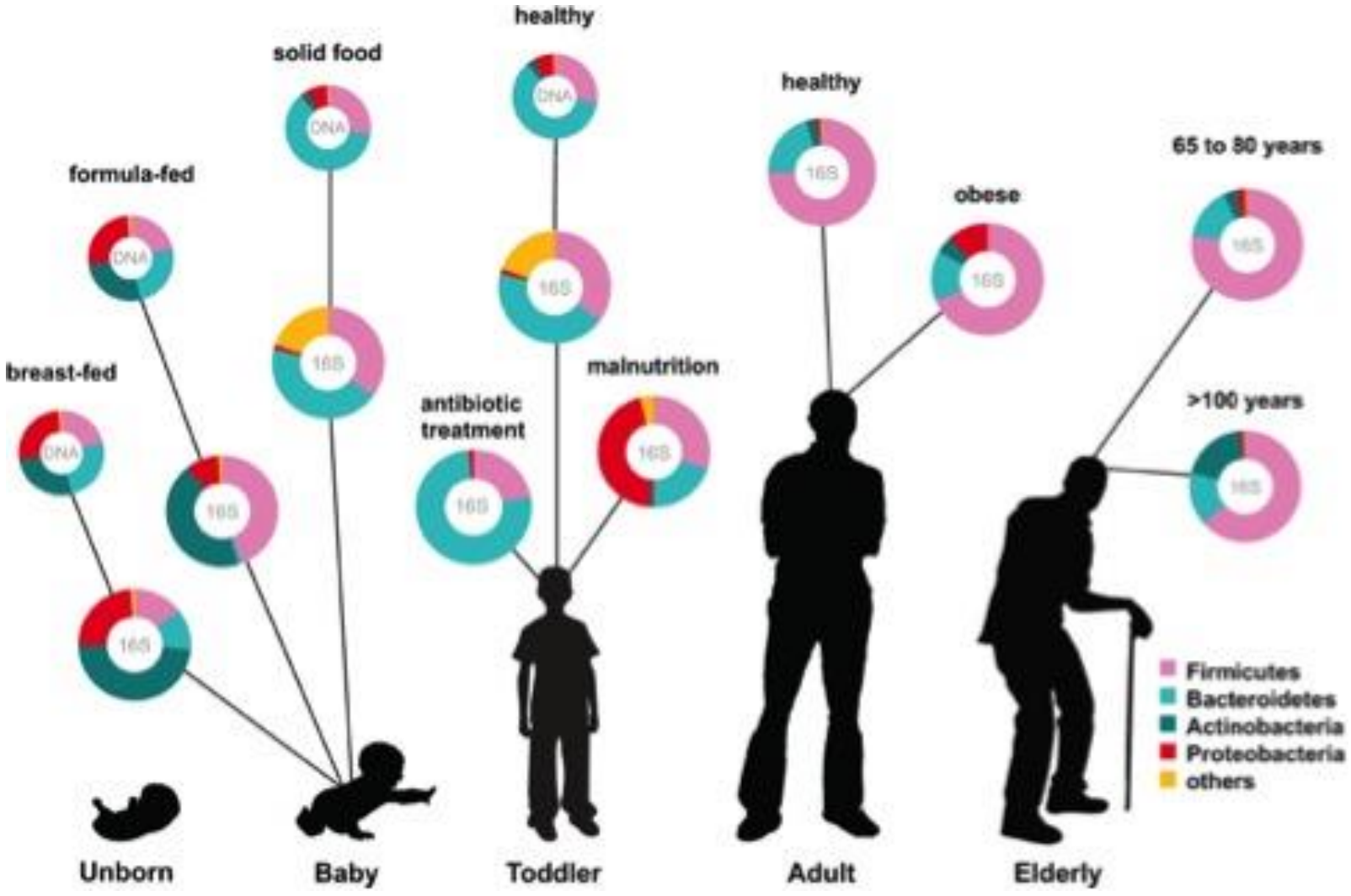


Figure 1 | **Compositional differences in the microbiome by anatomical site.**

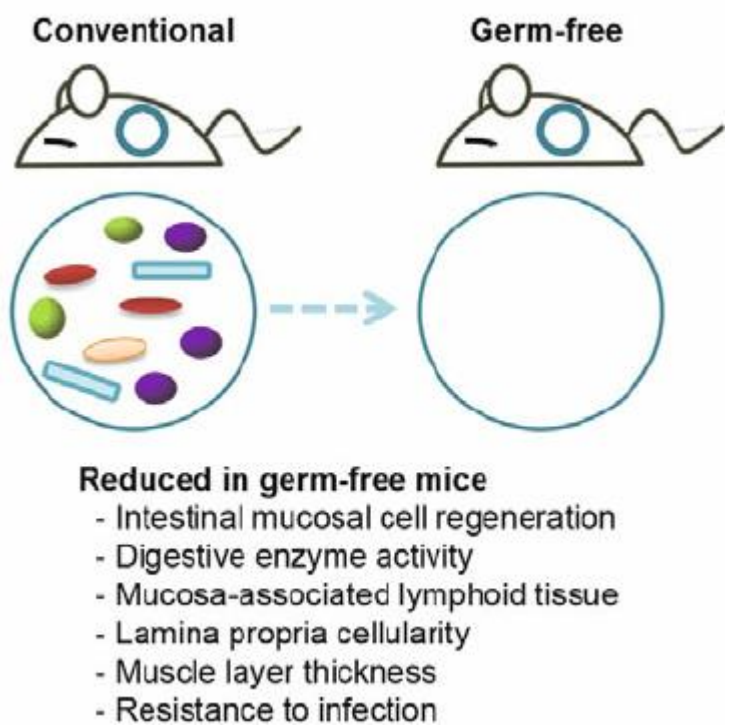
What is a "healthy human microbiome"?

*The human microbiome has **extensive diversity** between anatomical sites, but also per individual and over time*



Functions of the human microbiome: our second genome

- *The human gut microbiome contains **150 times more genes** than our own genome*
- *Can be viewed as a functional expansion of our own genome*
- *Adds a variety of enzymes that the human genome does not encode*
- *Germ-free mice, born and raised in sterile incubators, need to eat 30% more food to remain the same body weight than mice with a conventional microbiome.*



Functions of the human microbiome: our second genome

Functions of the human intestinal microbiome include:

- *Digestion of complex carbohydrates - extraction of energy from food*
- *modulation of the immune system*
- *vitamin synthesis*
- *lipid metabolism*
- *control of blood glucose levels*
- *brain-gut-axis mediation*

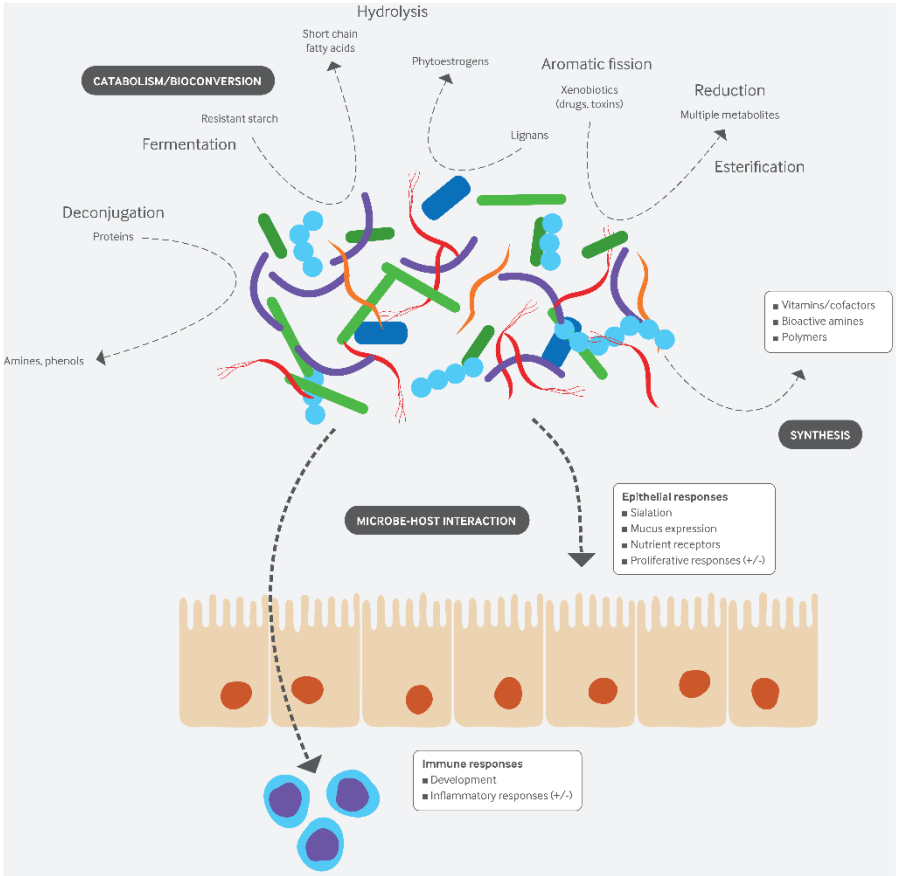


Fig 2| Potential functions of the indigenous microbiota. The microbiota can have effects through the microbes' synthetic or catabolic metabolic activity or through direct host-microbe interactions. Catabolism and bioconversion of dietary or host derived compounds can make nutrients more available to the host or alter the bioavailability of drugs. Some members of the microbiota can synthesize important cofactors or bioactive signaling molecules such as amines. Signaling between the microbiota and the host can trigger alterations in host function, such as altered expression of mucus or alteration of the immune response

Young2017

Microbiome and disease

- *Many diseases have been found to be associated with changes in the human microbiome*

- *Many of these associations might be correlations, not causations, with cause and effect often hard to distinguish*

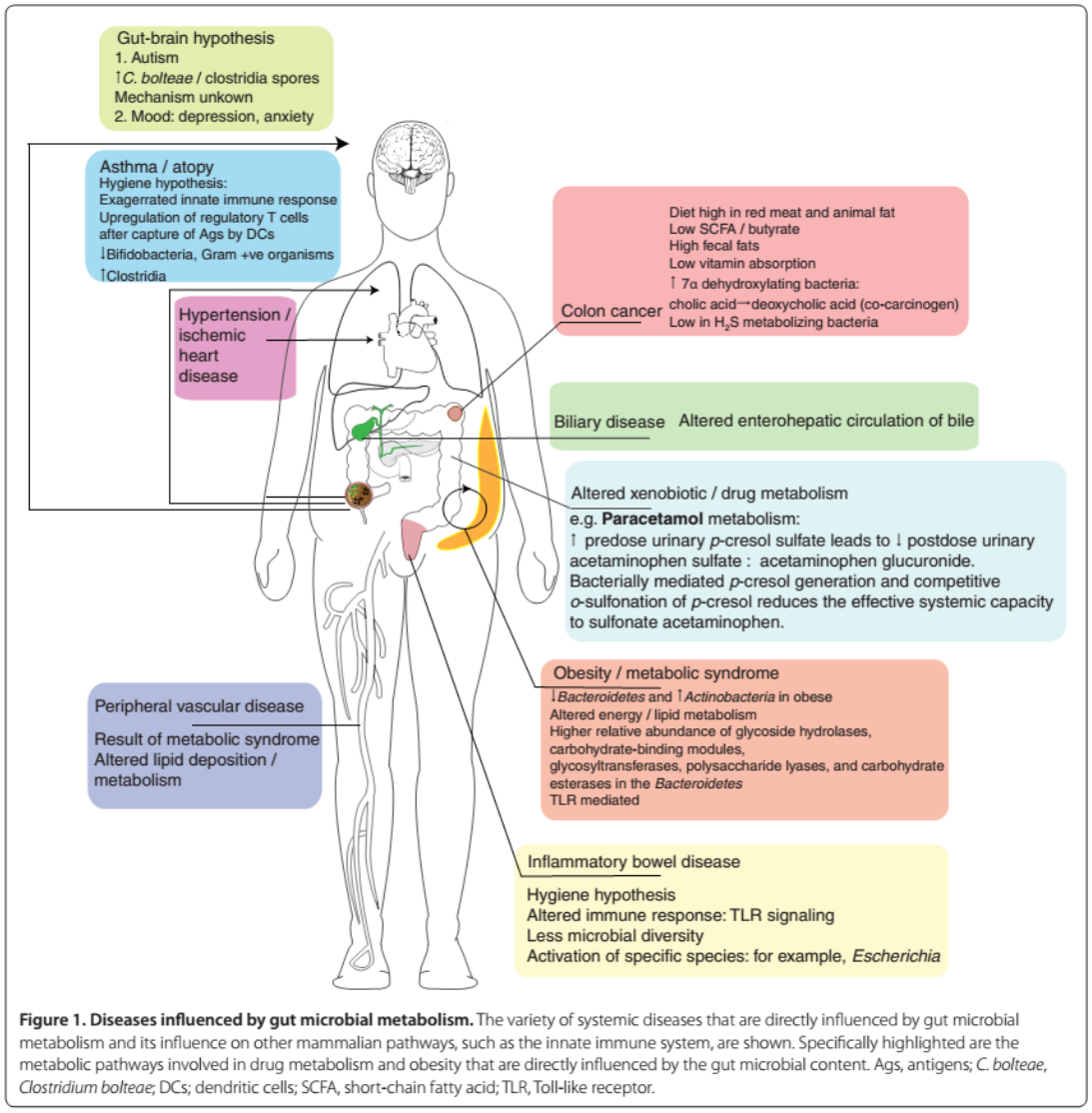
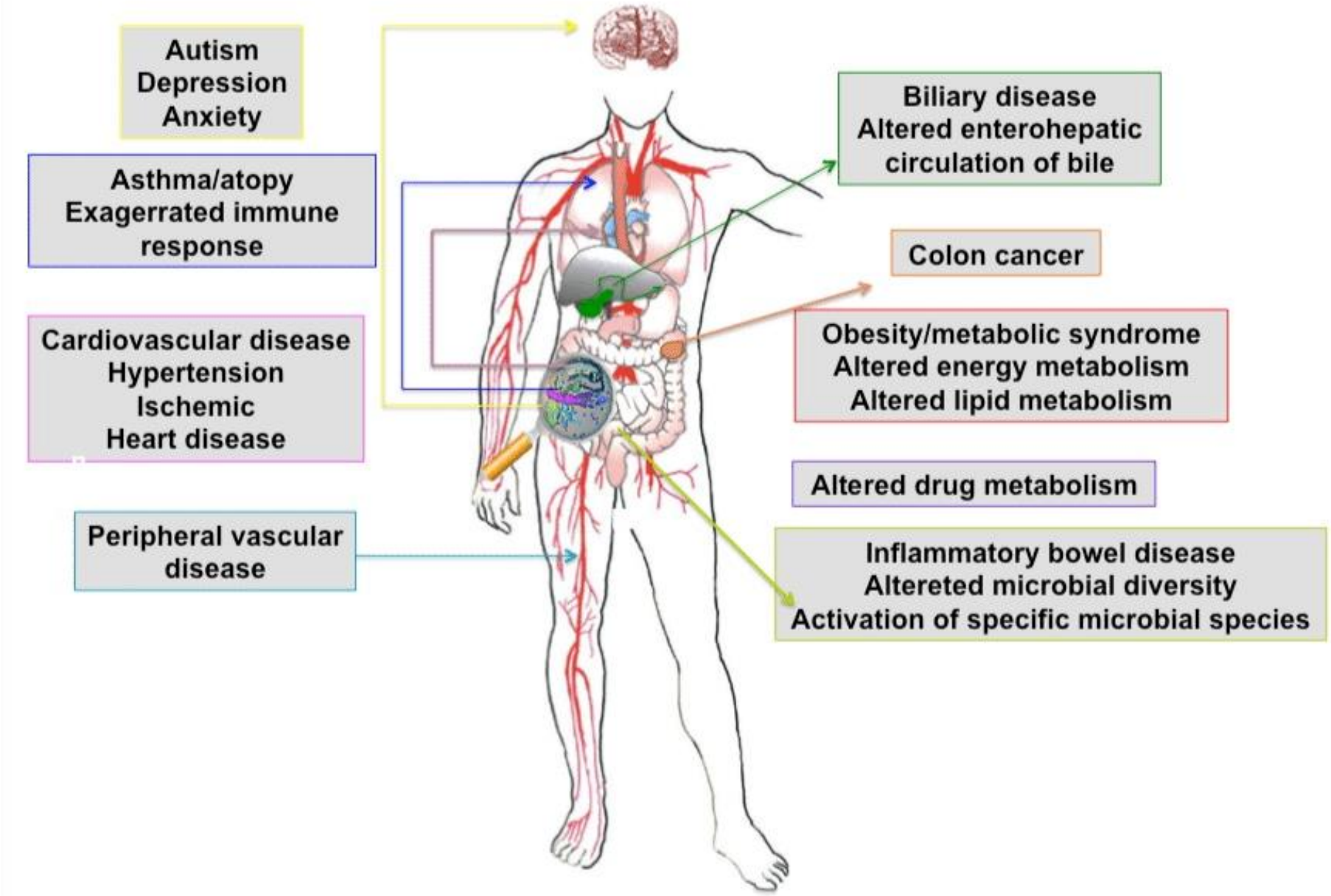
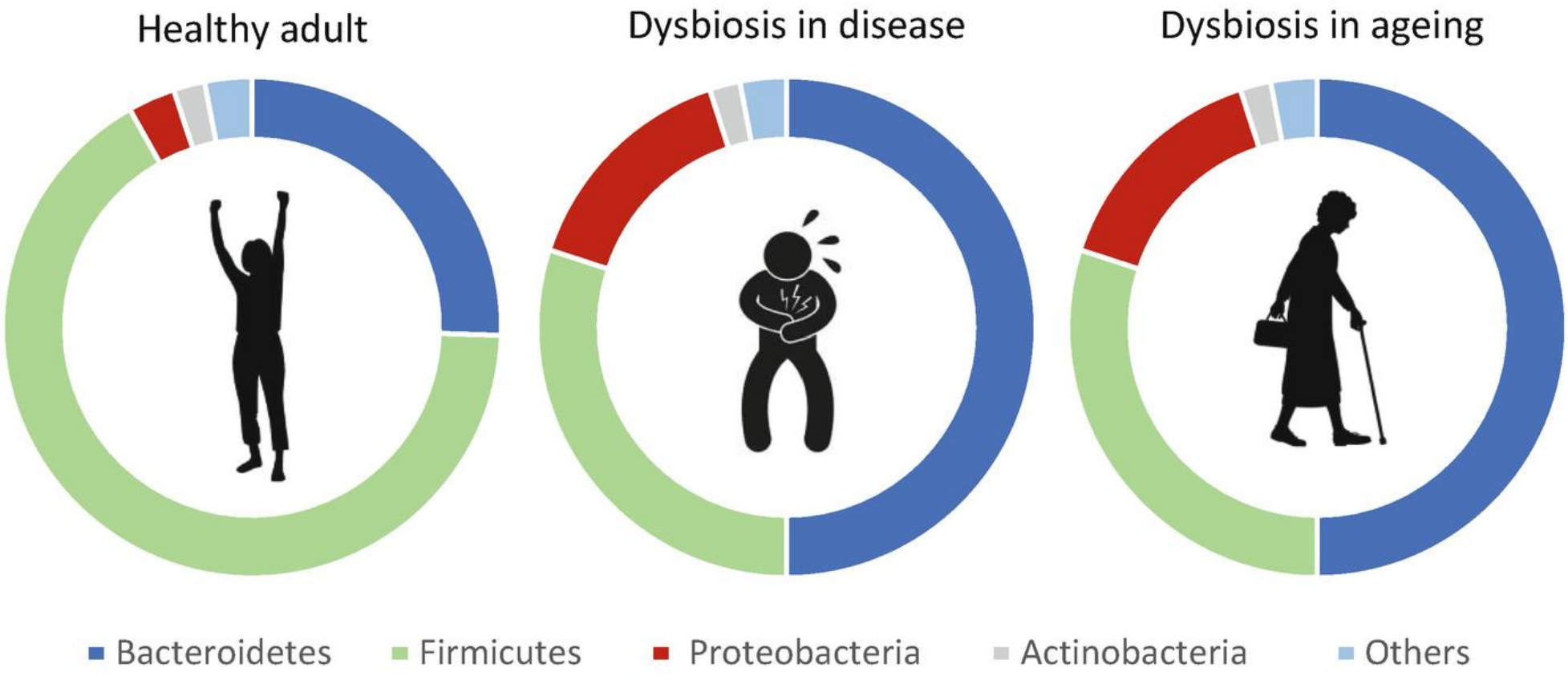


Figure 1. Diseases influenced by gut microbial metabolism. The variety of systemic diseases that are directly influenced by gut microbial metabolism and its influence on other mammalian pathways, such as the innate immune system, are shown. Specifically highlighted are the metabolic pathways involved in drug metabolism and obesity that are directly influenced by the gut microbial content. Ags, antigens; *C. bolteae*, *Clostridium bolteae*; DCs; dendritic cells; SCFA, short-chain fatty acid; TLR, Toll-like receptor.

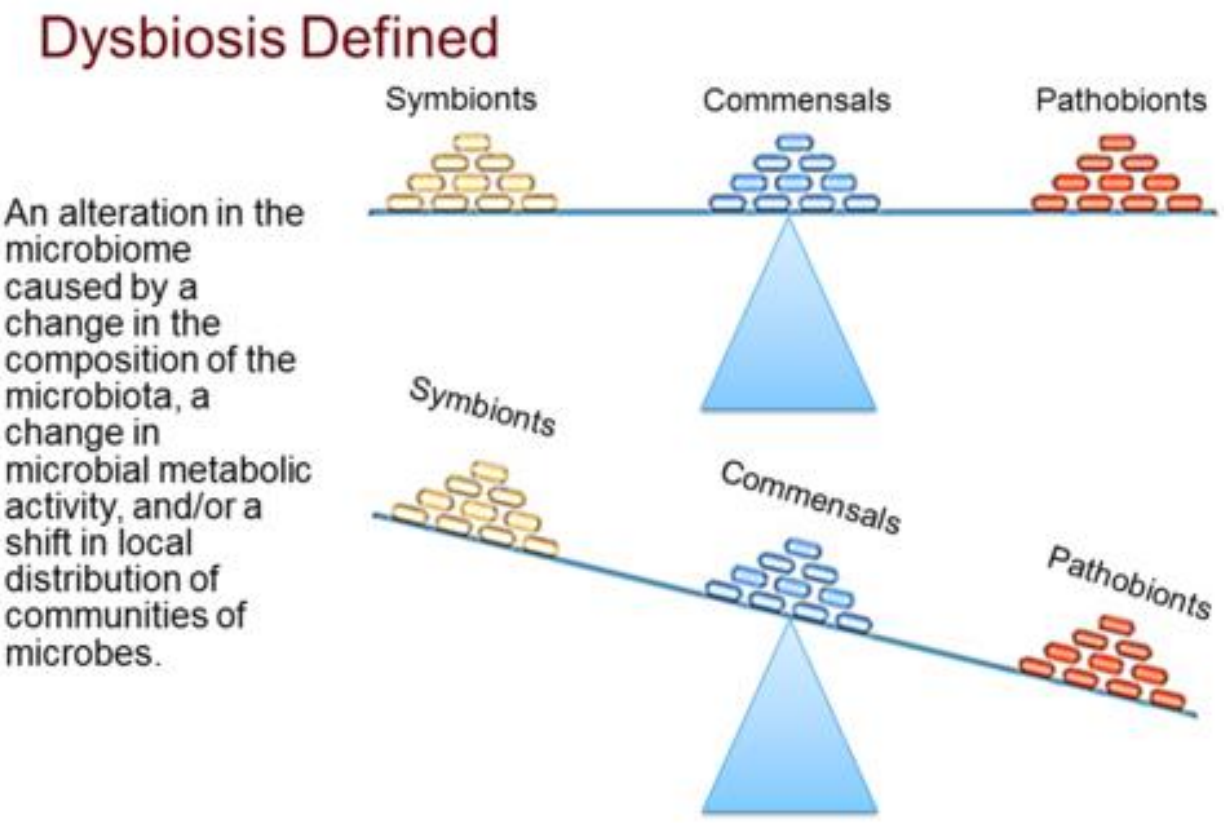
Microbiome and disease



Microbiome and disease

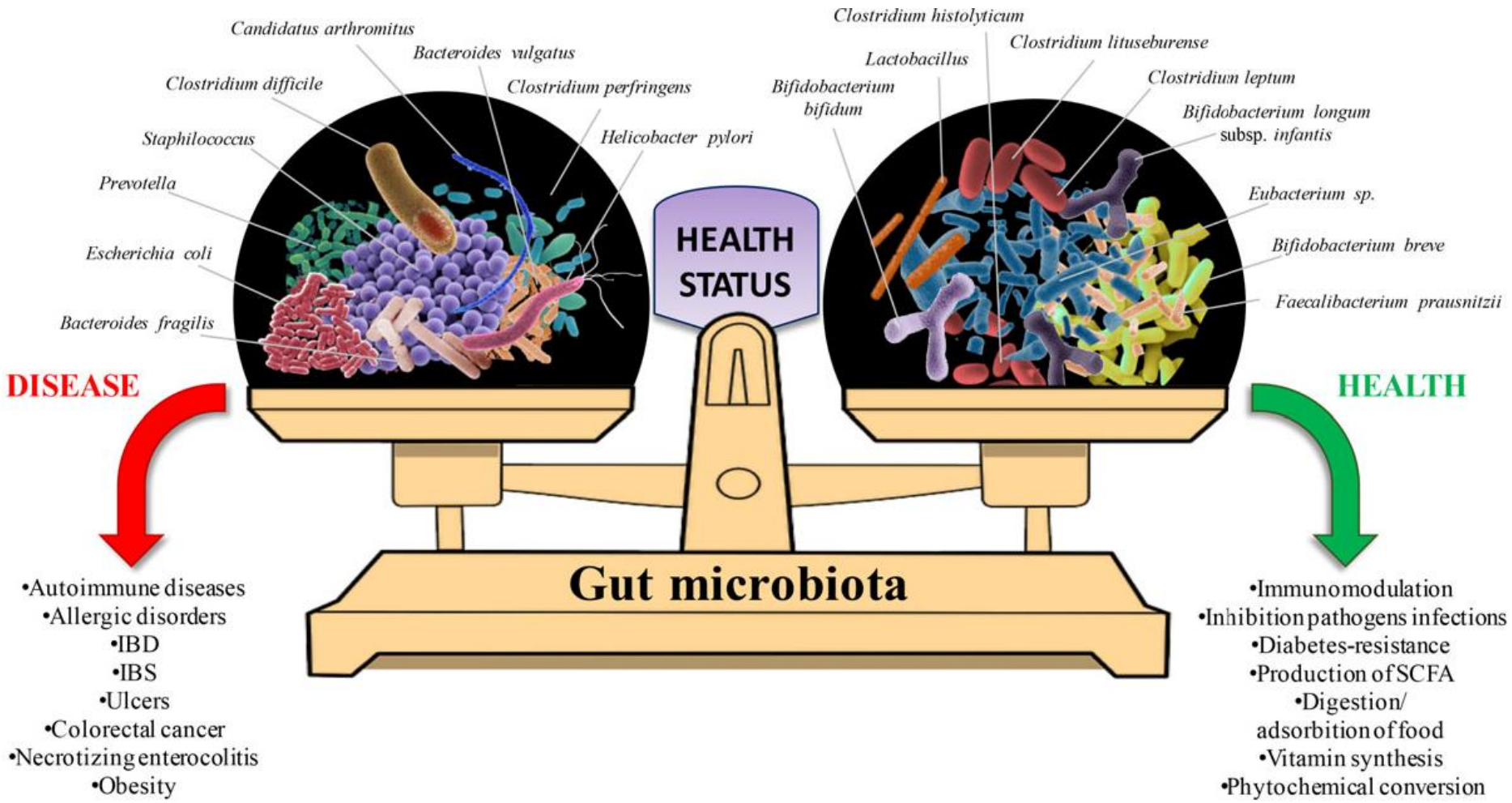


Microbiome and disease



Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 2009 May;9(5):313-23.

Microbiome and disease

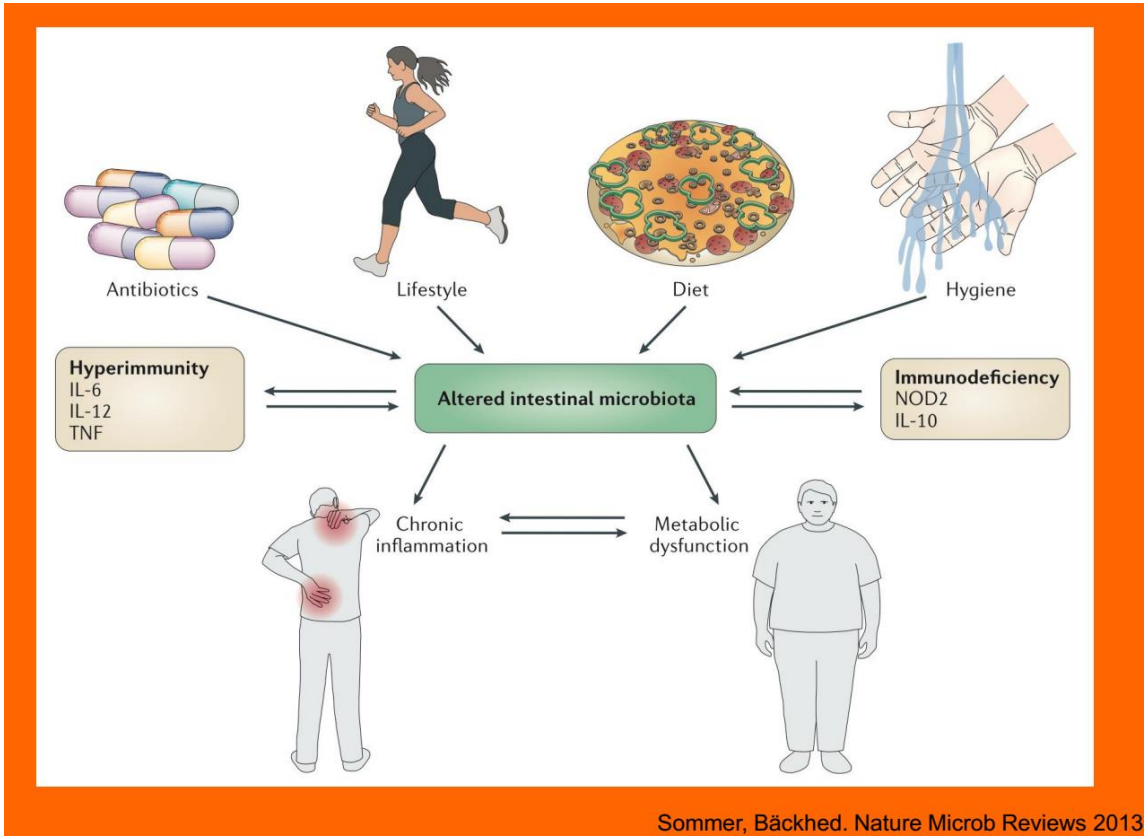


ure 1 Schematic representation of the functional roles of key members of the human gut microbiota in health and disease. IBD, inflammat

The gut microbiota and obesity

- *Different studies have shown that obesity is indeed partly determined by the composition of our microbiome.*

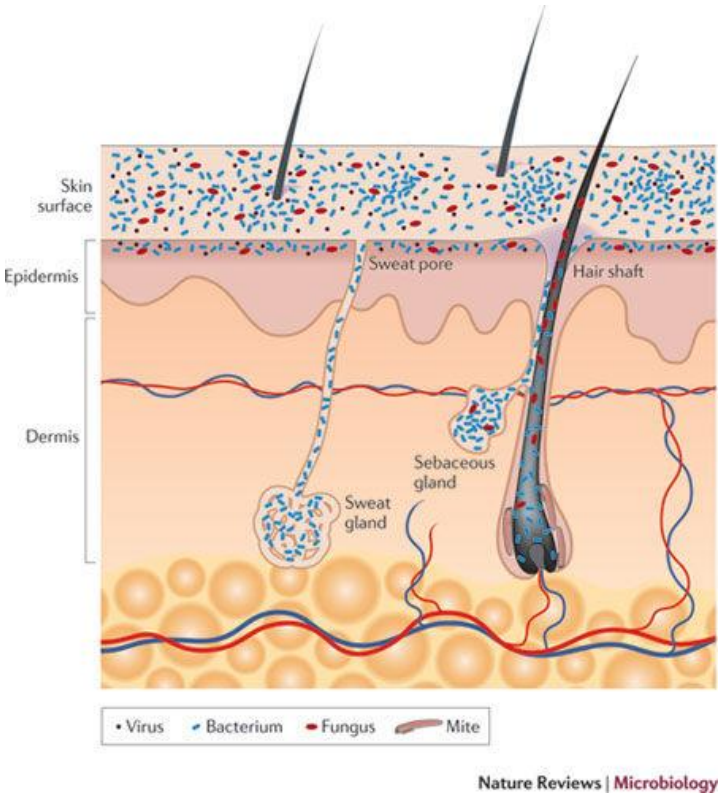
- ***Obesity** was associated with **decreased Bacteroidetes** and **diminished bacterial diversity**, with enrichment of genes related to lipid and carbohydrate metabolism.*



Cutaneous microbiome

- *Chronic inflammatory skin conditions such as Psoriasis, atopic dermatitis, acne and chronic skin ulcers have been associated to cutaneous microbiome changes*

- ***Psoriasis:*** *Firmicutes were significantly overrepresented and Actinobacteria were significantly under-represented*

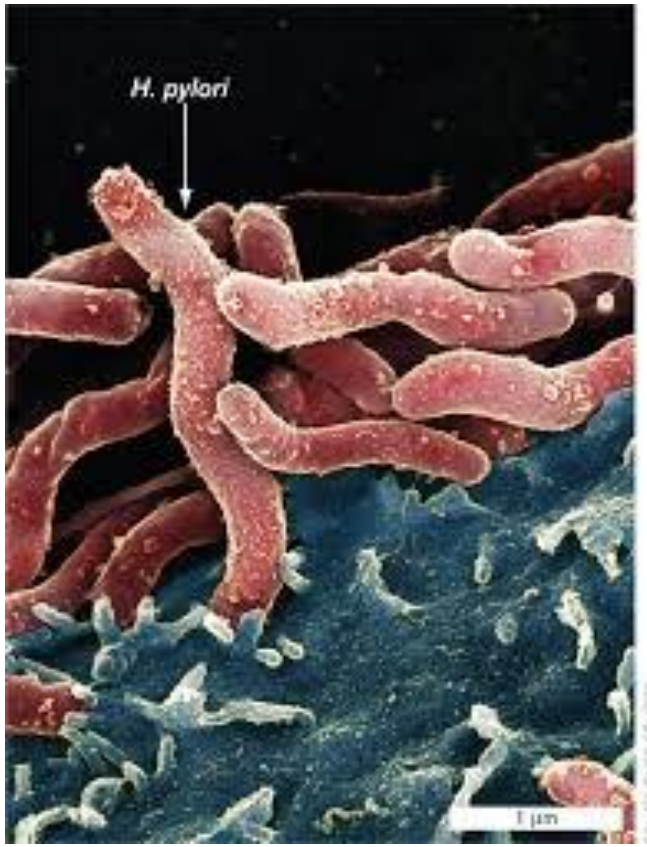


- ***Chronic skin ulcers:*** *increased abundance of Pseudomonadaceae in patients with chronic ulcers treated with antibiotics and an increased abundance of Streptococcaceae in diabetic ulcers*

Gastric microbiome

- *In H. pylori-positive persons, the presence of **H. pylori** dramatically reduces the overall diversity of the gastric microbiome.*

- *Its presence increases risk for developing peptic ulcer disease, gastric mucosa associated lymphoid tissue (MALT) tumors, and gastric adenocarcinoma.*



The colon microbiota and colorectal cancer

- *The **colonic microbiota** is suspected to be involved in the development of **colorectal cancers**, possibly by synthesizing short-chain fatty acids and other metabolites.*

- *In colorectal cancer samples, **Fusobacterium nucleatum**, a mucosally adherent, proinflammatory microbe, was significantly enriched, while both Bacteroidetes and Firmicutes were relatively depleted*

Published online: November 28, 2014

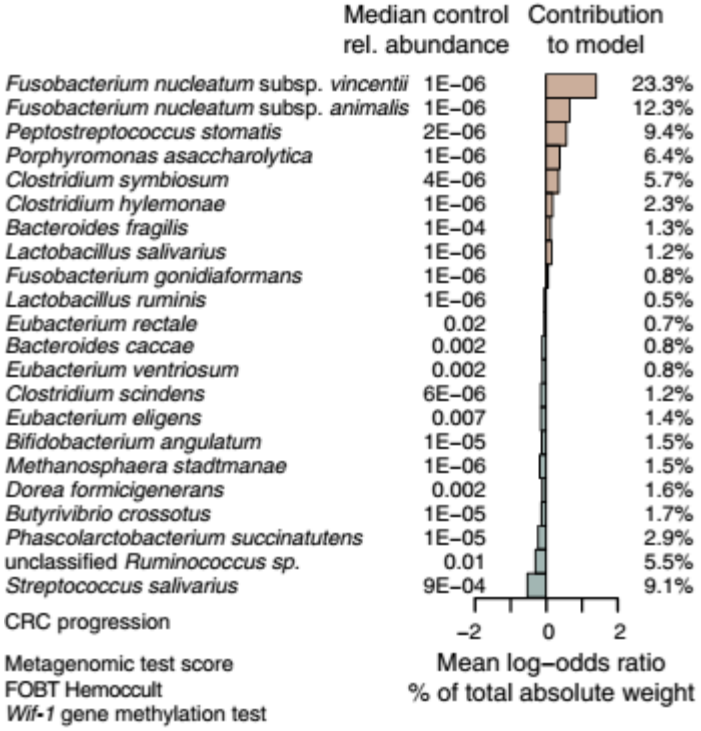
Article

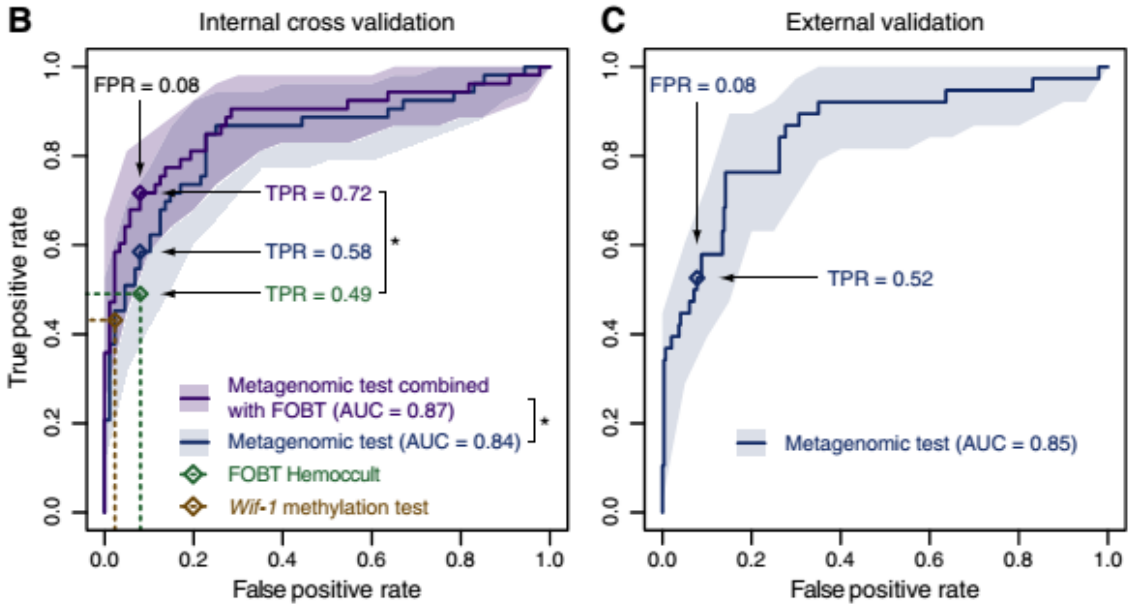
SOURCE DATA | TRANSPARENT PROCESS | OPEN ACCESS

molecular systems biology

Potential of fecal microbiota for early-stage detection of colorectal cancer

Georg Zeller[1,†], Julien Tap[1,2,†], Anita Y Voigt[1,3,4,5,†], Shinichi Sunagawa[1], Jens Roat Kultima[1], Paul I Costea[1], Aurélien Amiot[2], Jürgen Böhm[6,7], Francesco Brunetti[8], Nina Habermann[6,7], Rajna Hercog[9], Moritz Koch[10,‡], Alain Luciani[11], Daniel R Mende[1], Martin A Schneider[10], Petra Schrotz-King[6,7], Christophe Tournigand[12], Jeanne Tran Van Nhieu[13], Takuji Yamada[14], Jürgen Zimmermann[9], Vladimir Benes[9], Matthias Kloor[3,4,5], Cornelia M Ulrich[6,7,15], Magnus von Knebel Doeberitz[3,4,5], Iradj Sobhani[2,*] & Peer Bork[1,5,16,**]





The colon microbiota and inflammatory bowel disease

- ***Inflammatory bowel diseases*** *have long been associated to interactions between microbes and the host since the microbiome is essential for the activation of host immune responses*

- *Early childhood antibiotic exposure has been associated with significantly increased risk for* ***Crohn's disease****.*

- *Microbial diversity is significantly diminished in* ***Crohn's disease***

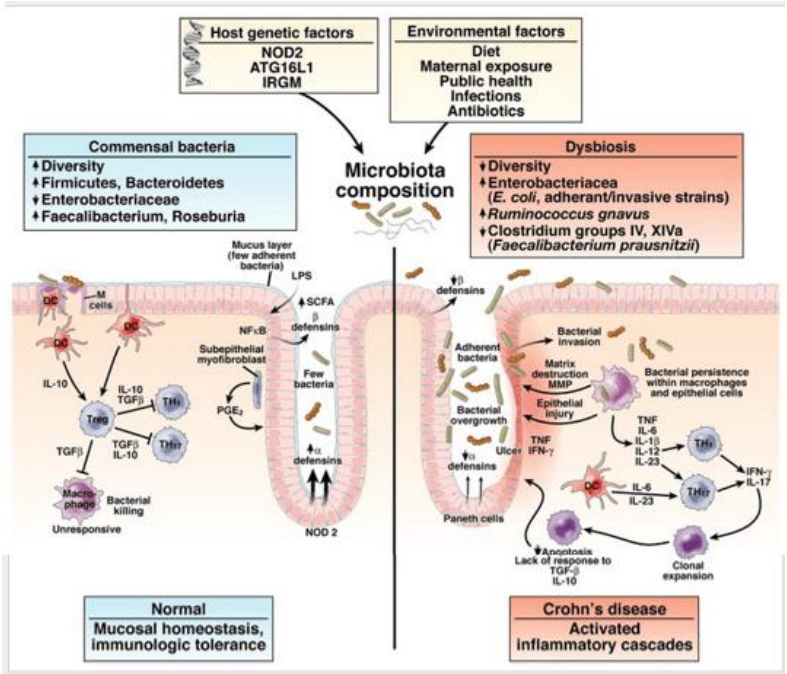


Figure 1

Genetic and environmental factors influence intestinal microbiota composition, which has an overall protective role in normal hosts, but contains an abnormal ratio of beneficial and aggressive bacterial species (dysbiosis) in Crohn's disease patients. This ratio of protective/ detrimental bacteria drives activated inflammatory cascades leading to Crohn's disease in genetically susceptible hosts and maintains mucosal homeostasis in normal subjects.

- ***Ulcerative colitis*** *affected patients had significantly reduced* ***bacterial diversity****, but increased proportions of* ***Actinobacteria*** *and* ***Proteobacteria****.*

Microbiome and HIV

Stockholm: External validation (STK)
84 subjects

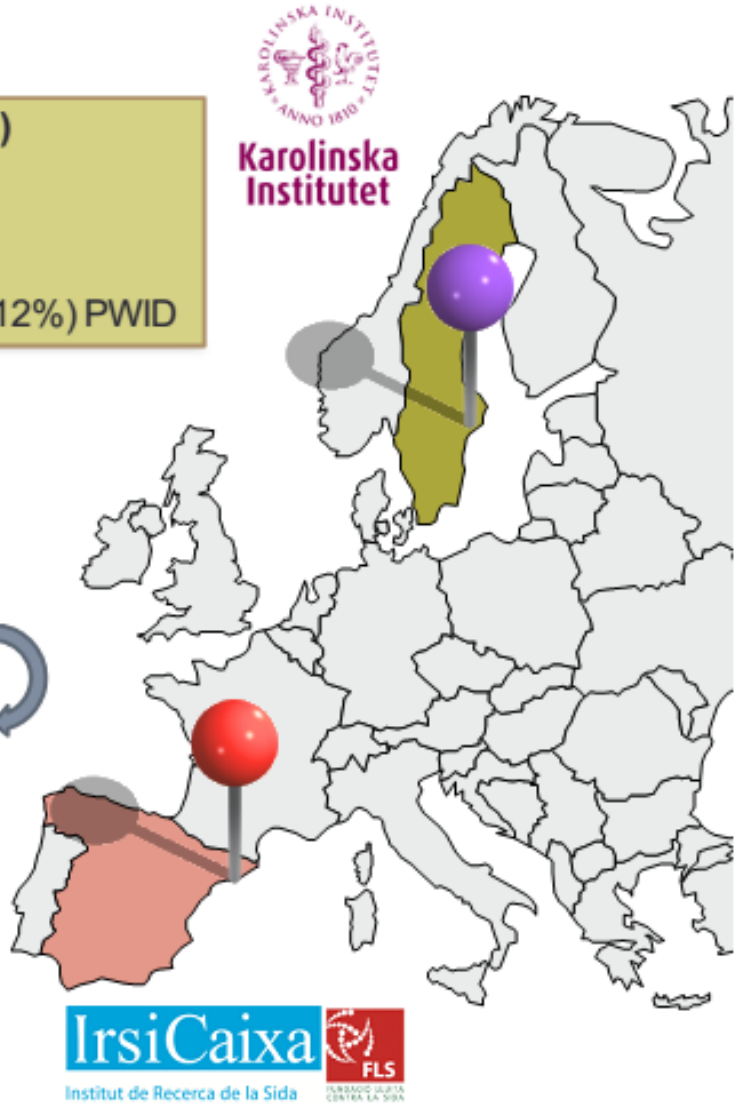
- 7 HIV-1-negative, all HTS
- 77 HIV-1-positive (including 3 EC)
- 19 (23%) MSM, **55 (66%) HTS**, 10 (12%) PWID

Barcelona: Internal validation (BCN1): same analysis 1 month later (n=110 subjects)!

Barcelona: Test dataset (BCN0)
156 subjects

- 27 HIV-1-negative, 85% MSM
- 129 HIV-1-positive with ≠ phenotypes
- **100 (64%) MSM**, 41 (26%) HTS, 15 (10%) PWID

** Study reviewed by the IrsiCaixa Community Advisory Committee*



Microbiome and HIV

SCIENTIFIC REPORTS

OPEN

ed: 26 January 2017
ed: 15 June 2017

Richer gut microbiota with distinct metabolic profile in HIV infected Elite Controllers

Jan Vesterbacka[1], Javier Rivera[2], Kajsa Noyan[3], Mariona Parera[2], Ujjwal Neogi[3], Malu Calle[4], Roger Paredes[2,4,5,6], Anders Sönnerborg[1,3], Marc Noguera-Julian[2,4] & Piotr Nowak[1]

- ***Elite controllers*** *(EC): the HIV-infected individuals with sustained viral suppression in absence of antiretroviral therapy (ART)*
- *The* ***microbiota of EC is diferent*** *from individuals with progressive infection and more similar to HIV negative individuals*
- *EC have* ***richer gut microbiota*** *than untreated HIV patients, with unique bacterial signatures and a distinct metabolic profile which may contribute to control of HIV*

How is the microbiome studied? Composition versus function

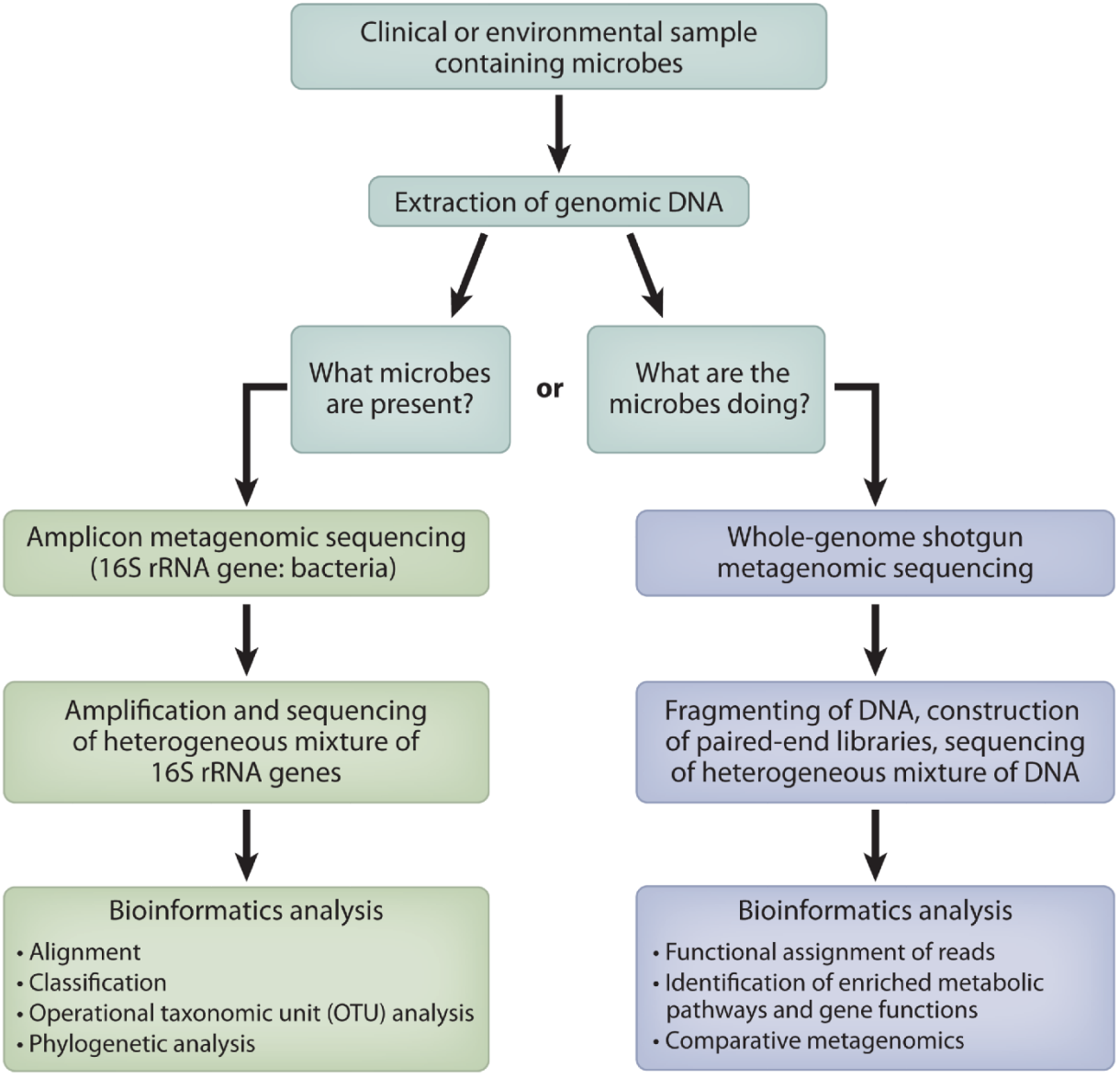


Figure 1.
Workflow for metagenomic sequencing and analysis projects.

Grice2013

How is the microbiome studied? Composition versus function

Amplicon sequencing -> Composition analysis:

- *Sequencing of an amplified (PCR)* ***marker gene***

Target	Gene/ region
Bacteria	16S
Archaea	16S
Fungi	ITS

- *Identification of the microbes that are present in a given community*
- *Raw and relative abundance of each microbe species or taxa*

➔ **OTU or ASV table**

How is the microbiome studied? Composition versus function

Shotgun sequencing -> Functional analysis:

- *Assess the functional potential of an entire microbial community*
- *Instead of using PCR to amplify a particular phylogenetic marker, the **DNA sequence of the entire community is sequenced** directly using high throughput techniques.*
- *This provides a catalog of all of the genes present in the microbial community.*
- *The relative abundance of specific metabolic pathways are used to predict the functional capacity of that community*

➔ **Pathway table**

Microbiome study

Microbial DNA extraction and sequencing

amplicon sequencing

shotgun sequencing

Marker gene data (16S, 18S, ITS)

Metagenomics data

Bioinformatics: sequence processing

qiime Quantitative Insights Into Microbial Ecology

mothur

PICRUSt

MetaPhlAn2

HUMAnN2

Microbiome abundance

Microbiome function

Species abundance (OTU table)

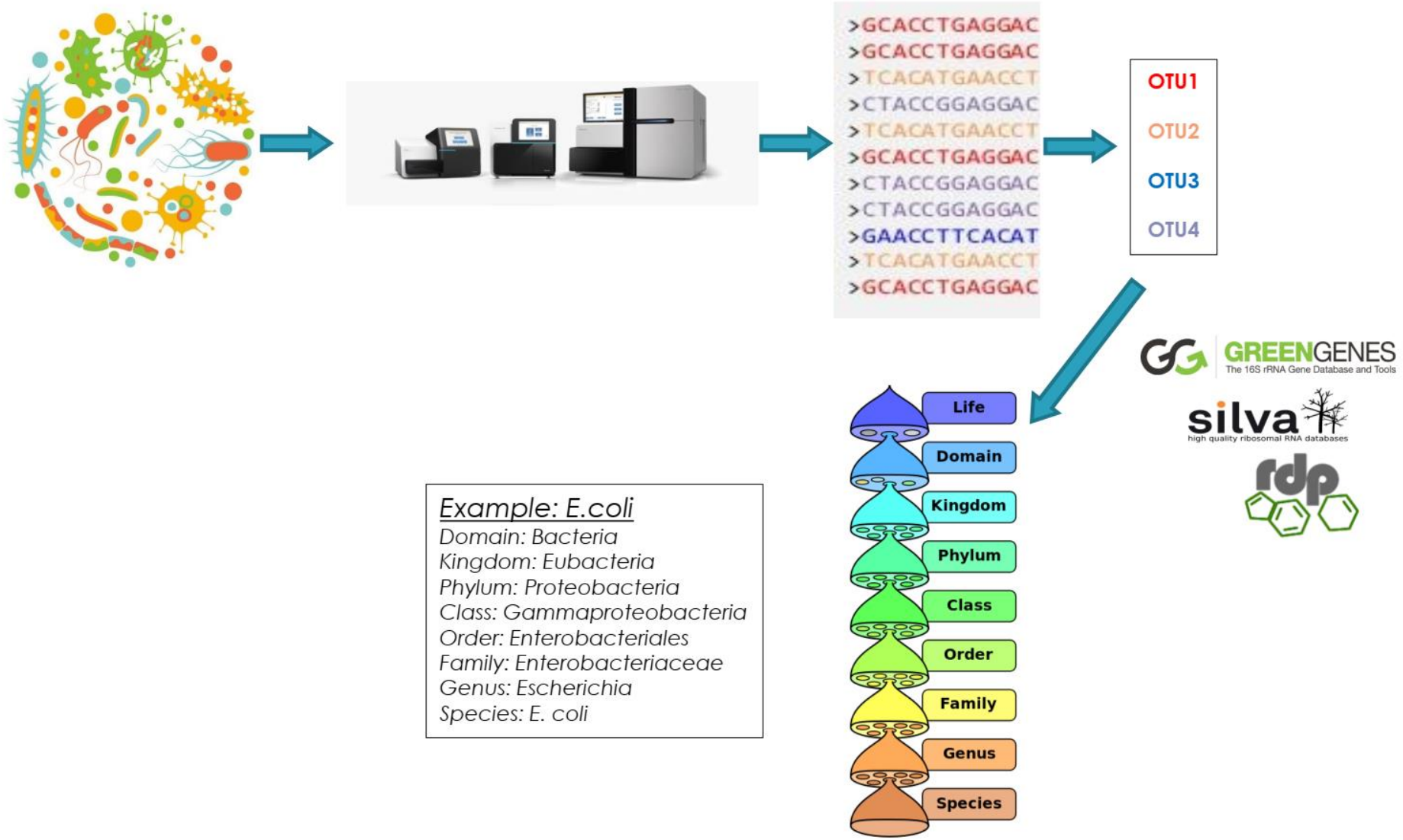
Gene abundance (pathway table)

Statistical analysis

R

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

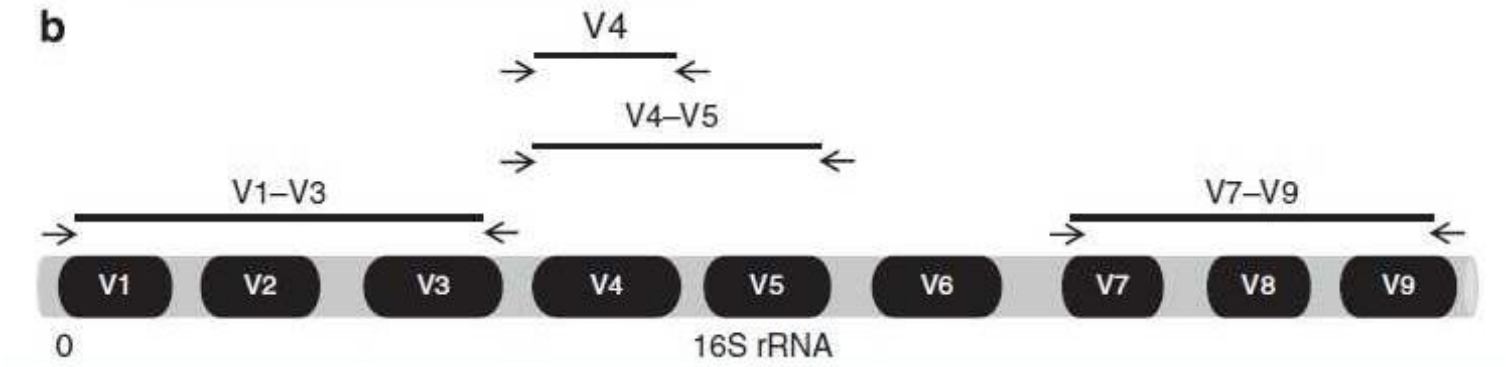
Microbiome analysis



16S ribosomal RNA gene

***16S rRNA gene**: is found in all bacteria and archaea and encodes the RNA component of the small ribosomal subunit*

- *Contains both conserved and hypervariable regions:*
  - *The conserved regions can be targeted with PCR primers*
  - *The hypervariable regions are taxa-specific and allow to distinguish the different microbes*
  - *The V1–V3 and V4 regions are most commonly targeted*



Black: Hypervariable regions; Gray: Conserved regions *Tyler2014*

Microbial DNA extraction and sequencing

- *DNA is extracted from a sample of the microbial community*
- *Polymerase chain reaction (PCR): Primers target conserved regions of the 16S gene to amplify most of the microbial species present*
- *PCR amplicons are subjected to high throughput DNA sequencing*

- *Selection of **primers** is relevant since some primers may be biased for or away from some species*

SEQUENCING PLATFORMS

- *Three major factors are important when choosing a sequencing platform:*
  - *sequencing depth*
  - *read length*
  - *error rates*

Microbial DNA extraction and sequencing

- ***Sequencing depth**: the number of sequences generated by a run*
  - *Large sequencing depth are required for the identification of rare microbial taxa.*
  - *Currently sequencing platforms produce between 10^6 and 3 x 10^9 reads per run*
- ***Read length**:*
  - *Longer reads are more reliable for identifying microbial taxonomy but tend to have higher error rates.*
  - *Read length and sequencing depth tend to be inversely related*
  - *Researchers currently sacrifice read length for sequencing depth when choosing a sequencing platform.*
- ***Error rate**:*
  - *Error rates on sequencing platforms range up to 1%*
  - *These errors can have a marked effect on datasets for which millions of reads are being generated.*

Microbial DNA extraction and sequencing

Illumina-sequencing platforms: *Most popular: more reads per sample and similar error rates compared to alternatives.*

MiSeq*: 12–15 million 150–350 base pair single reads per run*

HiSeq*: 5 billion 100–150 base pair single reads per run*

Pyrosequencing (e.g., Roche 454)

One of the first sequencing technologies used for microbiome analyses

- *Larger read length than illumina: 400–700 bp reads*
- *Lower sequencing depth: 1 million reads per run*
- *Higher error rate*

Use of 454 technology is currently limited to studies that depend heavily on long reads to address research questions

Other sequencing platforms: *IonTorrent, PacBio, and SOLiD*

Bioinformatic pipelines

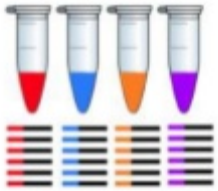
- *There are a number of bioinformatic pipelines available for processing microbiome sequence data, the two most popular are QIIME and mothur.*

- *Both pipelines are user-friendly and produce similar results, the choice is really a matter of user preference.*

1. *Preprocessing and quality control filtering*
2. *OTU binning*
3. *OTU table*

1. Preprocessing and quality control filtering:

- *Assign sequence to samples (demultiplexing)*



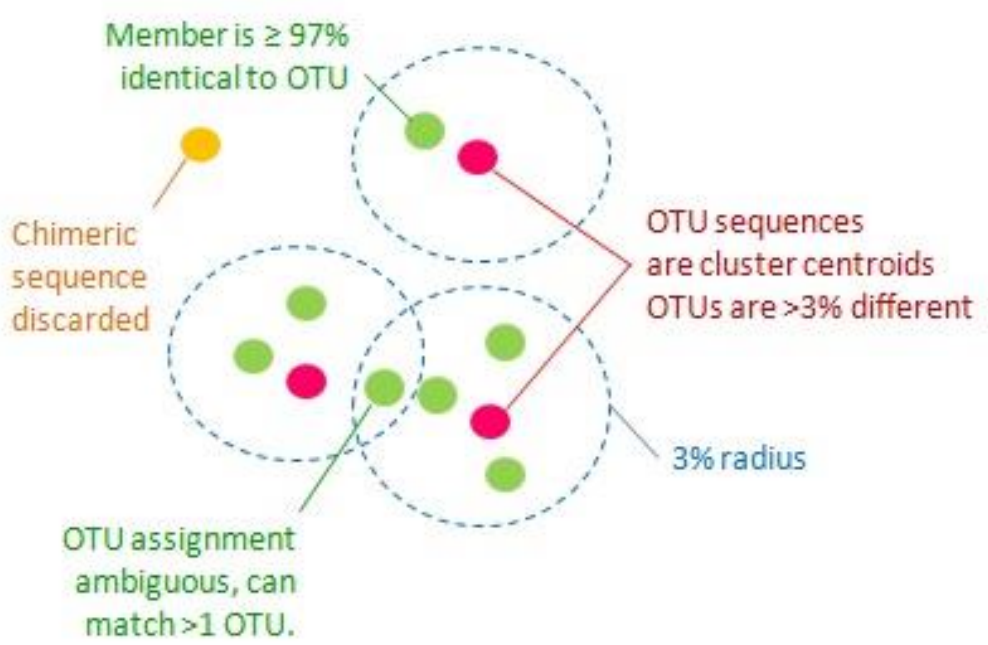
qiime.org

- *Raw sequence data must be quality filtered to reduce noise and avoid inflation of diversity estimates:*
- *Remove too short sequences, too many ambiguous base pairs, chimeras...*

2. OTU binning and OTU table

- *Quality filtered sequences are clustered into **operational taxonomic units** (OTUs).*

- *OTUs are analogous to microbial species but OTUs are defined based on DNA sequence similarity (usually, 97% similarity)*



Clustering strategies and algorithms

- *closed-reference*
- *de novo*
- *open-reference*

- ***Closed-reference OTU picking:***
  - *uses databases of known microbial 16S sequences such as GreenGenes, SILVA, or RDP.*
  - *Every sequence is compared to the sequences in the selected database and assigned to a certain microbial species if it passes the 97% similarity criteria.*
  - *Any sequence that does not match the database of known sequences is discarded.*
  - *The advantage of this method is speed and consistency.*
  - *An obvious disadvantage is that a portion of the sequence data is discarded: novel, uncharacterized microbes will not be included in final analyses.*

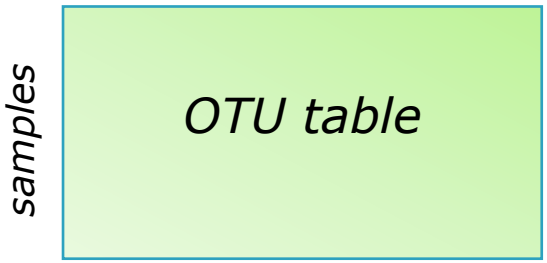
- ***De novo OTU-picking:***
  - *Clusters sequences without a database using a clustering algorithm*
  - *All sequences are compared to each other and assigned into clusters of 97% similarity.*
  - *This method is much slower than closed-reference OTU picking but no sequences are discarded and no biases in clustering are introduced from an external database.*

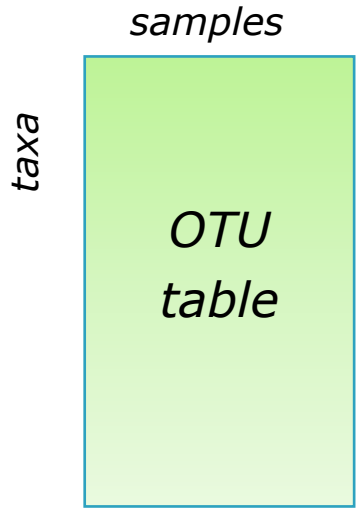
- ***Open-reference OTU-picking:***
  - *Sequences are first clustered using the closed-reference method, and those that do not match the database at 97% similarity are clustered de novo.*
  - *The closed reference step shortens the processing time substantially compared to de novo OTU-picking.*
  - *Therefore, open reference OTU-picking is commonly utilized.*

OTU table

- *Once OTU clustering is completed, an OTU abundance table is built.*
- ***OTU table**: table of counts = number of reads of the different OTUs for each sample.*

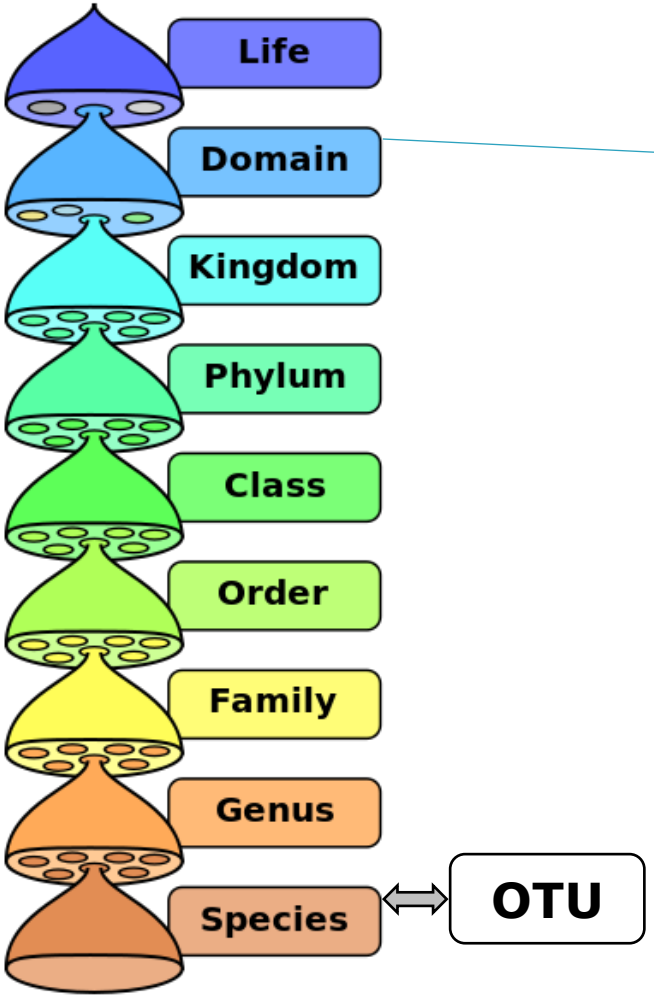
	OTU1	*OTU2*	*OTU3*	*...*	*OTUK*	*TOTAL*
Sample1	X_{11}	X_{12}	X_{13}	...	X_{1k}	N_1
Sample2	X_{21}	X_{22}	X_{23}	...	X_{2k}	N_2
...				...		
Samplep	X_{p1}	X_{p2}	X_{p3}	...	X_{pk}	N_p

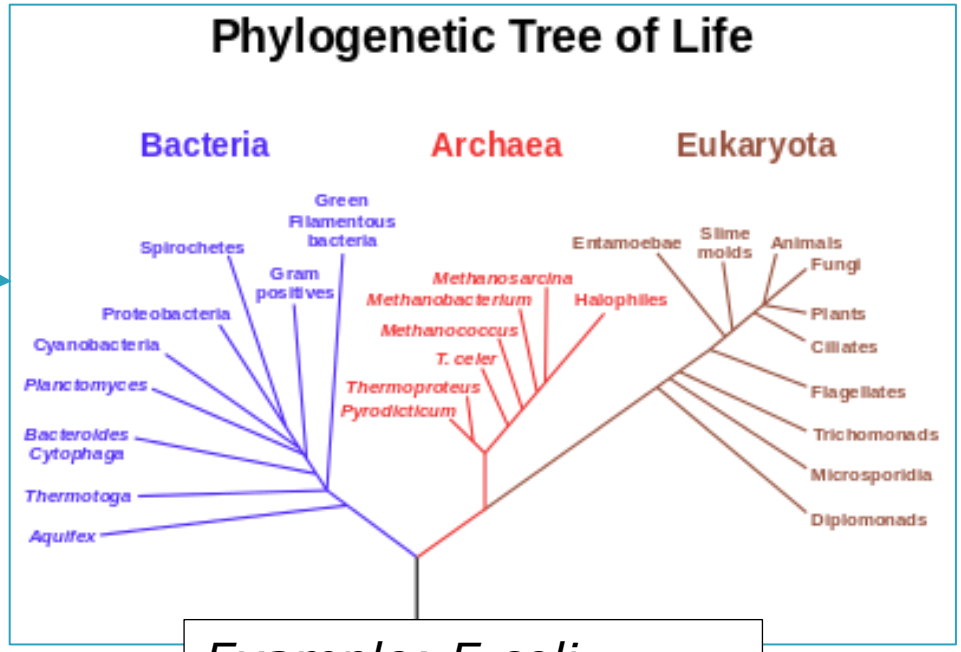




3. Taxonomy assignment

- *Representative sequences from each cluster are compared to a database to determine the taxonomic assignment.*





Example: E.coli

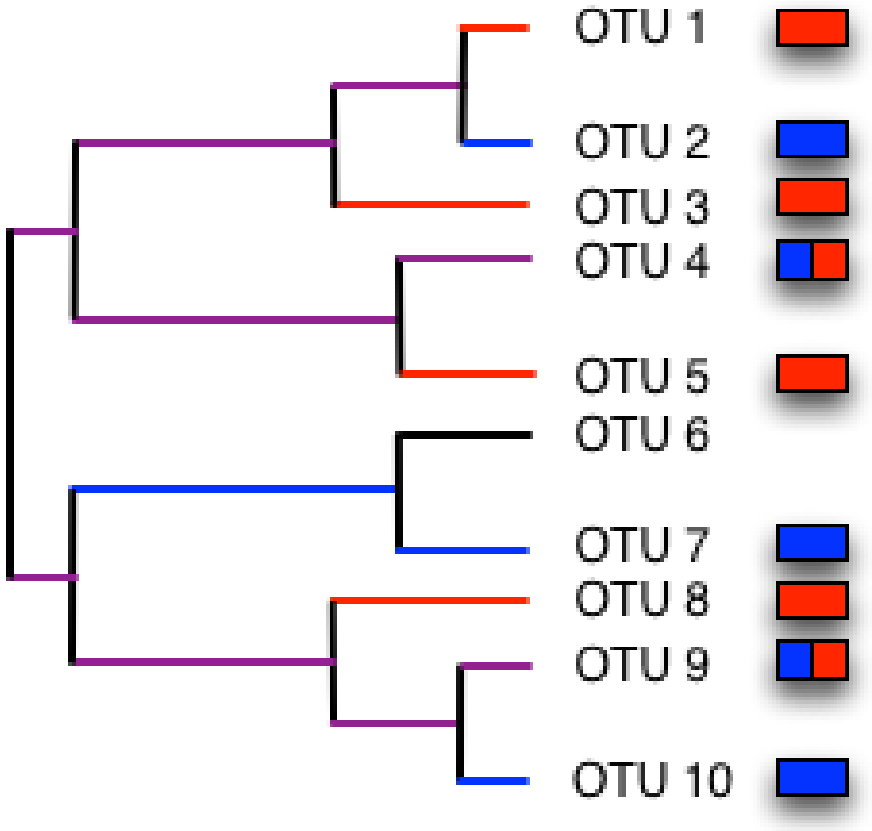
Domain:	*Bacteria*
Kingdom:	*Eubacteria*
Phylum:	*Proteobacteria*
Class:	*Gammaproteobacte*
Order:	*Enterobacteriales*
Family:	*Enterobacteriaceae*
Genus:	*Escherichia*
Species:	*E. coli*

Taxonomy assignment

Taxonomy Table: [6 taxa by 7 taxonomic ranks]:

	Kingdom	Phylum	Class	Order	Family	Genus	Species
239739	"Bacteria"	"Actinobacteria"	"Actinobacteria"	"Actinomycetales"	"Nocardioidaceae"	NA	NA
218299	"Bacteria"	"Actinobacteria"	"Actinobacteria"	"Actinomycetales"	"Nocardioidaceae"	NA	NA
158804	"Bacteria"	"Actinobacteria"	"Actinobacteria"	"Actinomycetales"	"Nocardioidaceae"	"Marmoricola"	NA
32247	"Bacteria"	"Actinobacteria"	"Actinobacteria"	"Actinomycetales"	"Nocardioidaceae"	"Marmoricola"	NA
305897	"Bacteria"	"Actinobacteria"	"Actinobacteria"	"Actinomycetales"	"Nocardioidaceae"	"Kribbella"	"Kribbellaaluminosa"
248838	"Bacteria"	"Actinobacteria"	"Actinobacteria"	"Actinomycetales"	"Nocardioidaceae"	"Kribbella"	NA

4. Phylogenetic tree



http://readiab.org/book/latest/

Phylogenetic trees are used to obtain phylogenetic distances between samples (UniFrac and weighted UniFrac)

Microbiome statistical analysis

Main goals of microbiome statistical analysis

- *Analysis and visualization of diversity of microbial communities*
- *Identification of possible data structures*
- *Microbiome differential abundance testing:*
  - *Multivariate community analysis: Are there global differences in microbial composition between sample groups?*
  - *Univariate testing: Which taxa are differentially abundant between sample groups? Which taxa is correlated with a given continuous variable?*

Main goals of microbiome statistical analysis

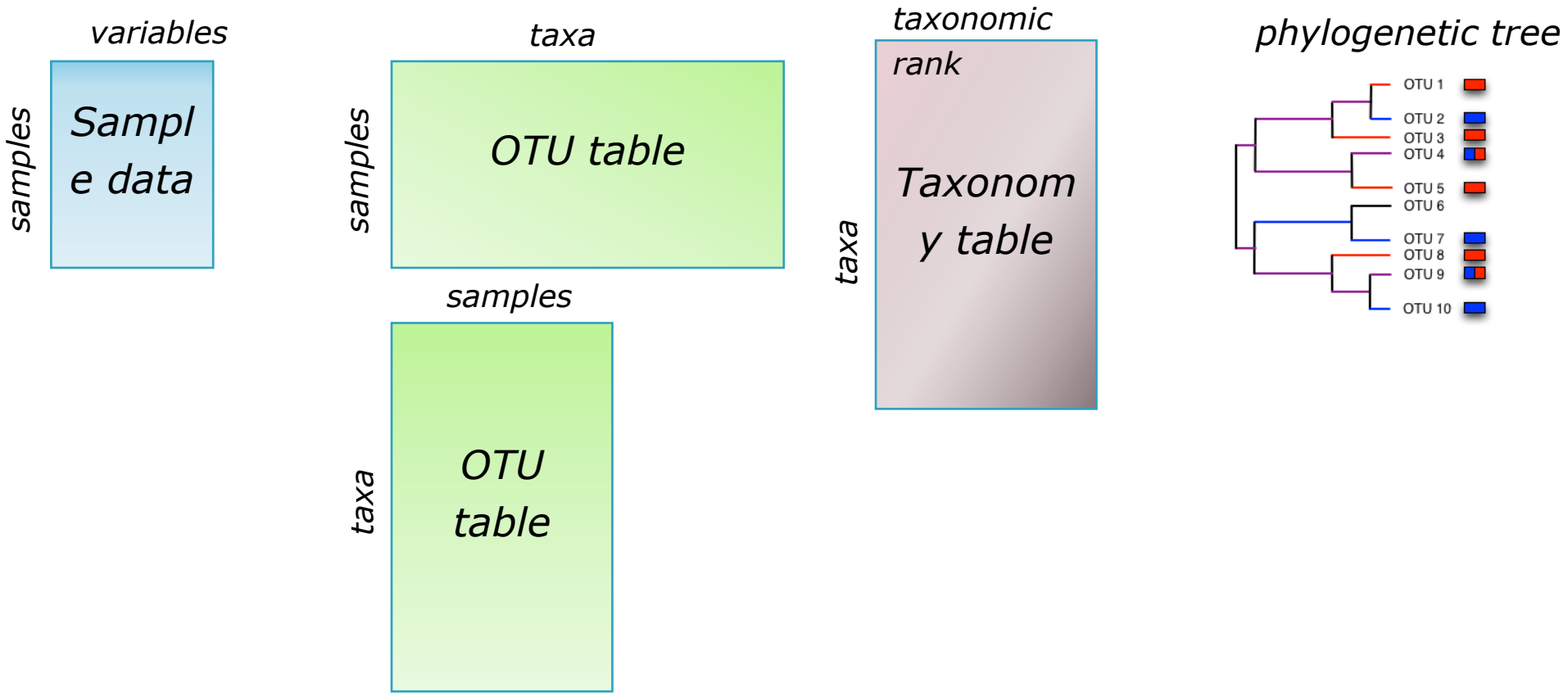
- *Analysis and visualization of diversity of microbial communities. Identification of possible data structures*
  - → *Abundance tables, alpha and beta diversity measures*
  - → *Ordination plots, heatmaps*
- *Microbiome differential abundance testing:*
  - *Multivariate community analysis: Are there global differences in microbial composition between sample groups?*
    - → *Adonis=PERMANOVA*
  - *Univariate testing: Which taxa are differentially abundant between sample groups?*
    - → *LEfSe, DESeq2, edgeR*

Challenges of microbiome statistical analysis

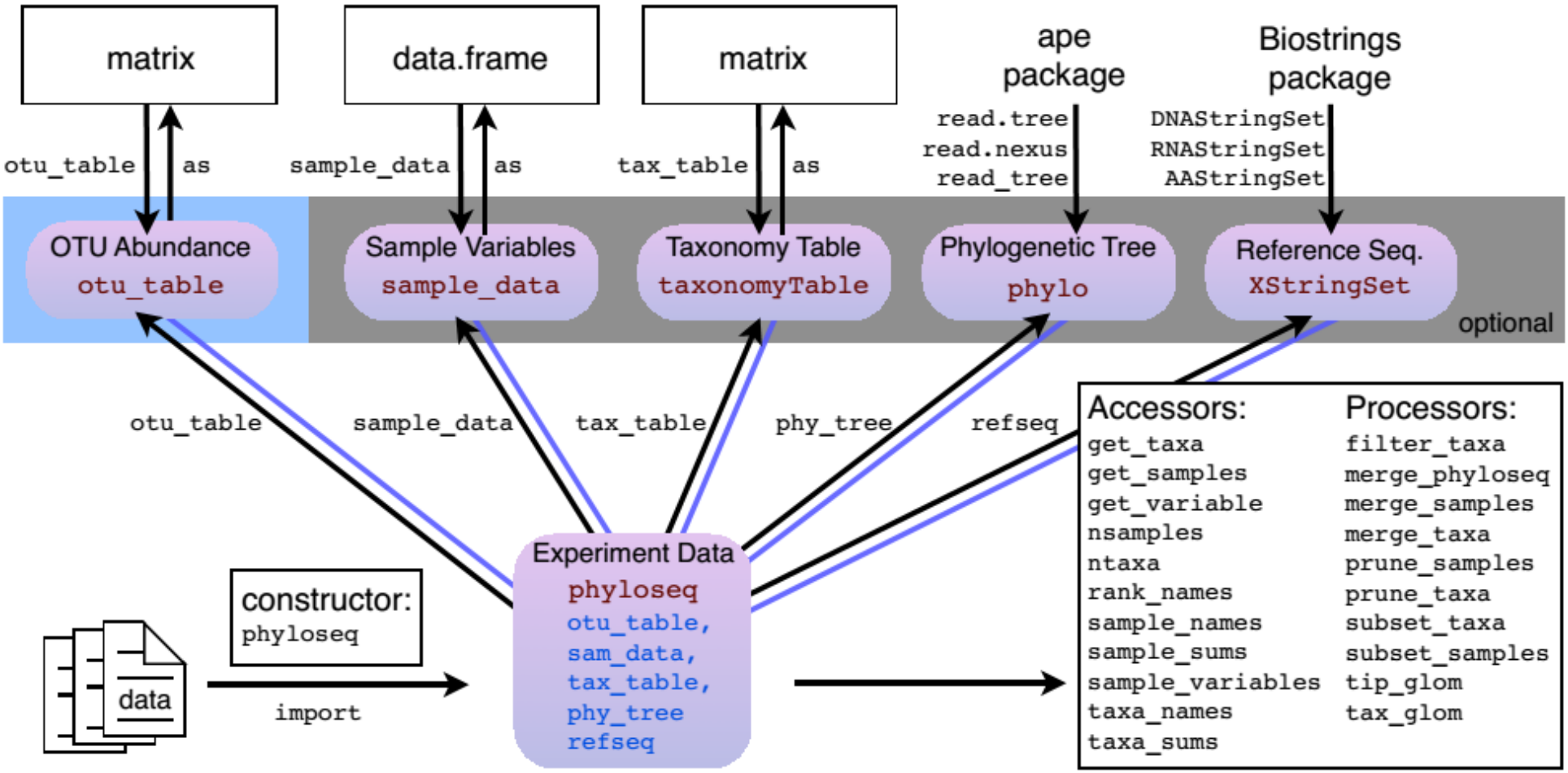
- *Sparsity: large proportion of zeros in OTU*
  → *Preprocessing*
- *Large differences in total counts per sample (sequencing depth, library size)*
  → *Normalization*
- *Heteroscedasticity: mean-variance relationship*
  → *Normalization*
- *Influence of dominant species*
  → *Transformation*
- *Complex phylogenetic structure*
  → *Multivariate analysis*
- *Compositional data*
  → *CODA methods:*
  ANCOM, Selbal, coda4microbiome

Elements of microbiome abundance data

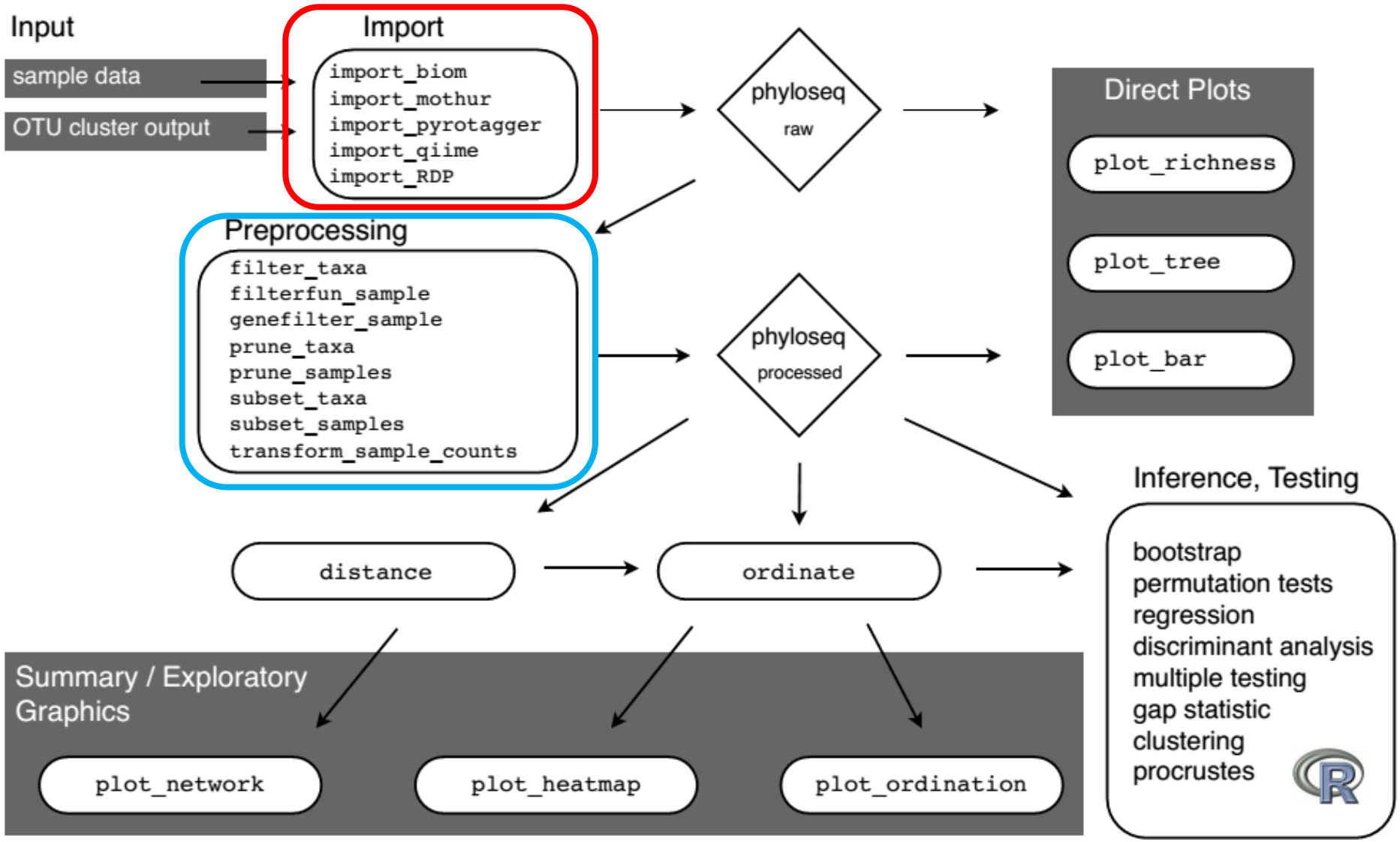
*Sample data, **OTU table**, Taxonomy table, Phylogenetic tree*



Phyloseq R package

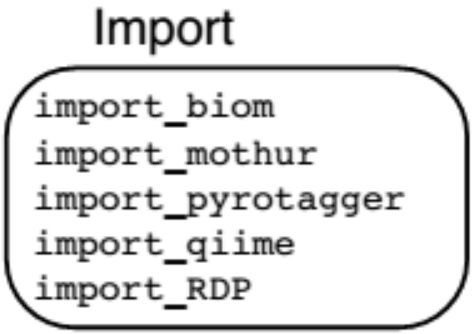


Import data and preprocessing



Import data:

Biom format http://biom-format.org/



`data-GlobalPatterns`	*(Data) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample (2011)*

Description

Published in PNAS in early 2011. This work compared the microbial communities from 25 environmental samples and three known "mock communities" – a total of 9 sample types – at a depth averaging 3.1 million reads per sample. Authors were able to reproduce diversity patterns seen in many other published studies, while also invesitigating technical issues/bias by applying the same techniques to simulated microbial communities of known composition.

Feces: 4 *Skin: 3* *Tongue: 2*

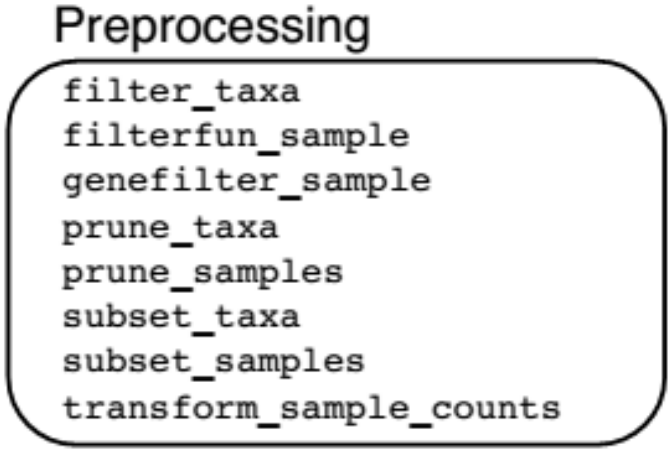
Freshwater: 2 *Freshwater (creek): 3* *Ocean: 3* *Sediment (estuary): 3* *Soil: 3*

Mock: 3

```
> data(GlobalPatterns)
> data<-GlobalPatterns
> dim(data@otu_table)

[1] 19216    26

OTU Table:          [6 taxa and 26 samples]
                     taxa are rows

       CL3 CC1 SV1 M31Fcsw M11Fcsw M31Plmr M11Plmr F21Plmr M31Tong M11Tong LMEpi24M SLEpi20M
549322   0   0   0       0       0       0       0       0       0       0         0        1
522457   0   0   0       0       0       0       0       0       0       0         0        0
951      0   0   0       0       0       0       1       0       0       0         0        0
244423   0   0   0       0       0       0       0       0       0       0         0        0
586076   0   0   0       0       0       0       0       0       0       0         0        0
246140   0   0   0       0       0       0       0       0       0       0         0        0

       AQC1cm AQC4cm AQC7cm NP2 NP3 NP5 TRRsed1 TRRsed2 TRRsed3 TS28 TS29 Even1 Even2 Even3
549322     27    100    130   1   0   0        0        0        0    0    0     0     0     0
522457      0      2      6   0   0   0        0        0        0    0    0     0     0     0
951         0      0      0   0   0   0        0        0        0    0    0     0     0     0
244423      0     22     29   0   0   0        0        0        0    0    0     0     0     0
586076      0      2      1   0   0   0        0        0        0    0    0     0     0     0
246140      0      1      3   0   0   0        0        0        0    0    0     0     0     0
```

79% zeros

Preprocessing:



Reduce extrem sparsity: *Remove extremely rare OTUs or samples with very few counts.*

No standard rules for preprocessing.

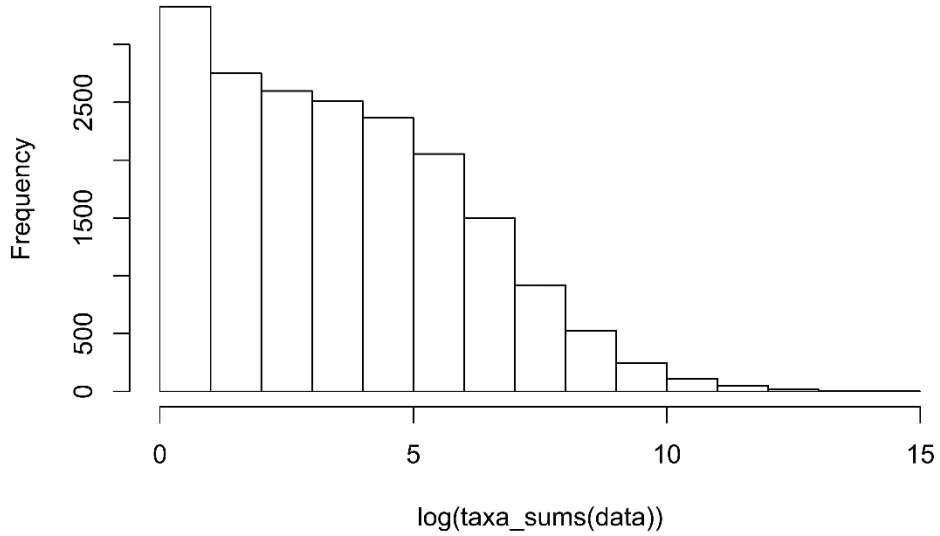
Examples: "Remove OTUs present in <1% of the samples and samples with less than 20 counts"

Normalization

Intended to enable meaningful comparison of data

with large variability in total counts pe

sample (sequencing depth, library size



Normalization:

- *None*
- *Relative abundances (total count s*
- *Size factors: median of ratios (DESeq2)*
- *Rarefaction: equal number of sequences are randomly selected from each sample such that all samples have the same number of total counts.*

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes*

Statistics Department, Stanford University, Stanford, California, United States of America

*"Despite its current popularity in microbiome analyses **rarefying biological count data is statistically inadmissible** because it requires the omission of available valid data."*

Weiss *et al. Microbiome* (2017) 5:27
DOI 10.1186/s40168-017-0237-y

Microbiome

RESEARCH Open Access

Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss[1], Zhenjiang Zech Xu[2], Shyamal Peddada[3], Amnon Amir[2], Kyle Bittinger[4], Antonio Gonzalez[2], Catherine Lozupone[5], Jesse R. Zaneveld[6], Yoshiki Vázquez-Baeza[7], Amanda Birmingham[8], Embriette R. Hyde[2] and Rob Knight[2,7,9*]

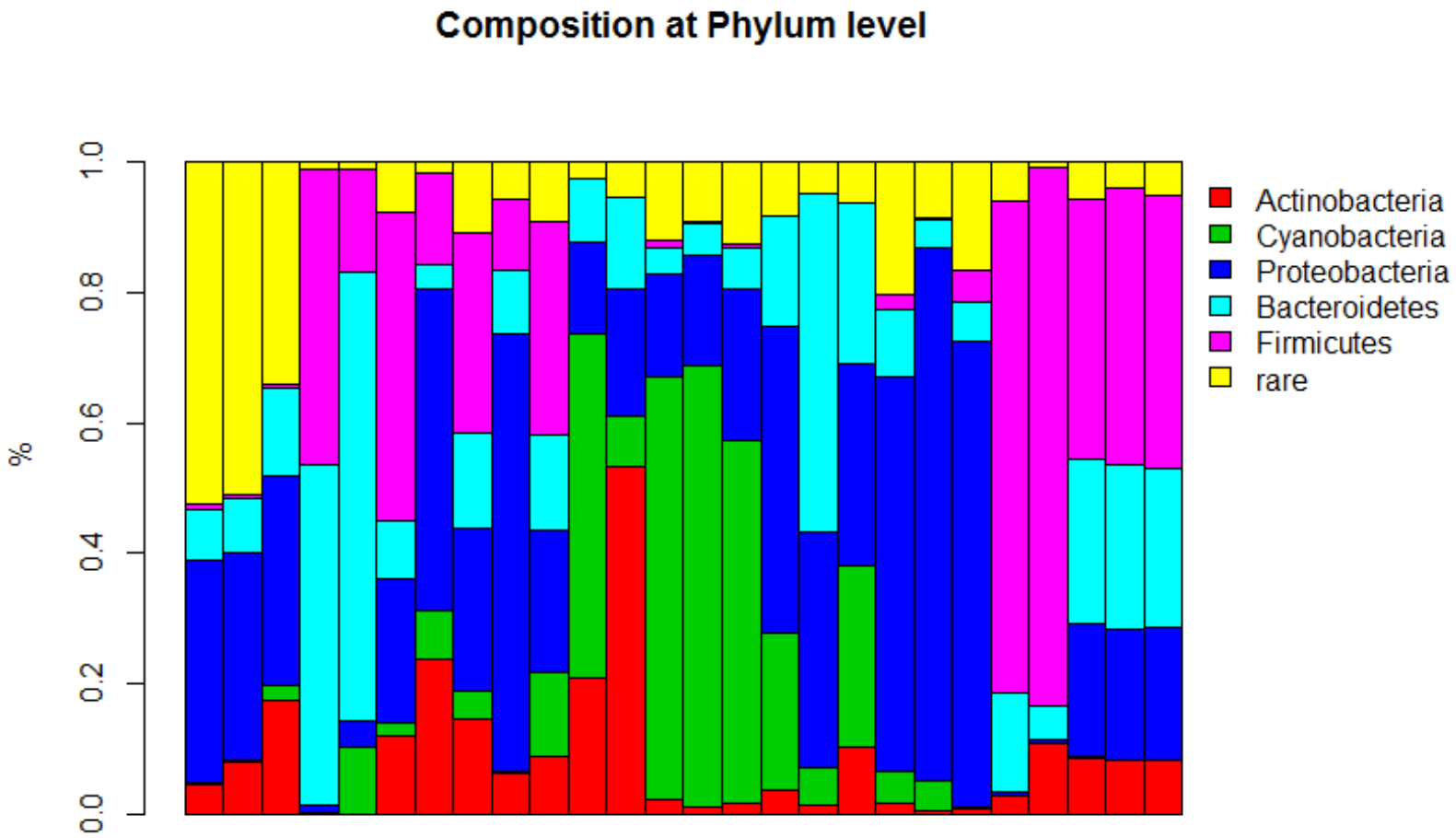
Transformation

Data can be transformed in order to reduce the strong influence of dominant species.

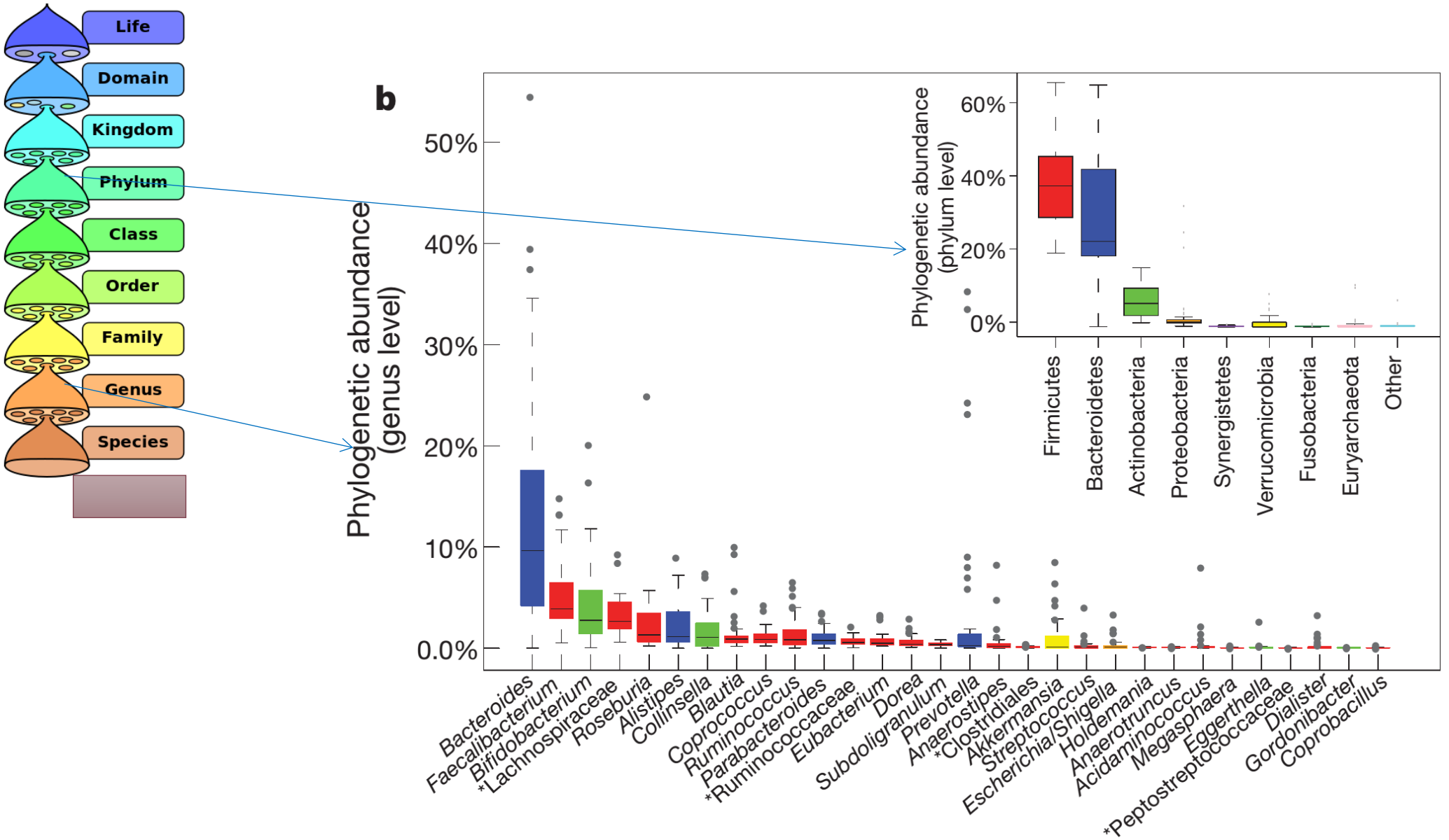
- *log(1+x) or other pseudocounts (0.01 or 0.0001). Values can be corrected by substracting the log of the pseudocount to preserve original zeros.*
- *square root*
- *cubic root*

Exploratory analysis - Abundance plots

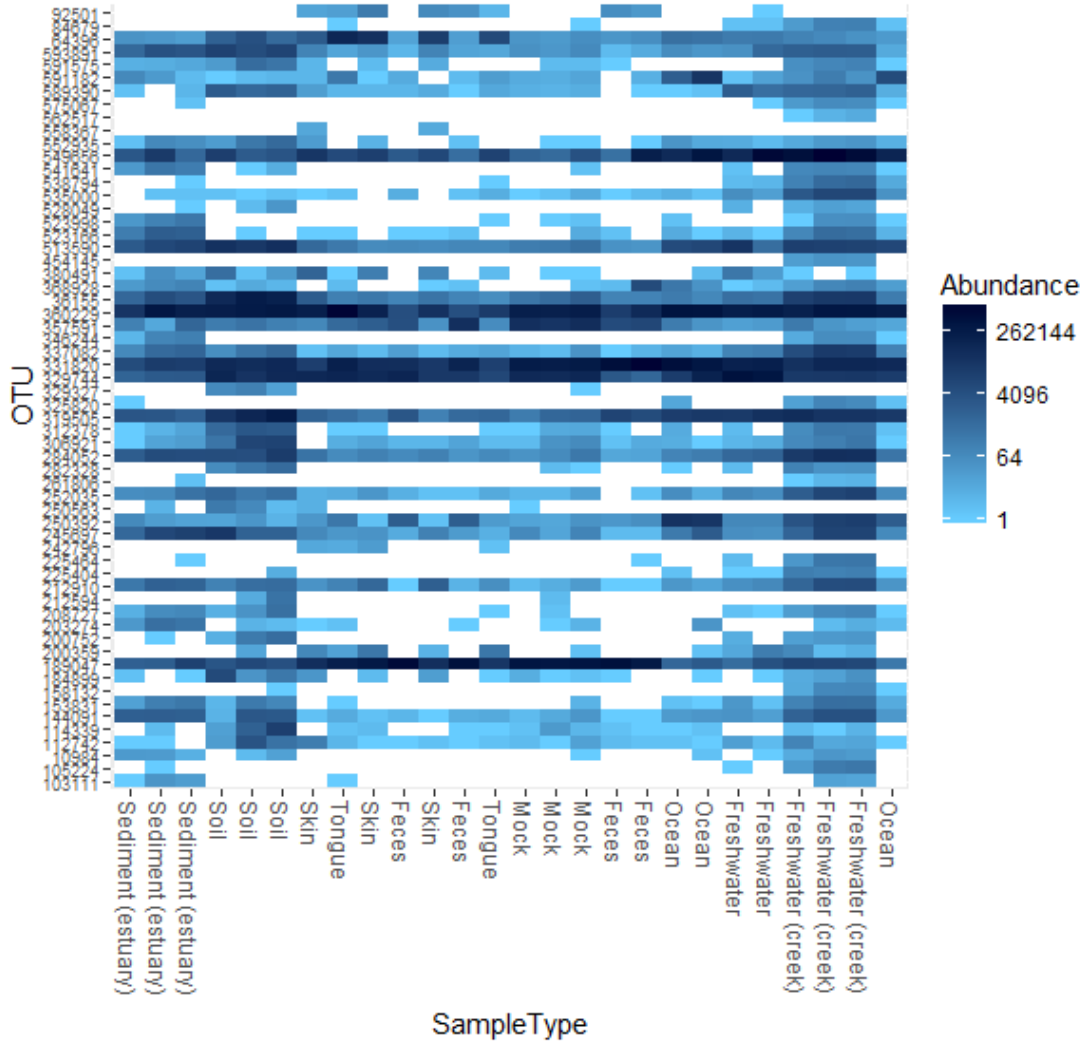
Bar plots



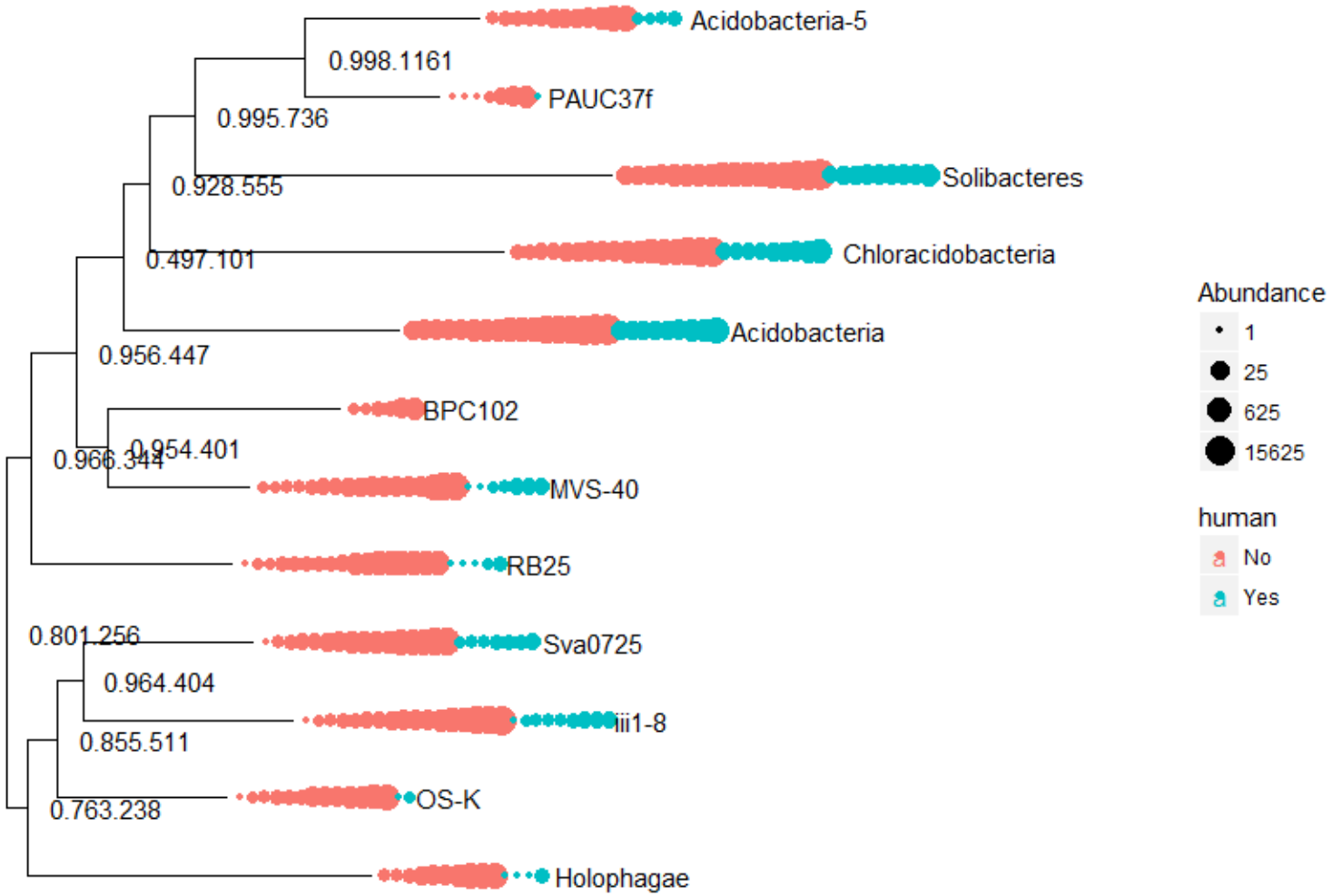
Boxplots: abundances for different taxonomic levels



Heatmaps: abundances of taxa by sample



Exploratory analysis - Plot phylogenetic tree



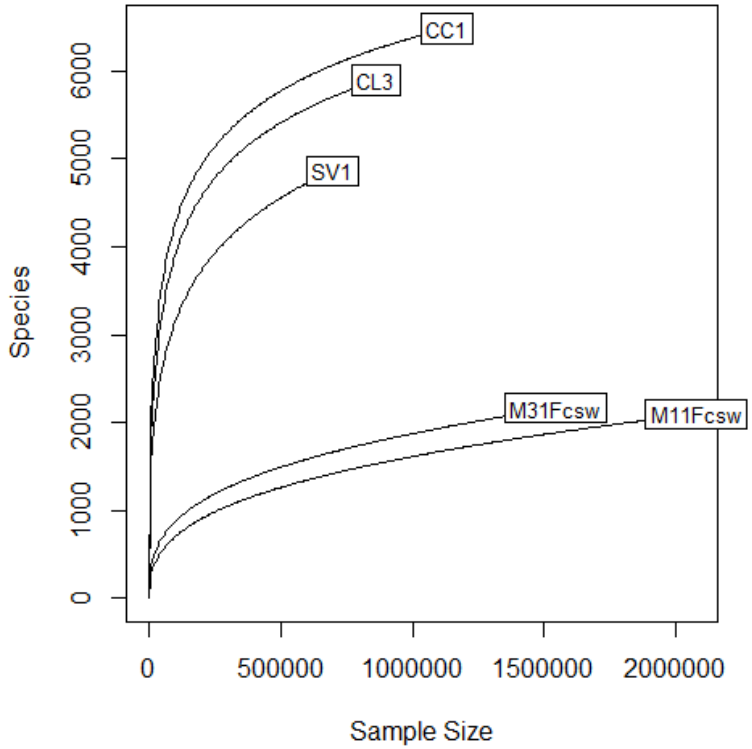
Ecological measures of richness and diversity

Alpha diversity: within sample diversity

***Richness**: R=Number of different taxa (OTU) in a sample*

***Observed**: R_{Obs} = Observed number of different taxa (OTU) in a sample*

Rarefaction curves: *Plot observed richness for different rarefaction depths. Provides information on the required sequencing depth.*



Chao1*: bias corrected richness estimator*

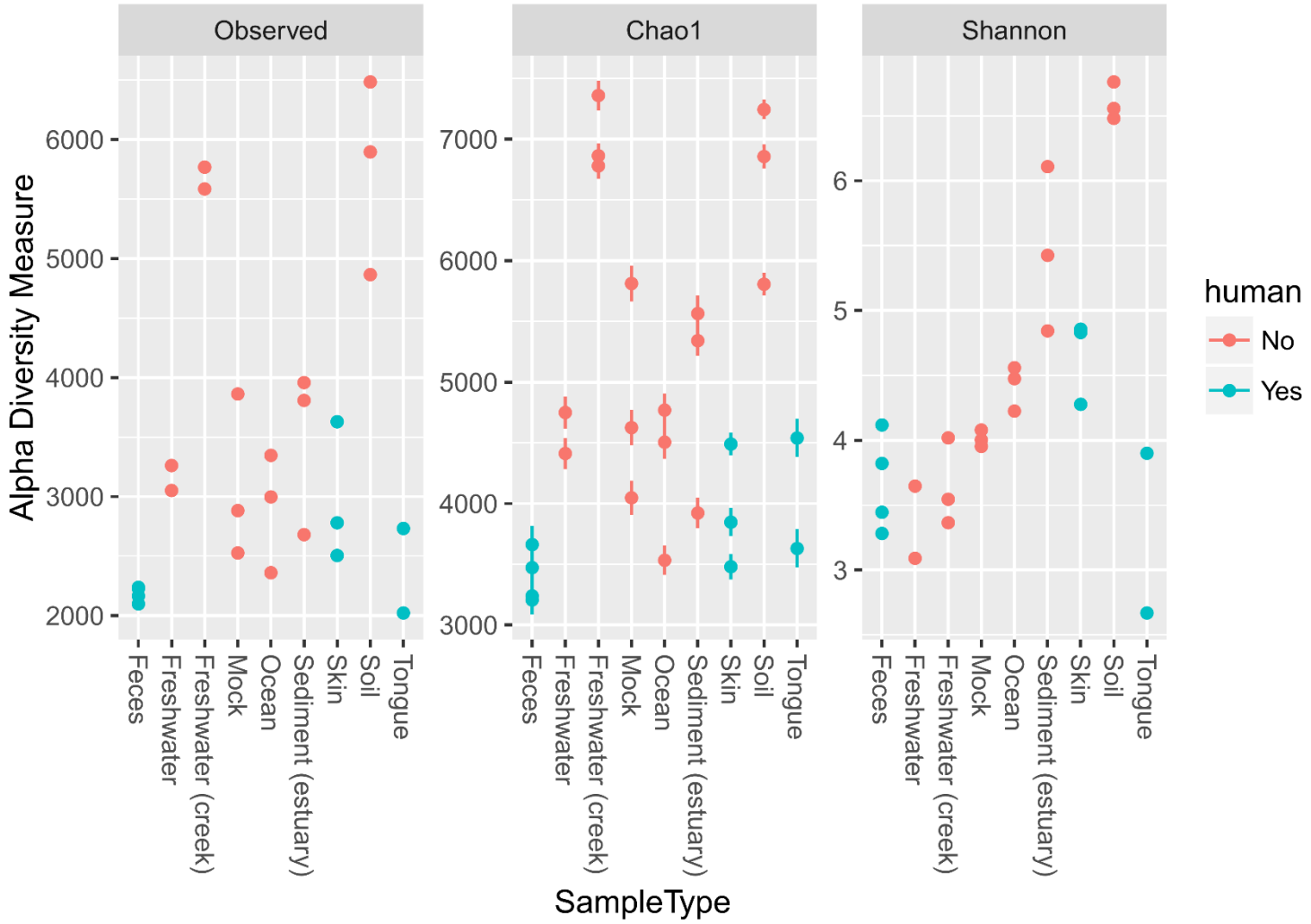
$$R_{Chao1} = R_{Obs} + \frac{f_1(f_1-1)}{2(f_2+1)}$$

f_1 *= number of species observed only once*

f_2 *= number of species observed only twice*

Shannon:

$$R_{Shannon} = -\sum_{i=1}^{k} p_i \, log \, p_i$$



Beta diversity: between samples diversity

***Ecological distance**: A good ecological distance describes the difference in species composition. For sites that share most of their species, the ecological distance should be small. When sites have few species in common, the ecological distance should be large[1].*

***Euclidean distance:** The Euclidean distance is not a good ecological distance if it is used on raw species matrices[1].*

Site	Species 1	Species 2	Species 3
A	1	1	0
B	5	5	0
C	0	0	1

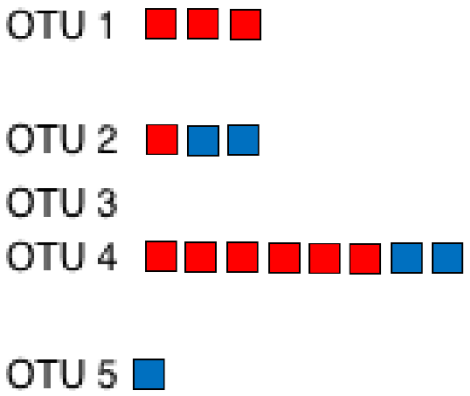
	A	B	C
A	0	5.656854	1.732051
B	5.656854	0	7.141428
C	1.732051	7.141428	0

Bray-Curtis dissimilarity

$$D_{BC}(A,B) = \frac{\sum_i |A_i - B_i|}{\sum_i (A_i + B_i)}$$

Jaccard dissimilarity

$$D_J(A,B) = 1 - \frac{\sum_i min\,(A_i, B_i)}{\sum_i max\,(A_i, B_i)}$$

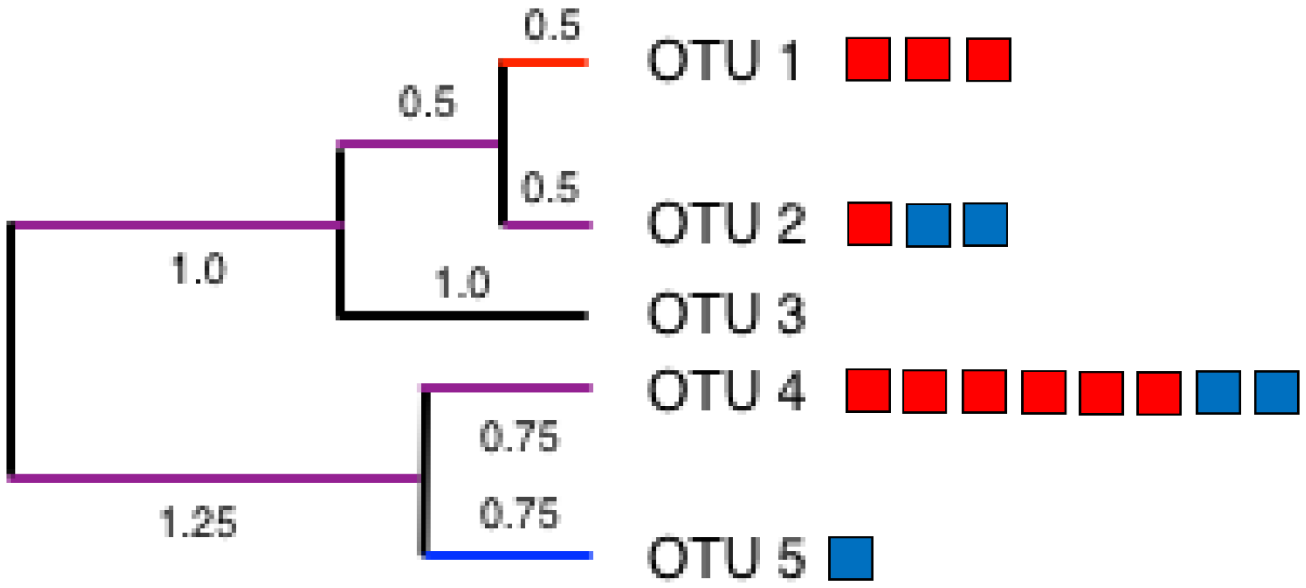


$A = (3,1,0,6,0)$ $B = (0,2,0,2,1)$

$D_{BC}(A,B) = \frac{9}{15}$ and $D_J(A,B) = \frac{9}{12}$

UniFrac dissimilarity (Lozupone and Knight, 2005)

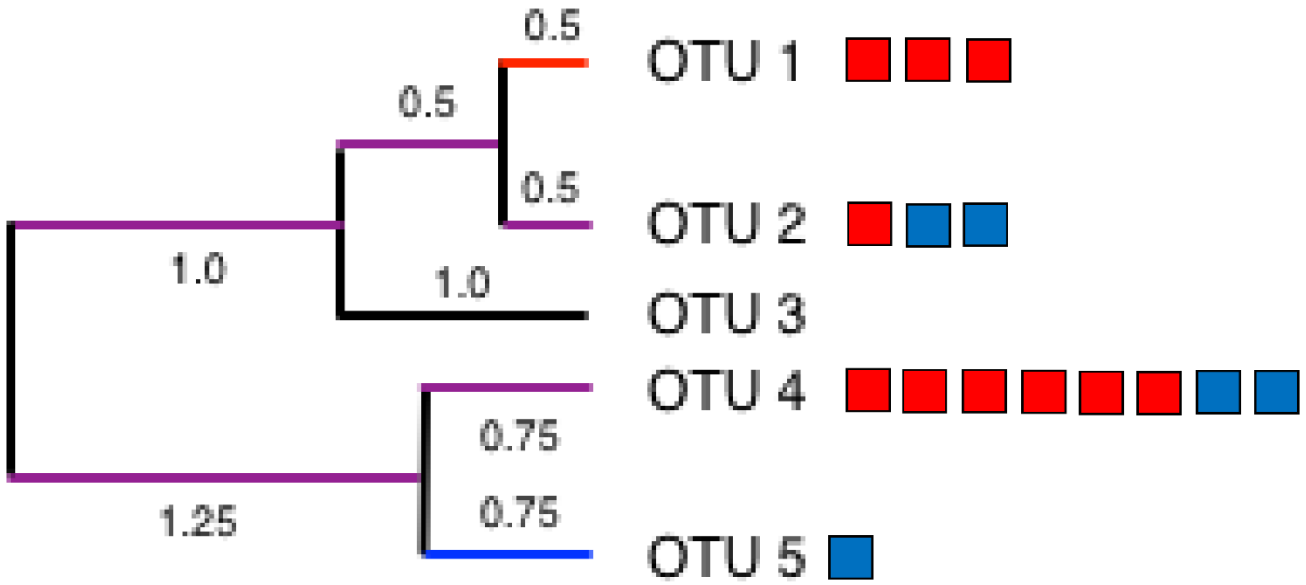
$$D_U(A,B) = \frac{unique\ branch\ length}{total\ branch\ length\ for\ A\ or\ B} = \frac{\sum_i l_i |1(A_i > 0) - 1(B_i > 0)|}{\sum_i l_i\, 1(A_i + B_i > 0)}$$



http://readiab.org/book/latest/3/1

Weighted UniFrac dissimilarity

$$D_{WU}(A,B) = \frac{\sum_i l_i \left|\frac{A_i}{n_A} - \frac{B_i}{n_B}\right|}{\sum_i l_i \left(\frac{A_i}{n_A} + \frac{B_i}{n_B}\right)} = \frac{\sum_i l_i \left|p_i^A - p_i^B\right|}{\sum_i l_i \left(p_i^A + p_i^B\right)}$$



Ordination: Visualization of beta diversity

***Ordination**: Visualization of beta diversity for identification of possible data structures*

Graphical representation of the data along a reduced number of orthogonal axes while keeping the main trends of the data, preserving their distance relationships as well as possible.

PCA preserves Euclidean distance, not appropriate for the analysis of species abundance

Principal Coordinates Analysis (PCoA) = Multidimensional Scaling (MDS)

PCoA is an extension of PCA. It does not require the original data X, but only a distance matrix D between the points.

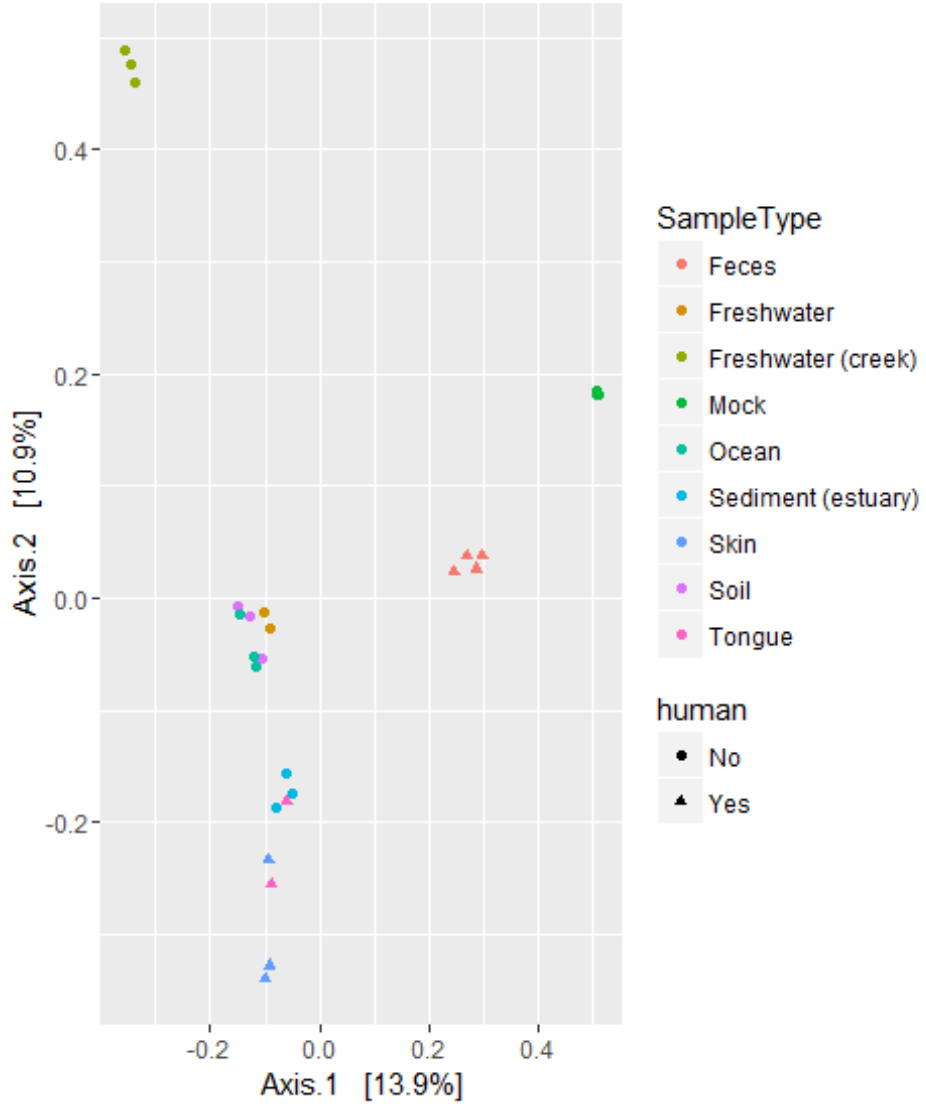
- *PCA performs eigenvalue decomposition of the covariance matrix* $C \propto X'X$
- *Given a distance D, PCoA performs eigenvalue decomposition of* $B \propto D_c' D_c$

  $D_c = D(I - \frac{1}{n} 11^{'})$ *is the centered distance matrix*

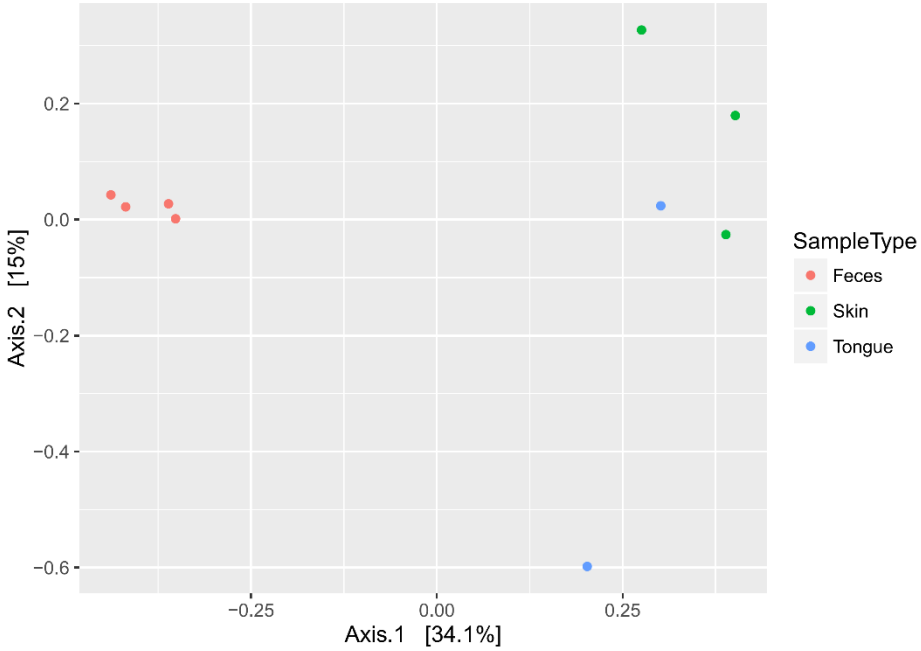
For Euclidean distance, PCoA = PCA.

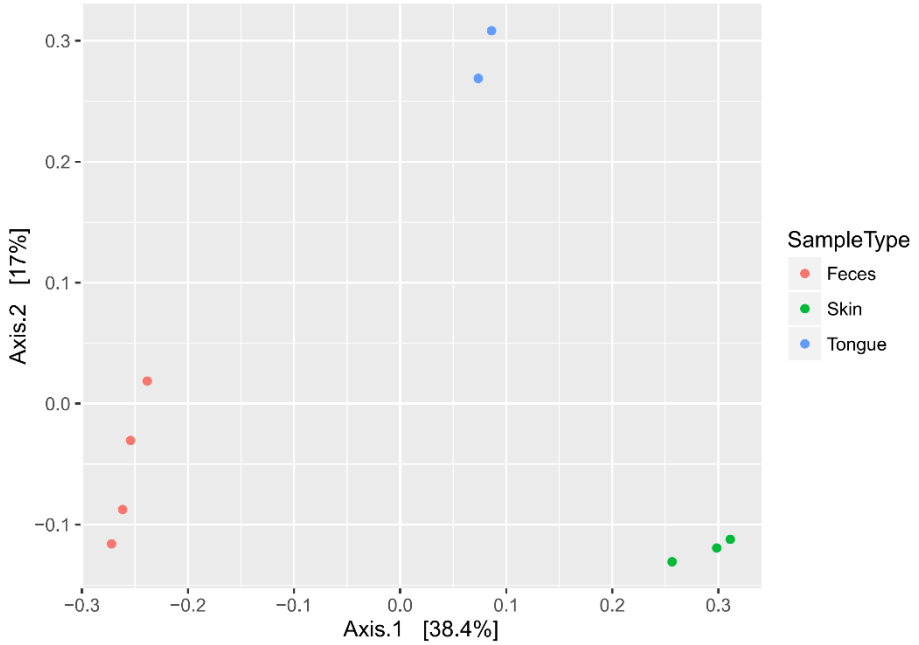
When the distances are not metric, some eigenvalues will be negative and it is not possible to obtain an ordination that exactly reproduces the distances of the distance matrix.

***Example**: MDS ordination of GlobalPatterns with BC distance:*



***Example**: MDS ordination of GlobalPatterns human samples with BC distance (left) and UF distance (right):*





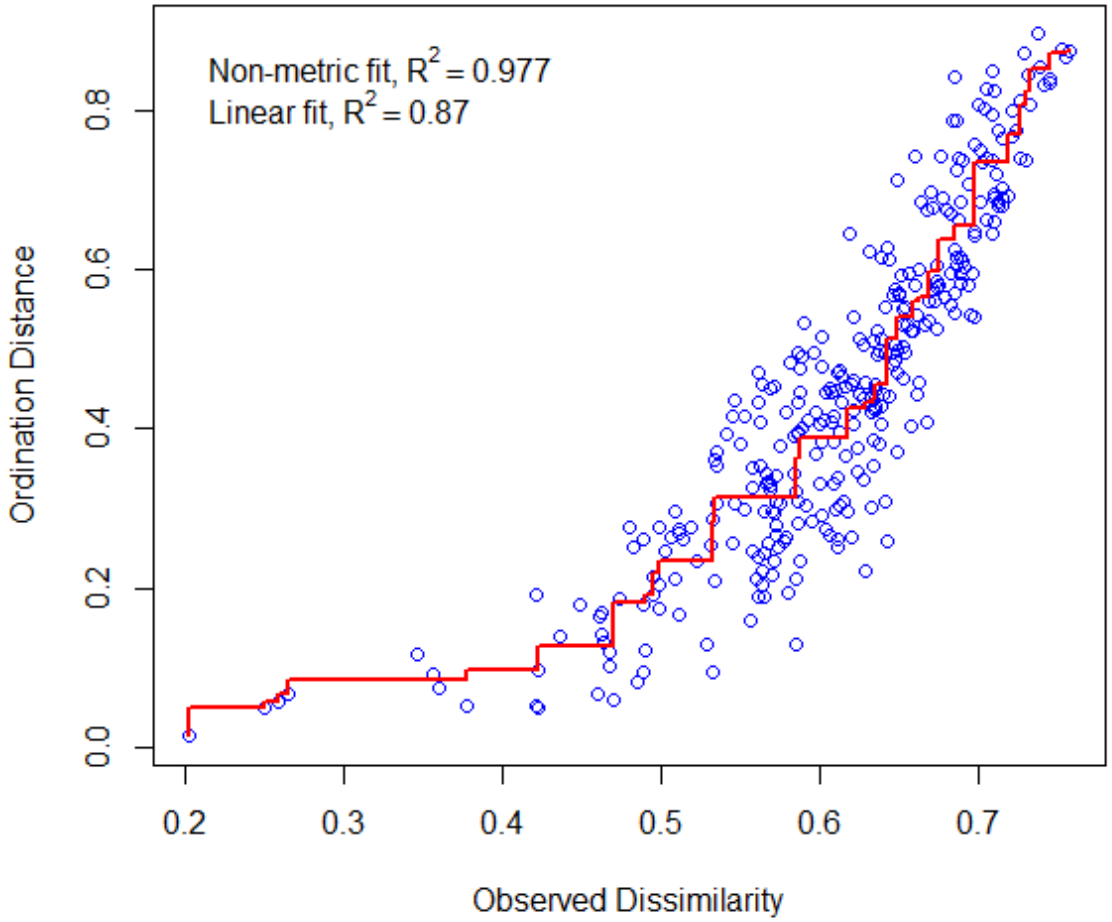
Nonmetric Multidimensional Scaling (NMDS)

NMDS is similar to PCoA in that the calculations are based on a distance matrix.

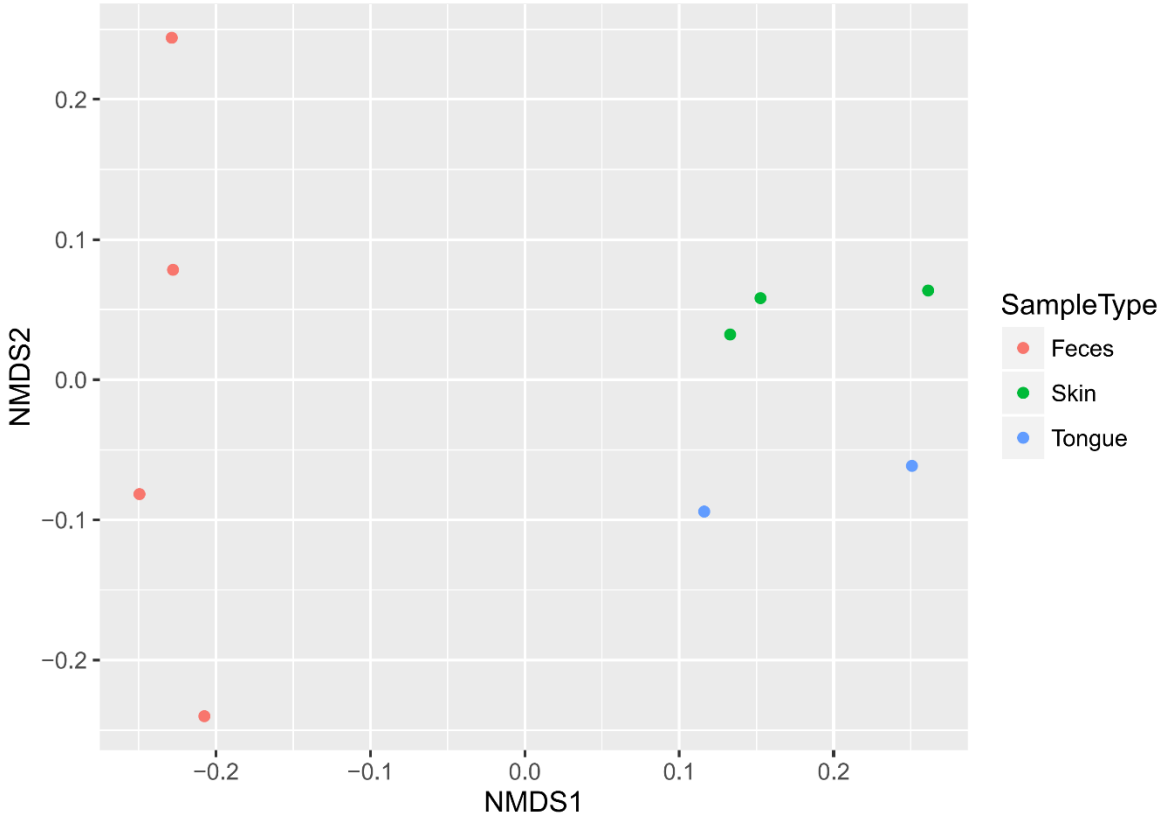
*Maximizes **rank-based correlation** between **original distances** and **distances in the new ordination space**. If samples 1 and 2 are the closest in the distance matrix, they will be in the ordination graph.*

NMDS is an iterative procedure, starts with a random configuration.

At each step the points are moved so that a "stress" function that measures the discrepancy between original and current ordination rank distances is minimized.


Non-metric fit, $R^2 = 0.977$
Linear fit, $R^2 = 0.87$
Ordination Distance
0.0
0.2
0.4
0.6
0.8
0.2
0.3
0.4
0.5
0.6
0.7
Observed Dissimilarity

Example: *NMDS ordination of GlobalPatterns human samples with wUF distance:*



Microbiome differential abundance testing

MICROBIOME COUNT DATA

	Taxa				
Sample	**1**	**2**	**...**	***K***	***Total***
1	X_{11}	X_{12}	...	X_{1K}	$N_{1\bullet}$
2	X_{21}	X_{22}	...	X_{2K}	$N_{2\bullet}$
⋮	⋮	⋮	⋱	⋮	⋮
P	X_{P1}	X_{P2}	...	X_{PK}	$N_{P\bullet}$
Total	$N_{\bullet 1}$	$N_{\bullet 2}$	...	$N_{\bullet K}$	$N_{\bullet\bullet}$

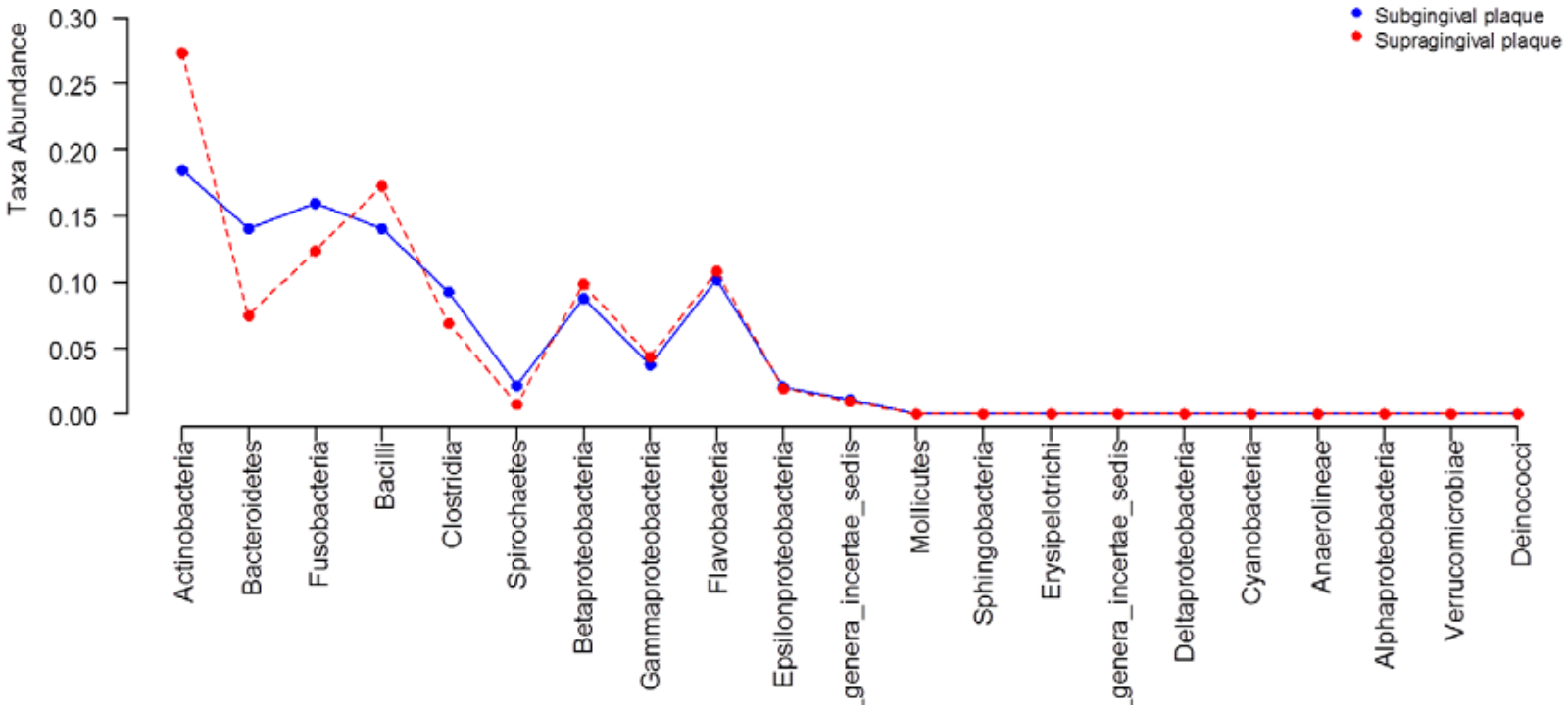
doi:10.1371/journal.pone.0052078.t001

A

B

Multivariate differential abundance testing

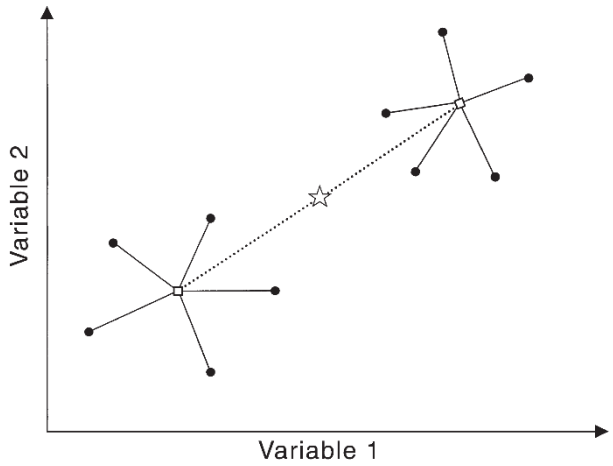
Multivariate community analysis: Global test of microbiome abundance differences between two (or more) classes. Are there global differences in microbial composition between sample groups?



adonis{vegan}: Multivariate ANOVA based on dissimilarities = PERMANOVA

Anderson2001

H0: no differences in composition among groups = same center of masses



Compare the variability within groups (SSW) against the variability between groups (SSA) but partition of sums-of-squares (SS) is applied directly to dissimilarities:

Pseudo F-ratio: $F = \frac{SSA/(a-1)}{SSW/(N-a)}$ *where* $SSA = SST - SSW$

$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \qquad SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij} \qquad (2)$$

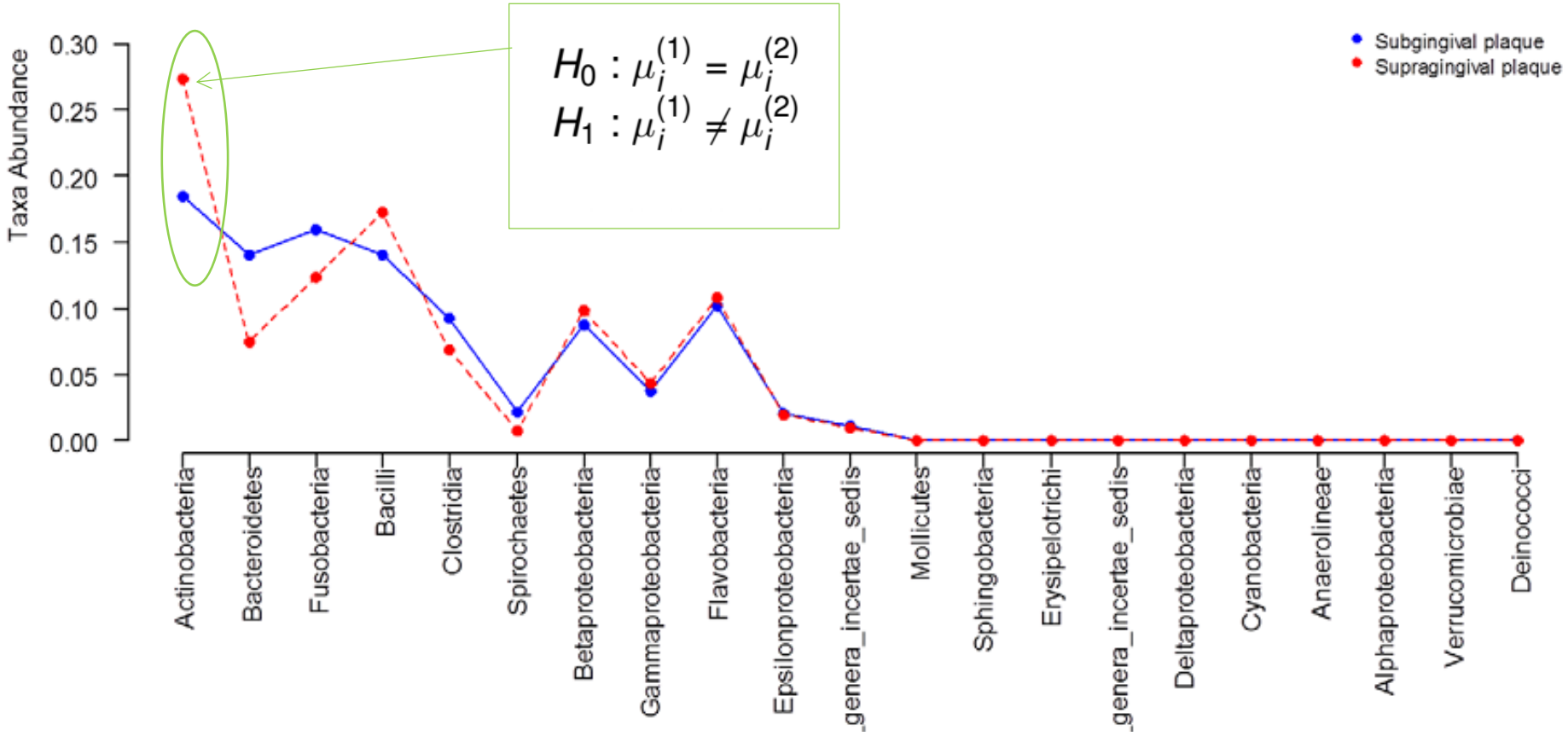
Significance is evaluated through permutations to generate a distribution of the pseudo F statistic under the null.

Example: *ADONIS test for differences in microbiome composition among human samples*

```
##
## Call:
## adonis(formula = BC.dist.human ~ sample_data(data.human)$SampleType)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                                    Df SumsOfSqs MeanSqs F.Model      R2
## sample_data(data.human)$SampleType  2    1.5534 0.77670  2.6042 0.46469
## Residuals                           6    1.7895 0.29825         0.53531
## Total                               8    3.3429                 1.00000
##                                    Pr(>F)
## sample_data(data.human)$SampleType  0.002 **
```

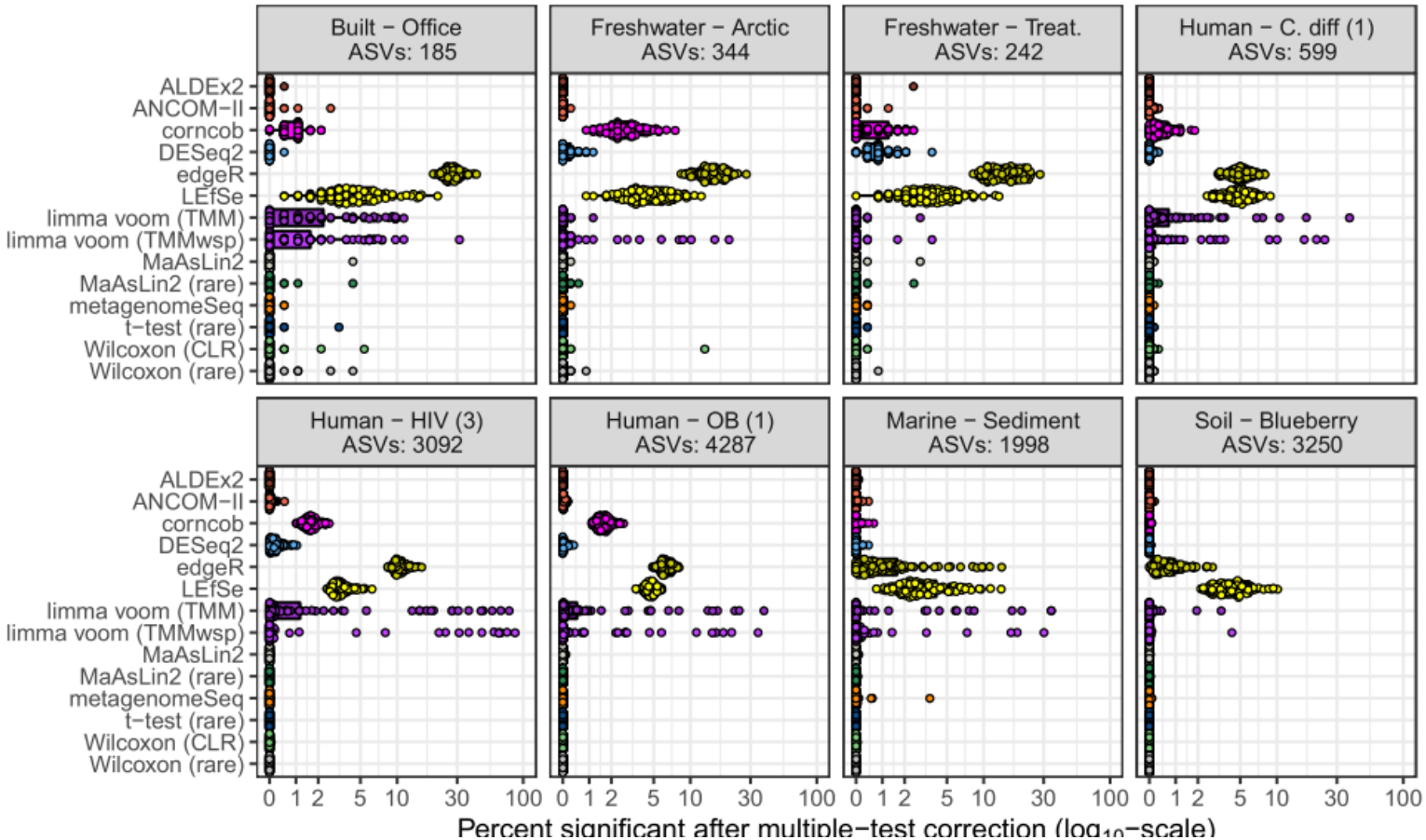
Univariate differential abundance testing

Univariate testing: Which taxa are differentially abundant between sample groups? Which taxa is correlated with a given continuous variable?



Univariate differential abundance testing

*Most used univariate tests (lefSe, edgeR, DESeq2) are biased because of the **compositionality** of microbiome data*



Nearing et al. Nat.Comm *2022*

References

1. Amato K.R. (2017) An introduction to microbiome analysis for human biology applications. Am. J. Hum. Biol. 2017;29: e22931.
2. Anderson M.J. (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecology, vol 26, 32–46
3. Borcard D., Gillet F. and Legendre P. (2011) Numerical Ecology with R. ISBN 978-1-4419-7975-9. DOI 10.1007/978-1-4419-7976-6. Springer New York Dordrecht London Heidelberg
4. Calle, M. (2019). Statistical Analysis of Metagenomics Data. Genomics & Informatics, 17(1), e6. https://doi.org/10.5808/gi.2019.17.1.e6.
5. Calle M.L. and Susin A. (2022) Identification of dynamic microbial signatures in longitudinal studies. BioRxiv. https://www.biorxiv.org/content/10.1101/2022.04.25.489415v1
6. Cho I. and Blaser M.J. (2012) The Human Microbiome: at the interface of health and disease. Nat Rev Genet. ; 13(4): 260–270.
7. Creer et al. (2016) The ecologist's field guide to sequence-based identification of biodiversity Methods in Ecology and Evolution, 7, 1008–1018
8. Gloor and Reid (2016) Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can. J. Microbiol. 62: 692–703 dx.doi.org/10.1139/cjm-2015-0821

9. Grice E. A. and Segre J. A. (2012) The Human Microbiome: Our Second Genome. Annu Rev Genomics Hum Genet. ; 13: 151–170.
10. Kindt R and Coe R. (2005) Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies. Nairobi: World Agroforestry Centre (ICRAF).
11. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, et al. (2012) Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLoS ONE 7(12): e52078. doi:10.1371/journal.pone.0052078
12. Lê Cao et al. (2016) MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. PLOS ONE 11(8): e0160169. doi: 10.1371/journal.pone.0160169
13. Mandal et al. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26:27663. doi:10.3402/mehd.v26.27663.
14. McMurdie P.J. and Holmes S. (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 10(4): e1003531. doi:10.1371/journal.pcbi.1003531
15. Thorsen et al. (2016) Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 4:62 DOI 10.1186/s40168-016-0208-8
16. Rivera-Pinto, J., Egozcue, J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. (2018). Balances: a New Perspective for Microbiome Analysis. Msystems, 3(4). doi: 10.1128/msystems.00053-18.

17. Weiss et al. (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 5:27 DOI 10.1186/s40168-017-0237-y
18. Young V. B. (2017) The role of the microbiome in human health and disease: an introduction for clinicians. BMJ;356:j831 doi: 10.1136/bmj.j831
19. Zhao et al. (2015) Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. Am J Hum Genet. 2015 May 7; 96(5): 797–807. doi: 10.1016/j.ajhg.2015.04.003

Links of interest

coda4microbiome website: https://malucalle.github.io/coda4microbiome/

coda4microbiome R package: https://cran.r-project.org/web/packages/coda4microbiome/index.html

Qiime: Quantitative Insights into Microbial Ecology: http://qiime.org/

Mothur: https://www.mothur.org/wiki/Main_Page

https://bioinformatics.ca/workshops/2015/analysis-metagenomic-data-2015

http://joey711.github.io/phyloseq-demo/phyloseq-demo.html

http://richardsprague.com/microbiome/2016/11/14/best-academic-papers.html

https://microbiomedigest.com/microbiome-papers-collection/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5119550/

Phyloseq tutorial: https://web.stanford.edu/class/bios221/labs/phyloseq/lab_phyloseq.html

Enterotype tutorial: http://enterotyping.embl.de/enterotypes.html

Deseq2 tutorial (regularized log transformation and variance stabilizing transformation)

http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

http://evomics.org/wp-content/uploads/2016/01/phyloseq-Lab-01-Answers.html#beta-diversity-distances

http://deneflab.github.io/MicrobeMiseq/demos/mothur_2_phyloseq.html

https://github.com/joey711/phyloseq-demo/blob/phyloseq-demo-snapshot/phyloseq-demo-old.Rmd

https://www.bioconductor.org/packages/devel/bioc/vignettes/phyloseq/inst/doc/phyloseq-analysis.html#multiple-testing-and-differential-abundance

From fastq to ASV

http://web.stanford.edu/class/bios221/MicrobiomeWorkflowII.html

https://benjjneb.github.io/dada2/tutorial.html

https://hanso3.github.io/TFG/Codi.html