Statistical Analysis of Microbiome Data with R

Dr. Eliana Ibrahimi

Department of Biology, University of Tirana, Albania

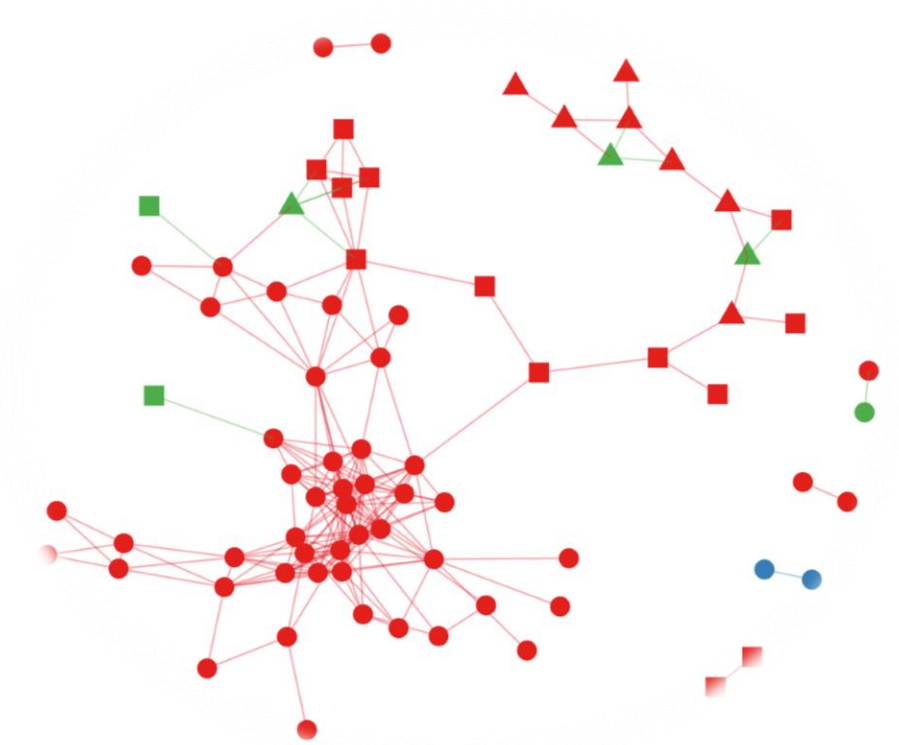
ML4Microbiome Workshop, October 15, 2021

Outline

- Exploration of microbiome data
- Statistical hypothesis testing
  - Sample size and power analysis
  - Univariate community analysis
  - Multivariate community analysis
- Compositional data analysis
- Statistical modeling of microbiome

Exploration of microbiome data

- Microbiome data description
- Graphical summary
- Ordination methods and plots



Microbiome data

Microbiome data generation and structure.

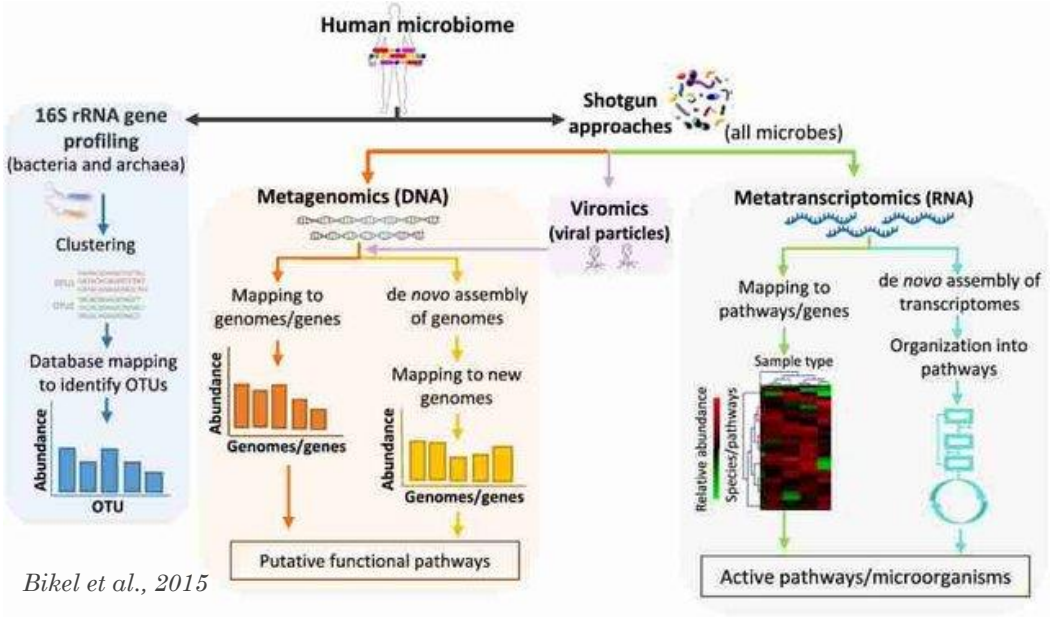
- Generated through 16S rRNA gene sequencing and shotgun metagenomic sequencing.

The 16S sequences are:

- mapped to an existing phylogenetic tree
- clustered into OTUs (operational taxonomic units)

➢ The final data that can be used for analysis

  - OTU table
  - Taxa count table
  - Taxa percent table



Bikel et al., 2015

Feature-by-sample contingency table used in microbiome, genomics, and other high-throughput data studies

Rows	Columns	Comments
Features	Samples	Used in all RNA-or DNA-sequencing experiments and contexts
OTUs (or taxa, i.e., class, genus, species)	Samples, libraries, microbiome or metagenomic samples	A feature is a species or OTU instead of a gene in the context of microbiome sequencing, DNA sequencing-based microbiome study
Genes (or tags or exons or transcripts, subsystems)	Samples, libraries	A feature is a gene in the RNA-sequencing context; the total reads per sample are called library size and sometimes referred to as depths of coverage
Observations (cases)	Variables (part of a composition)	Compositional data
Species	Sites	Ecological data

Xia et al., 2018. Springer Series in Statistics

Microbiome data

Open ecosystems and repositories

- Human Microbiome project datasets
  https://commonfund.nih.gov/hmp/databases
  https://portal.hmpdacc.org/

- Human Gut Microbiome Atlas
  https://www.microbiomeatlas.org/

- Microbiome Learning Repo (ML Repo)

  https://knights-lab.github.io/MLRepo/

- R packages have several datasets incorporated

Microbiome data

Microbiome data features

- High dimensional
- Sparse, large proportion of zero counts
- Compositional
- Complex covariance/correlation structures
- Over-dispersed

Over-dispersed and zero-inflated taxa (OTUs) abundance data

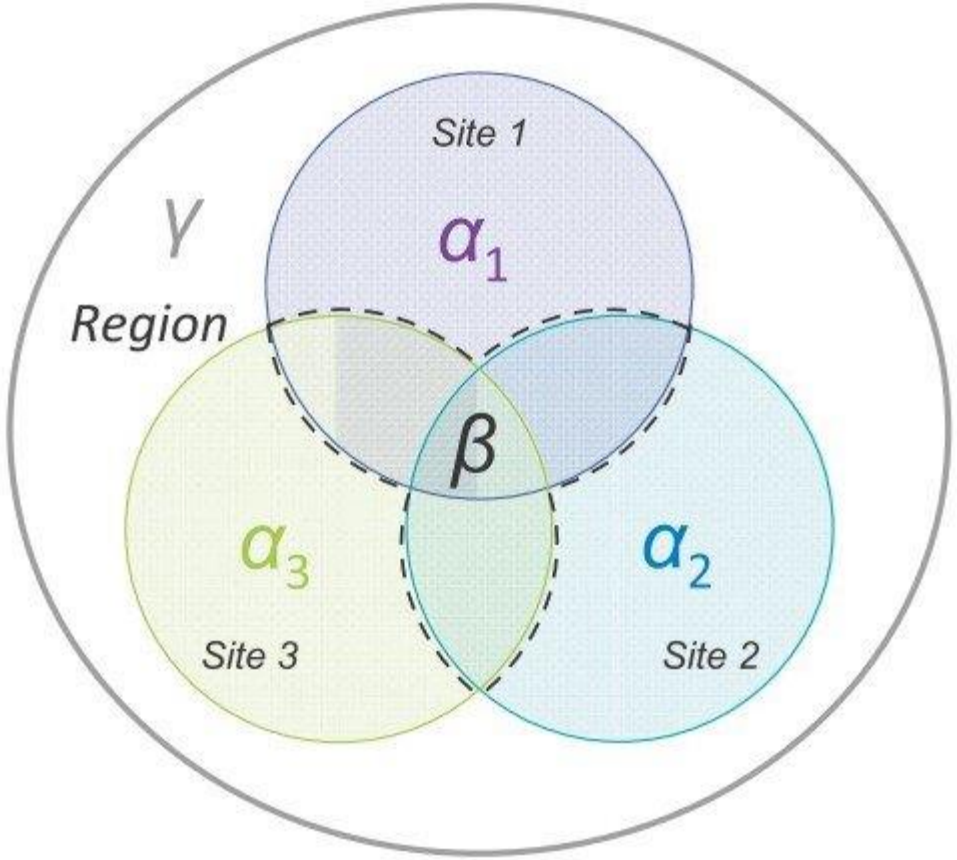
Species (OTUs)	Zero (percentage)	Median	Mean	Variance
Lactobacillus iners	15.11 (138/900)	238	1168	206E4
Lactobacillus crispatus	42.33 (381/900)	1	755.5	205E4
Atopobium vaginae	51.78 (466/900)	0	332.8	404E3
Lactobacillus	14.44 (130/900)	20	168.5	324E3
Lactobacillus jensenii	56.89 (512/900)	0	102.8	128E3
Lactobacillus gasseri	62.33 (561/900)	0	111.3	186E3
Clostridiales	58.56 (527/900)	0	49.8	22,535
Parvimonas micra	71.33 (642/900)	0	45.9	26,298
Leptotrichia amnionii	68.89 (620/900)	0	42.9	27,223
Prevotella genogroup 2	59.11 (532/900)	0	36.2	174,600
Actinomycetales	41.89 (377/900)	1	25.6	11,767
Gardnerella vaginalis	55.89 (503/900)	0	24.9	3322
Streptococcus anosus	73.78 (664/900)	0	18.5	30,230
Aerococcus christensenii	65.89 (593/900)	0	18.2	3710
Finegoldia magna	51.56 (464/900)	0	17.9	7606
Peptoniphilus	54.89 (494/900)	0	17.1	4050
Bifidobacteriaceae	61.11 (550/900)	0	16.6	5402

Romero et al. 2014

Microbiome data

Community diversity measures

- Alpha diversity
  - ✓ Richness, Phylogenetic diversity, evenness, dominance, rarity
- Beta diversity
  - ✓ Bray-Curtis index, Jaccard index, Aitchison distance, Unifrac distances
- Gamma diversity
  - ✓ estimates diversity within a region

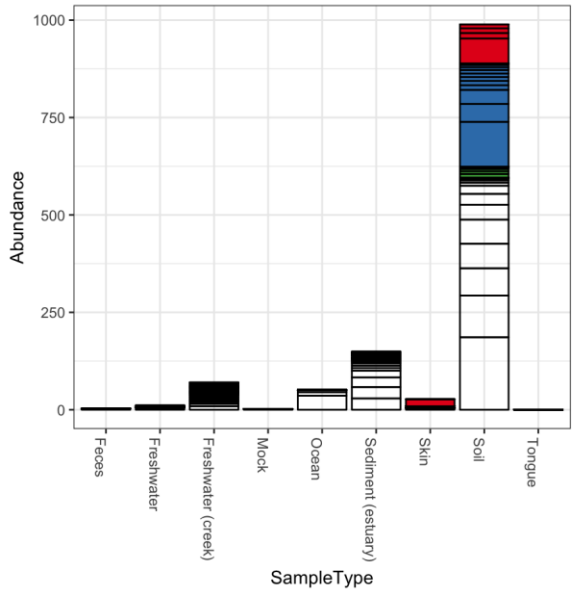


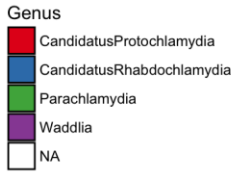
Infographic from Jurgens, 2018.

Lahti et al., Orchestrating Microbiome Analysis with Bioconductor, 2021

Visualization of microbiome

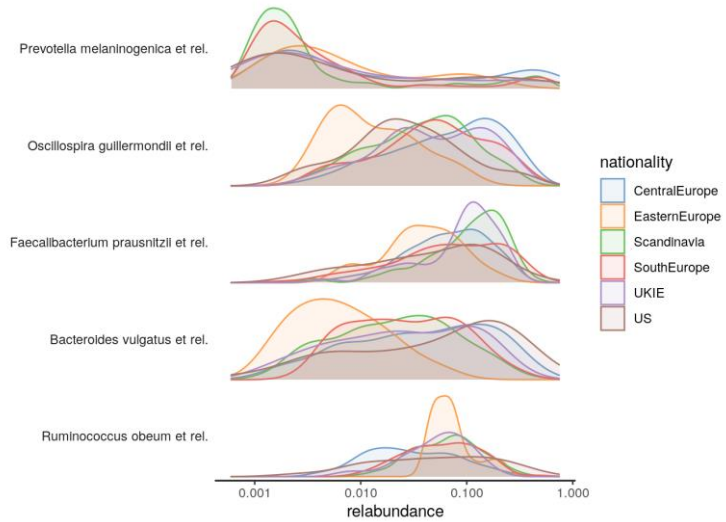
Graphical summary, abundance bar, richnees plot.



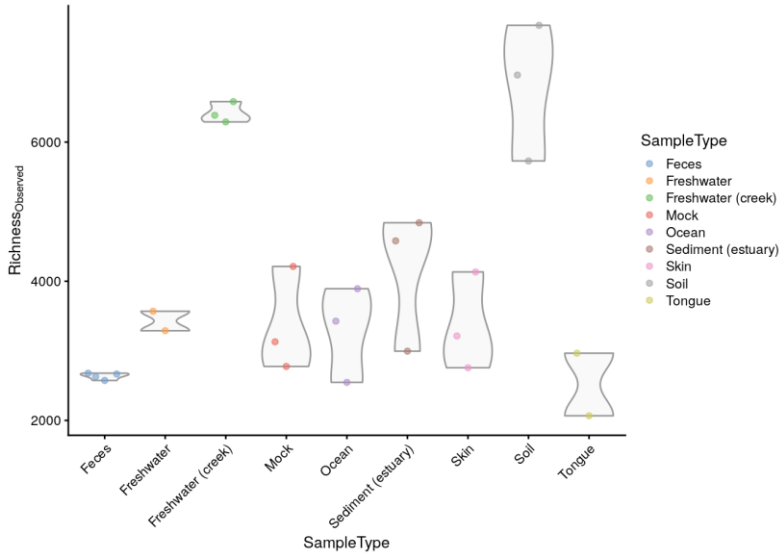


» Abundance bar

Globalpatterns data, Phyloseq package



» Relative abundance



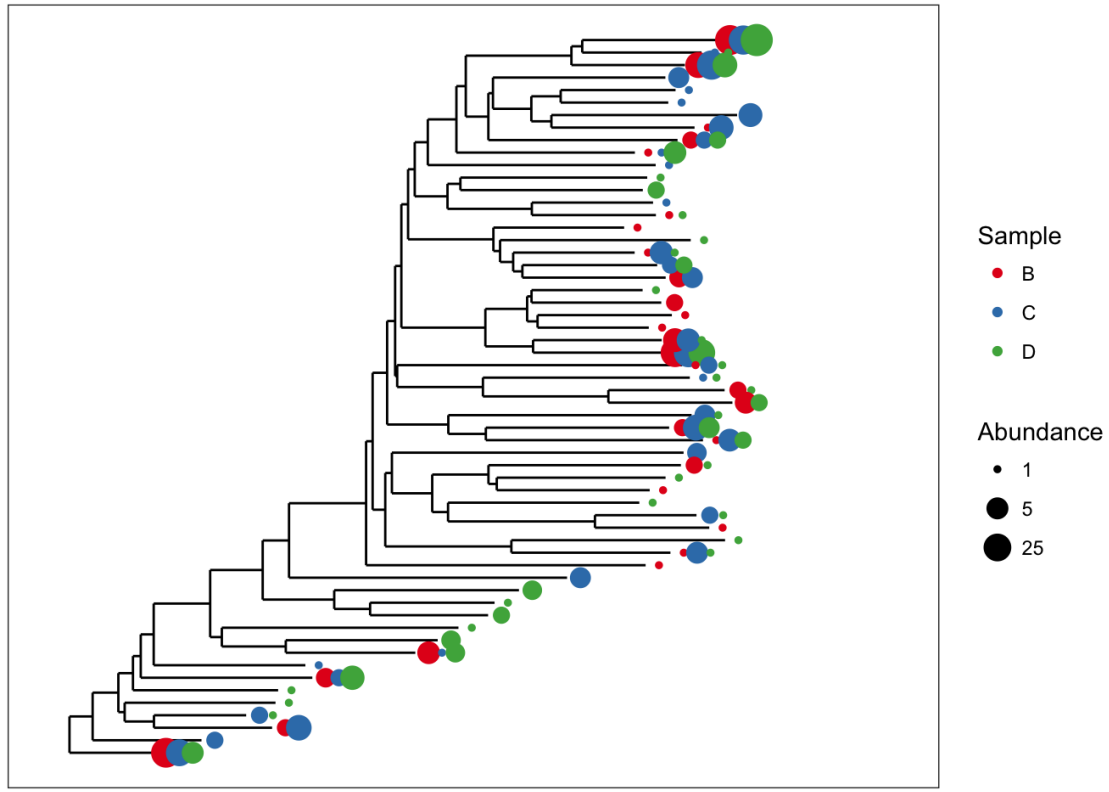
» Richnees plot

GlobalPatterns data, Package mia

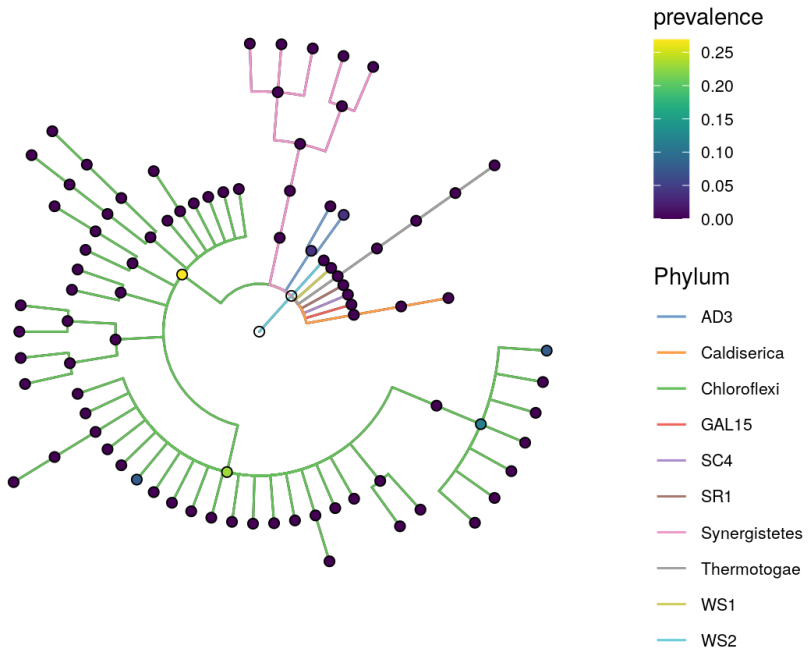
Lahti et al., Orchestrating Microbiome Analysis, 2021

Visualization of microbiome

Graphical summary, phylogenetic trees.



» Esophagus dataset tree, Phyloseq

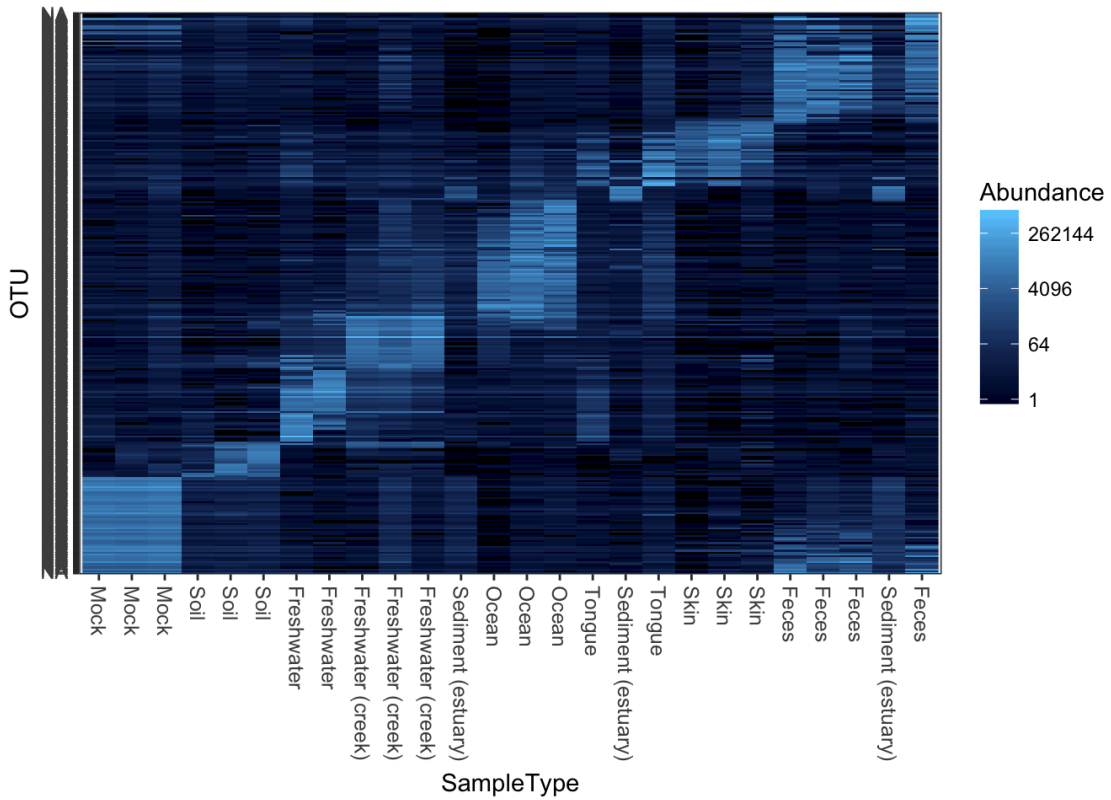


» Globalpatterns dataset tree, prevalence of top phyla

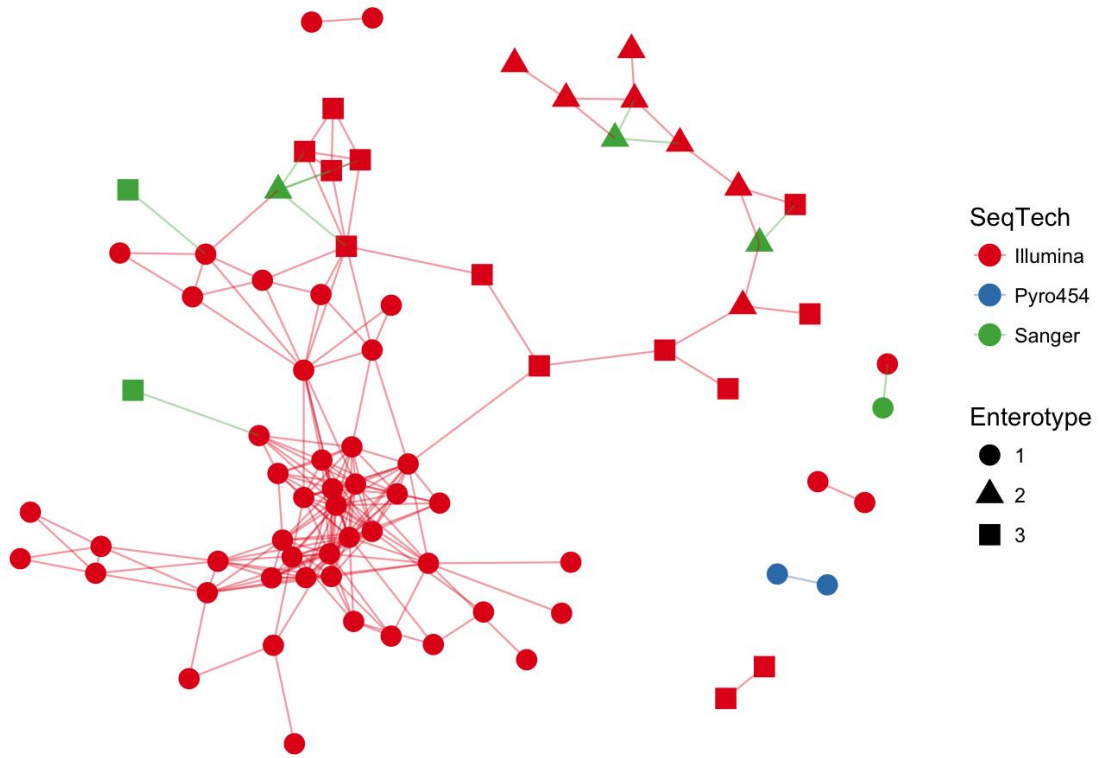
Lahti et al., Orchestrating Microbiome Analysis, 2021

Visualization of microbiome data

Graphical summary, heatmap and networks.



» Heatmap



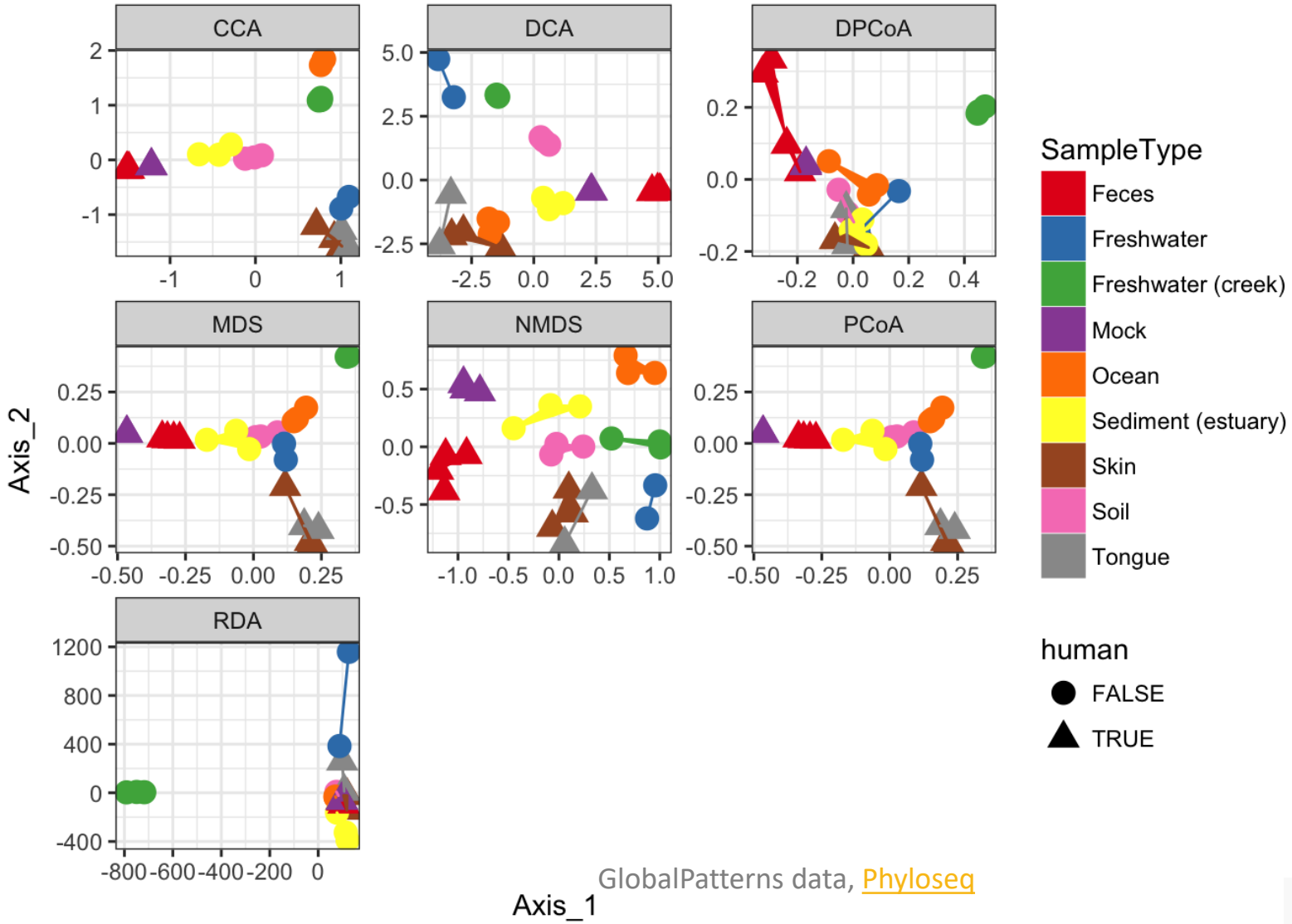
» Network

GlobalPatterns data, Phyloseq

Visualization of microbiome

Ordination methods

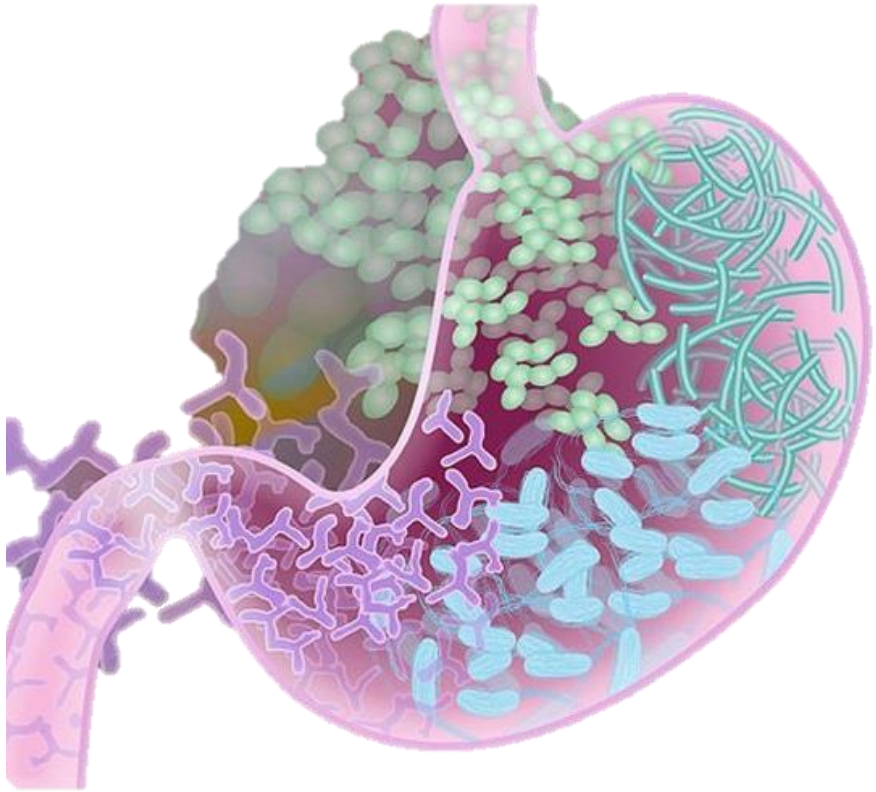
- Canonical Correspondence Analysis (CCA)
- Principal Cordinate Analysis (PCoA)
- Non-metric Multidimensional Scaling (NMDS)
- Redundancy Analysis (RDA)



GlobalPatterns data, Phyloseq

Statistical hypothesis testing

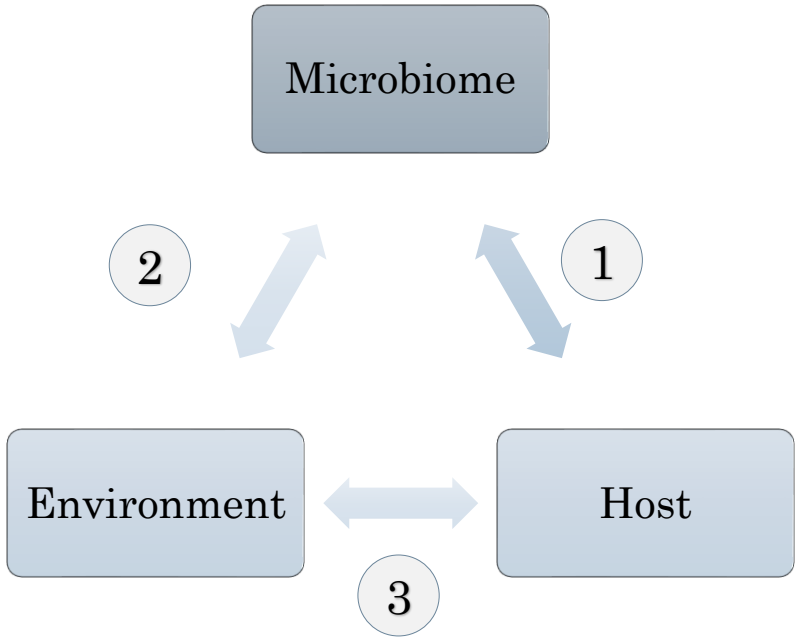
- Univariate analysis
- Multivariate analysis



Statistical hypothesis testing

Research Hypotheses

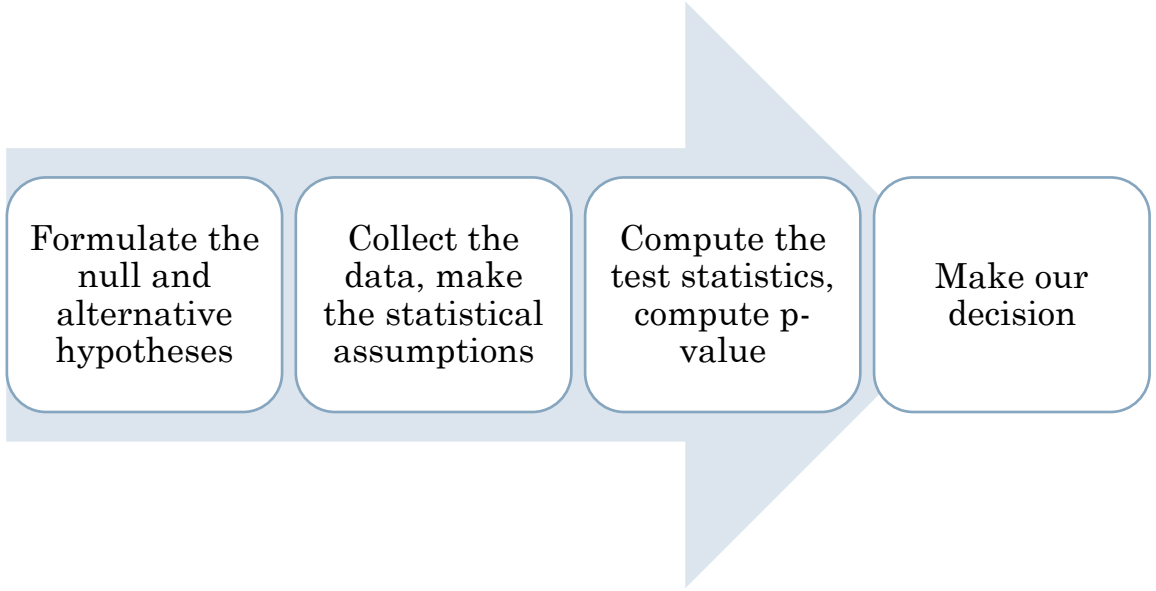
1. Association of microbiome with host.
2. Association of microbiome with environmental covariates.
3. Association between environment and host.



Xia & Sun, Genes and Diseases, 2017

Statistical hypothesis testing

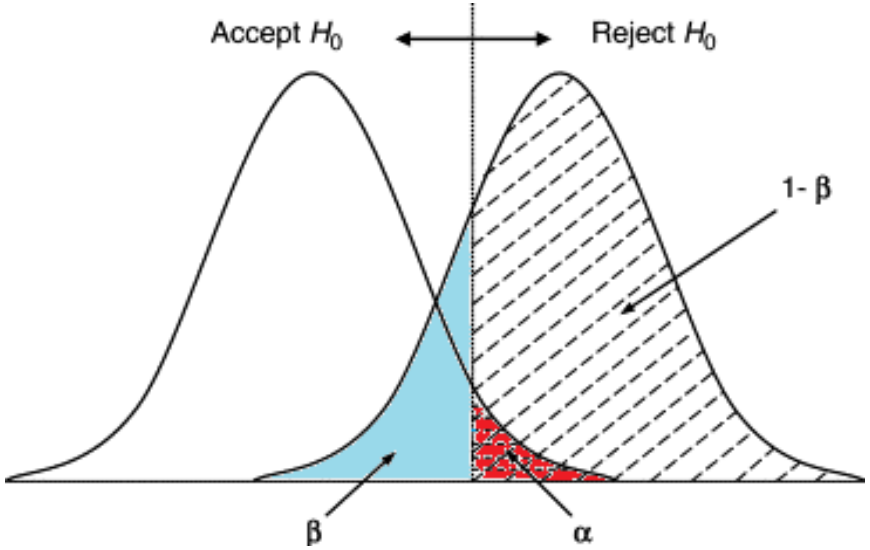
Hypothesis testing steps



$H_0 : \pi_1 = ... = \pi_j = ... = \pi_J$ versus H_A : at least two means are different

$\hookrightarrow \pi$ is the mean proportion

Type I and II errors

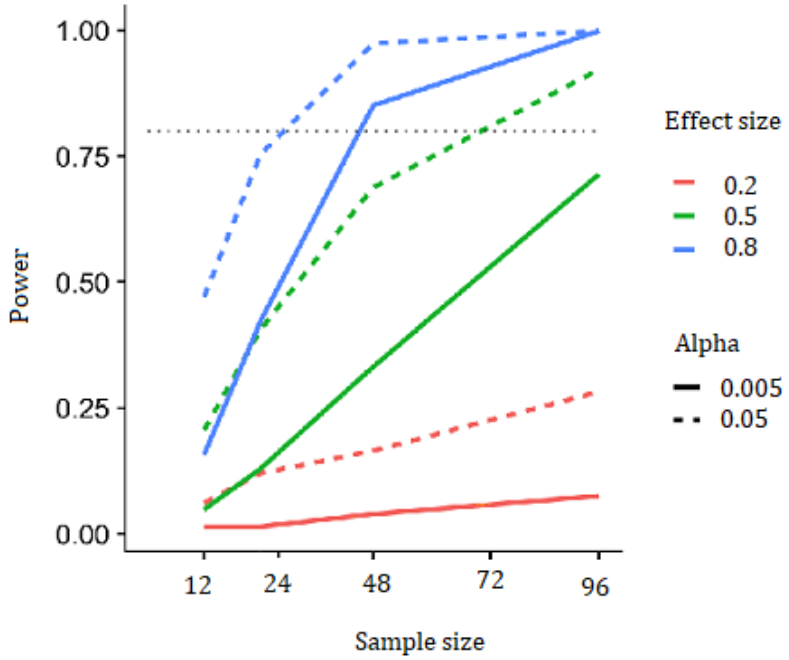


Power of the test = 1- β

Sample size and power analysis

How many subjects do we need?

Standard statistical tests are driven by sample size.



Graphic from stanford.edu

Other factors that affect power

- Experimental design
- Number of groups
- Statistical procedure and model
- Correlation between time points
- Missing data

R packages

- HMP (La Rosa et al. 2016)
- Micropower (Kelly et al. 2015)

Univariate community analysis

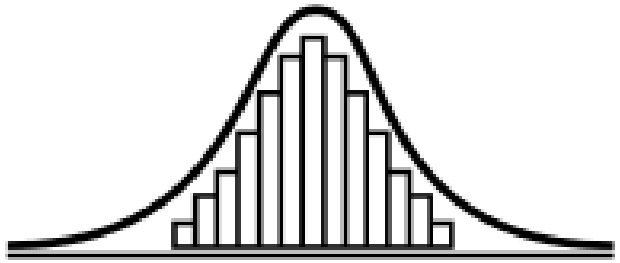
Parametric

- One-sample T-test
- Paired T-test
- Independent T-test
- Analysis of variance (ANOVA)
- Regression and Pearson Correlation

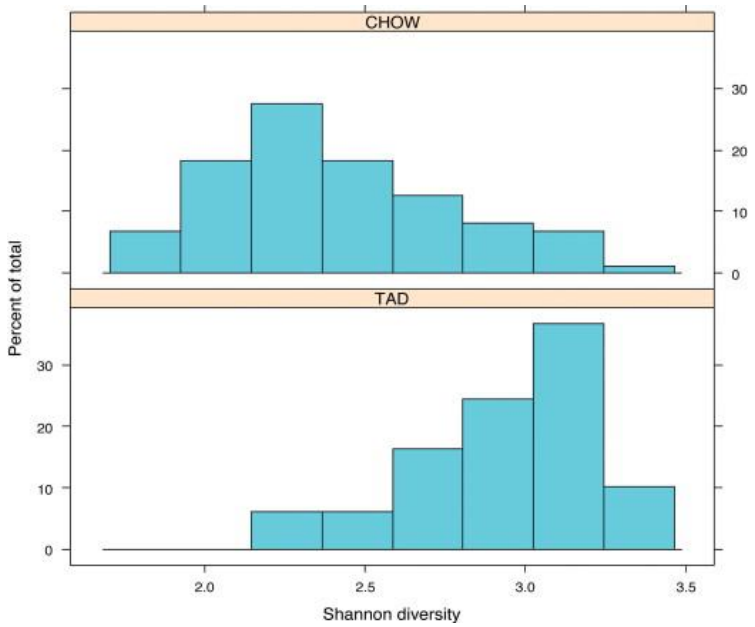
Nonparametric

- Wilcoxon signed rank test
- Mann Whitney
- Kruskal Wallis
- Spearman correlation

⇨ Parametric tests are based on the assumption of normality. Check graphically via histogram, QQ plot, boxplot, or perform Shapiro-Wilk test.

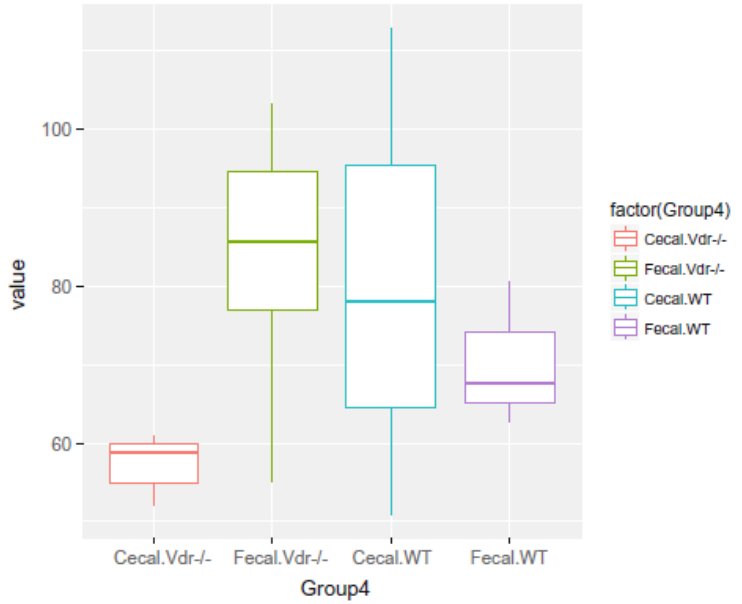


Univariate community analysis



- Compare the calculated Shannon diversity between two groups using t-test and Mann-Whitney test.

La Rosa et al, Metagenomics for Microbiology, 2015.



- Analysis of Chao 1 alpha diversity measures using ANOVA to see if Vdr status and intestinal location have an effect on the bacterial community in the gut.

Xia et al., 2018. Springer Series in Statistics

Univariate community analysis

Chi-square test: Comparing rates

Table . Distribution of the *Streptococcus* Rate Across Stool and Left-Retroauricular Crease Samples Obtained from the Human Microbiome Project

Body Site	Presence	Absence	Total
Left crease	167 (92%)	14 (8%)	181
Stool	135 (65%)	74 (35%)	209

- The chi-square test gave a p-value < 0.0001 ($X^2 = 40.9$, df = 2) leading to rejection of the null hypothesis. The groups have different rates of occurrence.

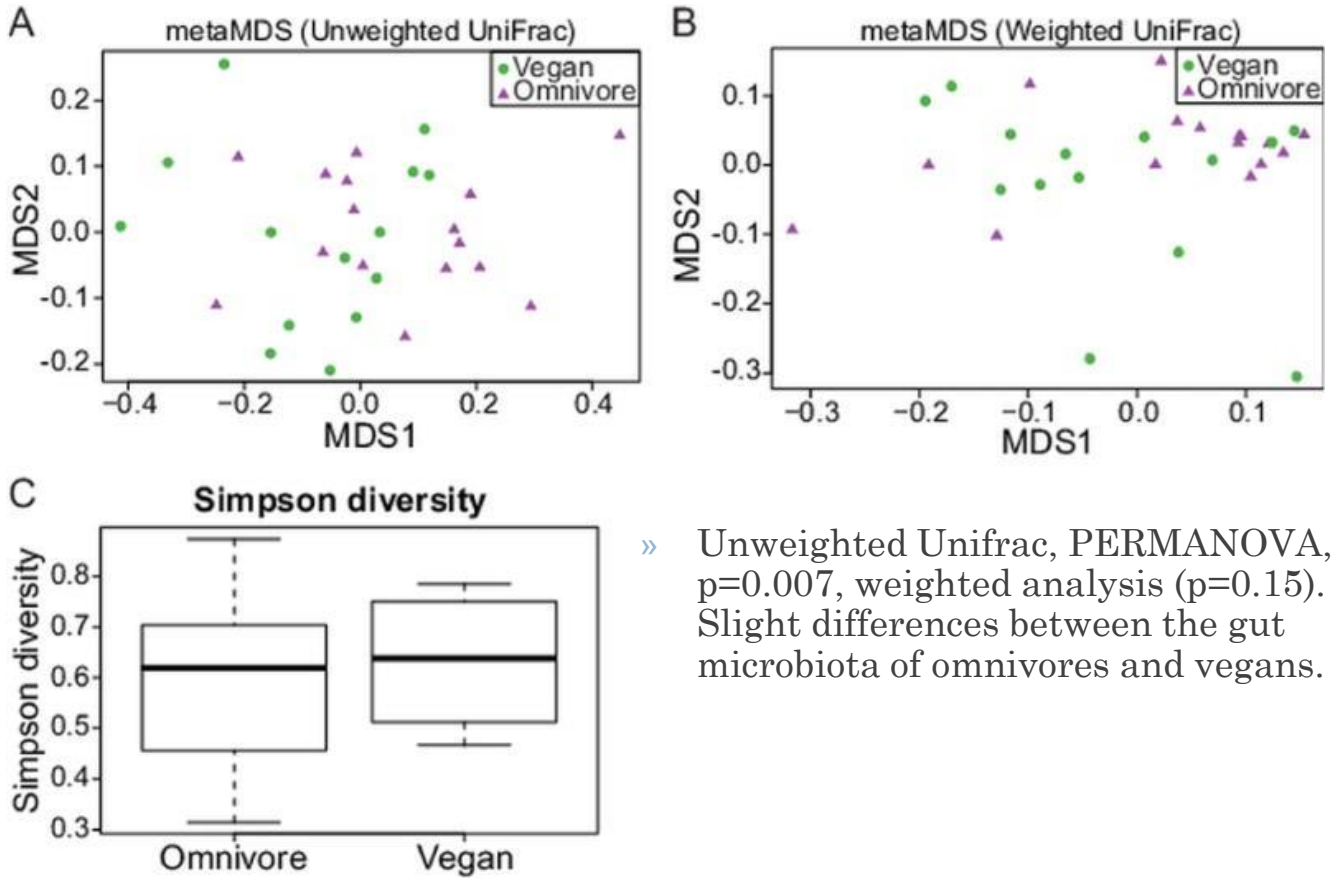
La Rosa et al, Metagenomics for Microbiology, 2015.

Multivariate community analysis

- Test the association of microbiome with environmental covariates
  - Choose one distance measure (i.e., UniFrac, Bray-Curtis, Jaccard, etc.) and then conduct the analysis of the estimated distances.
    - Multivariate analysis of variance with permutation (PERMANOVA).
    - Analysis of group similarities (ANOSIM)
    - Multi-response permutation procedures (MRPP)
    - Mantel's test (MANTEL)

Multivariate community analysis: PERMANOVA

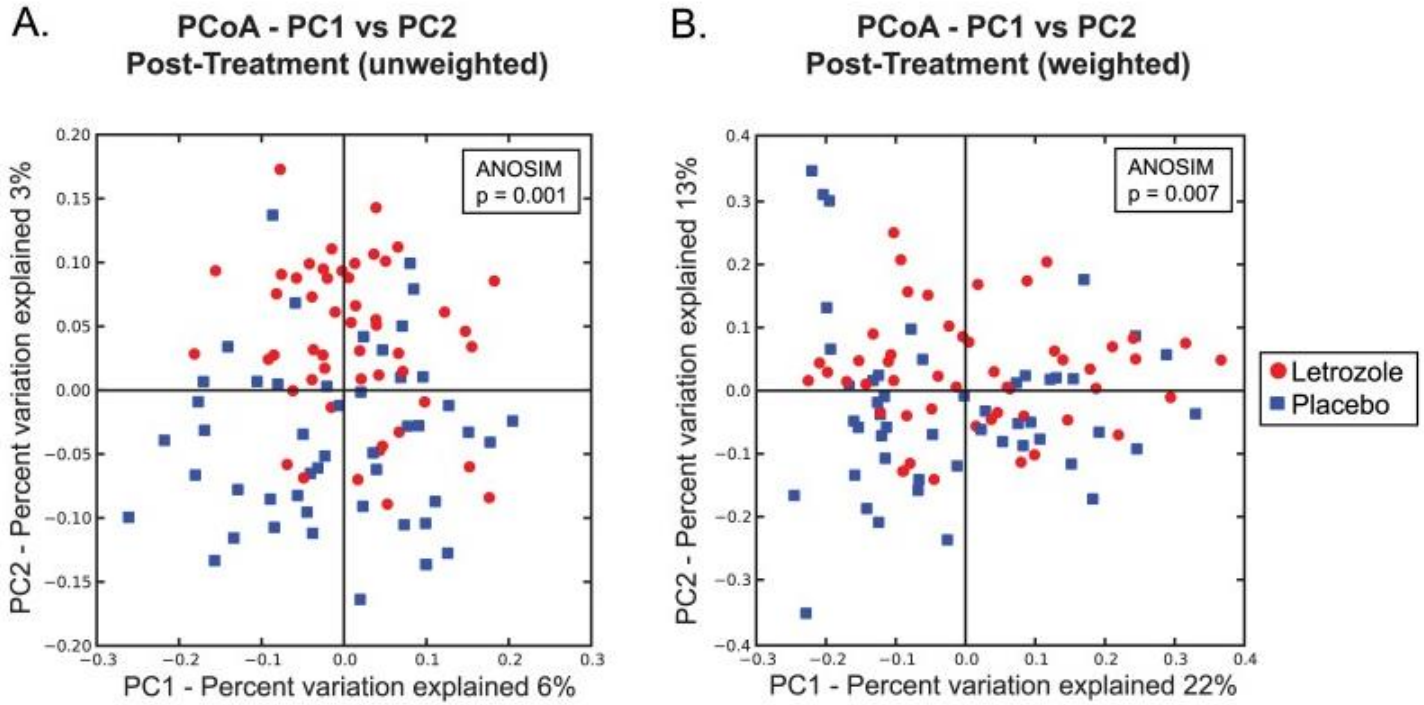
- Flexible to dissimilarity measure
- No assumption of multivariate normality.
- Not sensitive to differences in correlation structure among groups.
- Can include random effects, interaction terms, and hierarchical structures.



- Unweighted Unifrac, PERMANOVA, p=0.007, weighted analysis (p=0.15). Slight differences between the gut microbiota of omnivores and vegans.

Wu et al.. Gut, 2016, 65(1).

Multivariate community analysis: ANOSIM



- Nonparametric test which operates on a ranked dissimilarity matrix.
- The null hypothesis is that the similarities within sites are smaller or equal to the similarities between sites.

- Test the association of microbiome composition between treatments and among time points within treatments using weighted and unweighted UniFrac distances.

Kelley et al., 2016. PloS one, 11(1)

R packages that implement statistical analysis

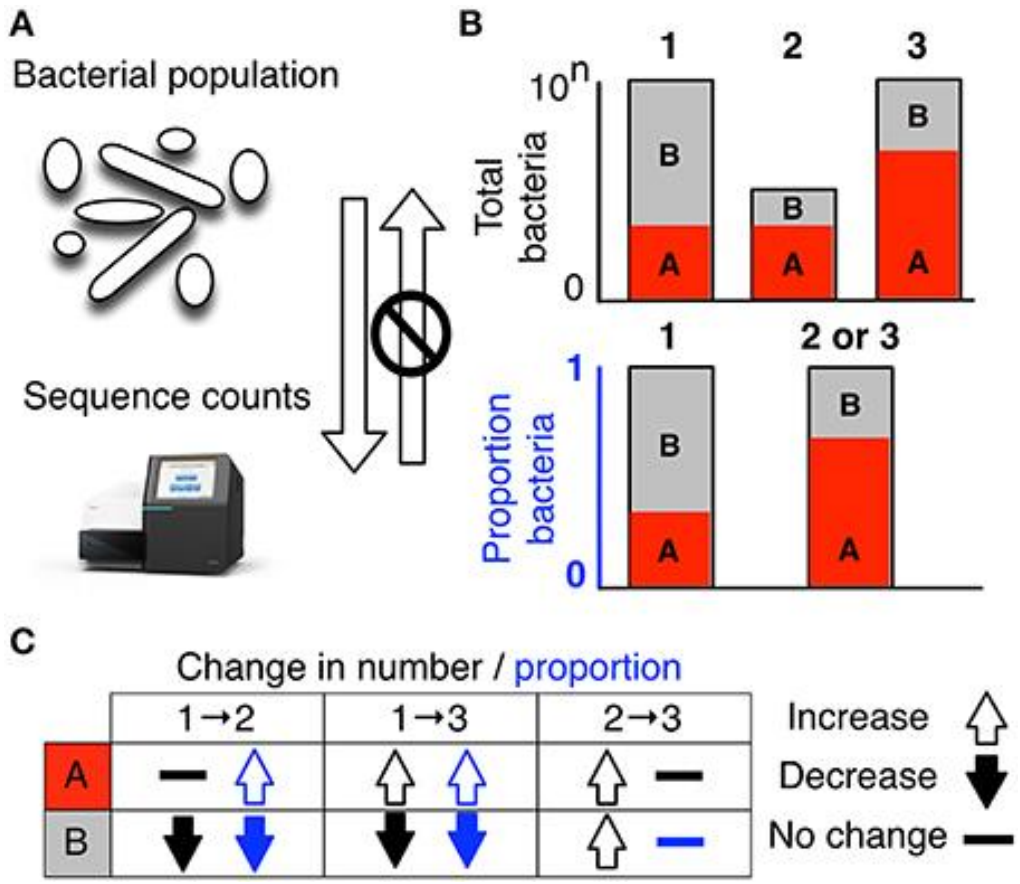
- Vegan
- biom
- DESeq
- DESeq2
- limma
- metagenomeSeq
- microbiome
- phyloseq

Microbiome and phyloseq are more comprehensive statistical tools.

Compositional analysis

Compositional data

Microbiome sequencing data are compositional



Gloor, Gregory B., et al. Frontiers in microbiology 8 (2017): 2224.

» Anything that can be represented as 'part of whole' is compositional.

- Sequencing data are compositional.
  - Proportions cause problems, breaking the statistical assumptions.
  - A zero is not necessarily a zero
    - True zeros: treat as NMAR, replace with small nonzero value.
    - Under sampling zeros: Bayesian-multiplicative methods to replace them

Compositional analysis

Compositional data, Log-Ratio Transformations

Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	Centered Log Ratio (CLR) Isometric Log Ratio (ILR) Additive Log Ratio (ALR)
Distance	Bray-Curtis UniFrac Jenson-Shanon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perMANOVA ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

Gloor, Gregory B., et al. Frontiers in microbiology 8 (2017): 2224.

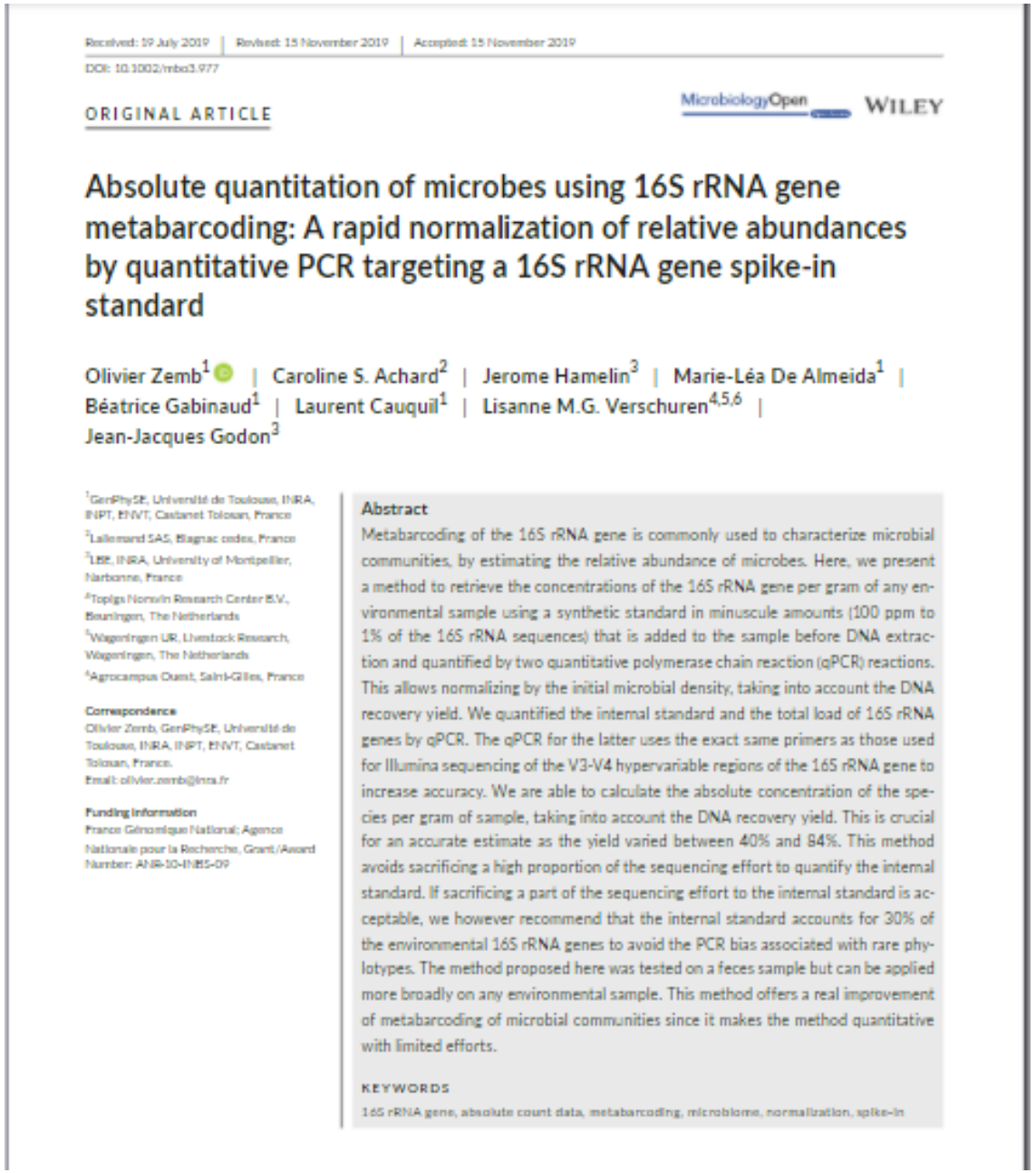
- » Ratios are the same whether the data are counts or proportions.
- » Logarithm of ratios (log-ratios) to achieve symmetry, linear relationship.

- ✓ Transformed data are suitable for most of the standard statistical methods.

Compositional analysis

R packages that analyze compositional data

- Compositions (van den Boogaart et al. 2014)
- robCompositions (Templ et al. 2011)
- zCompositions (Palarea-Albaladejo and Martin-Fernandez 2015).
- ANCOM (Mandal et al. 2015)
- ALDEx and ALDEx2 (Fernandes et al. 2013; Gloor et al. 2016), use Bayesian methods to replace t-test or ANOVA.



Received: 19 July 2019 | Revised: 15 November 2019 | Accepted: 15 November 2019
DOI: 10.1002/mbo3.977

ORIGINAL ARTICLE

MicrobiologyOpen WILEY

Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard

Olivier Zemb[1] | Caroline S. Achard[2] | Jerome Hamelin[3] | Marie-Léa De Almeida[1] | Béatrice Gabinaud[1] | Laurent Cauquil[1] | Lisanne M.G. Verschuren[4,5,6] | Jean-Jacques Godon[3]

[1]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France
[2]Lallemand SAS, Blagnac cedex, France
[3]LBE, INRA, University of Montpellier, Narbonne, France
[4]Topigs Norsvin Research Center B.V., Beuningen, The Netherlands
[5]Wageningen UR, Livestock Research, Wageningen, The Netherlands
[6]Agrocampus Ouest, Saint-Gilles, France

Correspondence
Olivier Zemb, GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France.
Email: olivier.zemb@inra.fr

Funding information
France Génomique National; Agence Nationale pour la Recherche, Grant/Award Number: ANR-10-INBS-09

Abstract

Metabarcoding of the 16S rRNA gene is commonly used to characterize microbial communities, by estimating the relative abundance of microbes. Here, we present a method to retrieve the concentrations of the 16S rRNA gene per gram of any environmental sample using a synthetic standard in minuscule amounts (100 ppm to 1% of the 16S rRNA sequences) that is added to the sample before DNA extraction and quantified by two quantitative polymerase chain reaction (qPCR) reactions. This allows normalizing by the initial microbial density, taking into account the DNA recovery yield. We quantified the internal standard and the total load of 16S rRNA genes by qPCR. The qPCR for the latter uses the exact same primers as those used for Illumina sequencing of the V3-V4 hypervariable regions of the 16S rRNA gene to increase accuracy. We are able to calculate the absolute concentration of the species per gram of sample, taking into account the DNA recovery yield. This is crucial for an accurate estimate as the yield varied between 40% and 84%. This method avoids sacrificing a high proportion of the sequencing effort to quantify the internal standard. If sacrificing a part of the sequencing effort to the internal standard is acceptable, we however recommend that the internal standard accounts for 30% of the environmental 16S rRNA genes to avoid the PCR bias associated with rare phylotypes. The method proposed here was tested on a feces sample but can be applied more broadly on any environmental sample. This method offers a real improvement of metabarcoding of microbial communities since it makes the method quantitative with limited efforts.

KEYWORDS
16S rRNA gene, absolute count data, metabarcoding, microbiome, normalization, spike-in

Avoiding Compositionality

From relative to absolute abundances

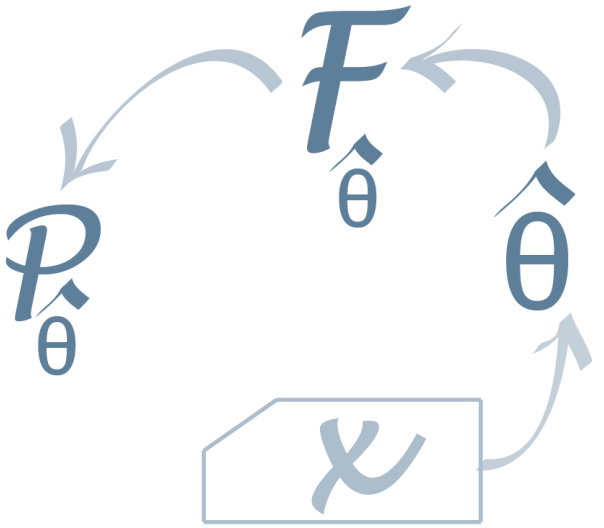
- qPCR with universal 16S primers
  - It works
  - It is cheap
- Spike-ins work too
  - Zemb et al. Microbiology Open 2019
  - Smets et al. PeerJ 2015
  - MB Jones, et al. PNAS 2015

Modeling microbiome data

Modeling microbiome data

Some models that can be fitted to microbiome data

- Over-Dispersed and Zero-Inflated Models
- Dirichlet-Multinomial Models
- Zero-Inflated Longitudinal Models
- Multivariate Bayesian Mixed-Effects Model



Infographic, Holmes & Huber, 2018

Over-Dispersed and Zero-Inflated Models

» **A zero inflated (ZI) model**, is a mixture of a Poisson or NB model with a point mass at zero to allow for the inclusion of structural zeros.

ZIP

$$P(Y_i|X_i,Z_i) = p_i + (1-p_i)\exp(-\mu_i) \quad \text{for } Y_i = 0,$$

$$P(Y_i|X_i,Z_i) = (1-p_i)\frac{\exp(-\mu_i)(\mu_i)^{Y_i}}{Y_i!} \quad \text{for } Y_i > 0,$$

ZINB

$$P(Y_i|X_i,Z_i) = p_i + (1-p_i)g(\mu_i), \quad \text{if } Y_i = 0,$$

$$P(Y_i|X_i,Z_i) = (1-p_i)f(\mu_i), \quad Y_i > 0,$$

» **A hurdle model**, also called a two-part model, with the first part being a binomial/Poisson probability and the second being count data truncated-at-zero.

ZHP

$$P(Y_i|X_i,Z_i) = p_i \quad \text{for } Y_i = 0,$$

$$P(Y_i|X_i,Z_i) = (1-p_i)\frac{\exp(-\mu_i)(\mu_i)^{Y_i}}{Y_i!(1-\exp(-\mu_i))} \quad \text{for } Y_i \geq 0.$$

ZHNB

$$P(Y_i|X_i,Z_i) = p_i \quad \text{for } Y_i = 0,$$

$$P(Y_i|X_i,Z_i) = (1-p_i)\frac{\Gamma(y_i+\alpha^{-1})}{\left(1-\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{1/\alpha}\right)\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{1/\alpha}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \quad \text{for } Y_i > 0,$$

Xia et al., 2018. Springer Series in Statistics

Analysis Pipeline	Description
1. Descriptive Statistics	• Summary statistics of demographic and clinical factors • Zero proportion of the OTU counts • Histograms of the total reads • Histogram of OTU counts
2. Model Setting	• OTU counts as dependent variable • Key predictor (i.e. gender in this study) • Adjusted covariates (i.e. age) • Offset variable (i.e. total reads)
3. Model Selection	• AIC calculation and comparison • Vuong test for nested models • Predictive ability comparison
4. Statistical Inference	• Parameter estimation and standard error • P-values for the zero and count components • Overall p-value for predictor effect
5. Conclusion	• Hypothesis testing • Robustness checking and sensitivity analysis • Zero proportion prediction

Over-Dispersed and Zero-Inflated Models

Assessment of competing models for microbiome

» We can use "pscl" R package, function zeroinfl() to fit a ZIP and ZINB, and function hurdle() to fit ZHP and ZHNB.

» Xu et al. use data for three organisms with proportion of zero counts 18%, 50% and 77% from the Genetic Environmental Microbial project.

- Fit the hurdle/ZI models, including covariates for the zero component, gender, and age.

Xu et al. Plos One 2015.

Over-Dispersed and Zero-Inflated Models

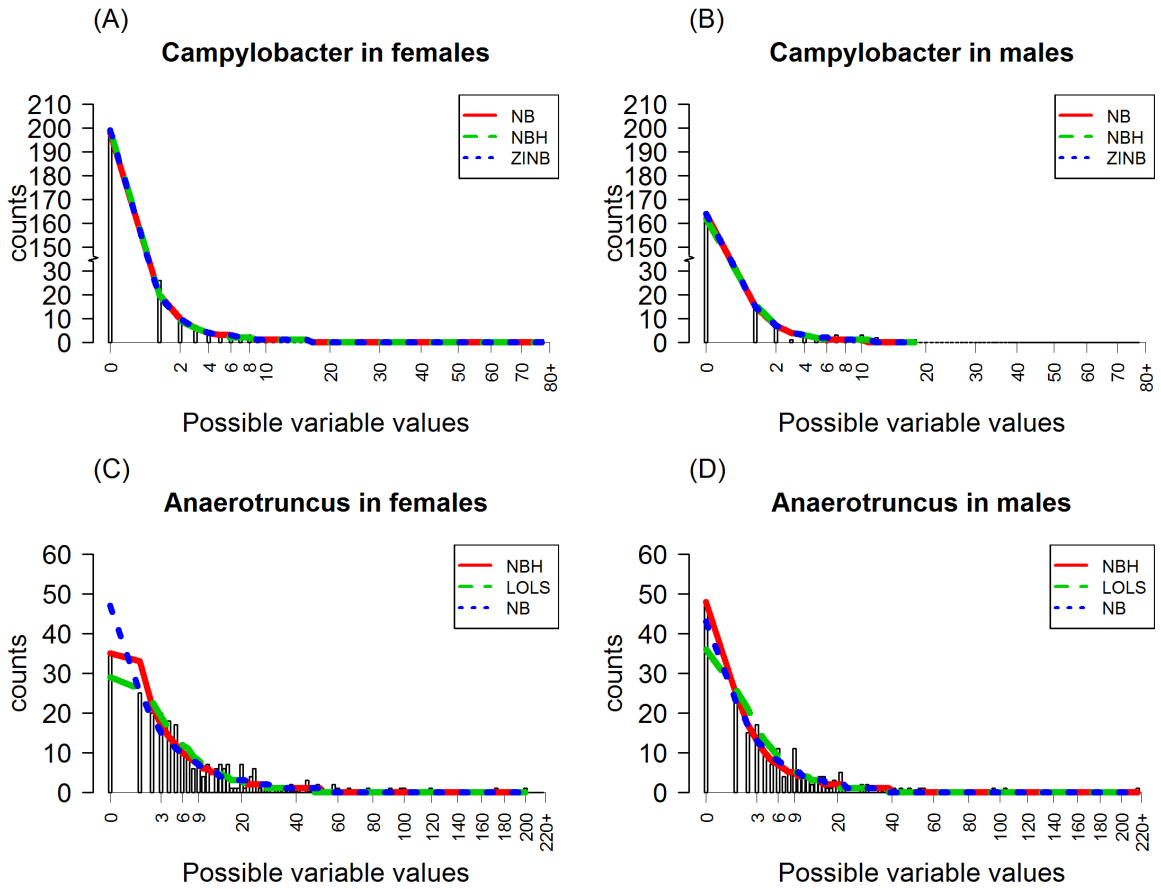
The parameter estimate of the gender effect and goodness of fit for bacteria Campylobacter (proportion of zeros: 77%).

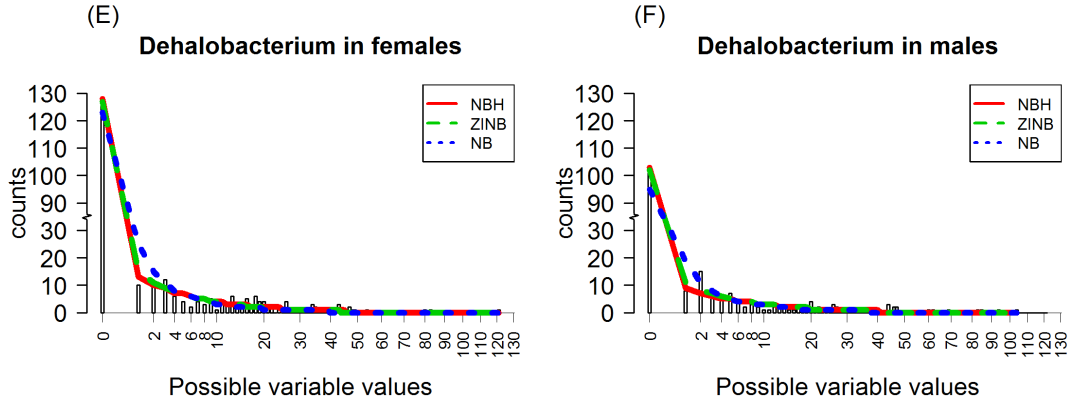
Model	Logit*		Count distribution		overall	AIC
	β_1 (SE)	$Pr(>\|t\|)$	γ_1 (SE)	$Pr(>\|t\|)$	p-value**	
LOLS	NA	NA	−0.074 (0.074)	0.316	0.316	1388
Poisson	NA	NA	−0.782 (0.091)	$< 10^{-6}$	$< 10^{-6}$	2781
NB	NA	NA	**−0.841 (0.306)**	**0.006**	**0.006**	**976**†
WRS	NA	NA	NA	NA	0.420	NA
2P-LOLS	0.335(0.236)	0.156	0.002 (0.220)	0.992	0.365	1051
PH	0.320(0.236)	0.174	−0.598 (0.096)	$< 10^{-6}$	$< 10^{-6}$	1792
ZIP	0.226(0.237)	0.342	−0.599 (0.096)	$< 10^{-6}$	$< 10^{-6}$	1793
NBH	0.320(0.236)	0.174	−0.923 (0.470)	0.049	0.059	978††
ZINB	0.022(3.567)	0.995	−0.813 (0.410)	0.047	0.047	981†††
2P-WRS	NA	NA	NA	NA	0.597	NA

The standard errors (SEs) of estimations are in parentheses. The first, second and third smallest AIC value among different models (except logistic regression) are displayed with superscript †, ††, and ††† respectively. The model with its name in bold font is the final selected model.

*: $logit(\phi_i) = log(\frac{\phi_i}{1-\phi_i}) = X_i^T \beta$, where ϕ is the probability of zeros/structural zeros as defined in hurdle/ZI models.

Over-Dispersed and Zero-Inflated Models





The comparison plots of the observed and expected counts of bacteria using the best three models judging by AIC criterion.

Xu et al. Plos One 2015.

Dirichlet-Multinomial Models

» La Rosa et al. 2012 proposed the Dirichlet-Multinomial distribution to perform hypothesis testing, power and sample size calculation.

$$P(\boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\pi}, \theta) = \frac{N_{i.}!}{x_{i1}! \ldots x_{iK}!} \frac{\Pi_{j=1}^{K} \Pi_{r=1}^{x_{ij}} \{\pi_j(1-\theta) + (r-1)\theta\}}{\Pi_{r=1}^{N_{i.}} (1-\theta) + (r-1)\theta}.$$

➢ R statistical package HMP (La Rosa et al. 2016) can fit these models.

» Differently from the multivariate non-parametric methods based on permutation test

- Can quantify the size of the difference between the groups in terms of bacterial taxa composition.
- Are usually more powerful than a nonparametric alternative to model metagenomic data.

La Rosa et al. Plos One 2012.

Zero-Inflated Longitudinal Models

Longitudinal designs and analyses of microbiome data

» ZINB mixed-effects model

$$\log(\lambda_{ij}|u_i) = \boldsymbol{X}_{ij}'\boldsymbol{\beta} + u_i,$$

$$\text{logit}(\pi_{ij}|v_i) = \boldsymbol{Z}_{ij}'\boldsymbol{\gamma} + v_i,$$

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix}\right).$$

X_{ij} and Z_{ij} are vectors of covariates for the NB and the logistic components, and β and γ are the vectors of regression coefficients.

❖ R statistical package NBZIMM (Nengjun Yi, 2021), function 'glmm.zimb', can fit this model.

Moghimbeigi et al. Journal of Applied Statistics, 2008

Fang et al. Epidemiology and Infection, 2016

Zero-Inflated Longitudinal Models

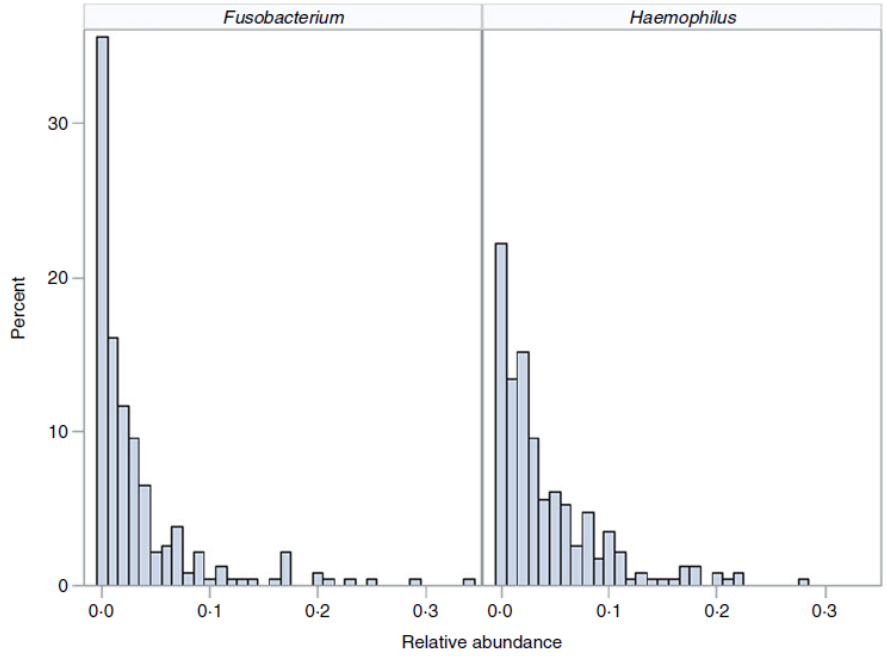
Longitudinal designs and analyses of microbiome data

- ZINB mixed-effects model is applied by Fang et al., to compare the relative abundance of two important organisms across disease states and sampling sites in a study of oesophageal microbiota.

$$\log(E(Y_{ij}|u_i)) = X_{ij}' + u_i + \log(\text{total}_{ij})$$

$$\log\left(\frac{E(Y_{ij})}{\text{total}_{ij}}|u_i\right) = X_{ij}' + u_i.$$

The left side of this equation is modelling
the log of the relative abundance as the outcome.

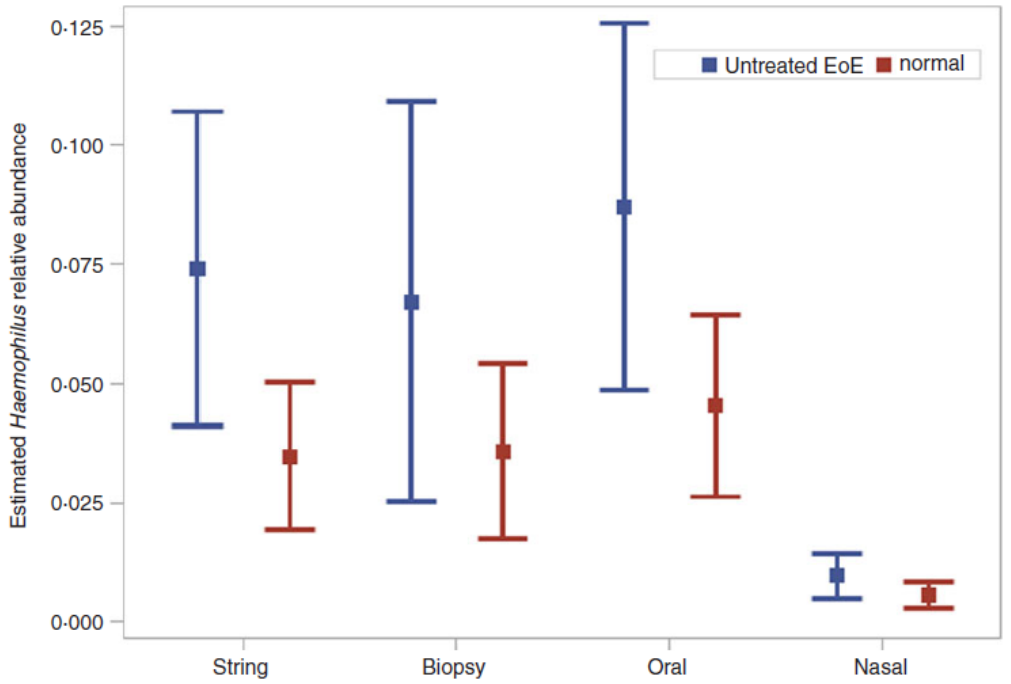


Fang et al. Epidemiology and Infection, 2016

Zero-Inflated Longitudinal Models

ZINB model parameters for Haemophilus

Parameter	Estimate	S.E.	*P* value	95% CI
Intercept	−3·51	0·27	<0·01	−4·05 to −2·96
String	−0·03	0·26	0·91	−0·55 to 0·49
Nasal	−1·85	0·29	<0·01	−2·43 to −1·26
Oral	0·23	0·26	0·38	−0·29 to 0·75
PPI	0·66	0·27	0·02	0·12 to 1·20
Steroid	−0·39	0·26	0·14	−0·92 to 0·14
EoE	0·83	0·42	0·05	−0·02 to 1·68
GORD	−0·38	0·33	0·26	−1·04 to 0·29
Active disease	−0·09	0·27	0·74	−0·62 to 0·44
EoE*PPI	−0·82	0·38	0·03	−1·57 to −0·07
EoE*string	0·13	0·41	0·75	−0·68 to 0·94
EoE*nasal	−0·07	0·44	0·87	−0·96 to 0·81
EoE*oral	0·03	0·41	0·95	−0·79 to 0·84
ZI intercept	−27·26	12·59	0·03	−52·38 to −2·14
ZI GORD	3·47	1·89	0·07	−0·29 to 7·24
ZI total sequence	3·02	1·63	0·07	−0·23 to 6·27
Overdispersion	0·63	0·09	<0·01	0·46 to 0·80
σ_u	0·57	0·09	<0·01	0·39 to 0·75
σ_v	2·49	0·69	<0·01	1·12 to 3·87



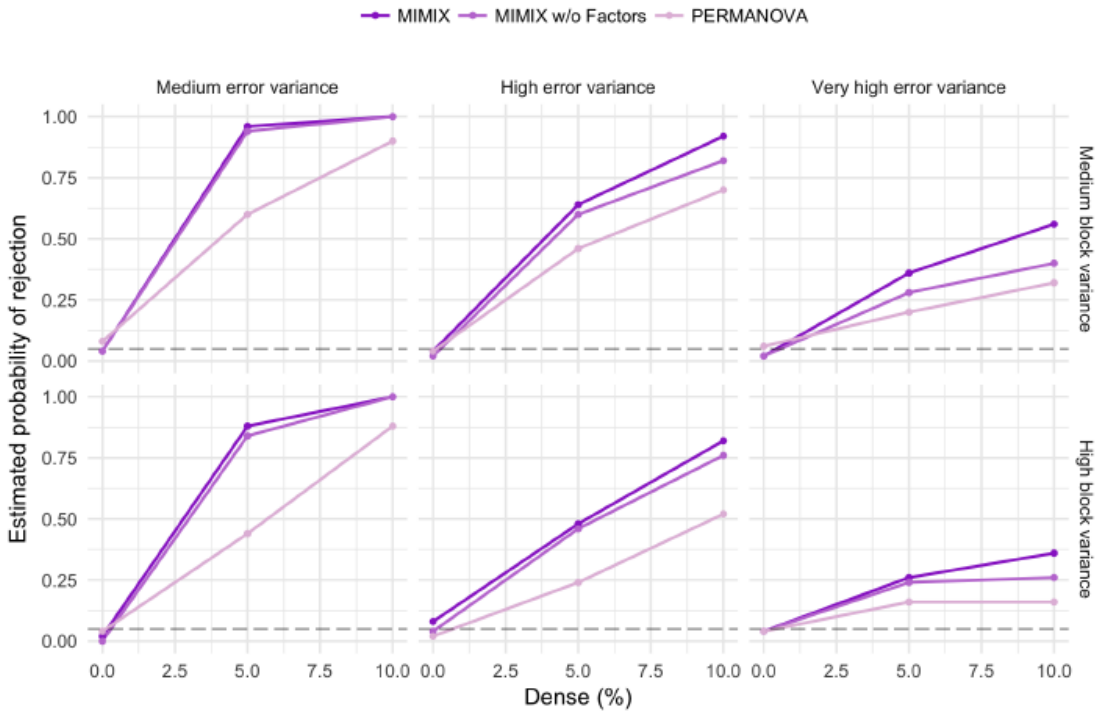
Least square mean estimates (points) and corresponding 95% confidence intervals (whiskers) of Haemophilus relative abundance by eosinophilic oesophagitis diagnosis across anatomical sites.

Fang et al. Epidemiology and Infection, 2016

Multivariate Bayesian Mixed-Effects Model

» Grantham et al. 2019 propose a Bayesian mixed-effects model to analyze microbiome data.

- MIMIX performs spike-and-slab variable selection to identify treatment effects on OTUs.
- Bayesian factor analysis with a Dirichlet-Laplace prior clusters OTUs into different factors.
- MIMIX is not currently suited for handling data from longitudinal studies



Bayesian MIMIX outperforms PERMANOVA and the non-Bayesian MIMIX.

Grantham et al. Journal of the American Statistical Association , 2019

Take home …

- Microbiome data are complex and sparse. Bias in microbiome data analysis can impact interpretation and discovery.
- A compositional data analysis can help identify and solve problems with microbiome complex data. Analysis with absolute abundances is better when possible.
- Zero inflated models and Dirichlet models can fit microbiome data quite well.
- ZINB mixed models can be fitted to data collected repeatedly from individual subjects.
- Bayesian form of MIMIX model outperforms both PERMANOVA and non-Bayesian MIMIX models.

Acknowledgements

- *Open science lovers*

  R and Bioconductor communities for the great open source materials and support. ☺

References and useful resources

References

- P.J. McMurdie, S. Holmes phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data PLoS One, 8 (4) (2013).
- P.S. La Rosa, Y. Zhou, E. Sodergren, G. Weinstock, W.D. Shannon Hypothesis Testing of Metagenomic Data J. Izard, M.C. Rivera (Eds.), Metagenomics for Microbiology, Academic Press, Waltham, MA, USA (2015), pp. 81-96.
- Pajau Vangay, Benjamin M Hillmann, Dan Knights, Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks, *GigaScience*, Volume 8, Issue 5, May 2019.
- Anderson, M.J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels).
- Lahti L., Salojarvi J. 2014. Microbiome R Package.
- Wu, G. D., Compher, C., Chen, E. Z., Smith, et al., (2016). Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. Gut, 65(1), 63
- Kelley, S. T., Skarra, D. V., Rivera, A. J., Thackray, V. G. (2016). The Gut Microbiome Is Altered in a Letrozole-Induced Mouse Model of Polycystic Ovary Syndrome. PloS one, 11(1), e0146509.

Useful resources

- Orchestrating microbiome analysis with Bioconductor https://microbiome.github.io/OMA/resources.html#data-containers-1
- Human Microbiome project datasets https://commonfund.nih.gov/hmp/databases , https://portal.hmpdacc.org/
- Microbiome Learning Repo (ML Repo) https://knights-ab.github.io/MLRepo/
- Microbiome Discovery Tutorial: https://www.youtube.com/watch?v=6564K4-_DBI&list=PLOPiWVjg6aTzsA53N19YqJQeZpSCH9QPc&index=2&ab_channel=DanKnights

Books

- Susan Holmes, Wolfgang Huber, Modern Statistics for Modern Biology. Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2019. https://web.stanford.edu/class/bios221/book/
- Xia Y., Sun J., Chen DG. (2018) Statistical Analysis of Microbiome Data with R. ICSA Book Series in Statistics. Springer, Singapore.

Thank You

Eliana Ibrahimi

eliana.ibrahimi@fshn.edu.al

www.fshn.edu.al