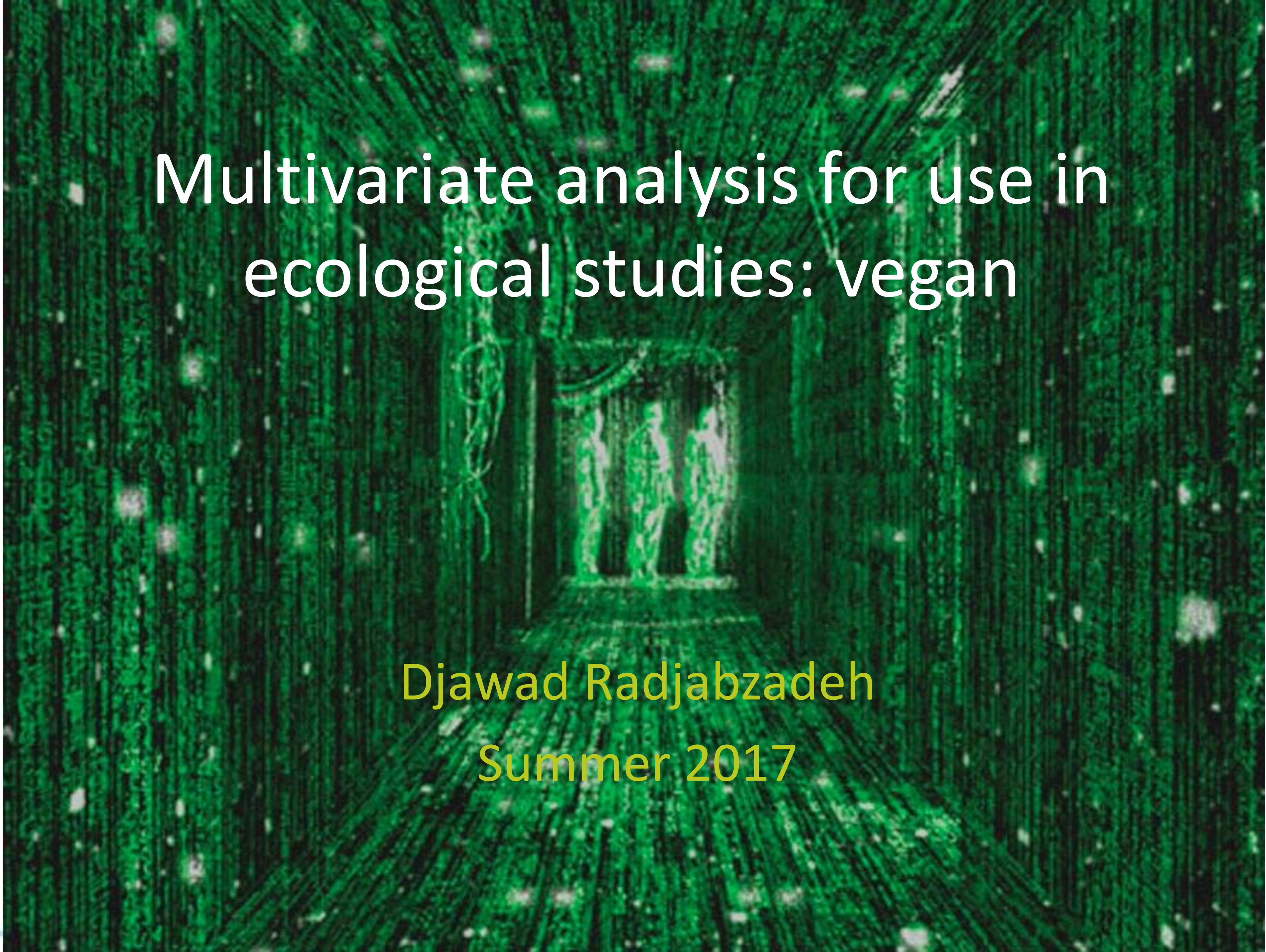


Multivariate analysis for use in ecological studies: vegan

Djawad Radjabzadeh

Summer 2017

OVERVIEW

- Analysis toolkit
- vegan
- Microbiome Variation in two cohorts
- GWAS
- Questions

OVERVIEW

- Analysis toolkit
- vegan
- Microbiome Variation in two cohorts
- GWAS
- Questions

Qiime

- QIIME is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data to publication quality graphics and statistics
- Written by the team of Gregory Caporaso at Colorado University (*Nature Methods* 7, 335 – 336, 2010)

Mothur

- Mothur is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data to statistical analysis.
- Written by the team of Patrick Schloss at University of Michigan (*Appl Environ Microbiol,* 2009. 75(23):7537-41)

Phyloseq

- Phyloseq package is a tool to import, store, analyze, and graphically display complex phylogenetic sequencing data that has already been clustered into Operational Taxonomic Units (OTUs)
- Written by the team of McMurdie and Holmes (*PLoS ONE.* 8(4):e6121; 2013) at Stanford University.

MaAsLin

- MaAsLin is a multivariate statistical framework that finds associations between clinical metadata and microbial community abundance or function (Huttenhower lab)

Vegan

- Multivariate Analysis of Ecological Communities in R: “vegan” package in R
  - vegan is developed by Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O’Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner.
  - The functions in the vegan package contain tools for diversity analysis ordination methods and tools for the analysis of dissimilarities.
  - It provides most standard tools of descriptive community analysis.

OVERVIEW

- Analysis toolkit
- vegan
- Microbiome Variation in two cohorts
- GWAS
- Conclusion
- Questions

Useful web resources

- Vegan tutorial:

  http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf
- The little book of r for multivariate analyses:

  http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html#means-and-variances-per-group
- Ordination Methods by Michael Palmer:

  http://ordination.okstate.edu/overview.htm#Nonmetric_Multidimensional_Scaling
- Community analyses lectures by Jari Oksanen:

  http://cc.oulu.fi/~jarioksa/opetus/metodi/

Univariate statistics to measure community dynamics

- Richness (*R* or *S*, Either local or regional)
- Shannon index (*H'*; Shannon &Weaver 1949): Incorporates richness as well as the relative abundances into a metric
- Simpsons index (*D* or *λ*; Simpson 1949): Emphasizes evenness

Species	low.light	mid.light	high.light
A	0.75	0.38	0.08
B	0.62	0.15	0.15
C	0.24	0.52	0.18
D	0.33	0.57	0.52
E	0.21	0.28	0.54
F	0.14	0.29	0.56

Metric	low.light	mid.light	high.light
Richness	6	6	6
H’	1.63	1.71	1.60
D	0.78	0.81	0.77

No information about individual species responses

- Species A and B are dominant in low light
- Species E and F are dominant in high light

Multivariate statistics

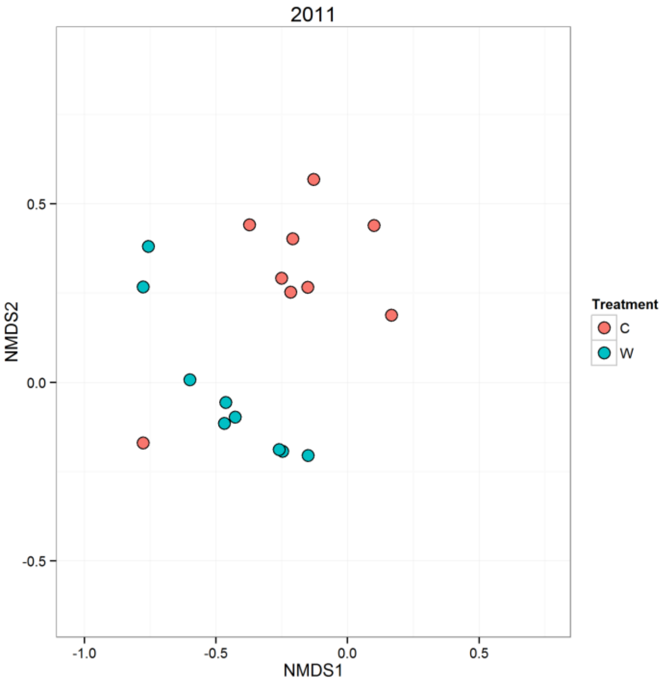
3 parts:

1. Dissimilarity matrices
2. Ordinations
3. Statistical tests of differences between or among communities

Dissimilarity metrics are the building blocks used in many multivariate statistics

- A dissimilarity matrix is simply a table that compares all local communities. The higher the number, the more dissimilar the communities are.

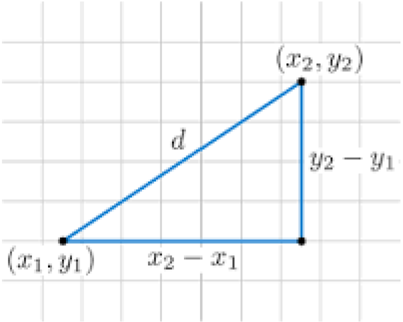
- Visualization of representation (ordination).
- Statistical tests.



Types of dissimilarity metrics

- Euclidean distance
  - Operates in species space
    - Meaning that each species gets its own orthogonal axis in multidimensional space.
  - Differences are squared → single large differences become very important
  - In vegan: function “vegdist” perform ED (method=euclidean).

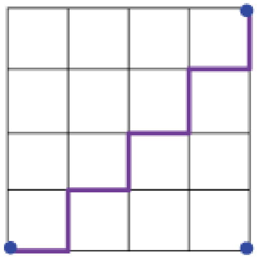
$$d_{jk} = \sqrt{\sum_{i=1}^{N}(x_{ij} - x_{ik})^2} \quad \text{Euclidean}$$



- Manhattan-type distances
  - Bray-Curtis (abundance data)
  - Jaccard (presence-absence)
  - Use sums or differences instead of squared terms making it less sensitive to single differences
  - In vegan function “vegdist” perform Bray-Curtis and Jaccard measurenments (method=bray or jaccard)

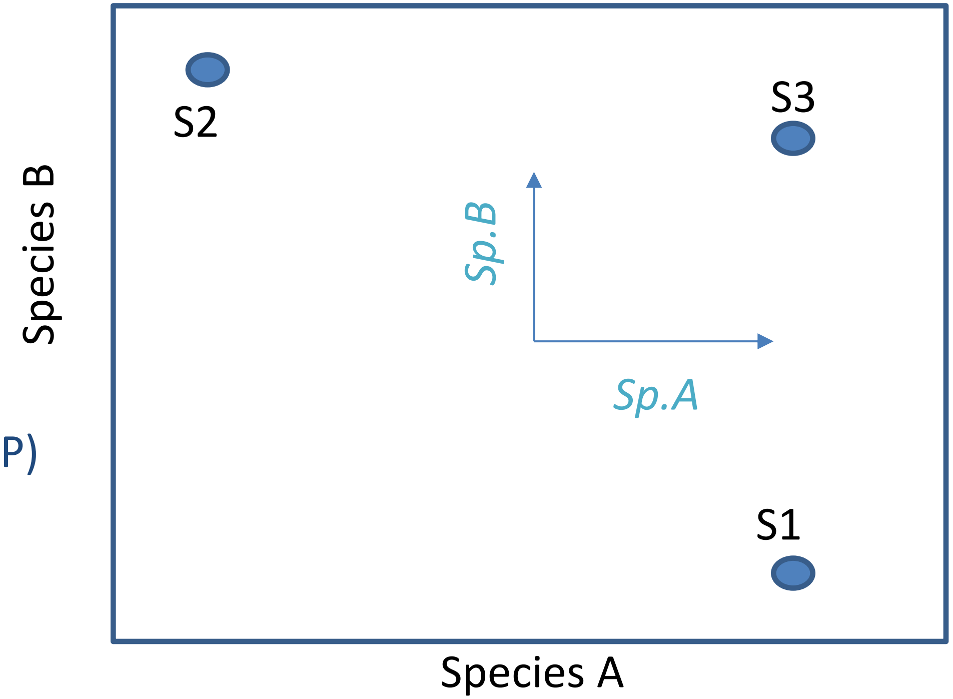
$$d_{jk} = \frac{A + B - 2J}{A + B}$$

$$A = \sum_{i=1}^{N} x_{ij} \qquad B = \sum_{i=1}^{N} x_{ik} \qquad J = \sum_{i=1}^{N} \min(x_{ij}, x_{ik})$$



Ordinations

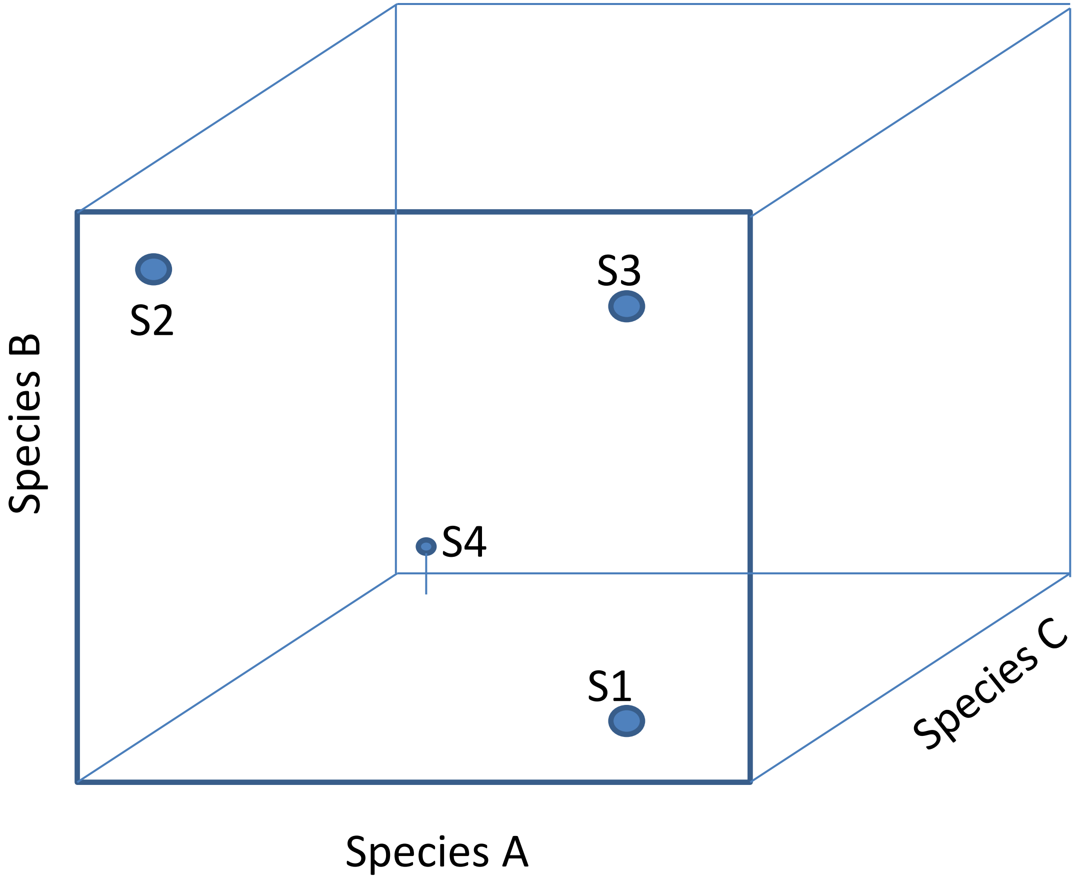
- Basically, ordinations plot the communities based on all response variables (e.g. species responses) and then squish this into 2 or 3 dimensions.
- Example 1: 2 species, 2 axes.



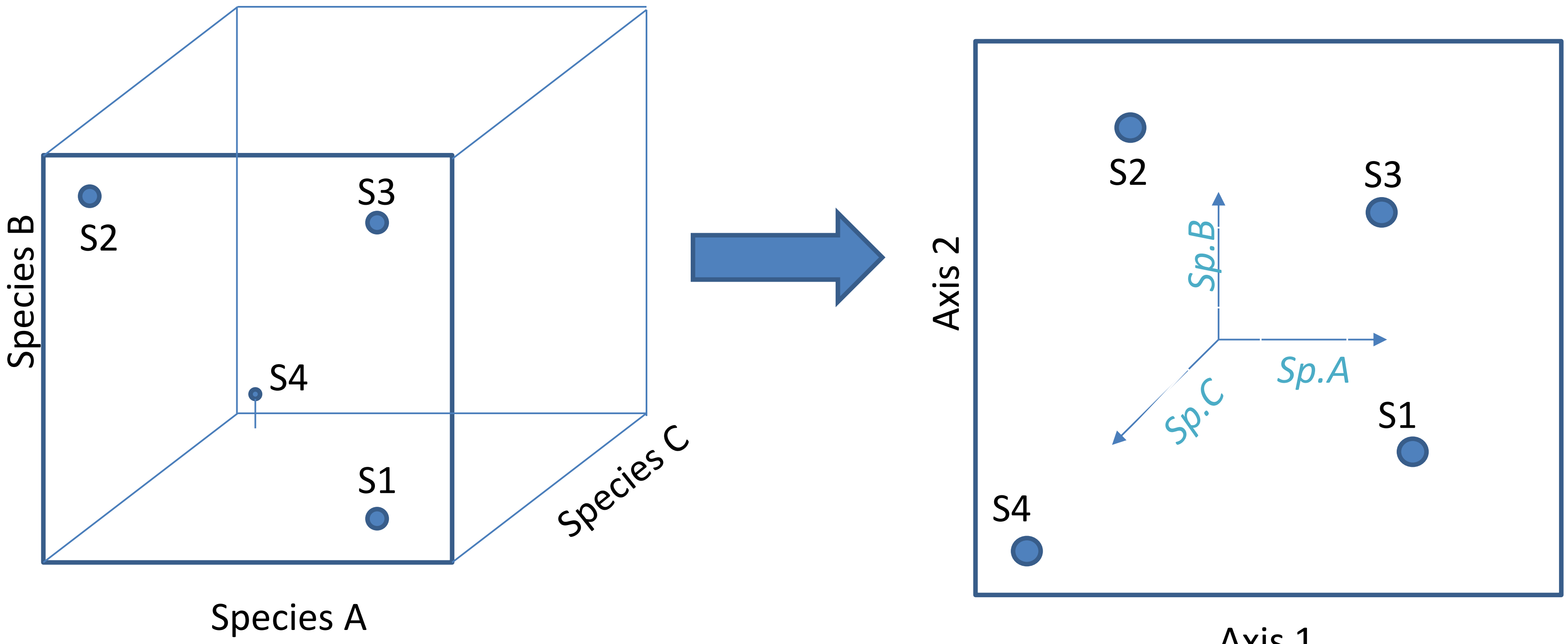
- Constrained analysis of proximities (CAP)
  - You can plug in any dissimilarity matrix into this (vegan: function (“capscale”)
- Redundancy analysis (RDA, vegan: function “rda”)
  - Constrained version of PCA
- Constrained correspondence analysis (CCA, vegan: function “cca”)
  - Based on Chi-squared distances
  - Verschil rda en cca

Ordinations

- Plots the communities based on the response variables and then squishing this into 2 or 3 dimensions.
- Example 2: 3 species, 3 axes
- Etc up to *n* response variables
  - We can't visualize this well after 3 axes but it happens



Ordinations



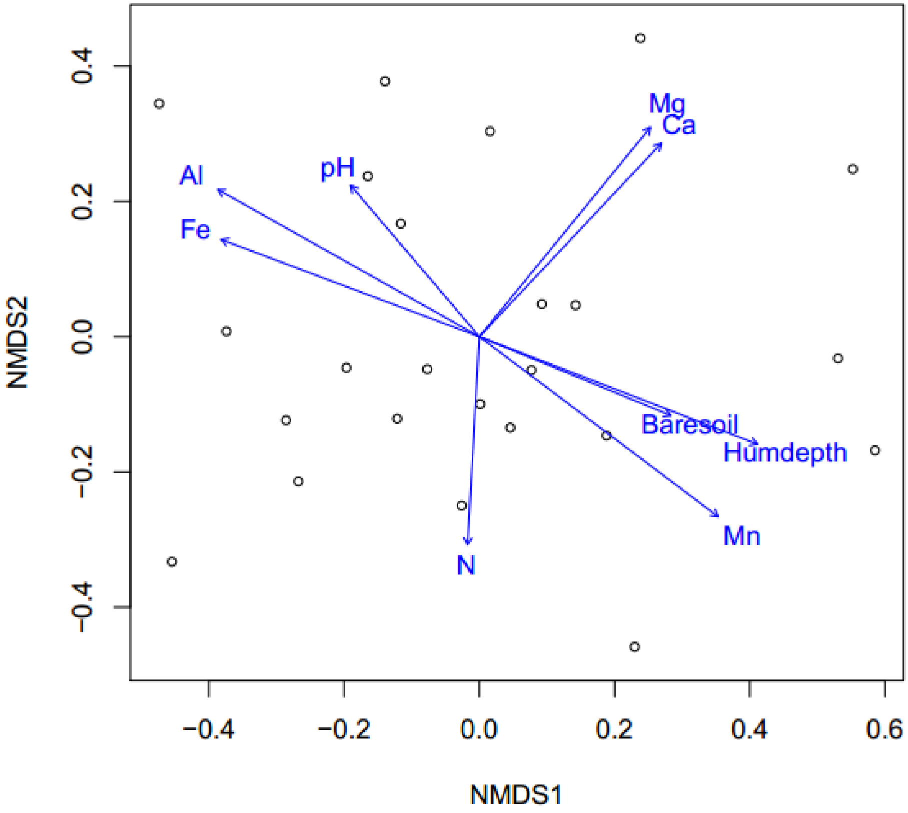
Ordination

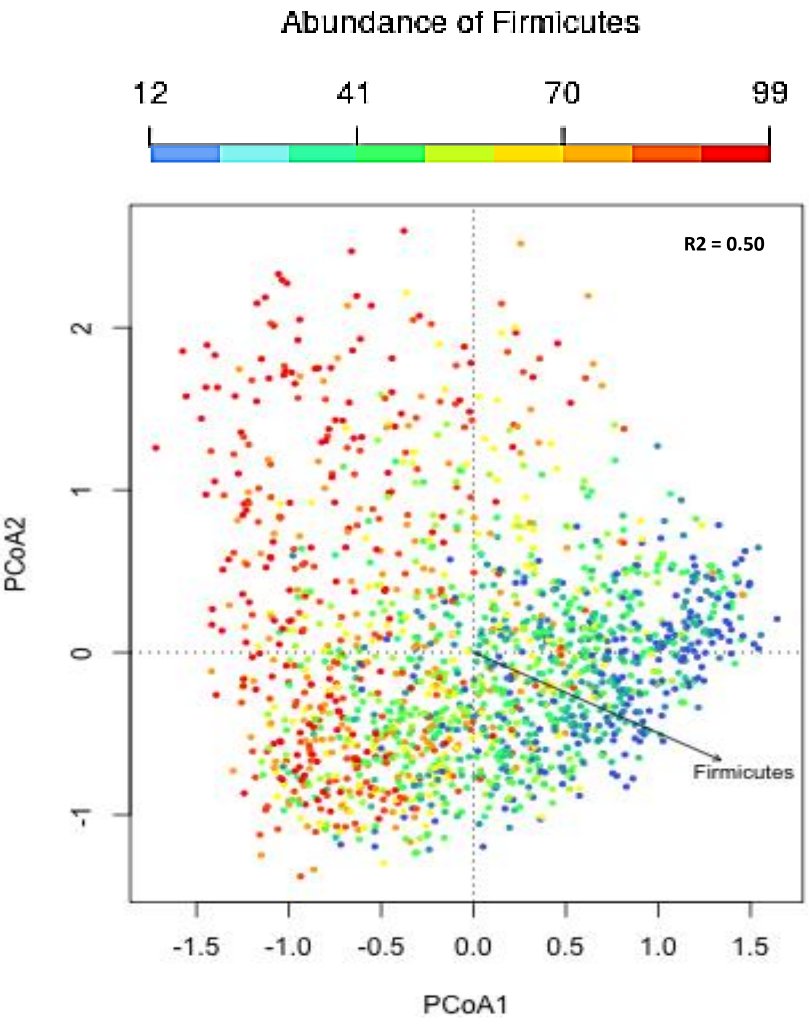
- Basically, ordinations plot the communities based on all response variables (e.g. species responses) and then squish this into 2 or 3 dimensions.
- Different way of ordinations: based on data used.

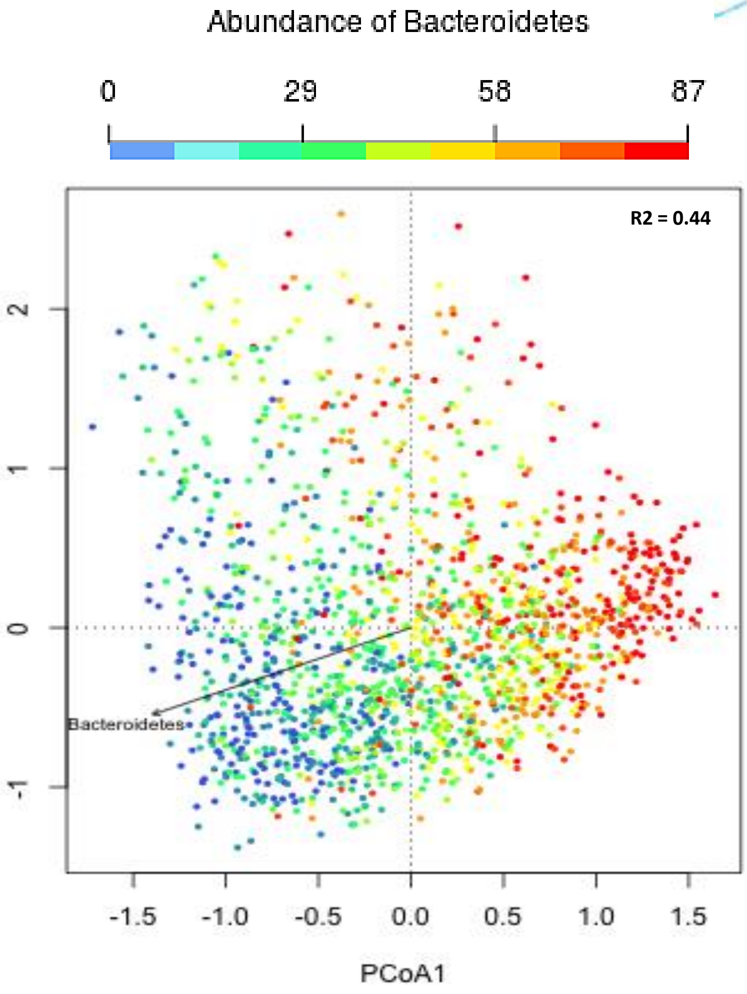
- Principle components analysis (PCA, vegan: function “euclidean”)
  - Uses Euclidean distances to map samples with the 2 or 3 axes that explain the majority of variation
  - Use with environmental/abiotic data
- Principle coordinates analysis (PCoA; vegan: function “capscale”)
  - Acts like PCA but uses a dissimilarity matrix instead of pulling straight from the data (CAP)
- Redundancy analysis (RDA, vegan: function “rda”)
  - Constrained version of PCA
- Constrained correspondence analysis (CCA, vegan: function “cca”)
  - Based on Chi-squared distances

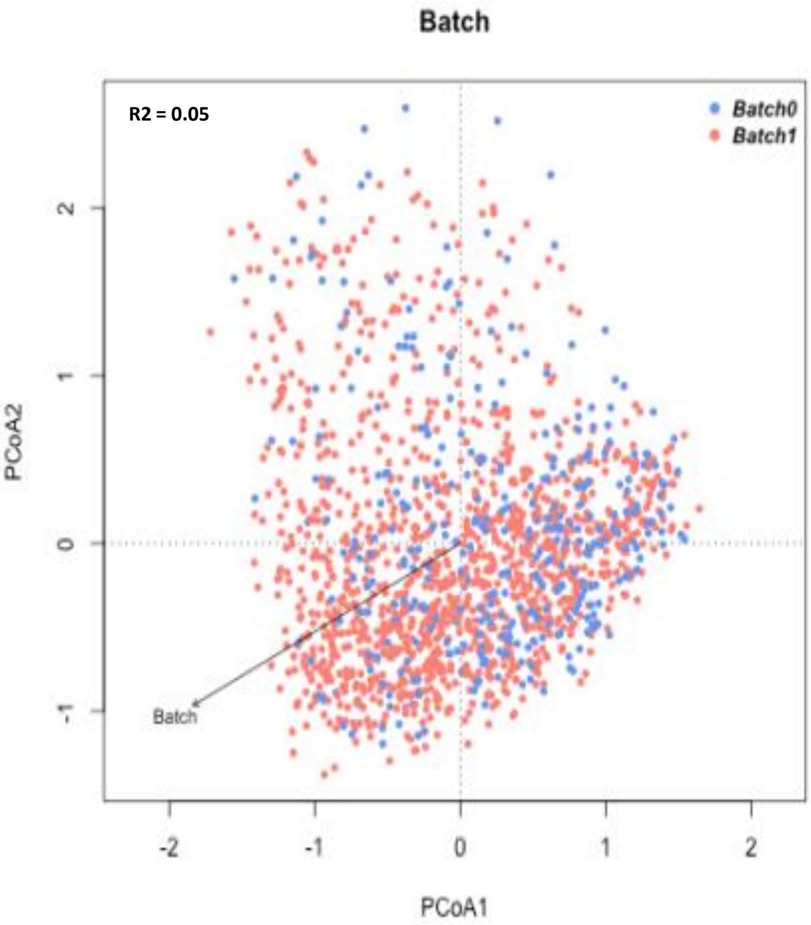
Incorporating environmental data into ordination

- Can overlay *vectors* of environmental data on top of community data
  - Vectors supply information about the direction and strength of environmental variables
  - Easy to interpret the effects of many variables
  - It assumes all relationships are linear. This might not be the case…
  - Vegan: function "envfit" and "bioenv".

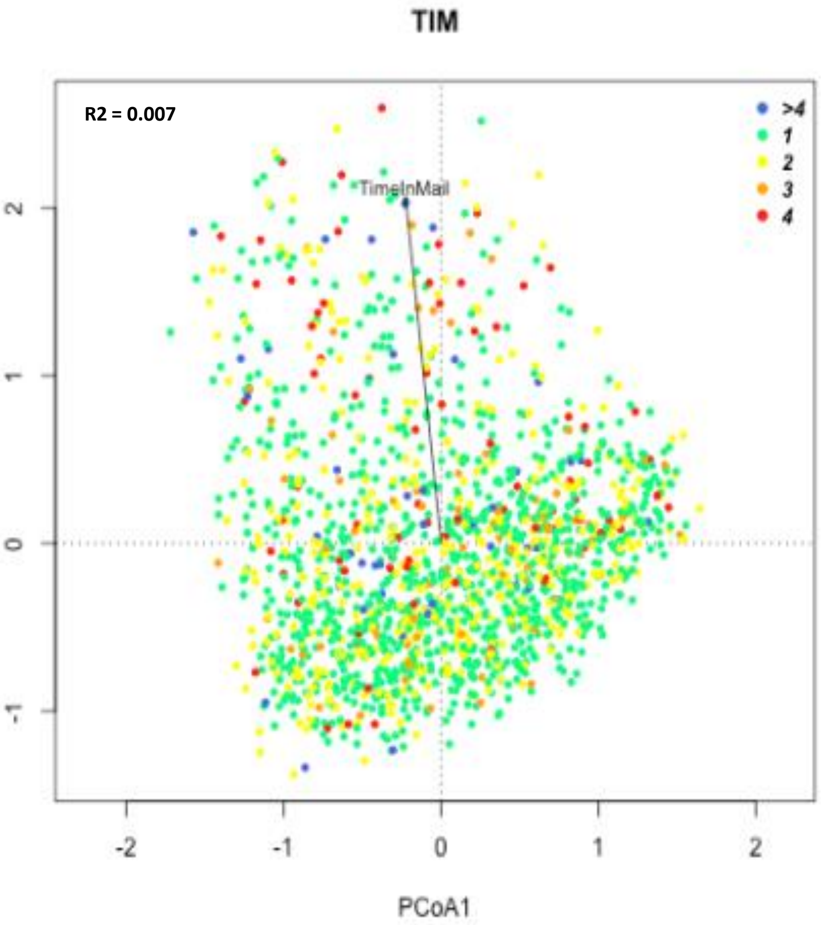



Abundance of Firmicutes
12
41
70
99
R2 = 0.50
Firmicutes
PCoA2
2
1
0
-1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
PCoA1


Abundance of Bacteroidetes
0
29
58
87
R2 = 0.44
Bacteroidetes
2
1
0
-1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
PCoA1


Batch

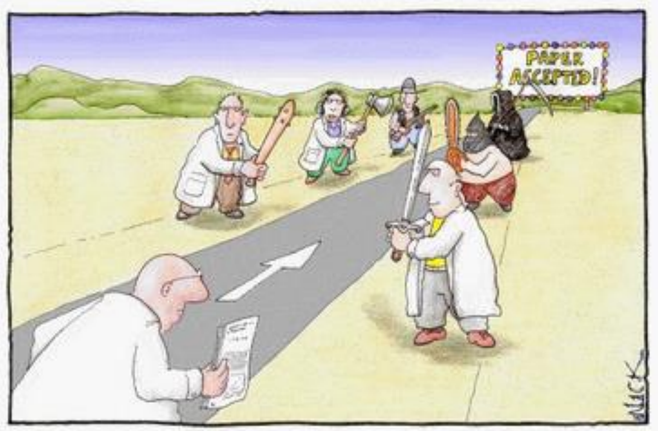
R2 = 0.05
● Batch0
● Batch1
Batch
PCoA2
PCoA1
-2
-1
0
1
2


TIM

R2 = 0.007
● >4
● 1
● 2
● 3
● 4
TimeInMail
PCoA1
-2
-1
0
1
2

Ordination by itself is not a robust statistical test

- Ordination is great for visualizing your data, BUT… we need to back it up.
- One way is to calculate confidence ellipses around the centroid
- Another way is to use resemblance-based permutation methods
  - They give P values…



Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

Resemblance-based permutation methods

- They compare *n* dimensional data instead of ordination data squished into 2 or 3D
- Many assumptions of regular MANOVAs are violated with ecological community data (see Clarke 1993) → creation of new methods for analyzing multivariate data
- 5 majorly used methods:
  - Permutational MANOVA (or PERMANOVA)
  - Analysis of similarities (ANOSIM)
  - Mantel's test
  - Permutational analysis of multivariate dispersions (PRMDISP)
  - Similarity percentages of component species or functional groups (SIMPER)

PERMANOVA

- Calculates a pseudo-F statistic
  - Pseudo-F is identical to a normal F statistic if there is only one response variable
- This pseudo-F is calculated using the original data and compared with a distribution of pseudo F statistics from many random permutations. This step is the same as ANOSIM.
- Vegan: function “adonis”

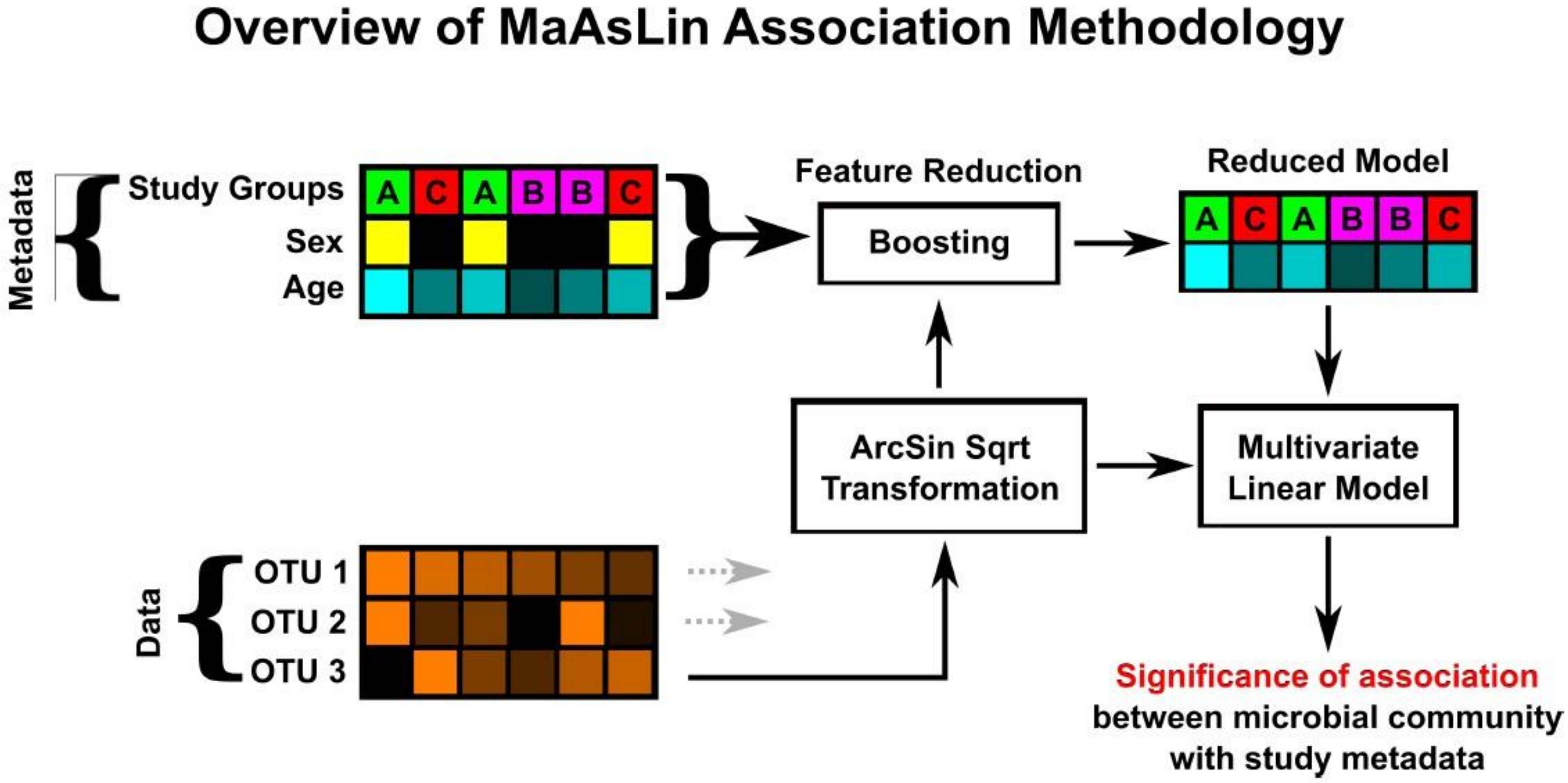
(See Anderson 2001, 2005 for more detail)

ANOSIM – Clarke 1993

- Ranks dissimilarities among local communities from 1 to the number of comparisons made.
- Then looks at averages of ranked dissimilarities within and among groups.
- Compares these averages to random permutations of the R values to get p-value.
- Vegan: function “anosim”

(See Clarke 1993 for more detail)

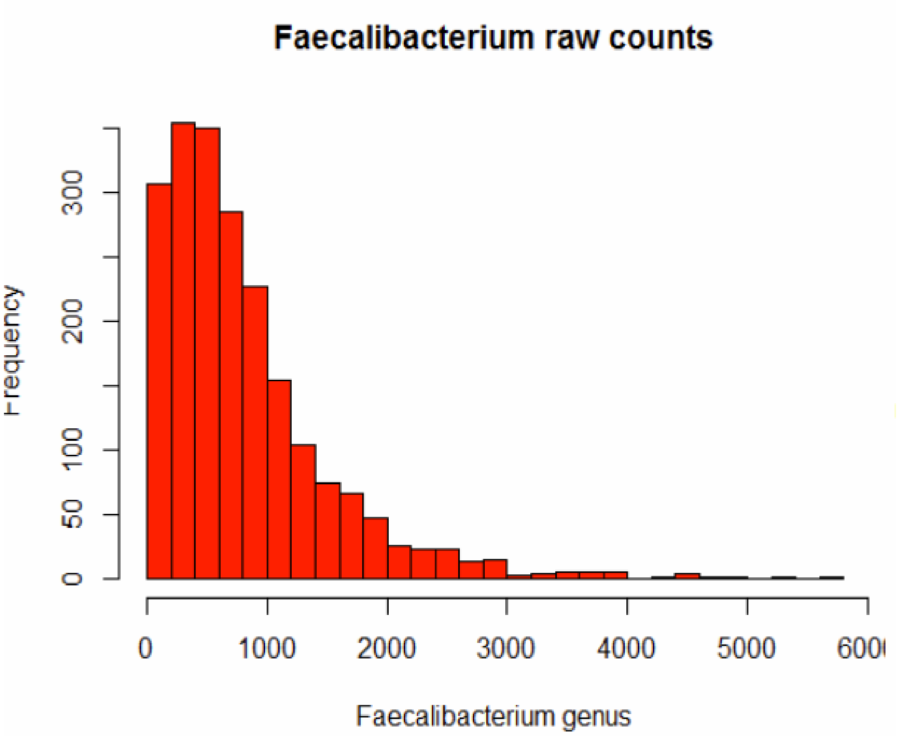
MaAsLin: Multivariate Association with Linear Models


Overview of MaAsLin Association Methodology
Metadata
Study Groups
Sex
Age
A
C
A
B
B
C
Feature Reduction
Boosting
Reduced Model
A
C
A
B
B
C
ArcSin Sqrt Transformation
Multivariate Linear Model
Data
OTU 1
OTU 2
OTU 3
Significance of association
between microbial community
with study metadata

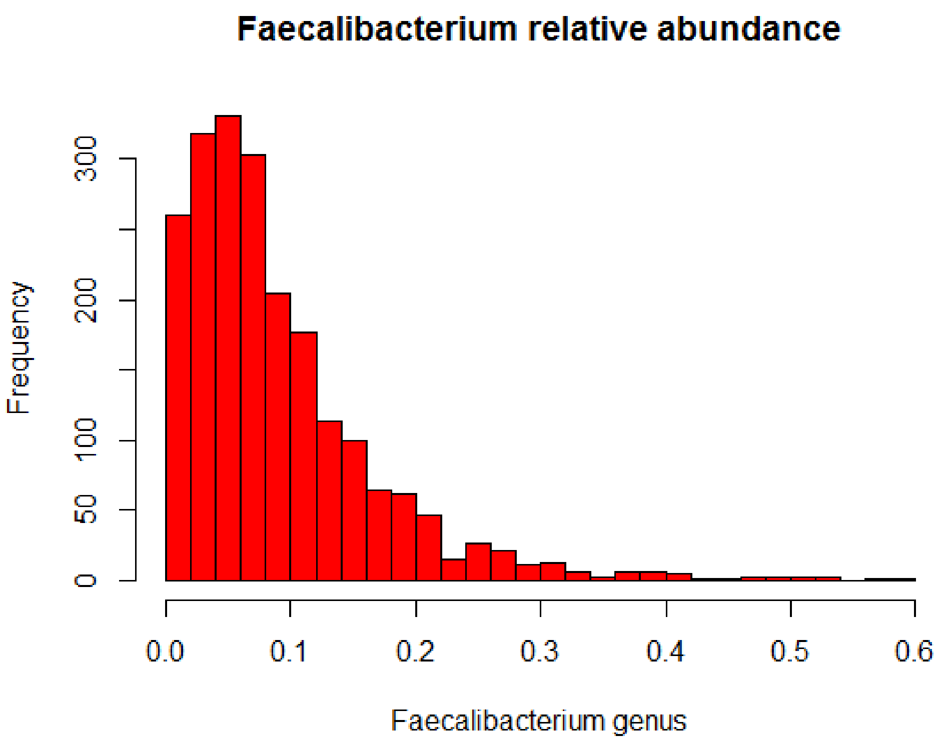
https://huttenhower.sph.harvard.edu/maaslin --

Department of Biostatistics, Harvard T.H. Chan School of Public Health

- MaAsLin works based on relative abundance data: #Counts OTUx/#Counts Total per individual



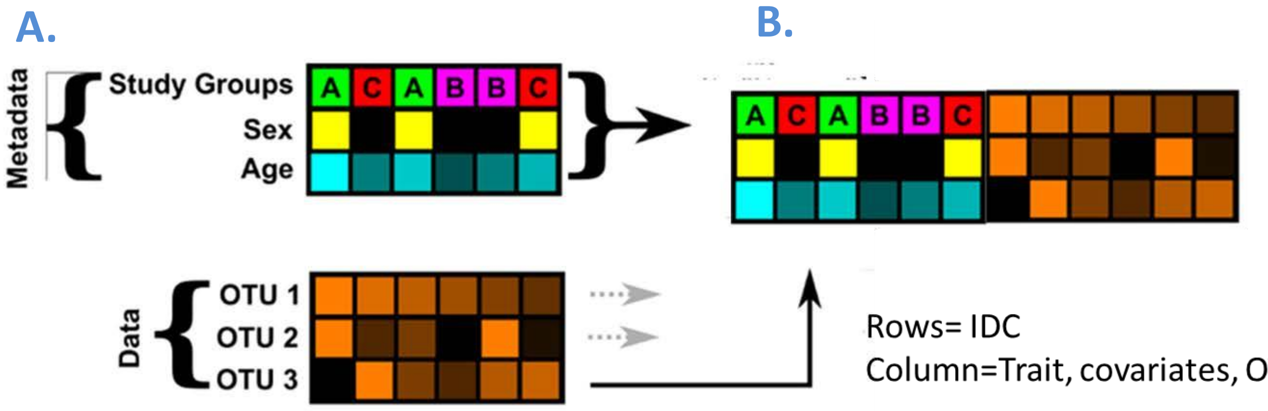
Faecalibacterium raw counts



Faecalibacterium relative abundance

Metadata - OTU_relative abundances, input MaAsLin

- Meta data and the obtained relative abundances are merged



Rows= IDC
Column=Trait, covariates, OTU RA

ID	sex	BMI	run	bmd	height	TimeInMa	agechild9	Bact1	Bact2
2541	Boy	19.4	1	0.753769	146.5	4	9.801506	0.163259	0.010891
2311	Boy	18.8	1	0.659904	136.2	4	9.489391	0.052258	0.217482
2840	Girl	17.2	1	0.616391	143.7	1	9.864476	0.144199	0.013276
2715	Boy	14.3	1	0.670339	136.6	1	9.711157	0.417467	0.028614
981	Boy	14.6	1	0.646141	136.7	2	9.667351	0.077223	0.023987
705	Boy	15.5	1	0.62355	135.6	5	9.54141	0.028879	0.171855
211	Boy	15.8	0	0.63742	151.6	5	9.675565	0.170183	0.150593
154	Girl	14.7	0	0.62141	150	3	9.659138	0.090777	0.134765
3028	Girl	16.1	1	0.590749	142.5	3	9.672827	0.039035	0.351317
217	Girl	15.5	0	0.721635	145.3	4	9.71937	0.03964	0.189926
449	Boy	16.5	1	0.746637	146.6	4	9.817933	0.095402	0.077844
2784	Boy	18.4	1	0.628587	140.6	5	9.596167	0.224372	0.115795

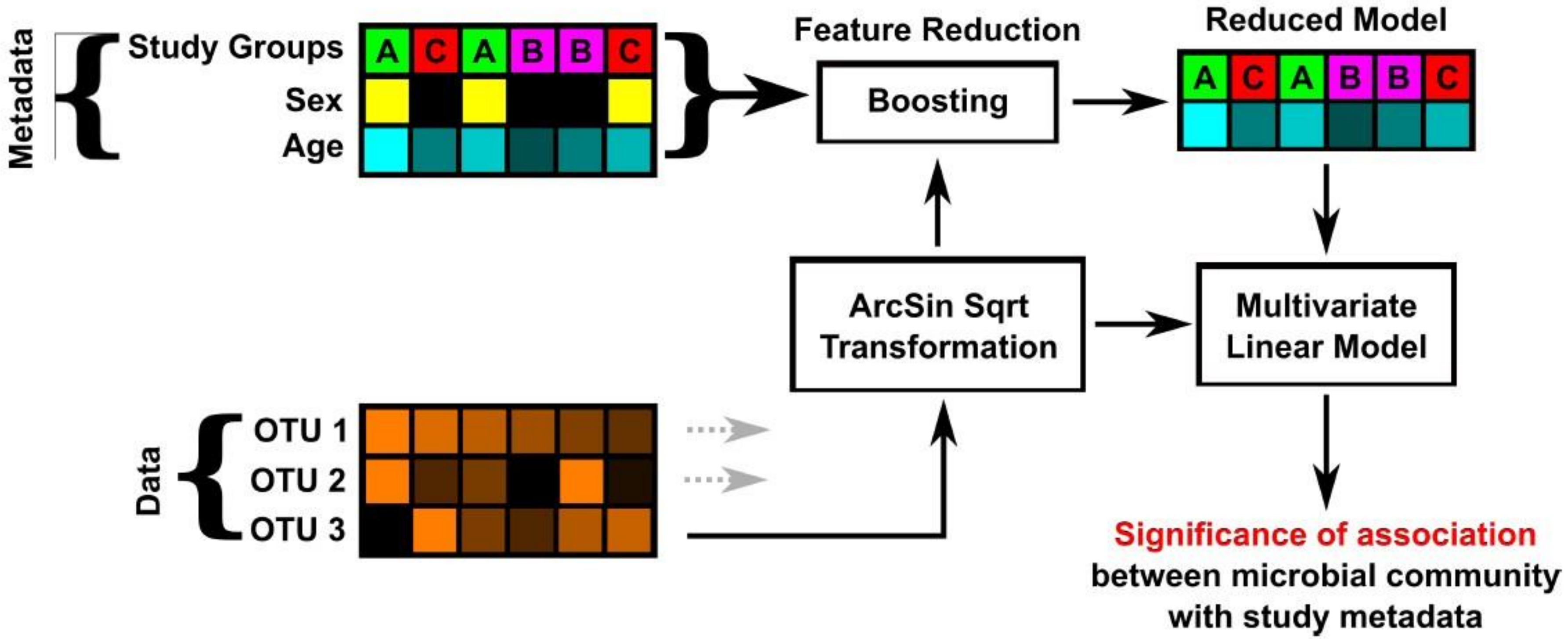
Input files – R command file

In R:

```
library(Maaslin)
Maaslin(‘inputfile.txt',
‘output suffix',
strInputConfig=‘Configuration.read.config',
fAllvAll = TRUE,
strForcedPredictors = c(“all_covariates"),
strModelSelection =  "none")
```

MaAsLin: Multivariate Association with Linear Models

Overview of MaAsLin Association Methodology



Metadata

Study Groups

Sex

Age

Feature Reduction

Boosting

Reduced Model

ArcSin Sqrt Transformation

Multivariate Linear Model

Data

OTU 1

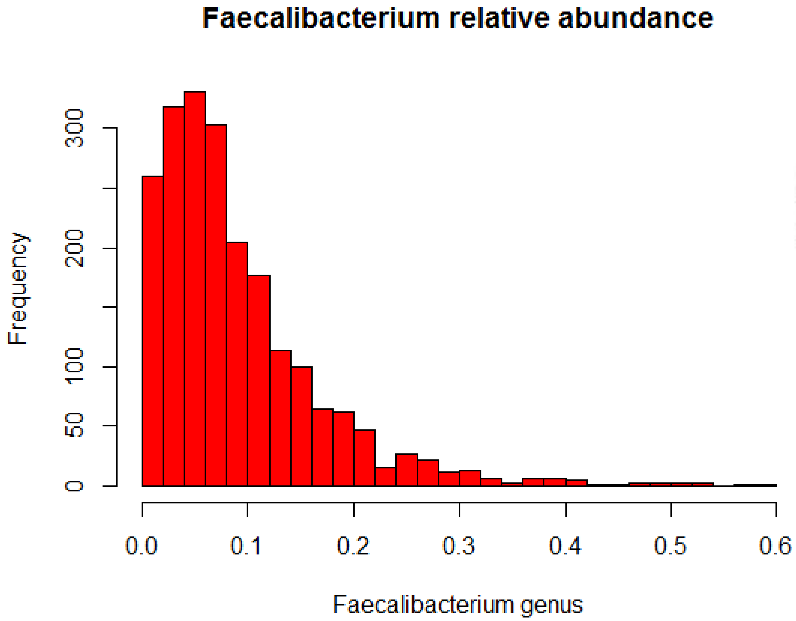
OTU 2

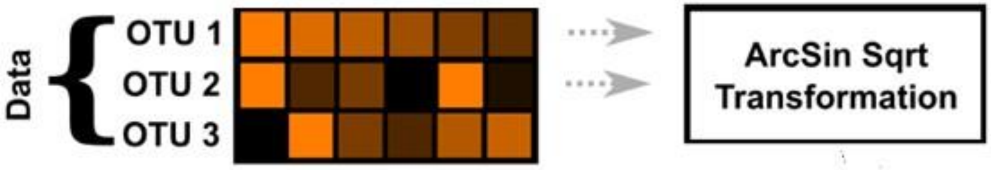
OTU 3

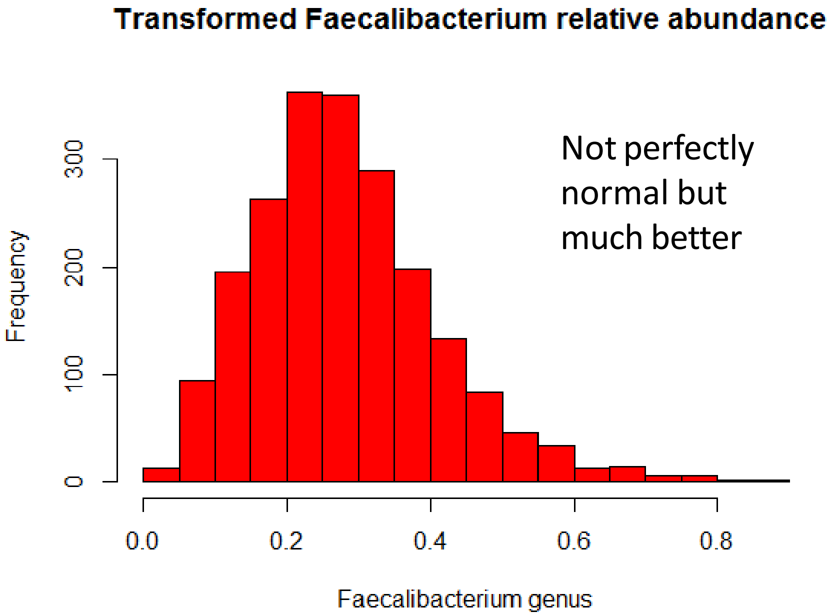
Significance of association between microbial community with study metadata

https://huttenhower.sph.harvard.edu/maaslin

Relative Abundances– Arcsine transformation





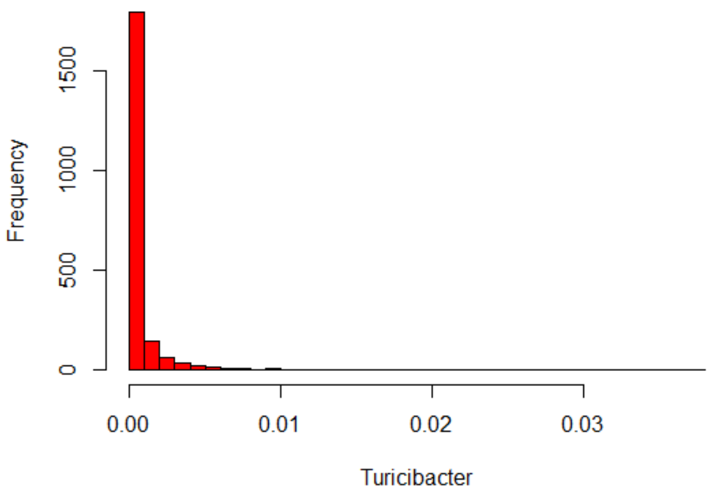


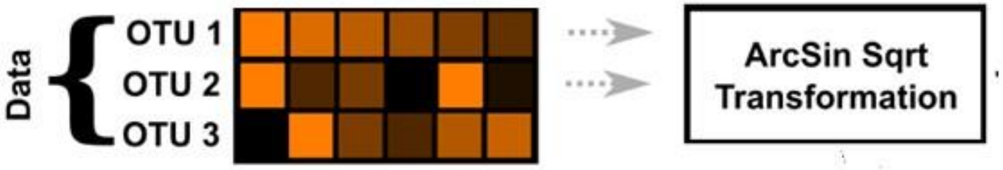
Not perfectly normal but much better

ArcSin transformation is default to MaAslin but you have an option to change this default transformation

Relative Abundances– Arcsin transformation

Turicibacter relative abundance





Data

OTU 1

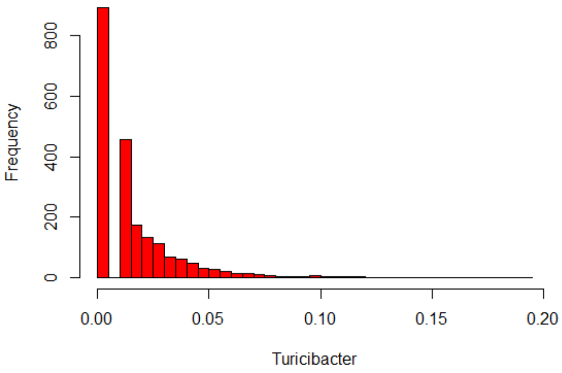
OTU 2

OTU 3

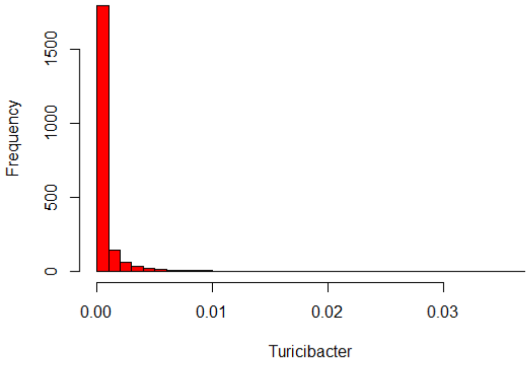
ArcSin Sqrt Transformation

Zero inflated models!

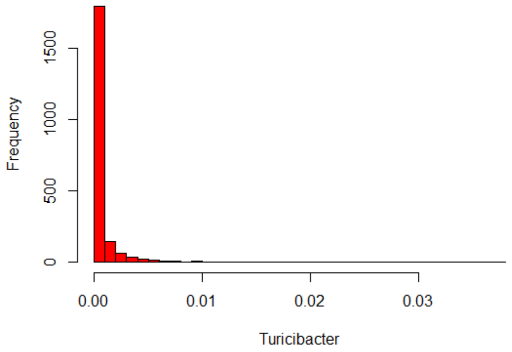
Arcsin Turicibacter relative abundance



Log Transformed Turicibacter relative abundance



ICH Transformed Turicibacter relative abundance



MaAsLin Multivariate linear Model

Arcsin(OTUxx) ~ age + sex + BMI + PCs + Technical covariates + …

The model and link function can be modified, mixed models are allowed (random variables can be defined).Different parameters of QC can be chosen

- All missing data will be imputed to the median value, unless specified

  strNoImpute= c( )

- By default relative abundances less than 0.0001 or outliers will be substituted by median of the sample

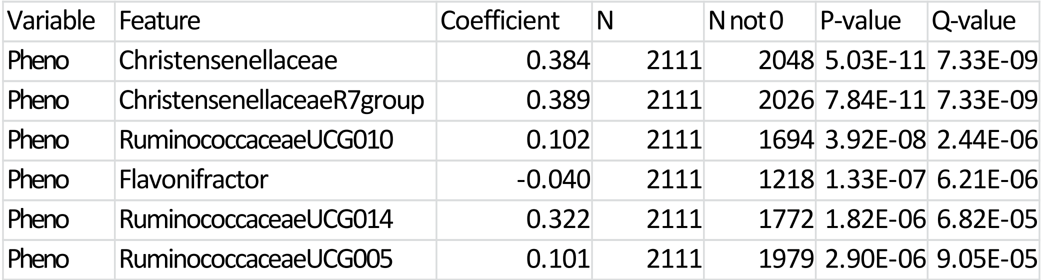
  dMinAbd = XX, dOutlierFence = XX, dPOutlier = XX

- By default if 10% of the data is missing for a variable this will not be consider in the analysis, also, for analysis at least 10% samples need to have >0.0001 R.A for OTU in order to be analyzed 10% of the samples

  dMinSamp =XX

MaAsLin Output Results

- **QC folder:** Run parameters used, final dataset used, final read configuration used
- **Log file:** save the *R.history* of the analysis
- If multiple phenotypes analyzed simultaneously then **multiple result files** will be generated



Variable	Feature	Coefficient	N	N not 0	P-value	Q-value
Pheno	Christensenellaceae	0.384	2111	2048	5.03E-11	7.33E-09
Pheno	ChristensenellaceaeR7group	0.389	2111	2026	7.84E-11	7.33E-09
Pheno	RuminococcaceaeUCG010	0.102	2111	1694	3.92E-08	2.44E-06
Pheno	Flavonifractor	-0.040	2111	1218	1.33E-07	6.21E-06
Pheno	RuminococcaceaeUCG014	0.322	2111	1772	1.82E-06	6.82E-05
Pheno	RuminococcaceaeUCG005	0.101	2111	1979	2.90E-06	9.05E-05

187/205

- N is the sample size in your data (not necessarily the N used in analysis)
- Q value, the FDR adjusted P value
- Coefficient is in Arcsin(sqrt(R.A OTU)) units

OVERVIEW

- Analysis toolkit
- vegan
- Microbiome Variation in two cohorts
- GWAS
- Conclusion
- Questions

Microbiome variation in RS and GenR: analysis plan

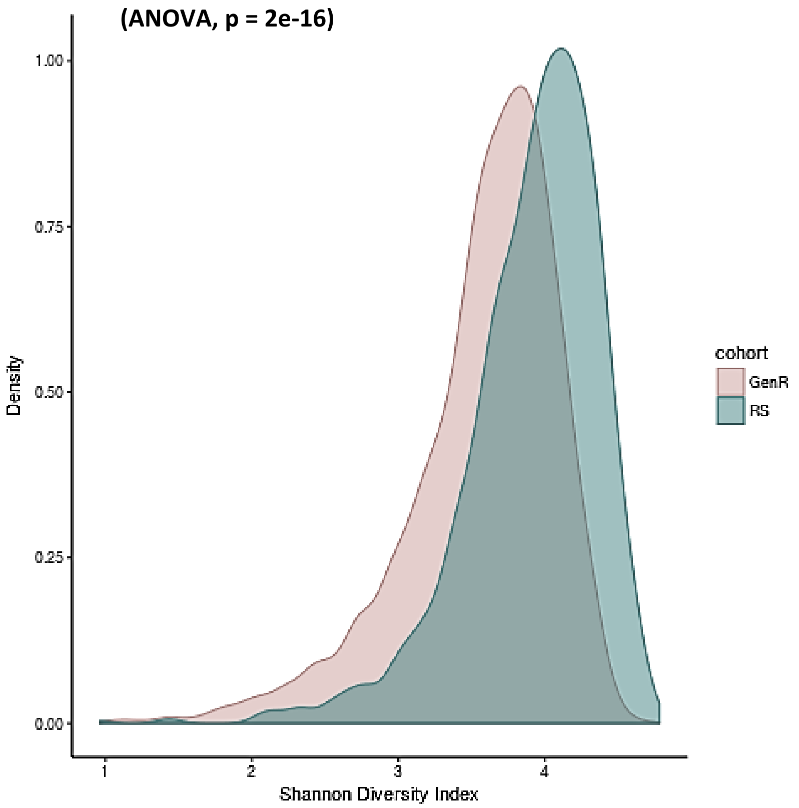
- Ethinicity
  - Only North-European subjects in both cohorts
- Age
  - Limited range of age in GenR (~10 years old)
  - Only subjects within range of 52-62 years old in RS

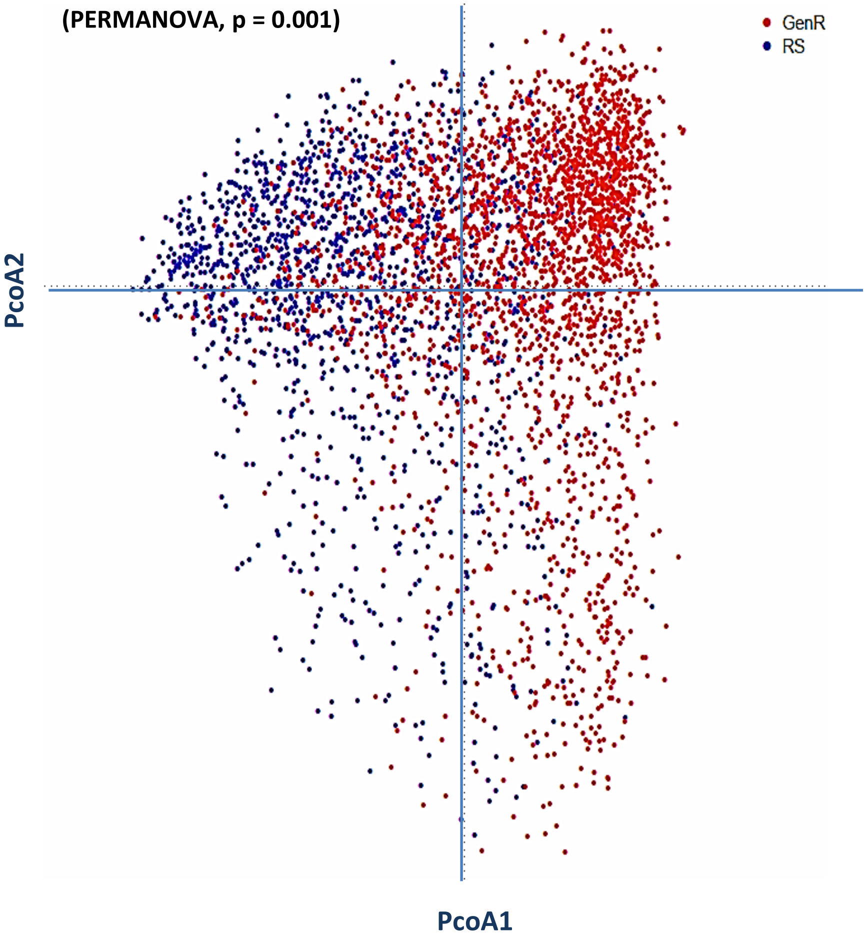
- Resulted in 1,081 subjects in RS and 1,463 subjects in GenR
- The reads (10K) of all samples piled up → pipeline → OTU table

Microbiome variation in RS and GenR: analysis plan

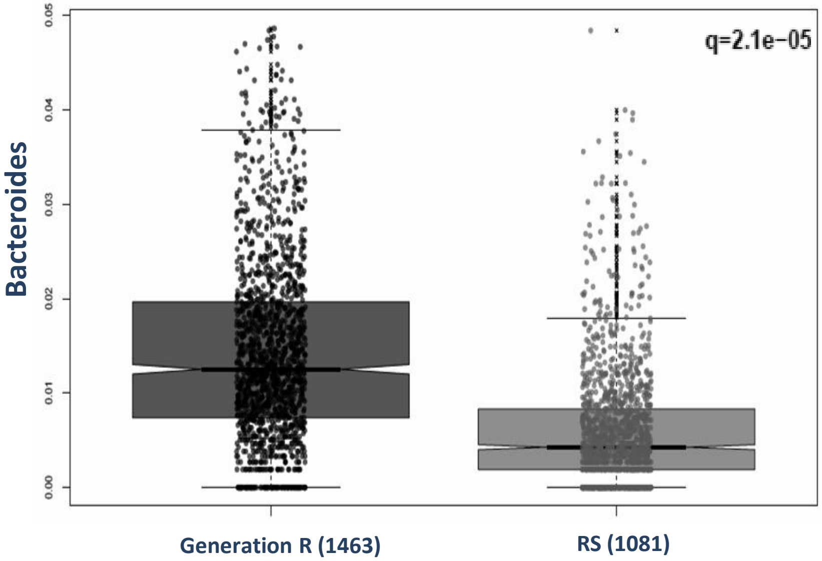
- Beta-diveristy(in vegan):
  - Calcualting Bray-Curtis
  - Running ordination on dissimilaritiy metric
  - PERMANOVA
- Alpha-diversity (in vegan):
  - Calculating alpha diversity (Shannon, Richness, InvSimpson)
  - ANOVA
- Individual OTU response (MaAsLin):
  - Preparing relative OTU table for analysis
  - Running MaAsLin: adjusting for multiple testing by FDR ($q < 0.05$)

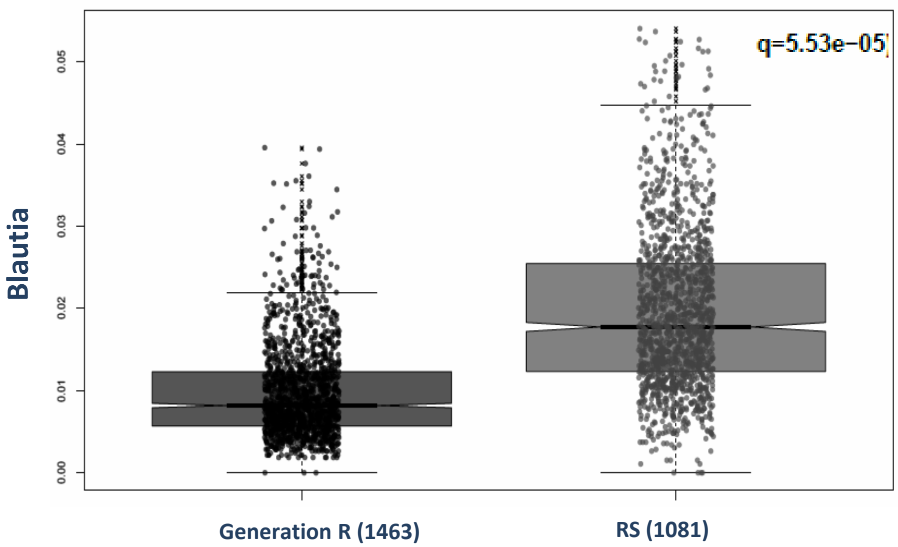
Differecnes in diversity in the two cohort





Major genera characterizing each cohort





Erasmus MC
University Medical Center Rotterdam

Acknowledgments

- Robert Kraaij
- Pelle van der Wal
- Cindy G. Boer
- Carolina Medina
- Fernando Rivadeneira
- Joyce van Meurs
- André Uitterlinden

GTTACACATGGACTGGAGATCACACOMPLEX GENETICSGTGTACGAATTTTAAGGGGATAGT
GTCAGCTTAAGENETIC LABORATORYATGGTCACCATGATCGATGGTTACGAACTTAAGAAA
TACCGAAATTCGGATGINTERNAL MEDICINEGTACTTGGCCCATGAGCTGAGTTCCCAAACGG
CGATTGCACCATAATCGCGACCAATATGWASGCATAATAGCCGGGGSNPATTGCTGAGATAC
GATAACCTTTAAACACACAEXOME SEQUENCINGGATCCGTAGGGACCGCGAATCCGTTTTTT

GENETISCH LABORATORIUM

RAPtorμ
Rotterdam Annotation Pipeline

Questions?

GWAS cohorts (miQTL)

Cohort name	Ethnicity	population subjects	age participants	Imputation done HRC?	microbiome	variable region
SHIP / SHIP-TREND	European (Germany)	2000	adults	yes	16S	V1-V2
CARDIA (Coronary Artery Risk Development in Young Adults)	USA (African-Americans and European-Americans--roughly even split)	550	adults	yes	16S	V3-V4
COPSAC2010	European (from Denmark)	700	children, multiple time points	yes	16S	V4
GEM (Genetic Environmental Microbial) Project	Canada, Israel, US, and UK	1561	6-35 years old	no	16S	V4
LifeLines-DEEP (LLD)	European (Duthc)	1200	adults (>18)	yes	16S, MGS	V4
Metabolic Syndrome in Men (METSIM)	European (Finnish)	938	adults (male?)	no	16S	V4
Rotterdam Study	European (Dutch)	1440	adults	yes	16S	V3-V4
Generation R	Multi-ethnic	2400	children, 9 years old	yes	16S	V3-V4
PopGen	European(Germany)	914	adults	no	16S	V1-V2
FoCus	European(Germany)	1535	1158 population cohort + 377 obesity	no	16S	V1-V2
	Israel	700				
FGFP	Belgian	1000?	adults	?	16S	

Total: 13,938 subjects

GWAS cohorts (miQTL): analysis plan

Four major steps:

- Processing of 16S data
- Processing of SNP microarray data
- Performing the association study
- Performing meta-analysis

Genome-Wide Association Study itself will be performed:

Cutoffs and transformations

- Taxonomies:

  Abundance cutoff: presence in 10% of the samples

  Log (base **e**) transformation on the counts

- SNPs:

  MAF > 1%

  Imputation quality > 0.4

  Genotypes represented in dosages

- Models used (two part model):

  Taxonomy absence/presence as binary trait: logistic regression with Chisquared-based p-value estimation

  For non-zero samples: linear regression model on log-transformed counts with Fisher test-based p-value estimation

- Meta-analysis

  Performed separately, for binary and quantitative models