Microbiome Data Visualization

Justin Gibbons, PhD

How is the microbiome measured?

1. Marker gene studies: Taxa composition
2. Metagenomics: Gene composition
3. Metatranscriptomics: Gene expression

The R Microbiome Package

- Wrapper that provides uniform interface for multiple analysis packages useful for microbiome research
  - ape—phylogenetic analysis
    - Phylogenetic relatedness of samples
  - vegan—ecological analysis
    - Diversity analysis
    - Ordination—Dimensionality reduction
  - igraph—network creation and analysis
    - Analysis/visualization of relationships between taxa or samples
    - Analysis/visualization of relationship between taxa and gene expression
  - Survival—survival analysis
    - Microbiome association with disease progression
  - Many more…
- Utilizes the Phyloseq data format

Phyloseq data format

- Standard representation of microbiome data
- There are 3 components to a phyloseq object
  1. Abundance data (OTUs or ASVs)
     - OTU—Group of sequences with a predefined level of sequence similarity (97%)
       - Done to minimize the effects of sequencing errors
       - Results in the lose of real variation
       - Mostly outdated
       - Useful for combining data from different technologies or primers
     - ASV—Error correction applied to get exact sequence features
  2. Phylogenetic tree or taxonomic assignments for abundance data
  3. Sample meta information (age, treatment, disease, ethnicity, etc)

Phyloseq Object (Data container)

Abundance estimates

	Sample-1	Sample-2	Sample-3	Sample-4
Actinomycetaceae	0	1	0	1
Aerococcus	0	0	0	0
Aeromonas	0	0	0	0
Akkermansia	18	97	67	256
Alcaligenes faecalis et rel.	1	2	3	2
Allistipes et rel.	336	63	36	96
Anaerobiospirillum	0	0	0	0
Anaerofustis	0	1	0	0
Anaerostipes caccae et rel.	244	137	27	36
Anaerotruncus colihominis et rel.	12	108	203	68
Anaerovorax odorimutans et rel.	6	73	30	60
Aneurinibacillus	0	0	0	0
Aquabacterium	0	0	0	0
Asteroleplasma et rel.	0	0	0	0
Atopobium	0	1	0	1
Bacillus	1	2	1	1
Bacteroides fragilis et rel.	443	21	73	29
Bacteroides intestinalis et rel.	12	1	3	6

Phylogenetic tree or taxonomy

	Phylum	Family	Genus
Actinomycetaceae	Actinobacteria	Actinobacteria	Actinomycetaceae
Aerococcus	Firmicutes	Bacilli	Aerococcus
Aeromonas	Proteobacteria	Proteobacteria	Aeromonas
Akkermansia	Verrucomicrobia	Verrucomicrobia	Akkermansia
Alcaligenes faecalis et rel.	Proteobacteria	Proteobacteria	Alcaligenes faecalis et rel.
Allistipes et rel.	Bacteroidetes	Bacteroidetes	Allistipes et rel.
Anaerobiospirillum	Proteobacteria	Proteobacteria	Anaerobiospirillum
Anaerofustis	Firmicutes	Clostridium cluster XV	Anaerofustis
Anaerostipes caccae et rel.	Firmicutes	Clostridium cluster XIVa	Anaerostipes caccae et rel.
Anaerotruncus colihominis et rel.	Firmicutes	Clostridium cluster IV	Anaerotruncus colihominis et rel.
Anaerovorax odorimutans et rel.	Firmicutes	Clostridium cluster XI	Anaerovorax odorimutans et rel.
Aneurinibacillus	Firmicutes	Bacilli	Aneurinibacillus
Aquabacterium	Proteobacteria	Proteobacteria	Aquabacterium
Asteroleplasma et rel.	Firmicutes	Asteroleplasma	Asteroleplasma et rel.
Atopobium	Actinobacteria	Actinobacteria	Atopobium
Bacillus	Firmicutes	Bacilli	Bacillus
Bacteroides fragilis et rel.	Bacteroidetes	Bacteroidetes	Bacteroides fragilis et rel.
Bacteroides intestinalis et rel.	Bacteroidetes	Bacteroidetes	Bacteroides intestinalis et rel.
Bacteroides ovatus et rel.	Bacteroidetes	Bacteroidetes	Bacteroides ovatus et rel.

Meta information

	subject	sex	nationality	group	sample	timepoint	timepoint.within.group	bmi_group
Sample-1	byn	male	AAM	DI	Sample-1	4	1	obese
Sample-2	nms	male	AFR	HE	Sample-2	2	1	lean
Sample-3	olt	male	AFR	HE	Sample-3	2	1	overweight
Sample-4	pku	female	AFR	HE	Sample-4	2	1	obese
Sample-5	qjy	female	AFR	HE	Sample-5	2	1	overweight
Sample-6	riv	female	AFR	HE	Sample-6	2	1	obese
Sample-7	shj	female	AFR	HE	Sample-7	2	1	obese
Sample-8	tgx	male	AFR	HE	Sample-8	2	1	overweight
Sample-9	ufm	male	AFR	HE	Sample-9	2	1	lean
Sample-10	nms	male	AFR	HE	Sample-10	3	2	lean
Sample-11	olt	male	AFR	HE	Sample-11	3	2	overweight
Sample-12	pku	female	AFR	HE	Sample-12	3	2	obese
Sample-13	qjy	female	AFR	HE	Sample-13	3	2	overweight
Sample-14	riv	female	AFR	HE	Sample-14	3	2	obese
Sample-15	shj	female	AFR	HE	Sample-15	3	2	obese
Sample-16	tgx	male	AFR	HE	Sample-16	3	2	overweight
Sample-17	ufm	male	AFR	HE	Sample-17	3	2	lean
Sample-18	nms	male	AFR	DI	Sample-18	4	1	lean
Sample-19	olt	male	AFR	DI	Sample-19	4	1	overweight

Types of analysis performed

- Diversity
  - Is there a relationship between health status and microbiome diversity?
- Sample ordination
  - Do changes in health status, diet or drug exposure result in characteristic microbiomes?
- Differential microbe abundance
  - Are there differences in microbe abundance based on health status?
- Generalized linear model (GLM)
  - Is there an association between the microbiome and:
    - Gene expression
    - Metabolite concentration
    - Disease state
    - Phenotype

Properties of 16S microbiome data

1. Variation in amount of reads sequenced between samples
   - The more you sequence the more species you will find
   - Low abundance species not consistently detected
2. Sparse (Many zeros)
   - Low counts are unreliable
3. Counts do not reflect the absolute number of microbes present
   - Data is compositional (sum to 1)
   - Spurious correlations common
4. High-dimensional (many taxa)
   - Difficult to determine sample similarity
5. High variability
   - Makes normalization difficult

Weiss et al 2017
McMurdie and Holmes 2014

Diversity and sample ordination

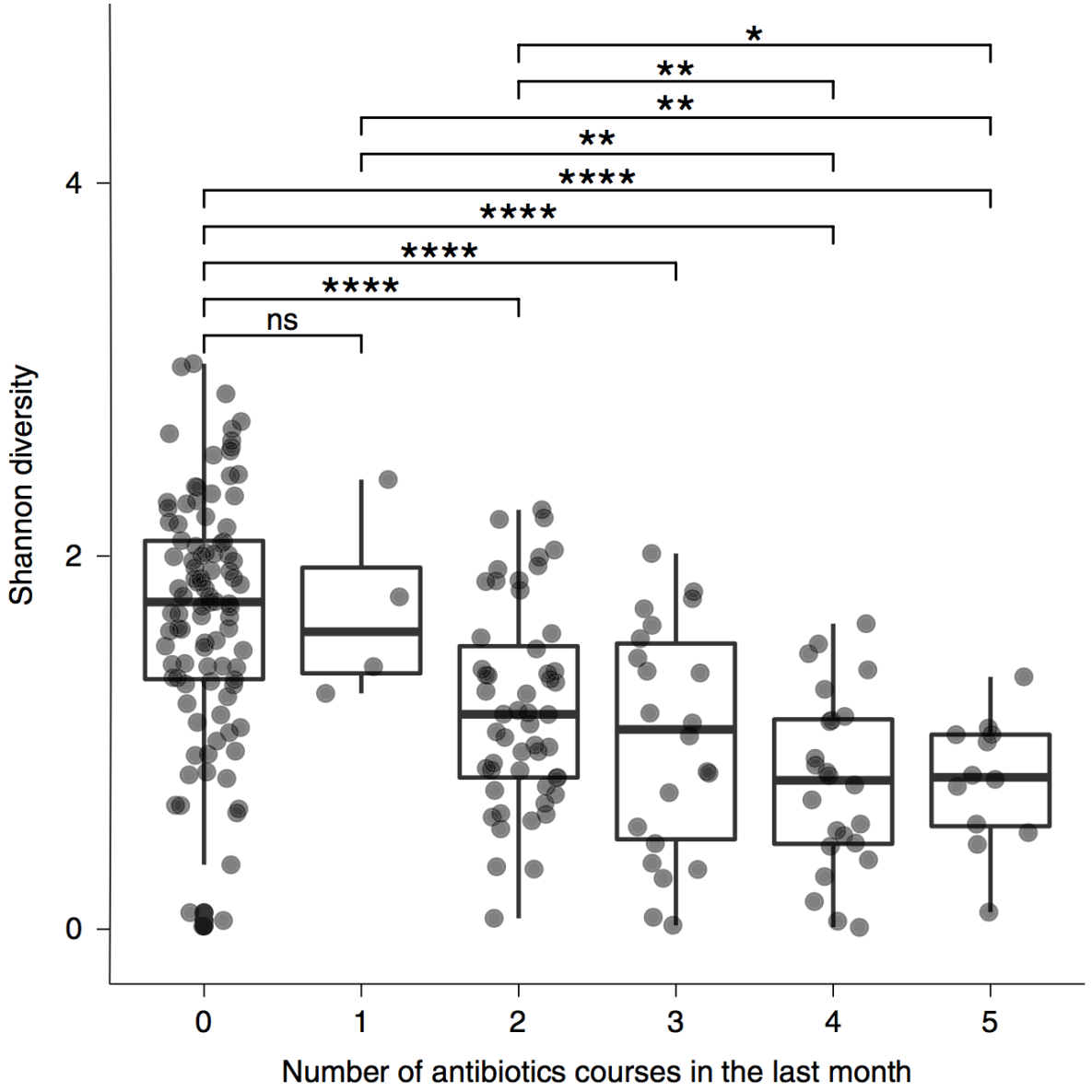
What is diversity?

- There are 2 components to diversity:
  1. How many taxa are present?
  2. How evenly distributed are they?
- Measures the unpredictability of the species identity of a randomly chosen individual
- There are many ways to measure diversity. None are perfect

Types of diversity

1. Alpha diversity: Within sample diversity
   - Can be compared between groups (i.e. before and after antibiotic treatment or before and after an oil spill)
   - Do not filter data before calculating alpha diversity!
     - Many diversity measures model the probability of low abundance species being shared between samples
   - Report Shannon index and a microbiome specific index (Chao1 or ACE)
2. Beta diversity: Between sample diversity
   - Distance measure used greatly influences results
     - Bray-Curtis most commonly used
   - Use all check ordination
   - Report the one that makes the most biological sense
   - Greatly influenced by differences in coverage

Antibiotic use decreases gut microbiome diversity in infants

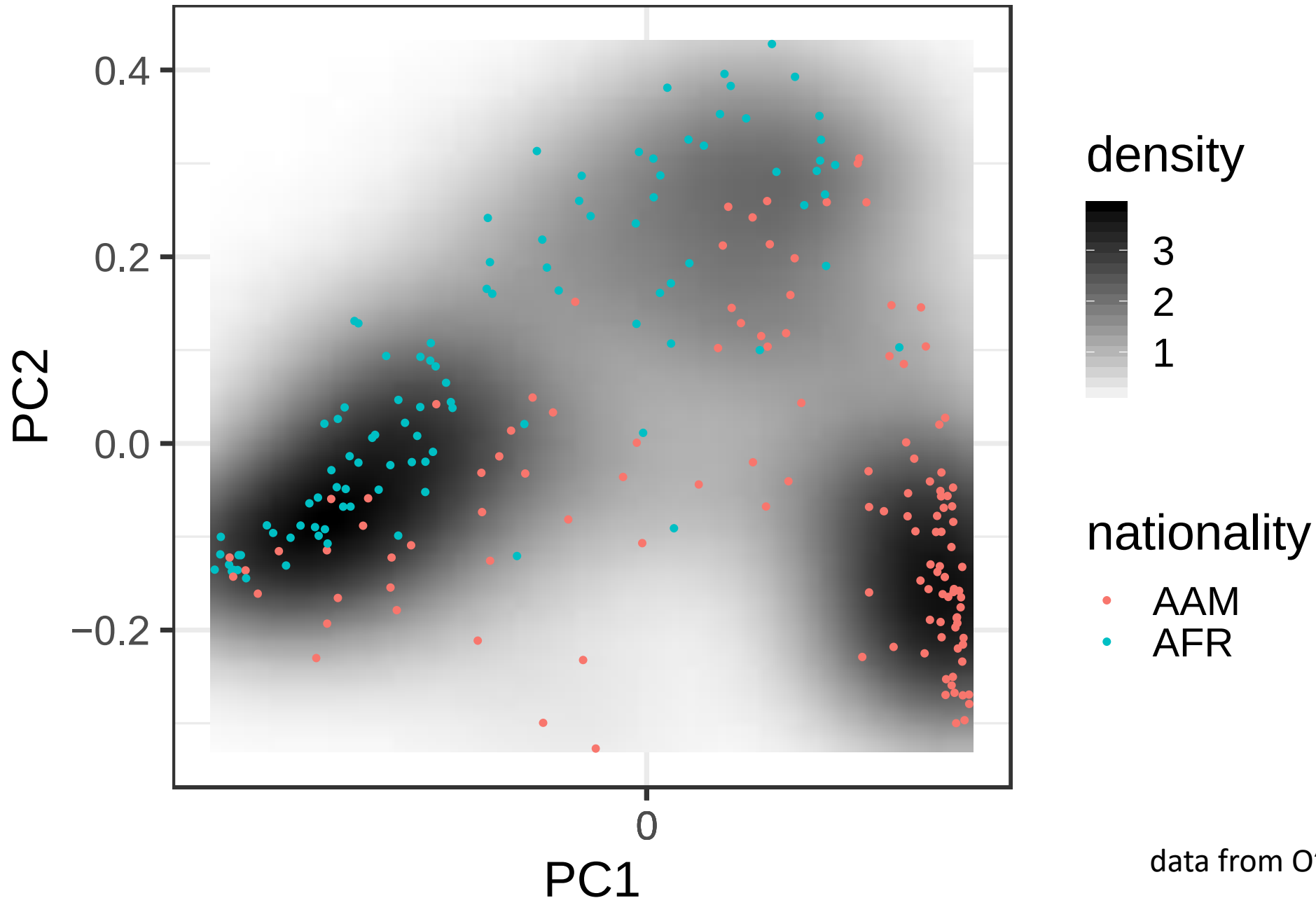


Gasparrini et al 2019

Sample Ordination with beta diversity

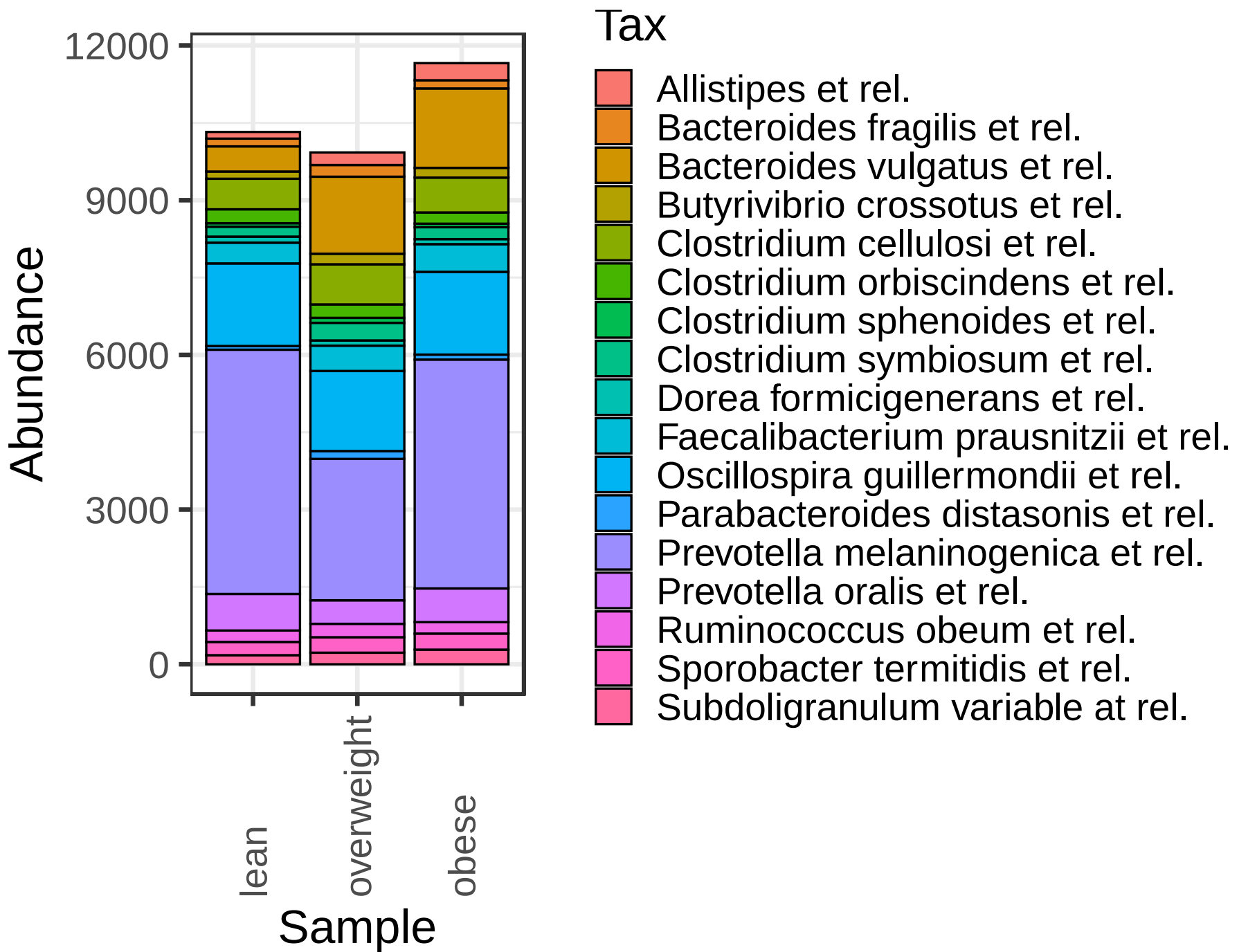
- Microbiome data has high dimensionality
  - There are many different taxa
- This makes it difficult to determine sample similarity
- Ordination techniques group similar samples together in a lower dimensional space (i.e. 2D or 3D)
  - PCA is a commonly used ordination technique, but does not work well with microbiome data
- Commonly used ordination methods in microbiome analysis
  - PCoA—Principal coordinates analysis
    - PCA modified to work with microbiome type data
    - Sometimes produces results that cannot be graphed
  - NMDS—Non-metric multidimensional scaling
    - Compares samples based on rank similarity
    - Can be used if PCoA does not work
- Measure statistical significance using PERMONVA

Gut microbiome samples cluster by nationality



data from O'Keefe et al 2015

Example: Composition bar plots



Testing statistical significance of community differences

- PERMONVA: Tests for differences in composition and/or relative abundances of different species in samples from different groups or treatments
  - Nonparametric
  - Input is dissimilarity matrix
    - i.e. Bray-Curtis dissimilarity measures between samples
  - P-values calculated from permutations of dissimilarity matrix

Anderson 2001

Rarefying data before sample ordination

- Rarefying is random subsampling without replacement
- Rarefying data is controversial
- Problems with rarefying
  - Inflates variance
  - Adds artificial uncertainty
- Not something you want to do before every analysis
- Benefit of rarefying
  - Only known method that corrects for library size bias for sample ordination

Weiss et al 2017
McMurdie and Holmes 2014

Differential microbe abundance

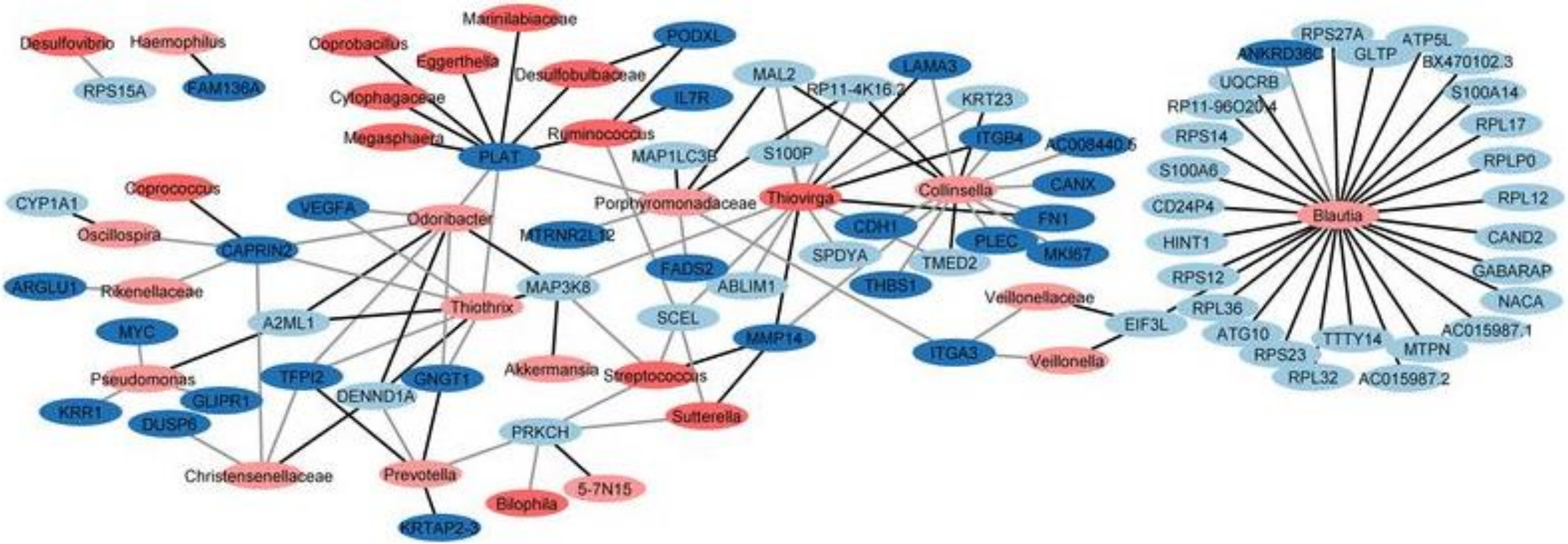
Differential microbe abundance using: metagenomeSeq

- Zero-inflated generalized linear model (GLM)
  - Corrects for sparsity and unequal sampling depth

Generalized Linear Model

- Generalized linear models that can be used when:
  - The range of the measurement is restricted (e.g. binary or count)
  - The variance of the measurement depends on the mean
- DESeq2: Differential analysis of count data
  - Developed for RNA-seq
  - Statistical model works with any type of count data
  - Can be used to:
    - Identify differences in microbe abundance between conditions
      - I would be wary of doing this since microbiome data does not meet test assumption of most features not differentially expressed
    - Test for association between microbe level and:
      - Gene expression
      - Metabolite concentration
      - Disease state
      - Other…

Abundance of microbiome taxa is associated with specific host gene expression changes.



Allison L. Richards et al. mSystems 2019;
doi:10.1128/mSystems.00323-18

Journals.ASM.org
This content may be subject to copyright and license restrictions.
Learn more at journals.asm.org/content/permissions
mSystems

Summary

- The microbiome is an important component of ecosystems and health
- 4 of the main analysis techniques used are:
  1. Diversity analysis
  2. Sample ordination
  3. Differential microbe abundance
  4. Generalized linear models
- The gene content of the microbiome is more important than who is there
  - PICRUSt: predict gene composition from taxa composition
  - Metagenomics: sequencing of microbiome genomes