Microbiome Data Analysis Workflow

Nolen Joy Perualila
CenStat/I-BioStat, Hasselt University, Belgium

January 16, 2017

UHASSELT
KNOWLEDGE IN ACTION

Overview UHasselt

- Data analysis workflow: Nolen.
- Research lines: Ziv.
- Specific case study: Rudradev.

Bioconductor Workflow

F1000Research

F1000Research 2016, 5:1492 Last updated: 25 DEC 2016

Check for updates

RESEARCH ARTICLE

REVISED Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; referees: 3 approved]

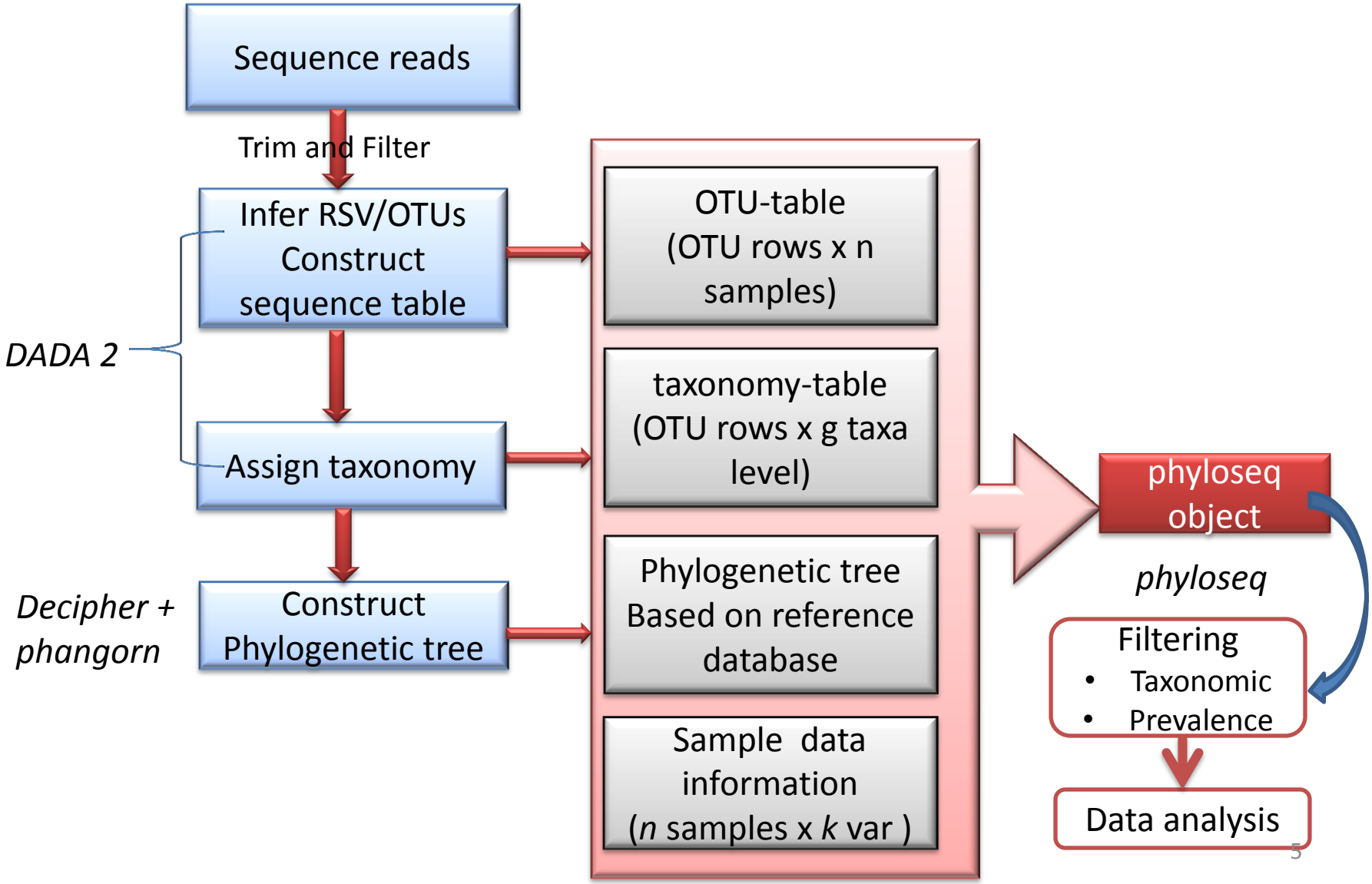
Ben J. Callahan[1], Kris Sankaran[1], Julia A. Fukuyama[1], Paul J. McMurdie[2], Susan P. Holmes[1]

[1]Statistics Department, Stanford University, Stanford, CA, 94305, USA
[2]Whole Biome Inc., San Francisco, CA, 94107, USA

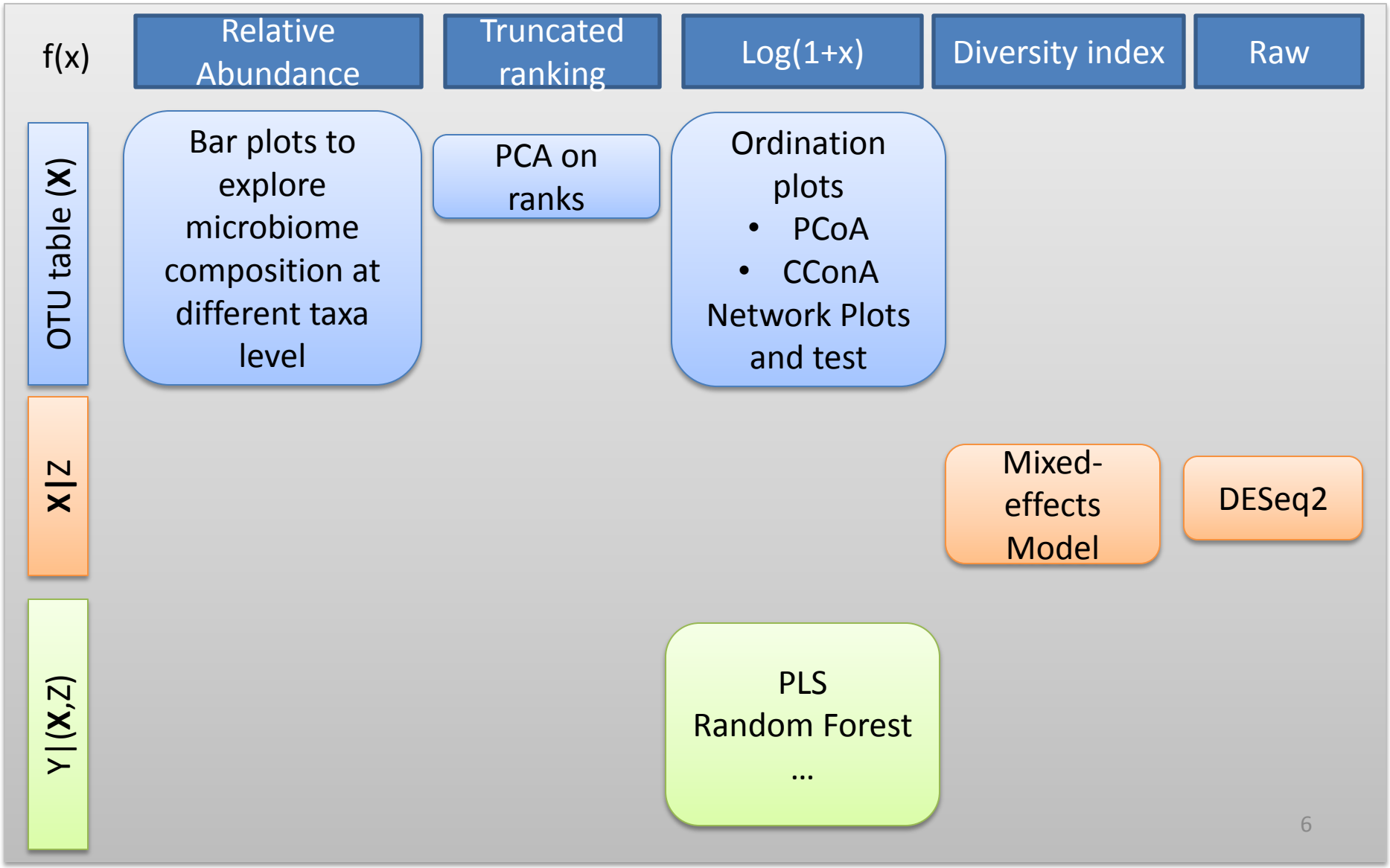
Goal

- To demonstrate the many features of an R/Bioconductor amplicon analysis workflow
  - denoising, filtering, data transformations, visualization, supervised learning analyses, community network tests, hierarchical testing and linear models

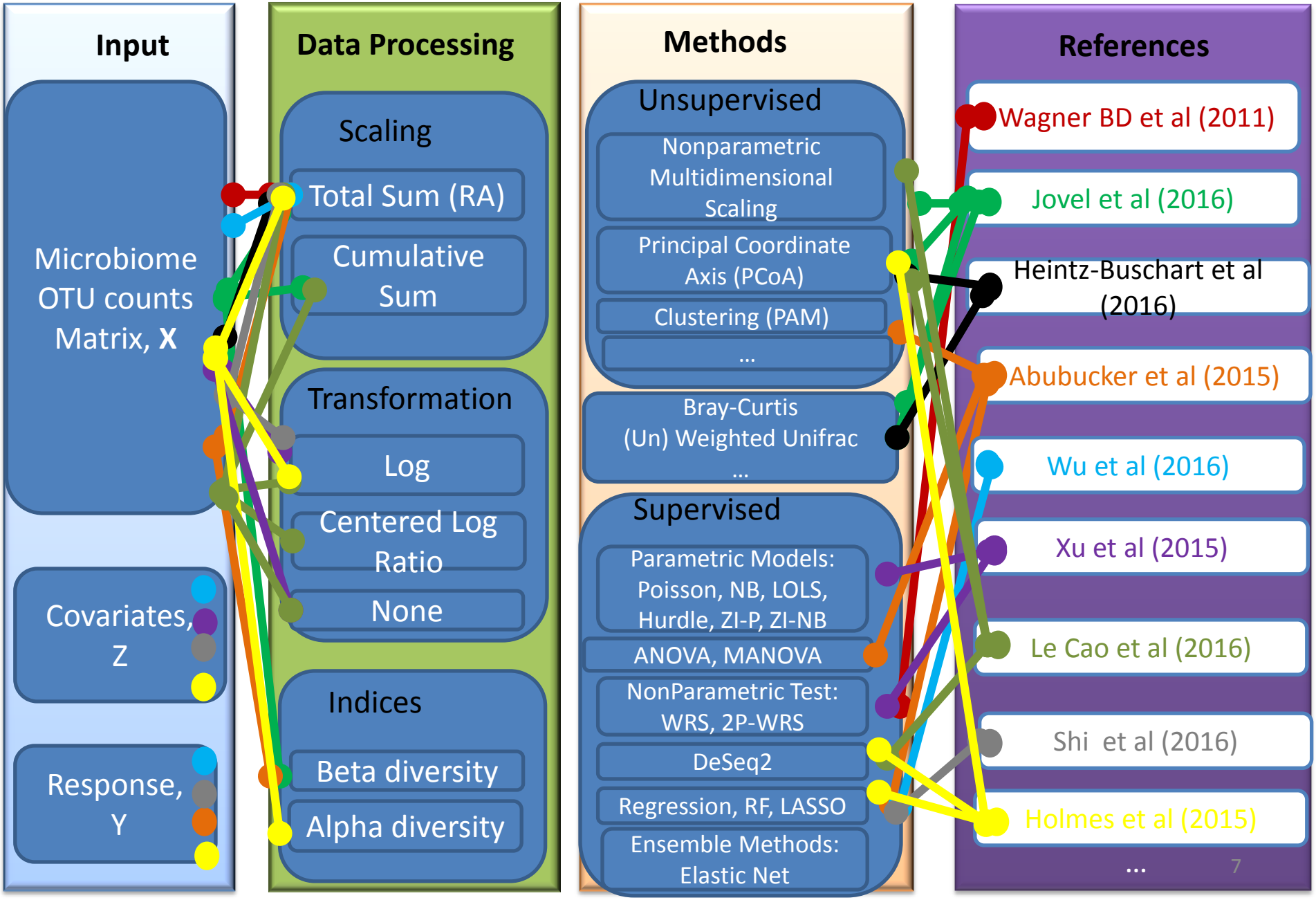
Bioinformatics: raw reads to tables



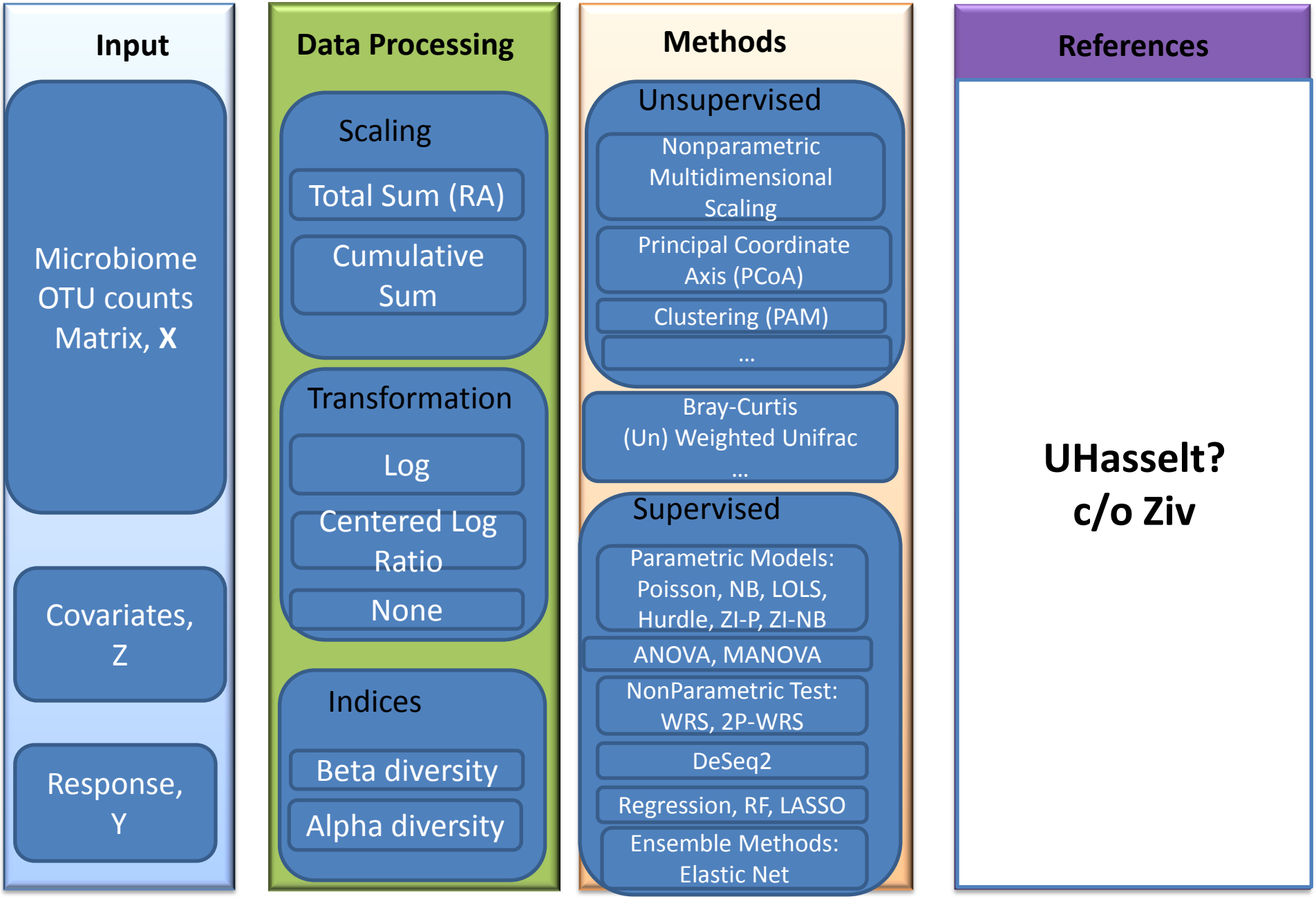
Data Analysis



Recent Works on Microbiome Data Analysis



Recent Works on Microbiome Data Analysis



Appendices

Supporting details

- Callahan et al 2016

Dataset used for illustration

- highly-overlapping Illumina Miseq 2×250 amplicon sequences from the V4 region of the 16S gene6
- 360 fecal samples from 12 mice longitudinally over the first year of life

Pre-processing

- DADA2
  - to infer ribosomal sequence variants (RSV) instead of clustering sequences into OUT
  - Construct OTU-table
  - implements the naive Bayesian classifier method to taxonomically classify the sequence variants (Assign taxonomy)
- Construct phylogenetic tree*(Phangorn R package)*

Pre-processing

- Combine data into a phyloseq object (phyloseq R pacage)
  - to import, store, analyze, and graphically display complex phylogenetic sequencing data that has already been clustered into Operational Taxonomic Units (OTUs) with sample data, phylogeny, and/or taxonomic assignment of each taxa.
  - Shiny-phyloseq for data exploration

Filtering of OTUs

- Filtering
  - Spurious taxa - taxa that were seen only rarely among samples
  - Taxonomic filtering- it is reasonable or even advisable to filter taxonomic features for which a high-rank taxonomy (e.g. phylum) could not be assigned

- Prevalence filtering- the number of samples in which a taxa appears at least once (for each feature)
  - (Supervised) Taxonomic filtering- Are there phyla that are comprised of mostly low-prevalence features?
  - (Unsupervised) plot taxa prevalence versus total counts to selecet filtering parameter (e.g. define a prevalence threshold in a range of zero to 10 percent or so)
    - Prune taxa if taxa prevalence is below threshold

Abundance value transformation

- Relative abundance
  - Corrects for library size
  - transform_sample_counts(ps3, function(x){x / sum(x)}))
  - Sample-wise
  - Explore composition

- Log-transform
  - an approximate variance stabilizing transformation
  - transform_sample_counts(ps, function(x) log(1 + x))
  - **analysis**

Unsupervised exploratory analysis

- To develop a representation of many bacteria with respect to sample characteristics
- PCoA with either the Bray-Curtis dissimilarity or the weighted Unifrac distance
  - Aspect ratio of ordination plots - normalize the axis norm ratios to the relevant eigenvalue ratios
  - To document outliers

Unsupervised exploratory analysis

- PCA on ranks
  - ignore the raw abundances altogether, and work instead with ranks
  - rank-transformed version of the data
    - the microbe with the smallest in a sample gets mapped to rank 1
  - truncated-ranking transformation
    - many bacteria are absent or present at trace amounts, an artificially large difference in rank could occur for minimally abundant taxa.
    - those microbes with rank below some threshold are set to be tied at 1 (plot abundance and rank)

- Canonical Correspondence Analysis
  - ps_ccpna <- ordinate(pslog, "CCA", formula = pslog ~ age_binned + family_relationship)

Supervised Learning

- Predict age group from microbiome composition
  - Partial Least Squares
    - biplot
  - Random forest
    - Proximity
      - PCoA+distance between samples based on tree-partition co-occurrence

Network-based visualization and testing

- Network + permutation test
  - H0: the two samples come from the same distribution
  - Network plot with Jaccard similarity (between samples) colored by litter (minimal spanning tree, nearest neighbor,..)
  - Jaccard, Bray-Curtis, …

Linear Modeling

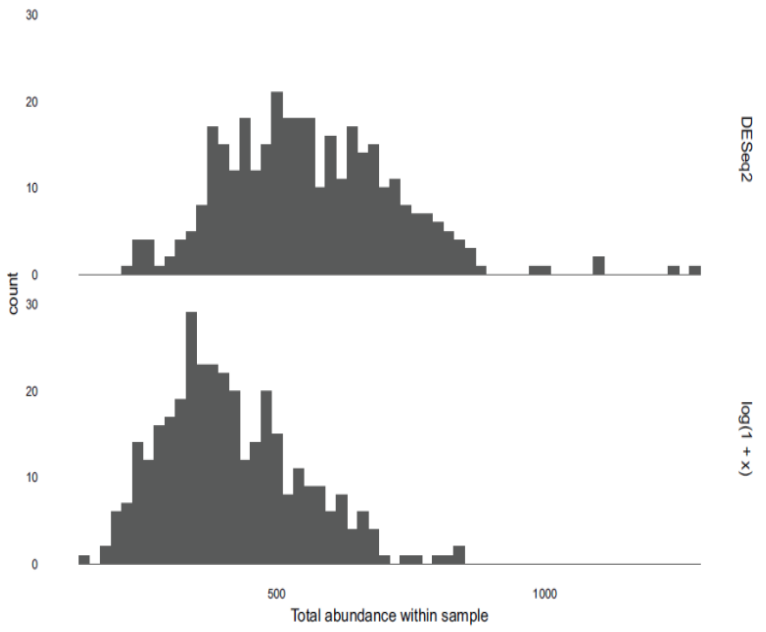
- To describe how a single measure of overall community structure (it need not be limited to diversity) is associated with sample characteristics

- A mixed-effects model to study the relationship between mouse microbial community diversity and the age and litter variables

```
model <- lme(fixed = alpha_diversity ~ age_binned, data
= ps_samp, random = ~ 1 | host_subject_id)
```

Hierarchical testing procedure

- Taxonomic groups are only tested if higher levels are found to be associated.
- Deseq2's variance stabilising transformation to test for differential abundance.
- The DESeq2 model internally corrects for library size, so transformed or normalized values such as counts scaled by library size should not be used as input.



Multitable techniques

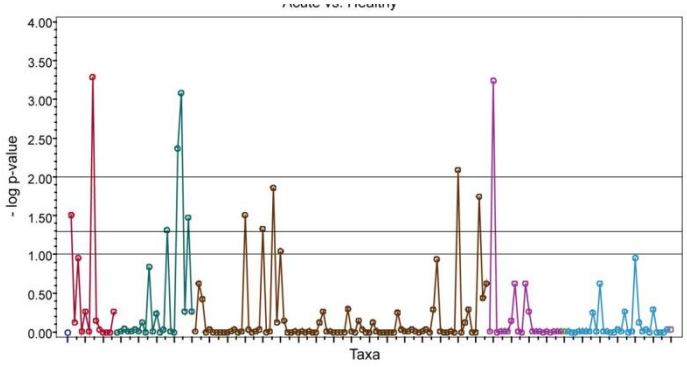
- Bacteria and another with metabolites.
- 12 samples were obtained, each with measurements at 637 m/z values and 20,609 OTUs.
- CCA+PCA - chooses a subset of available features that capture the most covariance and then PCA to this selected subset of features.

Supporting details

- Other papers

Application of Two-Part Statistics for Comparison of Sequence Variant Counts (Wagner BD, et al (2011))

- Data: CF sputum samples obtained during active disease (acute pulmonary exacerbation, n= 16) and sputum obtained from healthy controls (n= 10) by induction.
- Filtering :
  - taxonomic filtering
  - Prevalence thresholding: number of samples in which a taxa appears at least once
    - E.g. prevalence threshold=5% of total samples
- Transformation
  - Relative abundance, the percent of the total number of sequences obtained for each taxa within a sample
- Analysis
  - Two-part Wilcoxon
- Results
  - Manhattan plot, commonly used in genetic stud magnitude of the p-values for each comparison taxonomy line, and color-coded by phylum.



Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics (Jovel, 2016)

- Input: gut microbiome from 3 healthy and 2 non-healthy individuals
- Transformation: TSS and Upper quartile
- Analysis: beta-diversity comparison
  - phylogenetic β-diversity: that take into account the evolutionary differences between communities (Unifrac)
    - Unweighted UniFrac considers presence/absence of OTUs and therefore emphasizes rare species, while weighted also considers the abundance of OTUs.
  - taxon-based or non-phylogenetic methods
    - Bray Curtis dissimilarity
- NMDS and PCoA
- Pathway analysis: Using software like MEGAN5, each sequence can be directly mapped to KO representative sequences and the sum of KO counts that belongs to the same pathway can be computed.

Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes (Heintz-Buschart et al, 2016)

- Data: 4 families, healthy and T1DM members, microbiome,metagenome,..
- Filtering :
  - none
- Transformation
  - Scaling wrt to total sum (sum-normalized abundances)
- Analysis
  - Bray Curtis + MANOVA
  - Total Soerensen dissimilarity indices (betapart R package) on sum-normalized relative abundances.
  - Jensen–Shannon divergences (phyloseq),
  - Principal coordinate (ade4)
  - Multiple co-inertia analysis (omicade4)

Metabolic and metagenomic outcomes from earlylife pulsed antibiotic treatment (Abubucker2015)

- Data: Antibiotic treatment.
  - 4 antibiotic groups
  - 14 timed samples , a total of 338 samples.
- Transformation
  - The obtained phylogenetic tree and abundance tables were used to calculate unweighted and weighted UniFrac beta diversity indices.
  - The OTU absolute abundance table and UniFrac beta diversity matrices were extracted from the pipeline
- Analysis
  - PAM Clustering using square root of the Jensen–Shannon divergence distances
  - Random forest: age~f(microbiota)
  - The maturity index model was used to predict day of life based on microbiota composition.
  - The mouse age predicted by the model (microbiota age) was used to calculated microbial maturity and MAZ as described18, using the following formulae:
    - Microbial maturity (MM)=microbiota age-median microbiota age of control mice of similar age.
  - MAZ=MM/s.d. of microbiota age of control mice of similar age.
  - Significant differences in average MAZ for control and PAT mice at each time point were calculated with one-way ANOVA, followed by Fisher's least significant difference tests with false-discovery rate error correction.

Vertically transmitted faecal IgA levels determine extra-chromosomal phenotypic variation (Letter, 2015)

- Data: mice grouped by IgA level
- Filtering :
  - For culture/culture inoculate samples, a minimum relative abundance of 0.005 in at least one sample was set to filter out very rare operational taxonomic units before subsequent analysis.
- Transformation
  - Scaling wrt to total sum (relative abundance)
- Analysis
  - Statistical significance between two groups
    - unpaired Student's t-test if the data passed the D'Agostino–Pearson normality test
    - Mann–Whitney U-test
  - The relative abundance was used to identify discriminant biomarkers
  - After Kruskal–Wallis analysis (with an a value of 0.05) of all features, a linear discriminant analysis model was used to rank discriminant features by the effect size with which they differentiated classe s(IgA-high versus IgA-low samples)
  - Biomarkers are depicted as phylum.class.order.family.genus, to the level to which the taxonomic units were assignable by QIIME. 'IgA' indicates whether the biomarker was enriched in IgA-high samples ('High') or IgA-low samples ('Low').
  - Means and standard deviations of assigned order.family.genus operational taxonomic units are depicted. Each taxon was compared between groups by one-way ANOVA followed by Tukey's multiple comparisons test (P,0.5).

An adaptive association test for microbiome data (Wu2016)

- Data: Y= outcome of interest, **Z**=OTU matrix**, X**=covariates
  - IBD disease status(Y) and the overall Gut microbiome composition (Z)
    - 40 twin pairs who were concordant or discordant for CD or UC were collected and the compositions of microbial communities in feces samples
  - Throat microbiome data set (Z) for smoking effects (Y):
    - excluded the samples with fewer than 500 reads and OTUs with only one read, leading to 60 samples remaining and 856 OTUs.
    - Gender (X) and antibiotic use (X) within the last 3 months were collected.
- **Goal:** to test for a possible association between the overall microbial community composition and the outcome of interest after adjusting for the covariates.
- Transformation
  - Scaling wrt to total sum (relative abundance)
- Analysis
  - Compute for weighted and unweighted generalized taxon proportions, Qik for the ith sample, kth taxon based on phylogenetic tree
  - Model: logistic or linear regression: g(Y)= $\beta_0 + \beta' X_i + \sum_{k=1}^{m} Q_{ik}\varphi_k,$
- H_0: There is no association between any taxa and the outcome of interest under H_0, *i.e.,* $(\varphi_1, \ldots, \varphi_m)' = 0$
  - Test Stat: see next slide
- Results: rank the importance of taxa

An adaptive association test for microbiome data (Wu2016)

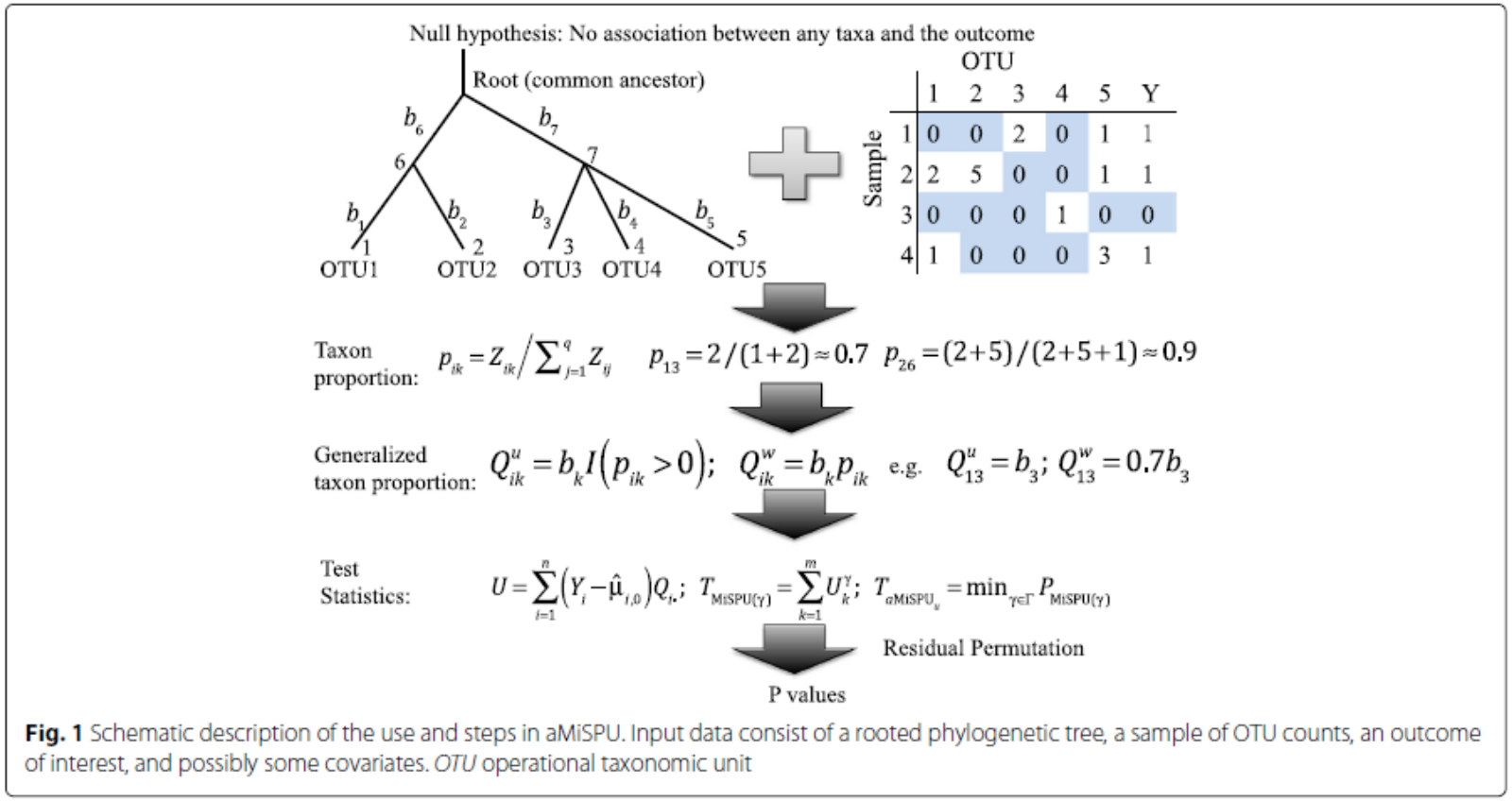


Fig. 1 Schematic description of the use and steps in aMiSPU. Input data consist of a rooted phylogenetic tree, a sample of OTU counts, an outcome of interest, and possibly some covariates. *OTU* operational taxonomic unit

Gamma parameter: to weight the taxa differentially.

Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data (Xu,2015)

- Data: Genetic Environmental Microbial (GEM)
  - Gender (M=204, F=262) and age as covariates
  - 3 selected taxa with different proportions of zeroes
  - Y= counts (dependent)
- Transformation
  - None
- Analysis
  - Competing models:
    - One Part: Poisson model, NB model, ordinary least squares on logarithmic transformed data (LOLS), and the non-parametric WRS test
    - Zero-Inflated Models: ZI-Poisson, ZI-Negative Binomial
    - Two-Parts: Poisson-Hurdle, Neg. Bin. Hurdle, 2P-LOLS, 2P-WRS
  - Model Selection: Vhoung Test, AIC, BIC
- Results
  - Selected model per taxa
  - Simulation results on power, type 1 error

mixMC: a multivariate statistical framework to gain insight into Microbial Communities (Le Cao 2016)

- Data: HMP case studies: oral
- Transformation
  - Aitchison (1982) - centered log ratio (CLR). CLR consists in dividing each sample by the geometric mean of its values and taking the logarithm. Standard univariate and multivariate methods can then be applied on the CLR data Mandal et al. (2015); Kalivodová et al. (2015).
  - Centered Log Ratio transformation (CLR):
  - Here, CLR on TSS data
  - Log counts then Cumulative sum scaling (CSS)

$$\boldsymbol{y} = (y_1, \ldots, y_p)' = (log\frac{x_1}{\sqrt[p]{\prod_{i=1}^{p} x_i}}, \ldots, log\frac{x_p}{\sqrt[p]{\prod_{i=1}^{p} x_i}})'.$$

- Analysis
  - Unsupervised: PCoA with no normalisation, PCA on ILR transformed data (*robCompositions* R package)
  - sPLS-DA for microbiome data using either CSS normalised data, or CLR transformed TSS data
  - Univariate + multiplicity adjustment : DESeq2 (normalisation in DESeq2 does not address the issue of compositional data) and Zero-Inflated Gaussian – metagenomeSeq R package (OTU counts are first log transformed and then CSS normalised)

- Results
  - Multivariate plots
  - Differentially abundant OTUs

Regression Analysis for Microbiome Compositional Data (Shi 2016)

- Data: Y=BMI, X= log-transformed compositions of the 45 genera as the covariates.
  - total fat intake and total caloric intake were also included as the covariates in the model.
  - The goal of this analysis is to identify the bacteria genera that are associated with bmi.
- Transformation
  - Log(Relative abundance)

$$E(\text{BMI}) = \sum_{g=1}^{45} \beta_g \log(X_g) + \gamma_1 \text{Fat} + \gamma_2 \text{Calorie},$$

where $\sum_{g=1}^{45} \beta_g = 0$, and $\log(X_g)$ is the logarithm of the relative abundance of the gth genus.

- Analysis
  - LASSO
    - By genus
    - By phylum

$$E(\text{BMI}) = \sum_{g=1}^{4} \sum_{s=1}^{m_g} \beta_{gs} \log(X_{gs}) + \gamma_1 \text{Fat} + \gamma_2 \text{Calorie},$$

where $\sum_{s=1}^{m_g} \beta_{gs} = 0$ for $g = 1, \cdots, 4$, and $\log(X_{gs})$ is the logarithm of the relative abundance of the sth genus of the gth phylum.

- Results

A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses (Shanka 2015)

- Data: Mouse
  - Y=colonization level
  - Xg = log relative abundance gth genus
  - Xc= log normalized mRNA expression levels of cytokines
  - Xa = indicator of antibiotic treatment group
- Transformation
  - Log(RA)
- Analysis
  - Elastic net with cross-validation (ENC)
  - Ensembles based on subsamples (PS, LS, SS, SSW)
  - Ensembles based on resamples (PR, LR, SR, SRW)
  - *Bayesian ensembles* $Y = X_g\beta_g + X_c\beta_c + X_a\beta_a + \epsilon,$
- Results
  - Rank of OTUs by method

Bioinformatics: raw reads to tables