Statistical methods for microbiome data

Ni Zhao, Assistant Professor of Biostatistics
Johns Hopkins University

3rd Term, 2020-2021

General Information

- Classroom: N. Wolfe building. W3031
- Lecture: Friday 10:30AM -11:50AM (EST)
- Instructor: Ni Zhao (nzhao10@jhu.edu)
- Assessment: critique on a recent preprint paper

In-person environment

- Face covering are required for all students & faculties during in-door environment
- *If you feel sick, please stay at home and get tested*
  - Even if you feel it is just a "common cold".
  - Zoom recording will be provided.
- In rare cases, virtual classes may be scheduled. Pay attention to class emails.

Course content

Chapter 1. Introduction

- sequencing procedures
- data structure

Chapter 2. Ecological Methods

- alpha diversities, beta diversities.
- PERMANOVA, MiRKAT

Chapter 3. Compositional models

- Aitchison distance
- Log-contrast models

Chapter 4. Bias in microbiome data and how to address it.

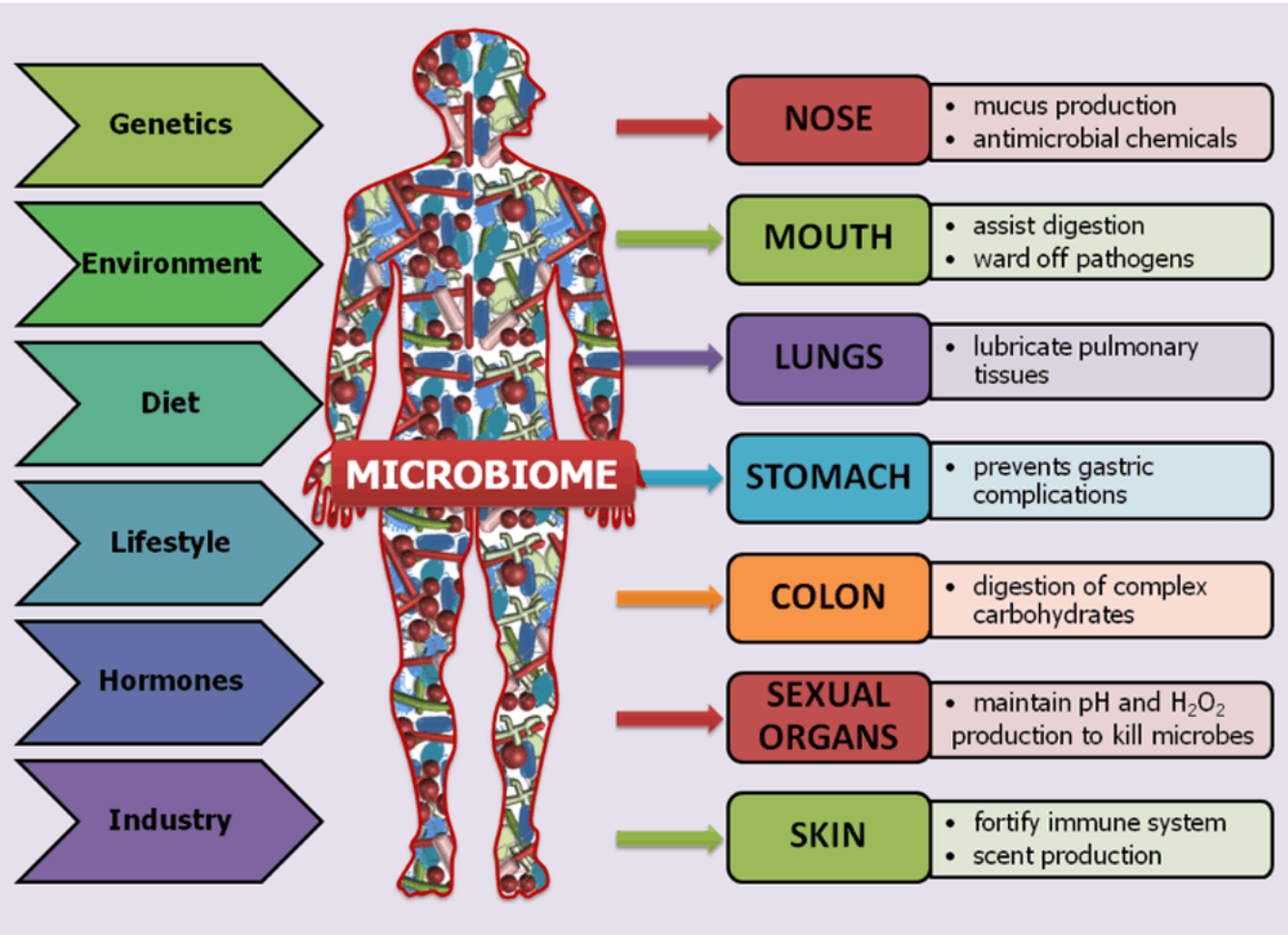
More about the course

- Material available at my github site?

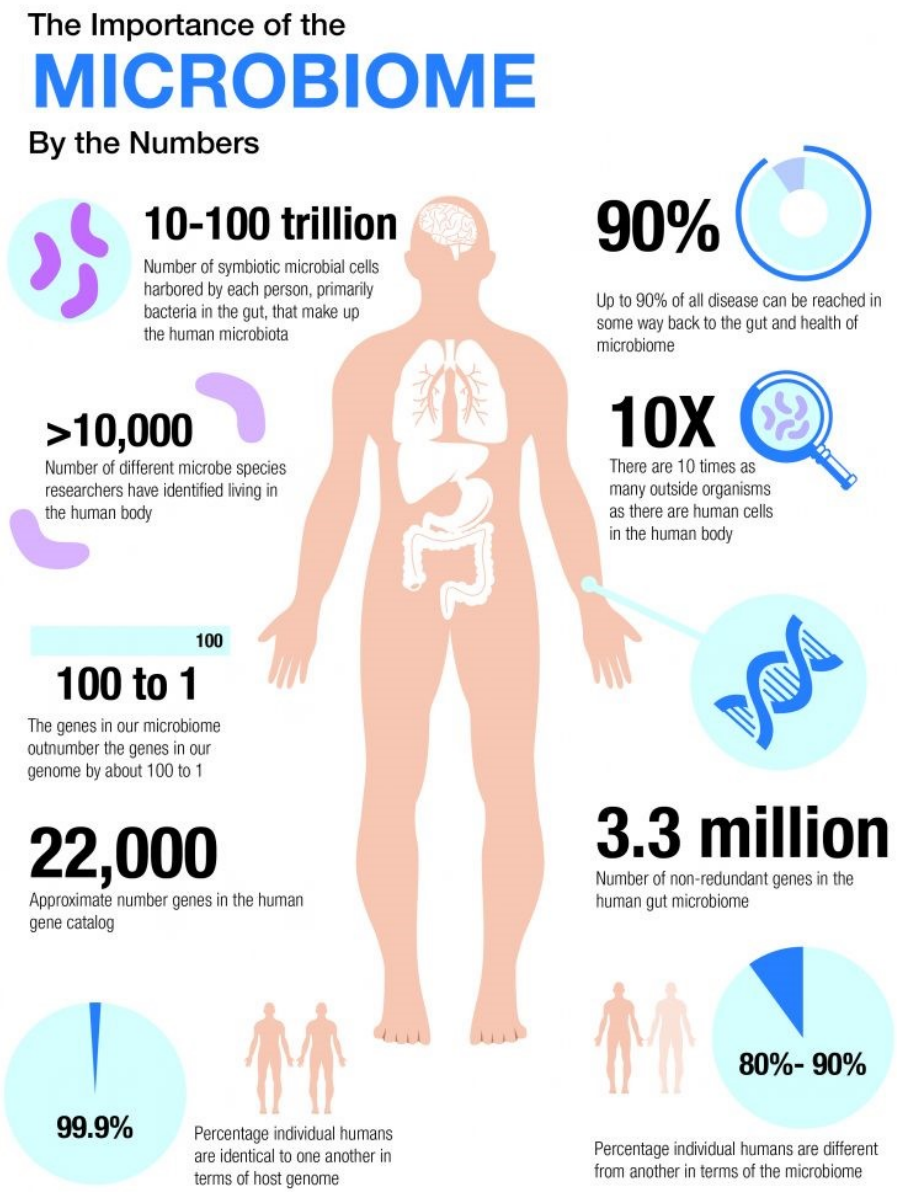
Chapter 1: Introduction

What is microbiome?

The collection of all microorganisms that co-exist with human being.



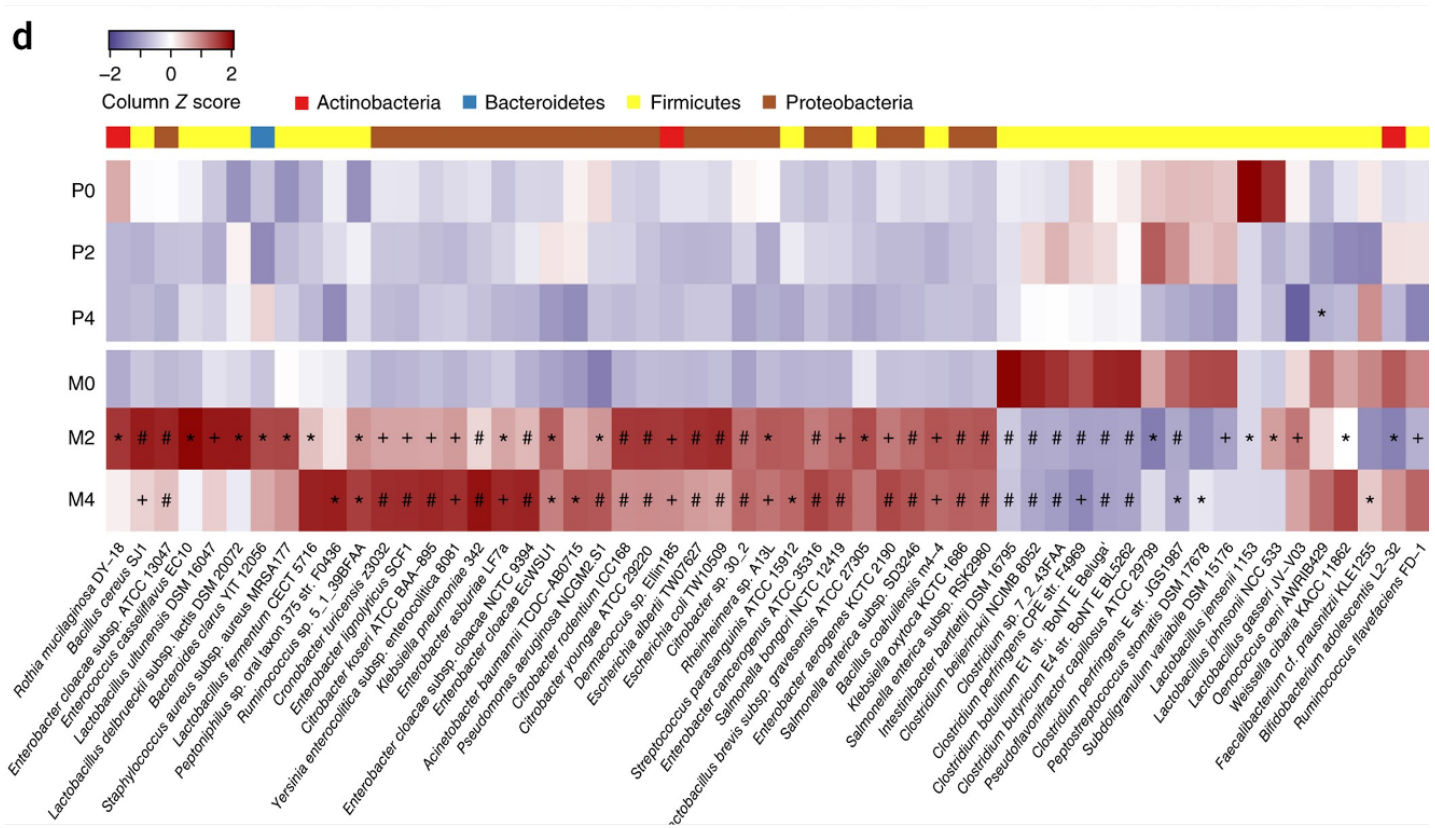
https://theconversation.com/the-human-microbiome-is-a-treasure-trove-waiting-to-be-unlocked-118757



https://www.sciencerepository.org/the-invisible-us-the-human-microbiome-in-health-and-disease

Example 1: Diabetes and Metformin treatment

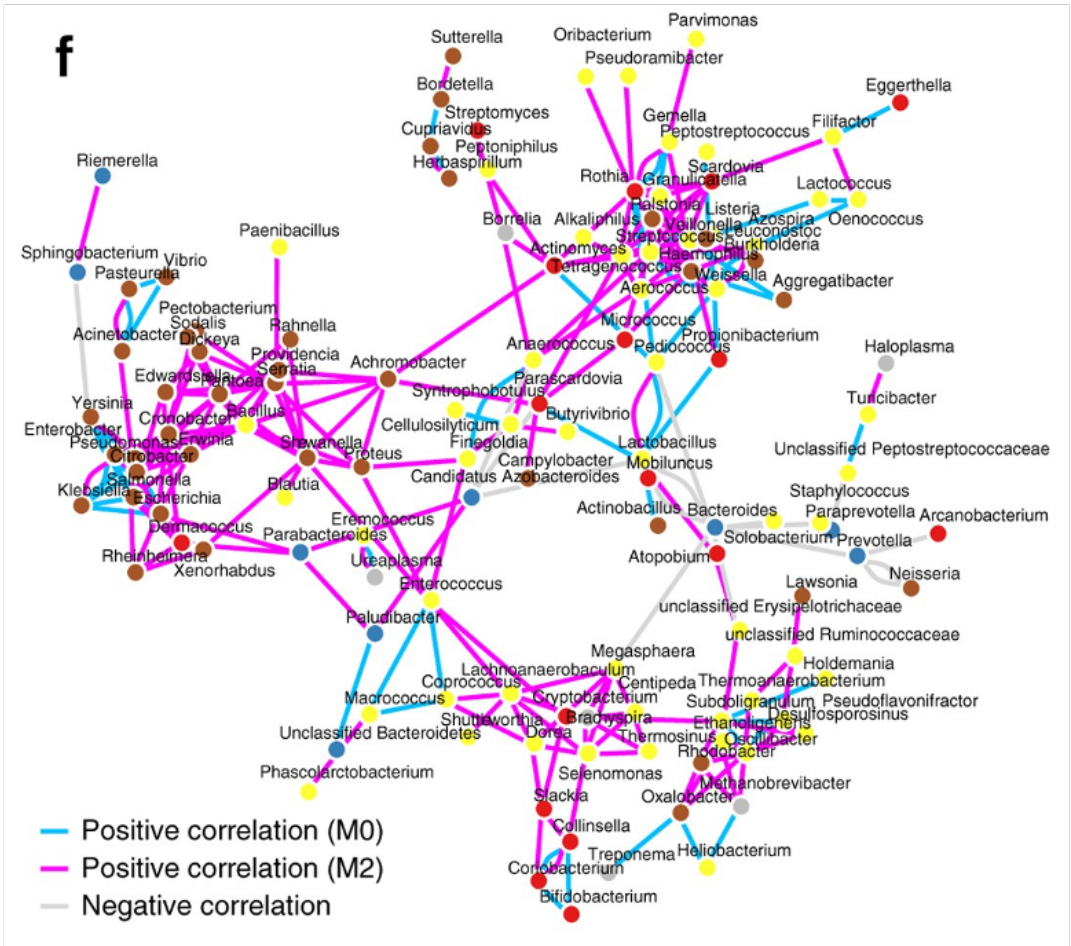
Metformin treatment promotes rapid changes in the composition of the gut microbiota, contributing to the therapeutic effects of the drug.



Hao Wu et al., *Nature Medicine*, 2017

Example 1: Diabetes and Metformin treatment

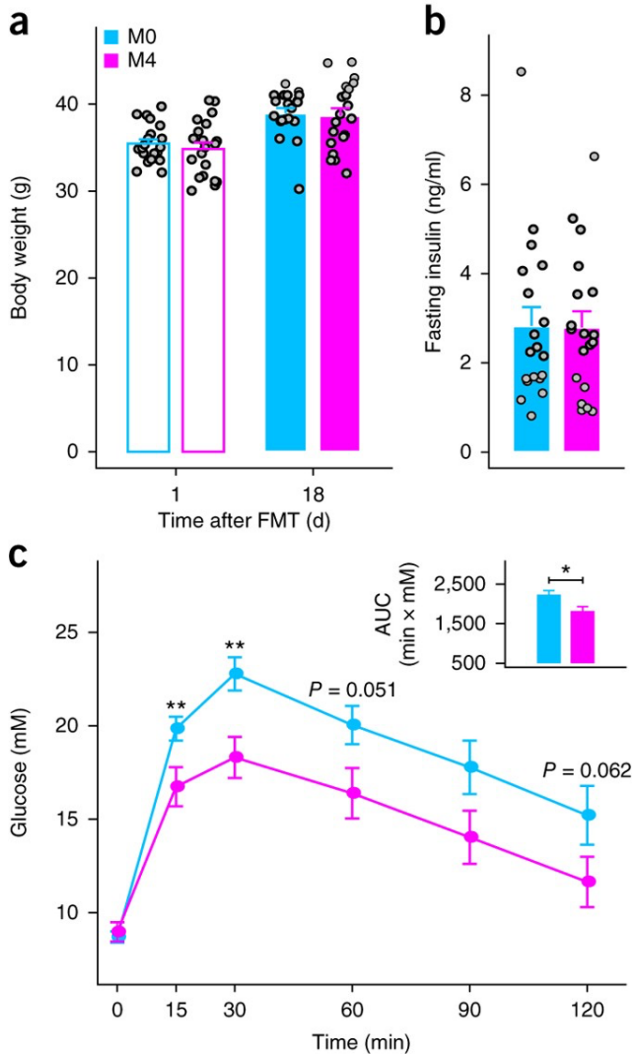
Metformin treatment promotes rapid changes in the structure of the gut microbiota.



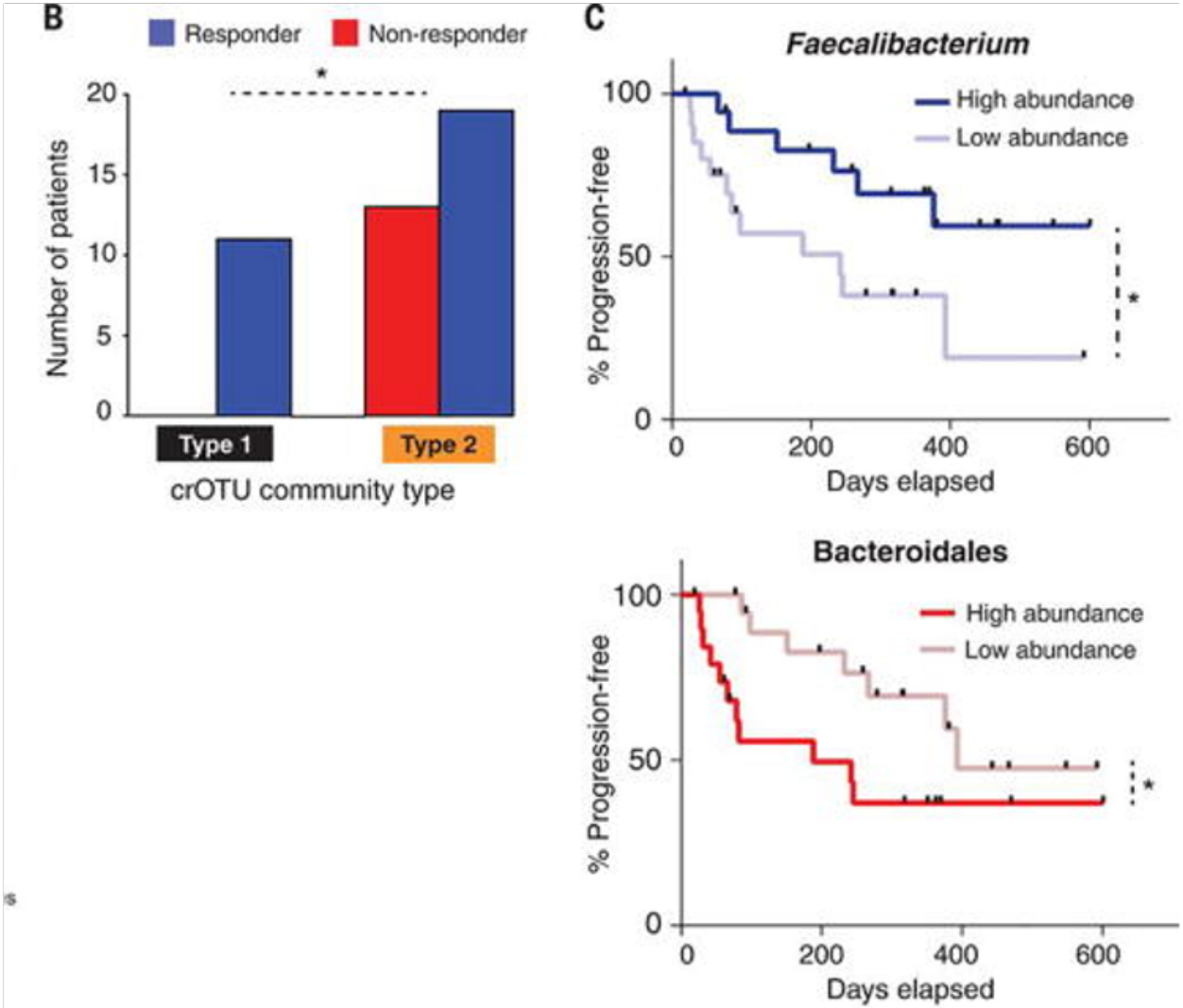
Hao Wu et al., *Nature Medicine*, 2017

Example 1: Diabetes and Metformin treatment

Metformin-altered microbiota improves glucose tolerance.

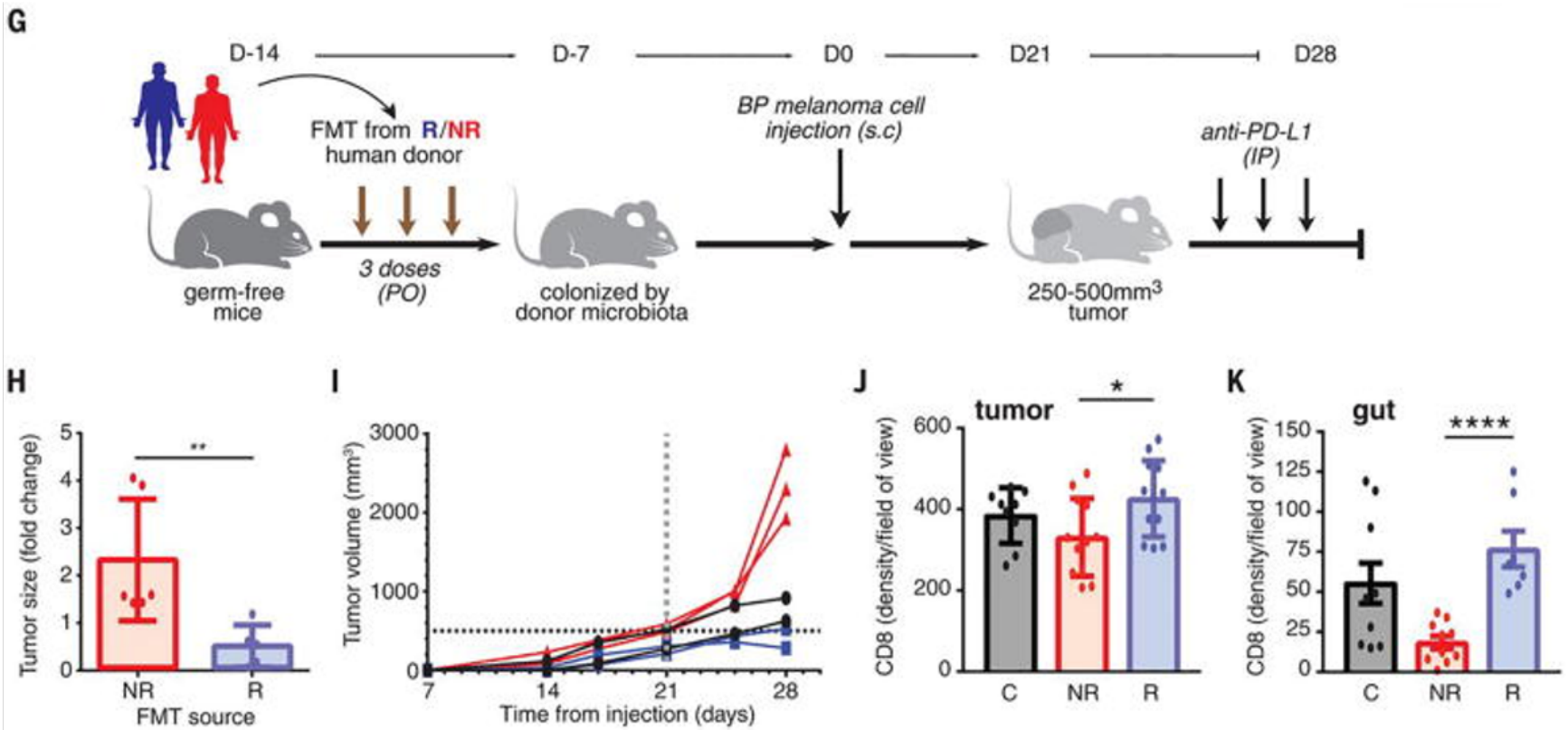


Example 2: Immonotherapy and gut microbiome

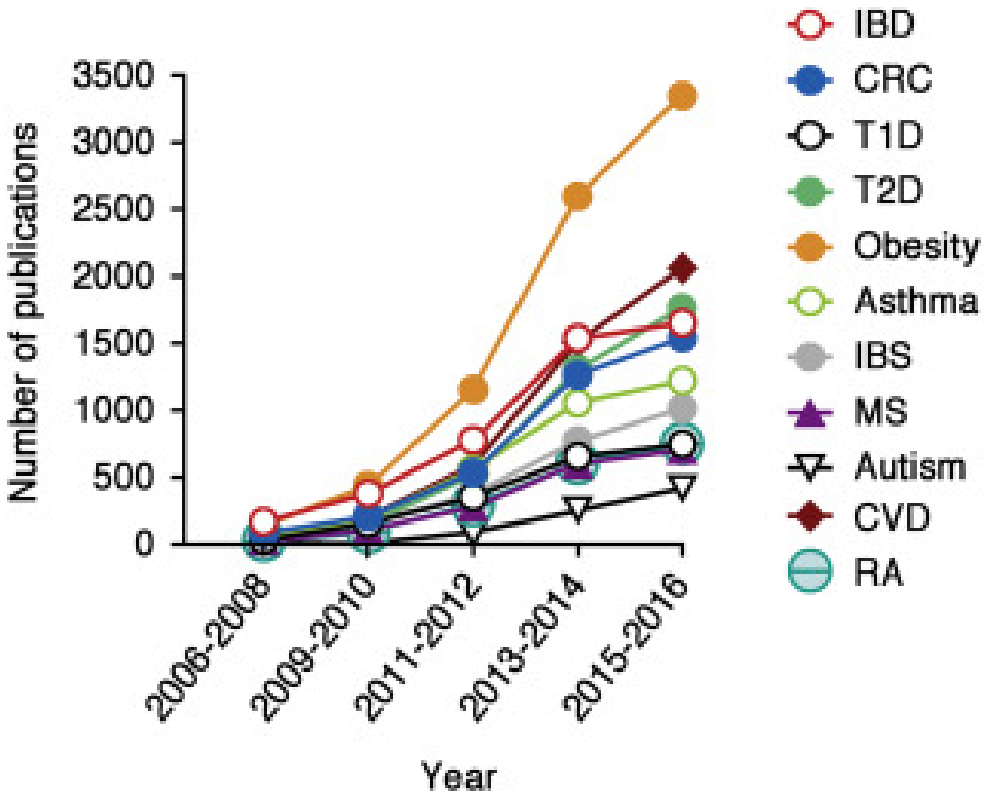


Gopalakrishnan et al., *Science*, 2018

Example 2: Immunotherapy and gut microbiome

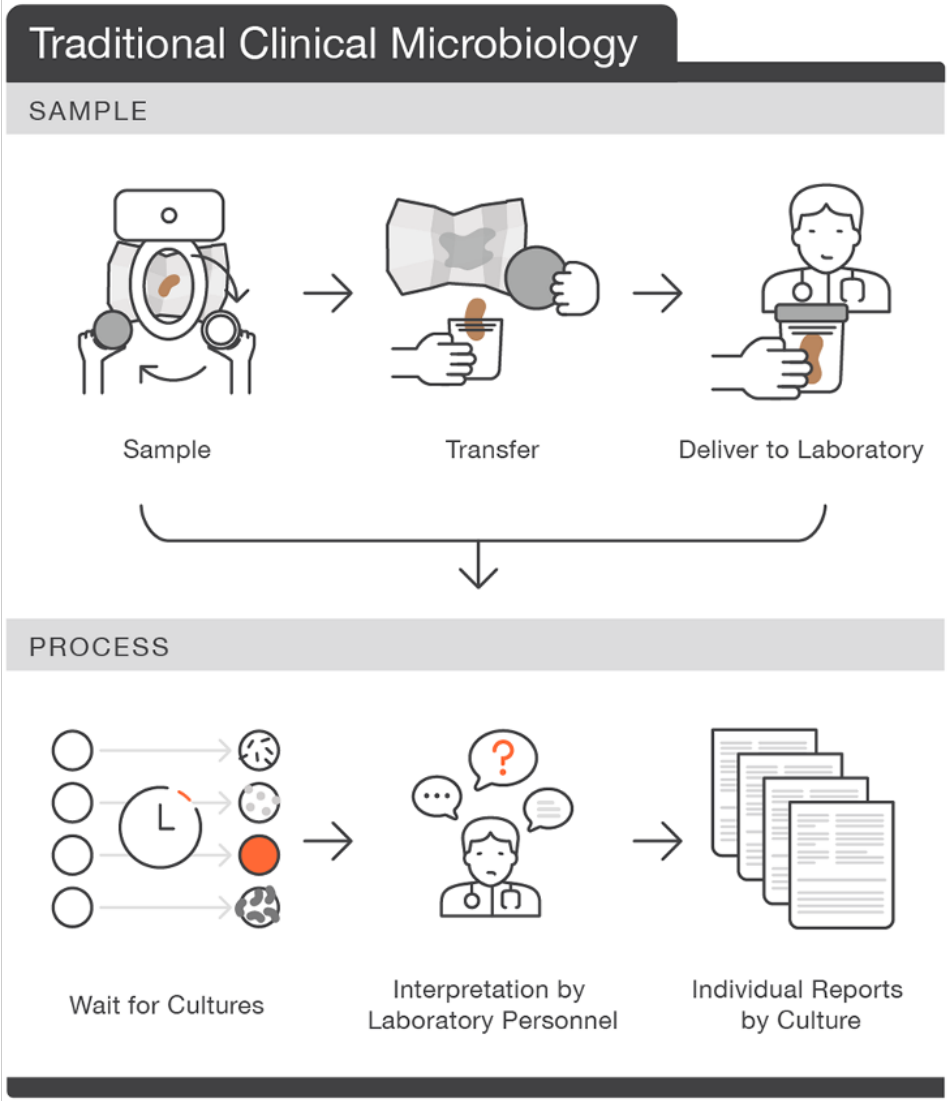


Gopalakrishnan et al., *Science*, 2018



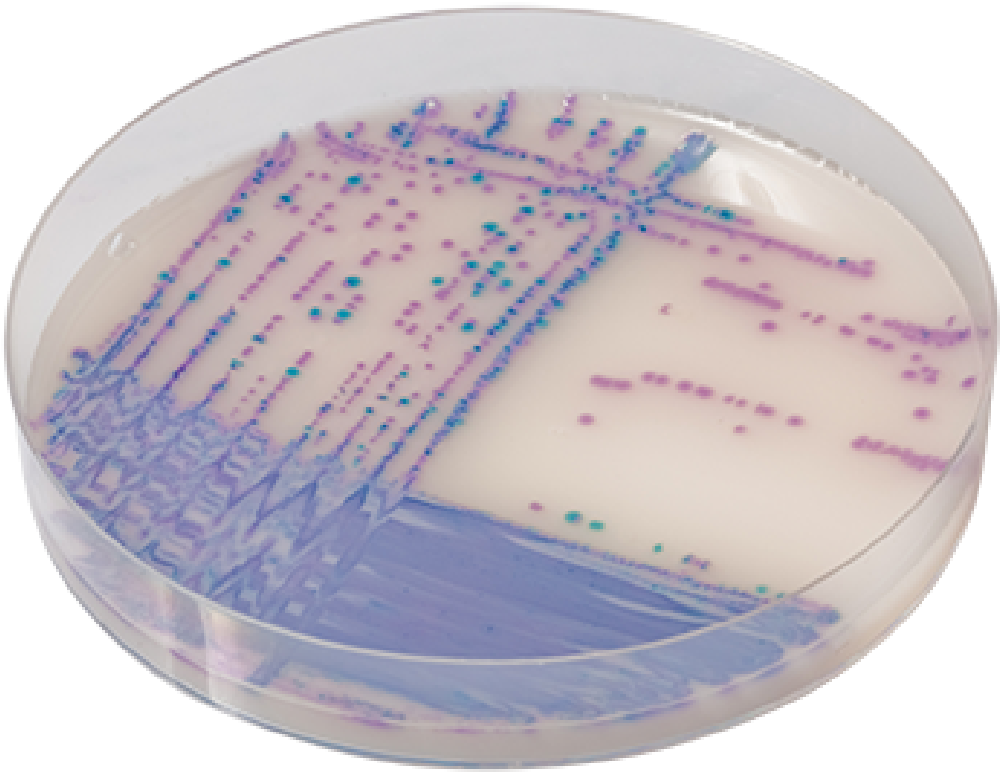
Butto et al., *Journal of Allergy and Clinical Immunology*, 2017

Traditional Microbiology



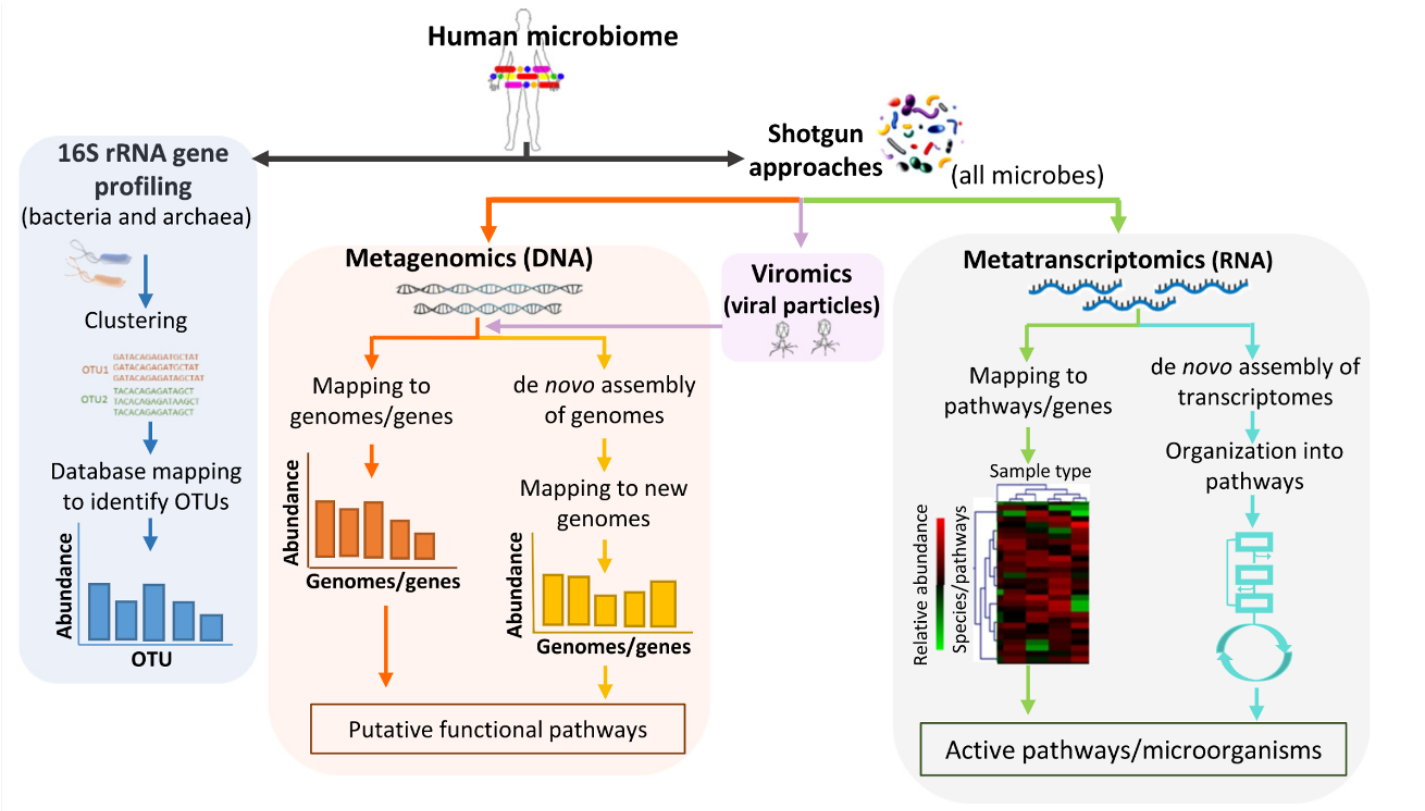
Daniel E. Almonacid et al., *PLos-One*, 2017

Cell Culture



https://asm.org/Articles/2020/September/How-CHROMagar-TM-Revolutionized-Bacterial-Identifi

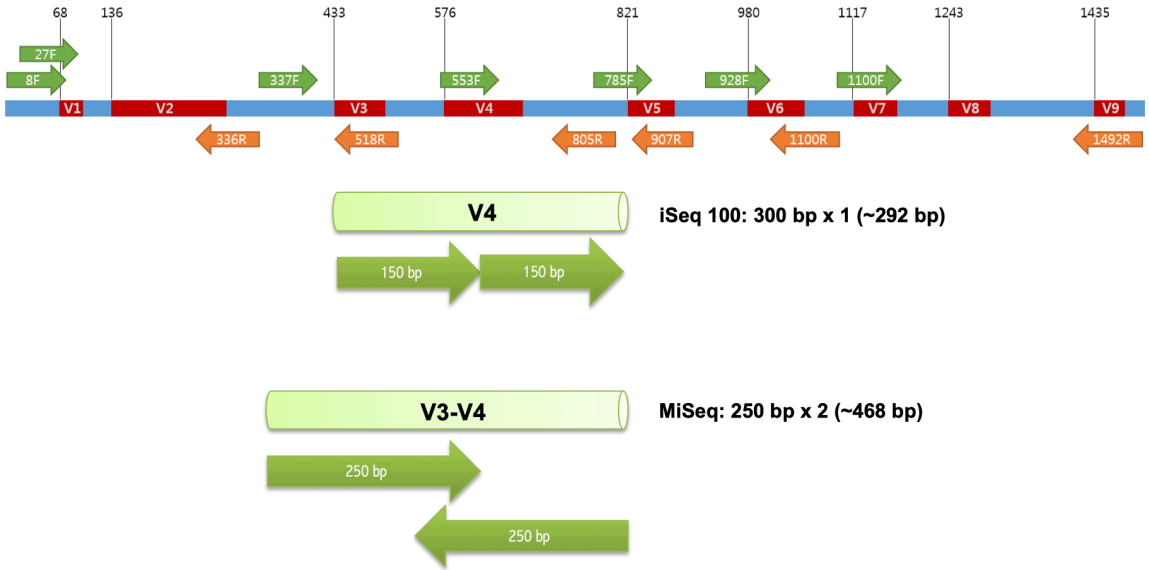
Next-generation sequencing for microbiome assessment



Bikel et al., *Computational and Structural Biotechnology Journal*, 2015

Marker gene sequencing, a.k.a., amplicon sequencing

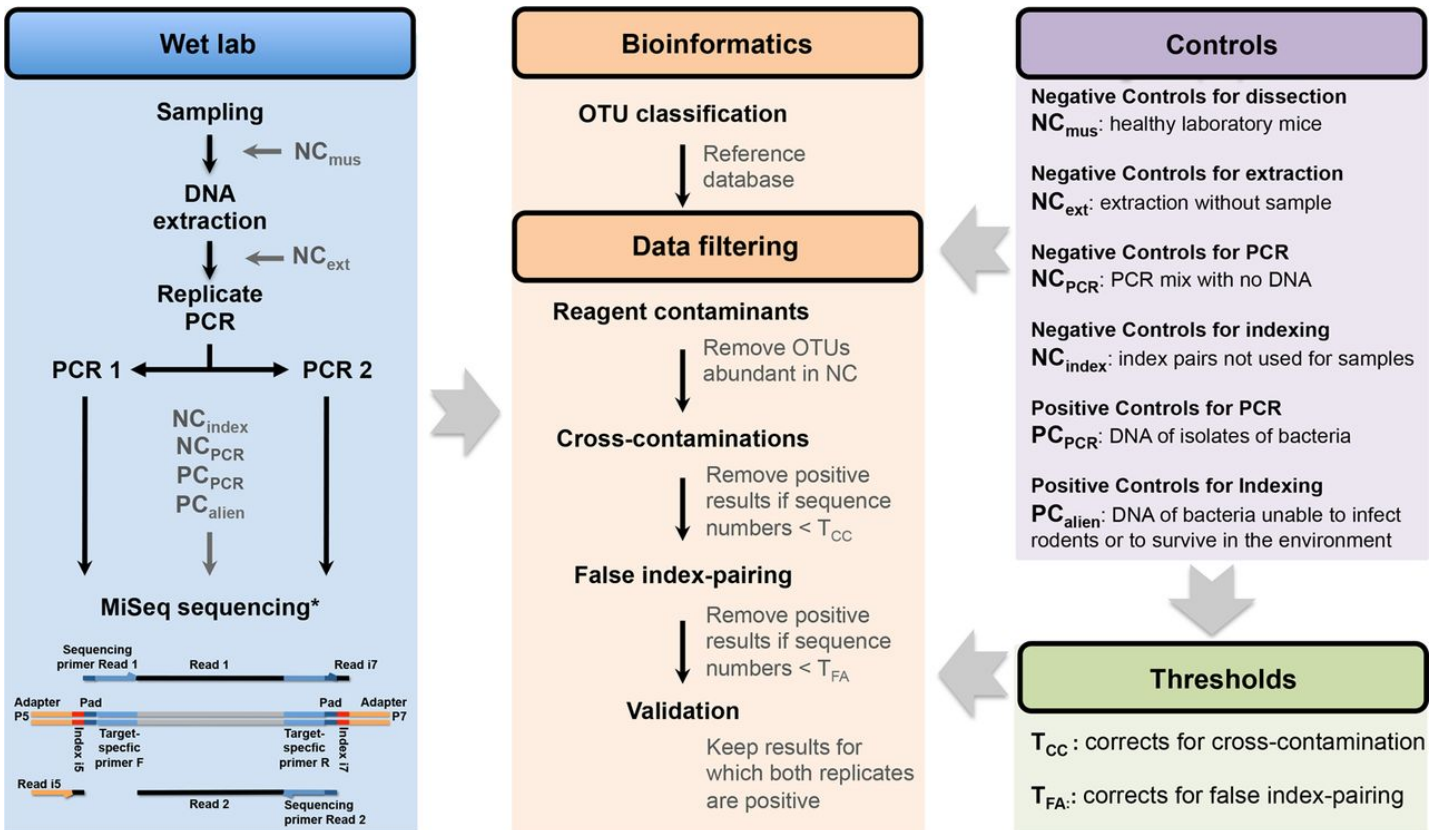
Most commonly used marker gene: 16s rRNA gene



https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/

Other marker genes include: 18S rRNA, ITS (internal transcribed spacer) for fungi identification

Marker gene sequencing, a.k.a., amplicon sequencing



Galan, et al., *mSystems*, 2016

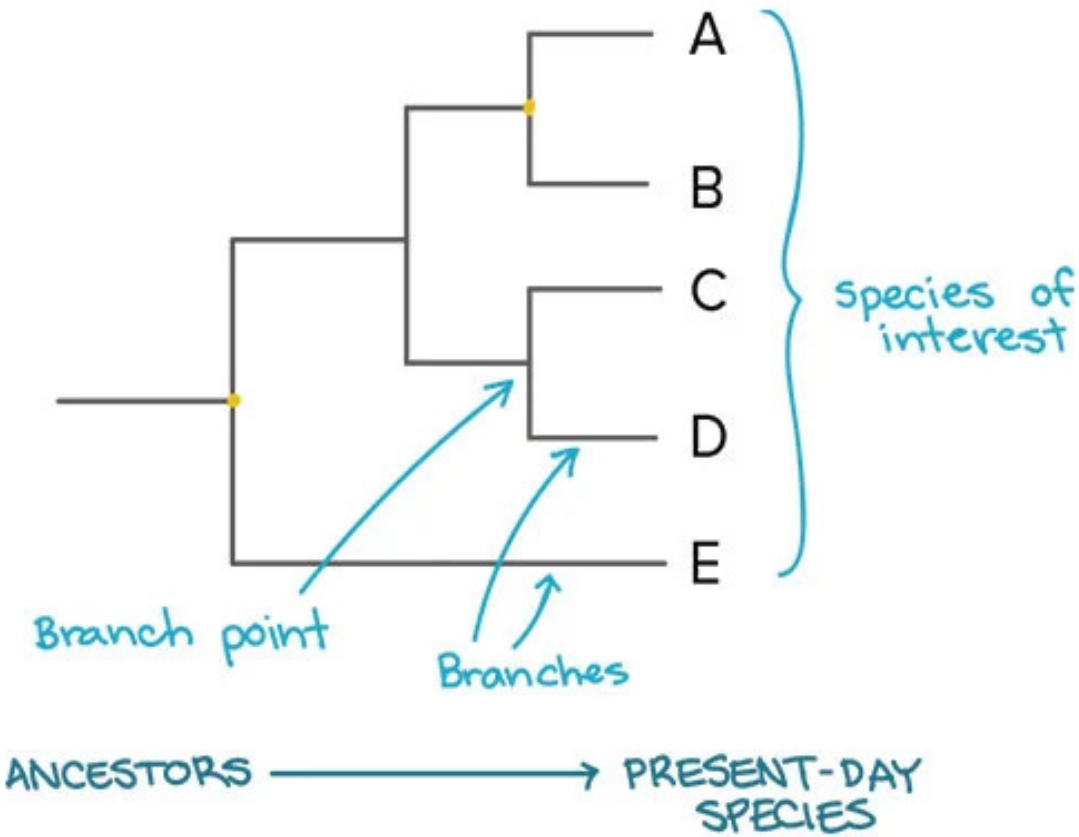
OTUs vs ASVs

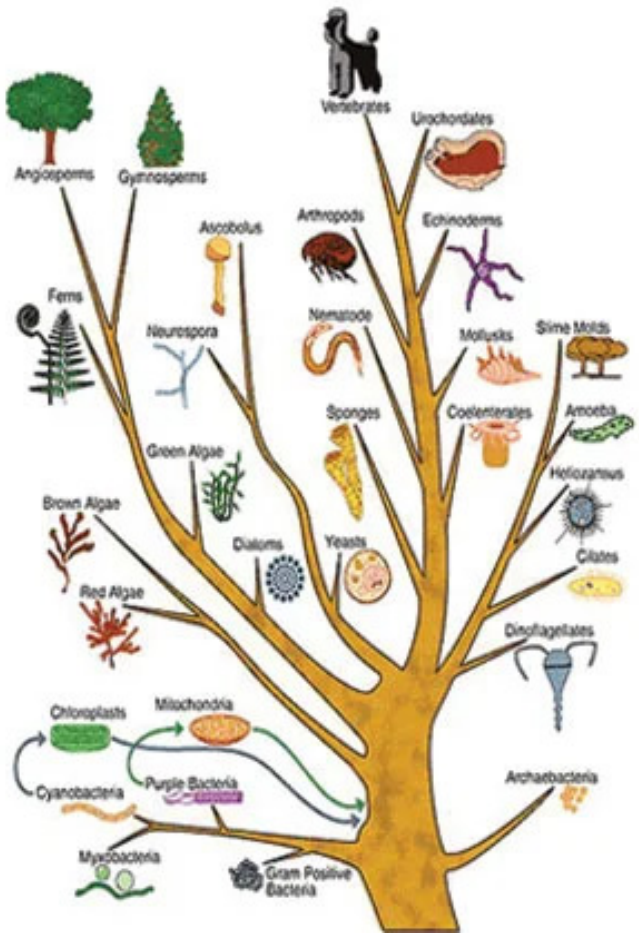
- Operational Taxonomic Units (OTUs): clustering based approach
  - Usually, 97% similarity for genus level.
  - Species level information is usually not possible for 16s data.
  - Possible to match to existing database: SALIVA, greengene etc.
  - de novoOTUs defined in two different data sets cannot be compared.
- Amplicon sequence variants (ASVs): Callahan, et al., *Nature Methods*, 2016
  - Non-clustering based approach
  - ASVs defined in two different data sets cannot be compared.
- Both are provided in Qiime2 pipeline.

Typical data from 16s sequencing

# Constructed from biom file											
#OTU ID	66058686		66165152		66167056		66158660		66167050		66167120	
4386761	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1800048	0.0	0.0	0.0	1.0	3.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0
358030	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.0	5.0	1.0	0.0	0.0
309284	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0
4030157	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
296165	152.0	9.0	36.0	16.0	2.0	0.0	1.0	4.0	6.0	0.0	6.0	1.0
309873	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
337735	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
592160	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
189110	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
327049	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
193591	0.0	7.0	6.0	4.0	1.0	2.0	4.0	8.0	2.0	3.0	3.0	1.0
4355379	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
782984	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
330294	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	55.0	0.0	0.0	507.0
177224	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0
177222	1.0	0.0	4.0	0.0	0.0	3.0	0.0	1.0	0.0	1.0	2.0	0.0
569210	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0
3589405	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
949863	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1026778	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0

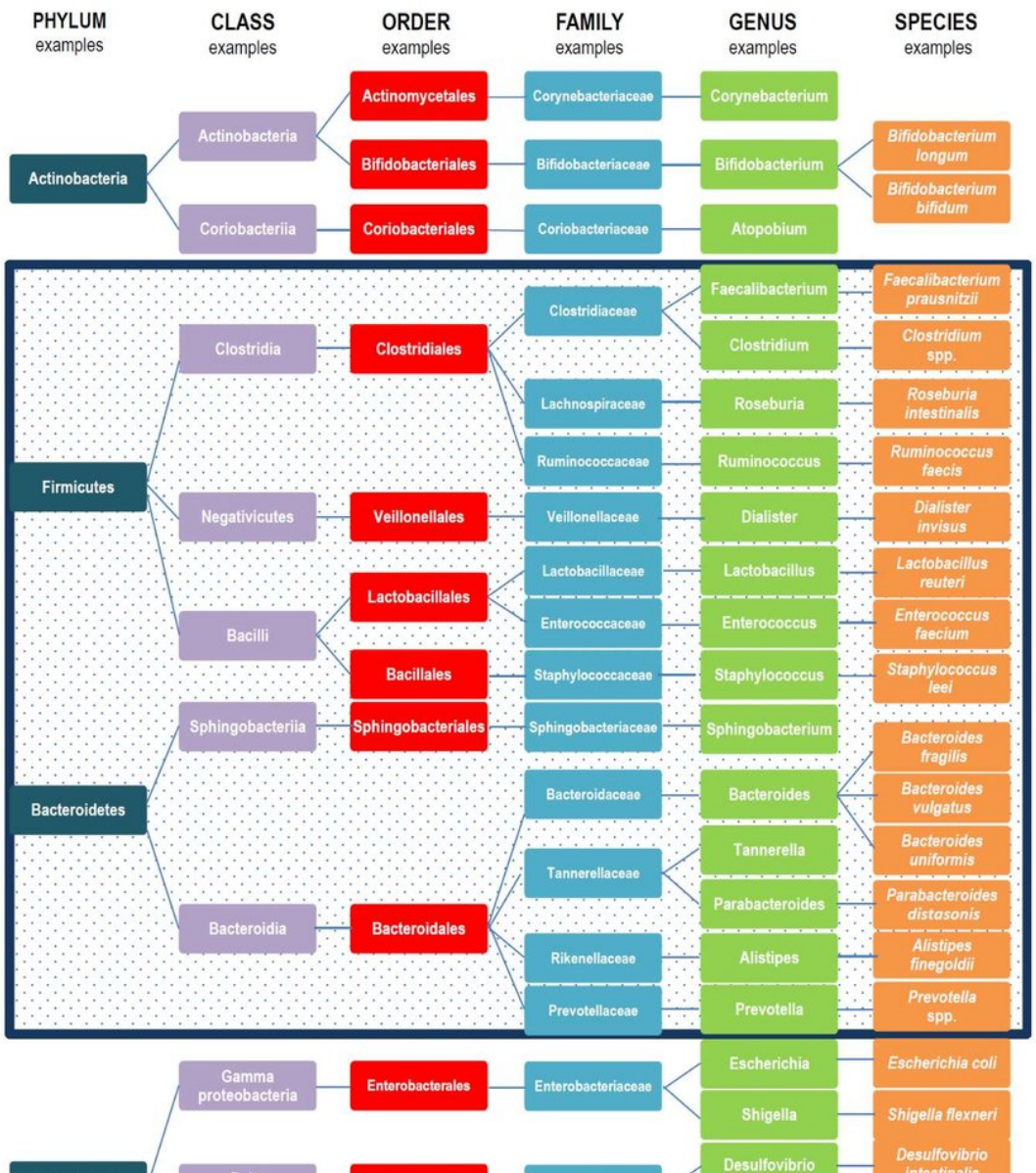
Phylogenetic tree



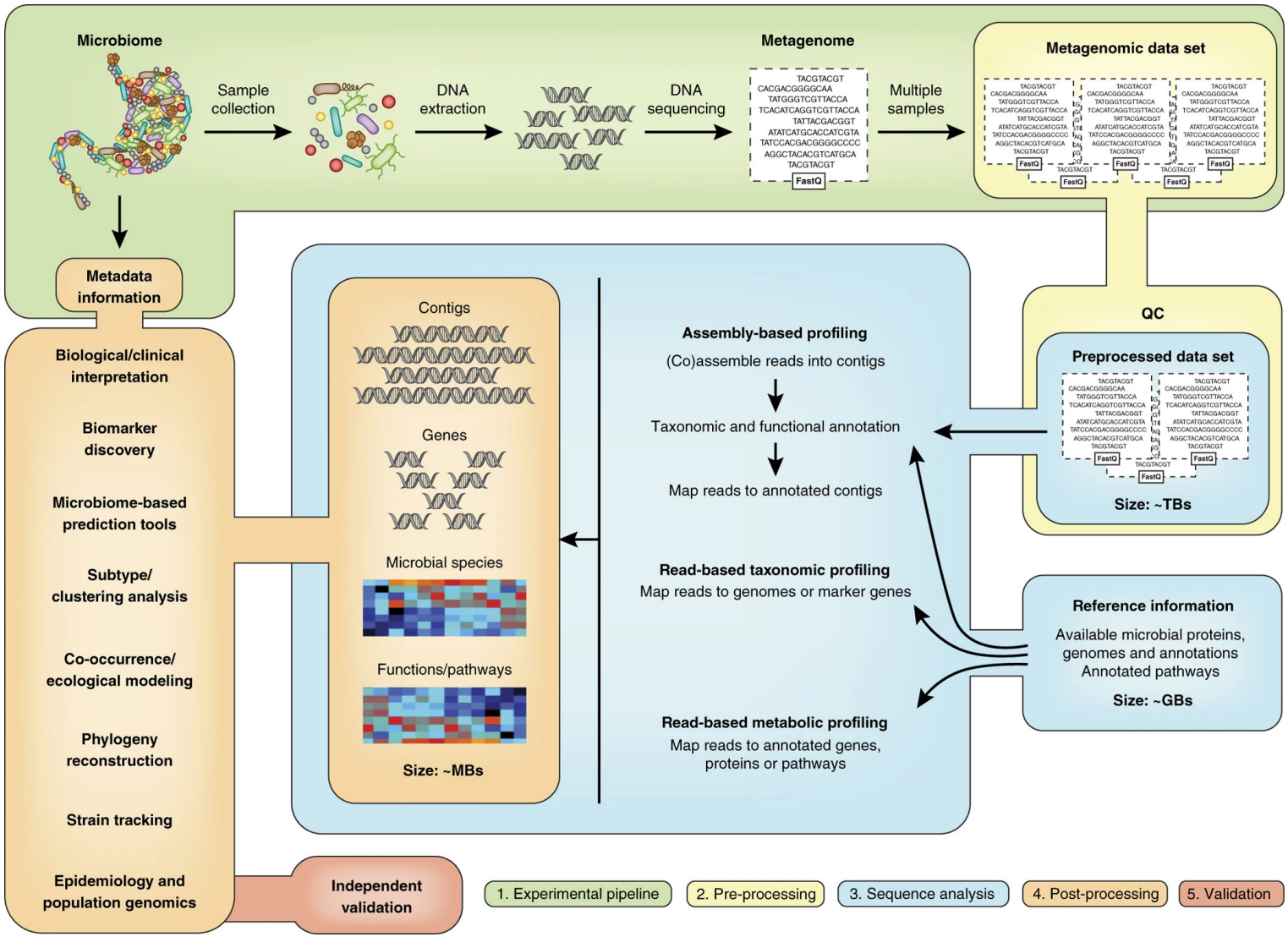


https://microbenotes.com/how-to-construct-a-phylogenetic-tree/

Taxonomy tree



Shotgun metagenomics sequencing



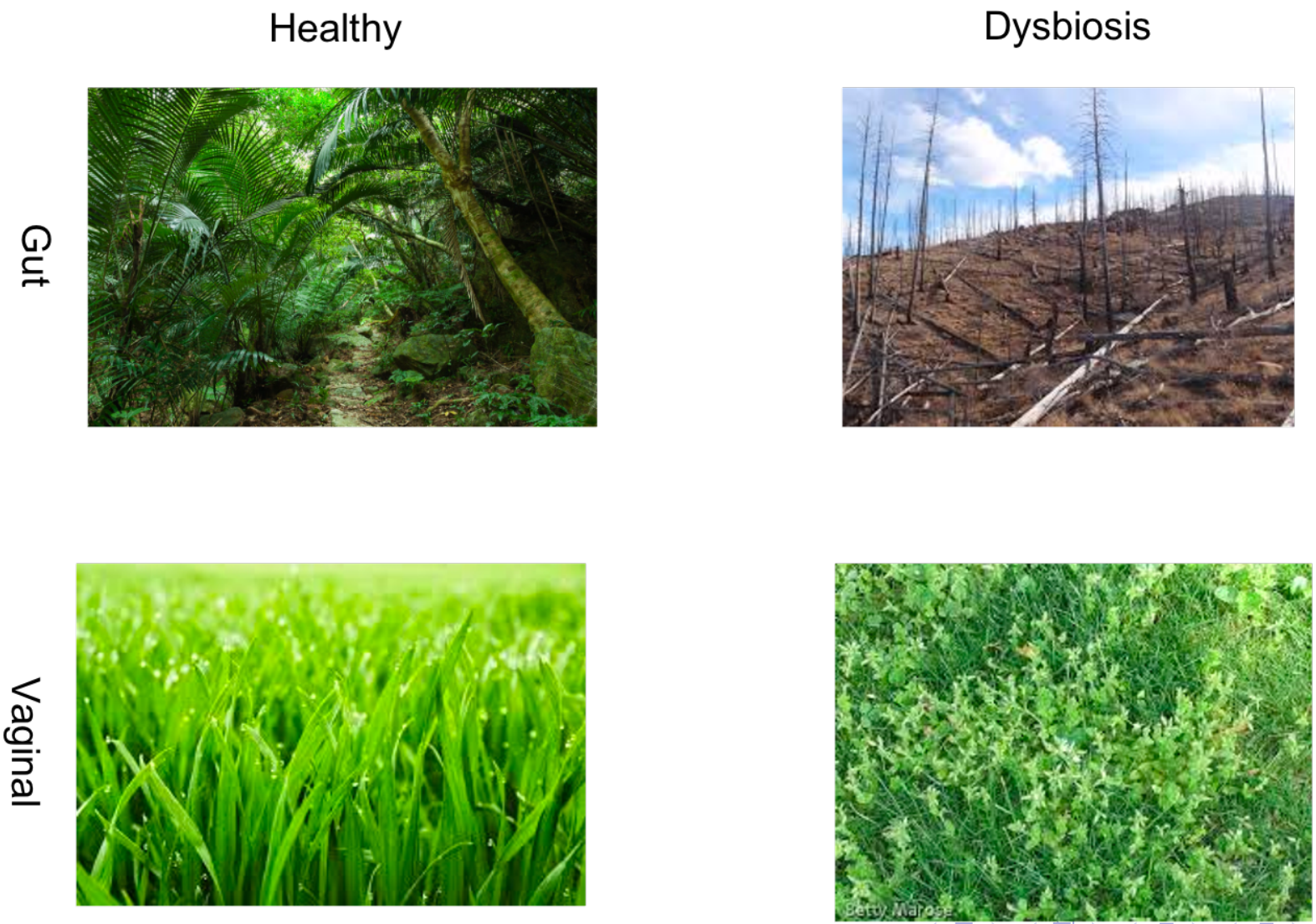
Quince et al., *Nature Biotechnology*, 2017

Compared to marker gene sequencing, shotgun sequencing

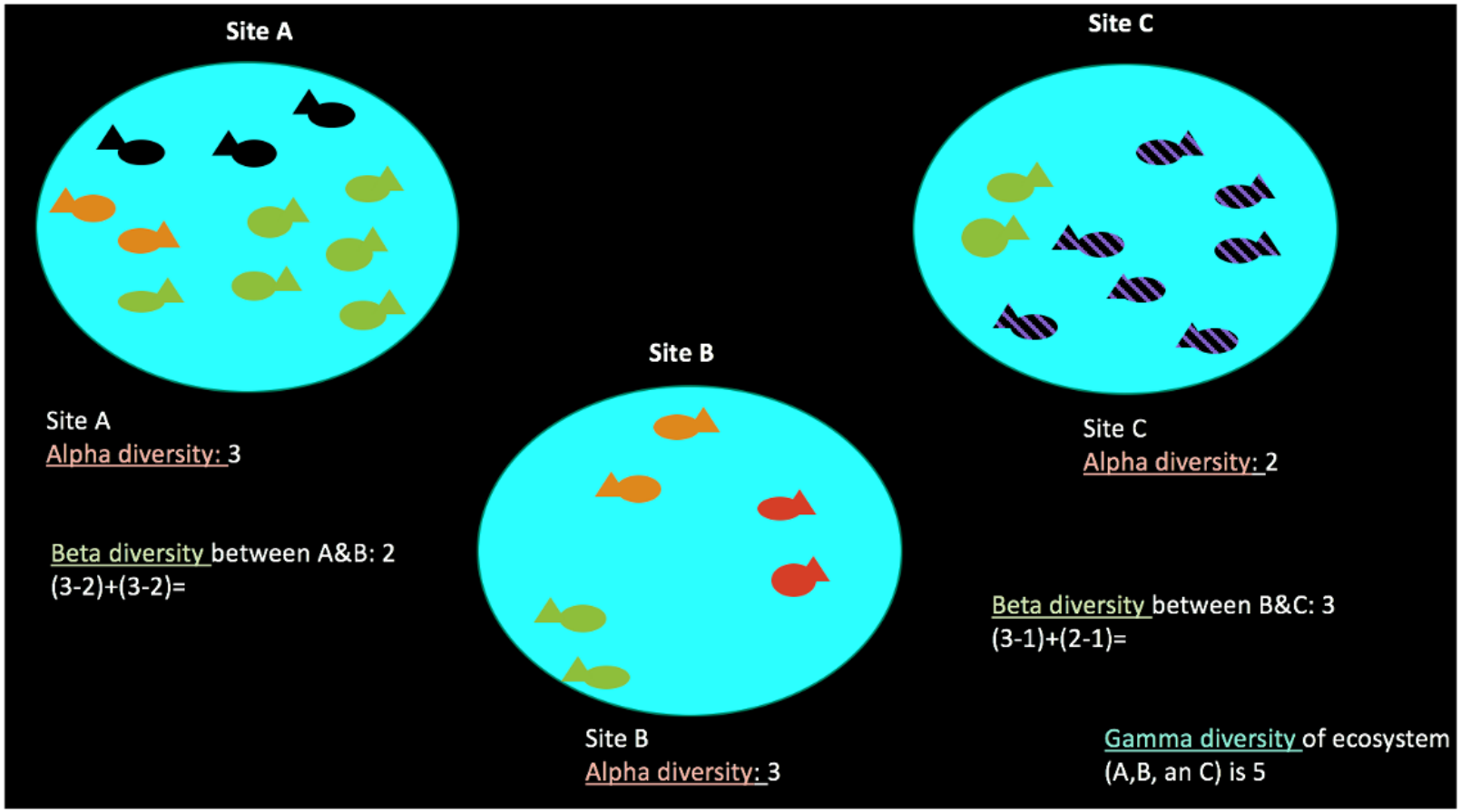
- sequences all genomic DNA in your sample.
- able to obtain species level resolution.
- obtain genomic and functional content in addition to taxa information. (inferring functional pathway is possible for marker gene sequencing: PiCRUSt)
  - Antibiotic-resistant genes.
  - Strain specific information.
- Other possible use: inference of bacteria growth rate.

- More expensive
- Higher false positives
- Host DNA interference

Community structure



Diversities

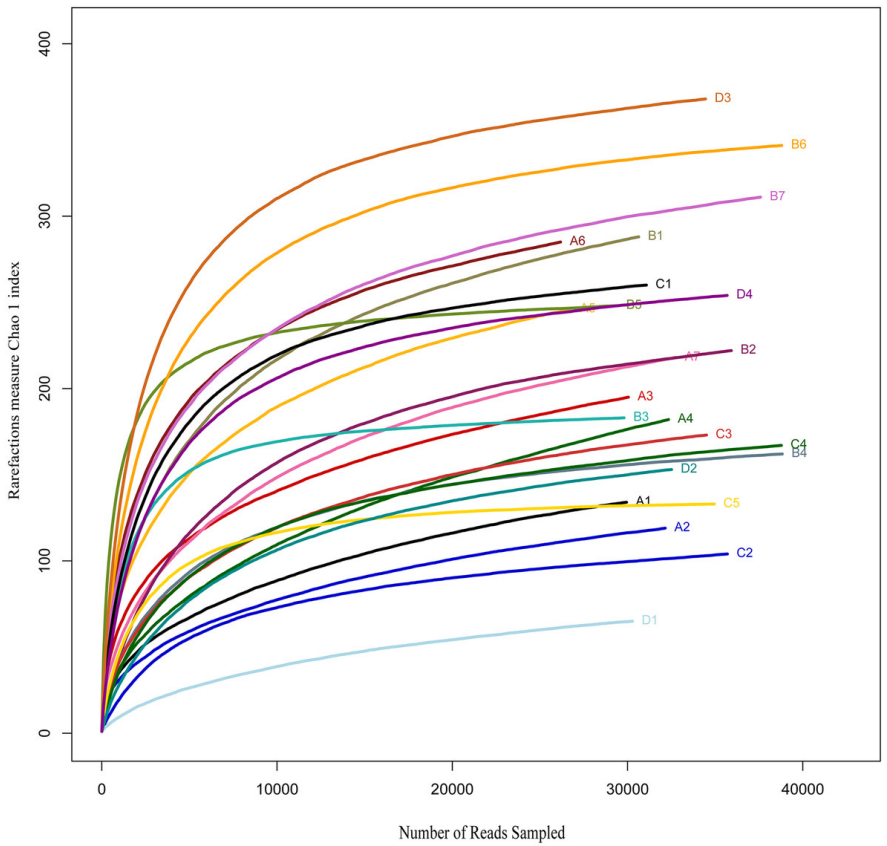


Alpha diversity

Diversity	Description	Formula
Non-phylogenetically based		
Observed Species	No. of unique taxa in a sample	Number of species
Chao 1	Adding a correction to observed species	$S_{obs} + \frac{n_1^2}{2n_2}$
Shannon	Both richness and evenness	$-\sum_{i=1}^{s} p_i \log p_i$
Simpson reciprocal	Both richness and evenness	$\frac{N(N-1)}{n(n-1)}$
Pielou's evenness	evenness	$\frac{H}{H\_max}$
Phylogenetically based		
PD-Tree	Based on where the bacteria are in the evolution tree	

Diversity depends on the intensity of sampling

- Rarefaction



Xuemei Li et al., *PeerJ*, 2018

Rarefying: controversial yet popular

- Appears in most standard analysis pipelines.
- A nonparametric robust approach that appears to work well in real data.
- McMurdie and Holmes, 2014
  - Rarefying throws away data.
  - Inadmissible in both clustering of communities and DA analysis.
- Weiss et al, 2017
  - "Rarefying more clearly clusters samples according to biological origin than other normalization techniques do for ordination metrics based on presence or absence."

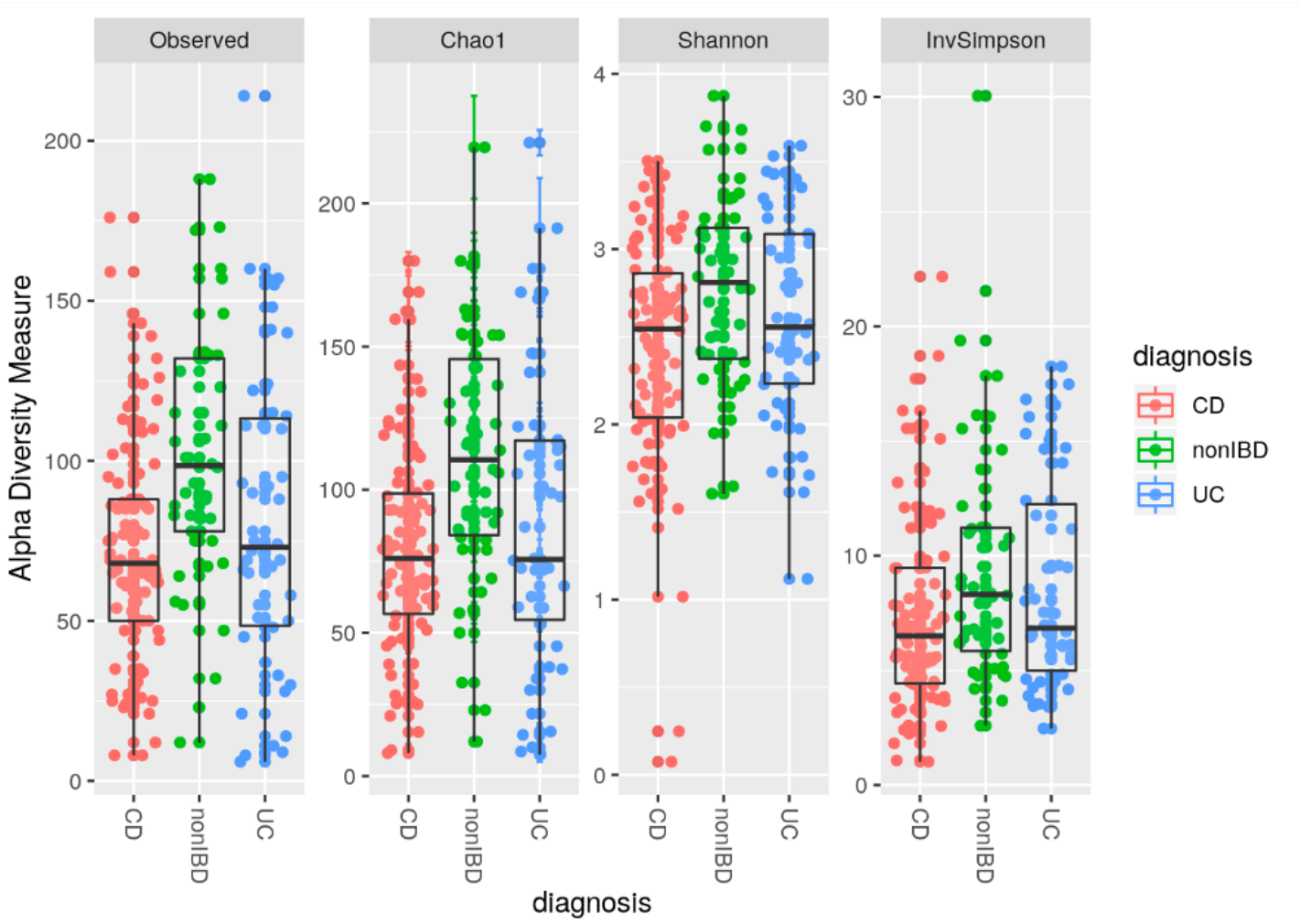
Other approaches: estimate the unobserved taxa.

- W_{ij} read counts in sample i, taxon j.
- N_i: total number of (all, including nonobserved) taxa in sample i.
- Assumes a Poisson Multinomial distribution: $N_i \sim \text{Pois}(\nu_i)$
- $f^*_{X_i}(W_{i1}, \cdots, W_{ip}|N_i) = \frac{N_i!}{\prod_{j=1}^p W_{ij}!} \prod_{j=1}^p {X^*_{ij}}^{W_{ij}}$
- MLE of $\hat{X}ij = W_{ij} / \sum_{k=1}^p W_{ik}$, which is not very suitable due to dropout.
- Regularized approach provided.

Yuanpei Cao et al., *Biometrika*, 2020

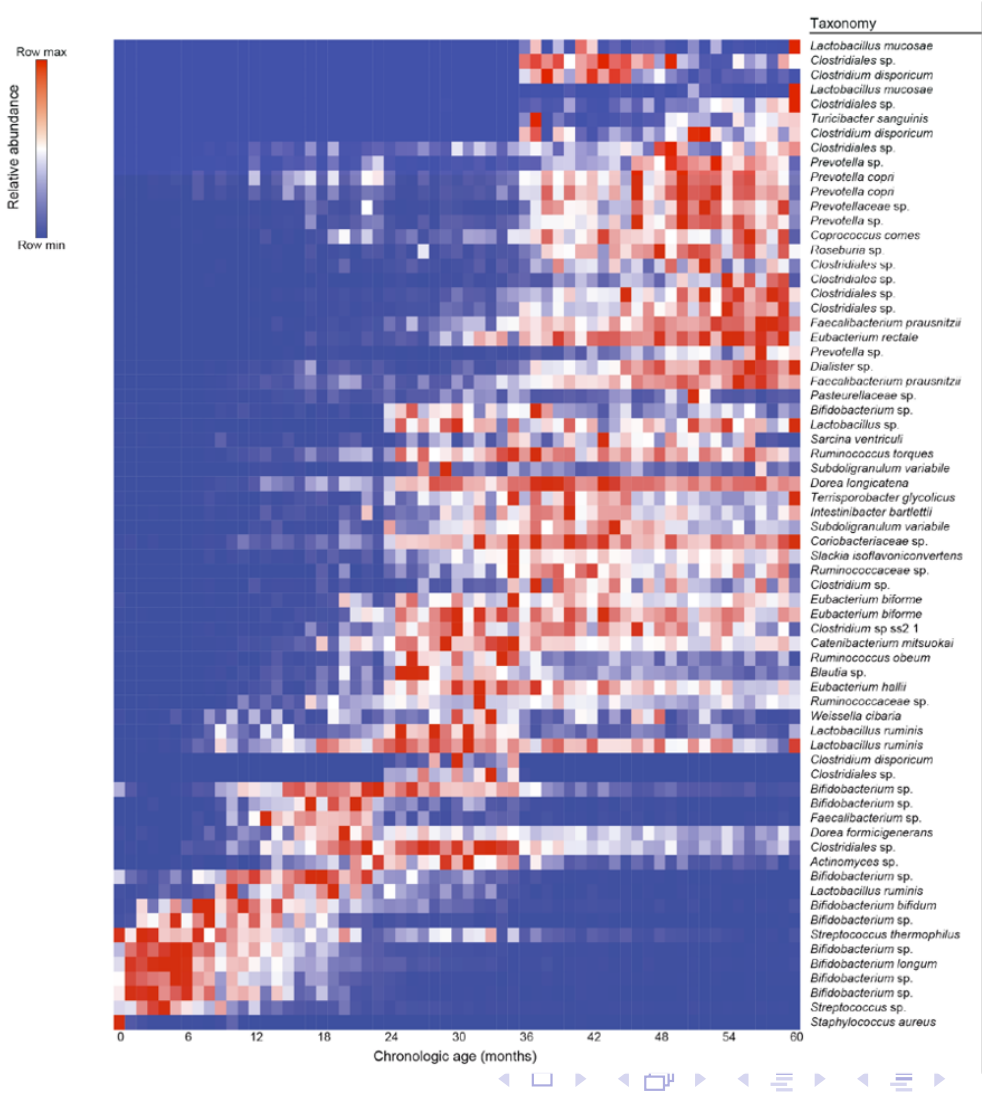
Alpha diversity analysis

As alpha diversity is univariate, after it is calculated, apply your STATS 101.



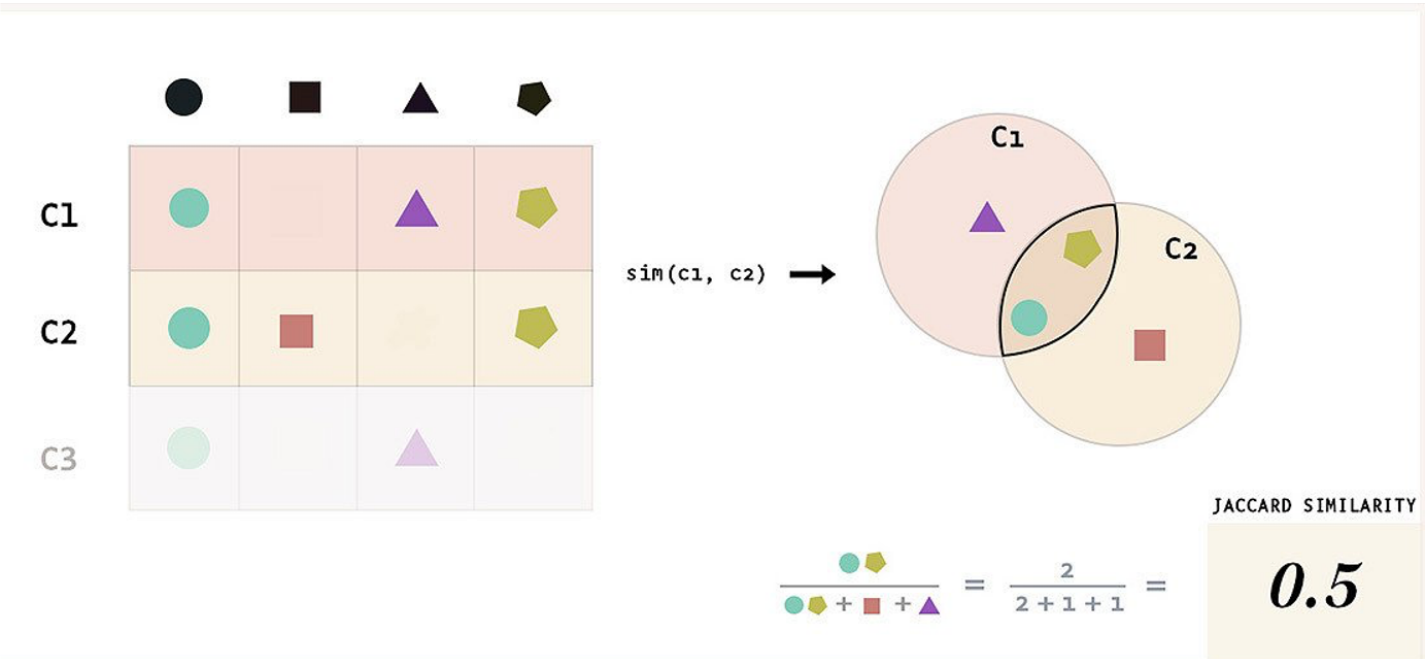
Defining normal gut microbiota development

- Healthy members of Bangladeshi birth cohort
- Monthly sampling (1-60 months)
- 16S rDNA-based analysis
- AAAS2022, Jeff Gordon



Beta diversities: captures the dis-similarity between samples.

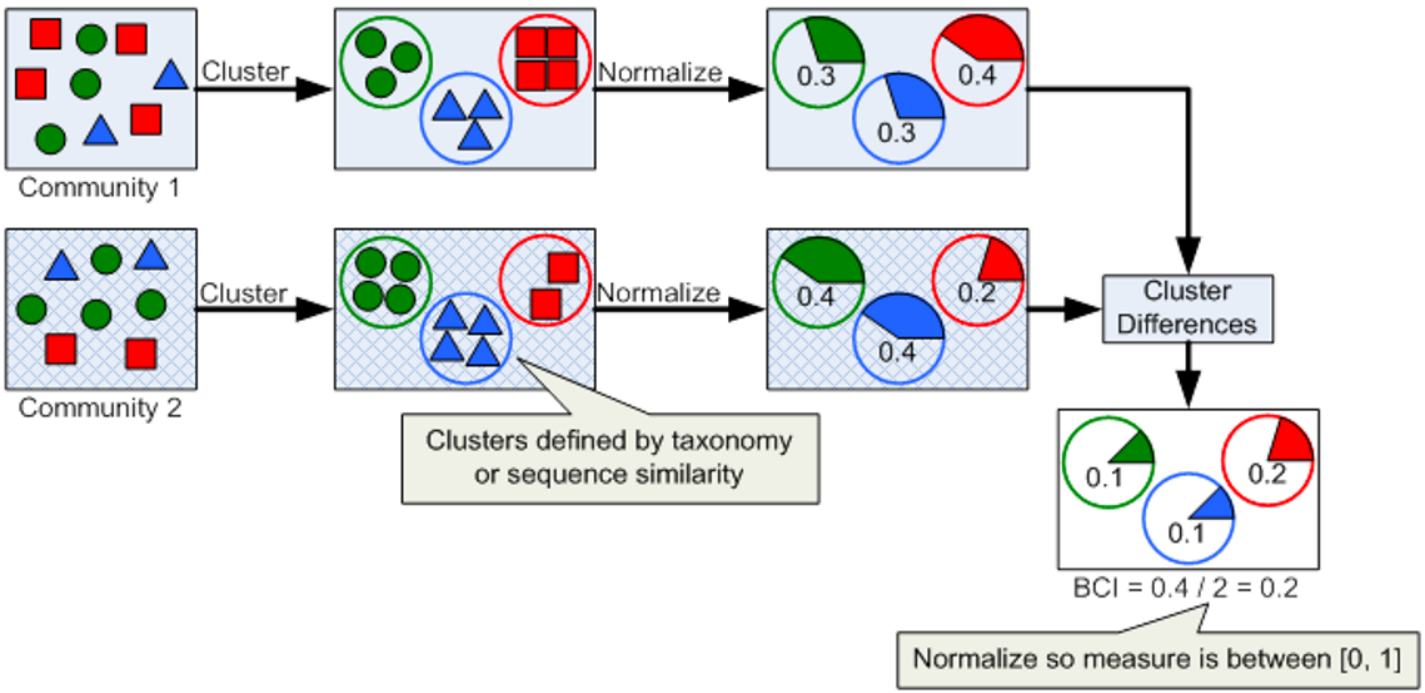
Jaccard distance



Jaccard distance: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

https://www.learndatasci.com

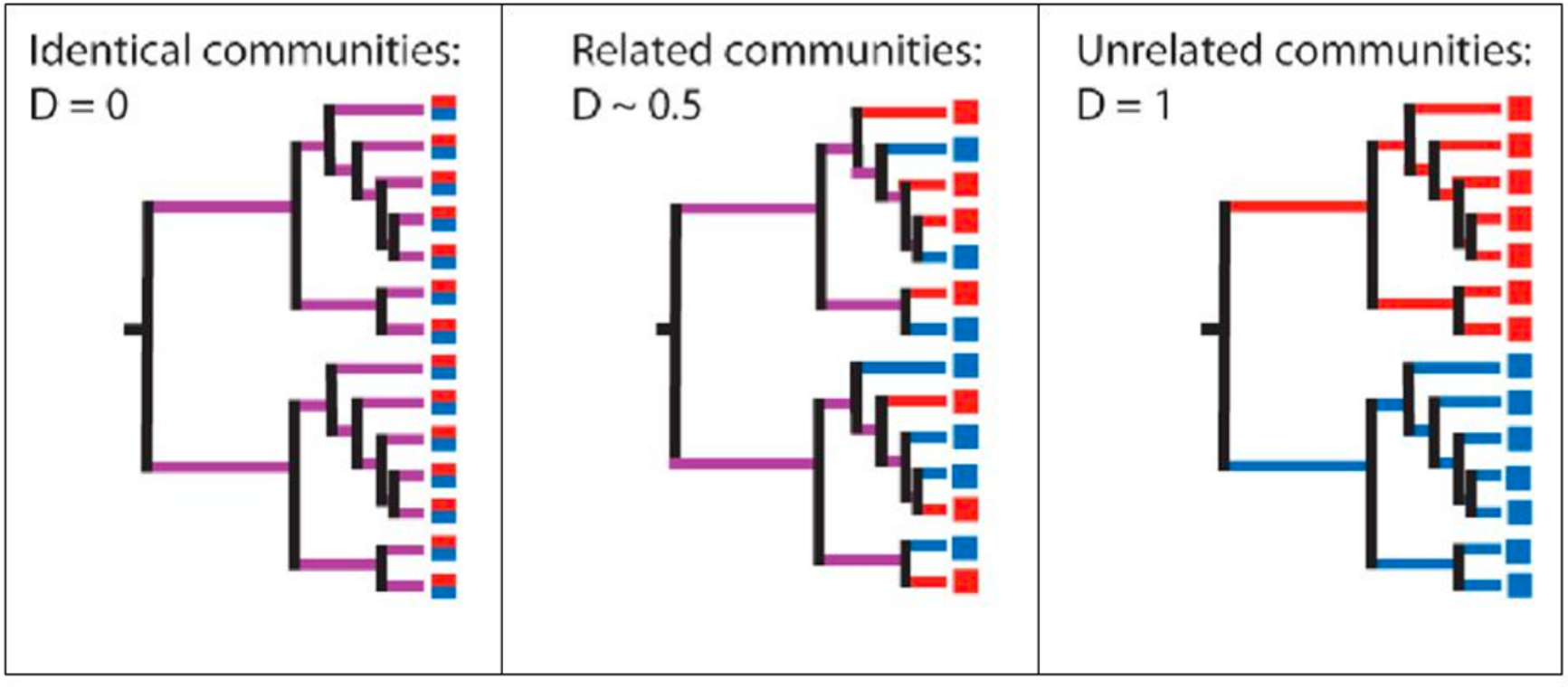
Bray-Curtis Distance



$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j} = 1 - 2 \times \frac{0.3+0.3+0.2}{2} = 0.2$$

Unique Fraction (UniFrac) metric

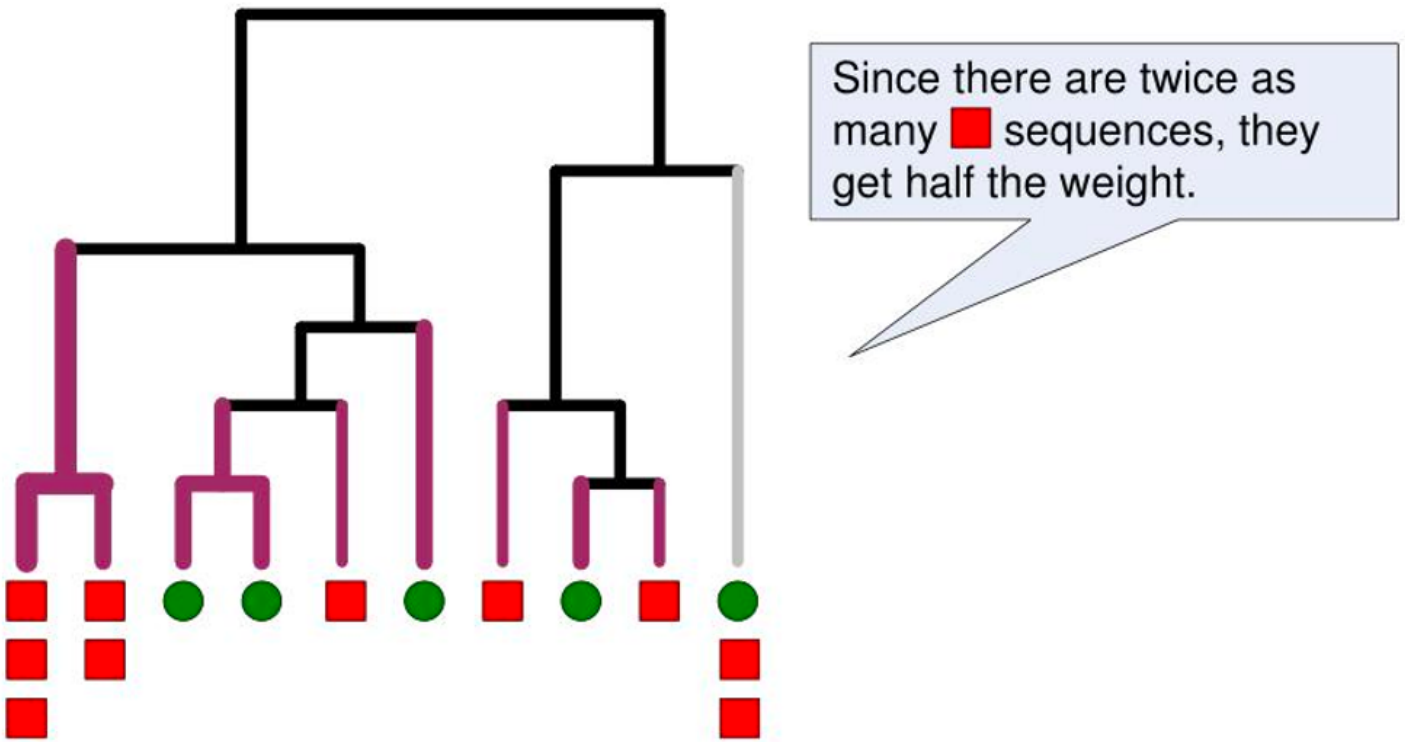
- Qualitative phylogenetic β diversity.
- Distance = fraction of the total branch length that is unique to any particular environment.



Lozupone and Knight, 2005, Appl Environ Microbiol 71:8228

Weighted UniFrac

- What about the relative abundance of sequences (i.e., evenness)?

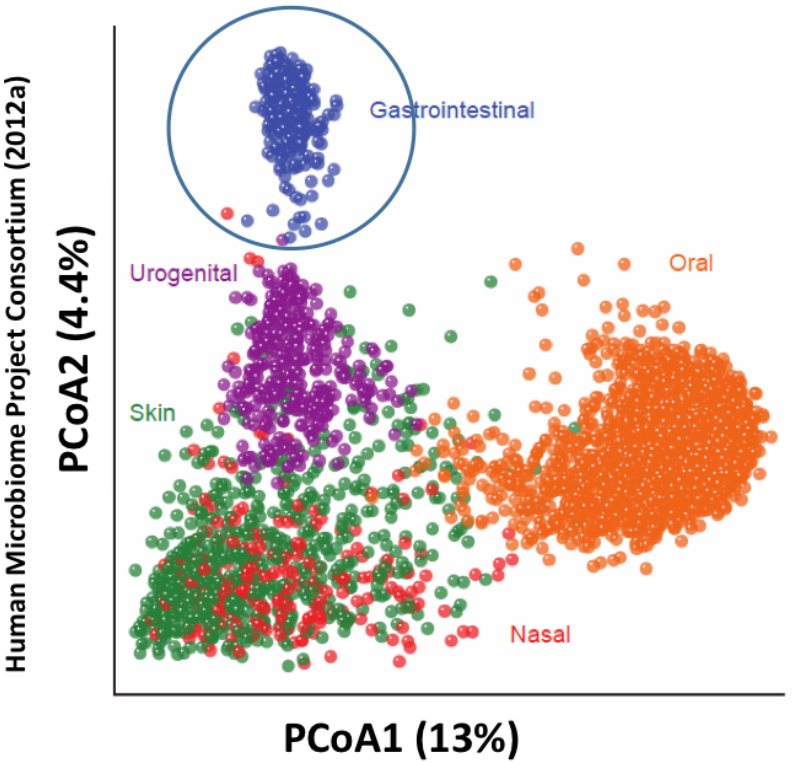


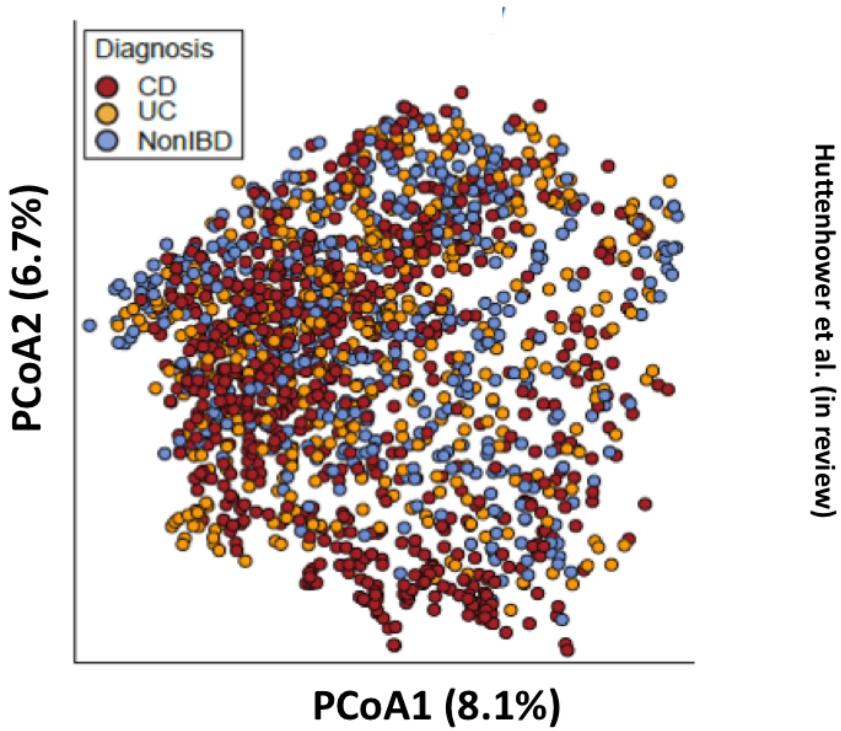
Generalized UniFrac: Chen J.,et al., *Bioinformatics*. 2012

Commonly used beta diversities

Diversity	Description	Formula
UniFrac	Qualitative, Phylogenetics-based	$\sum_{i=1}^{n} \frac{b_i \lvert I(p_i^A > 0) - I(p_i^B > 0) \rvert}{\sum_{i=1}^{n} b_i}$
Weighted UniFrac	Quantitative, Phylogenetics-based	$\frac{\sum_{i=1}^{n} b_i \lvert p_i^A - p_i^B \rvert}{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)}$
Generalized UniFrac	Compromise between the previous two	$\frac{\sum_{i=1}^{n} b_i \left(p_i^A + p_i^B\right)^{\alpha} \lvert \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \rvert}{\sum_{i=1}^{n} b_i \left(p_i^A + p_i^B\right)^{\alpha}}$
Bray-Curtis	Quantitative	$1 - \frac{2C_{ij}}{S_i + S_j}$
Jaccard	Qualitative	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$

Principal Coordinate Plot





Microbial community composition in each body region is distinct.

But large-scale community composition alone cannot differentiate host phenotypes.

PCoA

- Distance to similarity $D = [D_{hi}]$, $A = -\frac{1}{2}[D_{hi}^2]$
- Gower centering (1966).

$$G = (I - \frac{11'}{n})A(I - \frac{11'}{n})$$

- Eigen-decomposition of G
- Project the distance metric into lower dimensional space

Note: metric multidimensional scaling vs nonmetrix multidimensional scaling (MDS vs NMDS)

PERMANOVA

Austral Ecology (2001) **26**, 32–46

A new method for non-parametric multivariate analysis of variance

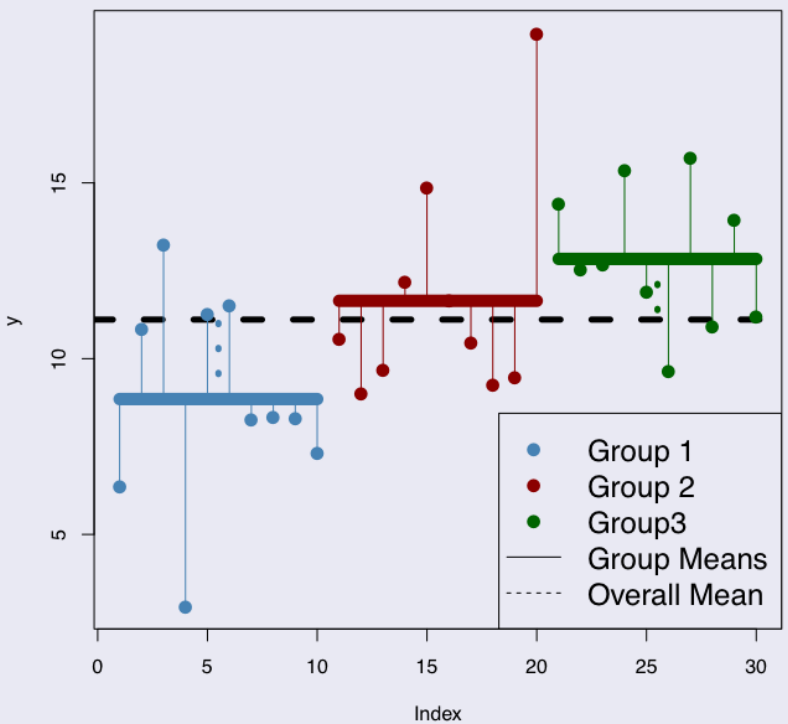
MARTI J. ANDERSON
Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology L
University of Sydney, New South Wales 2006, Australia

Abstract Hypothesis-testing methods for multivariate data are needed to make rigo
about the effects of factors and their interactions in experiments. Analysis of variance

- Non-parametric approach combined with ecological distance measures!

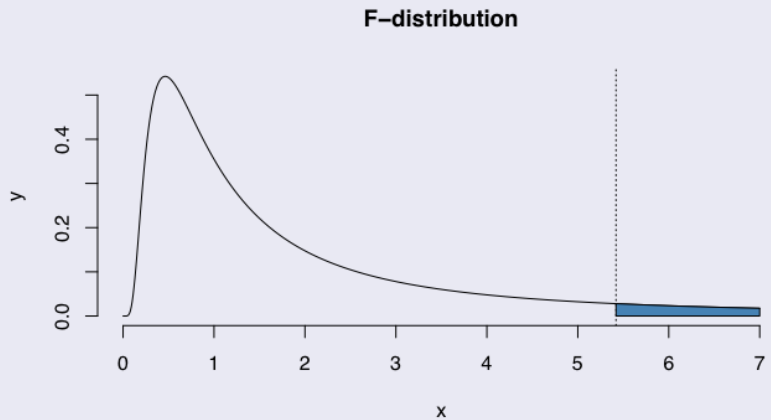
Question : How is univariate ANOVA calculated?

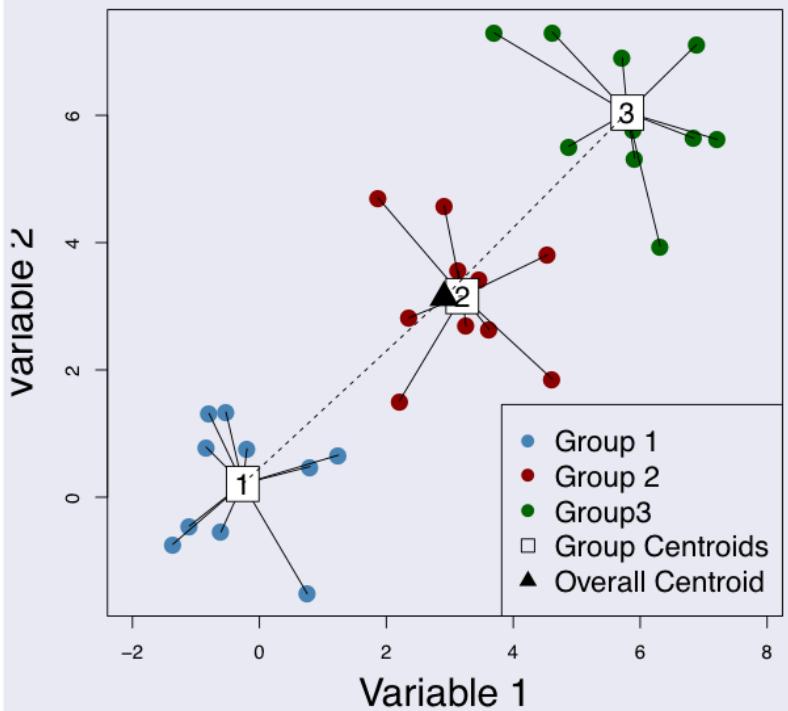
From univariate...



$$SS_{total} = SS_{residual} + SS_{group}$$

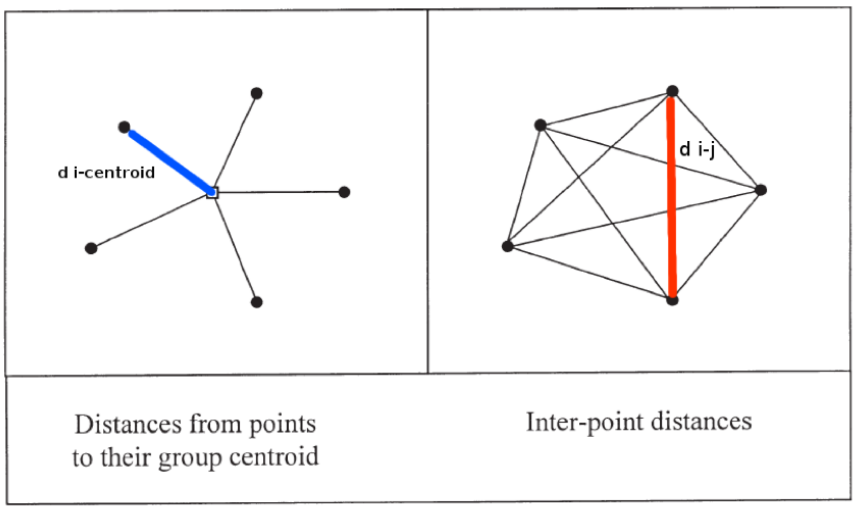
$$F - ratio = \frac{SS_{group}}{SS_{residual}} \cdot \frac{df_{residual}}{df_{group}}$$





- Partitioning into variance components:
  $SS_{total} = SS_{group} + SS_{residual}$
- **centroids**
- p-value by **permutations**

- We can use any distance matrix to partition the variance
- Sum of squared distances from individual points to their centroid equals the sum of squared interpoint distances divided by the number of points.



$$\sum d^2_{i-centroid} = \frac{1}{n} \sum d^2_{i-j}$$

PERMANOVA

- Distance to similarity $D = [D_{hi}]$, $A = -\frac{1}{2}[D_{hi}^2]$
- Gower centering (1966).

$$G = (I - \frac{11'}{n})A(I - \frac{11'}{n})$$

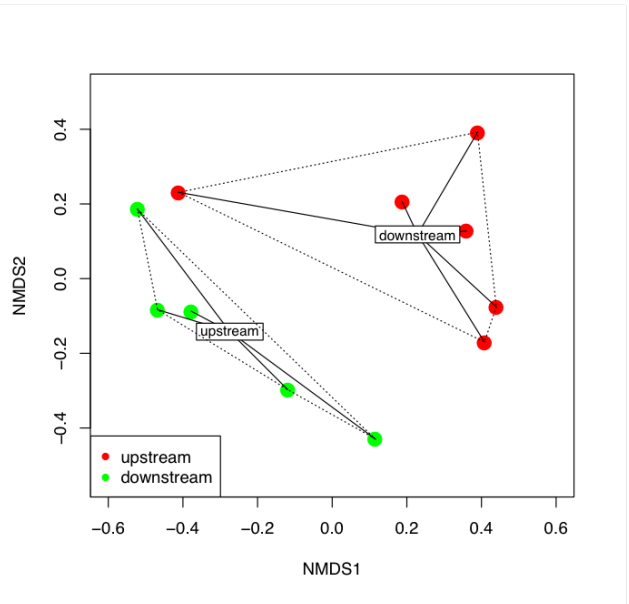
- $F = \frac{tr(HGH)}{tr(I-H)G(I-H)}$, with $H = X(X'X)^{-1}X'$
- Testing via permutations

Assumptions behind PERMANOVA

- Equal dispersions
- Visual inspection

Pros and Cons

- Very powerful and easy to implement (adnois in *vegan* package)
- Extension to other studies more difficult
- Multiple distance test is challenging



Kernel Machine Regression–A Similarity based Approach

$$y = \mathbf{X}\boldsymbol{\beta} + h(\mathbf{Z}) + \boldsymbol{\varepsilon}$$

where $h(\cdot)$ is a function in a function space $\mathcal{H}_K$ generated by positive semidefinite kernel function $K(\cdot, \cdot)$

- $\phi_1, \cdots, \phi_J$(J might be infinite) form a basis of $\mathcal{H}_K$, then

$$h(\mathbf{Z}_i) = \sum_{j=1}^{J} \phi_j(\mathbf{Z}_i)\gamma_j = \phi(\mathbf{Z}_i)\boldsymbol{\gamma},$$

- Dual representation: $h(\mathbf{Z}_i) = \sum_{i'=1}^{n} \alpha_{i'} K(\mathbf{Z}_i, \mathbf{Z}_{i'})$.
- Simple linear kernel $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \sum_{j=1}^{p} Z_{ij} Z_{i'j} \Rightarrow \phi_j(\mathbf{Z}) = \mathbf{Z}_j$

Kernels Specific to Microbiome Data

Key Idea: use the microbiome distance metrics to define similarities.

$$\mathbf{K} = -\frac{1}{2}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{D}^2(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})$$

Kernels Specific to Microbiome Data

Key Idea: use the microbiome distance metrics to define similarities.

$$\mathbf{K} = -\frac{1}{2}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{D}^2(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})$$

Kernels for Microbiome Data

- UniFrac distance $\rightarrow$ UniFrac kernel
  - Fraction of branches of the phylogenetic tree present in sample i or sample i', but not both
- Weighted UniFrac distance $\rightarrow$ Weighted UniFrac kernel
  - Incorporate abundance information
- Others: composite distances, Bray-Curtis, Rao's root of patristic distances, etc...

Testing $h(\mathbf{Z}) = 0$

- KMR is equivalent to linear mixed model

$$y = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon}$$

  $\mathbf{h} \sim F$ with mean 0 and variance $\tau\mathbf{K}$

- $H_0 : h(\mathbf{Z}) = 0 \Leftrightarrow H_0 : \tau = 0.$
- Variance component statistic: $Q = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{K}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}^2$
- Under H_0, Q asymptotically follows a mixture of χ^2 distribution
- Small Sample Exact Distribution (*Chen et al,. Gen Epi 2016)*

Equivalent to PERMANOVA

- No additional covariates
- $\mathbf{K}$ corresponds to $\mathbf{D}$

Omnibus Test

- The previous model uses one kernel
- Different Ks are targeted toward different scenarios

Omnibus Test

- The previous model uses one kernel
- Different Ks are targeted toward different scenarios

Multiple Kernel Testing

- Construct single-kernel testing and get their p-values p_k
- Use $p^o = \min_{1 \le k \le l} p_k$ as the test statistic
- A residual permutation approach to obtain the null distribution of p^o considering the correlations
- Compare p^o with $p_1^o, \cdots, p_b^o \cdots, p_B^o$ for the final p-value