Microbiome Data Analysis

Yijuan Hu

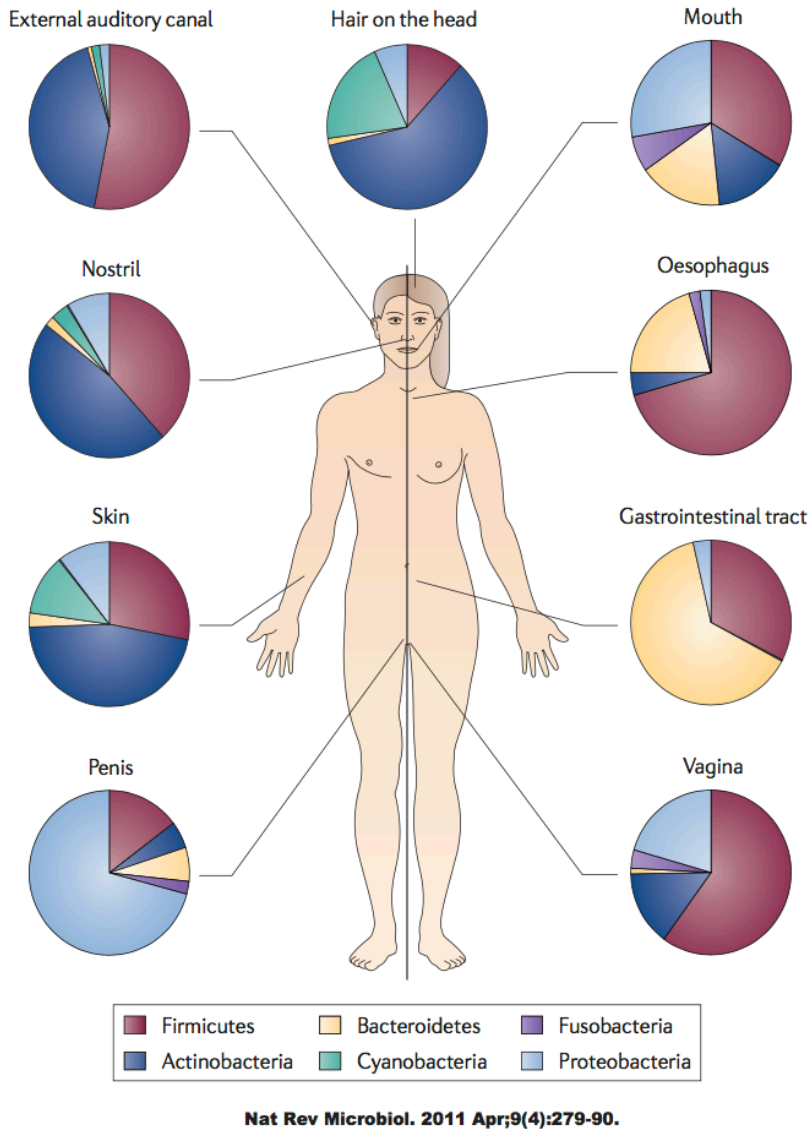
2017-10-12

Outline

- Microbiome and Human Diseases
- Sequencing + Bioinformatics Pipelines
- Human Microbiome Project (HMP), Metagenomics of the Human Intestinal Tract (MetaHIT)
- Statistical Analyses
- Study Design and Power

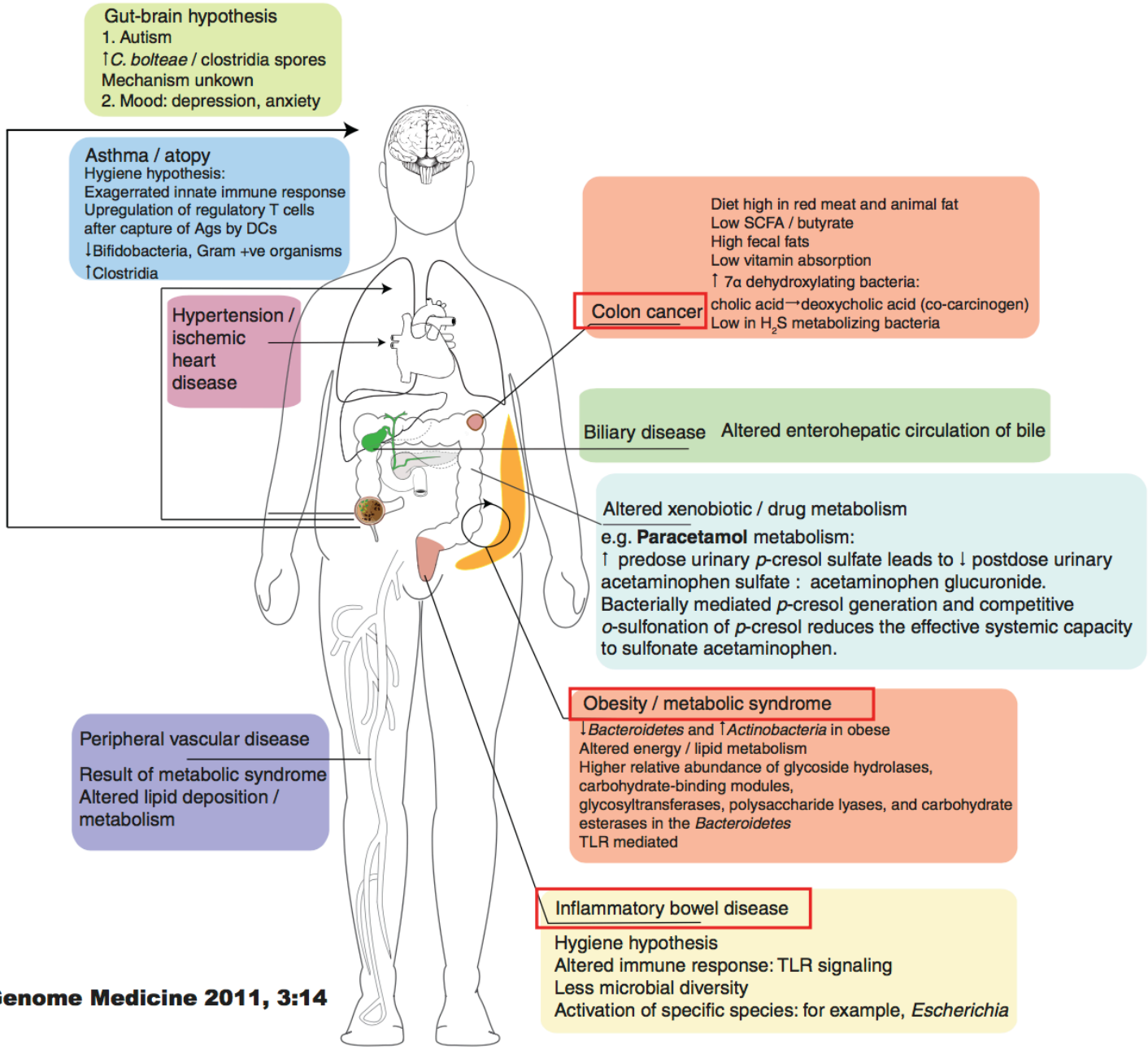
The Human Microbiome

- Microbiome as extended human genome
  - 10^{13} human cells vs 10^{14} bacterial cells
  - Consist of bacteria, fungi, and viruses
  - More than 3×10^6 genes provided by our gut microbiome
  - Distinctive microbiomes at different body sites
  - The human microbiome may explain the missing link between genetic variation and disease
- The human microbiome in health
  - Digestive enzyme activity
  - Synthesis of vitamins
  - Interaction with the immune system
  - Protection from pathogens, etc.



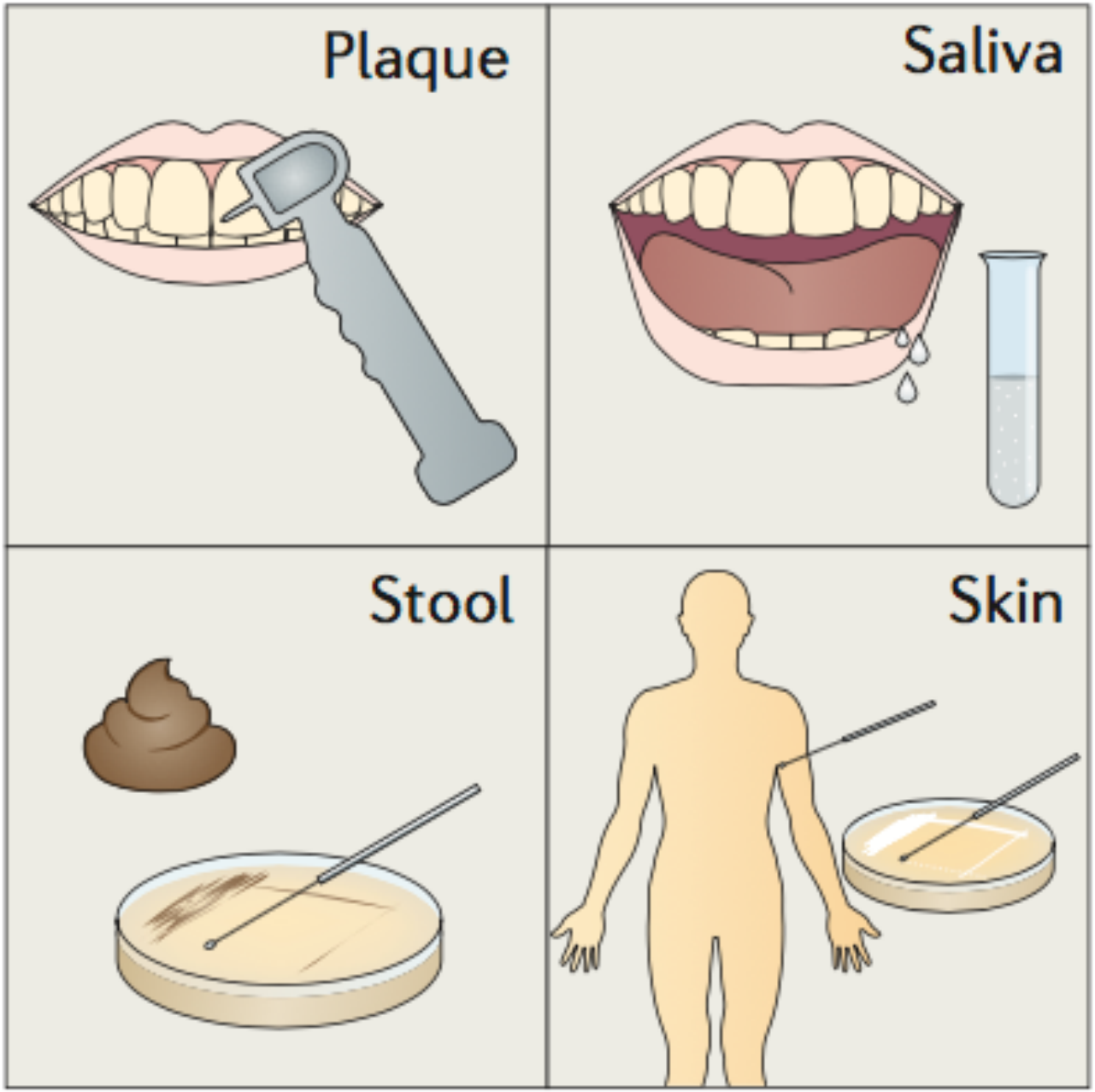
Nat Rev Microbiol. 2011 Apr;9(4):279-90.

Microbiome and Human Diseases

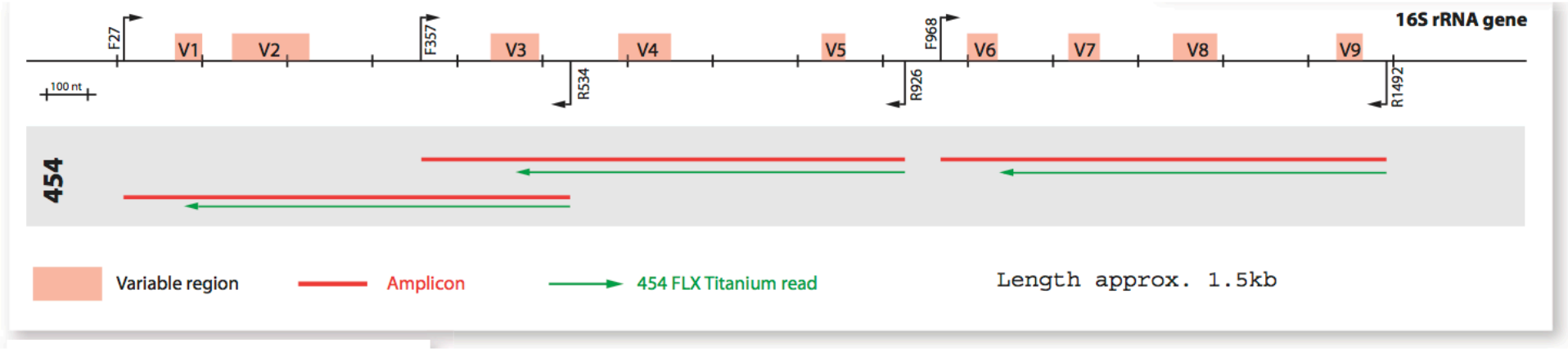


Kinross et al. Genome Medicine 2011, 3:14

Sampling



- 16S rRNA gene targeted sequencing
  - Specific to bacteria, not in fungi and viruses
  - Omnipresent in bacteria
  - Some regions are constant, allowing amplification
  - Some regions are variable, allowing identification of a particular genus and species
  - Reveals “who is there” in terms of relative abundances of bacterial taxa

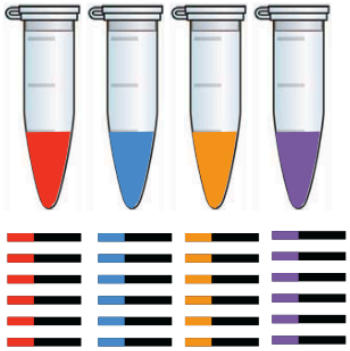


Technologies – Next-Generation Sequencing

- Metagnomic (whole-genome) shortgun sequencing
  - The total extracted DNA is fragmented and sequenced
  - Reveals “what can they do” in terms of the encoded functions of the sequenced microbial DNA
  - 20–30 times more expensive than 16S rRNA gene sequencing, as well as requiring additional computational costs and high-level expertise for performing metagenomic analyses
  - We focus on 16S rRNA gene sequencing data here

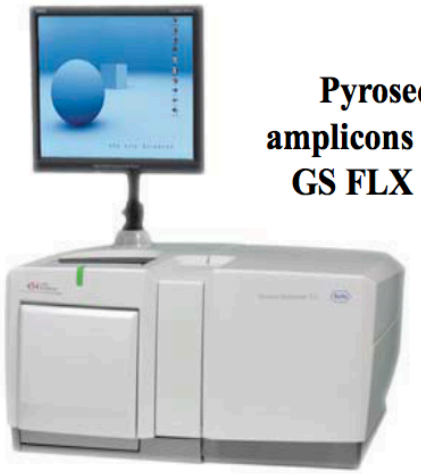
16S Gene Sequencing + Bioinformatics Pipeline

Extract DNA and PCR amplify with barcoded primer



Pool amplicons

Pyrosequence amplicons using 454's GS FLX instrument



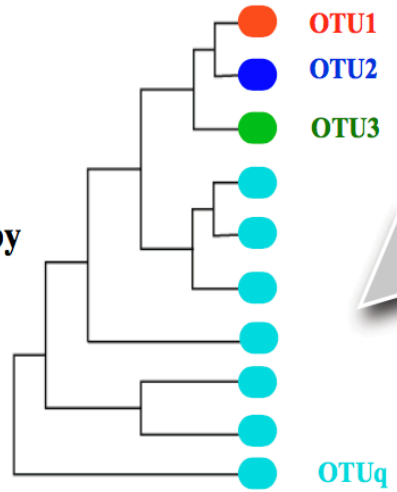
Quality control, assign sequences to samples using barcode and denoise

>AGTGAGAGAAGCAGGGTCGTAATGTT . . . OTU1
>AGTGCGATGCGTAGGGTCGTAATGCG . . .

>AGTGCGATGCGTAGGGTCGTAATGTA . . . OTU2
>AGTGGATGCTCTAGGGTCGTAATGCA . . .

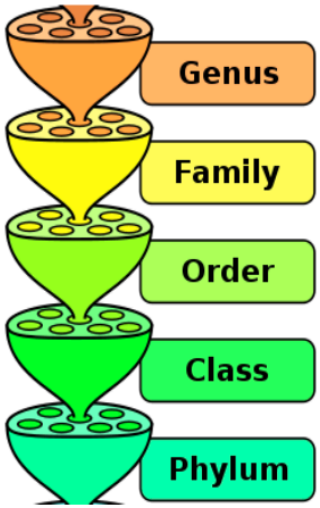
>AGTGTCACGGTGAGGGTCGTAATGGG . . . OTU3
>AGTGGATGCTCTAGGGTCGTAATGTT . . .
>AGTGTCACGGTGAGGGTCGTAATGCC . . .
>AGTGAGAGAAGCAGGGTCGTAATCAC . . .
. . .

Cluster sequences into OTUs (97% level), align to reference alignment (e.g., using PYNAST), infer phylogeny (FastTree)



Thousands of OTUs, 100's of samples

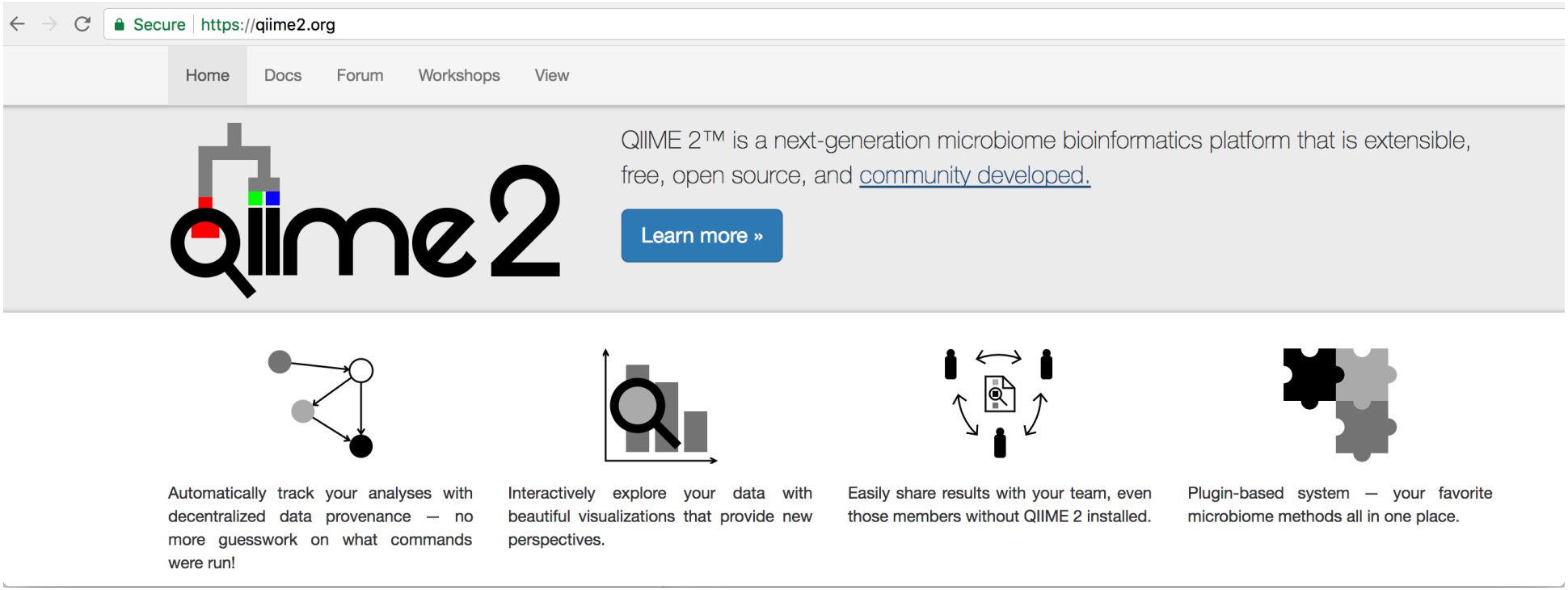
Assign lineages by RDP classifier



Species level OTUs can be summarized into higher levels

OTU: Operational Taxonomic Unit

Bioinformatics Platform – QIIME 2



OTU Table

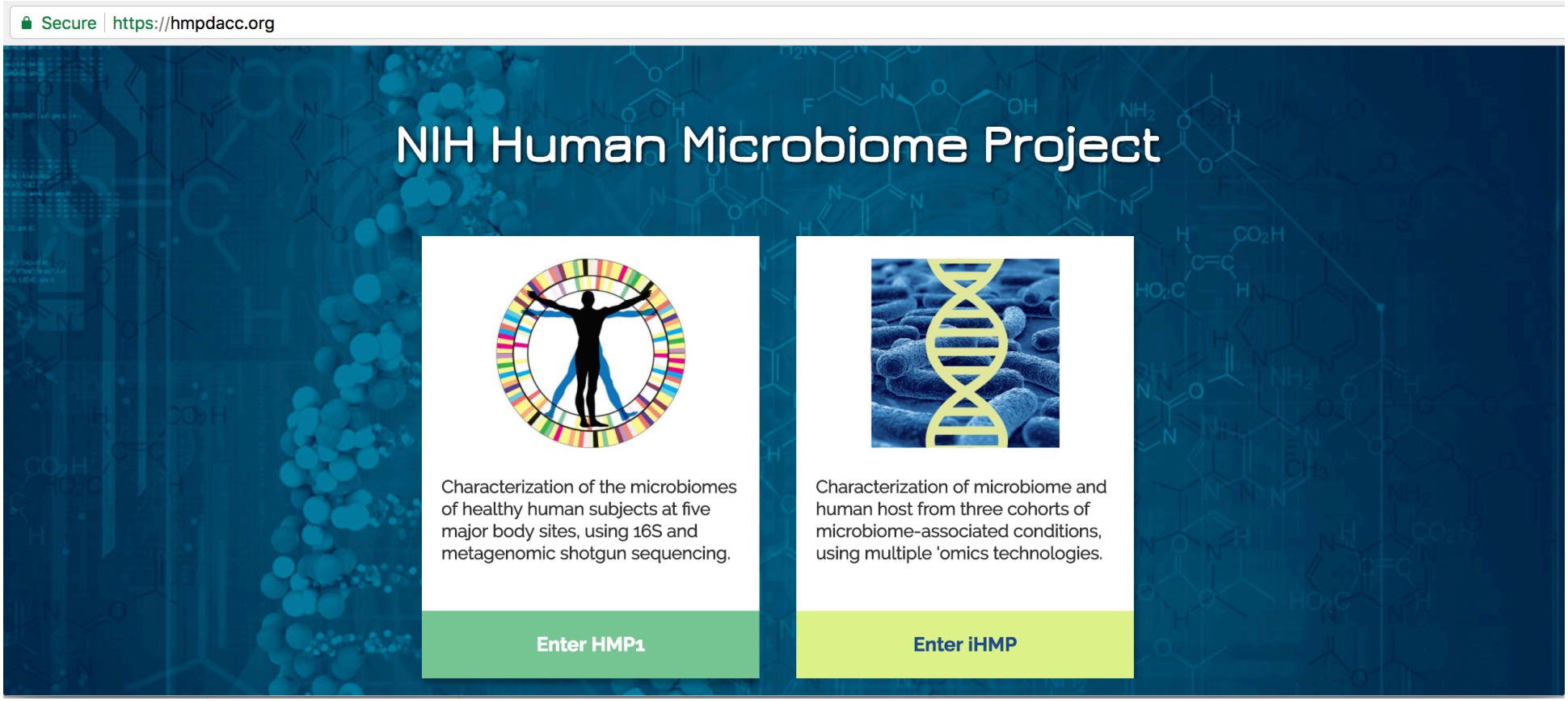
Constructed from biom file

#OTU ID	A1	A2	B1	B2	C1	C2	D1	D2	ConsensusLineage
denovo0	1	0	0	0	0	0	0	0	k__Bacteria
denovo1	0	1	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillospira; s__
denovo2	1	0	1	0	0	1	0	0	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides
denovo3	0	0	0	0	0	2	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Dialister; s__
denovo4	0	1	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus
denovo5	2	0	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillospira; s__
denovo6	0	0	0	0	1	1	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae
denovo7	0	0	0	0	3	1	10	11	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
denovo8	1	7	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Blautia; s__
denovo9	0	0	0	1	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae
denovo10	1	0	0	2	0	1	1	0	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Desulfovibrionales; f__Desulfovibrionaceae; g__; s__
denovo11	0	0	0	0	0	0	0	3	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__[Tissierellaceae]; g__Finegoldia; s__
denovo12	0	0	0	0	0	0	0	1	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales
denovo13	0	0	0	0	0	1	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae
denovo14	12	13	6	13	121	58	1	12	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Dialister; s__
denovo15	30	16	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae
denovo16	0	0	0	1	0	0	0	0	k__Bacteria; p__Firmicutes; c__Bacilli
denovo17	8	4	0	3	1	0	1	2	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales
denovo18	0	0	1	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales
denovo19	0	0	0	0	1	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales

>denovo0 A1_21775

TACGTAGGTGGCAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGAGATCAAGTCGGCTGTGACAACTACAGGCTTAACTTGTAGACTGCGGTCGAAACTGGTTTTCTTGAGTGAAGTATAGG

Human Microbiome Project (HMP)



National Institutes of Health (NIH) Common Fund supported

- Phase I (HMP1): established in 2008
- Phase II (iHMP): ongoing

Metagenomics of the Human Intestinal Tract (MetaHIT) — 11/25 —

- to establish associations between the genes of the human intestinal microbiota and our health and disease
- focus on two disorders of increasing importance in Europe, Inflammatory Bowel Disease (IBD) and obesity

Components of a Typical Statistical Analysis

- Quality control
  - Filtering of OTUs and samples
  - Library size and Rarefaction
- Exploratory analysis
  - Relative abundance (e.g., heatmap, painter plot)
  - Alpha diversity (e.g., boxplot)
  - Beta diversity (e.g., PCoA)
- Global testing
  - Compare the overall microbiome composition across different clinical groups
- OTU-based testing
  - Detect differentially abundant OTUs across different clinical groups

Quality Control – Filtering

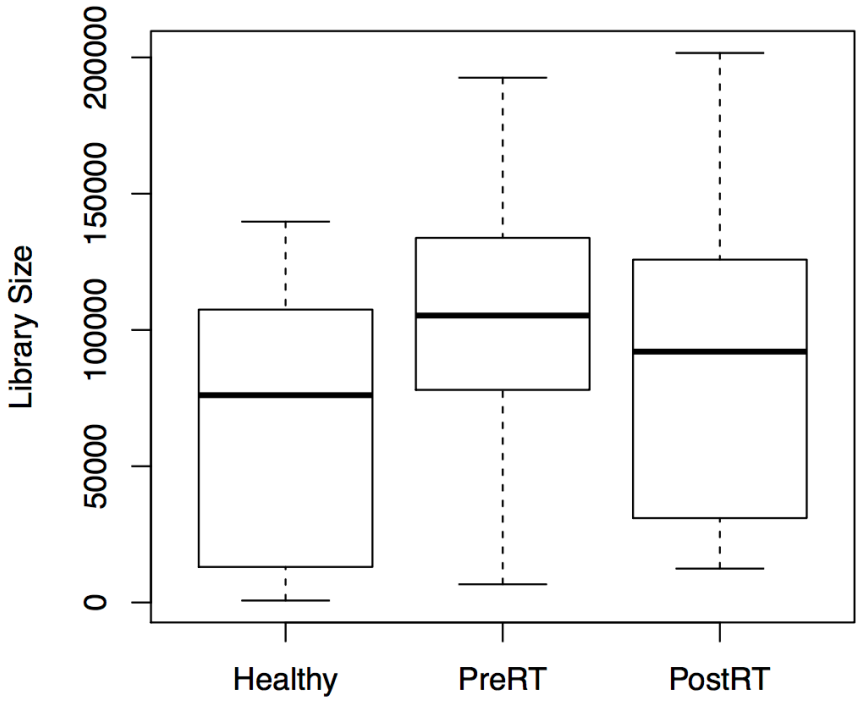
There is no gold standard yet!

For example:

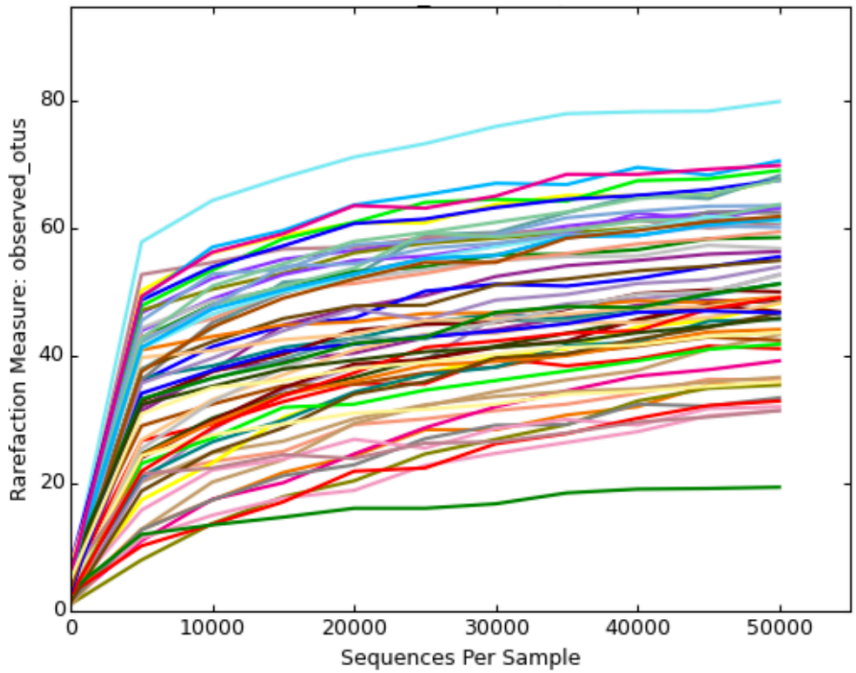
- OTU filter:
  - remove singletons (i.e., exist only in one sample)
  - remove OTUs that are present in $< 10\%$ of samples
  - remove OTUs with relative abundance $< 0.5\%$
- Sample filter
  - remove samples with < 500 sequencing reads

Quality Control – Library Size

Library Size

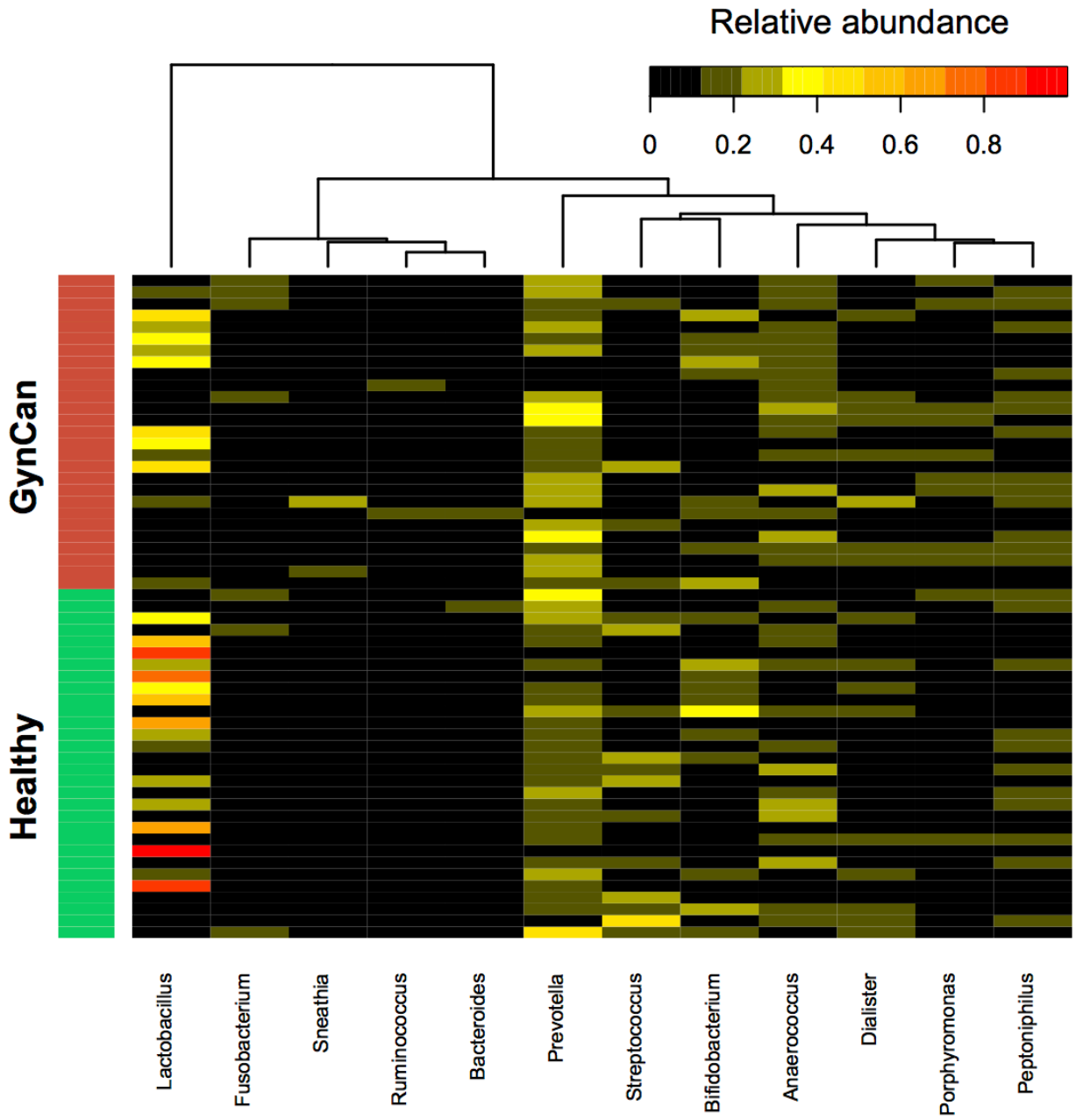


Rarefaction

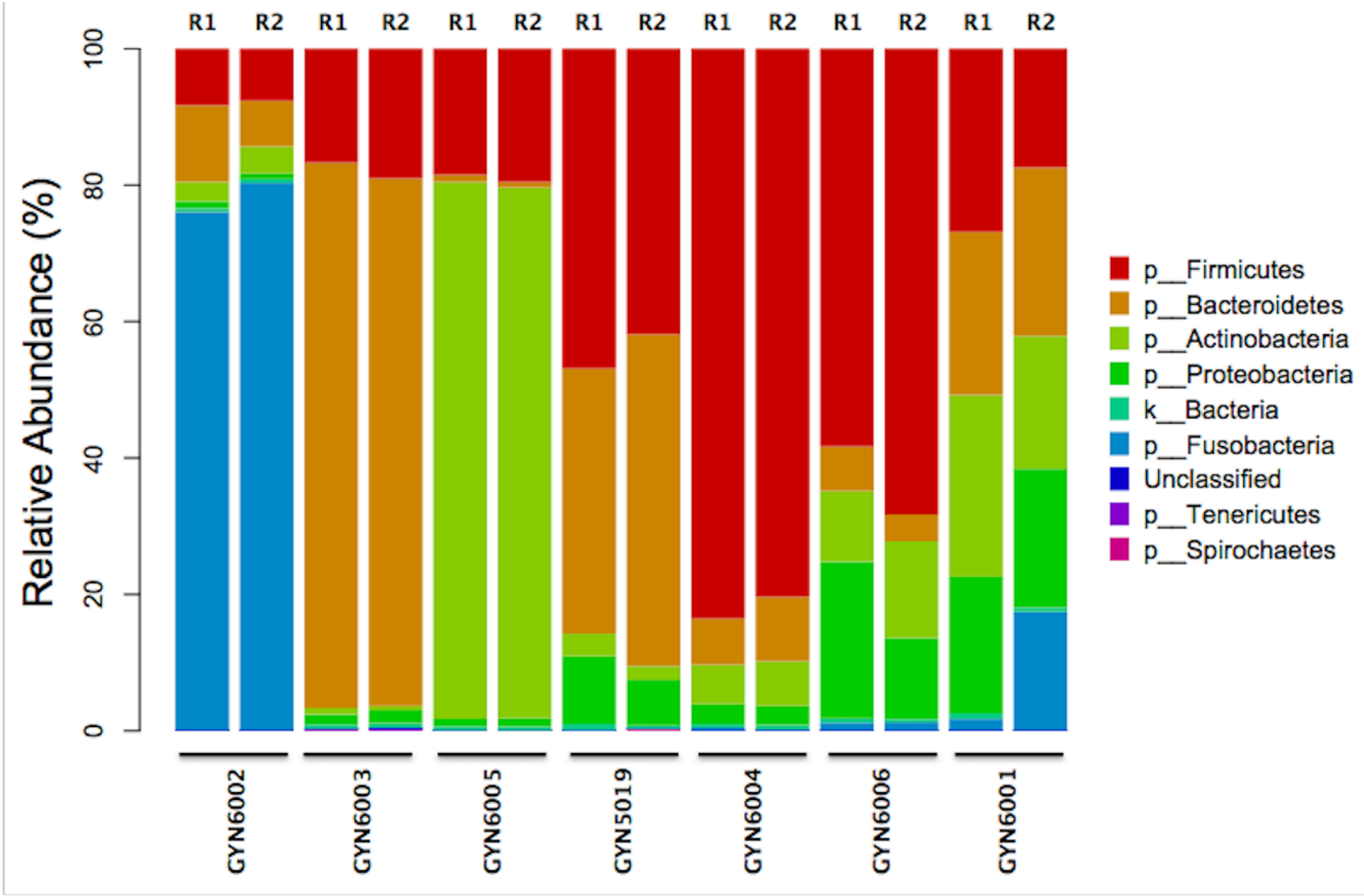


- Sequencing experiments lead to an arbitrary total number of sequence reads per sample (**library size**); strong batch effect on library size
- Uneven library size is a strong confounder for microbiome analysis
- Rarefaction curves are used to determine the library size that all samples are rarefied to.

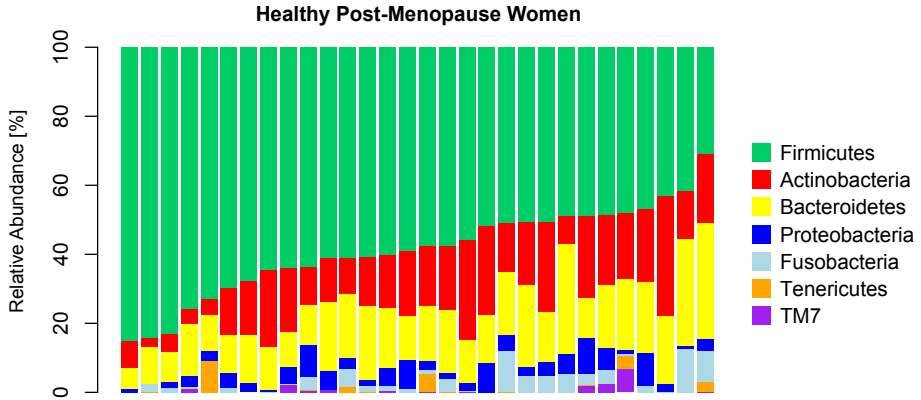
Exploratory Analysis – rel. abundance (heatmap)

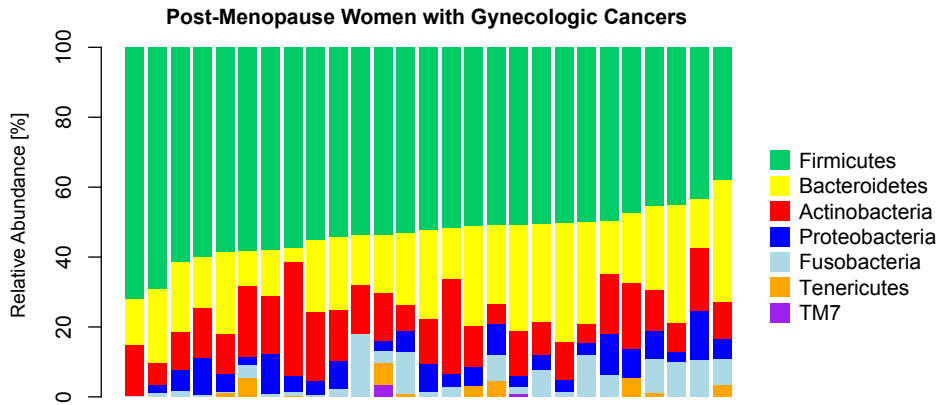


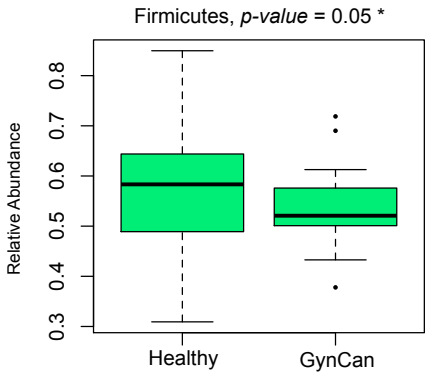
Exploratory Analysis – rel. abundance (painter plot)

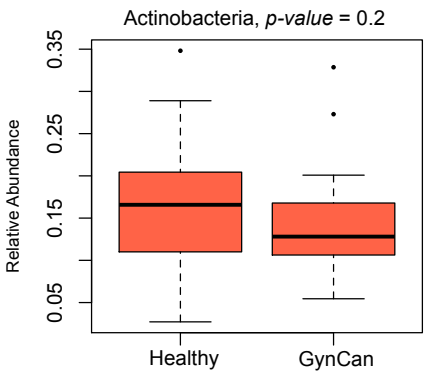


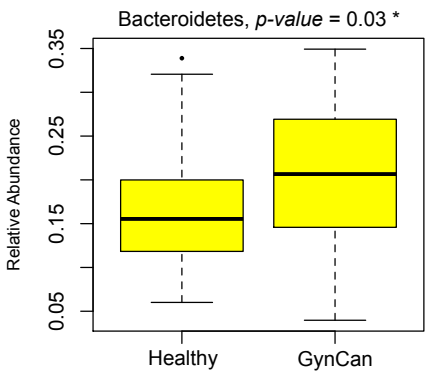
Exploratory Analysis – rel. abundance

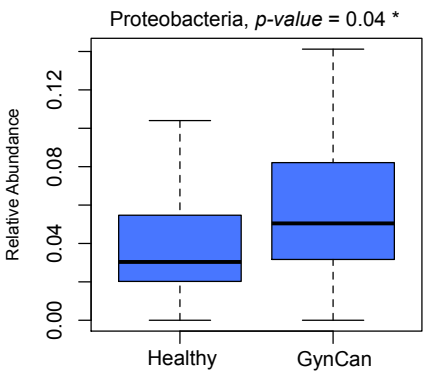










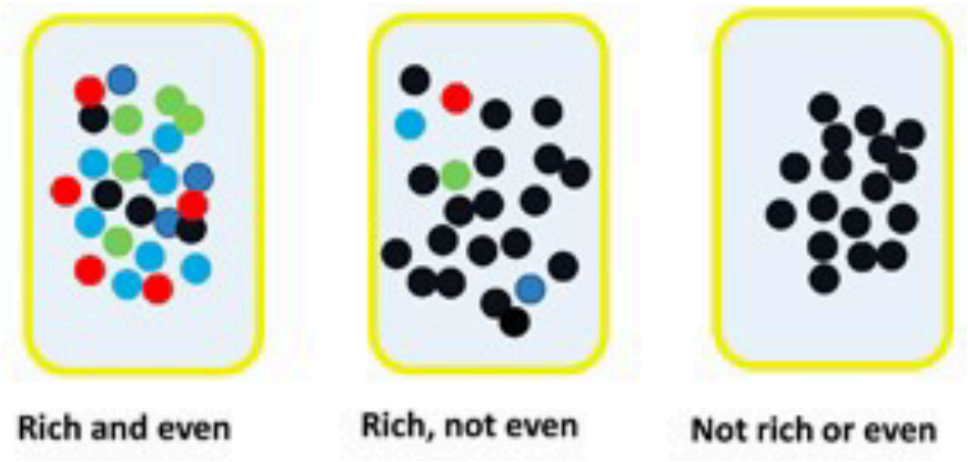


- Wilcoxon rank-sum test (nonparametric) for comparing two groups
- Krustal-Wallis test (nonparametric) for comparing more than two groups

Exploratory Analysis – Alpha Diversity

Alpha diversity: measure the diversity within a sample

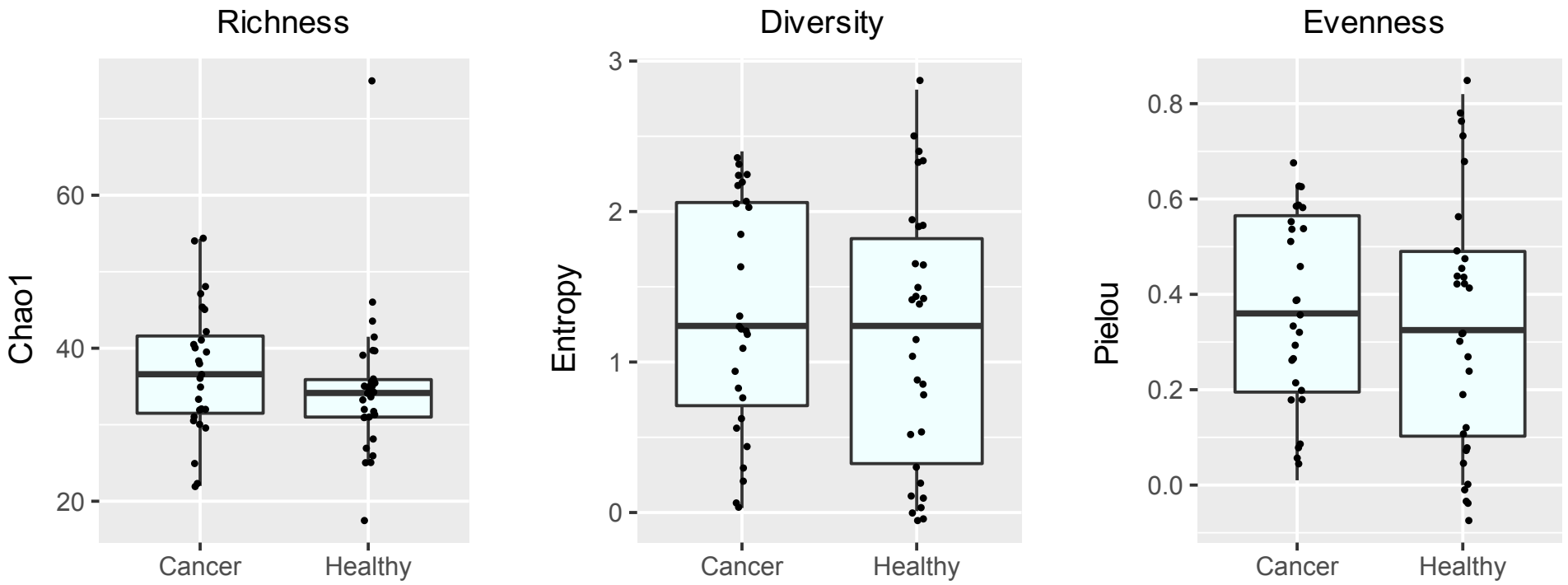
- Richness: a measure of number of species present in a sample
- Evenness: distribution of different microbes



Common used alpha diversity metrics

- **Observed species**: measure richness only, S_{obs}
- **Chao1**: measure richness only, $S_{\text{obs}} + (1/2)S^2_{\text{singleton}}/S_{\text{doubleton}}$
- **Shannon**: measures richness and evenness, $H' = -\sum_{j=1}^{J} p_j \ln p_j$
- **Pielou**: measure evenness only, $H'/H'_{\max} = H'/\ln S_{\text{obs}}$

Exploratory Analysis – Alpha Diversity



- Wilcoxon rank-sum test (nonparametric) for comparing two groups
- Krustal-Wallis test (nonparametric) for comparing more than two groups

Exploratory Analysis – Beta Diversity

Beta diversity: measure the distance or dissimilarity between each sample pair
$\Rightarrow$ distance/dissimilarity matrix

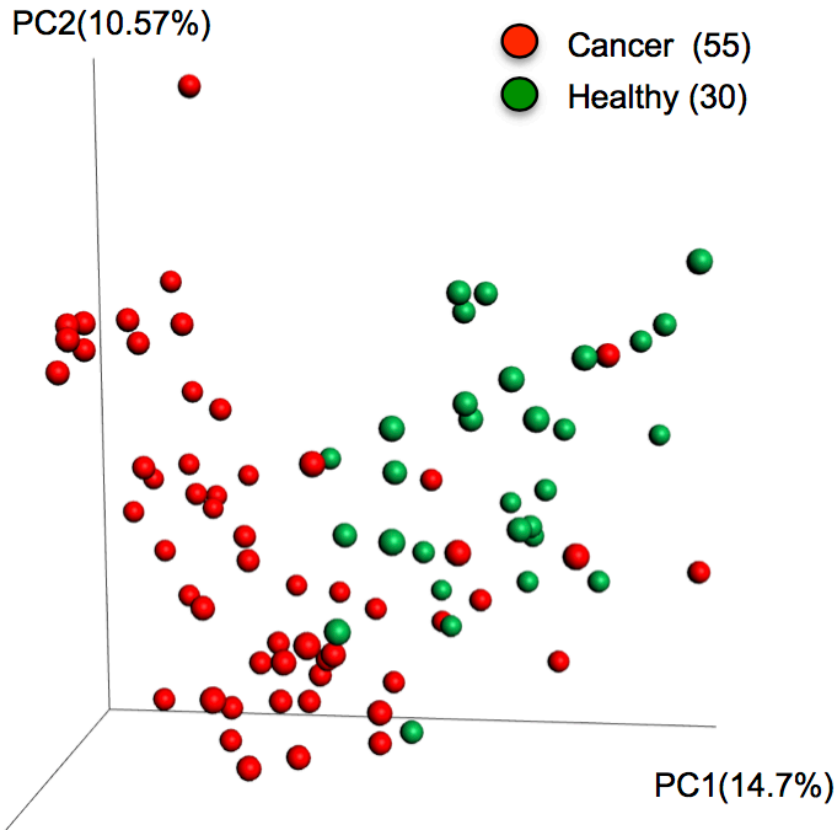
Common used beta diversity metrics

- Non-phylogeny based
  - **Bray-Curtis**: based on abundance
    - based on absolute abundance n_{ij}, $b_{ii'} = \frac{\sum_{j=1}^{J} |n_{ij} - n_{i'j}|}{n_{i+} + n_{i'+}}$
    - based on relative abundance p_{ij}, $b_{ii'} = \sum_{j=1}^{J} |p_{ij} - p_{i'j}|$
  - **Jaccard**: based on presence-absence
    - : $J_{ii'} = \frac{|S_i \bigcap S_{i'}|}{|S_i \bigcup S_{i'}|}$, S_i is the set of present OTUs in sample i
- Phylogeny based (b_j the length of branch j)
  - **Weighted UniFrac**: based on abundance, $d_{W,ii'} = \frac{\sum_{j=1}^{J} b_j |p_{ij} - p_{i'j}|}{\sum_{j=1}^{J} b_j (p_{ij} + p_{i'j})}$
  - **Unweighted UniFrac**: based on presence-absence, $d_{U,ii'} = \frac{\sum_{j=1}^{J} b_j |I(p_{ij} > 0) - I(p_{i'j} > 0)|}{\sum_{j=1}^{J} b_j}$

Exploratory Analysis – Beta Diversity (PCoA plot)

Principal Coordinates Analysis (PCoA) can be used for visualization of the data present in the beta diversity distance matrix in the form of 2-dimensional or 3-dimentional plots known as PCoA plots.

Perform eigen-decomposition of a pre-specified distance matrix and obtain eigenvectors (PC1, PC2, ...)



Statistical hypothesis: the microbiome compositions are different in the healthy and in the diseased group

PERMANOVA (Permutation-based ANOVA): based on a pre-specified distance matrix $(d_{ii'})$

- Square of distance matrix: $A = (a_{ii'})$, where $a_{ii'} = -\frac{1}{2}d_{ii'}^2$
- Gower standardization: $G = \left(I - \frac{11'}{n}\right) A \left(I - \frac{11'}{n}\right)$
- Hat matrix of the design matrix X: $H = X(X^T X)^{-1} X^T$
- The pseudo-F statistic (m covariates and n samples):

$$F = \frac{\text{tr}(HGH)/(m-1)}{\text{tr}[(I-H)G(I-H)]/(n-m)}$$

- The significance of the pseudo-F statistics is assessed based on permutations

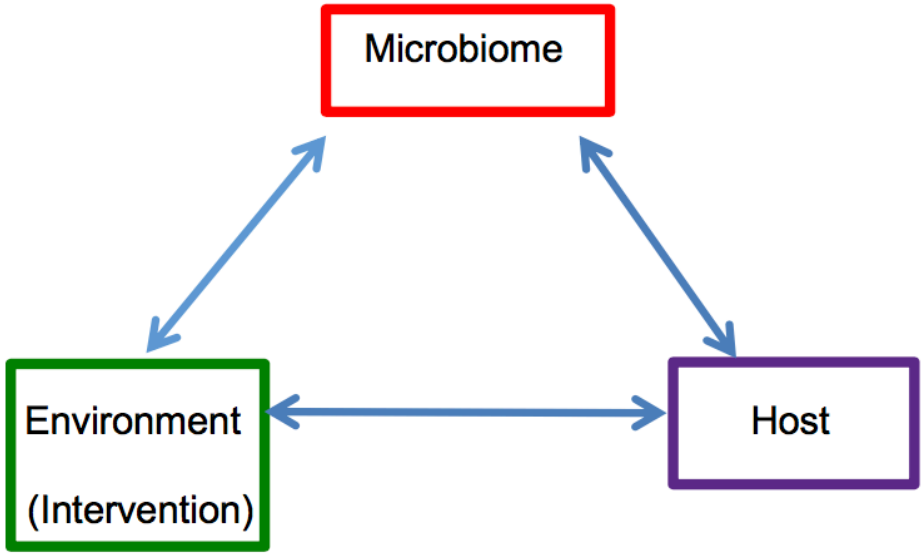
OTU-based Testing

There is no gold standard!

- **DESeq2** (Love et al., 2014)
  - Developed for detecting differentially expressed genes using RNA-seq data
  - Normalization for gene expression data (non-sparse data)
  - Assume Negative-Binomial model
- **MetagenomeSeq** (Paulson et al., 2013)
  - Developed for detecting differentially abundant OTUs using 16S sequencing data
  - Normalization accounts for sparse data
  - Assume a zero-inflated Gaussian (ZIG) distribution mixture model
- **ANCOM** (ANalysis of Composition Of Microbiomes, Mandal et al., 2015)
  - Developed for detecting differentially abundant OTUs using 16S sequencing data
  - make no distributional assumptions; use log-ratios

Advanced Statistical Analyses

- Analysis of paired, clustered, or longitudinal data
- Adjustment of confounders (e.g., gender, ancestry)
- Adjustment of batch effects (e.g., library size)
- Causal inference
  - Randomized clinical trails
  - Mediation analysis



- Network analyses identify co-varied OTUs

Study Design and Power

- To control batch effects
  - Randomization
  - Use control samples with known composition in each batch
  - Replicate some samples across sequencing batches
- Paired sample designs will increase power
- Longitudinal design help reveal dynamics or even causality