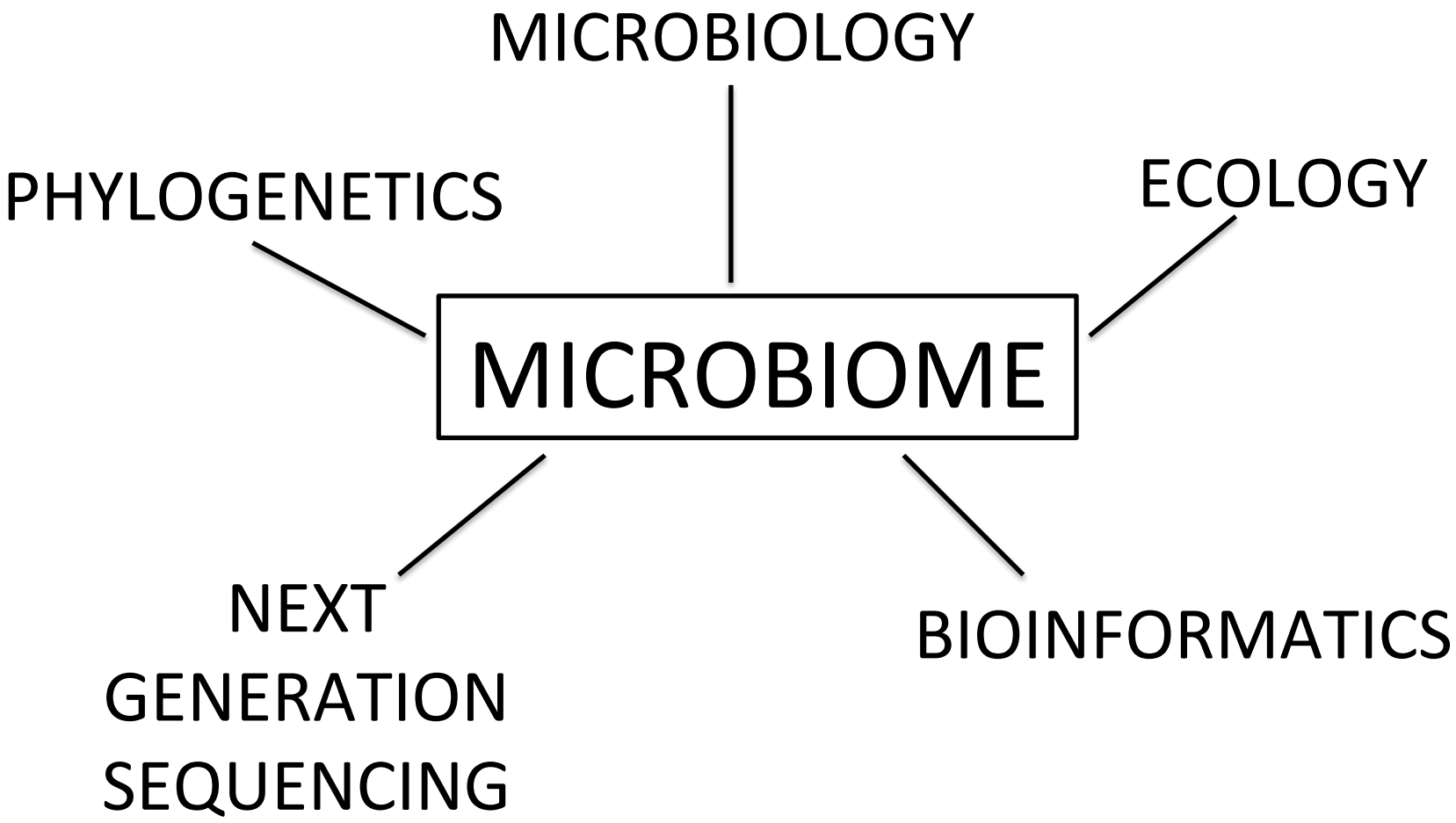
Microbiome 101

Tracy Meiring
Medical Virology
University of Cape Town
Institute of Infectious Diseases and Molecular Medicine

13-18 Sept 2015
Hermanus
South Africa

FOURTH ADVANCED SUMMER SCHOOL IN AFRICA

THE ROLE OF THE MICROBIOME IN DISEASE


MICROBIOLOGY
PHYLOGENETICS
ECOLOGY
MICROBIOME
NEXT GENERATION SEQUENCING
BIOINFORMATICS

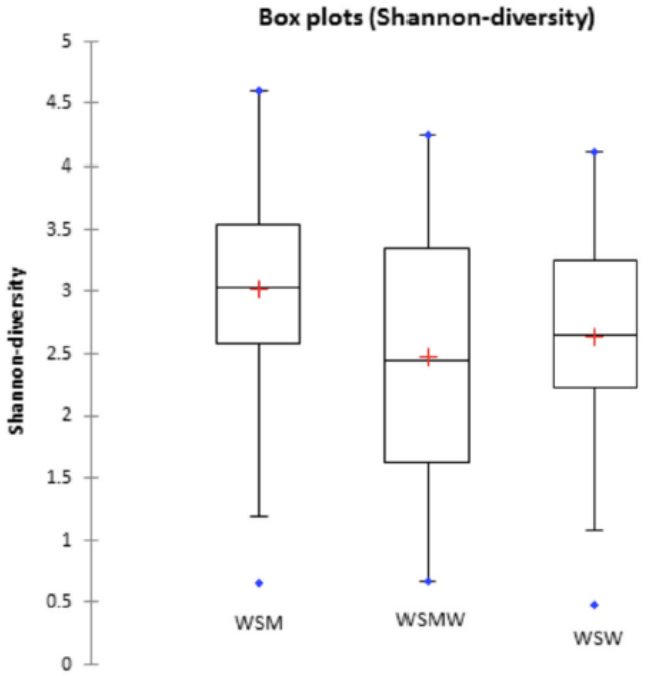


Figure 5. Box plot of Shannon's diversity for each sexual risk behavior group. The red plus signs depict the average value of Shannon's diversity for each sexual risk behavior group; blue dots represent outliers. The WSM group had the highest diversity compared to the other sexual risk behavior groups.

doi: 10.1371/journal.pone.0080254.g005

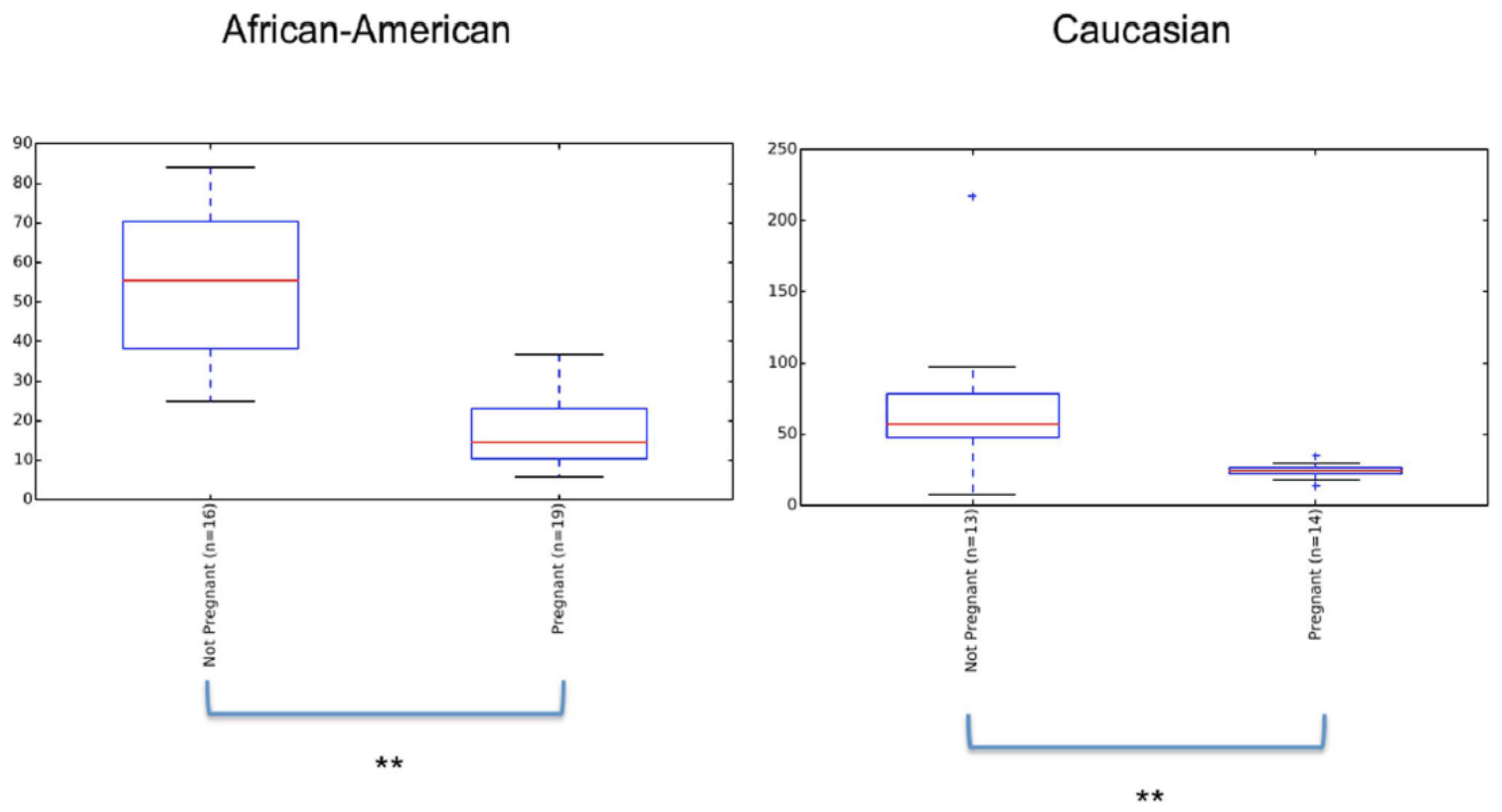


Figure 4. Pregnancy effect in African-American and Caucasian subjects as measured by Chao1 diversity Index. Diversity is significantly reduced during pregnancy in both ethnicities (**p<0.01, Monte Carlo analyses, 999 permutations).
doi:10.1371/journal.pone.0098514.g004

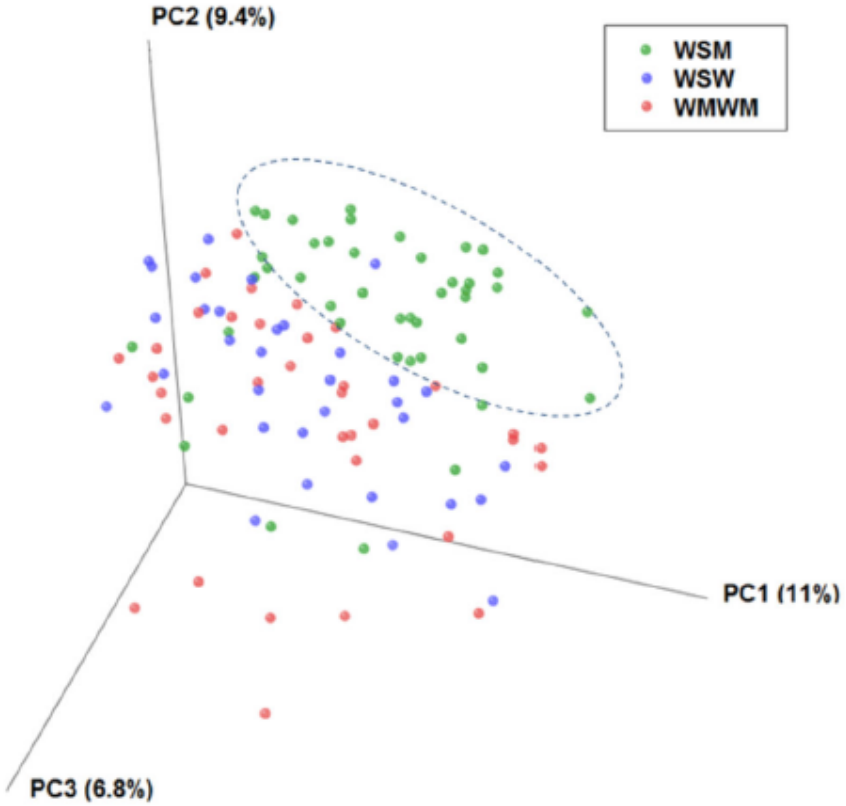
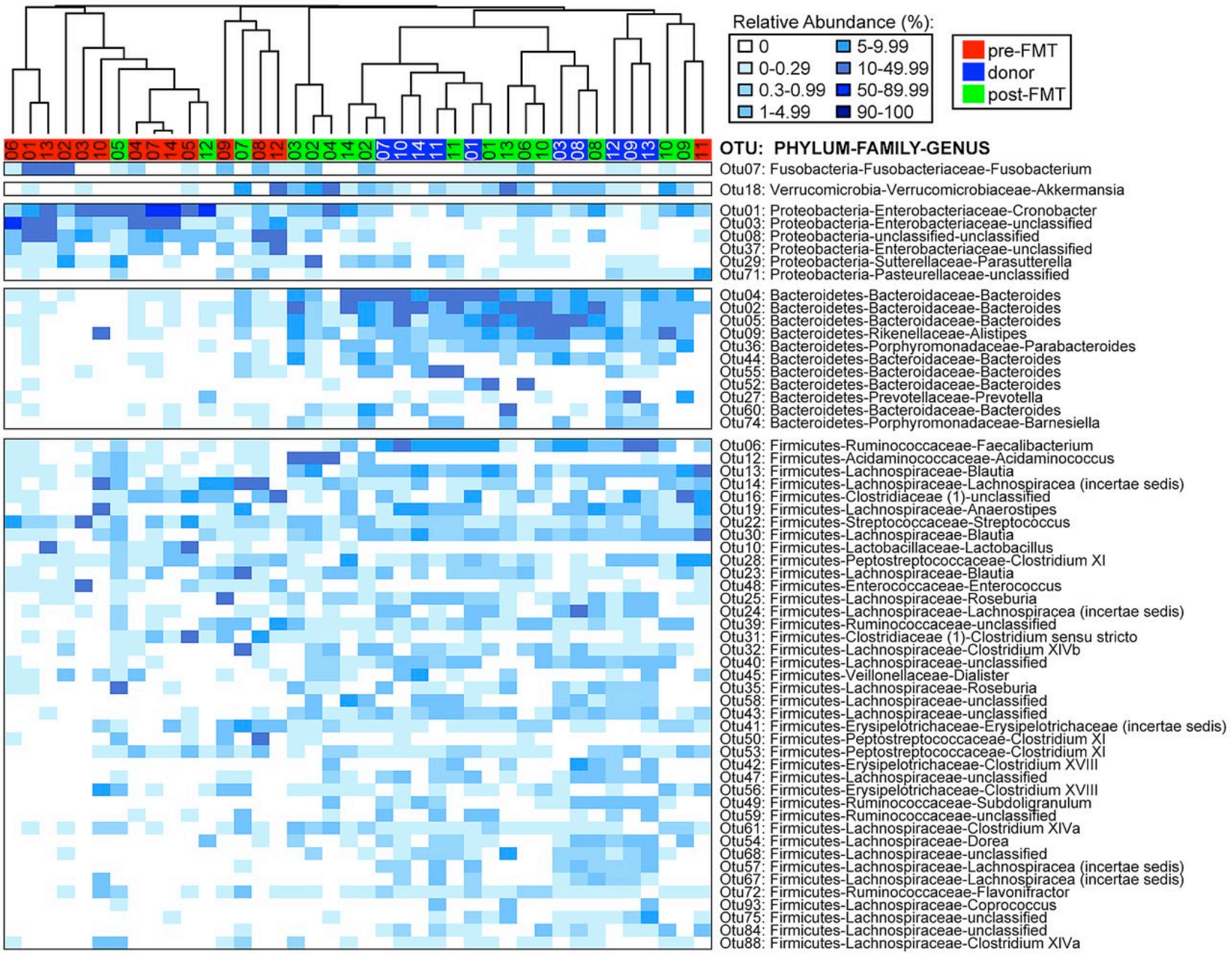


Figure 6. Unweighted beta-diversity Unifrac analysis. This figure depicts a distinct cluster within the WSM group (indicated in green) that is not present within the WSW or WSWM groups (indicated in blue and red, respectively).

doi: 10.1371/journal.pone.0080254.g006

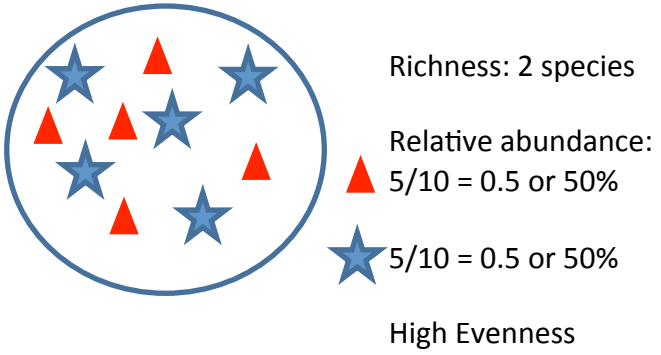

Relative Abundance (%):
0
0-0.29
0.3-0.99
1-4.99
5-9.99
10-49.99
50-89.99
90-100
pre-FMT
donor
post-FMT
OTU: PHYLUM-FAMILY-GENUS
Otu07: Fusobacteria-Fusobacteriaceae-Fusobacterium
Otu18: Verrucomicrobia-Verrucomicrobiaceae-Akkermansia
Otu01: Proteobacteria-Enterobacteriaceae-Cronobacter
Otu03: Proteobacteria-Enterobacteriaceae-unclassified
Otu08: Proteobacteria-unclassified-unclassified
Otu37: Proteobacteria-Enterobacteriaceae-unclassified
Otu29: Proteobacteria-Sutterellaceae-Parasutterella
Otu71: Proteobacteria-Pasteurellaceae-unclassified
Otu04: Bacteroidetes-Bacteroidaceae-Bacteroides
Otu02: Bacteroidetes-Bacteroidaceae-Bacteroides
Otu05: Bacteroidetes-Bacteroidaceae-Bacteroides
Otu09: Bacteroidetes-Rikenellaceae-Alistipes
Otu36: Bacteroidetes-Porphyromonadaceae-Parabacteroides
Otu44: Bacteroidetes-Bacteroidaceae-Bacteroides
Otu55: Bacteroidetes-Bacteroidaceae-Bacteroides
Otu52: Bacteroidetes-Bacteroidaceae-Bacteroides
Otu27: Bacteroidetes-Prevotellaceae-Prevotella
Otu60: Bacteroidetes-Bacteroidaceae-Bacteroides
Otu74: Bacteroidetes-Porphyromonadaceae-Barnesiella
Otu06: Firmicutes-Ruminococcaceae-Faecalibacterium
Otu12: Firmicutes-Acidaminococcaceae-Acidaminococcus
Otu13: Firmicutes-Lachnospiraceae-Blautia
Otu14: Firmicutes-Lachnospiraceae-Lachnospiracea (incertae sedis)
Otu16: Firmicutes-Clostridiaceae (1)-unclassified
Otu19: Firmicutes-Lachnospiraceae-Anaerostipes
Otu22: Firmicutes-Streptococcaceae-Streptococcus
Otu30: Firmicutes-Lachnospiraceae-Blautia
Otu10: Firmicutes-Lactobacillaceae-Lactobacillus
Otu28: Firmicutes-Peptostreptococcaceae-Clostridium XI
Otu23: Firmicutes-Lachnospiraceae-Blautia
Otu48: Firmicutes-Enterococcaceae-Enterococcus
Otu25: Firmicutes-Lachnospiraceae-Roseburia
Otu24: Firmicutes-Lachnospiraceae-Lachnospiracea (incertae sedis)
Otu39: Firmicutes-Ruminococcaceae-unclassified
Otu31: Firmicutes-Clostridiaceae (1)-Clostridium sensu stricto
Otu32: Firmicutes-Lachnospiraceae-Clostridium XIVb
Otu40: Firmicutes-Lachnospiraceae-unclassified
Otu45: Firmicutes-Veillonellaceae-Dialister
Otu35: Firmicutes-Lachnospiraceae-Roseburia
Otu58: Firmicutes-Lachnospiraceae-unclassified
Otu43: Firmicutes-Lachnospiraceae-unclassified
Otu41: Firmicutes-Erysipelotrichaceae-Erysipelotrichaceae (incertae sedis)
Otu50: Firmicutes-Peptostreptococcaceae-Clostridium XI
Otu53: Firmicutes-Peptostreptococcaceae-Clostridium XI
Otu42: Firmicutes-Erysipelotrichaceae-Clostridium XVIII
Otu47: Firmicutes-Lachnospiraceae-unclassified
Otu56: Firmicutes-Erysipelotrichaceae-Clostridium XVIII
Otu49: Firmicutes-Ruminococcaceae-Subdoligranulum
Otu59: Firmicutes-Ruminococcaceae-unclassified
Otu61: Firmicutes-Lachnospiraceae-Clostridium XIVa
Otu54: Firmicutes-Lachnospiraceae-Dorea
Otu68: Firmicutes-Lachnospiraceae-unclassified
Otu57: Firmicutes-Lachnospiraceae-Lachnospiracea (incertae sedis)
Otu67: Firmicutes-Lachnospiraceae-Lachnospiracea (incertae sedis)
Otu72: Firmicutes-Ruminococcaceae-Flavonifractor
Otu93: Firmicutes-Lachnospiraceae-Coprococcus
Otu75: Firmicutes-Lachnospiraceae-unclassified
Otu84: Firmicutes-Lachnospiraceae-unclassified
Otu88: Firmicutes-Lachnospiraceae-Clostridium XIVa

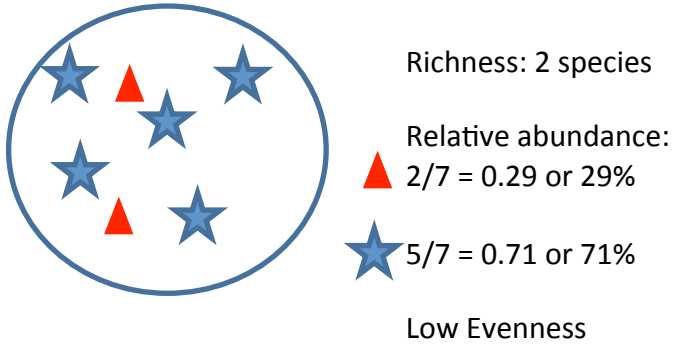
- How do we describe microbiomes?
- How do we compare microbiomes?

Important ecological concepts

- How is biodiversity defined and measured?
- Component of biodiversity:
  - RICHNESS
  - EVENNESS
- Species richness: number of different species in a habitat/sample
- Species relative abundance: number of each species relative to total number of all species in a sample (number of reads per OTU in a sample relative to total number of reads in that sample)
- Species evenness: how close in numbers each species in an environment are; distribution

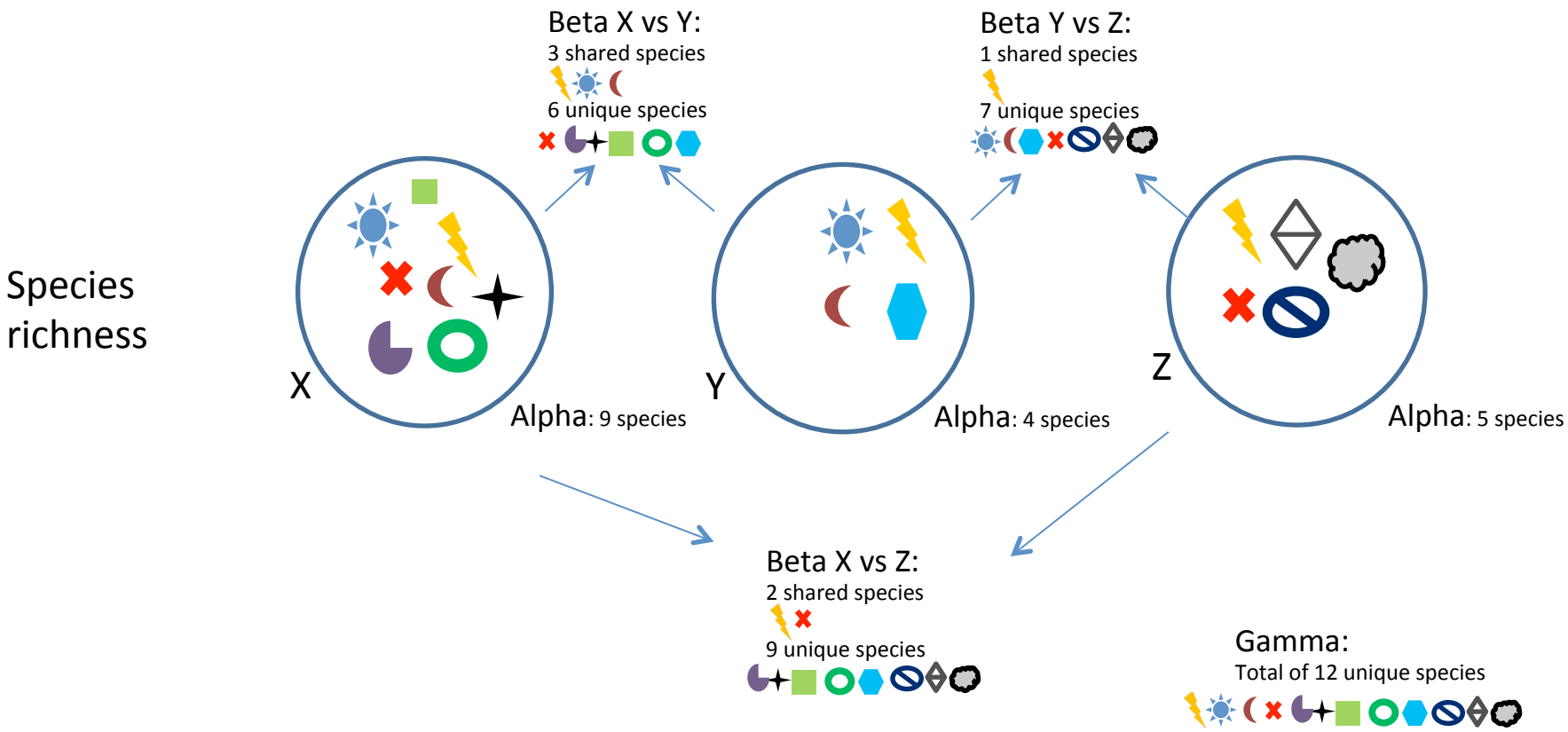
Simple example:





- Alpha diversity: Diversity within a single sample (Alone)
- Beta diversity: Diversity Between samples
- Gamma diversity: total diversity in a landscape

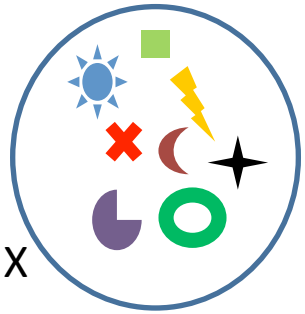
SIMPLE EXAMPLE

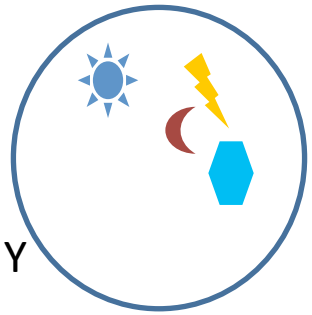


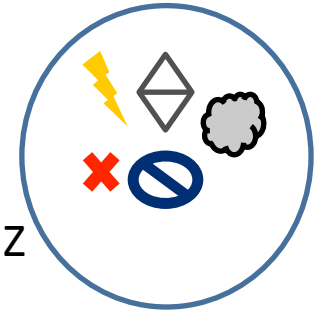
(modified from http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209%28Composition&Diversity%29/9_2_Biodiversity.htm)

PROBLEM: Doesn't take abundance of each species OR relatedness of each species into account

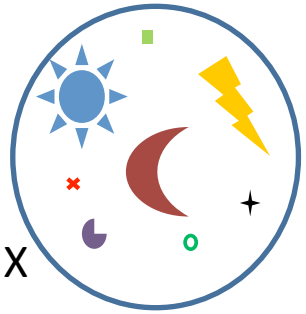
Species richness

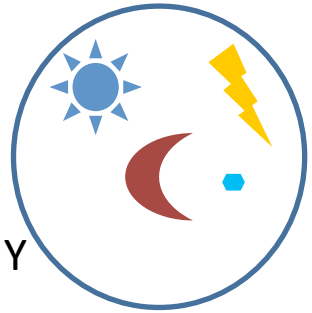


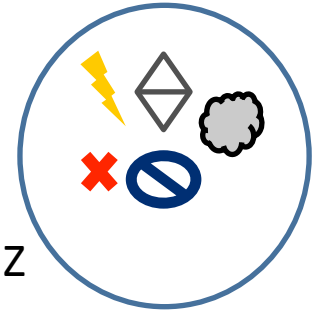




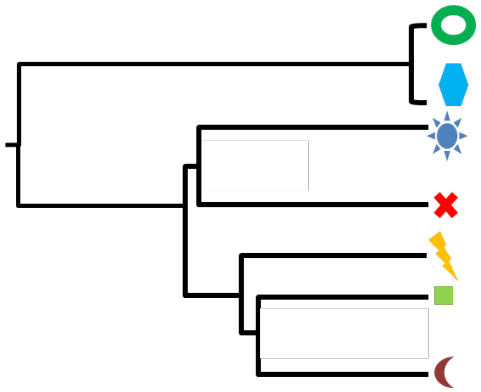
Size adjusted according to abundance







Phylogenetic relationship



Metrics used to describe diversity measure different aspects of the community

Rarefaction

50 individuals

2 species

250 individuals

4 species

Proc. R. Soc. Lond. B (2002) **269**, 2401–2405
DOI 10.1098/rspb.2002.2116

500 individuals

8 species

Felix Borner / AP

Rarefaction

- Collector's curves
- NGS: individuals = reads
- Evaluate sample size: is sequencing depth(reads per sample) enough?
- Comparing the richness and diversity observed in different samples
- Note rarefaction is not the same as rarefying

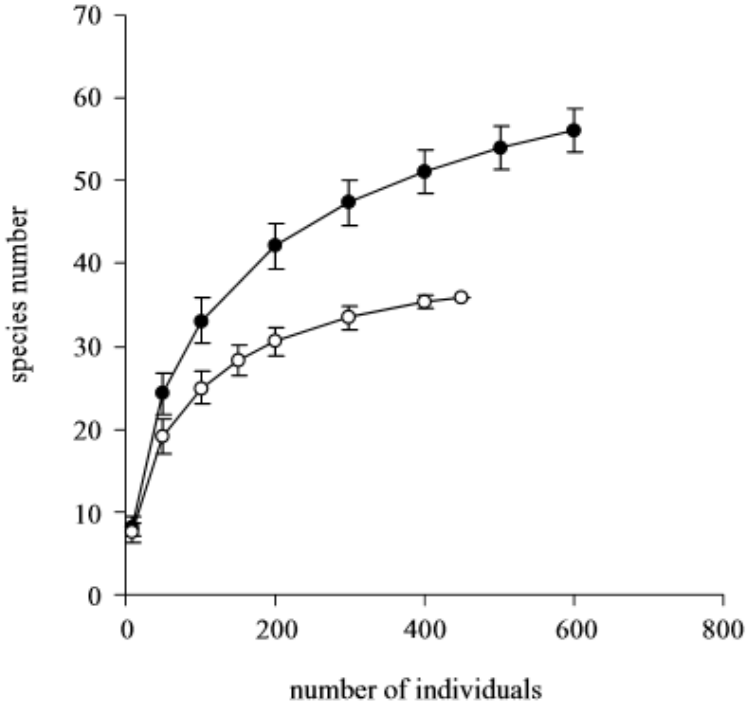
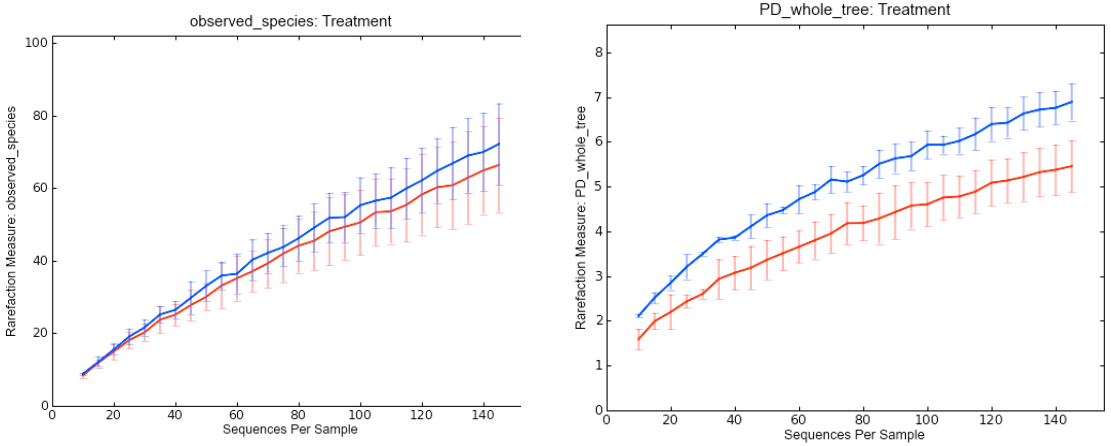
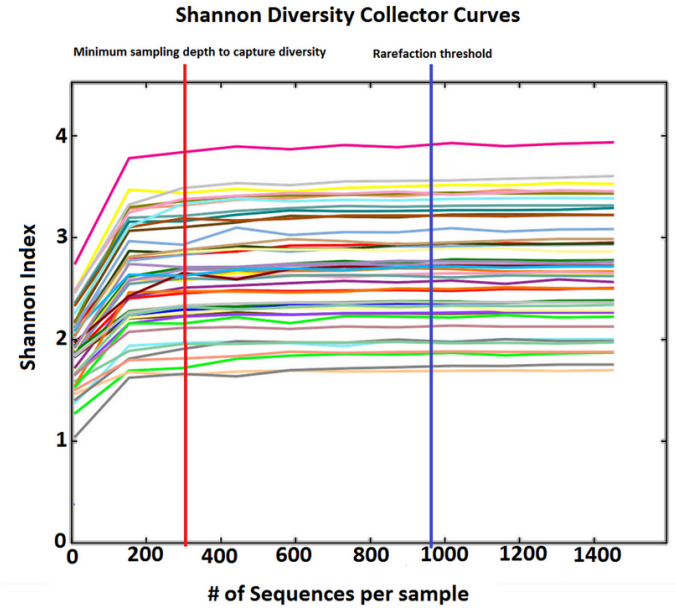


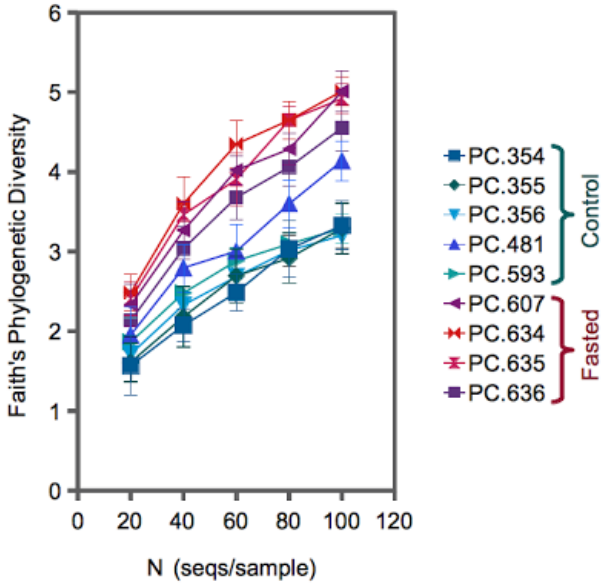
Figure 2. Rarefaction curves for numbers of species of ground dwellers in native savannah (filled circles) and in adjacent agriculture (open circles). For the same sample of individuals (450) there were 30% fewer species in agriculture. The curve for agriculture is approaching an asymptote while the savannah curve climbed to 73 species. Vertical lines indicate one s.d.

Alpha diversity rarefaction QIIME tutorial and some examples



http://qiime.org/tutorials/tutorial.html





Cardenas PA, Cooper PJ, Cox MJ, Chico M, Arias C, et al. (2012) Upper Airways Microbiota in Antibiotic-Naıve Wheezing and Healthy Infants from the Tropics of Rural Ecuador. PLoS ONE 7(10): e46803. doi:10.1371/journal.p

http://www.wernerlab.org/teaching/qiime/overview/e

Alpha metrics

- Richness: observed species, chao1
- Diversity (Richness and Abundance / Evenness): Shannon, Simpson
- Phylogenetic: PD, Faith's PD or PD_whole tree

Beta diversity

- Diversity between samples
- Single metric to describe difference or similarity between samples
- Non-phylogenetic metrics
- Phylogenetic metrics

Phylogenetic beta diversity: UniFrac distance

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2005, p. 8228–8235 Vol. 71, No. 12
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.12.8228–8235.2005

UniFrac: a New Phylogenetic Method for Comparing Microbial Communities

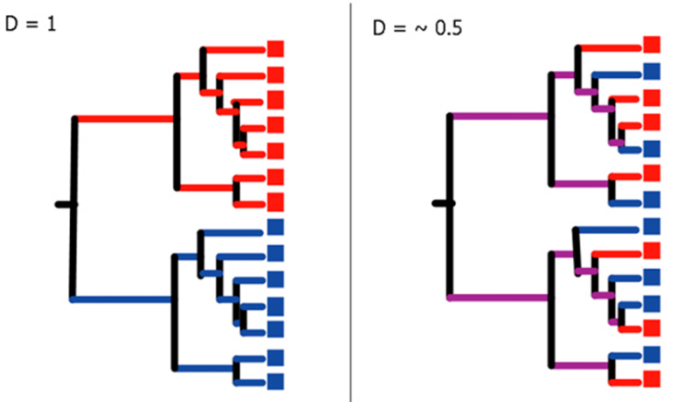
Catherine Lozupone[1] and Rob Knight[2]*

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado 80309,[1] and Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309[2]

UNIFRAC help: http://bmf.colorado.edu/unifrac/help.psp

UNIFRAC help: http://bmf.colorado.edu/unifrac/help.psp

- Raw unweighted Unifrac: sum of branch length that is unique to one environment or the other

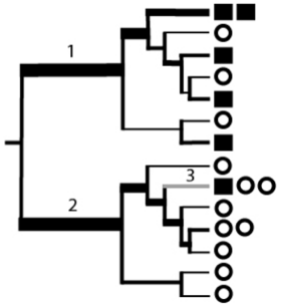


$$u = \frac{\sum_{i=1}^{N} l_i |A_i - B_i|}{\sum_{i=1}^{N} l_i \max(A_i, B_i)}$$

l_i is the branch length between node i and its parent, and A*i* and B*i* are indicators equal to 0 or 1 as descendants of node i are absent or present in communities A and B respectively

A = red, B= blue, branches in common are purple, branches unique to A are red and unique to B are blue. Presence/absence metric.

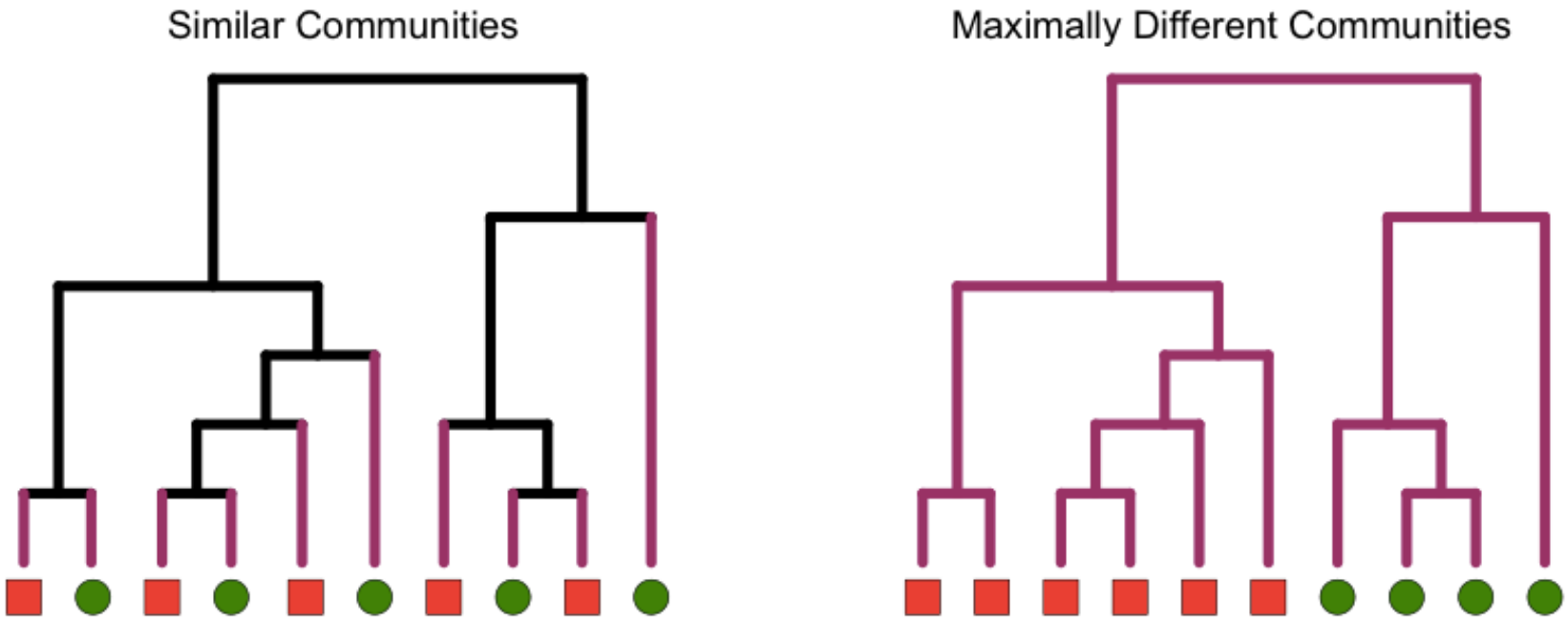
- Raw weighted Unifrac: Branch lengths are weighted by the relative abundance of sequences



$$u = \sum_{i}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

Here, n is the total number of branches in the tree, b_i is the length of branch i, A_i and B_i are the number of descendants of branch i from communities A and B respectively, and A_T and B_T are the total number of sequences from communities A and B respectively. In order to control for unequal sampling effort, A_i and B_i are divided by A_T and B_T.

- Normalised weighted Unifrac: takes abundance and normalises branch length
  - Rapidly evolving lineages (with long branch length can skew unifrac)



UniFrac Distance Measure = (—) / (— + —)

Branch length of unique branches

Branch length of common branches

Unweighted Unifrac does not take abundance into account
Weighted Unifrac takes abundance into account – branch lengths weighted by relative abundance

Completely genetically different communities D=1
Exactly same communities D=0

Principal Coordinate Analysis (PCoA or PCO)

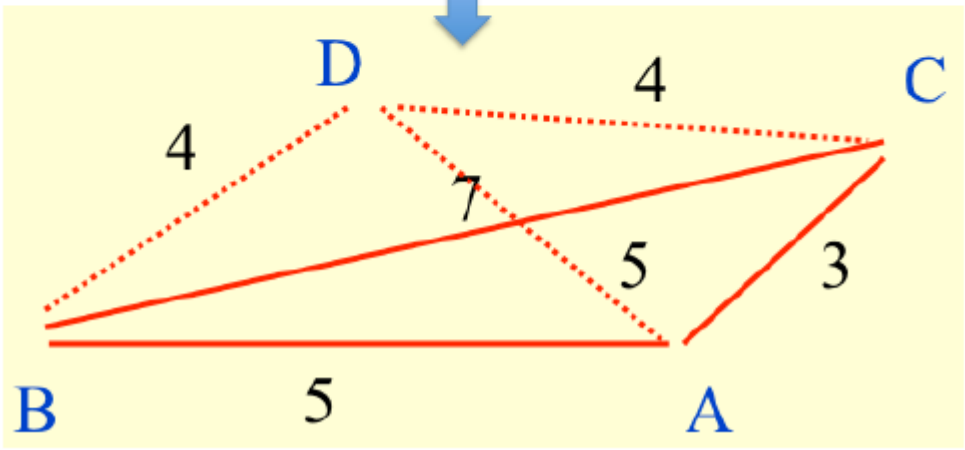
- Also sometimes called classical MDS (multidimensional scaling)
- Can use any distance matrix (must obey triangle inequality), in this case the Unifrac distance matrix
- Assumes linear relation
- represent distance between samples graphically in multidimensional space (n-1 dimension, n = number samples)

- A new set of reduced variables is derived from the original distances and used to scale samples
- Samples now represented on 2D or 3D plot with these new variables as axes and the relationship between the sample on the plot should reflect their underlying distance

- Ordinates data on plot so that axis 1 (PC1) explains the greatest amount of variance, axis 2 (PC2) explains the next greatest amount of variance, etc.

Principal Coordinates Analysis

	A	B	C	D
A	0	5	3	5
B		0	7	4
C			0	4
D				0

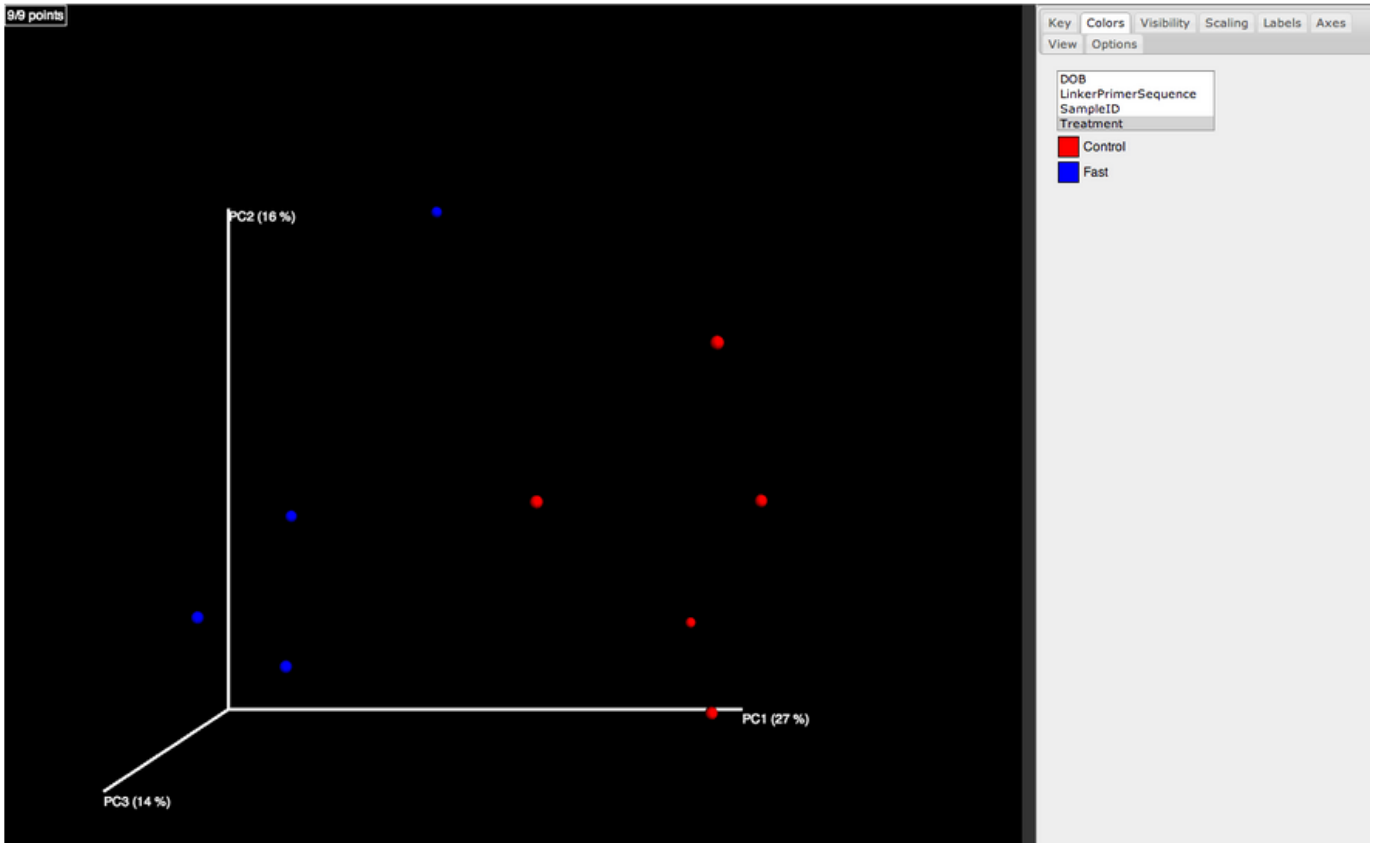


Adapted from H. Whitehead

NESCent QIIME Tutorial

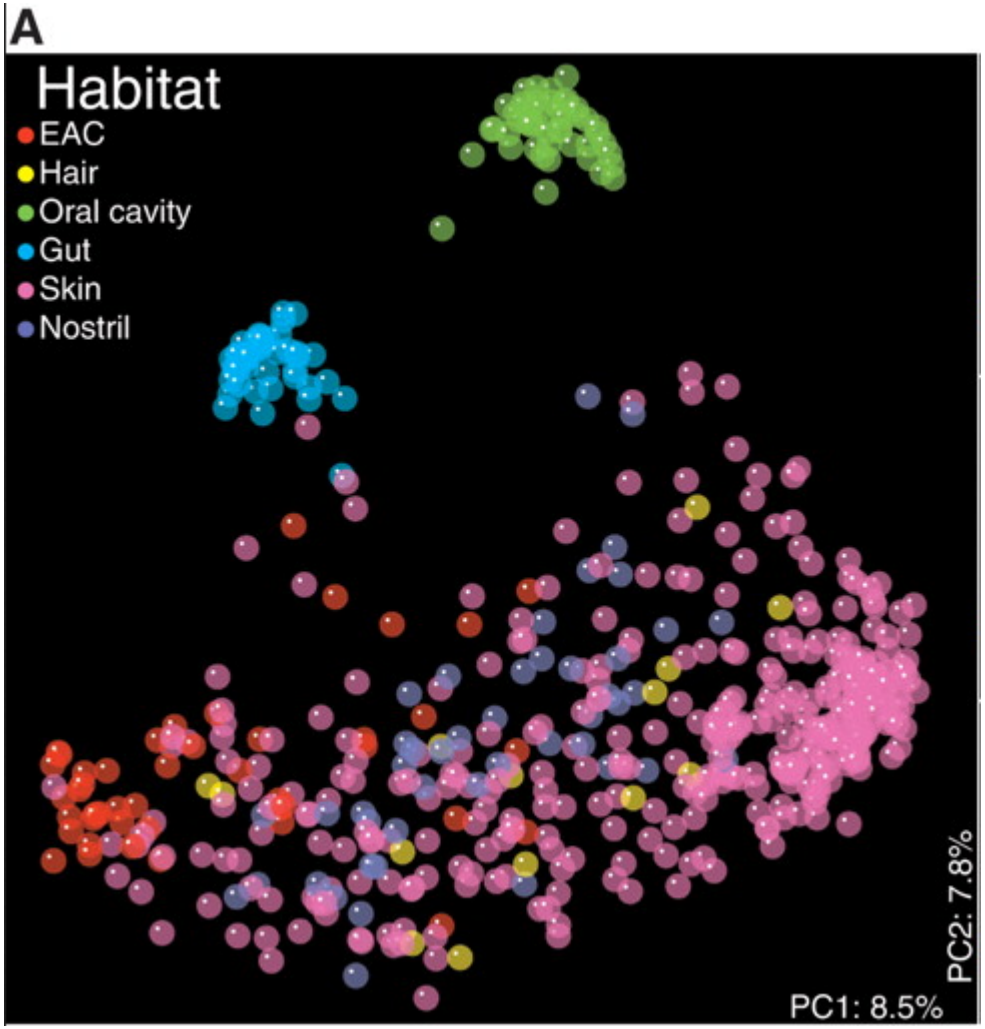
Jose Clemente
Daniel McDonald
6.12.12

https://www.nescent.org/sites/academy/File:Nescent_qiime_tutorial_june2012.pdf



QIIME tutorial data “Five control samples are all red and the four Fast samples are all blue. This lets you easily visualize “clustering” by metadata category. The 3d visualization software allows you to rotate the axes to see the data from different perspectives.” http://qiime.org/tutorials/tutorial.html

qiime
Quantitative Insights Into Microbial Ecology



Communities clustered using PCoA of the unweighted UniFrac distance matrix

Science 18 December 2009:
Vol. 326 no. 5960 pp. 1694-1697
DOI: 10.1126/science.1177486

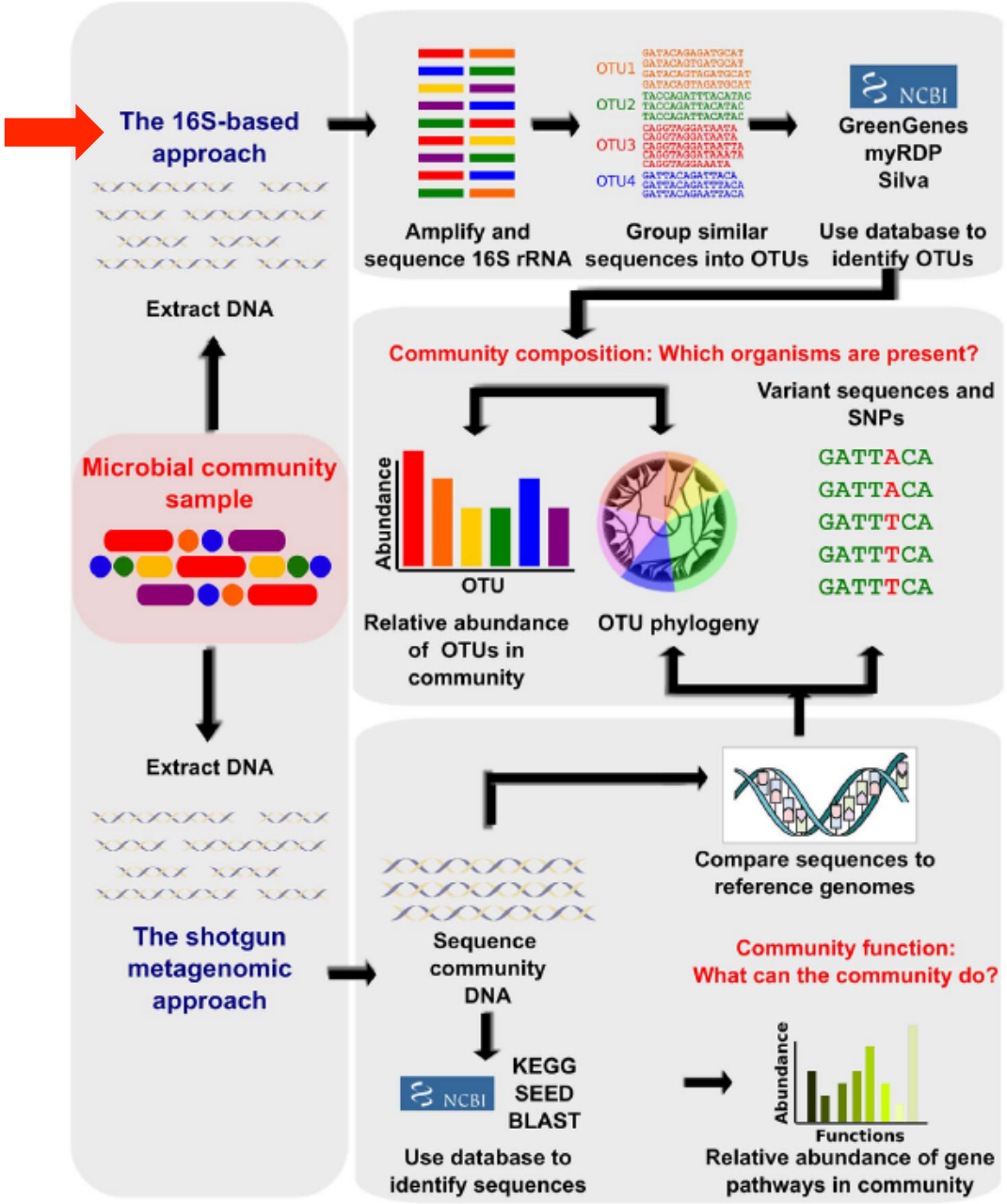


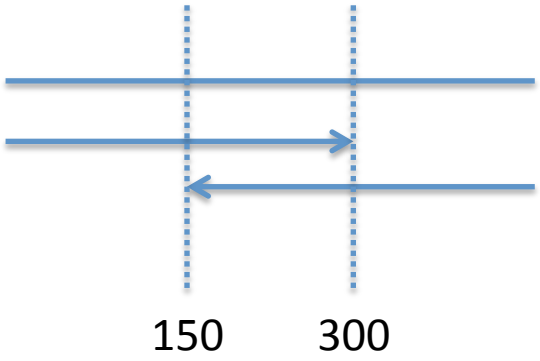
Figure 1. Bioinformatic methods for functional metagenomics. Studies that aim to define the composition and function of uncultured microbial communities are often referred to collectively as "metagenomic," although this refers more specifically to particular sequencing-based assays. First, community DNA is extracted from a sample, typically uncultured, containing multiple microbial members. The bacterial taxa present in

Which 16S region(/s)?

- Which sequencing platform?

illumina

16S V3-V4



Max Output	Max Read Number	Max Read Length
15 Gb	25 M	2 x 300 bp

Max Output	Max Read Number	Max Read Length
1500 Gb	5000 M	2 x 150 bp

ion torrent
Sequencing for all.™

Max Output	Max Read Number	Max Read Length
15 Gb	**25 M**	**2 x 300 bp**

Ion PGM™ System Performance Specifications				
		Ion 314™ Chip v2	Ion 316™ Chip v2	Ion 318™ Chip v2
Output*	200 base 400 base	30-50 Mb 60-100 Mb	300-600 Mb 600 Mb-1 Gb	600 Mb-1 Gb 1.2-2 Gb
Reads		400-550 thousand	2-3 million	4-5.5 million
Run time	200 base 400 base	2.3 hr 3.7 hr	3.0 hr 4.9 hr	4.4 hr 7.3 hr

PCR and sequencing controls

- Controls to detect possible contamination
- Contamination can occur at multiple stages:
  - Sample collection
  - DNA extraction
  - 16S PCR
  - Barcoding and Adaptor PCR
  - Amplicon purification
  - Sequencing

Research article

Highly accessed Open Access

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter[1]*, Michael J Cox[2], Elena M Turek[2], Szymon T Calus[3], William O Cookson[2], Miriam F Moffatt[2], Paul Turner[4,5], Julian Parkhill[1], Nicholas J Loman[3] and Alan W Walker[1,6]*

* Corresponding authors: Susannah J Salter sb18@sanger.ac.uk - Alan W Walker alan.walker@abdn.ac.uk

▾ Author Affiliations

[1] Pathogen Genomics Group, Wellcome Trust Sanger Institute, Hinxton, UK

[2] Molecular Genetics and Genomics, National Heart and Lung Institute, Imperial College London, London, UK

[3] Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK

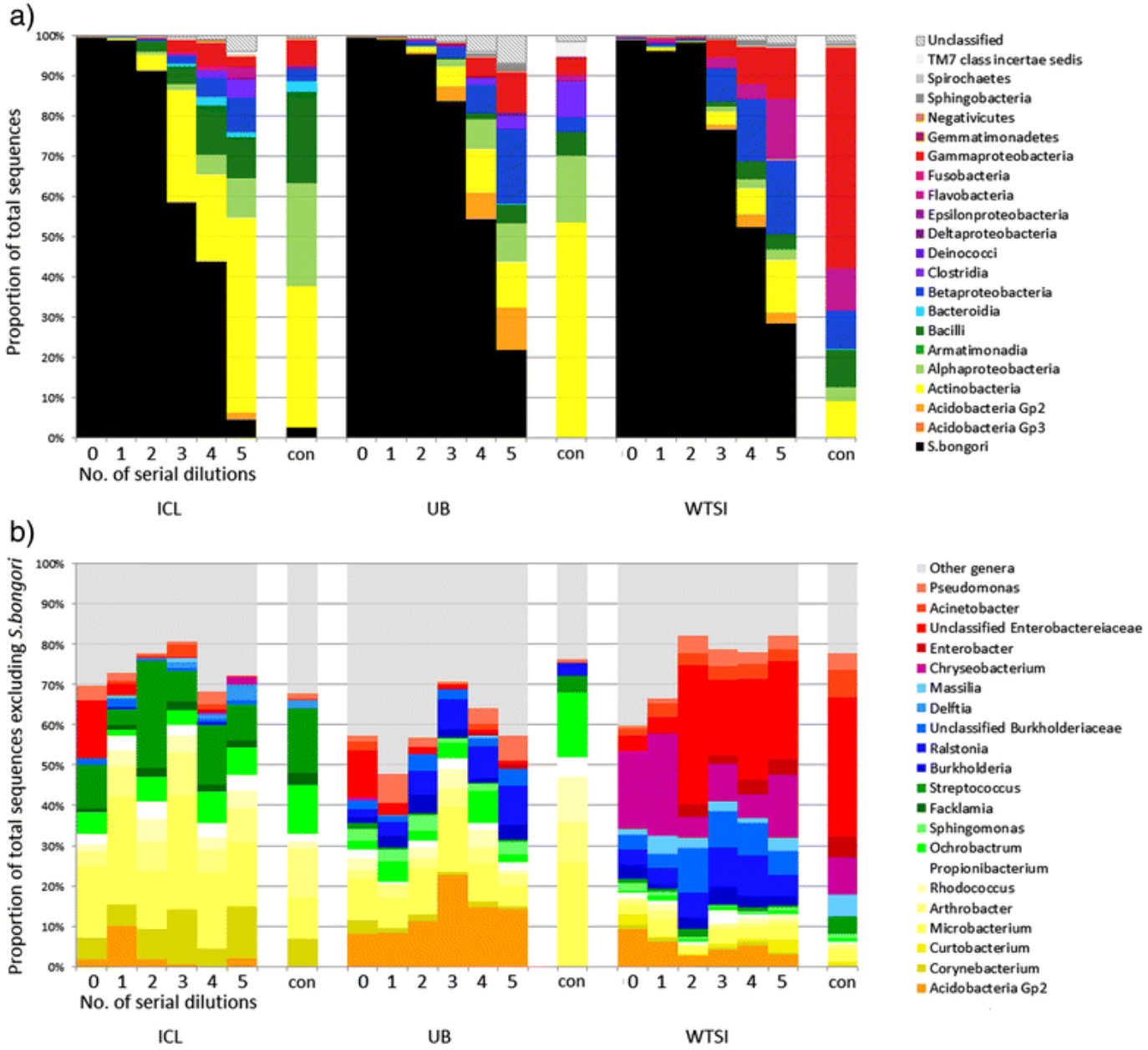
[4] Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Thailand

[5] Centre for Tropical Medicine, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[6] Microbiology Group, Rowett Institute of Nutrition and Health, University of Aberdeen, Aberdeen, UK

For all author emails, please log on.

BMC Biology 2014, **12**:87 doi:10.1186/s12915-014-0087-z



Summary of 16S rRNA gene sequencing taxonomic assignment from ten-fold diluted pure cultures and controls. Undiluted DNA extractions contained approximately 108 cells, and controls (annotated in the Figure with 'con') were template-free PCRs.

Research highlight

Highly accessed Open Access

Tracking down the sources of experimental contamination in microbiome studies

Sophie Weiss[1], **Amnon Amir**[2], **Embriette R Hyde**[2], **Jessica L Metcalf**[2], **Se Jin Song**[2] and **Rob Knight**[2,3,4]*

* Corresponding author: Rob Knight rob.knight@colorado.edu ▾ Author Affiliations

[1] Department of Chemical and Biological Engineering, University of Colorado at Boulder, Boulder 80309, CO, USA

[2] BioFrontiers Institute, University of Colorado at Boulder, Boulder 80309, CO, USA

[3] Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder 80309, CO, USA

[4] Howard Hughes Medical Institute, Boulder 80309, CO, USA

For all author emails, please log on.

Genome Biology 2014, **15**:564 doi:10.1186/s13059-014-0564-2

Positive controls

- Mock microbial communities
- BEI Resources (ATCC)

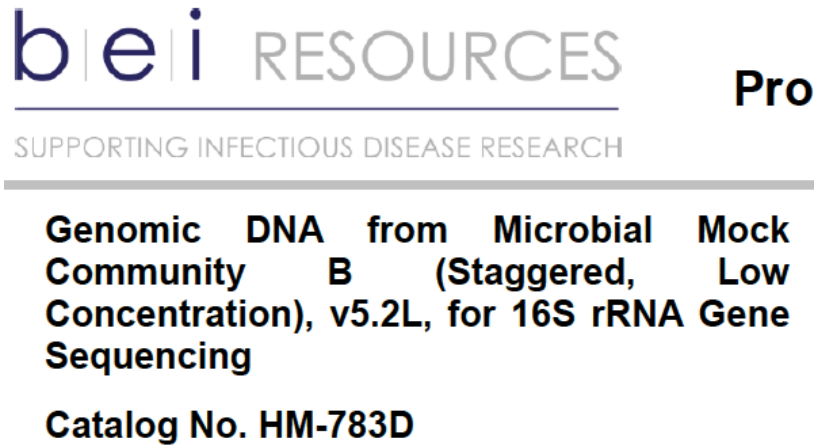
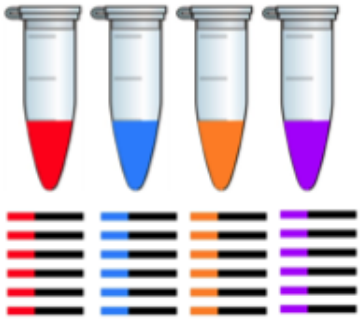


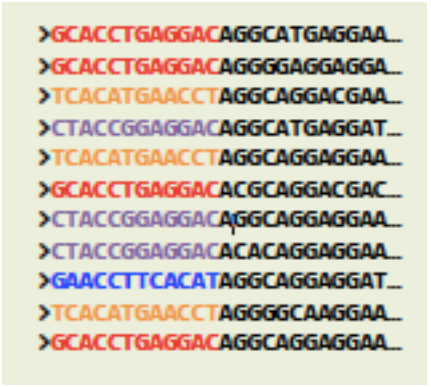
Table 1: Microbial Mock Community B

Organism	NCBI Reference Sequence
Acinetobacter baumannii, strain 5377	NC_009085
Actinomyces odontolyticus, strain 1A.21	NZ_AAYI02000000
Bacillus cereus, strain NRS 248	NC_003909
Bacteroides vulgatus, strain ATCC® 8482™	NC_009614
Clostridium beijerinckii, strain NCIMB 8052	NC_009617
Deinococcus radiodurans, strain R1 (smooth)	NC_001263, NC_001264
Enterococcus faecalis, strain OG1RF	NC_17316
Escherichia coli, strain K12, substrain MG1655	NC_000913
Helicobacter pylori, strain 26695	NC_000915
Lactobacillus gasseri, strain 63 AM	NC_008530
Listeria monocytogenes, strain EGDe	NC_003210
Neisseria meningitidis, strain MC58	NC_003112
Propionibacterium acnes, strain KPA171202	NC_006085
Pseudomonas aeruginosa, strain PAO1-LAC	NC_002516
Rhodobacter sphaeroides, strain ATH 2.4.1	NC_007493, NC_007494
Staphylococcus aureus, strain TCH1516	NC_010079
Staphylococcus epidermidis, FDA strain PCI 1200	NC_004461
Streptococcus agalactiae, strain 2603 V/R	NC_004116
Streptococcus mutans, strain UA159	NC_004350
Streptococcus pneumoniae, strain TIGR4	NC_003028



Amplify 16S region with barcoded primers

Sequencing



Demultiplex

Merge paired end reads

Quality control:
Remove primer sequences
Remove chimeras
Trim reads based on Phred score
Remove singletons (and doubletons)

High quality sequences

Associate sequence data with samples and sample metadata

Clustering reads

- Sequence clustering into OTUs (Operational Taxonomic Units)
- Pick representative sequence for each OTU

OTU picking (clustering)

- QIIME has three methods for OTU picking
  - *de novo*
  - closed-reference
  - open reference

    http://qiime.org/tutorials/otu_picking.html
- Sequences grouped together based on sequence identity (de novo) or alignment to reference sequence (closed = reads discarded if they don't match a reference, open = reads that don't hit reference form de novo cluster)

Assign taxonomy

Tools

- Blast
- RDP classifier
- UCLUST is now preferred method

16S database