Metagenome bioinformatics

“If you don't like bacteria, you're on the wrong planet.
This is the planet of the bacteria.”

Craig Venter

New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

NCGR
National Center for Genome Resources

Why study bacteria?

- They influence everything:
  - Disease…
  -
  -
  -
  -
- What microbes are in a community?
- What are their relative proportions?

Definitions

- Microbiome:
  - All of the microbes in a community
- Metagenome:
  - All of the genes in a community

What are we talking about today?

- Metagenomes
- Microbiomes
- Microbiota
- Flora

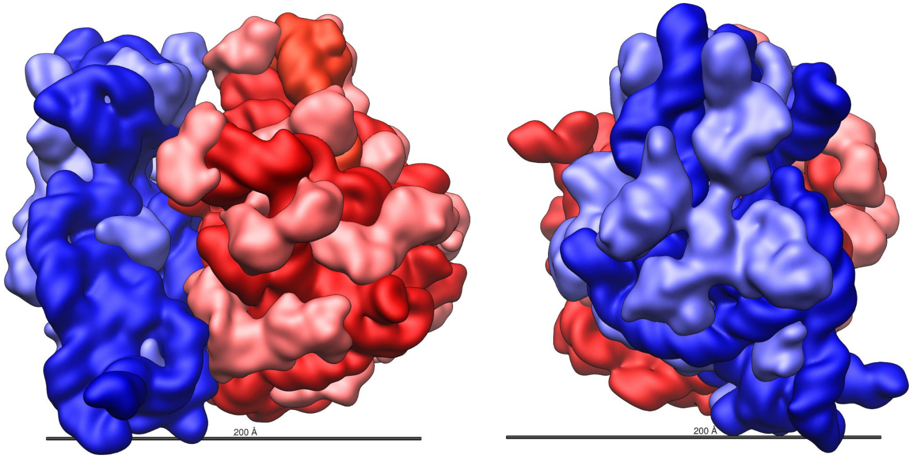
Today's goals

- Why study the microbiome?
- Basic concepts
- Methodologies
- Terminology

What do we mean by microbes?

- Bacteria
- Archaea
- Protists
- Fungi
- Viruses

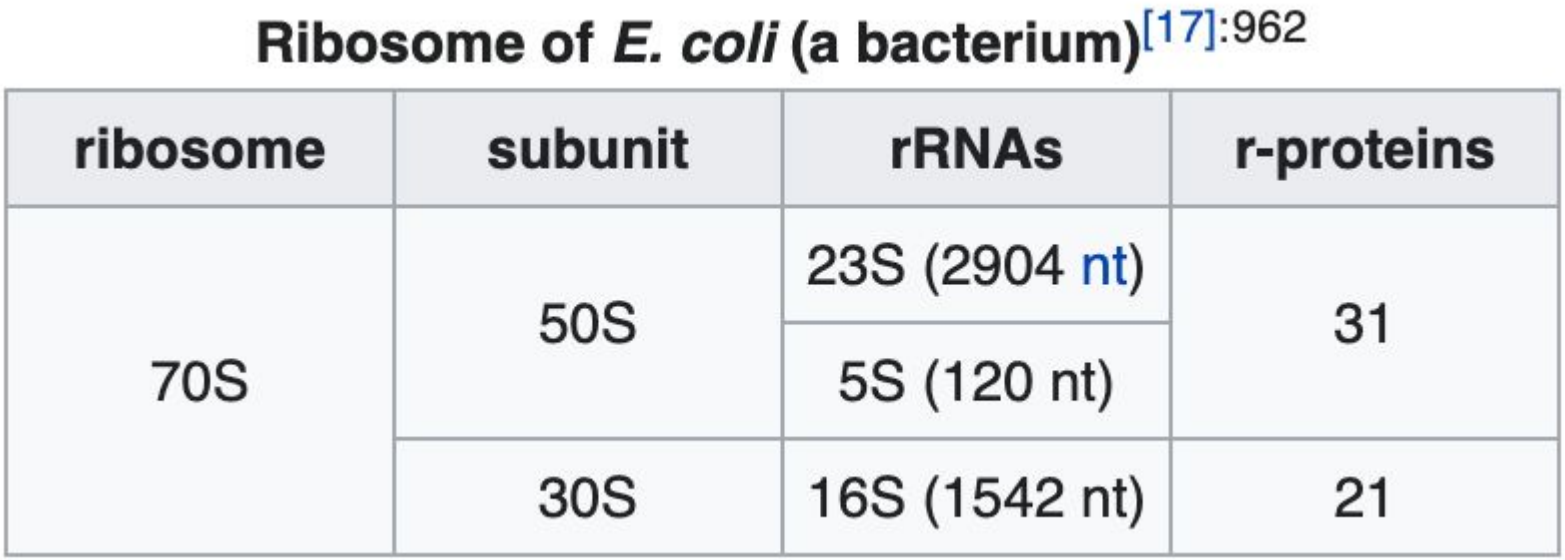
E. coli ribosome



Vossman - Own work

Structure and shape of the *E.coli* 70S ribosome. The large 50S ribosomal subunit (red) and small 30S ribosomal subunit (blue) are shown with a 200 Ångstrom (20 nm) scale bar. For the 50S subunit, the 23S (dark red) and 5S (orange red) rRNAs and the ribosomal proteins (pink) are shown. For the 30S subunit, the 16S rRNA (dark blue) and the ribosomal proteins (light blue) are shown.

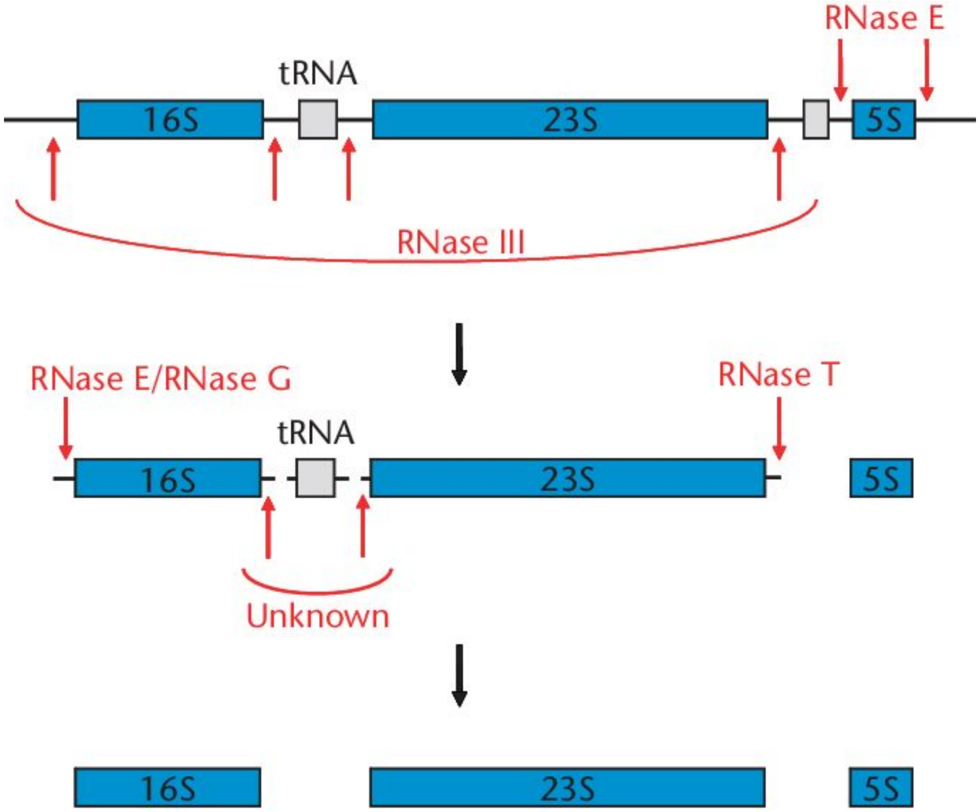
Ribosome composition



Ribosome of *E. coli* (a bacterium)[17]:962

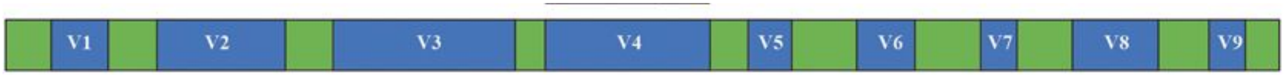
ribosome	subunit	rRNAs	r-proteins
70S	50S	23S (2904 nt)	31
		5S (120 nt)	
	30S	16S (1542 nt)	21

Bacterial rRNA processing



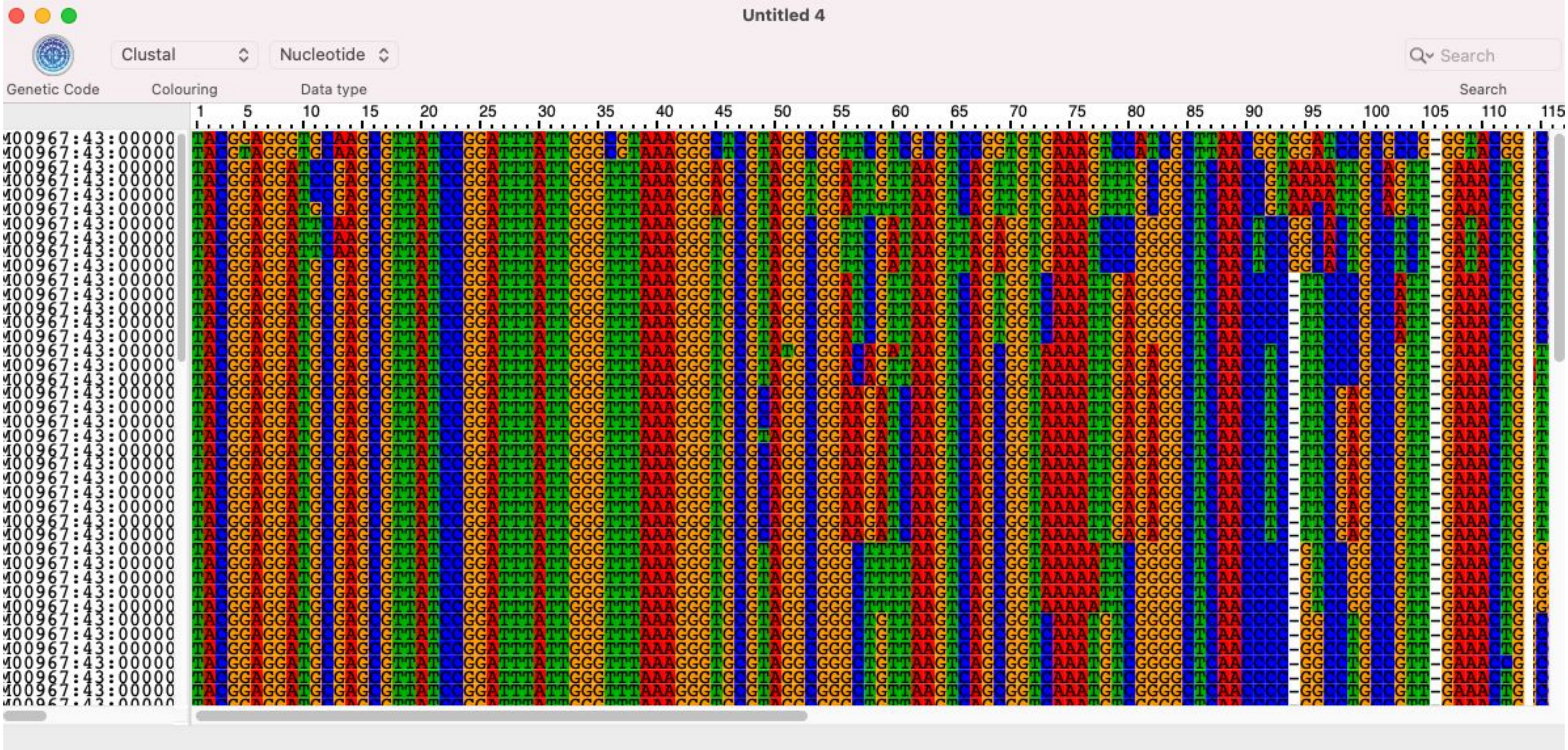
https://doi.org/10.1002/9780470015902.a0021836

16s rRNA gene sequencing



- Green sequences are common to most bacteria
- V = Variable, differ among bacteria
- Identical V sequences bong to the same bacterium
- Similar V sequences are closely related
- Dissimilar V sequences are more distantly related
- Clusters of 97% similarity generally groups genera

16s rRNA gene sequencing



Untitled 4

Colouring: Clustal

Data type: Nucleotide

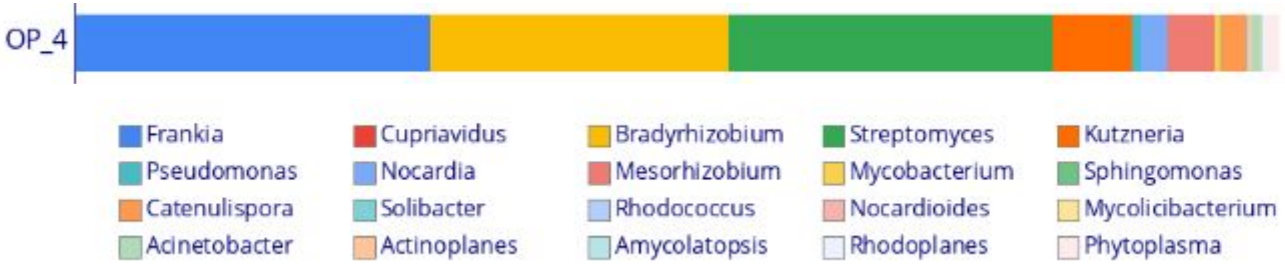
https://doi.org/10.1186/s13104-016-1927-4

New Mexico INBRE IDeA Networks of Biomedical Research Excellence

NCGR National Center for Genome Resources

The Operational Taxonomic Unit

- PCR the 16s rRNA region
- Sequence the product
- Cluster the reads at 97% identity
- Each cluster is an OTU
- Count the reads in each OTU
- Match the OTU to an organism
- Read proportions are your abundances



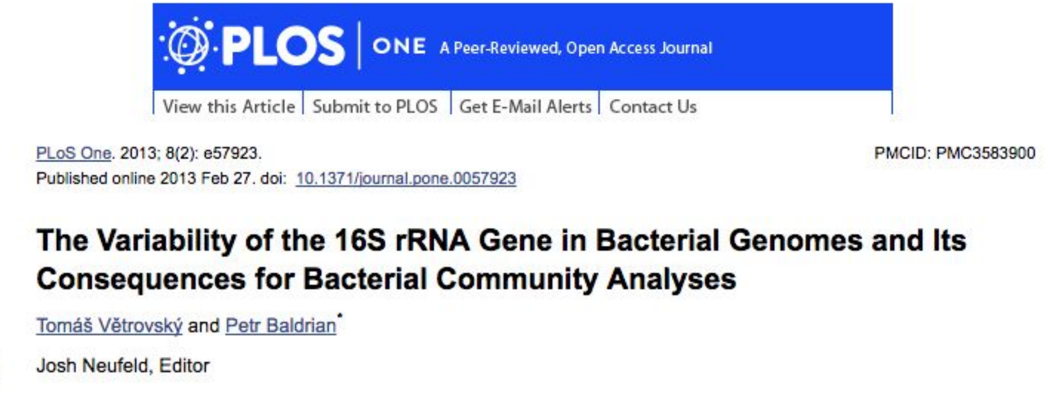
Benefits

- Cheaper
- Straightforward
- Well understood analysis pipelines
- Informative

Problems

- Samples bacteria (and some archaea) only
- Primers may not be universal
- Databases are not complete
- Limited resolution
- The OTU problem
- The copy number problem

The copy number problem



PLoS One. 2013; 8(2): e57923. PMCID: PMC3583900

Published online 2013 Feb 27. doi: 10.1371/journal.pone.0057923

The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses

Tomáš Větrovský and Petr Baldrian*

Josh Neufeld, Editor

- Sampled 1,690 bacterial genomes
- 1-16 16s rRNA gene copies per genome
- Sequences can differ within a genome
- Many species have identical copies

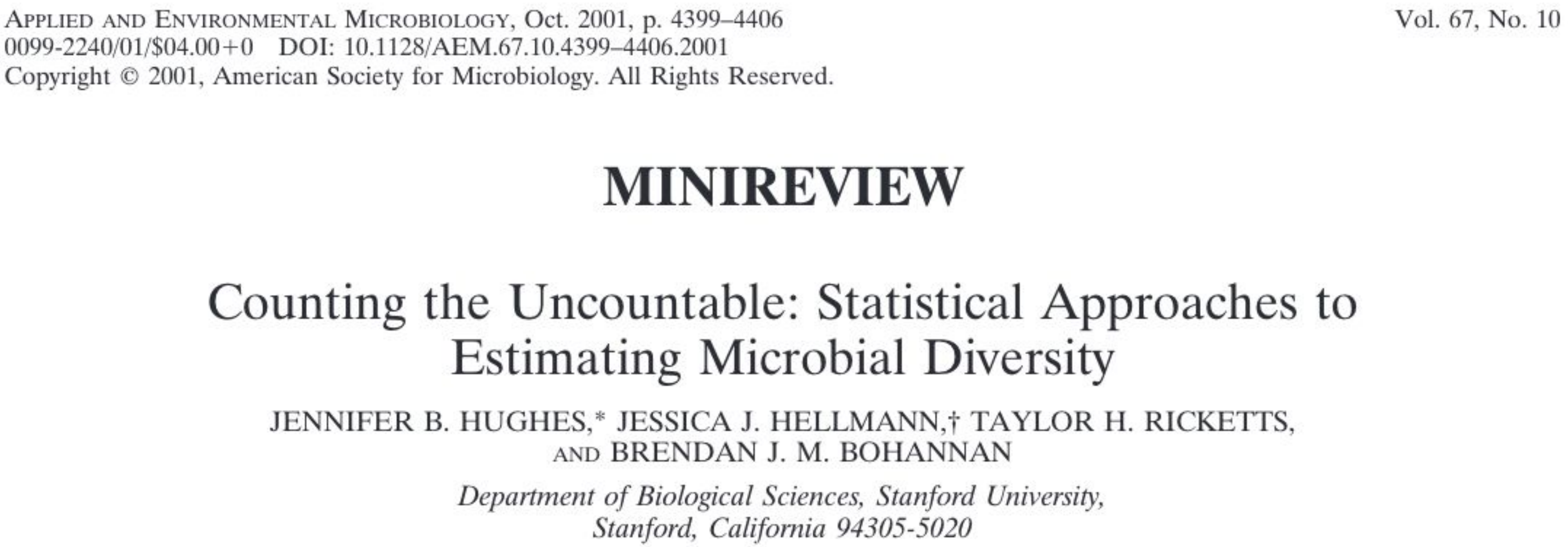
This means

- A species may have >1 OTU
- Many species may belong to the same OTU

Some terminology

- Alpha diversity
  - Number of OTUs in a sample
  - Their relative abundance

How to estimate microbial diversity?



Applied and Environmental Microbiology, Oct. 2001, p. 4399–4406 Vol. 67, No. 10
0099-2240/01/$04.00+0 DOI: 10.1128/AEM.67.10.4399–4406.2001

MINIREVIEW

Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity

JENNIFER B. HUGHES,* JESSICA J. HELLMANN,† TAYLOR H. RICKETTS, AND BRENDAN J. M. BOHANNAN

*Department of Biological Sciences, Stanford University,
Stanford, California 94305-5020*

- Whiteboard

Richness estimates

- Fit a curve and estimate Y max from the asymptote

Richness estimates

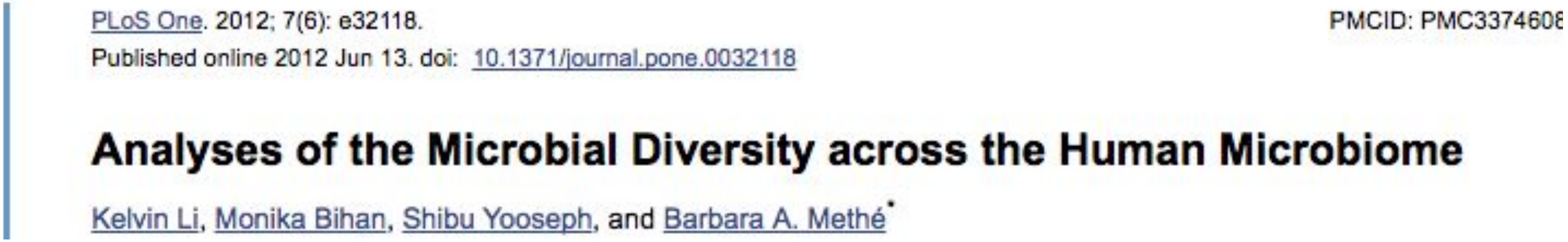
- Chao1 index:

$$S_{EST} = S_{OBS} + \frac{N_1^2}{2N_2}$$

Richness estimates

PLOS ONE A Peer-Reviewed, Open Access Journal

View this Article | Submit to PLOS | Get E-Mail Alerts | Contact Us



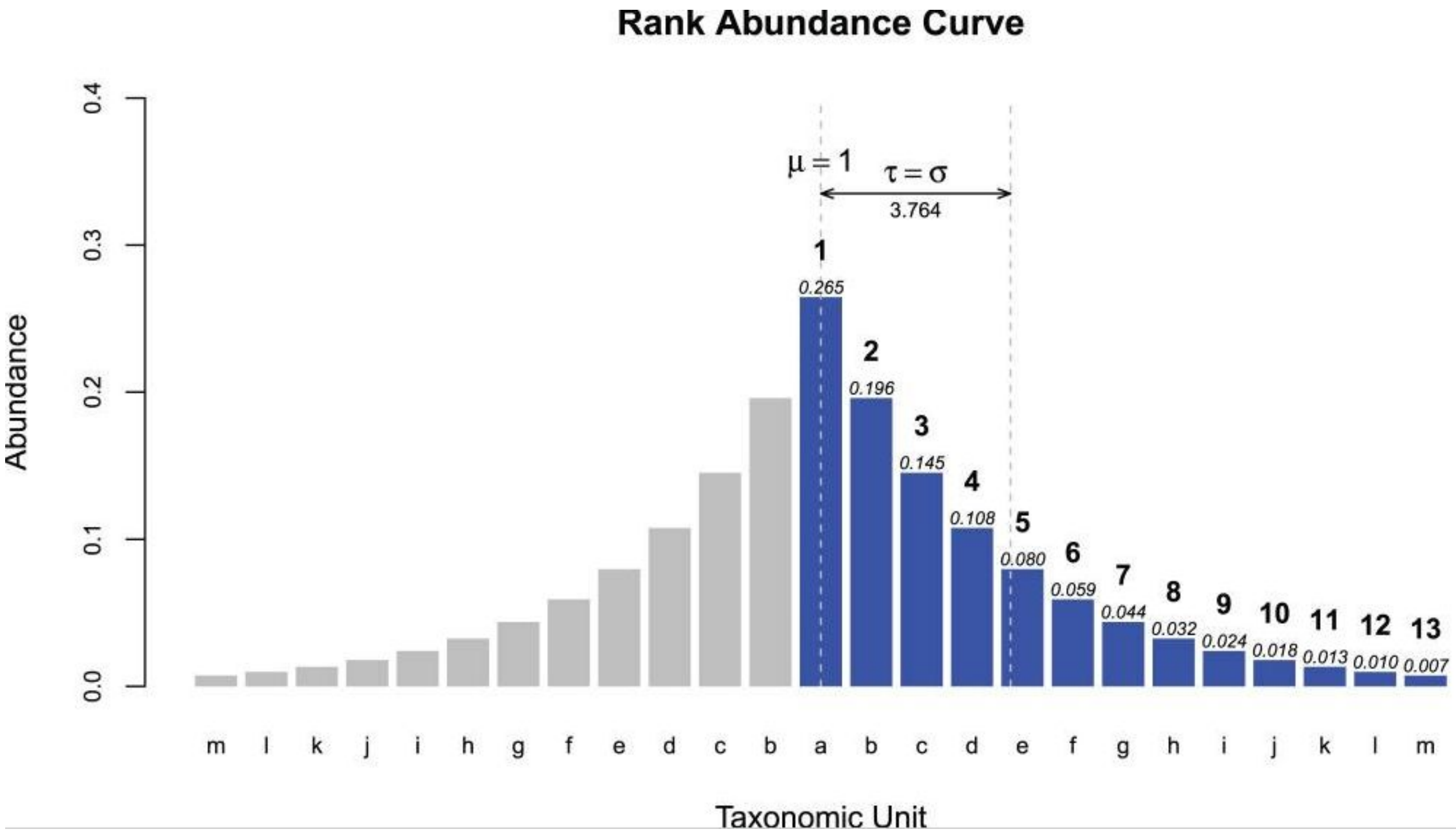
PLoS One. 2012; 7(6): e32118. PMCID: PMC3374608

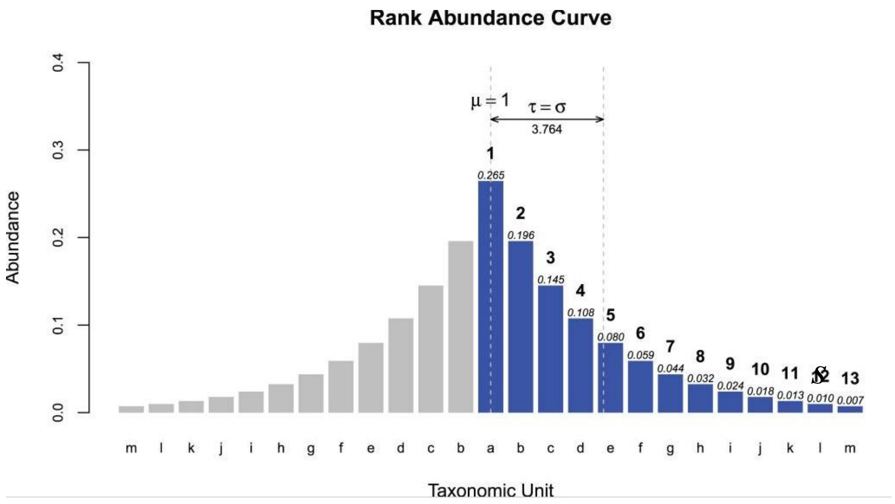
Published online 2012 Jun 13. doi: 10.1371/journal.pone.0032118

Analyses of the Microbial Diversity across the Human Microbiome

Kelvin Li, Monika Bihan, Shibu Yooseph, and Barbara A. Methé*

- Tail statistic


Rank Abundance Curve
Abundance
0.0
0.1
0.2
0.3
0.4
$\mu = 1$
$\tau = \sigma$
3.764
1
0.265
2
0.196
3
0.145
4
0.108
5
0.080
6
0.059
7
0.044
8
0.032
9
0.024
10
0.018
11
0.013
12
0.010
13
0.007
m l k j i h g f e d c b a b c d e f g h i j k l m
Taxonomic Unit



Fraction of OTUs discovered

$$F \geq \left(\frac{\tau^2}{D}\right)$$

Richness versus Evenness

- Pop. 1: 19 ants and 1 centipede
- Pop. 2: 10 ants and 10 centipedes
- Both have 20 organisms
- Both have 2 species
- How to represent the difference?
- Need something that scales with complexity.

Shannon Diversity Index

$$D = -\sum_{i=1}^{s} p_i \ln p_i$$

- Where *p* is the proportion of species *i* in the community
- More even = greater value
- Quantifies uncertainty in predicting identity of a species chosen at random

Simpson Index

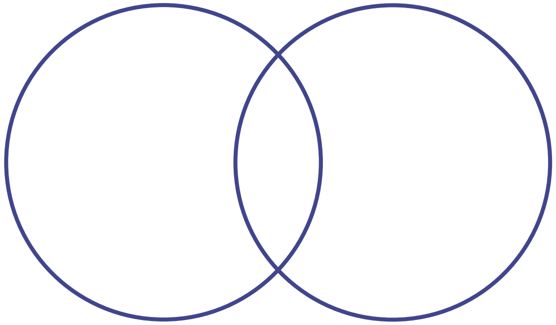
$$D = \sum_{i=1}^{s} p_i^2$$

- Where p is the proportion of species i in the community
- Less even = greater value
- Probability of two random members of the population being the same type

Beta diversity

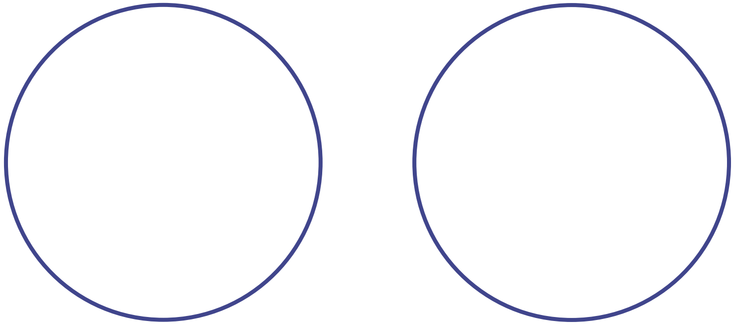
- A measure of the differences in richness and evenness between two communities

Jaccard Index



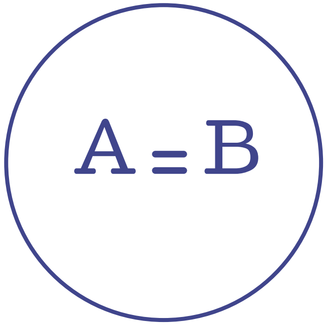
$$\frac{A \cap B}{A \cup B}$$

Jaccard Index



$$\frac{A \cap B}{A \cup B} = 0$$

Jaccard Index

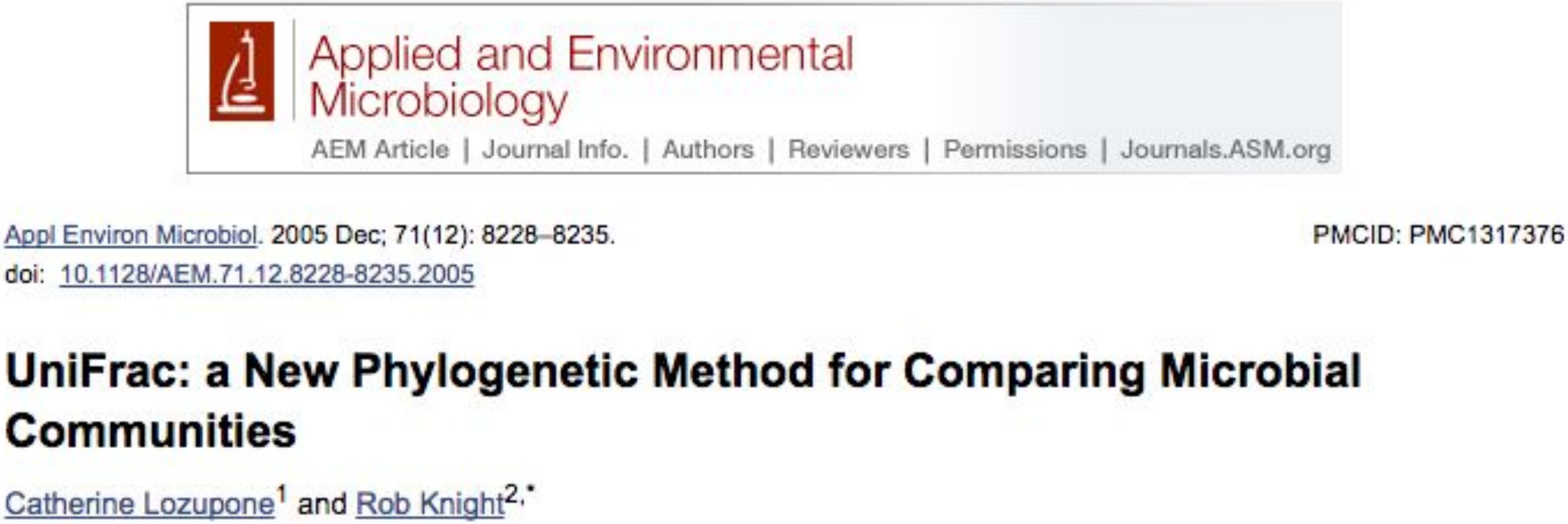


$$\left|\frac{A \cap B}{A \cup B}\right| = 1$$

Jaccard Index

- Problem: Relatedness/phylogeny is ignored
- As long as the union and intersection OTU numbers are the same, two highly related communities will have the same index as two distantly related communities

Unifrac



Applied and Environmental Microbiology

AEM Article | Journal Info. | Authors | Reviewers | Permissions | Journals.ASM.org

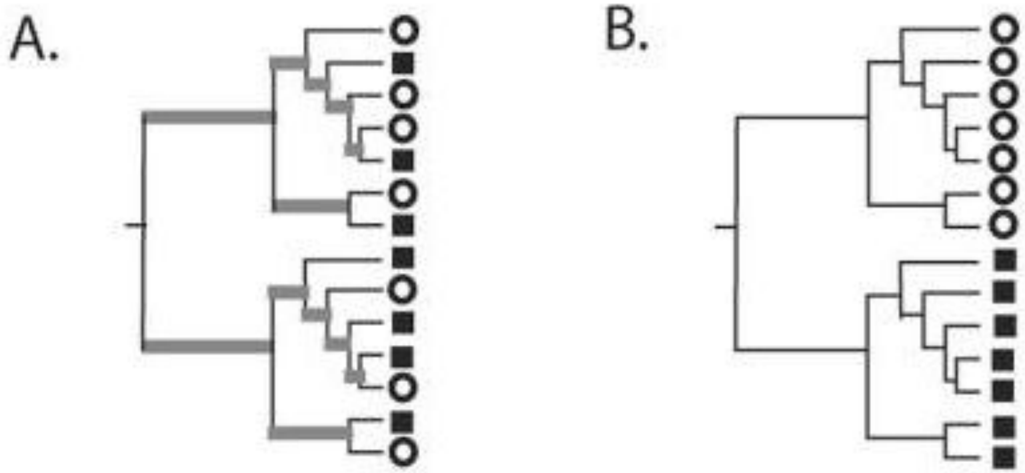
Appl Environ Microbiol. 2005 Dec; 71(12): 8228–8235. PMCID: PMC1317376

doi: 10.1128/AEM.71.12.8228-8235.2005

UniFrac: a New Phylogenetic Method for Comparing Microbial Communities

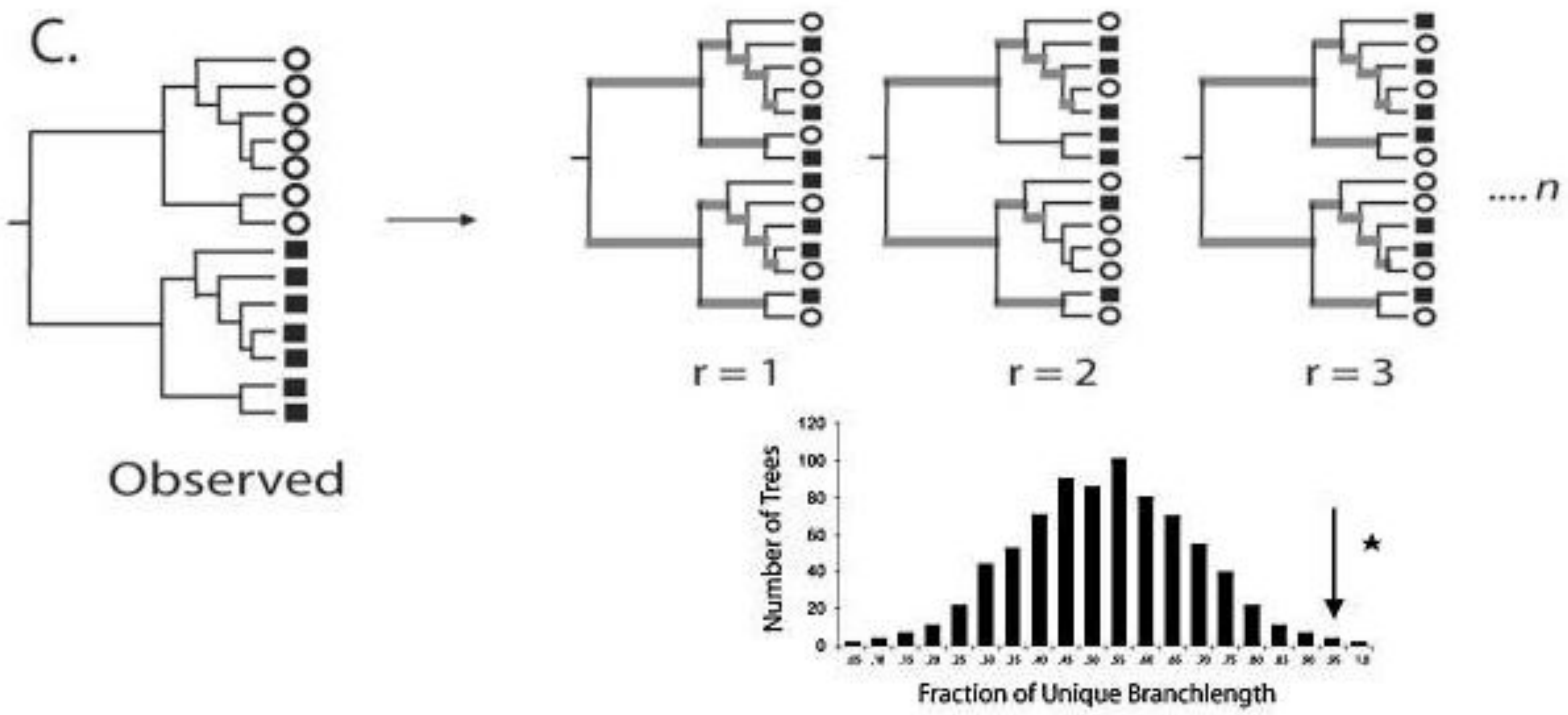
Catherine Lozupone[1] and Rob Knight[2,*]

- UNIFRAC measures the phylogenetic relatedness of communities



- Squares and circles are OTUs from two communities
- A has all of the branch length shared by squares and circles
  - The two communities are highly related
- B has zero branch length shared
  - The two communities are distantly related

Computing significance



Problems with 16s sequencing

- Samples bacteria (and some archaea) only
- Primers may not be universal
- Databases are not complete
- Limited resolution (genus, usually)
- The OTU problem
- The copy number problem

Problems with 16s sequencing

- What if we were to sequence all the DNA in the sample?
- What problems would be solved?
- What new problems would occur?
- What new opportunities would the data offer?