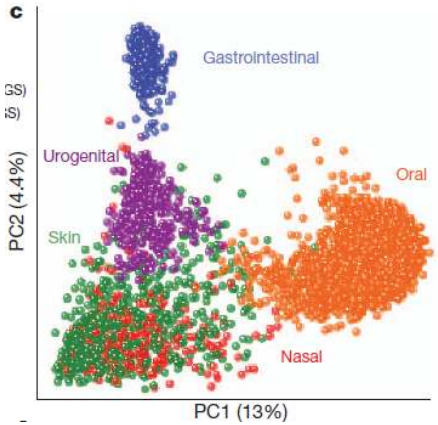
An Introduction to Metagenomics

Paul "Joey" McMurdie, PhD
Second Genome, Inc.



Schedule for today

Sec	Day	Start	End	Topic	Lead Instr.
1	Mon	09:00	10:00	Introduction to Metagenomics. Culture independent techniques, 16S rRNA, etc. (60 -75 min)	Joey
2	Mon	10:00	11:00	Introduction to microbiome analysis concepts -- Exploratory data analysis, Distances, PCoA, Ordination, taxa & sample-level inferences (75 min)	Joey
3	Mon	11:00	11:59	Introduction to microbiome analysis practices: QIIME, phyloseq, reproducible research (30 min)	Joey
---	Mon	12:00	14:00	Lunch (120min)	---
4	Mon	14:00	17:00	QIIME Lab (180min)	Daniel
---	Mon	17:00	19:00	Dinner (120min)	---
5	Mon	19:00	22:00	phyloseq Lab (180min)	Joey

Acknowledgements

Susan Holmes	Former postdoc advisor, mentor, co-author
Benjamin Callahan	DADA2 first author, slides, discussions, feedback, etc.
Holmes Group	Helpful advice and feedback
Wolfgang Huber	Helpful advice and feedback, creator of DESeq and DESeq2
BioC and CRAN	Support, Feedback, Distribution of phyloseq and biom
Rob Knight	QIIME, UniFrac, etc.
Huttenhower grp	Biobakery suite, slides, etc.
Hadley Wickham	ggplot2, reshape2, plyr R packages, Rstudio

An Introduction to Metagenomics

Outline for Today:

- What is metagenomics?
  - What methods, theoretical basis?
  - Why is it useful?
  - Where is it headed?
- How can I use it?
  - wet lab procedures (dry workshop)
  - computational protocols, practices

An Introduction to Metagenomics

Outline for Today:

- What is metagenomics?
  - What methods, theoretical basis?
  - Why is it useful?
  - Where is it headed?

} morning lecture

- How can I use it?
  - wet lab procedures (dry workshop)
  - computational protocols, practices

} afternoon + evening labs

An Introduction to Metagenomics

Outline for morning lecture:

- Microbiomes and metagenomics
  - What is a microbiome?
  - Why are they important?

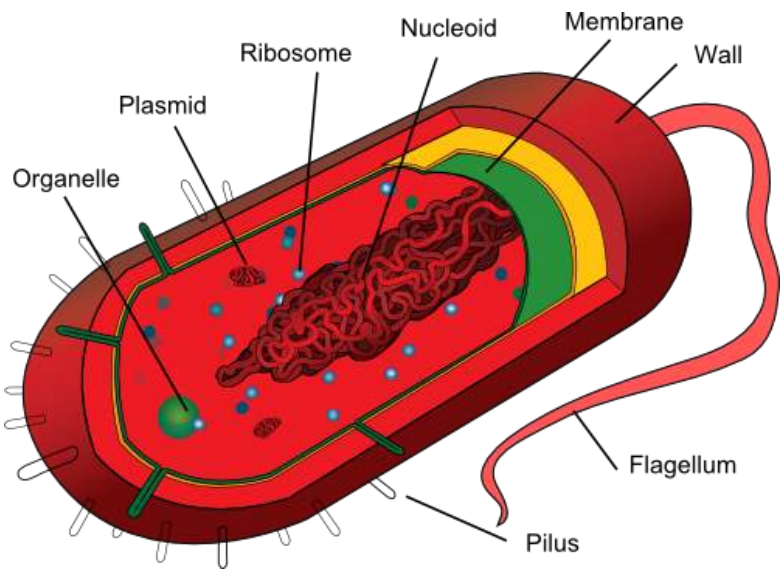
} Biological motivation

- Methods
  - Experimental methods
  - Analysis theory
  - Analysis tools, practices

} Methods

What are microbes?

Cell structure



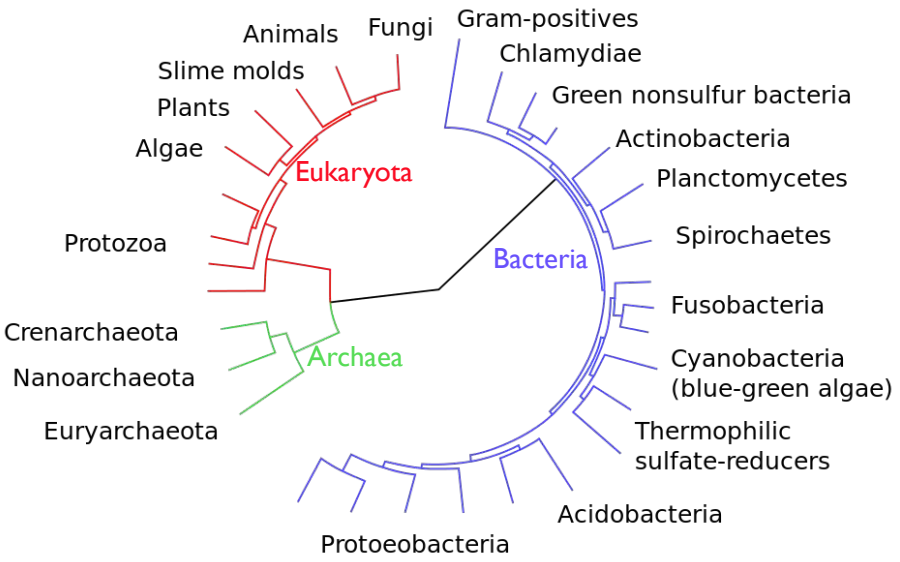
What are microbes?

Some key differences from eukaryota (e.g. humans, plants)

- Haploid genome
- Single circular chromosome, sometimes plasmids
- Genetic malleability, metabolic diversity
- Usually no nucleus (prokaryotes)
- Relatively easy interspecies gene transfer

What are microbes?

Ancestry of Life



http://en.wikipedia.org/wiki/Tree_of_life_(biology)

What is a microbiome?

The totality of microbes in a defined environment, especially their genomes and interactions with each other and surrounding environment.

- A population of a single species/strain is a culture, extremely rare outside of lab, some infections
- A microbiome is a mixed population of different microbial species (microbial ecosystem)

A mixed community is the norm!

Why Study Microbiomes?

Environmental Science
- Critical elemental cycles (carbon, nitrogen, sulfur, iron, …)
- Pollution control, cleanup
- Ecology / Evolution (chloroplasts, mitochondria, genetic evolution, …)

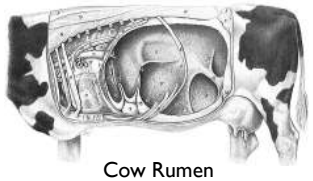
Industrial Applications
- Wastewater treatment (*V. cholera,* algal blooms, etc.)
- Bioprospecting (novel enzymes, compounds)
- Novel biosynthesis
- Fermentations: Consortia (yogurt) / wild (kombucha, Belgian ales)

Human Health
- Protection from pathogens (e.g. *Clostridium difficile*)
- Absorption/Production of nutrients in the gut
- Possible Role in chronic diseases
  (obesity, Crohn's/IBD, other autoimmune, UTIs, periodontitis, …)

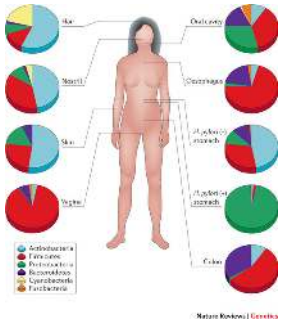
What is a microbiome?

Wastewater Treatment

Oceans, soils, waterways



Cow Rumen



Human Microbiomes

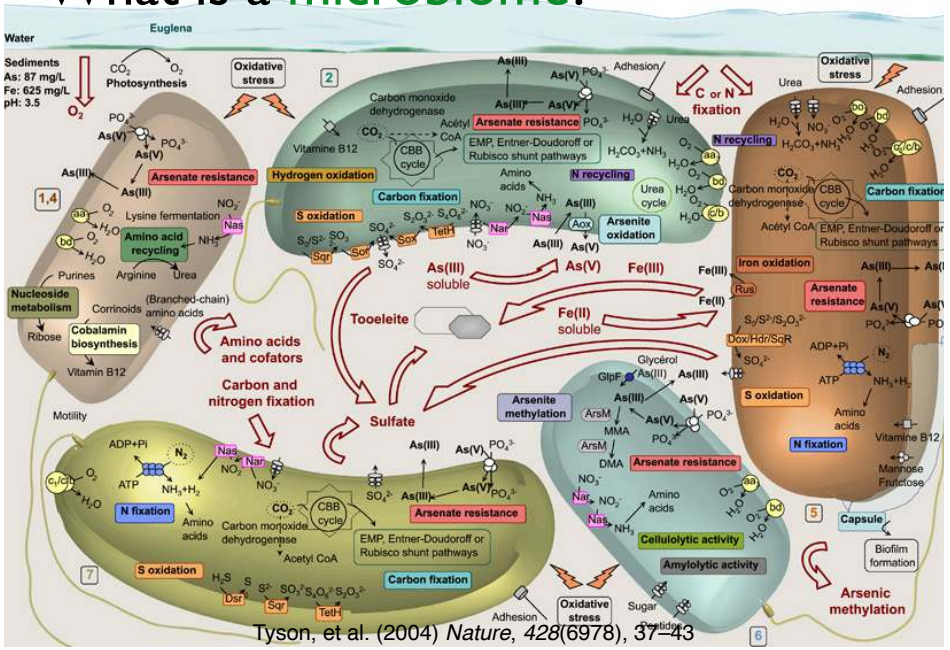
What is a microbiome?

acid mine biofilm

Tyson, et al. (2004) *Nature*, *428*(6978), 37–43

What is a microbiome?

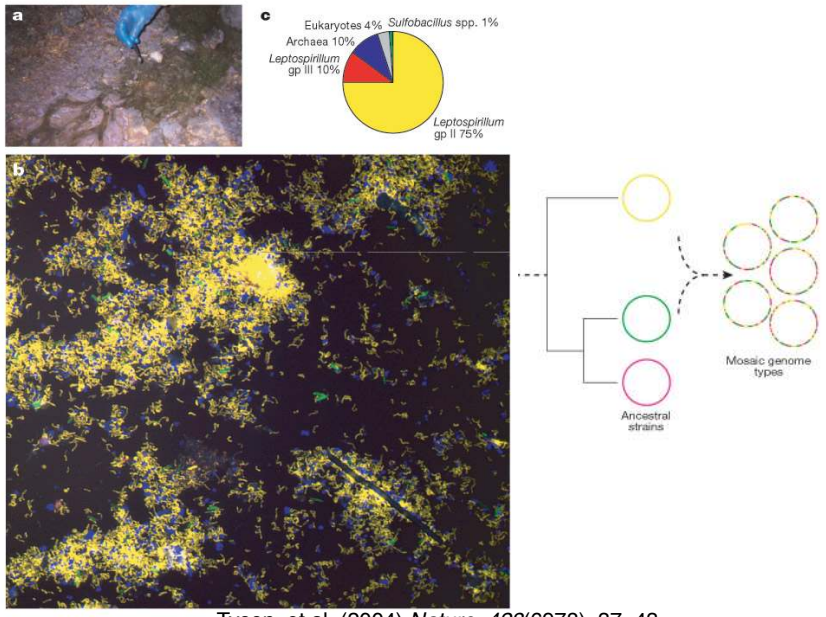
acid mine biofilm



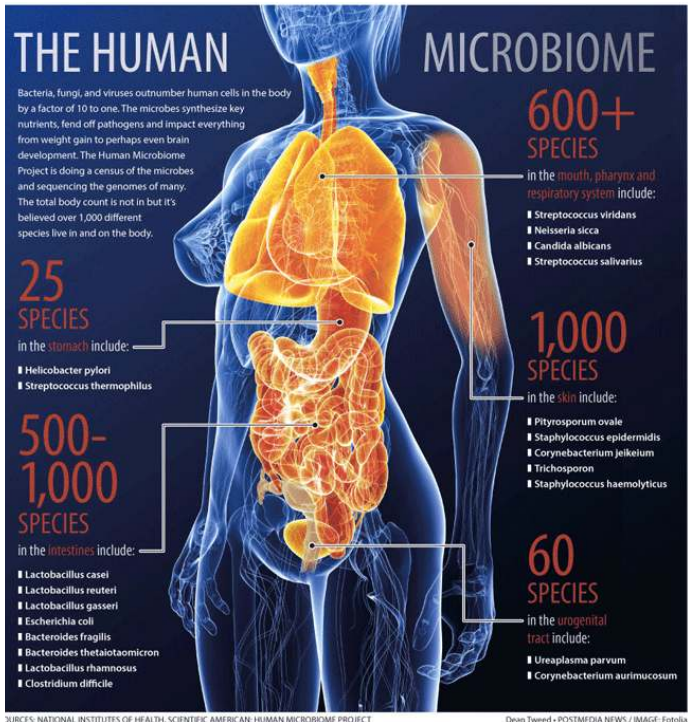
Tyson, et al. (2004) *Nature*, *428*(6978), 37–43

What is a microbiome?

acid mine biofilm

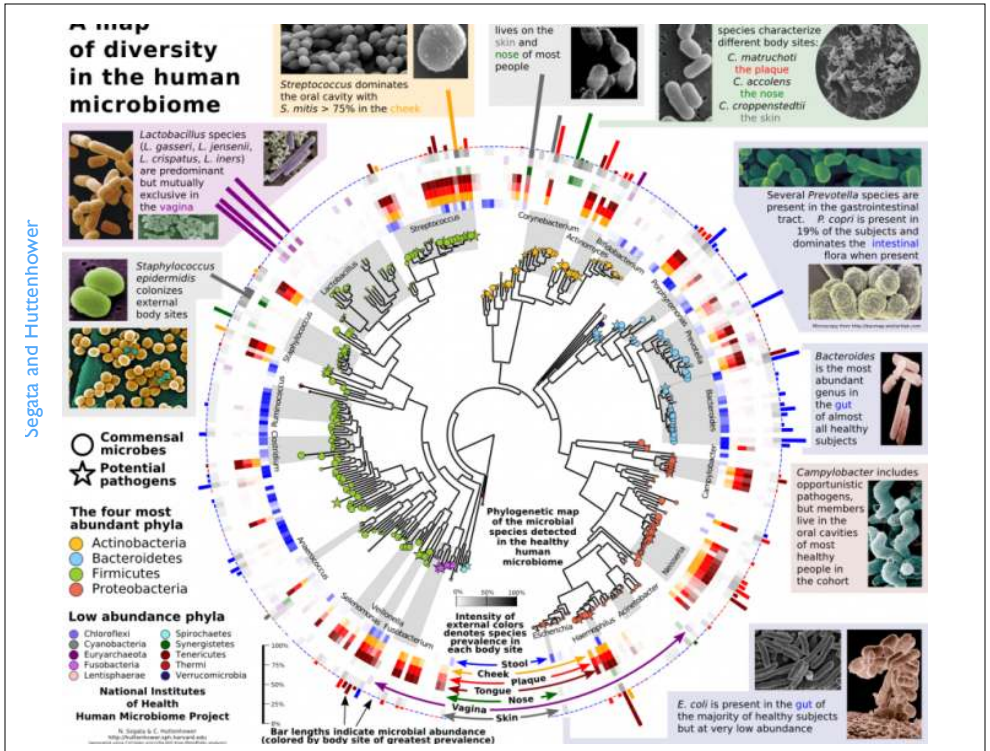


Tyson, et al. (2004) *Nature*, *428*(6978), 37–43



1–10 times more microbial cells than human cells… depends on timing of your last bowel movement

Typical human microbiome < 2 kg



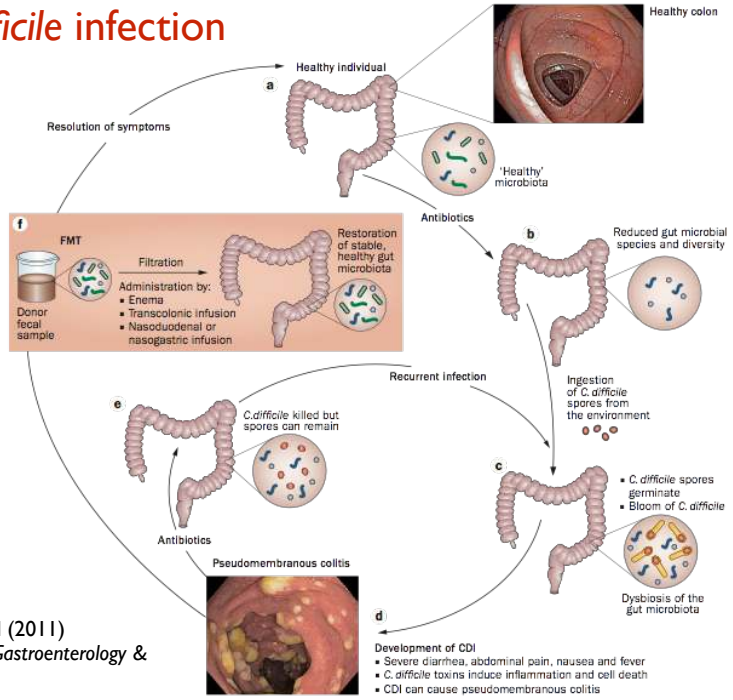
Some provocative oversimplifications…

Microbes can…

1. "Kill you by acute infection"
2. "Prevent same infection"
3. "Make you fat(ter)"
4. "Give you a heart attack"
5. "Give you cancer"
6. "Rescue you from cancer"

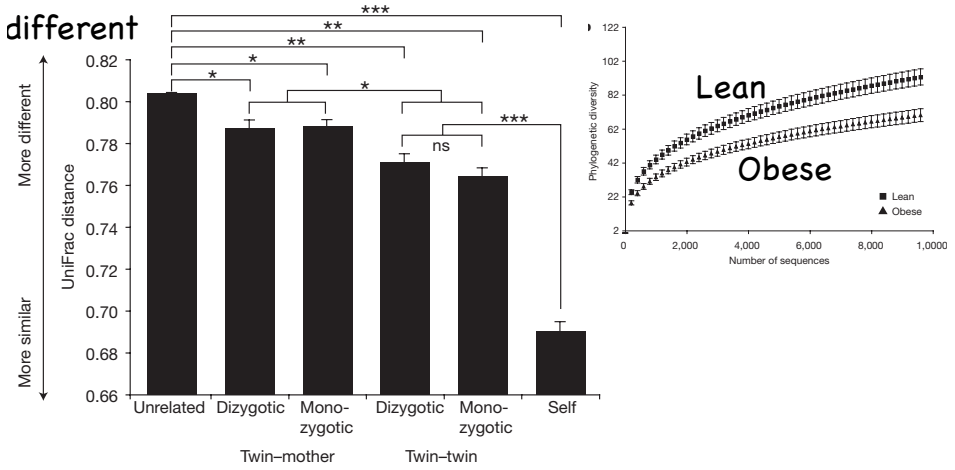
Can you guess the condition / scenario?

C. difficile infection



Borody, et al (2011)
Nature Rev Gastroenterology & Hepatology

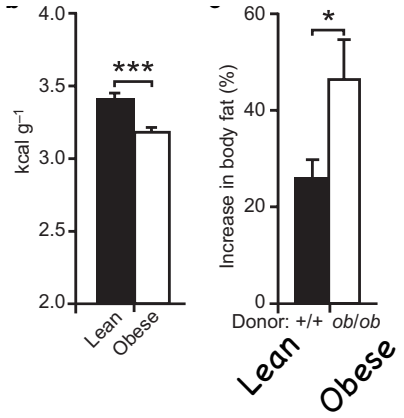
Microbes can make you fat(ter)…



Turnbaugh, et al. (2009). A core gut microbiome in obese and lean twins. *Nature*

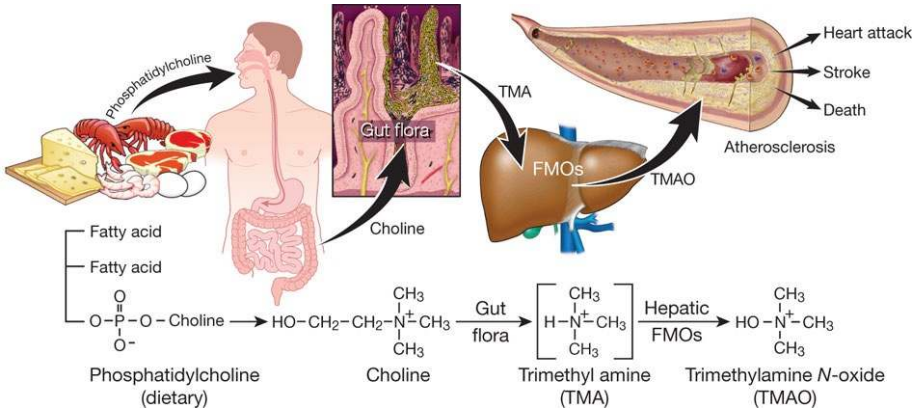
Microbes can make you fat(ter)…

- Lean (n = 10) & obese donors (n=9)
- Colonization of germ-free wild-type mice with microbiota from obese donors causes significant increase in total body fat
- Total body fat content was measured before and after a 2-week colonization
- Confirm that the ob/ob microbiome has an increased capacity for dietary energy harvest



Turnbaugh, et al. (2006). An obesity-associated gut microbiome … *Nature*

Gut microbes promote cardiovascular disease

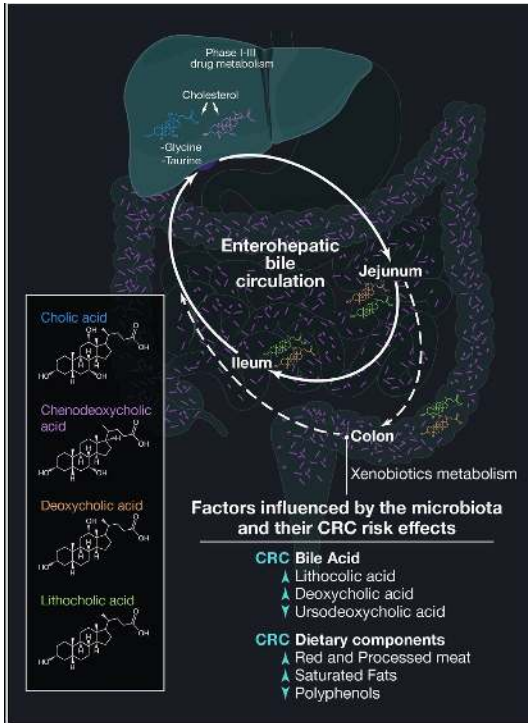


- Gut flora required for production of TMAO
- Supplementing diet with choline or TMAO promotes atherosclerosis (mouse)
- Gut flora suppression (Abx) inhibits dietary choline enhanced atherosclerosis
- TMAO is also a renal (kidney) toxin. Fogelman, A. M. (2015). *Circulation Research.*

ZN Wang, …, Stanley Hazen. *Nature* **472**, 57-63 (2011)
Fogelman, A. M. (2015). TMAO Is Both a Biomarker and a Renal Toxin. *Circulation Research*.

Colorectal Cancer (CRC)

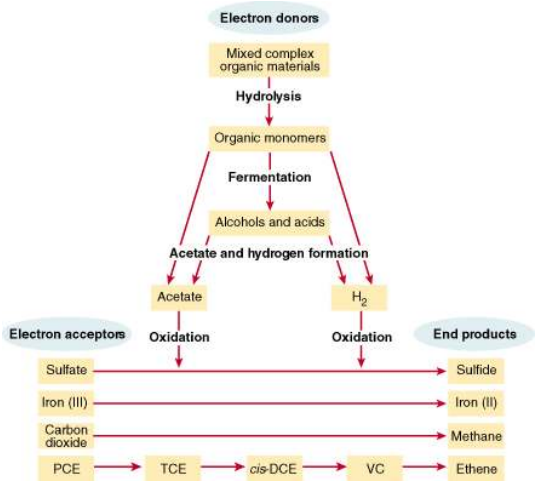
- Microbes affect colonic bile pool exposure, drug metabolism, and mortality-correlated compounds
- Microbe-produced secondary bile acids are among these.
- Gut microbial metabolism may play role in beneficial or detrimental effects of certain foods



Sears, C. L., & Garrett, W. S. (2014). Microbes, Microbiota, and Colon Cancer. Cell Host & Microbe, 15(3), 317–328.

Groundwater: Chlorinated Solvents

Dehalococcoides

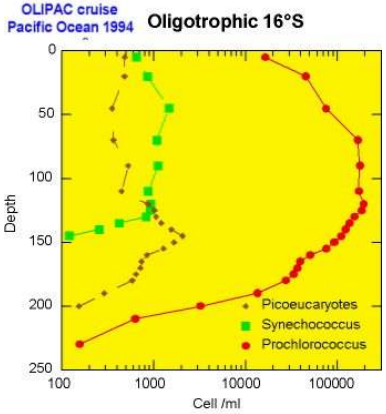


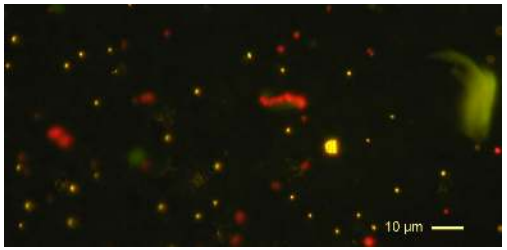
McCarty, P. L. (1997). Breathing with chlorinated solvents. *Science*

Marine picoplankton

most abundant organism on Earth?

- *Prochlorococcus* appears to be the **most abundant organism on the planet**
- Huge light harvesting proteins
- its density can reach up to **100 million cells per liter**
- it can be found down to a **depth of 150 m** in all of the intertropical belt
- picoplankton synchronize cell division at the same time every day —> **biological clock**





Vertical distribution of the photosynthetic picoplankton populations determined by flow cytometry in the tropical Pacific (OLIPAC cruise, 1994).

Yellowstone National Park

Octopus Spring

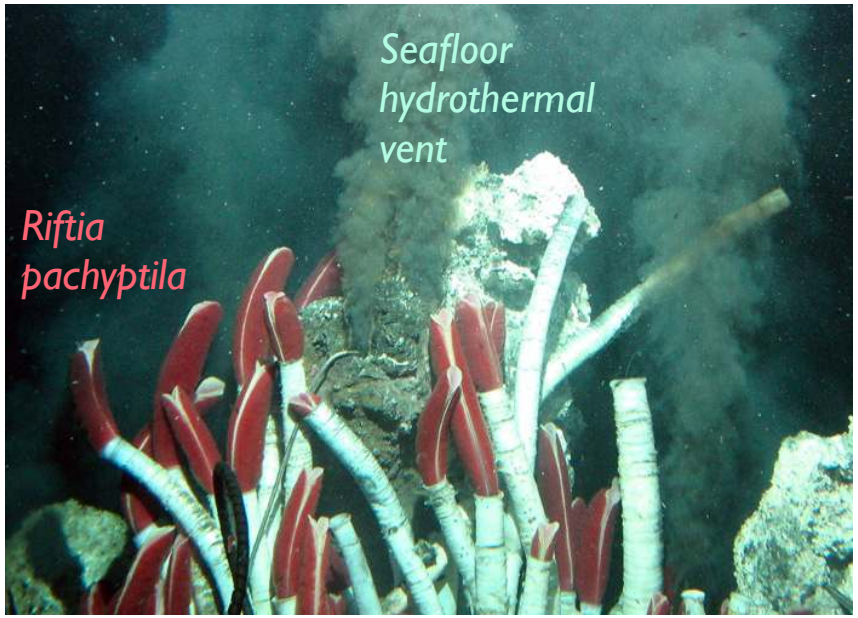
- 90° to 93°C
- extremely low in nutrients
- contains abundant biomass
- home to "oldest" known bacteria

Obsidian Pool

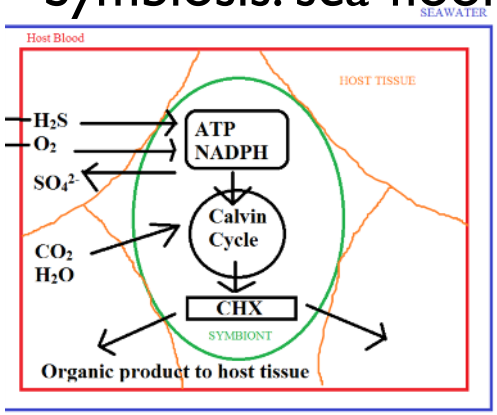
- 75° – 95°C
- high iron (II) hydrogen sulfide
- extensive diversity (previously unknown)

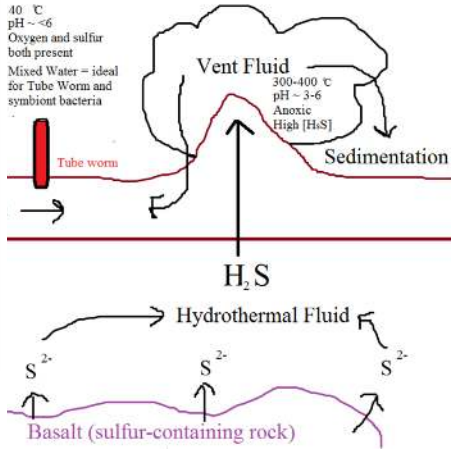
Ward, D. M., Weller, R., & Bateson, M. M. (1990). Nature, 345(6270), 63–65.
Barns, S. M., Fundyga, R. E., Jeffries, M. W., & Pace, N. R. (1994). PNAS 91(5), 1609–1613.

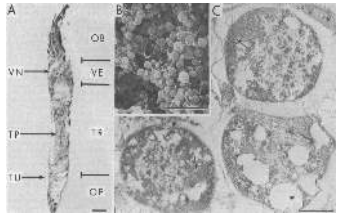
Symbiosis: sea-floor vent tube worm



Symbiosis: sea-floor vent tube worm







Cavanaugh, C. M. (1983). Nature, 302(5903), 58–61.
Cavanaugh, C. M., et al. (1981). Science. 213(4505), 340–342

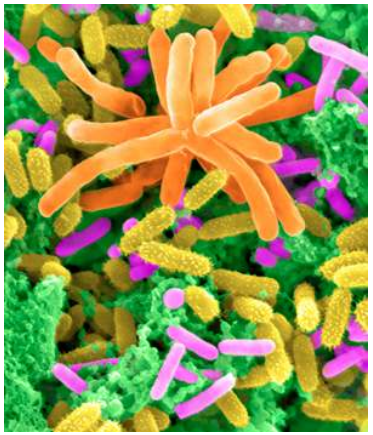
End: Biological Motivation

Questions before moving on?

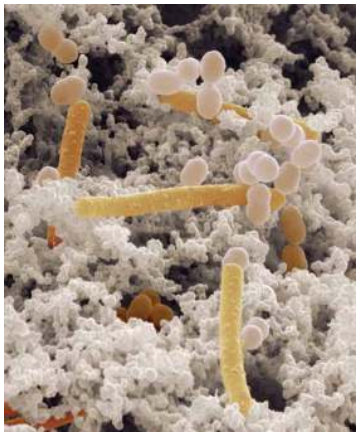
Metagenomics Experimental Methods

Exercise: How many species are present?

1



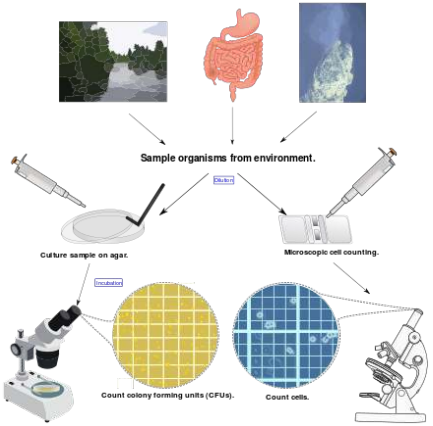
2



Confer amongst yourselves. We'll take a poll.

The great "plate count" anomaly

- Cultivation-based cell counts are orders of magnitude lower than direct microscopic observation.
- This is because microbiologists are able to **cultivate only a small minority** of naturally occurring microbes
- Our nucleic-acid derived understanding of microbial diversity has rapidly **outpaced our ability to culture new microbes**

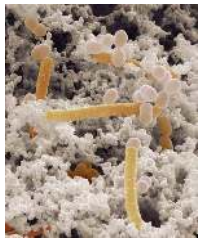


Staley, J.T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, 39, 321–346.

Why is microbiome research new?

Considering that…

- We have a bacterial endosymbiont in all our cells!
- Humans have always coexisted with bacteria
- We've known about bacteria for a few hundred years



- Historically prokaryotic biology has been focused on microbes that can be grown to large quantities/densities in the lab, especially pathogens; or can be distinguished under the microscope.
- An example of "searching where the light is"…

Why is microbiome research new?

Bias for cultivable microbes, especially pathogens

- Culture-based methods fail to detect most microbes
- Microbes are easy to miss (except pathogens)
- Most microbes are NOT pathogens (even the human-associated)

Availability of tools limited to last 3 decades

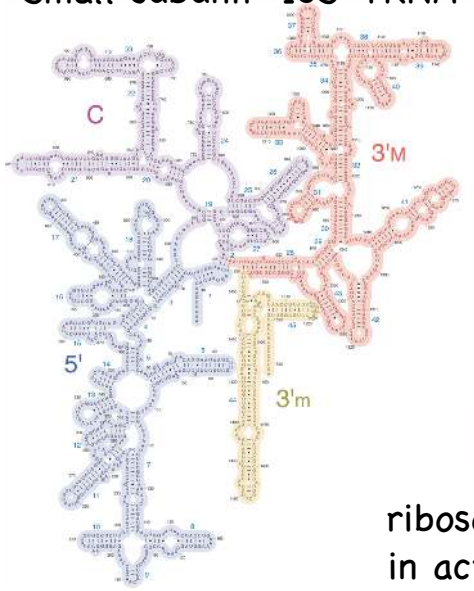
- Discovery of culture-independent techniques
- PCR, fast & cheap DNA sequencing, microarrays, etc

Discovery of *Culture Independent* Techniques

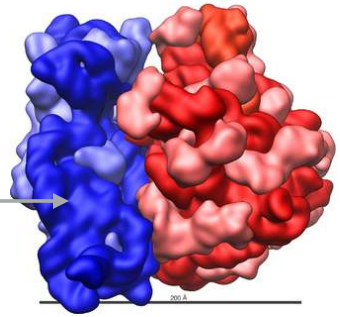
- 1977 rRNA as evolutionary marker - Woese & Fox *PNAS*
- 1985 Polymerase Chain Reaction (PCR) - K. Mullis *Science*
- 1985 "Universal" Primers for rRNA sequencing - N. Pace *PNAS*
- 1989 PCR amplification of 16S rRNA gene - Böttger *FEMS Microbiol.*
- 1996 Large, curated rRNA database (RDP) - Maidak *Nuc. Acids Res*
- 2001 term "microbiome" coined by Joshua Lederberg

Discovery of *Culture Independent* Techniques

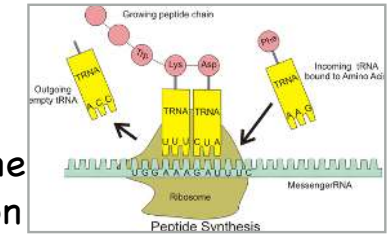
Small subunit "16S" rRNA



ribosome



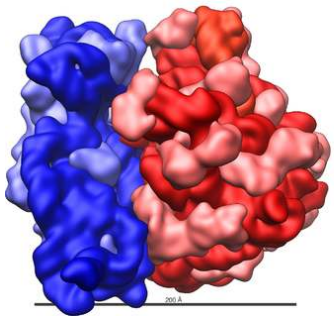
ribosome in action

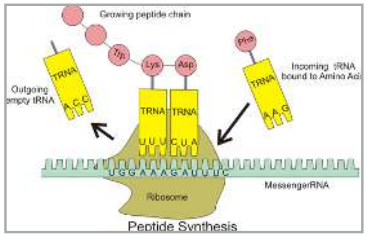


Discovery of *Culture Independent* Techniques

ribosome

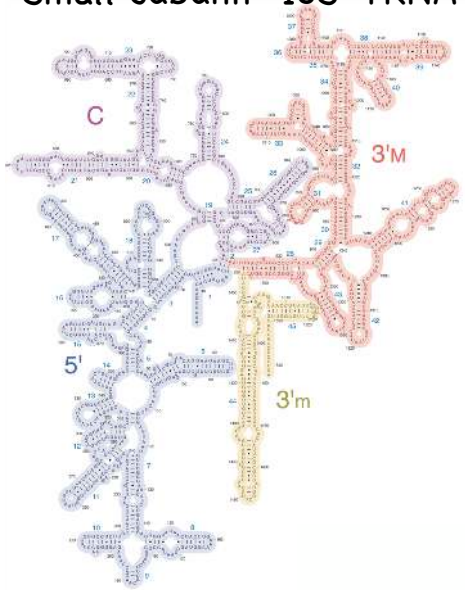
- rRNA has both catalytic and structural function.
- The small and large subunits have different lengths, 2nd-structure, 3D shape; but must work together.
- All of the catalytic activity of the ribosome is carried out by the RNA; the proteins reside on the surface and seem to stabilize the structure.





Discovery of *Culture Independent* Techniques

Small subunit "16S" rRNA

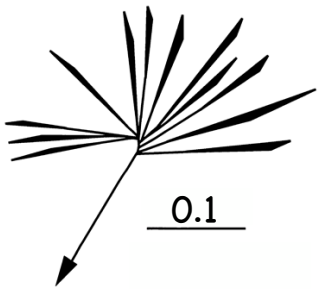


- **Ubiquitous** - present in all known life (viruses don't count)
- **Functionally constant** translation, 2°-structure
- **Evolves slowly** - mutations more rare than for protein-coding genes
- **Large** - information for evolutionary inference
- **No exchange** - Limited examples of rRNA gene-sharing between organisms

Discovery of *Culture Independent* Techniques

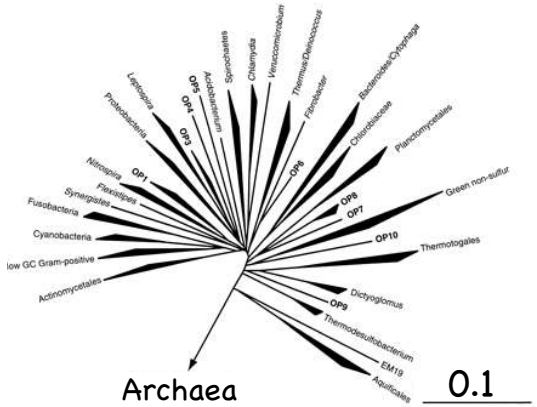
Evolutionary Tree, Known Bacteria

1987



Archaea
Eukarya

1997



Archaea
Eukarya

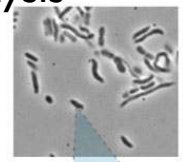
0.1

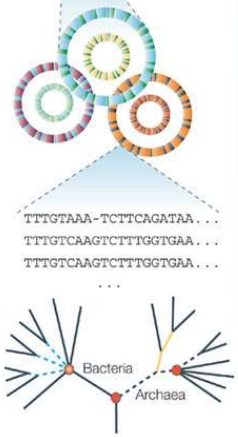
Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313), 734–740.

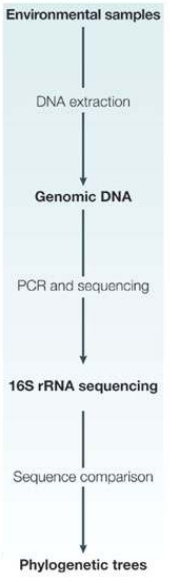
Metagenomics: Nucleic acid sequencing as a tool for microbial community analysis

Single microbiome:

1. Break all cells, extract all DNA (gDNA)
2. PCR-amplify a universal gene from gDNA
3. DNA sequencing from pool of amplified genes
4. Cluster sequences according to species
5. Count each species and make a tree





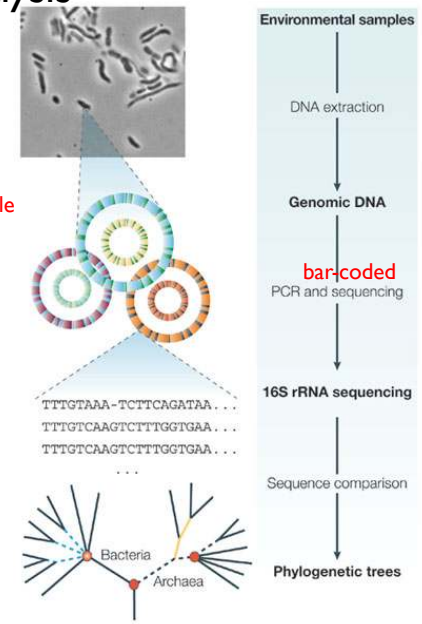


Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, *6*(11), 805–814.

Metagenomics: Nucleic acid sequencing as a tool for microbial community analysis

Many microbiomes in parallel:

1. Break all cells, extract all DNA (gDNA)
2. PCR-amplify a universal gene from gDNA
   using bar-coded primers, diff code for each sample
3. DNA sequencing from pool of amplified genes
   4a. "De-multiplex" barcode, ID source sample
4. Cluster sequences according to species
5. Count each species and make a tree

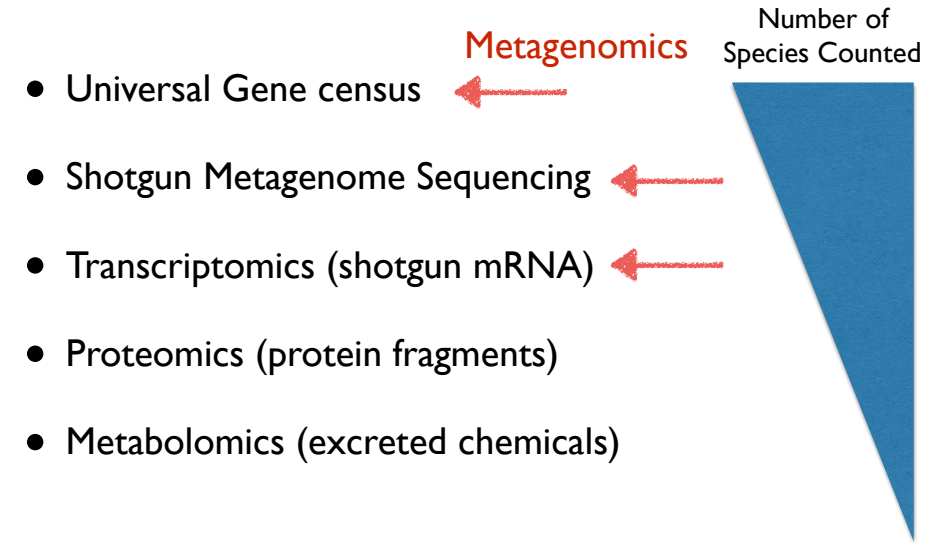


Culture Independent Techniques:

Metagenomics

Number of Species Counted

- Universal Gene census
- Shotgun Metagenome Sequencing
- Transcriptomics (shotgun mRNA)
- Proteomics (protein fragments)
- Metabolomics (excreted chemicals)

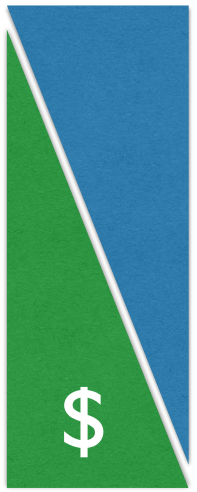


Culture Independent Techniques:

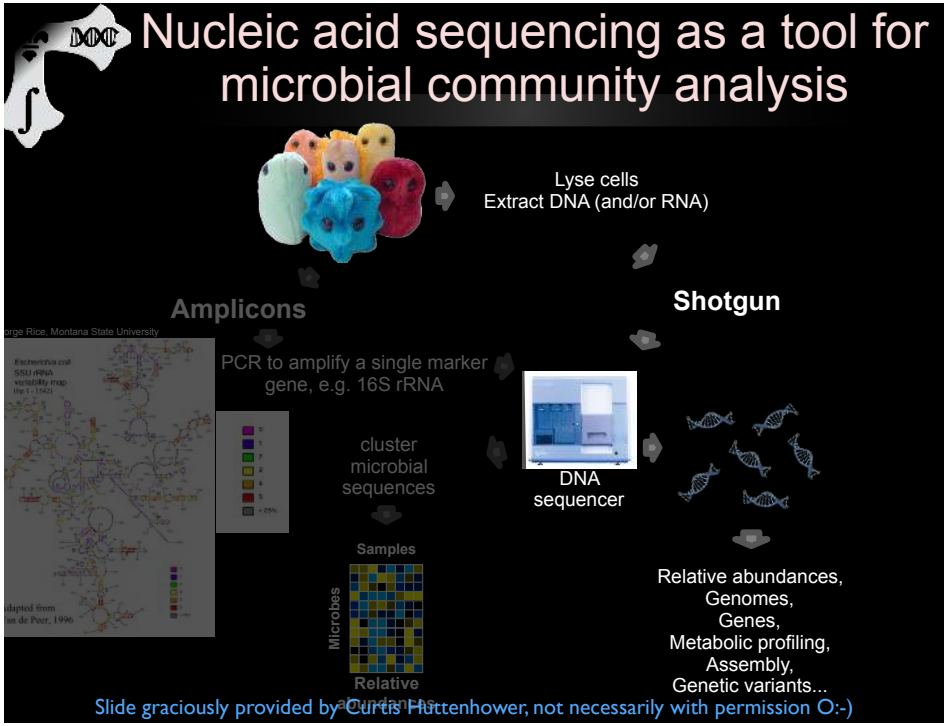
Metagenomics

Number of Species Counted

- Universal Gene census
- Shotgun Metagenome Sequencing
- Transcriptomics (shotgun mRNA)
- Proteomics (protein fragments)
- Metabolomics (excreted chemicals)



Nucleic acid sequencing as a tool for microbial community analysis



Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

Sequencing as a tool for microbial community analysis

Who's there?
(Taxonomic profiling)

What are they doing?
(Functional profiling)

What does it all mean?
(Statistical analysis)

Relative abundances

Assembly, Genetic variants...

Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

A Summary of Meta'omics

- Piles of short DNA/RNA reads from >1 organism
- You can...
  - Ecologically profile them
  - Taxonomically or phylogenetically profile them
  - Functionally profile them – gene/pathway catalogs
  - Assemble them
- Prior knowledge is helpful
- Caution: Correlation ≠ Causation
  - Most 'omics results require lab confirmation

Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

Working toward high-impact outcomes from meta'omic microbial community profiling

Epidemiology

Privacy and ethics
- Disease risk/pathogen exposure
- Tracking

Health policy
- Early life exposures
- Pharma. best practices

Translation

Phenotype association for diagnostics
- Human disease risk: lifetime, activity, outcome
- Longitudinal analysis and study design
- Dense longitudinal measures, multiple nested outcomes

Systems analysis for intervention
- More and simpler model systems
- Systematic understanding of current models
- Ecological models for ecosystem restoration

Basic biology and molecular mechanism

Microbial experiments
- Quantitative methods
- Integration/meta-analysis of genomes and metagenomes

Host-microbe-microbiome interactions
- Immunity in specific host tissues
- Non-immune mechanisms (metabolites, peptides)
- Model system perturbations, "knock ins" and "knock outs"

Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

An Introduction to Metagenomics

Outline for morning lecture:

- Microbiomes and metagenomics
  - What is a microbiome? ✓
  - Why are they important? ✓
- Methods
  - Experimental methods ✓
  - Analysis theory
  - Analysis tools, practices

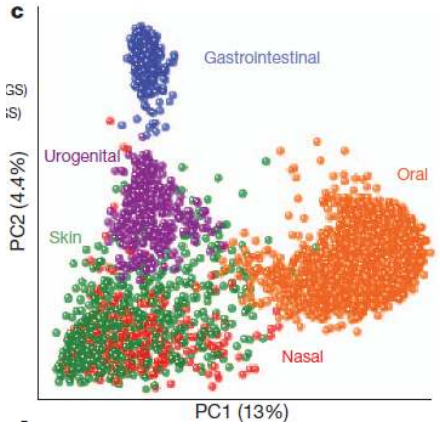
Biological motivation

Methods

End Metagenomics Lecture 1

Questions?

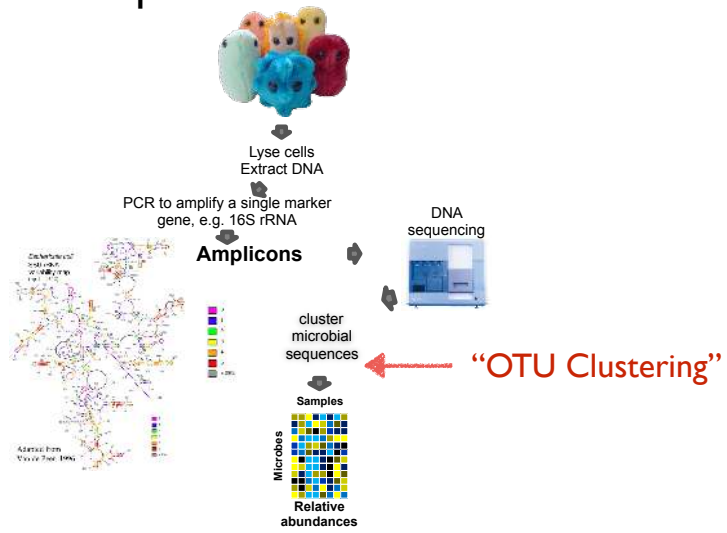
Introduction to Microbiome / Metagenome Analysis Concepts



- **Sequence Processing (OTUs)**
  - Denoising
  - Chimera detection
  - Construction of sequence clusters (OTUs)
- **Comparing microbiomes**
  - Distances, Diversity
  - Exploratory Data Analysis
    - Ordination Methods
    - hierarchical dendrogram
    - extract patterns from a plot
      - clusters - gap statistic
      - gradient - regression, modeling, etc.
- **Identifying important microbes/taxa**
  - projected points, coinertia (plots)
  - inferential testing
  - modeling

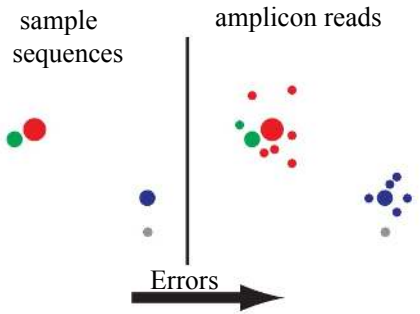
- **Sequence Processing (OTUs)**
  - Denoising
  - Chimera detection
  - Construction of sequence clusters (OTUs)
- **Comparing microbiomes**
  - Distances, Diversity
  - Exploratory Data Analysis
    - Ordination Methods
    - hierarchical dendrogram
    - extract patterns from a plot
      - clusters - gap statistic
      - gradient - regression, modeling, etc.
- **Identifying important microbes/taxa**
  - projected points, coinertia (plots)
  - inferential testing
  - modeling

OTUs - Operational Taxonomic Unit



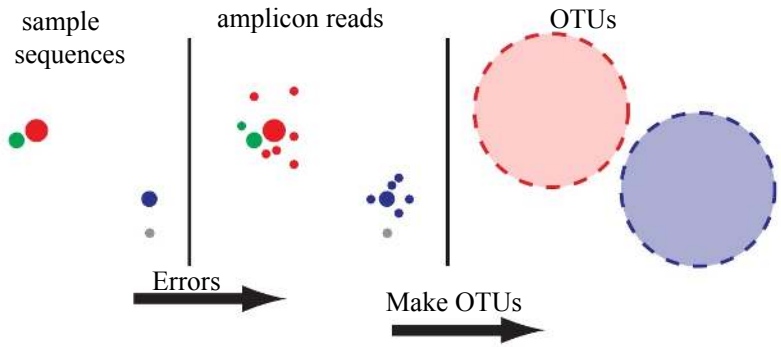
Slide adapted from slide by Curtis Huttenhower, not necessarily with permission O:-)

Sample Inference from Noisy Reads



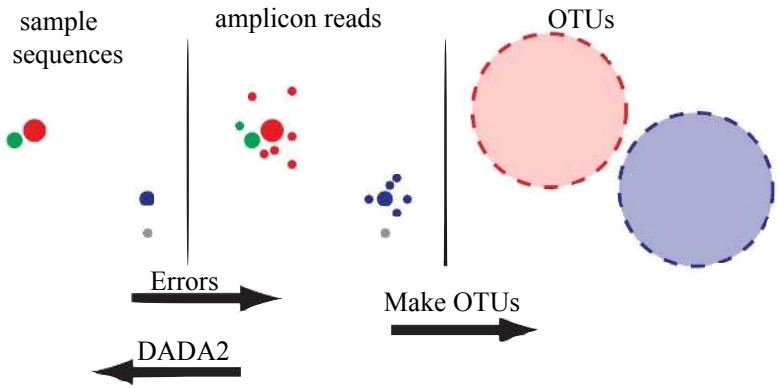
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Sample Inference from Noisy Reads



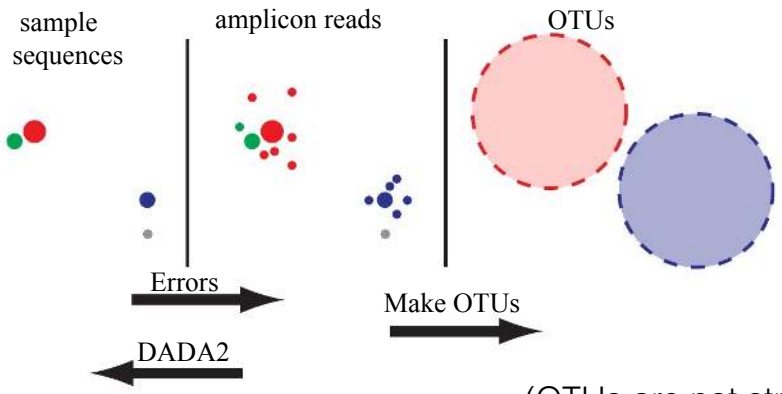
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Sample Inference from Noisy Reads



Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Sample Inference from Noisy Reads

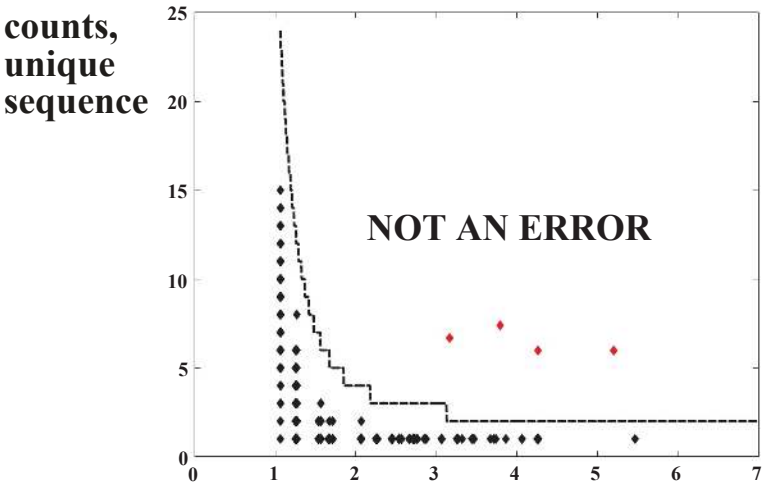


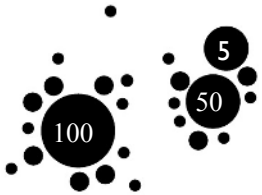
(OTUs are not strains)

OTUs: Lump similar sequences together
DADA2: Statistically infer the sample sequences

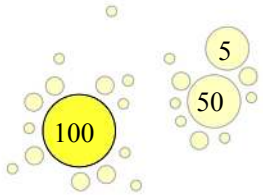
The true shape of an error cloud

DADA2: Error Model





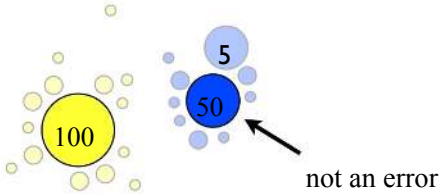
Initial guess: one real sequence + errors



Infer initial *error model* under this assumption.

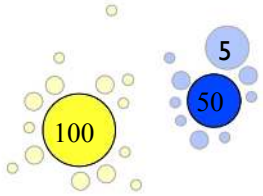
$\Pr(i \to j) =$

	A	C	G	T
A	0.97	10^{-2}	10^{-2}	10^{-2}
C	10^{-2}	0.97	10^{-2}	10^{-2}
G	10^{-2}	10^{-2}	0.97	10^{-2}
T	10^{-2}	10^{-2}	10^{-2}	0.97



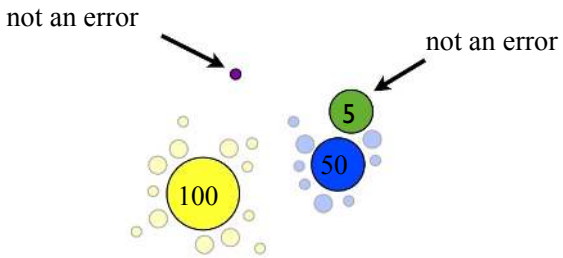
Reject unlikely error under model. **Recruit** errors.

	A	C	G	T
A	0.97	10^{-2}	10^{-2}	10^{-2}
C	10^{-2}	0.97	10^{-2}	10^{-2}
G	10^{-2}	10^{-2}	0.97	10^{-2}
T	10^{-2}	10^{-2}	10^{-2}	0.97



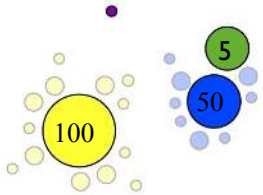
Update the model.

	A	C	G	T
A	0.997	10^{-3}	10^{-3}	10^{-3}
C	10^{-3}	0.997	10^{-3}	10^{-3}
G	10^{-3}	10^{-3}	0.997	10^{-3}
T	10^{-3}	10^{-3}	10^{-3}	0.997



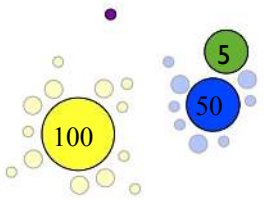
Reject more sequences under ***new*** model

	A	C	G	T
A	0.997	10^{-3}	10^{-3}	10^{-3}
C	10^{-3}	0.997	10^{-3}	10^{-3}
G	10^{-3}	10^{-3}	0.997	10^{-3}
T	10^{-3}	10^{-3}	10^{-3}	0.997



Update model again

	A	C	G	T
A	0.998	$1x10^{-4}$	$2x10^{-3}$	$2x10^{-4}$
C	$6x10^{-5}$	0.999	$3x10^{-6}$	$1x10^{-3}$
G	$1x10^{-3}$	$3x10^{-6}$	0.999	$6x10^{-5}$
T	$2x10^{-4}$	$2x10^{-3}$	$1x10^{-4}$	0.998



Convergence: all errors are plausible

	A	C	G	T
A	0.998	1×10^{-4}	2×10^{-3}	2×10^{-4}
C	6×10^{-5}	0.999	3×10^{-6}	1×10^{-3}
G	1×10^{-3}	3×10^{-6}	0.999	6×10^{-5}
T	2×10^{-4}	2×10^{-3}	1×10^{-4}	0.998

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

DADA2: Why is this possible?

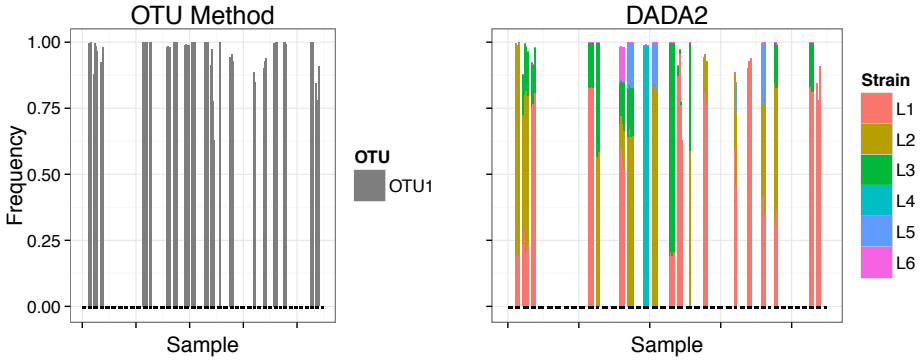
Uses more of the information than traditional OTU clustering

	DADA2	OTUs
Abundance	✓	**Ranks only**
Sequence Differences	✓	**Count only**
Quality	✓	**No**
Error Model	✓	**No**

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

DADA2 Advantages: Real Data

Lactobacillus crispatus sampled from vaginal microbiome 42 pregnant women



Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

DADA2 Advantages

Analytical

Single nucleotide resolution
- genotypes/strains instead of 97% OTUs

Lower false positive rate
- Better error model, easier to ID chimeras

Computational

Linear scaling of computational costs
- Exact sequences are inherently comparable, so samples can be processed independently.

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

DADA2

Divisive Amplicon Denoising Algorithm - ver.2

DADA2: High resolution sample inference from amplicon data

Benjamin J Callahan[1,*], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo Johnson[2] and Susan P Holmes[1]

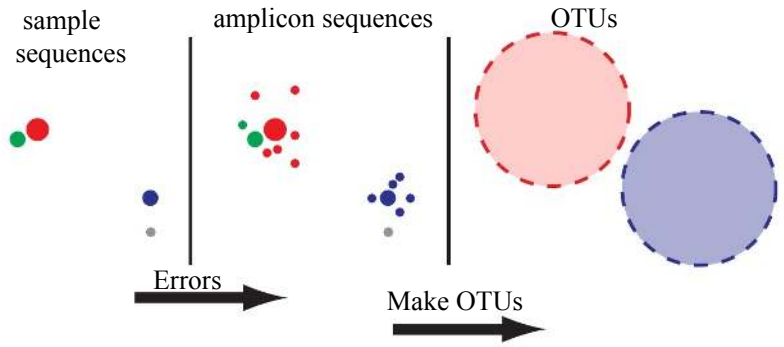
[1]Department of Statistics, Stanford University
[2]Second Genome, South San Francisco, CA
[3]Department of Applied Physics, Stanford University
[*]Corresponding Author: benjamin.j.callahan@gmail.com

http://dx.doi.org/10.1101/024034 Manuscript draft on bioRxiv (Revisions in review)

http://benjjneb.github.io/dada2/ Basic R package available on GitHub

DADA1: Rosen MJ, Callahan BJ, Fisher DS, Holmes SP (2012) Denoising PCR-amplified metagenome data. BMC bioinformatics, 13(1), 283.

That said, we are going to use OTUs!

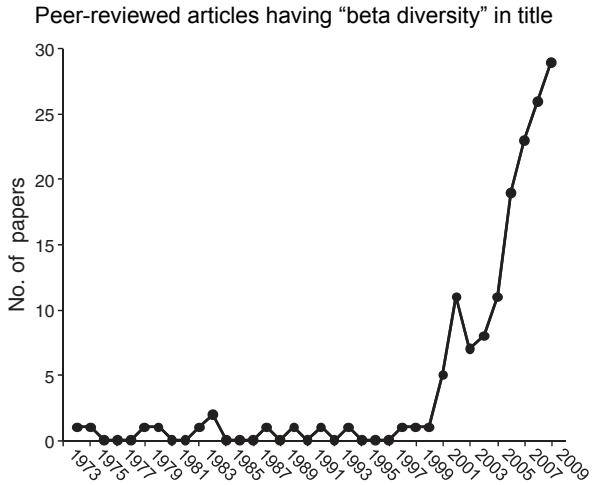


Diversity

Diversity of diversity
(diversity of greek letters used in ecology)

- α – diversity within a community, # of species
- β – diversity between communities (differentiation), species identity is taken into account
- γ – (global) diversity of the site, $\gamma = \alpha \times \beta$, but only this simple if α and β are independent
- Probably others, but α and β are most common

Beta-Diversity



Anderson, M. J., et al. (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology Letters, 14(1), 19–28.

Beta-Diversity

- Microbial ecologists typically use beta diversity as a broad umbrella term that can refer to any of several indices related to compositional differences

  (Differences in species content between samples)
- For some reason this is contentious, and there appears to be ongoing (and pointless?) argument over the possible definitions
- For our purposes, and microbiome research, when you hear “beta-diversity”, you can probably think:

“Diversity of species composition”

http://en.wikipedia.org/wiki/Beta_diversity

Distances between microbiomes

Community Distance

Communities are a vector of abundances:

$\mathbf{x} = \{x_1, x_2, x_3, \ldots\}$

E. coli: ●●●
P. fluorescens: ●
B. subtilis: ●
P. acnes:
D. radiodurans:
H. pylori: ●●●●●●●
L. crispatus:

$\mathbf{x} = \{3,1,1,0,0,7,0\}$

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Community Distance Properties

- Range from 0 to 1
- Distance to self is 0
- If no shared taxa, distance is 1
- Triangle inequality (metric)
- Joint absences do not affect distance (biology)
- Independent of absolute counts (metagenomics)

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

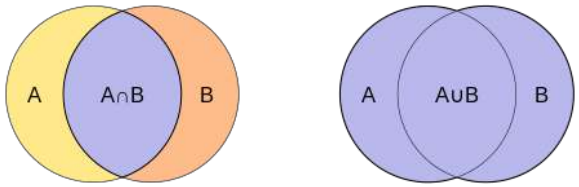
The Distance Spectrum

	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Jaccard

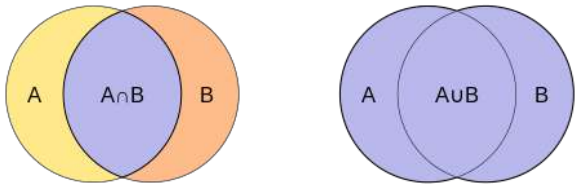
$$\text{Dist}(A, B) = 1 - (A \cap B)/(A \cup B)$$
$$= ((\mathbf{x_A}>0) \,\&\, (\mathbf{x_B}>0))/((\mathbf{x_A}>0) \mid (\mathbf{x_B}>0))$$



Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Jaccard

$$\text{Dist}(A, B) = 1 - (A \cap B)/(A \cup B)$$
$$= ((\mathbf{x_A}>0) \,\&\, (\mathbf{x_B}>0))/((\mathbf{x_A}>0) \mid (\mathbf{x_B}>0))$$

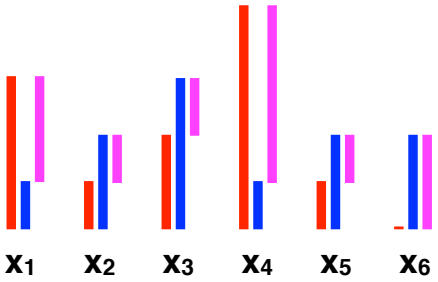


Intuition: Fraction of shared **types** unique to one of the communities

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Bray-Curtis

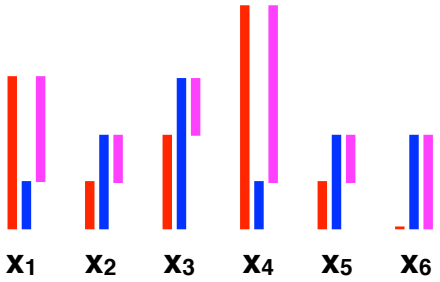
$$Dist(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} = \frac{\quad}{\quad + \quad}$$



Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Bray-Curtis

$$Dist(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} = \frac{\quad}{\quad + \quad}$$

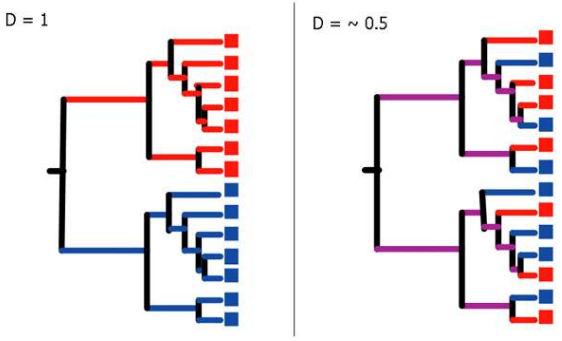


Intuition: *City block distance*. Sum of absolute differences over total abundance.

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Unifrac

$$Dist(x, y) = \frac{\quad + \quad}{\quad + \quad + \quad}$$

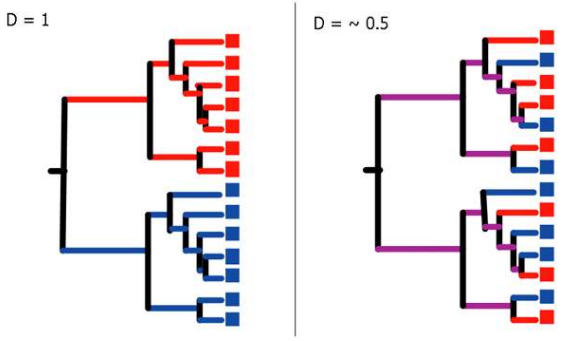


Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone and Knight (2008

Unifrac

$$Dist(x, y) = \frac{\quad + \quad}{\quad + \quad + \quad}$$

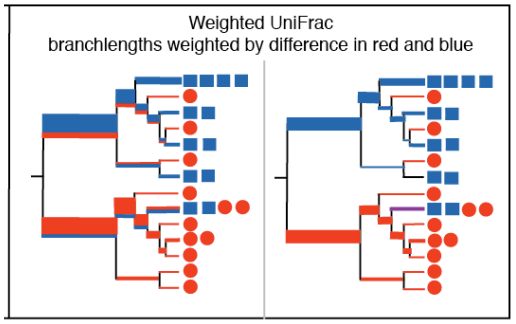


Intuition: Fraction of shared **tree** unique to one of the communities

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone and Knight (2008

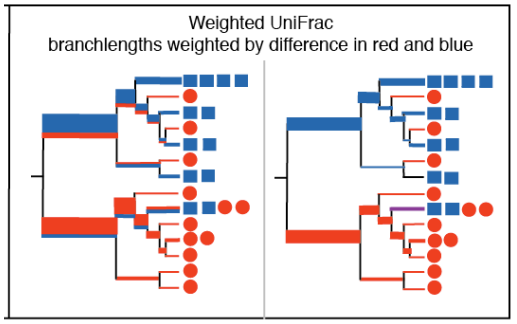
Weighted Unifrac



Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone et al. (2007)

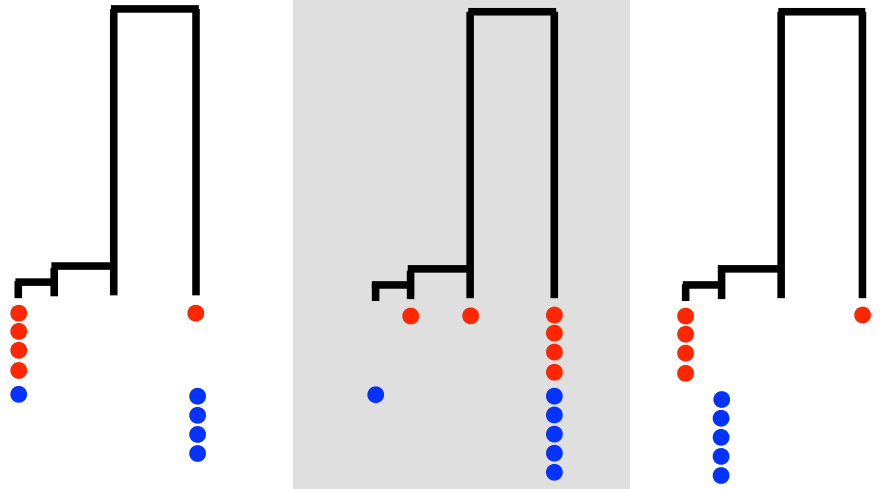
Weighted Unifrac



Intuition: The cost of turning one distribution into the other; where the cost is the amount of "dirt" moved times the distance by which it is moved.

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone et al. (2007)

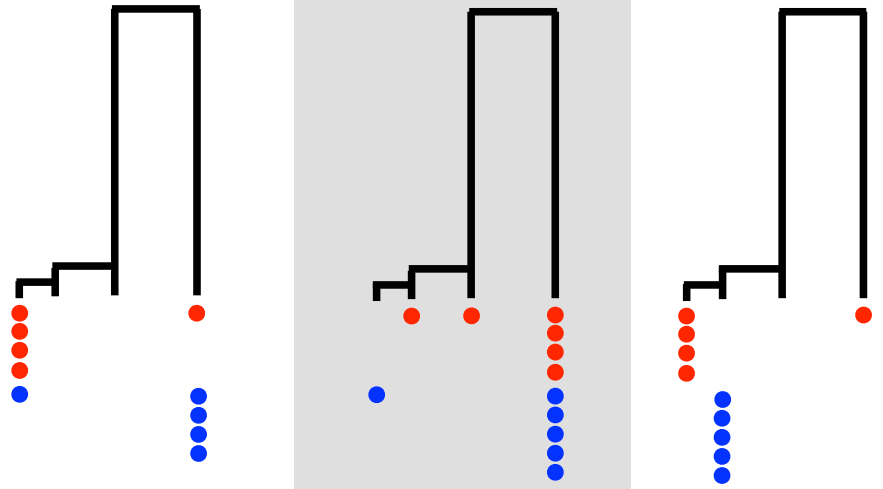


Jaccard:
Bray:
Unifrac:
W-Unifrac:

Jaccard:
Bray:
Unifrac:
W-Unifrac:

Jaccard:
Bray:
Unifrac:
W-Unifrac:

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

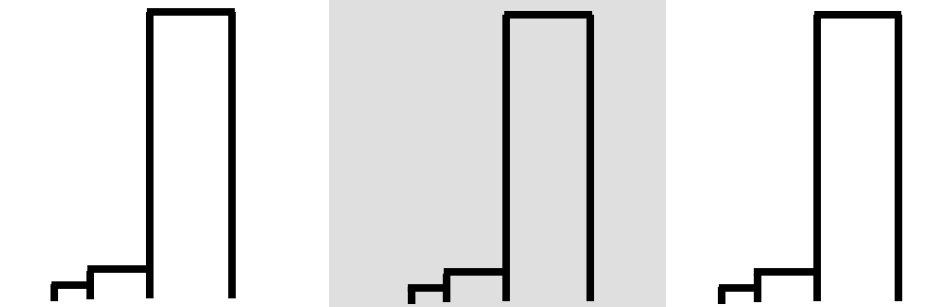


Jaccard: d=0
Bray:
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray:
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray:
Unifrac:
W-Unifrac:

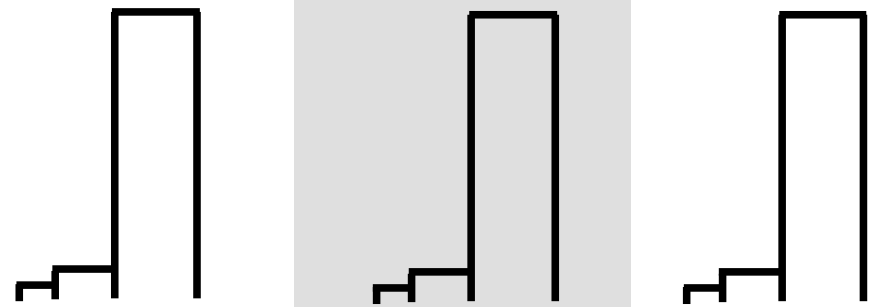
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)



Jaccard: d=0
Bray: Distant
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray: Similar
Unifrac:
W-Unifrac:

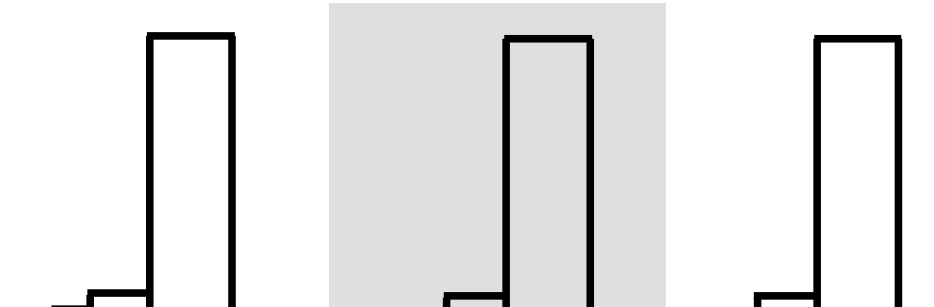
Jaccard: Distant
Bray: Distant
Unifrac:
W-Unifrac:



Jaccard: d=0
Bray: Distant
Unifrac: d=0
W-Unifrac:

Jaccard: Distant
Bray: Similar
Unifrac: Similar
W-Unifrac:

Jaccard: Distant
Bray: Distant
Unifrac: Distant
W-Unifrac:

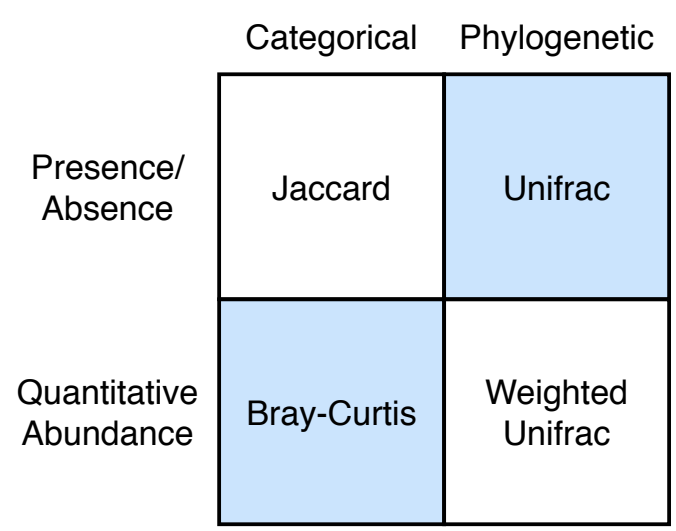


Jaccard: d=0
Bray: Distant
Unifrac: d=0
W-Unifrac: Distant

Jaccard: Distant
Bray: Similar
Unifrac: Similar
W-Unifrac: Similar

Jaccard: Distant
Bray: Distant
Unifrac: Distant
W-Unifrac: Similar

The Distance Spectrum



phyloseq distances

manhattan
euclidean
canberra
bray
kulczynski
jaccard
gower
altGower
morisita-horn
mountford
raup
binomial
chao
cao
jensen-shannon
unifrac
weighted-unifrac
...

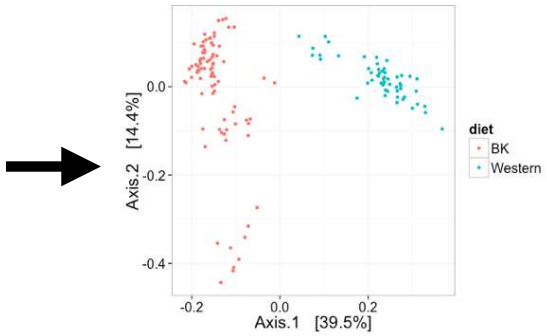
Ordination Methods

Project high-dimensional data onto lower dimensions

P taxa

N samples

0,1,5,1,0,1,2,1,0,0,9,…
7,2,0,0,0,0,0,0,1,0,0,…
0,0,0,0,0,0,8,0,0,0,1,…
0,0,0,1,0,1,2,0,0,0,5,…
0,1,0,2,0,0,0,1,0,0,4,…
0,0,0,1,9,1,2,5,2,0,1,…
0,0,0,0,0,1,2,1,8,0,0,…
0,0,0,0,9,4,0,0,0,0,1,…
.
.



P-dimensions 2-dimensions

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Ordination Methods

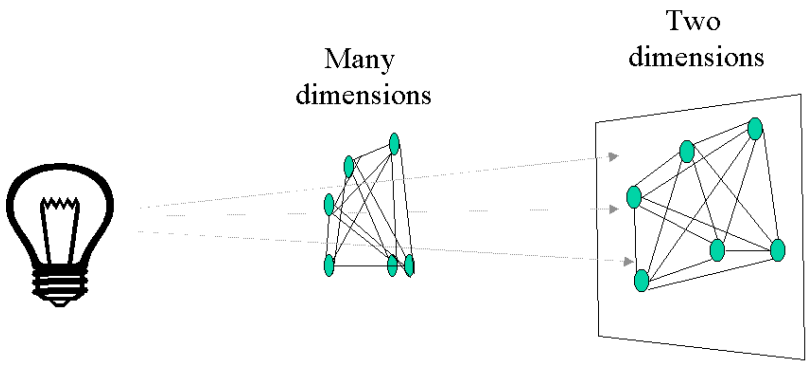
Intuition:

Each PC axis is projection that maximizes the area of the shadow
Equivalently - max(sum of square of distances between points)
Goal: "See" as much variation as possible

Slide graciously provided by Susan Holmes, not necessarily with permission O:-)

Multi-dimensional Scaling

Why MDS? It works with any distance!



Input distance matrix can by Bray-Curtis, Unifrac, …

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

MDS Details

Given distances between each observation (sample), MDS finds the closest approximation of that in lower dimensional Euclidean space.

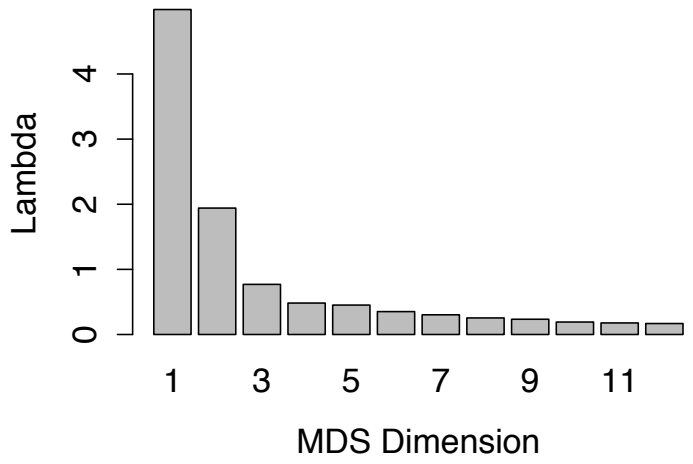
- Algorithm starts from **D** inter-point distances:
  - Center the rows and columns of the distance matrix: $\mathbf{S} = -1/2\ \mathbf{H}\ \mathbf{D}^{(2)}\ \mathbf{H}$
  - Compute SVD by diagonalizing S: $\mathbf{S} = \mathbf{U}\ \mathbf{\Lambda}\ \mathbf{U}^T$
  - Extract Euclidean representations: $\mathbf{X} = \mathbf{U}\ \mathbf{\Lambda}^{1/2}$
- The relative values of diagonal elements of Λ gives the proportion of variability explained by each of the axes.
- The valued of Λ should always be looked at in deciding how many dimensions to retain

---

NMDS is similar, but minimizes a different function (difference in distance ranks)

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

MDS Scree Plot



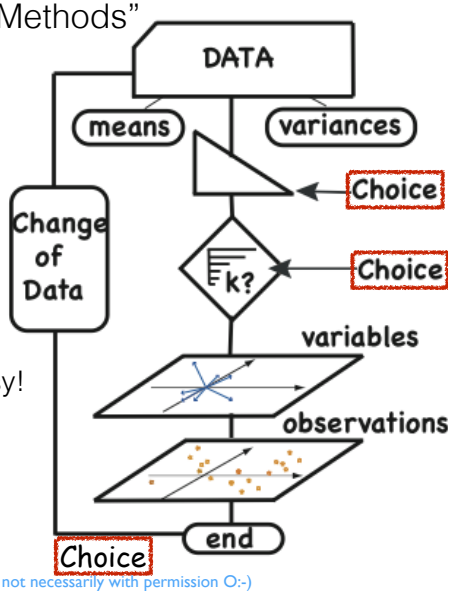
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

Best Practices

- Looking for patterns (the "I-test")
- Always look at scree plot
- Biplot (if legible)
- Use multiple distances
  - For which D is pattern strongest?
- phyloseq (and R/Rmd) make this easy!



Slide graciously provided by Susan Holmes, not necessarily with permission O:-)

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

What we "learn" depends on the data.

- How many axes are probably useful?
- Are their clusters? How many?
- Are their gradients?
- Are the patterns consistent with covariates
- (e.g. sample observations)
- How might we test this?

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

- Are their clusters? How many?

Technique:
Gap Statistic

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

- Are their **gradients**?
- Are they explained by one or more sample covariates?

Technique:
PC regression (statistics' PCR)

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

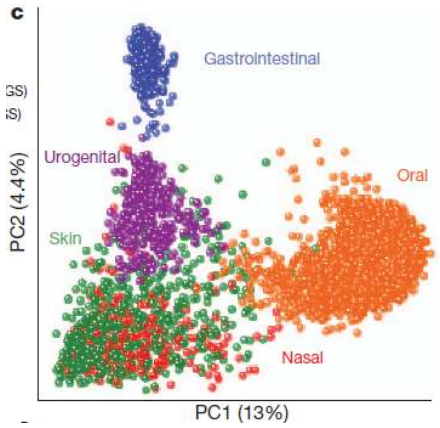
- Are the patterns consistent with covariates?

Technique:
Permutational Multivariate ANOVA
vegan::adonis( )

End:
Introduction to Microbiome / Metagenome Analysis Concepts

Questions?

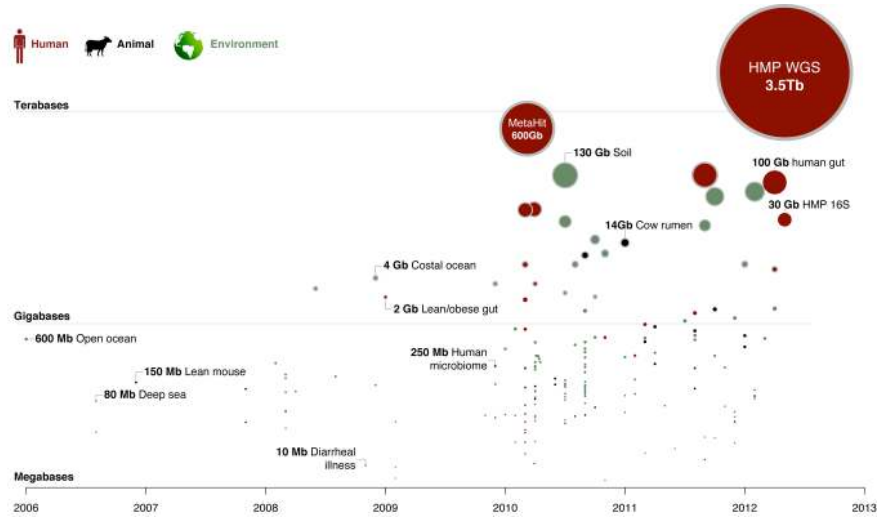
Introduction to Microbiome / Metagenome Analysis Tools and Practices



Introduction to Microbiome / Metagenome Analysis Tools and Practices

1. Probably-not-comprehensive summary of metagenomic tools
2. Short sermon on the virtues of reproducible analysis
3. Introduction to phyloseq & send-off this afternoon's lab

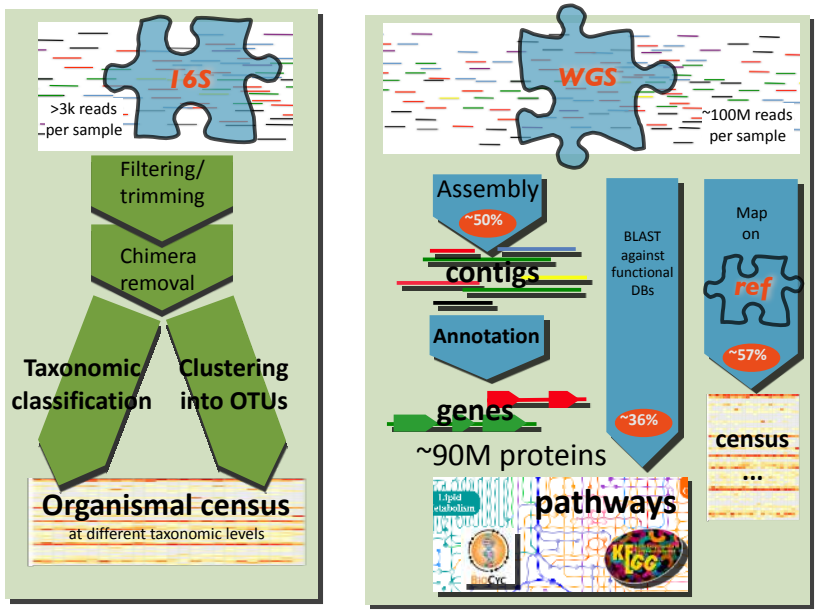
Figure 1. Timeline of microbial community studies using high-throughput sequencing.



Gevers D, Knight R, Petrosino JF, Huang K, et al. (2012) The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol 10(8): e1001377. doi:10.1371/journal.pbio.1001377
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001377

PLOS BIOLOGY
TENTH ANNIVERSARY

Slide graciously provided by Dirk Gevers, not necessarily with permission O:-)



Slide graciously provided by Dirk Gevers, not necessarily with permission O:-)

16S rRNA Databases

- GreenGenes - http://greengenes.secondgenome.com
- Silva - www.arb-silva.de
- Ribosomal Database Project (RDP) - https://rdp.cme.msu.edu

  - ~100Ks - millions of unique 16S rRNA genes
  - Curated taxonomy
  - Classification tools (e.g. RDP classifier, ARB, etc.)

(16S rRNA) Amplicon Sequence Processing Tools:

- QIIME (Soon 'Qiita') - http://qiime.org/
- mothur - www.mothur.org/
- usearch - www.drive5.com/usearch
- DADA2 - https://github.com/benjjneb/dada2

Afternoon will be spent using QIIME
Daniel has much more to say about it…

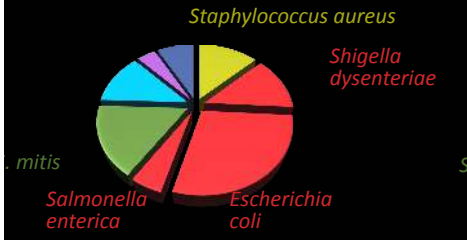
How to find biology in your meta'ome

- Looking for ecology?
  - Diversity metrics, k-mer analysis, curve fitting
- Looking for specific bugs?
  - Assembly: +novelty, -difficulty
  - Mapping: +speed/ease, -novelty
- Looking for specific processes?
  - Intrinsic annotation: +novelty, -difficulty
  - Extrinsic annotation: +sensitivity, -novelty
- Looking for variants?
  - Clustering: +specificity, -difficulty
  - Mapping: +sensitivity, -novelty
- **What else?**

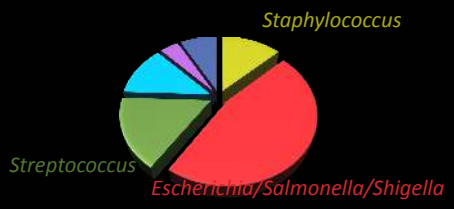
Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

Goal: expose strain level features

The picture using shotgun metagenomics and MetaPhlAn



The picture using 16S rRNA sequencing



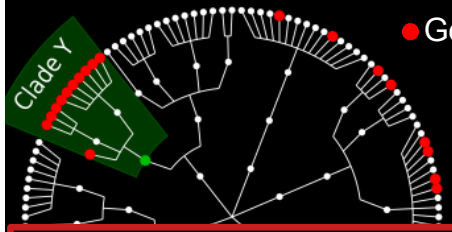
Next step: strain-level profiling

(i) Identify
(ii) Track (e.g. across samples)
(iii) Characterize (genomically)

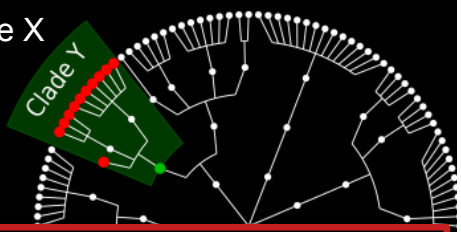
Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

MetaPhlAn: Taxonomic profiling using unique marker genes

X is a core gene for clade Y



X is a unique marker gene for clade Y

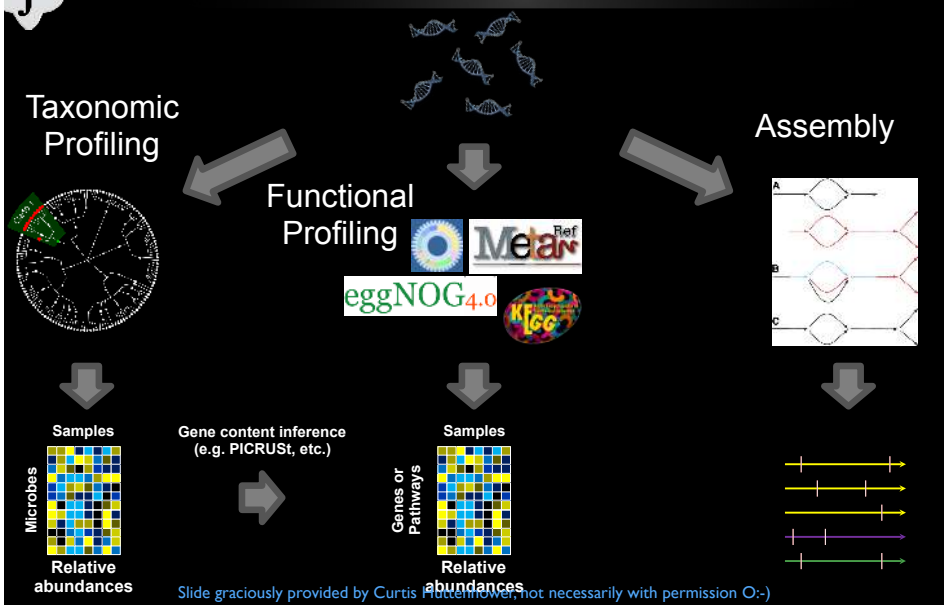


- ~1M most representative markers used for identification
  - 184±45 markers per species (target 200)
- ~7,100 species (excludes incomplete annotations, spp., etc.)
- False positive/False negative rates of ~1 in 10^6
- Profiles all domains of life: bacteria, viruses, euks, archaea
- Strain level profiling using marker barcodes and SNPs
- Quasi-markers used to resolve ambiguity in postprocessing

http://huttenhower.sph.harvard.edu/metaphlan

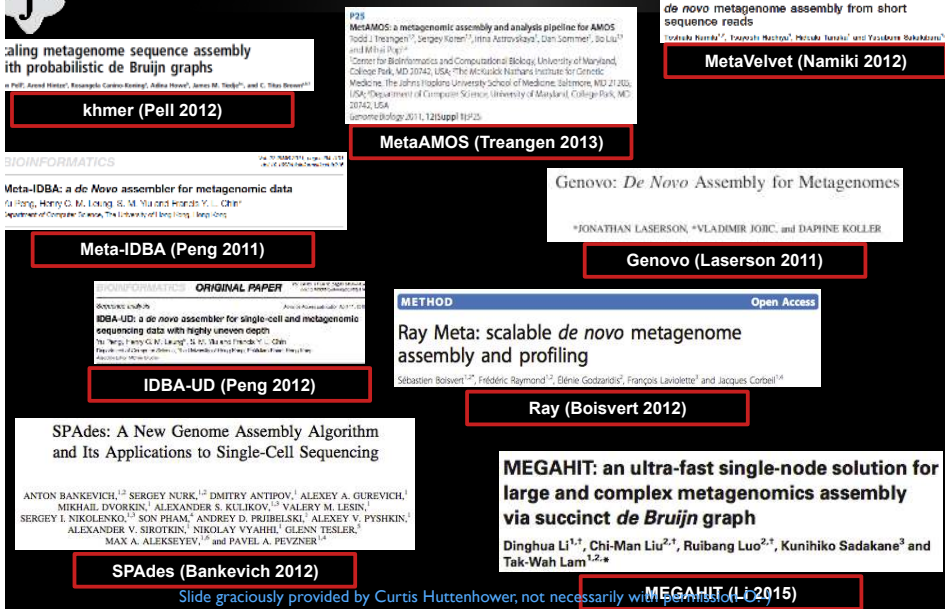
Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

Typical shotgun metagenome and metatranscriptome analyses



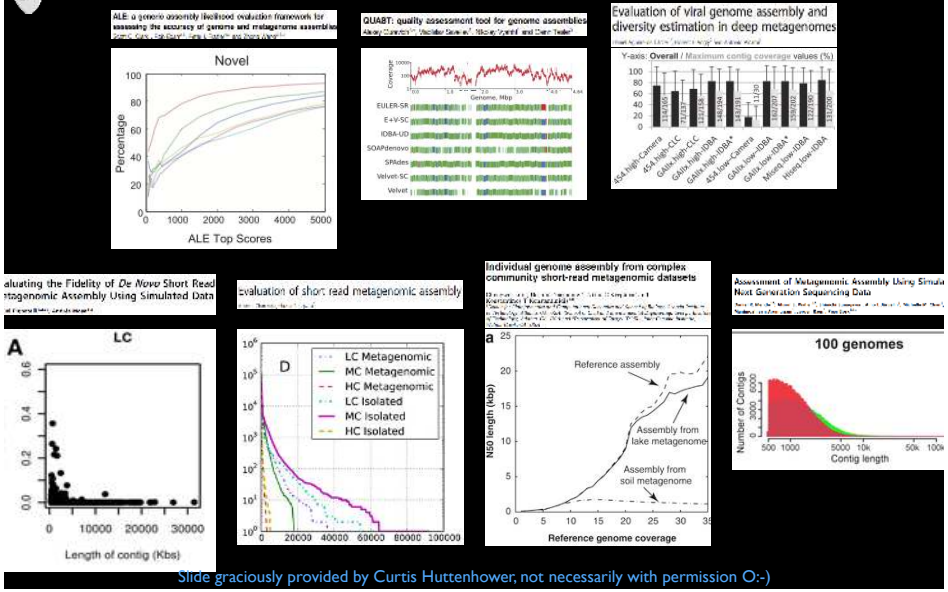
Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

Microbiome meta'omic analyses: assembly



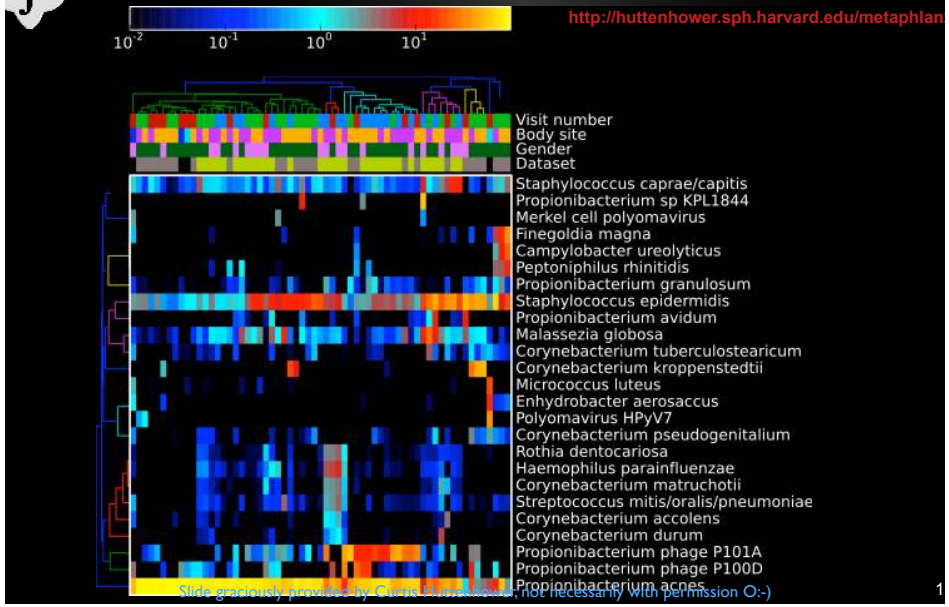
Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

Microbiome meta'omic analyses: assembly



Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

MetaPhlAn2: Trans-kingdom profiling



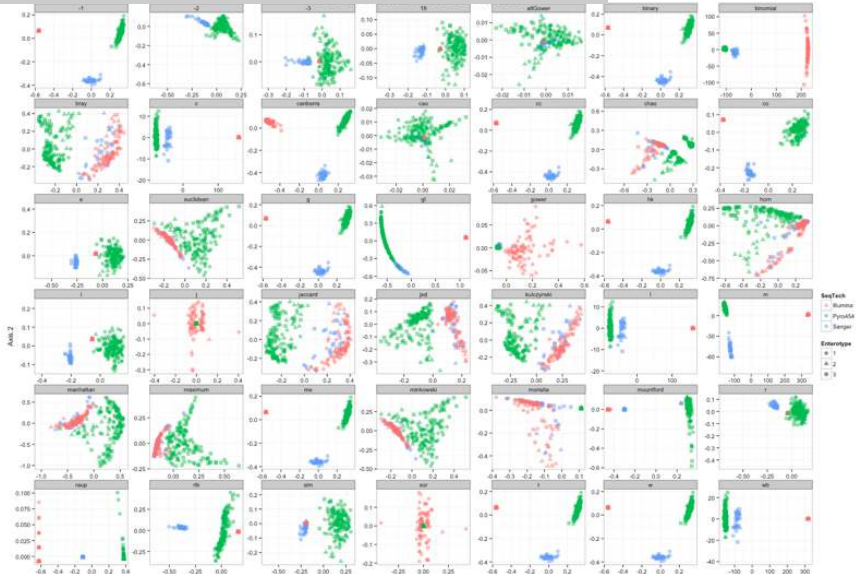
Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

Reproducible analysis of microbiome / metagenome data

- Why make the effort?
- What if I don't want someone else reproducing my analysis?
- What if I don't know how?
- Isn't it enough to provide a cursory description in the methods section with a light sprinkling of literature citations?

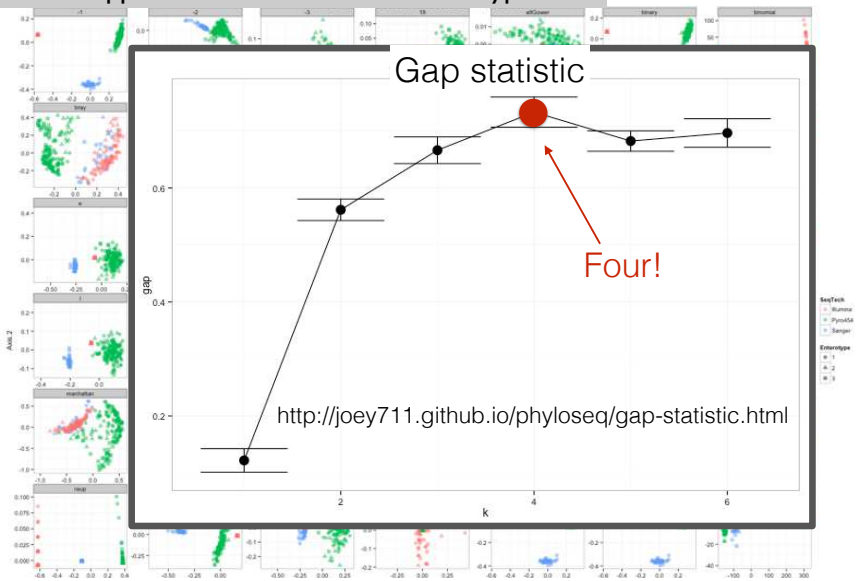
illustrative example favoring reproducible analysis: "Enterotypes of the human genome"

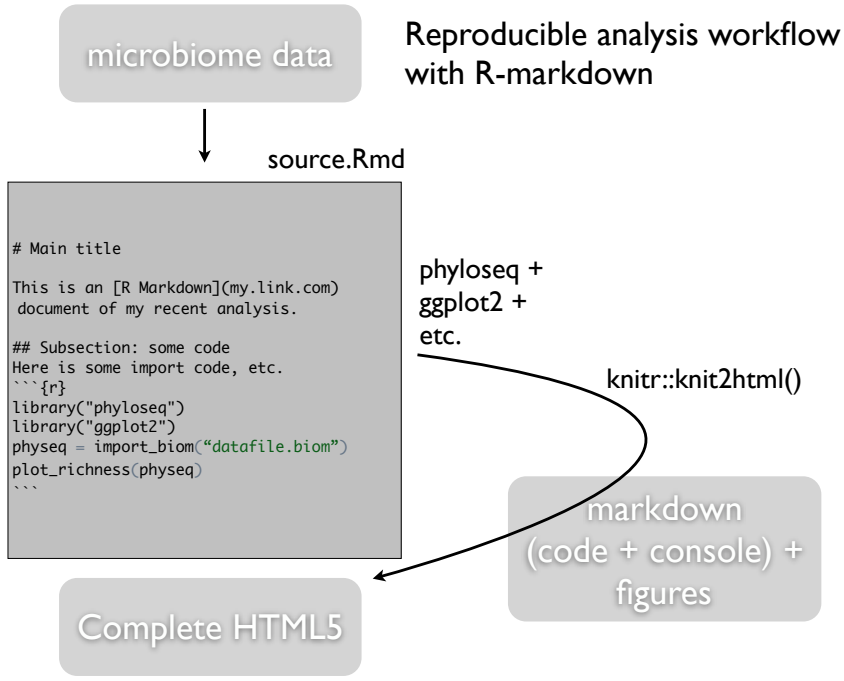
MDS on supported distance metrics: enterotype data



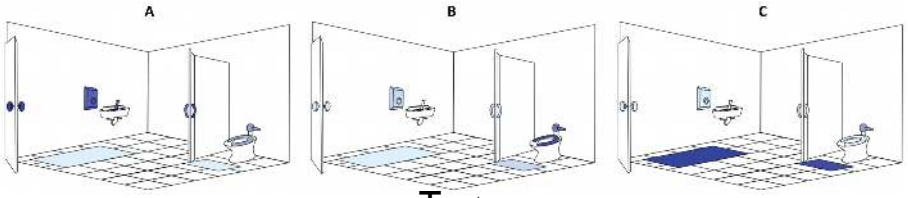
illustrative example favoring reproducible analysis: "Enterotypes of the human genome"

MDS on supported distance metrics: enterotype data





Reproducible analysis workflow with R-markdown



Text

joey711.github.io/phyloseq-demo/Restroom-Biogeography

phyloseq

OPEN ACCESS Freely available online

PLOS ONE

phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

Paul J. McMurdie, Susan Holmes*

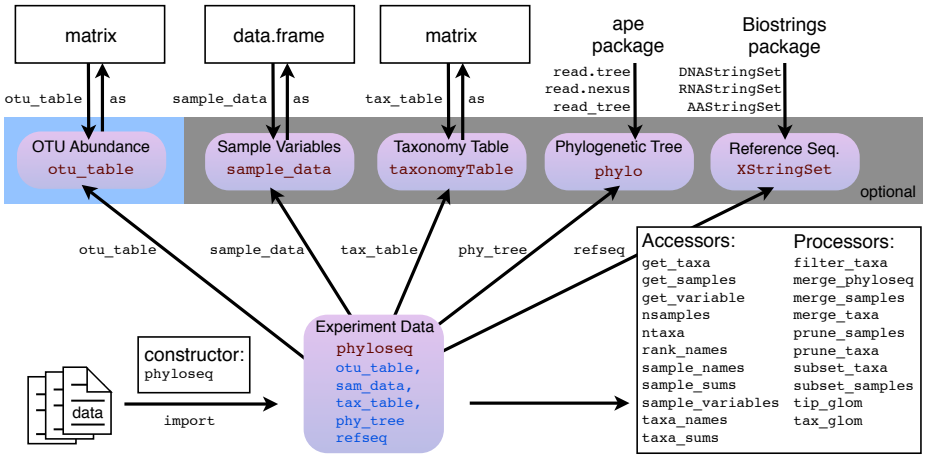
Department of Statistics, Stanford University, Stanford, California, United States of America

Key Packages:

vegan
ape
distory
phangorn
picante
metagenomeSeq

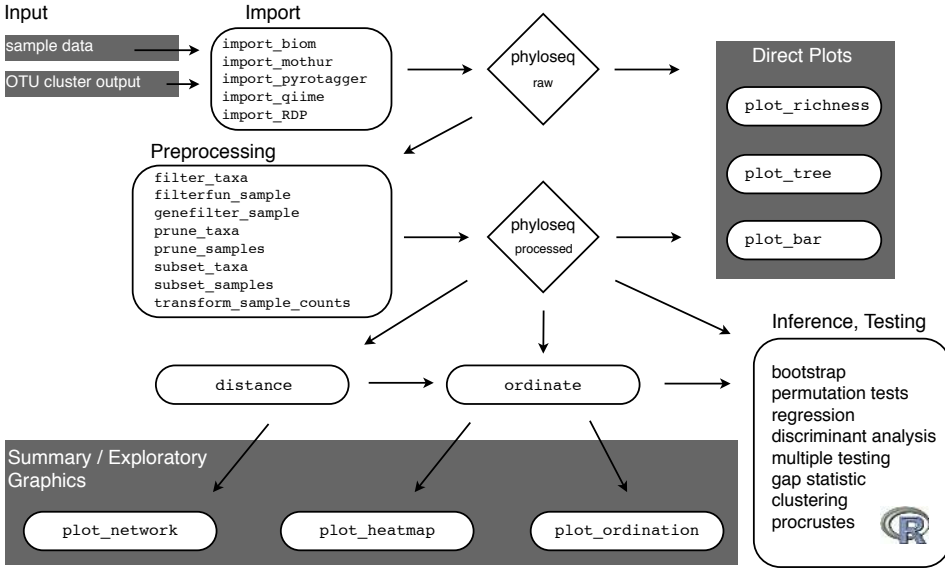
phyloseq

data structure & API



phyloseq

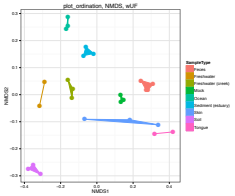
work flow



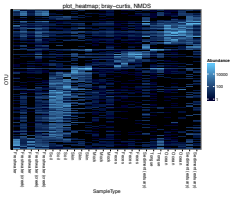
phyloseq

graphics

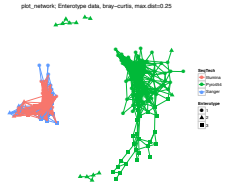
plot_ordination()



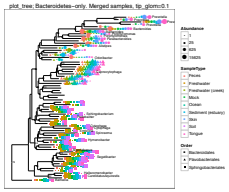
plot_heatmap()



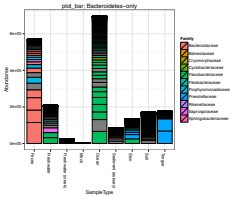
plot_network()



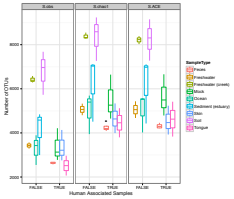
plot_tree()



plot_bar()



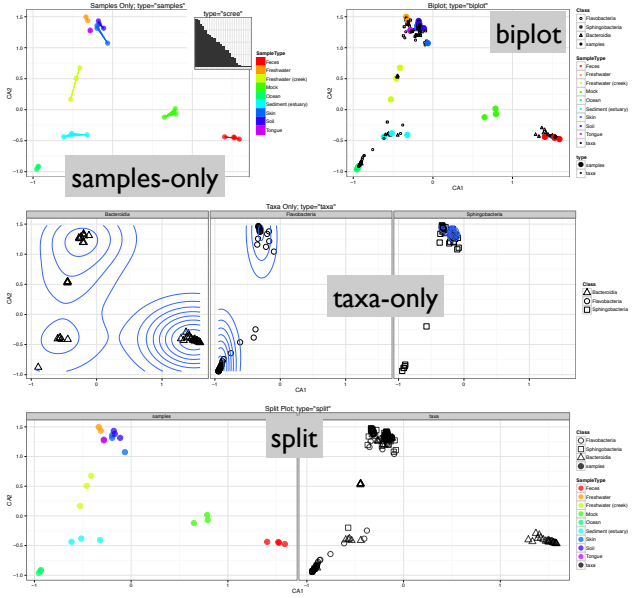
plot_richness()

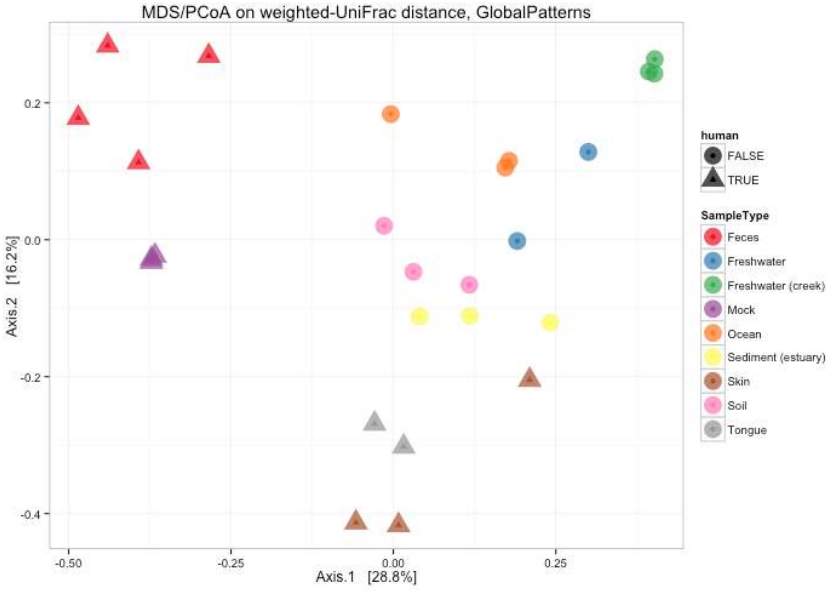


phyloseq

graphics

plot_ordination()

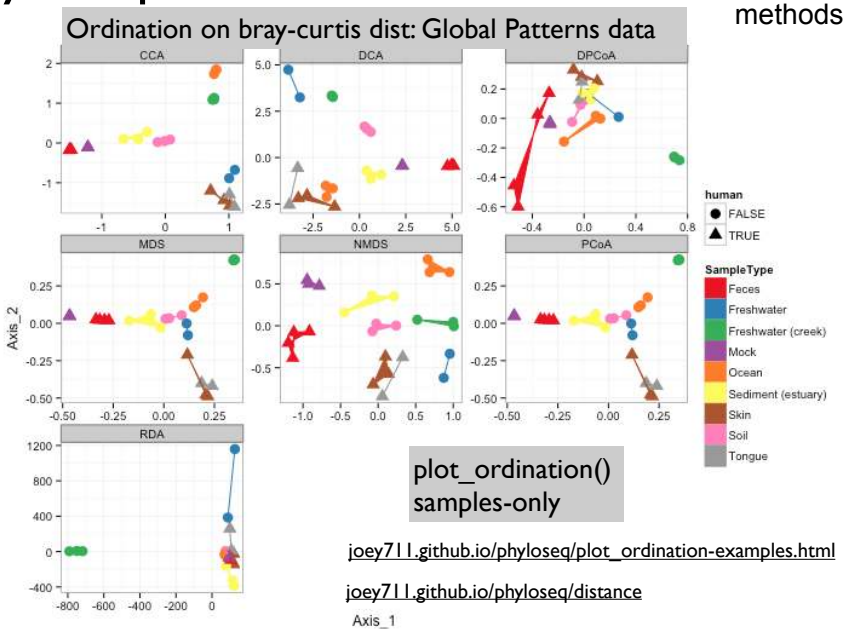





```
ordu = ordinate(GP1, "PCoA", "unifrac", weighted = TRUE)
plot_ordination(GP1, ordu, color = "SampleType", shape = "human")
```

phyloseq

supported ordination methods



plot_ordination()
samples-only

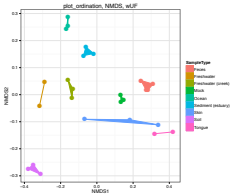
joey711.github.io/phyloseq/plot_ordination-examples.html

joey711.github.io/phyloseq/distance

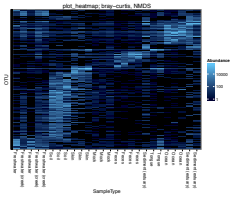
phyloseq

graphics

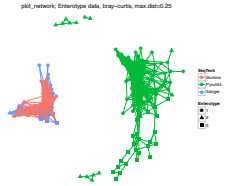
plot_ordination()



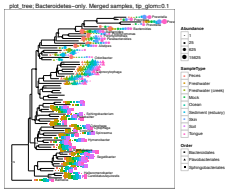
plot_heatmap()



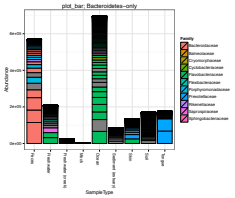
plot_network()



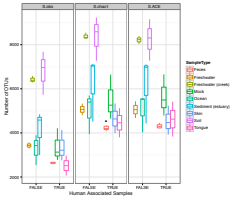
plot_tree()

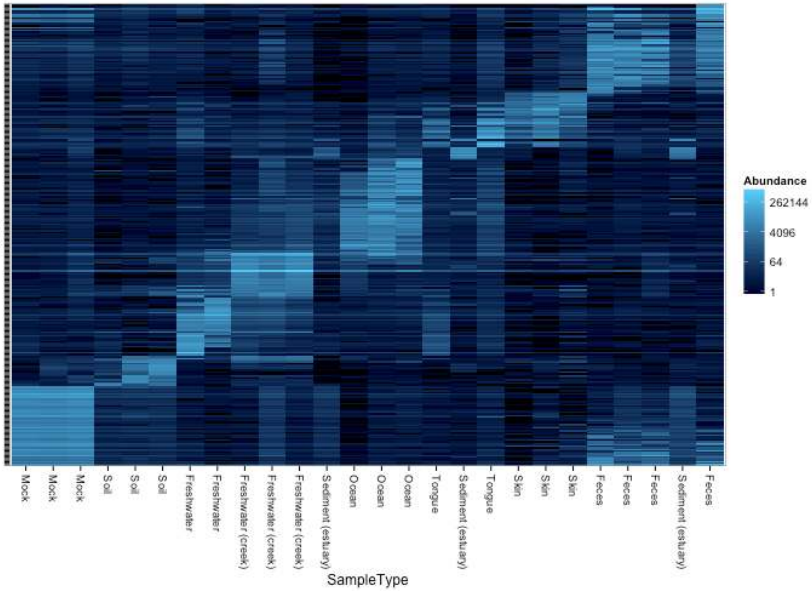


plot_bar()



plot_richness()

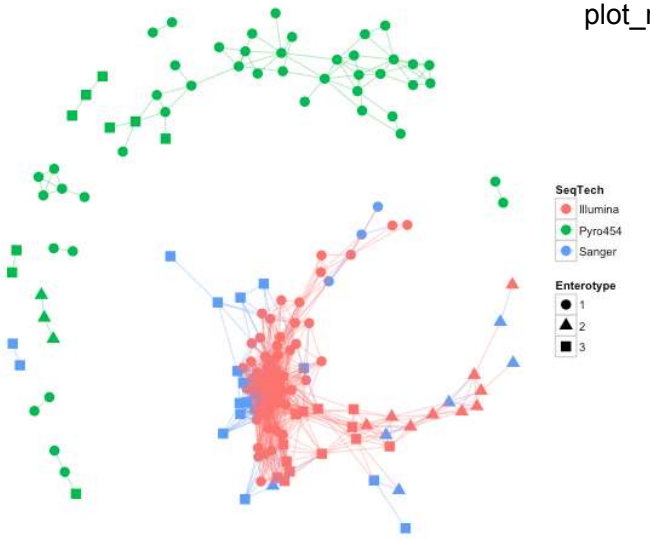





```
gpt <- subset_taxa(GlobalPatterns, Kingdom == "Bacteria")
gpt <- prune_taxa(names(sort(taxa_sums(gpt), TRUE)[1:300]), gpt)
plot_heatmap(gpt, sample.label = "SampleType")
```

joey711.github.io/phyloseq/plot_heatmap-examples.html

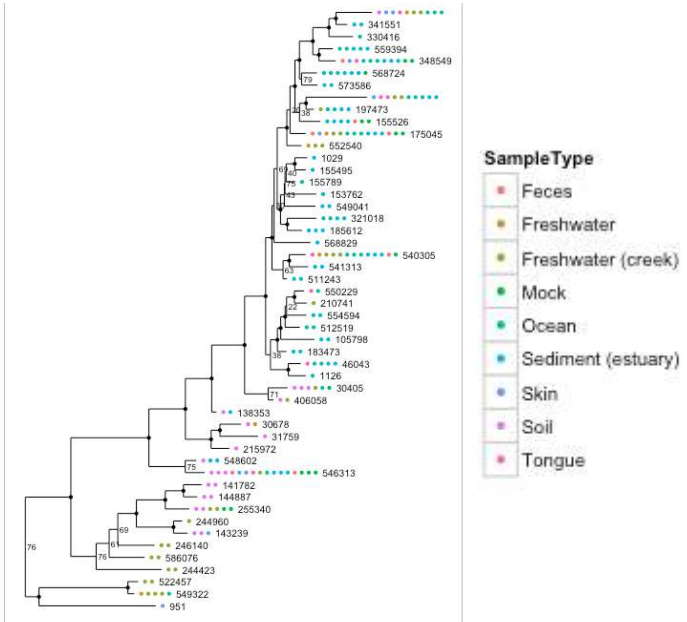
plot_network()




```
ig <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)
plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
 line_weight = 0.4, label = NULL)
```

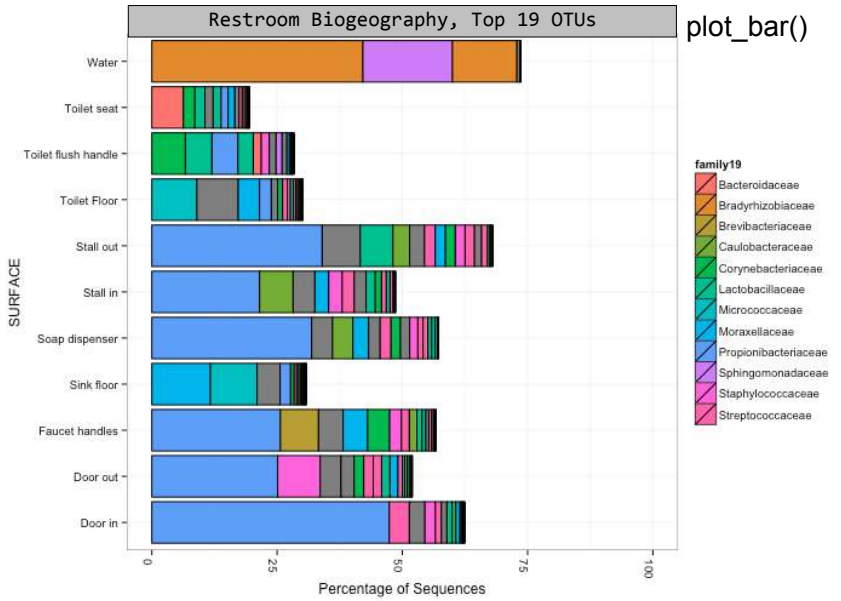
joey711.github.io/phyloseq/plot_network-examples.html

plot_tree()




```
plot_tree(physeq, nodelabf=nodeplotboot(80, 0, 3), color="SampleType",
          label.tips="taxa_names", ladderize="left")
```

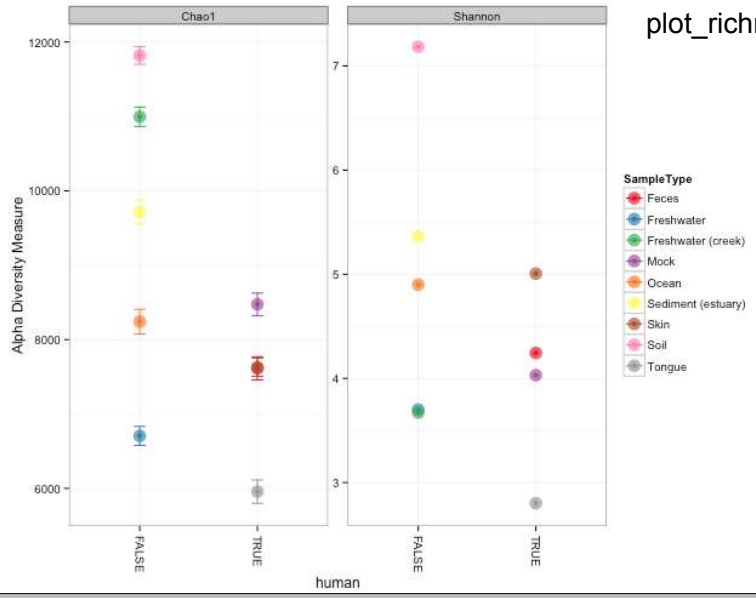
joey711.github.io/phyloseq/plot_tree-examples.html



plot_bar()

```
plot_bar(restroomRm19, "SURFACE", fill = "family19", title = title) + coord_flip()
+ ylab("Percentage of Sequences") + ylim(0, 100)
```

joey711.github.io/phyloseq-demo/Restroom-Biogeography.html



plot_richness()

```
GPst = merge_samples(GP, "SampleType")
 = plot_richness(GPst, x="human", color="SampleType", measures=c("Chao1", "Shannon"))
 + geom_point(size = 5, alpha = 0.7)
```

joey711.github.io/phyloseq/plot_richness-examples.html

Schedule for today

Sec	Day	Start	End	Topic	Lead Instr.
1	Mon	09:00	10:00	Introduction to Metagenomics. Culture independent techniques, 16S rRNA, etc. (60 -75 min)	Joey
2	Mon	10:00	11:00	Introduction to microbiome analysis concepts -- Exploratory data analysis, Distances, PCoA, Ordination, taxa & sample-level inferences (75 min)	Joey
3	Mon	11:00	11:59	Introduction to microbiome analysis practices: QIIME, phyloseq, reproducible research (30 min)	Joey
---	Mon	12:00	14:00	Lunch (120min)	---
4	Mon	14:00	17:00	QIIME Lab (180min)	Daniel
---	Mon	17:00	19:00	Dinner (120min)	---
5	Mon	19:00	22:00	phyloseq Lab (180min)	Joey