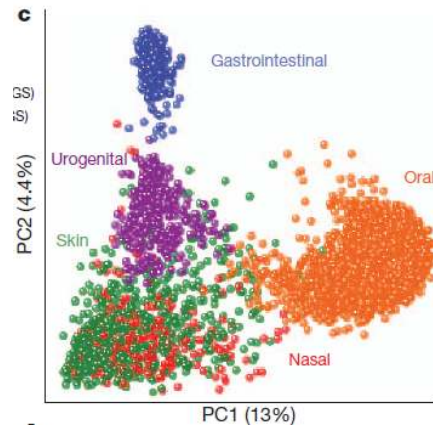


An Introduction to Metagenomics

Paul “Joey” McMurdie, PhD
Whole Biome, Inc.



1

Schedule for today

Sec	Day	Start	End	Topic	Lead Instr.
1	Tues	09:00	10:00	Introduction to Metagenomics. Culture independent techniques, 16S rRNA, etc. (60 -75 min)	Joey
2	Tues	10:30	11:29	Introduction to microbiome analysis concepts -- Exploratory data analysis, Distances, PCoA, Ordination, taxa & sample-level inferences (75 min)	Joey
3	Tues	11:30	11:59	Introduction to microbiome analysis practices: QIIME, phyloseq, reproducible research (30 min)	Joey
---	Tues	12:00	14:00	Lunch (120min)	---
4	Tues	14:00	17:00	QIIME Lab (180min)	Daniel
---	Tues	17:00	19:00	Dinner (120min)	---
5	Tues	19:00	22:00	phyloseq Lab (180min)	Joey

2

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

3

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

4

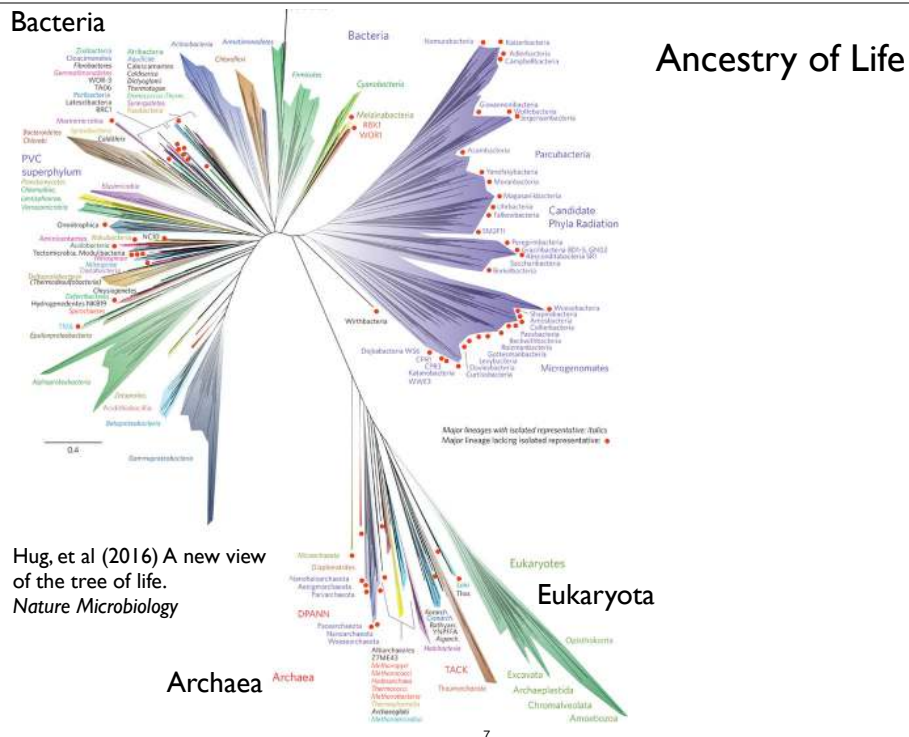
Outline for *morning* lecture:

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

5

“All of the visible organisms that we’re familiar with, everything that springs to mind when we think of ‘nature’, are latecomers to life’s story. They are part of the coda. For most of the tale, microbes were the only living things on Earth.”

6



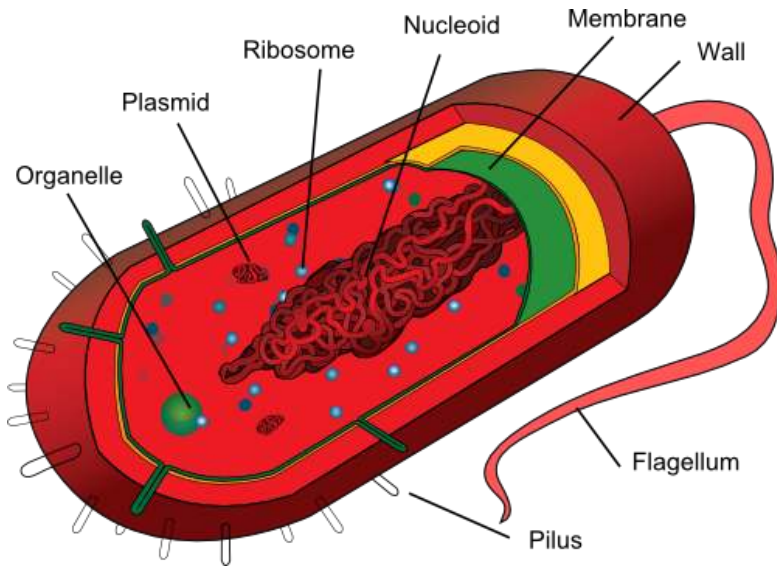
All of animal evolution and development has occurred in the presence of microbes.

- Germ-free mice:
 - grow slower,
 - live shorter,
 - have dysfunctional GI and immune systems
 - are more susceptible to stress and infections
 - 1965 Dubious, repeated many times since
 - This observation generalizes to virtually all animals, at varying degrees
- Without microbes:
 - Horrible maladies for most animals (esp. development, metabolism)
 - Most animal species would become extinct within a year (estimate)
 - There would be (almost) no oxygen in the atmosphere
 - ocean microbes alone account for ~half of your O₂
 - We’d all quickly die of CO₂ poisoning (and later global warming)
 - Most elemental cycles are predominantly microbe-driven

8

What are microbes?

Cell structure



9

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

10

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!

11

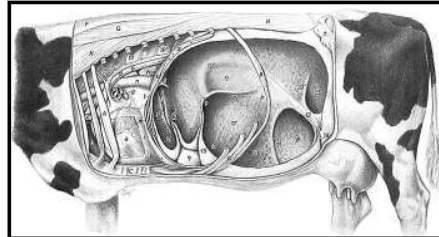
Environmental Science

- Critical elemental cycles (carbon, nitrogen, sulfur, iron, ...)
- Pollution control, cleanup
- Ecology / Evolution (chloroplasts, mitochondria, symbiosis, competition, ...)

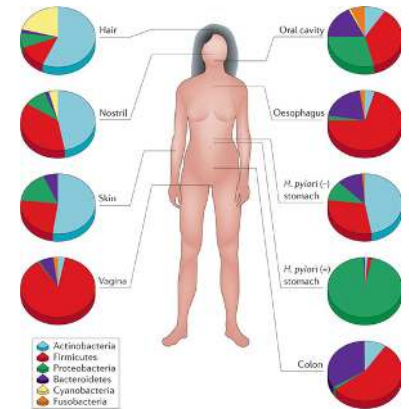
12

Industrial Applications

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



13



Nature Reviews | Genetics

Human Health

- Protection from pathogens (e.g. *Clostridium difficile*)
- Cancer
- Absorption/Production of nutrients in the gut (obesity, T2D)
- Development/regulation of immune system, e.g. chronic diseases (T1D, RA, IBD, other autoimmune, UTIs, periodontitis, ...)

14

of diversity in the human microbiome

Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina

Staphylococcus epidermidis colonizes external body sites

○ Commensal microbes
★ Potential pathogens

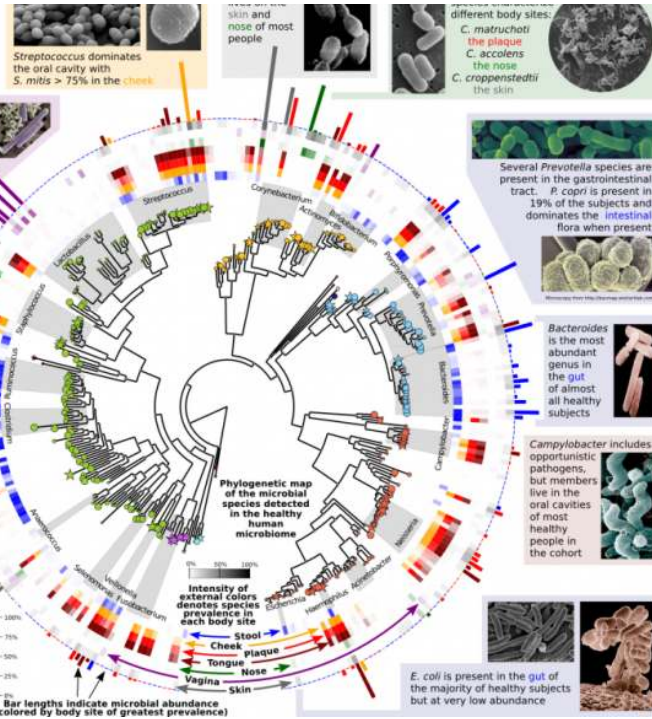
The four most abundant phyla

- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health Human Microbiome Project



15

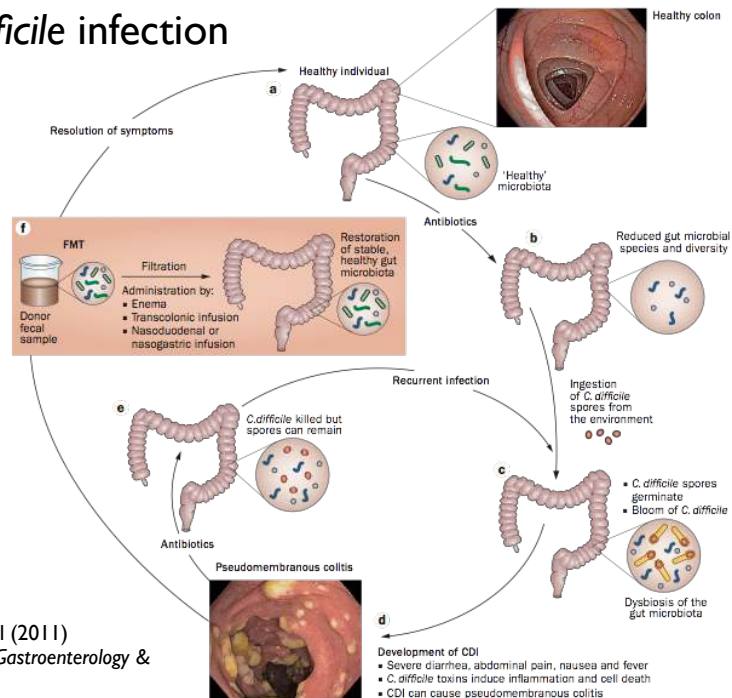
Microbes can...

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?

16

C. difficile infection

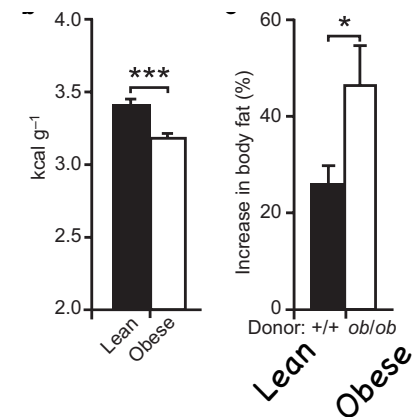


Borody, et al (2011)
Nature Rev Gastroenterology & Hepatology

17

Microbes can make you fat(ter)...

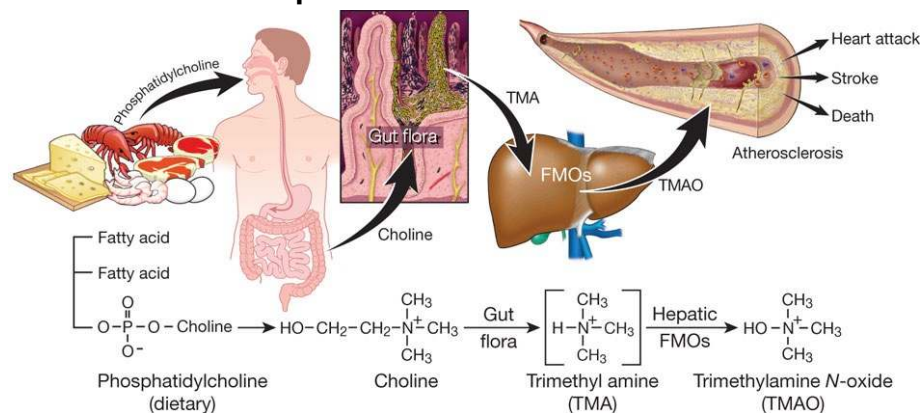
- Lean (n = 10) & obese donors (n=9)
- Colonization of germ-free wild-type mice with microbiota from obese donors causes significant increase in total body fat
- Total body fat content was measured before and after a 2-week colonization
- Confirm that the ob/ob microbiome has an increased capacity for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



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

18

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*.

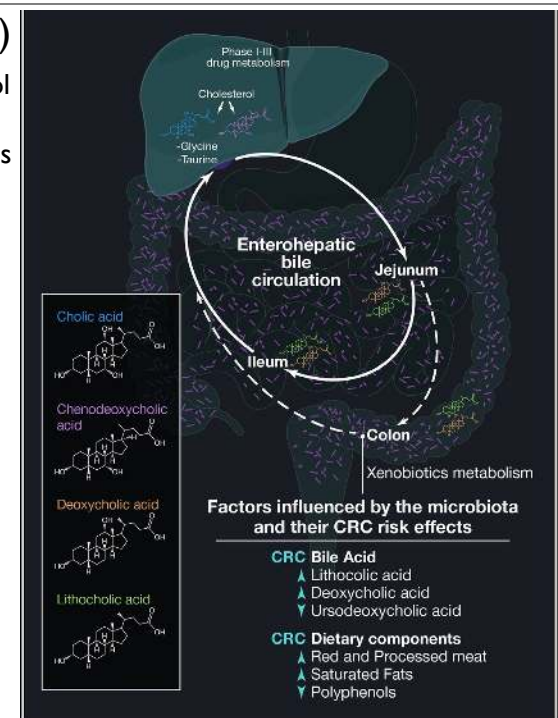
ZN Wang, ..., Stanley Hazen. *Nature* **472**, 57-63 (2011)

Fogelman, A. M. (2015). TMAO Is Both a Biomarker and a Renal Toxin. *Circulation Research*.

19

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

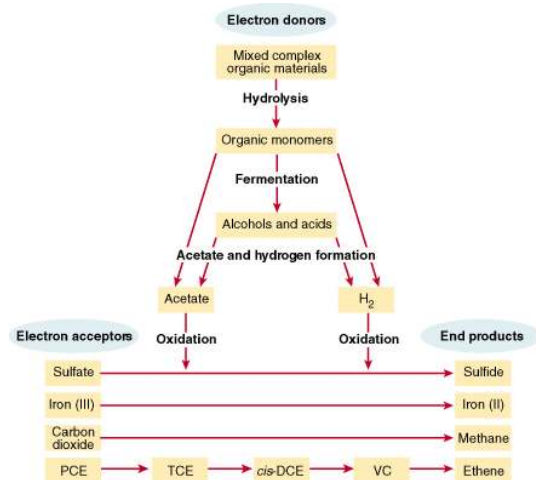


Sears, C. L., & Garrett, W. S. (2014). Microbes, Microbiota, and Colon Cancer. *Cell Host & Microbe*, 15(3), 317-328.

20

Groundwater: Chlorinated Solvents

Dehalococcoides



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

21

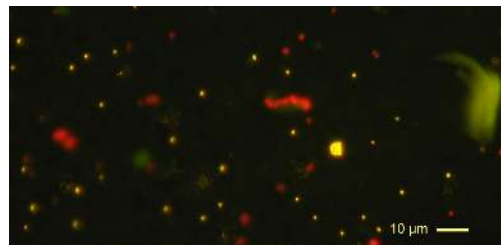
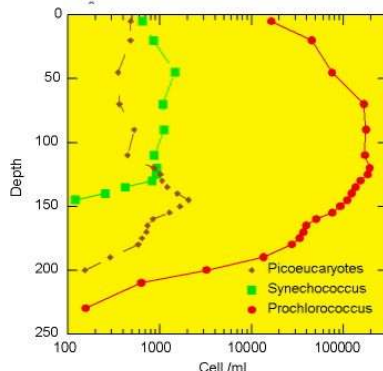
Bonus microbiome show-and-tell

22

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

OLIPAC cruise
Pacific Ocean 1994 Oligotrophic 16°S



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

23

Yellowstone National Park



Octopus Spring

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



Obsidian Pool

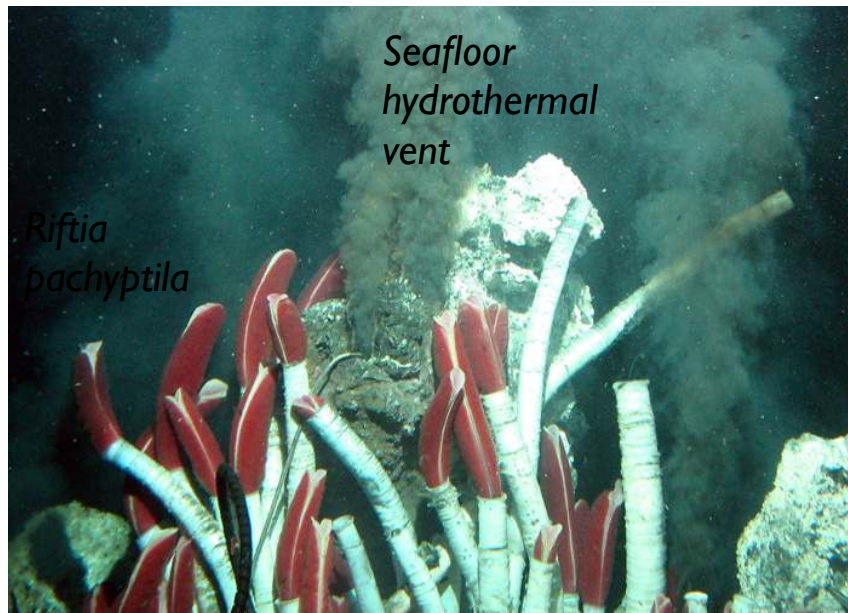
- 75° – 95°C
- high iron (II) hydrogen sulfide
- extensive diversity (previously unknown)

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.

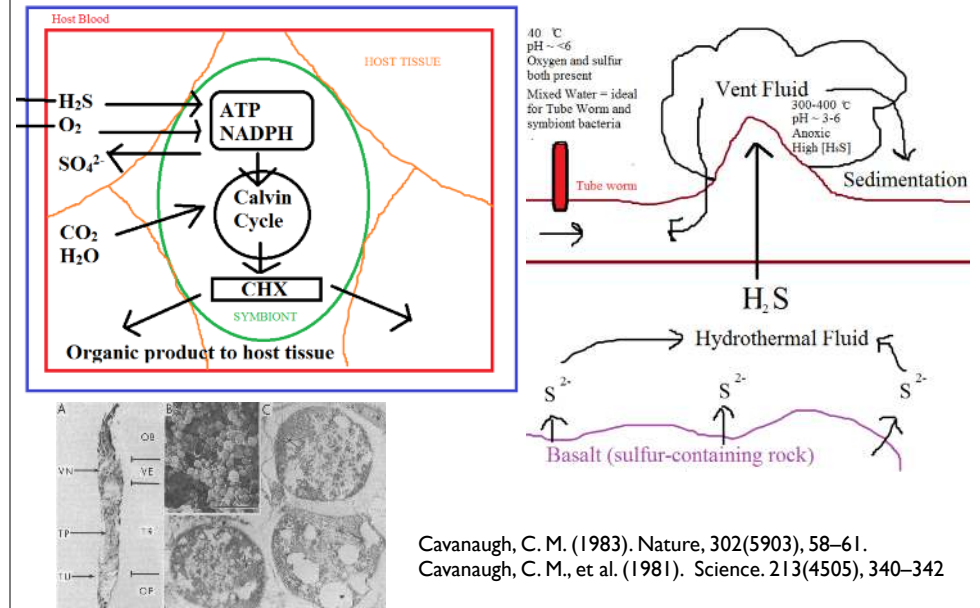
24

Symbiosis: sea-floor vent tube worm



25

Symbiosis: sea-floor vent tube worm



26

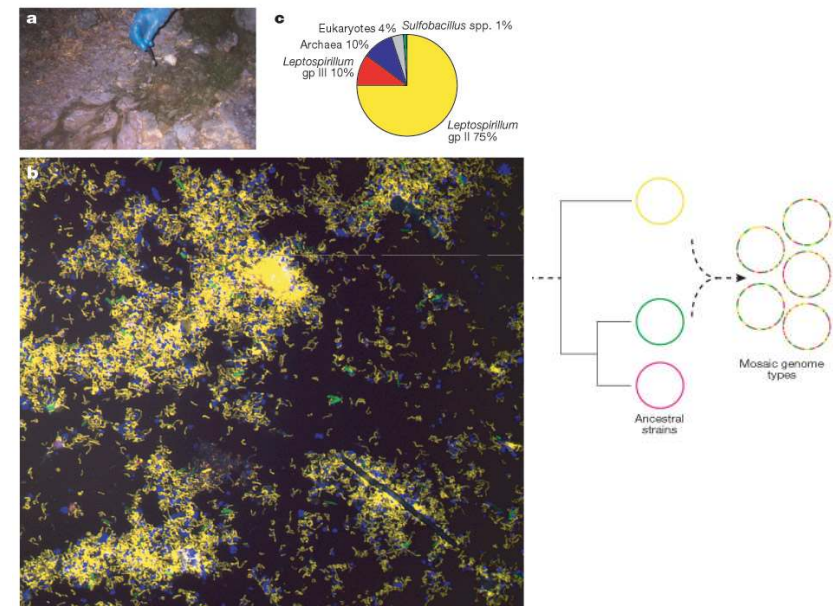
Example of “model” microbiome: acid mine biofilm



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

27

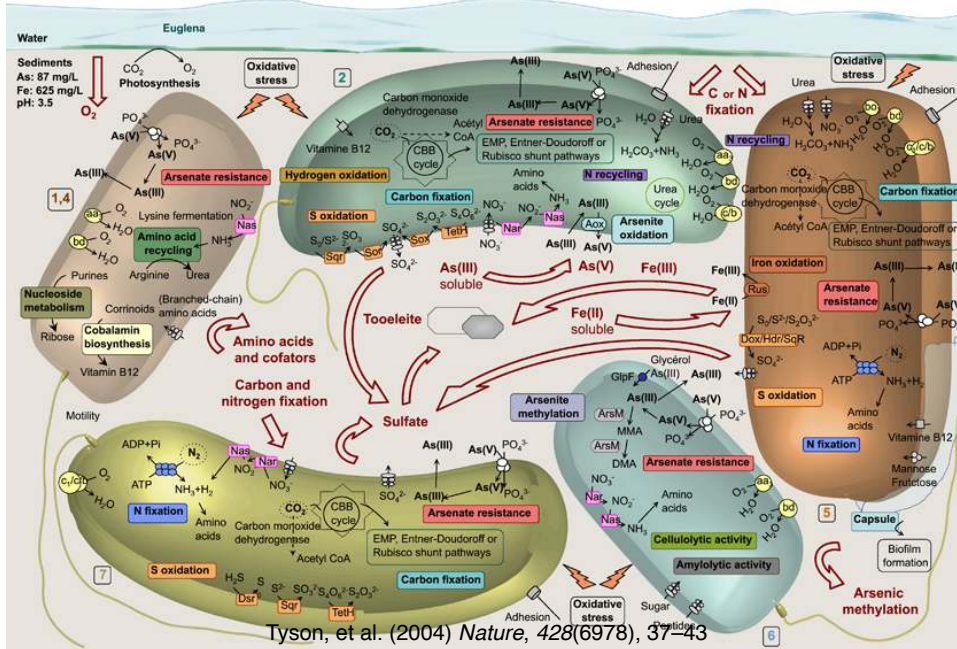
acid mine biofilm



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

28

acid mine biofilm



29

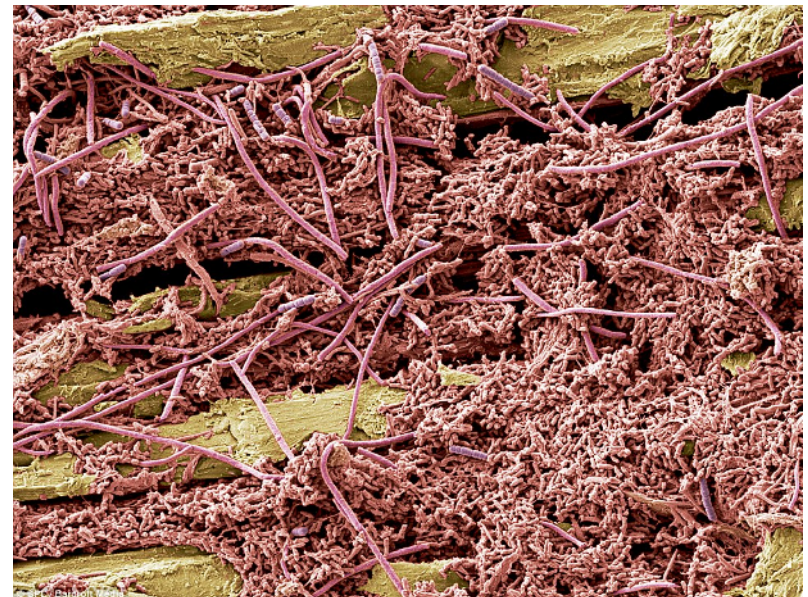
End: Biological Motivation

Questions before moving on?

30

Metagenomics Experimental Methods

Exercise: How many species are present?



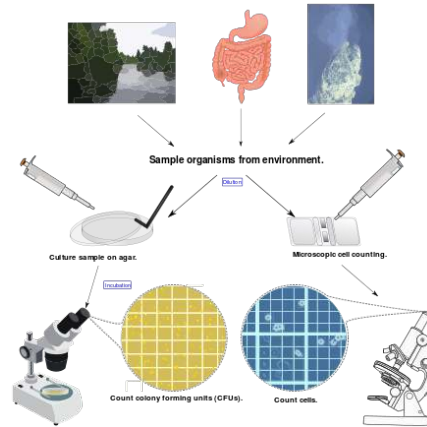
Confer amongst yourselves. We'll take a poll.

31

32

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**



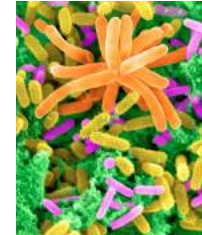
Staley, J.T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, 39, 321–346.

33

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”...

34

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

35

- 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*
- 1998 *metagenome* genomics of communities coined by Jo Handelsman
- 2001 *microbiome* coined by Joshua Lederberg

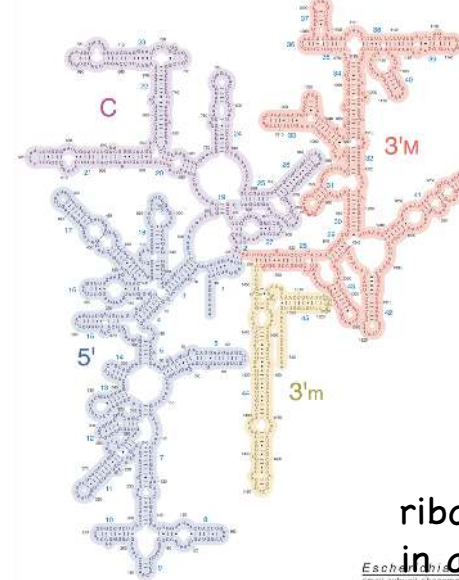
36

Woese was originally scorned at the discovery of archaea via rRNA gene (dis)similarity.

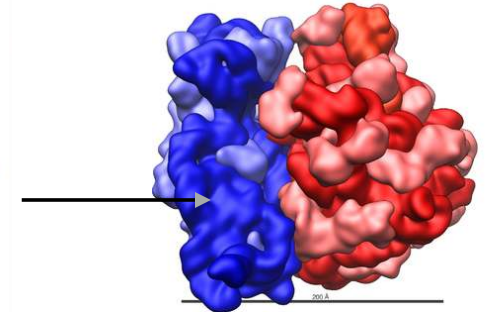
History of modern metagenomics/microbiome research is deeply tied to modern molecular ecology

37

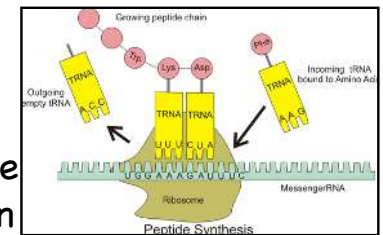
Small subunit "16S" rRNA



ribosome



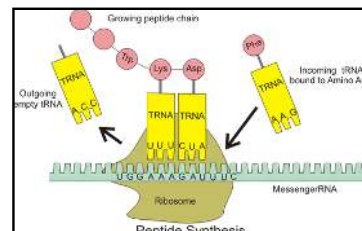
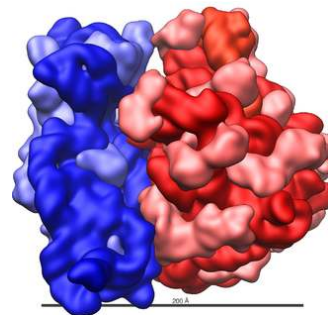
ribosome in action



38

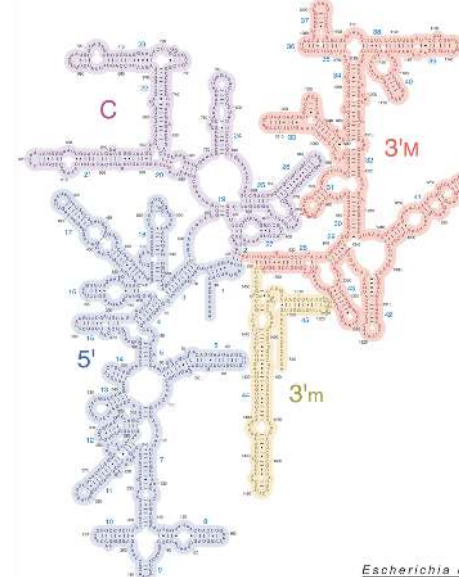
ribosome

- 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.



39

Small subunit "16S" rRNA

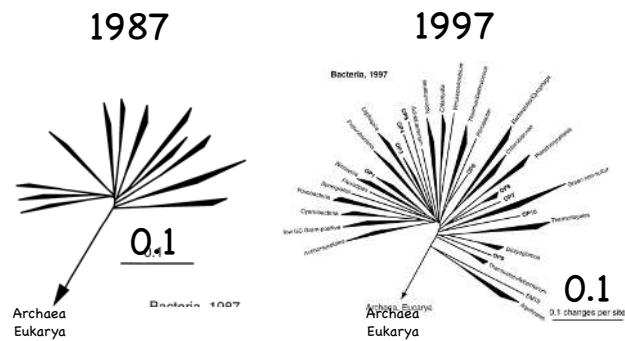


- **Ubiquitous** - present in all known life (viruses don't count)
- **Functionally constant** translation, 2^o-structure
- **Evolves slowly** - mutations more rare than for protein-coding genes
- **Large** - information for evolutionary inference
- **No exchange** - Limited examples of rRNA gene-sharing between organisms
- **Feasibility** - The right size for available sequencing technology (e.g. Sanger)

Escherichia coli
small subunit 16S ribosomal RNA

40

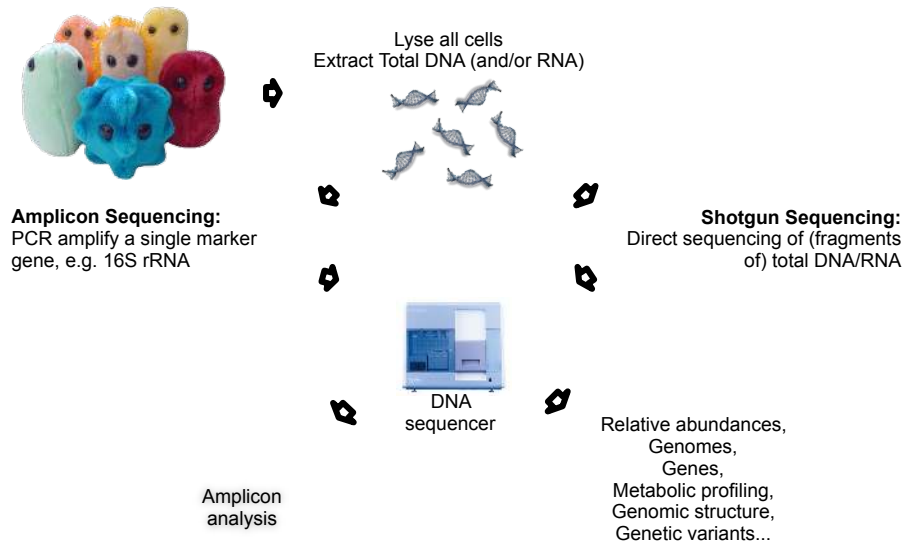
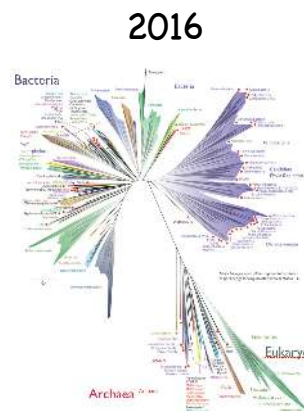
16S rRNA phylogeny, Known Bacteria



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

41

genome phylogeny



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

42

Who's there?

(Taxonomic profiling)

What are they doing?

(Functional profiling)

What does it all mean?

(Statistical analysis)

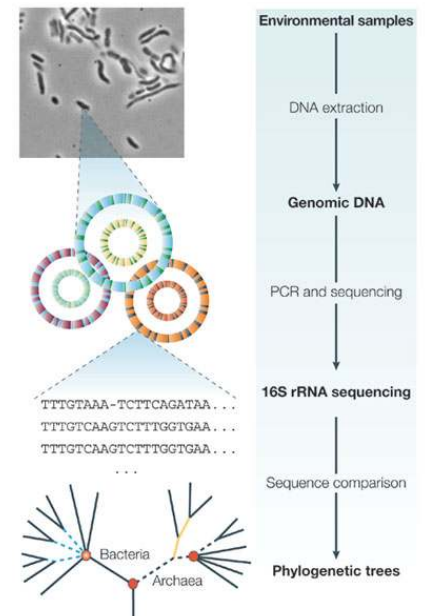
Slide adapted from Curtis Huttenhower, not necessarily with permission O:-)

43

Amplicon sequencing

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



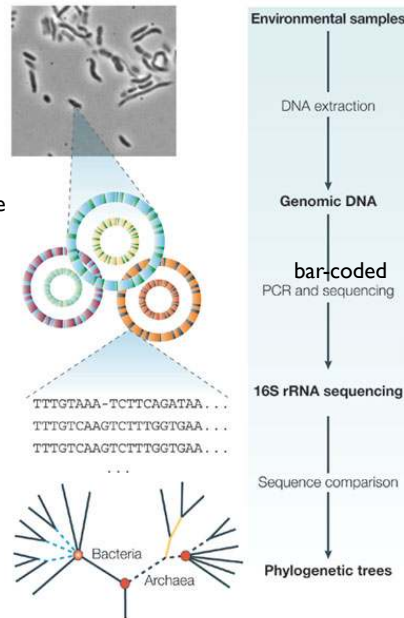
Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, 6(11), 805–814.

44

Amplicon sequencing

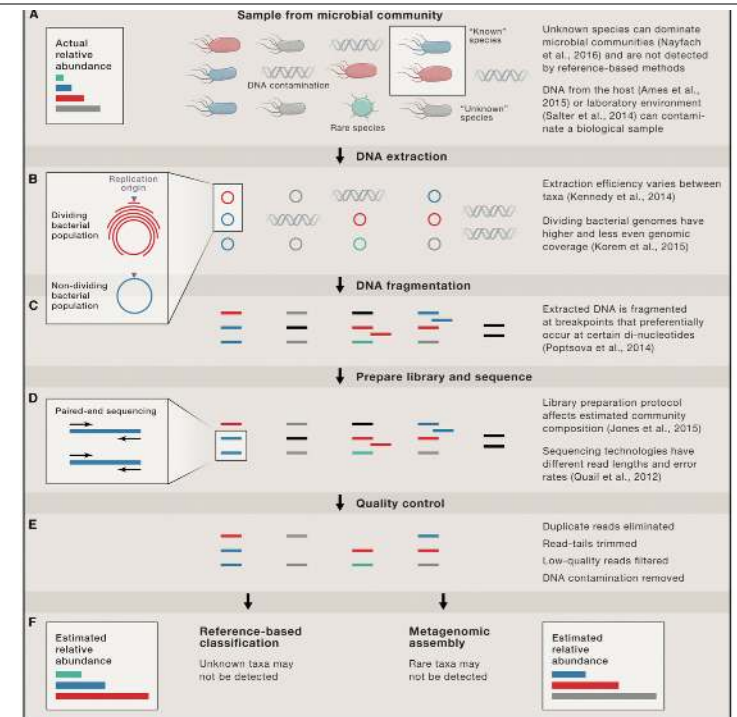
Many microbiomes in parallel:

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



45

"Shotgun" sequencing



46

Why not just always sequence entire (meta)genomes?

(As Bill described in motivation for RADSeq):

- still prohibitively expensive
- for many biological questions a full sequence isn't needed
- For low-abundance microbes, amplicon sequencing might be the only feasible option
- This is a different kind of "Reduced representation sequencing"
- Use ~~restriction enzyme digestion~~ PCR amplification to focus sequencing of multiple samples on [one] homologous regions across the genomes
- Cost is a fraction of the cost of re-sequencing the metagenomes

47

Culture Independent Techniques:

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

\$

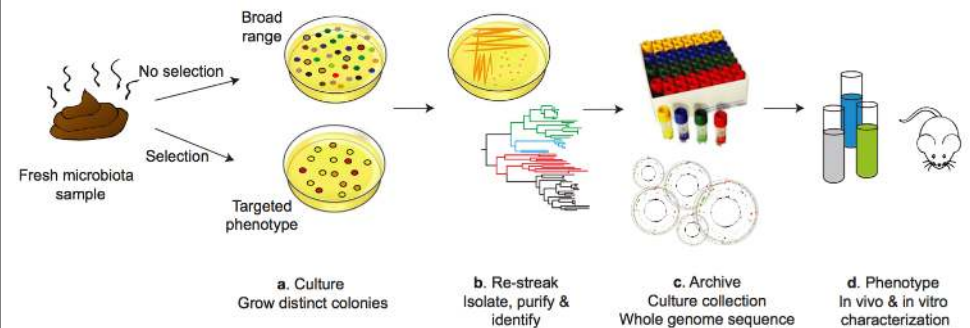
48

- 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
 - Comparative/structural genomics
- Prior knowledge is helpful
- Caution: Correlation $\neq$ Causation
 - Most 'omics results require *lab confirmation*

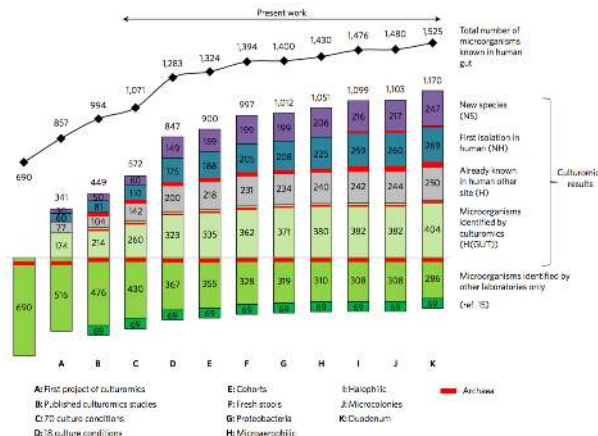
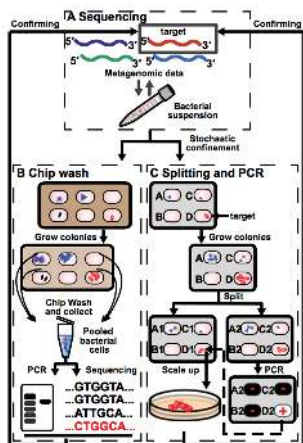
Slide adapted from Curtis Huttenhower, not necessarily with permission O:-)

49

"Bacterial culture was the first method used to describe the human microbiota [after the microscope], but this method is considered outdated by many researchers ... however, a 'dark matter' of prokaryotes, which corresponds to a hole in our knowledge and includes minority bacterial populations, is not elucidated by [metagenomic] studies..."



50



51

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

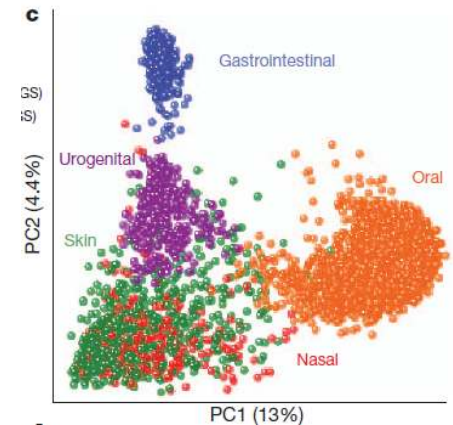
52

End Metagenomics Lecture I

Questions?

53

Introduction to Microbiome / Metagenome Analysis Concepts



54

- **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

55

- **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

56

Amplicon sequencing (esp. 16S rRNA gene) remains the first and most-common culture-independent method applied to new microbiome samples. (\$, time)

Some pervasive misunderstandings in the field:

- (1) Sequences *must be* processed through an *ad hoc* clustering procedure, generating “OTUs”, and
- (2) Resolution <3% sequence similarity not reliable, nor perhaps even useful

These presumptions are untrue.

There is enough information from current Illumina platforms to support *de novo* single-nucleotide resolution in practice.

57



58



59

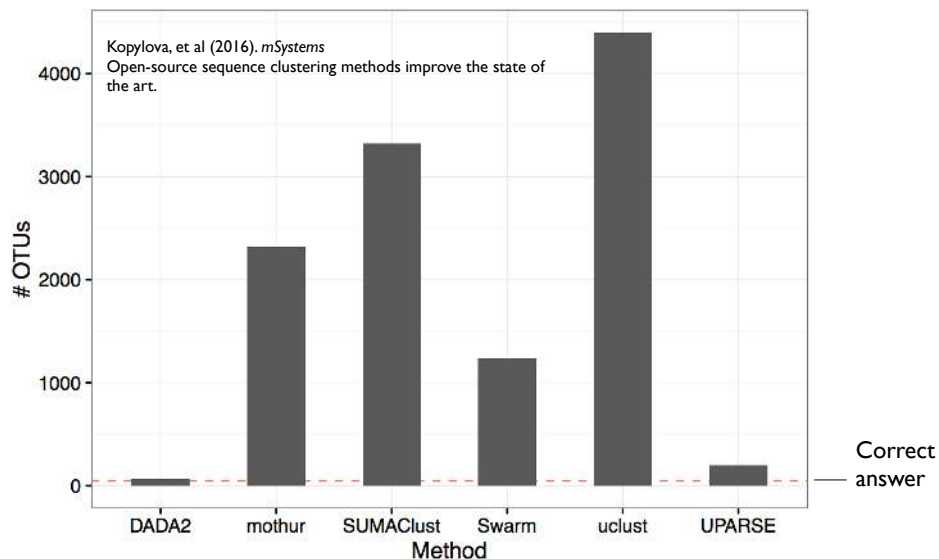
- False Positives - e.g. 1000s of OTUs when only 10s of sequences present
 - Consequently, richness appears to depend on library size
 - Microbiome distances that appear to depend on library size
- Poor Seq/Taxonomic Resolution - defined by arbitrary similarity radius
- Accuracy - Abundance estimates biased and noisier than necessary.
- Cost - Poor data efficiency ~ larger costs to achieve same inference.
- Cost - Computational scaling is quadratic ($\sim N^2$). Becomes costly or intractable as datasets get larger, or more numerous (meta analysis)
- Unstable - OTU sequence and count depend on input
 - must re-run clustering if any data added/removed, or
 - if you want to compare against an external dataset
- Recent open-source methods seem to focus on speed, are analytically worse than UPARSE (a 2012 OTU method)...
- OTU results appear to plateau/degrade with larger library
 - DADA2 improves with more data

"if getting the wrong answer as quickly as possible is important... then there are a number of options..."

—Jon Bentley (as conveyed by R. Gentleman, BioC 2016)

60

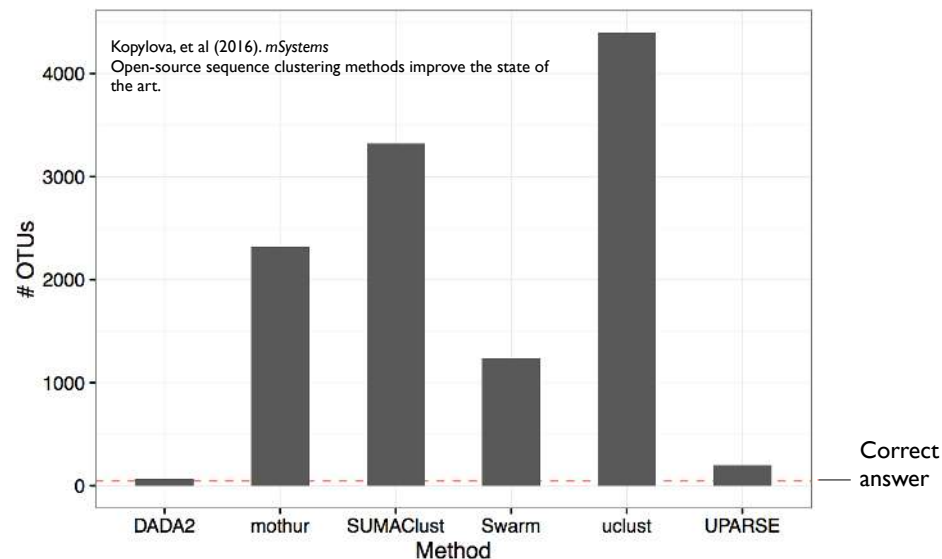
False-positive performance as measured in a microbial community of known composition ("mock community")



<http://benjjneb.github.io/dada2/R/SotA.html>

61

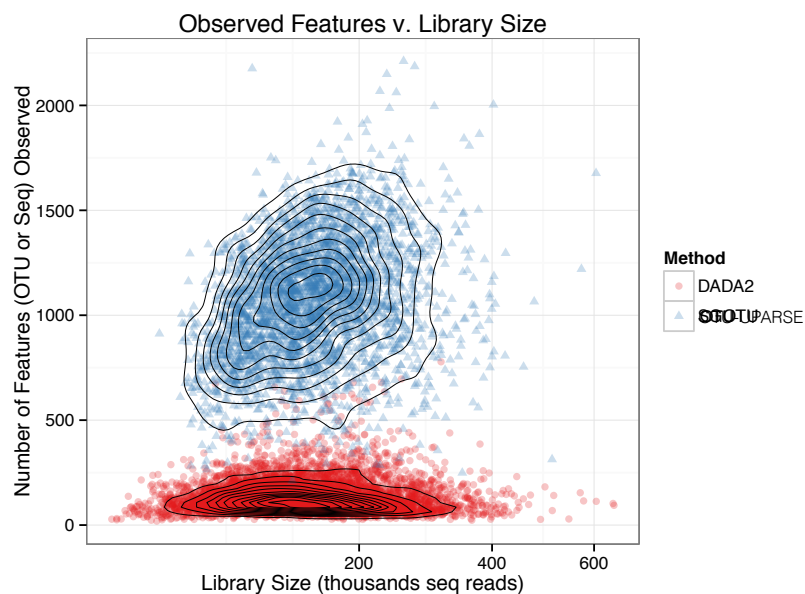
False-positive performance as measured in a microbial community of known composition ("mock community")



<http://benjjneb.github.io/dada2/R/SotA.html>

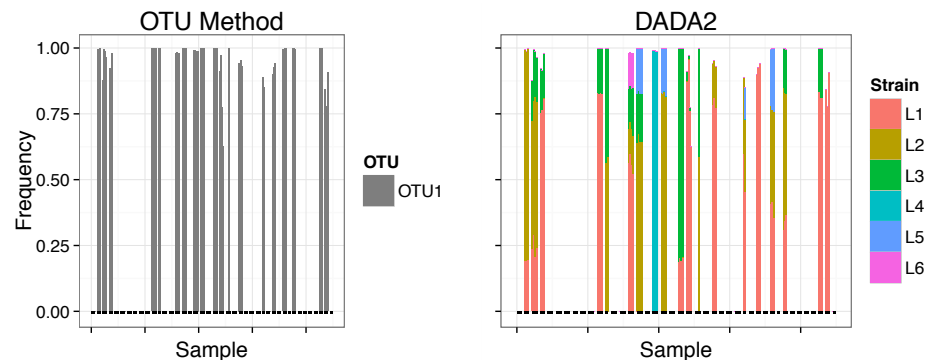
62

Anecdotal example of mitigated dependence of observed richness on sequencing effort



63

- Subject-discriminating strain-level resolution of *Lactobacillus crispatus*
- Repeated samples from vaginal microbiome of 42 pregnant women

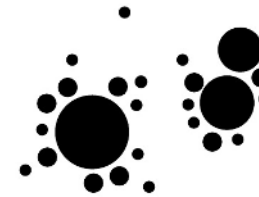


Data: MacIntyre et al. Scientific Reports, 2015.

64

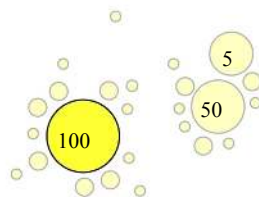
How does this work?

65



Initial guess: one real sequence + errors

66

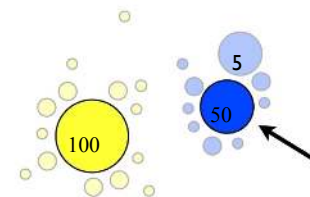


Infer initial *error model* under this assumption.

$\Pr(i \rightarrow j) =$

	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

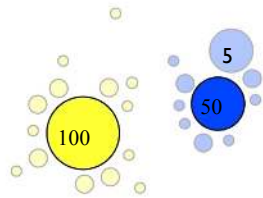
67



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

68

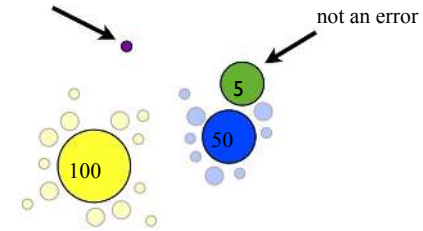


Update the model.

	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

69

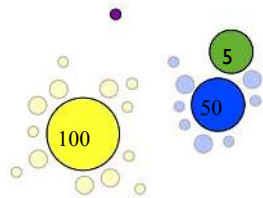
not an error



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

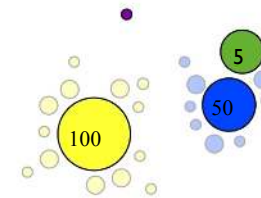
70



Update model again

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

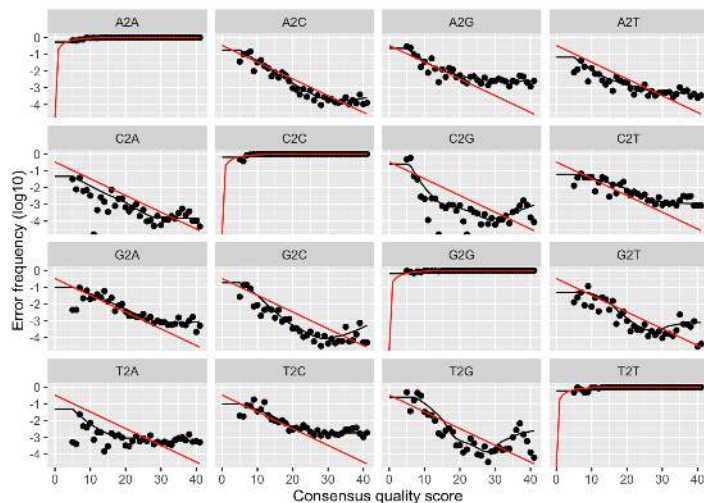
71



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

72



- *selfConsist* mode for DADA2 includes joint inference of error rates as function of quality score.
- red line is expected error rate if Q-scores were exactly correct
- black line is DADA2's empirical model (smooth)
- Notice especially underestimates of errors at high values, $Q > 30$
- For illumina these differences are specific to sequencing run and read direction
 - for small lib sizes, can aggregate estimate across libraries from the same run/direction

73

This sounds complicated.
Isn't it really expensive and
time-consuming to compute?

74

This sounds complicated.
Isn't it really expensive and
time-consuming to compute?

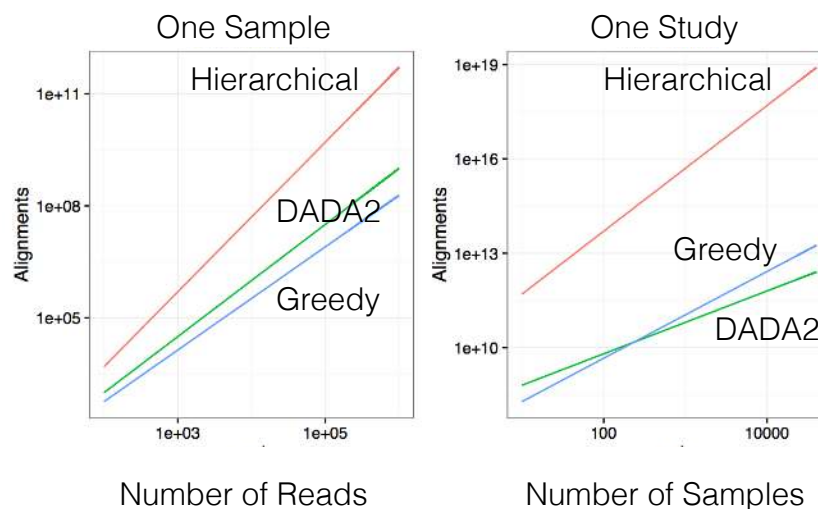
No.

Unlike OTU methods, DADA2 can work on each sequence library independently. The outputs are sequences themselves, which are intrinsically comparable. This has important bonus for computation: *embarrassingly parallel*

- “Horizontal Scaling”, each sample in parallel
- Much faster for large projects
- Can use **cheap** commodity hardware (e.g. your laptop), rather than expensive, high-memory clusters
- **Robust:** results don't change with new data
- Bad data or failure from one sample can't affect others

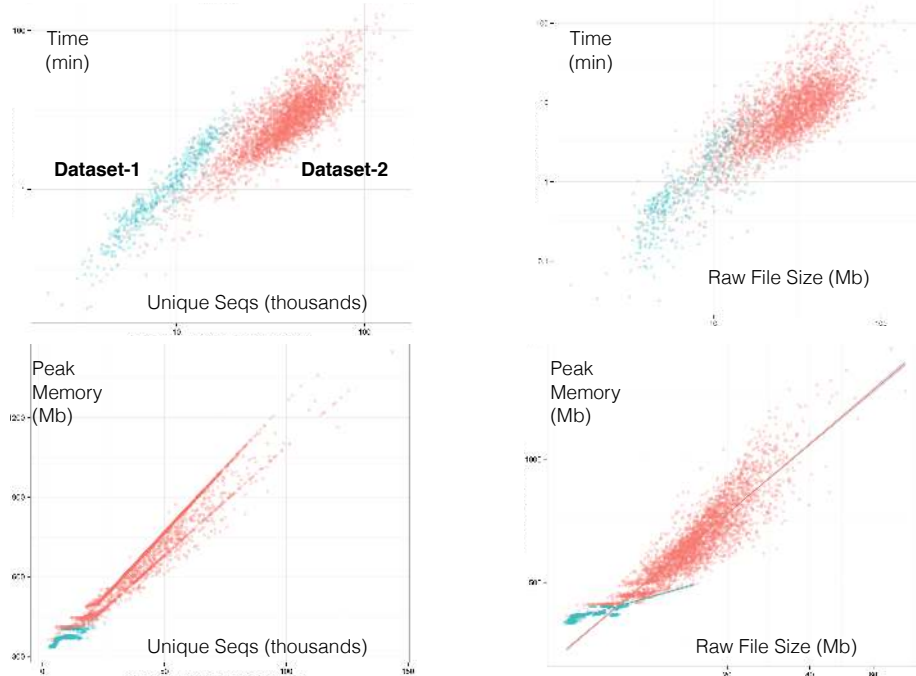
75

Compute performance, as the required
number of sequence alignments



76

Computational Performance - each point is a sample



77

Applications

- Any amplicon target... not just 16S rRNA or even microbiome
- Detection of low-abundance microbes
- Strains that are unique to an individual host
- Strains that are associated with a particular patient outcome
- Improved shotgun metagenomic inference (e.g. PiCRUST, etc.)
 - Mitigate ambiguity of representative genome to use
- Detecting pathogens (special cases)
- Bridging gap to world where shotgun is cheap enough

78

DADA₂

Divisive Amplicon Denoising Algorithm - ver.2

NATURE METHODS | BRIEF COMMUNICATION

July 2016

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan^{1,*}, Paul J McMurdie², Michael J Rosen³, Andrew W Han², Amy Jo Johnson² and Susan P Holmes¹

¹Department of Statistics, Stanford University

²Second Genome, South San Francisco, CA

³Department of Applied Physics, Stanford University

*Corresponding Author: benjamin.j.callahan@gmail.com

<http://benjjneb.github.io/dada2/>

R package available on BioConductor

79

Diversity

(That is, we are now switching to an overview of methods related to the formal analysis of ecological diversity)

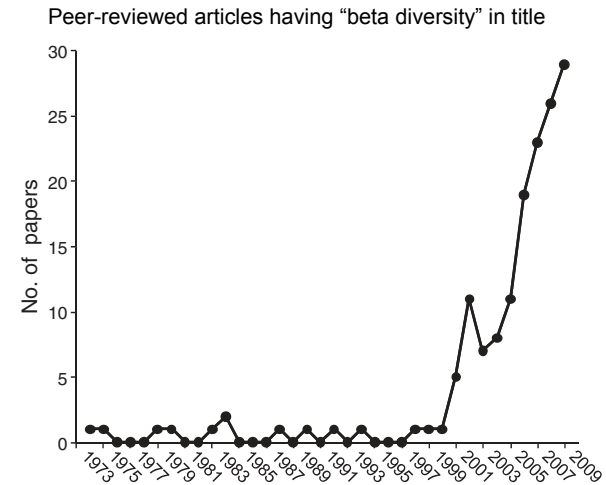
80

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

81

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.

82

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

or

Analysis comparing whole microbiomes to one another

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

83

Distances between microbiomes

84

Community Distance

Communities are a vector of abundances:

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

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:-)

85

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:-)

86

The Distance Spectrum

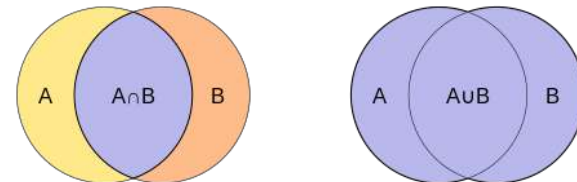
	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

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

87

Jaccard

$$\begin{aligned} \text{Dist}(A, B) &= 1 - (A \cap B) / (A \cup B) \\ &= ((\mathbf{x}_A > 0) \& (\mathbf{x}_B > 0)) / ((\mathbf{x}_A > 0) \mid (\mathbf{x}_B > 0)) \end{aligned}$$



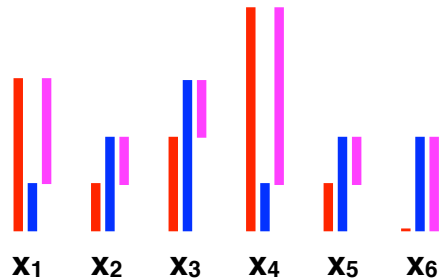
Intuition: Fraction of shared **types** unique to one of the communities

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

88

Bray-Curtis

$$\text{Dist}(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} = \frac{\text{pink}}{\text{red} + \text{blue}}$$



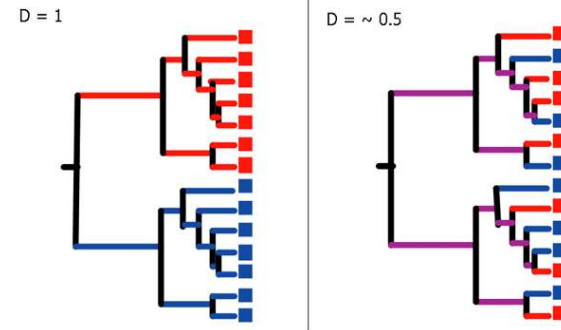
Intuition: *City block distance.* Sum of absolute differences over total abundance.

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

89

Unifrac

$$\text{Dist}(x, y) = \frac{\text{red} + \text{blue}}{\text{red} + \text{blue} + \text{pink}}$$



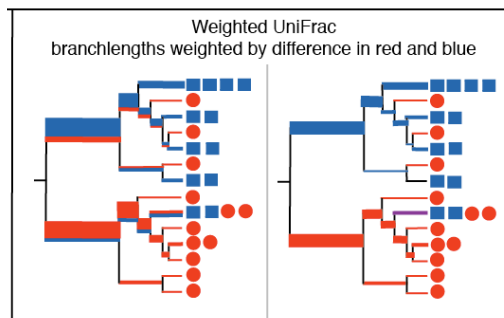
Intuition: Fraction of shared **tree** unique to one of the communities

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

Lozupone and Knight (2008)

90

Weighted Unifrac

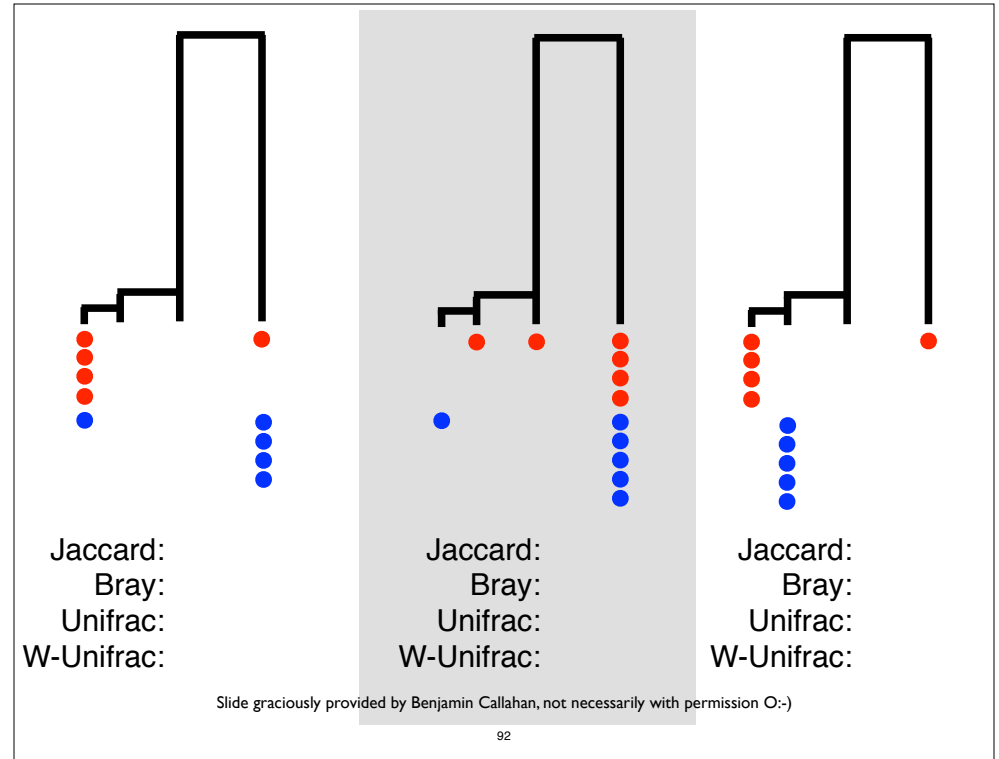


Intuition: The cost of turning one distribution into the other; where the cost is the amount of “dirt” moved times the distance by which it is moved.

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

Lozupone et al. (2007)

91



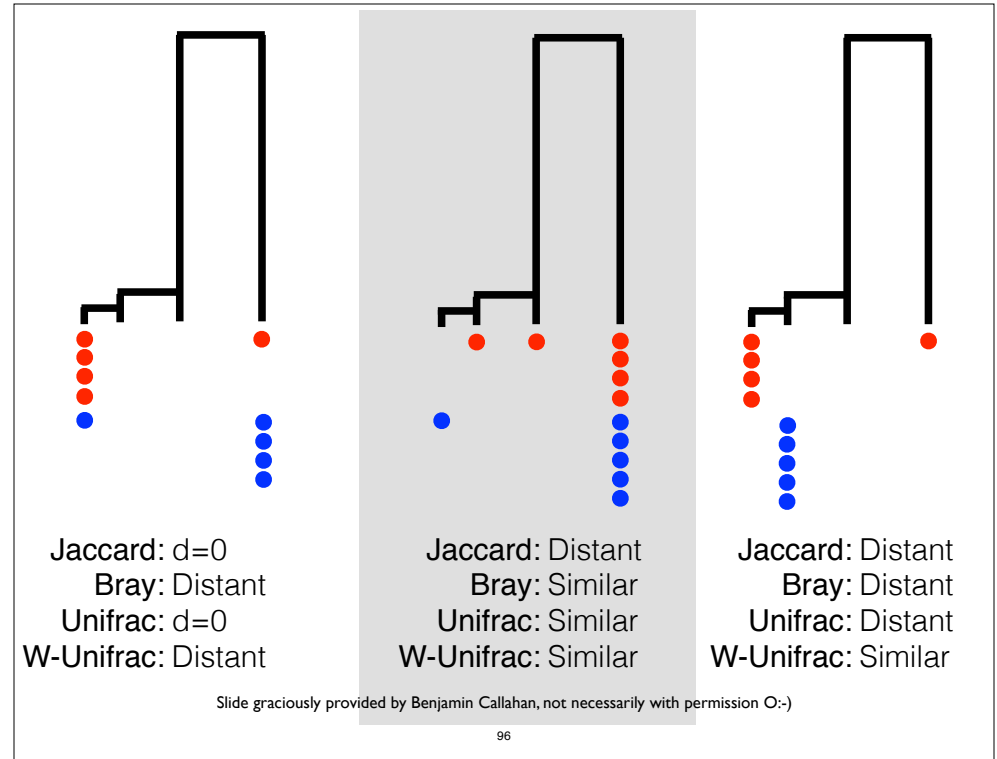
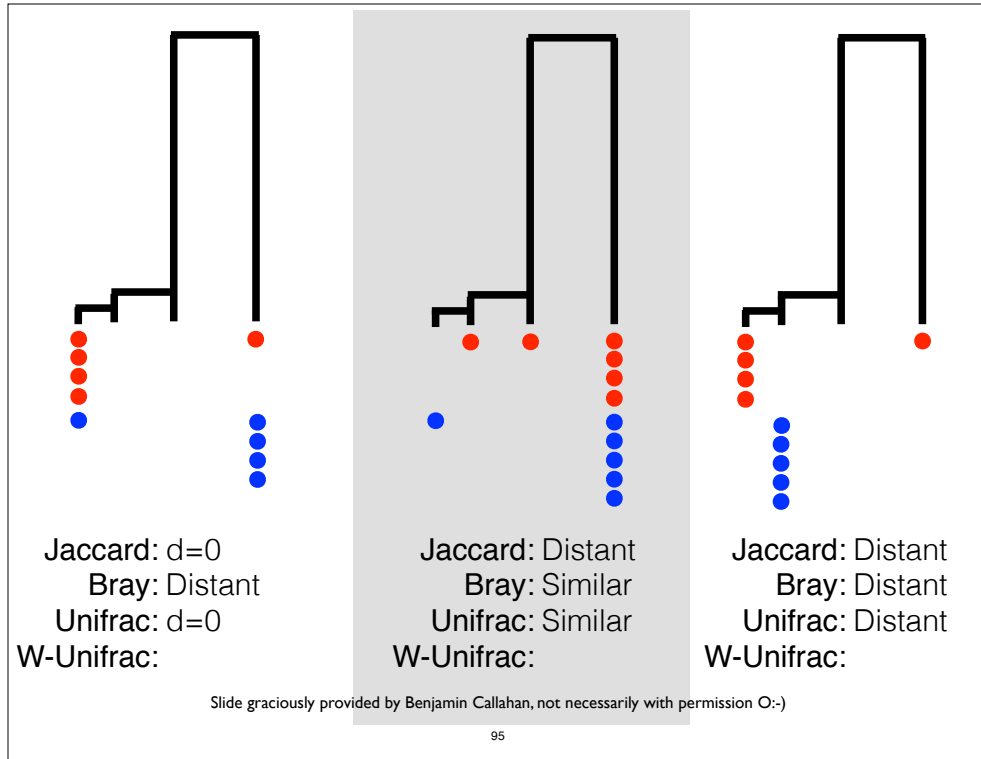
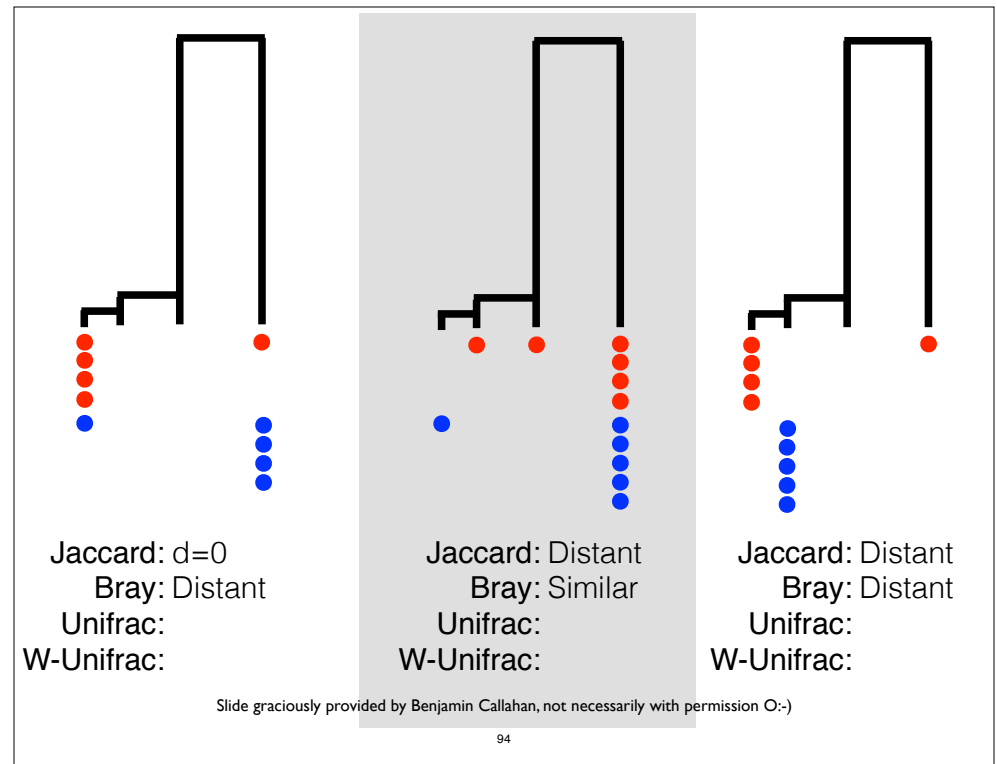
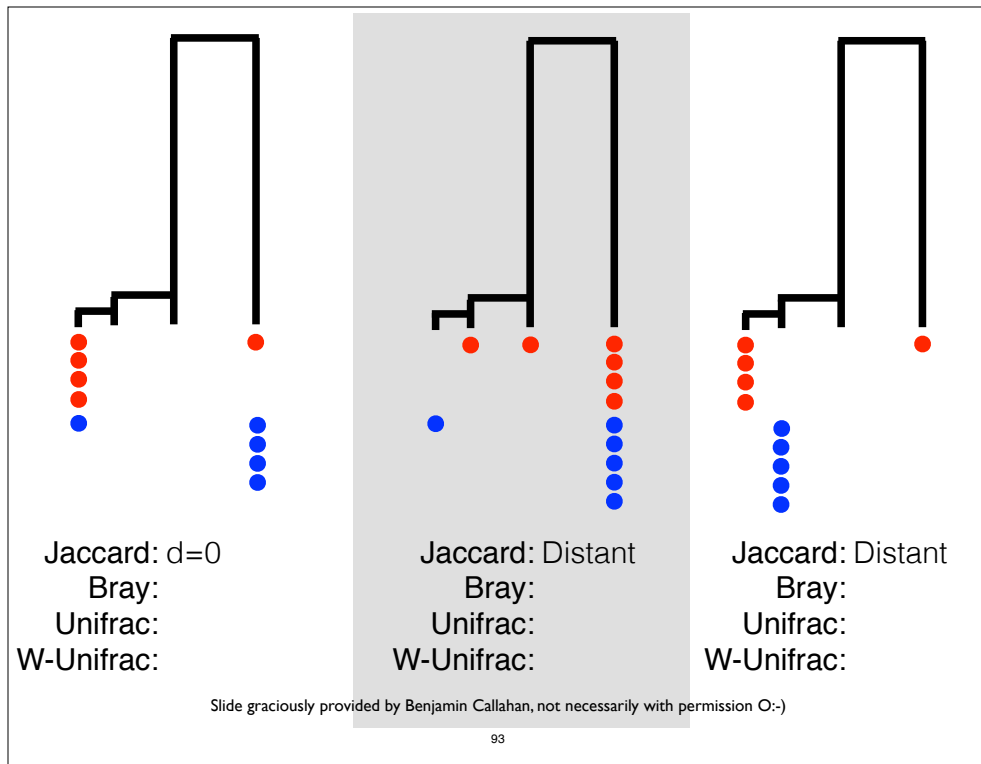
Jaccard:
Bray:
Unifrac:
W-Unifrac:

Jaccard:
Bray:
Unifrac:
W-Unifrac:

Jaccard:
Bray:
Unifrac:
W-Unifrac:

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

92



The Distance Spectrum

	Categorical	Phylogenetic	<u>phyloseq distances</u>
Presence/ Absence	Jaccard	Unifrac	manhattan euclidean canberra bray kulczynski jaccard gower altGower morisita-horn mountford raup binomial chao cao jensen-shannon unifrac weighted-unifrac ...
Quantitative Abundance	Bray-Curtis	Weighted Unifrac	

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

97

What do we do with distances between microbiome samples?

For starters: Plot / exploratory analysis

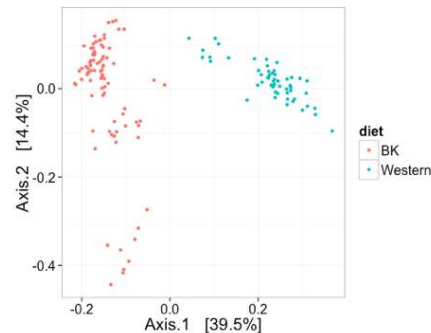
98

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,...
 0,0,0,0,1,2,1,8,0,0,...
 0,0,0,0,9,4,0,0,0,1,...
 .
 .



P-dimensions

2-dimensions

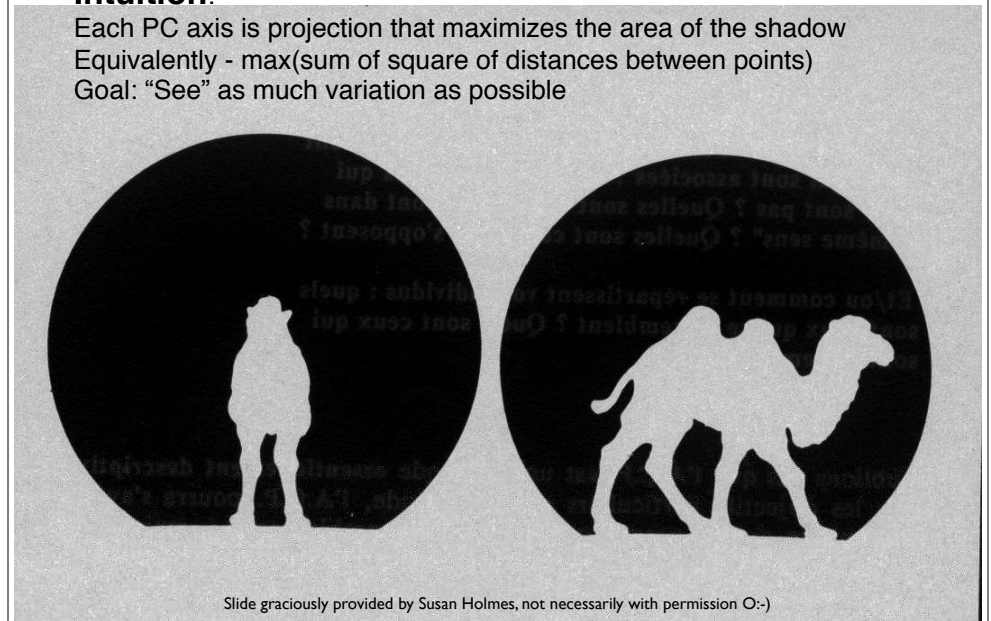
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

99

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

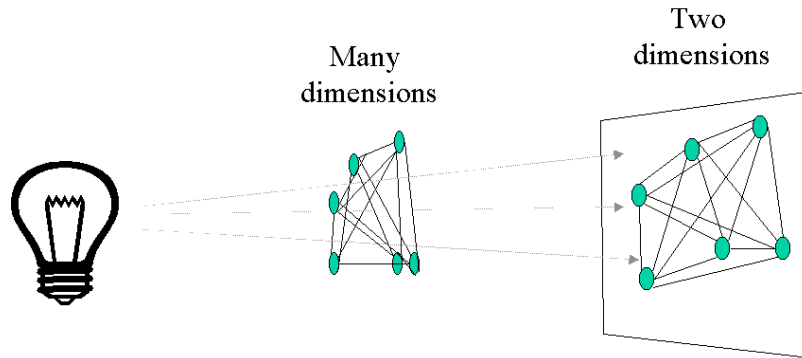


Slide graciously provided by Susan Holmes, not necessarily with permission O:-)

100

Multi-dimensional Scaling

Why MDS? It works with any distance!



Input distance matrix can be by Bray-Curtis, Unifrac, ...

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

101

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 $\mathbf{\Lambda}$ gives the proportion of variability explained by each of the axes.
- The value of $\mathbf{\Lambda}$ should always be looked at in deciding how many dimensions to retain

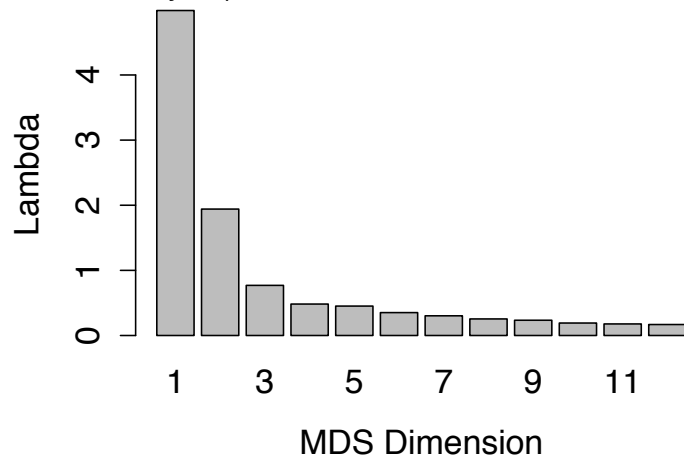
NMDS is similar, but minimizes a different function
(difference in distance ranks)

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

102

MDS Scree Plot

These values are the relative quantity of variability represented in each new dimension

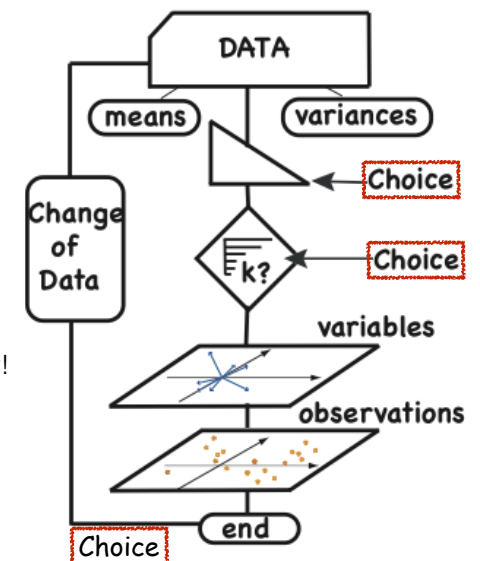


Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

103

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:-)

104

What we “learn” depends on the data.

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

105

- Are there clusters? How many?

Technique:
Gap Statistic

106

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

Technique:
PC Regression (statistics’ “PCR”)

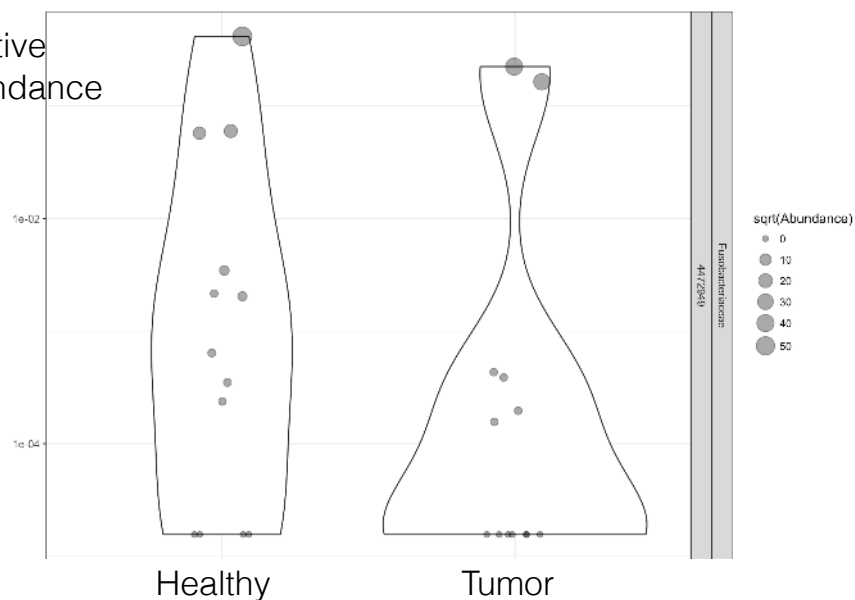
107

- Are the patterns consistent with covariates?

Technique:
Permutational Multivariate ANOVA
`vegan::adonis( )`
(note: this works with discrete and continuous variables)

108

Relative
Abundance



109

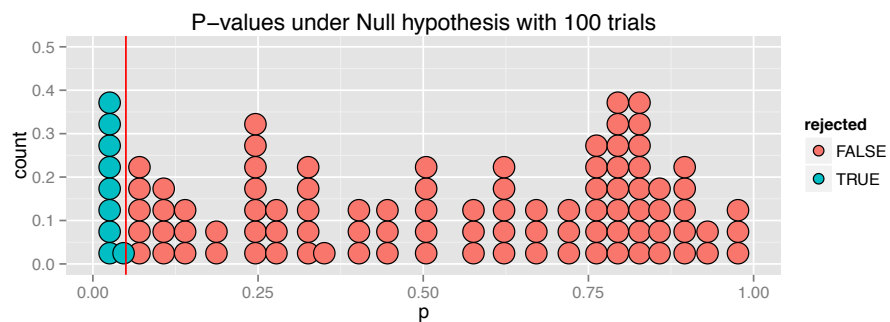
Clinically-motivated question:
"Does this microbe occur in different
abundance between healthy/disease tissue?"

Can we ask this for every microbe observed in
the dataset?

Is it cheating to perform the same test over and
over until *a hit*?

110

- In general, we often want to test many hypotheses at once.
- p-values are distributed uniformly when null hypothesis is true
- The expected number of rejections *by chance* is $m \cdot \alpha$



111

TABLE 1
Number of errors committed when testing m null hypotheses

	<i>Declared non-significant</i>	<i>Declared significant</i>	<i>Total</i>
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - \mathbf{R}$	R	m

Method	Category	Control	p.adjust
Bonferroni	FWER	$P(V \geq 1)$	"bonferroni"
Holm's	FWER	$P(V \geq 1)$	"holm"
B-H	FDR	$P(V/R)$	"BH", "fdr"

112

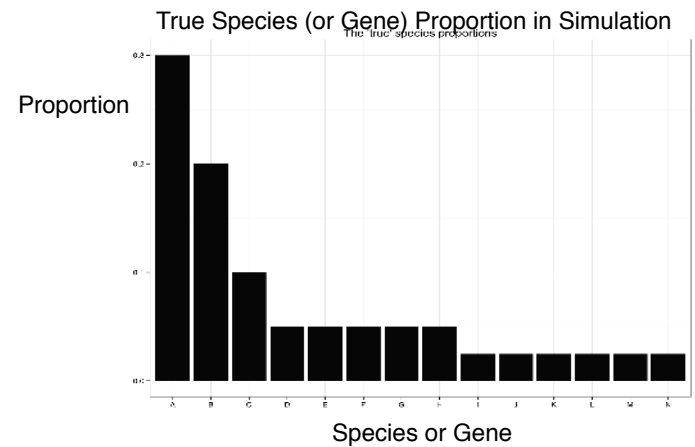
Independent filtering

- Is a general approach that can substantially increase the [statistical] efficiency of experiments
- Uses filter/ test pairs that are independent under the null hypothesis
- but correlated under the alternative

e.g. remove features with very low mean abundance

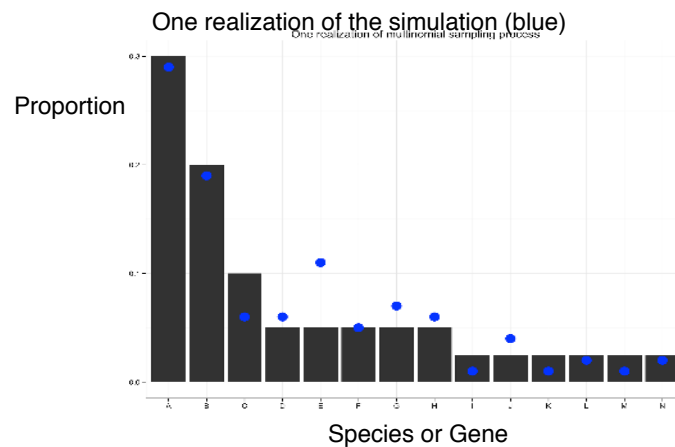
113

Poisson-only Count Simulation



114

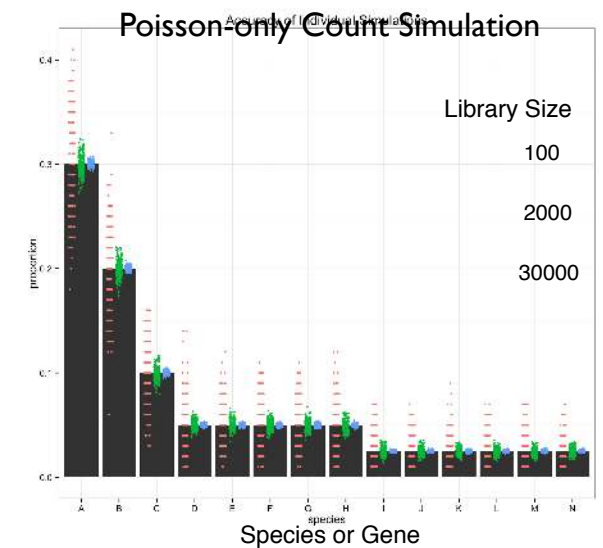
Poisson-only Count Simulation



115

Poisson-only Count Simulation

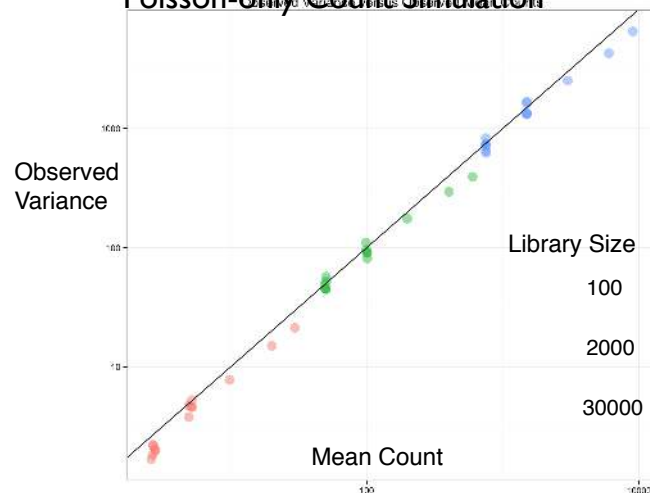
- Uncertainty Depends on Library Size
- This describes sequencing technical replicates quite well



116

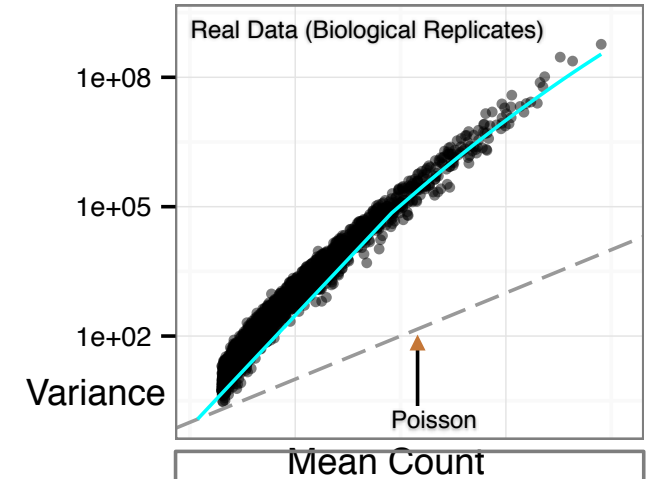
Poisson-only Count Simulation

- Uncertainty Depends on Library Size
- This describes sequencing technical replicates quite well



117

Est. Variance NGS Count Data



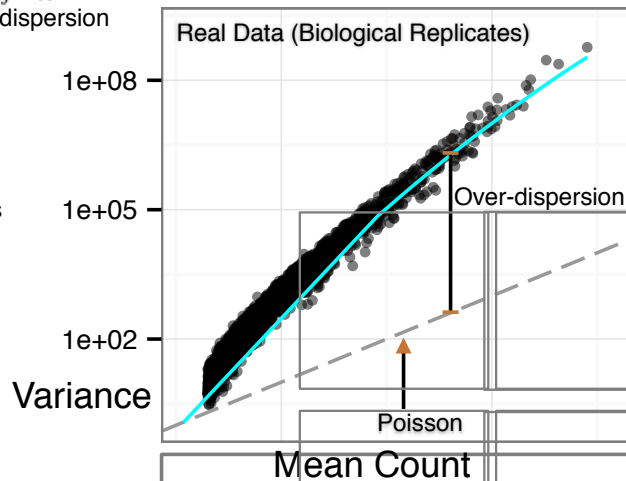
118

Negative Binomial

$$\text{Variance} = \underbrace{u_{ic}s_j}_{\text{Poisson}} + \underbrace{\phi_{ic}s_j^2 u_{ic}^2}_{\text{Overdispersion}}$$

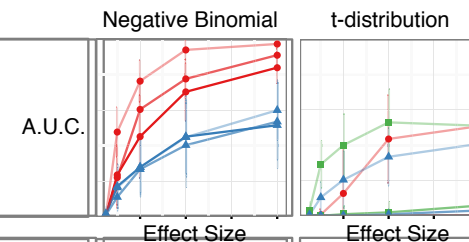
- Over-dispersion
- Strong Function of Mean
- Share Information Across Genes to Improve Fit (Performance)

Est. Variance NGS Count Data



119

- Negative Binomial is an infinite mixture of Poisson R.V.
- Intuition: "NB is relevant when we have (almost) as many different distributions (poisson means) as observations"
- Borrow from RNA-Seq analysis implementations? (Yes)



McMurdie & Holmes (2014). PLoS Computational Biology

- Robinson, Oshlack (2010). A scaling normalization... RNA-Seq data. *Genome Biology*
- Anders, & Huber (2010). Differential expression ... sequence count data. *Genome Biology*

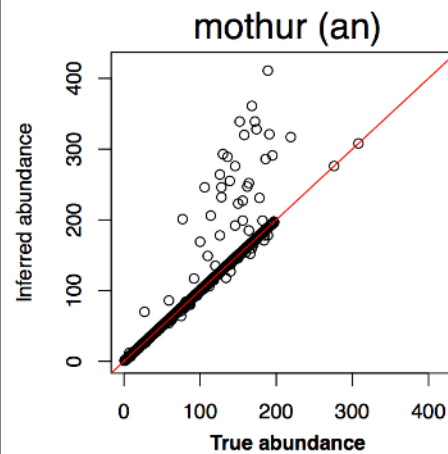
120

End: Introduction to Microbiome / Metagenome Analysis Concepts

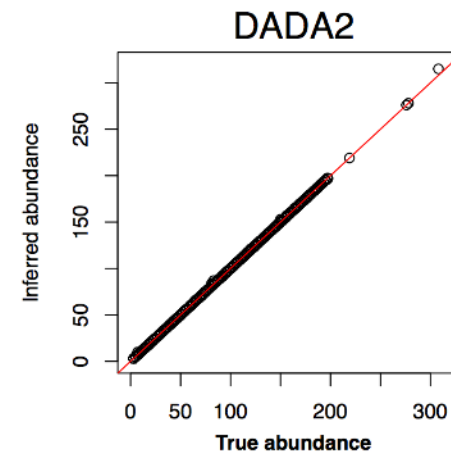
Questions?

121

Performance in a computationally simulated community



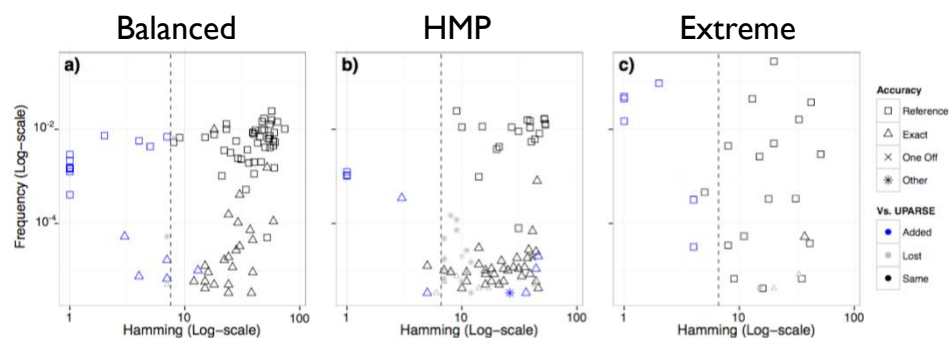
TP: 978
FP: 272
FN: 77



TP: 1042
FP: 0
FN: 13

122

Three other “mock” microbiome communities



Better FP and FN performance than UPPARSE...

123

DADA2 algorithm assumptions

DADA2 Error Model

- Errors independent b/w different sequences
- Errors independent b/w sites within a sequence
- Errant sequence i is produced from j with probability equal to the product of site-wise substitution probabilities:

$$\lambda_{j \rightarrow i} = \prod_{l=0}^L p(j(l) \rightarrow i(l), q(l))$$

- Each substitution probability depends on original nt, substituting nt, and quality score

124

DADA2 algorithm assumptions

DADA2 Abundance Model

- Errors are independent across reads
- Abundance of reads w/ sequence i produced from more-abundant sequence j is poisson distributed
- Probability of abundance equals error rate, $\lambda_{j \rightarrow i}$, multiplied by the abundance of “parent” sequence, j .
- i has count greater than or equal to one
- “Abundance p-value” for sequence i is thus:

$$p_A(j \rightarrow i) = \sum_{a=a_i}^{\infty} \rho_{\text{pois}}(n_j \lambda_{j \rightarrow i}, a) / (1 - \rho_{\text{pois}}(n_j \lambda_{j \rightarrow i}, 0))$$

- “Probability of seeing an abundance of sequence i that is equal to or greater than observed value, by chance, given sequence j .”
- A low p_A indicates that there are more reads of sequence i than can be explained by errors introduced during the amplification and sequencing of n_j copies

125

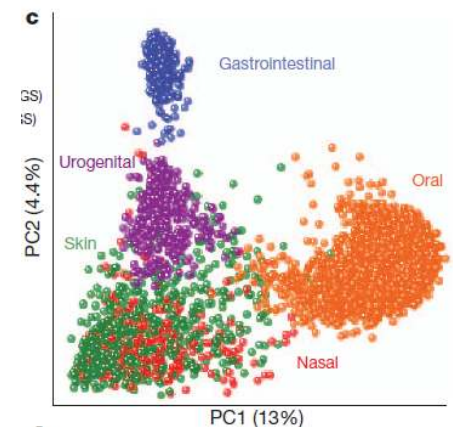
- To ensure overall significance at a given α , one performs each individual test at $\alpha' = \alpha/m$
- Useful when need to correct for just a few hypotheses
- Very stringent, results in “loss of power”
 - increase in Type II error, decreases sensitivity

126

- Rather than control probability of any errors, FDR instead controls the proportion of False Positives in the set of positives.
- Input: p-values for a set of univariate tests
- Output: p-values that are adjusted to FDR: “q-values”
- e.g. A collection of tests rejected at $P_{\text{FDR}} \leq 0.05$ will have 5% or fewer false positives
- This is what is meant by “controlling” the false positive rate

127

Introduction to Microbiome / Metagenome Analysis Tools and Practices



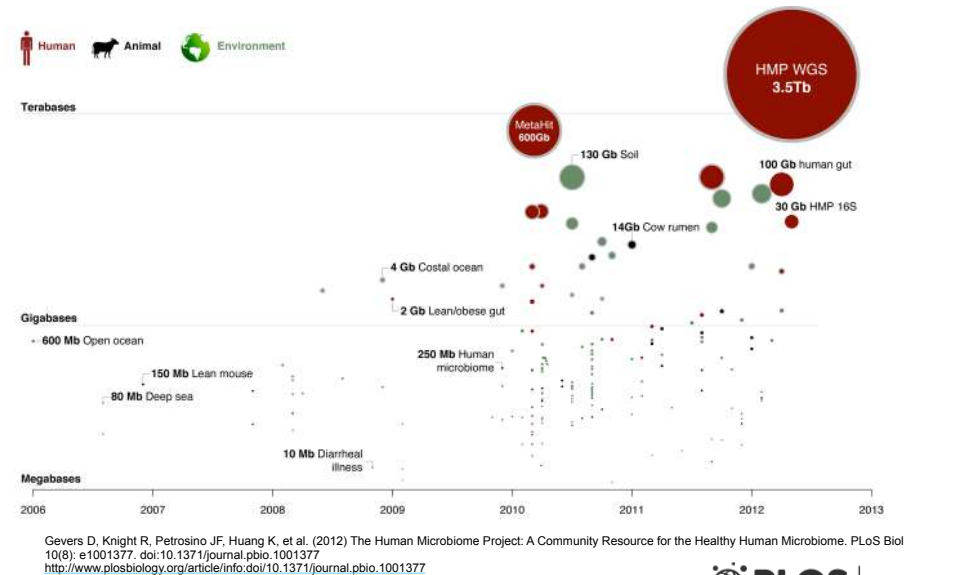
128

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

129

Timeline of microbial community studies using high-throughput sequencing.



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

130

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.)

131

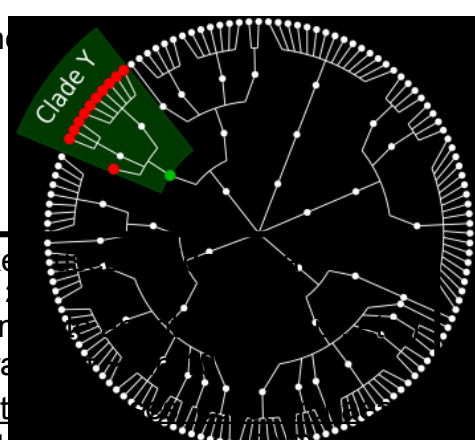
(16S rRNA) Amplicon Sequence Processing Tools:

- QIIME(2) (and 'Qiita'?) - <http://qiime.org/>
- mothur - www.mothur.org/
- usearch - www.drive5.com/usearch
- DADA2 - <https://github.com/benjjneb/dada2>

Afternoon will be spent using QIIME2
Daniel has much more to say about it...

132

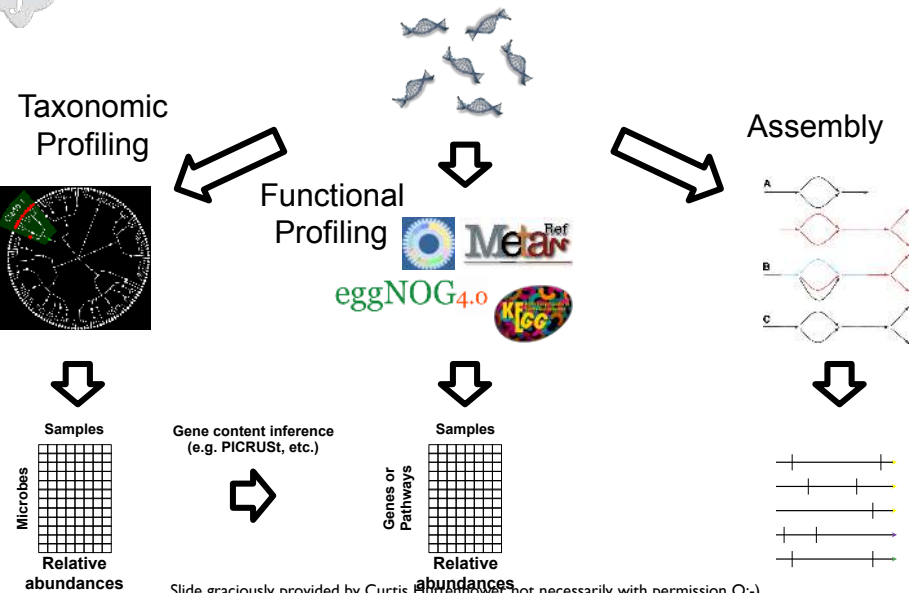
X is a unique marker gene for clade Y



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

133

133



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

134

Jason Pelf, Arend Hietzer¹, Rosangela Corino-Rensing¹, Adina Howe², James M. Tiedje^{3*}, and C. Titus Brown^{4,1}

MetaAMOS (Treangen 2013)

Genovo (Laserson 2011)

METHOD

Open Access

Sébastien Boissvert^{1,2*}, Frédéric Raymond^{1,2}, Flélie Godzardis², François Lavolette³ and Jacques Corbett^{1,4}

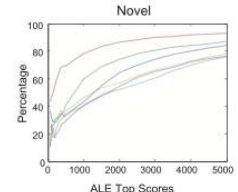
Dinghua Li^{1,†}, Chi-Man Liu^{2,†}, Ruibang Luo^{2,†}, Kunihiro Sadakane³ and Tak-Wah Lam^{1,2,*}

MEGAHIT (Li 2015)

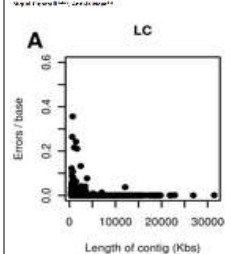
135

MEGAHIT (Li 2015)

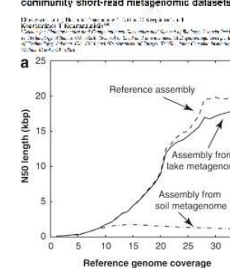
2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2680, 2681, 26



Evaluation of short read metagenomic assembly



Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data



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

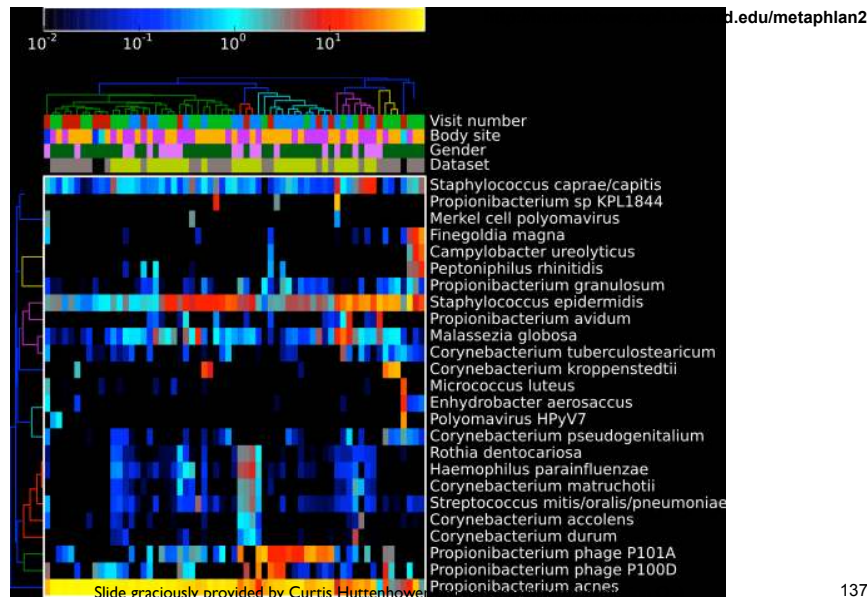
136

Slide graciously provided by Curtis Huttenhower, not necessarily with permission (C.)

MEGAHIT (Li 2015)



MetaPhlAn2: Trans-kingdom profiling



Slide graciously provided by Curtis Huttenhower

137

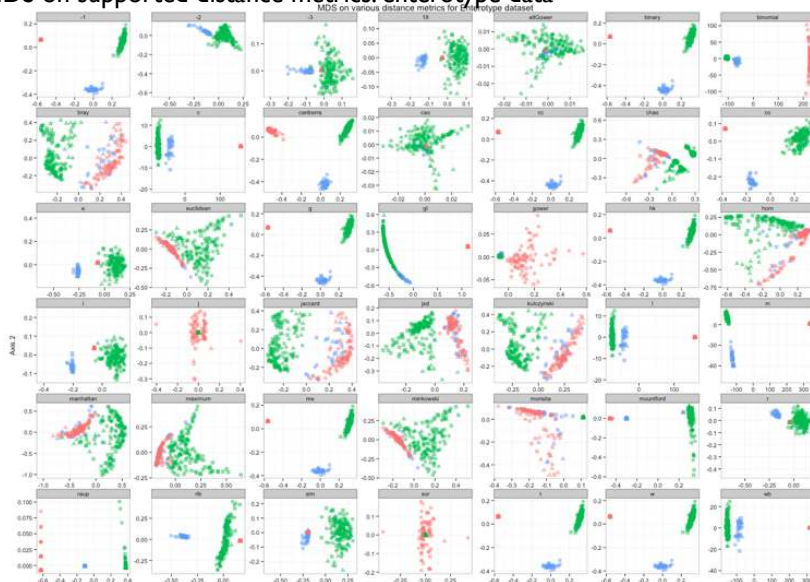
Reproducible analysis of microbiome / metagenome data

- 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?
 - (I call this a "science poem" of your analysis)

138

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

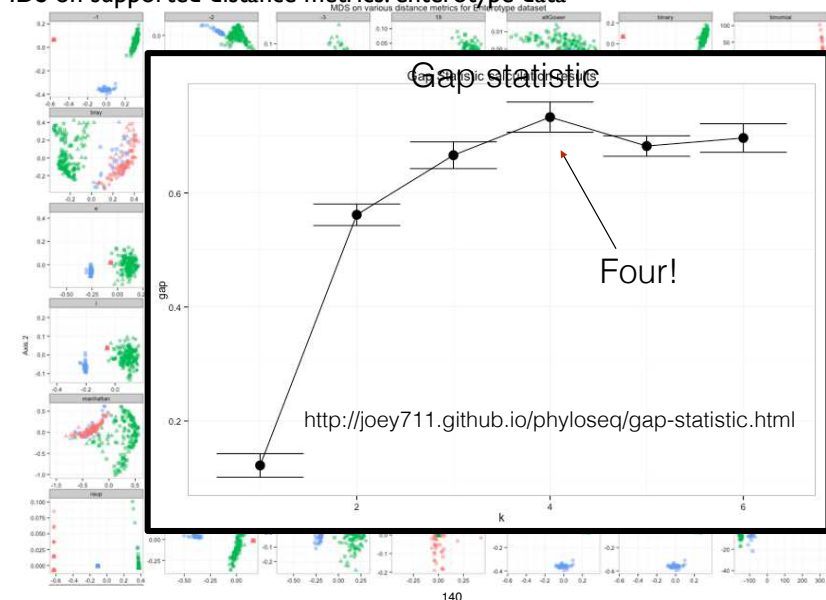
MDS on supported distance metrics: enterotype data



139

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

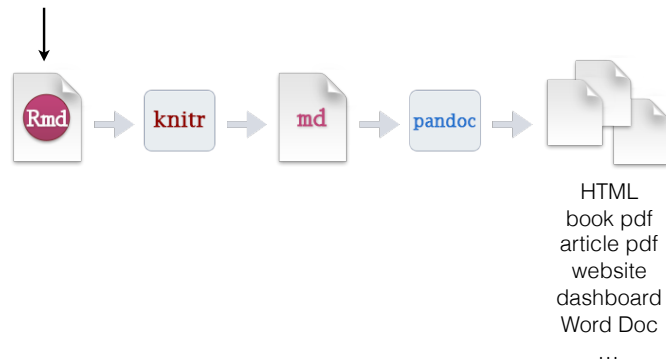
MDS on supported distance metrics: enterotype data



140

Reproducible analysis workflow with R-markdown

microbiome data



141

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*

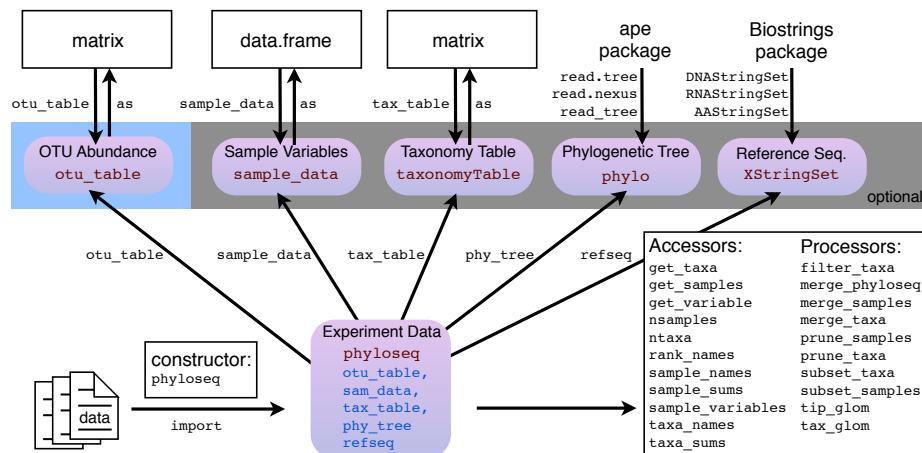
Department of Statistics, Stanford University, Stanford, California, United States of America

Key Packages:

vegan
ape
distory
phangorn
picante
metagenomeSeq
ggtree

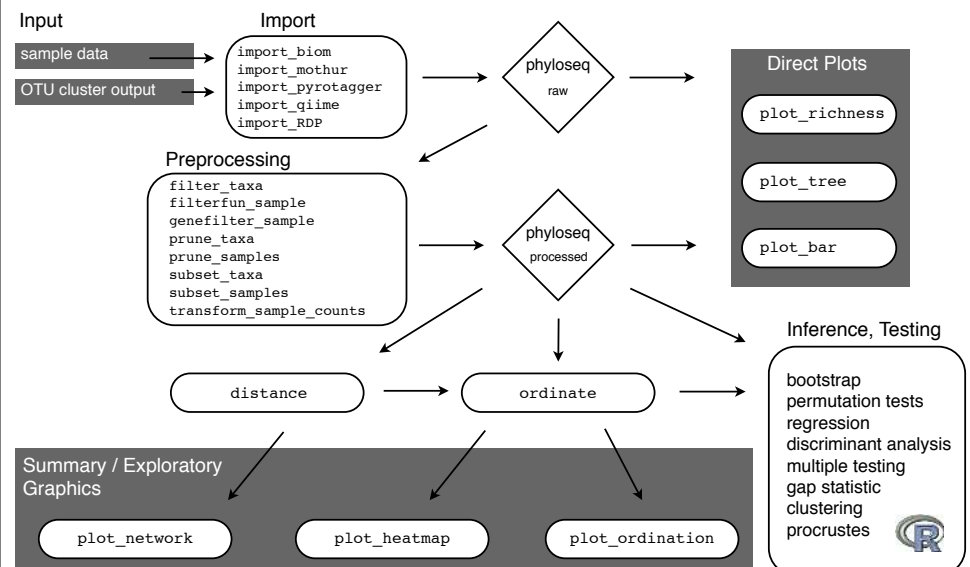
142

data structure & API



143

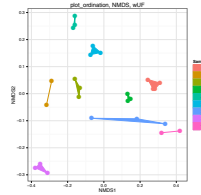
work flow



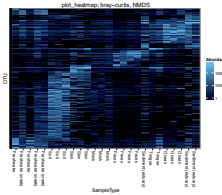
144

graphics

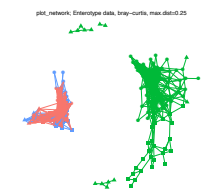
plot_ordination()



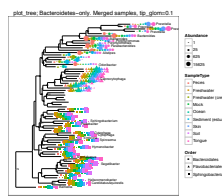
plot_heatmap()



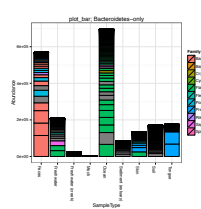
plot_network()



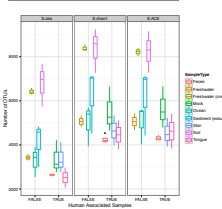
plot_tree()



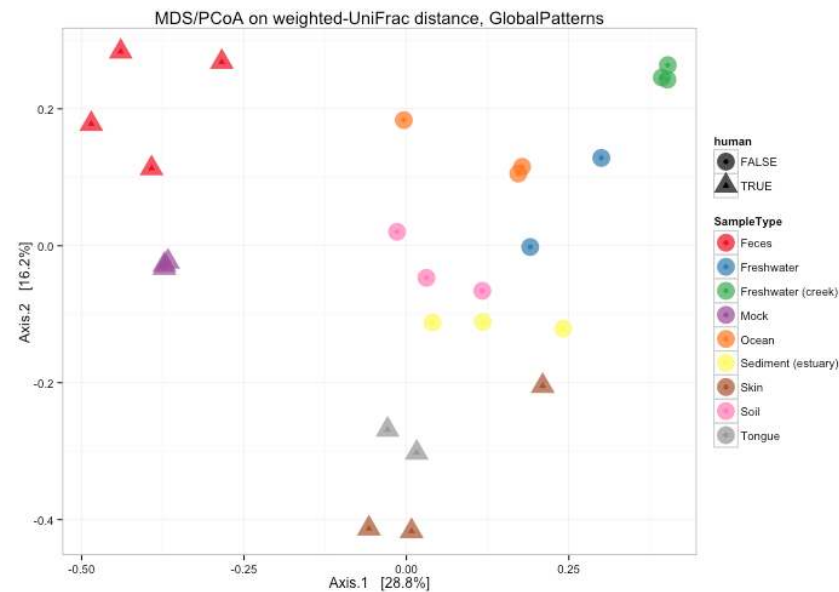
plot_bar()



plot_richness()



145

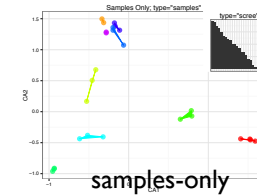


```
ordu = ordinate(GP1, "PCoA", "unifrac", weighted = TRUE)
plot_ordination(GP1, ordu, color = "SampleType", shape = "human")
```

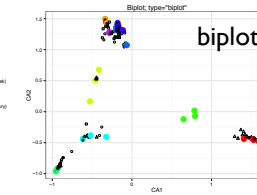
147

phyloseq

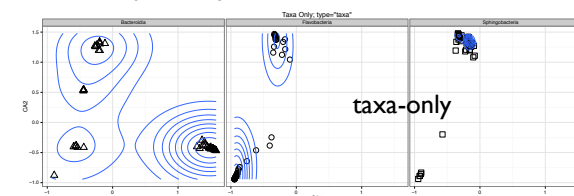
plot_ordination()



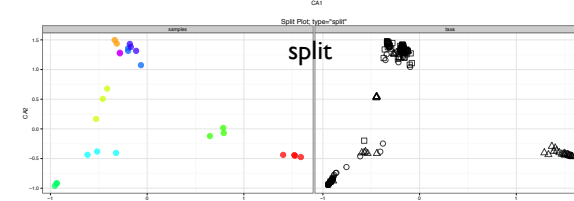
samples-only



biplot



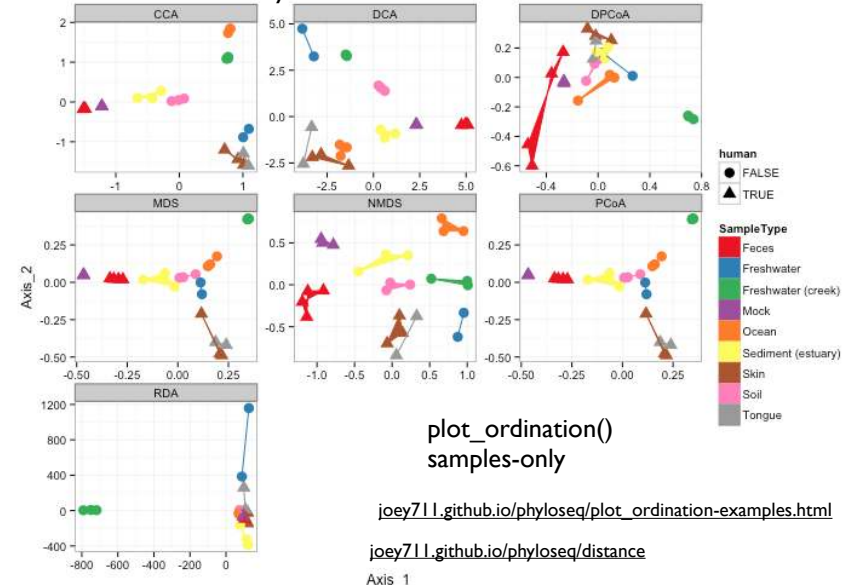
taxa-only



146

phyloseq

Ordination on bray-curtis dist: Global Patterns data



plot_ordination()
samples-only

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

joey711.github.io/phyloseq/distance

Axis_1

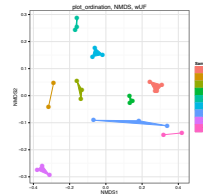
148

supported
ordination
methods

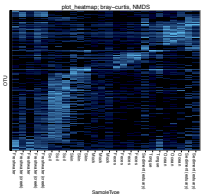
phyloseq

graphics

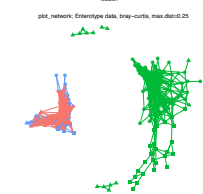
plot_ordination()



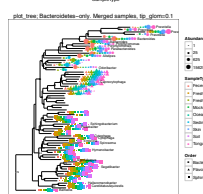
plot_heatmap()



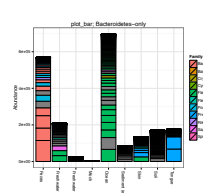
plot_network()



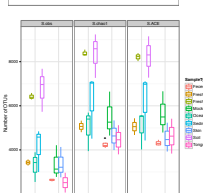
plot_tree()



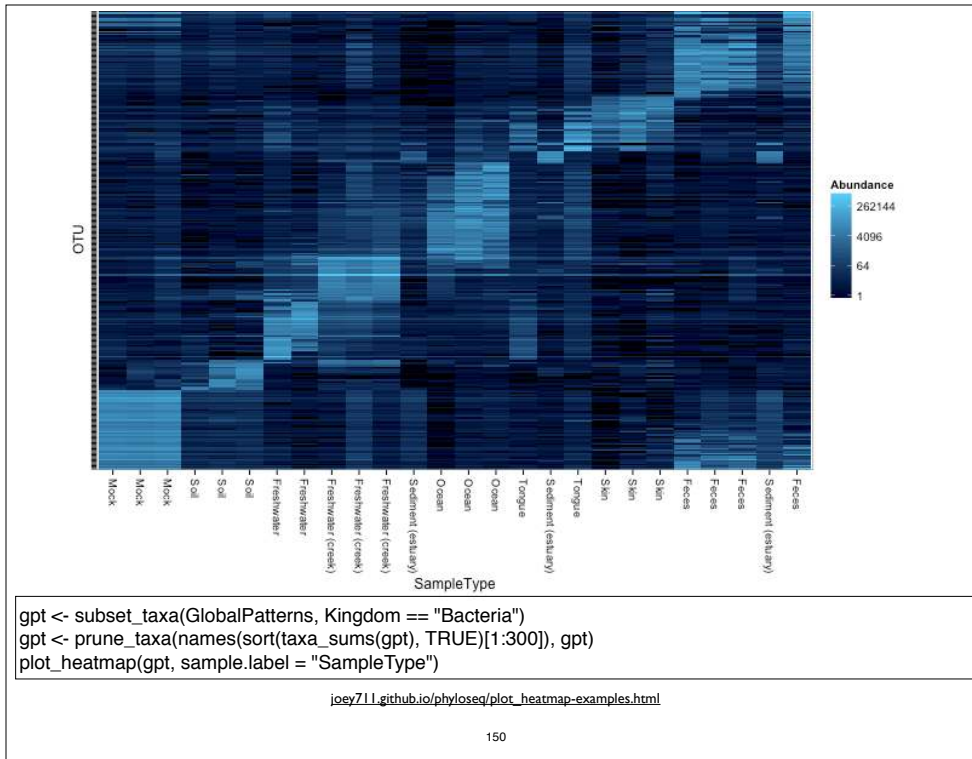
plot_bar()



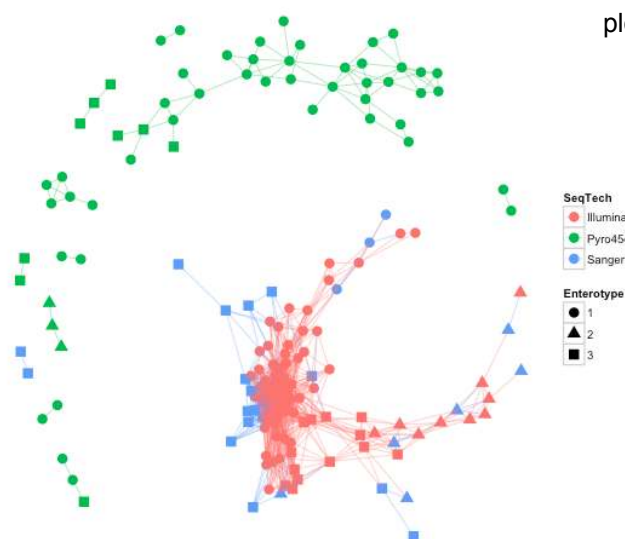
plot_richness()



149



plot_network()

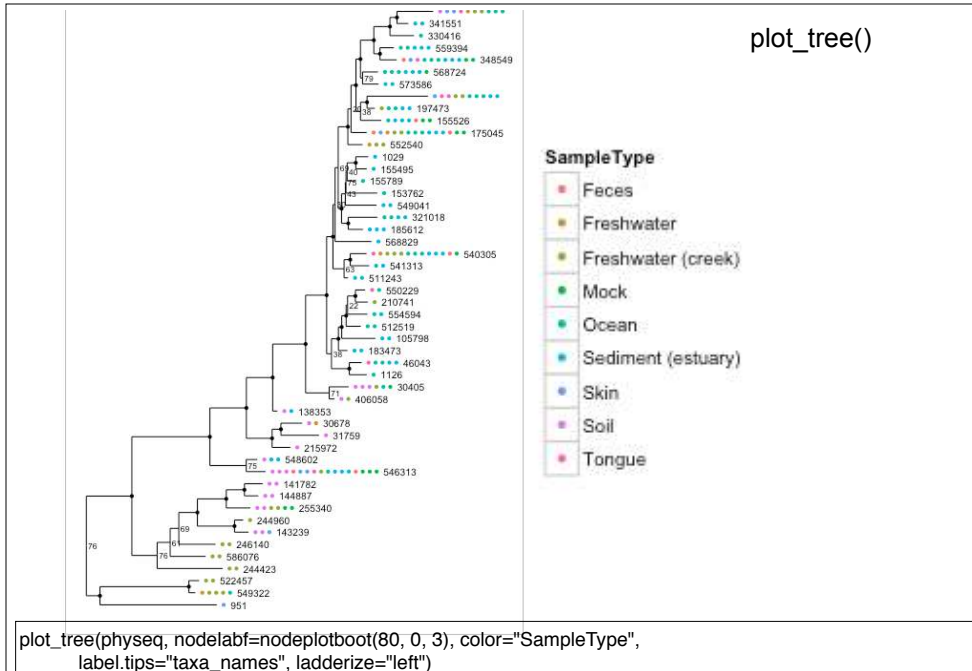


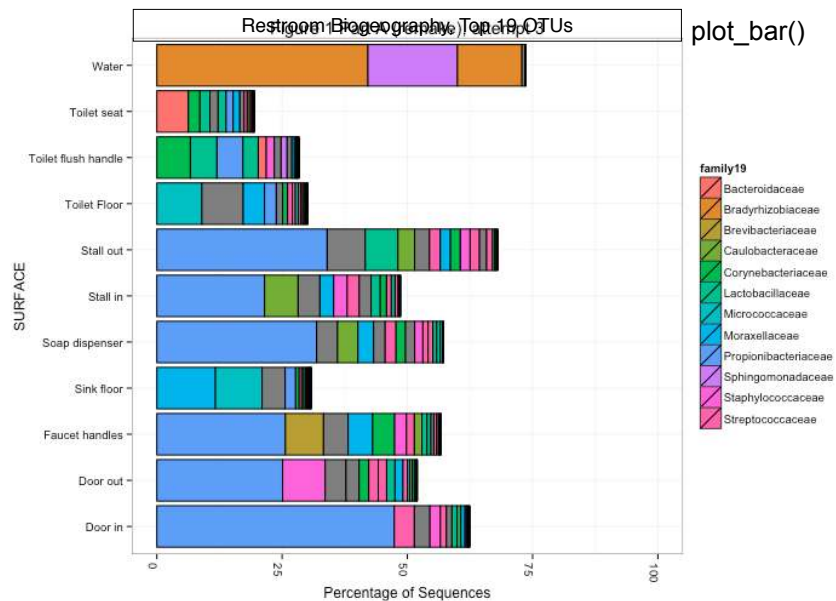
```
ig <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)
plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
  line_weight = 0.4, label = NULL)
```

[joey711.github.io/phyloseq/plot_network-examples.html](https://github.com/joey711/phyloseq/blob/master/inst/examples/plot_network-examples.html)

151

plot_tree()



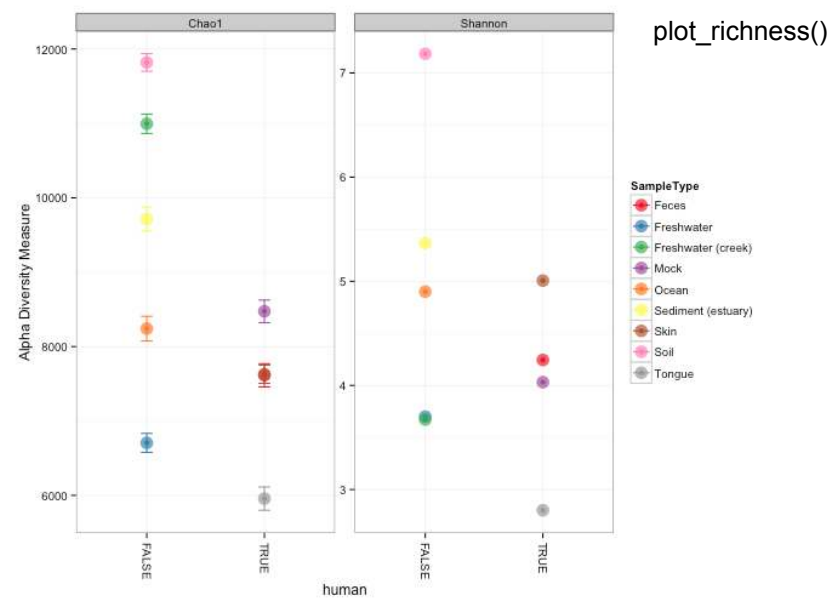


plot_bar()

```
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

153



plot_richness()

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

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

154

Schedule for today

Sec	Day	Start	End	Topic	Lead Instr.
1	Mon Tues	09:00	10:00	Introduction to Metagenomics. Culture independent techniques, 16S rRNA, etc. (60 -75 min)	Joey ✓
2	Mon Tues	10:00 10:30	11:00 11:29	Introduction to microbiome analysis concepts -- Exploratory data analysis, Distances, PCoA, Ordination, taxa & sample-level inferences (75 min)	Joey ✓
3	Mon Tues	11:00 11:30	11:59 12:00	Introduction to microbiome analysis practices: QIIME, phyloseq, reproducible research (30 min)	Joey ✓
---	Mon Tues	12:00	14:00	Lunch (120min)	---
4	Mon Tues	14:00	17:00	QIIME Lab (180min)	Daniel
---	Mon Tues	17:00	19:00	Dinner (120min)	---
5	Mon Tues	19:00	22:00	phyloseq Lab (180min)	Joey

155