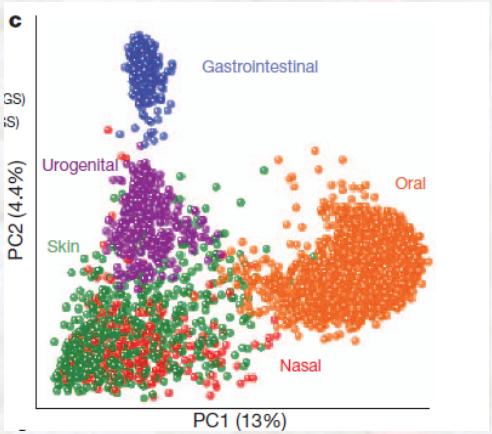
An Introduction to Metagenomics

Paul "Joey" McMurdie, PhD
Whole Biome, Inc.



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

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?

} 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

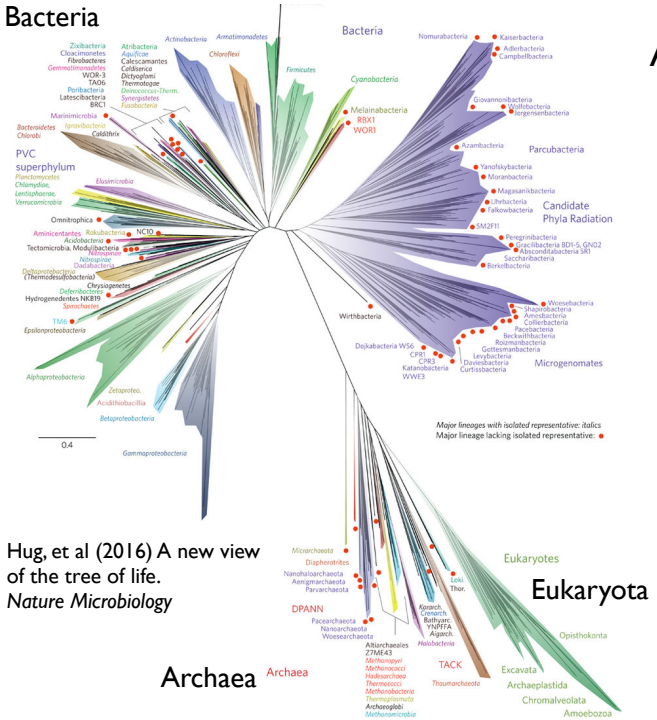
  } Methods

An Introduction to Metagenomics

> "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."

— I Contain Multitudes: The Microbes within Us and a Grander View of Life
Ed Yong 2016

Bacteria



Ancestry of Life

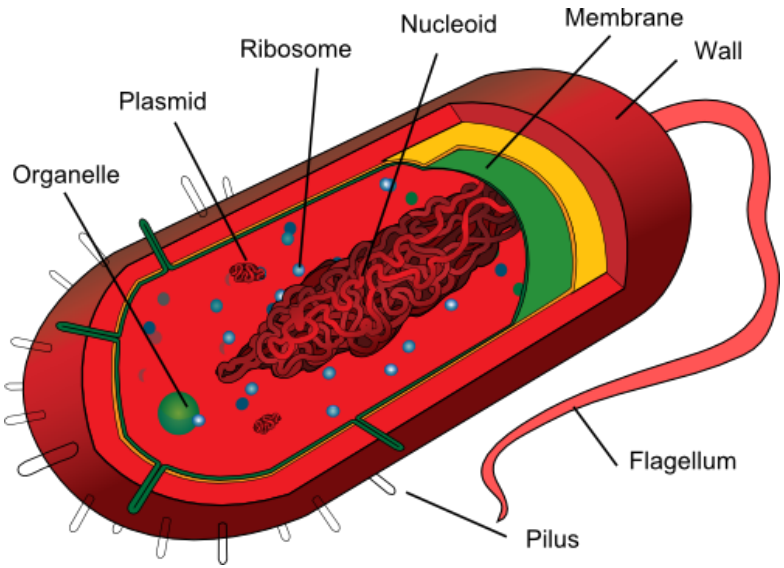
Hug, et al (2016) A new view of the tree of life. *Nature Microbiology*

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 O_2
    - We'd all quickly die of CO_2 poisoning (and later global warming)
    - Most elemental cycles are predominantly microbe-driven

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 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!

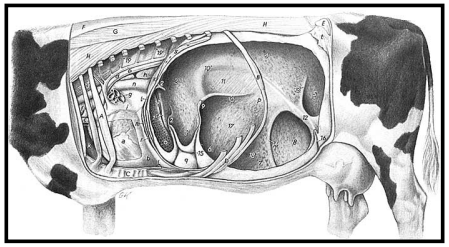
Why study microbiomes?

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

Why study microbiomes?

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

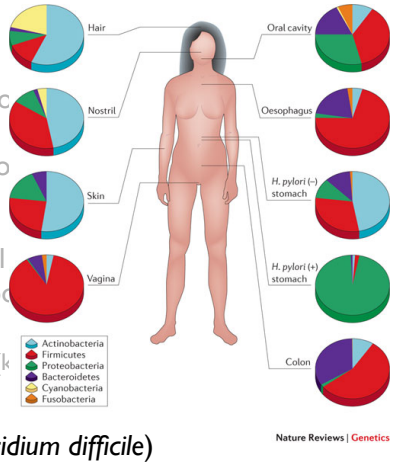
Industrial Applications
- Wastewater treatment (*V. cholera,* algal blooms, etc.)
- Bioprospecting (novel enzymes, compounds)
- Novel biosynthesis
- Fermentations: Consortia (yogurt) / wild (kombucha, Belgian ales)



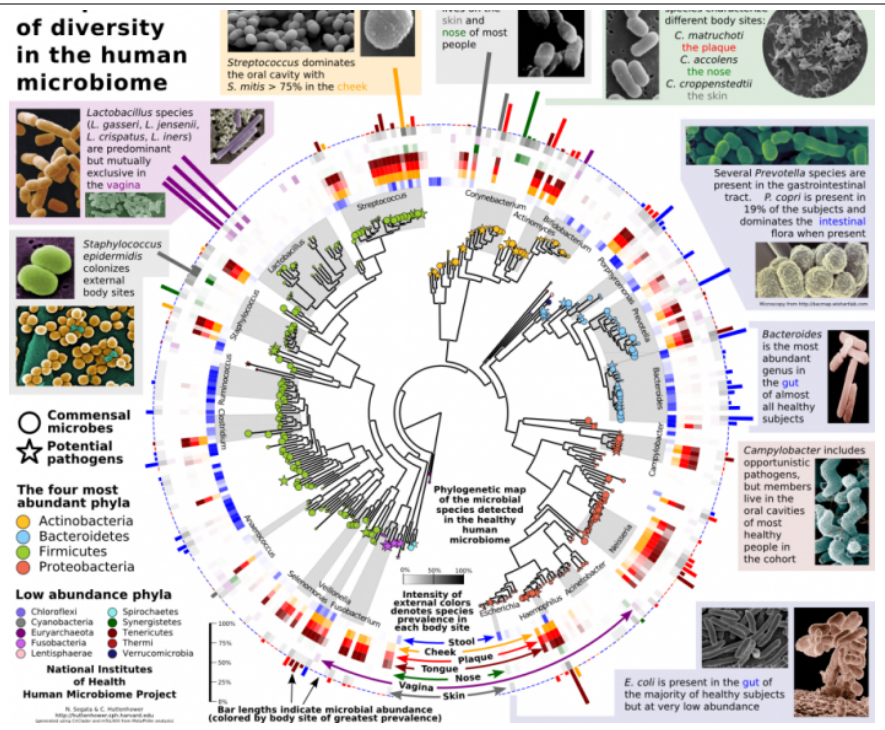
Why study microbiomes?

Environmental Science
- Critical elemental cycles (carbon, nitrc
- Pollution control, cleanup
- Ecology / Evolution (chloroplasts, mito

Industrial Applications
- Wastewater treatment (*V. cholera,* algal
- Bioprospecting (novel enzymes, compo
- Novel biosynthesis
- Fermentations: Consortia (yogurt) / wild (k



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, …)



Segata and Huttenhower

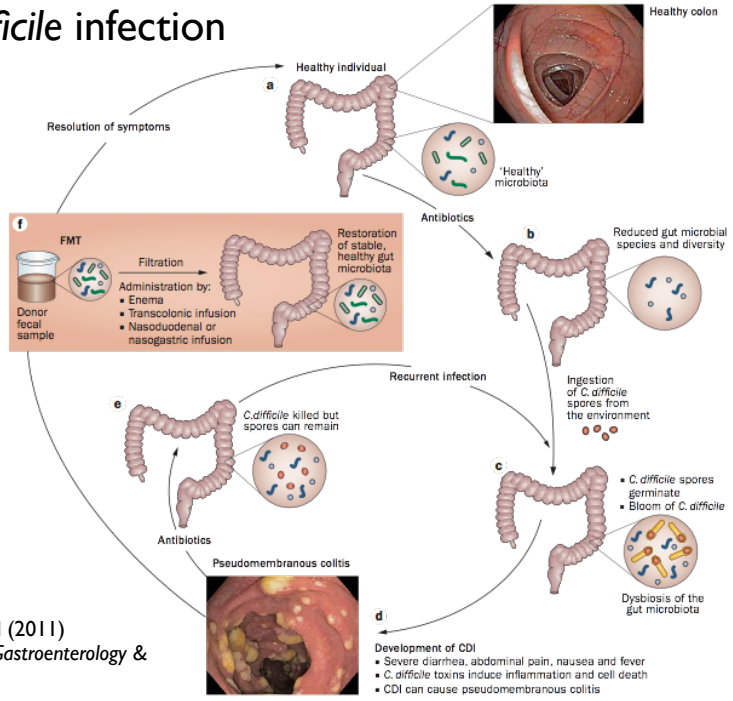
Some provocative oversimplifications…

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?

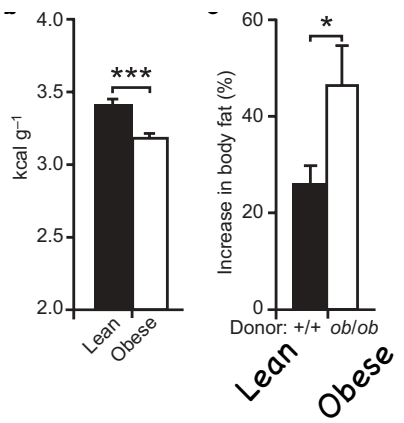
C. difficile infection



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

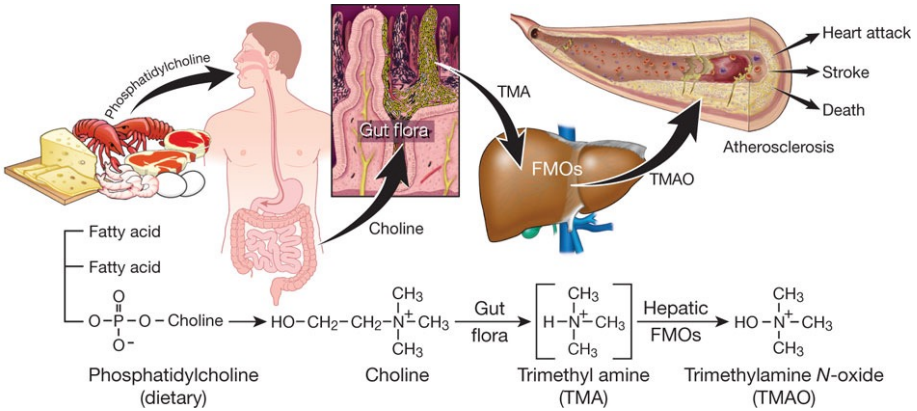
Microbes can make you fat(ter)…

- 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

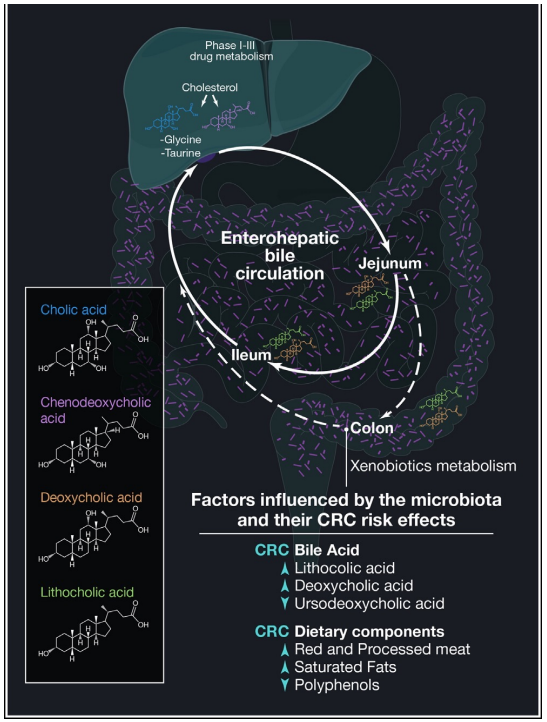


- Gut flora required for production of TMAO
- Supplementing diet with choline or TMAO promotes atherosclerosis (mouse)
- Gut flora suppression (Abx) inhibits dietary choline enhanced atherosclerosis
- TMAO is also a renal (kidney) toxin. Fogelman, A. M. (2015). *Circulation Research.*

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

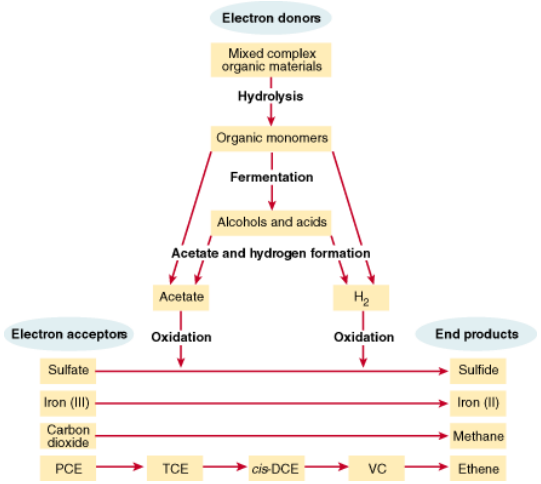
Colorectal Cancer (CRC)

- 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

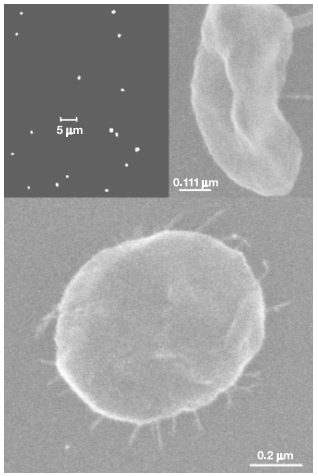


Sears, C. L., & Garrett, W. S. (2014). Microbes, Microbiota, and Colon Cancer.
Cell Host & Microbe, 15(3), 317–328.

Groundwater: Chlorinated Solvents



Dehalococcoides

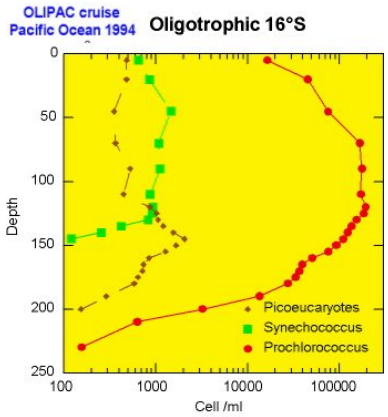


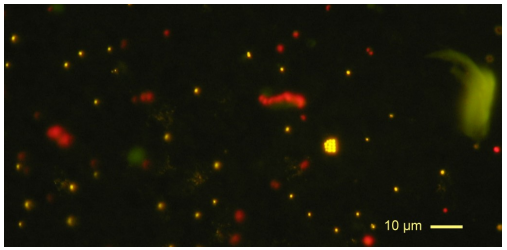
McCarty, P. L. (1997). Breathing with chlorinated solvents. *Science*

Bonus microbiome show-and-tell

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

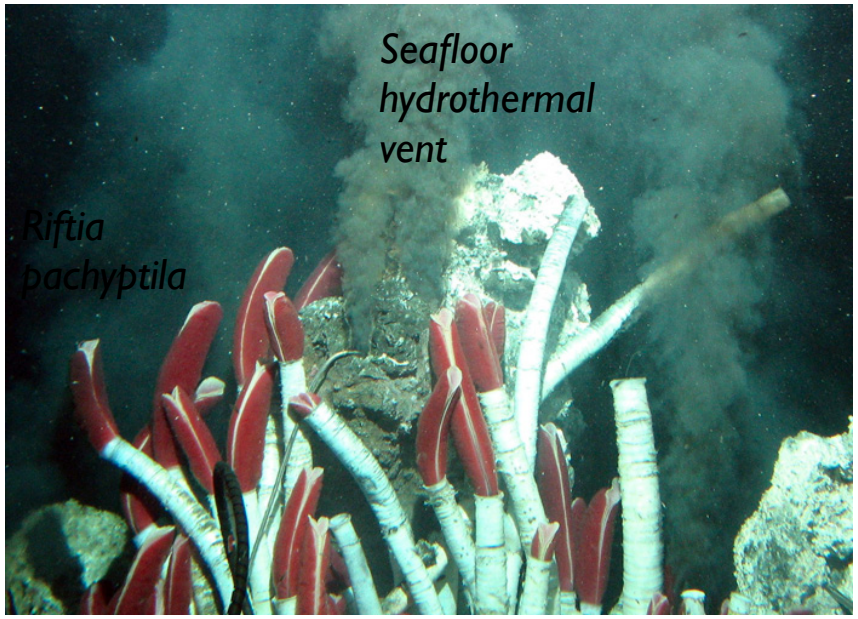
- 90° to 93°C
- extremely low in nutrients
- contains abundant biomass
- home to "oldest" known bacteria

Obsidian Pool

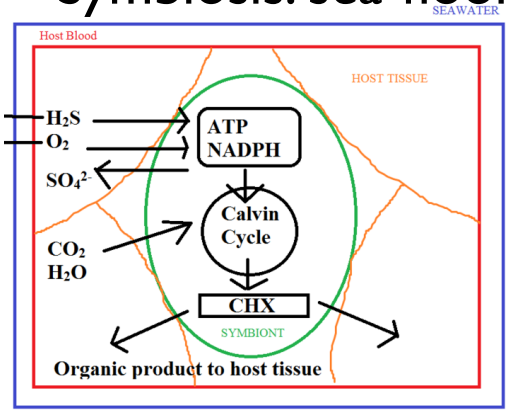
- 75° - 95°C
- high iron (II) hydrogen sulfide
- extensive diversity (previously unknown)

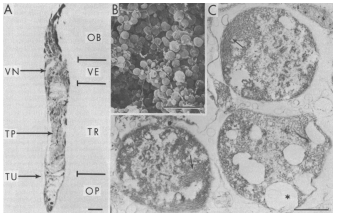
Ward, D. M., Weller, R., & Bateson, M. M. (1990). Nature, 345(6270), 63–65.
Barns, S. M., Fundyga, R. E., Jeffries, M. W., & Pace, N. R. (1994). PNAS 91(5), 1609–1613.

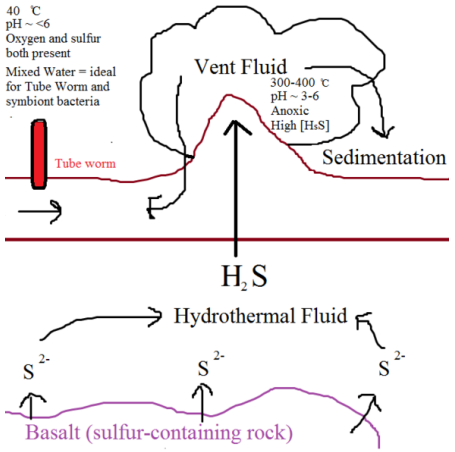
Symbiosis: sea-floor vent tube worm



Symbiosis: sea-floor vent tube worm





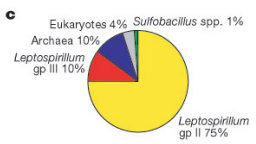


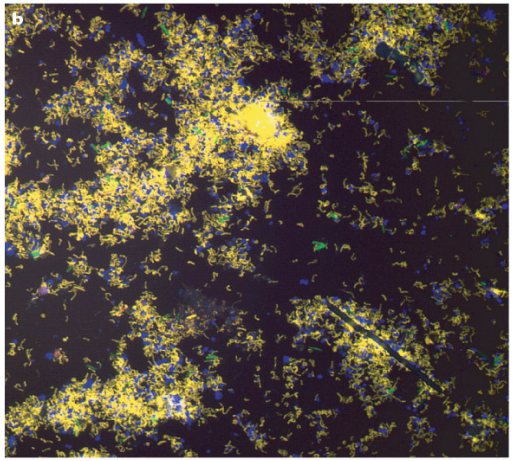
Cavanaugh, C. M. (1983). Nature, 302(5903), 58–61.
Cavanaugh, C. M., et al. (1981). Science. 213(4505), 340–342

Example of "model" microbiome: acid mine biofilm

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

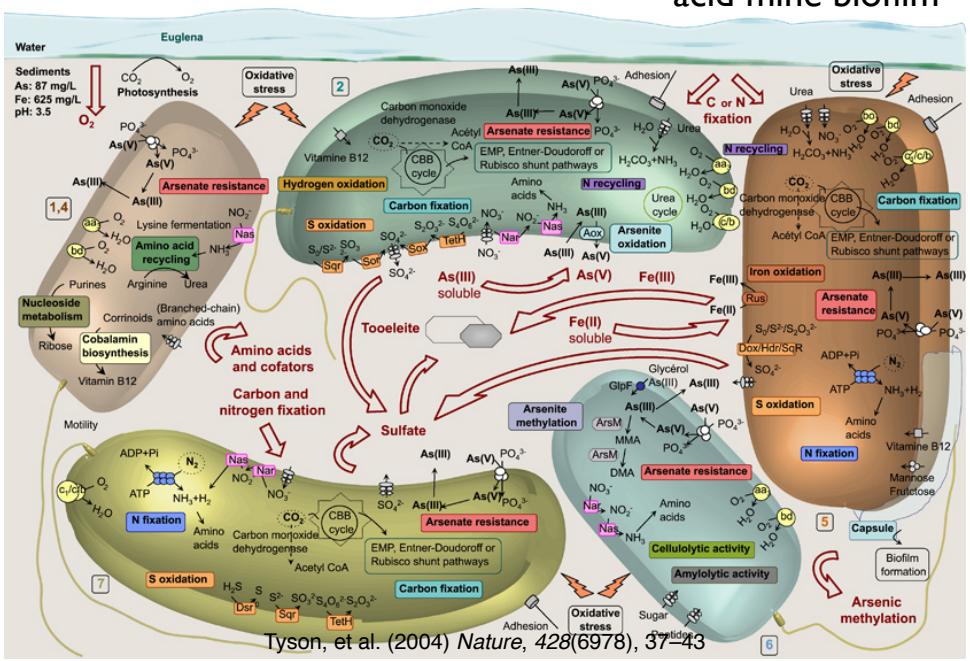
acid mine biofilm





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

acid mine biofilm



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

End: Biological Motivation

Questions before moving on?

Metagenomics Experimental Methods

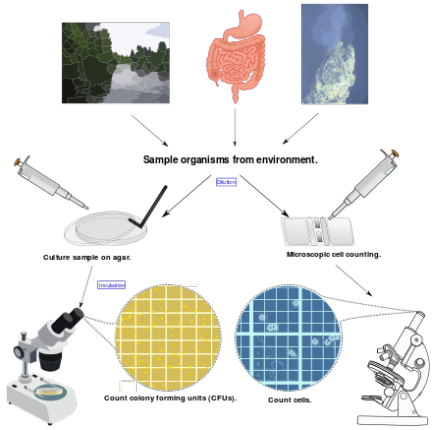
Exercise: How many species are present?

Confer amongst yourselves. We'll take a poll.

Discovery of *Culture Independent* Techniques

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**



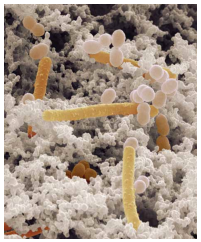
Staley, J.T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, 39, 321–346.

Discovery of *Culture Independent* Techniques

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"…

Discovery of *Culture Independent* Techniques

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

Discovery of *Culture Independent* Techniques

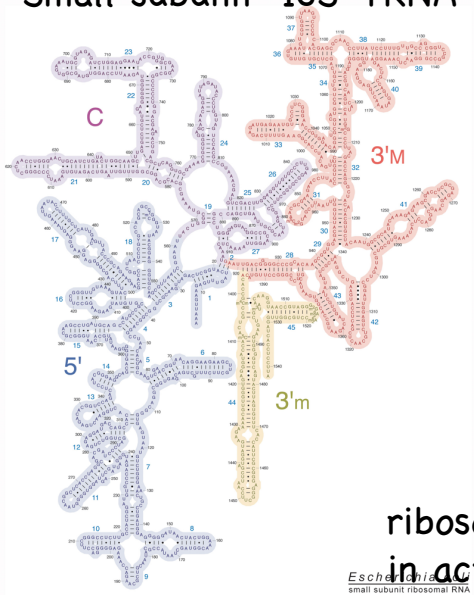
- 1977 rRNA as evolutionary marker - Woese & Fox *PNAS*
- 1985 Polymerase Chain Reaction (PCR) - K. Mullis
- 1985 "Universal" Primers for rRNA sequencing - N. Pace *PNAS*
- 1989 PCR amplification of 16S rRNA gene - Böttger *FEMS Microbiol.*
- 1996 Large, curated rRNA database (RDP) - Maidak *Nuc. Acids Res*
- 1998 *metagenome* genomics of communities coined by Jo Handelsman
- 2001 *microbiome* coined by Joshua Lederberg

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

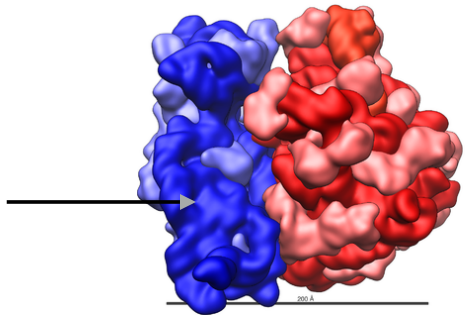
Discovery of *Culture Independent* Techniques

Small subunit "16S" rRNA

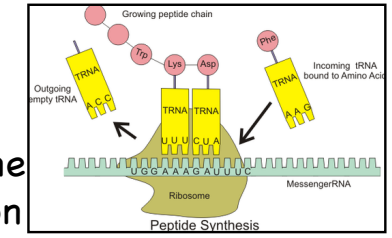


Escherichia coli small subunit ribosomal RNA

ribosome



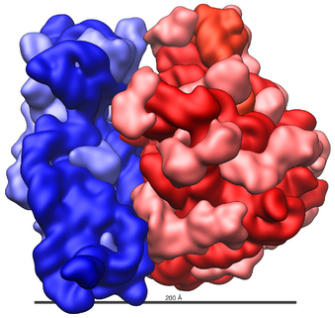
ribosome in action

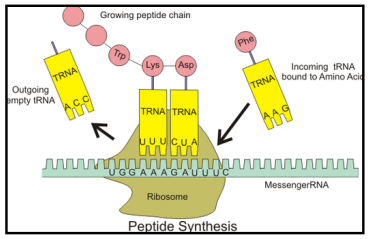


Discovery of *Culture Independent* Techniques

- 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.

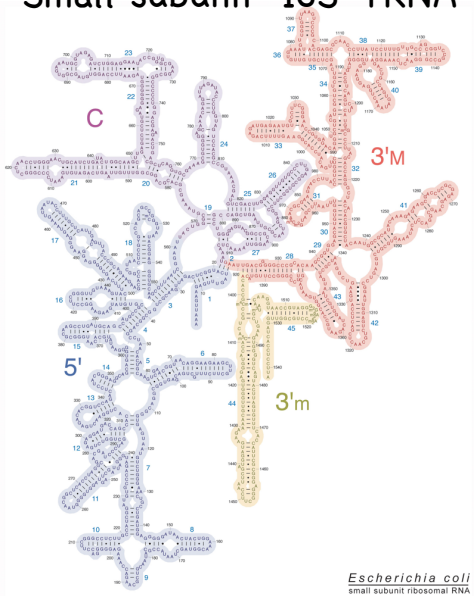
ribosome





Discovery of *Culture Independent* Techniques

Small subunit "16S" rRNA



Escherichia coli small subunit ribosomal RNA

- **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
- **Feasibility** - The right size for available sequencing technology (e.g. Sanger)

Discovery of *Culture Independent* Techniques

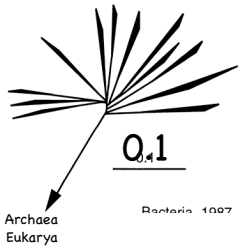
16S rRNA phylogeny, Known Bacteria

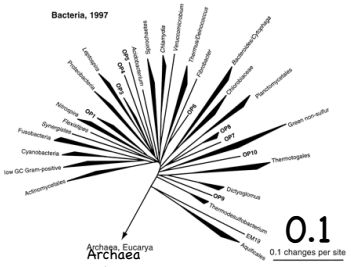
genome phylogeny

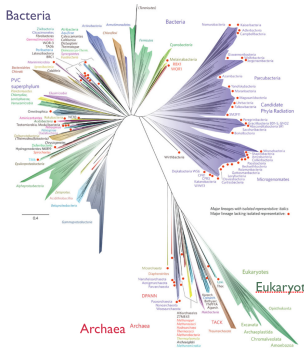
1987

1997

2016







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

A summary of metagenomics technique

Lyse all cells
Extract Total DNA (and/or RNA)

Amplicon Sequencing:
PCR amplify a single marker gene, e.g. 16S rRNA

Shotgun Sequencing:
Direct sequencing of (fragments of) total DNA/RNA

DNA sequencer

Relative abundances,
Genomes,
Genes,
Metabolic profiling,
Genomic structure,
Genetic variants...

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

A summary of metagenomics technique

Who's there?
(Taxonomic profiling)

What are they doing?
(Functional profiling)

What does it all mean?
(Statistical analysis)

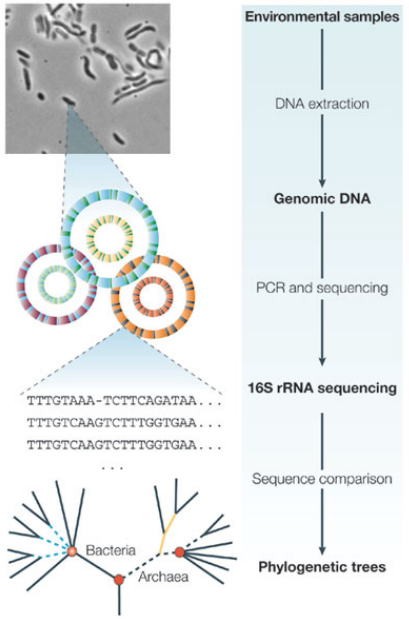
Slide adapted from Curtis Huttenhower, not necessarily with permission O:-)

A summary of metagenomics technique

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



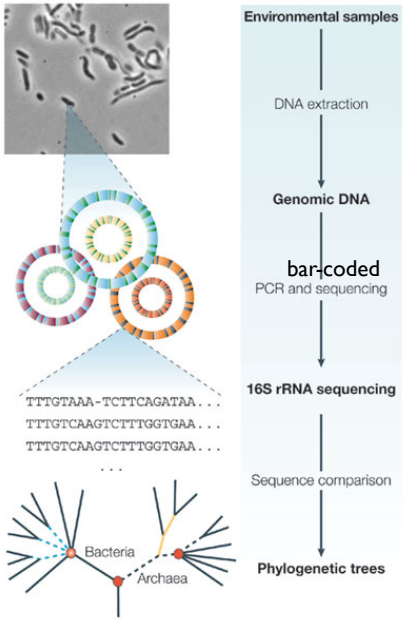
Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, *6*(11), 805–814.

A summary of metagenomics technique

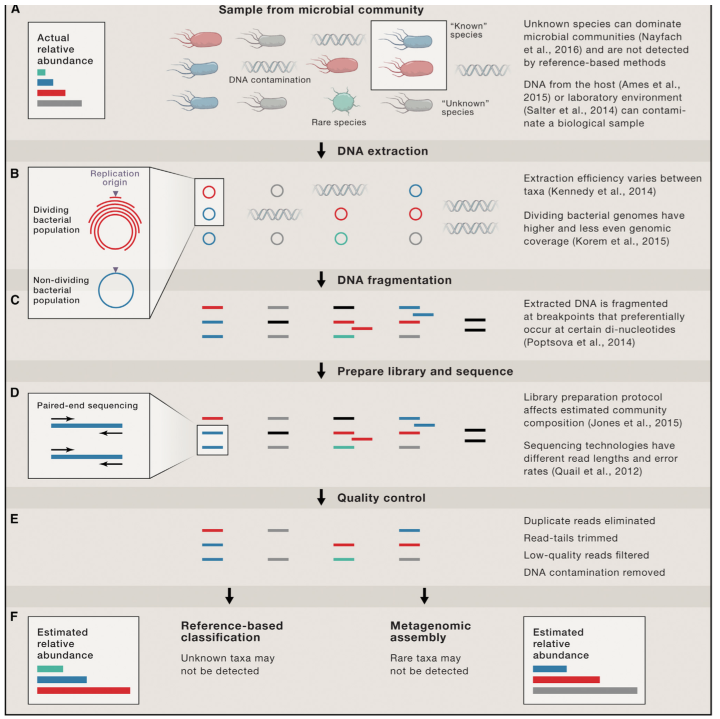
Amplicon sequencing

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



"Shotgun" sequencing



Nayfach, S., & Pollard, K. S. (2016). Perspective. Cell, 166(5), 1103–1116.

A summary of metagenomics technique

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 region~~s~~ across the genome<u>s</u>
- Cost is a fraction of the cost of re-sequencing the <u>meta</u>genome<u>s</u>

A summary of metagenomics technique

Culture Independent Techniques:

Metagenomics — Number of Species Counted

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

$

A summary of metagenomics technique

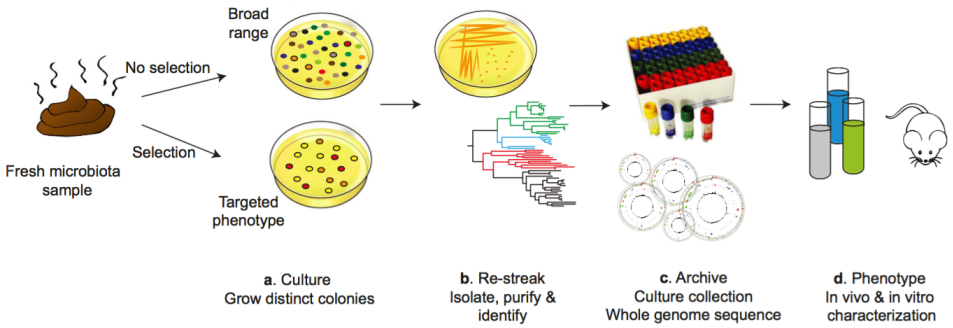
- Piles of short DNA/RNA reads from >1 organism
- You can...
  - Ecologically profile them
  - Taxonomically or phylogenetically profile them
  - Functionally profile them – gene/pathway catalogs
  - Comparative/structural genomics
- Prior knowledge is helpful
- Caution: Correlation ≠ Causation
  - Most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

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

Where things are headed: "Culturomics"

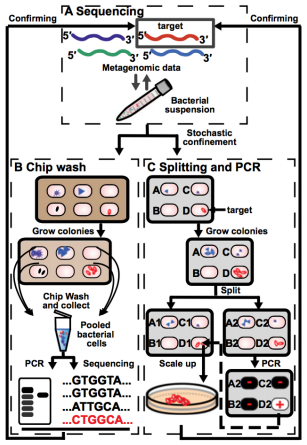
"Bacterial culture was the first method used to describe the human microbiota [after the microscope], but this method is considered outdated by many researchers ... however, a '*dark matter*' of prokaryotes, which corresponds to a hole in our knowledge and includes minority bacterial populations, is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

Lagier, J.-C., et al (2015). *The Rebirth of Culture in Microbiology… Culturomics…* Clinical Microbiology Reviews, 28(1), 237–264.

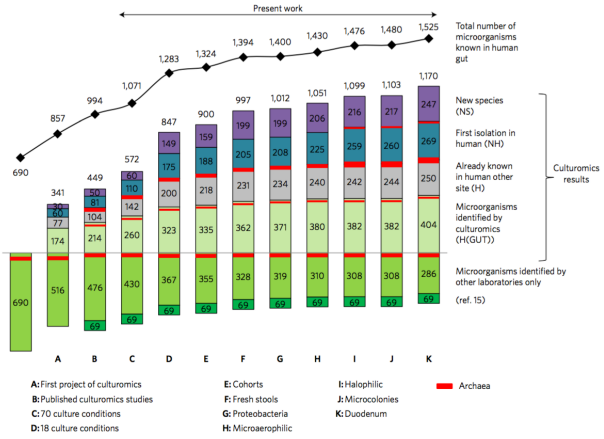


Browne, H. P., et al. (2016). Culturing of "unculturable" human microbiota… Nature, 533(7604), 543–546.

Where things are headed: "Culturomics"



Ma, L., et al. (2014). *Gene-targeted microfluidic cultivation…* PNAS,111(27), 9768–9773.



Lagier, J.-C., et al. (2016). *Culture of previously uncultured…* Nature Microbiology, 1(12), 1-8

An Introduction to Metagenomics

Outline for morning lecture:

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

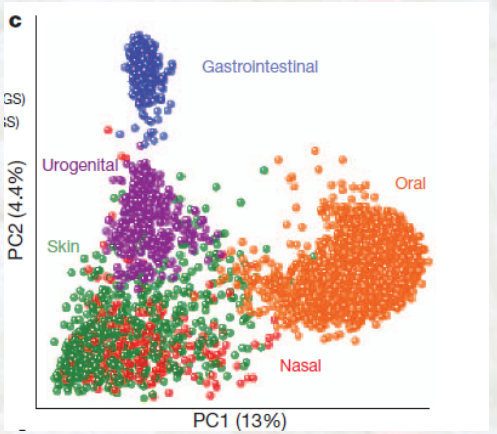
  } Biological motivation
- Methods
  - Experimental methods ✓
  - Analysis theory
  - Analysis tools, practices

  } Methods

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

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

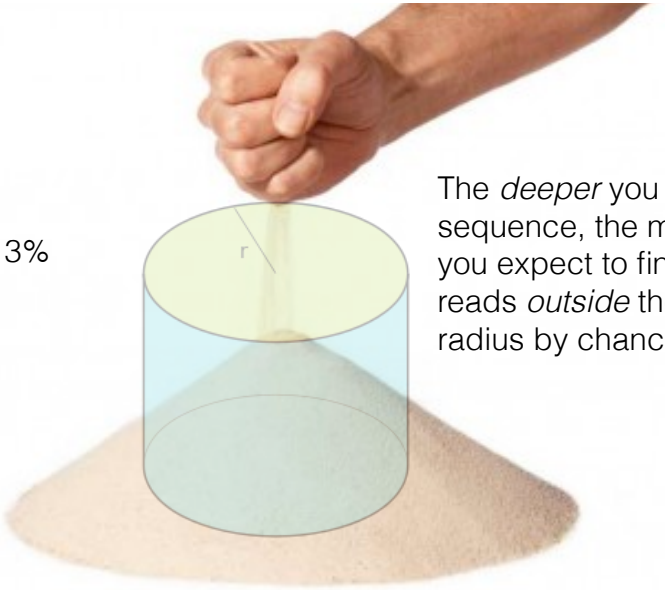
These presumptions are untrue.

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

Motivation: Lingering problems with "OTU"

imagine sequencing reads streaming from a single true sequence…

Motivation: Lingering problems with "OTU"



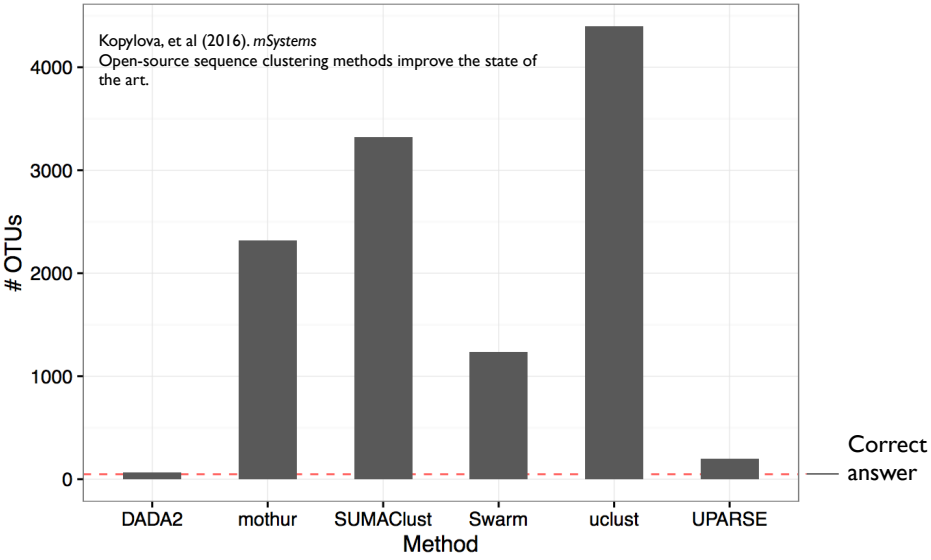
The *deeper* you sequence, the more you expect to find reads *outside* the radius by chance.

Motivation: Lingering problems with "OTU"

- 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 (~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

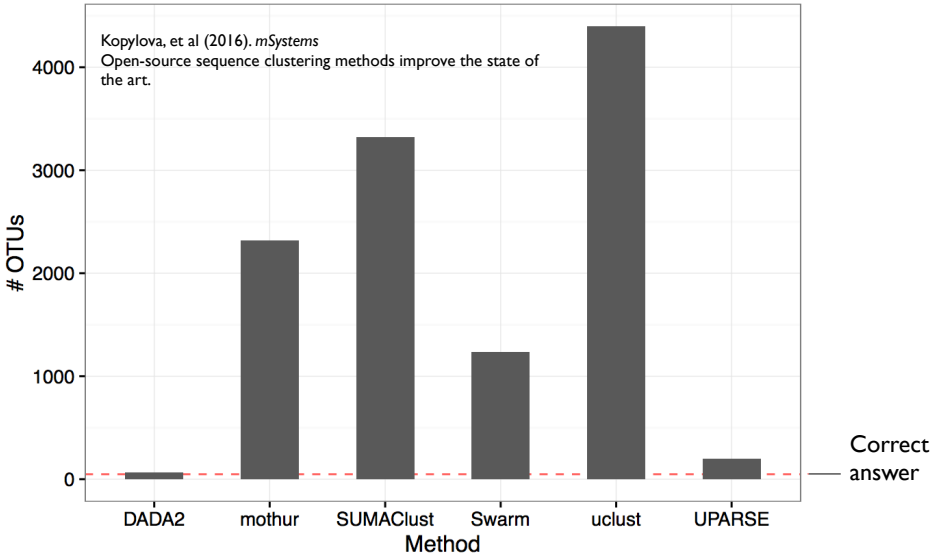
"if getting the wrong answer as quickly as possible is important... then there are a number of options…"
—Jon Bentley (as conveyed by R. Gentleman, BioC 2016)

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



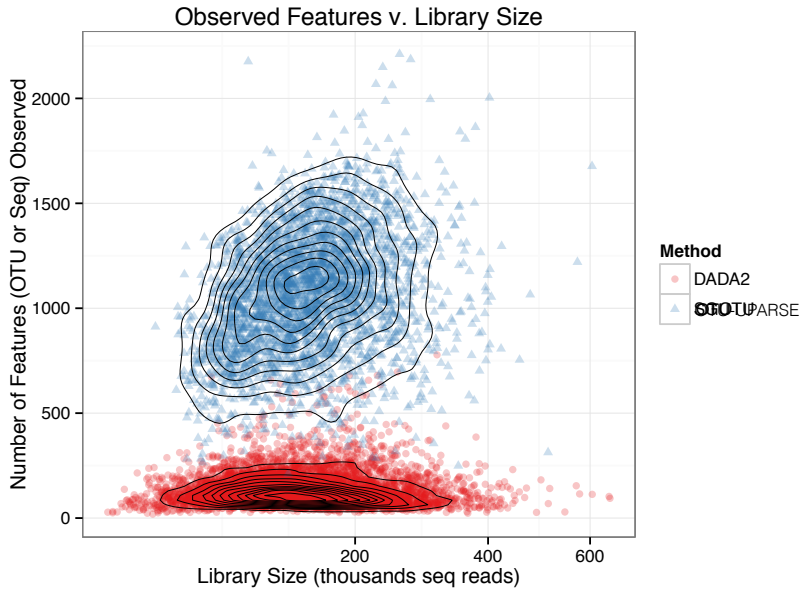
http://benjjneb.github.io/dada2/R/SotA.html

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



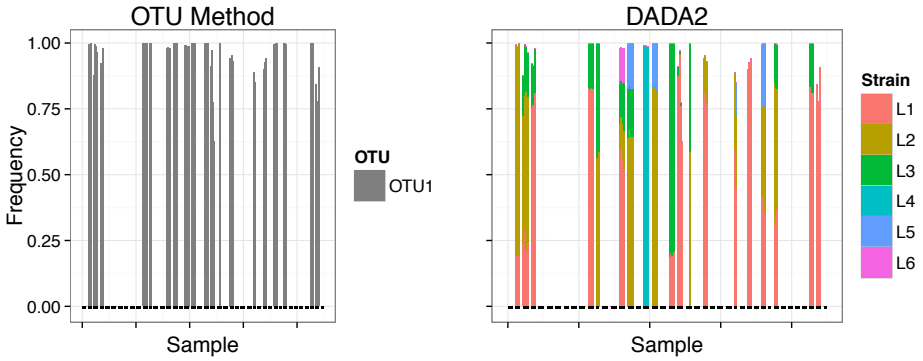
http://benjjneb.github.io/dada2/R/SotA.html

Anecdotal example of mitigated dependence of observed richness on sequencing effort



Example: vaginal microbiome

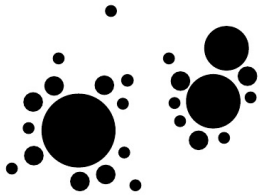
- Subject-discriminating strain-level resolution of *Lactobacillus crispatus*
- Repeated samples from vaginal microbiome of 42 pregnant women



Data: MacIntyre et al. Scientific Reports, 2015.

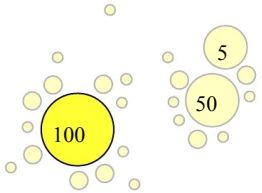
How does this work?

DADA2 algorithm cartoon



Initial guess: one real sequence + errors

DADA2 algorithm cartoon

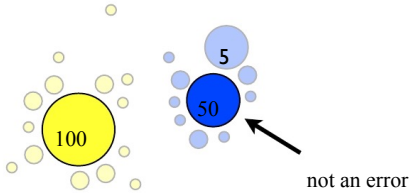


Infer initial *error model* under this assumption.

$Pr(i \rightarrow j) =$

	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

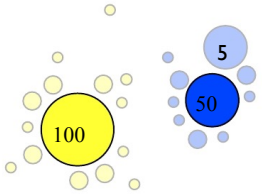
DADA2 algorithm cartoon



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

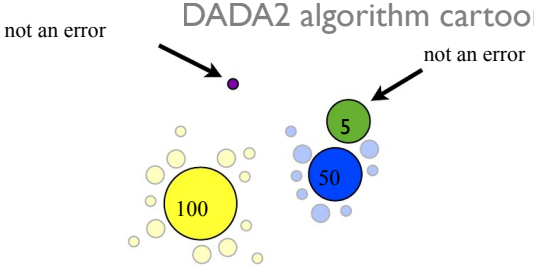
DADA2 algorithm cartoon



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

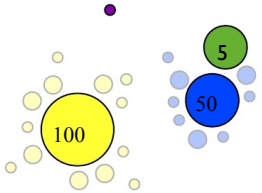
DADA2 algorithm cartoon



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

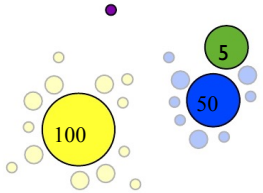
DADA2 algorithm cartoon



Update model again

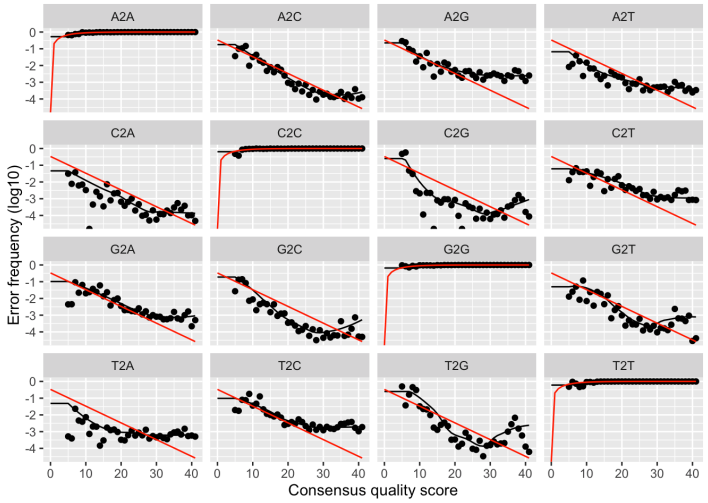
	A	C	G	T
A	0.998	$1x10^{-4}$	$2x10^{-3}$	$2x10^{-4}$
C	$6x10^{-5}$	0.999	$3x10^{-6}$	$1x10^{-3}$
G	$1x10^{-3}$	$3x10^{-6}$	0.999	$6x10^{-5}$
T	$2x10^{-4}$	$2x10^{-3}$	$1x10^{-4}$	0.998

DADA2 algorithm cartoon



Convergence: all errors are plausible

	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



- *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

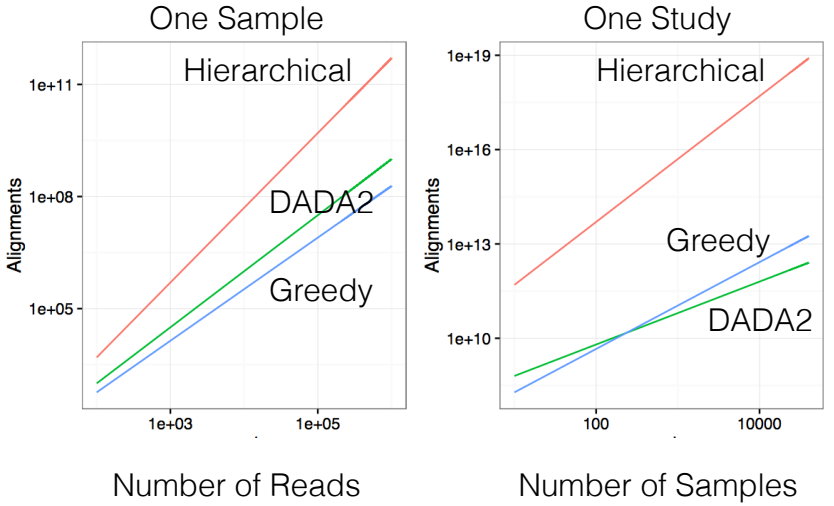
This sounds complicated.
Isn't it really expensive and time-consuming to compute?

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

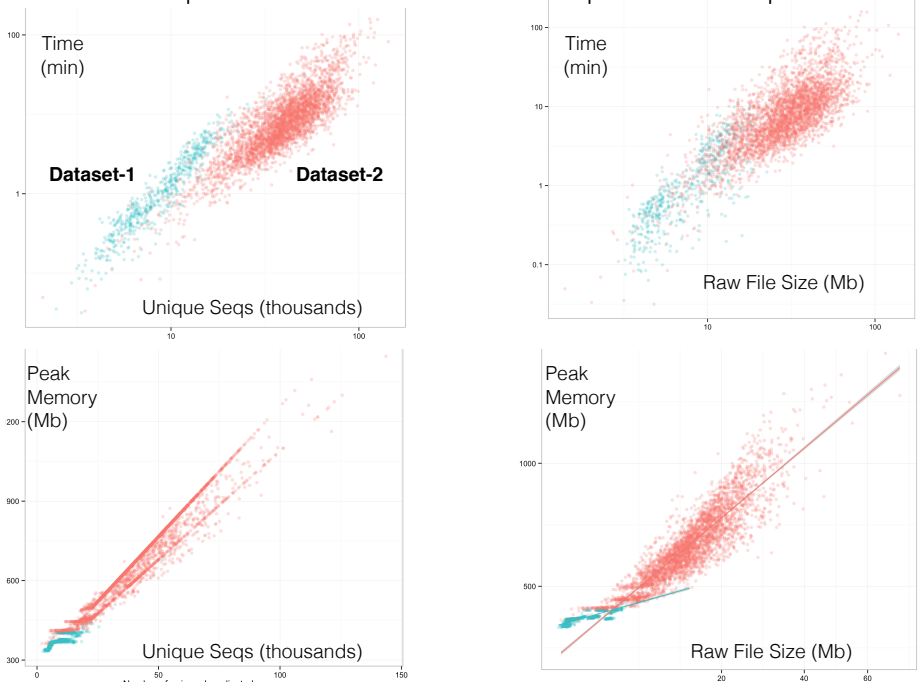
No.
Unlike OTU methods, DADA2 can work on each sequence library independently. The outputs are sequences themselves, which are intrinsically comparable. This has important bonus for computation: *embarrassingly parallel*

- "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

Compute performance, as the required number of sequence alignments



Computational Performance - each point is a sample



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

DADA$_2$

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 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://benjjneb.github.io/dada2/ R package available on BioConductor

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

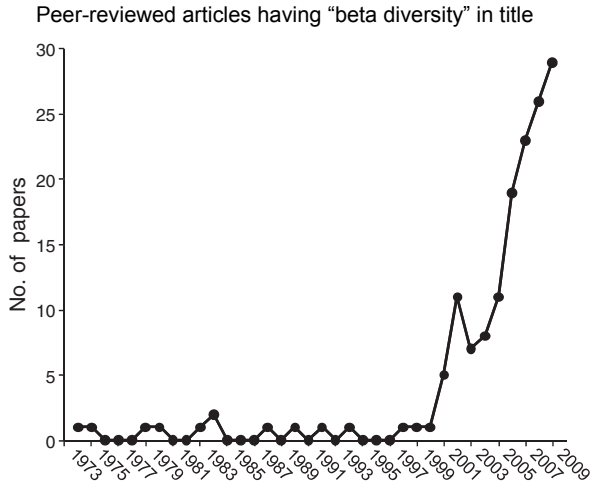
Diversity

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

or

Analysis comparing whole microbiomes to one another

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:-)

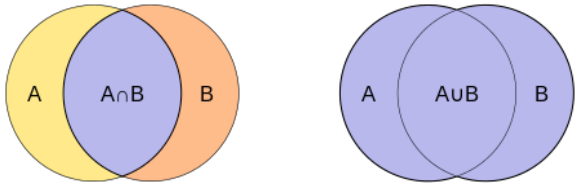
The Distance Spectrum

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

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

Jaccard

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

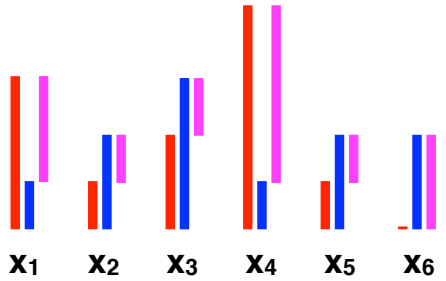


Intuition: Fraction of shared **types** unique to one of the communities

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

Bray-Curtis

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

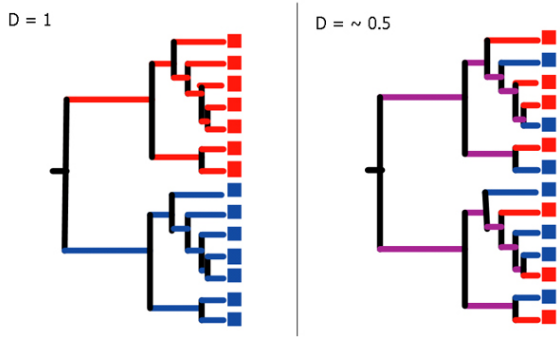


Intuition: *City block distance*. Sum of absolute differences over total abundance.

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

Unifrac

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

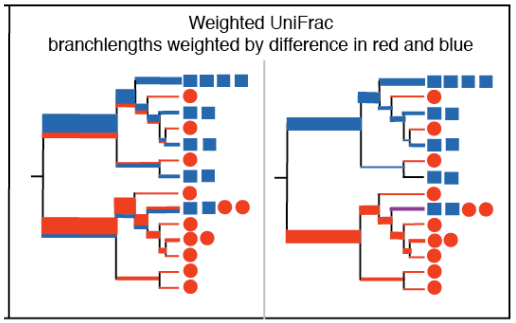


Intuition: Fraction of shared **tree** unique to one of the communities

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

Lozupone and Knight (2008)

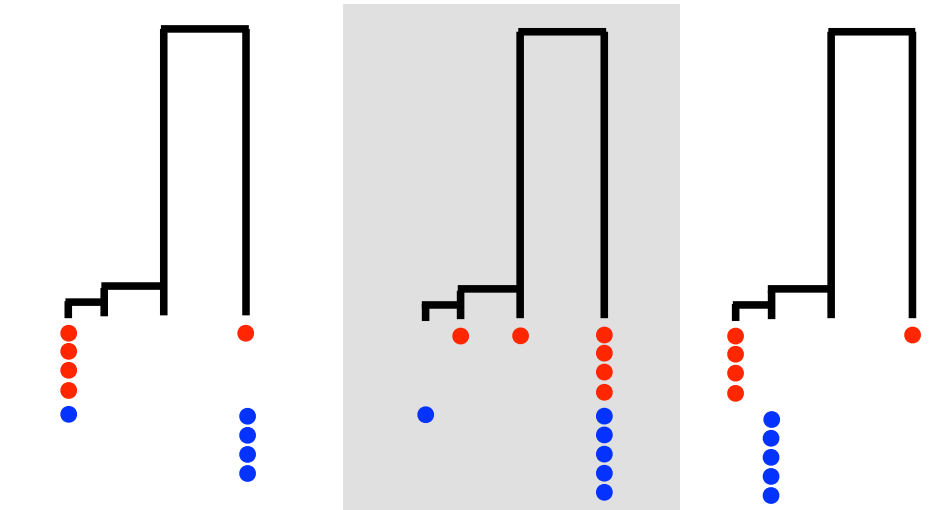
Weighted Unifrac



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)

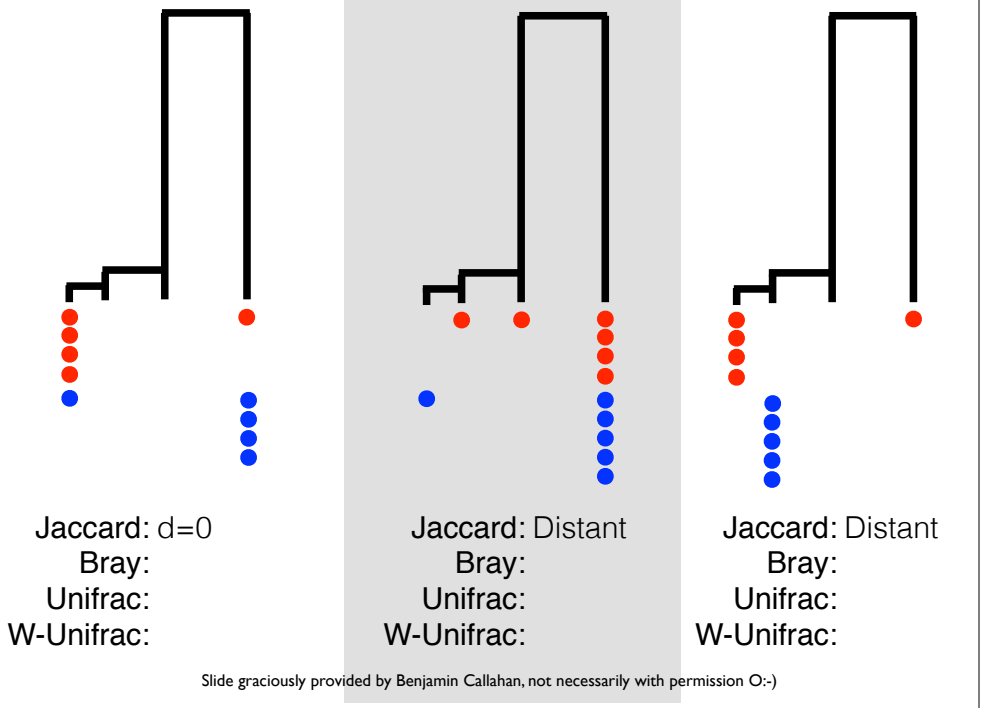


Jaccard:
Bray:
Unifrac:
W-Unifrac:

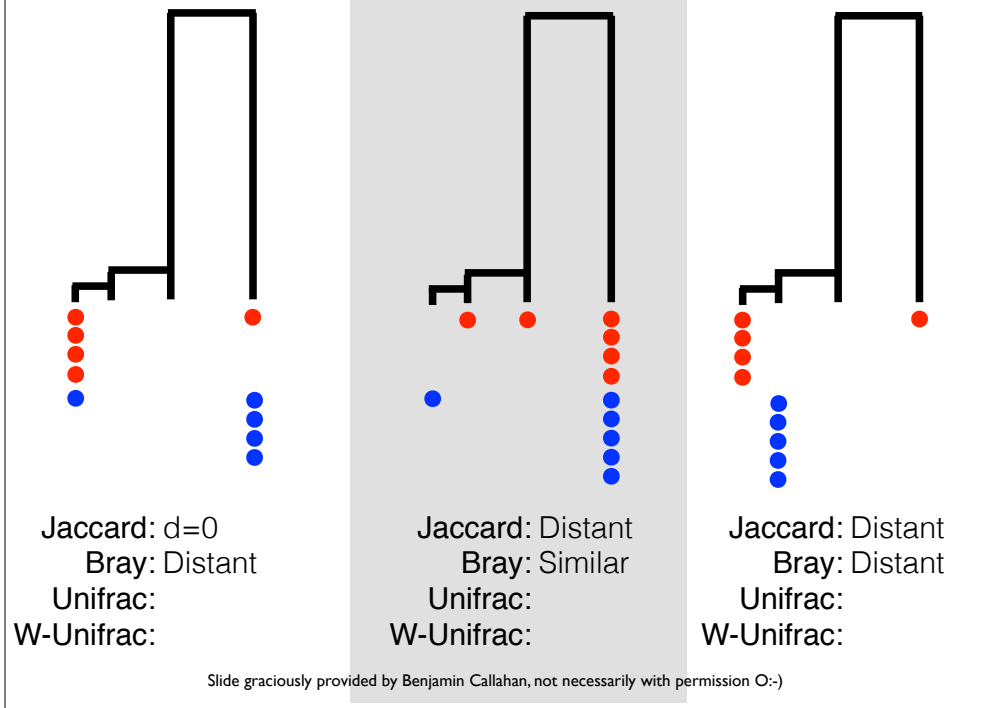
Jaccard:
Bray:
Unifrac:
W-Unifrac:

Jaccard:
Bray:
Unifrac:
W-Unifrac:

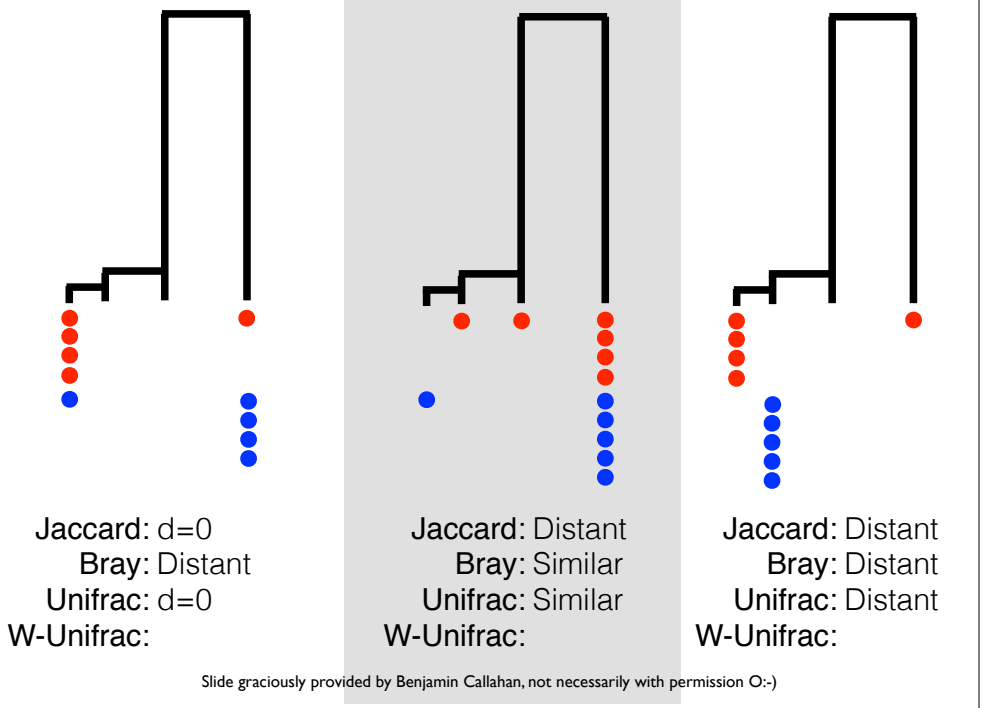
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)



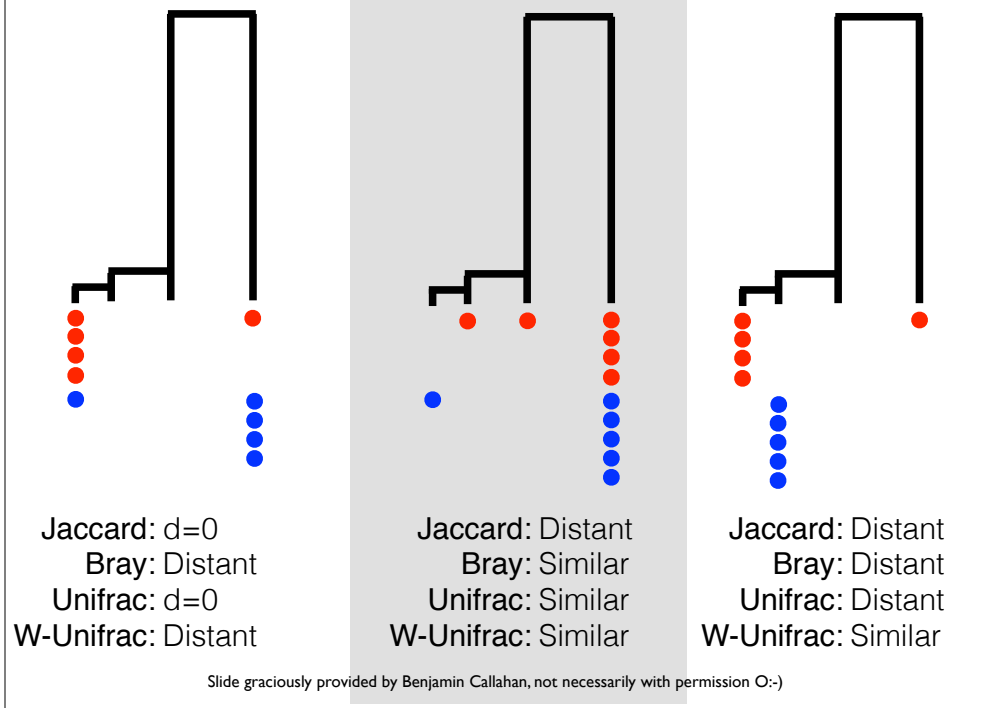
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)



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



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



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

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

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

What do we do with distances between microbiome samples?

For starters: Plot / exploratory analysis

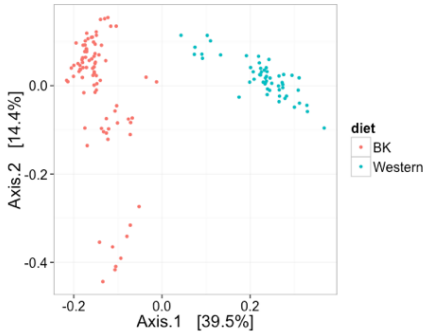
Ordination Methods

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,…
.
.



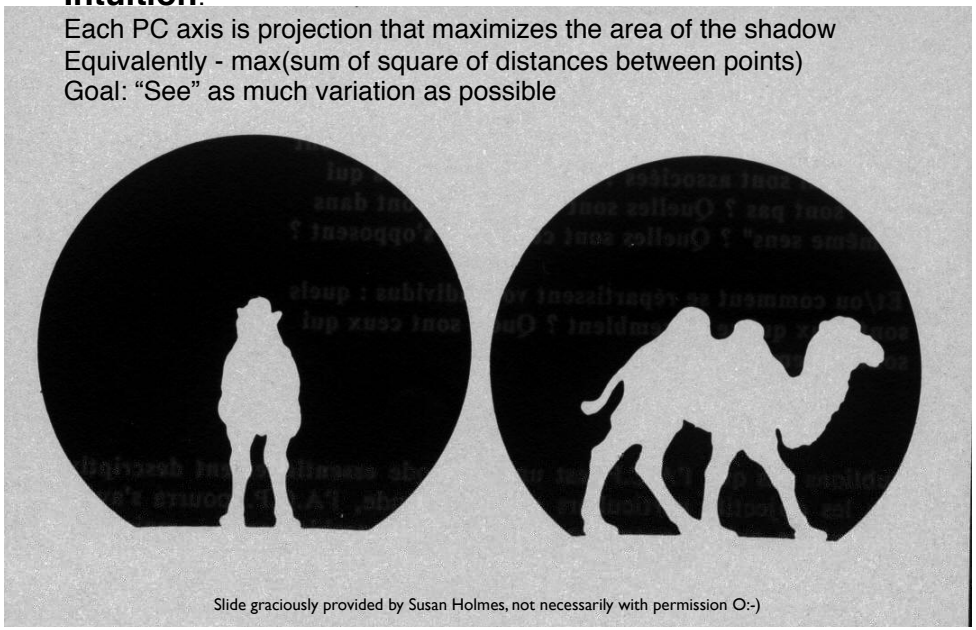
P-dimensions

2-dimensions

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

Ordination Methods

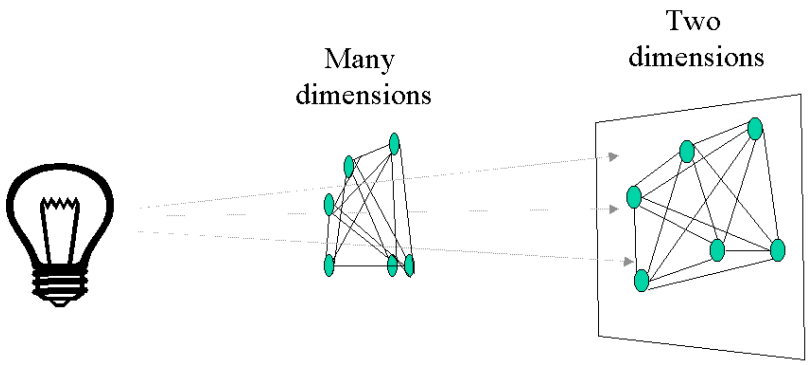
Intuition:
Each PC axis is projection that maximizes the area of the shadow
Equivalently - max(sum of square of distances between points)
Goal: “See” as much variation as possible



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

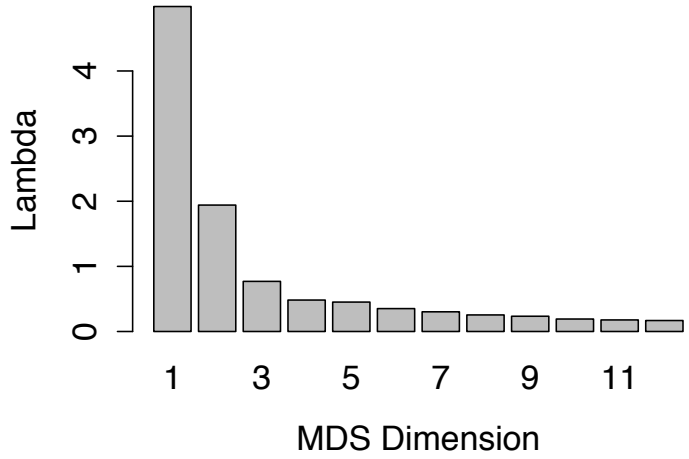
---

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

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



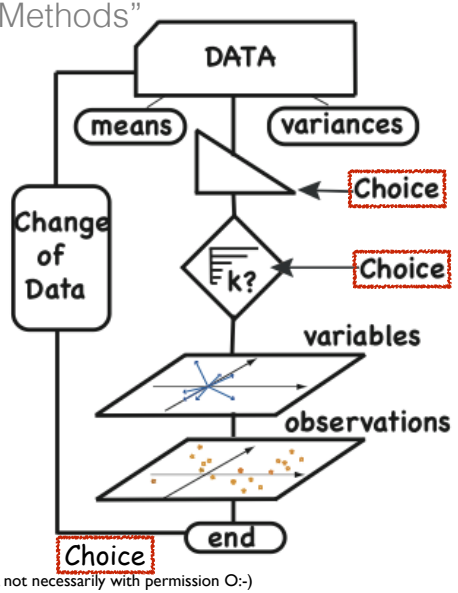
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

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

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

- Are there clusters? How many?

Technique:
Gap Statistic

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

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

Technique:
PC Regression (statistics' "PCR")

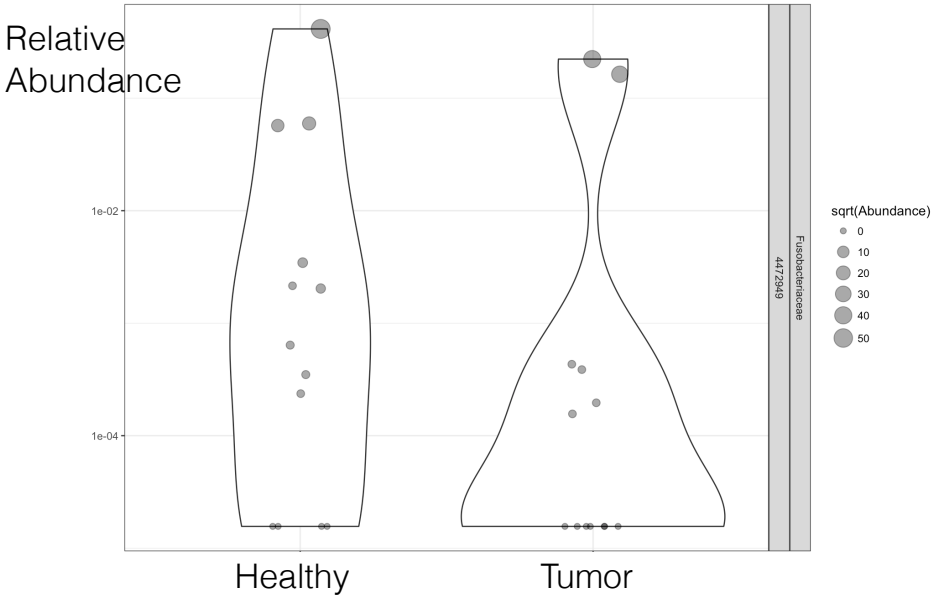
Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

- Are the patterns consistent with covariates?

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

(Multiple) Hypothesis testing of microbial differential abundance



(Multiple) Hypothesis testing of microbial differential abundance

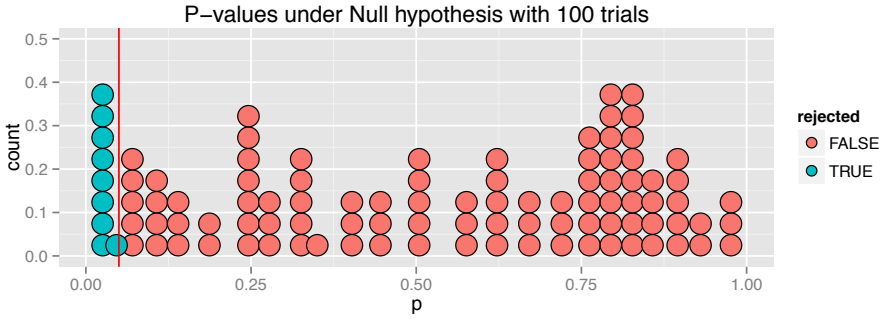
Clinically-motivated question:
"Does this microbe occur in different abundance between healthy/disease tissue?"

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*?

Multiple Hypothesis Testing

- 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*α`



Multiple Hypothesis Testing

TABLE 1
Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - R$	R	m

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

Multiple Hypothesis Testing

Independent filtering

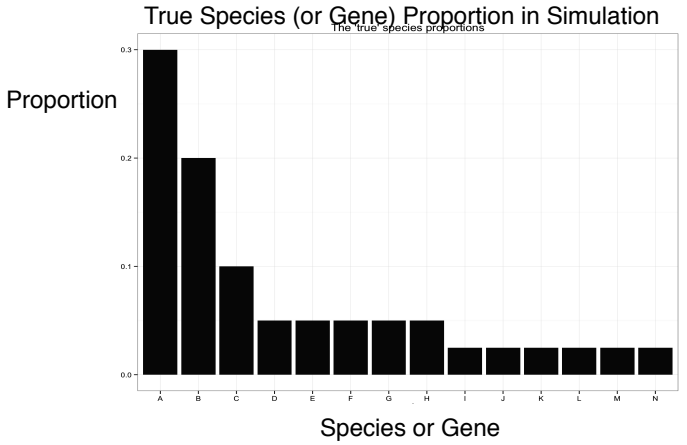
- Is a general approach that can substantially increase the [statistical] efficiency of experiments
- Uses filter/ test pairs that are independent under the null hypothesis
- but correlated under the alternative

e.g. remove features with very low mean abundance

Bourgon, Gentleman, & Huber (2010) Independent filtering increases detection power for high-throughput experiments. PNAS 107(21) 9546-9551

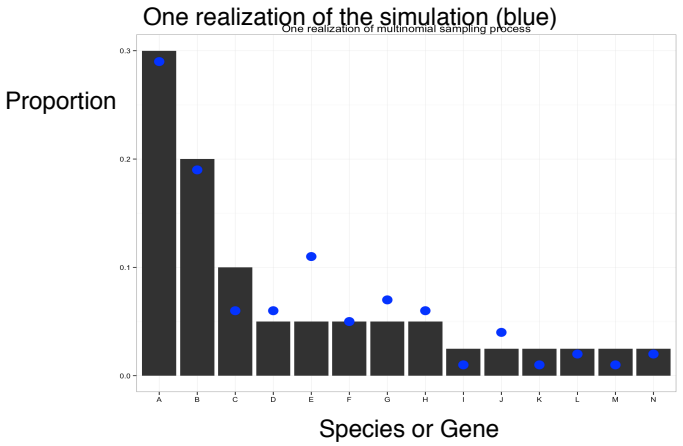
Model Uncertainty in NGS Count Data

Poisson-only Count Simulation



Model Uncertainty in NGS Count Data

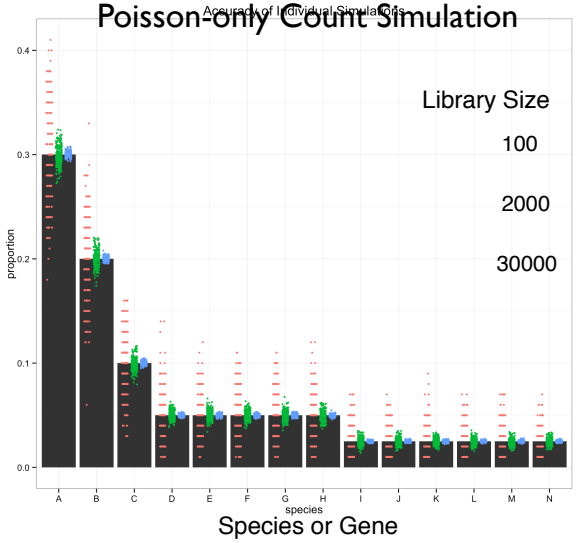
Poisson-only Count Simulation



Model Uncertainty in NGS Count Data

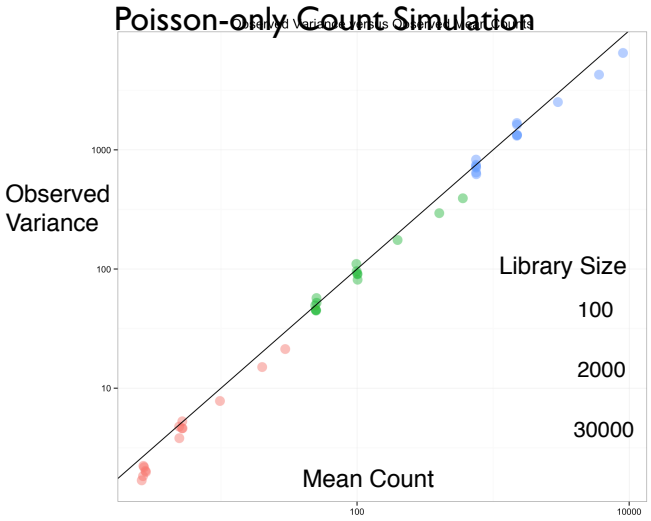
Poisson-only Count Simulation

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

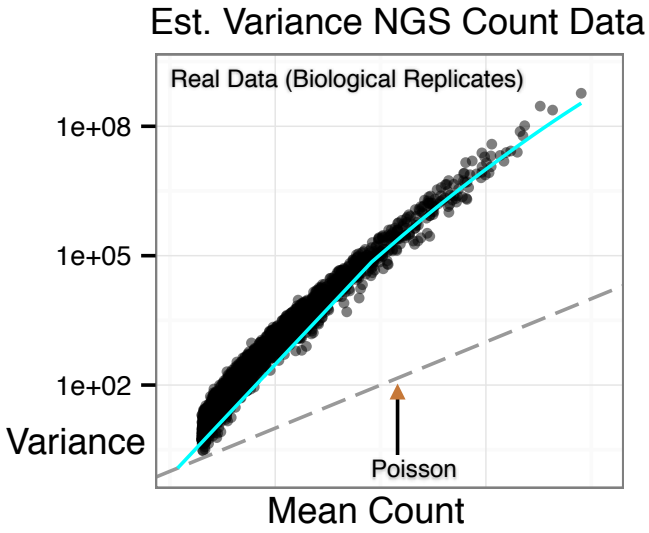


Model Uncertainty in NGS Count Data

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



Model Uncertainty in NGS Count Data



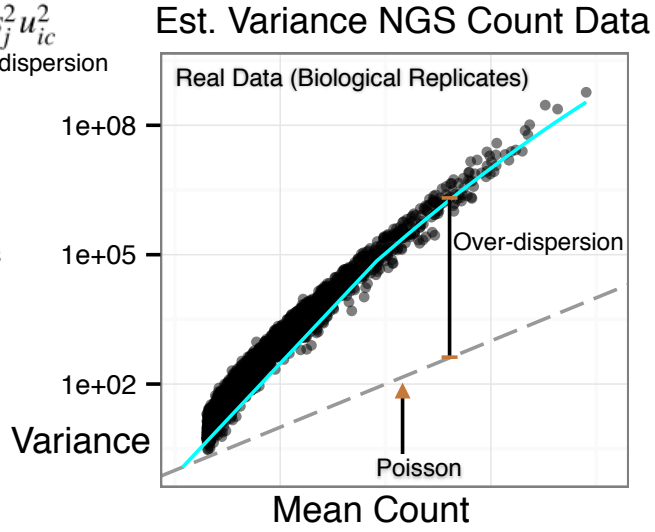
Model Uncertainty in NGS Count Data

Negative Binomial

Variance = $u_{ic}s_j + \phi_{ic}s_j^2 u_{ic}^2$

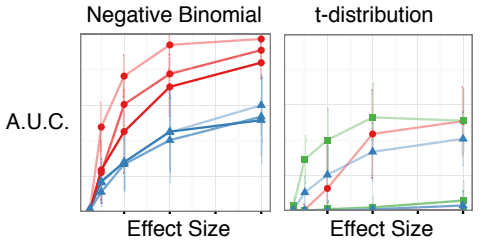
Poisson Overdispersion

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



Model Uncertainty in NGS Count Data

- 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)



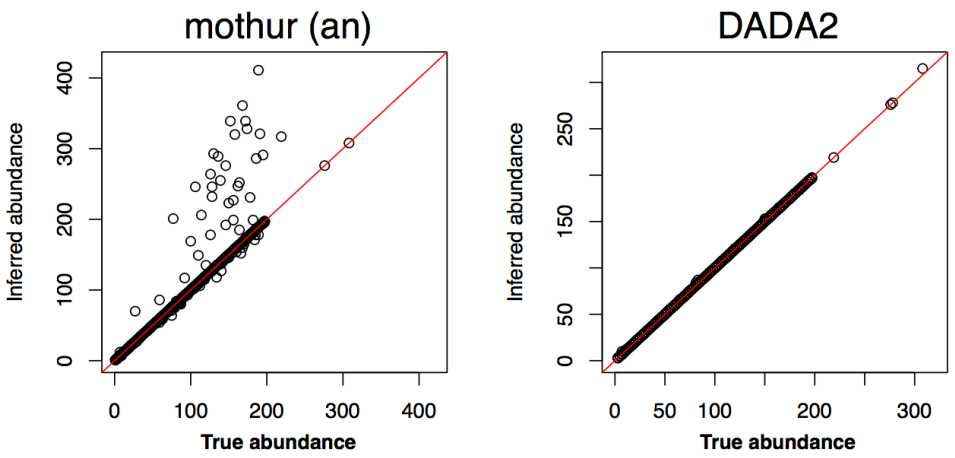
McMurdie & Holmes (2014). PLoS Computational Biology

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

End:
Introduction to Microbiome / Metagenome Analysis Concepts

Questions?

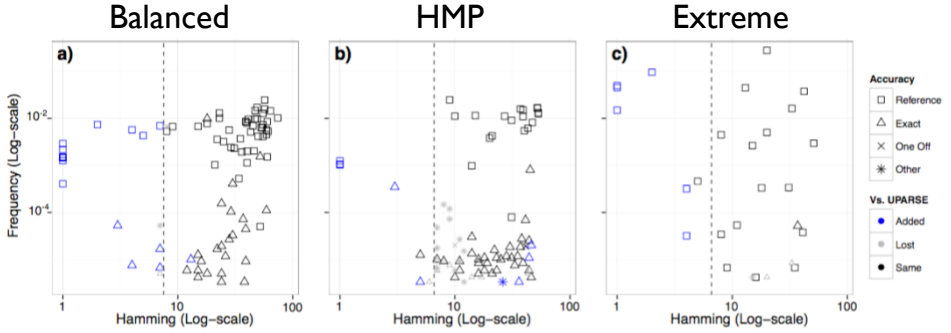
Performance in a computationally simulated community



TP:	978	**TP:**	1042
FP:	272	**FP:**	0
FN:	77	**FN:**	13

DADA2 Accuracy benchmarks

Three other "mock" microbiome communities



Better FP and FN performance than UPARSE…

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:

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

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

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:

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

- "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

Multiple Testing - Bonferroni

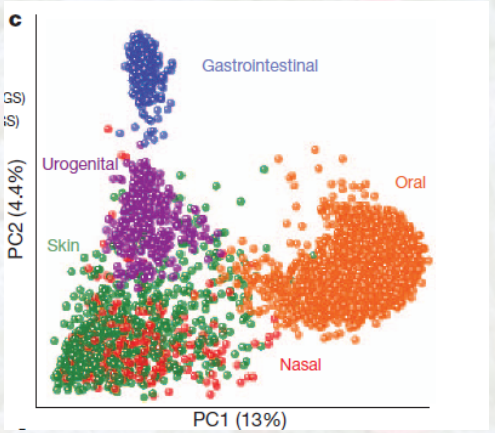
- To ensure overall significance at a given α, one performs each individual test at α' = α/m
- Useful when need to correct for just a few hypotheses
- Very stringent, results in "loss of power"
  - increase in Type II error, decreases sensitivity

Multiple Testing - Benjamini-Hochberg

- 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 $P_{FDR} <= 0.05$ will have 5% or fewer false positives
- This is what is meant be "controlling" the false positive rate

Benjamini & Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300.

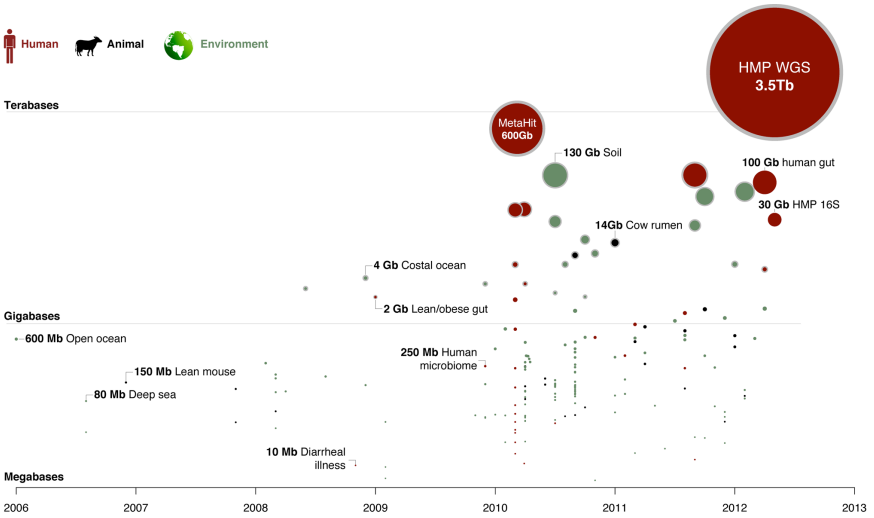
Introduction to Microbiome / Metagenome Analysis Tools and Practices



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

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:-)

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.)

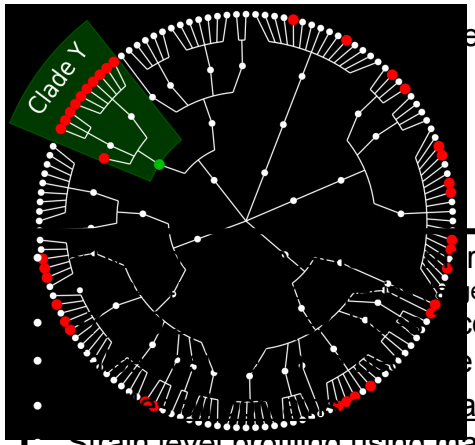
(16S rRNA) Amplicon Sequence Processing Tools:

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

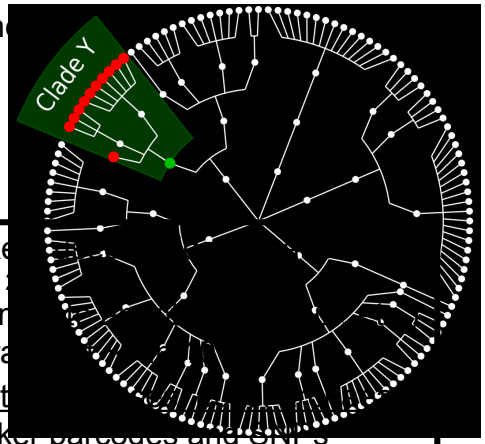
Afternoon will be spent using QIIME2
Daniel has much more to say about it…

MetaPhlAn: Taxonomic profiling using unique marker genes

X is a core gene for clade Y



X is a unique marker gene for clade Y

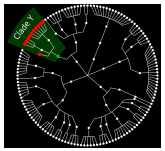


- Strain level profiling using marker barcodes and SNP's
- Quasi-markers used to resolve ambiguity in postprocessing

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

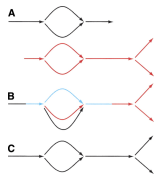
Typical shotgun metagenome and metatranscriptome analyses

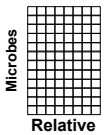
Taxonomic Profiling



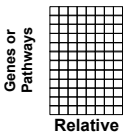
Functional Profiling

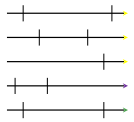
Assembly





Gene content inference (e.g. PICRUSt, etc.)





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

Microbiome meta'omic analyses: assembly

Scaling metagenome sequence assembly with probabilistic de Bruijn graphs

khmer (Pell 2012)

P25
MetAMOS: a metagenomic assembly and analysis pipeline for AMOS
Genome Biology 2011, 12(Suppl 1):P25

MetAMOS (Treangen 2013)

MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads

MetaVelvet (Namiki 2012)

BIOINFORMATICS
Meta-IDBA: a *de Novo* assembler for metagenomic data

Meta-IDBA (Peng 2011)

Genovo: *De Novo* Assembly for Metagenomes

*JONATHAN LASERSON, *VLADIMIR JOJIC, and DAPHNE KOLLER

Genovo (Laserson 2011)

BIOINFORMATICS ORIGINAL PAPER
IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth

IDBA-UD (Peng 2012)

METHOD Open Access
Ray Meta: scalable *de novo* metagenome assembly and profiling

Ray (Boisvert 2012)

SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing

SPAdes (Bankevich 2012)

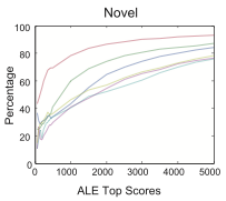
MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph

Dinghua Li[1,†], Chi-Man Liu[2,†], Ruibang Luo[2,†], Kunihiko Sadakane[3] and Tak-Wah Lam[1,2,*]

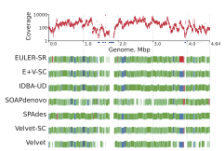
MEGAHIT (Li 2015)

Microbiome meta'omic analyses: assembly

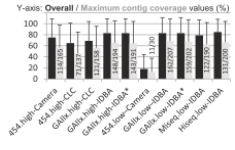
ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies



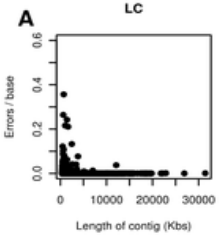
QUAST: quality assessment tool for genome assemblies



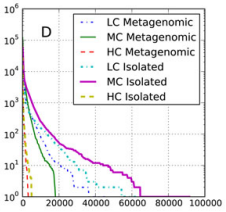
Evaluation of viral genome assembly and diversity estimation in deep metagenomes



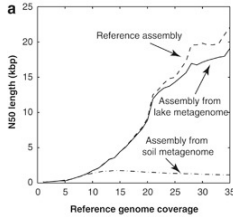
Evaluating the Fidelity of *De Novo* Short Read Metagenomic Assembly Using Simulated Data



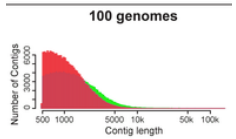
Evaluation of short read metagenomic assembly



Individual genome assembly from complex community short-read metagenomic datasets

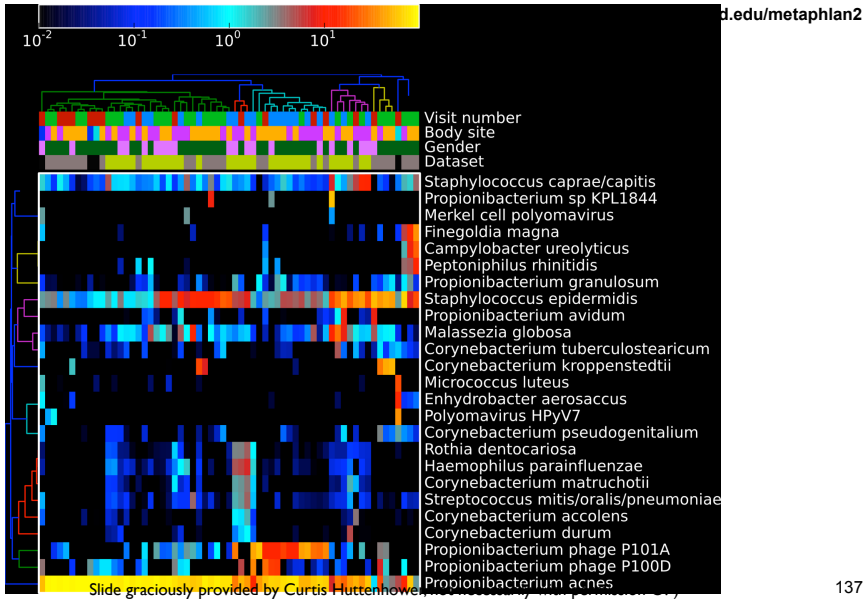


Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data



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

MetaPhlAn2: Trans-kingdom profiling



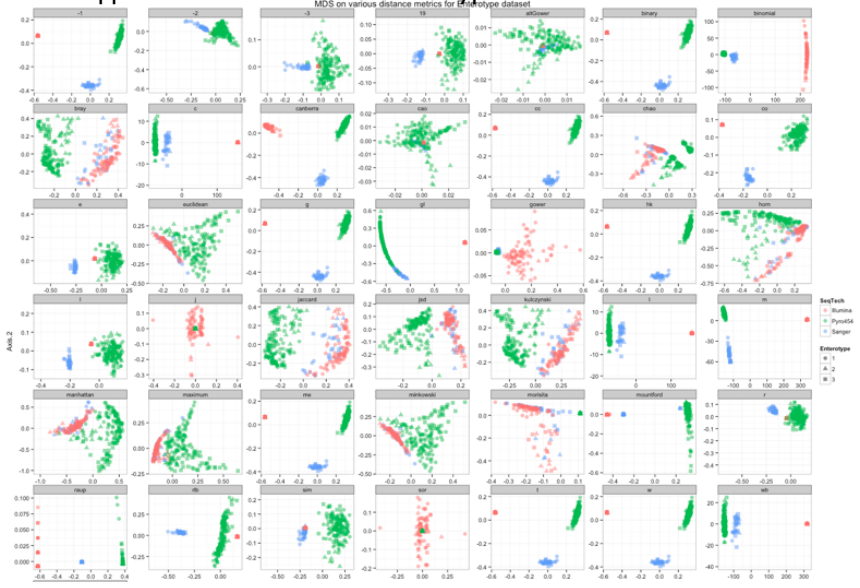
Slide graciously provided by Curtis Huttenhower

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)

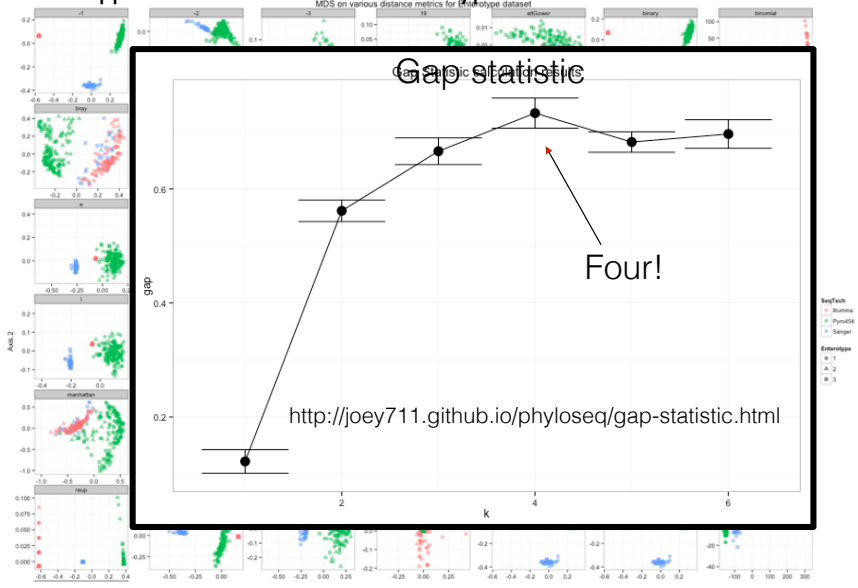
illustrative example favoring reproducible analysis: "Enterotypes of the human genome"

MDS on supported distance metrics: enterotype data



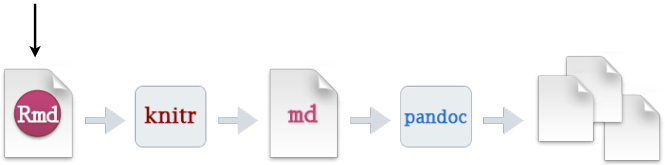
illustrative example favoring reproducible analysis: "Enterotypes of the human genome"

MDS on supported distance metrics: enterotype data



Reproducible analysis workflow with R-markdown

microbiome data



HTML
book pdf
article pdf
website
dashboard
Word Doc
…

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*

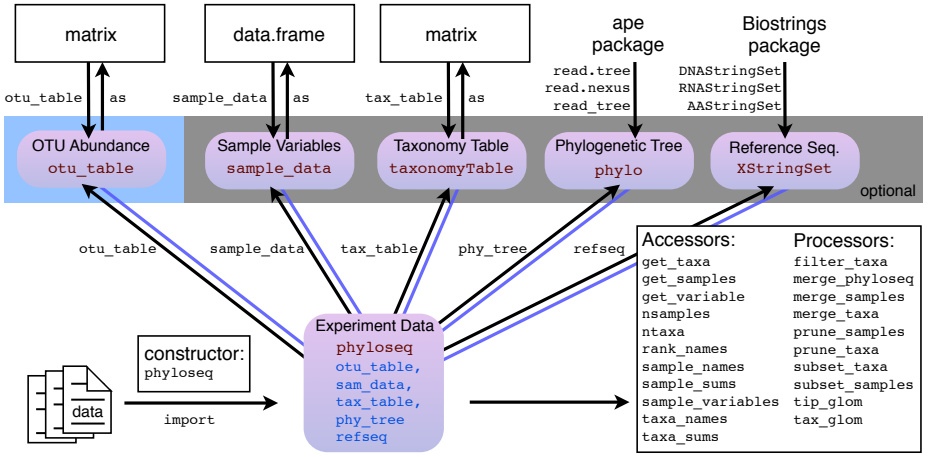
Department of Statistics, Stanford University, Stanford, California, United States of America

Key Packages:

vegan
ape
distory
phangorn
picante
metagenomeSeq
ggtree

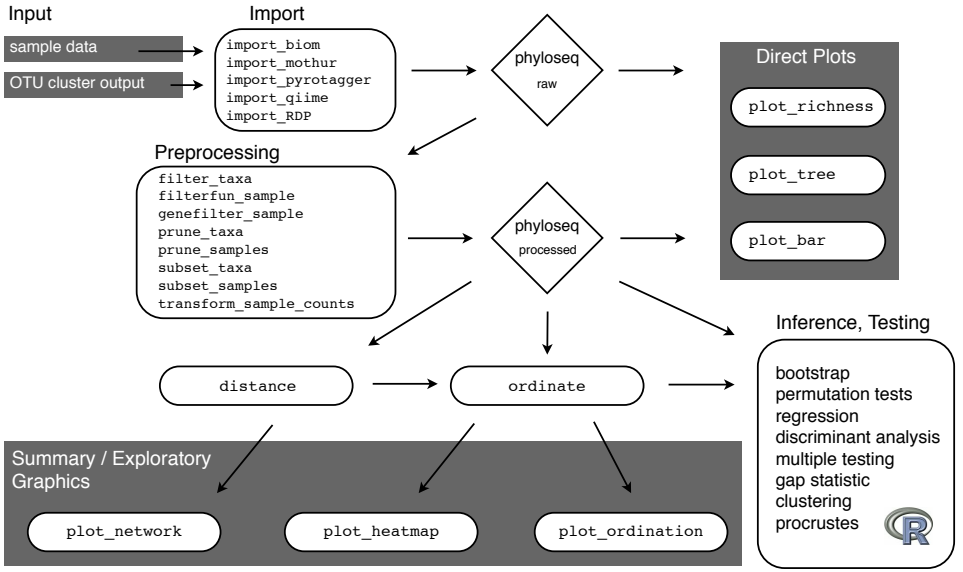
phyloseq

data structure & API



phyloseq

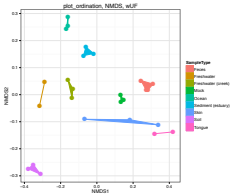
work flow



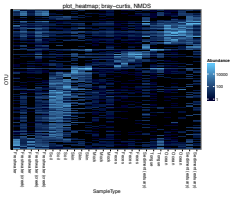
phyloseq

graphics

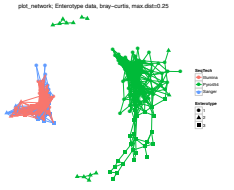
plot_ordination()



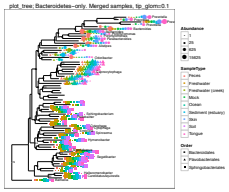
plot_heatmap()



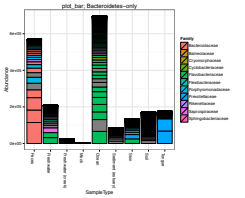
plot_network()



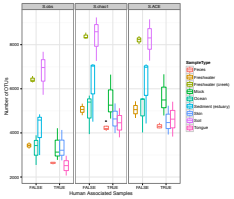
plot_tree()



plot_bar()



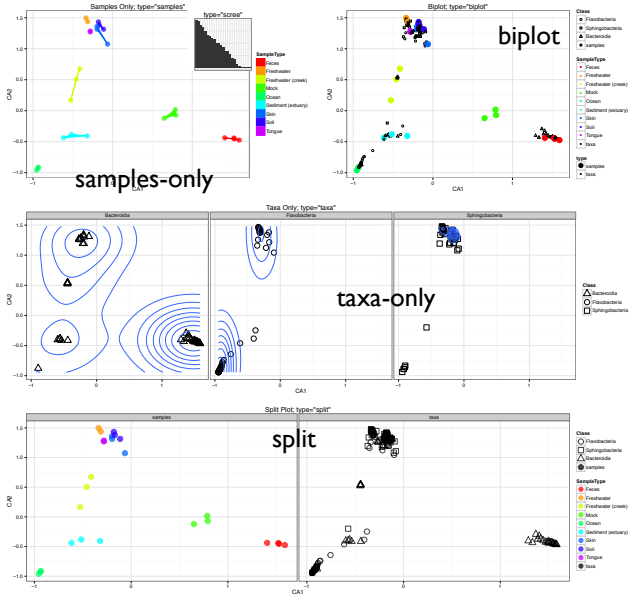
plot_richness()

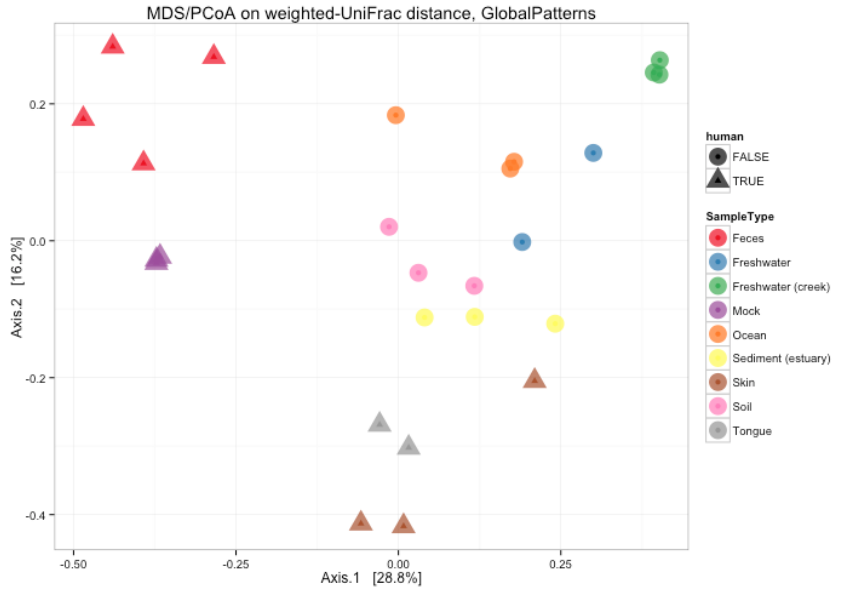


phyloseq

graphics

plot_ordination()

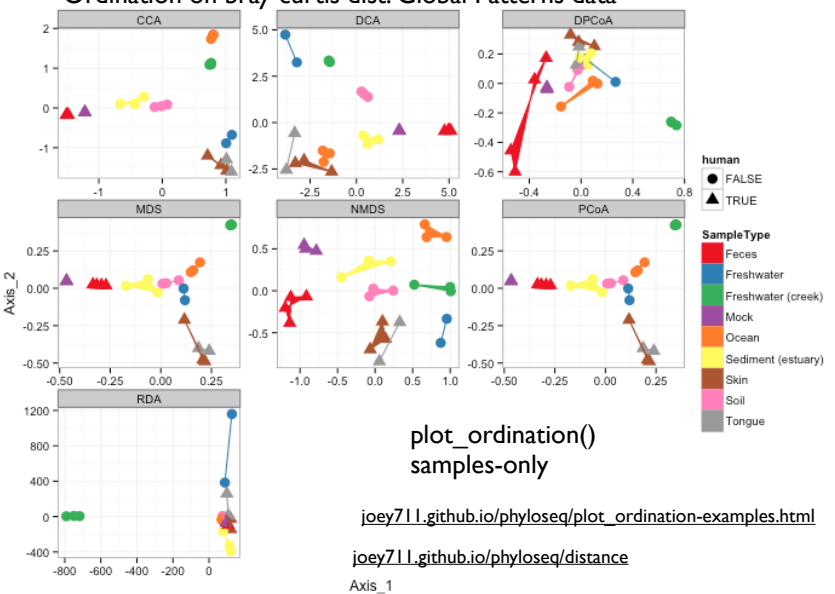





```
ordu = ordinate(GP1, "PCoA", "unifrac", weighted = TRUE)
plot_ordination(GP1, ordu, color = "SampleType", shape = "human")
```


phyloseq

supported ordination methods



plot_ordination()
samples-only

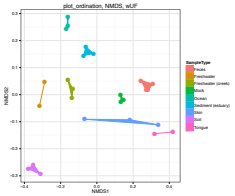
joey711.github.io/phyloseq/plot_ordination-examples.html

joey711.github.io/phyloseq/distance

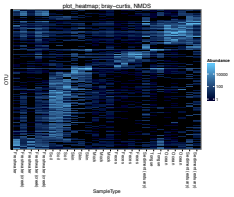
phyloseq

graphics

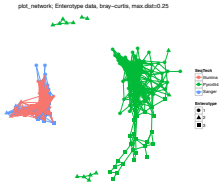
plot_ordination()



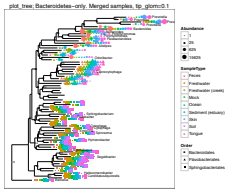
plot_heatmap()



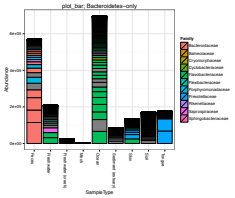
plot_network()



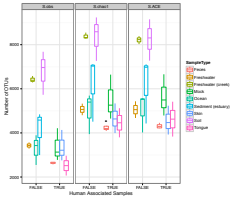
plot_tree()

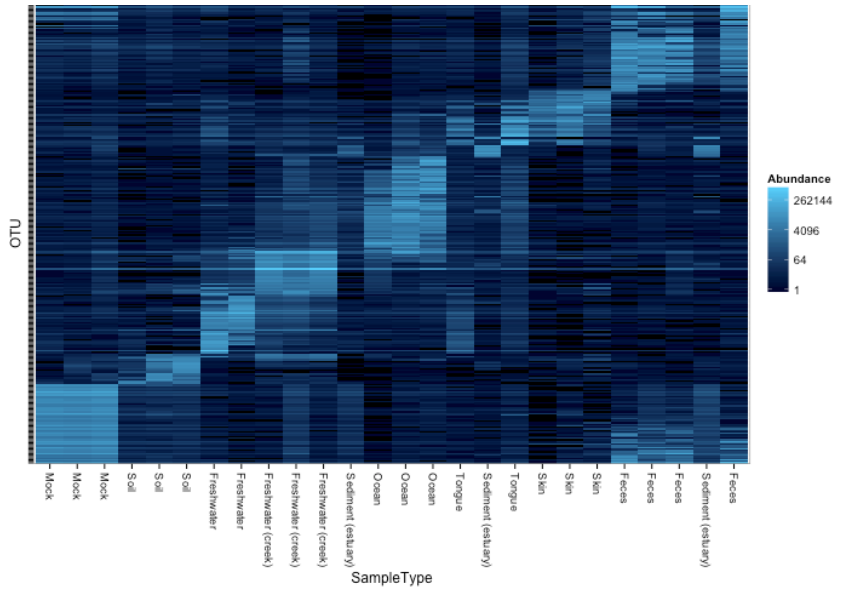


plot_bar()



plot_richness()

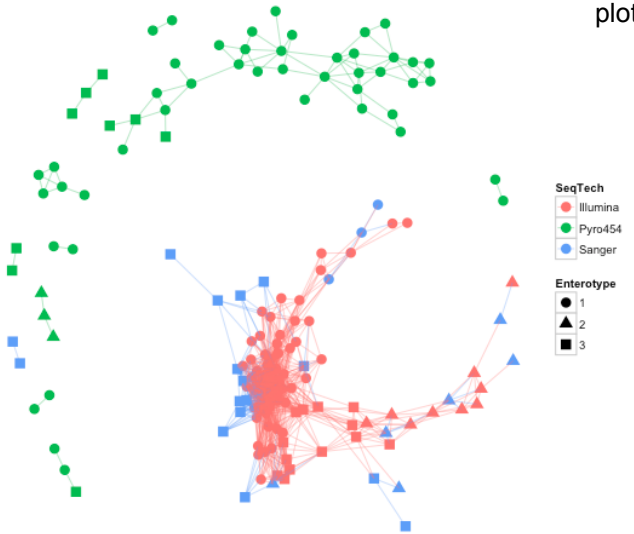





```
gpt <- subset_taxa(GlobalPatterns, Kingdom == "Bacteria")
gpt <- prune_taxa(names(sort(taxa_sums(gpt), TRUE)[1:300]), gpt)
plot_heatmap(gpt, sample.label = "SampleType")
```

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

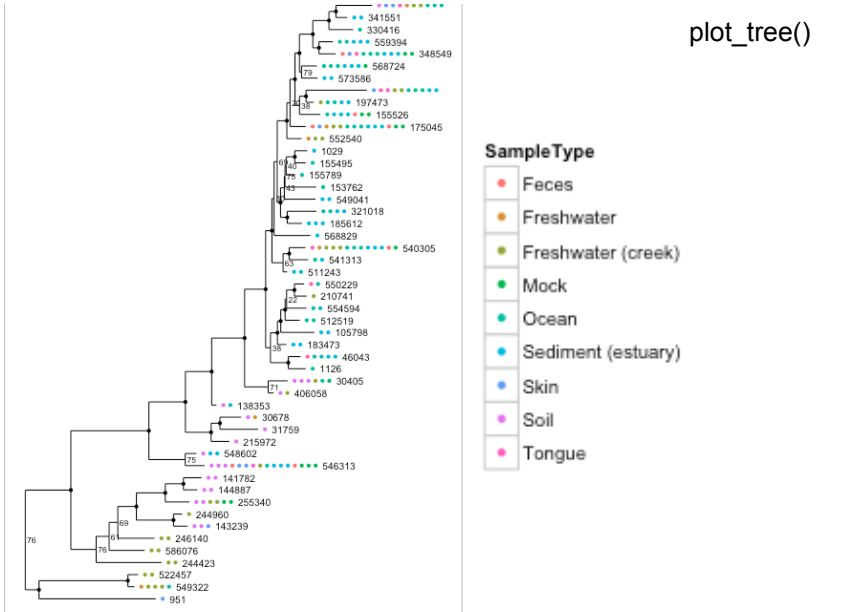
plot_network()




```
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)
```

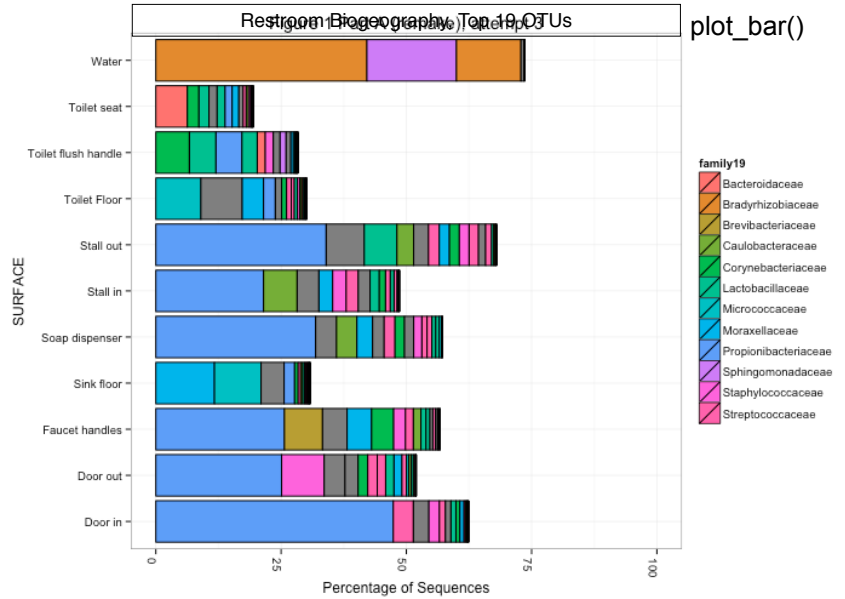
joey711.github.io/phyloseq/plot_network-examples.html

plot_tree()



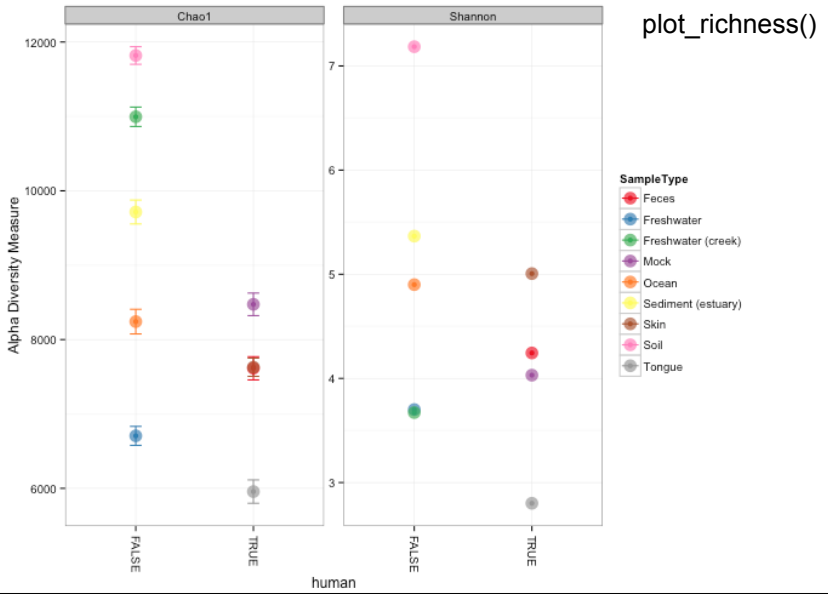

```
plot_tree(physeq, nodelabf=nodeplotboot(80, 0, 3), color="SampleType",
          label.tips="taxa_names", ladderize="left")
```

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




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

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




```
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

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	11:30	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