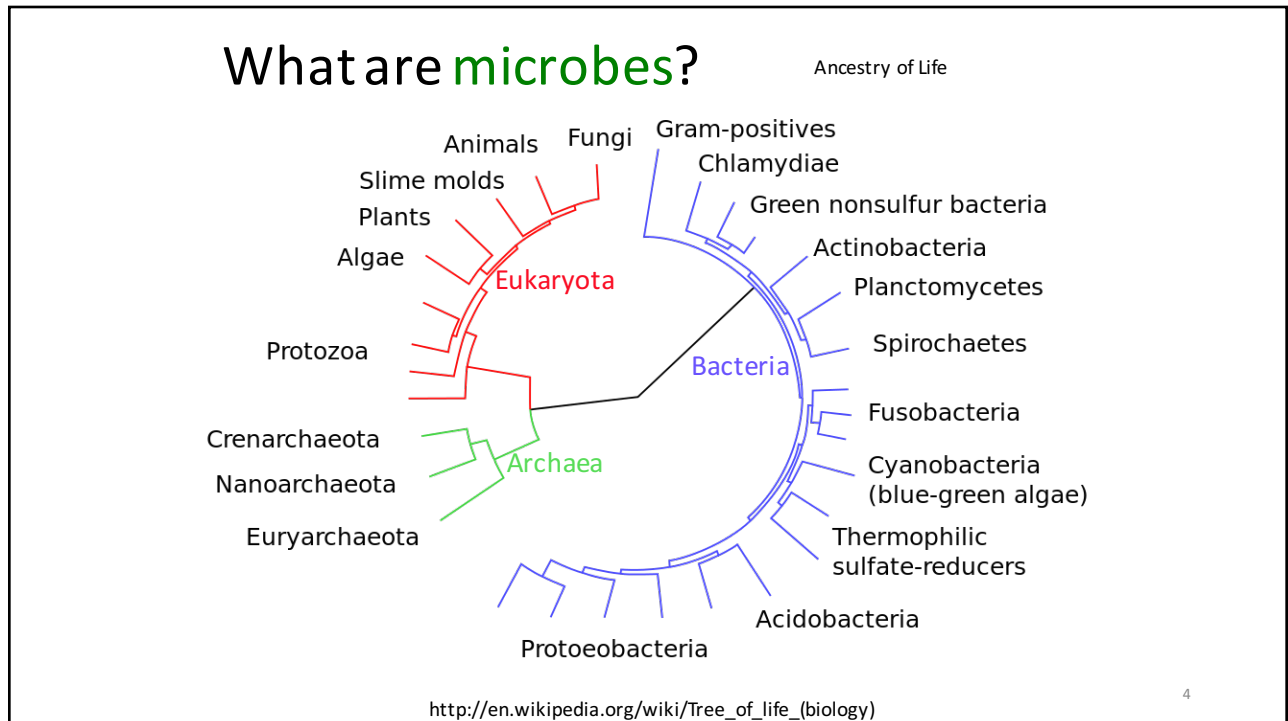
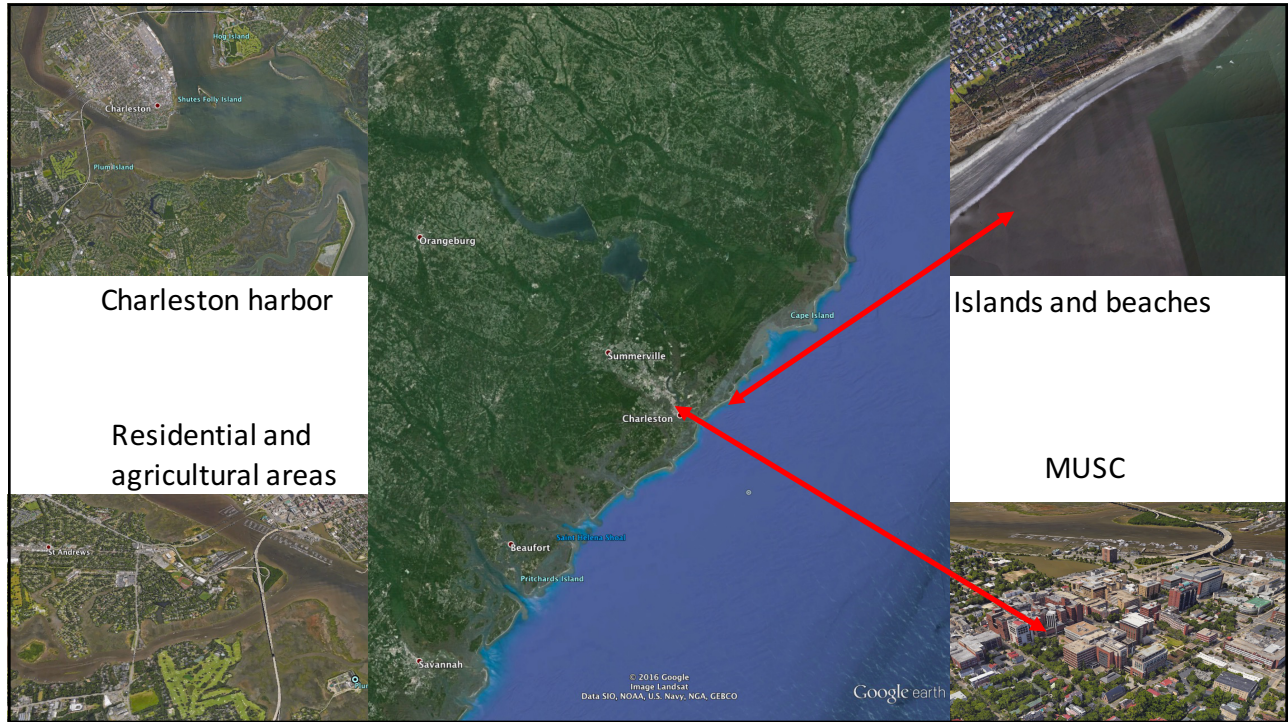
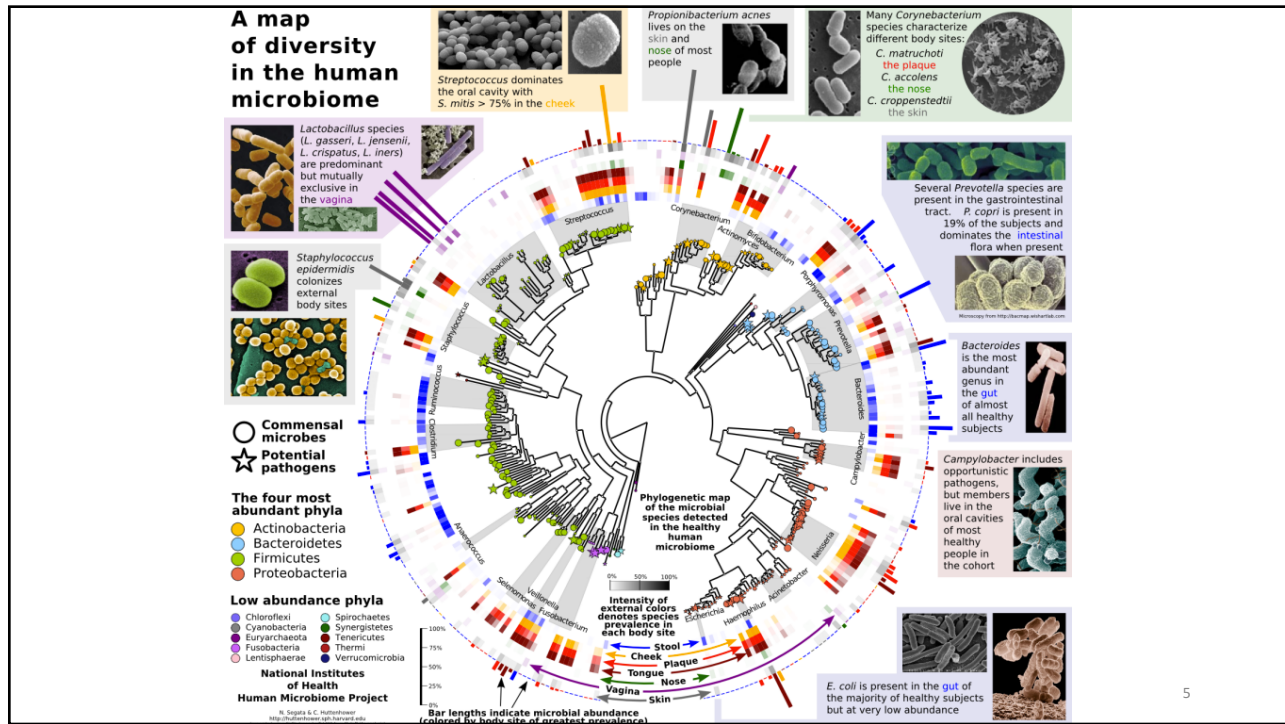


Introduction: Metagenomics in Biology and Medicine

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)
- Joshua Lederberg (1925 – 2008): “the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space” (Lederberg and McCray Scientist. 2001;15:8).

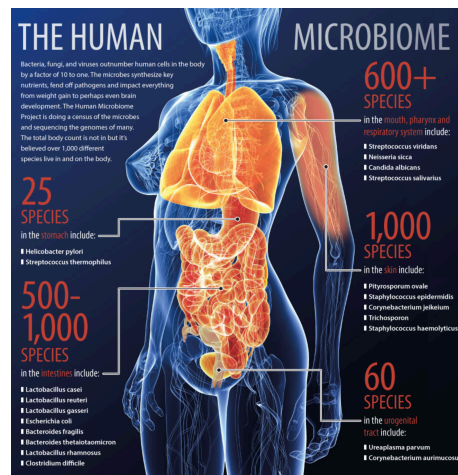




5

We are more microbes than we are humans?

- Human shelter 10 trillion microbes (10^{13} in their gut alone, (we are made of 10 trillion cells).
- **Only 1 in 10 cells in your body carries 'your' DNA. Recent evidence suggests as many bacterial cells as human.**
- It is estimated that there are 1000 species of bacteria living in the human gut.
- Compare also the number of human genes (~25,000) to the number of genes and variants that bacterial communities may carry (~4,000,000, see e.g. doi:10.1038/ncomms3151).

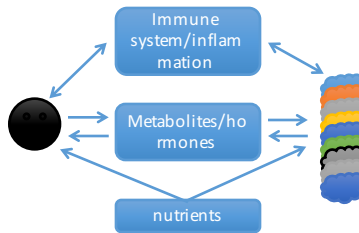


6

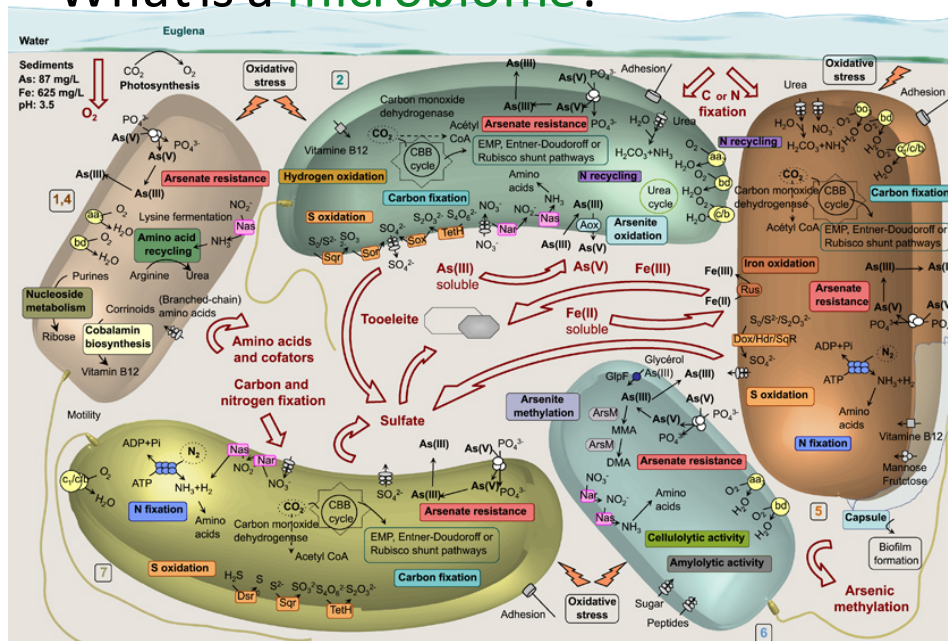
Mechanisms for host-microbe interactions

Interact:

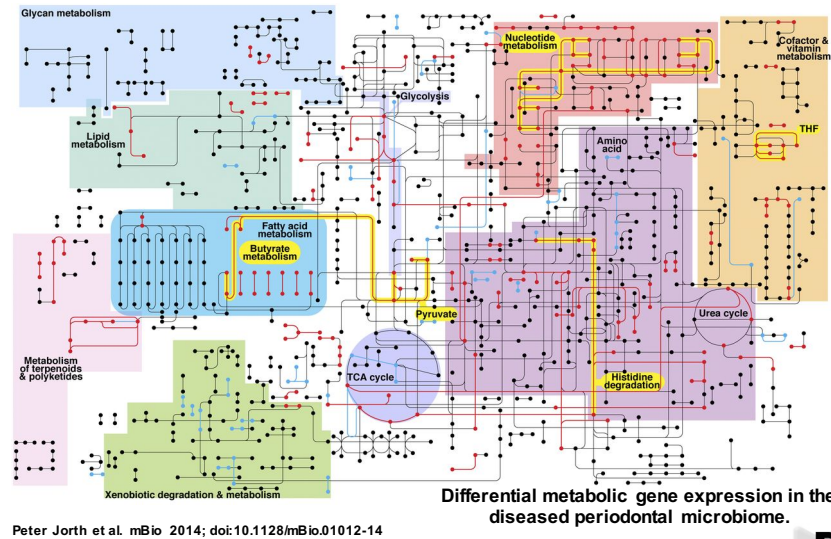
- With each other
 - Via regular ecological mechanisms (competition)
- With the host
 - Produce and metabolize hormones and common nutrients
 - Host immune system
- With the environment



What is a microbiome?



Mechanisms for host-microbe interactions



Journals.ASM.org | This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported license.

mBio[®]

9

Understanding the role of the microbiome in human disease, through Koch's postulates.

- Microbiomics effectively generalizes over the Koch's postulates:
 1. The microorganism must be found in abundance in all organisms suffering from the disease, *but should not be found in healthy organisms*.
 2. The microorganism must be isolated from a diseased organism and grown in pure culture*.
 3. The cultured microorganism should cause disease when introduced into a healthy organism.
 4. The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.
- Substitute *microbial community* for *microorganism*
- *How do we culture microbiomes if "it is estimated that as much as 20% to 60% of the human-associated microbiome, depending on body site, is uncultivable" (Genome Res. 2009 Dec; 19(12): 2317–2323)?

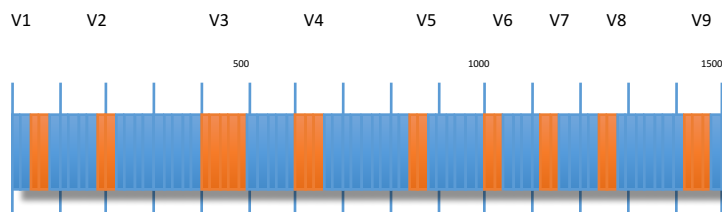
10

Discovery of *Culture Independent* Techniques

- 1980 – rRNA as evolutionary explained - C.Woese *Science*
- 1985 – Polymerase Chain Reaction (PCR) - K. Mullis *Science*
- 1985 – “Universal” Primers for rRNA sequencing - N. Pace *PNAS*
- 1989 – PCR amplification of 16S rRNA gene - Böttger *FEMS Microbiol.*
- 1996 – Large, curated rRNA database (RDP) - Maidak *Nuc.Acids Res*
- 2001 – term “microbiome” coined by Joshua Lederberg

11

Microbial community identification using targeted sequencing regions of 16S rRNA gene



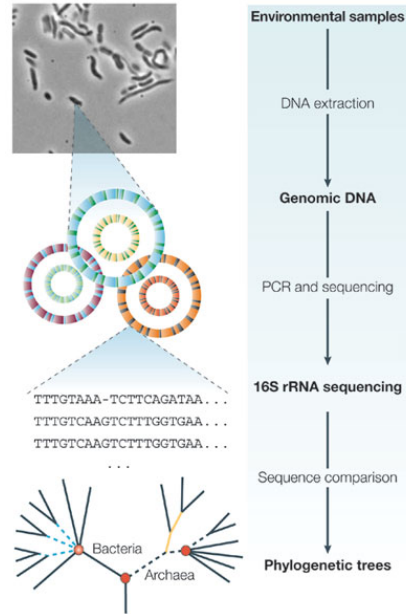
- Location of the hyper-variable regions of the 16S rRNA.
- Current technology does not allow for high-throughput sequencing of the entire 16S gene, only fragments can be sequenced.

12

How do we query a new microbiome?

- 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



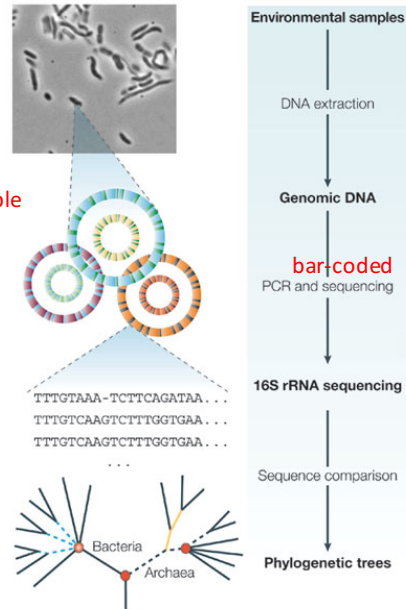
Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, 6(11), 805–814.

13

How do we query from many microbiomes??

- 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



14

Functional identification

- Sequencing based
 - Whole metagenome sequencing
 - Whole meta-transcriptome sequencing
 - Custom: e.g. IgA-Seq
- Mass spectrometry based
 - Metaproteomics
 - Metabolomics:
 - Small molecule
 - Glycomics
 - Lipidomics
- Microbial gene expression
 - NanoString codesets

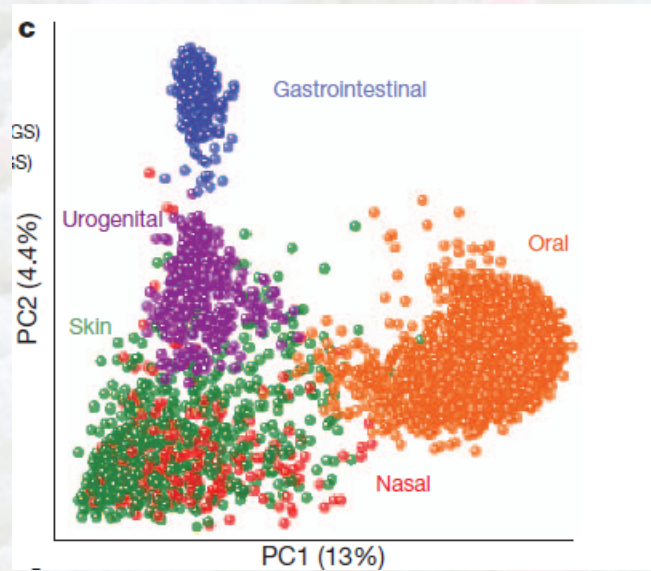
15

Some highlights of human microbiome research

- Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection, *Genome Medicine*, **8**:49 (2016);
- Gut microbiome–derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease, *Nature Immunology* **17**, 505–513 (2016);
- Cigarette smoking and the oral microbiome in a large study of American adults, *ISME J.* 2016;
- The maternal interleukin-17a pathway in mice promotes autism-like phenotypes in offspring, *Science* 2016, 351(6276):933-9;
- The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes, *Cell Host Microbe* 2015.

16

Lecture 01: Overview of Metagenomics



1

Culture Independent Techniques:

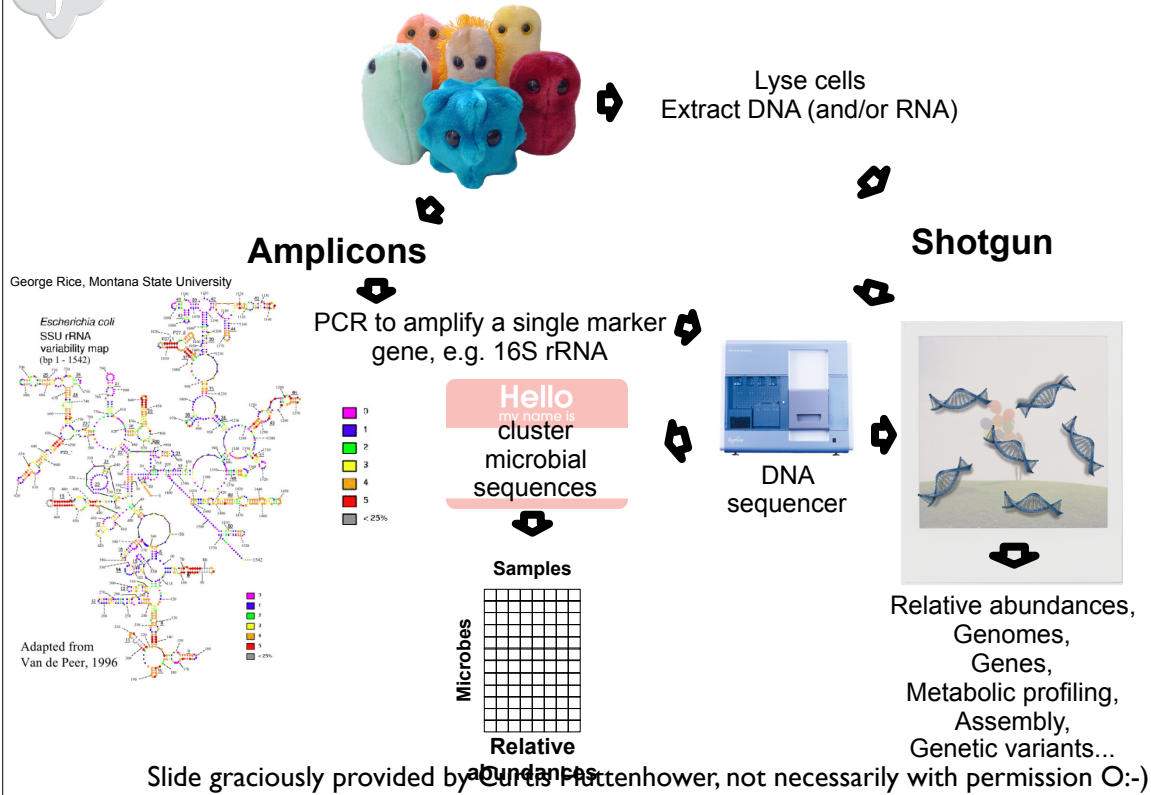
- | | Metagenomics | Number of Species Counted |
|-------------------------------------|--------------|---------------------------|
| ● Universal Gene census | ← | |
| ● Shotgun Metagenome Sequencing | ← | |
| ● Transcriptomics (shotgun mRNA) | ← | |
| ● Proteomics (protein fragments) | | |
| ● Metabolomics (excreted chemicals) | | |

\$

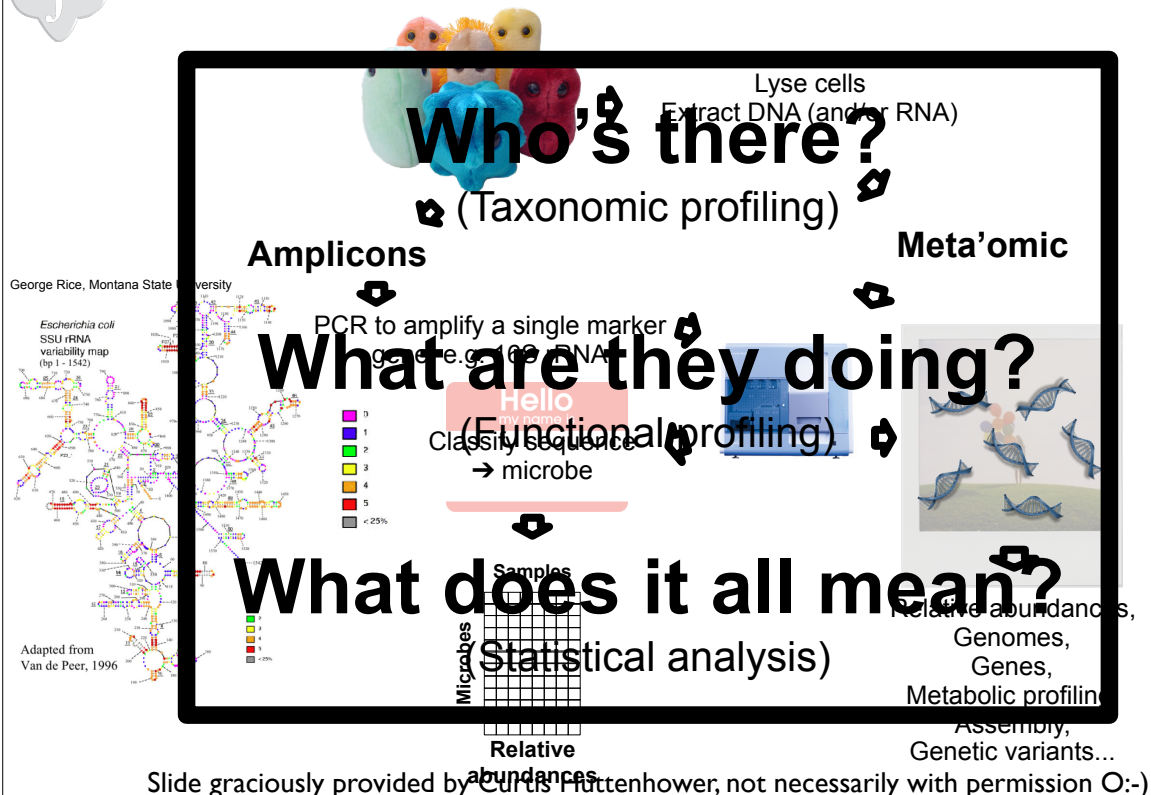
2



Nucleic acid sequencing as a tool for microbial community analysis



Sequencing as a tool for microbial community analysis





A Summary of Meta'omics

Piles of short DNA/RNA reads from >1 organism

You can...

- Ecologically profile them
- Taxonomically or phylogenetically profile them
- Functionally profile them – gene/pathway catalogs
- Assemble them

Prior knowledge is helpful

Caution: Correlation ≠ Causation

Most 'omics results require lab confirmation

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



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

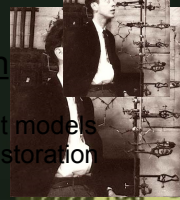
Translation

Phenotype association for diagnostics

Human disease risk: lifetime, activity, outcome
 Longitudinal analysis and study design
 Dense longitudinal measures, multiple nested outcomes

Systems analysis for intervention

More and simpler model systems
 Systematic understanding of current models
 Ecological models for ecosystem restoration



Host epidemiology

ecology ethics ecology

Privacy and

Disease risk/pathogen exposure

Tracking

Health policy

Early life, exposures

Pharma, best practices

Basic biology and molecular mechanism

Host microbiology

bio/immunology

Microbial experiments

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

Host-microbe-microbiome interactions

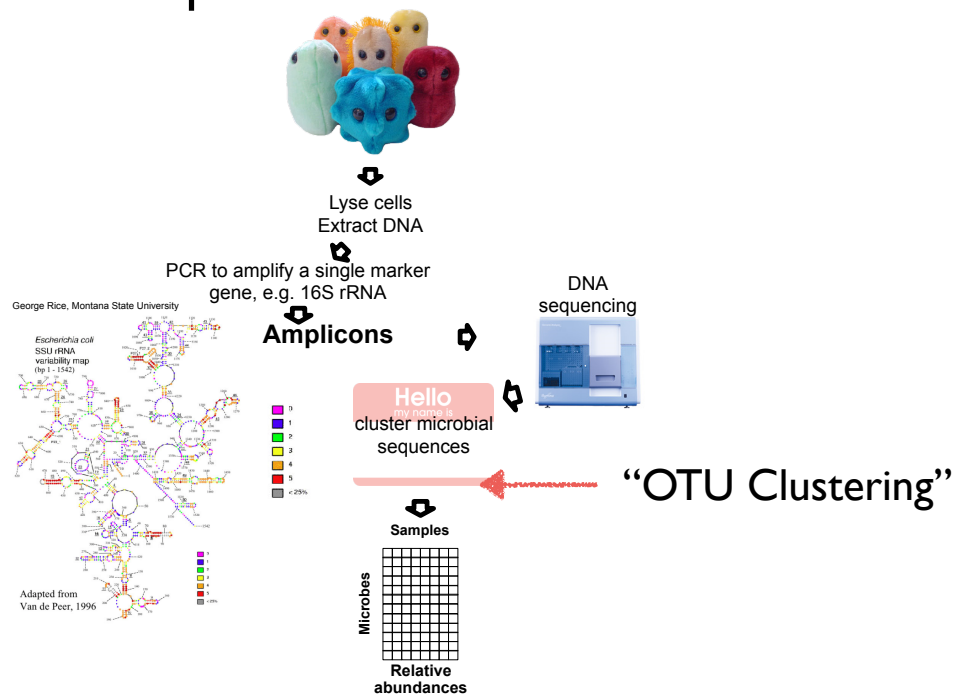
- Immunity in specific host tissues
- Non-immune mechanisms (metabolites, peptides)
- Model system perturbations, "knock ins" and "knock outs"

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

- **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

7

OTUs - Operational Taxonomic Unit



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

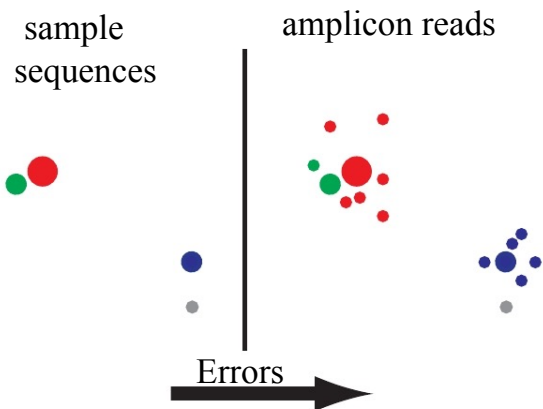
8

Motivation: Lingering problem with “OTUs”

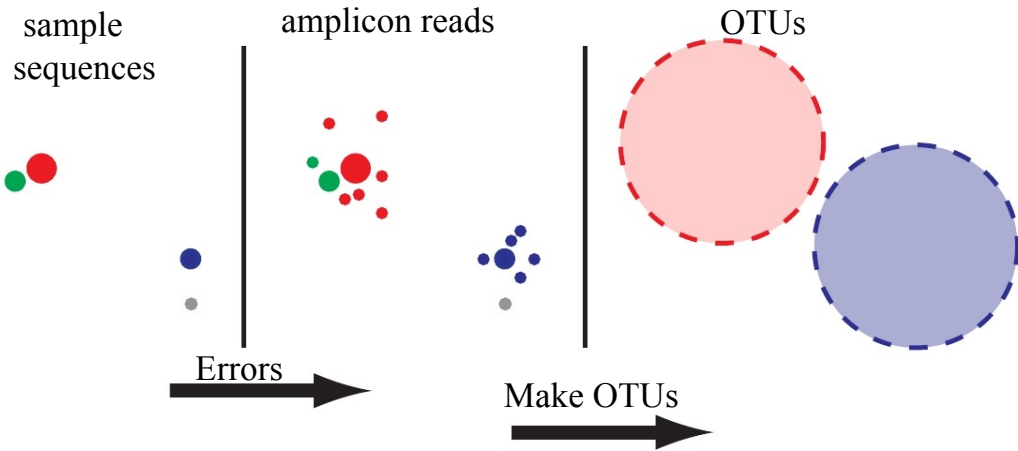
Some lingering major problems with OTU approaches:

- False Positives - e.g. 1000s of OTUs when only 10s of strains present
- Low Resolution - defined by arbitrary similarity radius
- Scaling to large datasets, comparisons
 - scales $\sim N^2$ unique sequences in dataset (all libraries)
- Unstable - OTU seq and count depends on input
 - must re-run clustering if any data added/removed, or
 - if you want to compare against an external dataset

9



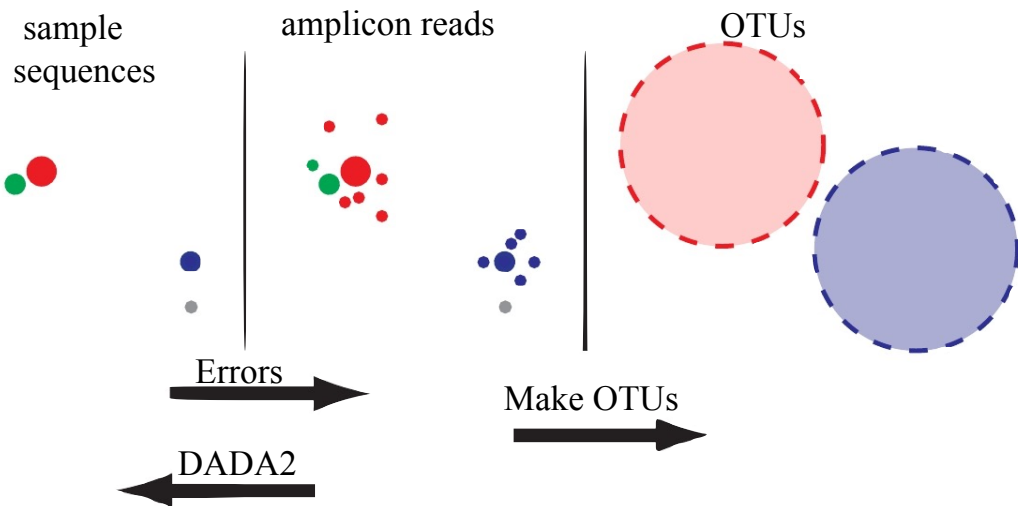
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)



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

11

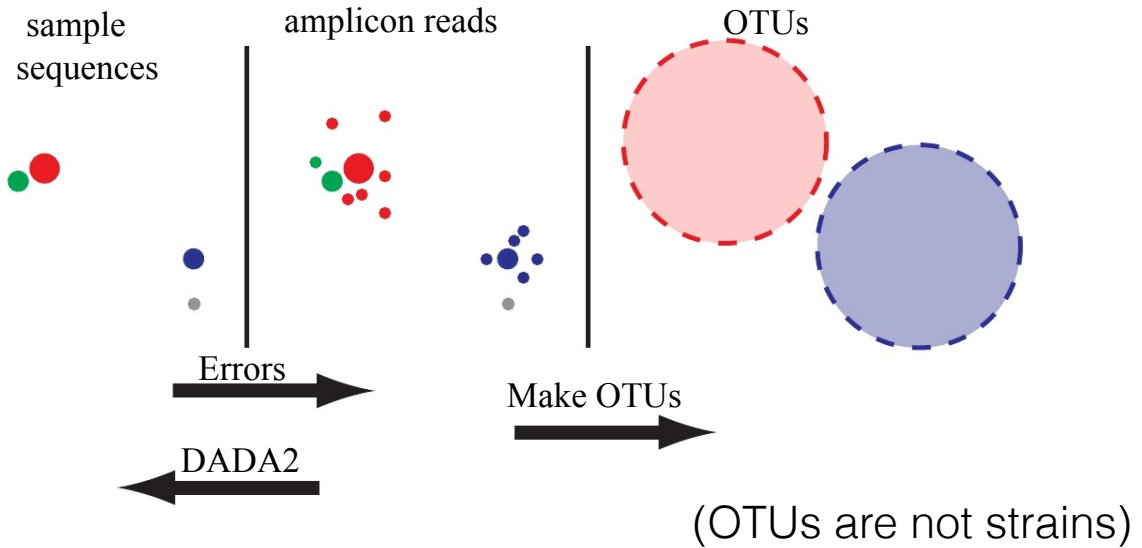
Sample Inference from Noisy Reads



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

12

Sample Inference from Noisy Reads



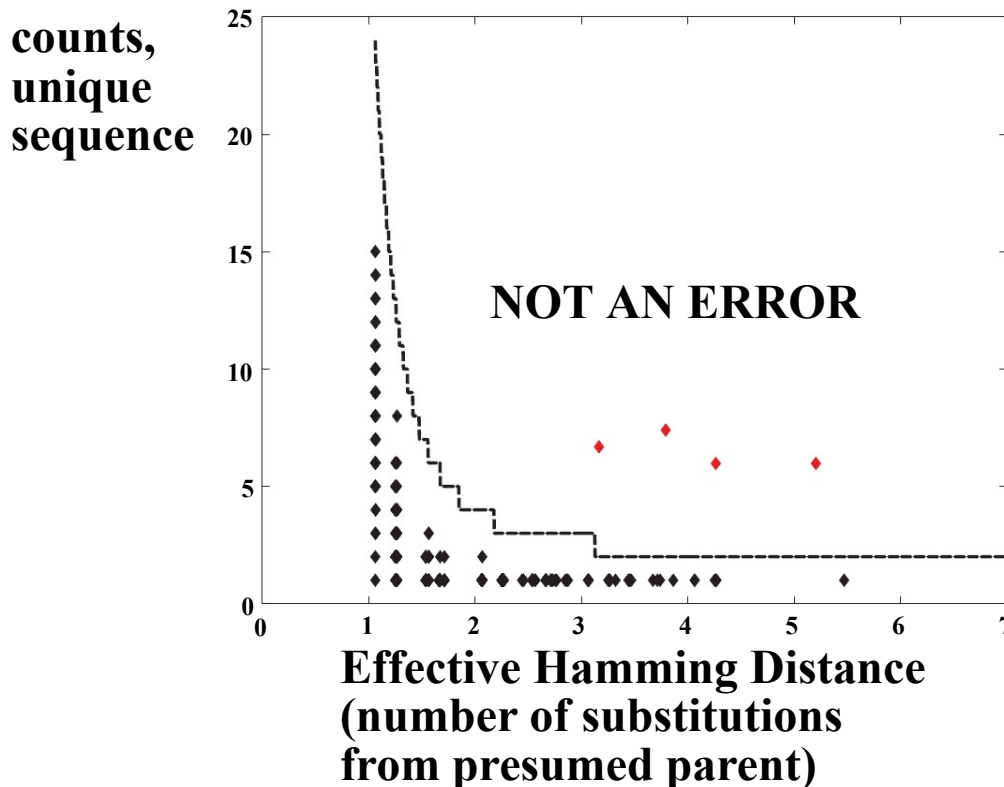
OTUs: Lump similar sequences together
DADA2: Statistically infer the sample sequences (strains)

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

13

The true shape of an error cloud

DADA2: Error Model



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

14

DADA2 algorithm assumptions

DADA2 Error Model

15

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 transition probabilities:

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

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

16

DADA2 algorithm assumptions

DADA2 Abundance Model

17

DADA2 algorithm assumptions

DADA2 Abundance Model

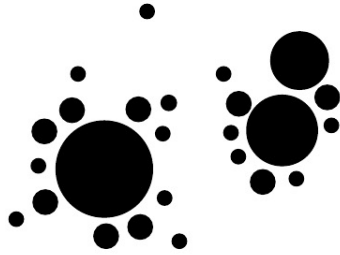
- Errors are independent across reads
- Abundance of reads w/ sequence i produced from more-abundant sequence j is poisson distributed
- Expectation of abundance equals error rate, $\lambda_{j \rightarrow i}$, multiplied by the expected reads of sample sequence j
- i has count greater than or equal to 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

18

DADA2 algorithm cartoon

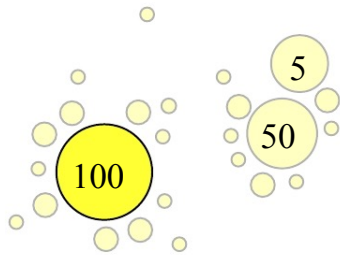


Initial guess: one real sequence + errors

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

19

DADA2 algorithm cartoon



Infer initial *error model* under this assumption.

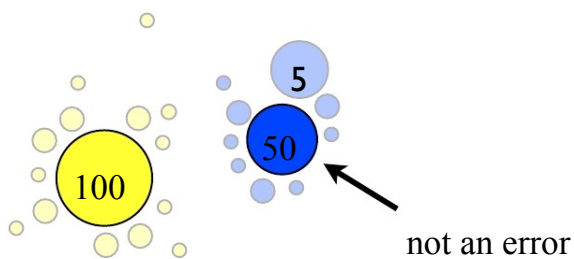
$$\Pr(i \rightarrow j) =$$

	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

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

20

DADA2 algorithm cartoon



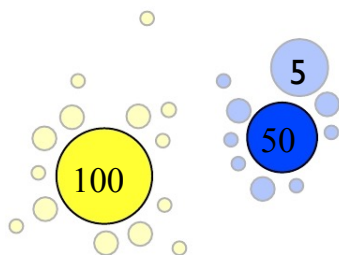
Reject unlikely error under model. **Recruit** errors.

	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

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

21

DADA2 algorithm cartoon



Update the model.

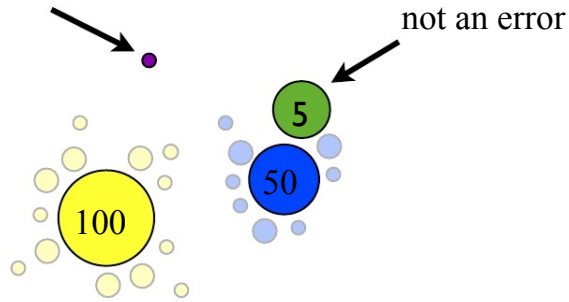
	A	C	G	T
A	0.997	10^{-3}	10^{-3}	10^{-3}
C	10^{-3}	0.997	10^{-3}	10^{-3}
G	10^{-3}	10^{-3}	0.997	10^{-3}
T	10^{-3}	10^{-3}	10^{-3}	0.997

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

22

DADA2 algorithm cartoon

not an error

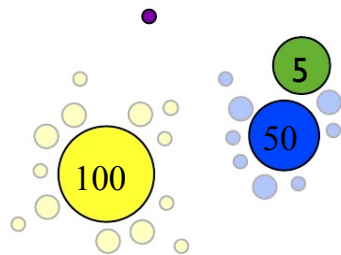


Reject more sequences under *new* model

	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

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

DADA2 algorithm cartoon

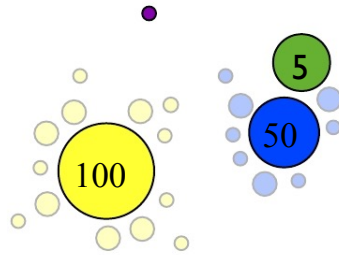


Update model again

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

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

DADA2 algorithm cartoon

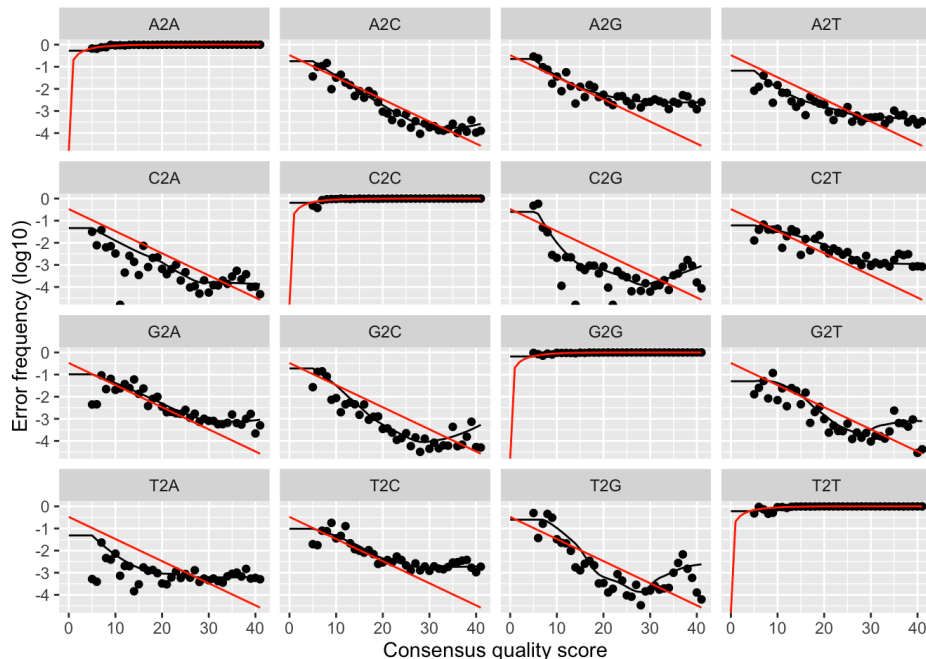


Convergence: all errors are plausible

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

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

25



- *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 overestimate 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

26

DADA2: Why is this possible?

Uses more of the information than traditional OTU clustering

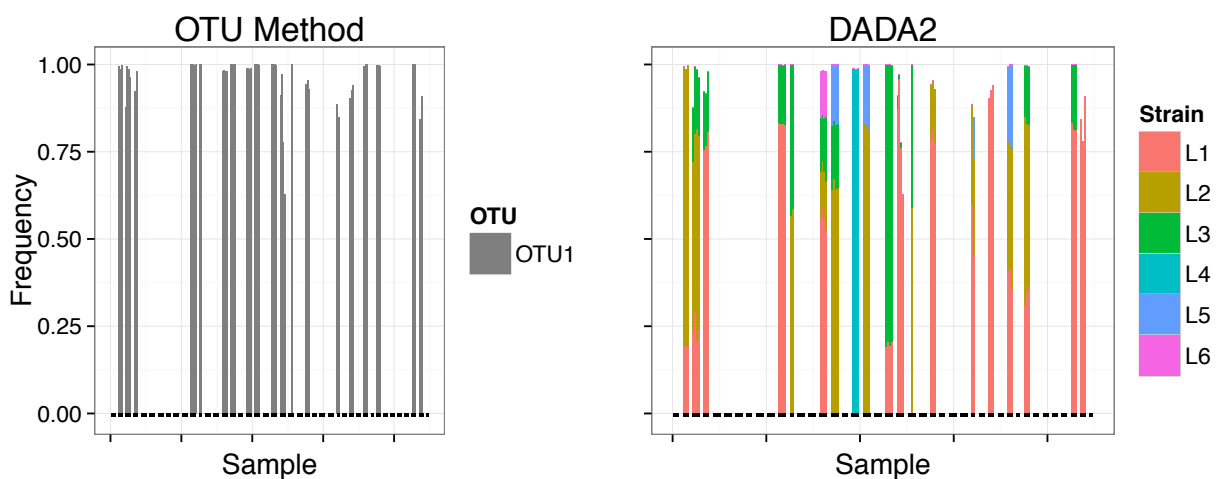
	DADA2	OTUs
Abundance	✓	Ranks only
Sequence Differences	✓	Count only
Quality	✓	No
Error Model	✓	No

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

27

DADA2 Advantages: Resolution

Lactobacillus crispatus sampled from vaginal microbiome 42 pregnant women

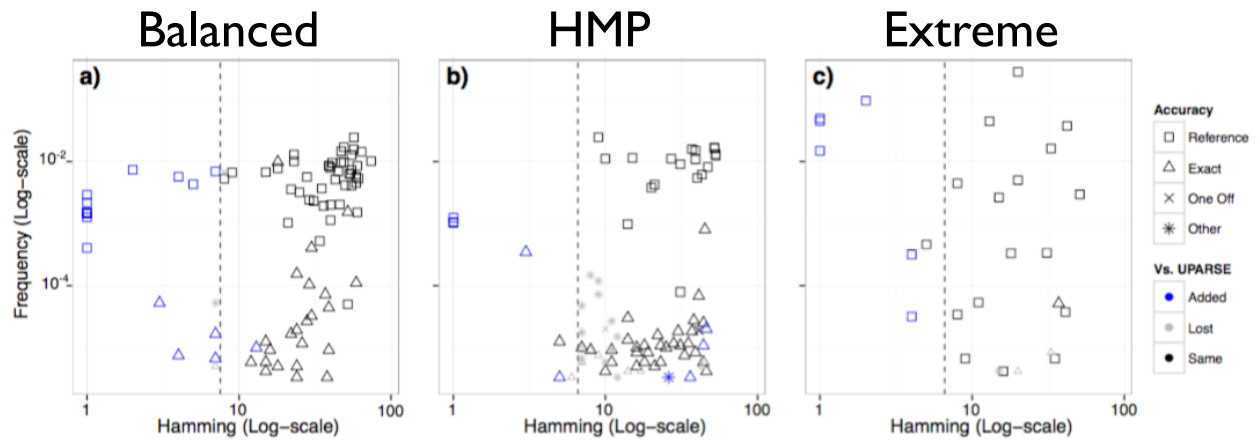


Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

28

DADA2 Advantages: Accuracy benchmarks

Mock community data for accuracy benchmarking



DADA2 performance relative to UPARSE
(best available alternative)

31

DADA2 Advantages

Analytical

Single nucleotide resolution

- genotypes/strains instead of 97% OTUs

Lower false positive rate

- Better error model, easier to ID chimeras

Computational

Linear scaling of computational costs

- Exact sequences are inherently comparable,
so samples can be processed independently.

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

32

Open-Source Sequence Clustering Methods Improve the State Of the Art

Evgenia Kopylova,^a Jose A. Navas-Molina,^{a,b} Céline Mercier,^c Zhenjiang Zech Xu,^a Frédéric Mahé,^d Yan He,^e Hong-Wei Zhou,^e Torbjørn Rognes,^{f,g} J. Gregory Caporaso,^h Rob Knight^{a,b}

February 2016

33

Four new open-source amplicon-clustering methods in last two years (since UPARSE):

- Swarm - very fast single-linkage clustering unsupervised
- SUMACLUSt - abundance-rank greedy clustering unsupervised
- OTUCLUSt - abundance-rank greedy clustering unsupervised
- SortMeRNA - clustering after reference alignment supervised

compared mainly against UPARSE (not open-source)

Kopylova, et al (2016).

Open-source sequence clustering methods improve the state of the art.

mSystems

<http://doi.org/10.1186/s12915-014-0069-1>

34

	Software	Data set																								
		Simulated						Mock						body_sites (V2)						Genuine						
		sim_even (V4)			sim_staggered (V4)			Bokulich_2 (V4)		Bokulich_3 (V4)		Bokulich_6 (V4)		OTUs		M ²		ρ		canadian_soil (V4)		global_soil (V9, 18S)				
<i>de.novo</i>	swarm	1,042	101.50	0.84	1,035	104.00	0.83	7,084	[4-50]	0.48	6,349	[4-35]	0.50	1,223	39.41	0.54	14,184	0.19	0.96	59,688	0.16	0.94	80,321	0.87	0.98	
	sumac1ust	1,031	104.06	0.83	1,022	109.92	0.83	9,575	[4-157]	0.38	13,982	[4-190]	0.41	3,317	90.80	0.52	7,103	0.18	0.99	74,284	0.14	0.87	60,781	0.50	0.96	
	uparse.q3	1,013	104.02	0.84	997	110.57	0.84							199	9.22	0.59	156	0.38	0.29	11,259	0.03	0.85				
	uparse.q16	972	100.74	0.84	806	93.28	0.78				57	[2-3]	0.79	31	3.53	0.45	108	0.36	0.26	6,275	0.06	0.75				
	uclust	1,045	105.37	0.83	1,035	110.42	0.83	20,084	[5-234]	0.40	21,929	[5-236]	0.40	4,397	105.37	0.52	11,204	0.00	1.00	91,143	0.00	1.00	82,642	0.00	1.00	
	usearch52	1,035	106.09	0.83	1,015	110.76	0.81	1,522	[3-22]	0.50	2,602	[4-28]	0.55	798	22.86	0.55	3,903	0.17	0.94	47,679	0.05	0.94	41,668	0.93	0.98	
	usearch61	1,049	104.85	0.84	1,034	110.68	0.83	22,987	[7-313]	0.39	24,704	[7-292]	0.41	4,635	123.04	0.51	14,483	0.18	0.99	102,435	0.06	0.99	102,211	0.48	0.98	
	otuc1ust.q3	996	111.03	0.84	953	106.88	0.81				438	[2-8]	0.61	228	10.36	0.61	2,753	0.18	0.85	18,373	0.08	0.82				
	otuc1ust.q20	996	111.03	0.84	953	106.88	0.81				314	[2-6]	0.65	113	7.20	0.58	2,654	0.16	0.85	18,373	0.07	0.81				
	mothur_near	957	110.09	0.82	949	110.45	0.81				1,600	[2-51]	0.44	447	23.63	0.54	806	0.45	0.12	31,546	0.06	0.76	11,440	0.53	0.76	
	mothur_fur	978	109.22	0.82	970	109.86	0.81				28,808	[5-263]	0.40	5,159	75.05	0.51	3,358	0.22	0.23	92,887	0.03	0.86	32,378	0.56	0.78	
	mothur_avg	963	109.99	0.82	959	110.98	0.82				13,255	[4-176]	0.41	2,314	55.90	0.51	2,491	0.26	0.11	83,664	0.05	0.86	20,809	0.49	0.72	
	<i>closed_ref</i>	usearch61	1,275	129.19	0.83	1,267	127.50	0.82	1,027	[5-26]	0.53	614	[4-18]	0.59	631	26.02	0.61	5,982	0.06	0.96	13,808	0.06	0.96	3,784	0.50	0.55
		uclust	1,238	127.59	0.83	1,225	126.02	0.84	1,053	[5-27]	0.53	557	[5-18]	0.57	547	25.03	0.60	5,446	0.00	1.00	13,659	0.00	1.00	305	0.00	1.00
sortmerma		1,072	122.75	0.82	1,067	121.89	0.81	396	[4-15]	0.53	290	[4-13]	0.61	382	19.47	0.57	6,174	0.06	0.99	13,281	0.06	0.98	255	0.34	0.75	
usearch52		1,001	115.38	0.80	980	113.39	0.78	571	[5-30]	0.54	331	[5-22]	0.64	315	18.24	0.59	3,355	0.08	0.97	4,121	0.04	0.79	5,763	0.48	0.19	
usearch61		1,275	129.19	0.83	1,267	127.50	0.82	1,027	[5-26]	0.53	614	[4-18]	0.59	631	26.02	0.61	5,982	0.06	0.96	13,808	0.06	0.96	3,784	0.50	0.55	
<i>open_ref</i>	uclust	1,262	106.12	0.83	1,245	111.29	0.83	10,169	[3-97]	0.40	4,170	[3-104]	0.42	4,109	93.67	0.48	12,442	0.00	1.00	87,936	0.00	1.00	37,380	0.00	1.00	
	sortmerma	1,072	104.77	0.82	1,085	111.80	0.81	9,272	[3-132]	0.39	2,649	[3-140]	0.41	2,727	88.56	0.51	10,242	0.06	0.98	79,363	0.03	0.82	35,345	0.12	0.92	
	sumac1ust	1,304	106.04	0.83	1,293	112.36	0.83	9,414	[3-108]	0.40	3,966	[3-126]	0.41	3,421	80.89	0.53	12,807	0.06	0.97	87,300	0.06	0.80	43,175	0.10	0.94	
	usearch61	1,304	106.04	0.83	1,293	112.36	0.83	9,414	[3-108]	0.40	3,966	[3-126]	0.41	3,421	80.89	0.53	12,807	0.06	0.97	87,300	0.06	0.80	43,175	0.10	0.94	

(OTU counts do not include singletons)

Kopylova, et al (2016).

Open-source sequence clustering methods improve the state of the art.

mSystems

<http://doi.org/10.1186/s12915-014-0069-1>

35

	Software	Data set																								
		Simulated						Mock						body_sites (V2)						Genuine						
		sim_even (V4)			sim_staggered (V4)			Bokulich_2 (V4)		Bokulich_3 (V4)		Bokulich_6 (V4)		OTUs		M ²		ρ		canadian_soil (V4)		global_soil (V9, 18S)				
<i>de.novo</i>	swarm	1,042	101.50	0.84	1,035	104.00	0.83	7,084	[4-50]	0.48	6,349	[4-35]	0.50	1,223	39.41	0.54	14,184	0.19	0.96	59,688	0.16	0.94	80,321	0.87	0.98	
	sumac1ust	1,031	104.06	0.83	1,022	109.92	0.83	9,575	[4-157]	0.38	13,982	[4-190]	0.41	3,317	90.80	0.52	7,103	0.18	0.99	74,284	0.14	0.87	60,781	0.50	0.96	
	uparse.q3	1,013	104.02	0.84	997	110.57	0.84							199	9.22	0.59	156	0.38	0.29	11,259	0.03	0.85				
	uparse.q16	972	100.74	0.84	806	93.28	0.78				57	[2-3]	0.79	31	3.53	0.45	108	0.36	0.26	6,275	0.06	0.75				
	uclust	1,045	105.37	0.83	1,035	110.42	0.83	20,084	[5-234]	0.40	21,929	[5-236]	0.40	4,397	105.37	0.52	11,204	0.00	1.00	91,143	0.00	1.00	82,642	0.00	1.00	
	usearch52	1,035	106.09	0.83	1,015	110.76	0.81	1,522	[3-22]	0.50	2,602	[4-28]	0.55	798	22.86	0.55	3,903	0.17	0.94	47,679	0.05	0.94	41,668	0.93	0.98	
	usearch61	1,049	104.85	0.84	1,034	110.68	0.83	22,987	[7-313]	0.39	24,704	[7-292]	0.41	4,635	123.04	0.51	14,483	0.18	0.99	102,435	0.06	0.99	102,211	0.48	0.98	
	otuc1ust.q3	996	111.03	0.84	953	106.88	0.81				438	[2-8]	0.61	228	10.36	0.61	2,753	0.18	0.85	18,373	0.08	0.82				
	otuc1ust.q20	996	111.03	0.84	953	106.88	0.81				314	[2-6]	0.65	113	7.20	0.58	2,654	0.16	0.85	18,373	0.07	0.81				
	mothur_near	957	110.09	0.82	949	110.45	0.81				1,600	[2-51]	0.44	447	23.63	0.54	806	0.45	0.12	31,546	0.06	0.76	11,440	0.53	0.76	
	mothur_fur	978	109.22	0.82	970	109.86	0.81				28,808	[5-263]	0.40	5,159	75.05	0.51	3,358	0.22	0.23	92,887	0.03	0.86	32,378	0.56	0.78	
	mothur_avg	963	109.99	0.82	959	110.98	0.82				13,255	[4-176]	0.41	2,314	55.90	0.51	2,491	0.26	0.11	83,664	0.05	0.86	20,809	0.49	0.72	
	<i>closed_ref</i>	usearch61	1,275	129.19	0.83	1,267	127.50	0.82	1,027	[5-26]	0.53	614	[4-18]	0.59	631	26.02	0.61	5,982	0.06	0.96	13,808	0.06	0.96	3,784	0.50	0.55
		uclust	1,238	127.59	0.83	1,225	126.02	0.84	1,053	[5-27]	0.53	557	[5-18]	0.57	547	25.03	0.60	5,446	0.00	1.00	13,659	0.00	1.00	305	0.00	1.00
sortmerma		1,072	122.75	0.82	1,067	121.89	0.81	396	[4-15]	0.53	290	[4-13]	0.61	382	19.47	0.57	6,174	0.06	0.99	13,281	0.06	0.98	255	0.34	0.75	
usearch52		1,001	115.38	0.80	980	113.39	0.78	571	[5-30]	0.54	331	[5-22]	0.64	315	18.24	0.59	3,355	0.08	0.97	4,121	0.04	0.79	5,763	0.48	0.19	
usearch61		1,275	129.19	0.83	1,267	127.50	0.82	1,027	[5-26]	0.53	614	[4-18]	0.59	631	26.02	0.61	5,982	0.06	0.96	13,808	0.06	0.96	3,784	0.50	0.55	
<i>open_ref</i>	uclust	1,262	106.12	0.83	1,245	111.29	0.83	10,169	[3-97]	0.40	4,170	[3-104]	0.42	4,109	93.67	0.48	12,442	0.00	1.00	87,936	0.00	1.00	37,380	0.00	1.00	
	sortmerma	1,072	104.77	0.82	1,085	111.80	0.81	9,272	[3-132]	0.39	2,649	[3-140]	0.41	2,727	88.56	0.51	10,242	0.06	0.98	79,363	0.03	0.82	35,345	0.12	0.92	
	sumac1ust	1,304	106.04	0.83	1,293	112.36	0.83	9,414	[3-108]	0.40	3,966	[3-126]	0.41	3,421	80.89	0.53	12,807	0.06	0.97	87,300	0.06	0.80	43,175	0.10	0.94	
	usearch61	1,304	106.04	0.83	1,293	112.36	0.83	9,414	[3-108]	0.40	3,966	[3-126]	0.41	3,421	80.89	0.53	12,807	0.06	0.97	87,300	0.06	0.80	43,175	0.10	0.94	

DADA2 performance:

- Mock:
 - Bokulich_6: 64 sequences, 25/26 taxonomies, 6 new
 - Bokulich_2: 17 sequences, 18/18 taxonomies, 11 new
- Simulations:
 - Even: 1055/1055 sequence variants, no Fps
 - Staggered: 1,042/1,055

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

36

DADA2

Divisive Amplicon Denoising Algorithm - ver.2

DADA2: High resolution sample inference from 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://dx.doi.org/10.1101/024034>

Manuscript draft on bioRxiv
(*Nature Methods*, in press)

<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.

37

Diversity

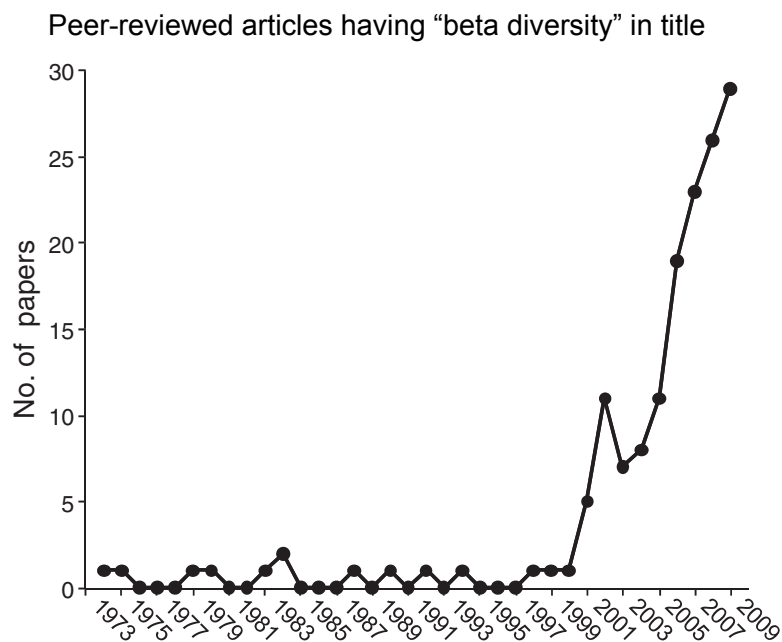
38

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

39

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.

40

Beta-Diversity

- Microbial ecologists typically use beta diversity as a broad umbrella term that can refer to any of several indices related to compositional differences
(Differences in species content between samples)
- For some reason this is contentious, and there appears to be ongoing (and pointless?) argument over the possible definitions
- For our purposes, and microbiome research, when you hear “beta-diversity”, you can probably think: “Diversity of species composition”

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

41

Distances between microbiomes

42

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:-)

43

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:-)

44

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:-)

45

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:-)

46

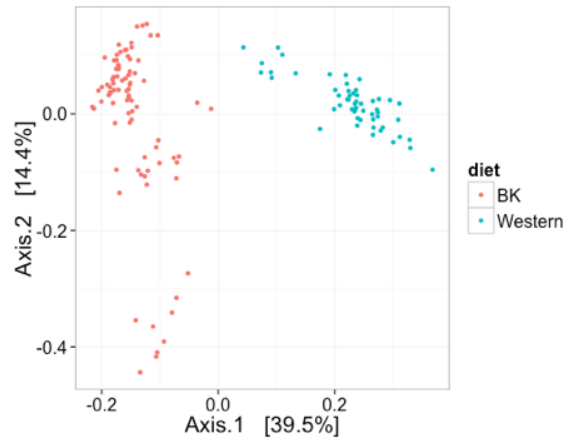
Ordination Methods

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

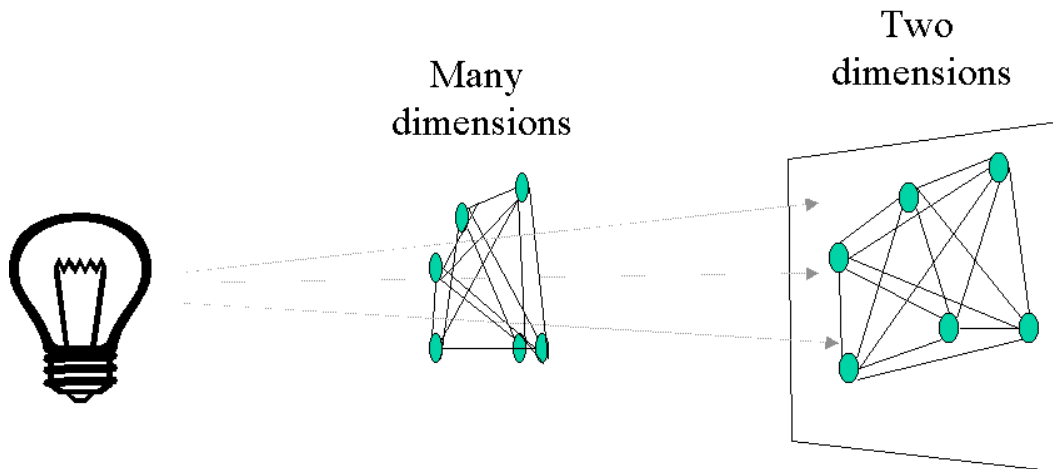
2-dimensions

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

47

Multi-dimensional Scaling

Why MDS? It works with any distance!



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

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

48

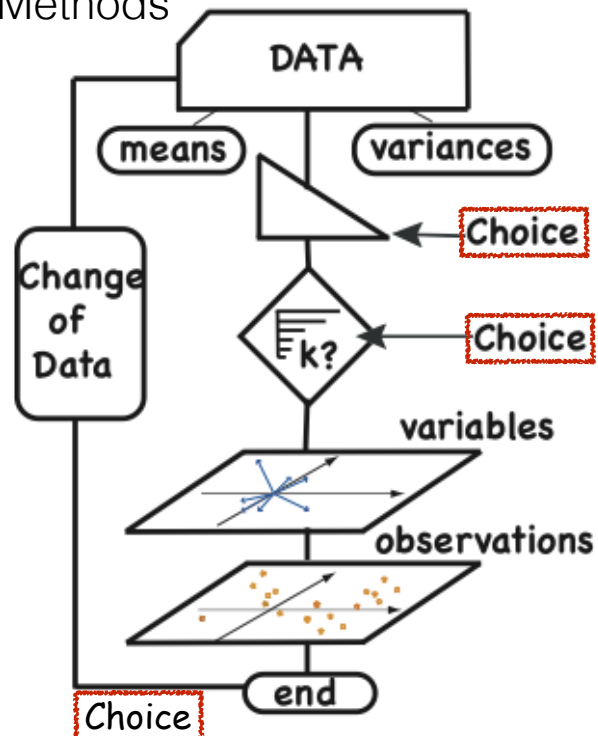
Exploratory Data Analysis

“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:-)

49

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?

50