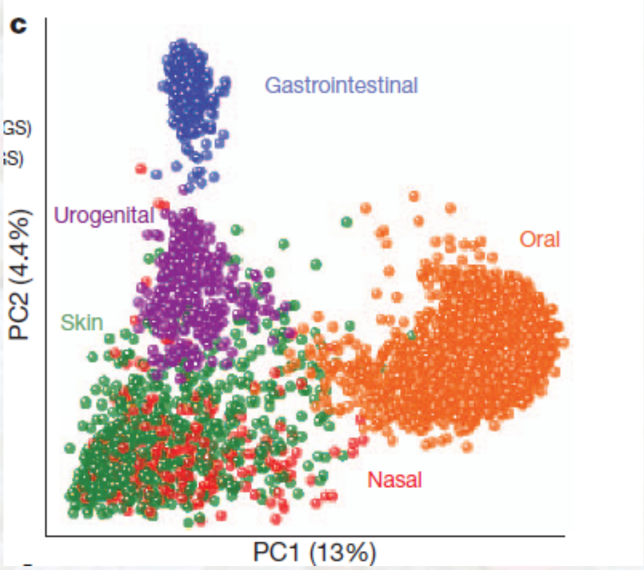
Lecture 01:
Overview of Metagenomics



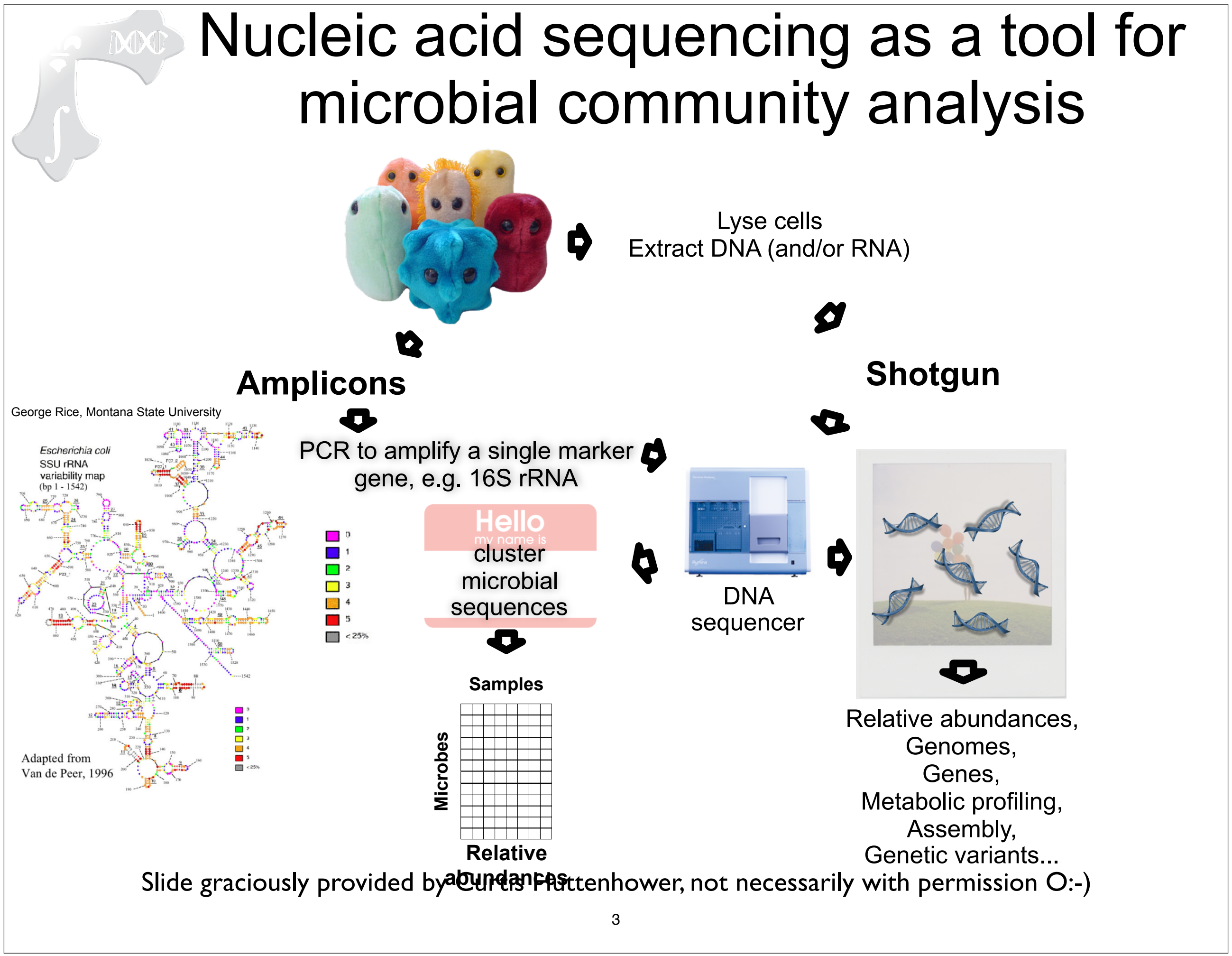
Culture Independent Techniques:

Metagenomics

Number of Species Counted

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

$



Nucleic acid sequencing as a tool for microbial community analysis

Lyse cells
Extract DNA (and/or RNA)

Amplicons

Shotgun

George Rice, Montana State University

Adapted from
Van de Peer, 1996

PCR to amplify a single marker
gene, e.g. 16S rRNA

cluster
microbial
sequences

DNA
sequencer

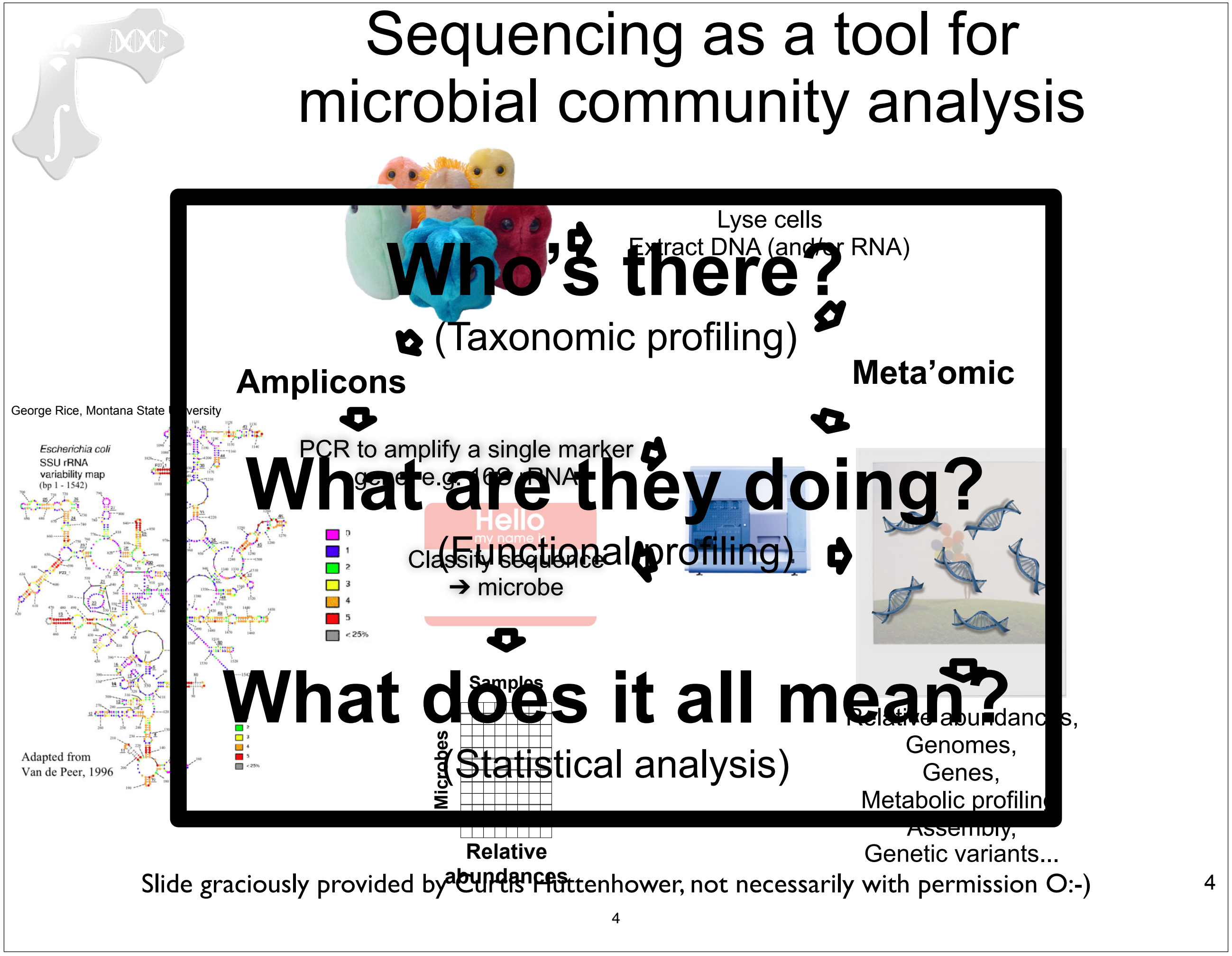
Samples

Microbes

Relative abundances

Relative abundances,
Genomes,
Genes,
Metabolic profiling,
Assembly,
Genetic variants...

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



Sequencing as a tool for microbial community analysis

Lyse cells
Extract DNA (and/or RNA)

Who's there?

(Taxonomic profiling)

Amplicons

Meta'omic

George Rice, Montana State University

Adapted from
Van de Peer, 1996

PCR to amplify a single marker
gene, e.g. 16S rRNA

What are they doing?

(Functional profiling)

Classify sequence
➔ microbe

What does it all mean?

(Statistical analysis)

Samples

Microbes

Relative abundances

Relative abundances,
Genomes,
Genes,
Metabolic profiling,
Assembly,
Genetic variants...

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

4

A Summary of Meta'omics

Piles of short DNA/RNA reads from >1 organism

You can...

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

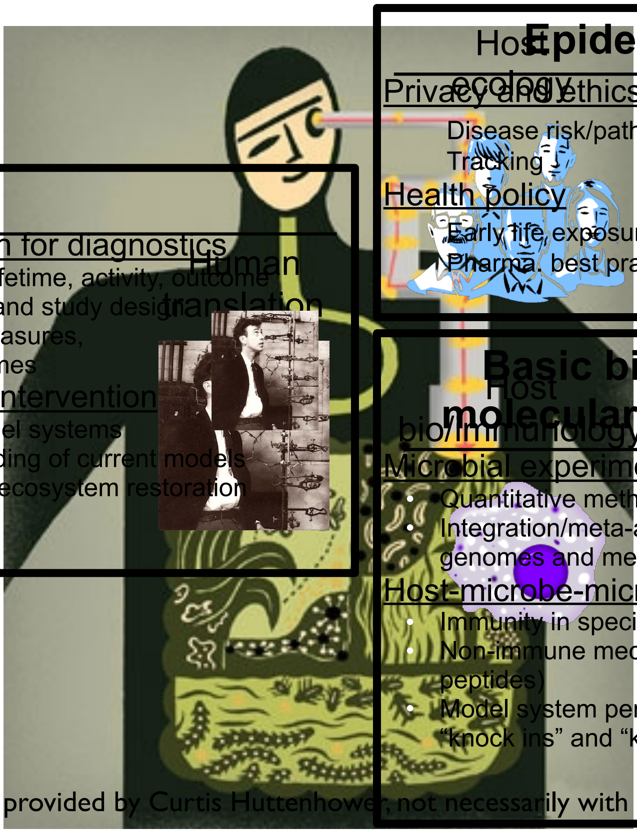
Prior knowledge is helpful

Caution: Correlation ≠ Causation

- Most 'omics results require lab confirmation

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

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



Epidemiology

Host ecology

Microbial ecology

Privacy and ethics

- Disease risk/pathogen exposure
- Tracking

Health policy

- Early life exposures
- Pharma. best practices

Translation

Human translation

Phenotype association for diagnostics

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

Systems analysis for intervention

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

Basic biology and molecular mechanism

Host bio/immunology

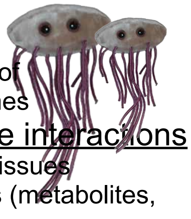
Microbiology

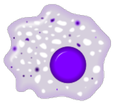
Microbial experiments

- 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"

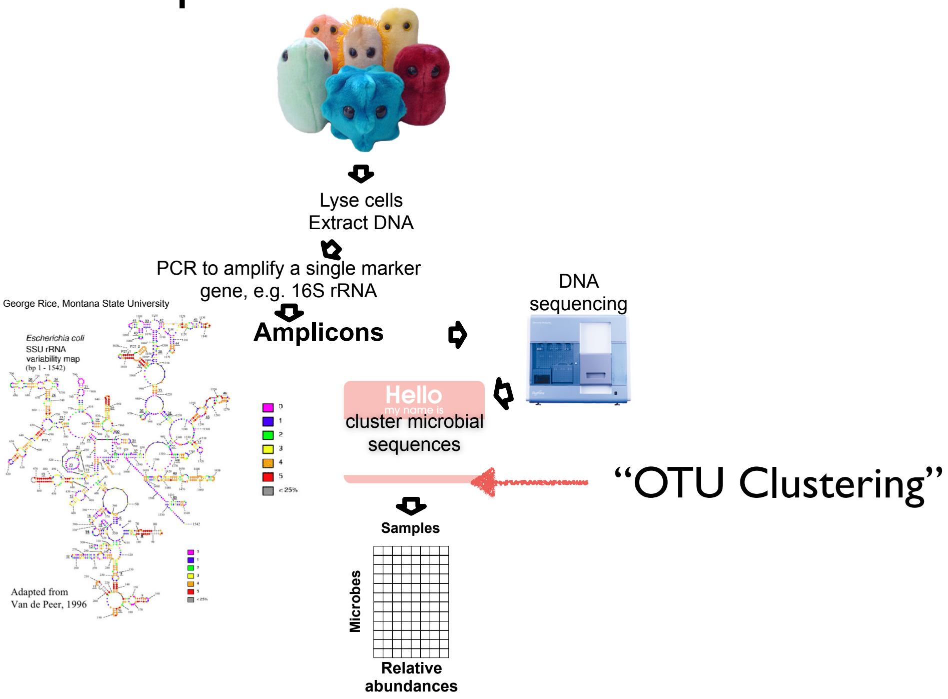




Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)

- **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

OTUs - Operational Taxonomic Unit

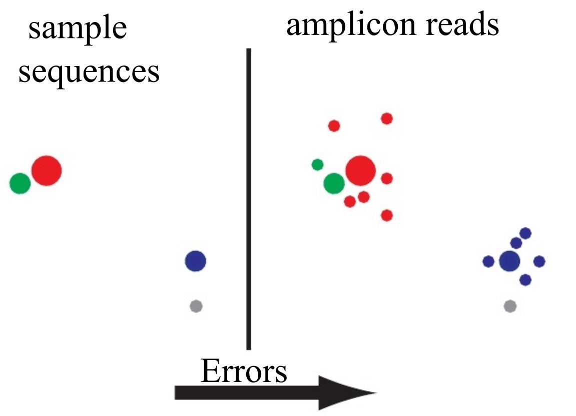

Lyse cells
Extract DNA
PCR to amplify a single marker gene, e.g. 16S rRNA
George Rice, Montana State University
Escherichia coli SSU rRNA variability map (bp 1 - 1542)
Adapted from Van de Peer, 1996
Amplicons
DNA sequencing
Hello
my name is
cluster microbial sequences
"OTU Clustering"
Samples
Microbes
Relative abundances

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

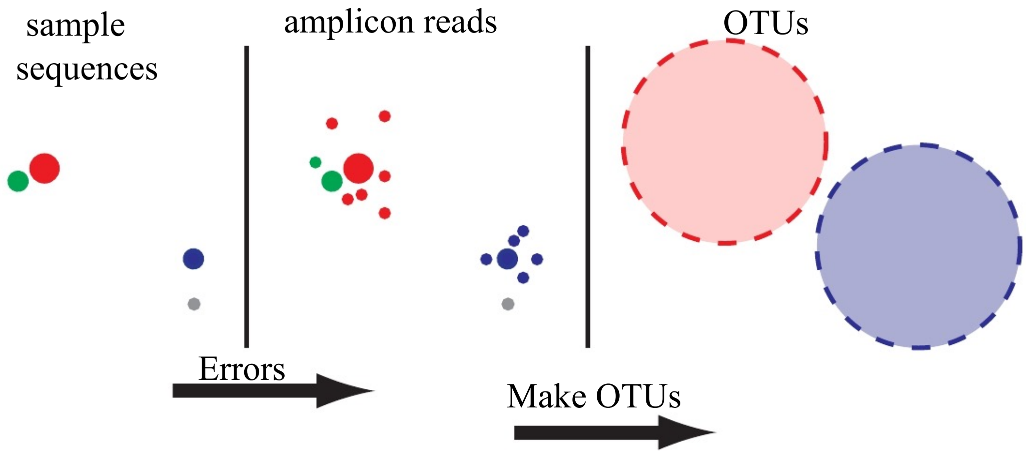
Motivation: Lingering problem with “OTUs”

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 ~ 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**



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



sample sequences

amplicon reads

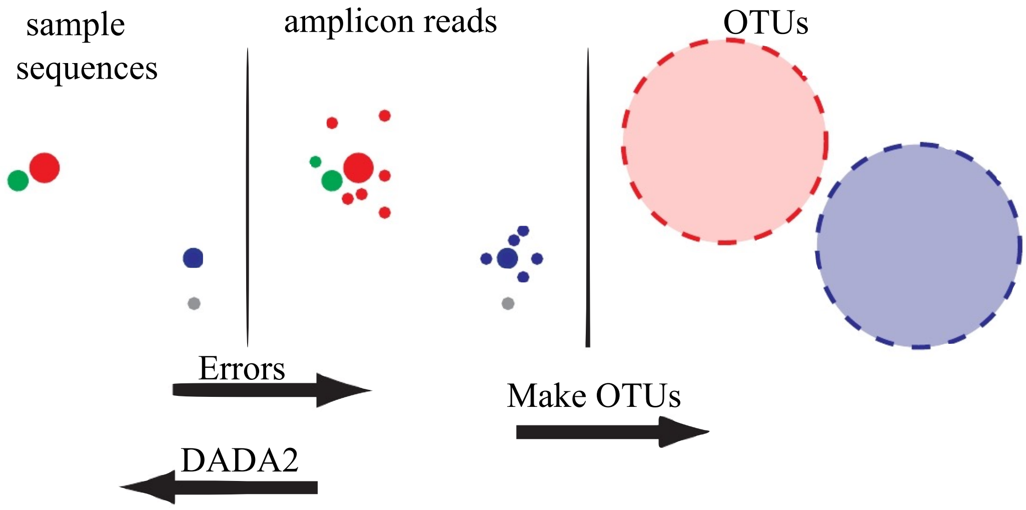
OTUs

Errors

Make OTUs

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

Sample Inference from Noisy Reads



sample sequences

amplicon reads

OTUs

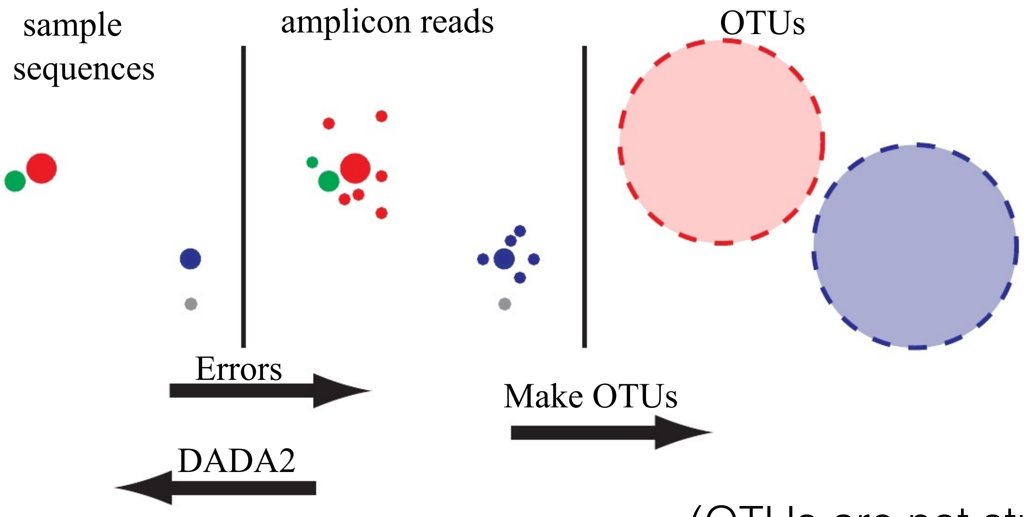
Errors

Make OTUs

DADA2

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

Sample Inference from Noisy Reads



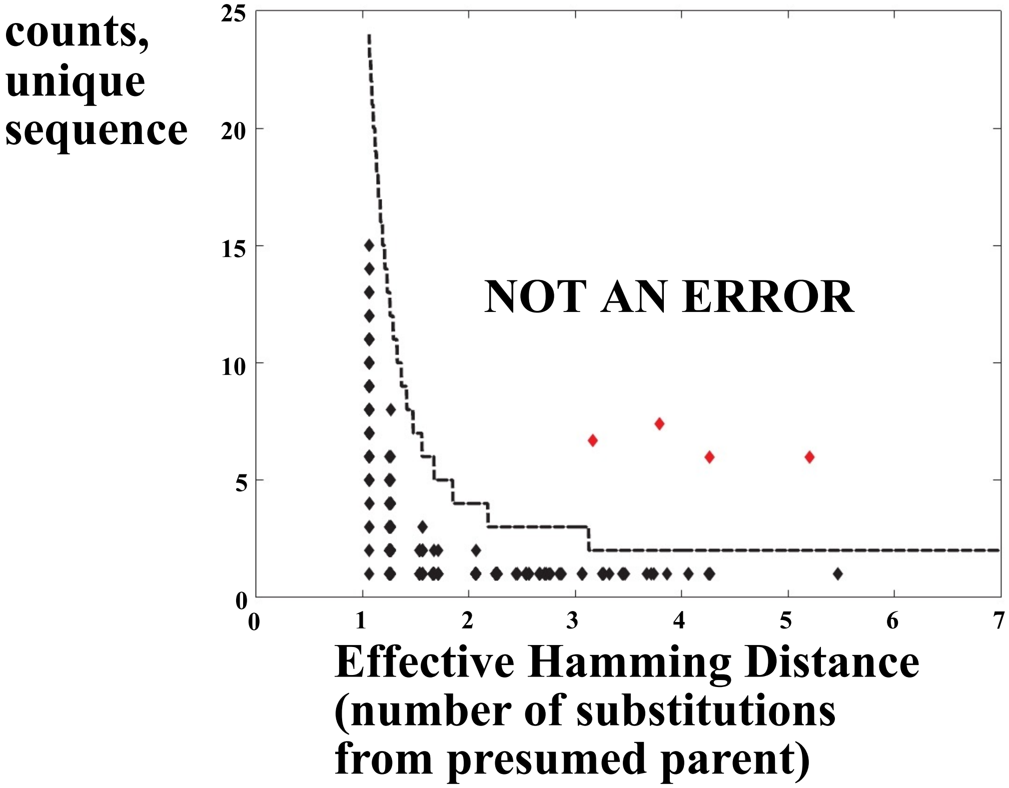
(OTUs are not strains)

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

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

The true shape of an error cloud

DADA2: Error Model



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

DADA2 algorithm assumptions

DADA2 Error Model

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:

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

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

DADA2 algorithm assumptions

DADA2 Abundance Model

DADA2 algorithm assumptions

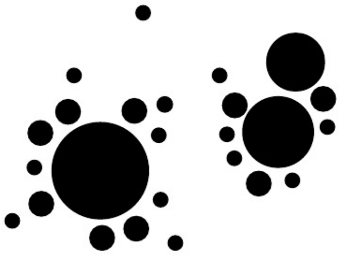
DADA2 Abundance Model

- 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, $\lambda j \to i$, 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:

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

- “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

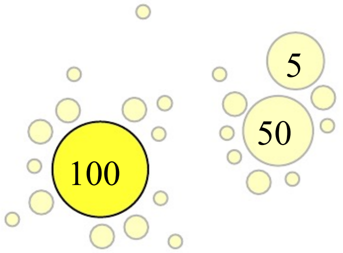
DADA2 algorithm cartoon



Initial guess: one real sequence + errors

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

DADA2 algorithm cartoon



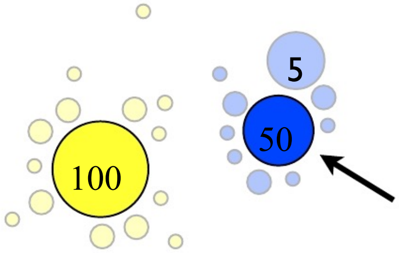
Infer initial *error model* under this assumption.

$\Pr(i \rightarrow j) =$

	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:-)

DADA2 algorithm cartoon



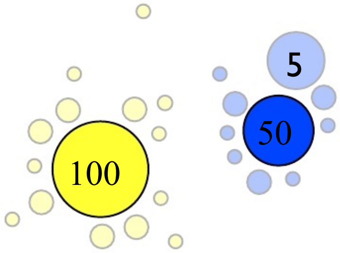
not an error

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:-)

DADA2 algorithm cartoon



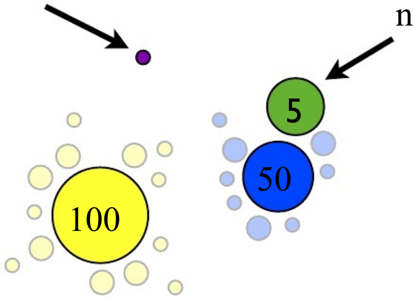
Update the model.

	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:-)

DADA2 algorithm cartoon

not an error



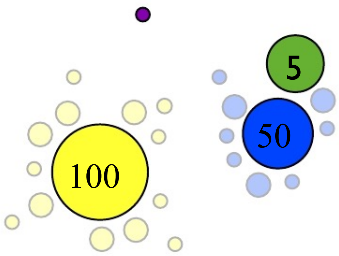
not an error

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

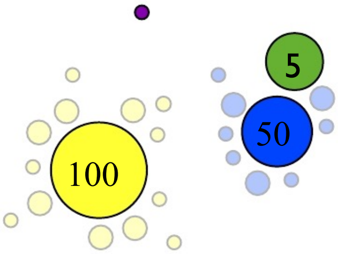


Update model again

	A	C	G	T
A	0.998	$1x10_{-4}$	$2x10_{-3}$	$2x10_{-4}$
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

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

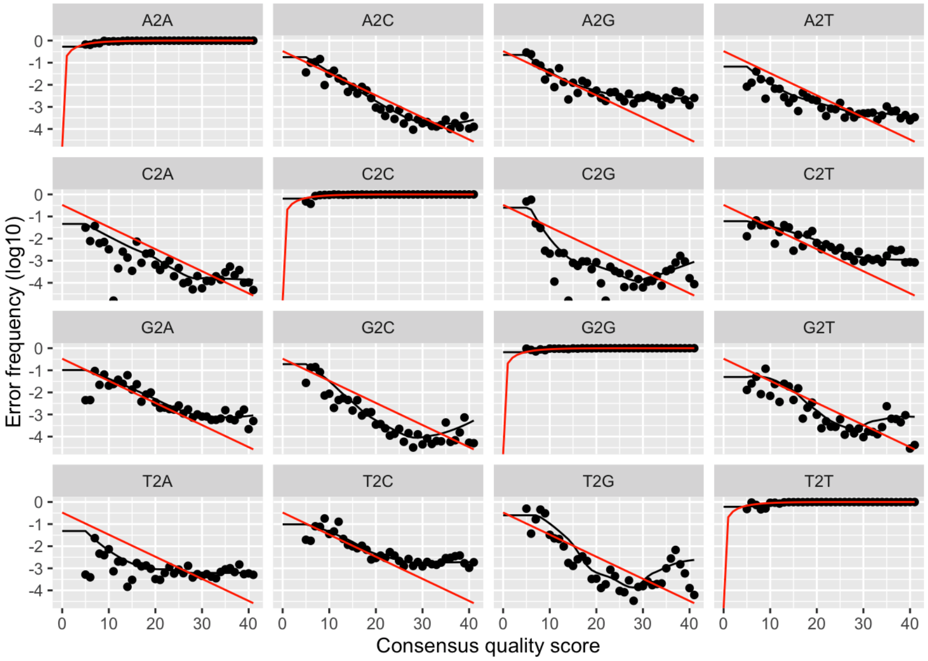
DADA2 algorithm cartoon



Convergence: all errors are plausible

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

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



- *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

DADA2: Why is this possible?

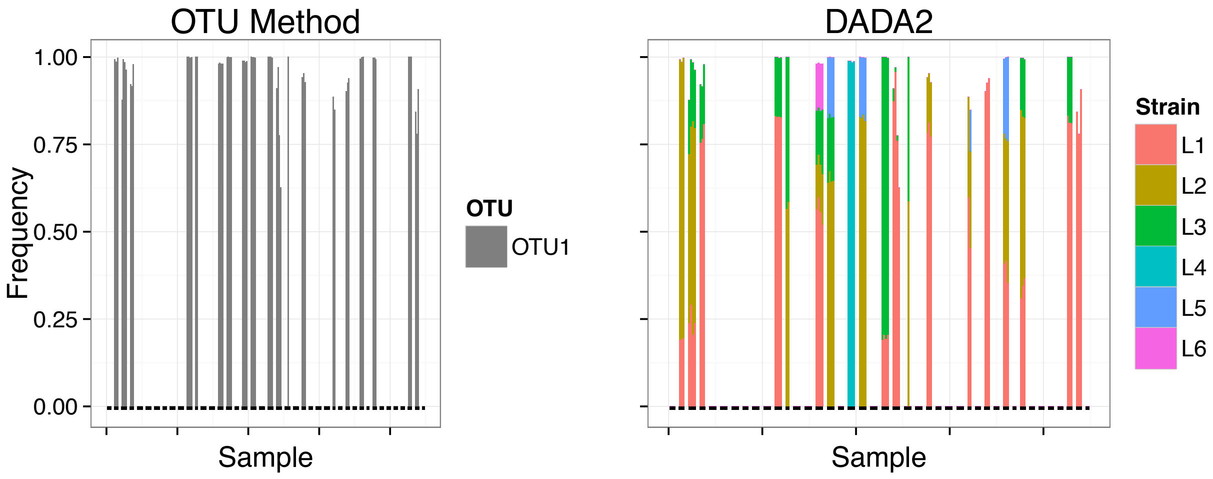
Uses more of the information than traditional OTU clustering

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

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

DADA2 Advantages: Resolution

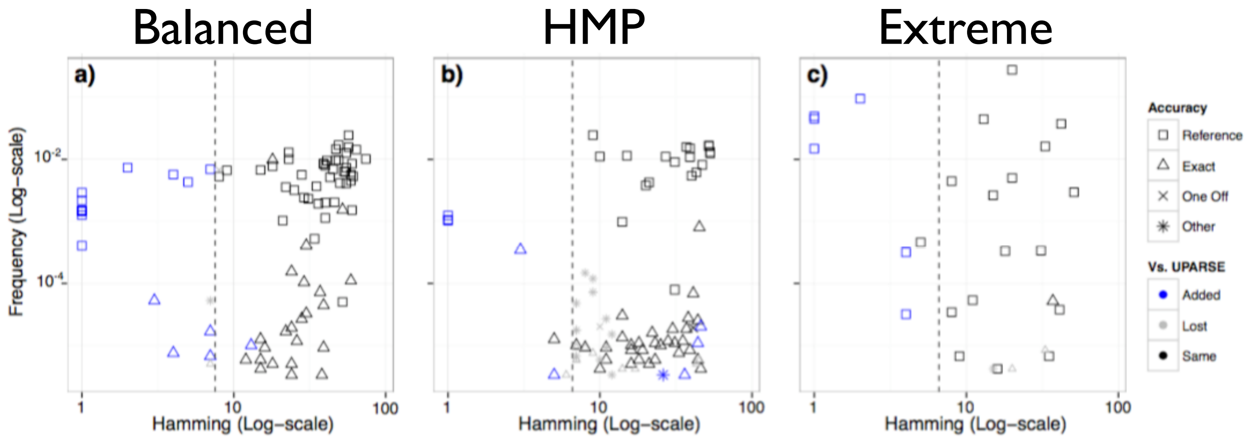
Lactobacillus crispatus sampled from
vaginal microbiome 42 pregnant women



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

DADA2 Advantages: Accuracy benchmarks

Mock community data for accuracy benchmarking



Balanced

HMP

Extreme

DADA2 performance relative to UPARSE
(best available alternative)

DADA2 Advantages

Analytical

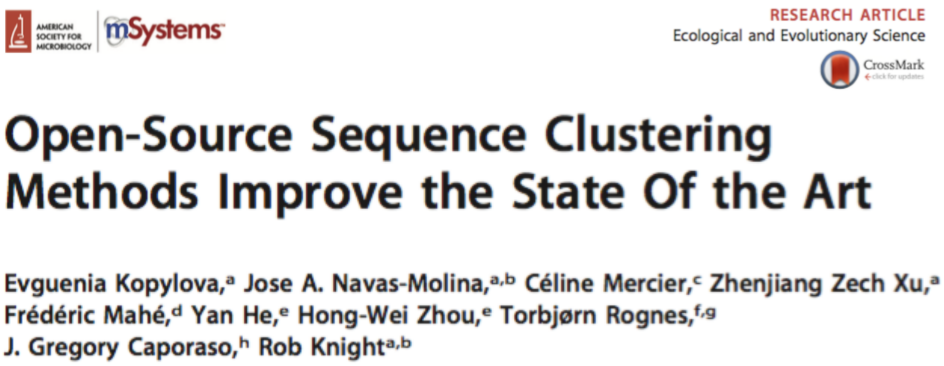
Single nucleotide resolution
- genotypes/strains instead of 97% OTUs

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:-)



AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

RESEARCH ARTICLE
Ecological and Evolutionary Science

CrossMark

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

Evguenia 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]

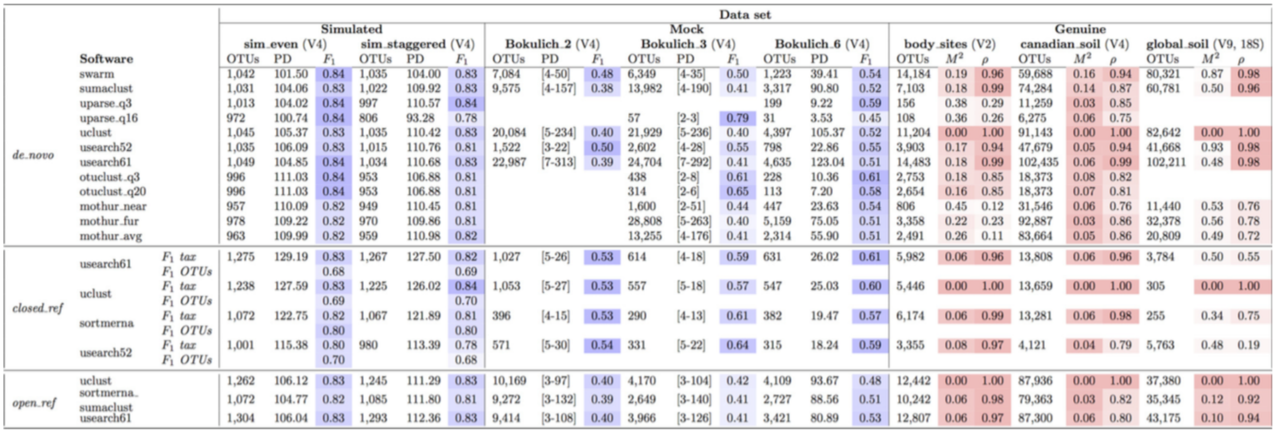
February2016

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)

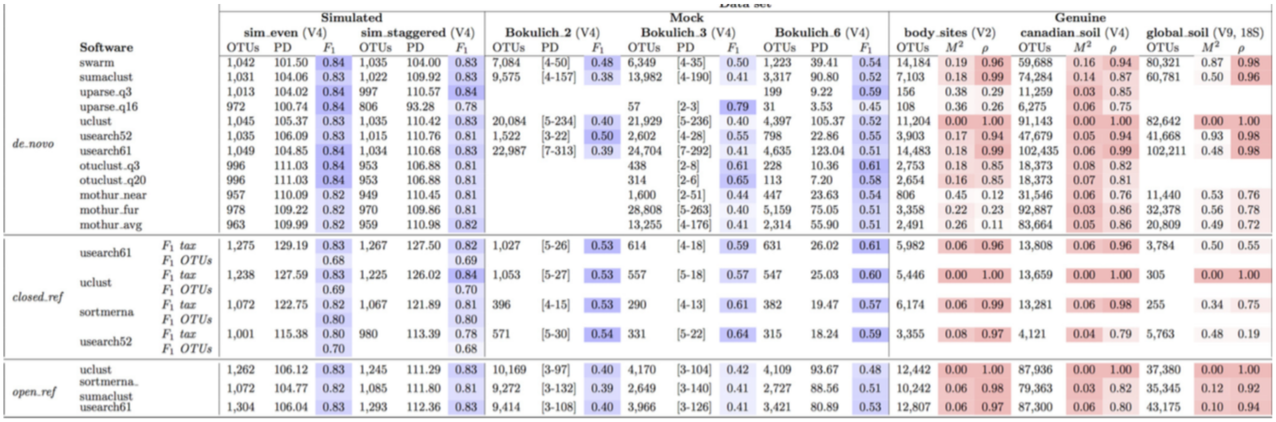
Kopylova, et al (2016).
Open-source sequence clustering methods improve the state of the art.
mSystems
http://doi.org/10.1186/s12915-014-0069-1



			Data set																							
			Simulated						Mock									Genuine								
			sim_even (V4)			sim_staggered (V4)			Bokulich_2 (V4)			Bokulich_3 (V4)			Bokulich_6 (V4)			body_sites (V2)			canadian_soil (V4)			global_soil (V9, 18S)		
	Software		OTUs	PD	F_1	OTUs	PD	F_1	OTUs	PD	F_1	OTUs	PD	F_1	OTUs	PD	F_1	OTUs	M^2	ρ	OTUs	M^2	ρ	OTUs	M^2	ρ
de_novo	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
	sumaclust		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
	otuclust_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			
	otuclust_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
closed_ref	usearch61	F_1 *tax*	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
		F_1 *OTUs*			0.68			0.69																		
	uclust	F_1 *tax*	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
		F_1 *OTUs*			0.69			0.70																		
	sortmerna	F_1 *tax*	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
		F_1 *OTUs*			0.80			0.80																		
	usearch52	F_1 *tax*	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
		F_1 *OTUs*			0.70			0.68																		
open_ref	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
	sortmerna_ sumaclust		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
	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



			Data set																							
			Simulated						Mock									Genuine								
			sim_even (V4)			sim_staggered (V4)			Bokulich_2 (V4)			Bokulich_3 (V4)			Bokulich_6 (V4)			body_sites (V2)			canadian_soil (V4)			global_soil (V9, 18S)		
	Software		OTUs	PD	F_1	OTUs	PD	F_1	OTUs	PD	F_1	OTUs	PD	F_1	OTUs	PD	F_1	OTUs	M^2	ρ	OTUs	M^2	ρ	OTUs	M^2	ρ
de_novo	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
	sumaclust		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
	otuclust_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			
	otuclust_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
closed_ref	usearch61	F_1 *tax*	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
		F_1 *OTUs*			0.68			0.69																		
	uclust	F_1 *tax*	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
		F_1 *OTUs*			0.69			0.70																		
	sortmerna	F_1 *tax*	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
		F_1 *OTUs*			0.80			0.80																		
	usearch52	F_1 *tax*	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
		F_1 *OTUs*			0.70			0.68																		
open_ref	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
	sortmerna_ sumaclust		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
	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

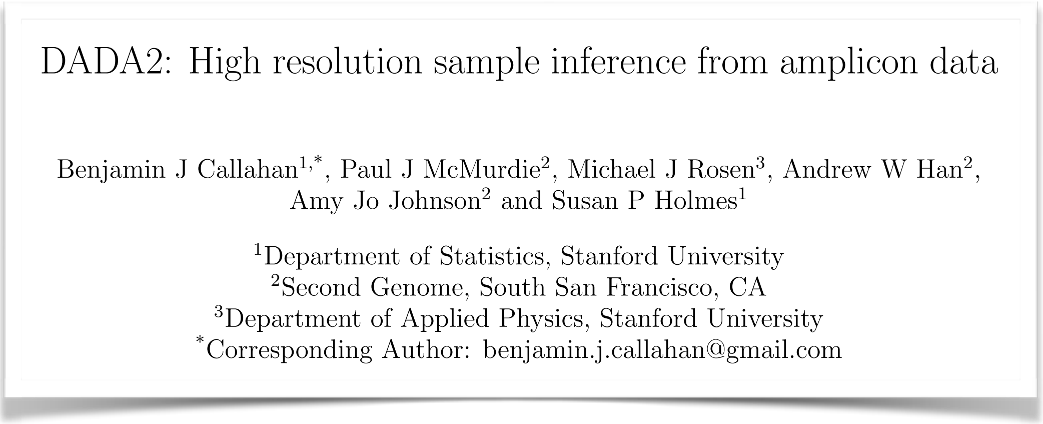
DADA2 performance:

- 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

DADA2

Divisive Amplicon Denoising Algorithm - ver.2



DADA2: High resolution sample inference from amplicon data

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

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

http://dx.doi.org/10.1101/024034 Manuscript draft on bioRxiv (*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.

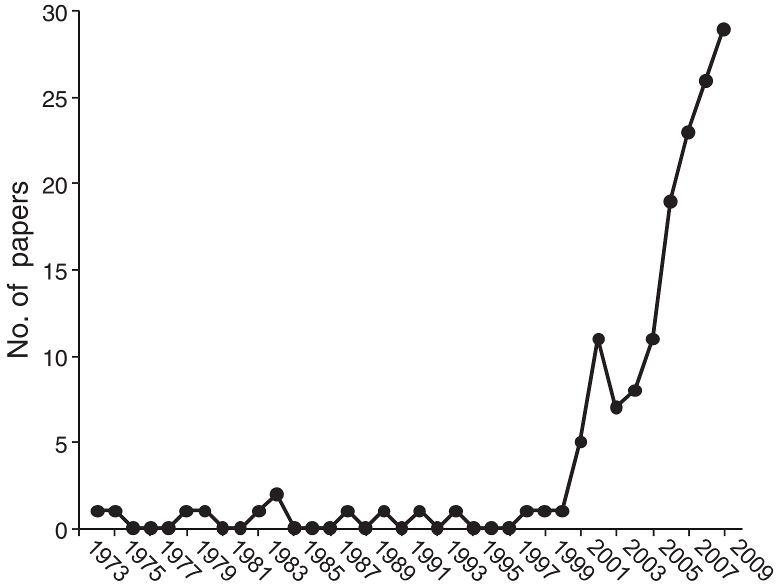
Diversity

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, γ = α × β, but only this simple if α and β are independent
- Probably others, but α and β are most common

Beta-Diversity

Peer-reviewed articles having “beta diversity” in title



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

Beta-Diversity

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

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

“Diversity of species composition”

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

Distances between microbiomes

Community Distance

Communities are a vector of abundances:
$\mathbf{x} = \{x_1, x_2, x_3, \ldots\}$

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

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

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

Community Distance Properties

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

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

The Distance Spectrum

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

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

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:-)

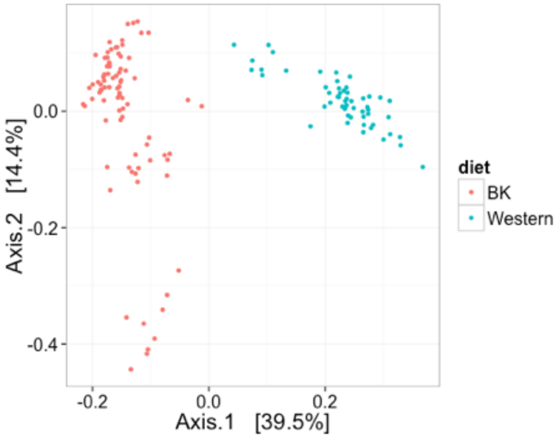
Ordination Methods

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,…
.
.



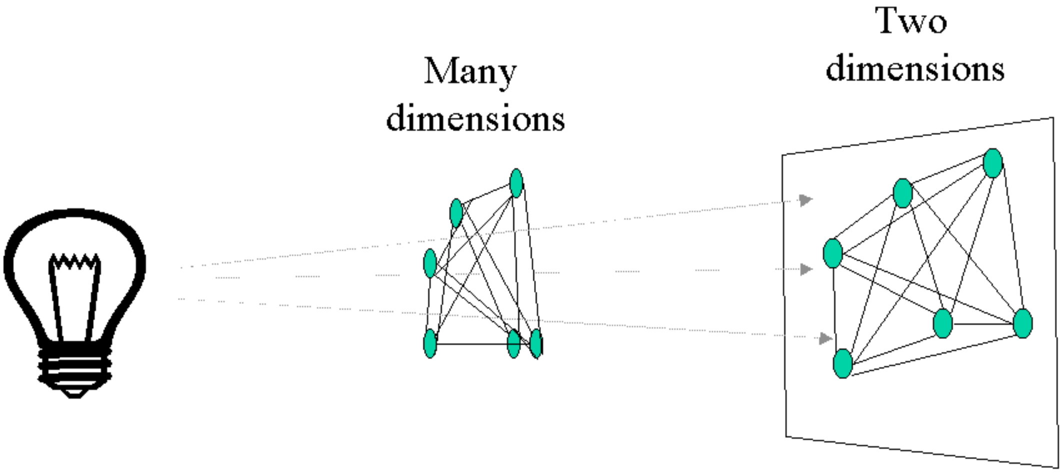
P-dimensions

2-dimensions

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

Multi-dimensional Scaling

Why MDS? It works with any distance!



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

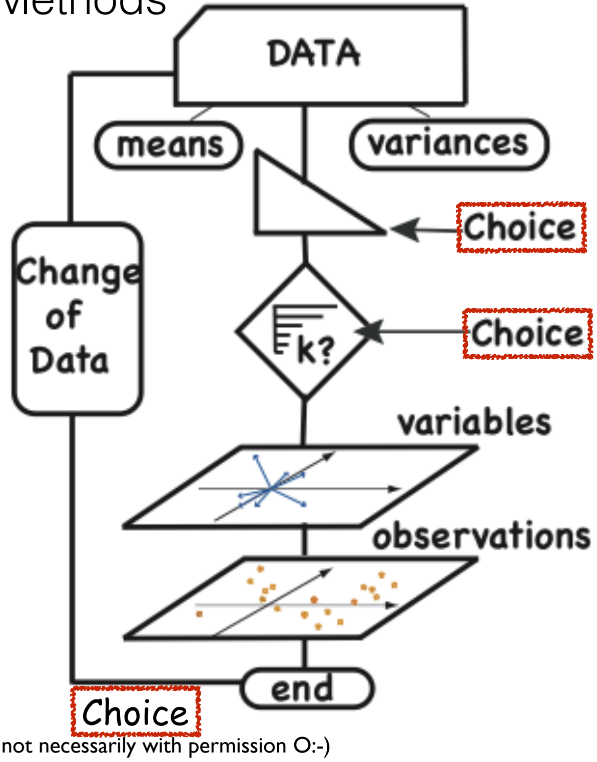
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

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?