

Introduction to Microbiome Analysis

Objectives:

- What is a microbiome?
- What is 'culture independent technique'?
- Why is it useful?
- What is amplicon sequencing?
- What do people mean when they say "16S"?

<BREAK>

- What do we do with our microbiome (amplicon) sequence data?
 - DADA2

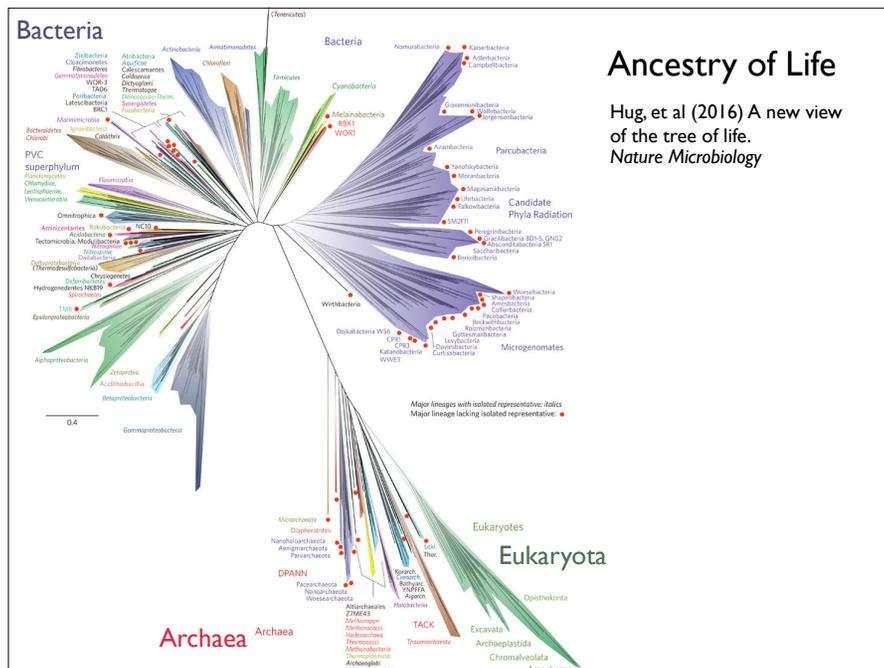
1

Introduction to Microbiome Analysis

“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

2



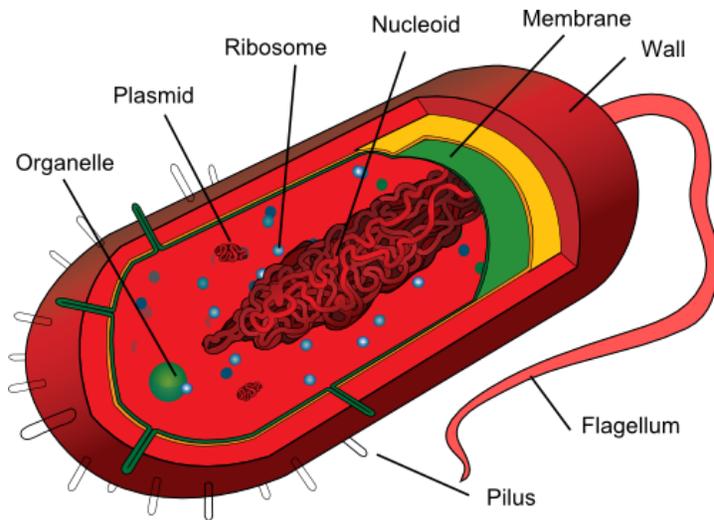
3

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

- **In 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
- **If microbes suddenly all disappeared...**
 - Horrible maladies for most animals (esp. development, metabolism)
 - Most animal species would become extinct within a year (estimate)
 - There would be (almost) no oxygen in the atmosphere
 - ocean microbes alone account for ~half of your O₂
 - We'd all quickly die of CO₂ poisoning (and later global warming)
 - Most elemental cycles are predominantly microbe-driven

4

What are microbes?



5

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

6

What is a microbiome?

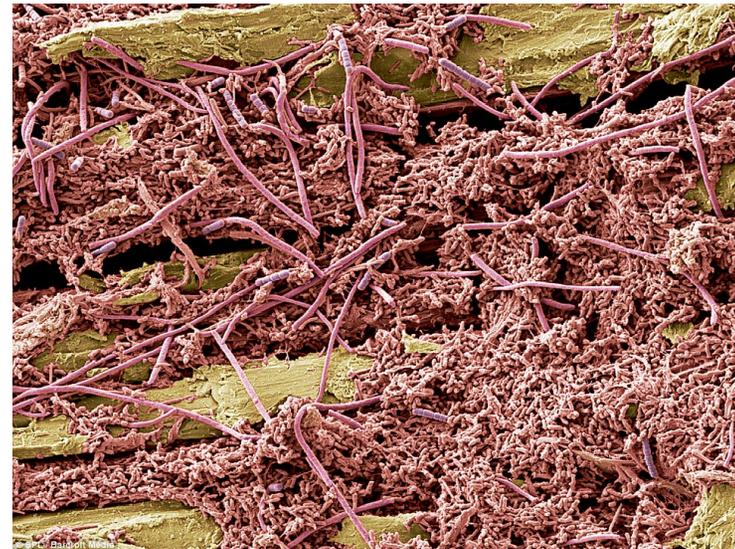
The totality of microbes in a defined environment, especially their genomes and interactions with each other and surrounding environment.

- 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!

7

Exercise: How many species are present?



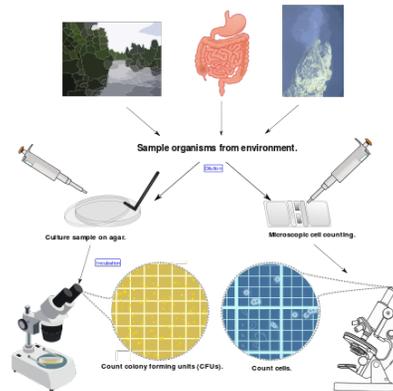
Confer amongst yourselves. We'll take a poll.

8

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**



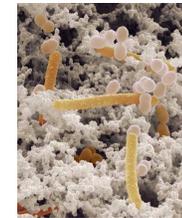
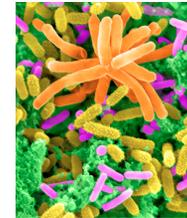
Staley, J.T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, 39, 321–346.

9

Discovery of *Culture Independent* Techniques

Why is microbiome research new? given...

- 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” ...

10

Discovery of *Culture Independent* Techniques

Why is microbiome research new? given...

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
- Accessible computing and algorithms

11

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

12

Discovery of *Culture Independent* Techniques

- 1977 rRNA as evolutionary marker - Woese & Fox *PNAS*

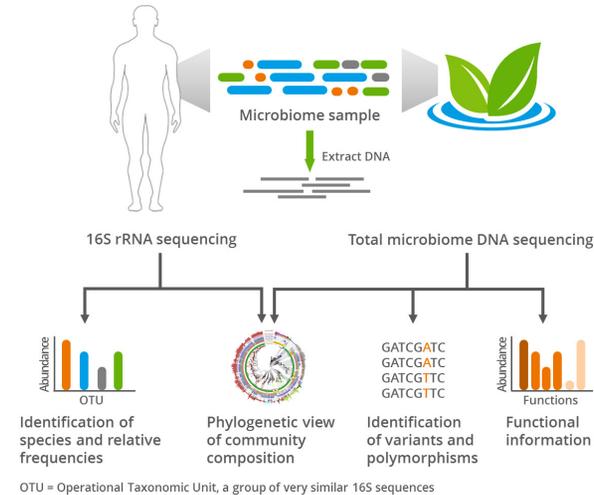
Woese was originally scorned at the discovery of archaea via rRNA gene (dis)similarity.

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

- 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

13

Culture Independent Techniques

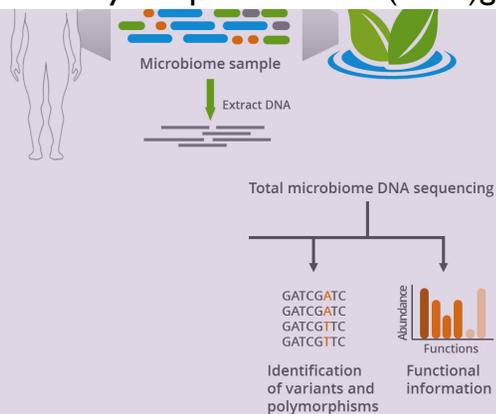


<https://www.gatc-biotech.com/en/expertise/targeted-sequencing/16s-rrna-analysis.html>

14

Culture Independent Techniques

Why not just always sequence entire (meta)genomes?



<https://www.gatc-biotech.com/en/expertise/targeted-sequencing/16s-rrna-analysis.html>

15

Culture Independent Techniques

Why not just always sequence entire (meta)genomes?

(similar motivation to RADSeq in pop-gen):

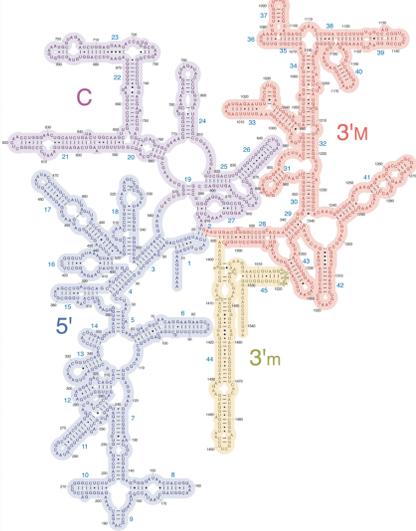
- still **prohibitively expensive (inefficient)**
- for many biological questions a full sequence isn't needed
- For low-abundance microbes, or high numbers of samples, amplicon sequencing might be the only feasible option

- This is a different kind of “Reduced representation sequencing”
- Use **restriction enzyme digestion** PCR amplification to focus sequencing of multiple samples on [one] homologous regions across the genomes
- Cost is a fraction of the cost of re-sequencing the metagenomes

16

The Small Subunit “16S” ribosomal RNA

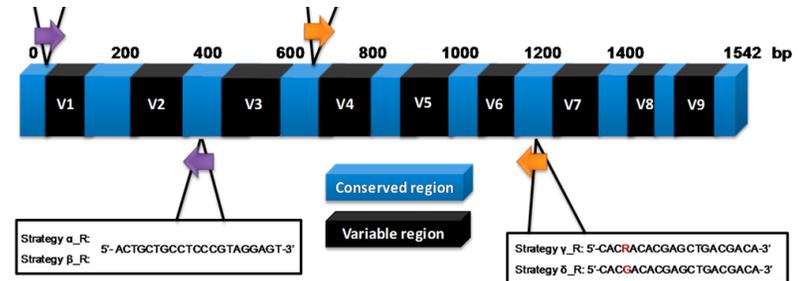
Small subunit “16S” rRNA



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

21

The Small Subunit “16S” ribosomal RNA



16S rRNA gene as target for amplicon sequencing

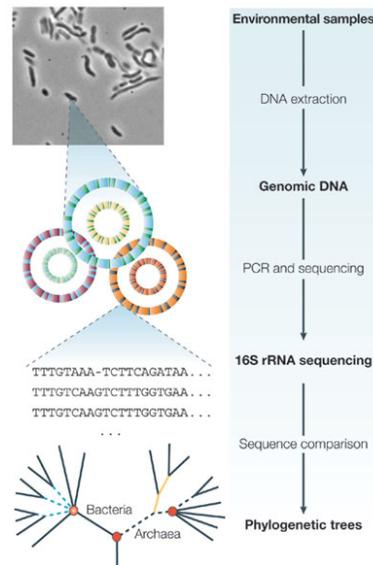
Int. J. Mol. Sci. 2014, 15(11), 21476-21491; doi:10.3390/ijms151121476

22

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



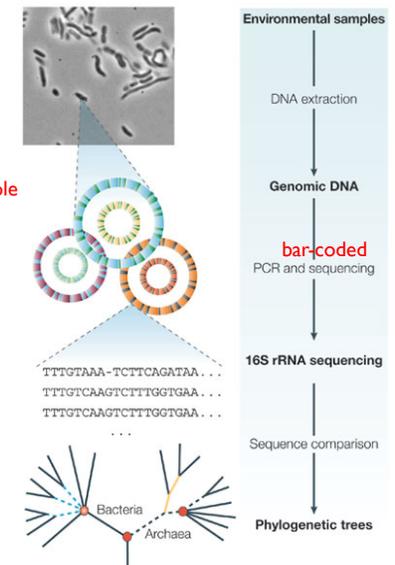
Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, 6(11), 805–814.

23

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



24

Divisive Amplicon-sequence Denoising Algorithm (DADA)

25

You just generated amplicon seq data...

You have a big pile of sequences that were amplified from the same genetic locus, simultaneous from the genomic DNA of many organisms...

- Separate real from error-containing sequences
- Count the abundances
 - True sequence + its errors

For many years now, the common practice was to solve this by UPGMA-style clustering at a fixed sequence distance (97% similarity).

“Operational Taxonomic Unit” - OTU

This was believed to approximate a species similarity, while also conveniently similar to the typical error rate from 454 sequencing, the popular platform at the time these methods proliferated...

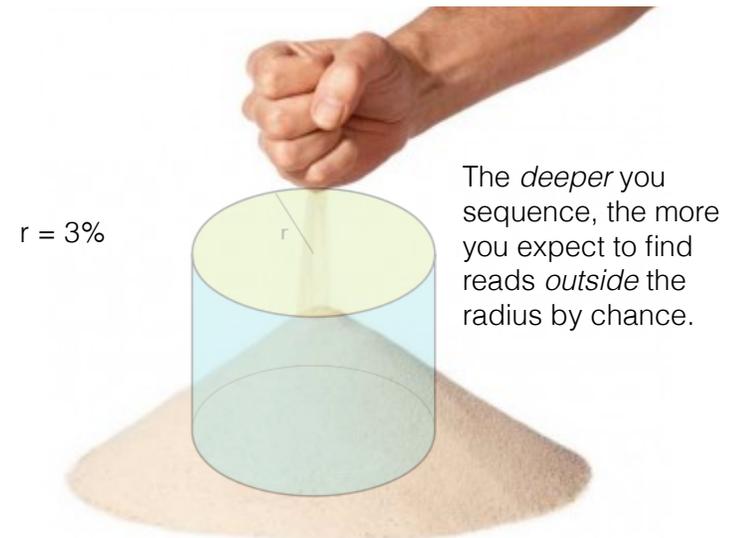
26

Motivation: Lingering problems with “OTU”



27

Motivation: Lingering problems with “OTU”



28

Motivation: Lingering problems with “OTU”

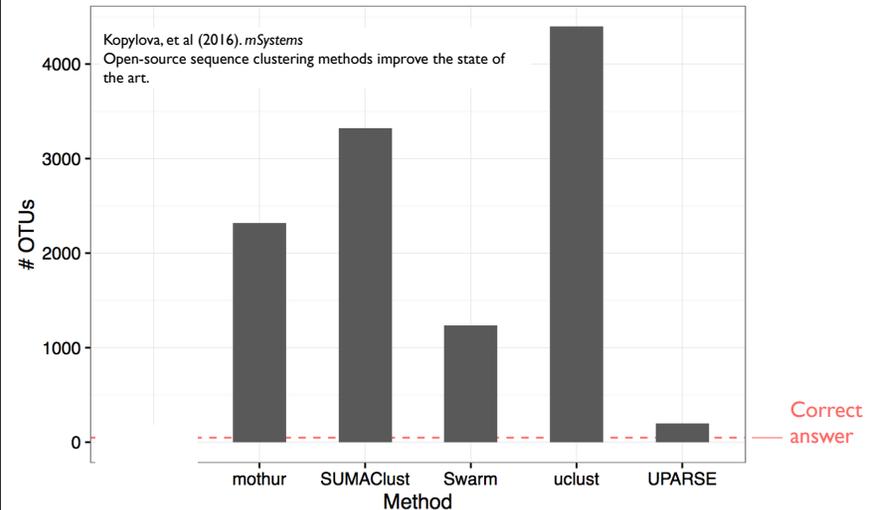
- False Positives - e.g. 1000s of OTUs when only 10s of sequences present
 - Consequently, richness appears to depend on library size
 - Microbiome distances that appear to depend on library size
- Poor Seq/Taxonomic Resolution - defined by arbitrary similarity radius
- Accuracy - Abundance estimates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

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

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

29

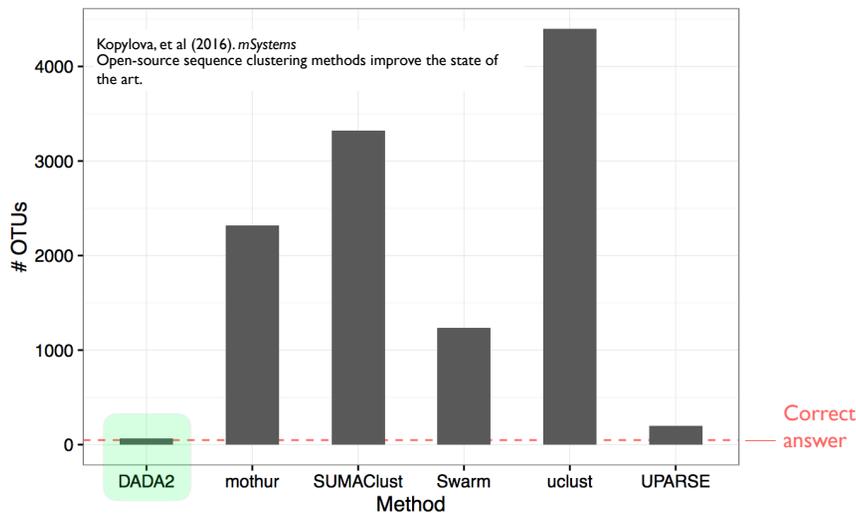
Typical “OTU” performance on validation data (“mock community”)



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

30

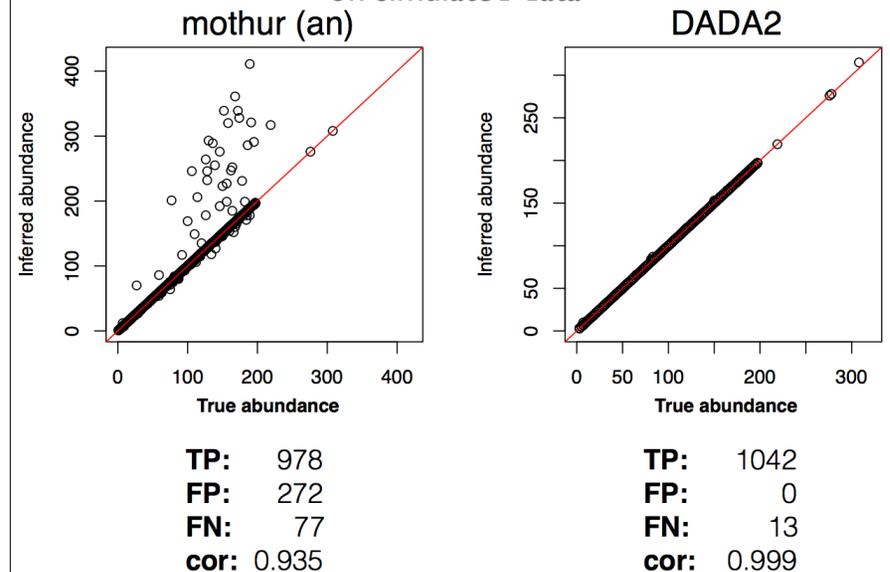
Typical “OTU” performance on validation data (“mock community”)



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

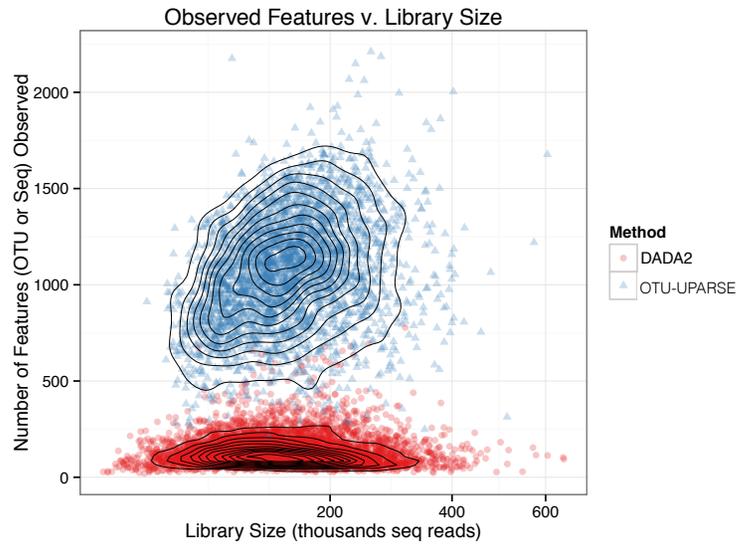
31

Typical “OTU” performance on simulated data

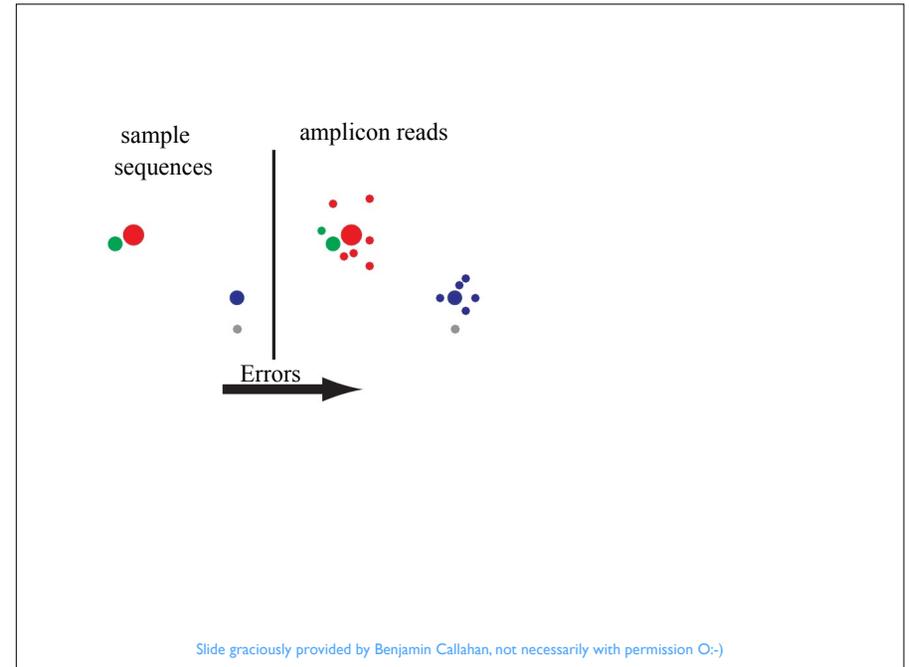


32

Anecdotal example of mitigated dependence of observed richness on sequencing effort

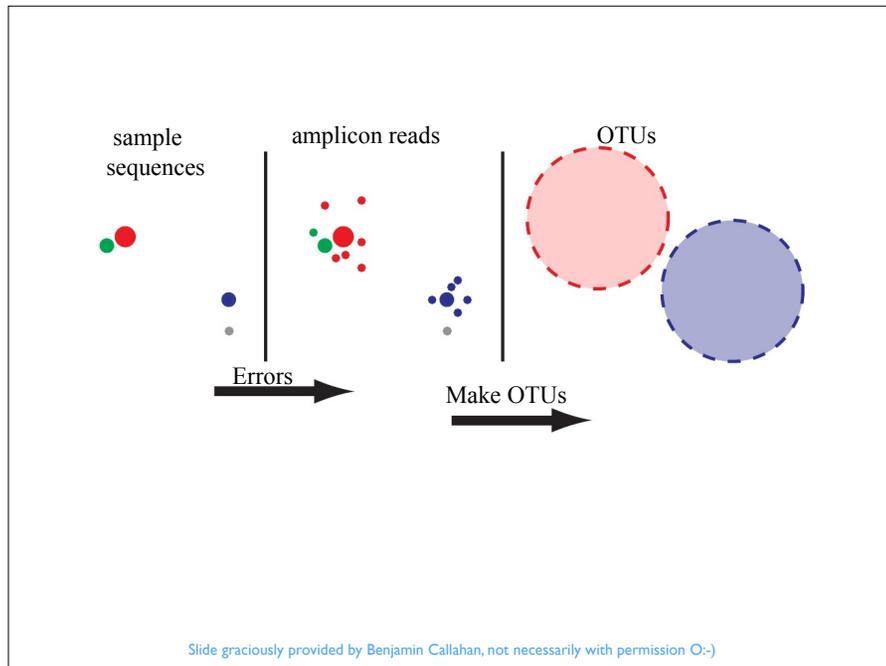


33



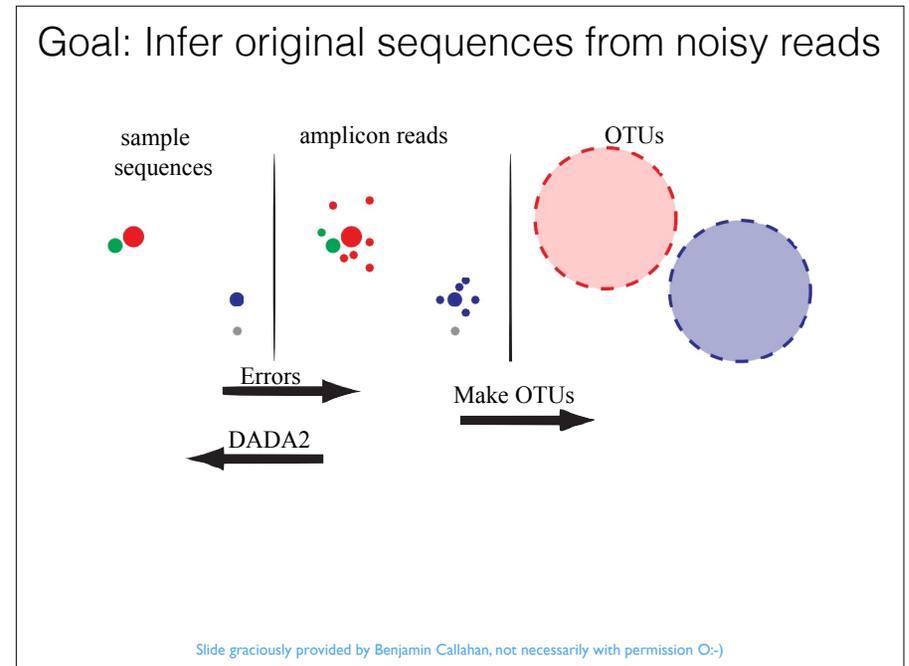
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

34



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

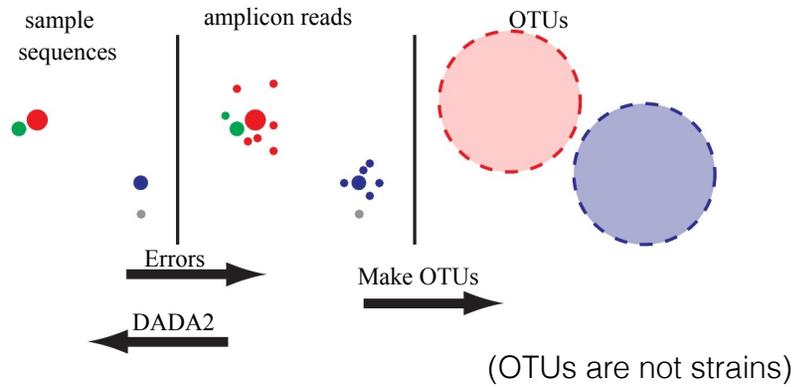
35



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

36

Goal: Infer original sequences from noisy reads

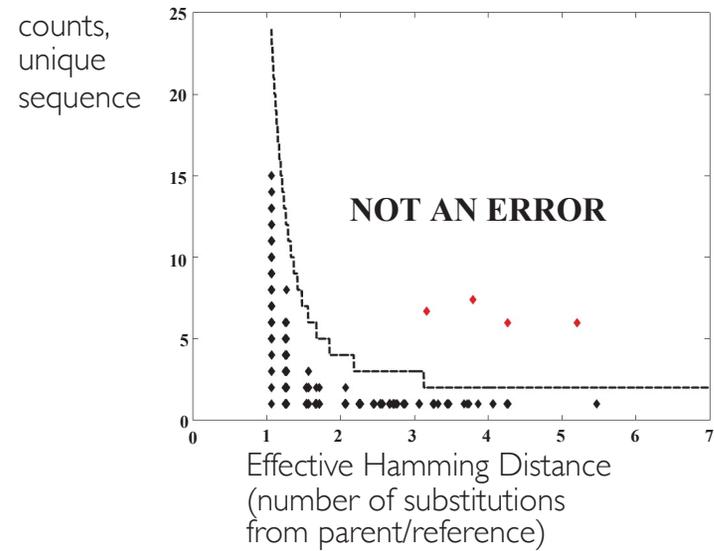


OTUs: Lump similar sequences together
 DADA2: Statistically infer the sample sequences (strains)

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

37

The shape of amplicon sequencing errors

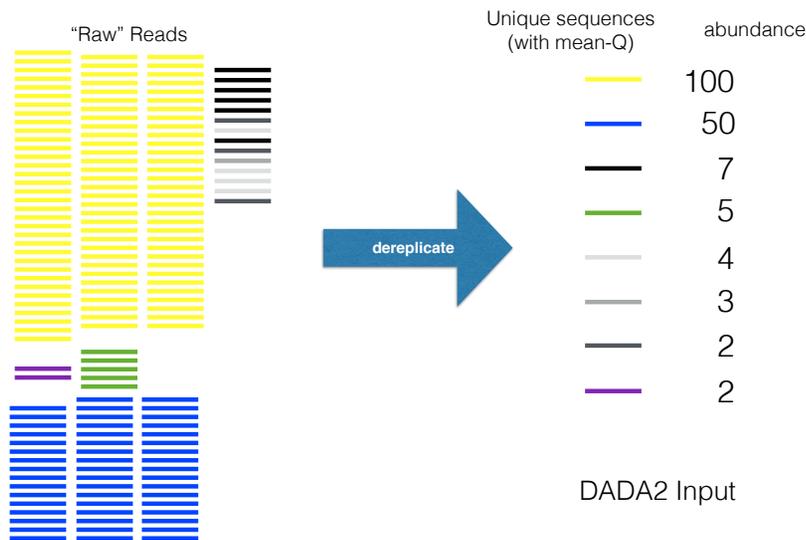


Slide adapted from Benjamin Callahan

38

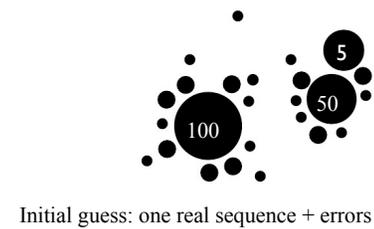
DADA2 algorithm cartoon

Input: unique sequences, their quality values, and abundances



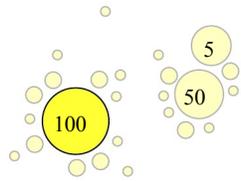
39

DADA2 algorithm cartoon



40

DADA2 algorithm cartoon



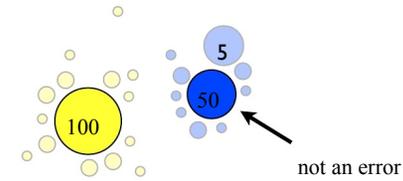
Infer initial *error model* under this assumption.

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

	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

41

DADA2 algorithm cartoon

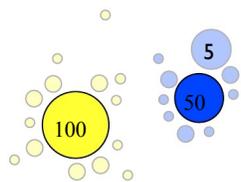


Reject unlikely error under model. **Recruit** errors.

	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

42

DADA2 algorithm cartoon

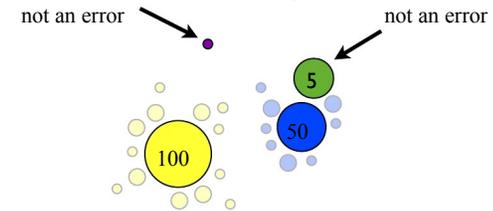


Update the model.

	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

43

DADA2 algorithm cartoon

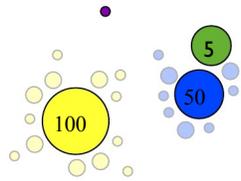


Reject more sequences under *new* model

	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

44

DADA2 algorithm cartoon

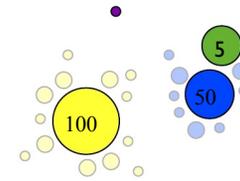


Update model again

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

45

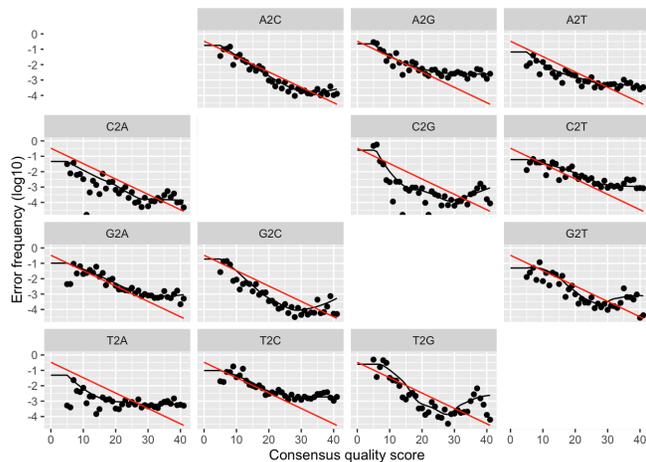
DADA2 algorithm cartoon



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^{-4}	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

46



- *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

47

DADA2 algorithm assumptions

48

DADA2 algorithm assumptions

DADA2 Error Model

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

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

- Each substitution probability depends on original nt, substituting nt, and quality score at position in i

49

DADA2 algorithm assumptions

DADA2 Abundance Model

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

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

- "Probability of seeing an abundance of sequence i that is equal to or greater than observed value, by chance, given sequence j ." (Bonferroni-corrected)
- A low p_A indicates there are more reads of sequence i than can be expected given n_j

50

performance, etc.

- Inferred sequences are *verifiable*
 - A major departure from OTU clustering methods
- Inferred sequences are *intrinsically comparable*
 - between samples
 - between experiments
 - A major departure from OTU clustering methods
- Computation on each specimen *independently*
 - *Embarrassingly parallel*
 - Much faster, accessible for large projects
 - Can use cheap commodity hardware (e.g. your laptop)
 - rather than \$\$ high-memory clusters
 - Robust: results don't change with new samples/projects
 - Artifact sample cannot affect others

51

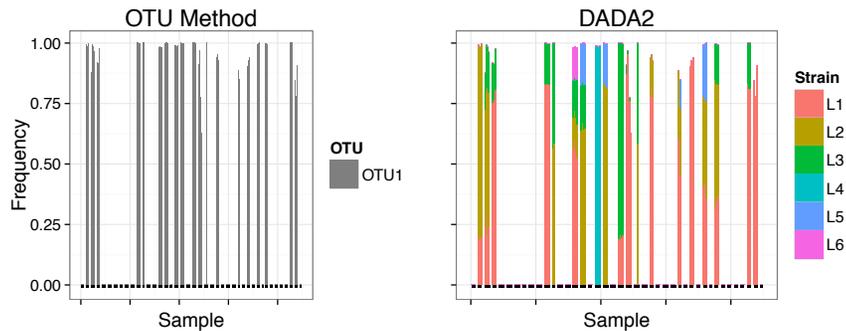
Applications

- Any amplicon-seq data, not just 16S rRNA or even microbiome
- Sequence variants unique to an individual host
- Sequence variants associated with a clinical outcome
- Improved meta-genomic inference (e.g. PICRUST)
 - Mitigate ambiguity of representative genome(s) to use
- Detecting pathogens (special cases)

52

Real example, exact sequence resolution

Lactobacillus crispatus sampled from vaginal microbiome 42 pregnant women



Data: MacIntyre et al. Scientific Reports, 2015.

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

53

Perspective

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

OPEN

Benjamin J Callahan¹, Paul J McMurdie² and Susan P Holmes³

¹Department of Population Health and Pathobiology, NC State University, Raleigh NC, USA

²Whole Biome Inc, San Francisco CA, USA

³Department of Statistics, Stanford University, Stanford CA, USA

The ISME Journal 21 July 2017; doi: 10.1038/ismej.2017.119

The **ISME** Journal
Multidisciplinary Journal of Microbial Ecology

Other relevant articles:

UNOISE2 — *bioRxiv* **Oct 2016** 081257

Deblur — *mSystems* **Mar 2017** 2 (2) e00191-16 **Unknown**

*MED — *The ISME Journal* **2015** 9, 968–979 **High FP!**

*DADA1 — *BMC bioinformatics* **2012** 13(1), 283 **Slow**

PyroNoise — *BMC Bioinformatics* Quince et al. **2011** 454 **only**

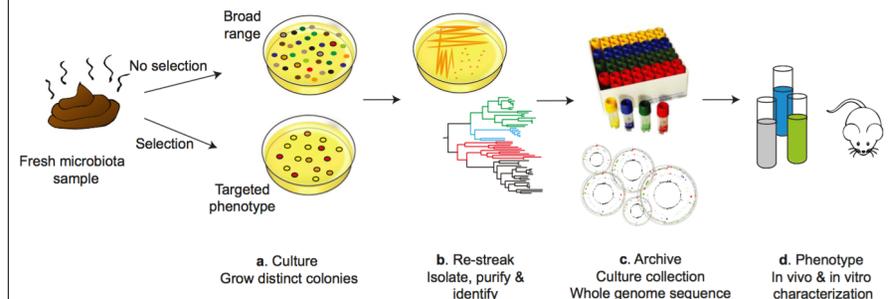
54

Where things are headed: “Culturomics”

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.

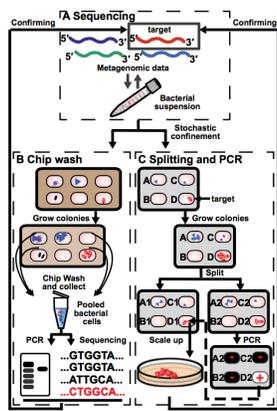


Browne, H. P., et al. (2016). Culturing of “unculturable” human microbiota... *Nature*, 533(7604), 543–546.

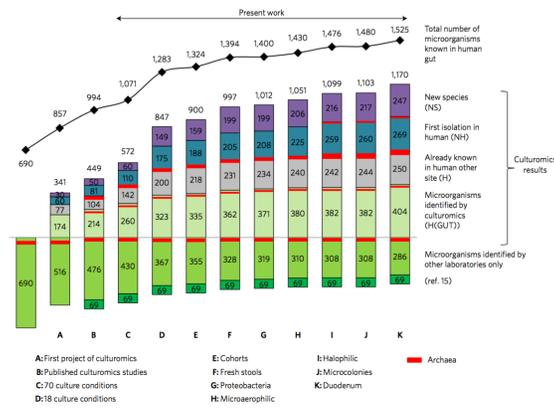
55

56

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

Next Up: Lab 01

We are going to run DADA2 on “raw” amplicon sequence data

Any lingering questions?