UCI Microbiome Initiative

Introduction to Microbiome Analyses in R

Alexander B. Chase
Postdoctoral Fellow
Paul Jensen Lab
October 23, 2019

Biological Question

↓

Design Experiment

↓

Conduct Study

Collect Samples
Store Samples
Extract DNA, PCR, Sequencing

↓

Microbial Community Analysis

16S rRNA marker gene
Metagenomics

↓

Statistical Analysis and Data Interpretation

Biological Question

↓

Design Experiment

↓

Conduct Study

Collect Samples
Store Samples
Extract DNA, PCR, Sequencing

↓

Microbial Community Analysis

16S rRNA marker gene
Metagenomics

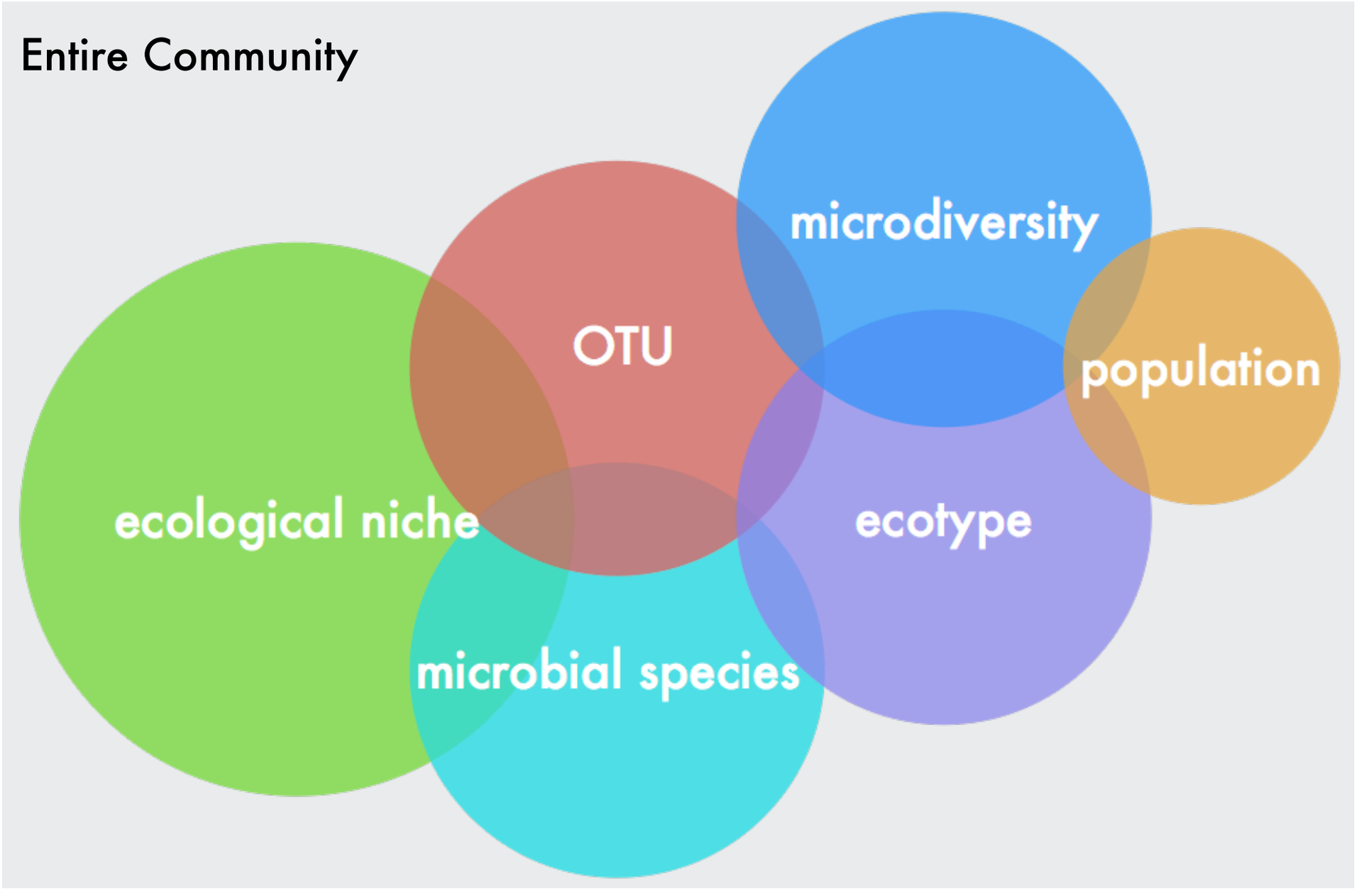
↓

Statistical Analysis and Data Interpretation

Microbial Community Analysis

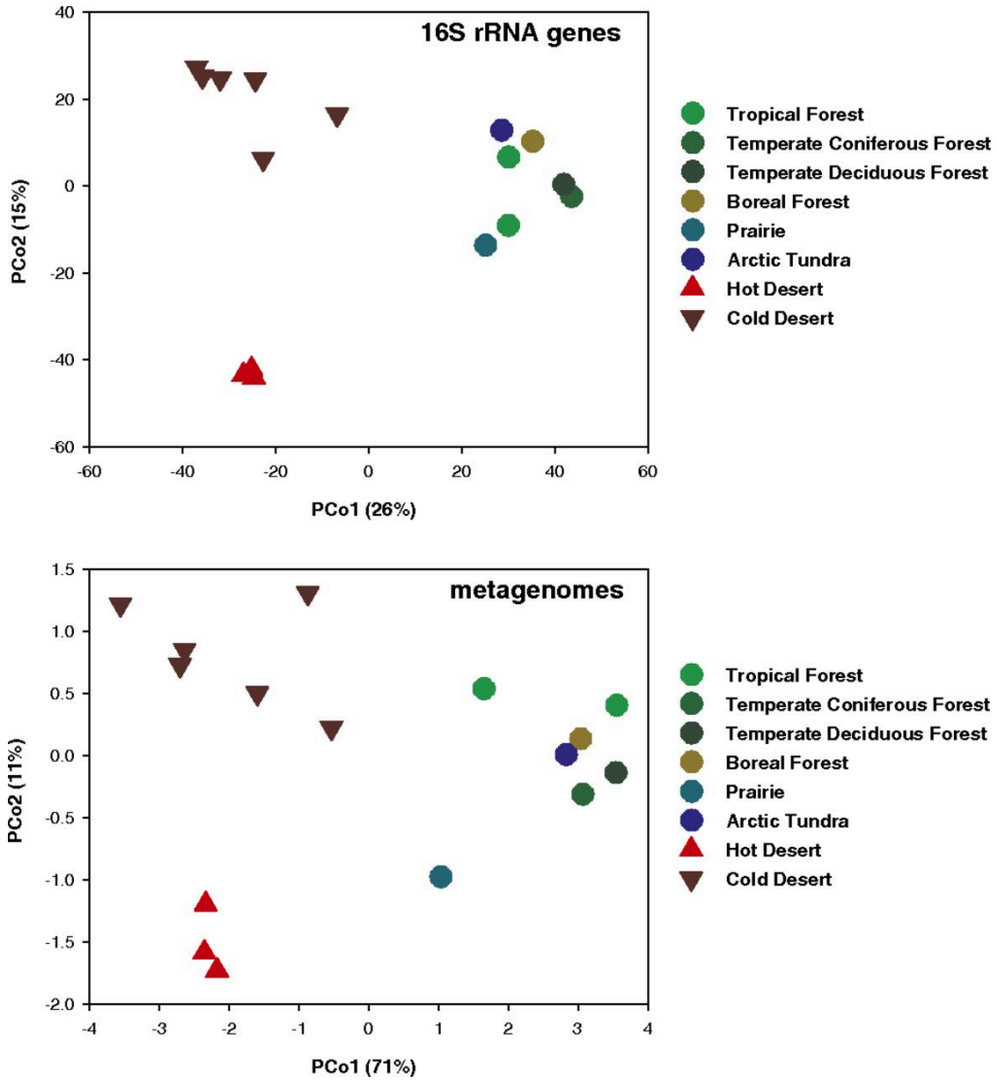
16S rRNA marker gene
Metagenomics

Which genes should you sequence? How much data do you need?

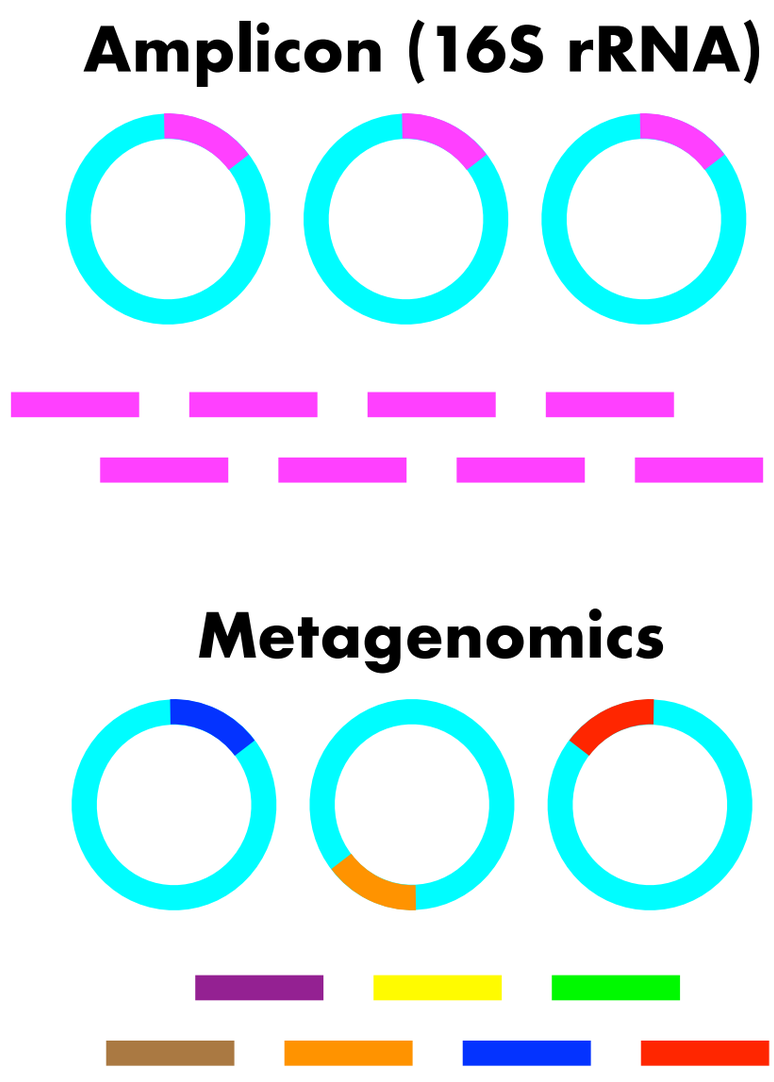


Microbial Community Analysis

16S rRNA marker gene
Metagenomics



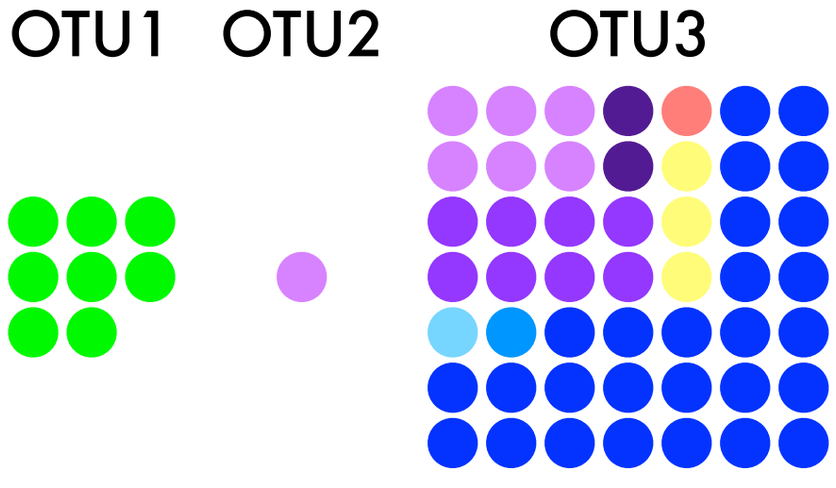
Fierer et al. PNAS. 2012



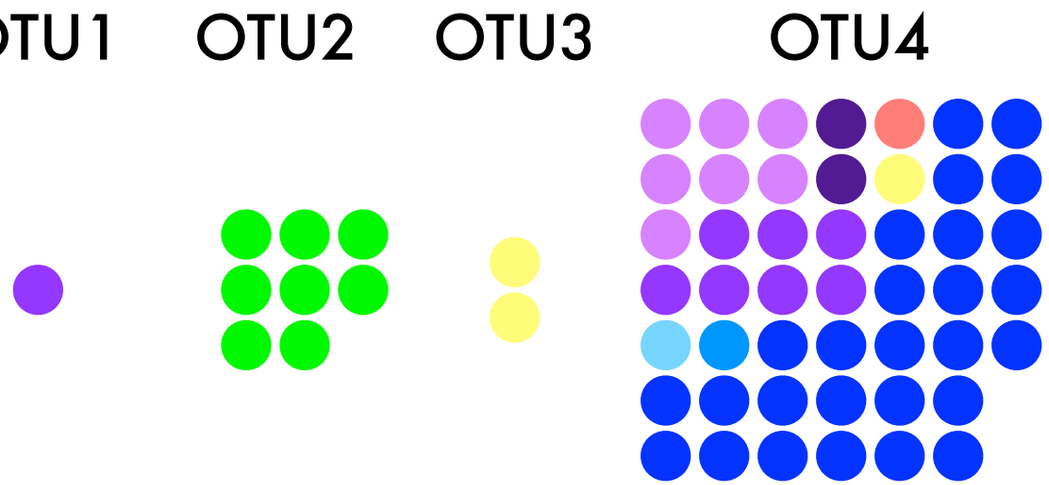
Microbial Community Analysis

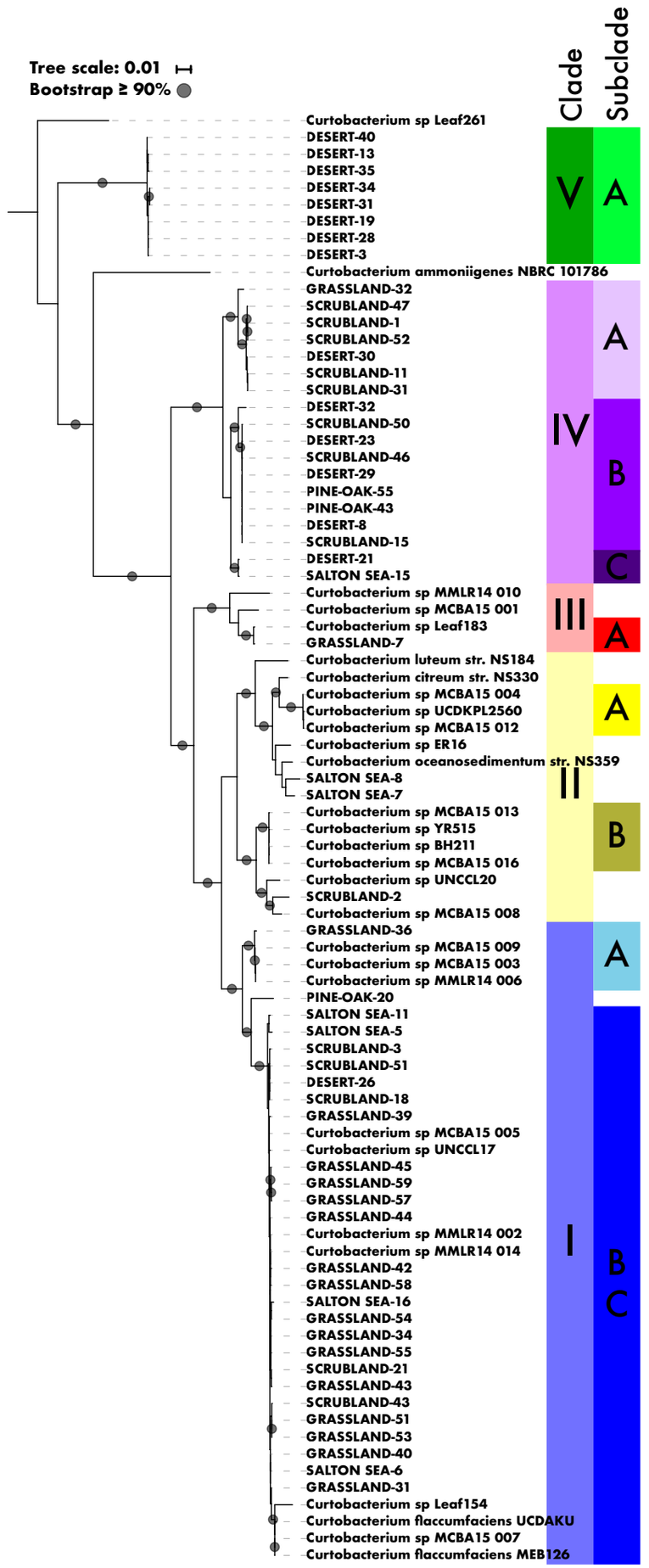
16S rRNA marker gene
Metagenomics

16S Full Length - Aligned
99% similarity



16S v4-v5 Region
100% similarity



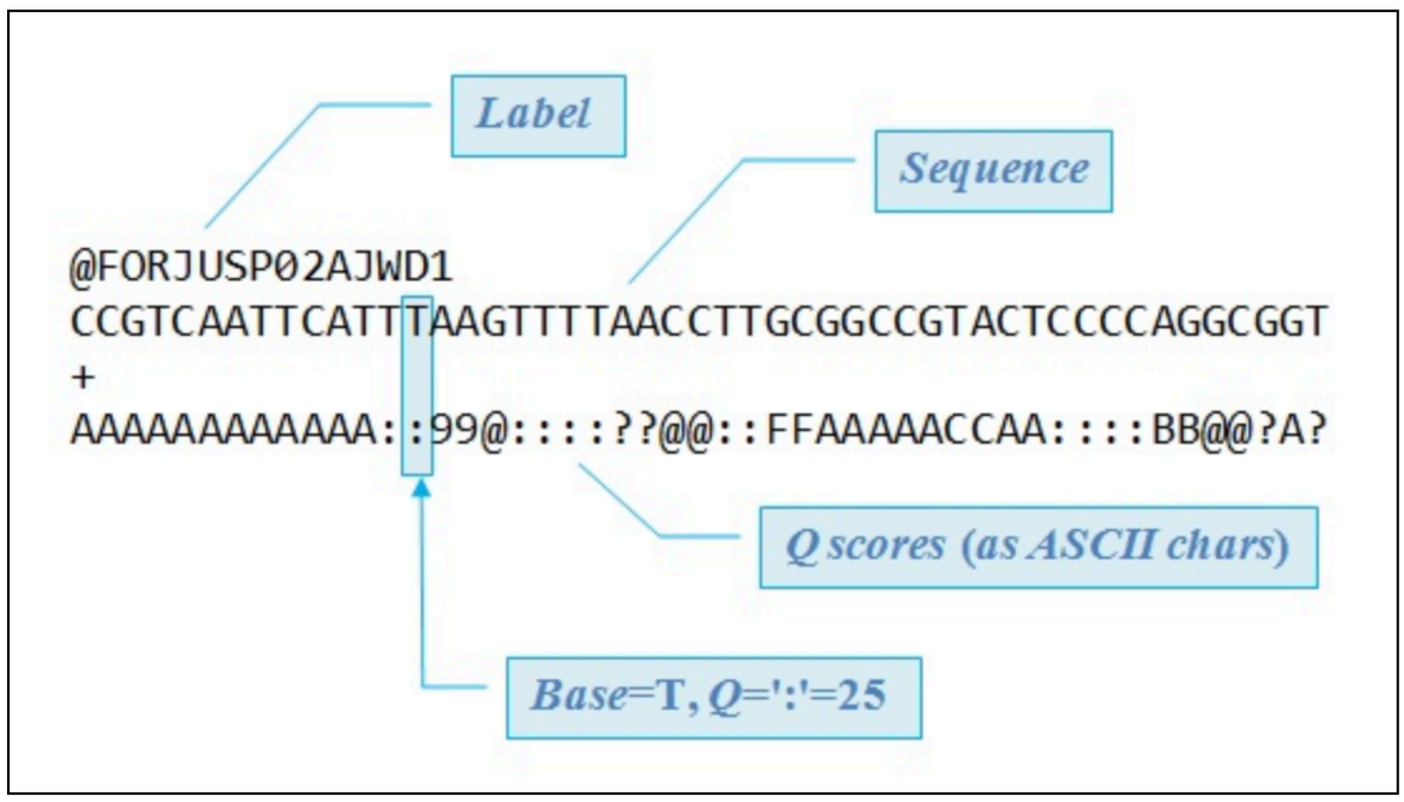


Chase et al. Environ. Microb. 2018

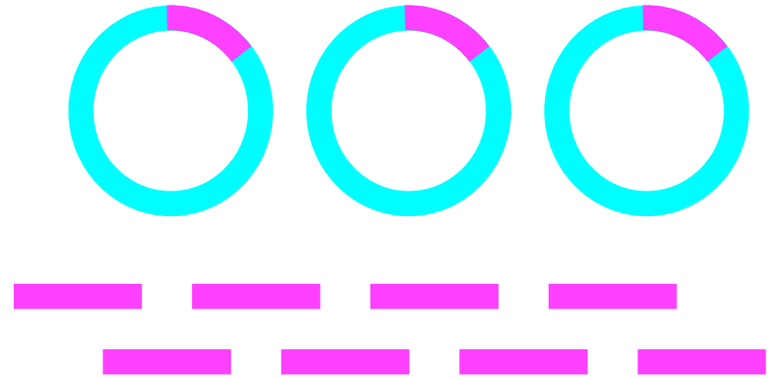
DATA

Fastq files (*.fastq , *.fq , *.fq.gz , *.txt.gz)

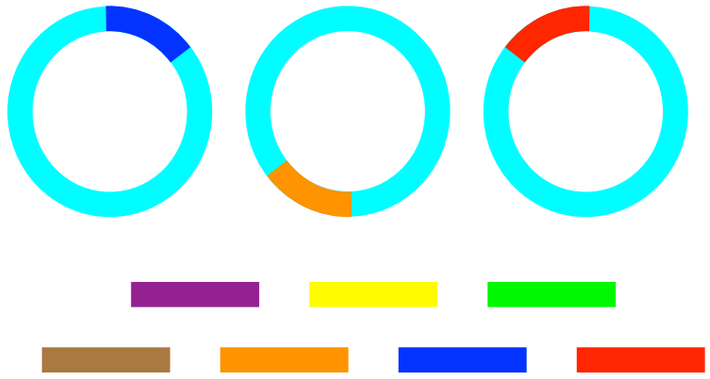
Example:
mR049-L1-READ1-Sequences.txt.gz
mR049-L1-READ2-Sequences.txt.gz



Amplicon (16S rRNA)



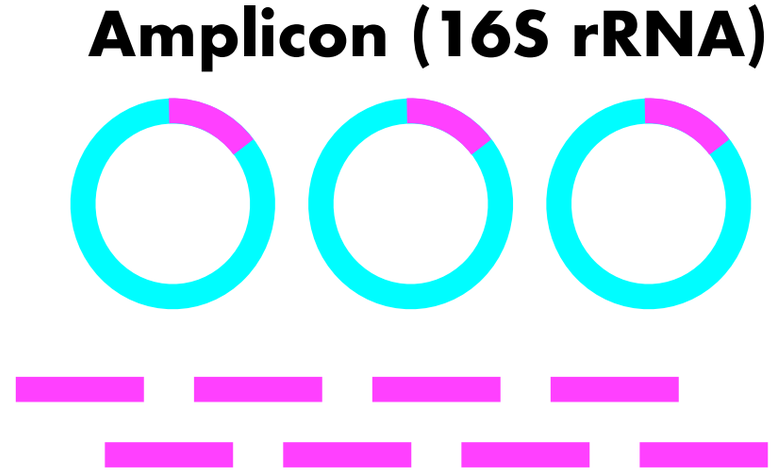
Metagenomics

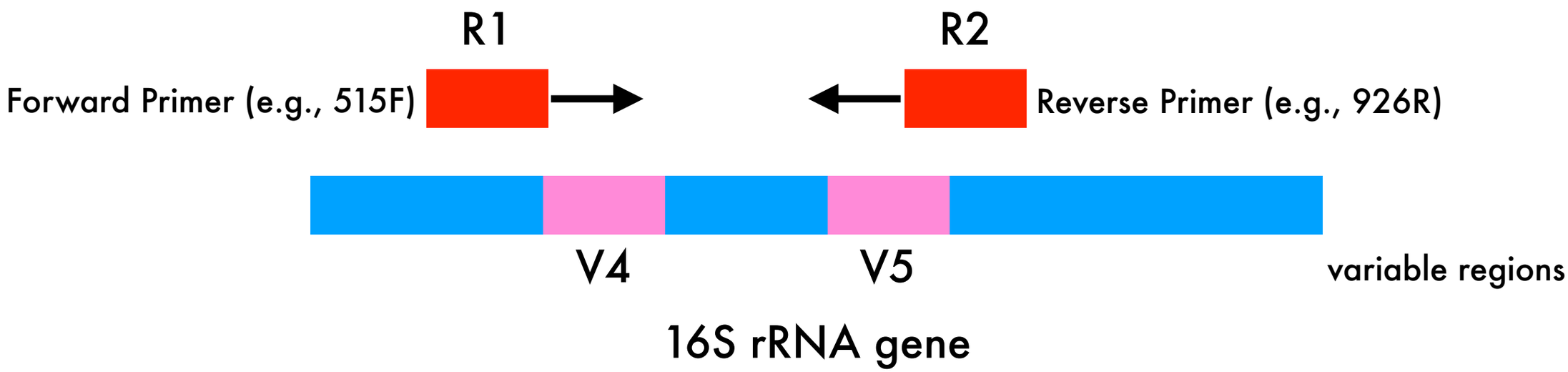


DATA

Fastq files (*.fastq , *.fq , *.fq.gz , *.txt.gz)

Example:
mR049-L1-READ1-Sequences.txt.gz
mR049-L1-READ2-Sequences.txt.gz

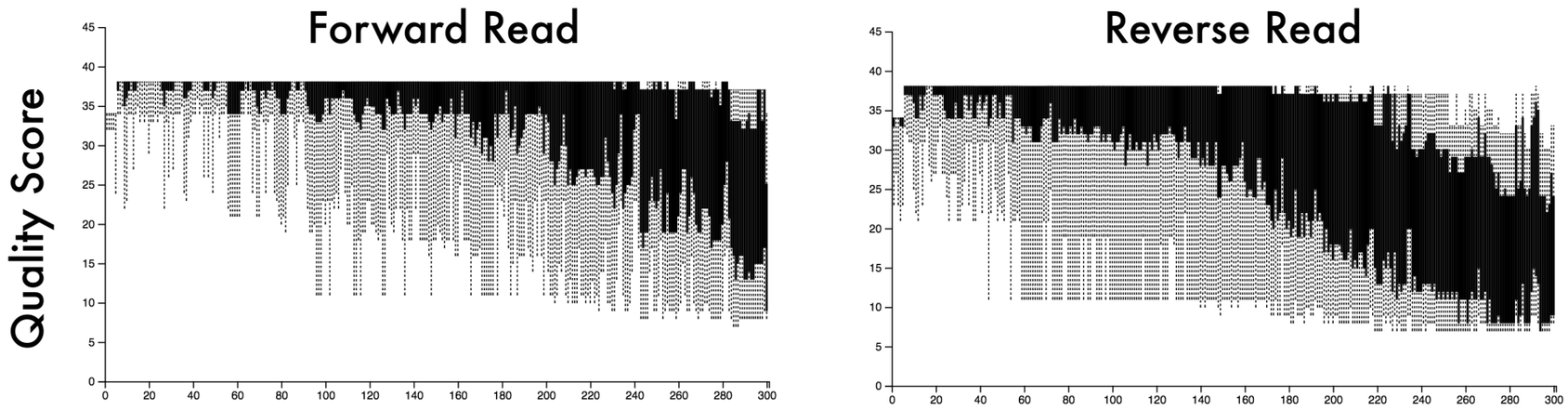


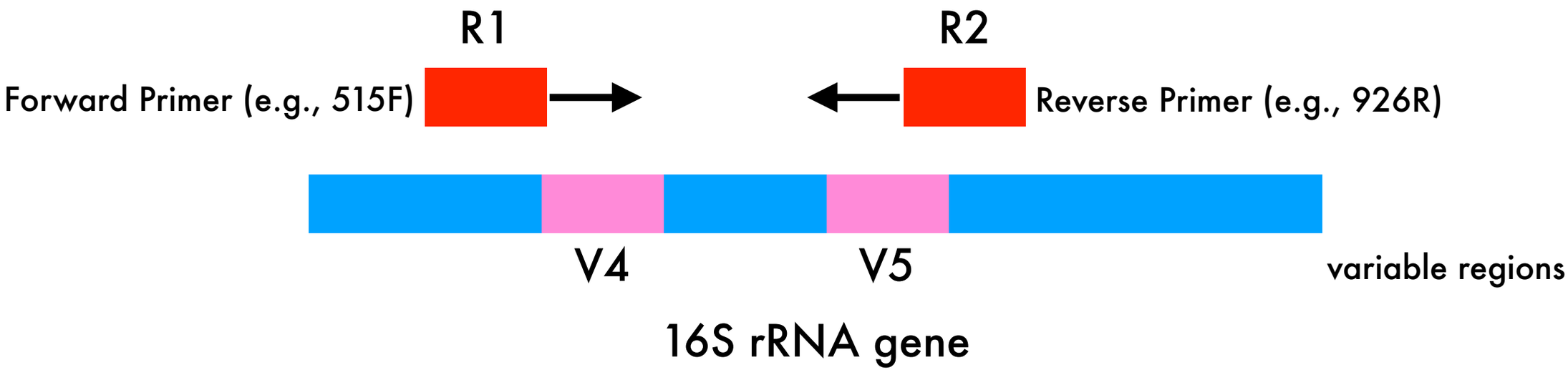


DATA

Fastq files (*.fastq , *.fq , *.fq.gz , *.txt.gz)

Example:
mR049-L1-READ1-Sequences.txt.gz
mR049-L1-READ2-Sequences.txt.gz





Sample 1

R1

R2

Forward Primer (e.g., 515F)

Reverse Primer (e.g., 926R)

V4

V5

variable regions

16S rRNA gene

Sample 2

R1

R2

Forward Primer (e.g., 515F)

Reverse Primer (e.g., 926R)

V4

V5

variable regions

16S rRNA gene

Sample N

R1

R2

Forward Primer (e.g., 515F)

Reverse Primer (e.g., 926R)

V4

V5

variable regions

16S rRNA gene

Sample 1

Sample 2

Sample N

Pooled Library

"Next-generation" Sequencing

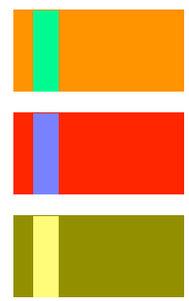
Data Processing

QIIME2
DADA2
Mothur
BBMap (metagenomes)
Anvi'o (metagenomes)

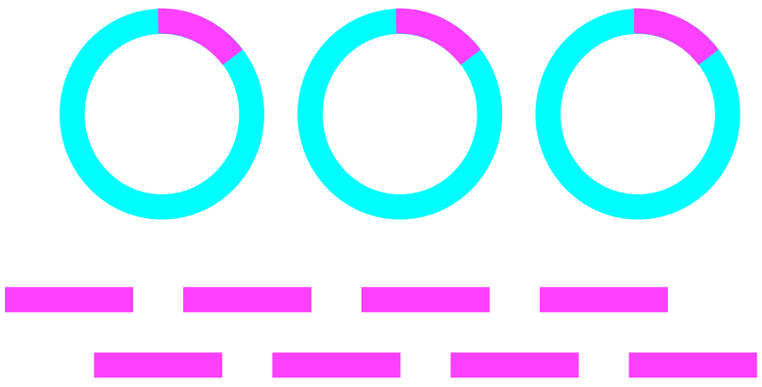
Data Processing

QIIME2 - Knight Lab UCSD
DADA2 - R-based pipeline
Mothur
BBMap (metagenomes)
Anvi'o (metagenomes)

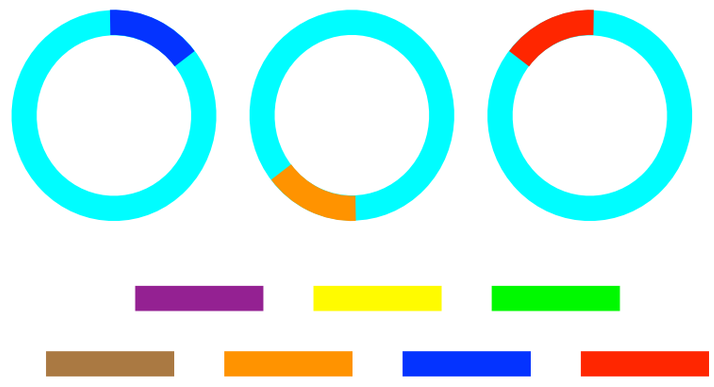
1. Demultiplex
2. Denoise (QC filtering, trimming)



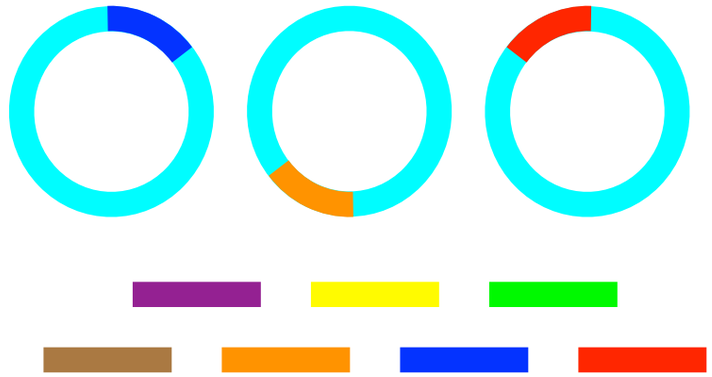
1. Clustering (OTU or ESV or ASV)
2. Feature Table
3. Community Analysis



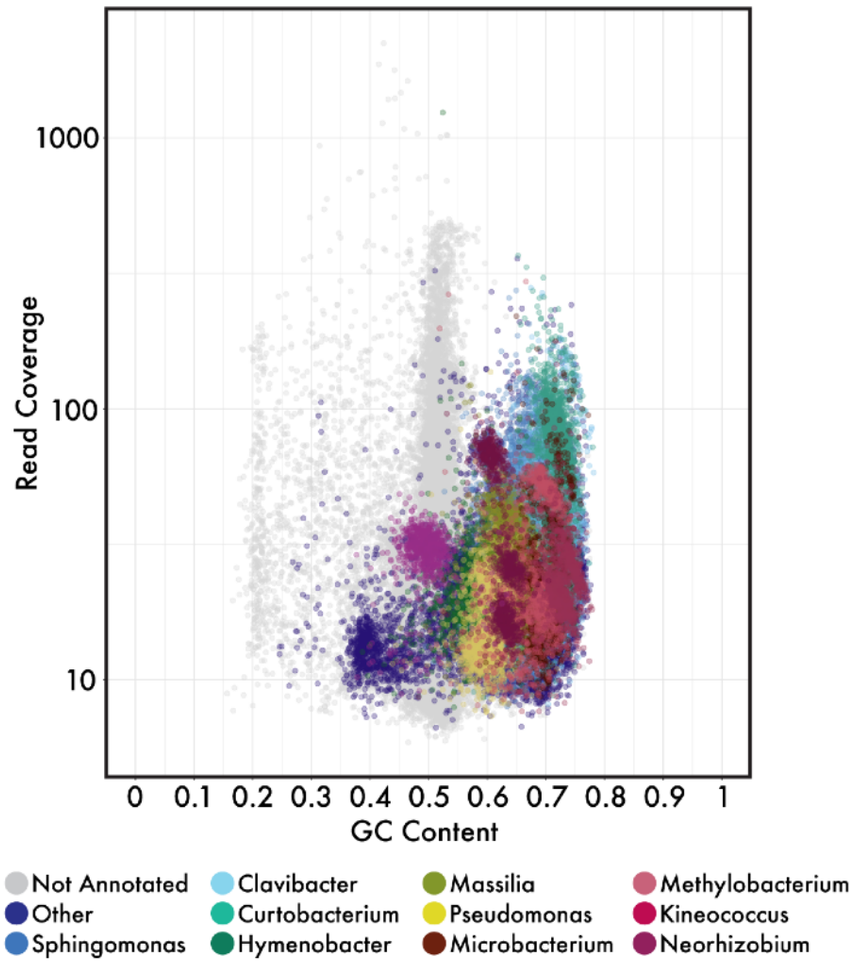
1. MAGs
    1. Assembly (MEGAHIT, metaSPAdes)
    2. Read mapping (bowtie, bwa)
    3. Binning (MetaBAT, CONCOCT)
    4. Bin Curation
2. Read-based Analysis
    1. Community Analysis (MIDAS)
    2. Functional genes (MetaQUBIC)



1. MAGs
    1. Assembly (MEGAHIT, metaSPAdes)
    2. Read mapping (bowtie, bwa)
    3. Binning (MetaBAT, CONCOCT)
    4. Bin Curation
2. Read-based Analysis
    1. Community Analysis
    2. Functional genes

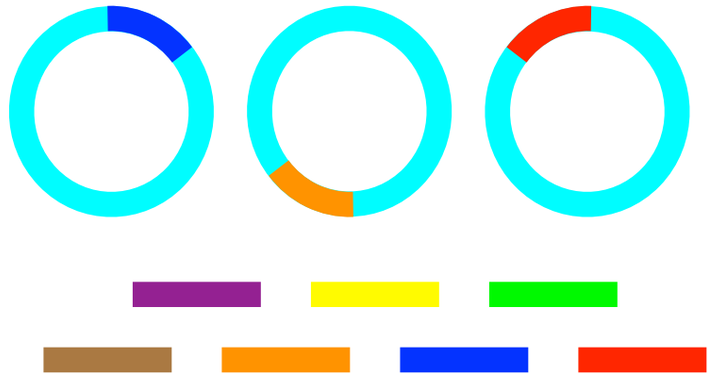


MAGs can be problematic and messy
in complex communities (e.g., soil/sediments)

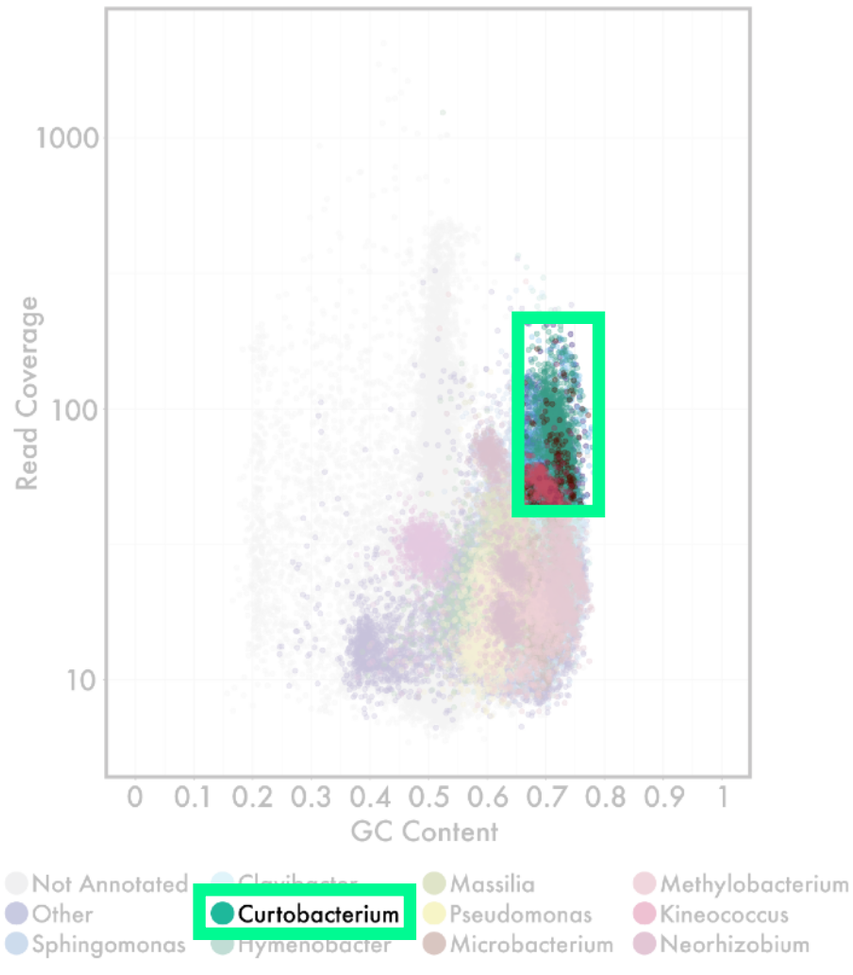


Chase et al. mBio. 2017

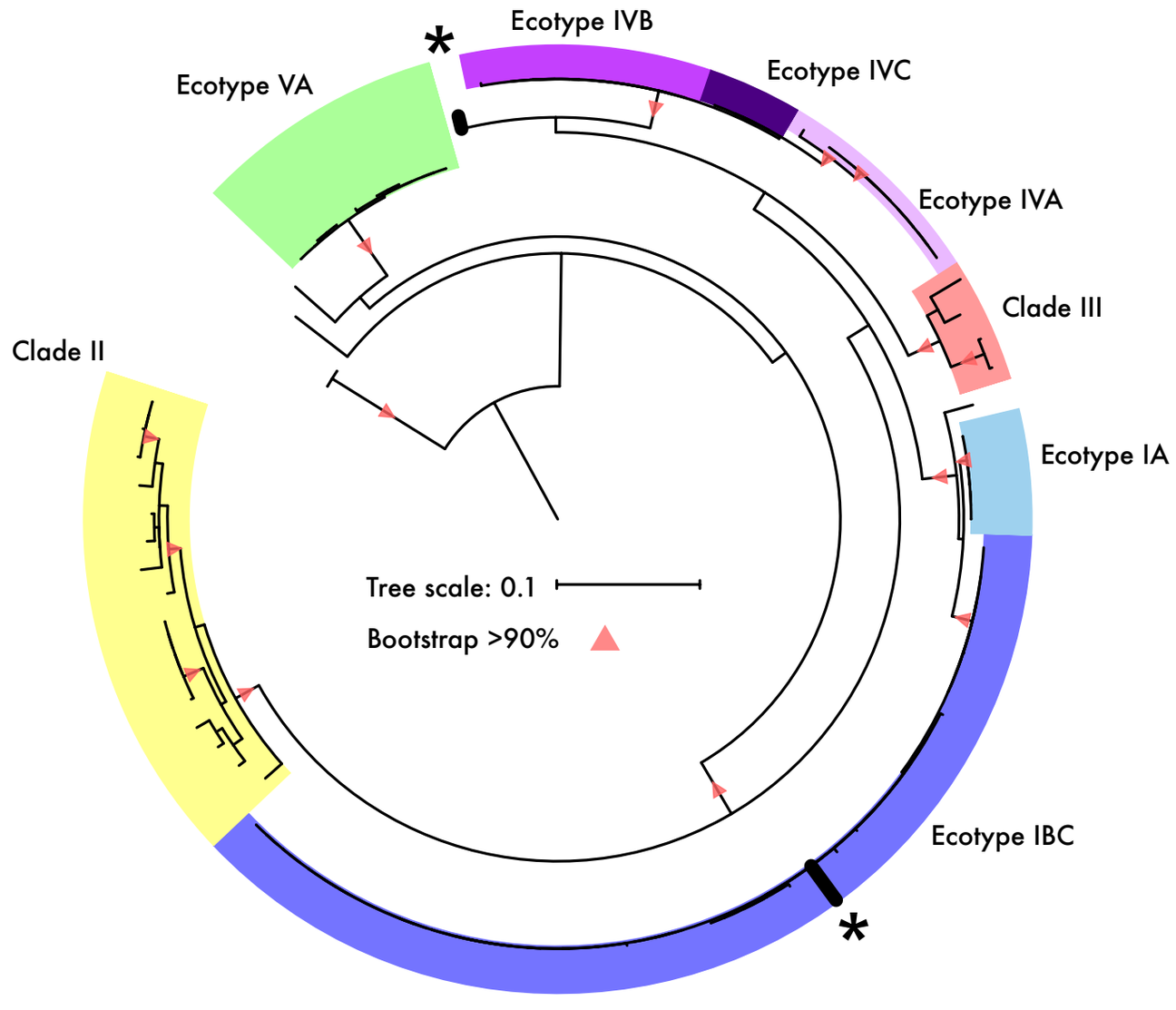
1. MAGs
    1. Assembly (MEGAHIT, metaSPAdes)
    2. Read mapping (bowtie, bwa)
    3. Binning (MetaBAT, CONCOCT)
    4. Bin Curation
2. Read-based Analysis
    1. Community Analysis
    2. Functional genes



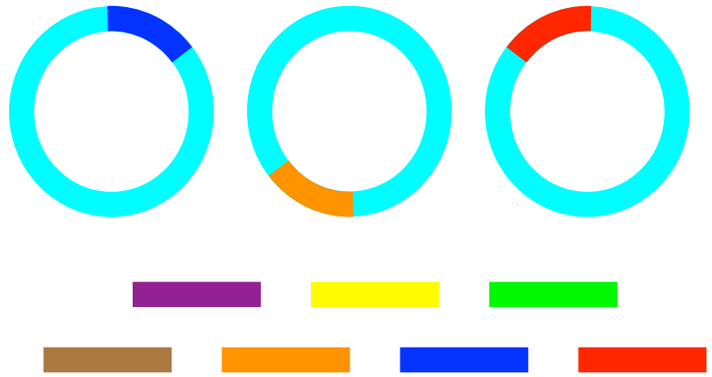
MAGs can be problematic and messy
in complex communities (e.g., soil/sediments)



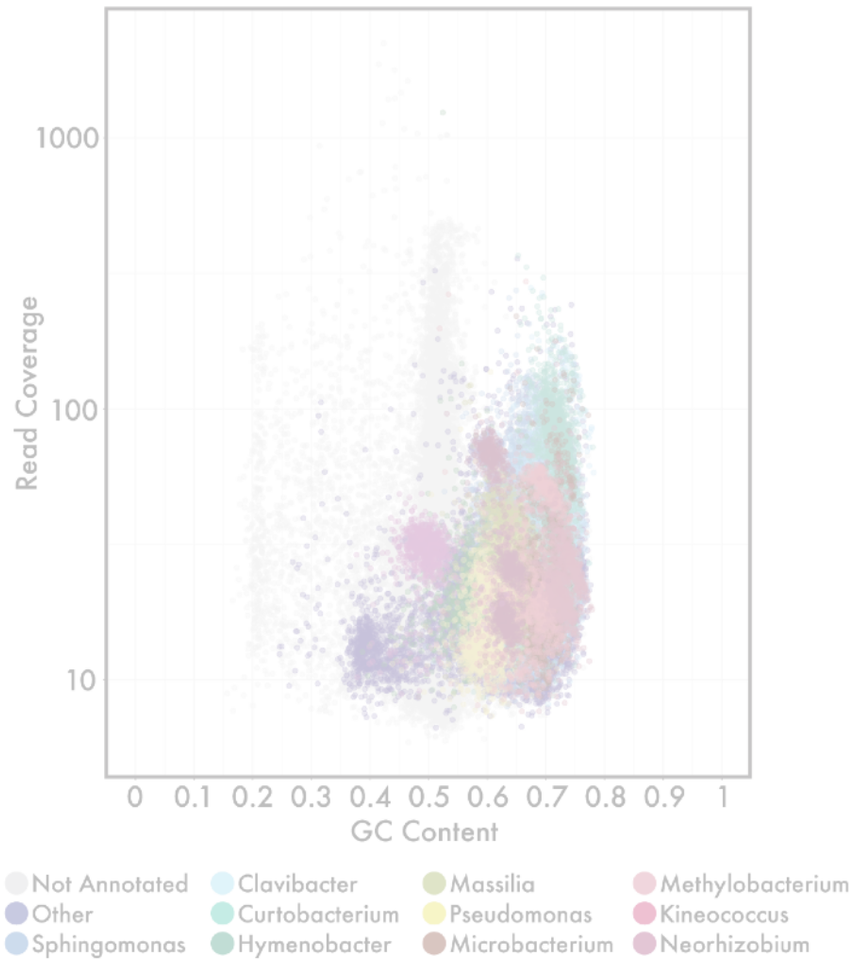
Chase et al. mBio. 2017



1. MAGs
    1. Assembly (MEGAHIT, metaSPAdes)
    2. Read mapping (bowtie, bwa)
    3. Binning (MetaBAT, CONCOCT)
    4. Bin Curation
2. Read-based Analysis
    1. Community Analysis
    2. Functional genes

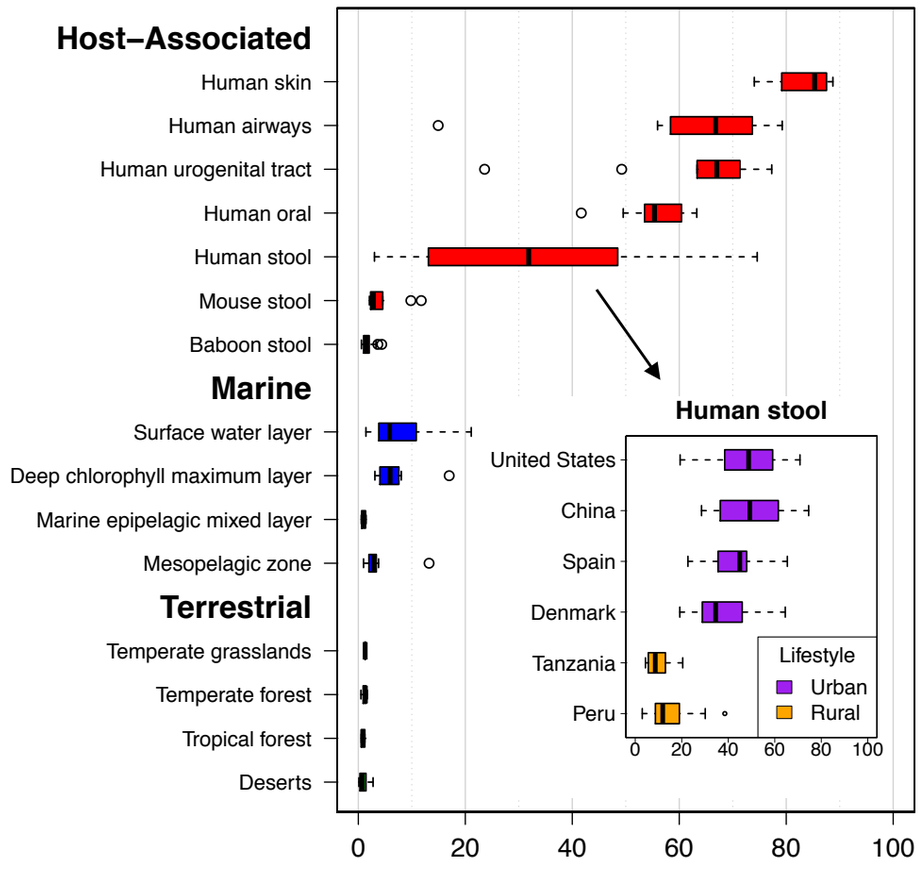


MAGs can be problematic and messy
in complex communities (e.g., soil/sediments)



Chase et al. mBio. 2017

Low representation in genomic databases



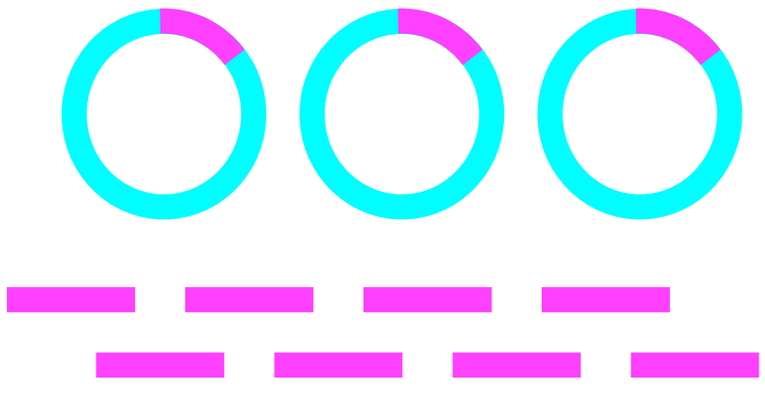
Nayfach et al. Genome Research. 2016

https://benjjneb.github.io/dada2/tutorial.html

https://docs.qiime2.org/2019.7/tutorials/

1. Demultiplex
2. Denoise (QC filtering, trimming)

1. Clustering (OTU or ESV or ASV)
2. Feature Table
3. Community Analysis

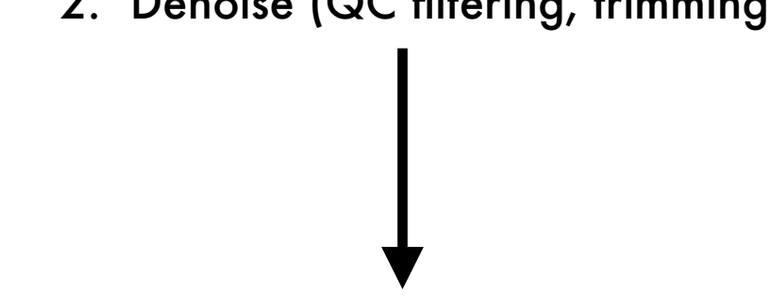


https://benjjneb.github.io/dada2/tutorial.html

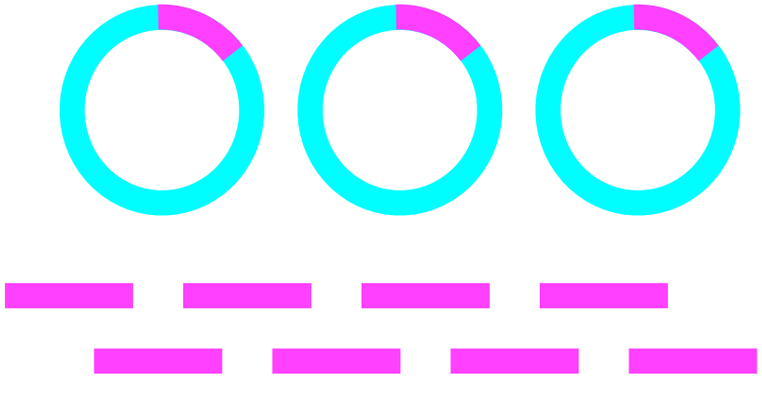
https://docs.qiime2.org/2019.7/tutorials/

"dada2 to denoise my pair-ended reads in qiime2-2019.7 and R (dada2 version, 1.12.1), and got 3533 and 2535 features in the raw ASV table, respectively"

1. Demultiplex
2. Denoise (QC filtering, trimming)



1. Clustering (OTU or ESV or ASV)
2. Feature Table
3. Community Analysis



DADA2 Amplicon Sequencing. Exactly. Version 1.12

https://benjjneb.github.io/dada2/tutorial.html

qiime2

https://docs.qiime2.org/2019.7/tutorials/

"dada2 to denoise my pair-ended reads in qiime2-2019.7 and R (dada2 version, 1.12.1), and got 3533 and 2535 features in the raw ASV table, respectively"

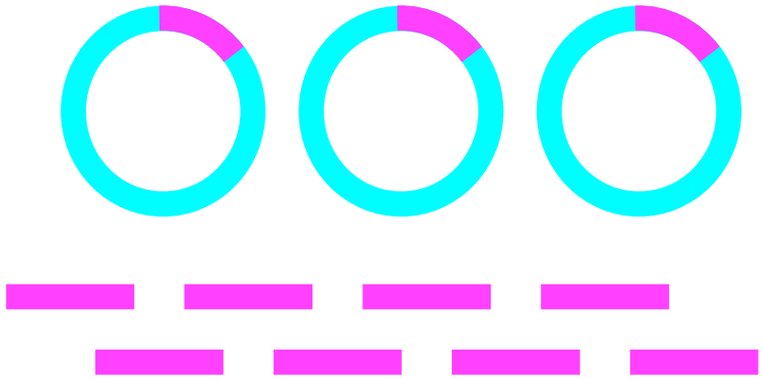
1. Demultiplex
2. Denoise (QC filtering, trimming)

1. Clustering (OTU or ESV or ASV)
2. Feature Table
3. Community Analysis

```
qiime demux emp-paired \
  --m-barcodes-file sample-metadata.tsv \
  --m-barcodes-column barcode-sequence \
  --p-rev-comp-mapping-barcodes \
  --i-seqs emp-paired-end-sequences.qza \
  --o-per-sample-sequences demux.qza \
  --o-error-correction-details demux-details.qza

qiime demux summarize \
  --i-data demux.qza \
  --o-visualization demux.qzv
```

```
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs demux.qza \
  --p-trim-left-f 13 \
  --p-trim-left-r 13 \
  --p-trunc-len-f 150 \
  --p-trunc-len-r 150 \
  --o-table table.qza \
  --o-representative-sequences rep-seqs.qza \
  --o-denoising-stats denoising-stats.qza
```

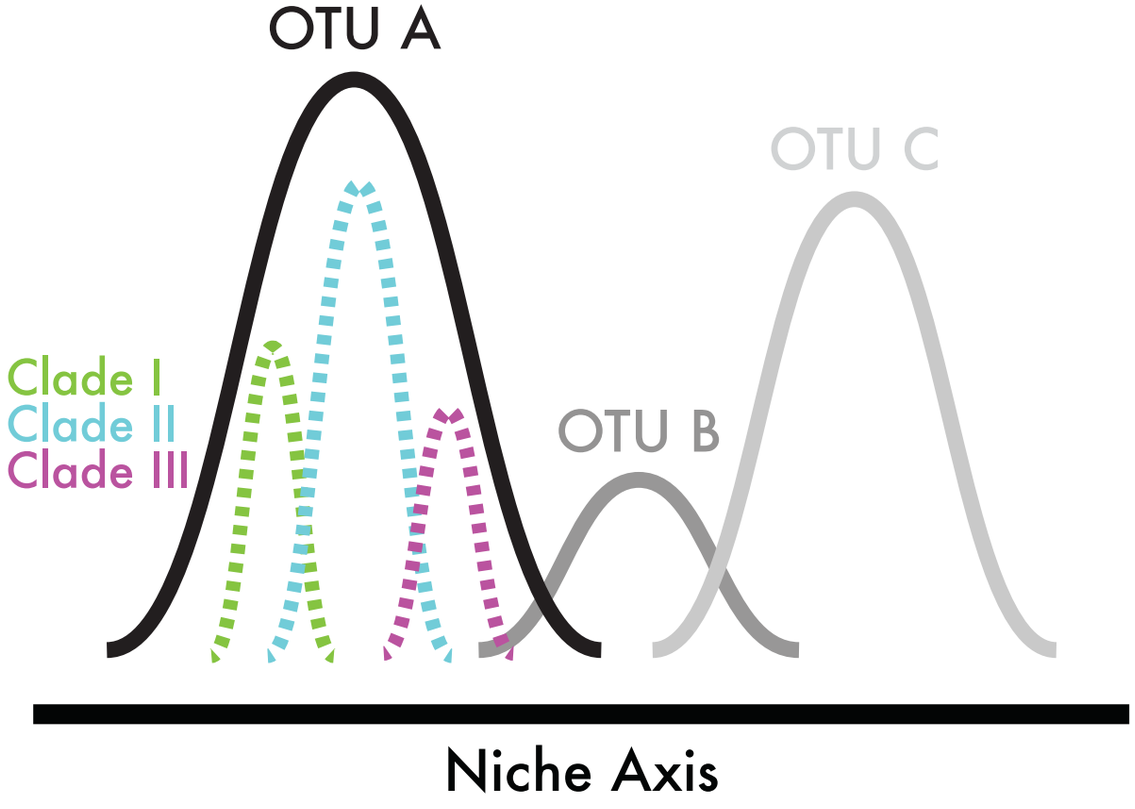


OTU Feature Table

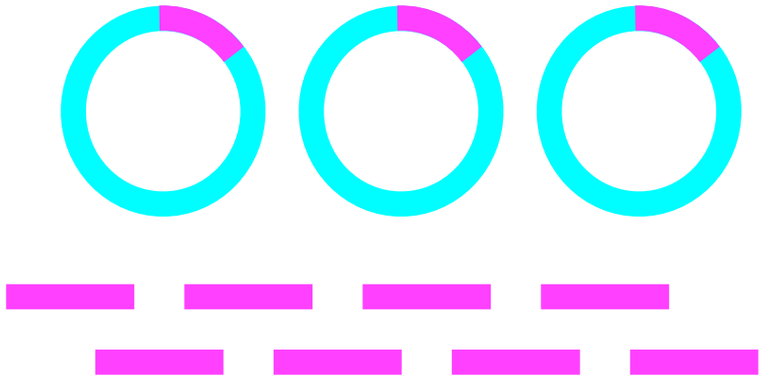
OTU = Operational Taxonomic Unit

Traditionally defined at 97% sequence similarity
Recently defined using exact sequence variants (ESV) or 100% OTUs

"one limitation of the 16S rRNA gene is that it is rather conserved and hence is **NOT** reliable for taxonomic identifiers at the species level" -J. Cole et al. 2010.



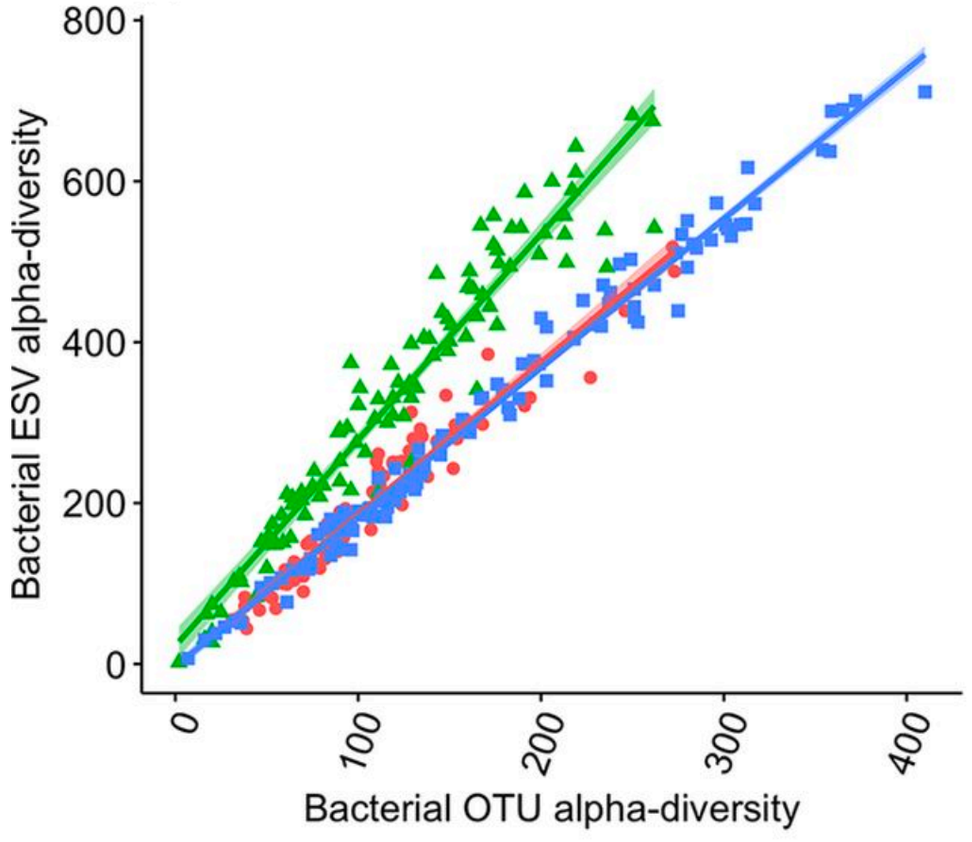
Chase & Martiny. Microbiology Australia. 2018



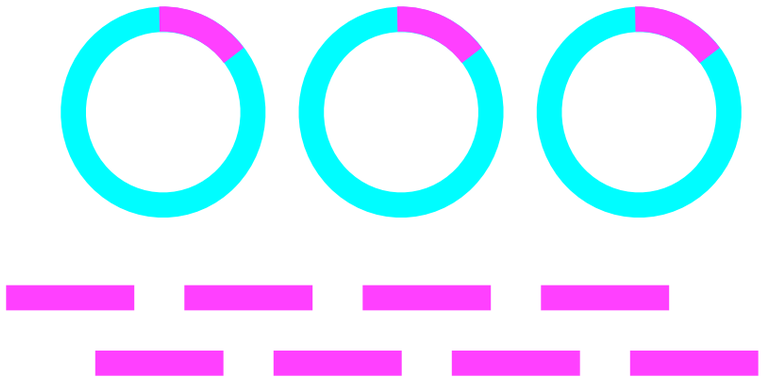
OTU Feature Table

OTU = Operational Taxonomic Unit

Traditionally defined at 97% sequence similarity
Recently defined using exact sequence variants (ESV) or 100% OTUs



Glassman & Martiny. mSphere. 2018



OTU Feature Table

OTU = Operational Taxonomic Unit

Traditionally defined at 97% sequence similarity
Recently defined using exact sequence variants (ESV) or 100% OTUs

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	…	Sample *N*
OTU 1	1	2	0	1	0		1
OTU 2	0	2	0	0	0		0
OTU 3	0	2	1	0	0		0
OTU 4	1	0	1	1	1		1
OTU 5	6	0	0	9	2		2
…							
OTU *i*	0	0	8	5	0		5

Goal - reduce tons and tons of sequence data to OTU table

Fastq files (*.fastq , *.fq , *.fq.gz , *.txt.gz)

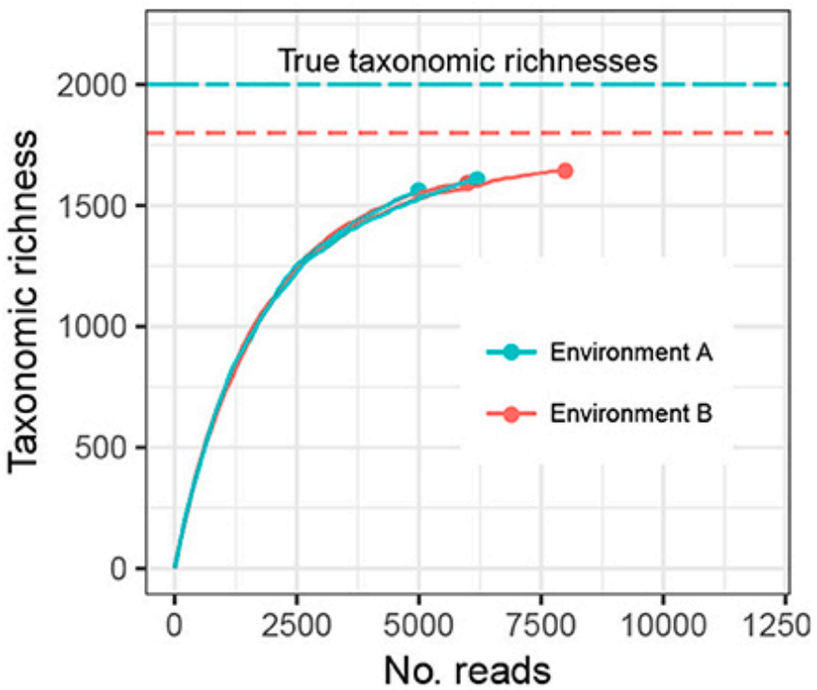
Example:
mR049-L1-READ1-Sequences.txt.gz
mR049-L1-READ2-Sequences.txt.gz

→ OTU Feature Table

Now what?

frontiers in Microbiology

PERSPECTIVE
published: 23 October 2019
doi: 10.3389/fmicb.2019.02407



Rarefaction, Alpha Diversity, and Statistics

*Amy D. Willis**

Department of Biostatistics, University of Washington, Seattle, WA, United States

Goal - reduce tons and tons of sequence data to OTU table

Fastq files (*.fastq , *.fq , *.fq.gz , *.txt.gz)

Example:
mR049-L1-READ1-Sequences.txt.gz
mR049-L1-READ2-Sequences.txt.gz

→ OTU Feature Table

Now what?

R - programming language for statistical computation

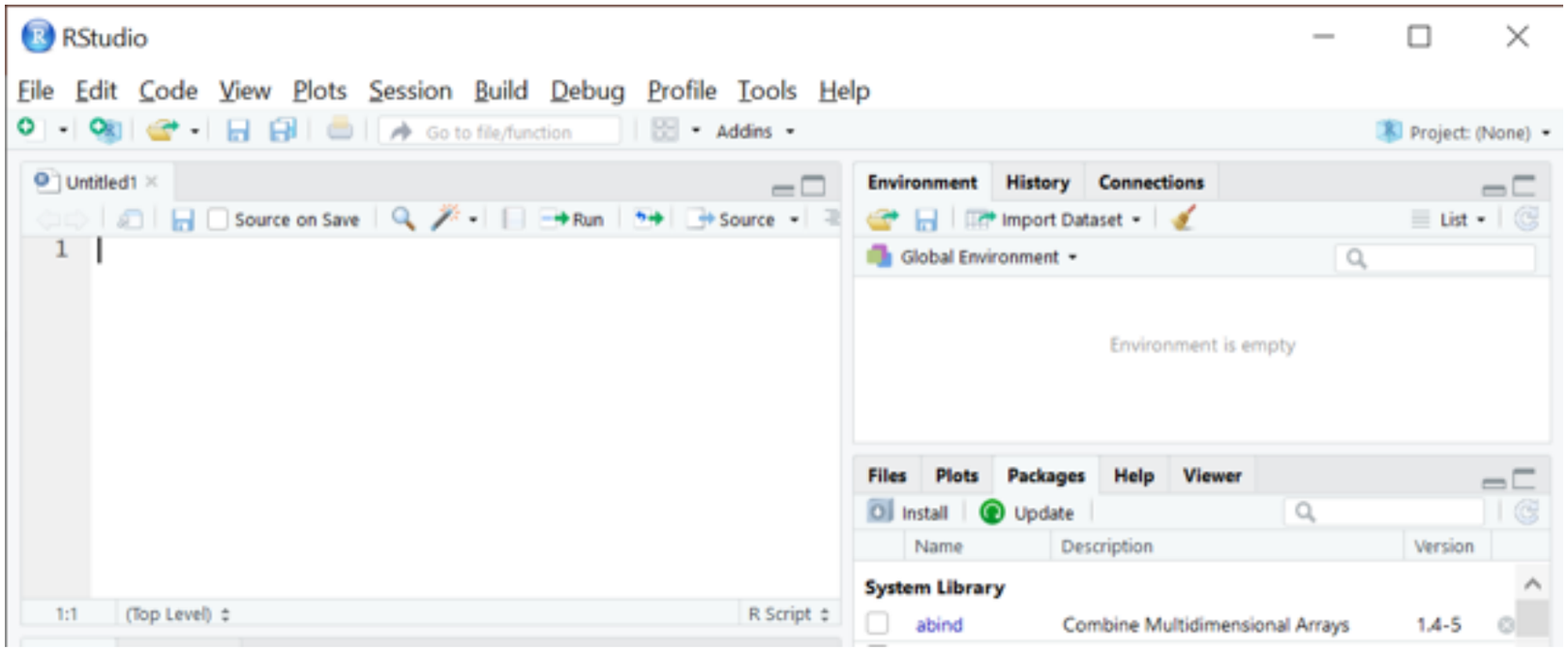
Benefits:

1. Robust - tons of user developed packages for data analysis
2. Reproducible - anyone with data and code can generate same results
3. Free and open-source - friendly and helpful user community
4. Publication quality figures are easy to generate

R - programming language for statistical computation

Benefits:

1. Robust - tons of user developed packages for data analysis
2. Reproducible - anyone with data and code can generate same results
3. Free and open-source - friendly and helpful user community
4. Publication quality figures are easy to generate



R - programming language for statistical computation

Benefits:

1. Robust - tons of user developed packages for data analysis
2. Reproducible - anyone with data and code can generate same results
3. Free and open-source - friendly and helpful user community
4. Publication quality figures are easy to generate