Amplicon Bioinformatic Analysis: DADA2

Josh Granek

Outline

Bioinformatic Goals

Get Data (pre-DADA2)

Validate Data (pre-DADA2)

Assemble Metadata Table (pre-DADA2)

Demultiplex (pre-DADA2)

Adapter Trimming (pre-DADA2)

Filter and Trim

Learn Error Rates

Dereplication

Sample Inference

Merge Paired Reads

Construct Sequence Table

Remove Chimeras

Assign Taxonomy

Generate Phyloseq Object

Save Phyloseq as RDS

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Bioinformatic Analysis

Input: Raw FASTQ File(s)

```
 M00698:36:000000000-AFBEL:1:1101:14738:1412 1:N:0:0
TTACGCTAACAGGCGGTAGCCTGGCAGGGTCAGGAAATCAATTAACTCATCGGAAGTGGTGATCTGTTCCATCAAGCGTGCGGCATCGTCA
+
ABBBABBBAFFFGGGGGGGGGGHGGHGGGCG2GF3FFGHHHHHHGGFGHEHHGGGEHHHHAGGHHGHHHFFDHFHHHGEGGGG@F@H?GHH
@M00698:36:000000000-AFBEL:1:1101:16483:1412 1:N:0:0
CTGCCAGTTGAACGACGGCGAGCAGTTATAAGCCAGCAGTTTGCCCGGATATTTCGCGTGGATAGCTTGTGCAAAGCGACGCGCCAGTTCC
+
AAABBFFFFFFFGGGGGGGGGGGGHHHHHHHHGHGHGHHHHHGHHHGGGGGHHHHGGGGGGGHHHGHHFFHHHHHGHGGGGGGGGGGHHHH
```

Output: Count Table

	Sample 1	Sample 2	...	Sample N
Bacteria 1	0	0		64
Bacteria 2	72	5		0
...				
Bacteria N	0	43		0

Naive Approach: Assumptions

- Library Prep is Perfect
- Sequencing is Perfect

Naive Approach: Counting

Naive Approach: Counting

1. Make an empty count table

Naive Approach: Counting

1. Make an empty count table
2. For each read in the FASTQ:

Naive Approach: Counting

1. Make an empty count table
2. For each read in the FASTQ:
    2.1 If read sequence is already in count table, add 1 to that row

Naive Approach: Counting

1. Make an empty count table
2. For each read in the FASTQ:
    2.1 If read sequence is already in count table, add 1 to that row
    2.2 Otherwise add a new row for the sequence and set its count to 1

Naive Approach: Counting Demo

Sequence	Count

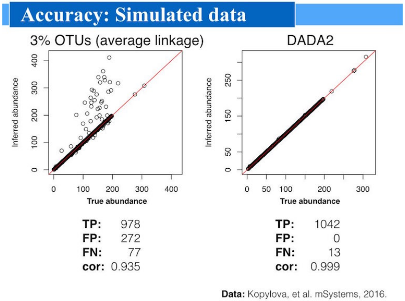
1. CAGCT
2. TATAA
3. TATAA
4. TGCGC
5. CGGGC
6. TGCGC
7. TGCGC
8. CAGCT
9. CGGGC
10. TGCGC

Naive Assumptions

- ~~Library Prep is Perfect~~
- ~~Sequencing is Perfect~~

Tools for Bioinformatic Analysis

- "Clustering"
  - Mothur
  - UCLUST
  - UPARSE
- "Denoising"
  - DADA2
  - UNOISE3
  - Deblur



Accuracy: Simulated data

3% OTUs (average linkage)

TP:	978
FP:	272
FN:	77
cor:	0.935

DADA2

TP:	1042
FP:	0
FN:	13
cor:	0.999

Data: Kopylova, et al. mSystems, 2016.

DADA2 Website

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Get Data: Sources

- Sequence Read Archive (SRA)
- MG-RAST (Metagenomic Rapid Annotations using Subsystems Technology)
- Sequencing Facility

Get Data: Tools

- curl
- wget
- ncftp
- rsync
- sftp
- SRA Toolkit

Get Data: Result

- FASTQ(s) (gzip'ed)
  - Undetermined_S0_L001_I1_001.fastq.gz
  - Undetermined_S0_L001_R1_001.fastq.gz
  - Undetermined_S0_L001_R2_001.fastq.gz
- Map File*
  - mydata_map.txt
- Checksum*
  - md5sum.txt

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Validate Data: Input

- FASTQ(s) (gzip'ed)
  - Undetermined_S0_L001_I1_001.fastq.gz
  - Undetermined_S0_L001_R1_001.fastq.gz
  - Undetermined_S0_L001_R2_001.fastq.gz
- Checksum*
  - md5sum.txt
- Map File*
  - mydata_map.txt

Validate Data: Output

```
$ md5sum -c md5sum.txt
mydata_map.txt: OK
Undetermined_S0_L001_I1_001.fastq.gz: OK
Undetermined_S0_L001_R1_001.fastq.gz: OK
Undetermined_S0_L001_R2_001.fastq.gz: OK
```

Validate Data: Tools

- md5sum

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Assemble Metadata Table: Why?

Associate barcode with Sample

- Label
- Animal
- Site
- Phenotype
- Treatment
- Date
- . . .

Assemble Metadata Table: Input

- Existing Map
- Publication
- Notes

Assemble Metadata Table: Output

Metadata Table (Mapping File)

SampleID	BarcodeSequence	Treatment	DOB	Description
PC.354	AGCACGAGCCTA	Control	20061218	Control_mouse__I.D._354
PC.355	AACTCGTCGATG	Control	20061218	Control_mouse__I.D._355
PC.356	ACAGACCACTCA	Control	20061126	Control_mouse__I.D._356
PC.481	ACCAGCGACTAG	Control	20070314	Control_mouse__I.D._481
PC.593	AGCAGCACTTGT	Control	20071210	Control_mouse__I.D._593
PC.607	AACTGTGCGTAC	Fast	20071112	Fasting_mouse__I.D._607
PC.634	ACAGAGTCGGCT	Fast	20080116	Fasting_mouse__I.D._634
PC.635	ACCGCAGAGTCA	Fast	20080116	Fasting_mouse__I.D._635
PC.636	ACGGTGAGTGTC	Fast	20080116	Fasting_mouse__I.D._636

Assemble Metadata Table: Tools

- Excel
- Text Editor
- Script

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Demultiplex: Why?

Split FASTQ File(s) by sample

Some data comes demultiplexed

Demultiplex: Input

- Sequence FASTQ(s)
  - Undetermined_S0_L001_I1_001.fastq.gz
  - Undetermined_S0_L001_R1_001.fastq.gz
- Barcode FASTQ or Trimmed Versions
  - Undetermined_S0_L001_R2_001.fastq.gz
- Map File
  - mydata_map.txt

Some facilities incorporate barcodes in the sequence FASTQ, these will need to be extracted

Demultiplex: Output

Demultiplexed FASTQs

- sampleA_R1.fastq.gz
- sampleB_R1.fastq.gz
- sampleC_R1.fastq.gz
- . . .
- sampleA_R2.fastq.gz
- sampleB_R2.fastq.gz
- sampleC_R2.fastq.gz
- . . .

Demultiplex: Tools

- split_libraries_fastq.py + split_sequence_file_on_sample_ids.py
- fastq_multx

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Adapter Trimming: Why?

Remove adapter contamination

- Necessary for amplicons with large variation in length (e.g. ITS)
- Generally unnecessary for 16S rRNA

Adapter Trimming: Input

Adapter Sequence

my_adapter.fasta

Demultiplexed FASTQs

- sampleA_R1.fastq.gz
- sampleB_R1.fastq.gz
- sampleC_R1.fastq.gz
- . . .
- sampleA_R2.fastq.gz
- sampleB_R2.fastq.gz
- sampleC_R2.fastq.gz
- . . .

Adapter Trimming: Output

Trimmed FASTQs

- sampleA_R1.trim.fastq.gz
- sampleB_R1.trim.fastq.gz
- sampleC_R1.trim.fastq.gz
- . . .
- sampleA_R2.trim.fastq.gz
- sampleB_R2.trim.fastq.gz
- sampleC_R2.trim.fastq.gz
- . . .

Synchronized Trimming

Depending on settings, some reads may be thrown out during trimming. It is essential that if a read is thrown out, its paired read is thrown out too. Most trimming software will do this for you if you input R1 and R2 files when you run.

Adapter Trimming: Tools

- fastq_mcf
- Trimmomatic
- cutadapt
- seqtk
- etc

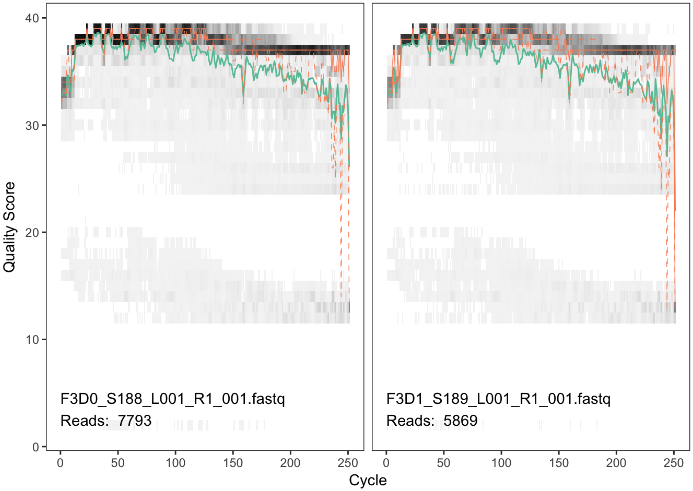
Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

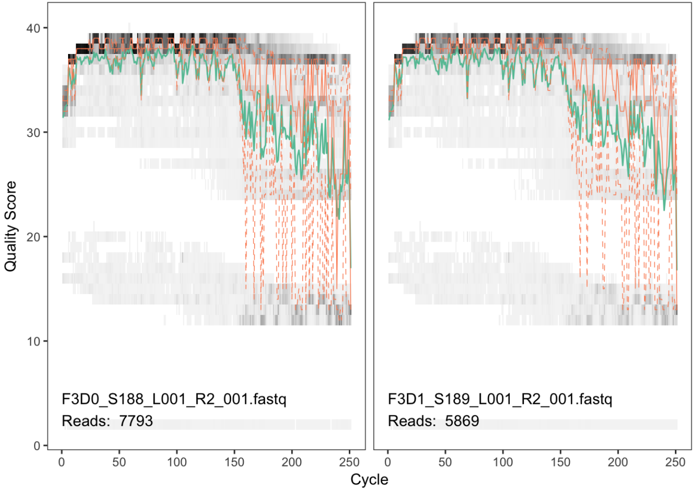
Filter and Trim: Why?

- Remove low quality parts of reads
- Remove reads that are low quality overall
- *Remove adapter contamination*

R1 Read Quality



R2 Read Quality



Filter and Trim: Input

Trimmed FASTQs (or Demultiplexed)

- sampleA_R1.trim.fastq.gz
- sampleB_R1.trim.fastq.gz
- sampleC_R1.trim.fastq.gz
- . . .
- sampleA_R2.trim.fastq.gz
- sampleB_R2.trim.fastq.gz
- sampleC_R2.trim.fastq.gz
- . . .

Filter and Trim: Output

Trimmed and filtered FASTQs

Filter and Trim: Tools

dada2::filterAndTrim()

Filter and Trim: Parameters

Filter and Trim: Parameters

- truncQ: Truncate reads at the first instance of a quality score less than or equal to truncQ.

Filter and Trim: Parameters

- truncQ: Truncate reads at the first instance of a quality score less than or equal to truncQ.
- truncLen: Truncate reads after truncLen bases. Don't use for ITS

Filter and Trim: Parameters

- truncQ: Truncate reads at the first instance of a quality score less than or equal to truncQ.
- truncLen: Truncate reads after truncLen bases. Don't use for ITS
- trimLeft: The number of nucleotides to remove from the start of each read.

Filter and Trim: Parameters

- truncQ: Truncate reads at the first instance of a quality score less than or equal to truncQ.
- truncLen: Truncate reads after truncLen bases. Don't use for ITS
- trimLeft: The number of nucleotides to remove from the start of each read.
- minQ: After truncation, reads contain a quality score less than minQ will be discarded.

Filter and Trim: Parameters

- truncQ: Truncate reads at the first instance of a quality score less than or equal to truncQ.
- truncLen: Truncate reads after truncLen bases. Don't use for ITS
- trimLeft: The number of nucleotides to remove from the start of each read.
- minQ: After truncation, reads contain a quality score less than minQ will be discarded.
- maxEE: After truncation, reads with higher than maxEE "expected errors" will be discarded.

  ```
  EE = sum(10^(-Q/10))
  ```

Filter and Trim: Parameters

- truncQ: Truncate reads at the first instance of a quality score less than or equal to truncQ.
- truncLen: Truncate reads after truncLen bases. Don't use for ITS
- trimLeft: The number of nucleotides to remove from the start of each read.
- minQ: After truncation, reads contain a quality score less than minQ will be discarded.
- maxEE: After truncation, reads with higher than maxEE "expected errors" will be discarded.
  `EE = sum(10^(-Q/10))`
- rm.phix: Discard reads that match against the phiX genome

Filter and Trim: Notes

Paired-End Reads need to be run simultaneously to keep them in sync

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Learn Error Rates: Why?

Build an error model from data

Phred	A:A	A:T	A:C	A:G	C:A	. . .	G:G
1	?	?	?	?	?	. . .	?
2	?	?	?	?	?	. . .	?
3	?	?	?	?	?	. . .	?
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
40	?	?	?	?	?	. . .	?

Learn Error Rates: Input

Filtered and Trimmed FASTQs

Learn Error Rates: Output

error model

Phred	A:A	A:T	A:C	A:G	...	G:G
1	0.27042	0.23546	0.24245	0.25167	...	0.24492
2	0.27927	0.23248	0.24764	0.24062	...	0.25699
3	0.26260	0.25353	0.24638	0.23749	...	0.23728
...	...	...	...	...	...	...
40	0.99894	0.00022	0.00022	0.00062	...	0.99724

Learn Error Rates: Tools

dada2::learnErrors()

Learn Error Rates: Notes

Separate error models need to be built for R1 and R2

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Dereplication: Why?

Summarize reads into unique observed reads, with quality summary and count

1. CAGCT
2. TATAA
3. TATAA
4. TGCGC
5. CGGGC
6. TGCcC
7. TGCGC
8. CAGCT
9. CGGGa
10. TGCGC

Sequence	Count	Quality
CAGCT	2	99989
TATAA	2	99998
TGCGC	3	99988
CGGGC	1	99999
TGCcC	1	99948
CGGGa	1	99993

Dereplication: Input

Filtered and Trimmed FASTQs

Dereplication: Output

Unique reads with summarized quality and counts

Dereplication: Tools

dada2::derepFastq()

Dereplication: Notes

Dereplication is done separately for R1 and R2

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Sample Inference: Why?

Attempt to determine the true sequences from which reads were derived

Sequence	Count	Quality
CAGCT	2	99989
TATAA	2	99998
TGCGC	3	99988
CGGGC	1	99999
TGCcC	1	99948
CGGGa	1	99993

Sequence	Count
CAGCT	2
TATAA	2
TGCGC	4
CGGGC	2

Sample Inference: Input

- Dereplicated Reads
- Error Model

Sample Inference: Output

Inferred read sequences with counts

Sample Inference: Tools

dada2::dada()

Sample Inference: Notes

Sample Inference is done separately for R1 and R2

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Merge Paired Reads: Why?

Collapse read pairs into a single sequence for each inferred amplicon

```
R1:       ATACCCTAGTGC
R2:          CCCTAGTGCCGT

Merged:   ATACCCTAGTGCCGT
```

Merge Paired Reads: Input

- R1
  - Inferred Sequences
  - Dereplicated Sequences
- R2
  - Inferred Sequences
  - Dereplicated Sequences

Merge Paired Reads: Output

Inferred amplicon sequences

Merge Paired Reads: Tools

dada2::mergePairs()

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- **Construct Sequence Table**
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Construct Sequence Table: Why?

Generate count table

	Sample 1	Sample 2	. . .	Sample N
Bacteria 1	0	0		64
Bacteria 2	72	5		0
. . .				
Bacteria N	0	43		0

Construct Sequence Table: Input

Merged sequences

Construct Sequence Table: Output

Count table

Construct Sequence Table: Tools

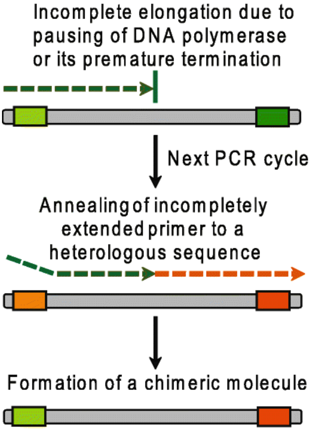
dada2::makeSequenceTable()

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Remove Chimeras: Why?

Library preparation is imperfect, so it generates chimeric amplicons


Incomplete elongation due to pausing of DNA polymerase or its premature termination
Next PCR cycle
Annealing of incompletely extended primer to a heterologous sequence
Formation of a chimeric molecule

Omelina, E.S., et al. BMC Genomics 20, 536 (2019)

Remove Chimeras: Input

Count Table

Remove Chimeras: Output

Count Table without chimeras

Remove Chimeras: Tools

dada2::removeBimeraDenovo()

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Assign Taxonomy: Why?

Relate sequences in our count table to specific bacteria

Assign Taxonomy: Input

Chimera-free merged sequences

Assign Taxonomy: Output

Mapping from sequences to specific bacteria

	Kingdom	Phylum	Class
Bacteria 1	Bacteria	Actinobacteria	Rubrobacteria
Bacteria 2	Bacteria	Gemmatimonadetes	Gemmatimonadetes
Bacteria 3	Bacteria	Actinobacteria	Actinobacteria
Bacteria 4	Bacteria	Proteobacteria	Alphaproteobacteria
Bacteria 5	Bacteria	Firmicutes	Bacilli
Bacteria 6	Bacteria	Actinobacteria	Thermoleophilia
Bacteria 7	Bacteria	Actinobacteria	Actinobacteria
Bacteria 8	Bacteria	Proteobacteria	Gammaproteobacter
Bacteria 9	Bacteria	Actinobacteria	Actinobacteria
Bacteria 10	Bacteria	Proteobacteria	Alphaproteobacteria
Bacteria 11	Bacteria	Actinobacteria	Nitriliruptoria
...			
Bacteria N	Bacteria	Proteobacteria	Gammaproteobacter

Assign Taxonomy: Tools

dada2::assignTaxonomy()

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- **Generate Phyloseq Object**
- Save Phyloseq as RDS

Generate Phyloseq Object: Why?

Phyloseq objects organize multiple aspects of our results and ease downstream analysis and visualization

Generate Phyloseq Object: Input

- Count Table
- Metadata Table
- Taxonomic Assignment
- Phylogenetic Tree (optional)

Generate Phyloseq Object: Output

Phyloseq Object

Generate Phyloseq Object: Tools

phyloseq::phyloseq()

Topic

- Bioinformatic Goals
- Get Data (pre-DADA2)
- Validate Data (pre-DADA2)
- Assemble Metadata Table (pre-DADA2)
- Demultiplex (pre-DADA2)
- Adapter Trimming (pre-DADA2)
- Filter and Trim
- Learn Error Rates
- Dereplication
- Sample Inference
- Merge Paired Reads
- Construct Sequence Table
- Remove Chimeras
- Assign Taxonomy
- Generate Phyloseq Object
- Save Phyloseq as RDS

Save Phyloseq as RDS: Why?

- Generating the final phyloseq object from raw FASTQs is time consuming, we would prefer to not repeat it everytime we want to play with the results
- The Phyloseq object is a very space efficient representation of the processed data

Save Phyloseq as RDS: Input

- Phyloseq object
- Name for RDS file

Save Phyloseq as RDS: Output

RDS file

Save Phyloseq as RDS: Tools

readr::write_rds()