



Always was, always will be...

I acknowledge the Traditional Owners of the lands on which the samples were collected and where we're running this workshop including the Wurundjeri, Wadawurrung, Djabwurrung, Gulidjan, Djargurdwurrung, Giraiwurrung, Guditjmara, Jardwadjali, Baraba Baraba, Kokatha, Meru, Danggali and Barkindji people.

I thank them for their continued custodianship of the land and animals and pay my respects to their Elders past and present.





Linux/Unix/macOS command line

- Tab: autofill (if it doesn't autofill something is incorrect)
- Ctrl-C: Abort command
- ls: list directory contents
- tree: visualize directories, recursively
- pwd: print working (i.e., current) directory
- cd: change directory
- mkdir: make directory
- rmdir: remove a directory
- nano: open a text editor
- cp: copy a directory or a file
- cat/more/less: print contents of a file to the terminal
- rm: remove a file (rm -r: removes a directory)
- mv: move (i.e., rename) a directory or a file
- head: print the first ten lines of a file to the terminal
- tail: print the last ten lines of a file to the terminal
- curl or wget: download a file from a URL (you will see this in other QIIME2 tutorials)
- man: learn about a command (also, most other cmds: -h; --help)

Background on data

Captive-breeding and translocations are standard tools in conservation and species management (IUCN 2013, Batson et al. 2015), but ...

- Captive-born animals often show **lower post-release survival** than wild-caught animals (Stepkovitch et al. 2022; Thomas et al. 2023)
- Translocated captive-born animals are particularly susceptible to **starvation, predation, and disease** (Jule et al. 2008; Stepkovitch et al. 2022; Thomas et al. 2023)

What changes during captivity that leave animals less equipped for wild survival?

Fat-tailed dunnart (*Sminthopsis crassicaudata*)

Wild



PC Dr. Millie Scicluna

Conditions of captivity:

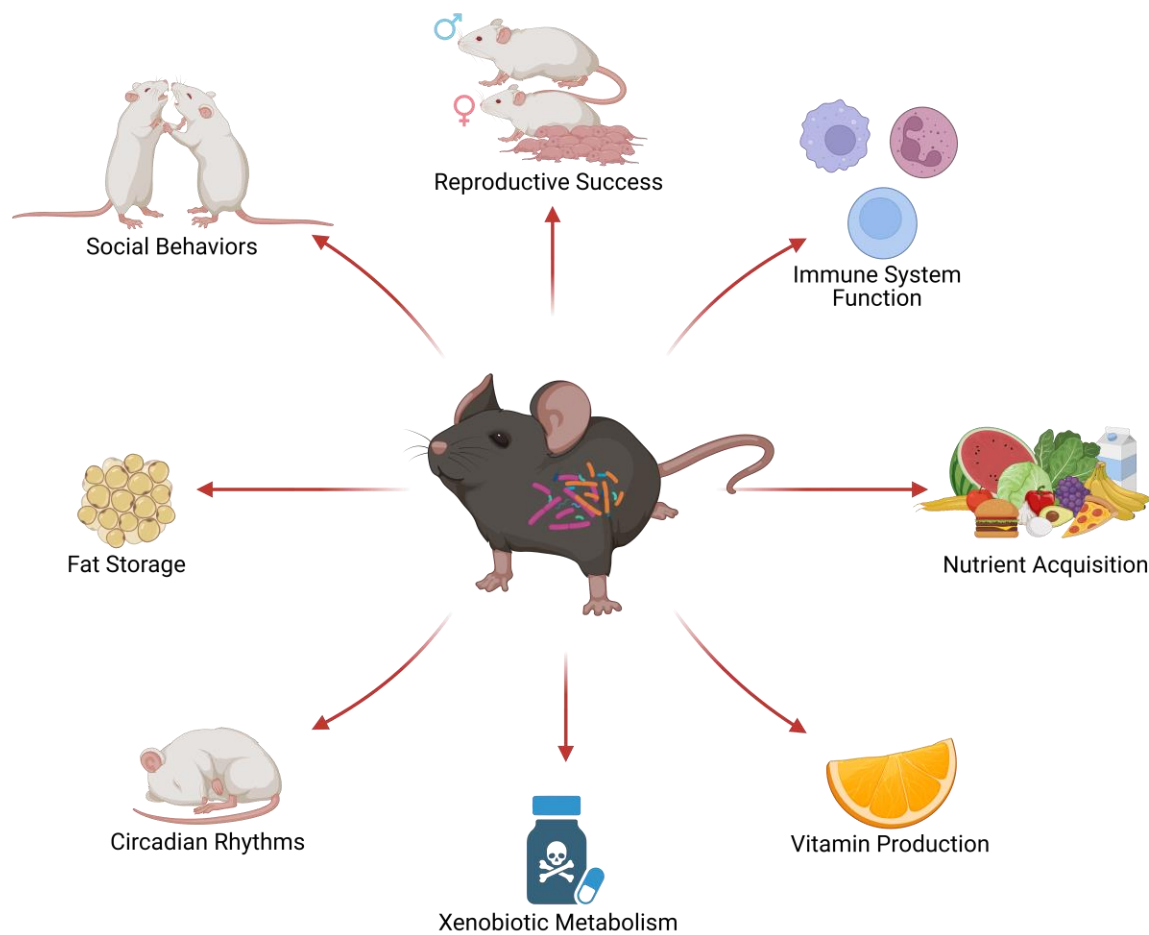


Captive

- Non-native, low diversity diets
- Antibiotic exposure
- Interactions with other species
- Potentially altered social structures
- Minimal exposure to natural environments



PC Dr. Millie Scicluna

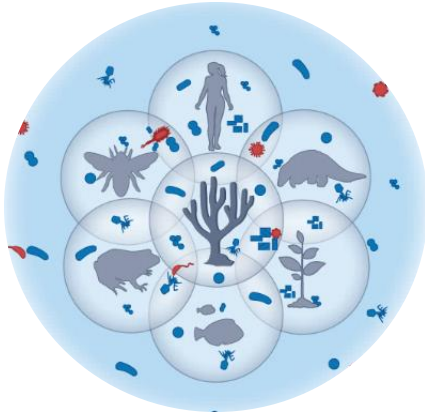


Animals are not just animals — they're ecosystems!

What is the influence of captivity on gut microbiota of the fat-tailed dunnart?

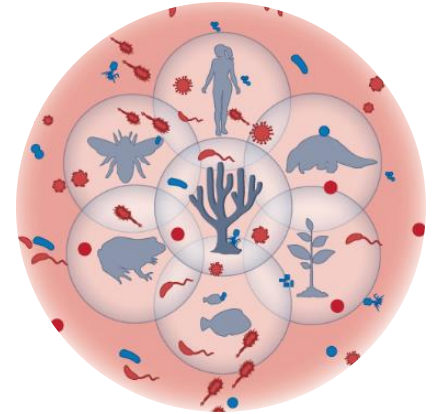
Conditions of captivity:

Wild Microbiome

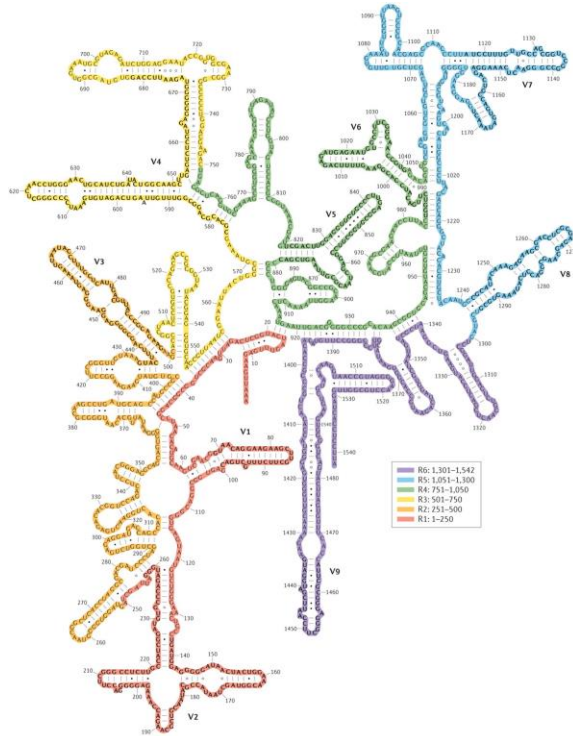


- Non-native, low diversity diets
- Antibiotic exposure
- Interactions with other species
- Potentially altered social structures
- Minimal exposure to natural environments

Captive Microbiome



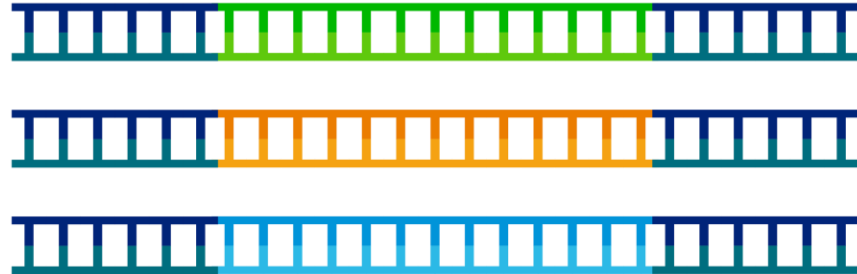
Variable Region 4 (V4) of the bacterial 16S rRNA gene

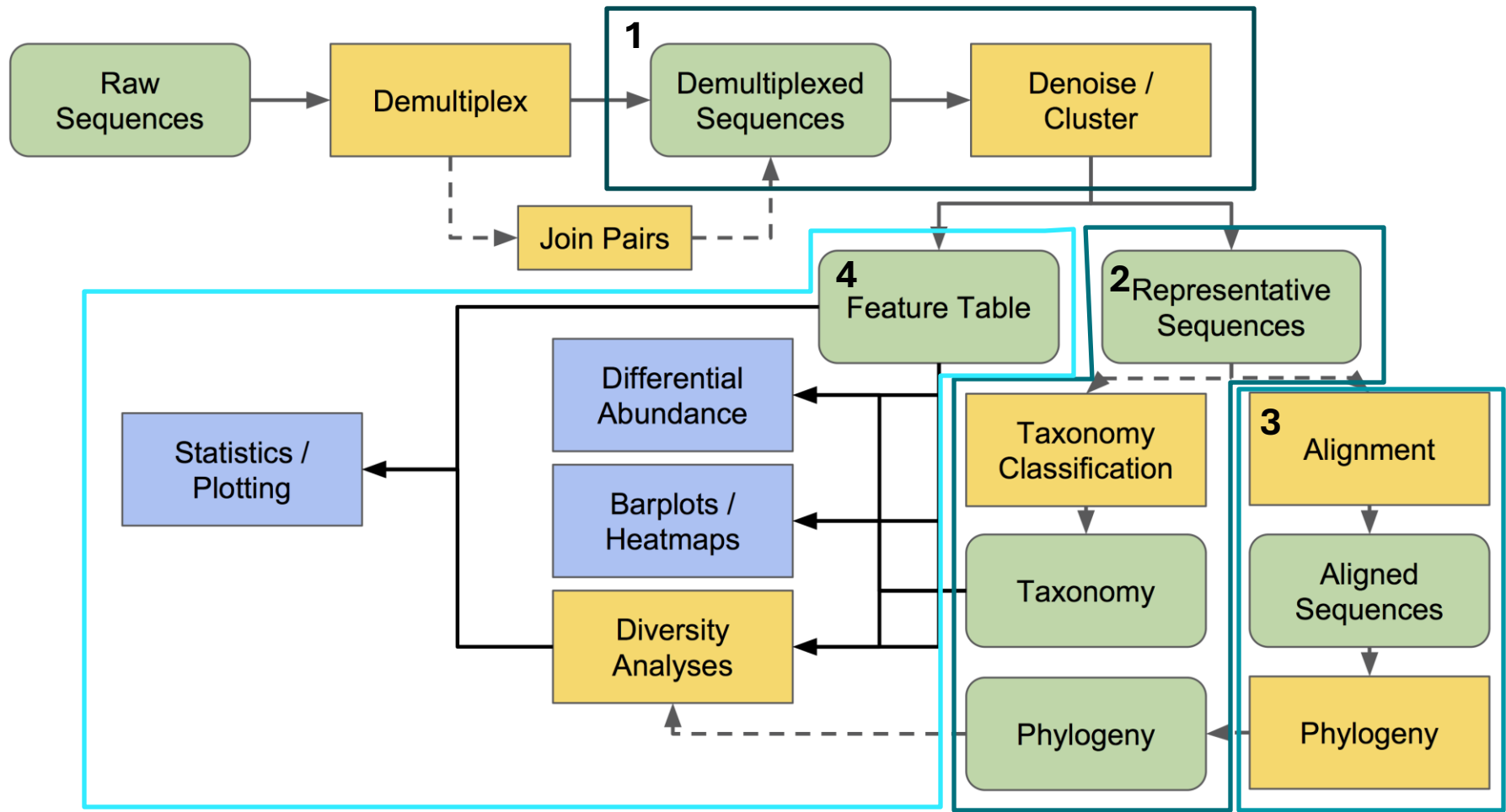


Nature Reviews | Microbiology

Other commonly sequenced marker genes/regions:

- ITS1/ITS2 (Fungi)
- 18S rRNA (Eukaryotes)
- CO1 (Animals, especially insects)
- 12S rRNA (Vertebrates)





Import data into QIIME2

What do you know about your data?

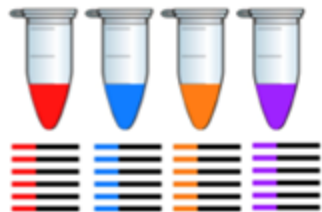
- Single vs paired end?
 - Single: one direction of sequencing
 - Paired: forward and reverse reads
- Multiplexed vs demultiplexed?
 - Multiplexed: fastq.gz file(s) for each read set and another that contains the associated barcodes
 - Demultiplexed: one fastq.gz file per sample

Import data into QIIME2

What do you know about your data?

- Single vs paired end?
 - Single: one direction of sequencing
 - **Paired: forward and reverse reads**
- Multiplexed vs demultiplexed?
 - Multiplexed: fastq.gz file(s) for each read set and another that contains the associated barcodes
 - **Demultiplexed: one fastq.gz file per sample**

Barcoded per-sample



Track per-sample barcodes (e.g., in spreadsheet)

Pool and sequence samples



sample-metadata.tsv	
SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

```
sequences.fastq(.gz)
```

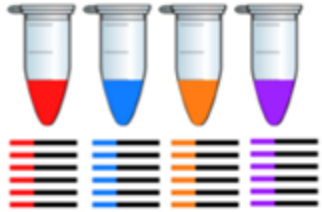
```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGT
GCGTGTGGTAAGTCATGGTGAAATCCTCCGGCTCAACCGAGGAAC TG
+
abaaaaa`^`a_]`\\``^`a`]]^^`a[VXGX`Z_\_\^a^SYOZWVSVGYVDXOV
T\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-6X_9267:1:1:25:267
TAGCTATGGGCCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGC GTAGGTG
GTGGCTTAAGCGCAGGGTTTAAAGGCA[ barcodes.fastq.gz)
```

barcodes.fastq(.gz)

aa^^[^_^^^_^[^[^____ZZ[^^[^']ZUZ XXRZRNVTRTNTWUUU^VJVOMIHQUU @HWI-6X_9267:1:1:25:609 TACGTAGGGGGCAAGCGTTATCCGGA GGATGGACAAGTCTGATGTGAAAGGCT + aaab`aaa`aaaaaaaaaaaaaaaa^aaaaa ^_`ZZ_T]XY^^^`ZXVJS[W[V^HVOYTE @HWI-6X_9267:1:1:25:519 GACGGAGGATGCAAGTGTATCCGGA GTTTACTAAGTCAACTGTTAAATCTTGAC + abaaaaaa`aaaaaa/aaaaaaa```aa_aaa ^`[XTVX]`T_VZ][ZXVYFX_VYJWWZL @HWI-6X_9267:1:1:25:1109 TACGGAGGGTGCGAGCGTTAATCGGA GTTTAGGTAAGTCAGATGTGAAAGCCC + aaaba^^`a^N_`_`a_a]Zaa^^Z`[M]a` [PTRPTRYUBBBBBBBBBBBBBBBBBBBB	@HWI-6X_9267:1:1:25:1051 AACGCAC + bbbbbbbb @HWI-6X_9267:1:1:25:267 AAGAGAT + bbbbbbbb @HWI-6X_9267:1:1:25:609 AACGCAC + bbbbbbbb @HWI-6X_9267:1:1:25:519 ACAGCAG + bbbbbbbb @HWI-6X_9267:1:1:25:1109 ACAGCTA + bbbbbbbb @HWI-6X_9267:1:1:25:434 ACACGAG +
--	--

Demultiplexed Data

Barcoded per-sample



Pool and
sequence
samples



Track per-sample
barcodes (e.g., in
spreadsheet)

sample-metadata.tsv	
SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

4ac2.fastq(.gz)

e375.fastq(.gz)

4gd8.fastq(.gz)

9872.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAATTAAGCGTAA
AGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAAAGCC
CCGGGCTCCACCTGGGAATGG
+
aaaba^`a^N_`_`a_a]Zaa^^\Z`[M]a`[VYa^X^Z]NZ`[T
Y]_^RVH_PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBB
BBBBBBBB
```

What do you know about your data?

- Single vs paired end?
 - Single: one direction of sequencing
 - Paired: forward and reverse reads
- Multiplexed vs demultiplexed?
 - Multiplexed: fastq.gz file(s) for each read set and another that contains the associated barcodes
 - Demultiplexed: one fastq.gz file per sample
- Have your adapters and primers been removed?
- Will your files come zipped? (ending in .gz)

Unsure? **Make sure you ask the sequencing facility and know the answers to these specific details.**

Import data

Software plugin action

--option 1

--option 2

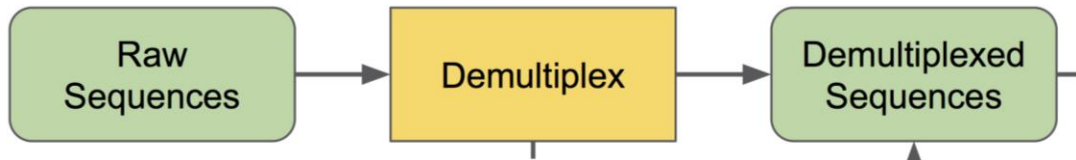
```
qiime tools import \
```

```
--type 'SampleData[PairedEndSequencesWithQuality]' \ #check out the import #QIIME2 page
```

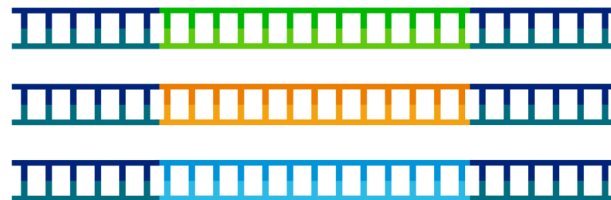
```
--input-path raw_data \ #path to data directory relative to current directory
```

```
--input-format CasavaOneEightSingleLanePerSampleDirFmt \ #from import tutorial
```

```
--output-path analysis/seqs/combined.qza #location and name for output file
```



Cutadapt code



```
qiime cutadapt trim-paired \ #we want to trim paired (F and R read) data
--i-demultiplexed-sequences analysis/seqs/combined.qza \ #location of
#demultiplexed sequences. This will match your output-path from import code.
--p-front-f GTGYCAGCMGCCGCGGTAA \ #F primer sequence
--p-front-r GGAATACNVGGGTWTCTAAT \ #R primer sequence
--p-error-rate 0.10 \ #maximum allowed error rate, range 0-1. Play with this!
--p-overlap 10 \ #Require at least X bases of overlap between read and adapter for an
adapter to be found. Play with this!
--p-match-adapter-wildcards \ #Interpret IUPAC wildcards (e.g. Y,N,V) in adapters
--p-discard-untrimmed \ #Discard reads in which no adapter was found.
--p-quality-cutoff-3end 20 \ #Trim nucleotides with quality lower than x from 3' end.
--p-quality-cutoff-5end 0 \
--p-cores 8 \ #Number of CPU cores to use
--output-dir analysis/seqs_trimmed \ #location of output file
--verbose #tell me when this action is done
```

Cutadapt = cutting off adapters (overhang+primer)

```
=== Summary ===
```

```
Total read pairs processed:          13,122
  Read 1 with adapter:              13,122 (100.0%)
  Read 2 with adapter:              13,122 (100.0%)
Pairs that were too short:           0 (0.0%)
Pairs written (passing filters):      13,122 (100.0%)
```

```
Overview of removed sequences
```

length	count	expect	max.err	error counts
43	1	0.0	3	1
45	1	0.0	3	1
46	19	0.0	3	14 3 0 2
47	106	0.0	3	62 27 17
48	1047	0.0	3	705 330 9 3
49	11931	0.0	3	11512 405 14
50	17	0.0	3	4 12 1

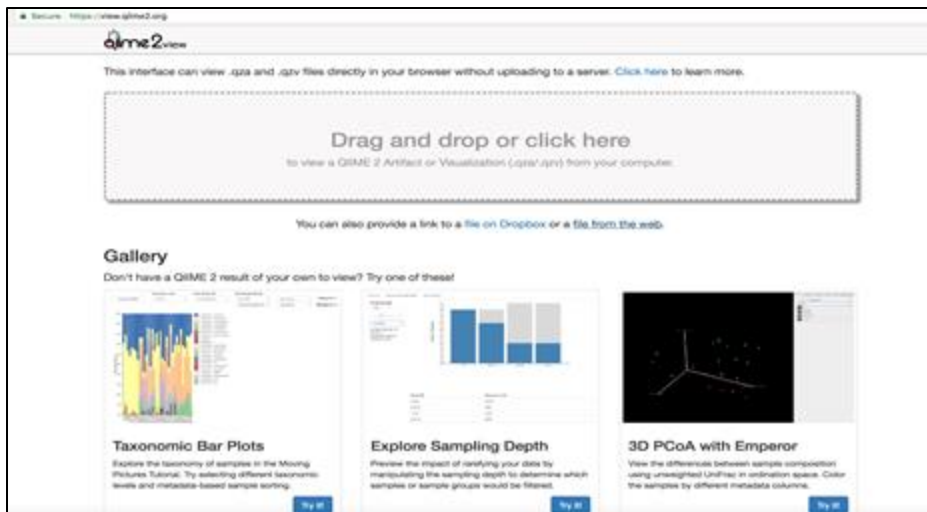
Summarize counts per sample

```
qiime demux summarize \ #we want to visualize the demultiplexed data
--i-data analysis/seqs_trimmed/trimmed_sequences.qza \ #location of data
--o-visualization analysis/visualisations/trimmed_sequences.qzv #output file
```

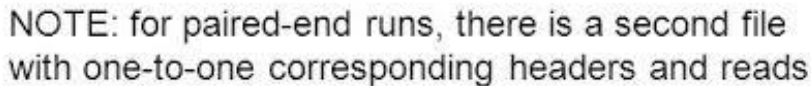
Accessing output files

Use FileZilla to transfer to your local drive

- Go to <https://view.qiime2.org/>
- Drag file into qiime2 view



FASTQ Format (Illumina Example)



Phred Quality Score = Q-score

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality Score Encoding

In FASTQ files, quality scores are encoded into a compact form, which uses only 1 byte per quality value. In this encoding, quality score is represented as the character with an ASCII code equal to its value + 33. The following table demonstrates relationship between the encoding character, its ASCII code, and the quality score represented.

NOTE

When Q-score binning is in use, the subset of Q-scores applied by the bins is displayed.

Table 2 ASCII Characters Encoding Q-scores (0-40)

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
\$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26
'	39	6	<	60	27
(	40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
,	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			

DADA2: What is it?

- **D**ivisive **A**mplicon **D**enoising **A**lgorithm, version 2 ([Callahan et al. 2016](#))
- DADA2 ...
 - ... is a software package (QIIME2 add-on) that models and corrects Illumina-sequenced amplicon errors
 - ... infers sample sequences exactly and resolves differences of as little as one nucleotide (ASVs). This allows for the identification of variants and reveal diversity in a given taxonomic group
 - ... is reference free and applicable to any amplicon on interest

DADA2: How does it do that?

- **Denoising (remove and/or correct noisy reads)**
 - Filtering - user defined. Trim/truncates sequences to a specified length, removes sequences shorter than that length
 - Model errors within a read and between reads
 - Abundance - sequences too abundant to be explained by errors in sequencing are kept
 - Sequence comparison (i.e. excluding reads whose pairs have >10% mismatch)
- **Clustering (collapse similar sequences)**
 - Reads with exact overlaps are merged by sample
 - Reads with the same sequence are grouped into unique sequences with an associated abundance and consensus quality profile (dereplication)
 - These are called **Amplicon Sequencing Variants (ASVs)** or Features in some tutorials
- **Chimera removal (identifying sequences that are two-parent chimeras of more abundant output sequences)**

How to apply summary plots to dada2 code

- 1. Understand Your Quality Profile Shape:** Modern Illumina data often shows flat, consistently high-quality profiles. If there are declines, you want to remove them (more likely in R reads).
- 2. Balance Quality Retention with Read Overlap:** Truncation is about keeping usable sequence whilst ensuring paired-end reads can merge. Minimum overlap is ~12 bp, but **>50 bp overlap is recommended** for robust merging and read retention. $\text{Overlap} = (\text{forward truncation length} + \text{reverse truncation length}) - \text{amplicon length}$
- 3. Inspect Both Forward and Reverse Reads:** Examine quality plots for both directions separately. Reverse reads often show lower quality at their 3' ends. Set truncation lengths (--p-trunc-len-f and --p-trunc-len-r) independently based on where *each* read's quality becomes problematic.
- 4. Test and Iterate:** There's no universal "correct" answer. **Try 2–3 truncation-length combinations, run DADA2, and compare output metrics (read retention, merge success rate, ASV count). Choose the combination that maximises sequence retention without sacrificing merging efficiency.**

DADA2 code

```
qiime dada2 denoise-paired \ #software-plugin-action
--i-demultiplexed-seqs analysis/seqs_trimmed/trimmed_sequences.qza \ #location
#of data
--p-trunc-len-f 210 \ #position to truncate forward reads due to decrease in quality
--p-trunc-len-r 170 \ #position to truncate reverse reads due to decrease in quality
--p-n-threads 0 \ #number of cores; 0 = all cores used = faster
--p-pooling-method 'pseudo' \ #pooling method (independent or pseudo)
--output-dir analysis/dada2out \ #output path
--verbose #tell me when the action is complete
```

Sample metadata: formatting

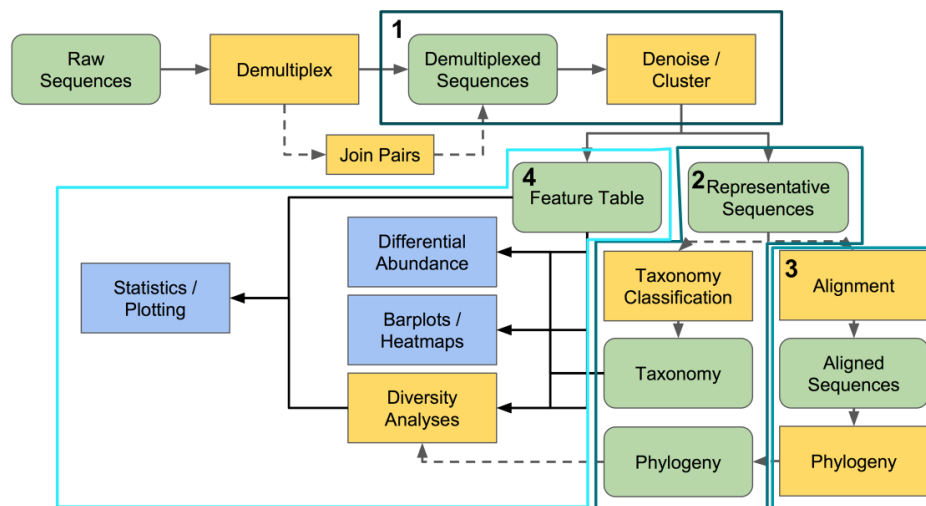
<https://keemei.qiime2.org>

Moving Pictures sample-metadata (QIIME 2.0.6) ☆ 📁										
File Edit View Insert Format Data Tools Add-ons Help Last edit was yesterday at 12:02 PM										
fx #SampleID										
	A	B	C	D	E	F	G	H	I	J
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	BodySite	Year	Month?	Day	Subject	ReportedAntibioticUsage	DaysSinceExperimentStart
2	L1S8	ERRORS: Duplicate sample ID. Duplicates in A2, A21		ut	2008	10	28	1	Yes	0
3	L1S140			ut	2008	10	28	2	Yes	0
4	L1S57			ut	2009	1	20	1	No	84
5	L1S208			ut	2009	1	20	2	No	84
6	L1S76	ACTACGTGTGC	GTGCCAGCMG	gut	2009	2	17	1	No	112
7	L1S105	AGTGCGATGC	GTGCCAGCMG	gut	2009	3	17	1	No	140
8	L1S257	CCGACTGAGAT	GTGCCAGCMG	gut	2009	3	17	2	No	140
9	L1S281	CCTCTCGTGAT	GTGCCAGCMG	gut	2009	4	14	2	No	168
10	L2S240	CATATCGCAGT	GTGCCAGCMG	left palm	2008	10	28	2	Yes	0
11	L2S155	ACGATGCGACC	GTGCCAGCMG	left palm	2009	1	20	1	No	84
12	L2S309	CGTGCAATTATC	GTGCCAGCMG	left palm	2009	1	20	2	No	84
13	L2S175	AGCTATCCACC	GTGCCAGCMG	left palm	2009	2	17	1	No	112
14	L2S204	ATGCAGCTCAC	GTGCCAGCMG	left palm	2009	3	17	1	No	140

Head to tutorial and complete Section 1

Section 1: Importing, cleaning and quality control of the data

The dada2 denoise-paired step must be run staggered.



Taxonomic assignment of observed sequences (ASVs)

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGTAAGTCCATGG
TGAATCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGCAAGCGTTATCCGGATTACTGGGTGTAAGGGAGCGTAGACGGATGGACAAGTCTGATGT
GAAAGGCTGGGGCTCAACCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGCGTAAAGAGTGCCTAGGTGGCTTAAGCGCAGGGT
TTAAGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTACTAAGTCAACTGTT
AAATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGCGGTTAGGTAAGTCAGATGT
GAAAGCCCGGGCTCCACCTGGGAATGG
```

Taxonomic assignment of observed sequences.

Reference Database

Silva, Greengenes, etc.

```
FeatureData[Sequence]
```

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCTAGGTGGCTTGGTAAGTCATGG
TGAATCTCTCGGCTCAACCGAGGAAC TG
>feature4
TAGCTAGGGGGCAAGCGTTATCGGAATTACTGGTGTAAGGGAGCGTAGACGAGTGGAACAAGTCATGATG
GAAAGGCTGGGGCTCAACCCCGGAGCGG
>feature2
TAGCTATGGGGCAAGCGTTATCCGGAATTATTGGCGTAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGT
TTAAGGC AATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGCGGTAAGCGTCTGTAGGTGTTTACTAAGTCAACTGTT
AAATCTTGAGGCTCAACCTCGAAATCG
>feature3
TCGGAAGGGTGCAGCGTTAATCGGAATTACTGGGCGTAAGCGTACGTAGGCGGTAGGTAAGTCAGATGT
GAAAGCCCGCGGCTCCACCTGGGAATGG
```

```
FeatureData[Sequence]
```

```
>reference-sequence-1
TTGAAGGTGGGACGACGCTTGCTCGGAATCACTGGGCATAAAGCGCGGTAGTGGTCTTGTAAGTCAACAT
GTGACTCAACCGACGGAAGTAATTAAGAGTGGACGACGCTTGCTCGGAATCACTGGGCATAAAGCGCGC
GTAGGTGGCTTGTAAGTCAACATGGTGACTCAACCGAGGAACGTAA

>reference-sequence-2
AACGTAGCGAAGCGTTATCGGATTACTGGGTGTAAGGGAGCGTAGACGGATGGACAAAGCTGATGTGAA
ACGGCTGGGGCTCAACCCCGGAGCGGTGTAAGGTGGACGACCGCTTGCTCGGAATCACTGGGCATAAAGCGC
C
5'-----3'
AGGACGTAA
```

FeatureData[Taxonomy]

reference-sequence-1 Bacteria; Proteobacteria; Gammaproteobacteria; Legionellales;

reference-sequence-2 Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales; Flavob

reference-sequence-3 Bacteria; Proteobacteria; Deltaproteobacteria; Desulfuromonad

reference-sequence-4 Archaea; Euryarchaeota; DSEG; 104A5

Taxonomic assignment of observed sequences.

Reference Database

Silva, Greengenes, etc.

```
FeatureData[Sequence]

>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGG
TGAATCCTCGGCTCAACCGAGGAACCTG
>feature4
TACGTAGGGGCAAGCGTTATCCGGAATTACTGGGTGTAAGGGAGCGTAGACGGATGACAACTCTGATGT
GAAAGGCTGGGGCTCAACCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGCGTAAGAAGTGCCTAGGTGGTGGCTTAAGCGCAGGGT
TTAAGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGTTTACTAAGTCAACTGTT
AAATCTTGAGGCTCAACCTCGAATCG
>feature3
TACGAGGGTGCGAGCGTTAATCGGAATTACTGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGT
GAAAGCCCGGGCTCCACCTGGGAATGG
```

```
FeatureData[Sequence]

>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACAT
GGTGACTCAACCGAGGAACCTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGC
GTAGGTGGCTTGGTAAGTCAACATGGTGAACCAACCGAGGAACCTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGAATTACTGGGTGTAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCGGGACGGTTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGC
C
>reference-sequence-3
AGGAACGTAA
>reference-sequence-4
AGGAACGTAA

FeatureData[Taxonomy]

reference-sequence-1 Bacteria; Proteobacteria; Gammaproteobacteria; Legionellales;
reference-sequence-2 Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales; Flavob
reference-sequence-3 Bacteria; Proteobacteria; Deltaproteobacteria; Desulfuromonada
reference-sequence-4 Archaea; Euryarchaeota; DSEG; 104A5
```

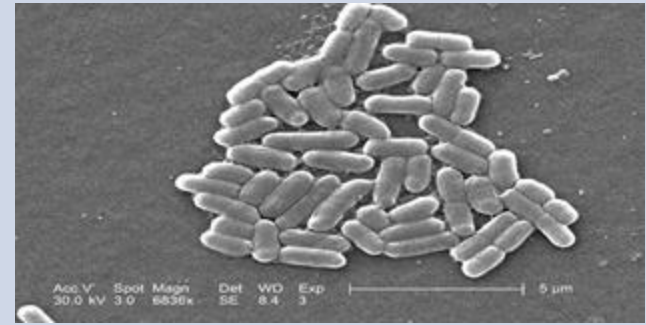
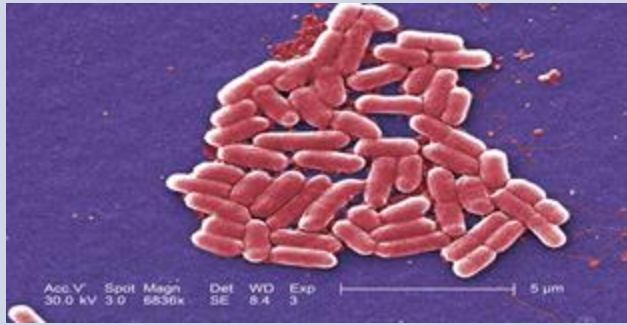
Compare observed sequences to annotated reference sequences to make taxonomic assignments.

```
FeatureData[Taxonomy]
```

```
feature5 Bacteria; Proteobacteria
feature4 Bacteria; Proteobacteria
feature2 Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales
feature1 Bacteria; Proteobacteria
feature3 Bacteria; Proteobacteria; Deltaproteobacteria
```

Ideal 16S

Real 16S



Kingdom

Bacteria

Bacteria

Phylum

Proteobacteria

Proteobacteria

Class

Gammaproteobacteria

Gammaproteobacteria

Order

Enterobacteriales

Enterobacteriales

Family

Enterobacteriaceae

Enterobacteriaceae

Genus

Eschericia

Species

coli

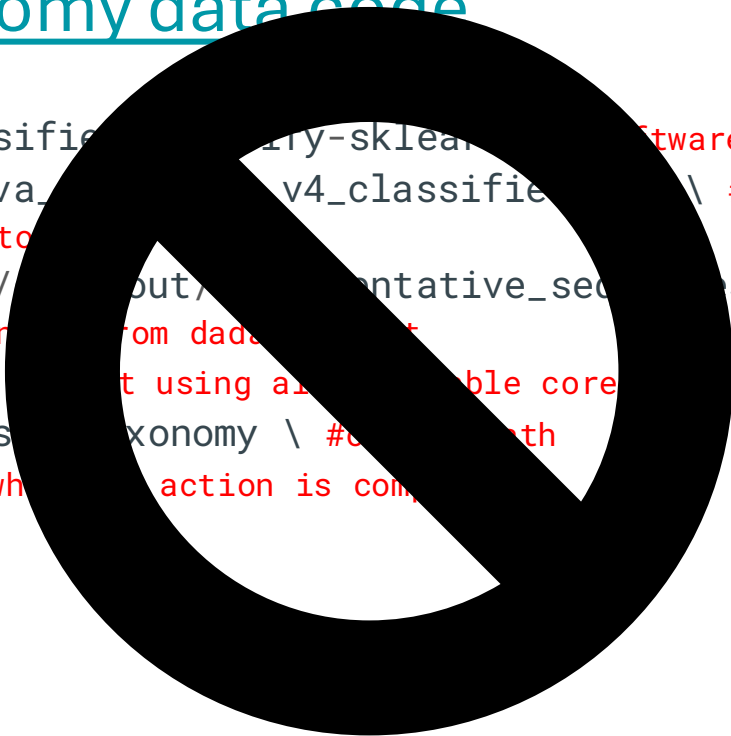
OTU 2445338

Strain

O157:H7

--

Assign taxonomy data code



```
qiime feature-classifier classify-sklearn --software-plugin-action
--i-classifier silva-132-tax4-classifier.qza \ #location classifier. This #file
was provided for you to
--i-reads analysis/output/representative_sequences.qza \ #dereplicated
#representative sequences from dada2
--p-n-jobs 1 \ #run using all available cores
--output-dir analysis/taxonomy \ #output path
--verbose #tell me when action is complete
```

Filtering actions

- Filter-table: taxonomy-based filtering of feature table
- Filter-features: filter specific features (ASVs) from feature table
- Filter-features-conditionally: filter features based on abundance and prevalence
- Filter-samples: filter samples from feature table

Filtering code

```
qiime taxa filter-table \ #software-plugin-action
--i-table analysis/dada2out/table.qza \ #feature table we are filtering
--i-taxonomy analysis/taxonomy/classification.qza \ #classification file that has
all of #the taxonomic assignments of the ASVs in our feature table
--p-exclude Mitochondria,Chloroplast \ #remove ASVs that have been identified as
Chloroplast or Mitochondria
--o-filtered-table analysis/taxonomy/16s_table_filtered.qza \ #output path
--verbose #tell me when the action is complete
```

```
qiime feature-table summarize \
--i-table analysis/taxonomy/16s_table_filtered.qza \
--m-sample-metadata-file dunnart_metadata.tsv \
--o-visualization analysis/visualisations/summary_table_filtered \ #gives three files:
feature_frequencies.qza, sample_frequencies.qza, summary.qzv
--verbose
```

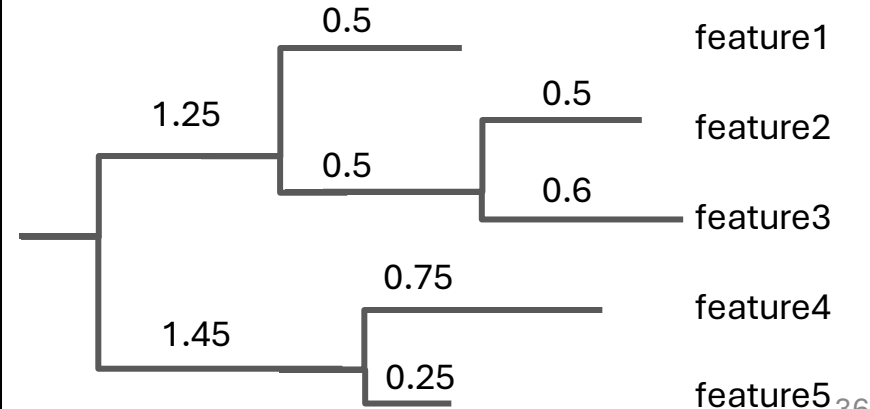
Phylogenetic reconstruction of observed sequences

FeatureData[Sequence]

```
>taxon5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCTAGGTGGCTTGTAAGTCATGG
TGAAATCCCTCGGCTCAACCGAGGAACTG
>taxon4
TACGTAGGGGGCAAGCGTTATCCGGATTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCGTATGT
GAAAGGCTGGGGCTCAACCCGGGACGG
>taxon2
TACGTATGGGGCAAGCGTTATCCGGAATTATGGGCGTAAAGAGTGCCTAGGTGGTGGCTTAAGCGCAGGGT
TTAAGGCAATGGCTTAACCTATTGTTCTC
>taxon1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGTTTACTAAGTCAACTGTT
AAATCTTGAGGCTCAACCTCGAAATCG
>taxon3
TACGGAGGGTGCAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGCGGTTAGGTAAGTCAGATGT
GAAAGCCCGGGCTCCACCTGGGAATGG
```

Align sequences,
filter highly variable
(i.e., randomly
evolving) positions,
and build
phylogenetic tree.

Phylogeny[Rooted]



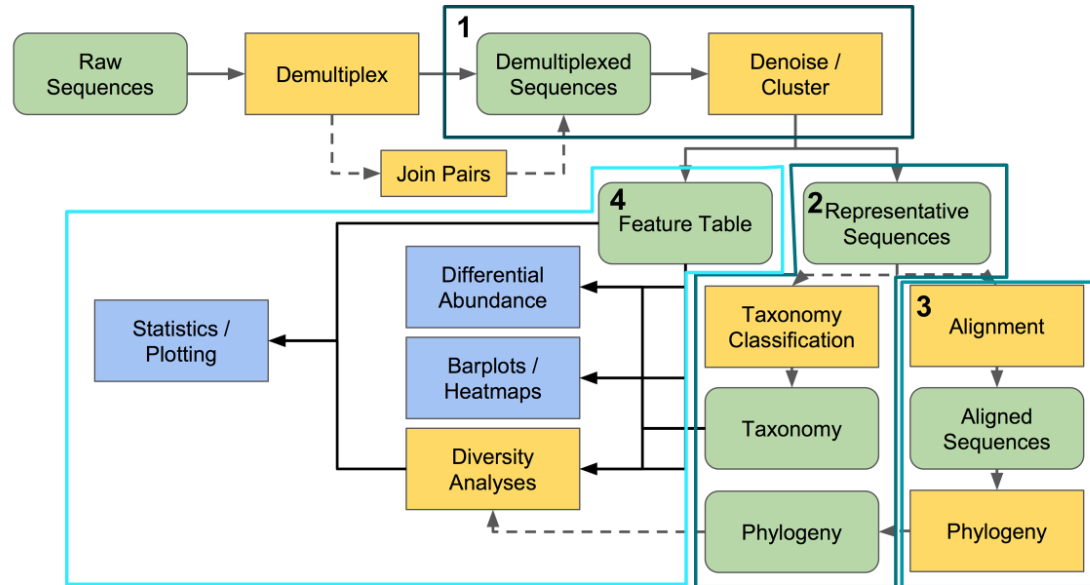
Build phylogenetic tree code

```
qiime phylogeny align-to-tree-mafft-fasttree \ #software-plugin-action
--i-sequences analysis/dada2out/representative_sequences.qza \ #sequences to
#align
--o-alignment analysis/tree/aligned_16s_representative_seqs.qza \ #perform an
alignment
--o-masked-alignment analysis/tree/masked_aligned_16s_representative_seqs.qza \
#Mask sites in the alignment that are not phylogenetically informative
--o-tree analysis/tree/16s_unrooted_tree.qza \ #Generate a phylogenetic tree
--o-rooted-tree analysis/tree/16s_rooted_tree.qza \ #Apply mid-point rooting to the
tree
--p-n-threads 1 \ #run script using all available cores
--verbose #tell me when the action is complete
```

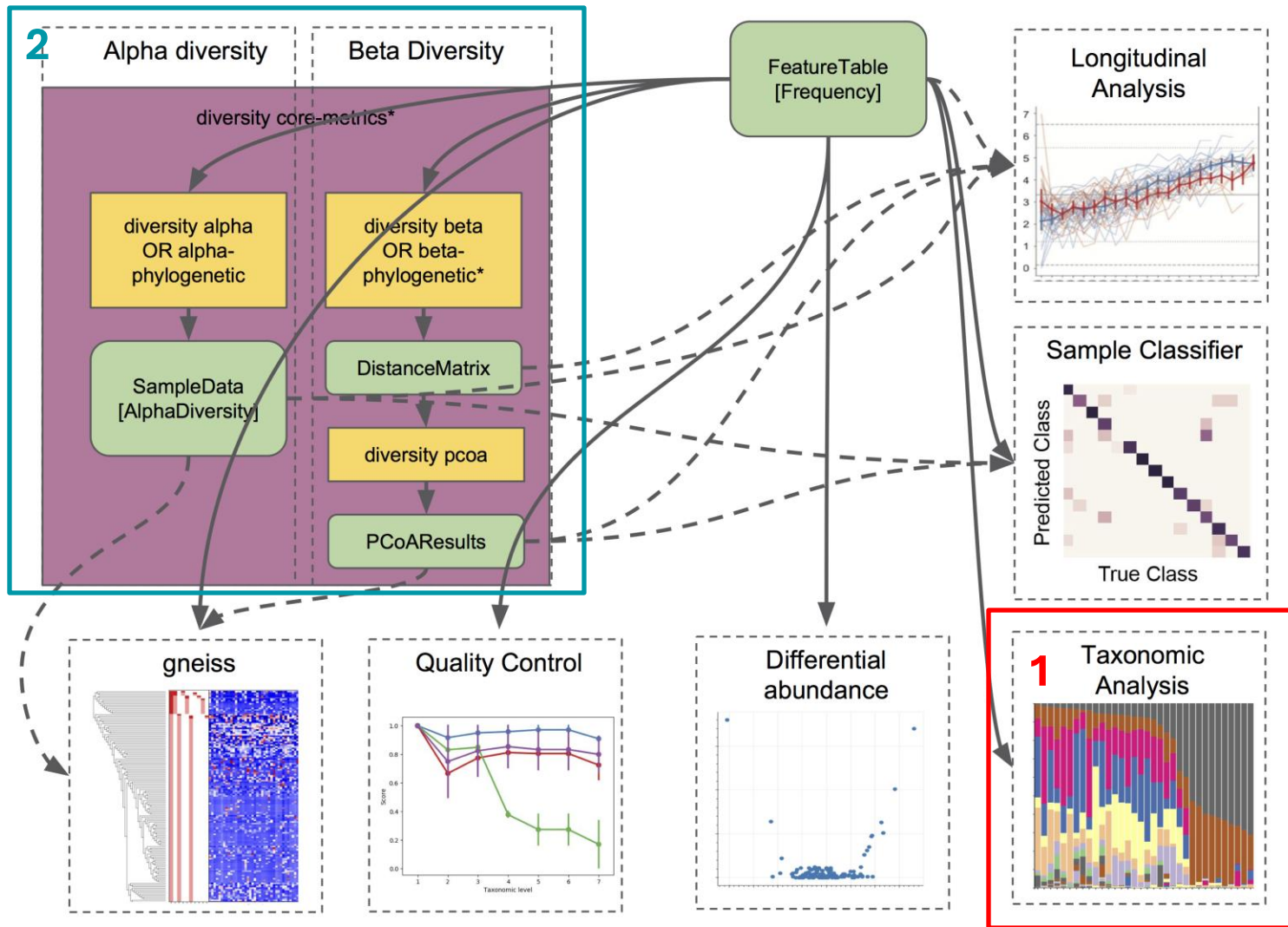
Head to tutorial and complete Section 2&3

Finish Section 2: Taxonomic Analysis

Classification.qza provided for you.



Section 3: Build a phylogenetic tree

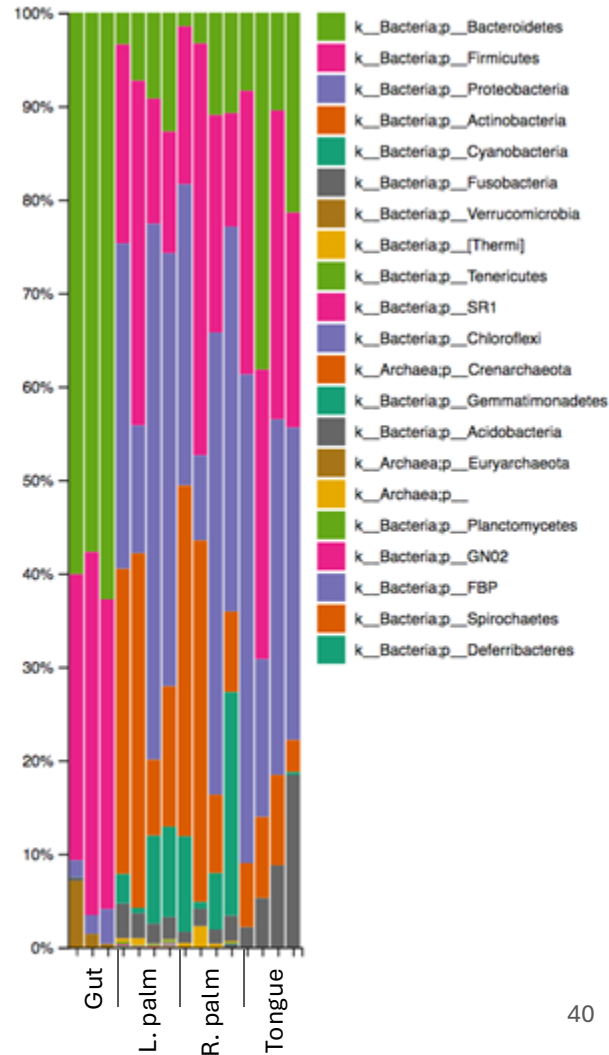


Visualizing taxonomic profiles

Interactive barplots support:

- Taxonomic level selection
- Multi-level sorting
- Filtering
- Coloring
- Exporting plots (SVG) and raw data

Relative
frequency



Barplot code

```
qiime taxa barplot \ #software-plugin-action
--i-table analysis/taxonomy/16s_table_filtered.qza \ #data to build barplot
--i-taxonomy analysis/taxonomy/classification.qza \ #classification file
--m-metadata-file dunnart_metadata.tsv \ #path to metadata file
--o-visualization analysis/visualisations/barchart.qzv \ #output file
--verbose #tell me when the action is complete
```

Comparing microbial communities

Alpha diversity metrics operate on a single sample (i.e., within sample diversity).

Common metrics include:

Richness (Observed ASVs, Chao1) — counts the number of distinct taxa

Evenness (Shannon, Simpson) — measures how evenly taxa are distributed

Phylogenetic diversity (Faith's PD) — incorporates evolutionary relationships

Comparing microbial communities

Beta diversity metrics operate on a pair of samples (i.e., between sample diversity).

Common metrics include:

Unweighted UniFrac — presence/absence of taxa; phylogenetically informed

Weighted UniFrac — incorporates taxon abundance; phylogenetically informed

Bray-Curtis — abundance-based; does not use phylogeny

Jaccard — presence/absence; does not use phylogeny

Comparing microbial communities

What is Rarefaction? Rarefaction normalises sequencing depth by randomly subsampling reads to a uniform level. Use the rarefaction curve to choose a depth that retains most samples whilst removing those with very low read counts.

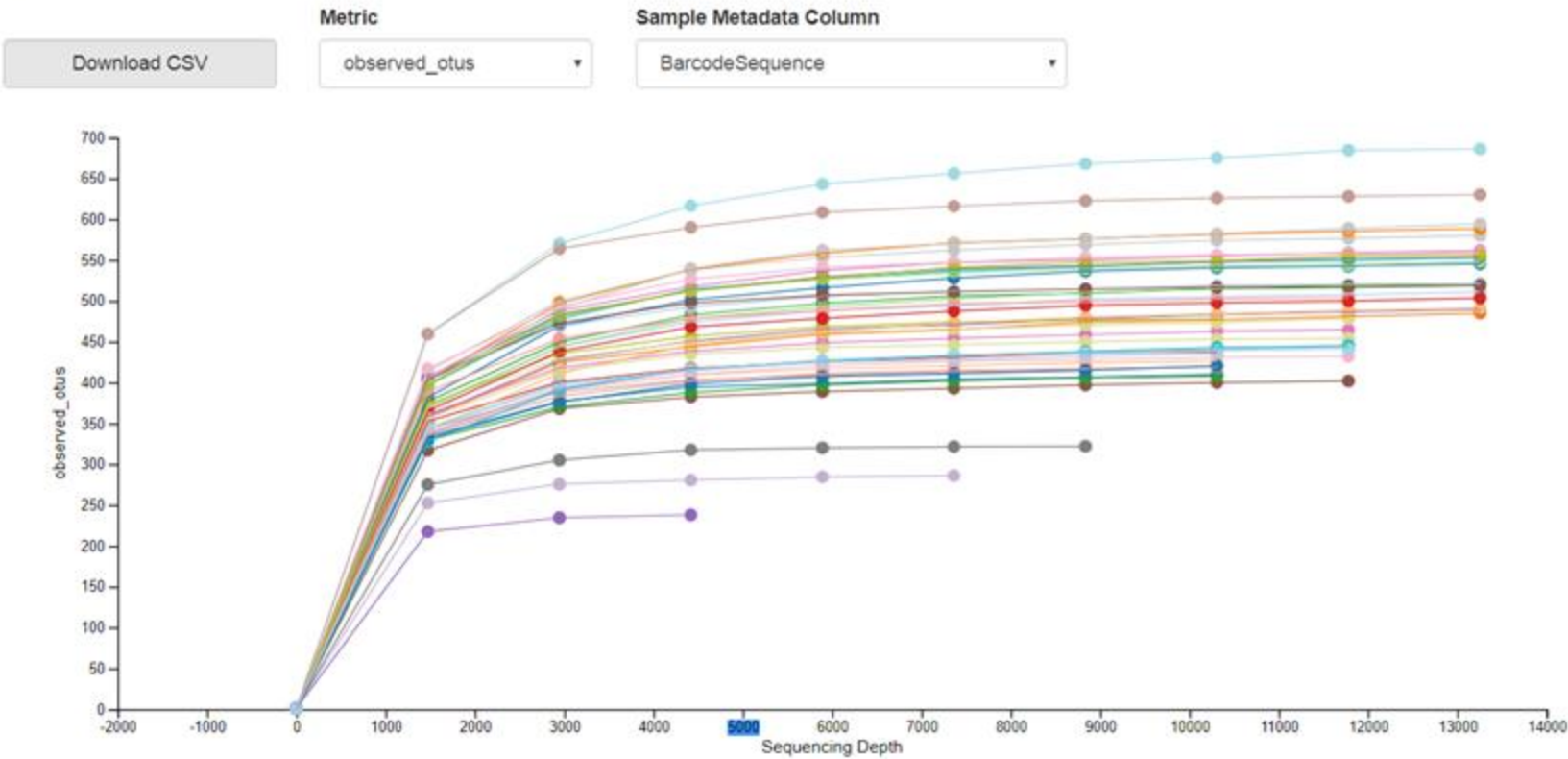
Rarefaction for Alpha Diversity = essential. Alpha diversity metrics are sensitive to sequencing depth—a sample with 10,000 reads appears more diverse than an identical sample with 1,000 reads. Rarefying removes this technical bias and enables fair comparisons.

Rarefaction for Beta Diversity = not necessary. Beta diversity metrics (Bray–Curtis, UniFrac) are compositionally based and naturally account for depth differences. Rarefying can discard information and reduce statistical power.

Rarefaction code (must be run consecutively)

```
qiime diversity alpha-rarefaction \ #software-plugin-action
--i-table analysis/taxonomy/16s_table_filtered.qza \ #path to data
--i-phylogeny analysis/tree/16s_rooted_tree.qza \ #phylogenetic tree required
for some analyses (i.e. unifrac)
--p-max-depth 200000 \ #maximum rarefaction depth. Typically use the median
number of reads from 16s_table_filtered.qzv file
--p-min-depth 500 \ #minimum rarefaction depth. Feel free to adjust this.
--p-steps 40 \ #number of points between min and max. Feel free to adjust this.
--m-metadata-file dunnart_metadata.tsv \ #path to metadata file
--o-visualization analysis/visualisations/16s_alpha_rarefaction.qzv \ #output
file
--verbose #tell me when the action is complete
```

Alpha rarefaction



Alpha and beta diversity code

```
qiime diversity core-metrics-phylogenetic \ #software-plugin-action
--i-phylogeny analysis/tree/16s_rooted_tree.qza \ #phylogenetic tree required
for #some analyses (i.e. unifrac)
--i-table analysis/taxonomy/16s_table_filtered.qza \ #path to data
--p-sampling-depth 100000 \ #selected based on rarefaction curves and read
counts in samples
--m-metadata-file dunnart_metadata.tsv \ #path to metadata file
--o-visualization analysis/diversity_metrics\ #output folder
```

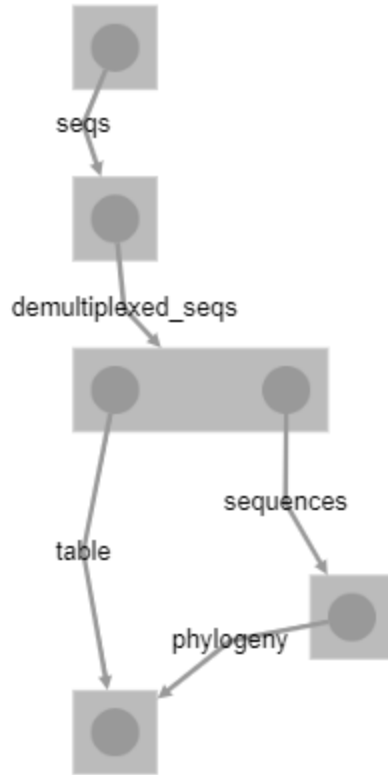
Stats code follows

Differential Abundance code

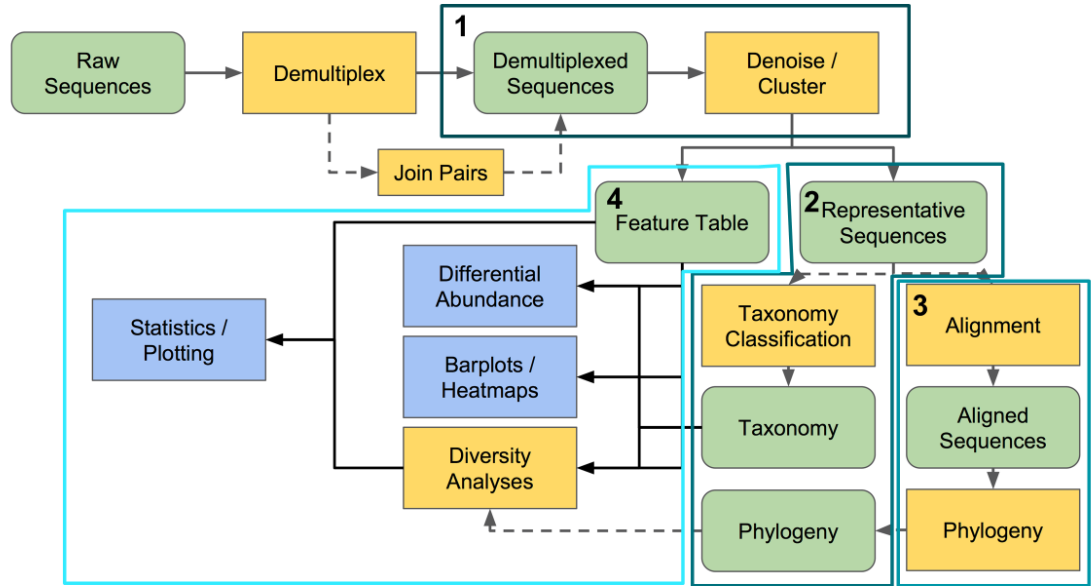
```
qiime composition ancombc2 \  
--i-table analysis/taxonomy/16s_table_filtered.qza \  
--m-metadata-file dunnart_metadata.tsv \  
--p-fixed-effects-formula Captivity \  
--o-ancombc2-output analysis/ancombc2-results.qza
```

```
qiime composition ancombc2-visualizer \  
--i-data analysis/ancombc2-results.qza \  
--i-taxonomy analysis/taxonomy/classification.qza \  
--o-visualization analysis/visualisations/ancombc2-barplot.qzv
```

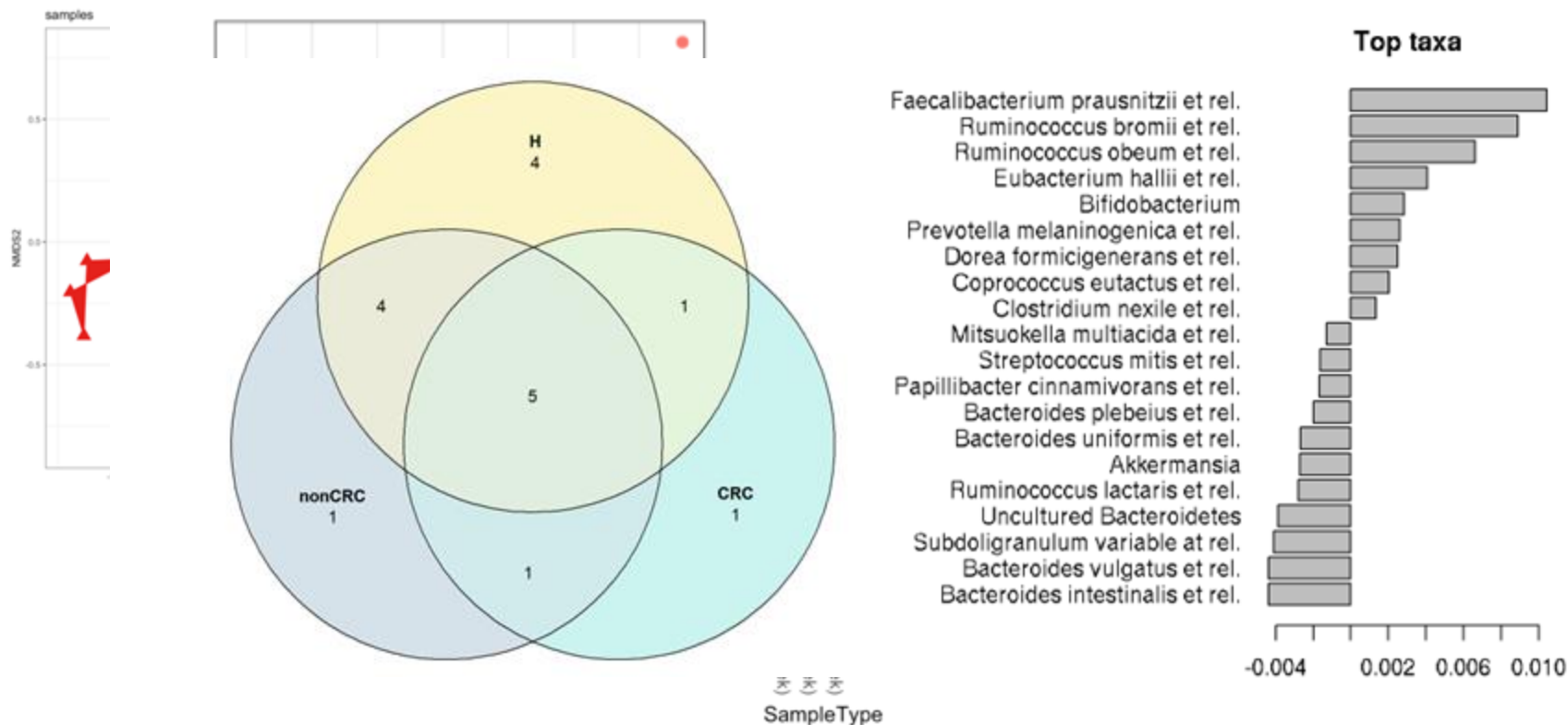
Provenance



Head to tutorial and complete Section 4



QIIME2 → R



Useful R packages

- [Phyloseq](#)
- [Microbiome](#)
- [Vegan](#)
- [Indicspecies](#)
- [Decontam](#)
- [Comprehensive list of R packages for microbiome analyses](#)

Head to tutorial and complete Section 5

Section 5: Exporting data for further analysis in R