Australian
BioCommons

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

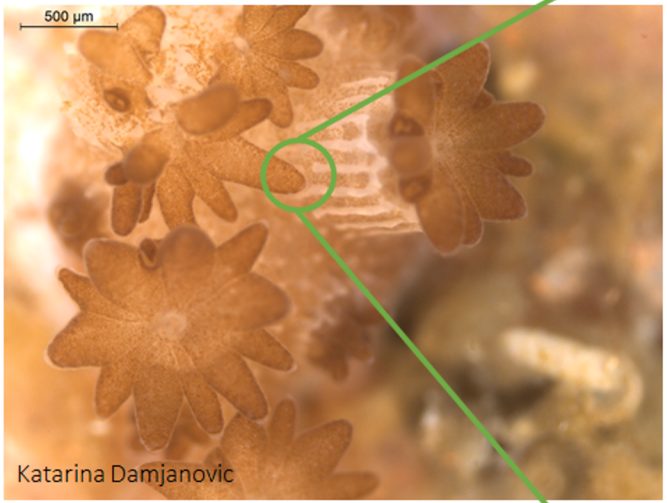
ARDC
Nectar
Research Cloud

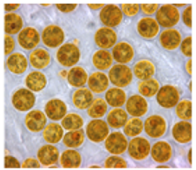
Linux/Unix/macOS command line

- 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)

Cnidarian **holobiont**

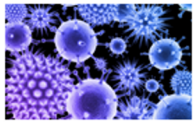
Coral



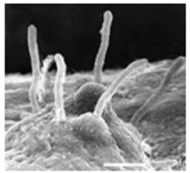


Symbiodiniaceae

Bacteria, Archaea



Viruses

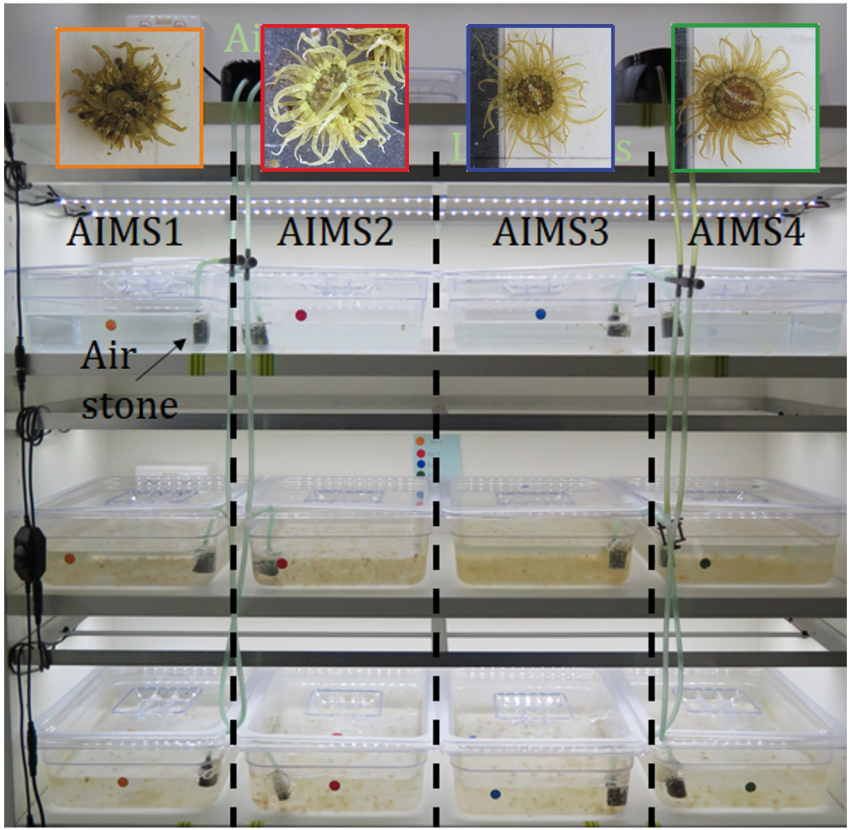


Fungi

Exaiptasia diaphana

Rohwer et al., 2002; Ricci et al., 2019

Background on data



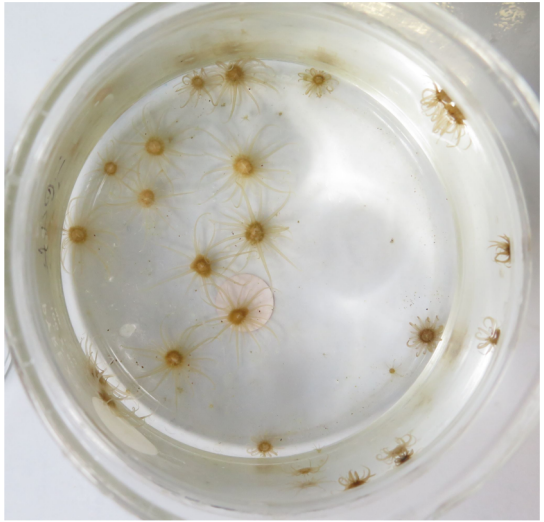
Short-Term Exposure to Sterile Seawater Reduces Bacterial Community Diversity in the Sea Anemone, *Exaiptasia diaphana*

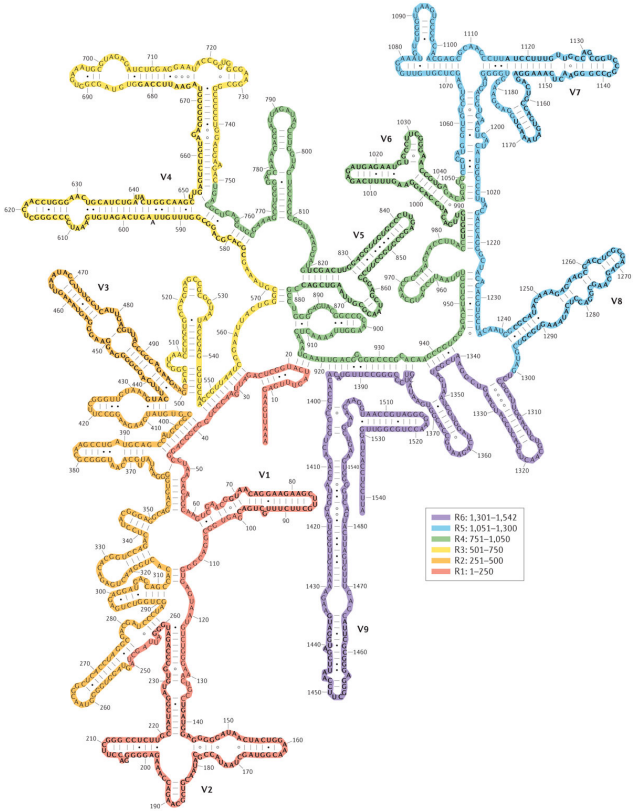
Ashley M. Dungan[1], Madeleine J. H. van Oppen[1,2] and Linda L. Blackall[1]*

[1] School of BioSciences, The University of Melbourne, Melbourne, VIC, Australia, [2] Australian Institute of Marine Science, Townsville, QLD, Australia

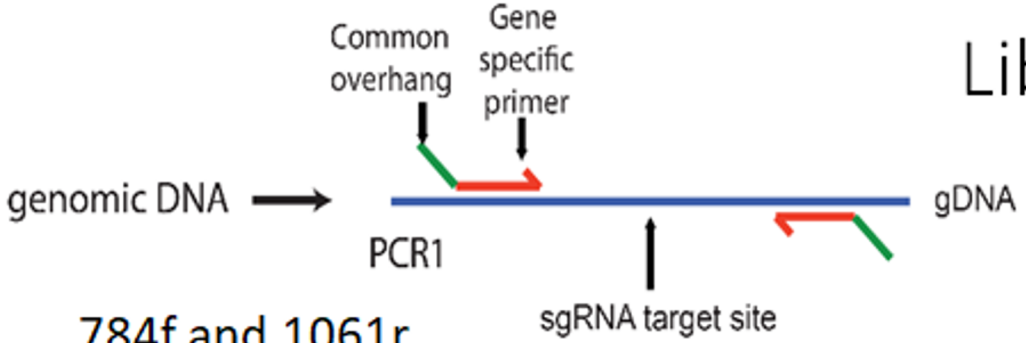
frontiers in Marine Science

Sterile SW
3 weeks



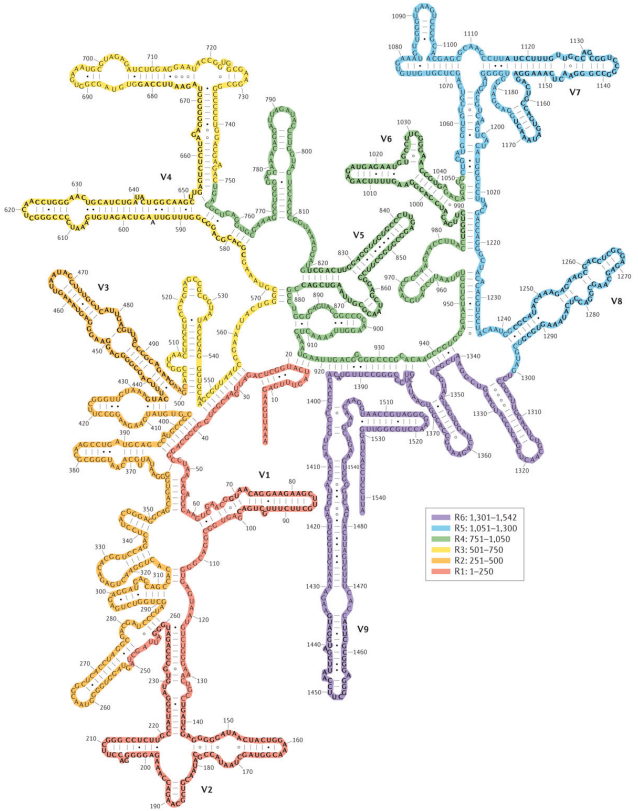

V1
V2
V3
V4
V5
V6
V7
V8
V9
R6: 1,301–1,542
R5: 1,051–1,300
R4: 751–1,050
R3: 501–750
R2: 251–500
R1: 1–250

Library prep @ WEHI



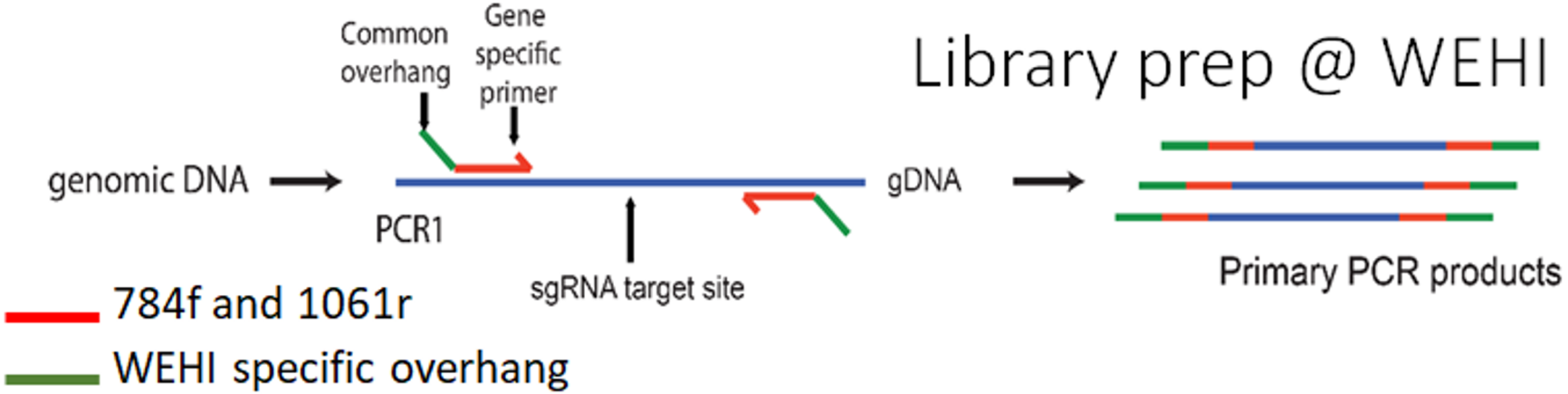
784f and 1061r

WEHI specific overhang

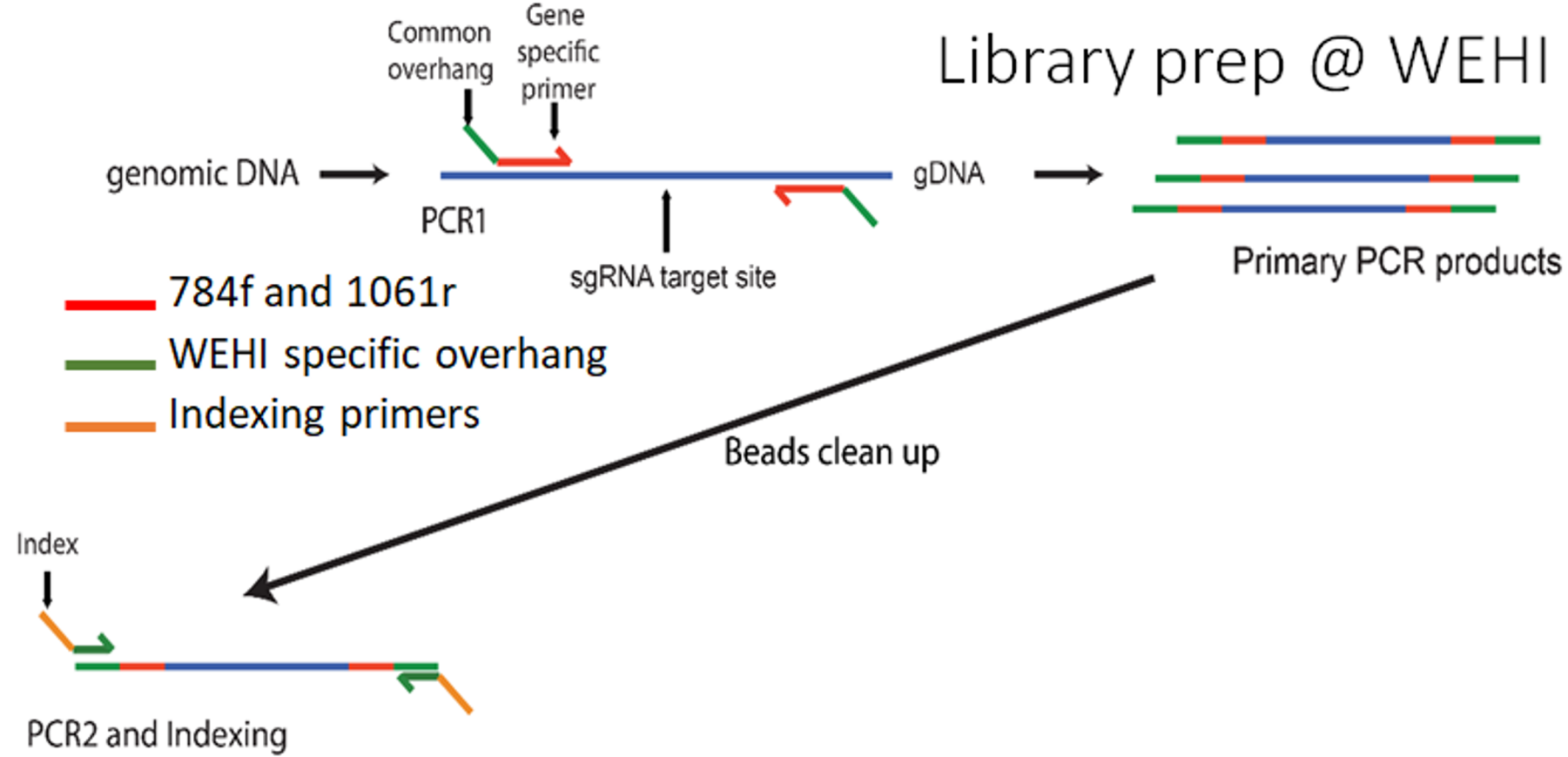


Nature Reviews | Microbiology

Library prep @ WEHI

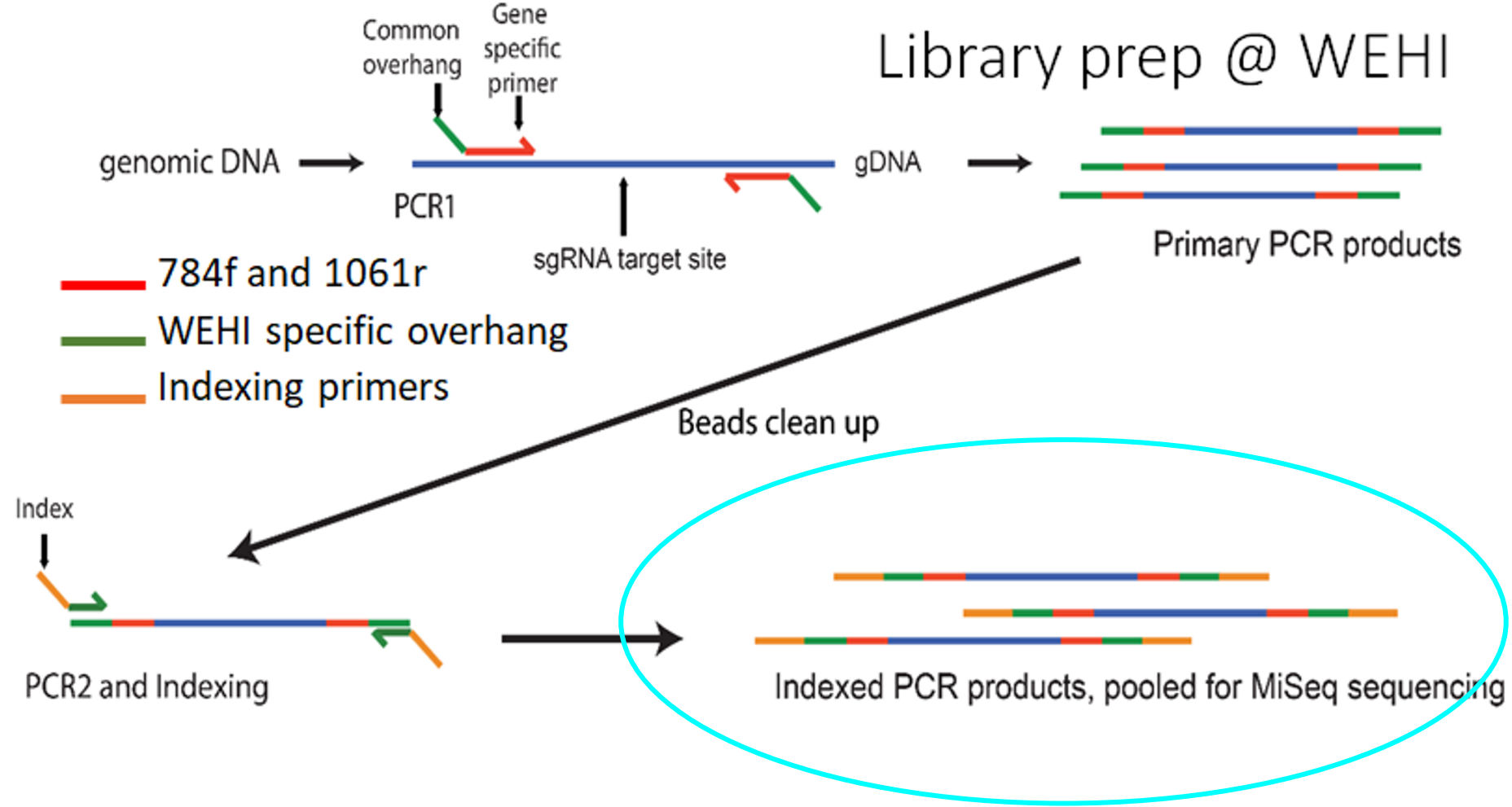


Library prep @ WEHI

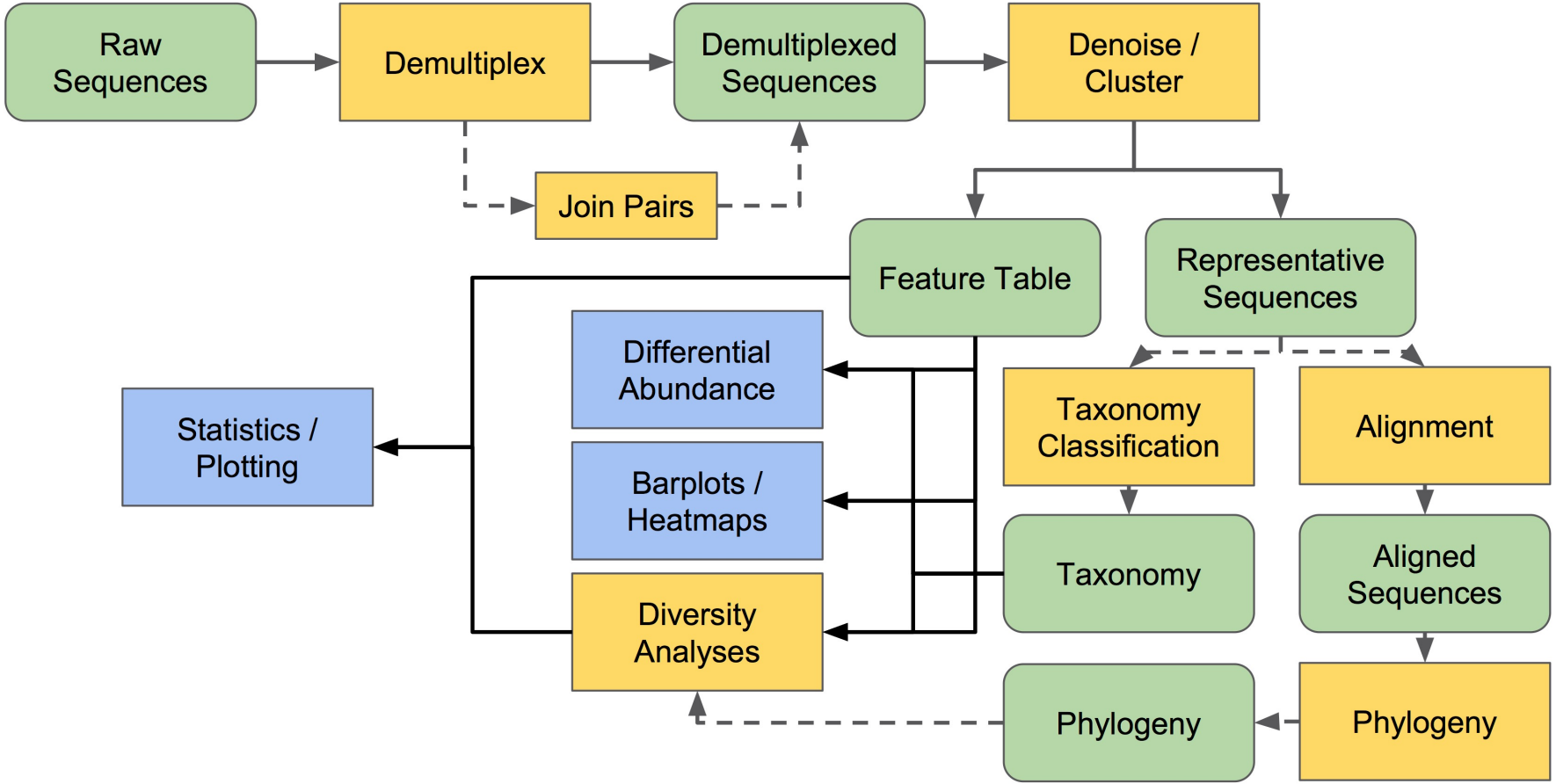


Adapted from Aubrey et al., 2014

Library prep @ WEHI



Adapted from Aubrey et al., 2014


Raw Sequences
Demultiplex
Demultiplexed Sequences
Denoise / Cluster
Join Pairs
Feature Table
Representative Sequences
Differential Abundance
Statistics / Plotting
Barplots / Heatmaps
Diversity Analyses
Taxonomy Classification
Alignment
Taxonomy
Aligned Sequences
Phylogeny
Phylogeny

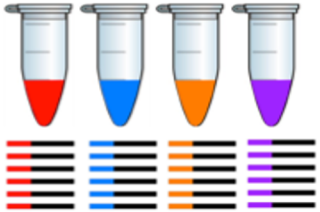
Import data into QIIME2

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

Multiplexed Data

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)

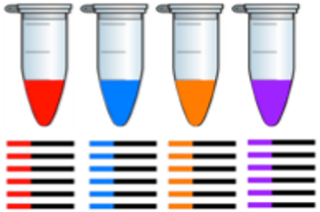
```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGC
TTGGTAAGTCCATGGTGAAATCCCTCGGCTCAACCGAGGAACTG
+
abaaaaa`^`a_]^\`\`\`a`^`]]]^^`a[VXGX``Z_\\\_^\a^SYOZVVSV
YGYVDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-6X_9267:1:1:25:267
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGT
GGCTTAAGCGCAGGGTTTAAGGCAATGG
+
aa^^[__^^^_^_^[^^[^^____ZZ[^
WWURZUYY]XXRZRNVTRTNTWUUU^VJ
@HWI-6X_9267:1:1:25:609
TACGTAGGGGGCAAGCGTTATCCGGATT
TGGACAAGTCTGATGTGAAAGGCTGGGG
+
aaab`aaa`aaaaaaaaaaaaaaa^aa
[I^^aZZ^WW^_^`ZZ_T]XY^^\^ZX\
@HWI-6X_9267:1:1:25:519
GACGGAGGATGCAAGTGTTATCCGGAAT
TTACTAAGTCAACTGTTAAATCTTGAGG
+
abaaaaaa`aaaaaa\aaaaaaa```aa
]]_Z_XX\\[]]]^`[\XTVX]`T_VZ[
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAAT
TAGGTAAGTCAGATGTGAAAGCCCCGGG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`
VH_PHOWZM[PTRPTRYUBBBBBBBBBB
```

barcodes.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:267
AAGAGAT
+
bbbbbbb
@HWI-6X_9267:1:1:25:609
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:519
ACAGCAG
+
bbbbbbb
@HWI-6X_9267:1:1:25:1109
ACAGCTA
+
bbbbbbb
@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
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAG
CGTACGTAGGCGGTTAGGTAAGTCAGATGTGAAAGCCCCGGG
CTCCACCTGGGAATGG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`[M]a`[VYa^_X^_
Z]NZ\`]TY\]_^RVH_PHOWZM[PTRPTRYUBBBBBBBBBB
BBBBBBBBBBBBBBBB
```

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)

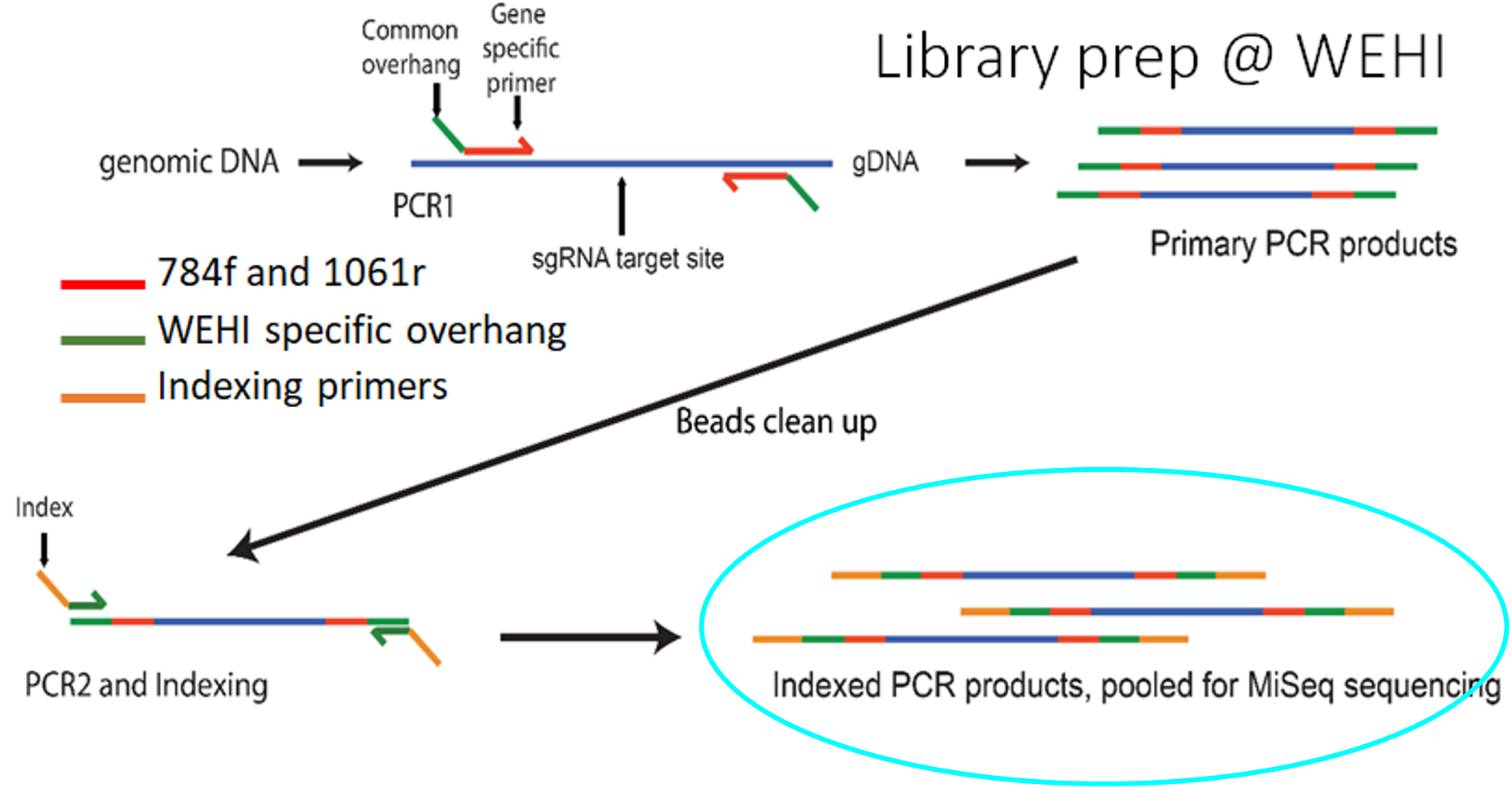
Unsure? **Make sure you ask the sequencing facility and know the answers to these specific details.**

Import data code

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

Library prep @ WEHI



Adapted from Aubrey et al., 2014

Cutadapt data 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 AGGATTAGATACCCTGGTA \ #F primer sequence (no overhang)
--p-front-r CRRCACGAGCTGACGAC \ #R primer sequence (no overhang)
--p-error-rate 0.20 \ #maximum allowed error rate, range 0-1. Play with this!
--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

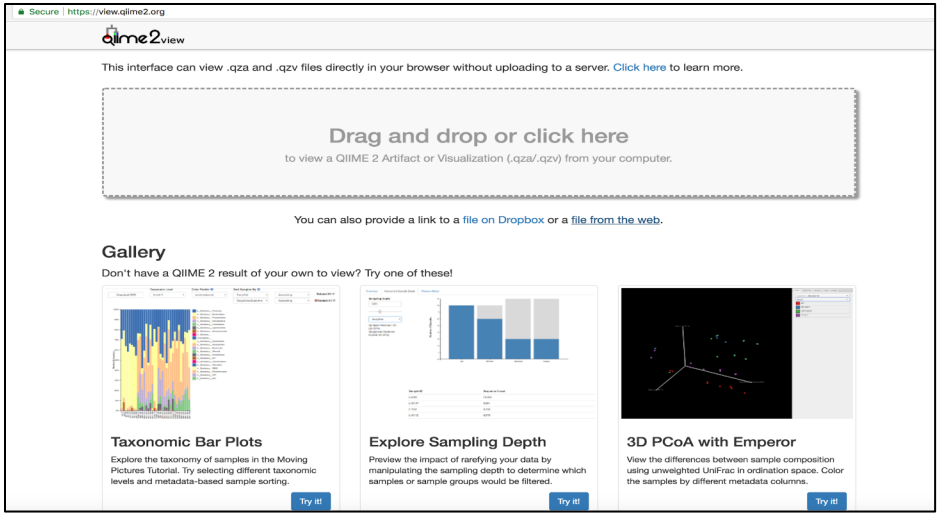
Summarize counts per sample

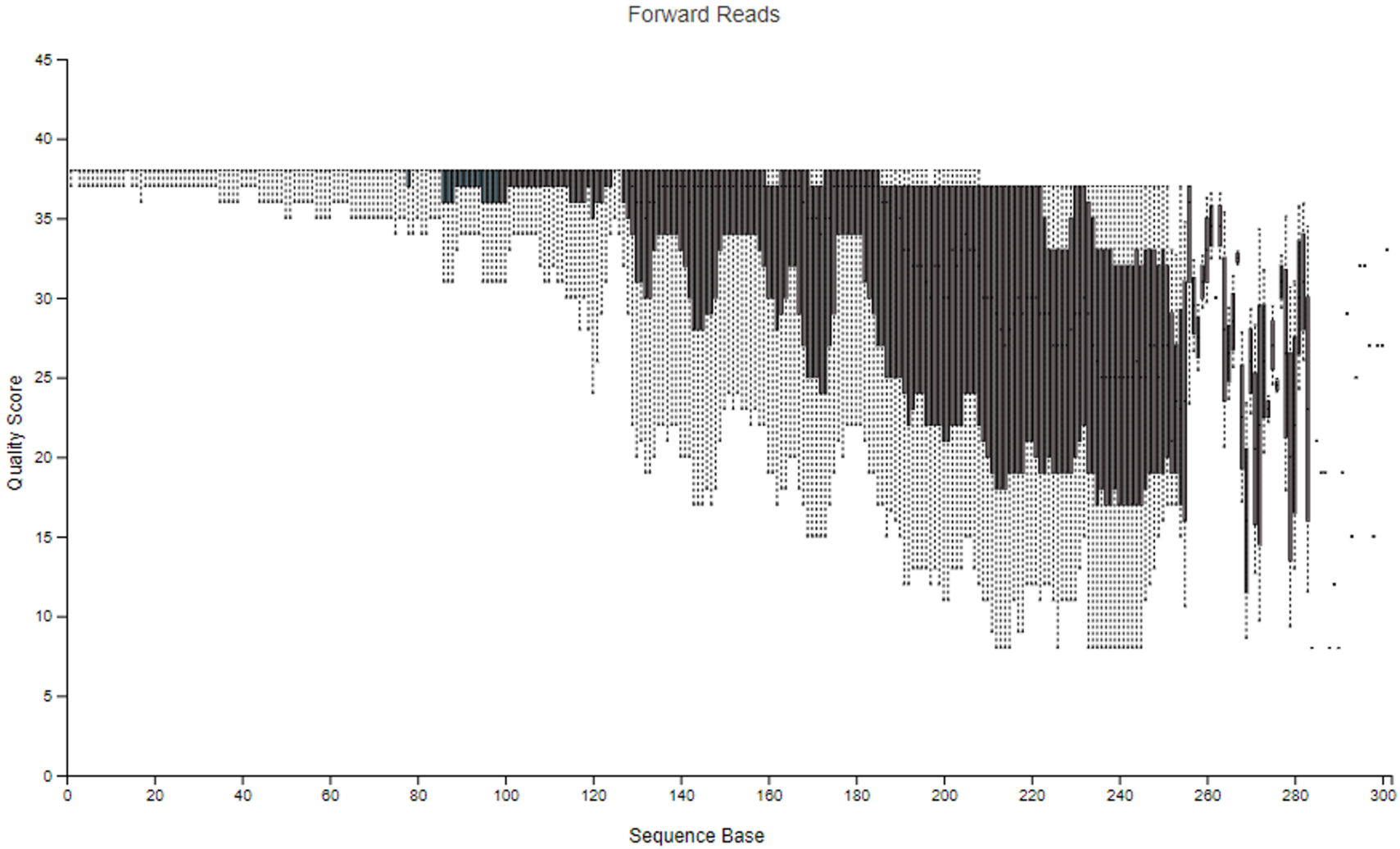
```
qiime demux summarize \ #we want to visualize the demultiplexed data
--i-data analysis/seqs_trimmed/trimmed_sequences.qza \ #location of data
--o-visualization analysis/visualisations/trimmed_sequences.qzv #output file
```

Accessing output files

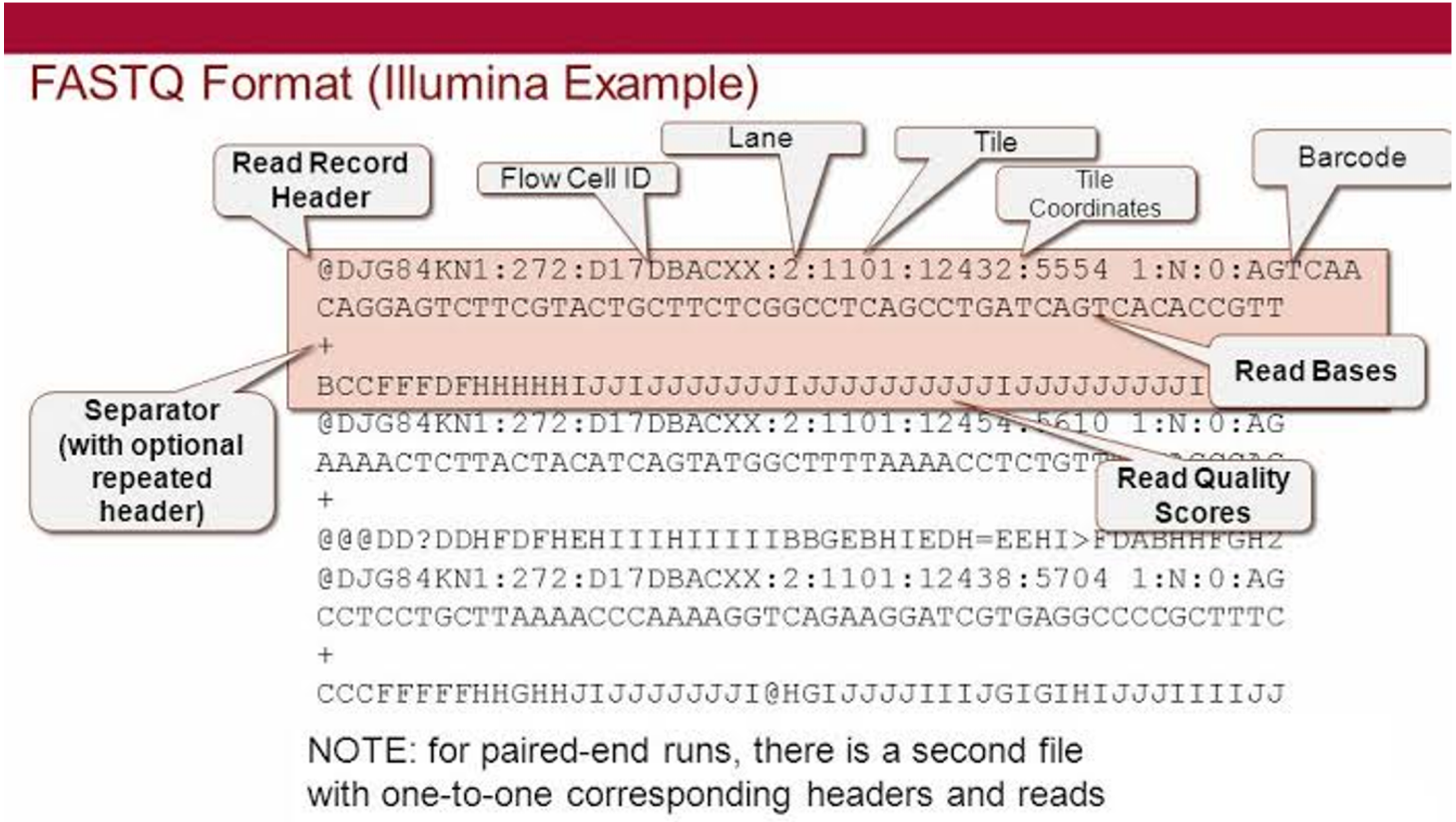
Use FileZilla to transfer to your local drive

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




Forward Reads
Quality Score
45
40
35
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Sequence Base

Quality Scores



NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads

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%

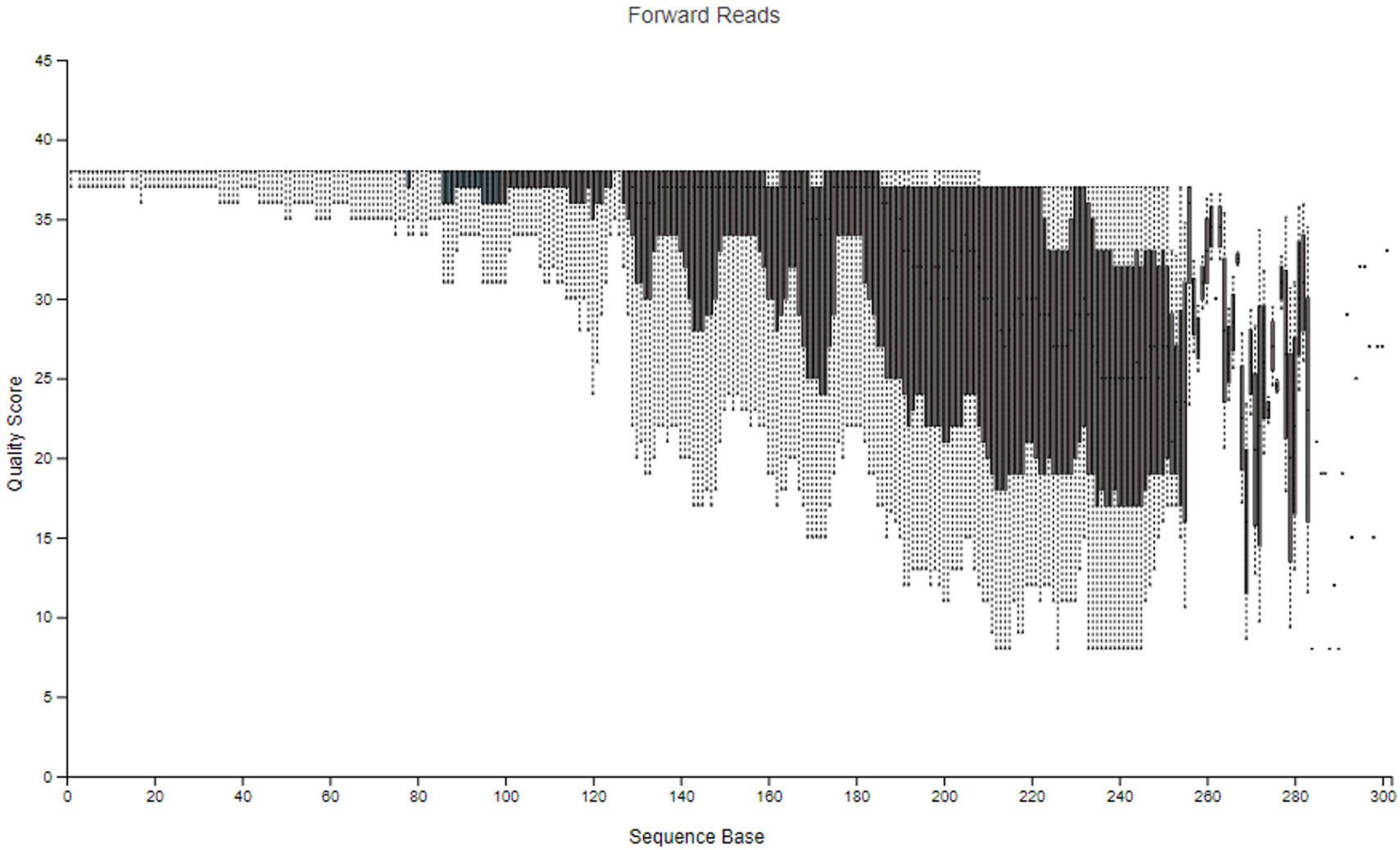
Quality Score Encoding

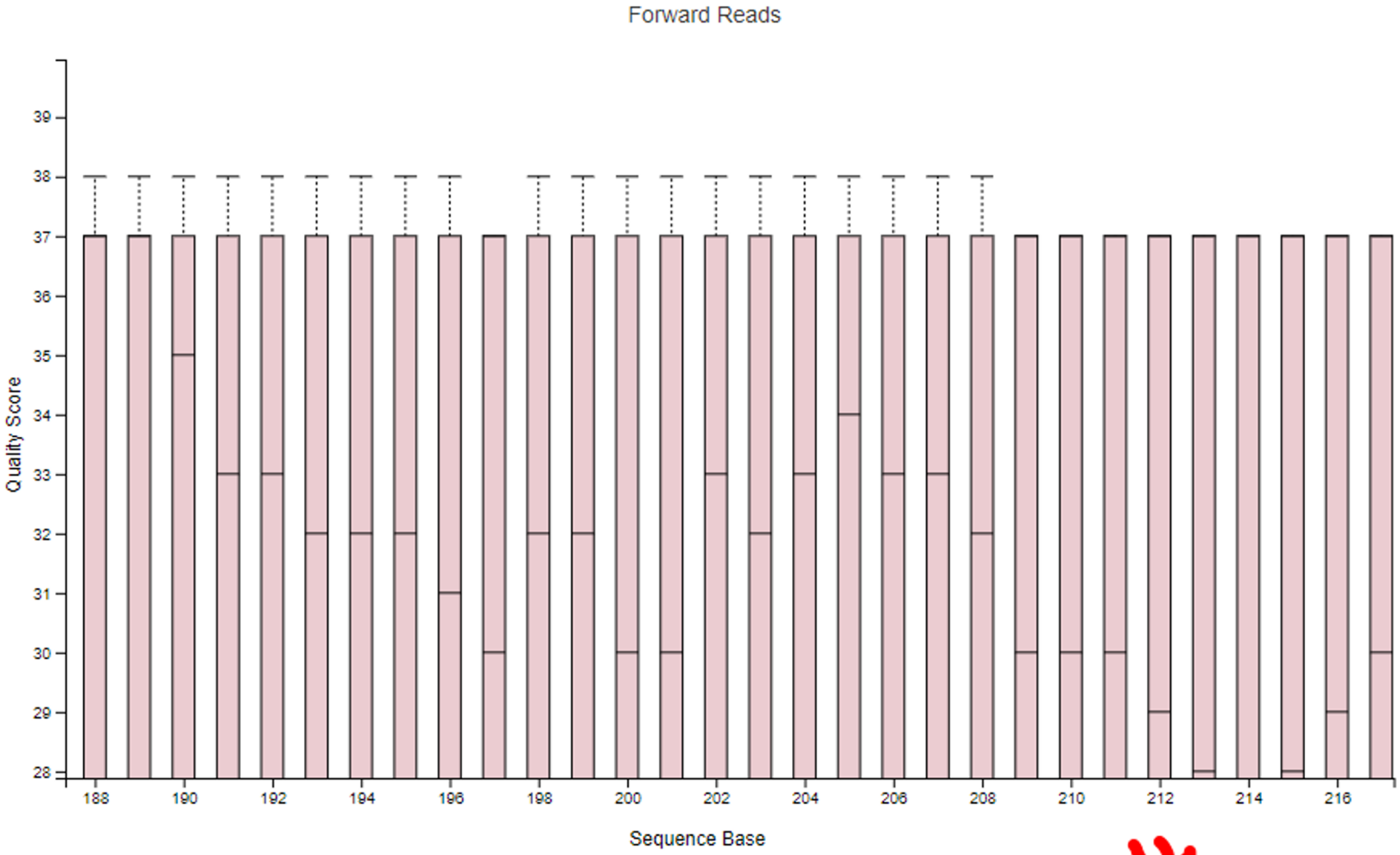
In FASTQ files, quality scores are encoded into a compact form, which uses only 1 byte per quality value. In this encoding, quality score is represented as the character with an ASCII code equal to its value + 33. The following table demonstrates relationship between the encoding character, its ASCII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


Forward Reads
Quality Score
45
40
35
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Sequence Base


Forward Reads
Quality Score
39
38
37
36
35
34
33
32
31
30
29
28
188
190
192
194
196
198
200
202
204
206
208
210
212
214
216
Sequence Base

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 genetic locus

DADA2: How does it do that?

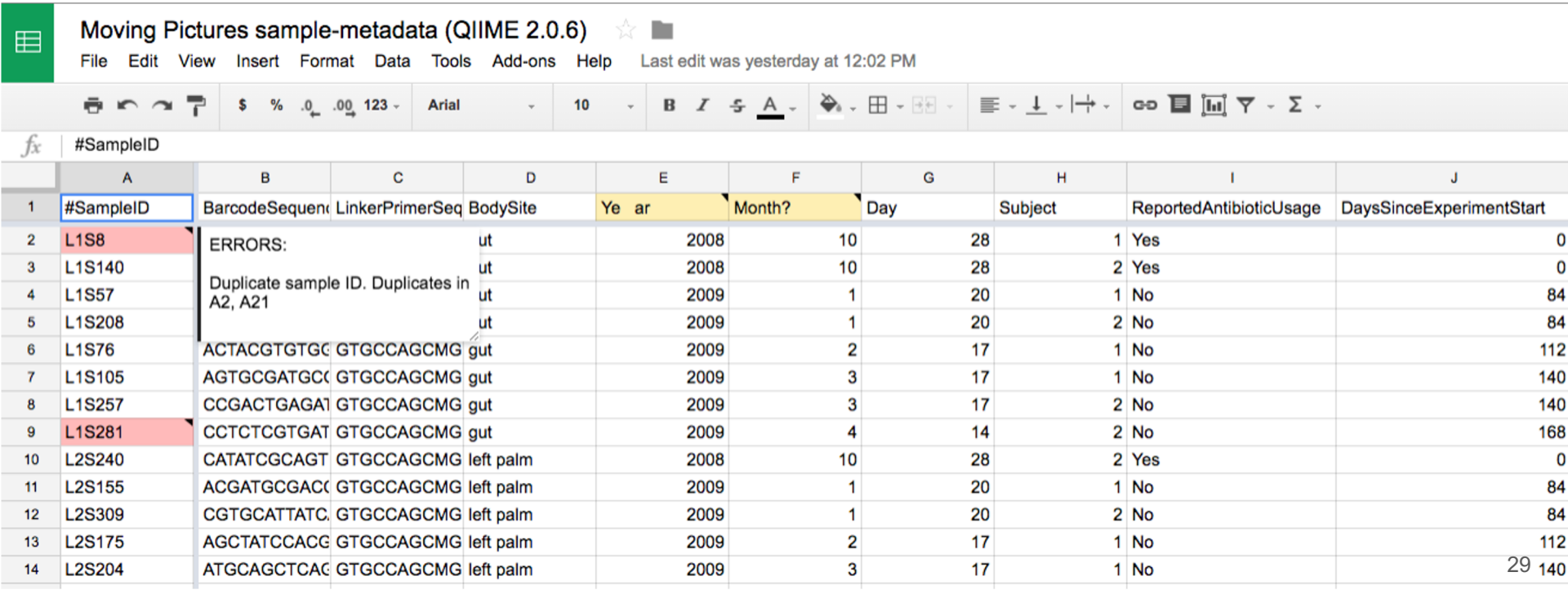
- Denoising (remove and/or correct noisy reads)
  - Filtering - user defined. Trims sequences to a specified length, removes sequences shorter than that length
  - Model errors within a read and between reads
  - Abundance - sequences too abundant to be explained by errors in sequencing are kept
  - 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 **A**mplicon **S**equencing **V**ariants (ASVs) or Features in some tutorials
- Chimera removal (identifying sequences that are two-parent chimeras of more abundant output sequences)

DADA2 data code

qiime dada2 denoise-paired \ #software-plugin-action
--i-demultiplexed-seqs analysis/seqs_trimmed/trimmed_sequences.qza \ #location #of data
--p-trunc-len-f **xxx** \ #position to truncate forward reads due to decrease in quality
--p-trunc-len-r **xxx** \ #position to truncate reverse reads due to decrease in quality
--p-n-threads 0 \ #number of cores; 0 = all cores used = faster
--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

Arial 10

#SampleID

	A	B	C	D	E	F	G	H	I	J
1	#SampleID	BarcodeSequen	LinkerPrimerSeq	BodySite	Ye ar	Month?	Day	Subject	ReportedAntibioticUsage	DaysSinceExperimentStart
2	L1S8			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	AGTGCGATGCC	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	CGTGCATTATC	GTGCCAGCMG	left palm	2009	1	20	2	No	84
13	L2S175	AGCTATCCACG	GTGCCAGCMG	left palm	2009	2	17	1	No	112
14	L2S204	ATGCAGCTCAC	GTGCCAGCMG	left palm	2009	3	17	1	No	140

ERRORS:

Duplicate sample ID. Duplicates in A2, A21

Head to tutorial and complete Sections 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
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGA
AATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGA
AAGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTT
AAGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTA
AATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGA
AAGCCCCGGGCTCCACCTGGGAATGG
```

Taxonomic assignment of observed sequences.

Reference Database
Silva, Greengenes, etc.

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGA
AATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGA
AAGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTT
AAGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTA
AATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGA
AAGCCCCGGGCTCCACCTGGGAATGG
```

FeatureData[Sequence]

```
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGG
TGACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGG
TGGCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAG
GCTGGGGCTCAACCCCGGGACGGTTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTA
G                                                                    GAA
>
T
A
G
>
T
A
G
```

FeatureData[Taxonomy]

reference-sequence-1	Bacteria; Proteobacteria; Gammaproteobacte
reference-sequence-2	Bacteria; Bacteroidetes; Flavobacteria; Fl
reference-sequence-3	Bacteria; Proteobacteria; Deltaproteobacte
reference-sequence-4	Archaea; Euryarchaeota; DSEG; 104A5

Taxonomic assignment of observed sequences.

Reference Database
Silva, Greengenes, etc.

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGA
AATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGA
AAGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTT
AAGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTA
AATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGA
AAGCCCCGGGCTCCACCTGGGAATGG
```

FeatureData[Sequence]

```
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGG
TGACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGG
TGGCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAG
GCTGGGGCTCAACCCCGGGACGGTTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTA
G                                                                    GAA
>
T
A
G
>
T
A
G
```

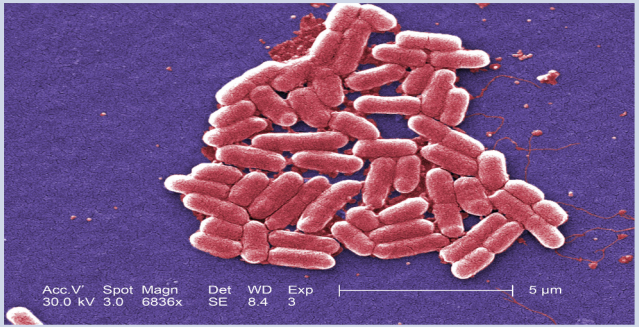
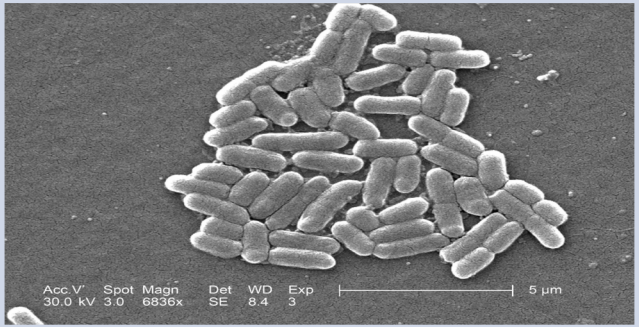
FeatureData[Taxonomy]

reference-sequence-1	Bacteria; Proteobacteria; Gammaproteobacte
reference-sequence-2	Bacteria; Bacteroidetes; Flavobacteria; Fl
reference-sequence-3	Bacteria; Proteobacteria; Deltaproteobacte
reference-sequence-4	Archaea; Euryarchaeota; DSEG; 104A5

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

FeatureData[Taxonomy]

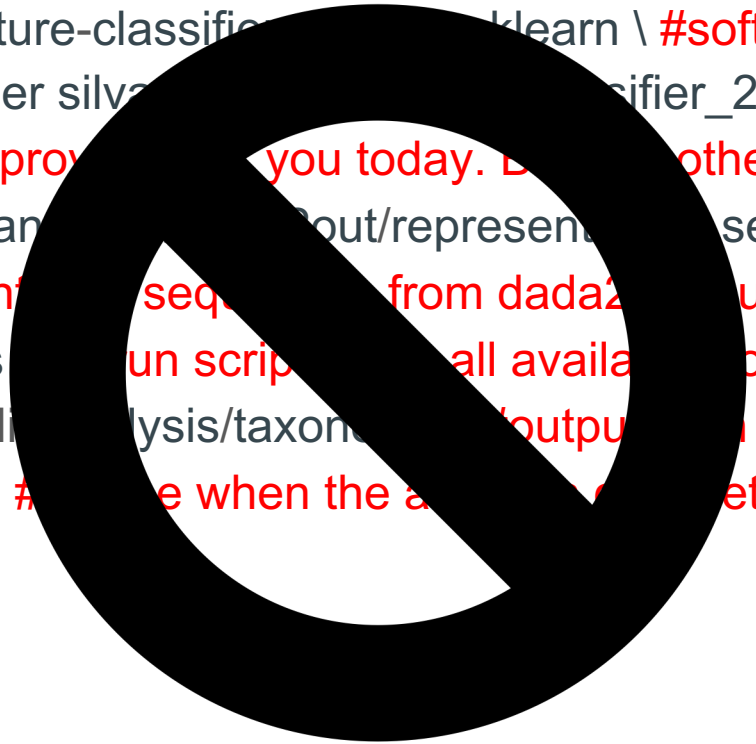
feature5	Bacteria; Proteobacteria
feature4	Bacteria; Proteobacteria
feature2	Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales
feature1	Bacteria; Proteobacteria
feature3	Bacteria; Proteobacteria; Deltaproteobacteria

	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	--

E. coli: http://media-3.web.britannica.com/eb-media//87/141087-050-24850517.jpg
http://upload.wikimedia.org/wikipedia/commons/d/d3/Staphylococcus_aureus_VISA_2.jpg

Assign taxonomy data code

```
qiime feature-classifi        klearn \ #software-plugin-action
--i-classifier silva          sifier_2021-4.qza \ #location classifier. This
#file was pro          you today. B          others here.
--i-reads an          out/represent          sequences.qza \ #dereplicated
#represent          seq          from dada2          ut
--p-n-jobs          un scrip          all availa          ores
--output-di          lysis/taxon          /outpu
--verbose #          e when the a          ete
```



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 data 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
```

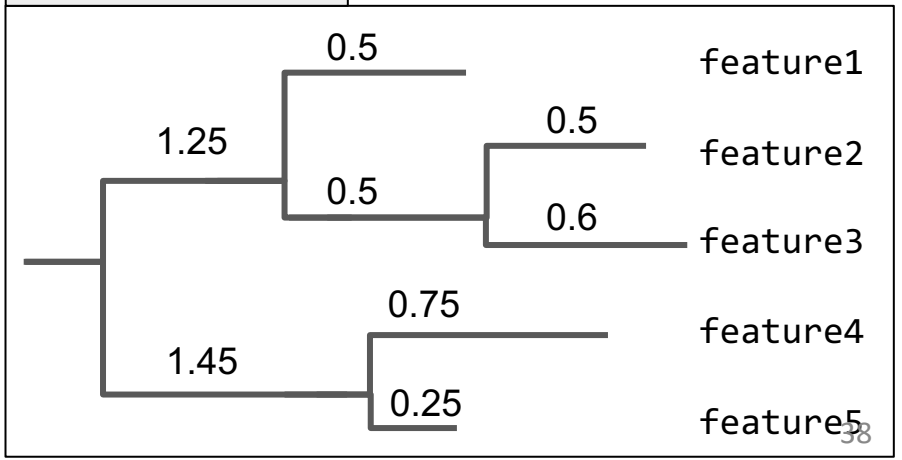
Phylogenetic reconstruction of observed sequences

FeatureData[Sequence]

```
>taxon5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGA
AATCCCTCGGCTCAACCGAGGAACTG
>taxon4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGA
AAGGCTGGGGCTCAACCCCGGGACGG
>taxon2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTT
AAGGCAATGGCTTAACTATTGTTCTC
>taxon1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTA
AATCTTGAGGCTCAACCTCGAAATCG
>taxon3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGA
AAGCCCCGGGCTCCACCTGGGAATGG
```

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

Phylogeny[Rooted]



Build phylogenetic tree code

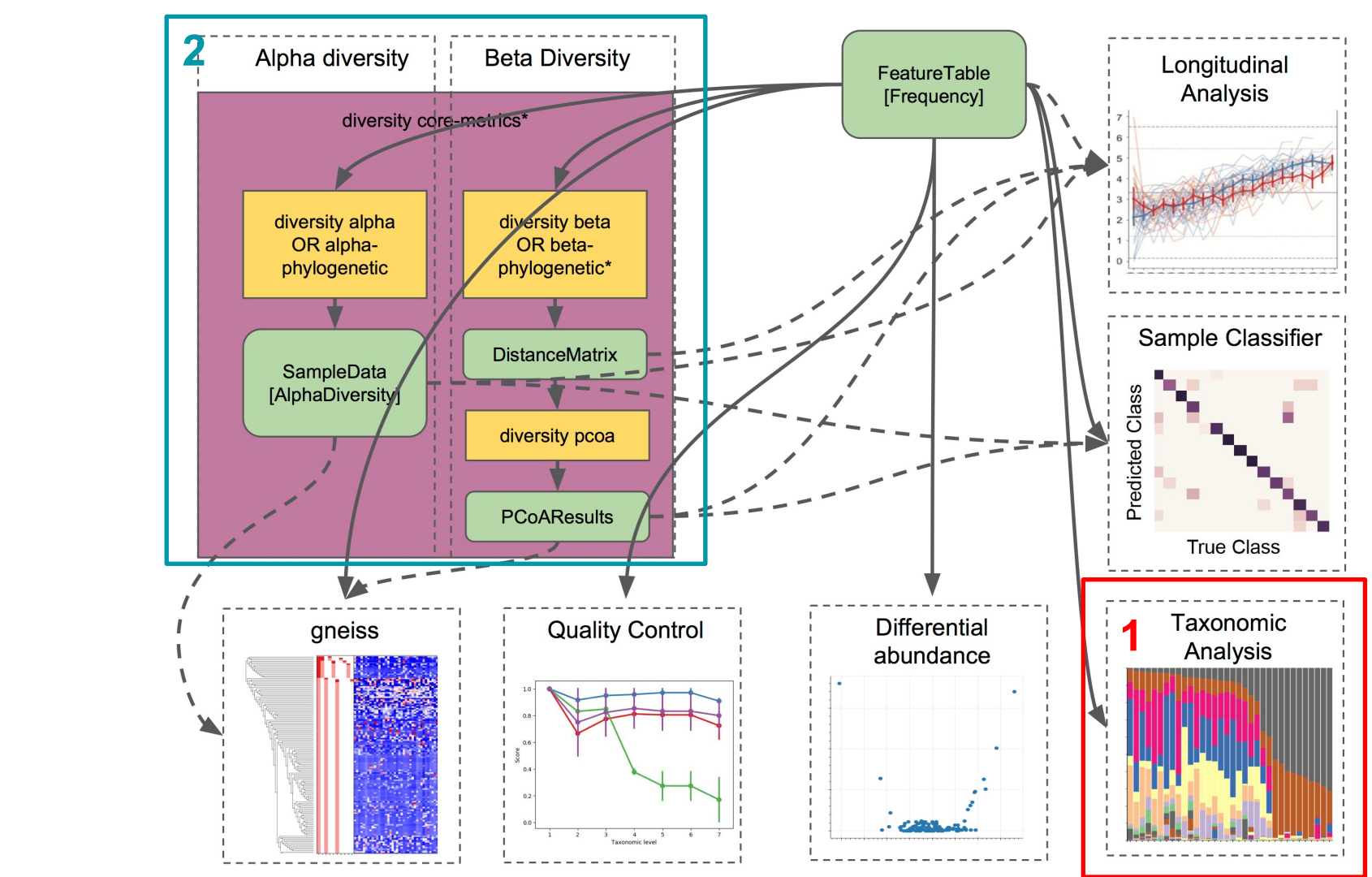
```
qiime phylogeny align-to-tree-mafft-fasttree \ #software-plugin-action
--i-sequences analysis/dada2out/representative_sequences.qza \ #sequences to #align
--o-alignment analysis/tree/aligned_16s_representative_seqs.qza \ #perform an alignment
--o-masked-alignment analysis/tree/masked_aligned_16s_representative_seqs.qza \ #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

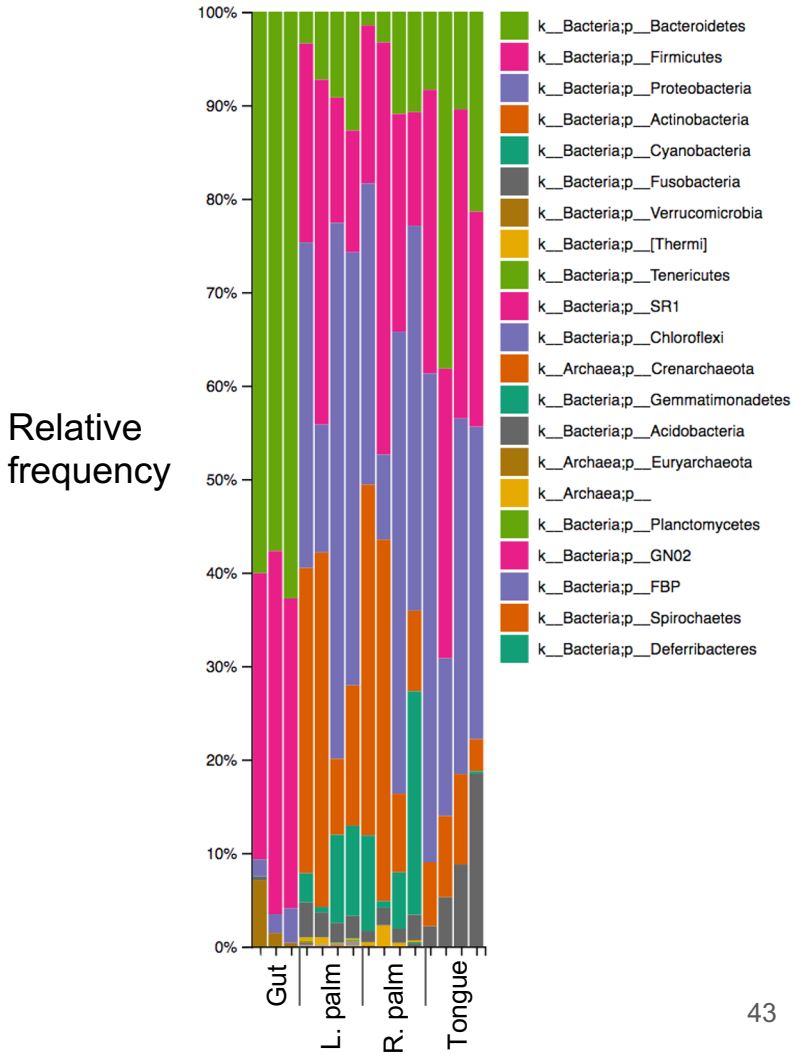
Basic visualizations and statistics


2
Alpha diversity
Beta Diversity
diversity core-metrics*
diversity alpha OR alpha-phylogenetic
diversity beta OR beta-phylogenetic*
SampleData [AlphaDiversity]
DistanceMatrix
diversity pcoa
PCoAResults
FeatureTable [Frequency]
Longitudinal Analysis
Sample Classifier
Predicted Class
True Class
gneiss
Quality Control
Taxonomic level
Differential abundance
1
Taxonomic Analysis

Visualizing taxonomic profiles

Interactive barplots support:

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



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 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).

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

Does anything concern you about this table?

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	84	1	73	198	2
e375	24	2	44	176	1
4gd8	11	0	10	30	0
9872	0	0	25	2	0

Diversity metrics are often impacted by the total frequency observed in samples, such that in this example `4gd8` might look more similar to `9872` than to `e375`.

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	84	1	73	198	2
e375	24	2	44	176	1
4gd8	11	0	10	30	0
9872	0	0	25	2	0

	Total frequency
4ac2	358
e375	247
4gd8	51
9872	27

This is most commonly handled by *rarefaction*, which is currently* a necessary evil. Frequencies are subsampled without replacement until all samples have the same total. Samples with fewer sequences than your *even sampling depth* will be filtered out of the feature table.

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
g345	11	1	10	29	0
c5d7	4	0	7	40	0
f6ee	11	0	10	30	0
~~cfd3~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~

	Total frequency
g345	51
c5d7	51
f633	51
~~cfd3~~	~~0~~

* A good project would be developing diversity metrics that are not sensitive to total frequency.

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 9062 \ #maximum rarefaction depth. Typically use the median
#number of reads from 16s_table_filtered.qzv file
--m-metadata-file 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

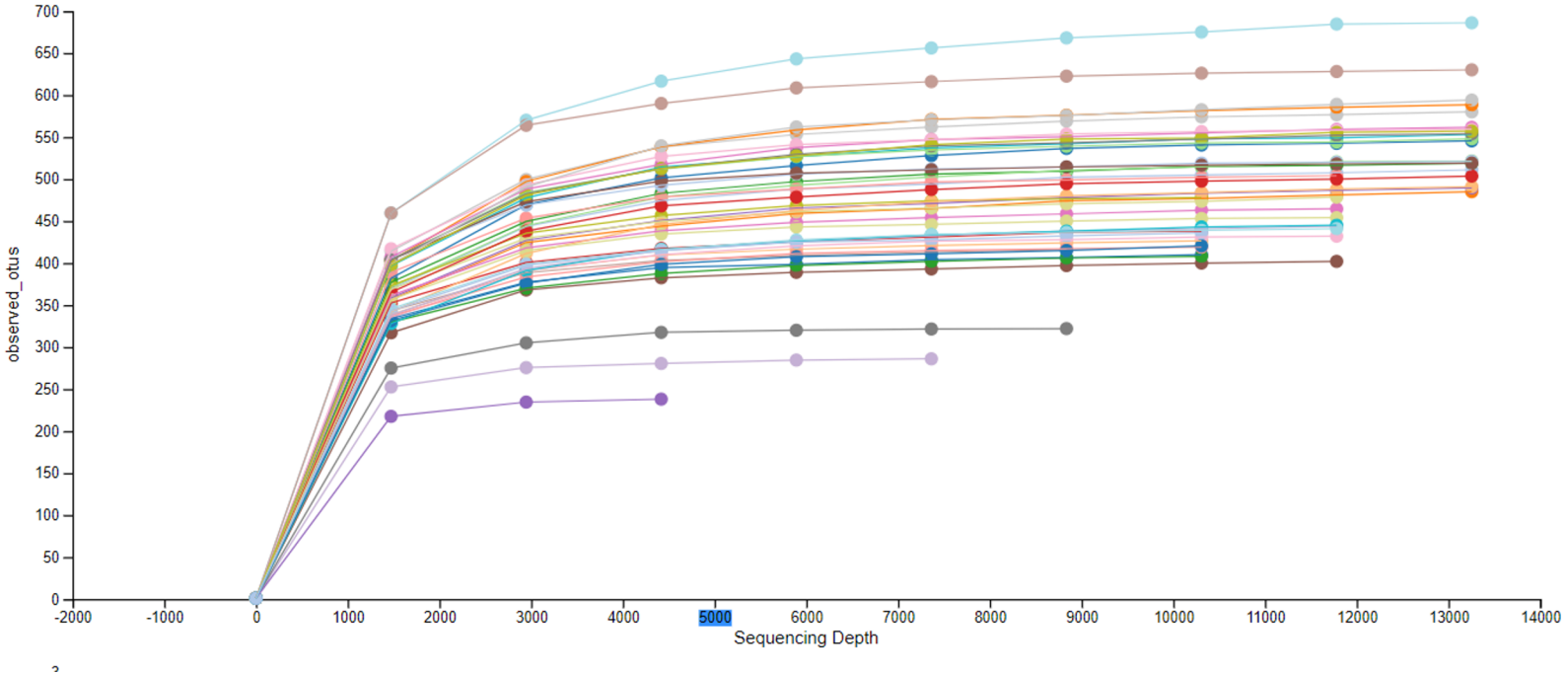
Download CSV

Metric

observed_otus

Sample Metadata Column

BarcodeSequence



Phylogenetic diversity metrics incorporate evolutionary relationships between taxa, but assume that we know what those relationships are. These require a phylogenetic tree.

- Weighted Unifrac
- Unweighted Unifrac*

Non-phylogenetic diversity metrics assume that all taxa are equally related and therefore make no assumptions about evolutionary relationships. No tree required.

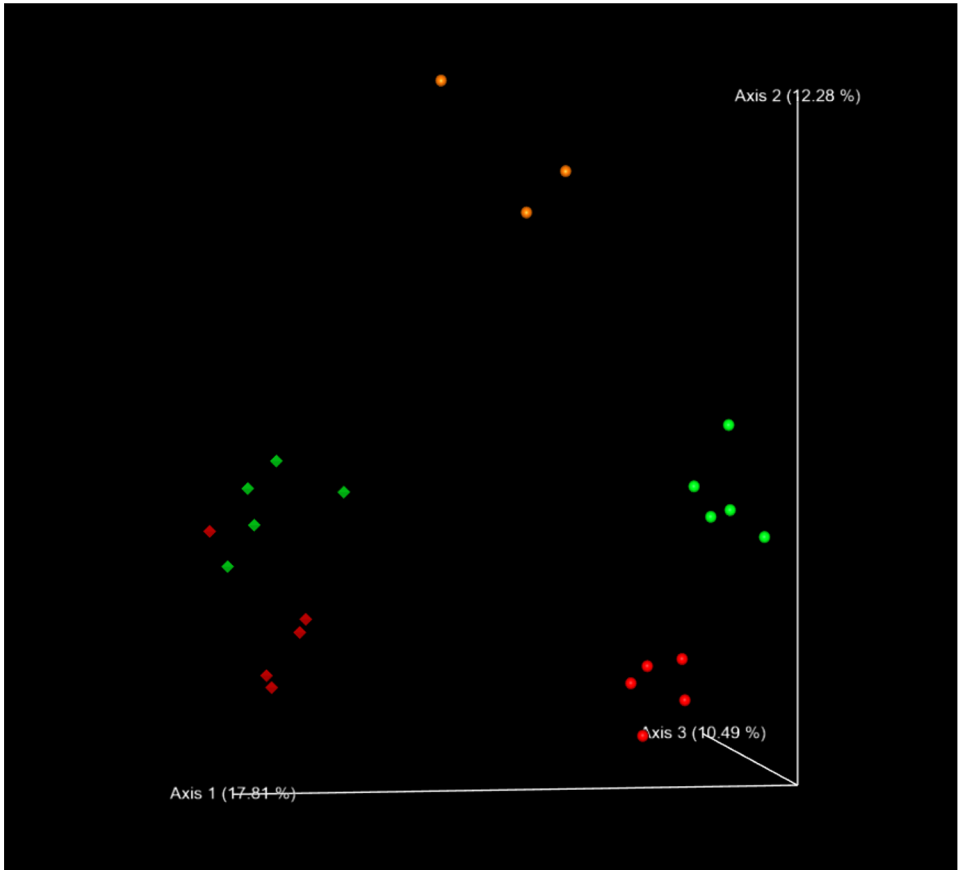
- Bray-Curtis
- Jaccard*

*Doesn’t consider abundance, just presence/absence

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 5583 \ #selected based on rarefaction curves and read counts
in samples
--m-metadata-file metadata.tsv \ #path to metadata file
--o-visualization analysis/diversity_metrics\ #output folder
```

Emperor Plots = PCoA



Color = Genotype

Shape = SW treatment

Alpha and Beta Diversity Stats

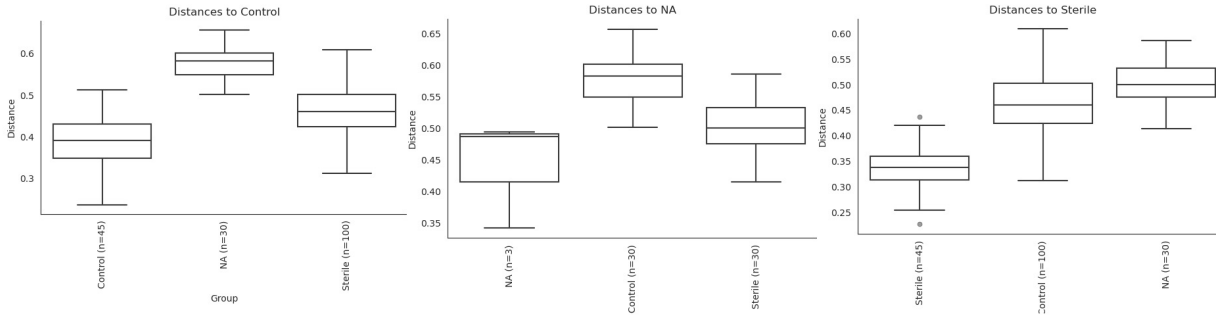
qiime2view File: unweighted-unifrac-environment-significance.qzv Visualization Details Provenance

Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	23
number of groups	3
test statistic	5.896316
p-value	0.001
number of permutations	999

Group significance plots

Download raw data as TSV

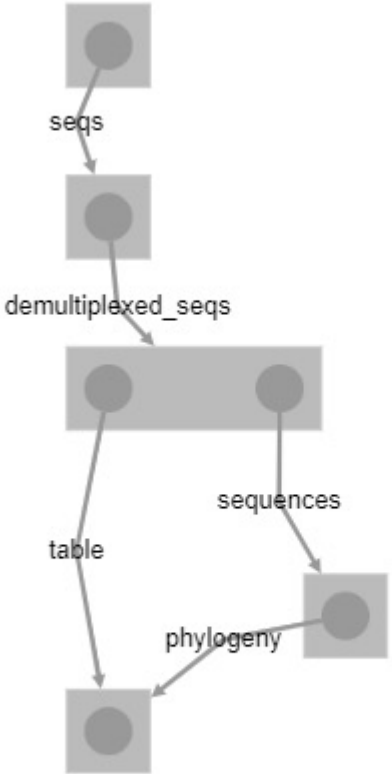


Download as PDF Download as PDF Download as PDF

Pairwise permanova results

Download CSV

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Control	NA	13	999	5.575155	0.002	0.003
	Sterile	20	999	6.895129	0.001	0.003
NA	Sterile	13	999	4.676336	0.009	0.009



Provenance

Head to tutorial and complete Section 4

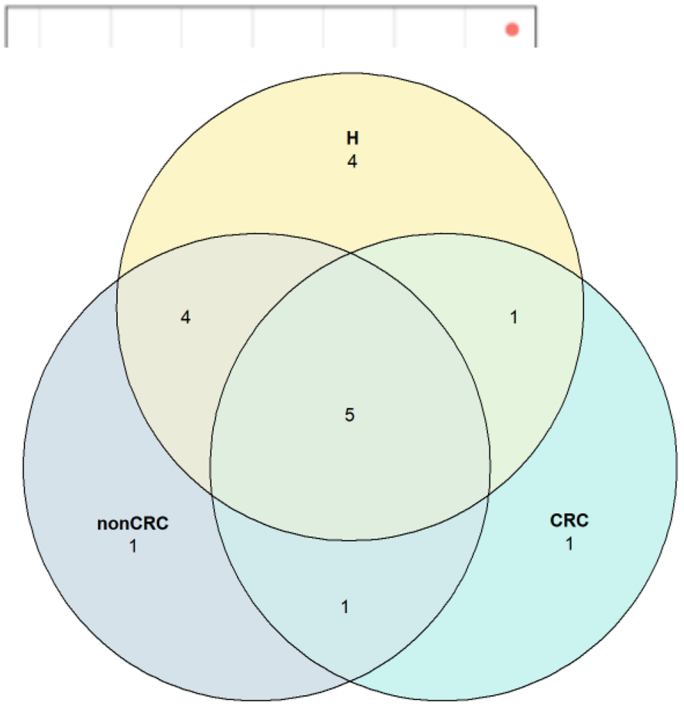
Section 4: Basic visualizations and statistics

Rarefaction code must be run consecutively (i.e. one person at a time within a group).

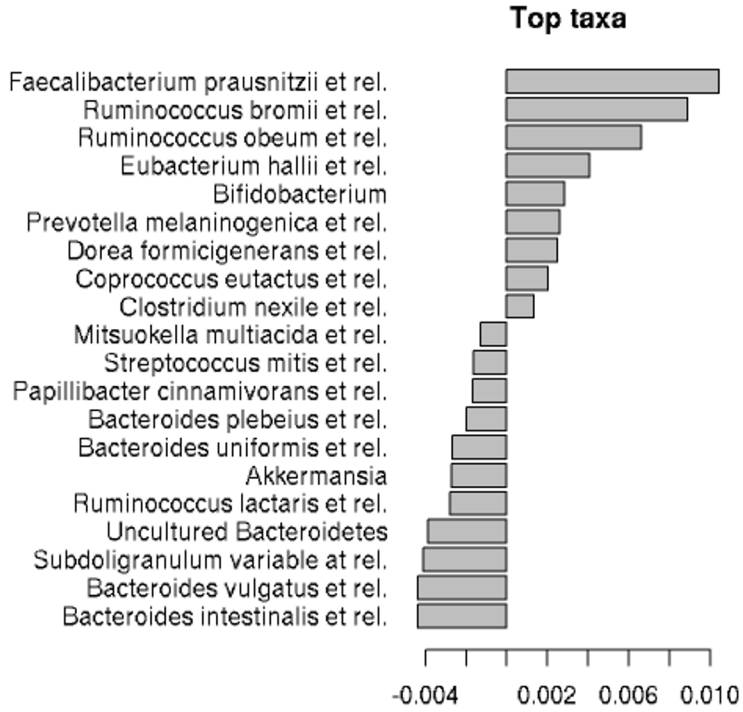
	A	B	C	D	E	F	G	H	I	J	K
1	# Constructed from biom file										
2	#OTU ID	AN002.M04_	AN003.M2	AN007.MC	AN022.MC	AN023.M2	AN025.M3	AN036.MC	AN038.M3	AN040.MC	AN045.MC	A
3	08da88cfc658fe0b3b360a213243a747	0	0	0	0	0	0	0	0	0	0	
4	c570d55b6c96a3393a101e2bed65872d	0	0	0	0	0	0	0	0	0	0	
5	981987ed4a2c01ff40ad458140d27949	0	0	0	0	0	0	0	0	0	0	
6	aab29ab9edbee32f63202a95b0090548	0	0	0	0	0	0	0	0	0	0	
7	d73ac03427201aa660bb14a84f053043	0	0	0	0	0	0	0	0	0	0	
8	fddab79ff073446b95c1532828a4d02e	0	0	0	0	0	0	0	0	0	0	
9	f7106e49bf3f73cb8dbf7ef7a4384f34	0	0	0	0	0	0	0	0	0	0	
10	c76f907623b1f0475eca537b9b70dd8b	0	0	0	0	0	0	0	0	0	0	
11	99664aa88271cbda49314da4a8eb7955	0	0	10	0	0	0	0	0	0	0	
12	42175a193304f0218973320abdac8e45	0	0	14	0	0	60	0	31	0	0	
13	99c46567fcb0002d3af444ce106a7f1d	0	0	0	0	0	0	0	0	0	0	
14	4dfb9be11f244c8b6554fd514fea6b20	0	63	0	0	179	70	0	0	0	8	
15	5adbe9ff29201074a091b243e33458fc	0	0	0	0	0	0	0	0	0	0	
16	7ca2d08e221882943253a52d7164e8db	0	0	0	0	0	0	0	0	0	0	
17	cf77d06c40fa9994c61d327ed719c72a	0	0	0	0	0	0	0	0	0	0	
18	6537101cb98fac0fe4bae47f368ea5d6	0	0	0	0	0	0	0	0	0	0	
19	7fc0b06b13fd939a3c80900b01bfa0ef	0	0	0	0	0	0	0	0	0	0	
20	eb29b79633aa2f26db590ecc9a3d2f3a	0	0	0	0	0	0	0	0	0	0	
21	284fdf2bd0470394cb34f8b7e7c0ac91	9	22	13	19	24	19	5	0	0	9	
22	189a68b4d66510db9e33e8b35d07fc94	0	0	0	0	0	0	0	0	0	0	
23	9c37ce1883e1babb8f405c43b81e2130	0	0	0	0	0	0	0	0	0	0	

QIIME2 → R


samples
0.5
0.0
-0.5
NMDS2


H
4
4
1
5
nonCRC
1
CRC
1
1


SampleType


Top taxa
Faecalibacterium prausnitzii et rel.
Ruminococcus bromii et rel.
Ruminococcus obeum et rel.
Eubacterium hallii et rel.
Bifidobacterium
Prevotella melaninogenica et rel.
Dorea formicigenerans et rel.
Coprococcus eutactus et rel.
Clostridium nexile et rel.
Mitsuokella multiacida et rel.
Streptococcus mitis et rel.
Papillibacter cinnamivorans et rel.
Bacteroides plebeius et rel.
Bacteroides uniformis et rel.
Akkermansia
Ruminococcus lactaris et rel.
Uncultured Bacteroidetes
Subdoligranulum variable at rel.
Bacteroides vulgatus et rel.
Bacteroides intestinalis et rel.
-0.004
0.002
0.006
0.010

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

Useful QIIME2 pages

- User Glossary
- Core concepts
- QIIME2 Overview with Flowcharts