MICROBIOME SEQUENCING AND ANALYSIS

NOLAN KEITH NEWMAN

IN-BIOS 5000 / 9000

NOVEMBER 6TH, 2024

UNIVERSITY OF OSLO

NCMM

SCHEDULE

9.15-12.00 – Microbiome sequencing and analysis (with 10 min break @ 10.45)

12.00-13.15 – Lunch

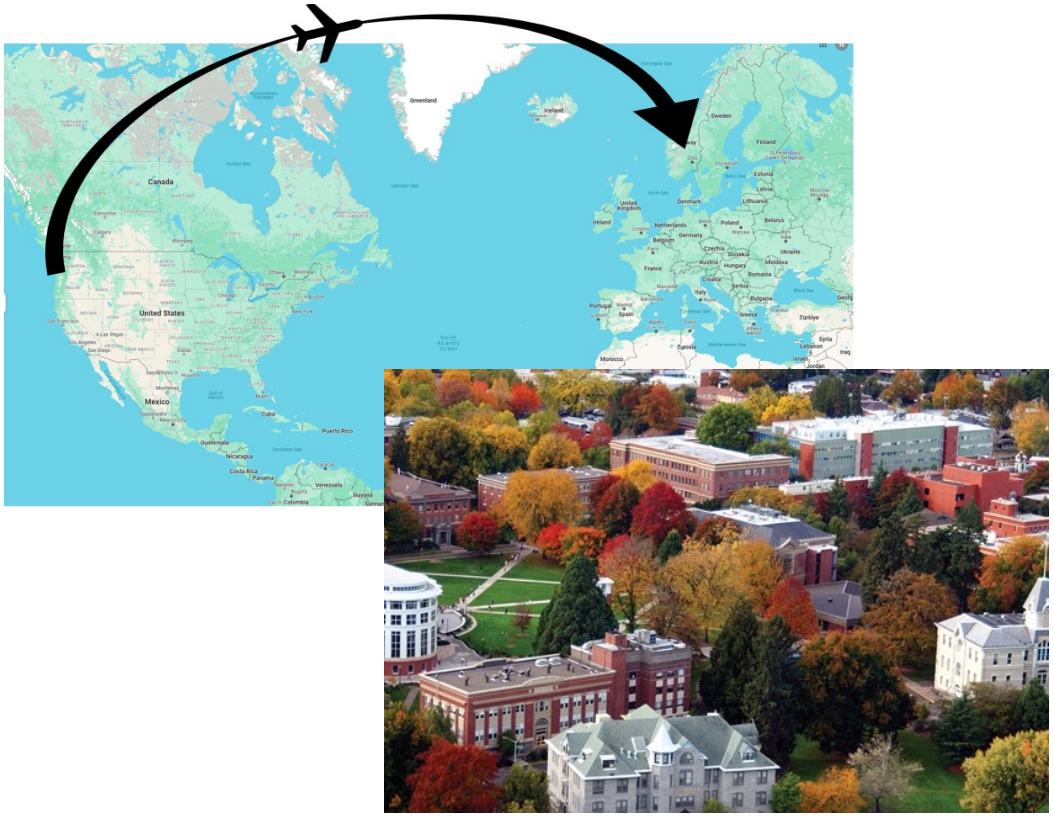
13.15-17.00 – Multi-omic network analysis (with 15 min break @ 15.00)

So you have to listen to me talk for a total of
approximately 6 hours today…

I am so sorry.

ABOUT ME

- PhD at Oregon State University
  - Computational biology
  - Microbiology
  - Pharmacology
  - Systems biology
- Main research topic: Using computational biology to identifying specific microbes that influence disease



QUICK DISCLAIMER

I will not know EVERYTHING there is to know about all techniques and software out there.

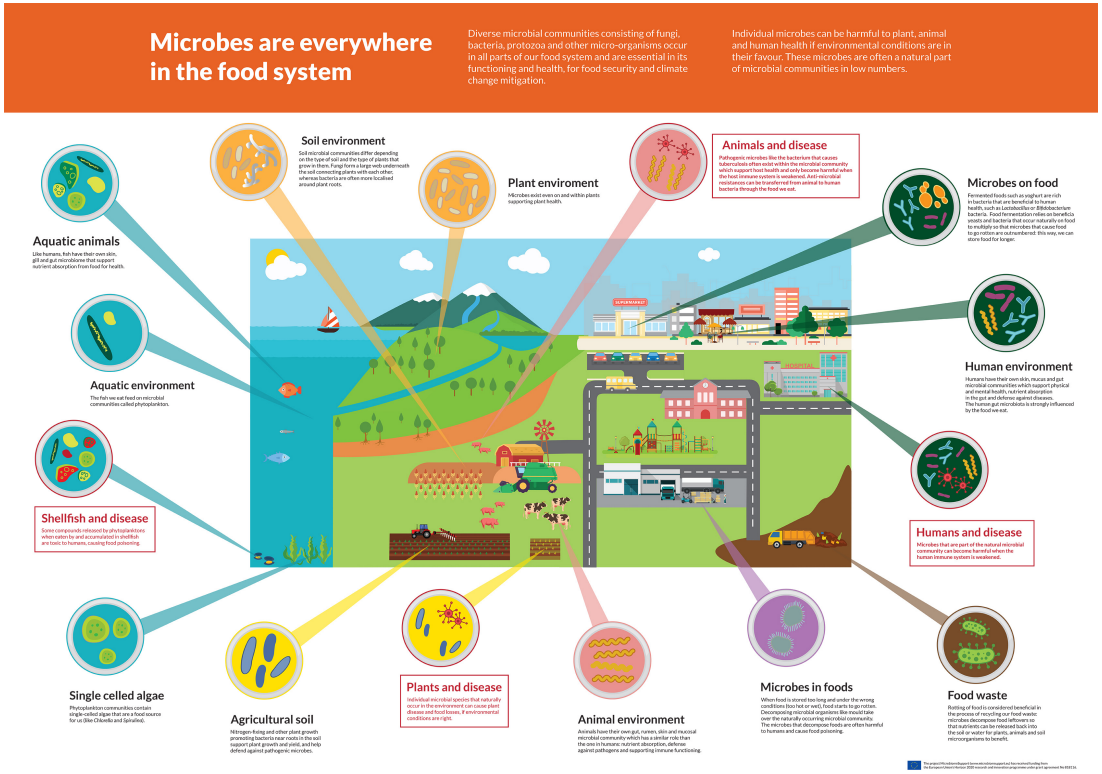
This lecture is designed to introduce you to the common techniques and software that many people in the field use, but there are always new tools being released!

I will do my best to answer any questions there are, but I cannot guarantee I will have a perfect answer for all of them, so please bear with me!

INTRODUCTION TO THE MICROBIOME

- Microbiome: A community of microscopic organisms that live together and interact with one another in an environment
  - Consists of viruses, fungi, and bacteria
- Historically, we were limited to visual identification of microbes and simple *in vitro/in vivo* techniques for analysis
- In this class, we will focus on analyzing the bacterial population from a bioinformatics perspective

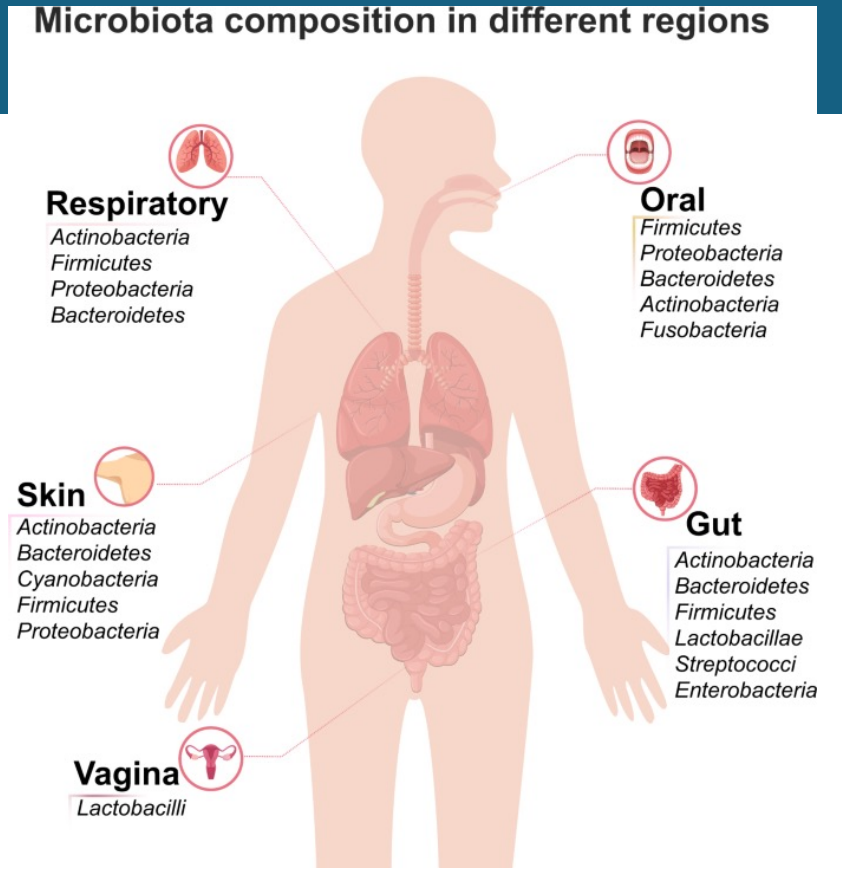
WHY SHOULD WE CARE ABOUT THE MICROBIOME?



https://www.microbiomesupport.eu/microbes-are-everywhere-in-the-foodsystem/

MICROBES IN HEALTH AND DISEASE

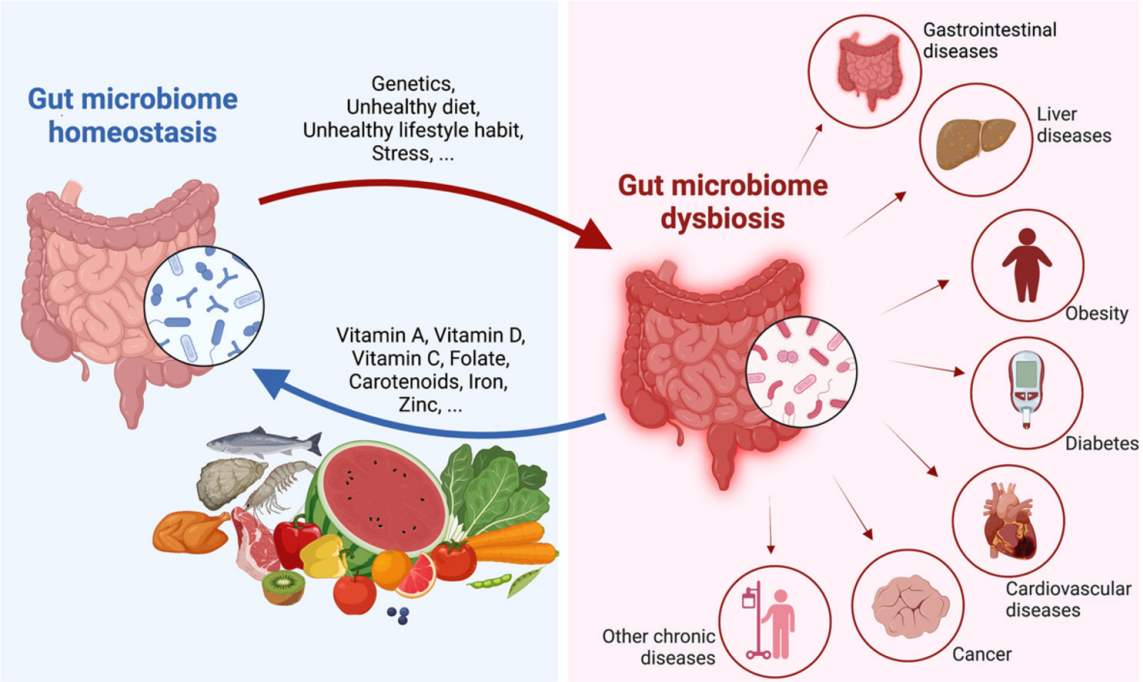
- The body consists of many different microbiomes
- These microbiomes are composed of different families/species/strains and thus contribute differently to disease



https://www.nature.com/articles/s41392-022-00974-4

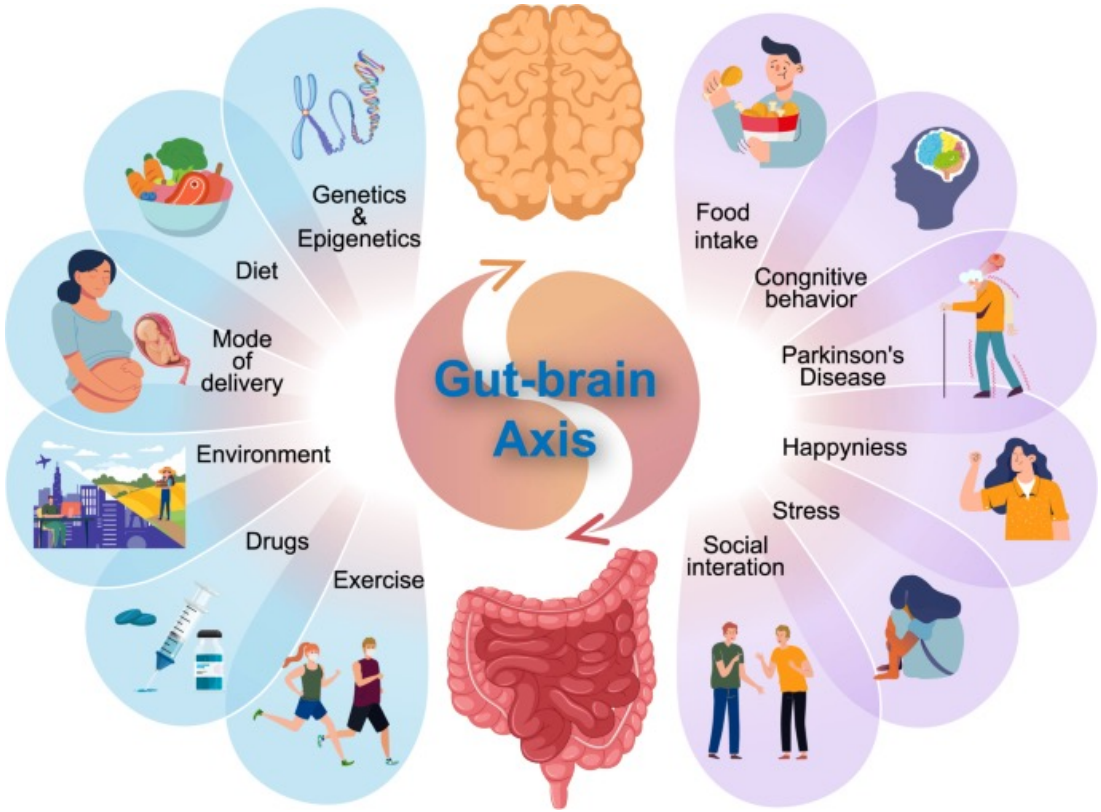
MICROBES IN HEALTH AND DISEASE

- The gut microbiome has been of particular interest in the past couple decades



https://www.sciencedirect.com/science/article/abs/pii/S0271531723000143

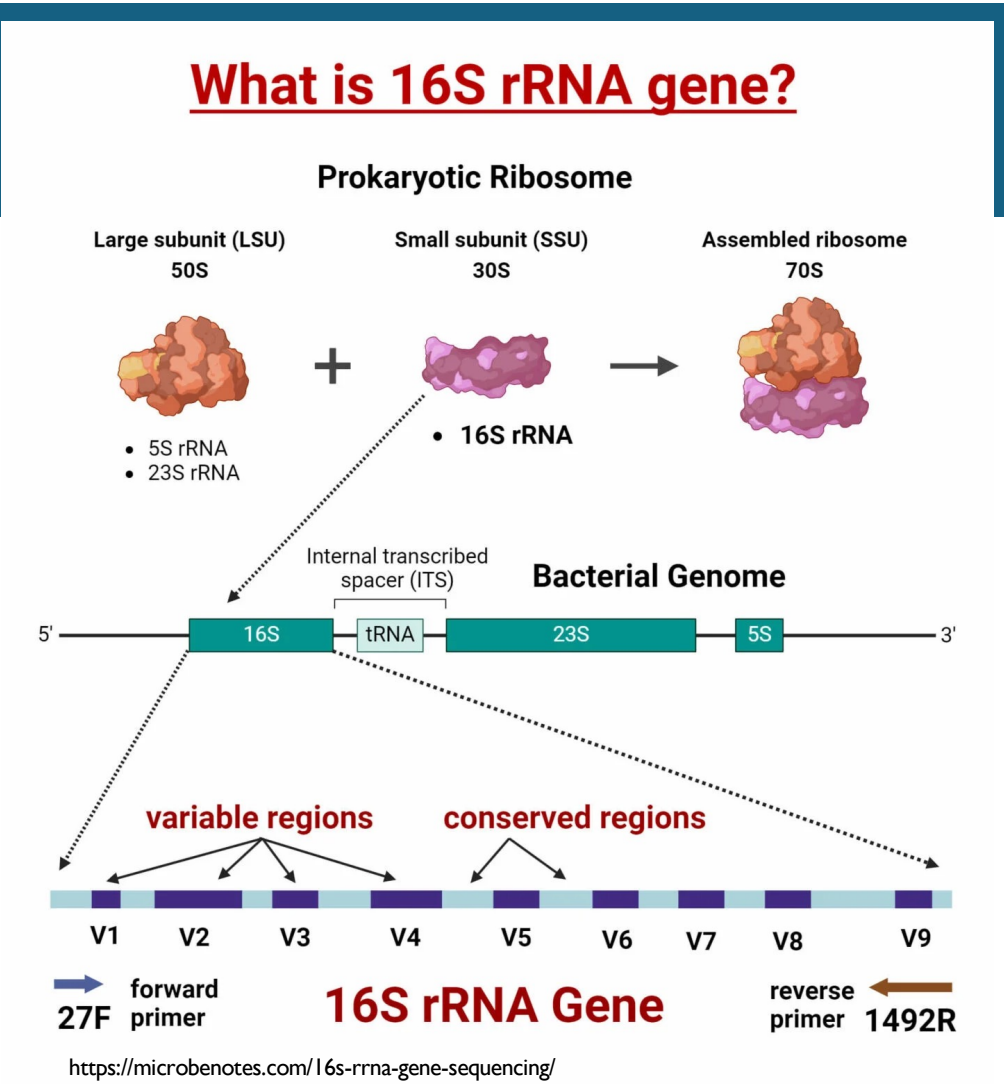
MICROBES IN HEALTH AND DISEASE



https://www.nature.com/articles/s41392-022-00974-4

16S SEQUENCING

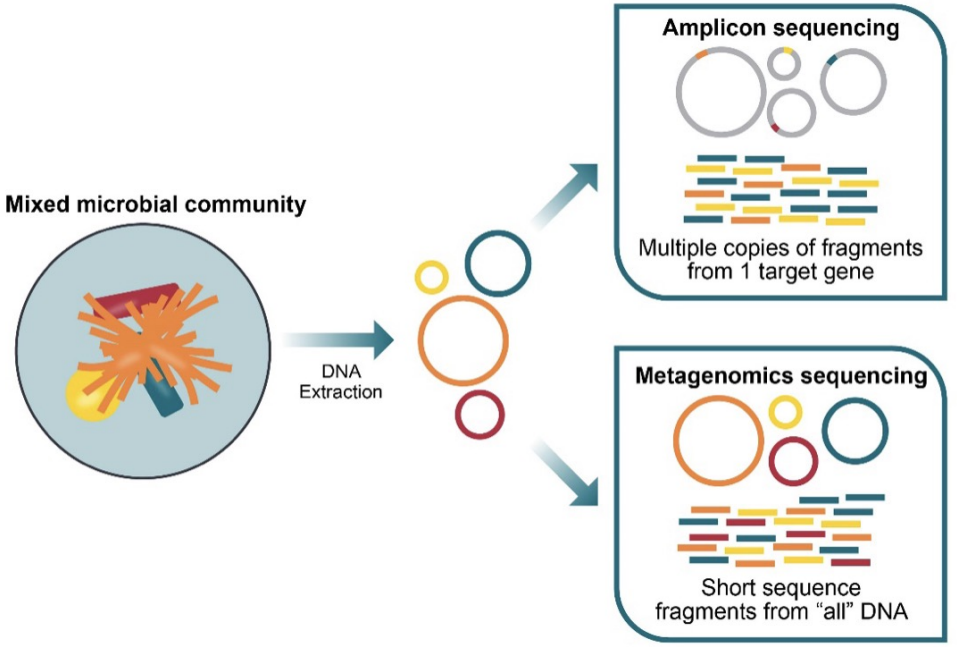
- 16S rRNA is a region of the 30S subunit of the ribosome in bacteria
- The 16S rRNA gene consists of both variable and conserved regions
- Question: Which regions do you think we should target for taxonomic identification of microbes? Why?



https://microbenotes.com/16s-rrna-gene-sequencing/

SHOTGUN SEQUENCING

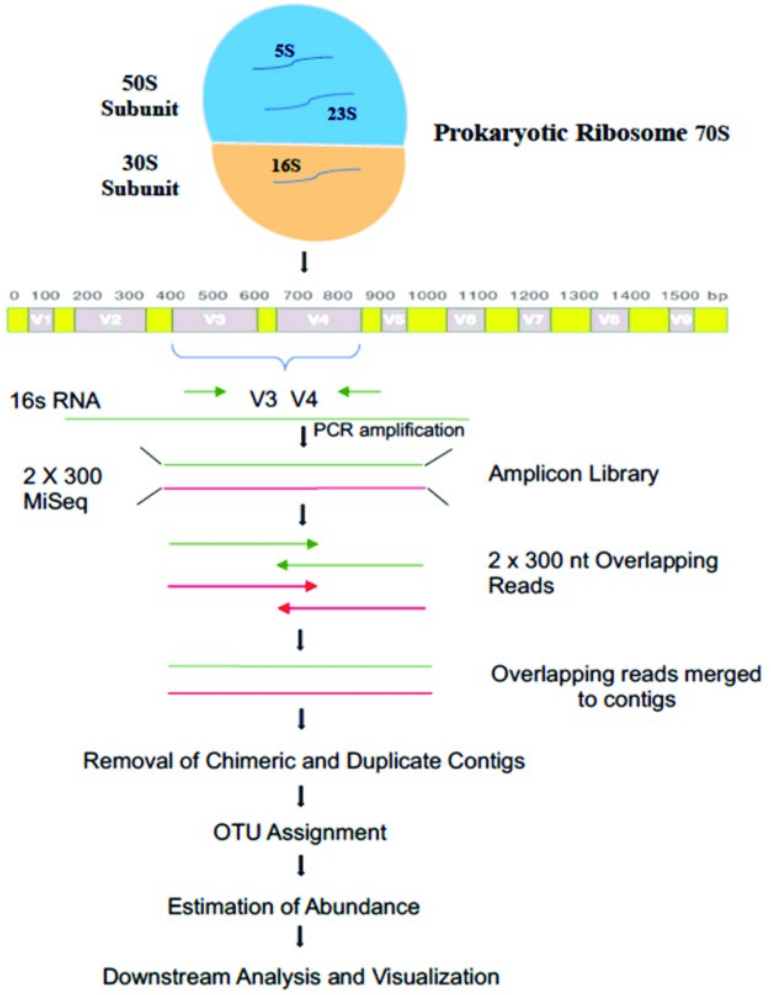
- We capture MUCH more information with shotgun sequencing
- Does not just rely on the 16S subunit, but rather the whole genome
- Question: Why is this not the standard in the field at this point then? What are the drawbacks of it?



https://blog.crownbio.com/understanding-the-microbiome-using-genomic-sequencing-and-analysis

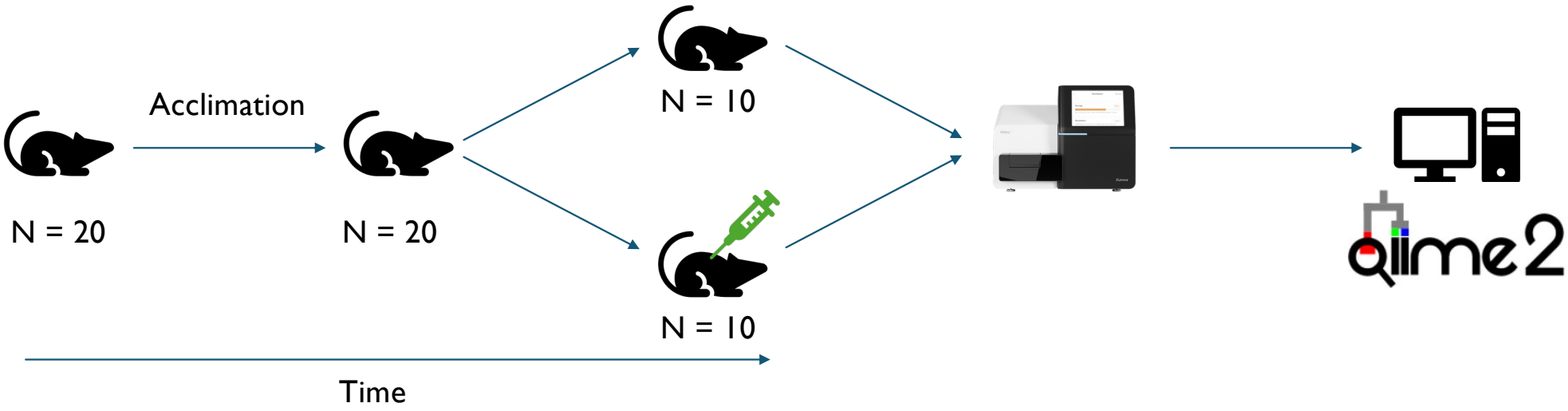
16S SEQUENCING WORKFLOW

1. Extract DNA from organism of interest
2. PCR amplification, where primers bind to conserved regions
3. Perform sequencing
4. Quality control
5. Taxonomic assignment
6. Create feature table



COMMON EXPERIMENTAL DESIGN FOR GUT MICROBIOME STUDIES

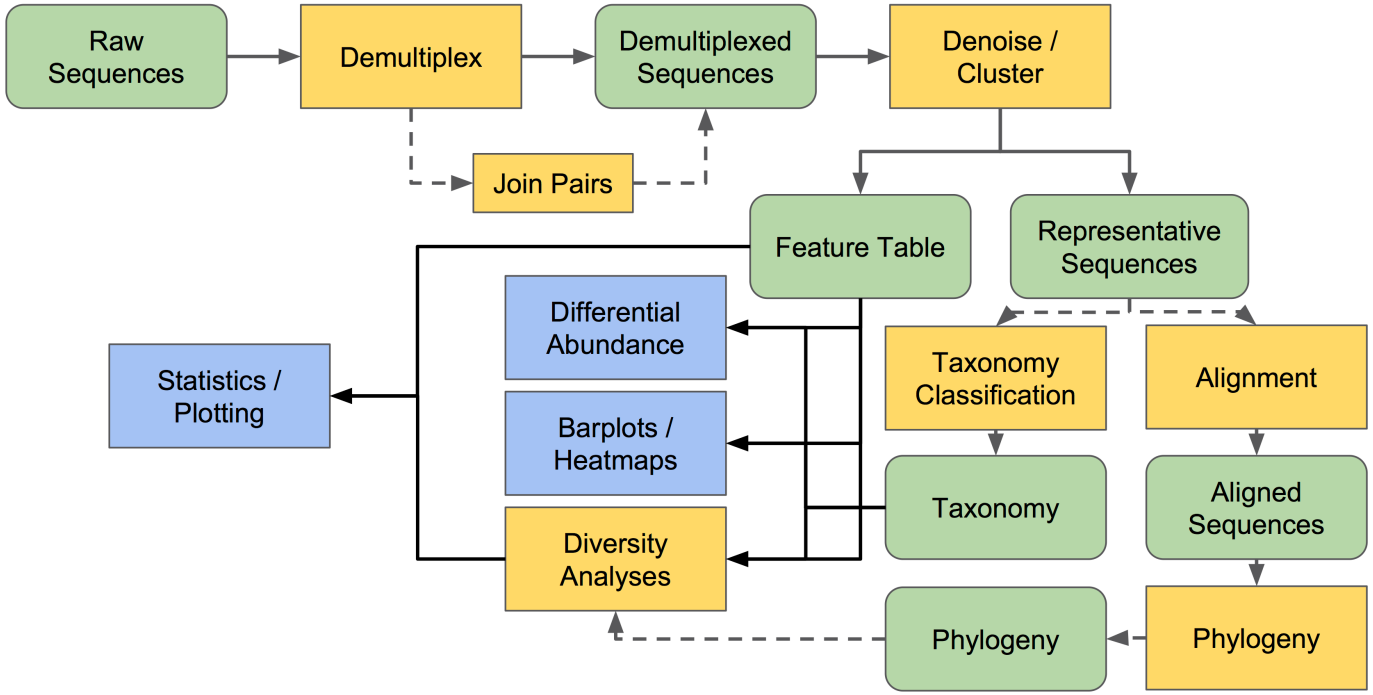
- How does the gut microbial community differ in individuals on medication Z when compared to individuals who are not on the medication?



QIIME2

- QIIME2 is a common microbiome analysis software
  - Takes as input single- or paired-end FASTQ files, along with a barcode file for demultiplexing and a metadata file
  - Performs taxonomic identification of features
  - Performs basic statistical analyses of samples, including plotting of results
  - Outputs a feature table, which we can use in other software

QIIME2 STEPS



https://docs.qiime2.org/2024.10/tutorials/overview/#useful-points-for-beginners

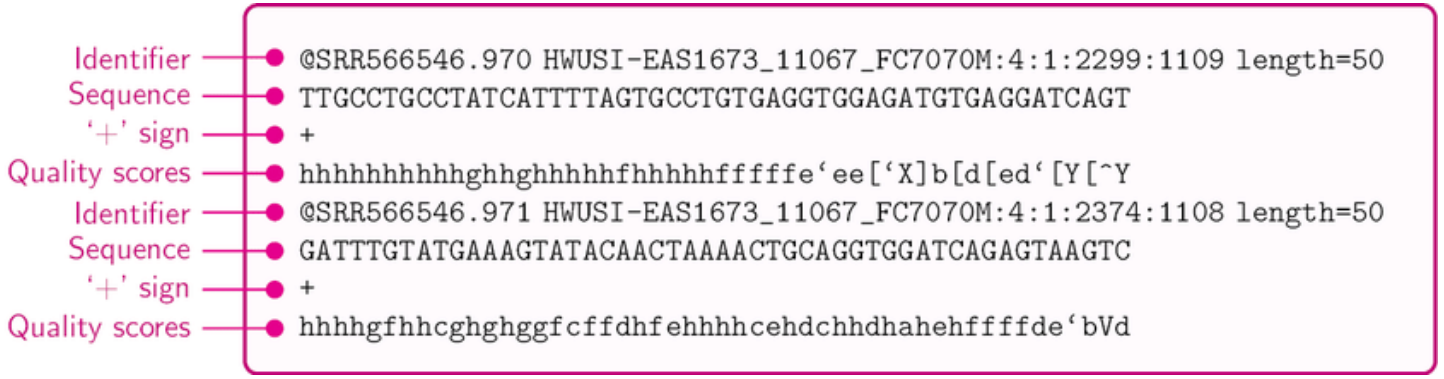
QIIME2 DATA FORMATS

- FASTQ: Contains the sequence ID, sequence itself, and quality information for each position
- QIIME2 artifact files
  - Intermediate files that are required by QIIME2
  - These contain both data and metadata
- QIIME2 visualization files
  - Stored in .qzv format
  - Contain the resulting plots from your QIIME2 analysis
  - Can be viewed in a web browser at https://view.qiime2.org

FASTQ DATA FORMAT

Each sequence consists of four main lines:

- Identifier: Unique ID for that sequence
- Sequence: The sequence itself
- A plus sign
- Quality scores



https://www.researchgate.net/figure/A-sample-of-the-FASTQ-file_fig2_309134977

FASTQ QUALITY SCORES

- Each ASCII character in the fourth line is associated with a **Phred score**, which is a score that denotes the accuracy (or quality) of that position
- The Phred score (Q) is calculated using the following formula, where P is the probability of the base being incorrectly called:

$$Q = -10 \log_{10}(P)$$

- Question: Let's assume the probability of the base being called incorrectly is 1/100. What would the Phred score be?

Q = -10 * log10(1/100) = 20

FASTQ QUALITY SCORES

- We then cross-reference this Phred score (Q) with this ASCII table, which tells us that the ASCII character denoting it would be '5'

- Question: What is the relationship between the error probability and Phred score?

convert ascii33 to error probability front

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

BITS VIB Bioinformatics Training and Service Facility

hg18-total-sequenced=2'858'034'764 (UCSC)

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
!	33	0	1.00E+00	0.000%	1	2,858,034,764
"	34	1	7.94E-01	20.567%	1	2,270,217,709
#	35	2	6.31E-01	36.904%	2	1,803,298,025
$	36	3	5.01E-01	49.881%	2	1,432,410,537
%	37	4	3.98E-01	60.189%	3	1,137,804,133
&	38	5	3.16E-01	68.377%	3	903,789,949
`	39	6	2.51E-01	74.881%	4	717,905,874
(	40	7	2.00E-01	80.047%	5	570,252,906
)	41	8	1.58E-01	84.151%	6	452,967,984
*	42	9	1.26E-01	87.411%	8	359,805,259
+	43	10	1.00E-01	90.000%	10	285,803,476
,	44	11	7.94E-02	92.057%	13	227,021,771
-	45	12	6.31E-02	93.690%	16	180,329,803
.	46	13	5.01E-02	94.988%	20	143,241,054
/	47	14	3.98E-02	96.019%	25	113,780,413
0	48	15	3.16E-02	96.838%	32	90,378,995
1	49	16	2.51E-02	97.488%	40	71,790,587
2	50	17	2.00E-02	98.005%	50	57,025,291
3	51	18	1.58E-02	98.415%	63	45,296,798
4	52	19	1.26E-02	98.741%	79	35,980,526
5	53	20	1.00E-02	99.000%	100	28,580,348
6	54	21	7.94E-03	99.206%	126	22,702,177
7	55	22	6.31E-03	99.369%	158	18,032,980
8	56	23	5.01E-03	99.499%	200	14,324,105
9	57	24	3.98E-03	99.602%	251	11,378,041
:	58	25	3.16E-03	99.684%	316	9,037,899
;	59	26	2.51E-03	99.749%	398	7,179,059
<	60	27	2.00E-03	99.800%	501	5,702,529
=	61	28	1.58E-03	99.842%	631	4,529,680
>	62	29	1.26E-03	99.874%	794	3,598,053
?	63	30	1.00E-03	99.900%	1,000	2,858,035
@	64	31	7.94E-04	99.921%	1,259	2,270,218
A	65	32	6.31E-04	99.937%	1,585	1,803,298
B	66	33	5.01E-04	99.950%	1,995	1,432,411
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253
G	71	38	1.58E-04	99.984%	6,310	452,968
H	72	39	1.26E-04	99.987%	7,943	359,805
I	73	40	1.00E-04	99.990%	10,000	285,803
J	74	41	7.94E-05	99.992%	12,589	227,022
K	75	42	6.31E-05	99.994%	15,849	180,330
L	76	43	5.01E-05	99.995%	19,953	143,241
M	77	44	3.98E-05	99.996%	25,119	113,780
N	78	45	3.16E-05	99.997%	31,623	90,379
O	79	46	2.51E-05	99.997%	39,811	71,791

back

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
P	80	47	2.00E-05	99.998%	50,119	57,025
Q	81	48	1.58E-05	99.998%	63,096	45,297
R	82	49	1.26E-05	99.999%	79,433	35,981
S	83	50	1.00E-05	99.999%	100,000	28,580
T	84	51	7.94E-06	99.999%	125,893	22,702
U	85	52	6.31E-06	99.999%	158,489	18,033
V	86	53	5.01E-06	99.999%	199,526	14,324
W	87	54	3.98E-06	100.000%	251,189	11,378
X	88	55	3.16E-06	100.000%	316,228	9,038
Y	89	56	2.51E-06	100.000%	398,107	7,179
Z	90	57	2.00E-06	100.000%	501,187	5,703
[	91	58	1.58E-06	100.000%	630,957	4,530
\	92	59	1.26E-06	100.000%	794,328	3,598
]	93	60	1.00E-06	100.000%	1,000,000	2,858
^	94	61	7.94E-07	100.000%	1,258,925	2,270
_	95	62	6.31E-07	100.000%	1,584,893	1,803
`	96	63	5.01E-07	100.000%	1,995,262	1,432
a	97	64	3.98E-07	100.000%	2,511,886	1,138
b	98	65	3.16E-07	100.000%	3,162,278	904
c	99	66	2.51E-07	100.000%	3,981,072	718
d	100	67	2.00E-07	100.000%	5,011,872	570
e	101	68	1.58E-07	100.000%	6,309,573	453
f	102	69	1.26E-07	100.000%	7,943,282	360
g	103	70	1.00E-07	100.000%	10,000,000	286
h	104	71	7.94E-08	100.000%	12,589,254	227
i	105	72	6.31E-08	100.000%	15,848,932	180
j	106	73	5.01E-08	100.000%	19,952,623	143
k	107	74	3.98E-08	100.000%	25,118,864	114
l	108	75	3.16E-08	100.000%	31,622,777	90
m	109	76	2.51E-08	100.000%	39,810,717	72
n	110	77	2.00E-08	100.000%	50,118,723	57
o	111	78	1.58E-08	100.000%	63,095,734	45
p	112	79	1.26E-08	100.000%	79,432,823	36
q	113	80	1.00E-08	100.000%	100,000,000	29
r	114	81	7.94E-09	100.000%	125,892,541	23
s	115	82	6.31E-09	100.000%	158,489,319	18
t	116	83	5.01E-09	100.000%	199,526,231	14
u	117	84	3.98E-09	100.000%	251,188,643	11
v	118	85	3.16E-09	100.000%	316,227,766	9
w	119	86	2.51E-09	100.000%	398,107,171	7
x	120	87	2.00E-09	100.000%	501,187,234	6
y	121	88	1.58E-09	100.000%	630,957,344	5
z	122	89	1.26E-09	100.000%	794,328,235	4
{	123	90	1.00E-09	100.000%	1,000,000,000	3
|	124	91	7.94E-10	100.000%	1,258,925,412	2
}	125	92	6.31E-10	100.000%	1,584,893,192	2
~	126	93	5.01E-10	100.000%	1,995,262,315	1

FASTQ QUALITY SCORES

- Question: What would the order of ASCII characters be (from most accurate to least accurate) for the following sequence then?

```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
```

Answer: hgfedb’^][YX

Note how in general, the original sequence started off at good quality, then degraded towards the end

convert ascii33 to error probability front

$$Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_e)$$

hg18-total-sequenced=2'858'034'764 (UCSC)

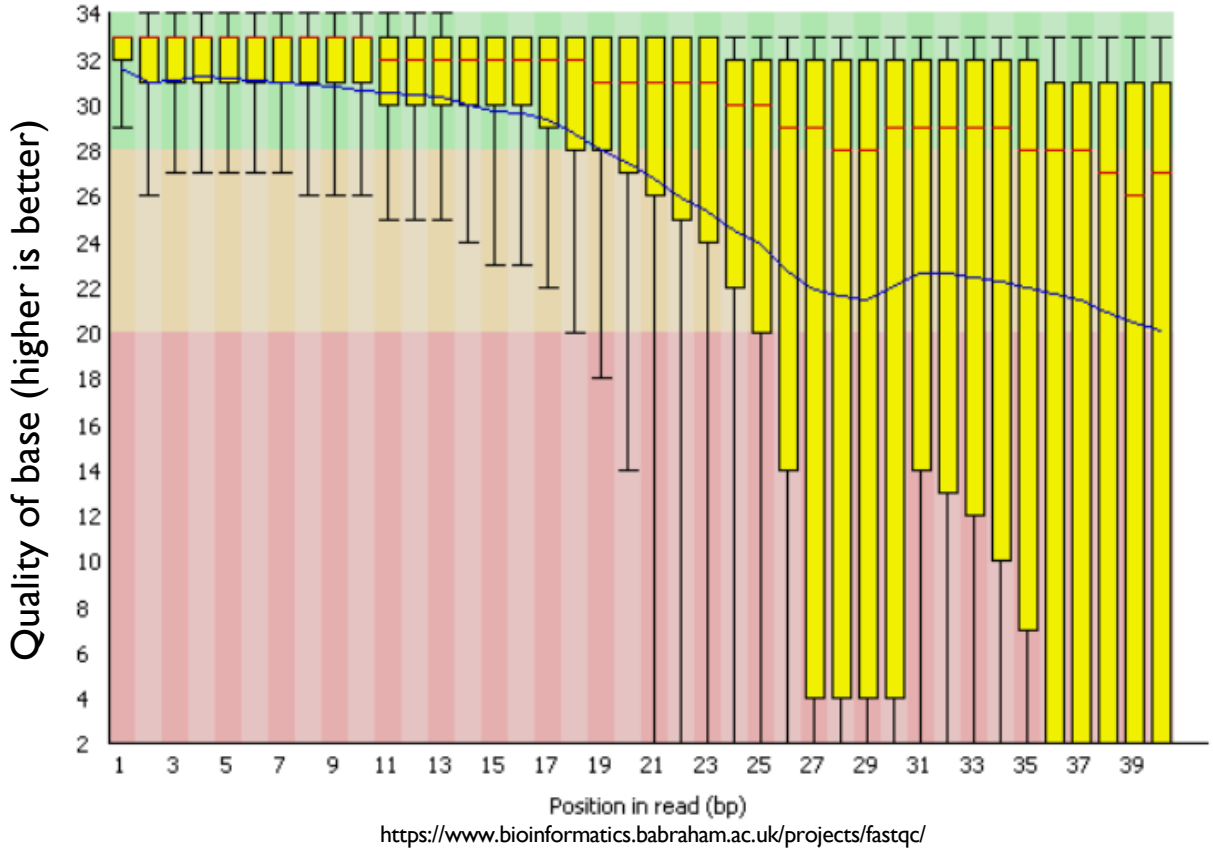
BITS VIB Bioinformatics Training and Service Facility

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
!	33	0	1.00E+00	0.000%	1	2,858,034,764
"	34	1	7.94E-01	20.567%	1	2,270,217,709
#	35	2	6.31E-01	36.904%	2	1,803,298,025
$	36	3	5.01E-01	49.881%	2	1,432,410,537
%	37	4	3.98E-01	60.189%	3	1,137,804,133
&	38	5	3.16E-01	68.377%	3	903,789,949
`	39	6	2.51E-01	74.881%	4	717,905,874
(	40	7	2.00E-01	80.047%	5	570,252,906
)	41	8	1.58E-01	84.151%	6	452,967,984
*	42	9	1.26E-01	87.411%	8	359,805,259
+	43	10	1.00E-01	90.000%	10	285,803,476
,	44	11	7.94E-02	92.057%	13	227,021,771
-	45	12	6.31E-02	93.690%	16	180,329,803
.	46	13	5.01E-02	94.988%	20	143,241,054
/	47	14	3.98E-02	96.019%	25	113,780,413
0	48	15	3.16E-02	96.838%	32	90,378,995
1	49	16	2.51E-02	97.488%	40	71,790,587
2	50	17	2.00E-02	98.005%	50	57,025,291
3	51	18	1.58E-02	98.415%	63	45,296,798
?	63	30	1.00E-03	99.900%	1,000	2,858,035
@	64	31	7.94E-04	99.921%	1,259	2,270,218
A	65	32	6.31E-04	99.937%	1,585	1,803,298
B	66	33	5.01E-04	99.950%	1,995	1,432,411
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253
G	71	38	1.58E-04	99.984%	6,310	452,968
H	72	39	1.26E-04	99.987%	7,943	359,805
I	73	40	1.00E-04	99.990%	10,000	285,803
J	74	41	7.94E-05	99.992%	12,589	227,022
K	75	42	6.31E-05	99.994%	15,849	180,330
L	76	43	5.01E-05	99.995%	19,953	143,241
M	77	44	3.98E-05	99.996%	25,119	113,780
N	78	45	3.16E-05	99.997%	31,623	90,379
O	79	46	2.51E-05	99.997%	39,811	71,791

back

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
P	80	47	2.00E-05	99.998%	50,119	57,025
Q	81	48	1.58E-05	99.998%	63,096	45,297
R	82	49	1.26E-05	99.999%	79,433	35,981
S	83	50	1.00E-05	99.999%	100,000	28,580
T	84	51	7.94E-06	99.999%	125,893	22,702
U	85	52	6.31E-06	99.999%	158,489	18,033
V	86	53	5.01E-06	99.999%	199,526	14,324
W	87	54	3.98E-06	100.000%	251,189	11,378
X	88	55	3.16E-06	100.000%	316,228	9,038
Y	89	56	2.51E-06	100.000%	398,107	7,179
Z	90	57	2.00E-06	100.000%	501,187	5,703
[	91	58	1.58E-06	100.000%	630,957	4,530
\	92	59	1.26E-06	100.000%	794,328	3,598
]	93	60	1.00E-06	100.000%	1,000,000	2,858
^	94	61	7.94E-07	100.000%	1,258,925	2,270
_	95	62	6.31E-07	100.000%	1,584,893	1,803
`	96	63	5.01E-07	100.000%	1,995,262	1,432
a	97	64	3.98E-07	100.000%	2,511,886	1,138
b	98	65	3.16E-07	100.000%	3,162,278	904
c	99	66	2.51E-07	100.000%	3,981,072	718
d	100	67	2.00E-07	100.000%	5,011,872	570
e	101	68	1.58E-07	100.000%	6,309,573	453
f	102	69	1.26E-07	100.000%	7,943,282	360
g	103	70	1.00E-07	100.000%	10,000,000	286
h	104	71	7.94E-08	100.000%	12,589,254	227
i	105	72	6.31E-08	100.000%	15,848,932	180
j	106	73	5.01E-08	100.000%	19,952,623	143
k	107	74	3.98E-08	100.000%	25,118,864	114
l	108	75	3.16E-08	100.000%	31,622,777	90
m	109	76	2.51E-08	100.000%	39,810,717	72
n	110	77	2.00E-08	100.000%	50,118,723	57
o	111	78	1.58E-08	100.000%	63,095,734	45
p	112	79	1.26E-08	100.000%	79,432,823	36
q	113	80	1.00E-08	100.000%	100,000,000	29
r	114	81	7.94E-09	100.000%	125,892,541	23
s	115	82	6.31E-09	100.000%	158,489,319	18
t	116	83	5.01E-09	100.000%	199,526,231	14
u	117	84	3.98E-09	100.000%	251,188,643	11
v	118	85	3.16E-09	100.000%	316,227,766	9
w	119	86	2.51E-09	100.000%	398,107,171	7
x	120	87	2.00E-09	100.000%	501,187,234	6
y	121	88	1.58E-09	100.000%	630,957,344	5
z	122	89	1.26E-09	100.000%	794,328,235	4
{	123	90	1.00E-09	100.000%	1,000,000,000	3
|	124	91	7.94E-10	100.000%	1,258,925,412	2
}	125	92	6.31E-10	100.000%	1,584,893,192	2
~	126	93	5.01E-10	100.000%	1,995,262,315	1

THE DEGRADATION IN QUALITY IS WHAT WE CAPTURE IN OUR FASTQC PLOTS



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

QIIME2 LEARNING OUTCOMES

1. Be able to install QIIME2
2. Understand the format of the QIIME2 input files
3. Perform the preprocessing steps for 16S data, including quality control
4. Perform taxonomic classification using QIIME2
5. Export the feature table

QIIME2 INSTALLATION

- QIIME2 can be ran a few different ways, but I prefer the command line interface
  - Also can be ran via Jupyter notebook and the Galaxy web server

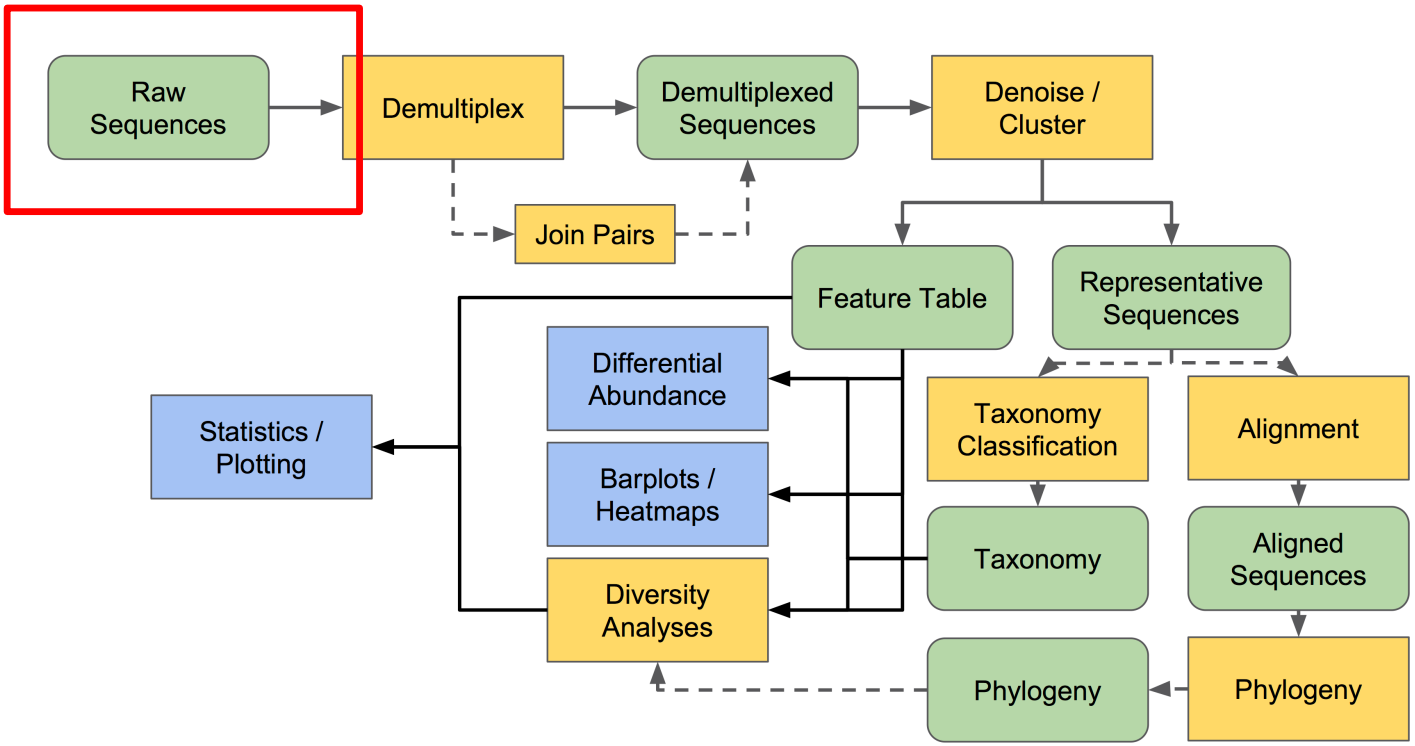
1. Install miniconda in your bash environment if you have not already done so
2. Create a new conda environment using your desired QIIME2 distribution

```
conda env create -n qiime2-amplicon-2024.10 --file
https://data.qiime2.org/distro/amplicon/qiime2-amplicon-2024.10-py310-linux-conda.yml
```

3. Activate your conda environment

```
conda activate qiime2-amplicon-2024.10
```

QIIME2 STEP 1 – IMPORTING DATA



https://docs.qiime2.org/2024.10/tutorials/overview/#useful-points-for-beginners

THE QIIME2 MOVING PICTURES TUTORIAL

https://docs.qiime2.org/2024.10/tutorials/moving-pictures/

- For this class, we will be following the simple and easy-to-understand “Moving Pictures” tutorial on the QIIME2 website
- I highly encourage you to work through this tutorial on your own time, but I will summarize the key points in this talk

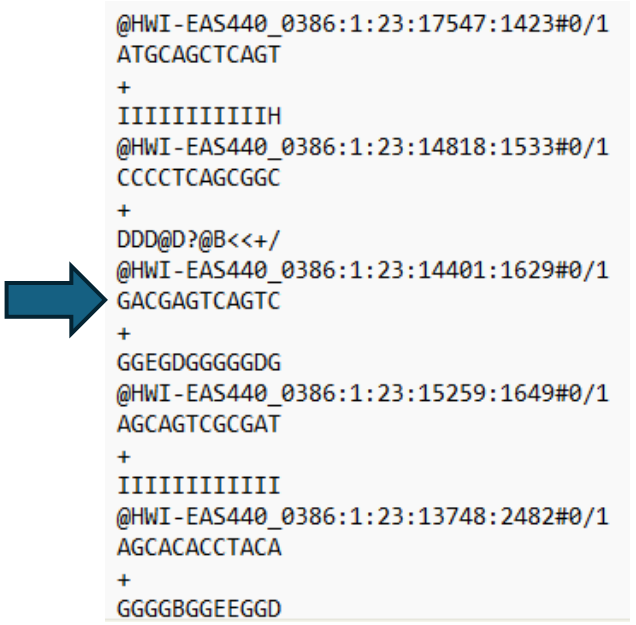
THE QIIME2 MOVING PICTURES TUTORIAL

This tutorial uses data formatted as EMP single end sequences, but you will need to check what format your data is by discussing whoever performed the sequencing. The commands will change slightly for this first step, but then everything later on will be the same.

If your data is in a different format, the commands for importing can be found here (https://docs.qiime2.org/2024.10/tutorials/importing/)

QIIME2 STEP 1 – IMPORTING DATA

Let's first take a look at the supplied tutorial data. One of our supplied FASTQ files looks like the following.



Questions:

1. Do you think it's the barcode file or the 16S sequence file? How can you tell?
2. What does the indicated line in the file tell us?
3. The sequence at the indicated line shows up many times in this file. Is that an issue? Why/why not?

QIIME2 STEP 1 – IMPORTING DATA

Now let's look at the head of the other FASTQ file.

```
@HWI-EAS440_0386:1:23:17547:1423#0/1
TACGNAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGGGGGGATTGGTGTG
+
IIIE)EEEEEEEEGFIIGIIIHIHHGIIIGIIHHHGIIHGHEGDGIFIGEHGIHHGHHGHHGGHEEGHEGGEHEBBHBBEEDCEDDD>B?BE@@B>@@@@@CB@ABA@@?@@=>?08;3=;==8:5;@6?######################
@HWI-EAS440_0386:1:23:14818:1533#0/1
CCCCNCAGCGGCAAAAATTAAAATTTTTACCGCTTCGGCGTTATAGCCTCACACTCAATCTTTTATCACGAAGTCATGATTGAATCGCGAGTGGTCGGCAGATTGCGATAAACGGGCACATTAAATTTAAACTGATGATTCCACTGCAACAA
+
64<2$24;1)/:*B<?BBDDBBD<>BDD#################################################################################################################################
@HWI-EAS440_0386:1:23:14401:1629#0/1
TACGNAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGACGCTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGGTGTCTTGAGTACAGTAGAGGCAGGGGGGGGGTTGGGGG
+
GGGC'ACC8;;>;HHHHGHDHHHHHEEHHEHHHHHECHEEEHCHFHHHAGGEHHFBCCBABBBE>>>E=>A>A<>>B8B:B=BBABA@AAAA@?>???>>>9>@AA@@@@AA#######################################
@HWI-EAS440_0386:1:23:15259:1649#0/1
TACGNAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGACGCTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGGTGTCTTGAGTACAGTAGAGGCAGGGGGGAGTTTGGGGG
+
IIIE)DEE?CCCBIIIIIIGIIIIIHIIIIIIIIIHIIIHFIDIGIIIHHIHIGIIIHFIHBGGBFDGEHHEI=CBGBEEEEEHEEGECD?>B@=?@BAA=9A?@A>ABBBCDB:C:@??9>?;:?;?BA@?B####################
```

Question: Would you be concerned about the quality of any of these sequences?

convert ascii33 to error probability

front

$$Q_{PHRED} = -10 \times \log_{10}(P_e)$$

BITS VIB Bioinformatics Training and Service Facility

hg18-total-sequenced=2'858'034'764 (UCSC)

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
!	33	0	1.00E+00	0.000%	1	2,858,034,764
"	34	1	7.94E-01	20.567%	1	2,270,217,709
#	35	2	6.31E-01	36.904%	2	1,803,298,025
$	36	3	5.01E-01	49.881%	2	1,432,410,537
%	37	4	3.98E-01	60.189%	3	1,137,804,133
&	38	5	3.16E-01	68.377%	3	903,789,949
`	39	6	2.51E-01	74.881%	4	717,905,874
(	40	7	2.00E-01	80.047%	5	570,252,906
)	41	8	1.58E-01	84.151%	6	452,967,984
*	42	9	1.26E-01	87.411%	8	359,805,259
+	43	10	1.00E-01	90.000%	10	285,803,476
,	44	11	7.94E-02	92.057%	13	227,021,771
-	45	12	6.31E-02	93.690%	16	180,329,803
.	46	13	5.01E-02	94.988%	20	143,241,054
/	47	14	3.98E-02	96.019%	25	113,780,413
0	48	15	3.16E-02	96.838%	32	90,378,995
1	49	16	2.51E-02	97.488%	40	71,790,587
2	50	17	2.00E-02	98.005%	50	57,025,291
3	51	18	1.58E-02	98.415%	63	45,296,798
4	52	19	1.26E-02	98.741%	79	35,980,526
5	53	20	1.00E-02	99.000%	100	28,580,348
6	54	21	7.94E-03	99.206%	126	22,702,177
7	55	22	6.31E-03	99.369%	158	18,032,980
8	56	23	5.01E-03	99.499%	200	14,324,105
9	57	24	3.98E-03	99.602%	251	11,378,041
:	58	25	3.16E-03	99.684%	316	9,037,899
;	59	26	2.51E-03	99.749%	398	7,179,059
<	60	27	2.00E-03	99.800%	501	5,702,529
=	61	28	1.58E-03	99.842%	631	4,529,680
>	62	29	1.26E-03	99.874%	794	3,598,053
?	63	30	1.00E-03	99.900%	1,000	2,858,035
@	64	31	7.94E-04	99.921%	1,259	2,270,218
A	65	32	6.31E-04	99.937%	1,585	1,803,298
B	66	33	5.01E-04	99.950%	1,995	1,432,411
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253
G	71	38	1.58E-04	99.984%	6,310	452,968
H	72	39	1.26E-04	99.987%	7,943	359,805
I	73	40	1.00E-04	99.990%	10,000	285,803
J	74	41	7.94E-05	99.992%	12,589	227,022
K	75	42	6.31E-05	99.994%	15,849	180,330
L	76	43	5.01E-05	99.995%	19,953	143,241
M	77	44	3.98E-05	99.996%	25,119	113,780
N	78	45	3.16E-05	99.997%	31,623	90,379
O	79	46	2.51E-05	99.997%	39,811	71,791

QIIME2 STEP 1 – IMPORTING DATA

Finally, let's look at the head of the metadata.

This study wanted to describe the normal microbial populations across different regions in the body, at different time points

	A	B	C	D	E	F	G	H	I
1	sample-id	barcode-sequence	body-site	year	month	day	subject	reported-antibiotic-usage	days-since-experiment-start
2	#q2:types	categorical	categorical	numeric	numeric	numeric	categorical	categorical	numeric
3	L1S8	AGCTGACTAGTC	gut	2008	10	28	subject-1	Yes	0
4	L1S57	ACACACTATGGC	gut	2009	1	20	subject-1	No	84
5	L1S76	ACTACGTGTGGT	gut	2009	2	17	subject-1	No	112
6	L1S105	AGTGCGATGCGT	gut	2009	3	17	subject-1	No	140
7	L2S155	ACGATGCGACCA	left palm	2009	1	20	subject-1	No	84
8	L2S175	AGCTATCCACGA	left palm	2009	2	17	subject-1	No	112
9	L2S204	ATGCAGCTCAGT	left palm	2009	3	17	subject-1	No	140
10	L2S222	CACGTGACATGT	left palm	2009	4	14	subject-1	No	168
11	L3S242	ACAGTTGCGCGA	right palm	2008	10	28	subject-1	Yes	0
12	L3S294	CACGACAGGCTA	right palm	2009	1	20	subject-1	No	84
13	L3S313	AGTGTCACGGTG	right palm	2009	2	17	subject-1	No	112
14	L3S341	CAAGTGAGAGAG	right palm	2009	3	17	subject-1	No	140
15	L3S360	CATCGTATCAAC	right palm	2009	4	14	subject-1	No	168
16	L5S104	CAGTGTCAGGAC	tongue	2008	10	28	subject-1	Yes	0
17	L5S155	ATCTTAGACTGC	tongue	2009	1	20	subject-1	No	84
18	L5S174	CAGACATTGCGT	tongue	2009	2	17	subject-1	No	112
19	L5S203	CGATGCACCAGA	tongue	2009	3	17	subject-1	No	140
20	L5S222	CTAGAGACTCTT	tongue	2009	4	14	subject-1	No	168

QIIME2 STEP 1 – IMPORTING DATA

Use the following command to import your data, assuming it is in an EMP single-end sequenced format

This is the name of the directory that contains all your input files

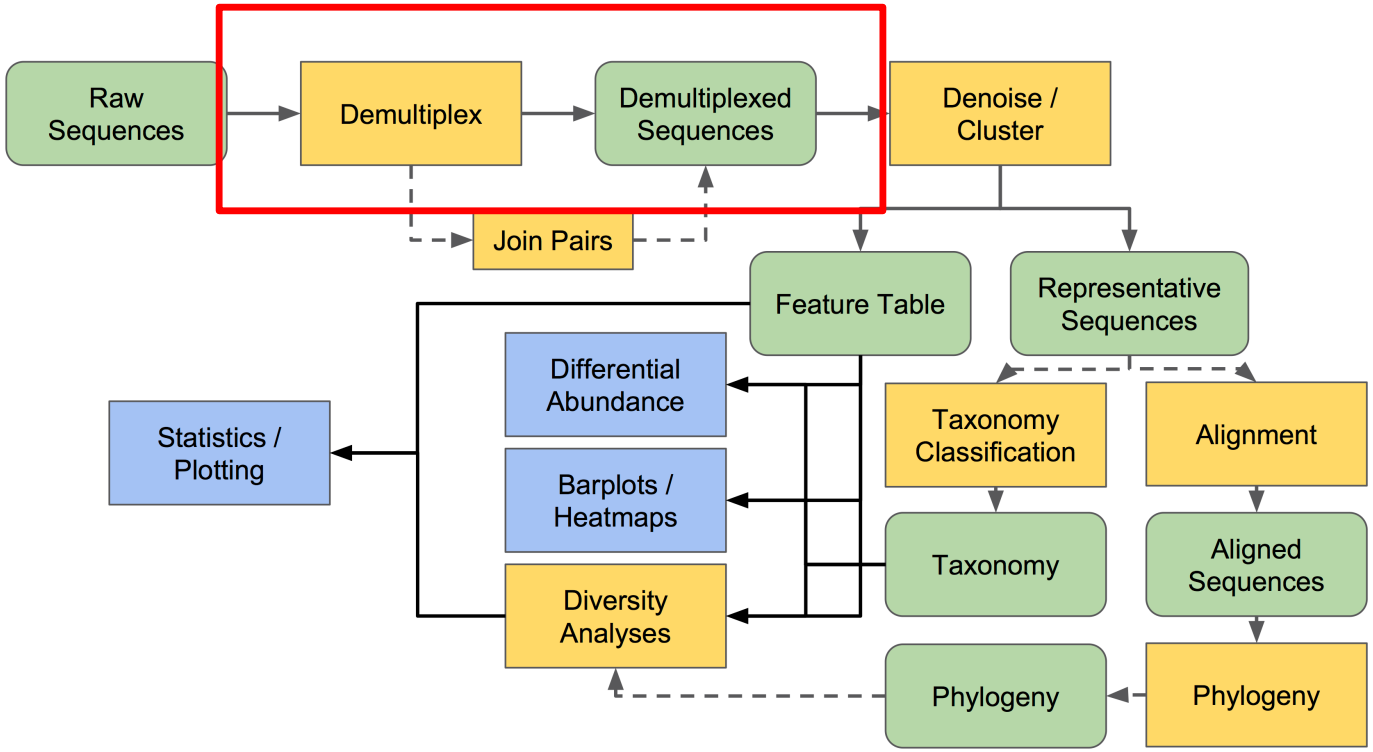
```
qiime tools import \
  --type EMPSingleEndSequences \
  --input-path emp-single-end-sequences \
  --output-path emp-single-end-sequences.qza
```

Note here that our output is a .qza file, so it is an **artifact**, but we can still view its contents with the online page https://view.qiime2.org. It's just not all that exciting.

Details of emp-single-end-sequences.qza

```
name: "emp-single-end-sequences.qza"
uuid: "593a4a7d-1cf5-46d6-b3c8-ca39810dc842"
type: "EMPSingleEndSequences"
format: "EMPSingleEndDirFmt"
```

QIIME2 STEP 2 – DEMULTIPLEXING SEQUENCES



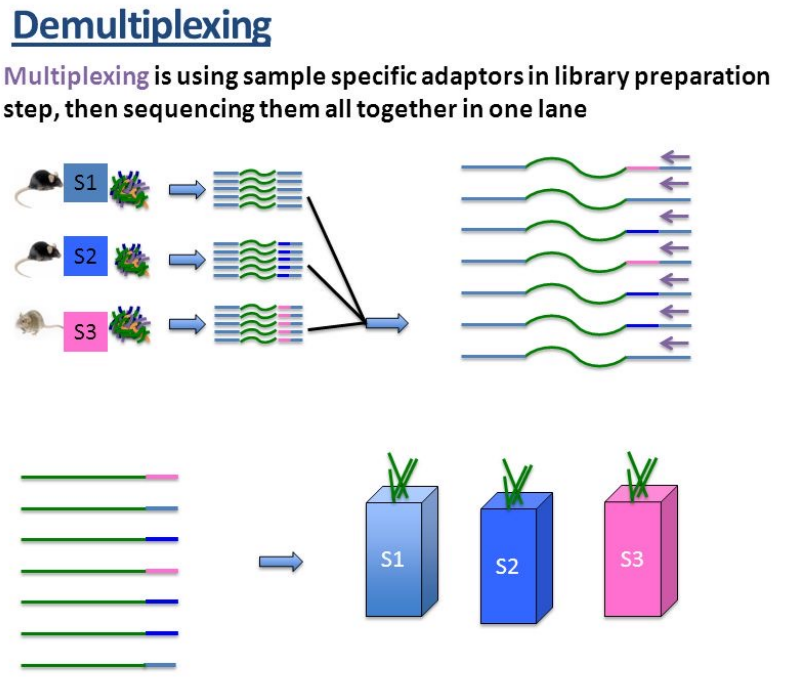
https://docs.qiime2.org/2024.10/tutorials/overview/#useful-points-for-beginners

QIIME2 STEP 2 – DEMULTIPLEXING SEQUENCES

We sequence multiple samples **together at the same time, all in one lane**

Each sequence gets a barcode that we can reference back to each sample, in a process called **demultiplexing**

Question: Why do we bother multiplexing sequences? Wouldn't it be less confusing and less work if we just sequenced everything individually?



https://awbrooks19.github.io/vmi_microbiome_bootcamp/rst/3_sequences_to_composition.html

QIIME2 STEP 2 – DEMULTIPLEXING SEQUENCES

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

This is the name of the column containing the barcode

Output artifacts:

- `demux-details.qza` : view | download
- `demux.qza` : view | download

QIIME2 STEP 2 – DEMULTIPLEXING SEQUENCES

This will output two qza files, each of which, again, are not very useful to look at

Details of demux-details.qza

```
name: "demux-details.qza"
uuid: "5998287b-d227-4904-aba3-bb9f4503b63e"
type: "ErrorCorrectionDetails"
format: "ErrorCorrectionDetailsDirFmt"
```

Details of demux.qza

```
name: "demux.qza"
uuid: "63a69bf7-6920-47e4-b02e-2d3b192f58eb"
type: "SampleData[SequencesWithQuality]"
format: "SingleLanePerSampleSingleEndFastqDirFmt"
```

QIIME2 STEP 2 – DEMULTIPLEXING SEQUENCES

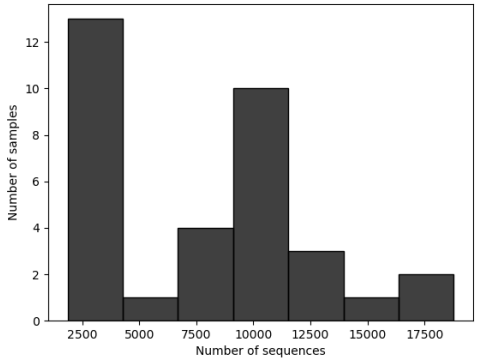
We can also view our results of the demultiplexing step by using the demux.qzv file

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

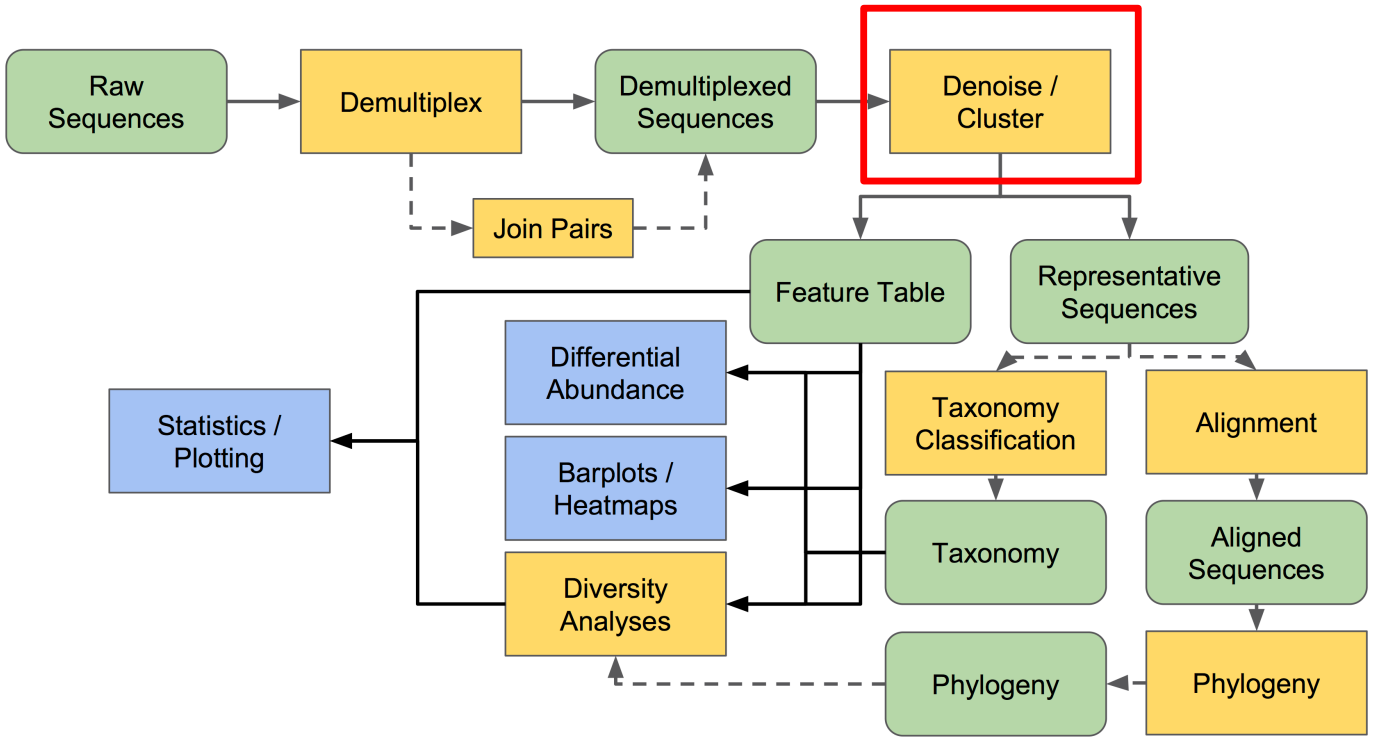
Demultiplexed sequence counts summary

	forward reads
Minimum	1854
Median	8646.5
Mean	7762.676471
Maximum	18787
Total	263931

Forward Reads Frequency Histogram



QIIME2 STEP 3 – QUALITY CONTROL



https://docs.qiime2.org/2024.10/tutorials/overview/#useful-points-for-beginners

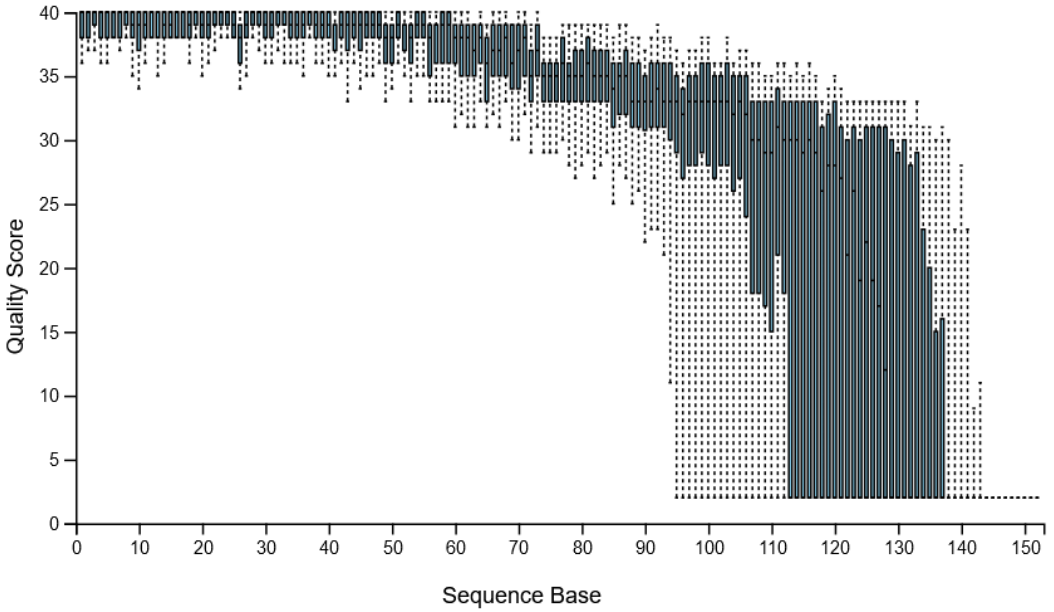
QIIME2 STEP 3 – QUALITY CONTROL

The demux.qzv file will also contain information on sequence quality

Note that in Illumina, sequence quality will sometimes be a bit lower for the first few reads, then increase. This is normal and is generally not a cause for concern.

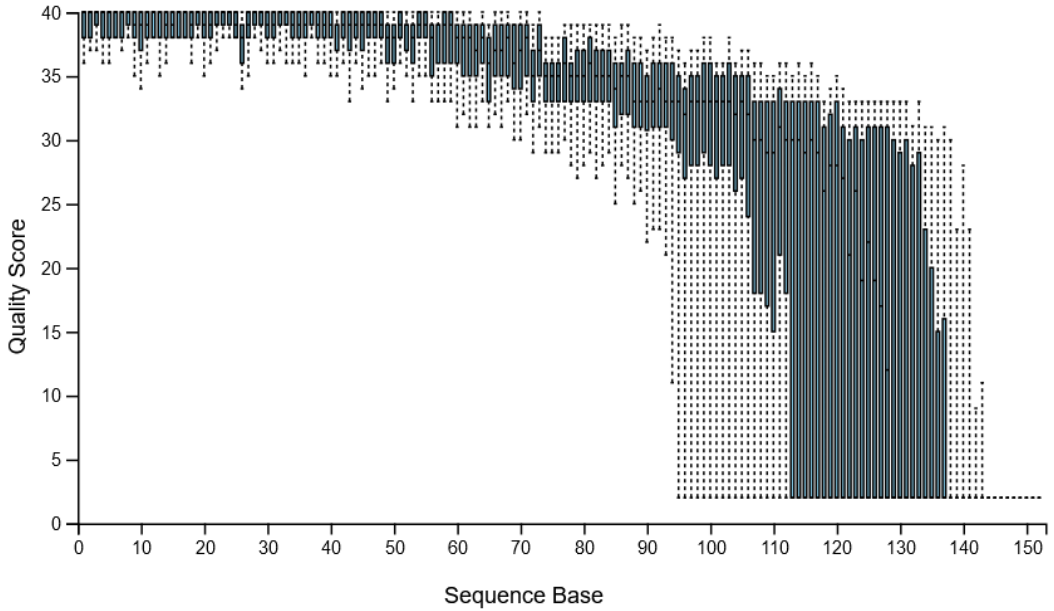
Bases that are of lower quality can choose to be **trimmed** from the sequence for downstream analysis.

- Trimming can occur at the beginning or end of a sequence



QIIME2 STEP 3 – QUALITY CONTROL

Question: If we want to ensure 99.9% accuracy for all positions in our sequences, which Phred score should we use as a threshold? And which base should we use to trim to from the right-hand side in the below plot?



convert ascii33 to error probability

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

front

BITS VIB Bioinformatics Training and Service Facility

hg18-total-sequenced=2'858'034'764 (UCSC)

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
!	33	0	1.00E+00	0.000%	1	2,858,034,764
"	34	1	7.94E-01	20.567%	1	2,270,217,709
#	35	2	6.31E-01	36.904%	2	1,803,298,025
$	36	3	5.01E-01	49.881%	2	1,432,410,537
%	37	4	3.98E-01	60.189%	3	1,137,804,133
&	38	5	3.16E-01	68.377%	3	903,789,949
`	39	6	2.51E-01	74.881%	4	717,905,874
(	40	7	2.00E-01	80.047%	5	570,252,906
)	41	8	1.58E-01	84.151%	6	452,967,984
*	42	9	1.26E-01	87.411%	8	359,805,259
+	43	10	1.00E-01	90.000%	10	285,803,476
,	44	11	7.94E-02	92.057%	13	227,021,771
-	45	12	6.31E-02	93.690%	16	180,329,803
.	46	13	5.01E-02	94.988%	20	143,241,054
/	47	14	3.98E-02	96.019%	25	113,780,413
0	48	15	3.16E-02	96.838%	32	90,378,995
1	49	16	2.51E-02	97.488%	40	71,790,587
2	50	17	2.00E-02	98.005%	50	57,025,291
3	51	18	1.58E-02	98.415%	63	45,296,798
4	52	19	1.26E-02	98.741%	79	35,980,526
5	53	20	1.00E-02	99.000%	100	28,580,348
6	54	21	7.94E-03	99.206%	126	22,702,177
7	55	22	6.31E-03	99.369%	158	18,032,980
8	56	23	5.01E-03	99.499%	200	14,324,105
9	57	24	3.98E-03	99.602%	251	11,378,041
:	58	25	3.16E-03	99.684%	316	9,037,899
;	59	26	2.51E-03	99.749%	398	7,179,059
<	60	27	2.00E-03	99.800%	501	5,702,529
=	61	28	1.58E-03	99.842%	631	4,529,680
>	62	29	1.26E-03	99.874%	794	3,598,053
?	63	30	1.00E-03	99.900%	1,000	2,858,035
@	64	31	7.94E-04	99.921%	1,259	2,270,218
A	65	32	6.31E-04	99.937%	1,585	1,803,298
B	66	33	5.01E-04	99.950%	1,995	1,432,411
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253
G	71	38	1.58E-04	99.984%	6,310	452,968
H	72	39	1.26E-04	99.987%	7,943	359,805
I	73	40	1.00E-04	99.990%	10,000	285,803
J	74	41	7.94E-05	99.992%	12,589	227,022
K	75	42	6.31E-05	99.994%	15,849	180,330
L	76	43	5.01E-05	99.995%	19,953	143,241
M	77	44	3.98E-05	99.996%	25,119	113,780
N	78	45	3.16E-05	99.997%	31,623	90,379
O	79	46	2.51E-05	99.997%	39,811	71,791

QIIME2 STEP 3 – QUALITY CONTROL

We will use the DADA2 tool (part of QIIME2) to perform the trimming

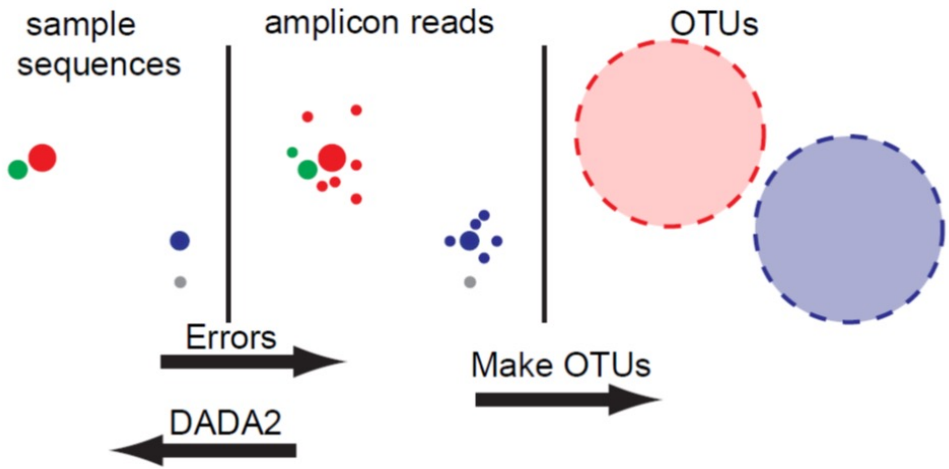
```
qiime dada2 denoise-single \
  --i-demultiplexed-seqs demux.qza \
  --p-trim-left 0 \
  --p-trunc-len 120 \
  --o-representative-sequences rep-seqs-dada2.qza \
  --o-table table-dada2.qza \
  --o-denoising-stats stats-dada2.qza
```

Output artifacts:

- `stats-dada2.qza` : view | download
- `table-dada2.qza` : view | download
- `rep-seqs-dada2.qza` : view | download

QIIME2 STEP 3 – QUALITY CONTROL

- During PCR amplification, there are inherent errors that occur.
- DADA2 then models these sequencing errors and uses the error rates within the data to infer sample composition
  - DADA2 = Divisive Amplicon Denoising Algorithm 2
- Utilizes the Needleman-Wunsch algorithm to align sequences. Then, based on the alignment and error model it creates, it determines whether a read is truly a unique read, or if it is simply a mistake that was made in the sequencing step

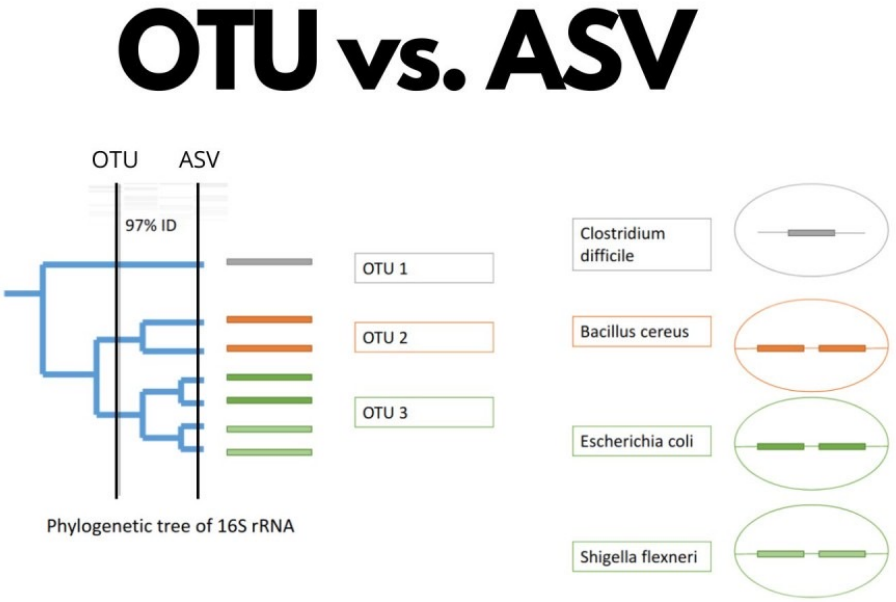


https://www.toolsbiotech.com/news_detail.php?id=81

OTU VS ASV… SO MANY ACRONYMS!

We can obtain multiple types of output data from 16S sequencing

- OTU: Observational Taxonomic Unit
  - Bins samples based on sequence similarity
- ASV: Amplicon Sequence Variant
  - Achieves features that differ at single nucleotide resolution



https://es.linkedin.com/pulse/otu-vs-asv-steven-criollo-msc-3icoe

Questions:

1. Why does nearly everyone use ASVs today instead of OTUs?
2. If someone was to generate an OTU/ASV in their dataset, then another group generates an OTU/ASV at a later time in a different dataset, would they be comparable to one another?

QIIME2 STEP 3 – QUALITY CONTROL

And then we can create the .qzv file to visualize the results of DADA2

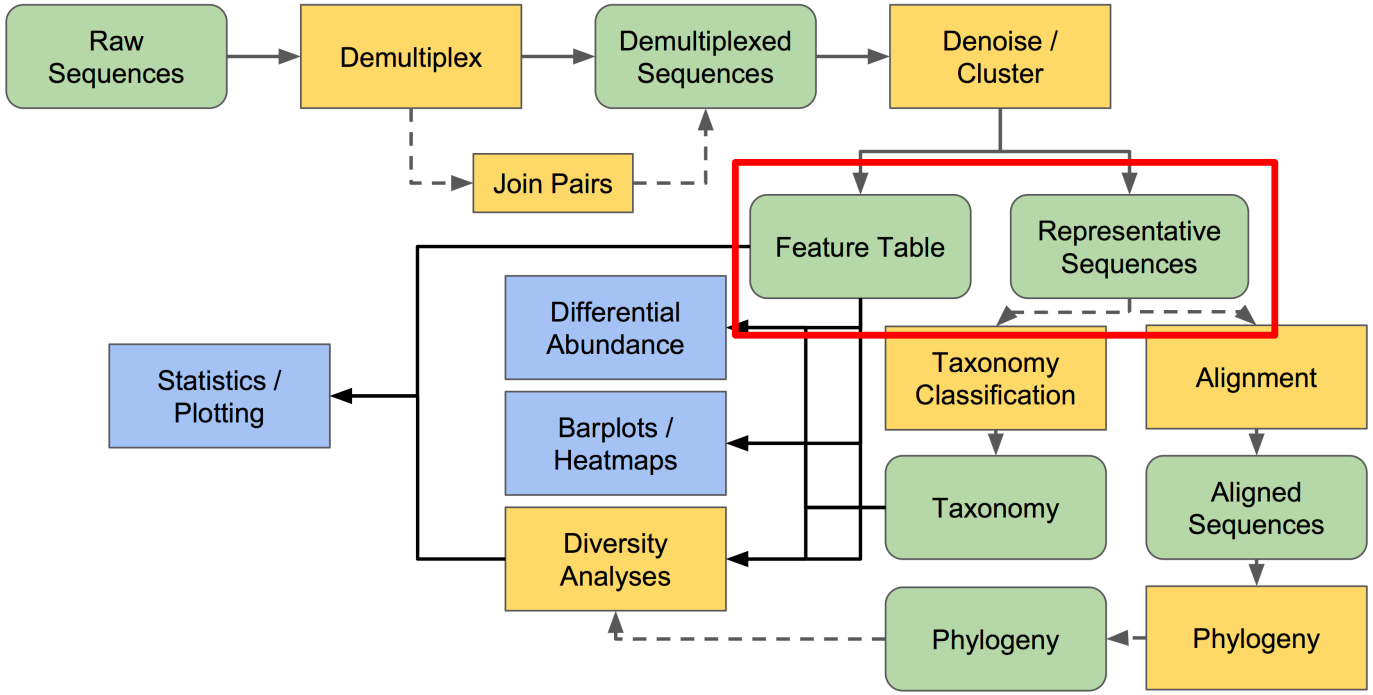
```
qiime metadata tabulate \
  --m-input-file stats-dada2.qza \
  --o-visualization stats-dada2.qzv
```

Output visualizations:

- `stats-dada2.qzv` : view | download

This will just give you basic information on how many sequences per sample were filtered out

QIIME2 STEP 4 – CREATING THE FEATURE TABLE



https://docs.qiime2.org/2024.10/tutorials/overview/#useful-points-for-beginners

QIIME2 STEP 4 – CREATING THE FEATURE TABLE

Next we can create two of the most important files in QIIME2:

```
qiime feature-table summarize \
  --i-table table.qza \
  --o-visualization table.qzv \
  --m-sample-metadata-file sample-metadata.tsv
qiime feature-table tabulate-seqs \
  --i-data rep-seqs.qza \
  --o-visualization rep-seqs.qzv
```

Output visualizations:

- `table.qzv` : view | download
- `rep-seqs.qzv` : view | download

QIIME2 STEP 4 – CREATING THE FEATURE TABLE

Let's first look at table.qzv

This contains both sample-based information as well as feature-based information.

Specifically, let's look at the "Feature Detail" tab.

	Frequency	# of Samples Observed In
4b5eeb300368260019c1fbc7a3c718fc	11,373	13
fe30ff0f71a38a39cf1717ec2be3a2fc	8,929	16
d29fe3c70564fc0f69f2c03e0d1e5561	8,622	25
868528ca947bc57b69ffdf83e6b73bae	7,663	10
154709e160e8cada6bfb21115acc80f5	7,412	13

Question: What do you think these strings of characters represent?

QIIME2 STEP 4 – CREATING THE FEATURE TABLE

Now let's take a look at ref-seqs.qzv

This contains information on the unique sequences identified with DADA2. At this point, we can call these Amplicon Sequence Variants (ASVs)

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
770	120	120	120.0	0	0.0

Question: Why is our minimum, mean, and maximum read length the same?

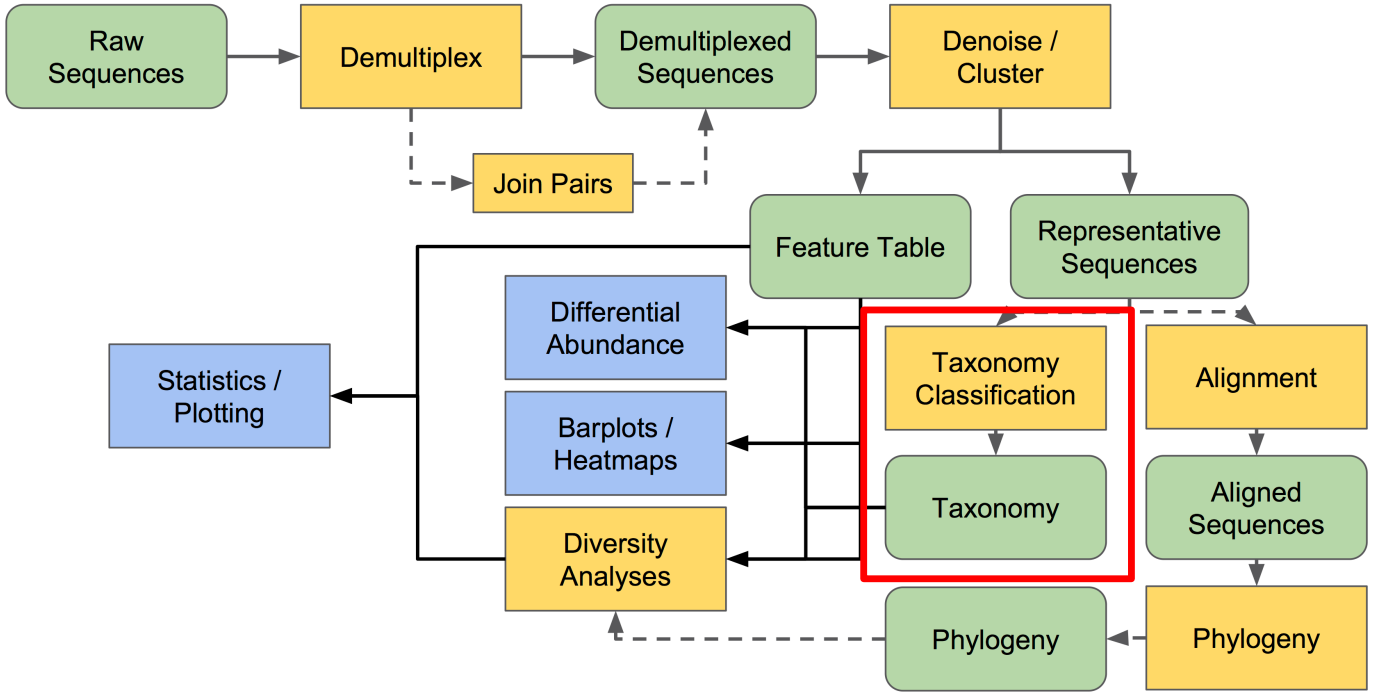
QIIME2 STEP 4 – CREATING THE FEATURE TABLE

ref-seqs.qzv also contains the actual sequence we can use for identifying the microbe

Feature ID	Sequence ▼	Sequence Length
c18826df5af5da174f580164c805a38a	TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTGAAAGGCAGTGGCTCAACCATTGTAGGCTTTGGAAACTGTTTAACTTG	120
bae6b03d63955f134b6e523285263226	TACGGAGGATCCAAGCGTTATCCGGAATCATTGGGTTTAAAGGGTCCGTAGGCGGCCCTGTAAGTCAGCGGTGAAATCTCCCGGCTCAACCGGGAAATGGCCGTTGATACTGCAGGGCTT	120
f99678fdc17d27a97c79faac3f2980a4	TACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGGGCGCAGACGGTTATTTAAGCAGGATGTGAAATCCCCGGGCTCAACCTGGGAACGGCGTTCTGAACTGGATAGCTA	120
c99439f78b73ffcf55858ae828aac608	TACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGAGCGTAGGTGGCTTGATAAGTCAGATGTGAAAGCCCCGGGCTCAACCTGGGAACGGCATCTGATACTGTTAGGCTA	120

Question: What is it that this sequence represents in the bacterial genome?

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION



https://docs.qiime2.org/2024.10/tutorials/overview/#useful-points-for-beginners

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION

Now that we have our features, we need to know which microbes they are!

In QIIME2, we use taxonomic **feature classifiers** to perform this task

- A classifier is an algorithm used in machine learning that assigns “features” to categories
- In our case, we are assigning our 16S sequences to taxonomic labels

As stated in the QIIME2 docs (https://docs.qiime2.org/2024.10/tutorials/feature-classifier/), to create a good feature classifier, we need **reference sequences** and the **taxonomic classifications** of these sequences

- Both can be obtained from the QIIME2 website
- SILVA reference sequences can also be obtained from there

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION

Training the classifier

Step 1: Import the reference and taxonomy files to create qza files

```
qiime tools import \
  --type 'FeatureData[Sequence]' \
  --input-path 85_otus.fasta \
  --output-path 85_otus.qza

qiime tools import \
  --type 'FeatureData[Taxonomy]' \
  --input-format HeaderlessTSVTaxonomyFormat \
  --input-path 85_otu_taxonomy.txt \
  --output-path ref-taxonomy.qza
```

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION

Training the Naïve Bayes classifier

Step 2: Extract the reference reads that are from the exact location used in your sequencing. This is to ensure the classifier is trained on the specific region of DNA that you sequenced.

```
qiime feature-classifier extract-reads \
  --i-sequences 85_otus.qza \
  --p-f-primer GTGCCAGCMGCCGCGGTAA \
  --p-r-primer GGACTACHVGGGTWTCTAAT \
  --p-trunc-len 120 \
  --p-min-length 100 \
  --p-max-length 400 \
  --o-reads ref-seqs.qza
```

Forward and reverse primer sequences you selected for sequencing

Number of nucleotides you truncated to in the DADA2 step

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION

Training the Naïve Bayes classifier

Step 3: Train the classifier

```
qiime feature-classifier fit-classifier-naive-bayes \
  --i-reference-reads ref-seqs.qza \
  --i-reference-taxonomy ref-taxonomy.qza \
  --o-classifier classifier.qza
```

Note: This is the new ref-seqs.qza from the previous step, not the original one

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION

Training the Naïve Bayes classifier

Step 4: Test to ensure the classifier works

```
qiime feature-classifier classify-sklearn \
  --i-classifier classifier.qza \
  --i-reads rep-seqs.qza \
  --o-classification taxonomy.qza

qiime metadata tabulate \
  --m-input-file taxonomy.qza \
  --o-visualization taxonomy.qzv
```

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION

We see that the step did work and we can now map the feature we got previously to the taxon!

Feature ID #q2:types	Taxon categorical	Confidence categorical
0023879846315b5eea4cc545513fb264	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__	0.9993652465188472
003803ad4793fac9ebe6a54c78ad5a93	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__	0.9044875670924011
0062cb1703222371ca8dba17a77f9c18	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales	0.9978597908154107
00ab9bd631ca3a99f0ff659dd68515ca	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__[Paraprevotellaceae]; g__; s__	0.7884958249837584
00ea325b7c0b57797ee084e17e068299	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__	0.9891279166858383
010b498be37bdab8841b42301fab093e	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__S24-7; g__; s__	0.9961521960224529

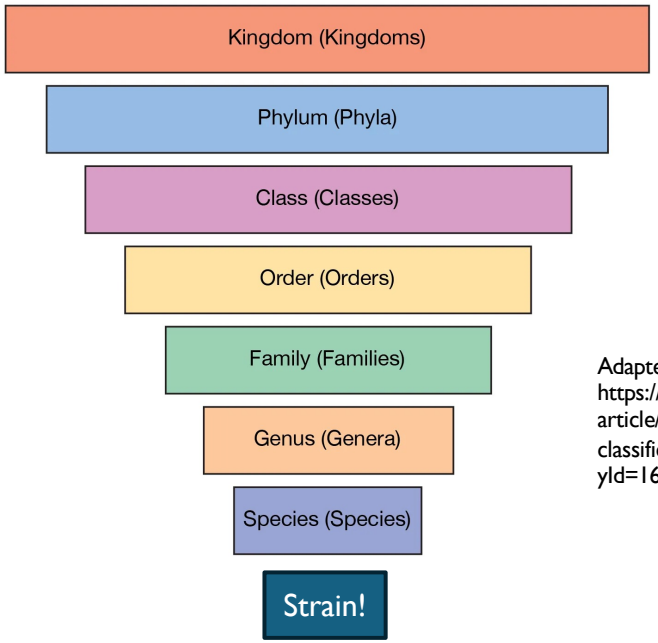
Question: How do you read these Taxon labels? What's up with the ones that just have something like "g__"?

QIIME2 STEP 5 – TAXONOMIC CLASSIFICATION

Reminder: Katy Perry Came Over For Good Soup

Ideally, we want all our ASV assignments to be at the species (or even strain) level

- Realistically though, this is not reasonable
- There are a lot of unknown bacteria!



Adapted from https://kids.britannica.com/students/article/biological-classification/611149/media?assemblyId=166807

USING BLAST TO (POSSIBLY) GET BETTER RESOLUTION

- BLAST: Basic Local Alignment Search Tool
  - Input a sequence of nucleotides and it will search a database for an organism with the closest sequence match

- Let's take this FASTA 16S sequence, for example, and BLAST it to see if we can find a match

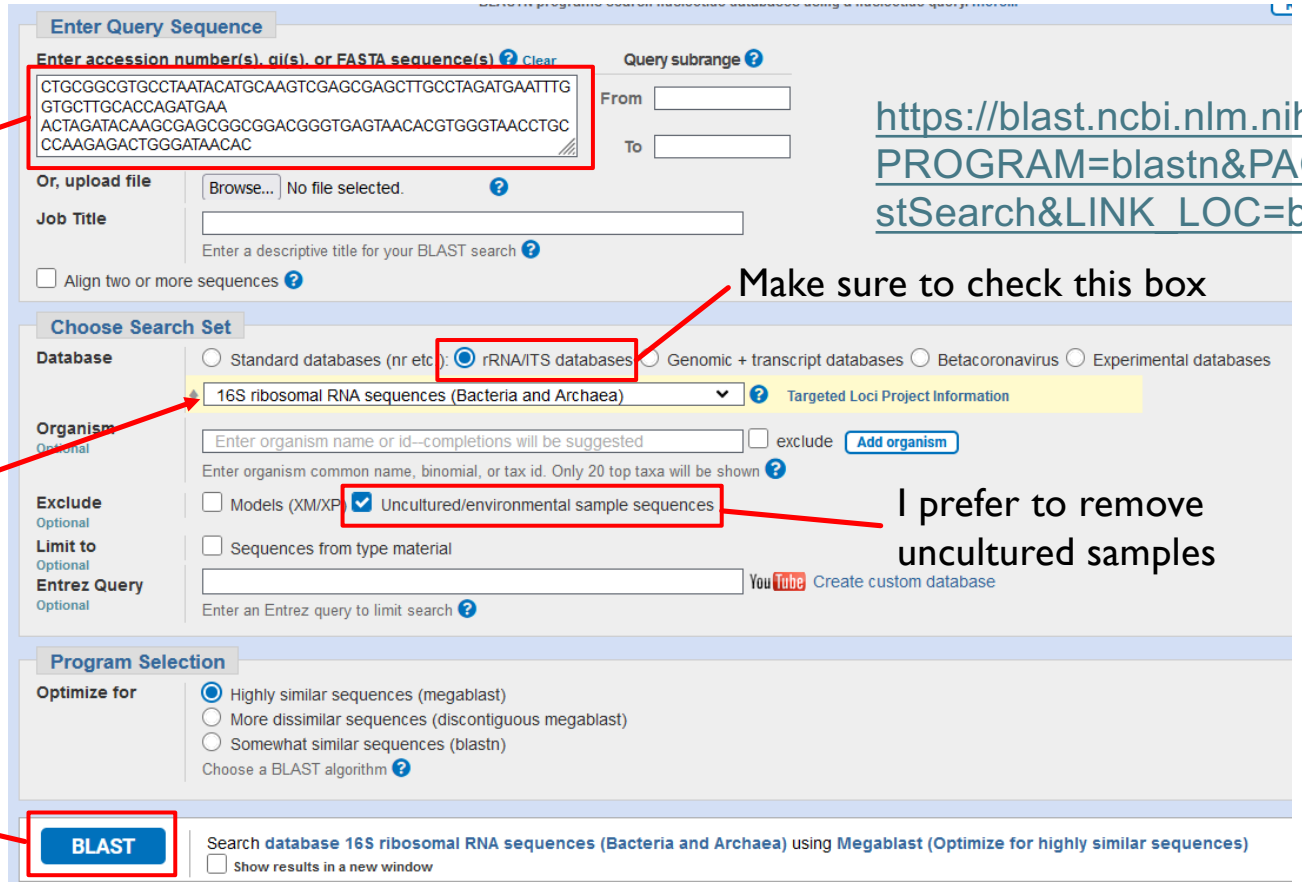
CTGCGGCGTGCCTAATACATGCAAGTCGAGCGAGCTTGCCTAGATGAATTT
GGTGCTTGCACCAGATGAAACTAGATACAAGCGAGCGGCGGACGGGTGAG
TAACACGTGGGTAACCTGCCCAAGAGACTGGGATAACACCTGGAAACAGAT
GCTAATACCGGATAACAACACTAGACGCAT

USING BLAST TO (POSSIBLY) GET BETTER RESOLUTION

16S sequence

Select this option in the drop-down

GO!



https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Make sure to check this box

I prefer to remove uncultured samples

USING BLAST TO (POSSIBLY) GET BETTER RESOLUTION

- We can see that the sequence is that of *Lactobacillus gasseri*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Lactobacillus gasseri strain CIP 102991 16S ribosomal RNA, partial sequence	Lactobacillus gasseri	337	337	100%	2e-92	100.00%	1529	NR_117573.1
☑	Lactobacillus gasseri ATCC 33323 = JCM 1131 16S ribosomal RNA, complete sequence	Lactobacillus gasseri ...	333	333	98%	2e-91	100.00%	1573	NR_075051.2
☑	Lactobacillus gasseri ATCC 33323 = JCM 1131 16S ribosomal RNA, partial sequence	Lactobacillus gasseri ...	333	333	98%	2e-91	100.00%	1572	NR_041920.1
☑	Lactobacillus paragasseri strain JCM 5343 16S ribosomal RNA, partial sequence	Lactobacillus paragas...	327	327	98%	1e-89	99.44%	1511	NR_179379.1
☑	Lactobacillus paragasseri strain JCM 5343 16S ribosomal RNA, partial sequence	Lactobacillus paragas...	327	327	98%	1e-89	99.44%	1495	NR_179257.1
☑	Lactobacillus johnsonii strain ATCC 33200 16S ribosomal RNA, partial sequence	Lactobacillus johnsonii	311	311	98%	1e-84	97.78%	1487	NR_025273.1
☑	Lactobacillus johnsonii strain CIP 103620 16S ribosomal RNA, partial sequence	Lactobacillus johnsonii	311	311	98%	1e-84	97.78%	1543	NR_117574.1

Question: We also see there are other Lactobacillus here in the output though. Why?

QIIME2 STEP 6 – EXPORTING THE FEATURE TABLE

We can now export our original feature table we made, then modify it so it's easier to work with downstream

```
qiime tools export \
  --input-path feature-table.qza \
  --output-path exported-feature-table
```

QIIME2 STEP 6 – EXPORTING THE FEATURE TABLE

This should produce a file in the resulting directory that looks something that looks like the following, where the rows are the ASV IDs and the columns are the sample names.

#ASV ID	lane1-s121-index-GTACCTAATTGC-121	lane1-s122-index-ACTCACGGTATG-122	lane1-s123-index-GTCTACACACAT-123	lane1-s124-index-ATACTTCGCAGG-124
1f64d820f	5576	5091	18831	3259
4c5683a6a	2293	5216	3011	5965
1da913198	227	250	743	133
ddfde9751	6887	8827	1055	7608
0343d97b8	8705	7617	6819	8132
c692b0bd8	1928	833	82	349
eda6eca30	11	16	10	10
641cf0060	452	605	1128	319
5abbbb944	792	1389	2314	2327
fae30c57b	1977	2093	1314	2491
f9318fdc58	1336	939	4758	1274
fcd9ec88d	470	407	377	469

QIIME2 STEP 6 – EXPORTING THE FEATURE TABLE

At this point, I would suggest converting the ASV IDs into something more manageable. **JUST MAKE ABSOLUTELY SURE TO SAVE A MAPPING FILE FOR THIS.**

Note: I did not do this in this example, but in hindsight I should have also included the 16S sequence in this file that corresponds to each ID

#ASV ID	Updated ID	taxonomy
1f64d820f7c0f162d75592f31b06ddcf	ASV1	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__S24-7; g__; s__
4c5683a6aabb837532ffb3d707784bbb	ASV2	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__; g__; s__
1da9131983b379a76004172f7e944abc	ASV3	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Lactococcus; s__
ddfde975148c98138bf71de53041bd79	ASV4	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__S24-7; g__; s__
0343d97b85092d9ff471b7aa75d35f4b	ASV5	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__S24-7; g__; s__
c692b0bd8e197177c9ef880e3388b2a0	ASV6	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus; s__
eda6eca307701892ae08b6cc45df00cf	ASV7	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae
641cf0060866bd5dc2a1d541cc181f2d	ASV8	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptostreptococcaceae; g__; s__
5abbbb9446dcf7f0de85c9564a1e9404	ASV9	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillospira; s__
fae30c57bbda49164e4c12c30ecf8e69	ASV10	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__; g__; s__
f9318fdc58cd51496b825706980f7250	ASV11	k__Bacteria; p__Firmicutes; c__Bacilli; o__Turicibacterales; f__Turicibacteraceae; g__Turicibacter; s__
fcd9ec88d408427cc3ad75bec342ee0c	ASV12	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
f6d486ba616dc626f44d227387fb9a1c	ASV13	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__; g__; s__
6c11efc9f352007dfbc2e220144b5bd2	ASV14	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__; g__; s__

I prefer to stop with QIIME2 at this point and use alternative software for normalization and downstream analysis

PREPROCESSING THE FEATURE TABLE

#ASV ID	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s13
ASV2	2293	5216	3011	5965	4344	12007	5950	8529	11536	7357
ASV1	5576	5091	18831	3259	4028	6240	13481	11432	11445	6767
ASV5	8705	7617	6819	8132	15599	2316	1228	4518	1723	2639
ASV4	6887	8827	1055	7608	4929	1879	742	2114	1481	1093
ASV10	1977	2093	1314	2491	1173	2492	1933	2055	1830	1568
ASV11	1336	939	4758	1274	1084	917	1280	1708	588	777
ASV9	792	1389	2314	2327	1991	2853	1372	1225	2505	1195
ASV21	0	78	681	730	0	2394	2553	835	1553	3403
ASV15	914	1560	2214	633	942	628	211	1640	355	1131
ASV13	844	449	899	1961	1462	1447	826	563	984	1142
ASV158	24	30	21	39	38	79	44	31	65	64
ASV126	41	16	28	13	17	47	34	30	63	32
ASV136	68	79	52	46	48	20	17	16	10	20
ASV160	0	27	26	6	0	49	45	25	29	76
ASV159	11	27	27	29	18	37	25	13	24	31
ASV175	98	480	0	71	164	0	0	0	0	0
ASV391	0	0	42	0	0	0	0	0	0	0
ASV23	0	0	0	0	0	100	200	200	0	0
ASV398	0	4	30	30	25	50	10	10	10	0
ASV302	0	0	0	0	0	0	0	0	0	20
ASV415	0	0	0	0	0	0	0	0	0	0
ASV280	0	0	0	0	1	0	0	0	0	0
ASV350	0	0	0	0	0	0	0	0	0	0
ASV371	2	0	0	0	0	0	0	0	0	0
ASV337	0	0	0	0	0	0	0	0	0	0
ASV344	0	0	0	0	0	0	0	0	0	0
ASV302	0	0	0	0	0	0	0	0	0	0
ASV415	0	0	0	0	0	0	0	0	0	0

I have condensed this feature table to only contain a handful of ASVs, but keep in mind that a real table will be much larger.

Questions:

1. What do you notice about this table?
2. Do you spot any issues with the data that may affect our downstream analysis?
3. How may we go about addressing this/these issue(s)?

PREPROCESSING THE FEATURE TABLE

- In a typical 16S sequencing run, you may have more than 1,000 unique ASVs in your feature table.
  - Many of these are lowly abundant, with there only being a single instance of it in a single sample.

- What we can do is remove these microbes with a low abundance

Question: What impact does removing low-abundance microbes have in our analysis? Is it safe to assume that microbes with a low abundance do not play a significant role in your study?

PREPROCESSING THE FEATURE TABLE

- We can first identify which ASVs have a low abundance by calculating the cumulative abundance of all ASVs (column CA in the table)
- Then, we choose a threshold we wish to use
  - 99%
  - 99.5%
  - 99.9%
  - Etc.

#ASV ID	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s13	Sum	Relativized	CA
ASV2	2293	5216	3011	5965	4344	12007	5950	8529	11536	7357	66208	19.70%	19.7015%
ASV1	5576	5091	18831	3259	4028	6240	13481	11432	11445	6767	86150	25.64%	45.3372%
ASV5	8705	7617	6819	8132	15599	2316	1228	4518	1723	2639	59296	17.64%	62.9820%
ASV4	6887	8827	1055	7608	4929	1879	742	2114	1481	1093	36615	10.90%	73.8775%
ASV10	1977	2093	1314	2491	1173	2492	1933	2055	1830	1568	18926	5.63%	79.5093%
ASV11	1336	939	4758	1274	1084	917	1280	1708	588	777	14661	4.36%	83.8720%
ASV9	792	1389	2314	2327	1991	2853	1372	1225	2505	1195	17963	5.35%	89.2172%
ASV21	0	78	681	730	0	2394	2553	835	1553	3403	12227	3.64%	92.8556%
ASV15	914	1560	2214	633	942	628	211	1640	355	1131	10228	3.04%	95.8992%
ASV13	844	449	899	1961	1462	1447	826	563	984	1142	10577	3.15%	99.0466%
ASV158	24	30	21	39	38	79	44	31	65	64	435	0.13%	99.1760%
ASV126	41	16	28	13	17	47	34	30	63	32	321	0.10%	99.2715%
ASV136	68	79	52	46	48	20	17	16	10	20	376	0.11%	99.3834%
ASV160	0	27	26	6	0	49	45	25	29	76	283	0.08%	99.4676%
ASV159	11	27	27	29	18	37	25	13	24	31	242	0.07%	99.5397%
ASV175	98	480	0	71	164	0	0	0	0	0	813	0.24%	99.7816%
ASV391	0	0	42	0	0	0	0	0	0	0	42	0.01%	99.7941%
ASV23	0	0	0	0	0	100	200	200	0	0	500	0.15%	99.9429%
ASV398	0	4	30	30	25	50	10	10	10	0	169	0.05%	99.9932%
ASV302	0	0	0	0	0	0	0	0	0	20	20	0.01%	99.9991%
ASV415	0	0	0	0	0	0	0	0	0	0	0	0.00%	99.9991%
ASV280	0	0	0	0	1	0	0	0	0	0	1	0.00%	99.9994%
ASV350	0	0	0	0	0	0	0	0	0	0	0	0.00%	99.9994%
ASV371	2	0	0	0	0	0	0	0	0	0	2	0.00%	100.0000%
ASV337	0	0	0	0	0	0	0	0	0	0	0	0.00%	100.0000%
ASV344	0	0	0	0	0	0	0	0	0	0	0	0.00%	100.0000%
ASV302	0	0	0	0	0	0	0	0	0	0	0	0.00%	100.0000%
ASV415	0	0	0	0	0	0	0	0	0	0	0	0.00%	100.0000%

NORMALIZATION OF 16S DATA

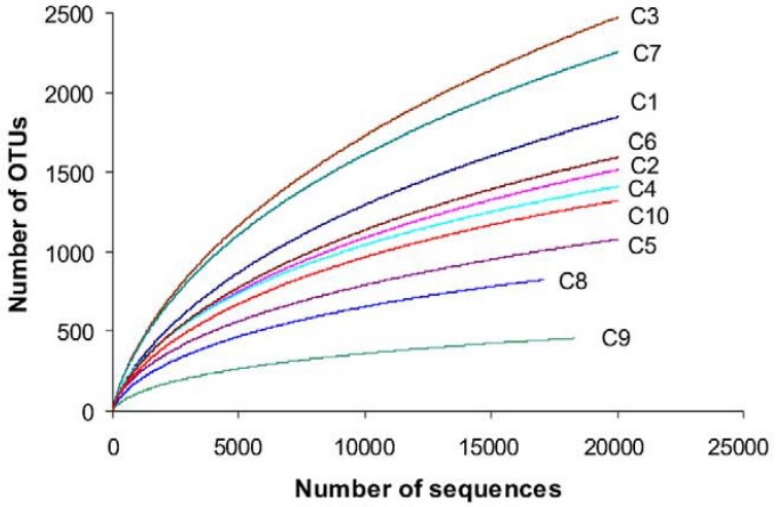
There are MANY types of normalization methods for microbiome data, but we will focus on some of the more common ones

- Rarefication
- Total sum scaling
- Log transformation
- Quantile normalization
- Upper quartile normalization
- DESeq2's median of ratios

RAREFACTION

- Tries to account for differences in sequencing depth by ensuring even sequencing depth among all samples
- Employs drawing without replacement

Question: The use of rarefaction is highly debated in the field. What do you think the reason is for this?



TOTAL SUM SCALING

- Converts reads to **relative abundances**
- Divides each read by the sum of all reads in the sample
- Can be performed after rarefaction

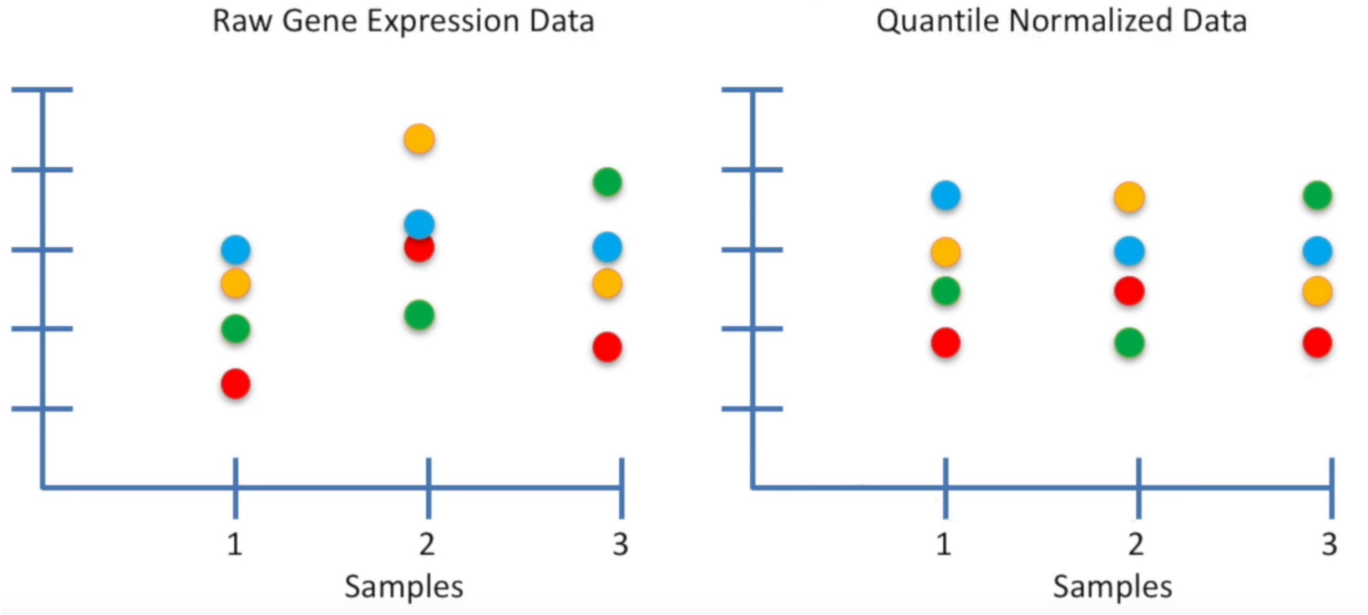
#ASV ID	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s13
ASV1	5576	5091	18831	3259	4028	6240	13481	11432	11445	6767
ASV2	2293	5216	3011	5965	4344	12007	5950	8529	11536	7357
ASV5	8705	7617	6819	8132	15599	2316	1228	4518	1723	2639
ASV4	6887	8827	1055	7608	4929	1879	742	2114	1481	1093
ASV10	1977	2093	1314	2491	1173	2492	1933	2055	1830	1568
ASV9	792	1389	2314	2327	1991	2853	1372	1225	2505	1195
ASV11	1336	939	4758	1274	1084	917	1280	1708	588	777
ASV13	844	449	899	1961	1462	1447	826	563	984	1142
ASV21	0	78	681	730	0	2394	2553	835	1200	2000
ASV15	914	1560	2100	633	942	540	211	1200	355	1131
ASV175	98	480	0	71	164	0	0	0	0	0
ASV23	0	0	0	0	0	100	200	200	0	0
ASV158	24	30	21	39	38	79	44	31	65	64
ASV136	68	79	52	46	48	20	17	16	10	20
ASV160	0	12	26	6	0	12	45	12	29	76
ASV126	10	16	10	13	17	11	10	15	10	12
ASV159	5	0	5	5	5	37	5	13	5	31

#ASV ID	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s13
ASV1	0.188831	0.150283	0.44947	0.0943	0.112439	0.18714	0.450915	0.331689	0.33895	0.261557
ASV2	0.077652	0.153973	0.071868	0.172598	0.121259	0.360095	0.199017	0.247461	0.341645	0.284361
ASV5	0.294795	0.224849	0.16276	0.235301	0.435434	0.069458	0.041074	0.131086	0.051028	0.102002
ASV4	0.233228	0.260568	0.025181	0.220139	0.137589	0.056352	0.024819	0.061336	0.043861	0.042246
ASV10	0.066951	0.061784	0.031363	0.072078	0.032743	0.074736	0.064655	0.059624	0.054197	0.060606
ASV9	0.026821	0.041002	0.055232	0.067332	0.055577	0.085563	0.045891	0.035542	0.074187	0.046189
ASV11	0.045244	0.027719	0.113567	0.036863	0.030259	0.027501	0.042814	0.049556	0.017414	0.030032
ASV13	0.028582	0.013254	0.021458	0.056742	0.040811	0.043396	0.027628	0.016335	0.029142	0.04414
ASV21	0	0.002303	0.016255	0.021123	0	0.071797	0.085393	0.024227	0.035539	0.077304
ASV15	0.030953	0.04605	0.050124	0.018316	0.026295	0.016195	0.007058	0.034817	0.010514	0.043715
ASV175	0.003319	0.014169	0	0.002054	0.004578	0	0	0	0	0
ASV23	0	0	0	0	0	0.002999	0.00669	0.005803	0	0
ASV158	0.000813	0.000886	0.000501	0.001128	0.001061	0.002369	0.001472	0.000899	0.001925	0.002474
ASV136	0.002303	0.002332	0.001241	0.001331	0.00134	0.0006	0.000569	0.000464	0.000296	0.000773
ASV160	0	0.000354	0.000621	0.000174	0	0.00036	0.001505	0.000348	0.000859	0.002938
ASV126	0.000339	0.000472	0.000239	0.000376	0.000475	0.00033	0.000334	0.000435	0.000296	0.000464
ASV159	0.000169	0	0.000119	0.000145	0.00014	0.00111	0.000167	0.000377	0.000148	0.001198

Note that each column sums to 1

QUANTILE NORMALIZATION

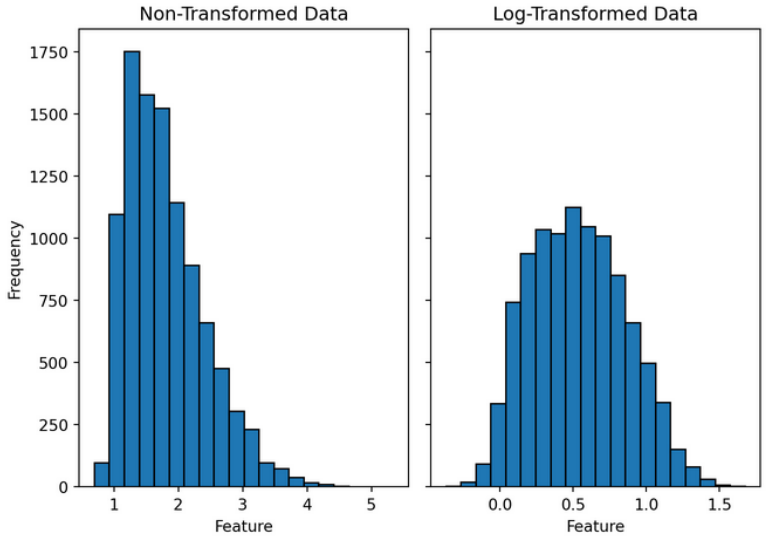
Normalizes data into their quantiles by first ranking each ASV per-sample, then finding the average abundance value for the ranks, and shifting the raw values to that average



https://www.youtube.com/watch?v=ecjN6Xpv6SE

LOG TRANSFORMATION

- Applied to right-skewed data
  - Meaning lots of samples with smaller expression/abundance and fewer with larger expression/abundance
- Typically, we do log(value)+1 since the data is zero-inflated, and you cannot take the log of 0



https://dataalltheway.com/posts/001-data-transformation/index.html

UPPER QUARTILE NORMALIZATION

Finds the value of an upper quartile (commonly 75%) and divides all features by that value

Question: How is this different from total sum scaling? What will the resulting range of values look like at the end?

DESEQ2'S MEDIAN OF RATIOS

DESeq2 calculates a scaling factor based on both read depth and the entire library composition

1. For each gene (row), calculate the geometric mean of the gene across **all** samples (columns)
2. In each sample, divide the raw value of this gene by the calculated geometric mean of that gene.
3. For each sample, find the median of the ratios and divide each value by that median

Calculate GeoMean

#ASV ID	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s13	GeoMean
ASV1	5577	5092	18832	3260	4029	6241	13482	11433	11446	6768	7445.405
ASV2	2294	5217	3012	5966	4345	12008	5951	8530	11537	7358	5867.329
ASV5	8706	7618	6820	8133	15600	2317	1229	4519	1724	2640	4494.155
ASV4	6888	8828	1056	7609	4930	1880	743	2115	1482	1094	2548.072
ASV10	1978	2094	1315	2492	1174	2493	1934	2056	1831	1569	1843.681
ASV9	793	1390	2315	2328	1992	2854	1373	1226	2506	1196	1671.644
ASV11	1337	940	4759	1275	1085	918	1281	1709	589	778	1225.164
ASV13	845	450	900	1962	1463	1448	827	564	985	1143	971.2456
ASV21	1	79	682	731	1	2395	2554	836	1201	2001	233.6066
ASV15	915	1561	2101	634	943	541	212	1201	356	1132	794.0034
ASV175	99	481	1	72	165	1	1	1	1	1	7.503427
ASV23	1	1	1	1	1	101	201	201	1	1	4.582172
ASV158	25	31	22	40	39	80	45	32	66	65	40.91437
ASV136	69	80	53	47	49	21	18	17	11	21	31.62323
ASV160	1	13	27	7	1	13	46	13	30	77	11.60022
ASV126	11	17	11	14	18	12	11	16	11	13	13.16638
ASV159	6	1	6	6	6	38	6	14	6	32	7.762416

Divide

Calculate median of ratios

#ASV ID	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s13
ASV1	0.749053	0.683912	2.529345	0.437854	0.541139	0.838235	1.810781	1.535578	1.537324	0.909017
ASV2	0.390979	0.889161	0.513351	1.016817	0.740541	2.046587	1.01426	1.453813	1.966312	1.254063
ASV5	1.937183	1.69509	1.517527	1.809684	3.471175	0.515559	0.273466	1.005528	0.383609	0.58743
ASV4	2.703221	3.464581	0.414431	2.98618	1.934796	0.737813	0.291593	0.830039	0.581616	0.429344
ASV10	1.072854	1.135771	0.713247	1.351644	0.63677	1.352186	1.048988	1.11516	0.993122	0.851015
ASV9	0.474383	0.831517	1.384864	1.392641	1.191641	1.707301	0.821347	0.73341	1.499123	0.715463
ASV11	1.091282	0.767244	3.884377	1.040677	0.885595	0.749287	1.045574	1.394915	0.480752	0.635017
ASV13	0.870017	0.463323	0.926645	2.020086	1.506313	1.490869	0.851484	0.580698	1.014162	1.176839
ASV21	0.004281	0.338175	2.919438	3.129192	0.004281	10.25228	10.93291	3.578666	5.141121	8.565682
ASV15	1.152388	1.965987	2.646084	0.798485	1.187652	0.681357	0.267001	1.512588	0.448361	1.425687
ASV175	13.19397	64.10404	0.133272	9.595615	21.98995	0.133272	0.133272	0.133272	0.133272	0.133272
ASV23	0.218237	0.218237	0.218237	0.218237	0.218237	22.04195	43.86566	43.86566	0.218237	0.218237
ASV158	0.611032	0.75768	0.537708	0.977652	0.95321	1.955303	1.099858	0.782121	1.613125	1.588684
ASV136	2.181941	2.529786	1.675983	1.486249	1.549494	0.664069	0.569202	0.53758	0.347846	0.664069
ASV160	0.086205	1.120669	2.327542	0.603437	0.086205	1.120669	3.965443	1.120669	2.586158	6.637806
ASV126	0.835461	1.291167	0.835461	1.063314	1.367118	0.911412	0.835461	1.215216	0.835461	0.987363
ASV159	0.772955	0.128826	0.772955	0.772955	0.772955	4.895383	0.772955	1.803562	0.772955	4.122428
Median	0.835461	0.889161	0.926645	1.063314	0.95321	1.120669	0.851484	1.120669	0.835461	0.909017

Divide

Normalized values

#ASV ID	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s12	lane1-s129-index-CG
ASV1	0.896574	0.769165	2.729573	0.411782	0.567702	0.747978	2.126618	1.370234	1.840091	1
ASV2	0.467979	1	0.553989	0.956272	0.776892	1.82622	1.191168	1.297273	2.353565	1.379581
ASV5	2.318699	1.906393	1.637657	1.701928	3.641563	0.460045	0.321164	0.897257	0.459159	0.646225
ASV4	3.235603	3.896461	0.447238	2.80837	2.029769	0.658368	0.342453	0.740665	0.696162	0.472317
ASV10	1.284146	1.277352	0.769709	1.271161	0.668026	1.206589	1.231953	0.995085	1.188711	0.936192
ASV9	0.56781	0.93517	1.494492	1.309717	1.250135	1.523467	0.964607	0.654439	1.794366	0.787074
ASV11	1.306203	0.862885	4.191871	0.978711	0.929066	0.668607	1.227943	1.244717	0.575433	0.698575
ASV13	1.041361	0.521078	1	1.899802	1.580253	1.330339	1	0.518171	1.213895	1.294628
ASV21	0.005124	0.380331	3.150546	2.942867	0.004491	9.148358	12.83983	3.193331	6.153634	9.423015
ASV15	1.379344	2.211058	2.855553	0.75094	1.24595	0.607992	0.313572	1.349719	0.536663	1.568383
ASV175	15.79244	72.09498	0.143823	9.024253	23.06936	0.118922	0.156518	0.118922	0.15952	0.146612
ASV23	0.261218	0.245442	0.235513	0.205242	0.22895	19.66857	51.51673	39.14241	0.261218	0.24008
ASV158	0.731371	0.852129	0.580274	0.919438	1	1.744765	1.291696	0.697906	1.93082	1.747694
ASV136	2.611661	2.845139	1.808657	1.397752	1.625553	0.592565	0.668482	0.479695	0.416352	0.730535
ASV160	0.103183	1.260366	2.511795	0.567506	0.090437	1	4.657096	1	3.095486	7.30218
ASV126	1	1.452119	0.901598	1	1.434225	0.813275	0.981182	1.084367	1	1.086188
ASV159	0.925184	0.144885	0.834144	0.72693	0.810897	4.36827	0.907774	1.609363	0.925184	4.535039

SO WHICH DO I USE?

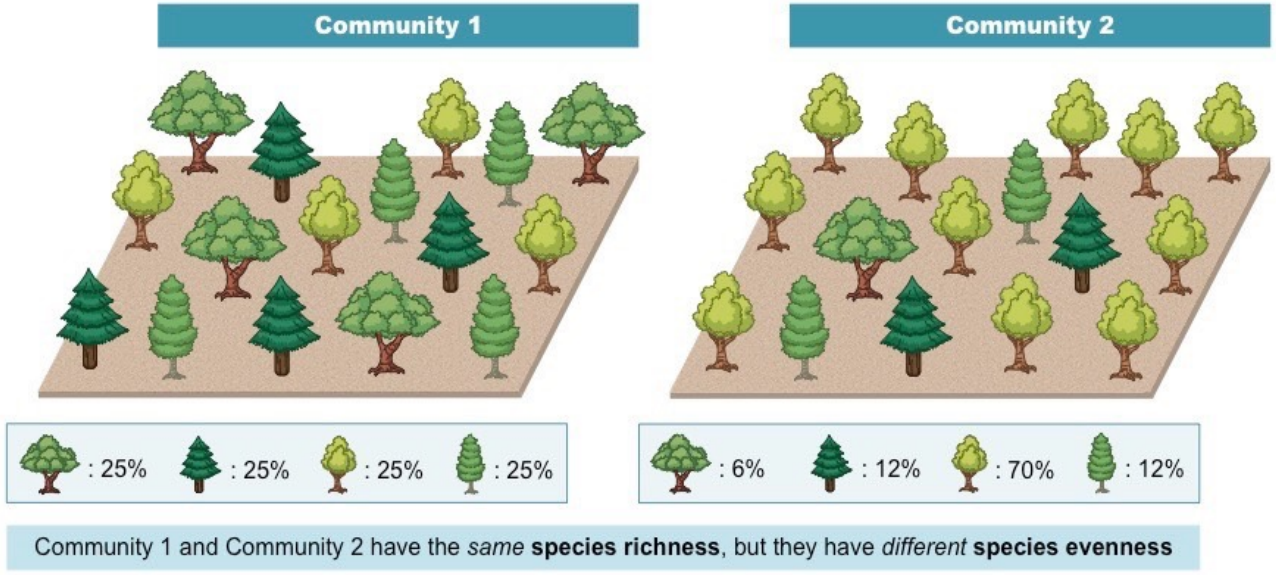
- In practice, there is no one single best answer for this, and I am certainly not qualified to tell you the best option
- Explore your data and work with experts in the field to determine the best course of action for you!

BASIC MICROBIOME ANALYSIS TECHNIQUES

- Species diversity
  - Alpha diversity
  - Beta diversity
- Functional analysis
- Pathway-based analysis
- Machine learning approaches
- Network-based techniques

ALPHA DIVERSITY

- A measure of the **species richness** and **species evenness** in a community
  - Species richness: the number of species in a community
  - Species evenness: a measure that takes into account the presence/absence of individuals as well as their abundance



https://old-ib.bioninja.com.au/options/option-c-ecology-and-conser/c4-conservation-of-biodiver/biodiversity.html

ALPHA DIVERSITY METRICS

- Note that in ecological terms, we are using the term "species" to belonging to a single feature, but in microbiome analysis, this is what we refer to as an ASV
- **Shannon diversity index (H)**
  - Accounts for both presence and abundance of species
- **Simpson's index (D)** also accounts for the same, but just uses a different formula
- Note: Shannon index values are high in communities with high evenness, while Simpson's values are heavily influenced by a few dominant species
- In practice, neither is necessarily better than the other. Just understand the underlying assumptions of each

$$H' = -\sum(p_i \cdot \ln(p_i))$$

Proportion of individuals belonging to species *i*

$$D = \sum(p_i^2)$$

BETA DIVERSITY

- While alpha diversity explains the diversity of species within a single community, **beta diversity** is used to compare two communities
- **Dysbiosis**: the abnormal change in microbial composition
- Described using different metrics
  - Bray-Curtis dissimilarity
  - Weighted UniFrac
  - Unweighted UniFrac
  - Jaccard distance

BRAY-CURTIS DISSIMILARITY

- Scaled from 0 to 1, where 0 means they are identical and 1 means they are completely dissimilar

1. Count the number of each species in each sample
2. For each species that is shared between samples, sum up the minimum of each species
3. Plug the values into the formula

Sum of minimums

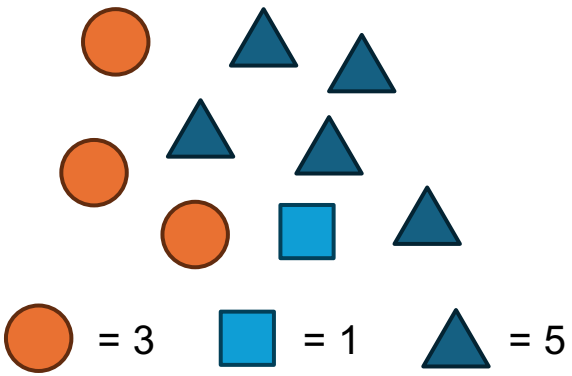
$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

Number of species in each community

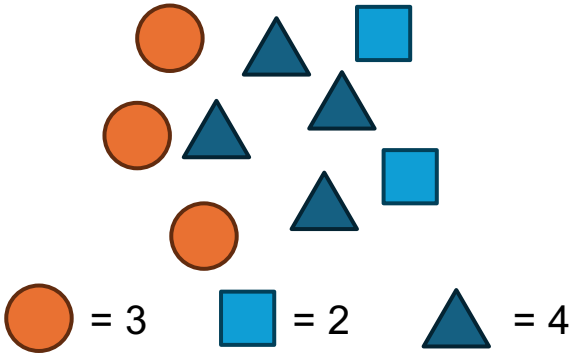
Adapted from https://docs.onecodex.com/en/articles/4150649-beta-diversity

BRAY-CURTIS DISSIMILARITY

Community 1



Community 2



Sum of minimums

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

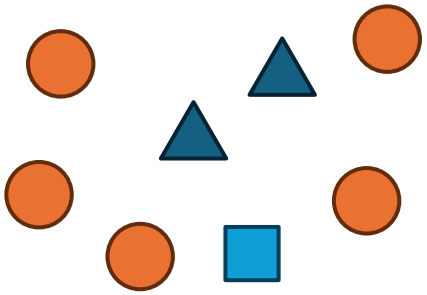
Number of species in each community

Question: By just looking at these two communities, do you think there is much dissimilarity between them?

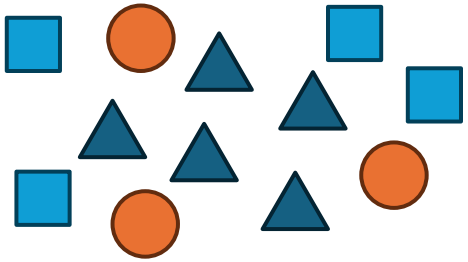
BCij = 1 – (2 * (3+1+4)) / ((3+1+5) + (3+2+4))
BCij = 1 – 16/(9 + 9)
BCij = 1 – 16/18
BCij = 0.11

BRAY-CURTIS DISSIMILARITY

Community 1



Community 2



Sum of minimums

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

Number of species in each community

Question: Now consider these communities. What would be the dissimilarity score between them?

BCij = 1 – (2 * (3+2+1)) / ((2+1+5) + (3+5+4))

BCij = 1 – 12/(8 + 12)

BCij = 1 – 12/20

BCij = 0.4

BRAY-CURTIS DISSIMILARITY

- Note that for BC dissimilarity, it is highly dependent on the number of species! A small change in the composition of a small community can drastically change the BC value

Sum of minimums

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

Number of species in each community

UNIFRAC METRICS

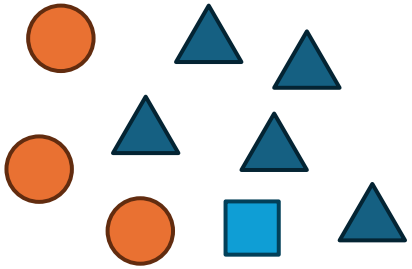
- UniFrac metrics takes into consideration phylogenies, scaling the dissimilarity score based on branch lengths between species
  - Unweighted UniFrac does not depend on microbial abundance
  - Weighted UniFrac does depend on microbial abundance

JACCARD DISTANCE

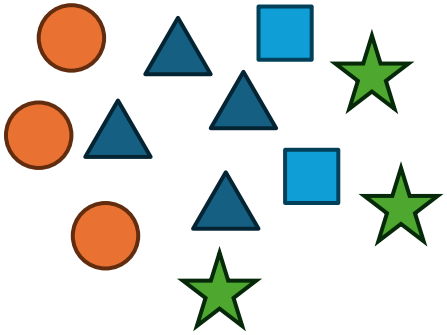
- Does not depend on microbial abundance, just presence of species

$$Jaccard_{AB} = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Community 1



Community 2



Jaccard = 1 – 3/4 = 0.25

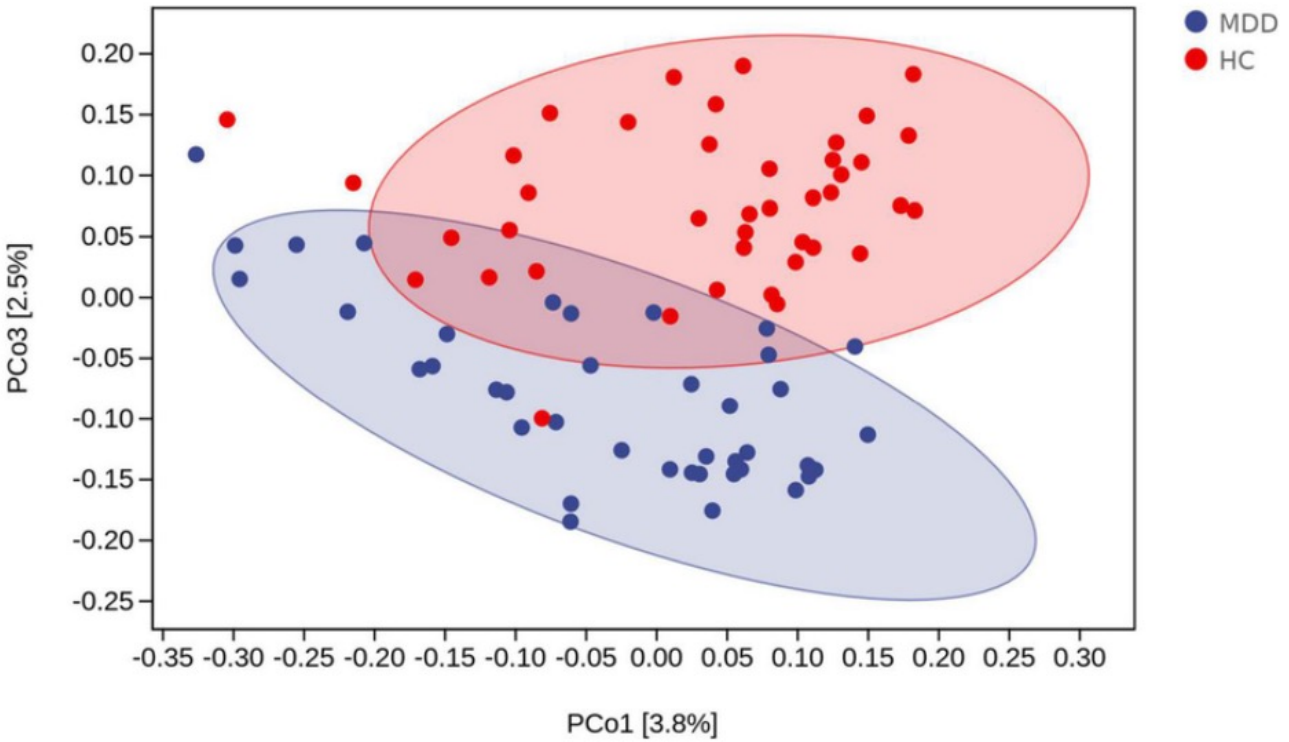
Question: How does keeping low-abundant microbes in your preprocessing step affect this score?

PCA/PCOA PLOTS

- 16S data is high dimensional (many features)
- PCoA: Principle coordinates analysis
  - PCA is a type of PCoA that uses Euclidean distances
  - We tend to visualize beta diversity by using PCoA in microbiome studies
- PCoA plots are limited in their interpretability

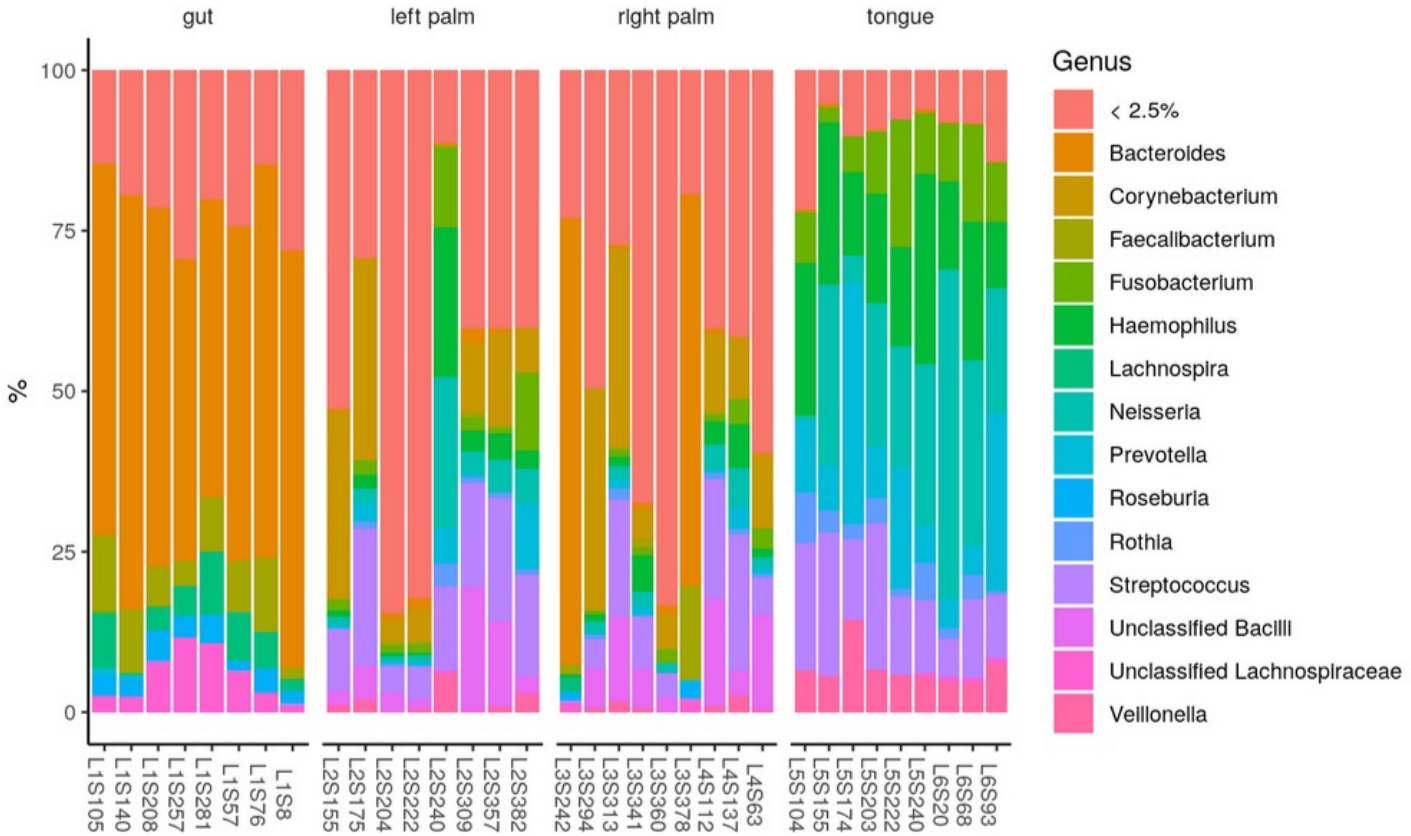
PCA/PCOA PLOTS

Question: In this study by Liu et al., they researchers were interested in what the differences were in gut microbiota composition between individuals with major depressive disorder (MDD) and healthy controls (HC). What can this plot tell us about their microbial populations?



https://www.researchgate.net/publication/358622955_Gut_Microbiome_Composition_Linked_to_Inflammatory_Factors_and_Cognitive_Functions_in_First-Episode_Drug-Naive_Major_Depressive_Disorder_Patients

RELATIVE ABUNDANCE PLOTS



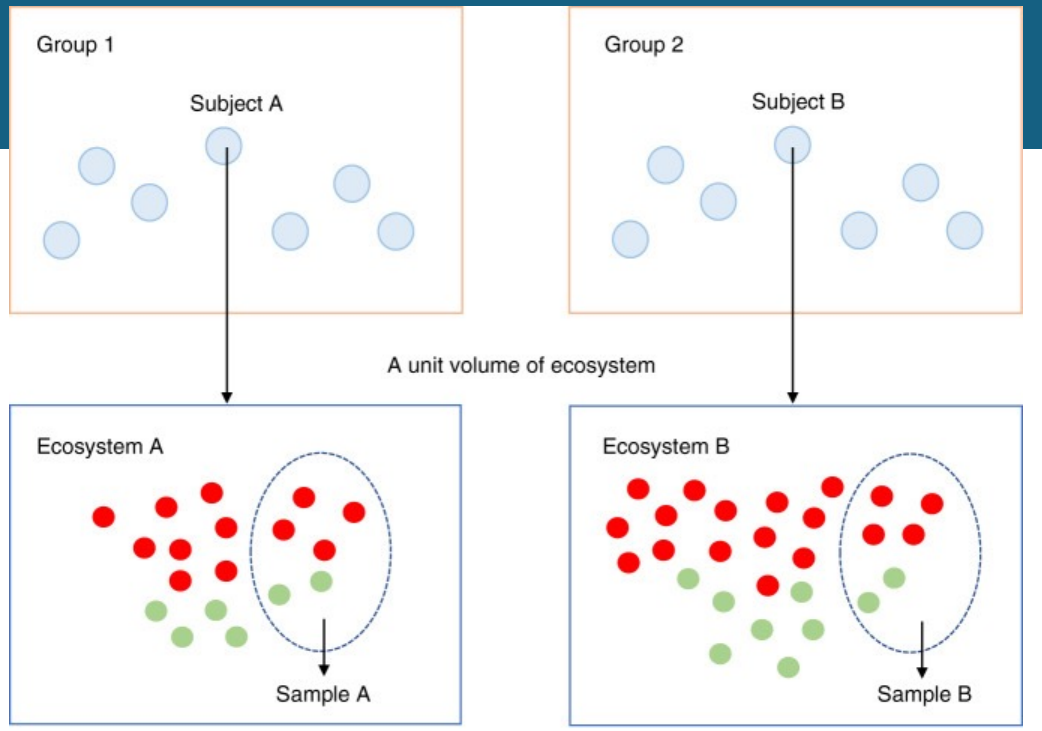
https://www.yanh.org/2021/01/01/microbiome-r/

DIFFERENTIAL ABUNDANCE TECHNIQUES

- t-test
  - Assumes a normal distribution in samples, which is not always a good assumption to make
- Mann-Whitney
  - Does not assume normality in sample distributions
  - Compares the abundance/expression of features based on their ranks instead of their raw values
- R packages: DESeq2, edgeR, metagenomeSeq (will not cover in-depth today)
- ANCOM-BC
- Machine learning methods
  - LEfSe (Linear discriminant analysis Effect Size) (Not the food)
  - Random forest models

ANCOM-BC

- Lin and Das Peddada, 2020
- “Analysis of Compositions of Microbiomes with Bias Correction”
- A new method that takes into account the sampling fraction
- In the image to the right, both subjects were identified as having the same relative abundance of red microbes, but in reality there is a significant difference in the microbial populations between the two ecosystems
- Introduces a per-sample offset term in a linear model, which it calculates based on the observed data

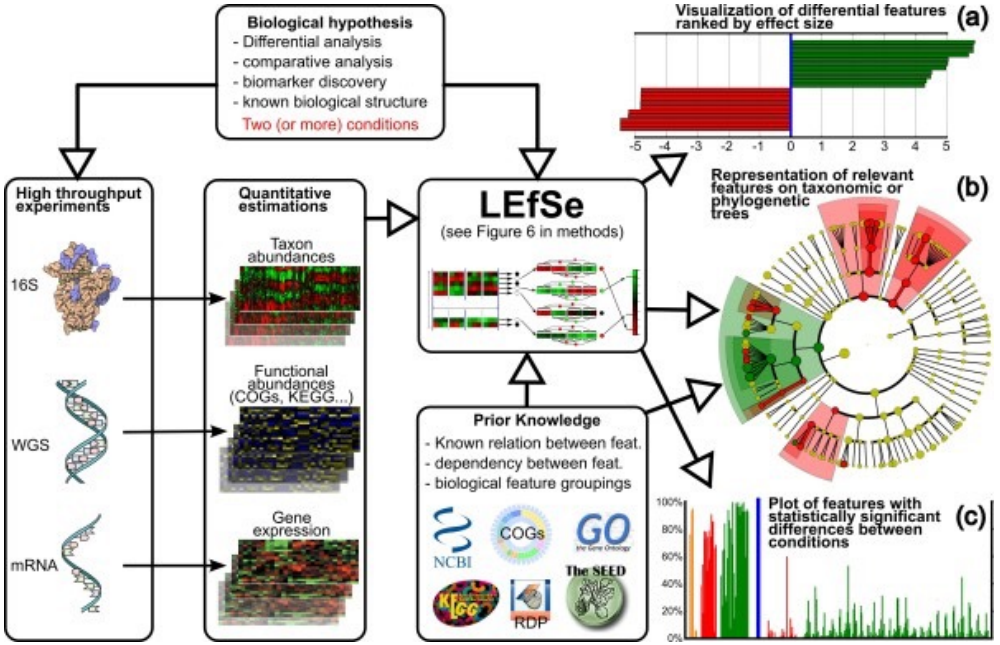


	Sample		Ecosystem	
	A	B	A	B
●	4	4	12	18
●	2	2	6	9
Sum	6†	6†	18‡	27‡

† Library size; ‡ microbial load.

LEFSE

- Linear discriminant analysis (LDA) effect size
- Available as a Galaxy module (http://galaxy.biobakery.org/)
- Finds the ASVs that are most likely to describe the differences between your groups
- Employs prior knowledge of biology, differential abundance tests (Kruskal-Wallis), and LDA to estimate the **effect size** of each feature



https://pmc.ncbi.nlm.nih.gov/articles/PMC3218848/

Metagenomic biomarker discovery and explanation

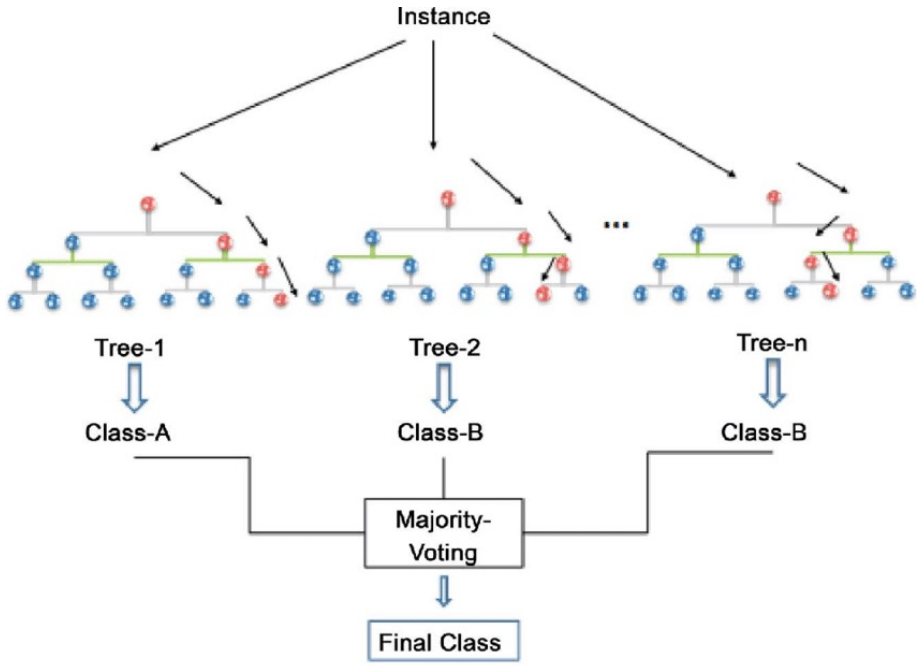
Nicola Segata [1], Jacques Izard [2,3], Levi Waldron [1], Dirk Gevers [4], Larisa Miropolsky [1], Wendy S Garrett [5,6,7], Curtis Huttenhower [1,✉]

▸ Author information ▸ Article notes ▸ Copyright and License information

PMCID: PMC3218848 PMID: 21702898

RANDOM FORESTS

- Utilize decision trees in their classification steps
- Randomly select samples from the original dataset
- Randomly select features and make many random decision trees to see which best separate individuals into their known classifications
- Make many, many trees, using this method and see which one most accurately classifies your samples that were not used as input
- These also provide the **<u>importance of each feature</u>** in the model, which is measured by taking the reduction in accuracy of the model if the feature is removed



Demonstration of the Random Forest methodology

https://www.researchgate.net/figure/Demonstration-of-the-random-Forest-methodology_fig1_322098019

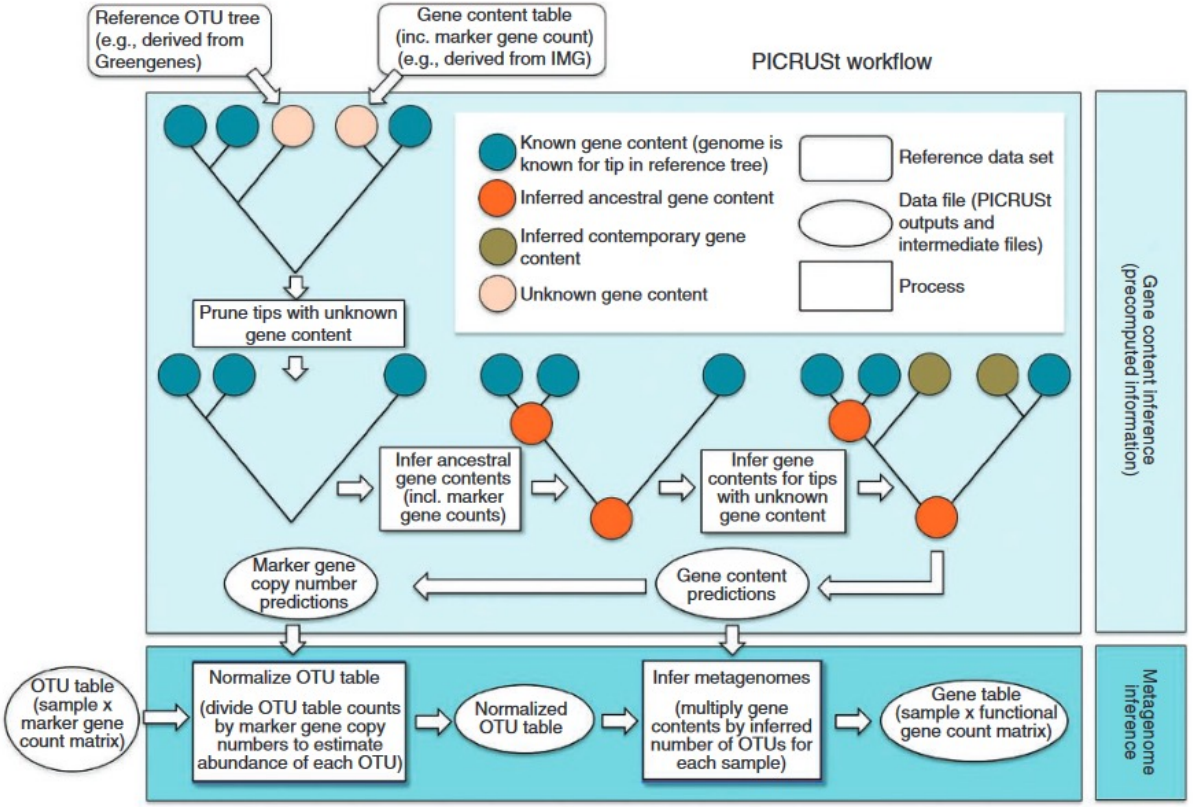
SOFTWARE FOR DOWNSTREAM MICROBIOME ANALYSIS

- PICRUSt 2.0 and Tax4Fun2 (will only discuss PICRUSt today)
- HUMAnN 3.0
- Vegan (R package, not discussed in detail in this lecture)
- MicrobiomeAnalyst (one of my favorites!)

PICRUST

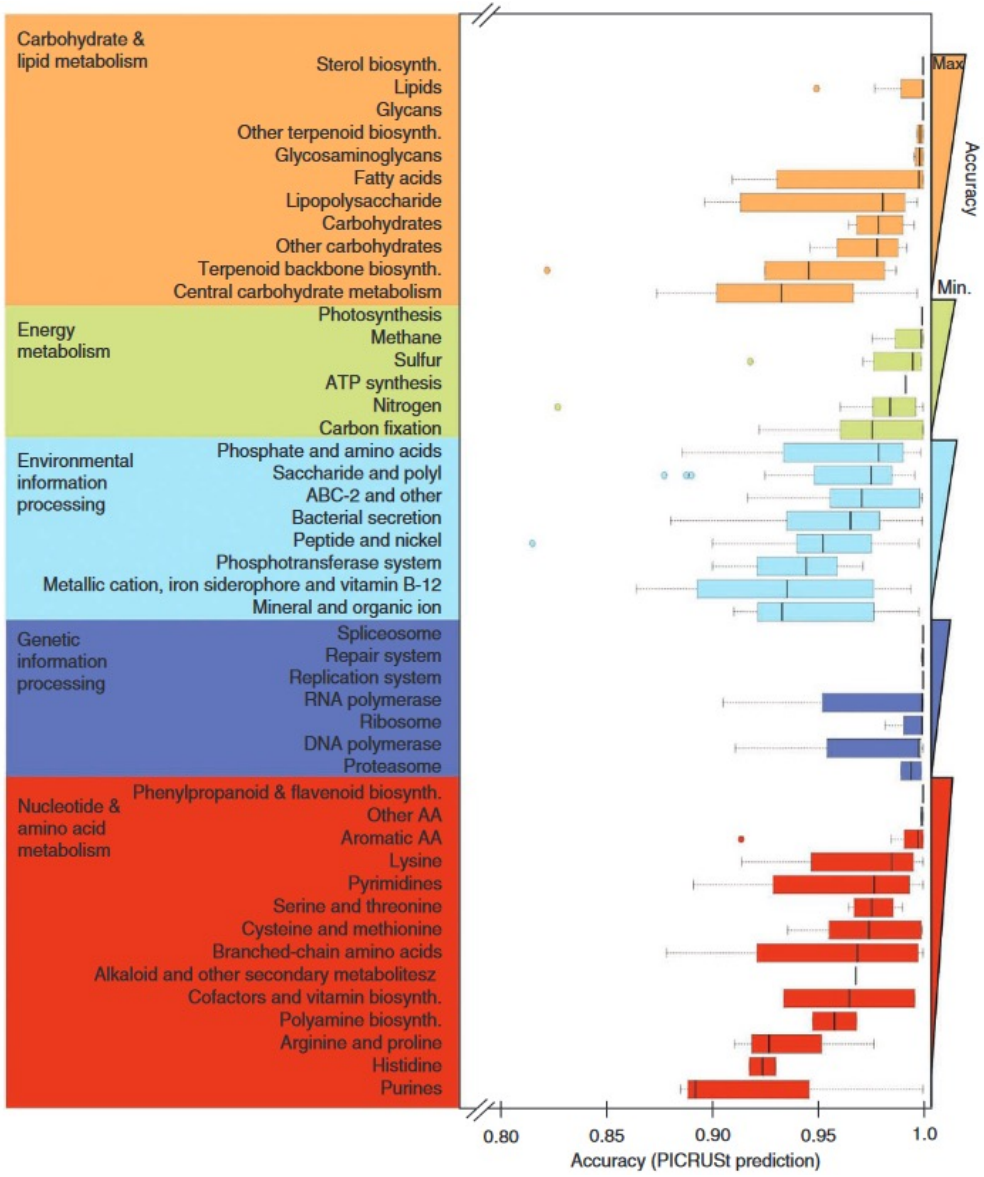
- Estimates the gene families and function based on 16S data
  - Each species has a different functional potential, depending on the proteins they can produce and their metabolism of lipids
- Very easy to use, but **make sure not to overinterpret results**
  - Results are suggestive only
- https://huttenhower.sph.harvard.edu/picrust/

PICRUST



Langille et al., 2013

PICRUST



Langille et al., 2013

PICRUST VS TAX4FUN

- According to Sun et al., who compared PICRUSt, PICRUST 2.0, and Tax4Fun:
  - They all work relatively well on human samples, but struggle with other organisms
  - They perform well for housekeeping gene functions, not so great for other genes
  - All have similar performance, do not recommend using one over another

Short report | Open access | Published: 02 April 2020

Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories

Shan Sun, Roshonda B. Jones & Anthony A. Fodor

Microbiome **8**, Article number: 46 (2020) | Cite this article

HUMANN 3.0

- https://learn.gencore.bio.nyu.edu/metgenomics/shotgun-metagenomics/functional-analysis/
- Similar to PICRUSt/Tax4Fun, but specifically for metagenomic data
  - Predicts gene family abundance, pathway abundance, and pathway coverage

MICROBIOMEANALYST

- https://www.microbiomeanalyst.ca/
- Online tool that takes as input the feature table and taxonomic information and performs many types of analyses
- Free, easy to use, and is frequently updated
- Also allows for raw data processing, like what we did with QIIME2 previously
- Let's pause here and take a look at MicrobiomeAnalyst

MICROBIOMEANALYST LEARNING OBJECTIVES

1. Familiarize yourself with the input to MicrobiomeAnalyst
2. Explore what happens as you vary the rarefaction threshold and perform different pre-processing steps
3. Perform alpha- and beta-diversity analyses
4. Identify features that are differentially abundant between treatment groups

PUTTING ALL THIS INTO PERSPECTIVE

Microbiome Research Reports

Home About Publish with us Articles Special Issues Volumes Features

Home > Articles > Article

Original Article | Open Access | 10 Jul 2024

Host response to cholestyramine can be mediated by the gut microbiota

Views: 423 | Downloads: 47 | Cited: Crossref 0

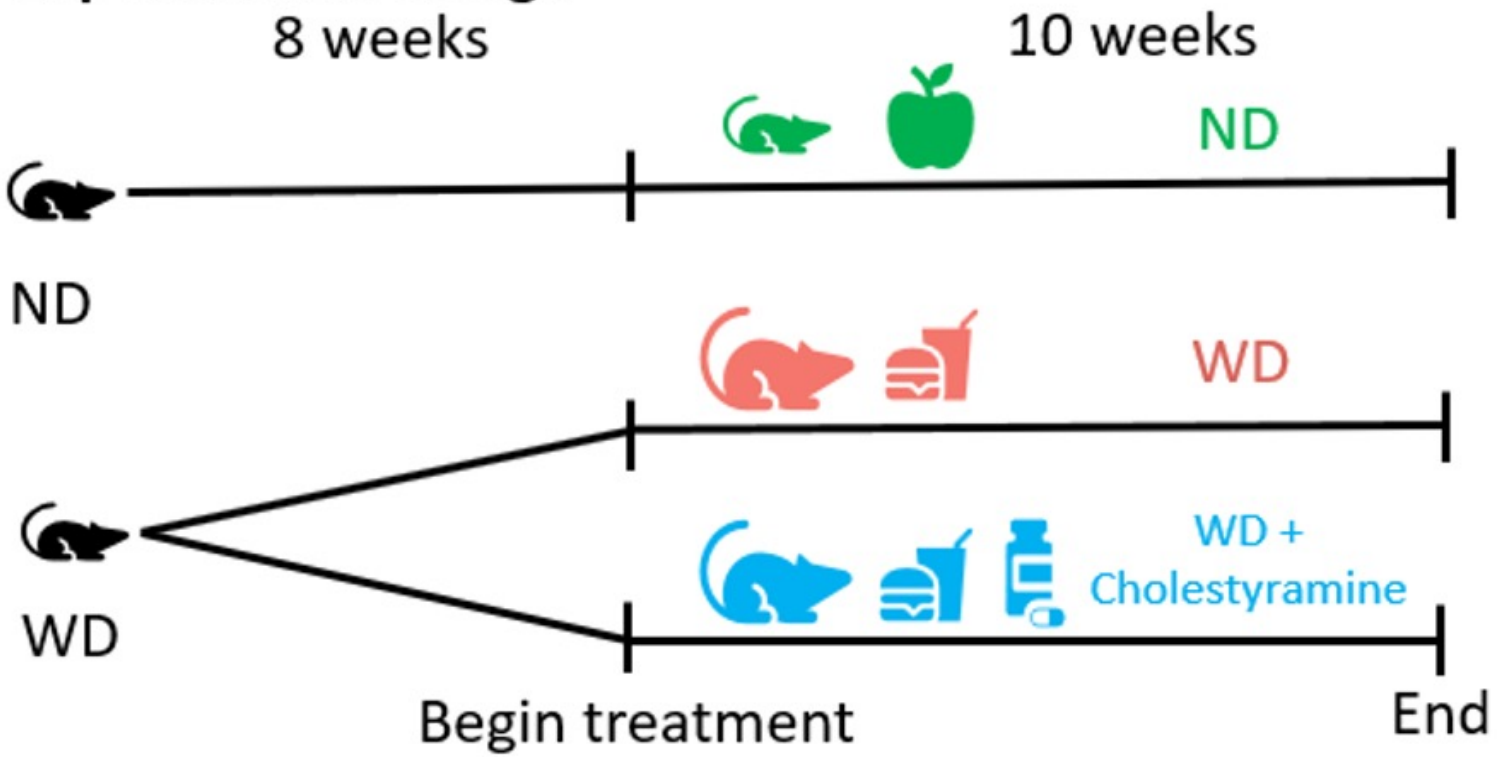
Nolan K. Newman[1], Philip M. Monnier[1], ... Natalia Shulzhenko[2] + Show Authors

Microbiome Res Rep 2024;3:40.

10.20517/mrr.2023.82 | © The Author(s) 2024.

Author Information Article Notes Cite This Article

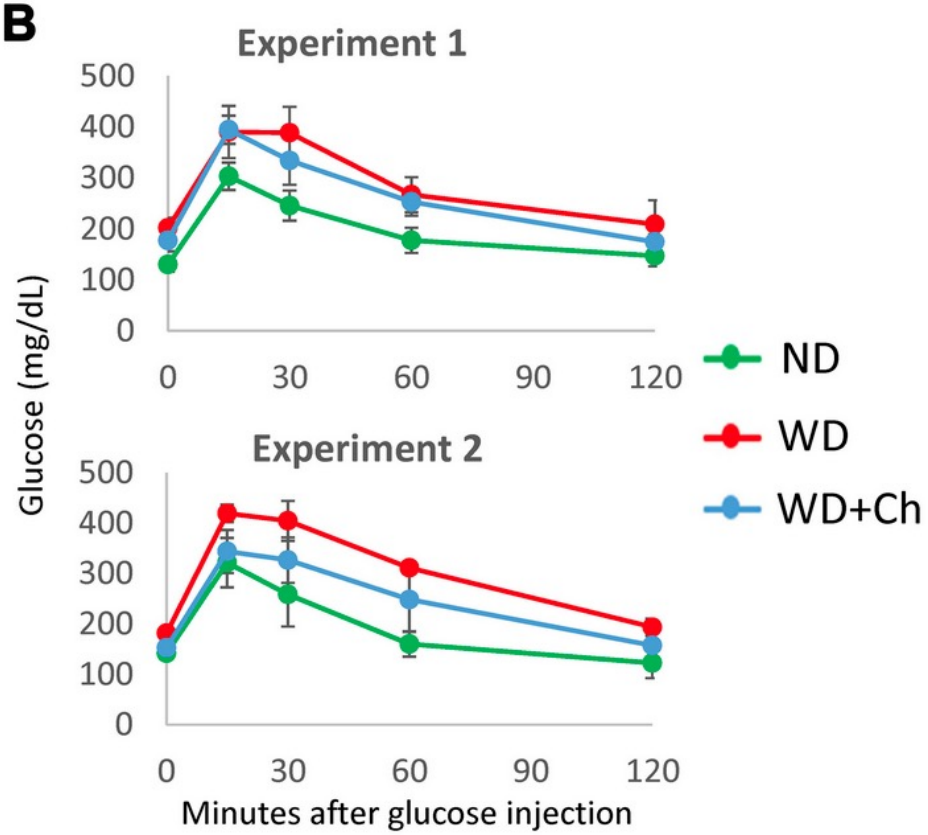
Experimental design

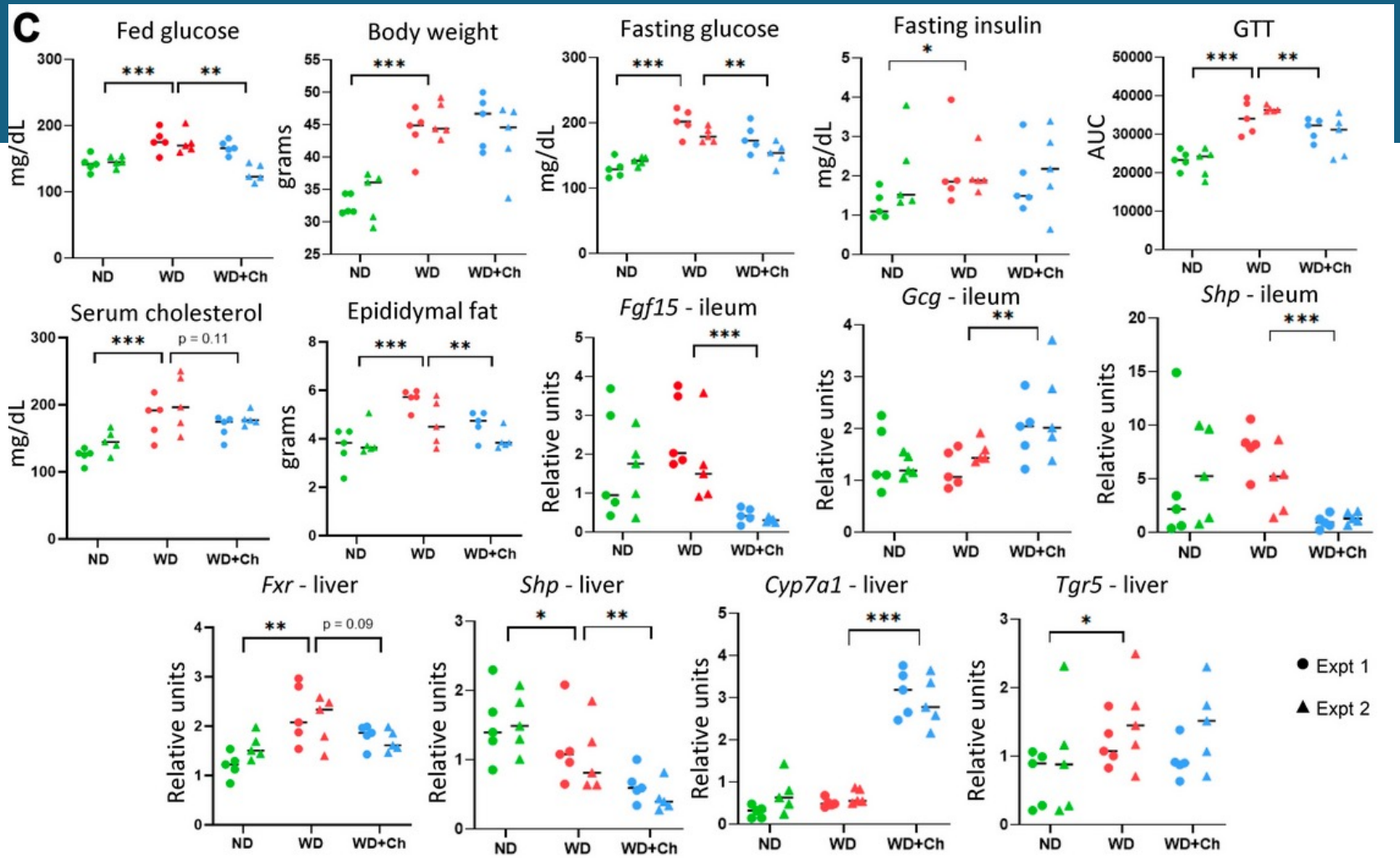


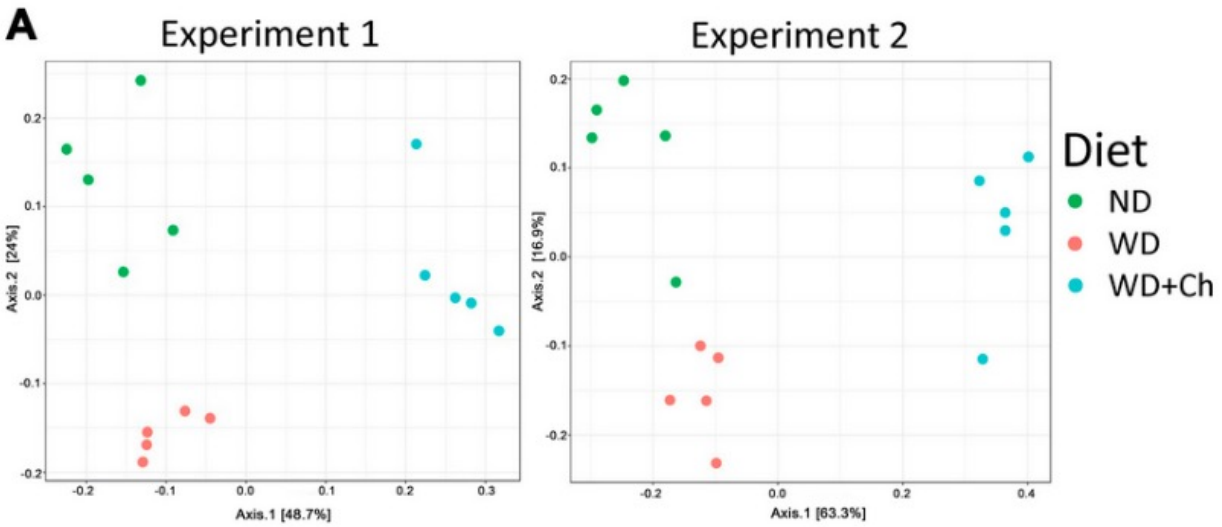
A

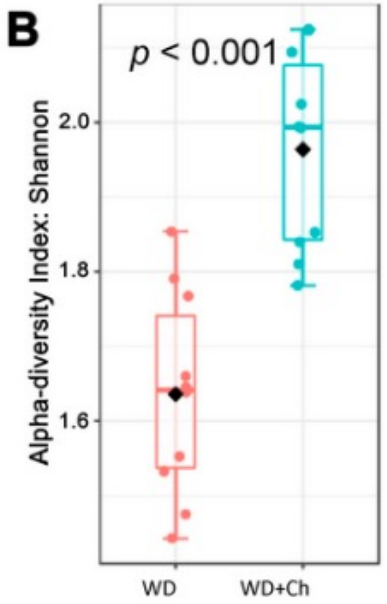
Host parameters	WD vs ND	WD+Ch vs WD
Fed glucose	UP	DOWN
Body weight	UP	
Fasting glucose	UP	DOWN
15 min glucose	UP	
30 min glucose	UP	DOWN
60 min glucose	UP	
120 min glucose	UP	DOWN
GTT-AUC	UP	DOWN
Fasting insulin	UP	
Serum cholesterol	UP	p = 0.11
Epididymal fat	UP	DOWN
Fgf15 - ileum		DOWN
Gcg - ileum		UP
Shp - ileum		DOWN
Fxr - liver	UP	DOWN†
Shp - liver	DOWN	DOWN
Cyp7a1 - liver		UP
Tgr5 - liver	UP	

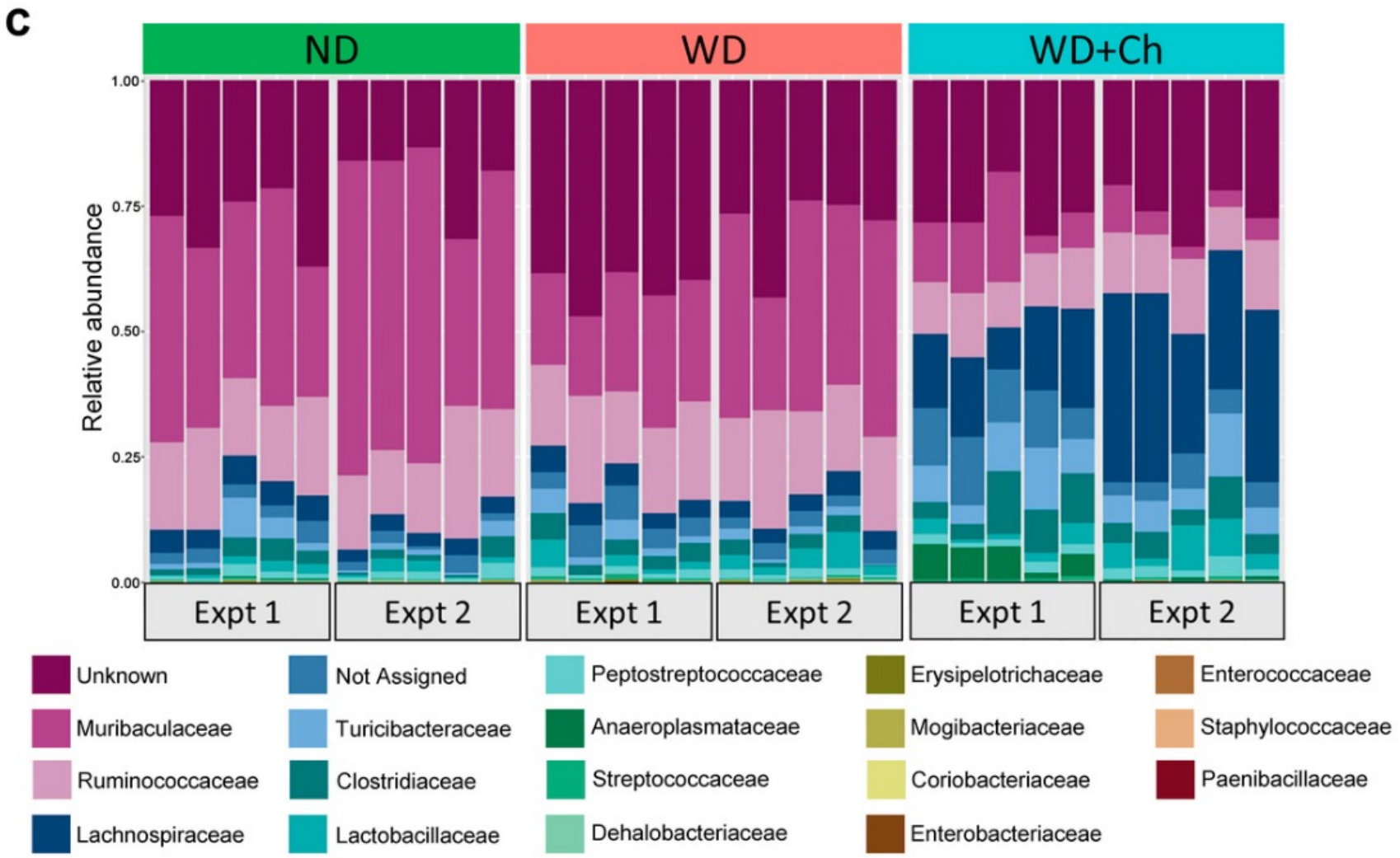
B

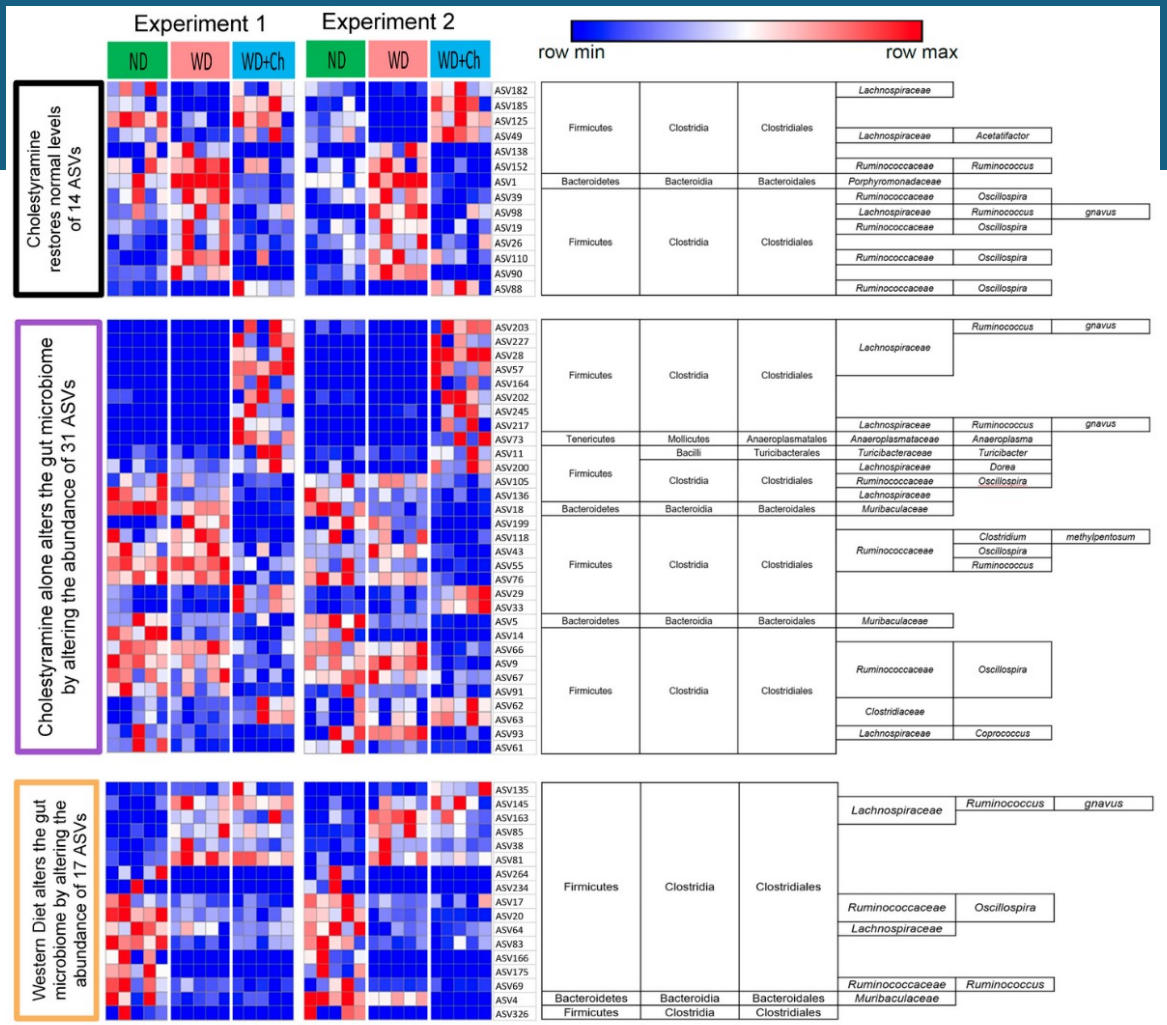



C
Fed glucose
mg/dL
Body weight
grams
Fasting glucose
mg/dL
Fasting insulin
mg/dL
GTT
AUC
Serum cholesterol
mg/dL
p = 0.11
Epididymal fat
grams
Fgf15 - ileum
Relative units
Gcg - ileum
Relative units
Shp - ileum
Relative units
Fxr - liver
p = 0.09
Relative units
Shp - liver
Relative units
Cyp7a1 - liver
Relative units
Tgr5 - liver
Relative units
ND
WD
WD+Ch
Expt 1
Expt 2


A
Experiment 1
Experiment 2
Axis.2 [24%]
Axis.1 [48.7%]
Axis.2 [16.9%]
Axis.1 [63.3%]
Diet
ND
WD
WD+Ch


B
p < 0.001
Alpha-diversity Index: Shannon
WD
WD+Ch


c
ND
WD
WD+Ch
Relative abundance
1.00
0.75
0.50
0.25
0.00
Expt 1
Expt 2
Expt 1
Expt 2
Expt 1
Expt 2
Unknown
Muribaculaceae
Ruminococcaceae
Lachnospiraceae
Not Assigned
Turicibacteraceae
Clostridiaceae
Lactobacillaceae
Peptostreptococcaceae
Anaeroplasmataceae
Streptococcaceae
Dehalobacteriaceae
Erysipelotrichaceae
Mogibacteriaceae
Coriobacteriaceae
Enterobacteriaceae
Enterococcaceae
Staphylococcaceae
Paenibacillaceae


Experiment 1
Experiment 2
ND
WD
WD+Ch
ND
WD
WD+Ch
row min
row max
Cholestyramine restores normal levels of 14 ASVs
Cholestyramine alone alters the gut microbiome by altering the abundance of 31 ASVs
Western Diet alters the gut microbiome by altering the abundance of 17 ASVs
ASV182
ASV185
ASV125
ASV49
ASV138
ASV152
ASV1
ASV39
ASV98
ASV19
ASV26
ASV110
ASV90
ASV88
ASV203
ASV227
ASV28
ASV57
ASV164
ASV202
ASV245
ASV217
ASV73
ASV11
ASV200
ASV105
ASV136
ASV18
ASV199
ASV118
ASV43
ASV55
ASV76
ASV29
ASV33
ASV5
ASV14
ASV66
ASV9
ASV67
ASV91
ASV62
ASV63
ASV93
ASV61
ASV135
ASV145
ASV163
ASV85
ASV38
ASV81
ASV264
ASV234
ASV17
ASV20
ASV64
ASV83
ASV166
ASV175
ASV69
ASV4
ASV326
Firmicutes
Clostridia
Clostridiales
Lachnospiraceae
Bacteroidetes
Bacteroidia
Bacteroidales
Porphyromonadaceae
Ruminococcaceae
Oscillospira
Ruminococcus
gnavus
Acetatifactor
Tenericutes
Mollicutes
Anaeroplasmatales
Anaeroplasmataceae
Anaeroplasma
Bacilli
Turicibacterales
Turicibacteraceae
Turicibacter
Dorea
Muribaculaceae
Clostridium
methylpentosum
Clostridiaceae
Coprococcus

FINALLY, WE GET TO NETWORKS, WHICH WE WILL DISCUSS AFTER LUNCH!