

The background of the slide is a dark blue field filled with numerous spherical virus particles. Each particle is covered in small, protruding spikes or glycoproteins, characteristic of coronaviruses. The particles are rendered in a semi-transparent, glowing blue color, creating a sense of depth and movement. A central white-bordered box contains the main title text.

# Microbiome Amplicon Sequencing Data Analysis

# The Research Technology Team



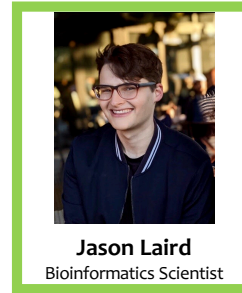
**Delilah Maloney**  
High Performance Computing Specialist



**Kyle Monahan**  
Senior Data Science Specialist



**Shawn Doughty**  
Manager, Research Computing



**Jason Laird**  
Bioinformatics Scientist



**Chris Barnett**  
Senior Geospatial Analyst



**Tom Phimmasen**  
Senior Data Consultant



**Patrick Florance**  
Director, Academic Data Services



**Jake Perl**  
Digital Humanities NLP Specialist



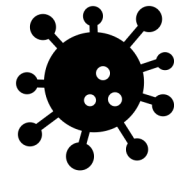
**Carolyn Talmadge**  
Senior GIS Specialist



**Uku-Kaspar Uustalu**  
Data Science Specialist

- ✓ Consultation on Projects and Grants
- ✓ High Performance Compute Cluster
- ✓ Workshops

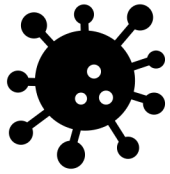
<https://it.tufts.edu/research-technology>







# Introduction to the Microbiome

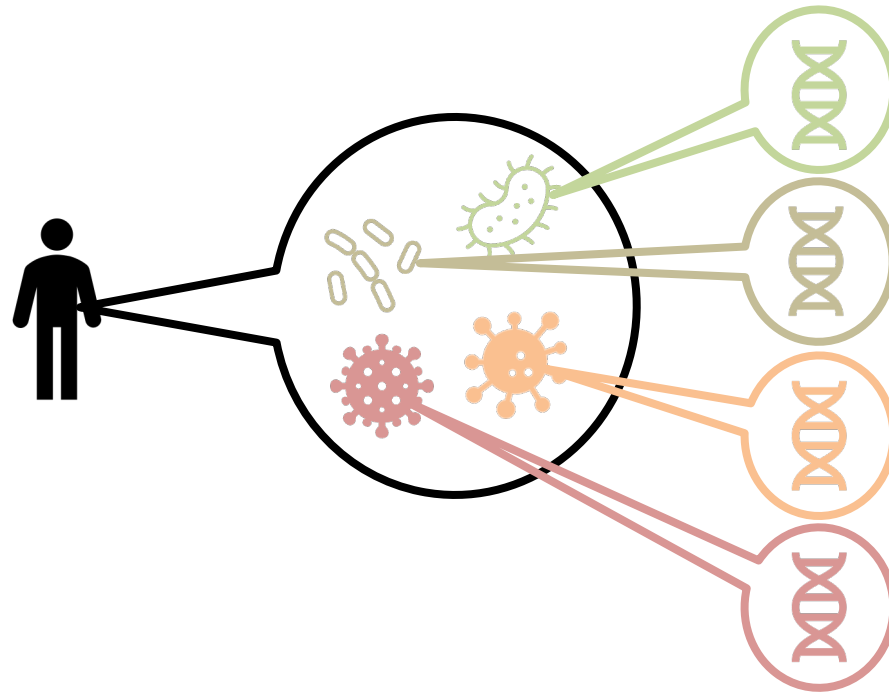


# What is the Microbiome?

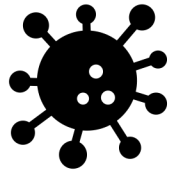
Sample

Microbiota

Microbiome



The microbiome are the set of genes belonging to the microbiota in a specimen. The term microbiome can also refer to the microbes themselves



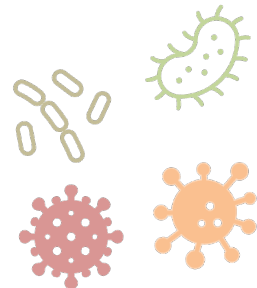
# Introduction to the Microbiome

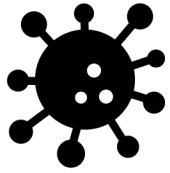
- Previously it was thought that the number of cells in the microbiome outnumbered human cells by 10:1. We now know that it is closer to 1:1.
- Disturbances in the microbiome are linked to obesity, inflammatory bowel disease, alcoholic and nonalcoholic fatty liver disease, and hepatocellular carcinoma

Number of  
Cells in a  
Human



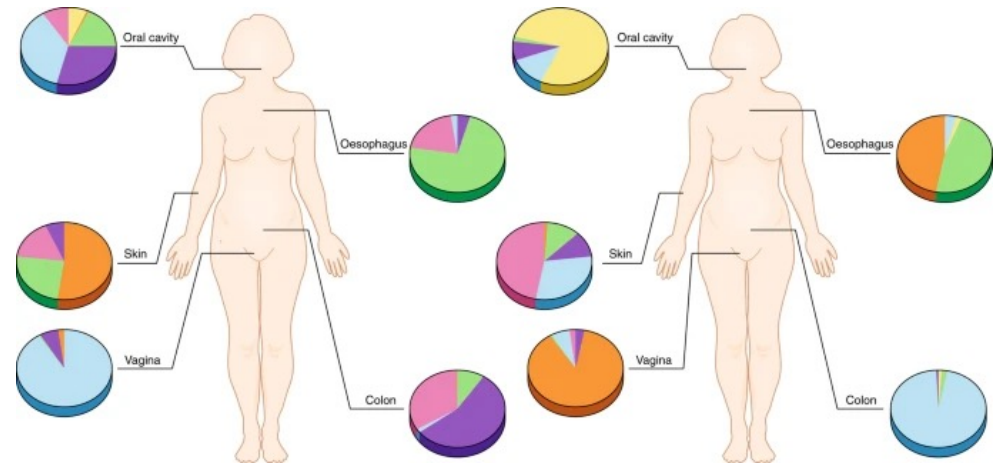
Number of  
Cells in the  
Microbiome

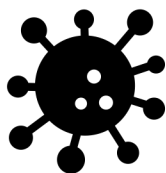




# Microbiome Variability

- Assessing a microbiome disturbance is not a trivial task as it is highly variable from person to person.
- Large sample sizes, hundreds of patients, are needed to overcome interindividual variability.

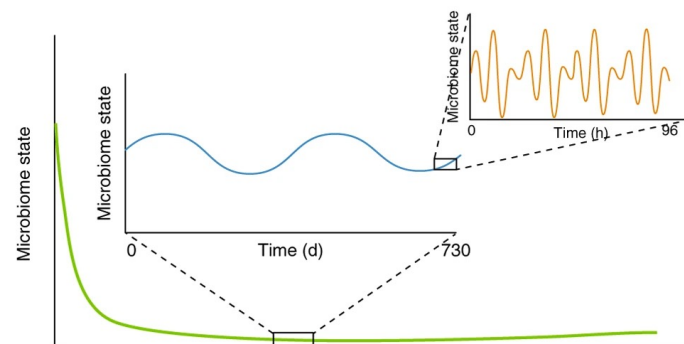




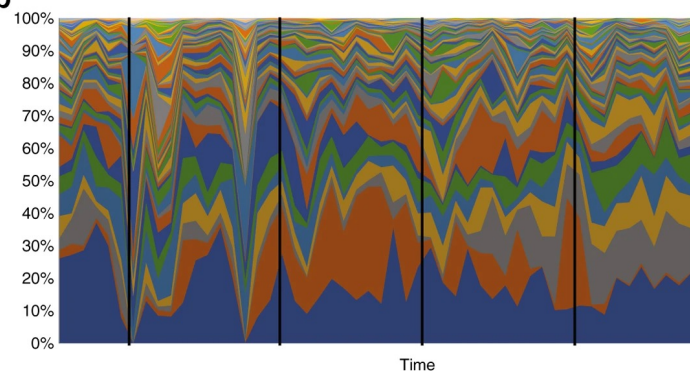
# Sample Collection

- Sample collection is also a difficult challenge and highly dependent on the study question.
- The microbiome can change in an individual over time, especially in diseases marked by flare ups like IBD.
- Samples might not be representative of the site in question. For example, a stool sample sits in the rectum – an environment that is undergoing dehydration and fermentation which might select for different bacteria than in the small intestine.

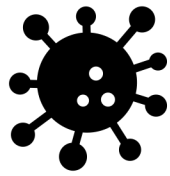
a



b

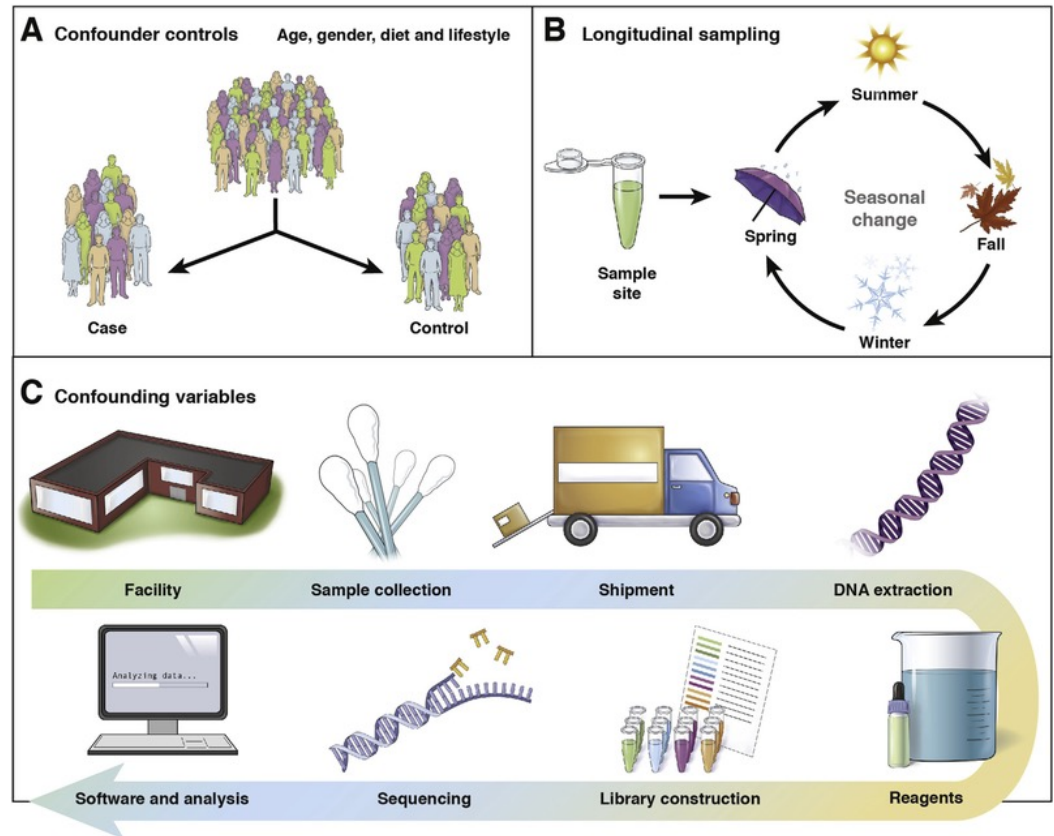




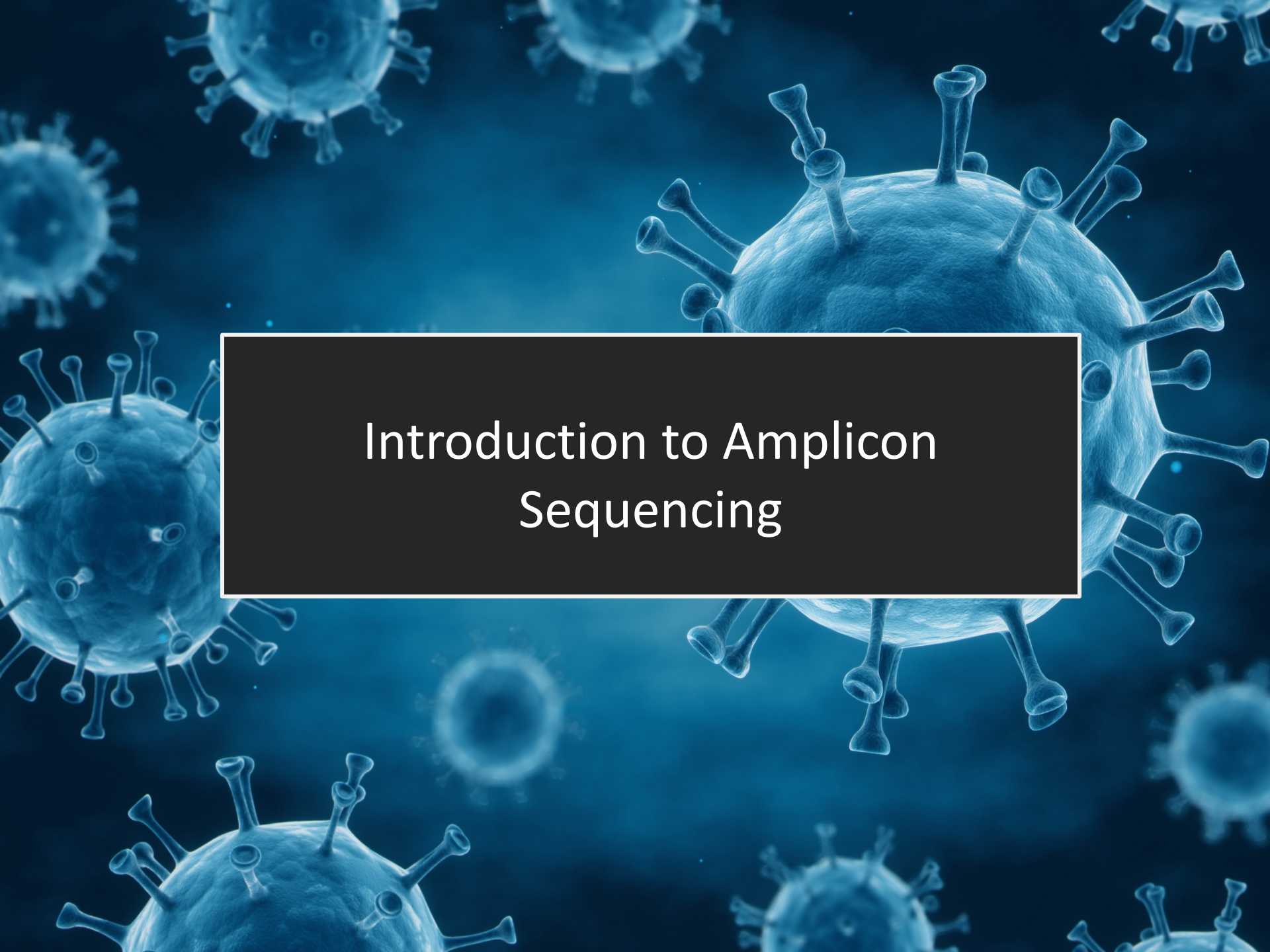


# Accounting For Confounding Variables

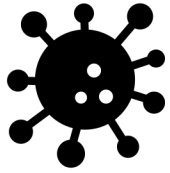
- When conducting a clinical experiment, it is pertinent to stratify accounting for age, gender, diet, etc.
- Sampling over time is incredibly valuable as you can better capture inpatient variability.
- Additionally, the way the sample is processed can also confound your results



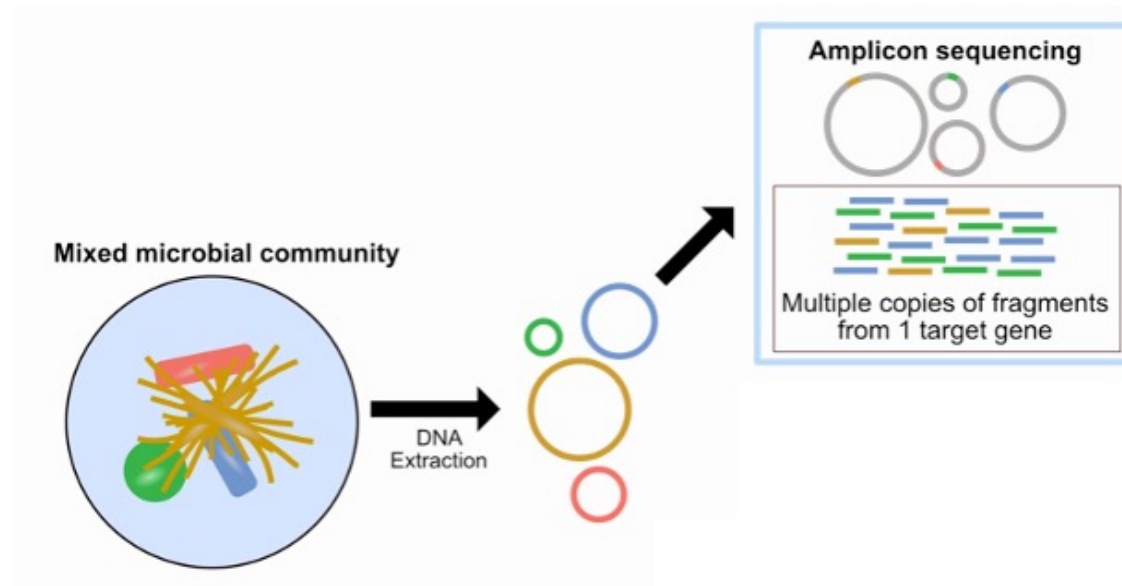


The background of the slide is a dark blue field filled with numerous spherical virus particles. Each particle is covered in small, protruding spikes or glycoproteins, characteristic of coronaviruses. The particles are rendered in a semi-transparent, glowing blue color, creating a sense of depth and movement. A central white-bordered box contains the title text.

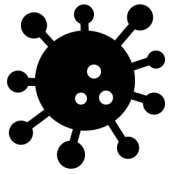
# Introduction to Amplicon Sequencing



# What is an Amplicon?



Microbiome Amplicon sequencing involves sequencing a specific gene from microbial community

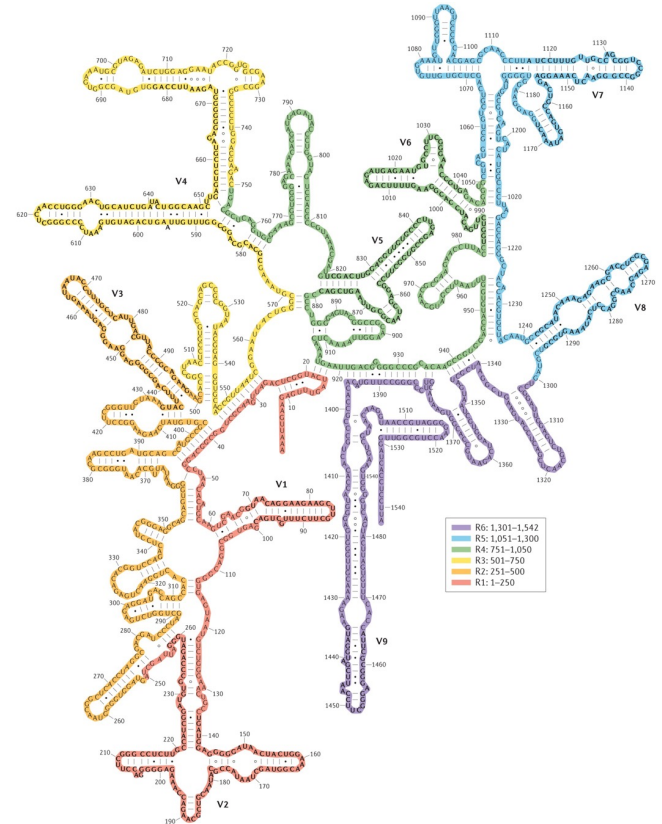


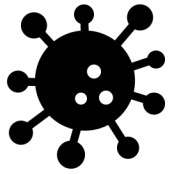
# Why Sequence One Gene?

Genes can vary per organism and may not be well conserved across species. To assess the microbial community composition, we need to sequence a conserved gene across organisms of interest:

- **16S ribosome DNA (rDNA)** for prokaryotes
- **18S rDNA and internal transcribed spacers (ITS)** for eukaryotes

16S rRNA

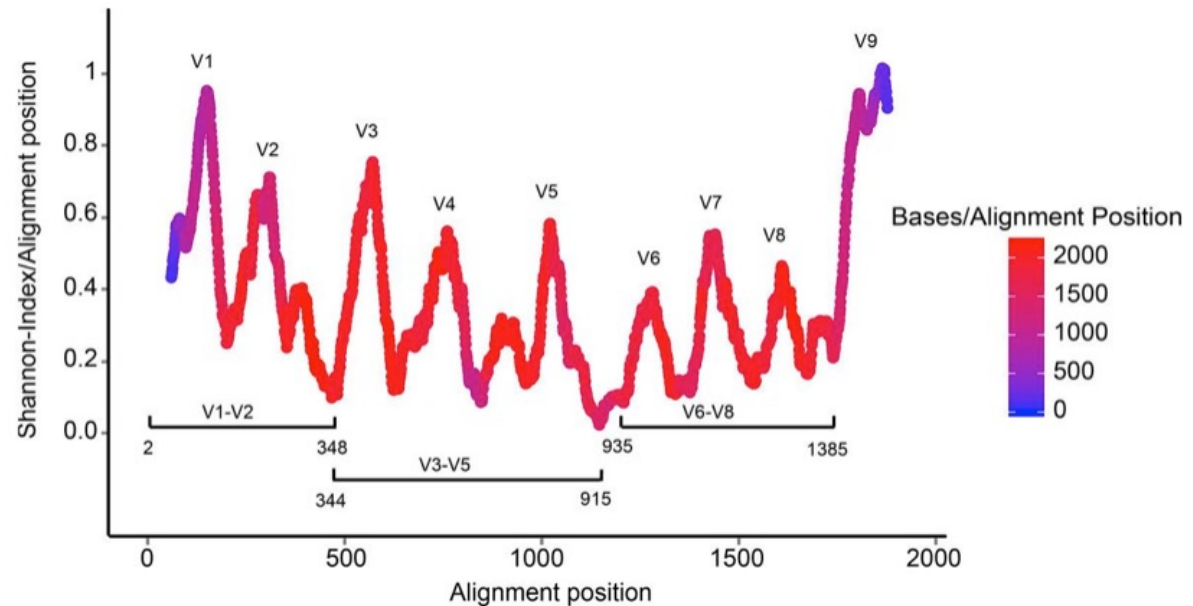




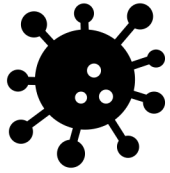
# Primers for Marker Gene

In the selected gene there are different levels of conservation across organisms. To circumvent this parts of the gene with high conservation (like the V4 region of 16S rRNA) are selected for

16S rRNA gene conservation







# 16S Analysis Pipelines

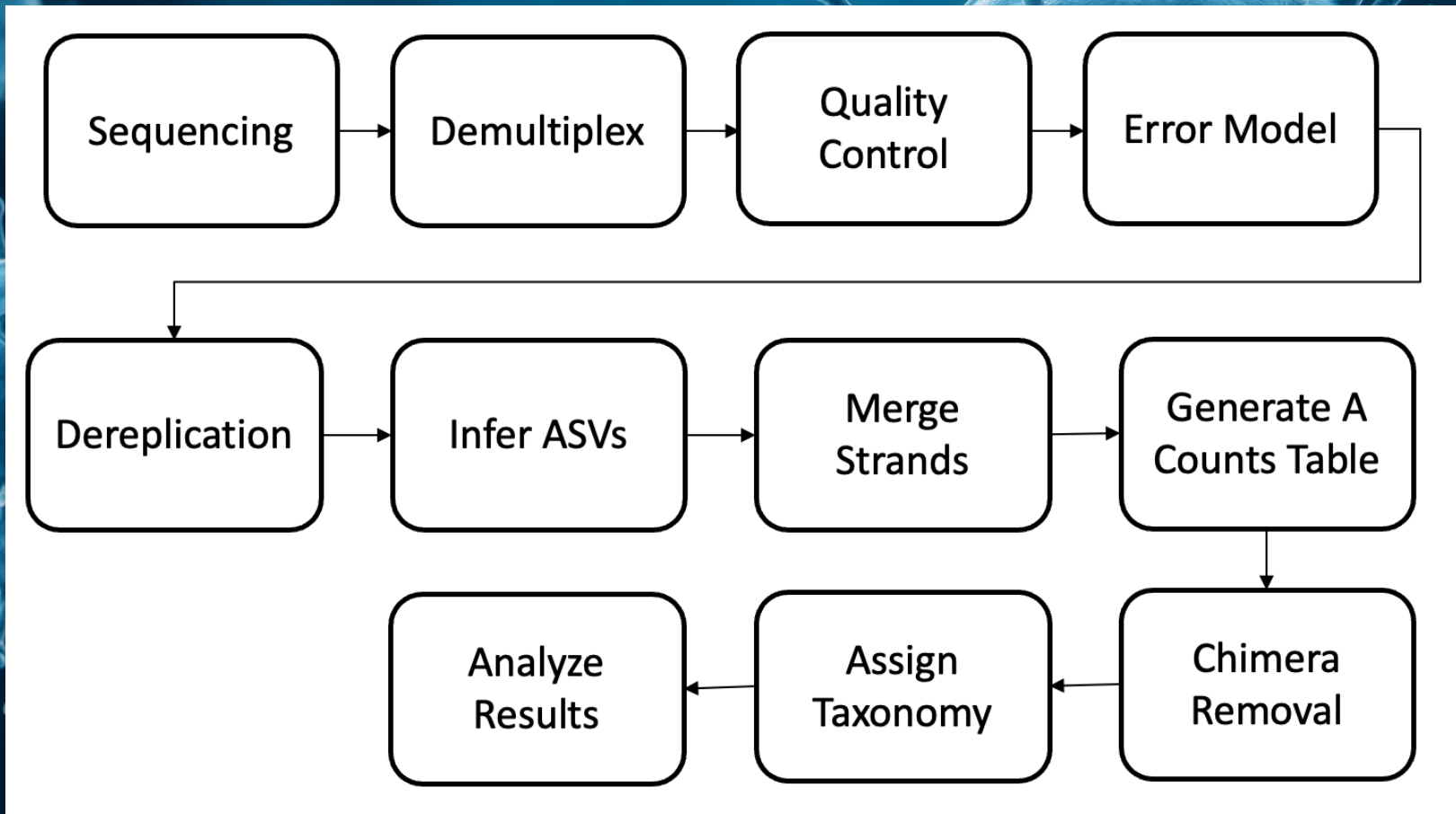


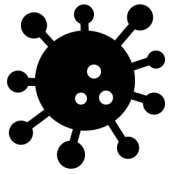
- There are many different pipelines/tools to apply to 16S data:
  - [DADA2](#)
  - [USEARCH](#)
  - [Mothur](#)
  - [QIIME](#)
- Today we will be using DADA2!





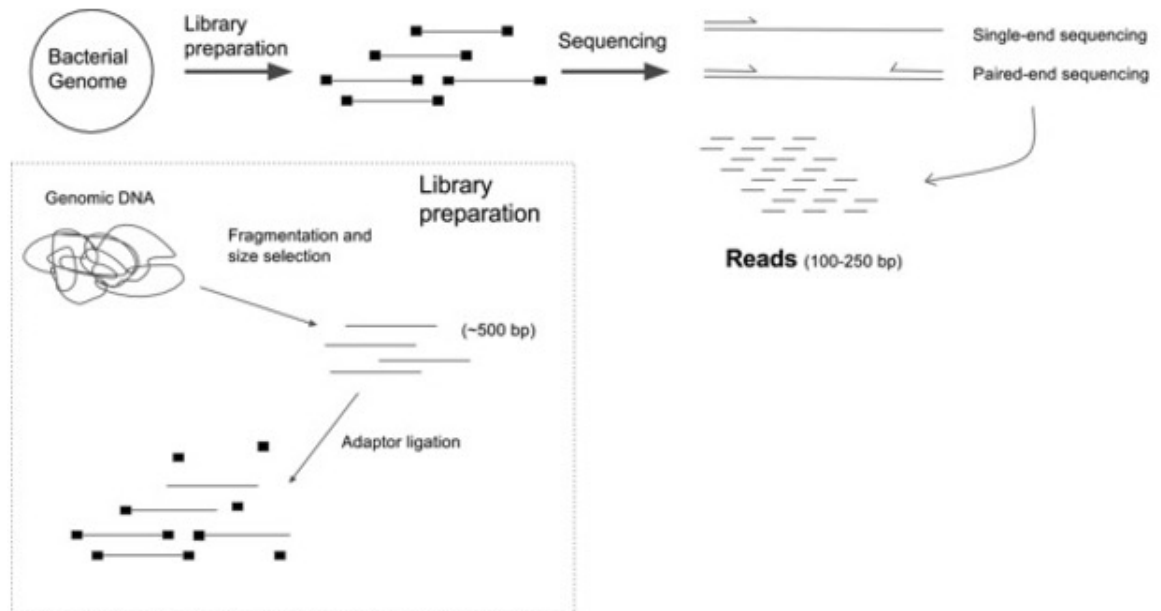
# Amplicon Workflow (DADA2)

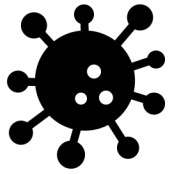




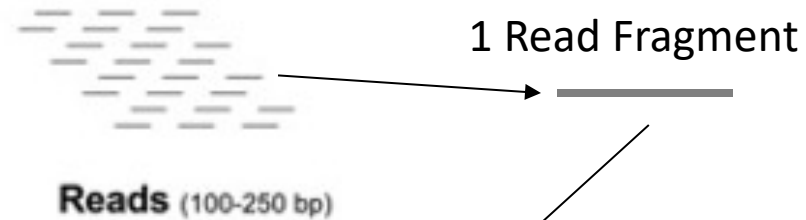
# Sequencing Overview

- Marker gene (16S, 18S, or ITS) is selected
- Primers target areas of high conservation in gene
- DNA is fragmented
- Adapters are added to help the DNA attach to a flow cell
- Barcodes may also be added to identify which DNA came from which sample
- The fragments are sequenced to produce reads
- Reads can be single-end (one strand sequenced) or paired-end (both strands sequenced)

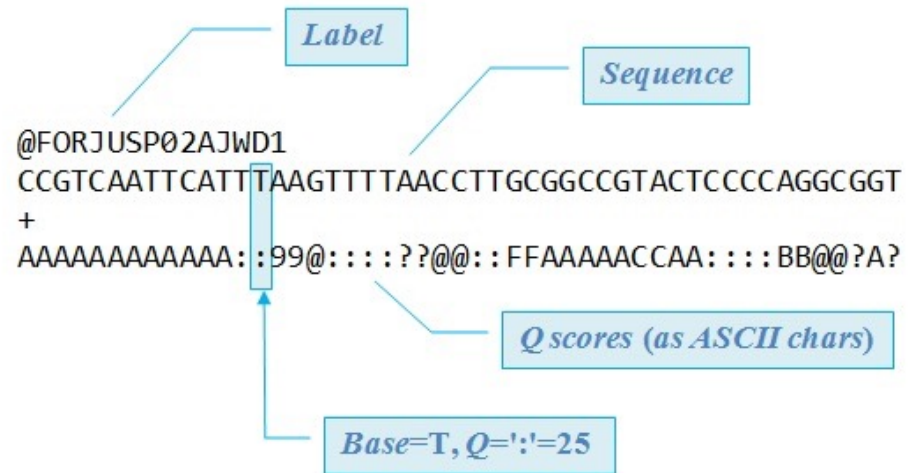




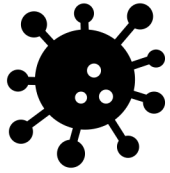
# Read Data is Stored in FASTQ Files



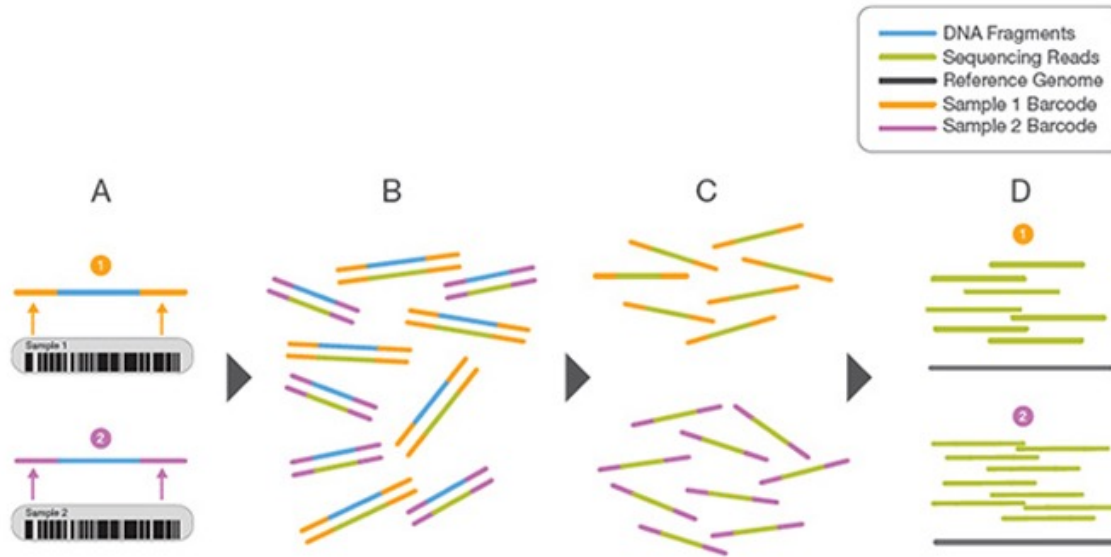
- After sequencing we end up with a FASTQ file which contains:
  - A sequence label
  - The nucleic acid sequence
  - A separator
  - The quality score for each base pair





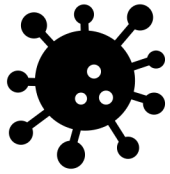


# Demultiplexing



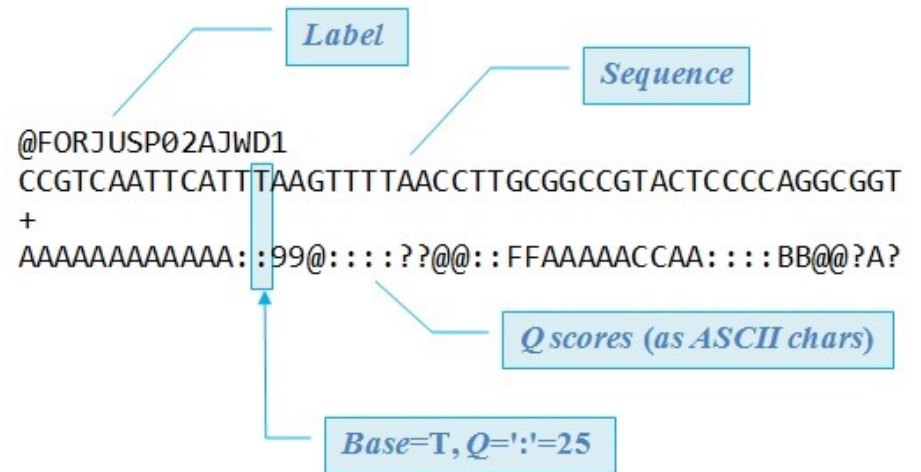
- Sometimes samples are mixed to save on sequencing cost
- To identify which DNA is from which sample Barcodes are added
- Before moving forward samples need to be separated and those DNA barcodes need to be removed
- Tools like [sabre](#) can demultiplex pooled FASTQ data

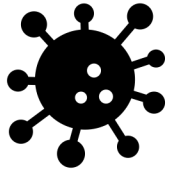




# Quality Scores

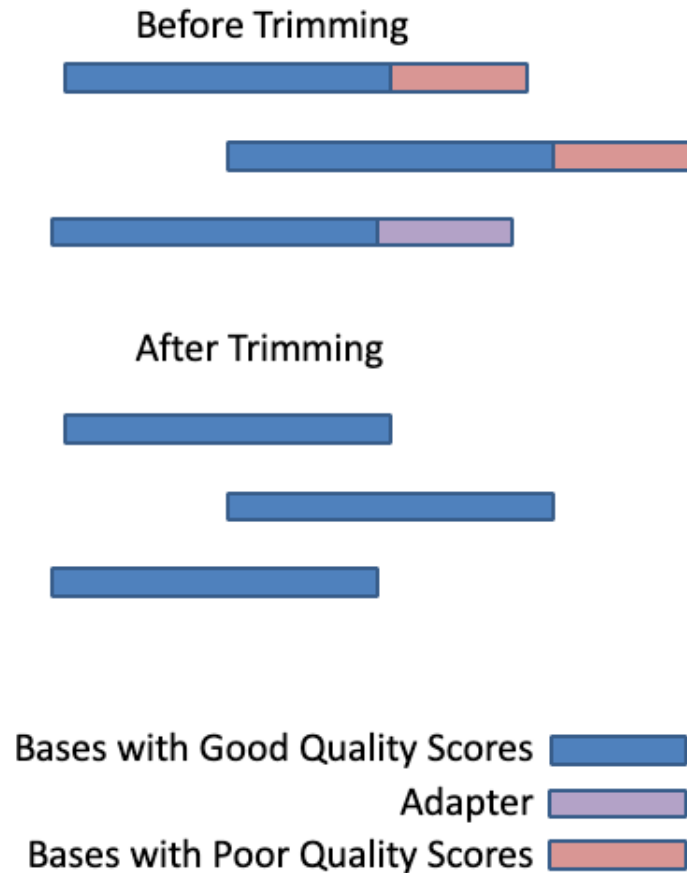
- Quality Scores are the probability that a base was called in error
- Higher scores indicate that the base is less likely to be incorrect
- Lower scores indicate that the base is more likely to be incorrect

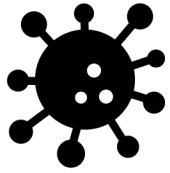




# Trimming

- To reduce noise in our data low quality bases and any adapters present are removed by trimming the sequence
- Tools like [Trimmomatic](#) and [Trim-Galore](#) can trim poor sequences and adapters





# DADA2 Error Model

- Here we ask: What is the error rate for an amplicon sequence read  $i$  that was produced from a sequence  $j$  over  $L$  aligned nucleotides with a quality score  $q$ ?
- Basically, this is a product of error probabilities given some quality score e.g.  $p(A > G, 35)$

$$\lambda_{ji} = \prod_{l=0}^L p(j(l) \rightarrow i(l), q_i(l))$$

$\lambda_{ji}$  Error Rate

$L$  Number of aligned nucleotides

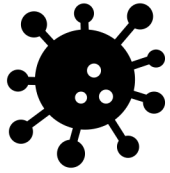
$j(l)$  Correct sequence at nucleotide  $l$

$i(l)$  Amplicon read at nucleotide  $l$

$q_i(l)$  Quality score for nucleotide  $l$

Error Model





# DADA2 p-value

$$p_A(j \rightarrow i) = \frac{1}{1 - \rho_{pois}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{pois}(n_j \lambda_{ji}, a)$$

- The error rate is fed into another function to collect the p-value
- This p-value assess if sequence **i** is too abundant for it to be explained by errors in amplicon sequencing
- **low p-value** = sequence **i** is too abundant to be some sequencing error

$p_A(j \rightarrow i)$  P-value for nucleotide in sequence **j** to sequence **i**

$\lambda_{ji}$  Error Rate

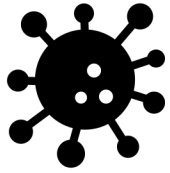
$\rho_{pois}$  Poisson Density Function

$n_j$  Number of **j** sequences

$a$  Abundance of sequence **i**



Error Model



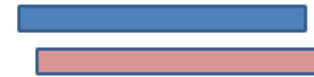
# Dereplication



- Microbiome samples will often contain large numbers of the same organism and as such we will find the same sequence repeated in our data
- To speed up computation only unique sequences are kept

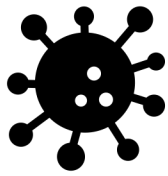
Before Dereplication



After Dereplication

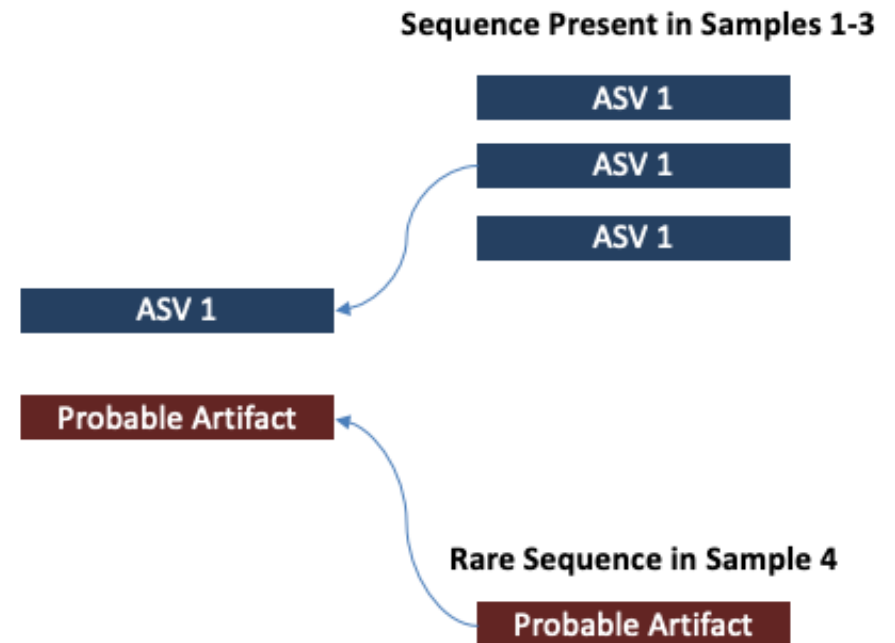


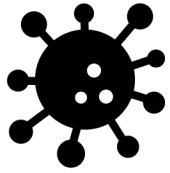
Sequence 1   
Sequence 2 



# Inferring Amplicon Sequence Variants (ASVs)

- So far, we have assigned p-values for each sequence in each sample
- DADA2 then tries to determine which sequences are of biological origin (ASVs) and which aren't by assessing which sequences are present in other samples
- If a sequence is present in another sample, it is more likely that it is a real biological sequence





# ASVs vs. OTUs

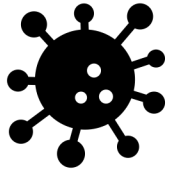
- Traditional 16S metagenomic approaches use OTUs or operational taxonomic units instead of ASVs
- So why does DADA2 use ASVs?

ASV



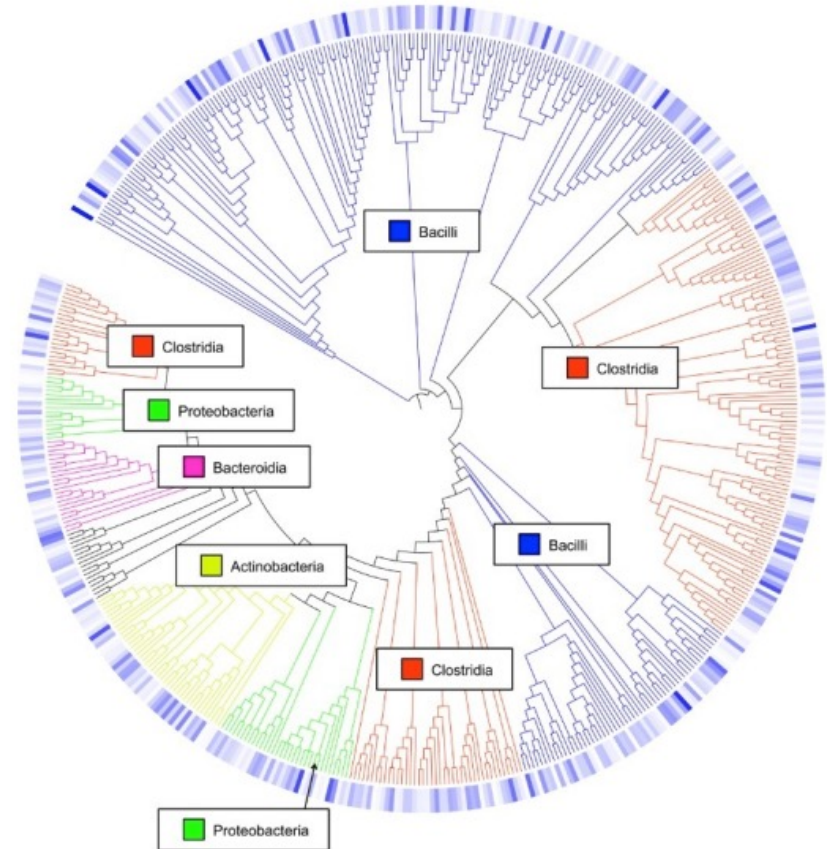
OTU



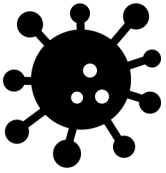


# What is an OTU?

- Methods that use OTUs, cluster sequences are clustered together by similarity
- Those sequences with above a 97% identity threshold are clustered into an OTU
- These OTUs are then combined into a consensus sequence and mapped to a reference database to determine which species it is from



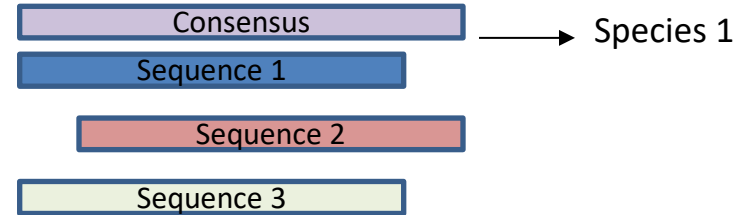




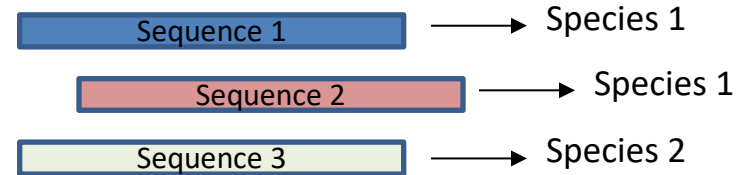
# ASV vs. OTU Debate

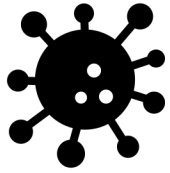
- Originally, OTUs were used to mitigate possible sequence errors by clustering similar sequences and getting a consensus sequence. However, this method has been found to inflate the number of unique sequences
- By contrast, ASV analysis derives an error term to assess the possibility of a sequencing error. These sequences are then mapped directly to the organism of interest - giving nucleotide resolution

## OTU



## ASV



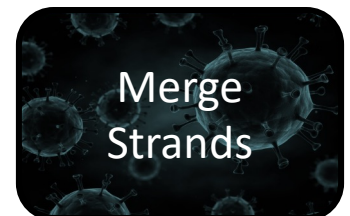


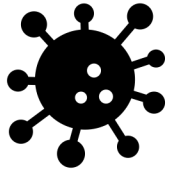
# Merging Strands

- For paired-end data there is a good deal of overlap between the forward and reverse read
- To resolve this redundancy, these reads are collapsed into contigs

C	A	T	T	G	A	C	A		
32	34	20	20	28	16	14	10	Forward read	
Reverse read		T	A	G	A	C	A	T	T
		2	5	4	8	12	20	38	40
								Base calls	
								Q scores	
C	A	T	T	G	A	C	A	T	T
32	34	22	16	35	28	30	34	38	40
								Consensus	
								Posterior Qs	

Mismatch ↑ Merged read

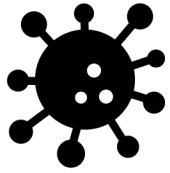




# ASV Counts Table

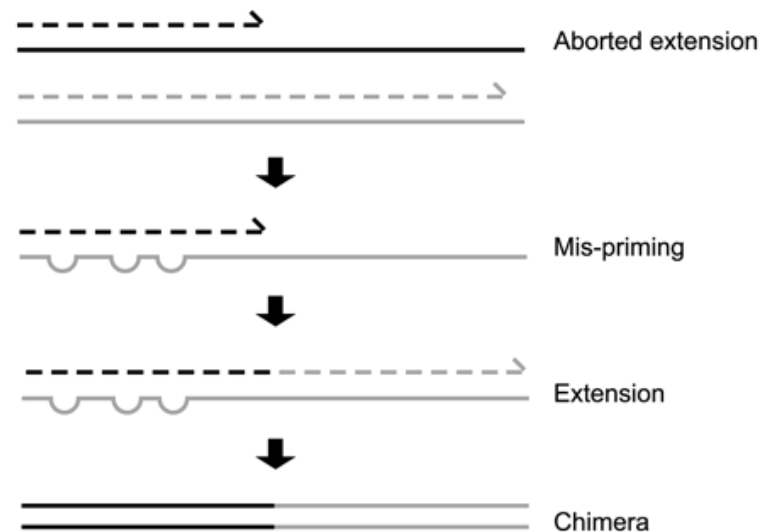
- Now that we have combined the forward and reverse strands into one ASV we can generate a counts table

	ASV 1	ASV 2	ASV 3
Sample 1	0	19	18
Sample 2	500	27	34
Sample 3	45	65	86



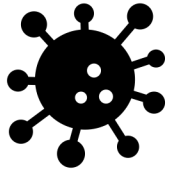
# Chimera Removal

- During Sequencing microbial DNA is subjected to PCR to amplify DNA
- During PCR it is possible for two unrelated templates to form a non-biological hybrid sequence
- DADA2 finds these chimeras by aligning each sequence to more abundant sequences and seeing if there are any low abundant sequences that can be created by mixing the left and right sides of the more abundant sequences



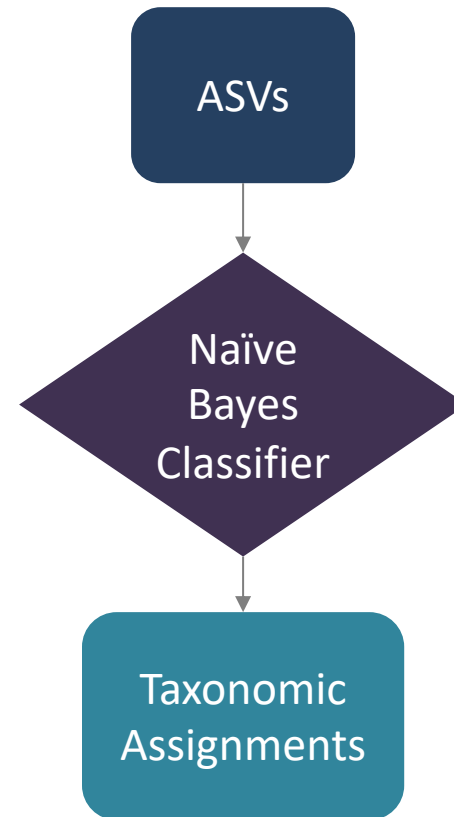
<https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html>  
<https://genome.cshlp.org/content/21/3/494/F1.expansion.html>  
[https://astrobiomike.github.io/amplicon/dada2\\_workflow\\_ex#merging-forward-and-reverse-reads](https://astrobiomike.github.io/amplicon/dada2_workflow_ex#merging-forward-and-reverse-reads)



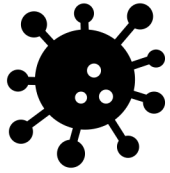


# Assigning Taxonomy

- To determine which taxon each ASV belongs to DADA2 uses a naïve bayes classifier
- This classifier uses a set of reference sequences with known taxonomy as the training set and outputs taxonomic assignments with bootstrapped confidence







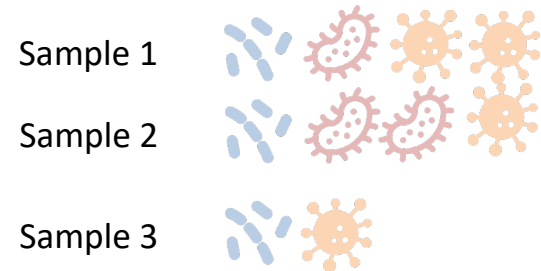
# Diversity Analysis

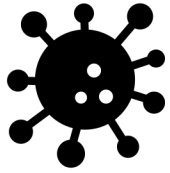
- Once you have taxonomical information, we can assess diversity. Typically, alpha or beta diversity
- **Alpha Diversity** - ecological complexity of a single sample
- **Beta Diversity** - ecological complexity between samples

## Alpha Diversity



## Beta Diversity





# Types of Diversity Analysis

- Diversity is not a standard term and there are different types of diversity to examine:
- **Species richness** = the number of different species in a community.
- **Species evenness** = how even in numbers each species in a community is.
- **Phylogenetic diversity** = how closely related the species in a community are.

## Species Richness



100x



3000x

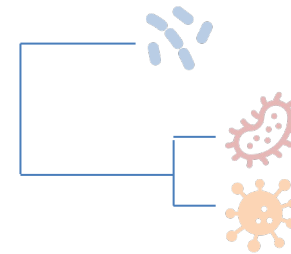


5x

## Species Evenness

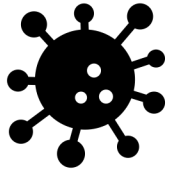


## Phylogenetic Diversity



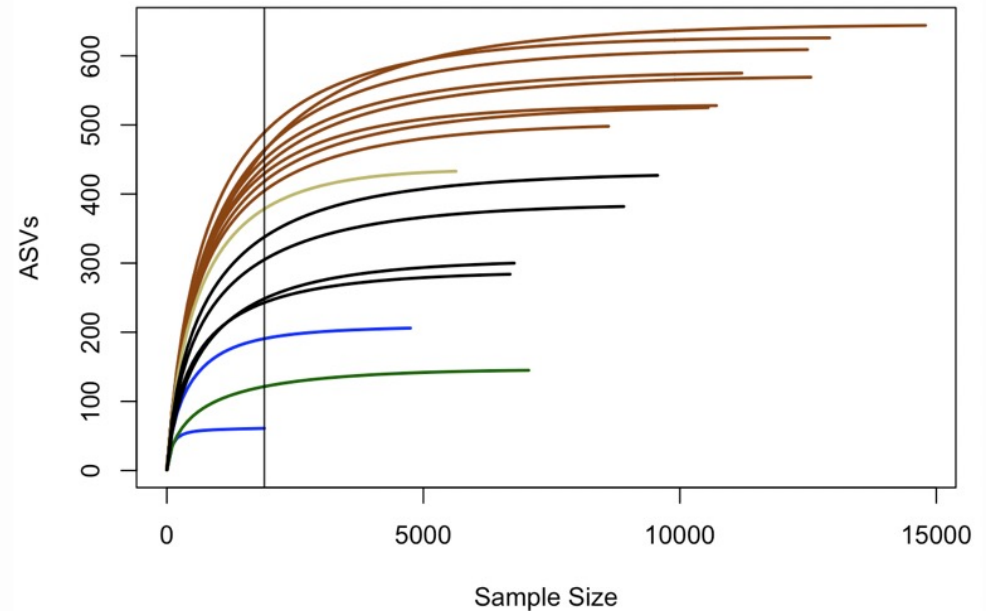
<https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html>

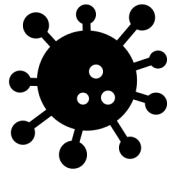
Analyze  
Results



# Alpha Diversity - Rarefaction Curves

- Rarefaction curves plot the number of ASVs (or OTUs if working with other methods) against the row sum of ASV counts for a particular sample
- The rarefaction curve to the right tells that samples in the brown group have more species present
- It is worth noting that this metric can be swayed by the presence of novel organism – so one sample might appear to have a lower number of species, but it could just in fact have more new species





# Alpha Diversity – Shannon Diversity Index

- Diversity Indices assess how diverse a community is

- **Shannon Diversity Index:** higher values = higher diversity

- **Simpson Diversity Index:** higher values = higher diversity

## Shannon Diversity Index

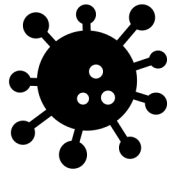
$$H = \sum_{i=1}^S - (P_i \times \ln P_i)$$

- H = Shannon Entropy,
- $P_i$  = fraction of population composed of a single species  $i$ ,
- $\ln$  = natural log,
- S = how many species encountered,
- $\Sigma$  = summation of species 1 to S

## Simpson Diversity Index

$$D = 1 - \frac{\sum n(n-1)}{N(N-1)}$$

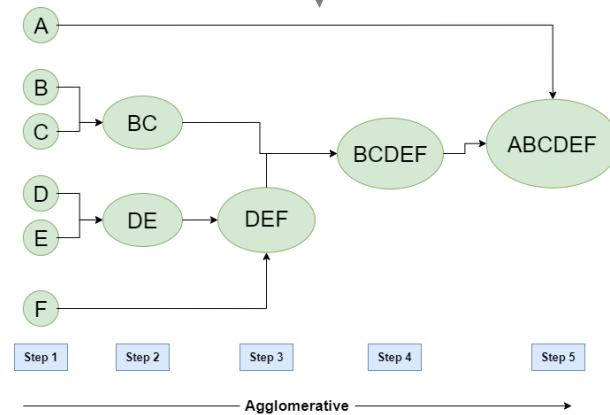
- n = number of individuals of each species
- N = total number of individuals of all species



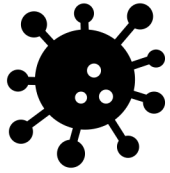
# Beta Diversity – Hierarchical Clustering

- We can use our counts matrix to determine how far apart each sample is from one another
- In Hierarchical clustering each sample starts off as its own cluster then grouped with the sample closest
- This is iterated until all the samples have been grouped into one cluster

	ASV 1	ASV 2	ASV 3
Sample A	0	19	18
Sample B	500	27	34
Sample C	45	65	86







# Beta Diversity – Ordination

- Aside from clustering we can visualize how our samples group together by ordination – a dimension reduction technique to help visualize sample to sample distance
- A commonly used metric is the **Bray-Curtis metric ( $BC_d$ )**

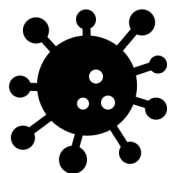
$$BC_d = \frac{\sum |x_i - x_j|}{\sum (x_i + x_j)}$$

- $S_i$  = Sample I
- $S_j$  = Sample J

Species	$S_i$	$S_j$	$ x_i - x_j $	$(x_i + x_j)$
Spp1	6	4	2	10
Spp2	5	3	2	8
Spp3	7	4	3	11
Spp4	2	6	4	8
Spp5	3	0	3	3
<b>Sum</b>			<b>14</b>	<b>40</b>

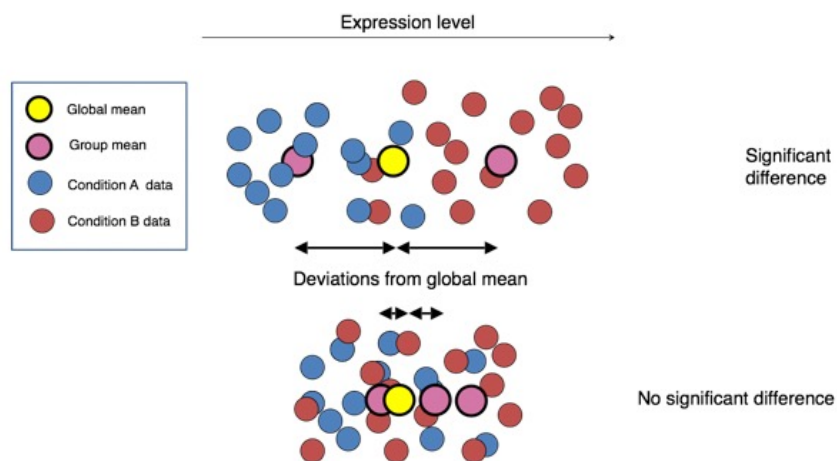
$$BC_d = 14/40$$
$$= 0.35$$

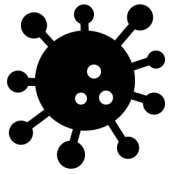




# Differential Abundance – DESeq2

- When assessing a microbial community, you might be interested to determine which species are differentially abundant between conditions
- Given that we have a counts matrix we can use DESeq2!





# Differential Abundance – DESeq2

## DESeq2 Normalization:

1. Geometric mean per ASV
2. Divide rows by geometric mean
3. Take the median of each sample
4. Divide all ASV counts by that median

## Normalization: DESeq2 Median of Ratios

1. Take a row-wise average to produce an average sample (geometric mean)  $\sqrt[n]{x_1 x_2 \dots x_n}$

	Sample A	Sample B	Avg. Sample
ASV 1	26	10	16
ASV 2	26	10	16
ASV 3	26	10	16
ASV 4	2	50	16

2. Divide all rows by the Average Sample for that gene (**Ratio**)

	Sample A/Avg.	Sample B /Avg.
ASV 1	26/16 = 1.6	10/16 = 0.6
ASV 2	1.6	0.6
ASV 3	1.6	0.6
ASV 4	0.2	5

3. Take the **median** of each column. Should be ~1 for all

Size factor	1.6	0.6
-------------	-----	-----

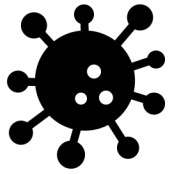
4. Divide all counts by sample specific size factor

	Sample A / S <sub>A</sub>	Sample B / S <sub>B</sub>
ASV 1	16.3	16.7
ASV 2	16.3	16.7
ASV 3	16.3	16.7
ASV 4	1.3	83.3

Normalized counts for non-DE ASVs are similar!

```
estimateSizeFactors(dds)
```





# Differential Abundance – DESeq2

## DESeq2 Normalization:

1. Geometric mean per ASV
2. Divide rows by geometric mean
3. Take the median of each sample
4. Divide all ASV counts by that median

## Normalization: DESeq2 Median of Ratios

1. Take a row-wise average to produce an average sample (geometric mean)  $\sqrt[n]{x_1 x_2 \dots x_n}$

	Sample A	Sample B	Avg. Sample
ASV 1	26	10	16
ASV 2	26	10	16
ASV 3	26	10	16
ASV 4	2	50	16

2. Divide all rows by the Average Sample for that gene (**Ratio**)

	Sample A/Avg.	Sample B/Avg.
ASV 1	26/16 = 1.6	10/16 = 0.6
ASV 2	1.6	0.6
ASV 3	1.6	0.6
ASV 4	0.2	5

3. Take the **median** of each column. Should be ~1 for all

Size factor	1.6	0.6
-------------	-----	-----

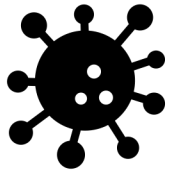
4. Divide all counts by sample specific size factor

	Sample A / S <sub>A</sub>	Sample B / S <sub>B</sub>
ASV 1	16.3	16.7
ASV 2	16.3	16.7
ASV 3	16.3	16.7
ASV 4	1.3	83.3

Normalized counts for non-DE ASVs are similar!

```
estimateSizeFactors(dds)
```

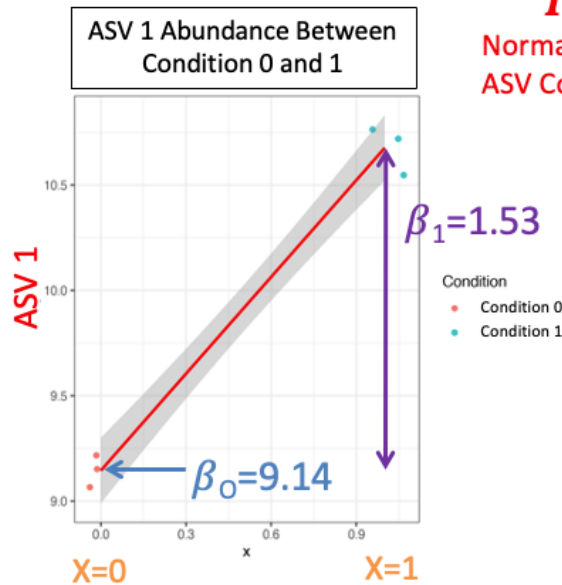




# Differential Abundance – DESeq2

## DESeq2 Model:

1. The normalized abundances of an ASV are plotted against two conditions
2. The regression line that connects these data is used to determine the p-value for differential abundance



$$Y = \beta_0 + \beta_1 X + e$$

Normalized ASV Counts

Intercept

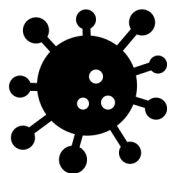
Condition (0 and 1)

Slope: difference between Condition 0 and 1

Error



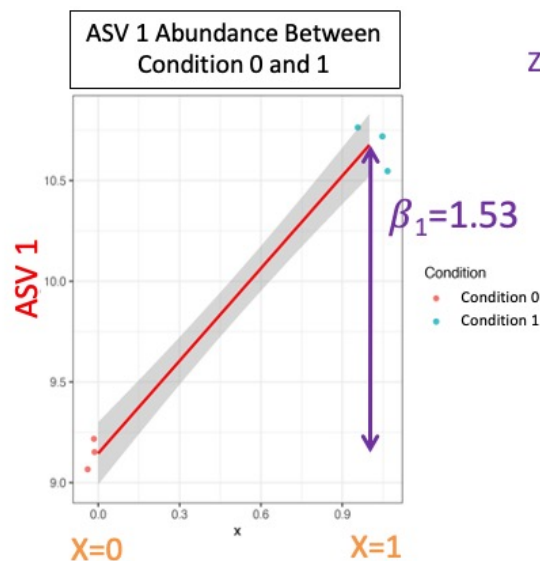




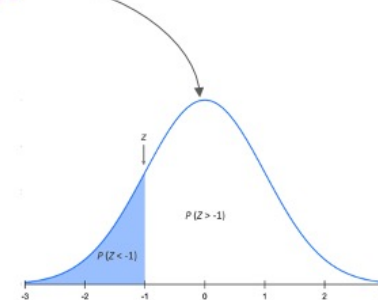
# Differential Abundance – DESeq2

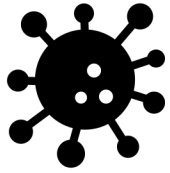
## DESeq2 P-Value:

1. The Slope or  $\beta_1$  is used to calculate a Wald Test Statistic  $Z$
2. This statistic is compared to a normal distribution to determine the probability of getting that statistic



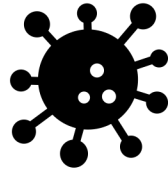
$$Z = \beta_1 / SE_{\beta_1}$$



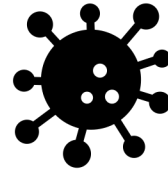


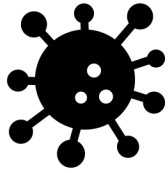
# Acknowledgement

Much of this tutorial has been adapted from [Astrobiomike's Amplicon Analysis Tutorial](#) and the [Galaxy Tutorial on Amplicon Analysis](#)



Setup





# References

1. <https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html>
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6391518/>
3. <https://www.nature.com/articles/nm.4517>
4. [https://astrobiomike.github.io/misc/amplicon\\_and\\_metagen](https://astrobiomike.github.io/misc/amplicon_and_metagen)
5. <https://www.nature.com/articles/nrmicro3330/figures/1>
6. [https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X\(17\)30709-7/fulltext](https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(17)30709-7/fulltext)
7. [https://www.drive5.com/usearch/manual/fastq\\_files.html](https://www.drive5.com/usearch/manual/fastq_files.html)
8. <https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing.html>
9. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4927377/>
10. <https://genome.cshlp.org/content/21/3/494/F1.expansion.html>
11. <https://benjjneb.github.io/dada2/tutorial.html>
12. <https://www.statisticshowto.com/>
13. <https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/>
14. <https://www.dataanalytics.org.uk/abundance-based-dissimilarity-metrics/>
15. [https://hbctraining.github.io/DGE\\_workshop/lessons/04\\_DGE\\_DESeq2\\_analysis.html](https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html)
16. [https://tuftsdatalab.github.io/Research\\_Technology\\_Bioinformatics/workshops/IntroToRNAseqGalaxy/slides/galaxyWorkshop\\_idgh1001\\_15Feb2022.pdf](https://tuftsdatalab.github.io/Research_Technology_Bioinformatics/workshops/IntroToRNAseqGalaxy/slides/galaxyWorkshop_idgh1001_15Feb2022.pdf)