Supplemental Materials

for

Sharing Notes Is Encouraged: Annotating and Cocreating with Hypothes.is and Google Docs

Carlos C. Goller1*, Micah Vandegrift2, Will Cross2, and Davida S. Smyth3

1Department of Biological Sciences, Biotechnology Program, North Carolina State University, Raleigh, NC 27695;
2NC State University Libraries, Raleigh, NC 27695;
3Eugene Lang College of Liberal Arts at The New School, New York, NY 10011

Table of Contents
(Total pages 36)

Appendix 1: Guidelines for annotations, summaries, and class notes
Appendix 2: Engagement and assessment data
Appendix 3: Additional suggestions for implementation

*Corresponding author. Mailing address: 6104 Jordan Hall, Campus Box 7512, North Carolina State University Biotechnology Program, Raleigh, NC 27695. Phone: 919-513-4135. E-mail: cgoller@ncsu.edu.
Received: 12 April 2020, Accepted: 28 November 2020, Published: 29 January 2021

©2021 Author(s). Published by the American Society for Microbiology. This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial-NoDerivatives 4.0 International license (https://creativecommons.org/licenses/by-nc-nd/4.0/ and https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode), which grants the public the nonexclusive right to copy, distribute, or display the published work.
# Table of Contents

Appendix 1: Guidelines for annotations, summaries, and class notes. ............. 3

Guidelines for Annotations and Summaries ............................................................. 3

Learning Outcomes ..................................................................................................... 3

Introduction .................................................................................................................. 3

Getting Started ............................................................................................................ 3

Annotations .................................................................................................................. 4

Grading ......................................................................................................................... 4

Summaries .................................................................................................................... 4

Rubric for Article Summary ......................................................................................... 4

Evaluation Method ....................................................................................................... 5

Annotations and Summaries for the Papers You Select ............................................. 5

Example Reading List ................................................................................................. 6

Guidelines for Class Notes ......................................................................................... 8

Introduction and Charge ............................................................................................. 8

Course Goals ............................................................................................................... 9

Learning Outcome for BIT 577 Students ................................................................... 9

Lectures ....................................................................................................................... 10

**Lecture 1. What is metagenomics?** Introduction to questions and methods. .......... 10

**Lecture 2. What is diversity?** Different metrics ....................................................... 14

**Lecture 3. What pipelines are available?** Pipelines and computational resources. .......................................................... 19

**Lecture 4. What are OTUs and ASVs?** Introduction to QIIME environment and DADA2 ......................................................................................................................... 21

What is QIIME 2? ........................................................................................................ 23

**Lecture 5. How can we plot this?** Data visualization with R. ......................... 25

**Lecture 6. How certain are we?** Methods and statistics found in selected papers ................................................................................................................................. 27

Appendix 2: Engagement and Assessment Data .................................................... 28

Credit Awarded on Individual Quizzes for Questions Related to Summarizing Main Points of Articles .................................................................................................................................................. 28
Quiz Question Key ........................................................................................................... 28
Quiz 1 ................................................................................................................................ 28
Quiz 2 ................................................................................................................................ 29
Quiz 3 ................................................................................................................................ 29
Bonus Question on Quiz 3 ................................................................................................. 30
Analyses of Class Notes and Selected Final Exam Questions ........................................... 33
Appendix 3: Additional Suggestions for Implementation ................................................. 35
Logistics .............................................................................................................................. 35
  Reading, Annotation, and Discussion of Articles ............................................................ 35
  Instructor videos reviewing the Class Notes ................................................................. 35
  Implementation ............................................................................................................... 35
Appendix 1: Guidelines for annotations, summaries, and class notes.

Guidelines for Annotations and Summaries

Learning Outcomes
- **Read** and **critically analyze** literature related to microbiomes and metagenomic approaches.
- **Annotate** and **summarize** salient information about the studies we discuss.
- **Create rich** annotations using Hypothes.is to engage in productive open discussions that focus on the methodology and implications of the work described.
- Actively **collaborate** in a scholarly community to share ideas, ask questions, and draw conclusions.

Introduction
We will read a series of recent studies in the BIT 477/577 Metagenomics course that will expose you to current trends in metagenomics and analyses of microbial communities. Each person will be required to provide at least ten (10) meaningful (relevant, thoughtful, and specific) annotations of each article discussed by the day the article is presented in class. This activity will allow us to have productive discussions of the content. Go beyond superficial comments: can you connect information, highlight limitations, link to other studies, include tags, respond to others...?

Getting Started
We will use the Hypothes.is tool to annotate the articles and organize our thoughts. Visit: web.hypothes.is to learn about this powerful tool.

1. Start by visiting the Quick Start Guide here: https://web.hypothes.is/help/quick-start-guide/

2. After you read the Quick Start Guide and sign up for an account, read about annotations: https://web.hypothes.is/help/annotation-basics/

3. We have formed a group for this course called Exploring Biodiversity. You can join the group by following this link: https://hypothes.is/groups/Ji9YBWvd/exploring-biodiversity
Annotations

You have been assigned three papers to read over the course of the class. The syllabus includes the dates for the discussions of these articles. Before each session, each one of you will use Hypothes.is to critically analyze the article and engage in discussion with others. Focus on the approaches the authors took and their assumptions. When possible, include links to other resources that may enrich the discussions. **Be sure to annotate as part of the Exploring Biodiversity group and refresh the pages frequently to see what others in the class are finding!** In class, we will discuss the papers and dive into the details. These studies were selected based on the topics, techniques, and analyses presented and they highlight cultural diversity in science and microbiome research.

Grading

Timely annotation (by the day of the discussion) of the articles will contribute 60 (20 for each assigned article) out of the 150 points for the Readings and Summaries (15% of course grade). Each of the two summaries will count for 45 points.

Summaries

In groups of four, you will work together on a GoogleDoc to summarize the comments from your peers and provide an overview of the study, significant findings, and, importantly, the methodology and its advantages and limitations. All members of the group should contribute to the GoogleDoc summary that will then be posted on our BIT Metagenomics Pressbooks site ([https://metagenomics.pressbooks.com/](https://metagenomics.pressbooks.com/); with your permission).

Rubric for Article Summary

A 2-3 page summary of the article is due **a week after the discussion of the paper**. A link to a GoogleDoc will be provided to each group for writing and editing. A member of the group will also post the Overview, Results, and Summary as Note Pages on the Hypothes.is page for the article.

Summaries will be used to cultivate your writing, summarizing, and critical thinking skills. **This writing exercise is your opportunity to interpret annotations and think about issues with this topic and then spend some time thinking and writing about your impressions.**
Evaluation Method

10 pts. – Overview
25 pts. – Results
10 pts. – Summary

The summary should be typed, 10-12 point font, and single-spaced. **Do not** copy text directly from the paper. **Use your own words! This document will be posted for public consumption.**

When writing your summary, address the following points, if applicable:

1. **Overview**
   - Include article title, authors, and journal **reference**.
   - A short 3-5 sentence statement on the **purpose** of the article and how it contributes to knowledge in the field.

2. **Results (pick **TWO or THREE KEY** figures/tables to summarize)**
   - For each figure/table summarize the following: **Note: for multi-panel figures you may need to break it down panel by panel.**
   - What is the hypothesis? (e.g., what question(s) are the authors addressing?)
   - How was the question addressed?
     - What experiment did they perform? Summarize the experimental details.
     - Be sure to state what they used as their controls (if applicable).
   - What did the data show?
   - Does what the authors say follow the evidence provided?

3. **Summary**
   - Overall what knowledge was acquired from the study?
   - What are some of the complexities of this situation?
   - Does all this make sense together?
   - What questions remain to be answered?
   - How relevant is this approach to other problems/samples/cases?
   - What are you left thinking?

**Annotations and Summaries for the Papers You Select**

The last article (article 4) you will read is one you select. The summary of the paper of your choice. Provide the Hypothes.is link. Each person should annotate the document, and the group should submit a summary to post using the rubric above.
Example Reading List

→ Join Biodiversity Hypothes.is group:
https://hypothes.is/groups/Ji9YWvd/exploring-biodiversity

| Week of… | Reading Assignment | Group(s) Summarizing |
|----------|--------------------|----------------------|
| 10/20    | **Lecture 1.** Discussion of **Paper #1.** Principal Antunes *et al.* 2016. *Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics.* *Scientific Reports.* 6. Article number: 38915. Hypothes.is link [here](#). More about the authors [here](#). | Group 1 Submit summary on Moodle one week after in-class discussion. |
| 11/3     | **Lecture 3.** Discussion of **Paper #2.** Che *et al.* (2019). *Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing.* *Microbiome.* 7. Article number: 44. Hypothes.is link [here](#). More about the authors [here](#). | Group 2 Submit summary on Moodle one week after in-class discussion. |
| 11/12    | **Lecture 5.** Discussion of **Paper #3.** Nowinski *et al.* (2019). *Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom.* *Scientific Data.* 6. Article number: 129. Hypothes.is link [here](#). More about the authors [here](#). | Groups 3 and 4 Submit summary on Moodle one week after in-class discussion. |
| 11/26 | Summary of paper of *your* choice due 12/2. Provide the Hypothes.is link. Each group member should annotate individually, and the group should submit a summary to post by 12/2.  

**With your permission, these summaries will be posted on our BIT Metagenomics website for others to use.** | **All groups.**  
Submit summary on Moodle by 12/2. |

**Note:** Hypothes.is works with Google Chrome browser.
Guidelines for Class Notes

Introduction and Charge

There is a massive amount of information about metagenomics! We can easily get overwhelmed by the number of tools, complex analyses, and large datasets. We have an opportunity to work together to summarize as a group, concepts that are central to the understanding of the power and applications of metagenomics.

This document is your notepad. You have access to edit it as you wish, and I encourage you to use a rich array of resources at your disposal. Include links, edit for clarity, and summarize. I have provided the framework in terms of the learning outcomes for each session. It is up to you to work collaboratively to produce a summary for each session. To work efficiently, we need some ground rules:

- Everyone is expected to contribute to the document during the session and up until Thursday after the lab that week.
- You can suggest comments to your peers by clicking on “Suggesting mode” in the top right corner. You can also make comments and respond to your peers. I will moderate and resolve comments.
- The material you incorporate into this document is meant to help all of you understand the concepts and make use of the tools and analyses appropriately. As mentioned previously, the number of resources can be overwhelming.
  - Select resources carefully: use reliable resources.
  - Cite your sources appropriately.
  - Provide links when available.
- The material you incorporate here will be turned into short summary videos for each lecture. I will send out the weekly summary video along with a link to the text.
- You are responsible for reviewing the notes we create: they will appear on quizzes to promote your ability to retrieve this knowledge and promote long term memory.
- At the end of the semester, this document will become a virtual e-book we all contributed to and, as such, will be credited as authors. This document will be found here: https://metagenomics.pressbooks.com/
  - You will be asked to sign a document acknowledging that it is ok for us to publish this content with or without your name as a contributor.
  - The current content of the site will be replaced with what we create, becoming V2019.

I hope this challenge and opportunity motivate you to create a resource that you and others will benefit from. We have the support and expertise of the NCSU OPEN Incubator project. We are a community of scholars and want to create an OPEN set of resources for the broader community. Start creating in the open.
Course Goals
Upon completion of the course, participants will be able to:

- **Demonstrate** laboratory skills required of a modern-day molecular biologist in the era of next-generation sequencing. This includes keeping detailed and accurate laboratory notes (*e.g.*, electronic records for sequence analyses) and choosing and using an appropriate sequence analysis tool.

- **Read** a scientific article and evaluate how bioinformatics methods were employed by the authors to explore a particular hypothesis. [from CourseSource framework]

- Given a scientific question, **develop** a hypothesis, and define computational approaches that could be used to explore the hypothesis. [from CourseSource framework]

- **Use** pre-existing tools to analyze a metagenomic data set to determine the set of organisms present in a metagenomic sample (*e.g.*, 16s rRNA, Greengenes, mothur, etc.). [from CourseSource framework]

- **Interpret** data and **identify** limitations related to metagenomic surveys.

Learning Outcome for BIT 577 Students

- **Design** a critical thinking scenario and **explain** analyses for hypothesis testing of metagenomic data.
Lectures

Lecture 1. *What is metagenomics?* Introduction to questions and methods.

- **Define** next-generation sequencing (NGS) or high-throughput sequencing. Massively parallel sequencing by synthesis (non-Sanger-based).
  - Examples: Illumina MiSeq is “2nd generation,” Nanopore ([https://nanoporetech.com/](https://nanoporetech.com/)) is “3rd generation” PacBio sequel ([https://www.pacb.com/](https://www.pacb.com/))?  

- **Illumina**: Sequencing by synthesis, building DNA model, and cataloging sequence by fluorescence
  - General Steps/ Method
    - PCR desired gene
    - Tag PCR products with bar codes and tags
    - Both forward and reverse reads are obtained but forward usually has greater precision
    - Generate library based on PCR
    - Sequence on MiSeq
    - Analyze via software (2)
  - Pitfalls/Limitations:
    - Read length
    - Read 2 has less quality (Q scores). See this for Illumina Q scores: [https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf](https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf)
    - Storage and utilization of a great amount of generated data
    - Expensive (run ~$2,000, reagents and barcodes $900, Smart controls $100)
  - Benefits/Advantages
    - Lots of reads
    - Can generate sequence data from multiple species in a given sample
    - Easy to use
    - Fast results
    - Barcoded library to separate pooled samples using bioinformatics as necessary
      - Benefits/Advantages
        - Lots of reads
        - Can generate sequence data from multiple species in a given sample
        - Easy to use
        - Fast results
Barcoded library to separate pooled samples using bioinformatics as necessary

- Previous methods
  - Capillary required individual bacteria be grown and cultures, thus individually sequenced.
  - Sanger Sequencing: terminally tagged nucleotides
    - Required primers

“Phased primers” for QIAseq 16S/ITS (for higher diversity in samples)
https://www.qiagen.com/us/products/next-generation-sequencing/qiaseq-16s-its-index-kits/#orderinginformation

- Nanopore: Still massively parallel, uses many pores to read each strand where each NT disrupts the ionic current and leaves a trace signature
  - Trace files: measures ionic disruptions to determine base present.
  - Read speed: 400bp/s
  - Sequence bins: add sample to beads to fragment DNA then add adaptor (from Nanopore) that enables barcode identification to separate samples by compost pile.

- General Steps/ Method
  -

- Pitfalls/Limitations:
  - Quality is less/ questionable (Q~ 11)
  - Storage and utilization of a great amount of generated data

- Benefits
  - Read length (longest 1mil bp)
  - Generates quickly across long strands
  - Completely mobile
  - Cheap(er) equipment ($900 flow and $500 reagents)
  - Shotgun sequencing

- Compare/Contrast w/ Illumina
  - Similar:
    - Massively parallel
  - Different:
    - Measuring disruption in ionic current → get trace files → software determines what base pairs
- Using a pore and stringing DNA through it
- Data quality
- 16S has both highly conserved and variable regions. This allows a single primer set to be used to identify many different species based on genomic variations.

Define **metagenomics**.
- Study of genomic samples from a defined microbial community
  - Genes that are present in the sample (shotgun sequencing)
  - Amplicons (16S, ITS)
- Could also be defined as the study of microbial genomics obtained directly from an environmental sample (eDNA) using sequencing approaches.

List two applications of metagenomics in health, industry, or medicine.
- *Role of the microbiome in humans* more specific? How do the communities of microbes affect our health?
- Is there a correlation of lifestyle to gut microbiome?
- Functional metagenomics for enzyme discovery
- Finding the relationship between environmental data and metadata
- Time-series analysis (reminds me of the composting article: [https://www.sciencedirect.com/science/article/pii/S1369527415000478](https://www.sciencedirect.com/science/article/pii/S1369527415000478))

In lab, we barcoded the end of the PCR products (16S and ITS)

**“Dirt is Good” by Jack Gilbert and Rob Knight**
Microbes probably associated with certain AIDs based on variations and commonalities with/between households

**American Gut Study**
- Objective: Establish relationships (correlation/associations) between bacteria with body and lifestyle, diseases, interactions, diet, etc.
- Observational Study via Survey and Sample collection
- Crowdsource for American public (e.g. hospitals, public spaces, ) to have access to microbiome within themselves.
**Microbiome:** a community of microbial genomes that live on/within organisms.
   ALT: Collection of genomes...

**Microbiota:** Depending on the author, interchangable with microbiome

**Agriculture** is just beginning research into the effect of microbes on plant development. The potential to identify microbes that assist with drought tolerance, heat stress, defense, and nutrient acquisition could be used to improve production as extreme weather becomes more prevalent.

Applications in the **industry:**
The Biofuel industry as an example, is always looking at the consortium of organisms that can generate methane, ethanol, or other compounds. Using metagenomics we can assess those microbes and characterize that microbial community. Food industry for instance is looking at microbes performing fermentation and how to characterize or manipulate that community to improve their products.

**Three major findings from Paper 1**
  1. The abundance of different organisms changed over time
  2. The potential identification of a new bacteria
  3. The workflow/method in paper 1 may be useful for our lab work or future projects.

**Approaches used in the paper:**
An introduction and a workflow of metagenomics from an online workshop:
[https://youtu.be/LUS62N7tSUU](https://youtu.be/LUS62N7tSUU)
Lecture 2. *What is diversity? Different metrics.*

- **Define** and explain the concepts of *metadata, OTU, rarefaction curve.*
- **Explain** three different diversity metrics.
- **Identify** and describe the limitations and assumptions of certain diversity metrics.

**What is diversity:**

**Define**

- **Metadata**
  - Data about the data. For example, date, location of sample collection, the concentration of DNA samples, etc.
  - Standards for metadata can be found on the Genomics Standards Consortium (gensc.org) → creates standard descriptors for metadata and sequencing approaches.
    - Genomics Standards Consortium used to help classify the data
  - Metadata can be distinct to specific fields (clinical microbiology has different metadata than environmental microbiology)
  - Indicates the “where, when and what” conditions of samples.

- **OTU (Operational Taxonomic Units):** defines a species (classify sequences together)- used to classify closely related groups based on sequence similarity
  - Not reproducible
  - *Traditionally used as a means of species identification or classifying sequence clusters*
    - *Generally 16S or 18S (ribotyping)*
  - *Reference*- Compares against known reference standards
    - Can miss species or misidentify. Only as good as the reference data set.
  - *de novo-* Compares against data in the set
    - Captures information based on what is in the sample. Not generalizable
  - *Amplicon Sequence Variants (ASV)*- Ben Calahann
    - Empirically based on the sequence
    - Used as the basis for DADA2 clustering

- **Determine:** Good use of classifying metagenomic samples
- **Our goal/application of OTUs**
  - Classify samples in various samples (in common or contrast)
○ **Rarefaction Curve:** a curve describing the growth of the number of species (y-axis) discovered as a function of individuals (x-axis) sampled.
  ■ Based on rarefaction curve
  ■ Allows researchers to assess *species richness* from sampling results
  ■ Added parameter: Read number (x) and sequence variability (specification)(y)
  ■ **Not to be confused with “rarifying”**
    ● Normalizing based on number of sequences present in various samples so that all samples have the same number of sequences
    ● Go back to sample to sub-sample take into account subsample for each sample
      ○ Controversy due to exclusion of some collected data.
  ■ High abundance organisms affect likelihood of finding low abundance organisms
  ■ Our experiment: Superimposing our data
    ● Comparing species present
    ● Used to determine whether or not we need additional sequencing
  ■ **Assumptions**
    ● Assumes differences are genuine and not errors
    ● Shotgun will be very difficult to capture the rare members
    ● How likely is it that more sequencing will help identify low abundance organisms (i.e. *not factored into graph*)
    ● Equal probability of identifying species in samples
    ● Rare organisms may have minimal effect
    ● A higher plateau has more variability. A plateau usually suggests that you sequenced “enough” to identify the majority of the organisms.
    ● Bioinformatics pipeline has minimized sequencing errors
    ● Every new read is a new organism

1) **Alpha diversity (focuses on one sample)**
   ○ **General: Within a sample**
     ■ **Species richness:** How many different species are present?
     ■ **Species diversity:** How different is the distribution?
   ○ Assumptions
   ○ Limitations
   ○ Applications
2) **Beta diversity (across many distinct samples)**
   - **General:** comparing microbial composition
     - microbial composition of one environment compared to another
     - Who is present, left-out, ___
     - “Is your sample different and how?”
   - **Assumptions**
   - **Limitations**
   - **Applications**

Videos
1. Alpha Diversity
   PD: Phylogenetic diversity, a phylogenetic alpha-diversity metric

**Measuring diversity metrics:**

1) **Shannon** - measures how evenly microbes distributed in a sample
   - **Advantages**
     - Accounts for number of species and abundance of species
   - **Limitations**
     - Single measure of richness rather than independent measurements
     - Does not account for the uniqueness of the biodiverse community (high biodiversity with undesirable organisms or low biodiversity with rare organisms)

2) **Chao1** - calculates the estimated true species diversity of a sample and is based on abundance (Alpha Diversity measure), requires the abundance of individuals belonging to a certain class in a sample.
   - Takes into account rare species

3) **Bray-Curtis:**
   - Beta diversity
   - 0 to 1 scale- easy to interpret: 0 is samples are identical, 1 completely different.
   - Dissimilarity index

**Limitations**
- Argued that this coefficient may provide misleading results for species abundance data containing zeros (e.g. Orlóci, 1972, Orlóci, 1978, Legendre and Gallagher, 2001)
- Often misinterpreted as a distance, it is only a dissimilarity index because it counts frequency

**Advantages**
- Intuitive scale for readers. Easy to calculate and scale is from 0 to 1 so meaning is clear
• calculate it you simply subtract the Bray Curtis dissimilarity (remember, a number between 0 and 1) from 1, then multiply by 100.

4) Jaccard: “Similarity between sites”
- Compares two samples based on the presence and absence of microbes.
- Size of the intersection divided by the size of the union of the samples.
- **Assumption:** The more species the samples have in common, the more similar they are to each other.
- **Limitations/ Disadvantages:** bad for small data sets. based on ranking, collaborative data sets, does not account for phylogeny.
  - Helpful visualizations: [https://www.oreilly.com/library/view/hands-on-convolutional-neural/9781789130331/a0267a8a-bd4a-452a-9e5a-8b276d7787a0.xhtml](https://www.oreilly.com/library/view/hands-on-convolutional-neural/9781789130331/a0267a8a-bd4a-452a-9e5a-8b276d7787a0.xhtml)
  - [https://thatware.co/jaccard-similarity/](https://thatware.co/jaccard-similarity/)

5) Euclidean Distance:
**Define:** The sqrt of the sum of the squared differences between two data sets (similar to pythagorean theorem).

**Builds on the most basic idea of differences between samples but can be applied across multiple variables (in a matrix).**

**Applications:** Beta diversity, Correlation is inversely related to Euclidean Distance

**Limitation:** Only appropriate for the same scales, not good for clustering, very bad at multidimensional analysis

**Advantages:** Simple Analysis, no elaborate analysis, good for absolute magnitudes

**Data Required:** Empirical values on two populations where the data is on the same scale.

6) UniFrac:
- Phylogenetic- based beta diversity
- Percent of observed branch length unique to either sample
- Identical communities D=0, related communities D=0.5, unrelated communities D=1

  **A. Unweighted:** only uses presence/absence emphasizes the minor species

**Limitations:** Does not take into account the abundance of populations like weighted unifrac
B. **Weighted**: takes into account relative abundance, emphasizing the more dominant species, quantitative

**Limitations:**
- not a reliable approach to measuring similarity
- small samples can inflate weighted UniFrac values

**Lab3**

Advantages of nanopore:

1. Real-time data
2. PCR is not very required
3. Cheaper
4. **Longer sequence reads**
5. Real-time
Lecture 3. *What pipelines are available?* Pipelines and computational resources.

- To critically **analyze** a published tool.
- To **describe** the importance of processing reads before use.
- To **examine** different approaches used for processing reads.

What pipelines are available:

- **MG-RAST** (Metagenomic Rapid Annotations using Subsystems Technology)
  - Suggests automatic phylogenetic and functional analysis of metagenomes
  - It is also a very large data repository for metagenomic data

- **Nephele**
  - Provides QIIME1, mothur, and DADA2 pipelines for amplicon data and the bioBakery pipeline for metagenome shotgun data.
  - Includes a quality control pipeline for demultiplexed
  - Some issues with Nephele may include low flexibility with the analyses that can be done

- In-Class Job
  - Lab: Taking short reads and analyzing possible coding sequences, functional**, and
  - Duplicate reads expected, given PCR used and 16S*
  - Omit job/samples with significant relatively low reads
  - Adapter removal (Illumina universal adapters) via GSL pipeline in QC
  - DaDa-2 Pipeline generates alpha diversity within each sample based on metadata
    - Different diversity metrics yield different outputs
    - Rarefaction curves show plateauing
  - Beta-Diversity
    - Goes back to subsample to generate same number of sequences per sample
    - *Heatmap showing relative abundance/richness* *

- **OneCodex**
  - Microbiome platform for genomic analysis and data management. Key features include:
    - Taxonomy analysis
    - Functional gene predictions and markers (identification of potential genes of interest such as AMRs)
    - Reporting capabilities to answer specific questions
- Does not do QC or paired-reads, but interactive plot produced within minutes thereafter.

- **CosmosID**
  - A bioinformatics platform that achieves strain-level analysis with industry-leading sensitivity and precision.
  - Helps identify and characterize isolated organisms for strain discovery and sub-typing
    - Emphasizes pathogens, antimicrobial resistance, and virulence
  - Expensive (but free 24 samples or 12 WGS samples upon signing up)
  - They work with Qiagen and they have CLC plug-in function

- **Amazon Web Services**
- Etc.

**Main points of article #2**
Three wastewater treatment facilities: influent, activated sludge, and effluent
Comparing Nanopore (long read) and Illumina (short read) readings
Cultivated species: took affluent + LB + antibiotics —> sequenced non-multidrug resistance
Illumina centrifuge: analyze sequences, get standardized pipelines
Plots: the abundance of antimicrobial genes in different samples
  1) Pathogens acquire multidrug resistance in wastewater treatments (plasmids facilitate this resistance acquisition)
  2) Higher prevalence of plasmids with antimicrobial resistance genes found in the effluent
  3) Combination of Nanopore + Illumina
  4) Limitations: No percentage abundance given for species
Lecture 4. *What are OTUs and ASVs?* Introduction to QIIME environment and DADA2.

- **Discuss** and critically **evaluate** each step of a 16S QIIME standard operating procedure (SOP).
- **Explain** the utility and fundamental tools of QIIME2.

Historically, two methods used to observe the microbes were:
1. Microscope
2. Culture

These methods are very limited-- microscopy can’t give any information about genes and the majority of microbes can’t be cultured. Sequencing can overcome both of these challenges and continues to get cheaper.

**Marker-gene Sequencing:**
- Targeted approach
- Analyzes composition of the community
- Has conserved flanking regions
- Uses barcodes

**Challenges:**
Errors

1. **Compare** and **contrast** OTUs and ASVs.

**OTUs-** operational taxonomic units
- Clustering sequences that fall into fixed similarity thresholds
- Closed-reference methods for defining OTUs: reads that are sufficiently similar to a sequence in a reference database are recruited into a corresponding OTU
- *De novo* methods of defining OTUs: reads are grouped into OTUs as a function of their pairwise similarities

**Closed reference OTUs: aka mapping against a reference database**
- Limitations:
  - Defined by sequences that were not directly observed in data/sample, missing some references for amplicons so amplicon data can be lost
  - Dependent on the reference data used. ie did you use human gut reference set when looking at the marine community
  - Samples → PCR and sequencing (includes a lot of errors)→ amplicon reads
- **Reference sequences**: Confirmed/supported previously characterized sequence(s)
- Maps data to closest reference
- Represented data: OTUs (closed reference)

OTUs are not as precise as ASVs, is tractable and reproducible, and non-comprehensive.

**de novo OTUs:**

Making from data directly, as opposed to by proxy via reference database - *we are not using a reference database*

Data is clustered to most common representatives, grouped to within 3%
Most common algorithm: Find the most abundant sequence in data, enclosed region (~70%), collapse, analyze, re-collapse

**A. Strengths**
- Working with actual data
- Groups errors with where they are closest
- More resolution/precision

**B. Weaknesses**
- Lost resolution from collapsing - No reference database is perfect - so data can be lost
- Very sensitive to detail (okay for Phylum level, but not as much the granular levels/lower taxa)
  - Increased occurrence of random error
- Not consistent labels
- Not replicable- every run with new data may generate new OTUs

**C. Amplicon Sequence Variants**
- Removes errors, can see true populations even close to together
- Consistent labels to identify de novo that can be reproduced between samples versus just being clustered*
- Continuous data integration because data can be analyzed discretely and compared to whole.

**d. Distinguishing Signal from Noise**
- **DADA2**
  1. Effective Hamming Distance
  2. Allows continuous data integration
    - Eliminates the need for joining reprocessing of raw data.
  3. Unlimited dataset size
- OTUs
1. Threshold reference

A limitation of OTUs is that this approach groups the data in such a way that may result in it (the dataset) being unusable in the future if, say the OTUs identified in the past are not identified in the future dataset (i.e., the OTUs are a secondary dataset to the actual data generated and it may not be relevant or applicable in the future; potentially inconsistent labeling).

- De novo is not as precise and tractable as ASVs, is comprehensive and non-reproducible.

**ASVs (Amplicon sequence variants):**
- Inferred unique sequences present in the original sample, after correcting for sequencing/sample preparation (e.g., PCR) errors.
- Does not use the arbitrary dissimilarity thresholds that define molecular OTUs
- These methods infer the biological sequences in the sample prior to the introduction of amplification and sequencing errors, and distinguish sequence variants that differ by as little as one nucleotide

- ASVs are precise, tractable, reproducible, comprehensive, continuous labels.

Also known as exact sequence variants (ESVs), sub-OTUs (zOTUs), haplotypes, oligotypes …

- **Discuss** and critically **evaluate** each step of a 16S QIIME standard operating procedure (SOP).
- **Explain** the utility and fundamental tools of QIIME2.

**What is QIIME 2?- microbiome analysis pipeline**

QIIME 2 is a powerful, extensible, and decentralized microbiome analysis package with a focus on data and analysis transparency. QIIME 2 enables researchers to start an analysis with raw DNA sequence data and finish with publication-quality figures and statistical results.

QIIME2 SOP- [https://chmi-sops.github.io/mydoc_qiime2.html](https://chmi-sops.github.io/mydoc_qiime2.html)

**QIIME2 workflow overview:**
Connect to server (linux)
Obtain and import fastq files and metadata
Demultiplexing (to determine which sample each read came from)
Sequence quality/ denoising and clustering/feature tables (that has counts (frequencies) of each unique sequence in each sample in the dataset) → uses DADA2 plugin or you can use Deblur

  - Feature table summary
  - Phylogenetic diversity analyses & weighted and unweighted UniFrac
  - Alpha and Beta diversity analysis
  - Alpha rarefaction plotting
  - Taxonomic analysis
  - Differential abundance across samples → uses ANCOM

A YouTube video that discuss the QIIME: https://youtu.be/nWeRN2IKIto
Lecture 5. How can we plot this? Data visualization with R.

- Given a formatted dataset and an appropriate visualization tool, the participant will be able to accurately **summarize** the output for different measures of diversity.
- **Explain** the **UniFrac** metric of distance and diversity.
  - **Unique Fraction Metric (Unifrac)** measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both. [https://aem.asm.org/content/71/12/8228](https://aem.asm.org/content/71/12/8228)
- **Interpret** and **evaluate** a 2D representation of multidimensional data (**ordination plot**)
- **Visualize** metagenomic data with **Phyloseq**
  - Phyloseq allows you to create graphs and analyze data once data in inputted and formatted for R studio
  - Data has already been clustered into OTUs

Ordination plot- visualization of beta diversity for identification of possible data structures.
  - PCoA is the most commonly used plot for microbiome data
  - Can be obtained via R and phyloseq
    - Ordination guide for phyloseq: [https://joey711.github.io/phyloseq/plot_ordination-examples.html](https://joey711.github.io/phyloseq/plot_ordination-examples.html)

**Paper 3:** Authors used metatranscriptomics, pipelines, RNAseq, Illumina, DADA2 for this study.

- Sampling was performed using an Environmental Sample Processor (ESP) which provides on-site (in situ) collection and analysis of water samples from the ocean.
  - The ESP had two filters: 5.0 uM pore for eukaryotic organisms and 0.22 uM pore to capture bacterial and archaeal microbes.
  - Environmental information/measurements were also taken by a CTD (conductivity, temperature, and depth) instrument mounted on the ESP.
- They constructed relative abundance of bacterial and archaeal maps by family.
  - Also constructed relative abundance of eukaryotic tax maps
- Why/how did they validate the study?
  1) **Quality control:** removed contaminants by removing and filtering of reads
     ■ BBDuk software
  2) Validated to other similar datasets by microbial community standards
  3) Used spike on controls- **used to calibrate abundance organisms by the amount of water filtered**
- Calculate relative abundance normalized by volume.

- This paper focused mainly on methods; missing results and discussion. It is a **data descriptor article**, which provides the scientific community with elaborative information and “high-quality materials data” to be able to understand and use in further research.
Lecture 6. *How certain are we?* Methods and statistics found in selected papers.

- **Describe** what published metagenomic survey figures represent.
- **Identify** statistical methods used for hypothesis testing in metagenomics.

What are some of the statistical tests that we saw in the three papers we discussed?

What statistical tests were done by QIIME2 as part of the Moving Pictures tutorial?  
https://docs.qiime2.org/2019.10/tutorials/moving-pictures/#differential-abundance-testing-with-ancom

What are some tools on Microbiome Analyst that you can use?

- Taxonomic profiling
- Diversity metric box plots
- Differential abundance levels e.g. stacked bar charts
- Visualization of similarities/ dissimilarities: e.g., PCoA plot
- Functional predictions/ profiles
- Classification
- Variation b/w samples through clustered heatmaps

Heatmaps: https://youtu.be/7xHsRkOdVwo and https://youtu.be/oMtDyOn2TCc
Appendix 2: Engagement and Assessment Data

Credit Awarded on Individual Quizzes for Questions Related to Summarizing Main Points of Articles

Figure S1. Students successfully summarized the main findings and key points for all three assigned readings. Students \( n = 15 \) took individual quizzes. One question on each quiz asked students to summarize the main point(s) of each article. Readings are listed in Supplemental 1.

Quiz Question Key

Quiz 1

How would you summarize the main point(s) of Paper #1 (Antunes et al. 2016. Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. Scientific Reports. 6. Article number: 38915.)?

Answers will vary but must include (each worth 0.5 points):

- use of NGS or high-throughput sequencing approaches to study the microbial composition of composting bins in a Brazilian zoo.
- Found potentially new species.
- Performed transcriptomics,
- and transcriptomics results match DNA sequencing.
Quiz 2

How would you summarize the main point(s) of Paper #2 (Che et al. (2019). Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. Microbiome. 7. Article number: 44.)?

Answers will vary but must include:

- use of Illumina and Nanopore technologies to sequence wastewater treatment plant samples,
- and identify antibiotic resistance genes and organisms.
- Must mention technologies and AMRs, 1 point each.

Quiz 3

How would you summarize the main point(s) of Paper #3 (Nowinski et al. (2019). Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. Scientific Data. 6. Article number: 129.)?

Answers will vary. 1.0 point for each main point of the article.

- The authors estimated the relative abundance of bacteria and archaea
  ○ constructed relative abundance of eukaryotic taxonomy maps
- Why/how did they validate the study?
  ○ Quality control: removed contaminants by removing and filtering of reads
    ■ BBDuk
  ○ Validated to other similar datasets by microbial community standards
    ■ Used spike-in controls: used to calibrate abundance organisms by the amount of water filtered
  ○ Calculated relative abundance normalized by volume.
- This paper focused mainly on methods; missing results and discussion
BONUS Question on Quiz 3

BONUS (1 point). When using Hypothes.is, what did you find yourself annotating/highlighting? Did you use keywords? Which ones? You must answer all three questions to obtain full points.

- I mostly found myself asking questions about methods or answering others questions if I knew the answer.
- I used keywords to search the papers.
- e.g., nanopore, luminescence, read depth, OTU, taxonomy

- Highlighting main points for sections and overall paper/pages (e.g., objectives, concerns, considerations, methods of troubleshooting & clarifications).
- Yes, "conclusion," "purpose," overall purpose, significance, tolerance, concepts, limitations, future direction, ideas, caution/notes.

- I tend myself annotating either points that clarify the author's analysis, points that did not make much sense, or points that could be supplemented by external resources.
- I did not use any keywords. I did not find them particularly helpful unless they were helpful in generalizing the papers.
BONUS (1 point). When using Hypothes.is, what did you find yourself annotating/highlighting? Did you use keywords? Which ones? You must answer all three questions to obtain full points.

I found myself annotating highlighting concepts I did not know about, so I would look them up and post what I found. I did not use keywords, but I should have used acronyms such as "database" for an unfamiliar database.

BONUS (1 point). When using Hypothes.is, what did you find yourself annotating/highlighting? Did you use keywords? Which ones? You must answer all three questions to obtain full points.

I found myself highlighting a lot of background information and results. Since I have only taken one microbiology course, I found a lot of the background was relevant to my understanding of the article.

I used keywords such as WGS, nanopore, or 16S/ITS Illumina to discern what types of sequences were collected.

Another set of keywords for these papers was diversity vs. abundance as it is important to include both metrics in microbial analysis to gain the full scope of the relevance.

BONUS (1 point). When using Hypothes.is, what did you find yourself annotating/highlighting? Did you use keywords? Which ones? You must answer all three questions to obtain full points.

I usually highlighted methods and sampling techniques the authors used. I didn’t use many keywords other than what was present in the context. Although, I usually had to define some words I defined acronyms like BBDUK and software techniques that I wasn’t familiar with.
BONUS (1 point). When using Hypothes.is, what did you find yourself annotating/highlighting? Did you use keywords? Which ones? You must answer all three questions to obtain full points.

1) I noticed several people annotating based on definitions, keywords, etc. but I find this a cop-out. I prefer to annotate where a claim is unsupplied or I do not understand something.

2) No, I'm still not sure what it meant by this. But I know I didn't know about them to use them.

3) I didn't use keywords, so none.

BONUS (1 point). When using Hypothes.is, what did you find yourself annotating/highlighting? Did you use keywords? Which ones? You must answer all three questions to obtain full points.

I would sometimes annotate words or methods I didn't know and had to look up or would highlight parts I had questions on or wanted further explanation.

Typical keywords I used were methodologies, ARM (autonomous robots), validation, microbes.

BONUS (1 point). When using Hypothes.is, what did you find yourself annotating/highlighting? Did you use keywords? Which ones? You must answer all three questions to obtain full points.

I'd like to annotate some terms that may be helpful to understand the paper clearly. Also, I'd like to annotate the results and methods that are interesting to learn. I personally did not use the keywords function. But I saw some keywords shared by other students. If I remembered correctly, someone use 'metagenomics' as a keyword.
Analyses of Class Notes and Selected Final Exam Questions

Learning Outcome:

Class Notes:

Q1.
Response: %

6. Why were the 16S/ITS primers “phased” or “staggered”?
   A. to decrease sequencing errors
   B. to increase read length
   C. to increase sequencing depth
   D. to increase diversity and quality scores

7. __________ is a metric for comparing membership and structure of different habitats.
   A. alpha diversity
   B. beta diversity
   C. rarefaction
   D. Q score

9. de novo generated OTUs are generalizable.

10. The UniFrac metric of distance and diversity requires a phylogenetic tree.

11. Amplification sequencing of the ITS is routinely used to identify bacteria.

12. A rarefaction curve shows abundance of specific taxa.

13. In CLC Genomics Workbench, OTU clustering used the databases Greengenes and UNITE as references to create taxonomy tables.

14. An advantage of the use of Amplicon Sequence Variants (ASVs) is that they allow for continuous data integration and data can be analyzed discretely and compared to the whole.

15. In QIIME2, the programs DADA2, and Deblur were used to measure the quality of the data.

17. Compare and contrast OTUs and ASVs.
OTU (Operational Taxonomic Units) - uses 16s or 18S amplicons (ribotyping) to classify closely related groups based on sequence similarity. OTUs can be determined using reference databases or de novo method. OTUs assembled using de novo methods are not reproducible while OTUs assembled using a reference database is.

ASV (Amplicon Sequence Variants) are reproducible across experiments. Methods using ASV remove potential PCR and sequencing errors from each sample. Groups are then defined by unique sequences present in the original sample. ASV does not use similarity threshold to determine groups and are able to resolve sequences that differ by one nucleotide.

21. What is/are the benefits of annotating with Hypothes.is and documenting collaboratively with the goal of following “OPEN practices”?

Answers will vary.
Appendix 3: Additional Suggestions for Implementation

Logistics

Reading, Annotation, and Discussion of Articles
All students annotate articles and then groups of students are assigned to summarize specific articles. In parallel, course content is addressed during class sessions and time is spent every other week discussing articles. Therefore, students are first assigned readings, they annotate individually, the entire class discusses in a class discussion, and groups submit summaries that are evaluated and shared with all students. Collaborative notes continue to be written after and often during each class session.

Instructor videos reviewing the Class Notes
The motivation for the instructor video is to demonstrate to students that the instructor reviews and comments on the collective notes before students are quizzed on the material. This serves as a form of feedback to the students and ensures that they feel supported and heard by the instructor. The short videos are easy to produce and convey the instructor’s interest in the collective voices of the participants that are represented in the notes. The instructor videos also offer an opportunity to draw connections between concepts and student contributions as well as clear misconceptions. An example of a video can be accessed:
https://youtu.be/OXobo_rxOHM (6:44 min)

Implementation

In our experience, all students participated in the creation of class notes by accessing the GoogleDoc and making edits. However, we did have a fraction of students that did not provide substantial contributions (extensive edits or new text). No grade was attached to this activity. In future offerings of the course, we have made class notes a graded activity and are currently evaluating outcomes.