Exploring Native Microbiota Lab Manual

BIO398, Winter 2024

Labs 9 & 10 : Preparation for Class Presentation, OOI Report, and Data/Sample Archiving

---

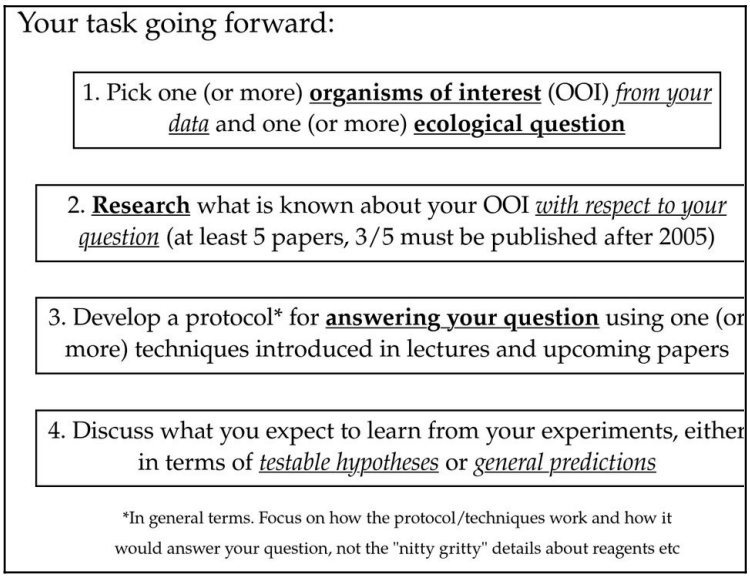
Overview:

We are now nearly at the end of the course! Coming up after the long weekend, we will have our final presentations which will be followed by your organism of interest (OOI) report. Today's lab is an opportunity to discuss your ideas with me and get feedback, suggestions of references, etc.

In addition, 10% of your final mark for this course will be allocated to laboratory record-keeping, findability of data / samples. The rationale behind this is that your DNA extracts and the resulting data provide a unique snapshot of the environment you targeted that may be useful to someone in the future for reasons you cannot foresee at the moment. Therefore, we should be archiving both our samples and data in a way that future researchers can take advantage of your hard work. However, it's very easy for samples or data to be lost or become unusable. This could be because of unreadable sample labels, samples that cannot be found, a lack of appropriate metadata, etc.

Part I - Developing your OOI report:

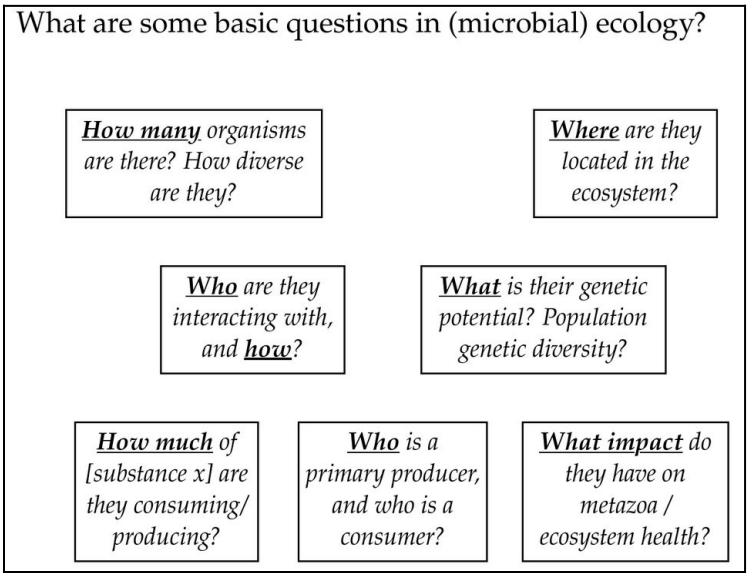
As discussed in class (and shown in the diagram to the right), this report is an opportunity to propose a unique research project based off of one (or more) organism(s) of interest (OOI) in your sample. I am using the word "organism" broadly - it could be a specific ASV (the finest resolution possible with our dataset) or it could be a broader group such as "Gammaproteobacteria". That being said, very broad groupings like "Bacteria" and "Archaea" are too coarse to be meaningfully ecologically. Talk to me if you are unsure whether the taxonomic level you are targeting is too broad.



The way in which you will accomplish this is to identify a ***scientific question*** (can be hypothesis-driven or more exploratory) that you want to address with ***a general protocol*** (inspired by readings in class or discussions with me) that uses your

dataset as preliminary results. Although I will not be marking you on whether your scientific question is earth-shattering or not (i.e. whether you are likely to get a "Nature paper" out of it) but I will be looking to see whether your scientific question is ***specific, answerable, and you have chosen the appropriate methods and understand them sufficiently.***

Although research in microbial ecology tends to be full of rather complicated methods (especially these days), it's important to remember that in most cases these methods aim to answer rather simple questions about ecosystems (see diagram to the right). Pick a question or set of questions that *interest you!* It need not be esoteric or complex, and can have links to the environment / human health if that is where your interest lies.



Finally, I have structured the final report as an imaginary grant proposal. Money is no object, and you can propose any technique or project that you can imagine. In both your presentation and final report you should aim to be persuasive. Imagine I'm a philanthropist interested in supporting scientific research and if you write a good enough proposal, I will fund you to do the research!

Suggestions for how to get started (if you don't know where to begin):

- Option 1: Pick an organism from your sample that has an interesting pattern in relation to your samples. Use the qiime2 plots or your spreadsheet to see whether it's one or more ASVs and then decide what level of taxonomic resolution you want to target.
- Option 2: Pick an organism that has some property you're interested in. Perhaps you want to target chemolithoautotrophs, heterotrophs, or mixotrophic algae. Do some research on this organism and find out what is known, and what is not known. Then design an experiment to advance the field.
- Option 3: Review the lectures for a method that interests you, and build your project around the application of this method (e.g. metagenomic binning into MAGs, metatranscriptomics, isotope probing, HTC, etc).
- Option 4: Discuss with me and I can give you some general ideas of what each taxa might be doing in the sample based on their name and you can choose one that appeals to you.

Part II - Data archiving / sample archiving / cleanup:

For this 10% of your grade, I will be marking you on the following:

1. Whether you cleaned up (2%)
    1. Cleaned & dried centrifuge tubes, and removed permanent marker labels using alcohol (if possible)
    2. Cleaned benchtop removing any empty tubes, reagents, etc, cleaning tools, and wiping down bench
    3. Discarded extra water, sediment, etc, and cleaned bottles as relevant, and put back any supplies you borrowed from the biology supply room
2. Whether your DNA extracts are findable and labelled clearly (3%):
    1. DNA extracts are contained in a box with clear labels and stored in the -80°C freezer. Label should have course name (Bio398), your name, year and semester. Please label the box with tape so it can be easily removed in the future for reorganization.
    2. Each tube has a clear label that corresponds directly to the data below.
    3. There are no unneeded intermediates (0.5ng, 3ng strip tubes).
3. Whether your data is uploaded to the Open Science Foundation (OSF; osf.io) in a format that can be used by me or others in the future (5%):
    1. A unique DOI / link is provided to the instructor by the end of the course ***by email*** that links to an OSF repository you created and is publicly accessible.
    2. The OSF repository has a description of the study sites targeted ("Wiki"), the general methods (with references as appropriate), and the overall goals of the study. This can be taken from your other reports.
    3. The OSF repository contains the first ***.zip file*** of processed data I sent to you. *To upload to OSF, first click the storage location (somewhere in Canada) and then you will see an "upload" button show up).*
    4. The OSF repository contains a ***spreadsheet file*** (tsv, csv, excel, ods all OK, but PDF or Word ***not acceptable****)* with the following contents:
        1. A "SampleID" column that ***exactly matches*** the beginning of the demultiplexed output file name. This can be taken from our demultiplexing sheet we filled in together (second column where I added your name). Please include all possible samples, even if they did not yield enough reads to be usable. This will allow me, or anyone else in the future, to link your data with where the samples were collected, associated measurements, etc.
        2. A "SampleQC" column that indicates whether sequences from a sample should be used or not. Use the table below to put in a number corresponding to the quality of your sample:

1. "1" if the sample appears fine without any major contamination.
2. "4" if the sample is grossly contaminated (appears to be only contaminant taxa) or does not match the environment at all (e.g. seawater taxa in a soil sample).
3. "2" if there is partial contamination (>5% of negative control ASVs in a given sample; make note in following column for likely reason).
4. "6" if not enough sequences / not pooled.

3. A "SampleQC_comment" column that indicates why you chose the above flag.
4. A "DNAid" column that ***exactly matches*** the labels found on your DNA extracts. Even if this is identical with your "SampleID", please still fill this column in. You will probably also have cryovials with lysate. These should be kept too, but make sure the IDs match to your DNA extracts. If they do not match, either relabel the cryotubes or put another column into your spreadsheet with their labels.
5. Metadata, with appropriate headers (indicating units and measurement value), values that do not contain units (i.e. for temperature put "1" not "1°C"). ***At minimum, you must have latitude, longitude, time, and elevation.***

An example of an OSF repository can be found here: https://osf.io/z8arq/

Key	Entry Term	Abbreviated term	Term definition
0	no quality control	none	No quality control procedures have been applied to the data value. This is the initial status for all data values entering the working archive.
1	good value	good	Good quality data value that is not part of any identified malfunction and has been verified as consistent with real phenomena during the quality control process.
2	probably good value	probably_good	Data value that is probably consistent with real phenomena but this is unconfirmed or data value forming part of a malfunction that is considered too small to affect the overall quality of the data object of which it is a part.
3	probably bad value	probably_bad	Data value recognised as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena.
4	bad value	bad	An obviously erroneous data value.
5	changed value	changed	Data value adjusted during quality control. Best practice strongly recommends that the value before the change be preserved in the data or its accompanying metadata.
6	value below detection	BD	The level of the measured phenomenon was too small to be quantified by the technique employed to measure it. The accompanying value is the detection limit for the technique or zero if that value is unknown.
7	value in excess	excess	The level of the measured phenomenon was too large to be quantified by the technique employed to measure it. The accompanying value is the measurement limit for the technique.
8	interpolated value	interpolated	This value has been derived by interpolation from other values in the data object.
9	missing value	missing	The data value is missing. Any accompanying value will be a magic number representing absent data.
A	value phenomenon uncertain	ID_uncertain	There is uncertainty in the description of the measured phenomenon associated with the value such as chemical species or biological entity.

Table 1: SeaDatNet Quality Flagging System (source: https://www.geotraces.org/geotraces-quality-flag-policy/)