Ecology of Marine Microbes (First Week)

Course Plan

May, 2025

Contents

Preface

The purpose of this document is to share the details of the first week lectures and hands-on exercises for the "**Ecology of marine microbes I (EMM)**". In the following sections you will find an hour-by-hour plan for our activities, learning objectives we will cover, as well as suggested reading material and datasets. All the other key information is at the very end of the page to ensure quick access to the schedule – if this is the first time you are looking at this document, please take a look at those sections as well.

Course Plan

Day 1, Monday (02/06/25)

The day of course logistics, discussions over microbial diversity, and a general overview of data-driven strategies to survey environmental microbiomes.

The location for Day 1 is V03 0-D001.

09:00 - 09:30: Course Logistics

Discussion over what will happen throughout the next week and beyond. A great time to take a look at the course syllabus online together:

https://merenlab.org/courses/EMM/

09:30 - 10:00: Installation check

Making sure everyone who wishes to have anvi'o (https://anvio.org) installed on their computers for hands-on exercises have it ready.

10:00 - 12:00: A brief introduction to microbial life on Earth

Please remember, we will have an activity break at 11:00.

- ***Learning Objectives***:
  - Describe the extent of microbial diversity on Earth, their involvement in key biogeochemical processes, as well as human health and disease
  - Remember seminal studies that contributed to our understanding of the diversity, functioning, and metabolic potential of naturally occurring microbial communities and approaches to study them
  - Explain the old and new intellectual and technical challenges that prevent us from defining fundamental units of microbial life, and the art of moving forward without any answers
- ***Suggested Reading***:
  - Gilbert JA, Neufeld JD (2014) Life in a World without Microbes. *PLoS Biology.*
  - Pace NR (2018) The small things can matter. *PLoS Biology.*
  - Falkowski, PG, Fenchel T, Delong EF (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science.*
- ***Even More to Suggested Resources for the Ambitious***:
  - Seeing the Invisible, Op-Docs, The New York Times (a short and cute video on microbial life for a lay audience).
  - Meet Your Microbes, Jonathan Eisen, TED Talk (an easy-to-follow introduction to microbes for a lay audience).
  - How Giant Tube Worms Survive at Hydrothermal Vents, Ed Yong, PBS Digital Studies (Ed Yong is a very successful science journalist, who talks about hydrothermal vent microbes with Colleen Cavanaugh, who made significant contributions to our understanding of microbial symbioses).

  - Overview of how next-generation sequencing works, Eric Chow (a very useful and easy-to-follow lecture on the general principles of sequencing with Illumina, Oxford Nanopore, and PacBio).

13:00 - 15:00: An overview of data-driven strategies to survey environmental microbiomes

- *Learning Objectives*:
  - Recognize currently available 'omics data types (such as metagenomics, and metatranscriptomics), approaches (such as pangenomics, and phylogenomics), and questions they can *and* can not answer
  - Recognize the available computational solutions to gain insights into fundamental questions in microbiology and their brief history
  - Explain the power of **metagenomic read recruitment** and interpret ecological and evolutionary insights we can infer through this strategy
- *Suggested Reading*:
  - Eren AM, Banfield JF (2024). Modern microbiology: Embracing complexity through integration across scales. *Cell.*
  - Franzosa EA, et al (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nature Reviews Microbiology.*

15:00 - 17:00: A read recruitment exercise to warm up

> The purpose of this exercise is to help you have a direct exposure to individual analysis steps and tools that enables one to recruit reads from metagenomes, and profile the read recruitment results to investigate gene distribution patterns of a given population.
>
> Throughout this exercise you will use a mock dataset to (1) familiarize yourself with commonly used file formats such as FASTA, FASTQ, SAM, and BAM, (2) learn the basic steps of read recruitment through Bowtie2 and samtools, (3) learn how to profile read recruitment results using anvi'o, and (4) familiarize yourself with downstream steps of the analysis of recruited reads. Please try to be mindful about individual steps, make notes of those steps that did not make much sense to you so we can discuss them further if we have time.
>
> You will find the exercise here: https://merenlab.org/tutorials/read-recruitment/

Day 2, Tuesday (03/06/25)

The day of genome resolved metagenomics, and pangenomics.

The location for Day 2 is V03 0-D001.

09:00 - 11:00: Genome-resolved metagenomics: opportunities and pitfalls

- *Learning Objectives*:
  - Recognize the difference between microbial isolates, enrichments, single-cell amplified genomes, and metagenome-assembled genomes
  - Explain the importance of the ability to acquire genomic information from microbes we have not yet cultivated
  - Tell the basics of algorithms and strategies to reconstruct microbial genomes directly from metagenomes
  - Appreciate the limitations and opportunities associated with genome-resolved workflows
- *Suggested Reading*:
  - Paoli L, et al (2022). Biosynthetic potential of the global ocean microbiome. *Nature*.
  - Chen LX, et al (2020). Accurate and complete genomes from metagenomes. *Genome Research*.
  - Shaiber A, Eren AM (2019). Composite metagenome-assembled genomes reduce the quality of public genome repositories. *mBio*.
  - Meren and Scott JJ (2020). Visualizing the fate of contigs across metagenomic binning algorithms. *A blog post on merenlab.org*.
- *Even further material to understand assembly*:
  - Assembly is always a difficult topic that we don't cover extensively during our lectures. But here is a VERY short lecture on de Bruijn graphs, and here is a slightly lengthier one. Even watching these, please remember that they are not covering the assembly of shotgun metagenomes. Even though the same principles apply, it is a much more difficult case.

11:00 - 12:00: Open lab

So you can take a look at some of the papers linked above (or below), and/or ask questions.

13:00 - 15:00: Pangenomics: comparative genomics in the era of genomic explosion

- *Learning Objectives*:
  - Explain the concepts of core and accessory genome, as well as open and closed pangenomes
  - Define gene clusters in pangenomes through sequence homology
  - Interpret ecological and evolutionary insights pangenomes offer
- *Suggested Reading*:
  - Tettelin H, et al (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *PNAS*.
  - McInerney MO, et al (2017). Why prokaryotes have pangenomes. *Nature Microbiology*.
  - Zhou Z, et al (2018). Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Current Biology*.

15:00 - 17:00: Pangenomic analysis of a bacterial genus

This is a small exercise with pangenomics. Please download the data pack for this exercise at this Dropbox link.

The data pack contains 15 genomes for you to work with. While each genome belongs to the bacterial genus *Bifidobacterium*, you don't know which species they assign. Please take a look at the anvi'o pangenomics tutorial and/or the pangenomics exercise to find out how to create a pangenome for all these 15 genomes using the program `anvi-pan-genome` with default parameters, and answer the following questions in your short report:

- How many **single-copy core genes** did you find?
- When you organize genomes based on gene cluster frequencies, how many **main groupings of genomes** do you observe?
- Which **'species' name** would you annotate these genomes with?
- According to gene clusters, which two species of *Bifidobacterium* in this mixture are **most closely related**?

Please include a screenshot of your final display you achieved through `anvi-display-pan`, and get cookie points for your pretty displays :)

Some optional questions for the overly enthusiastic:

- What are some of **common features of the genomic islands** that seem to be variable across individual genomes in this pangenome? Tip: you can have quick insights into genomic islands that occur only in some genomes by organizing gene clusters based on enforced synteny per genome.
- What **functions seem to differ between the main groups of genomes**? Tip: you can use functional enrichment analyses to figure out if there are functions that systematically occur in one clade of *Bifidobacterum* but not the other.

Day 3, Wednesday (04/06/25)

The day of genome metapangenomics (building on what we discussed the day before) and phylogenomics.

The location for Day 3 is V03 0-D001 until 14:00, after which we will be at V03 0-C001 until 17:00!

09:00 - 11:00: Metapangenomics: integrated interpretations of pangenomes and metagenomes

- *Learning Objectives*:
  - Explain the emerging opportunities to investigate the functioning and the ecology of microbial populations by linking pangenomes and metagenomes
  - Comprehend the power of characterizing a single genome across metagenomes
- *Suggested Reading*:
  - Delmont TO, Eren AM (2018). Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*.
  - Utter DR, et al (2020). Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biology*.
  - Boeuf D, et al (2021). Metapangenomics reveals depth-dependent shifts in metabolic potential for the ubiquitous marine bacterial SAR324 lineage. *Microbiome*.

11:00 - 14:00: Phylogenomics: inferring evolutionary relationships between microorganisms

- *Learning Objectives*:
  - Identify commonly used genes, statistics, and heuristics to infer phylogenomic relationships across distantly related organisms
  - Recognize historical events that led to the emergence of the current Tree of Life, and why scientists can't even
  - Appreciate technical and theoretical limitations of inferring deep branching patterns confidently
- *Suggested Reading*:
  - Woese CR, Fox GE (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *PNAS*.
  - Hug LA, et al (2016). A new view of the tree of life. Nature Microbiology.
  - Shaiber A, et al (2020). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biology*.
  - Gaïa M, et al (2023). Mirusviruses link herpesviruses to giant viruses. *Nature*.
- *Even More Suggested Reading for the Ambitious*:
  - Spang A, et al (2015). Complex Archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*.
  - Da Cunha V, et al (2017). Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLOS Genetics*.
  - Spang A, et al (2018). Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLOS Genetics*.

14:00 - 17:00: Phylogenomic analysis of a bacterial genus

This is a small exercise in phylogenomics. Please use the same data pack from the pangenomics exercise to complete this one. Since you already have your contigs-db files for the genomes in that data pack, this should be extremely fast for you. But please start early to avoid any last minute challenges :)

To solve this exercise, please apply phylogenomics principles to calculate a tree for the *Bifidobacterium* clade.

You can benefit from the tutorial on anvi'o phylogenomics workflow and see examples on how to get the necessary genes from genomes for phylogenomics. Reconstructing a final tree for these genomes with phylogenomics, and being able to explain why you have made certain choices to generate it, is the answer to this exercise.

Once you are done, please compare your phylogenomic tree to the dendrogram you have obtained from the pangenomic analysis. If you want to get fancy, feel free to include 'additional' *Bifidobacterium* genomes from other species in this genus :)

Day 4, Thursday (05/06/25)

The day of metabolism.

The location for Day 4 is V03 0-D003 until noon, after which we will be at V03 0-C002 until 17:00!

09:00 - 12:00: Inferring microbial metabolism in genomes and metagenomes

- *Learning Objectives*:
  - Recognize the difference between microbial genes, functions, and metabolism.
  - Explain the ways by which microbial metabolism can be recovered from genomes and metagenomes
  - Tell the difference between understanding microbial diversity and understanding metabolic potential in a given environment
- *Suggested Reading*:
  - Watson AR, Füssel J, Veseli I, et al (2023). Metabolic independence drives gut microbial colonization and resilience in health and disease. *Genome Biology.*
  - van Kessel MAHJ, et al (2015). Complete nitrification by a single microorganism. *Nature.*
  - Liu R, et al (2022). Novel Chloroflexi genomes from the deepest ocean reveal metabolic strategies for the adaptation to deep-sea habitats. *Microbiome.*

13:00 - 17:00: Comparative microbial metabolism

This is a small exercise in microbial metabolism analysis. Please find the data pack for this exercise on at this Dropbox link.

The data pack contains four microbial genomes, and your task is to investigate which of these organisms (if any) are capable of nitrogen cycling. Please use anvi'o to annotate these genomes with KOfams, and then run `anvi-estimate-metabolism` to calculate the completeness of metabolic pathways in the KEGG MODULE database. You should examine the output of that program to identify the completeness scores for nitrogen cycling pathways in each genome. You will find a list of all KEGG modules for nitrogen metabolism at this link. This list contains seven pathways for nitrogen fixation, nitrate reduction, denitrification, and nitrification.

Your short report should answer the following questions:

- Which nitrogen metabolism pathways are 'complete' in each genome? Please include in your answer their path-wise completeness scores and the score threshold that you are using (ie, the value of the `--module-completion-threshold` parameter).
- For the nitrifying organisms, which of the two nitrification reactions – the first conversion from ammonia to nitrite, or the second conversion from nitrite to nitrate – can they do? What evidence supports this?
- When you've analyzed all of the genomes, please summarize your findings with a few sentences describing the following points:
- which part(s) of the nitrogen cycle you found to be complete, and which part(s) were missing across all genomes
- which genome(s) were capable of carrying out multiple nitrogen metabolism pathways, and which genome(s) had no nitrogen metabolism capabilities at all
- other observations or hypotheses (if you have any) about these nitrogen cycle pathways, or the enzymes/gene annotations in these pathways, or why these genomes might >have these capabilities or not, etc

And here are some optional things to include in your report, if you have the time or interest :)

- Determine the taxonomic identity of each genome. Does the genome's metabolic capacity match to what you would expect, based on known research about its taxonomic clade?
- Visualize the metabolism estimation results across the four genomes as a heat map, and add a screenshot of the heat map to your report. You can find examples of how to create the heat map in the tutorials linked below (but feel free to use a different way to do it, too)

You might find some of the resources below helpful as you do this exercise:

- A recent tutorial on metabolism estimation in anvi'o
- Documentation for anvi-estimate-metabolism
- An older (and much simpler) tutorial on metabolism estimation

Day 5, Friday (06/06/25)

The day of genome microbial population genetics.

The location for Day 5 is V03 0-C002.

09:00 - 10:30 Microbial population genetics: tools, terminology, and open questions

- *Learning Objectives*:
  - Learn ecological and evolutionary implications of clonality and heterogeneity within environmental populations
  - Identify approaches to study single-nucleotide variants, and methods to reconstruct *haplotypes*
  - Comprehend differences and overlaps between population genetics approaches in animal populations and microbial populations
  - Characterize variation within a metagenomic sample and make use of it for exploratory analyses or hypothesis testing.
- *Suggested Reading*:
  - Simmons SL and Dibartolo G, et al (2008). Population genomic analysis of strain variation in Leptospirillum group II Bacteria involved in acid mine drainage formation. *PLOS Biology.*
  - Denef VJ (2018). Peering into the genetic make up of natural microbial populations using metagenomics. *Springer Publishing.*
  - Delmont TO, et al (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife.*

10:30 - 12:00 :: Structure-informed interpretations of microbial population genetics

- *Learning Objectives*:
  - Learn about the new generation of computational strategies to predict protein structures from sequences
  - Comprehend the implications of structure-informed interpretations of genomic variation in our ability to determine targets of distinct evolutionary processes
- *Suggested Reading*:
  - Jumper J, et al (2021). Highly accurate protein structure prediction with AlphaFold. *Nature.*
  - AlQuraishi M (2021). Protein-structure prediction revolutionized. *Nature News and Views.*
  - Robinson SL (2023). Structure-guided metagenome mining to tap microbial functional diversity. *Current Opinion in Microbiology.*
  - Kiefl E, et al (2023). Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. *Science Advances.*

13:00 - 15:00: Population genetics of a cryptic plasmid

This is a small exercise on microbial population genetics. The exercise aims to help you familiarize yourself with the population genetic signal recovered from metagenomes through single nucleotide variants, and sharpen your ability to answer some key questions using such data. You can download the data pack from here, in which you will find an anvi'o profile database and a contigs database that contains all the data you will need to be able to solve the following puzzle.

The contigs database is generated from a single plasmid, and the merged profile database contains the metagenomic read recruitment data that puts this plasmid in the context of 12 human gut metagenomes. The gut metagenomes are a subset of the data published in this study in case you are interested to take a look. But briefly, the subset of the data that is profiled here includes 6 gut metagenomes from mothers, and 6 gut metagenomes from their infants. But you don't know the real infant-mother pairs :)

Your task is to investigate single-nucleotide variants (SNVs) found in read recruitment results to and answer the following questions:

- As far as this dataset goes, would one argue that the plasmid is acquired from random sources upon birth, or is there evidence to suggest it is vertically transmitted from mothers to infants?
- If it is vertically transferred, can one identify mother infant pairs confidently?

To answer these questions you can get inspiration from strategies mentioned in this tutorial. If you want a refresher on SNVs, you may want to take a look at this blog post.

You can (and should) inspect the coverage plots for all of the mothers and infants (using the program `anvi-interactive`), but if you determine that the plasmid is vertically transmitted and you think you can identify mother-infant pairs, you are invited to create a final figure that summarizes the evidence for it.

If you believe there is signal to determine the answer for it, please try to figure out which mother matches which infant and be prepared to prove your conclusion!

15:00 - 17:00: Open lab

Discussions, revisiting old topics, and preparations for the next week.

Faculty and Communication

Lectures and exercises during this week will be led by Prof. Dr. A. Murat Eren (Meren) and Prof. Dr. Sarahi Garcia. While the lectures in this week will be primarily delivered by Meren, additional experts will take part in the design and/or delivery of various sections, and can be reached out to for questions. The following table lists individuals who will be involved in the first week of the course, and their contact information:

Name	Role	Expertise	Contact information
Meren	Professor	Microbial Ecology, Computer Science	meren@hifmb.de
Sarahi	Professor	Microbiology, Microbial Ecology	sarahi.garcia@uni-oldenburg.de
Iva Veseli	Postdoc	Microbial Ecology, Computer Science	iva.veseli@hifmb.de
Jessika Füssel	Postdoc	Microbial Metabolism, Biogeochemistry	jessika.fuessel@uol.de
Florian Trigodet	Postdoc	Microbiology, Bioinformatics	florian.trigodet@hifmb.de

Throughout the course (and beyond) you can reach out via email with any question to Meren, who should be your first contact for anything related to the course activities unless specified otherwise.

Description and Learning Objectives

Every ecological niche our planet has to offer is home to an astonishing number of microbial cells that form complex communities. The last several years witnessed tremendous advances in molecular and computational approaches which now offer unprecedented access to these communities through new 'omics strategies. Developing an overall understanding of these strategies, including the ability to identify their appropriate applications and shortcomings, has quietly become a de facto necessity in the journey of an independent life scientist. **The primary aim of the first week** is to help you appreciate the basics of commonly used 'omics strategies to study ecology, evolution, and functioning of naturally occurring microbial populations and recognize the current **conceptual framework** that helps us wrap our collective mind around the most diverse form of life on our planet.

Throughout our sessions you wil hear about the theoretical underpinnings of popular 'omics data types and their contemporary applications. These data types include **genomics**, **metagenomics**, **metatranscriptomics**, **metaepitranscriptomics**, as well as various 'omics data analysis approaches such as **metabolic reconstruction** in genomes and metagenomes, **metagenomic read recruitment**, **pangenomics**, **phylogenomics**, and **microbial population genetics**.

The learning objectives of the course includes the following:

- To explain **microbial diversity** in naturally occurring systems and their **evolution**.
- To recognize data-enabled means to study **microbial ecosystems**.
- To introduce **state-of-the-art 'omics approaches** and data types to characterize naturally occurring microbial diversity.
- To improve discussion, analytical, presentation and writing skills.

Prerequisites

To maximize benefit, the participants of this course are expected to be familiar with the central dogma of molecular biology, and able to answer what is a **gene**, a **genome**, a **transcript**, or a **protein**, and

have at least a **preliminary understanding of the principles in ecology and evolution**, such as the basics of **taxonomy** and broad ecological principles that maintain complex ecosystems.

You are also expected to be familiar with the the UNIX shell (also known as the 'terminal environment', or 'command line interface'), especially if you are interested in following the hands-on exercises by yourselves (if you are not familiar with the terminal environment, you can team up with those who do, and follow them and contribute to data analyses and interpretations). You can use some of the following material to familiarize yourself with the command line interface, and Meren will be happy to help you with any questions:

- Beginner's Guide to the Bash Terminal (a video introduction to the Linux command line environment – although Joe Collins is talking about Linux, the topics are relevant to anyone who uses a command line environment and Meren strongly recommends everyone to watch this in its entirety, and try to replicate commands).
- Learning the Shell (a chapter from the open book "*The Linux Command Line*" by William Shotts – Meren highly recommends).

The course will require its participants to read and understand contemporary literature written in English.

Content Delivery

The primary mode of course content delivery will be through slides, where Meren will explain **core concepts**, **data types**, and **analysis strategies**. There will often be extensive discussions over these slides, which will require **active, verbal participation** by the attendees.

This document will provide attendees with **suggested readings** related to each topic from the recent literature that covers relevant topics and/or their real-world applications to contemporary questions in marine microbiology, oceanography, and beyond.

Please note that **preparation** and **participation** will play a key role in your success. For an effective learning experience please consider taking a brief look at the suggested reading material whenever you have a chance and participate the lecture actively by asking questions and attending discussions.

Attendance Policy

Each participant is expected to attend each lecture in person. The attendance will be recorded by a strategy that we call **class citizenship**, which aims to help the course director to have an overall understanding of the evolution of the course.

The class citizenship demands every participant to send a **class citizenship** email to meren@hifmb.de **AND** sarahi.garcia@uni-oldenburg.de at the end of **each day**. The class citizenship email must be composed of two parts:

- A **brief summary** of the main concepts discussed during the day, interpreted by the attendee in their own words.
- One or more **short questions** that is/are relevant to concepts or ideas discussed throughout the day, yet remained unclear.

The last 15 minutes of every day will be dedicated to class citizenship emails, therefore the attendees will end their day without having to remember doing it later.

The title of the class citizenship email must follow this pattern **word-by-word**:

> **EMM Class Citizenship: DD/MM/YY**

For instance, the following would be the appropriate title for this email for the first day:

> **EMM Class Citizenship: 10/06/24**

The best class citizenship emails are those that are brief, genuine, and insightful. In an ideal world the emails should be no less than 50 words, and no more than 250 words. Please do not send notes you take throughout the class. You should use the last 15 minutes of the lecture to gather your thoughts, and come up with a summary of what you can remember. Here is an example class citizenship email:

> Summary: Today we discussed what is phylogenomics, how phylogenomic trees are built, and why single-copy core genes are suitable for building phylogenomics trees. We also discussed the relationship between phylogenetics, phylogenomics, and pangenomics with respect to the fraction of genome used and the evolutionary distance that they can cover.
>
> Question: Since phylogenomics and pangenomics are both useful for inferring evolutionary distances, it seems to me that integrating both methods in a systematic way would yield a more reliable tree. But it looks like the field only uses phylogenomics and pangenomics separately, is there a reason for that?