Applied Microbial 'Omics

Course Plan

October, 2024

Contents

Preface

The purpose of this document is to share the details of the course "**Applied Microbial 'Omics**". In the following sections you will find the course description, learning objectives, plan, schedule, attendance and grading policies, as well as other key information that is critical for the course attendees to consider.

Course Details

This course is a part of the module "**Applied Molecular Ecology**" (**mar260**) taught by Dr. Iva Veseli and Prof. Dr. Iliana Baums.

Course Details	
Name	**Applied Microbial 'Omics**
Meeting Location	**W04 1-171**
Number	5.12.263
Type	Seminar/Exercise
Credits	1
Language	English

In addition to this course, the module **mar260** contains three additional components that each participant of this course is expected to also attend,

- **Introduction to Popular 'Omics Strategies** (5.12.262, Lecture, Iva, Course Plan: HTML, PDF)
- **Coastal Conservation in the 'Omics Age** (5.12.260, Lecture, Iliana Baums)
- **Readings/Exercises in Coastal Conservation** (5.12.261, Seminar, Iliana Baums)

Please familiarize yourself with the details of the remaining courses using the appropriate documentation provided for the other components.

Faculty and Communication

Exercises throughout **Applied Microbial 'Omics** will be primarily delivered by Iva. However, additional experts will take part in the design and/or delivery of various sections. The following table lists individuals who will be involved in the course, and their contact information:

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

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

Iva's office is **Room 2118** at the Helmholtz Institute for Functional Marine Biodiversity. If you need to meet with her here in person, please schedule an appointment by email first.

Description and Learning Objectives

Generating hundreds of millions of sequences or tens of thousands of genomes to study naturally occurring microbial communities has become commonplace for many microbiologists. The ability to benefit from the ongoing data revolution demands the new generation of microbiologists to be familiar with the arsenal of 'omics tools that enable in-depth investigations of the new data streams that offer detailed snapshots of the microbial lifestyles. **The primary aim of this course** is to help its participants gain **hands-on experience** with some of the most popular data types and approaches in microbial 'omics and put some of the strategies detailed in the lecture **Introduction to Popular 'Omics Strategies** into practice.

Applied Microbial 'Omics is a *seminar* that is designed to introduce its participants to 'omics data analyses to answer real-world questions with often simplified datasets. Participants will learn about the practical aspects of working with popular 'omics data types and their contemporary applications. The 'omics data types and strategies that will be covered throughout the semester include **genomics**, **metagenomics**, **metagenomic read recruitment**, **metabolic reconstruction**, **pangenomics**, **phylogenomics**, and **microbial population genetics**.

The learning objectives of the course includes the following:

- To gain experience in the **UNIX shell** and its utility in working with open-source software and large datasets.
- To apply **state-of-the-art 'omics approaches** to various data types to make sense of complex datasets.

- To communicate data analyses through **reproducible bioinformatics workflows**, figures, and written reports.

Prerequisites

Throughout the course we will use anvi'o for 'omics analyses. Anvi'o is an open-source software platform that brings together many aspects of today's cutting-edge computational strategies of data-enabled microbiology, including genomics, metagenomics, metatranscriptomics, pangenomics, meta-pangenomics, phylogenomics, and microbial population genetics in an integrated and easy-to-use fashion through extensive interactive visualization capabilities. Anvi'o is cited over 1,000 times in the literature, and is actively maintained. The latest release of anvi'o is v8 (marie). While it is not a prerequisite, it will be most beneficial if the participants have access to personal computers with which they can work on the course exercises, preferably laptops that they can bring to the classroom.

Most of our work in this course will take place in the UNIX shell (also known as the 'terminal environment', or 'command line interface'). If you have no prior experience with the command line interface, that is OK, as you will generate those skills throughout the course - we will go through the basics together in class, and everything beyond the basics can be picked up along the way. Arguably, the exposure to the command line environment and developing a level of mastery of it will be one of the most impactful gains you will have from this course that will help you throughout your professional journey almost regardless of which career path you choose that involves data; so if you are not familiar with the command line environment, see this as an opportunity to invest time into developing some skills in it. You can use some of the following material to familiarize yourself with the command line interface, and Iva 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. It is strongly recommended to watch this in its entirety, and try to replicate the demonstrated commands).
- Learning the Shell (a chapter from the open book "*The Linux Command Line*" by William Shotts).
- The Unix Shell (an online lesson on basic BASH skills designed by the Software Carpentry non-profit. You can download the data pack and follow along with the commands.)

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 data analysis assignments and discussions during the class time related to their implementation. The vast majority of activities will be informed by the **core concepts**, **data types**, and **analysis strategies** explained in the companion lecture **Introduction to Popular 'Omics Strategies (ITPOS)**. There will often be extensive discussions over best analysis practices, and the observations we can make through our analyses, to which participants will be encouraged to show **active, verbal participation**.

Whenever possible, the theoretical lecture in ITPOS corresponding to the current analysis exercise will be given the same week (on Monday) that we meet to discuss the exercise (on Wednesday). The analysis report for the exercise will be due **1 week later** on Wednesday. For example, if we discuss pangenomics in ITPOS on Monday 18/11/24, then we will work on the pangenomics exercise in this class on Wednesday 20/11/24. Ideally, this setup will ensure that the theoretical background is fresh when you start the analysis, any initial questions or hiccups can be quickly cleared up in our Wednesday discussion, and there is sufficient time to finish the analysis outside of class before it is due.

Please note that **preparation** and **participation** will play a key role in your success as usual. For an effective learning experience please consider (1) taking a brief look at the new analysis assignment (~30 minutes of study) on the Monday before class, (2) take time early on to try to complete the assignment by yourself (~2-4 hours of study), and (3) ask questions to Iva or others before it is too late (using

Wednesday's discussion time wisely). Analysis tasks and data will be available on Stud.IP two weeks before the next session.

Attendance Policy

Each participant is expected to attend each lecture in person (unless a legitimate reason for absence that is recognized by the University is in effect). The attendance will be recorded by analysis reports.

Analysis Report Submission Guidelines

Please follow these guidelines carefully to format your reports. You can submit your report by email to iva.veseli@hifmb.de before the deadline. Each analysis report should be submitted as Markdown file as well as a PDF document in a single compressed directory. The directory structure and file names should look like this:

```
Eren_AM_EX_01/
├── Eren_AM.md
├── Eren_AM.pdf
├── files/
│   ├── auxiliary-data.db
│   ├── contigs.db
│   └── profile.db
└── images/
    ├── Figure_01.png
    ├── Figure_02.png
    └── Figure_03.png
```

where,

- `Eren_AM` is the 'last name' and 'initials' of the person who prepared the report.
- `EX_01` is the label shared for the exercise in the course plan.
- `Eren_AM.md` is the flat-text report file formatted in *markdown*. Here you can find a syntax guide for markdown, and MacDown is an open-source markdown editor with a WYSIWYG editor, which makes life easier.
- `Eren_AM.pdf` is a PDF document generated from `Eren_AM.md` using pandoc. Once you have your markdown file, you can easily convert it to a PDF document using the following command:

```
pandoc Eren_AM.md --toc --pdf-engine=xelatex -o Eren_AM.pdf
```

- `files` directory can include files necessary to share along with the report (they should be the absolute minimum set of files that are necessary to reproduce your nicely curated analysis workflow).
- and `images` directory is to keep images cited from the report.

Once the directory is ready, you can run the following command to compress it into a single archive file and submit the new file:

```
tar -zcf Eren_AM_EX_01.tar.gz Eren_AM_EX_01/
```

Please also consider the following formatting guidelines for your reports:

- Make sure your report is a stand-alone text, so that anyone who reads the report can understand its purpose without having to read the assignment itself. For instance, this could mean writing a little introduction that summarizes the context of your analysis.
- Use headers and sub-headers to separate sections from each other – make your report pretty and nicely formatted. For each section, briefly explain your analysis strategy and reasoning behind it.
- Use code blocks to separate command lines from free text (for which here is a markdown syntax).
- Make your analyses reproducible. So if someone reads your report and runs every command line one by one, they should be able to reproduce your analysis if they are in a directory that contains the original input data.

- Screenshots of key observations are encouraged :) Make sure you have figures to communicate your observations clearly.

Course Plan

Please note that each lecture takes place on **Wednesdays** (bi-weekly), between **10:15 - 12:00**, at **W04 1-171**.

23/10/24 :: Introduction to anvi'o and installation check

The primary purpose of this session is to **discuss the course format**, future exercises, and how to return reports.

We will also **discuss anvi'o and its features**, and will make sure everyone has a **working copy of anvi'o installed** on their computers. Towards the end of this first session those who have attempted yet not been able to install the platform will receive hands-on help. Please try to install the development version of anvi'o on your computers *before* you come to this exercise session. You will find the installation instructions on https://anvio.org/install

- *Suggested Reading*:
  - Eren AM, et al (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology*.

06/11/24 :: EX 01: A read recruitment exercise to warm up

Please try to accomplish this exercise by midnight on **13/11/23**. You will not need to return a report for this particular week.

> 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 (essentials of which are covered in the 04/11/24 lecture in the companion course), 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 that we can discuss them during class.
>
> You will find the exercise here: https://merenlab.org/tutorials/read-recruitment/

20/11/24 :: EX 02: Pangenomic analysis of a bacterial genus

Please read the assignment below carefully, and **return your reports by midnight on 27/11/24**:

> This is a small exercise with pangenomics. Please find the data pack for this exercise on stud.IP, or using this Dropbox link.
>
> This 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 figures :)

Some optional questions for the extra 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 enrichemnt analyses to figure out if there are functions that systematically occur in one clade of Bifidobacterum but not the other.

04/12/24 :: EX 03: Phylogenomic analysis of a bacterial genus

Please read the assignment below carefully, and **return your reports by midnight on 11/12/24**:

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 :)

18/12/24 :: EX 04: Comparative microbial metabolism

Warning! This exercise has an off-pattern schedule and is due earlier than usual. To avoid having an assignment due during the winter holiday, this exercise will be due on Friday 20/12/24, *two days after* our session on 18/12/24. To accommodate this, the corresponding theoretical lecture for this assignment will take place on 09/12/24, one week earlier than usual. Please get started on the assignment as early as you can.

Please read the assignment below carefully, and return your reports by midnight on **Friday, 20/12/24** (before the winter break):

This is a small exercise in microbial metabolism analysis. Please find the data pack for this exercise on stud.IP, or using 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 pathwise 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 heatmap, and add a screenshot of the heatmap to your report. You can find examples of how to create the heatmap 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

15/01/25 :: EX 05: Population genetics of a plasmid

Please read the assignment below carefully, and **return your reports by midnight on 22/01/25**:

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 avariants, and sharpen your ability to answer some key questions using such data. You can download the datapack 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 mother-infant 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 the 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.

Please make sure your report includes,

- A list that describes which mother matches which infant, if there is signal to determine that.
- A screenshot of your final visualization with a brief description of how you interpreted it.
- And a summary of your workflow with commands you have used.

Thank you and good luck!

29/01/25 :: EX 06: Proposal Discussion

During this session we will overview everything we have covered and discuss how you integrate your learnings into your final proposal for the module.

Evaluation and Grading

The evaluation of the attendee performance in this course (along with all the other three in the module "**Applied Molecular Ecology**" (**mar260**) will be based on two items to be returned by each attendee individually: (1) a **research pre-proposal** (which will provide the basis for the full proposals due at the end of class) and a final **research proposal**. Please see the Evaluation and Grading section for ITPOS for full details.

Examination Policy

Please find all relevant university policies here: https://uol.de/studiengang/pruefungen/umweltwissenschaften-fach-bachelor-136

Academic Integrity

All University policies regarding academic integrity, ethics and honorable behavior apply to this course. Academic integrity is the pursuit of scholarly activity free from fraud and deception and is an educational objective of this class. Academic dishonesty includes, but is not limited to, cheating, plagiarizing, fabricating of information or citations, facilitating acts of academic dishonesty by others, having unauthorized possession of examinations, submitting work of another person or work previously used without informing the instructor, or tampering with the academic work of other students. For any material or ideas obtained from other sources, such as the course reading materials or things you see on the web, in the library, etc., a source reference must be given. Direct quotes from any source must be identified as such.

Disability Access Statement

UOL welcomes students with disabilities and students with care obligations for their children or close relatives into the University's educational programs. In order to receive consideration for reasonable accommodations, you must contact the *Prüfungsausschuss*. Please let Iva and/or Iliana know at the beginning of the semester what accommodations were approved for you.