NFDI4MICROBIOTA Workshop

Unlocking the Power of Omics in Anaerobic Microbiology

25 September 2025 afternoon (13:00 – 17:00 incl. coffee breaks)

Agenda:

Experimental design to include (multi-)omics data

Goal: Understand that omics isn't an afterthought but must be integrated into experimental design

- Defining clear biological questions first (hypothesis- or discovery-driven)
- Sample numbers, replicates, randomization, and batch effects (especially crucial for machine learning downstream)
- Sample collection, storage, and DNA/RNA/protein extraction protocols
- Budget and time: Multi-omics increase not just sequencing costs, but metadata tracking and analysis complexity
- Depth vs Breadth — when to go deep (e.g., full genome reconstructions) vs broad (e.g., surveying community diversity with 16S/ITS)
- Legal and ethical aspects of data if human or environmental samples are involved (Nagoya Protocol)

Selecting the omics - focusing on hypothesis-driven research

Goal: Make clear that you don't have to do "all the omics" — choose wisely

- Different omics answer different questions: metagenomics, metatranscriptomics, metaproteomics, metabolomics
- Matching the omics layer to the hypothesis
- Combinations are powerful but complex: how to integrate multi-layer data
- Practical reality check: Each extra omics layer exponentially increases computational and statistical challenges

Machine-learning and omics experiments: What is the fuss?

Goal: Demystify machine learning (ML) while being honest about its challenges

- Omics data is high-dimensional, low-sample — ML techniques like random forests, SVMs, and more recently deep learning are better at handling this than traditional stats
- Where ML can help: Predicting sample traits (e.g., disease state, environmental conditions), Feature selection (identifying key genes/proteins/metabolites), Pattern discovery (unsupervised clustering)
- Important: Without proper experimental design (balanced data, proper controls), ML models overfit and give misleading results
- Tools: scikit-learn (Python), caret (R), TensorFlow/Keras for more ambitious applications
- Recent advances in interpretable AI (e.g., SHAP values) to make ML models less of a "black box"

The first step to make my data useful: (meta)data submission to public repositories

Goal: Instill the mindset that FAIR data (Findable, Accessible, Interoperable, Reusable) is no longer optional

- Importance of metadata: without good metadata (sampling date, conditions, sequencing platform, protocols) even the best omics data is nearly useless
- Main repositories:
    - Genomics/metagenomics: NCBI SRA, ENA, MG-RAST
    - Transcriptomics: GEO, ArrayExpress
    - Proteomics: PRIDE
    - Metabolomics: MetaboLights
- Submission tools: SRA Toolkit, ENA Webin-CLI, etc.
- Standards: MIxS (Minimum Information about any (x) Sequence), MIMARKS, MINSEQE
- Real-world examples: How sloppy metadata delayed discoveries during major projects (e.g., Tara Oceans, Human Microbiome Project)