UNIVERSITÄT BAYREUTH

Bayreuther Zentrum für Ökologie und Umweltforschung
Bayceer

SCRIPT

Practical course

FUNCTIONAL MICROBIOME RESEARCH

M.Sc. Students of Bio- and Life Sciences

Supervisors:

Tillmann Lüders, Daniel Thomas, Anita Gössner, Ralf Mertel

Practical course and seminars take place in the labs of the
Chair of Ecological Microbiology (ÖMIK), Dr.-Hans-Frisch-Str. 1-3

Time frame:
Mo. 06. – Fr. 17. March. 2023, ca. 9:00 – 16:00

To do in advance:

Please print this Script (A4) and bring it along on every day of the practical course.

Bring along your lab coat and laptop. Prepare for each day by reading the respective sections.

Prepare and practice your literature seminars

Register for an Elixir account (https://elixir-europe.org/)

CONTENT

1. COMPONENTS OF THE PRACTICAL COURSE

A. Stable Isotope Probing of methanol-, glucose- and acetate-oxidizing microbes in soil

B. FISH-Microscopy of sewage sludge and biofilm samples

C. Analysis and interpretation of (meta)genomic and amplicon sequencing data

2. LITERATURE RECOMMENDATIONS

The text books recommended here can help you in preparing for the course and while writing the reports. Support in access to primary literature can be provided by members of the ÖMIK team.

Clark, Stahl, Martinko, Madigan (Eds.) **Brock-Mikrobiologie.** Pearson, 2013

Dumont, Hernández García (Eds.) **Stable Isotope Probing - Methods and Protocols.** Humana Press, New York, NY. 2019. https://doi.org/10.1007/978-1-4939-9721-3

Charles, Liles, Sessitsch (Eds.) **Functional Metagenomics: Tools and Applications.** Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-61510-3

More literature and references will be provided during the seminars!

3. LAB SECURITY

Each student must participate in the **lab security introduction** at the beginning of the practical course. This will be held by the responsible persons of the ÖMIK Chair on **Monday, Mar. 6. at 9:00**. The participation in the introduction must be confirmed by every student by signing the respective security certificate. **Participation is obligatory!**

Do not forget to bring your own lab coat!

Microbes can be pathogenic! They can cause infectious diseases and produce toxins. Each microbial sample and microbial culture handled during the practical course, as long as not certified to be pathogen-free, must be handled as if containing potential pathogens! The following basic security measures are to be implemented:

- Wear you lab coat at all times during the practical course.
- If advised, eye protection must also be worn.
- Long hair are to be worn in a plait, do not carry long pendulous jewellery.
- Shoes must be solid and closed.
- Eating, drinking, smoking, chewing gum, putting on make-up etc. in the labs are not allowed!
- Utmost cleanliness is to be maintained in all labs and on all working materials.
- Apply techniques of sterile sample handling wherever necessary
- Apply mechanical pipetting wherever possible
- Any accidents and injuries must immediately be reported to the course supervisors!

4. REQUIRED PERFORMANCE, DELIVERABLES AND EVALUATION

A successful participation in the entire module requires:

1. Regular participation: It is required that you arrive on time and participate continuously for all parts of the practical course. Upon previous agreement by the supervisory staff, and if urgent justifications are warranted, it is possible to allow for not more than **one day of absence** during course duration. One second day of absence is possible, but must be justified via a doctoral attestation. Absence for more than two days does no longer warrant the successful completion of the practical course.

2. Final exam: Please register for the final written exam. The examination includes all topics taught during the lecture “Functional Microbiome Research”, during the practical course, during the seminars and presented in the script. Date and time of the final written exam is:
Wed. 29.03.2023, 10:00 – 11:30 in H7 (DHF-Str.)

5. WRITTEN REPORT

Each student must submit a written report for the entire practical course. Rationale, experimental procedures and results of all elements of the practical course must be documented in a contiguous manner. The aim is to provide training in putting together a written scientific report. The report should be structured in the following manner:

1. Introduction (scientific background and research questions)
2. Methods (can be kept brief, as long as not significantly extending beyond the procedures described in the script).
3. Results: all steps and results obtained during the course must be put documented. Pay attention to sample naming, proper illustration of axes of graphs and units. Figures (with legends) must be numbered continuously, same as tables (with headers).
4. Discussion of the results and conclusions. Refer to the literature where adequate and avoid simply repeating the results.
5. References

Further, more specific instructions for reporting on the metagenomics part of the course will be given below. The supervisory team can be approached at any time during the course and while preparing the written reports for additional advice on the reporting and illustrations. A joint discussion and presentation of the results of the course will be held on the last day of the practical course.

The written report must be submitted as printed versions. Accompanying E-mails with additional Excel tables etc. containing original data, calibrations, or more extensive metagenomic displays may also be submitted.

Submit your reports latest by Mo., April 17th 2023 in print to the ÖMIK secretary Melanie Nützel or directly to Prof. Lüders.

6. TIME PLAN – FUNCTIONAL MICROBIOME RESEARCH

Timing is always somewhat flexible, updates and changes are possible and will be communicated during the course.
Seminar presentations should be prepared and practiced before the start of the course.

Mo. 6.3.	**Tue. 7.3.**	**Wed. 8.3.**	**Thu. 9.3.**	**Fri. 10.3.**
9^{00} - Welcome and security introduction 10^{00} - Introduction to the SIP experiment (TL) 11^{00} – Setting up of SIP microcosms (TL) 12^{30} – Lunch break 13^{30} – Getting started: Linux and deNBI cloud operations (DT)	9^{00} – Introduction to Metagenomics, read QC and MG assembly (DT) 12^{00} – Lunch 13^{00} –CO_2 sampling of SIP microcosms (TL) 13^{30} – Student seminars (Arnold / Hertel) 14^{30} – Mock community genome handling (DT)	9^{00} - DNA extraction from SIP microcosms, CO_2 quantification (TL) 12^{00} – Lunch 13^{00} – Mock community MG assembly check (DT) 14^{00} – DNA purification contd., loading of SIP gradients and start of ultracentrifugation (TL)	9^{00} – IRMS analytics of $^{13}CO_2$ (GEO I, Campus) 11^{00} – Mock community binning and refining (DT) 12^{30} – Lunch 13^{30} – Student seminars (Hausmann / Merz) 14^{30} – Mock MG taxonomy and pathway annotation (DT)	9^{00} – SIP gradient fractionation (TL) 12^{00} – Lunch 13^{00} – preparation of Illumina sequencing amplicons (TL) 15^{00} – Biofilm fixation for FISH (TL)
Mo. 13.3.	**Tue. 14.3.**	**Wed. 15.3.**	**Thu. 16.3.**	**Fri. 17.3.**
9^{00} – 16S amplicon data handling (DT) 11^{30} – Sample preparation for FISH (TL) 12^{00} – Lunch 13^{00} – FISH Hybridize (TL) 14^{00} – 16S amplicon data cont. (DT) 16^{00} – FISH washing (TL)	9^{00} – SIP amplicon label interpretation (TL) 10^{00} FISH microscopy (TL) 11^{00} – Student seminars (Scharrer / Schneider) 12^{00} – Lunch 13^{00} – Start of sewage metagenome workshop (DT)	9^{00} – ^{13}C flux calculations (TL) 11^{00} – Sewage metagenome contd. (DT) 12^{00} – Lunch 13^{00} - FISH microscopy contd. (TL) 14^{00} – Metagenome workshop contd. (DT)	9^{00} – Sewage MG workshop contd. (DT) 12^{00} – Lunch 13^{00} - FISH microscopy contd. (TL) 14^{00} – Sewage MG workshop contd. (DT)	9^{00} – Final joint data interpretation and discussion (all) 11^{00} – Joint Q&A round Mock exam questions (all) 12:00 – Farewell lunch and course wrap-up

PART A: STABLE ISOTOPE PROBING

A.1 THEORETICAL BACKGROUNDS

Stable isotope probing (SIP) of isotopically labelled RNA and DNA is well-established in environmental microbiology. The concept of a labelling-based detection of process-relevant microbes independent of cellular replication or growth allows for a direct handle on functionally relevant microbiome components. Even over a decade after its introduction, stable isotope probing (SIP) of nucleic acids is still considered as a prime strategy for the targeted identification of microbial key-players in defined environmental processes. Since its conception for the identification of methylo- and methanotrophic bacteria [4], nucleic acid-based SIP has undergone a considerable development and has been applied to a variety of ^{13}C-, ^{15}N- and recently also ^{18}O-labeled compounds. The general strengths of SIP are that it allows for a substrate-based query, for a targeted identification of microbes involved in a specific biodegradation process within a complex community, as well as for the unraveling of involved catabolic genes and carbon flows across microbial kingdoms. All of this is possible in an undirected manner, without essential à-priori probes or genomic information. The basic approach is also a relatively "low-tech" method, which discriminates it from most other technologies capable of tracing isotopic labeling in microbes.

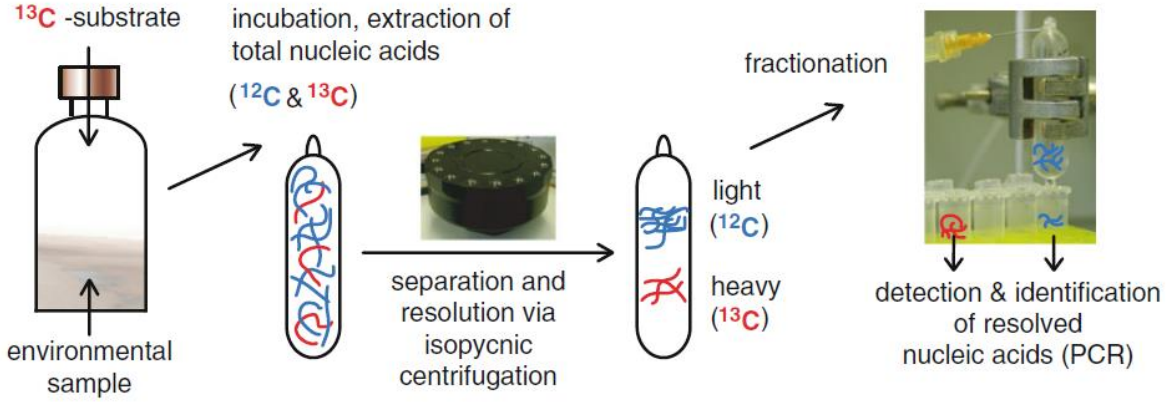


Fig. 1 Schematic view of a typical stable isotope probing (SIP) experiment to identify degraders of ^{13}C-labeled hydrocarbons

A.1.2 TECHNICAL CONSIDERATIONS

SIP uses ultracentrifugation to resolve isotopically labeled ("heavy") nucleic acids from unlabeled ("light") ones. In this isopycnic centrifugation, freely diffusing cesium salts form a density gradient driven by the delta between minimal and maximal centrifugation force (g_{min} & g_{max}) during the run. Nucleic acids loaded into the gradient then arrange themselves (or "band"), also by diffusion, at the buoyant density (BD) matching their own. CsCl media are commonly used for centrifugation of genomic DNA, which bands at a BD of ~1.70 g ml^{-1} in unlabeled form.

It is not possible to band rRNA in CsCl, as the salt would precipitate during centrifugation at the required BD (>1.9 g ml^{-1}). Instead, rRNA is centrifuged in CsTFA, where it bands at ~1.80 g ml^{-1} in unlabeled form. For both DNA and rRNA, fully ^{13}C-labeled nucleic acids have been observed to be ~0.04 g ml^{-1} "heavier" than their unlabeled counterparts. The maximal effect of full ^{15}N-labelling on nucleic acid BD has been reported to be roughly half of that, and thus less pronounced than the

distinction in BD between a low-GC and a high-GC genome. When conducting isopycnic centrifugation, the reader should be aware of the following principles:

(i) SIP centrifugation runs are very long. As already demonstrated much earlier, genomic DNA requires >24 h of ultracentrifugation for optimal banding. Focusing of bands has been observed to be even less efficient for rRNA, a molecule characterized by extensive secondary structure and self-affinity. Thus, long centrifugation runs of >36 h for DNA, and of 42 – 65 h for rRNA are generally recommended. These long centrifugation times are not detrimental to rRNA, because rRNA is protected during centrifugation by the chaotropic nature of CsTFA.

(ii) Banding is never absolute. Because banding of nucleic acids in SIP is a diffusion process, it never reaches an absolute stage. Even if uniformly labeled pure culture nucleic acids are centrifuged, heavy fractions will always contain backgrounds of light nucleic acids, and vice versa. Thus, the quantitative allocation of specific templates to a certain BD is the most crucial information obtainable from SIP gradients, and not the absolute detectability of certain templates at a given BD. Due to diffusion banding, neighboring gradient fractions will always harbor similar templates. Even if distinctions in template distribution are apparent over entire gradients, they will always occur gradually, over 3-4 neighboring fractions.

(iii) The rotor controls gradient resolution. Because of the small differences in BD between unlabeled and fully ^{13}C-labeled nucleic acids (~0.04 g ml^{-1}), very shallow gradients are required for optimal spatial resolution of light and heavy templates. The steepness of a centrifugation gradient is controlled mainly by the difference between effective g_{min} and g_{max} acting on a gradient, and thus by the difference between the inner and outer radius (r) of centrifugation. For SIP, this difference is optimally as small as possible. In essence, any large vertical, near-vertical or small ("table-top") fixed-angle rotor will produce gradients shallow enough for a good resolution of ^{12}C- and ^{13}C-nucleic acids. In contrast, although they have been occasionally used in the literature, large fixed-angle rotors generate much steeper gradients, where light and heavy nucleic acids band much more close to each other, with an unsatisfactory number of fractions to be resolved in-between. Even more so, SIP is technically impossible in classical swing-out rotors.

A.1.3 Strategies to substantiate label incorporation

As argued above, it is evident that the detection of a given template in a heavy gradient fraction alone does not substantiate label incorporation. For this, it is essential to show that a given template in heavy fractions is absent or less abundant in light fractions of the same gradient, that the appearance of heavy templates in gradients becomes visible over time (coupled to the consumption of the ^{13}C-substrate), and that this appearance is not observed to the same extent in gradients of unlabeled ^{12}C-control treatments. This evidence can be substantiated as follows:

(i) Entire gradients should be evaluated. Although early SIP studies relied on the use of ethidium bromide to visualize bands of unlabeled and labeled nucleic acids, it is now accepted that the fractionation of entire gradients offers a superior means to access the full information of density-resolved nucleic acids in SIP. Especially for intermediate ratios of label incorporation into given populations, either directly or via metabolic cross-feeding, results can be lost without gradient fractionation. After collection, gradient fractions can be screened by both quantitative (qPCR) and qualitative (fingerprinting, gene sequencing) methods to identify labeled taxa. The comparative allocation of specific nucleic acid templates to distinct BDs in gradients of ^{13}C-treatments vs. unlabeled controls provides the most relevant evidence for isotopic labeling. Since the advent of

high-throughput sequencing, the interpretation of SIP gradient fractions is now becoming frequently based on sequencing libraries, or even direct targeted metagenomics of density- resolved DNA.

(ii) Time series incubations should be conducted. The analysis of several successive time points allows monitoring the appearance of labeled templates in 'heavy' fractions over time, and to discriminate primary vs. secondary labeling effects in SIP (i.e., crossfeeding). Primary substrate consumers will always be more directly labeled, whereas metabolites, and even the degraders themselves, may be the basis for secondary label distribution via trophic interactions. Although a potential problem for data interpretation of single time points, crossfeeding can clearly be identified by time series incubations. In fact, this is a strength of SIP, allowing for the unraveling of trophic interactions and food web links in complex microbiota.

(iii) ^{12}C-controls must be regarded. As mentioned above, differences in BD between unlabeled and fully ^{13}C-labeled nucleic acids are very small (~0.04 g ml^{-1}), but can be resolved in SIP. However, BD differences of nucleic acids can be just as large due to distinct GC-content, which is especially problematic in DNA-SIP. Microbial genomes can vary between 35 and 75% GC-content, connected to intrinsic BD variation also of up to ~0.04 g ml^{-1} in CsCl. In effect, fully ^{13}C-labeled DNA of a low-GC bacterium can be of the same BD than unlabeled DNA of a high-GC microbe. This clearly explains the need for including ^{12}C-controls in SIP experiments, especially if labeling efficiency is low. The discrimination of true labeling effects (appearance of labeled nucleic acids) vs. GC-effects (growth of high-GC bacteria) is only possible if gradient results from labeled and unlabeled treatments are compared.

Literature:

Lueders, T. DNA- and RNA-Based Stable Isotope Probing of Hydrocarbon Degraders. In: Hydrocarbon and Lipid Microbiology Protocols; Springer Protocols Handbooks; Humana Press, 2015; p doi: 10.1007/8623_2015_74. https://doi.org/10.1007/8623_2015_74.

A.2 PRACTICAL PROCEDURES FOR SIP

A.2.1 INCUBATION OF SOIL MICROCOSMS

- The ÖMIK Team will provide the students with a representative local soil sample named "Grüner Hügel" (GH)
- The soil will be sieved (2 mm) and 5 g of soil will be weighed into 125 ml MK serum bottles (8 x).
- Add unlabelled (^{12}C) or^{13}C-labelled substrates to one each of paired microcosms:
  - 0.2 ml g^{-1} soil of a 50 mM glucose solution
  - 0.2 ml g^{-1} soil of a 100 mM methanol solution
  - 0.2 ml g^{-1} soil of a 100 mM acetate solution optional
  - no amendment controls
- Add by cautious distribution of droplets across soil with a 1-ml syringe, 0.4 mm needle
- Close serum bottles, add 20 ml of air as overpressure reservoir.
- Take starting point gas samples (3 ml in Exetainers) for subsequent CO_2 quantification
- Incubate for 48 h at room temperature
- Take further CO_2 samples after 24 and 48 h.
- Sacrifice microcosms after 48 h and continue with DNA extraction (optional: freeze soil at -20°C)

A.2.2 Quantification of CO_2 by gas chromatography

Gas chromatography (GC) is used to measure soil respiratory activity via the quantification of CO_2 production. Measurements will be done on a Hewlett-Packard 5980 series II gas chromatograph (Palo Alto, California, USA), equipped with a thermal conductivity detector (TCD), a Hewlett-Packard 3396 series II signal integrator and a Chromosorb 102 column (Alltech, Unterhaching). Column details: Length: 2 m; inner diameter: 3.2 mm) with 100% helium as the carrier gas (15 ml per min). The injector temperature is at 150°C, the column temperature at 40°C, detector temperature at 175°C.

To determine CO_2 concentrations, defined standards of gas mixtures are prepared in bottles in a range of 0.5- 5 vol.-% CO_2 with argon as pressure gas. Standards will always be measured at the beginning and end of each measurement series.

- For each measurement, inject 100 µl of the gas phase of each microcosm sampled with a gas syringe (glass) and inject into the GC.
- Measure each microcosm 2-3 times for technical replicates.
- After measurement, the baseline of the peak areas of the gas samples is straightened with the chromatograph editing program (Eurochrom Version 3.05 P5, Knauer Gmbh, Berlin, Germany) for correct output of the actual peak areas.
- Convert peak areas to CO2concentrations in Excel.

A.2.3 Soil DNA extraction

- Each student should extract 2 replicates of DNA from one distinct soil microcosm.
- Prepared in advance: Add 0.2 ml (= 1 PCR cup full) of ~1:1 mixed 0.1 mm and 0.7 mm Zirconia/Silica beads (www.biospec.com) to 2 ml bead beating vial with screw cap, autoclave.
- add 800 µl PTN buffer (pH 8) and 100 µl 20% SDS
- add soil or sediment sample (~500 mg, <500 µl), so that vial is filled maximally to below gripping ring
- FROM NOW ON WORK IN FUME HOOD (PHENOL!) AND SWITCH GLOVES REGULARLY
- Add 200 µL Phenol/Chloroform/Isoamylalcohol (25:24:1) pH8.
- Check that vial is not filled to more than 1.75 ml with >250 µl headspace remaining
- Tissue Lyzer: 1 min at 30 Hz
- spin down 4 min at 14000 rpm & 4°C (as all subsequent centrifugation steps)
- take 850 µl supernatant, place in 2 ml “Phase Lock Gel Heavy” tube (Eppendorf)
- Extract by vigorous manual shaking with 1 vol (850 µl) Phenol/Chloroform/Isoamylalcohol (25:24:1) pH8, spin 4 min at 14000 rpm & 4°C
- take 800 µl (or as much as possible) supernatant, place in 2 ml “Phase Lock Gel Heavy” tube (Eppendorf)
- Extract by vigorous manual shaking with 1 vol (800 µl) Chloroform/Isoamylalcohol (24:1), spin 4 min at 14000 rpm
- take 650 µl (or as much as possible) supernatant, mix thoroughly with 2 volumes PEG (1300 µl)
- You can stop working in fume hood now
- Precipitate DNA by spinning at max rpm (16000) and 4°C for 30 min

- remove liquid with pipette (or decant if you dare, best only if pellet is clearly visible).
- Add 150 µl ice cold (-20 °C) 70 % EtOH, gently wash pellet, spin down (4 min, 16000)
- remove EtOH carefully by pipetting with a 200 µl pipette tip, dry pellet briefly on lab table (max. 5 min)
- Resuspend each DNA pellet in 50 µl EB buffer (can vary depending on expected yield), pool duplicates
- Suspend by gentle flipping, do not vortex,
- Briefly spin down, transfer to 0.5 ml Eppendorf-Cup, store at -20°C

Buffers etc.:

PTN Puffer: 120mM $NaPO_4$, 125 mM Tris, 0.25 mM NaCl, pH8	16,02 g/l Na_2HPO_4, 0,86 g/l NaH_2PO_4, 11,2 g/l Tris-HCl, 6,6 g/l Tris-Base; 1,46 g/l NaCl, adjust to pH 8 with HCl, filter sterilize, autoclave
20% SDS	20 g SDS in 100 ml H_2O, prepare with sterilized H_2O in baked glassware
30% PEG, 1.6 M NaCl; pecipitation solution	150 g polyethylene glycol 6000 + 46,76 g NaCl in RNAse free water, 500 ml final volume. Prepare in baked glassware, first dissolve PEG in microwave, adjust to final volume, autoclave
EB Buffer	10 mM Tris, pH 8.5, prepare with RNAse free water, filter sterilize, autoclave. Better: take from purchased Kits (Qiagen, etc.).

A.2.4 Optional: Removal of humic acids by Sephadex Gel Filtration

This step is optional and will be decided during the course. Only for impure DNA extracts with much co-extracted humics, iron oxides, etc. Depending on the soil, DNA extracts have to be purified to remove PCR inhibitors like humic acids.

Either use original Sephadex G50 powder (custom-made, higher purification capacity) or ready-to-use Sephadex DyeEx Spin Columns from Qiagen.

Protocol for DyeEx Spin Columns (Quiagen)

- All centrifugation steps at 750g (~2800 rpm), 2 minutes and 20°C (RT)
- gently vortex the spin column to resuspend the resin
- loosen cap a quarter turn, snap off the bottom closure
- place the column in a 2ml collection tube
- centrifuge 2 min (to remove the storage buffer)
- place the column in a fresh sterile 1,5 ml microcentrifuge tube
- fill 50 µl nucleic acids slowly onto resin-bed, do not touch the resin-bed surface with the pipet
- centrifuge 2 min, transfer flow-through DNA into fresh 0.5 ml cup, label, continue

Protocol for custom-made Sephadex Spin Columns

- Sephadex G50 must be pre-hydrated (15 min) in RNAse free water (1g resin in 25 ml H_2O)
- All centrifugation steps at 750g (~2800 rpm), 2 min and 20°C (RT)
- load 500 µl of hydrated Sephadex resin into BioSpin-Columns (BioRad, with collection tubes) and centrifuge 2 min
- Discard flow-through and repeat above step to add more resin
- equilibrate column 2x with 50 µl EB buffer (spin after adding)
- load 50µl of raw DNA extract in EB onto column and spin
- discard column and continue working with DNA effluent

A.2.5 Agarose gel electrophoresis of DNA

Agarose gel electrophoresis is commonly used for visualization, separation and identification of DNA extracts and fragments. When agarose polymerizes (after melting in an electrophoresis buffer), a 'molecular sieve' structure forms within the gel that can be used to separate DNA fragments according to their length. Applying voltage, DNA fragments move through the gel towards the anode due to their negative charge. Short DNA fragments move faster through the 'molecular sieve' than long DNA fragments.

Agarose gels often prepared in a concentration ranging between 0.5% to 2.0%, depending on the desired separation efficiency of the target DNA fragment lengths (see table 1). Commonly used electrophoresis buffers are Tris base, e.g. TBE (Tris Borate EDTA) or TAE (Tris Acetate EDTA). To visualize DNA, a fluorescent dye (GelRed) that intercalates with the DNA double helix is added to the agarose gel. When exposed to ultraviolet light, it will fluoresce with an orange color that strongly intensifies after binding to DNA. Although less toxic than other popular dyes (e.g. ethidium bromide), it should be handled with care.

For security: Apply "good hand bad hand" principle! One glove only!

To increase the density of the DNA samples and make them sink into the wells of the gel, each DNA sample is mixed with a loading buffer that contains 30% Glycerol and a blue dye, Bromphenolblue (BPB, Bromphenolblau). In addition to the samples, a DNA marker / ladder is added into a separate well of the gel. The DNA marker contains DNA fragments of known size and concentration, acting as a reference for all unknown DNA samples.

A gel picture is generated after every DNA extraction or PCR step to visualize success of the step and integrity of nucleic acids.

Reagents (already prepared):

- 1x TAE (4 mM Tris-Acetat, 0.1 mM EDTA)
- 6x Loading buffer [0.1% w/v Bromphenolblue (BPB), 30% v/v Glycerol]
- Sybr Safe Nucleic acid stain dilution (1:1 dilution in TAE, 5000 x working conc.).
- DNA Marker: Gene Ruler 1 kb DNA-Ladder (Thermo)

Gel casting:

- Wear protective glasses and use special "hot-hand" glove!
- Agarose: Serva, 1.5 % gels in 1 x TAE.
- Weigh 0.6 g agarose into 200 ml-Erlenmeyer flask
- Fill up with 40 ml of 1 x TAE buffer, press "zero" on balance
- Boil in the microwave until agarose is dissolved (~2 x 1 min), mix after 1 min by swirling manually using a "hot-hand"
- Check volume by weight, compensate evaporation with MilliQ water
- Cool flask further under tap water
- Add 2 µl of Sybr Safe working solution, mix by swirling manually
- Pour the "not so hot" gel into gel cast tray on a flat surface avoiding bubbles, insert 2 combs

Electrophoresis:

- When gel has fully cooled, remove combs and put gel (in gel tray) into electrophoresis chamber
- Cover gel with 1 x TAE (1-2 mm of buffer over gel)
- Prepare the sample by mixing 5 µl of PCR-DNA with 2 µl 6 x loading dye (spots on parafilm)
- Load 5 µl sample into the appropriate wells carefully.
- Load 5 µl "ready to use" DNA ladder on both sides of each gel
- Run gel at 80 V for 45 min
- Note: DNA and RNA are negatively charged and thus migrate from cathode to anode (or from black to red colour)
- Visualize in UV-Chamber, check the length and intensity of your PCR-amplicons, take photo

A.2.6 NANO DROP QUANTIFICATION OF DNA

NanoDrop is a UV-Vis (220 nm to 750 nm) spectrophotometer designed for microvolume (0.5 – 2 µL) analysis of DNA, RNA and protein samples. Besides direct concentration measurement at 260 nm (A_{260}), ratios of sample absorbance at 230, 260 and 280 nm provide information about the quality and purity of a nucleic acid sample. For the '260/280' ratio, ~2.0 is generally accepted as "pure" for RNA. If the ratio is appreciably lower in either case, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. The 260/230 values for "pure" nucleic acid are often higher than the respective 260/280 values. They are commonly in the range of 1.8-2.2. If the ratio is appreciably lower, this may indicate the presence of co-purified contaminants, e.g. salts.

Procedure:

1. Start the NanoDrop (ND-ONE), select 'dsDNA' measurement on the home screen
2. Make sure that both "Auto-blank" and "Auto-Measure" functions are switched OFF!
3. Clean the surface of the measurement pedestal with 3 µl H_2O and a Kleenex
4. Add 3 µl H_2O, close the arm of the measurement pedestal, initiate blank measurement
5. Remove water blank with a Kleenex
6. Add again 2 µl of water and measure as "zero" sample, check if results and baseline are truly close to zero. If not, repeat cleaning and blank measurement.
7. Remove water with a Kleenex

8. Add 2 µl of DNA-extract and measure DNA. Document measurement, including concentration (ng μl^{-1}) as well as 260/280 ratios
9. Afterwards, again clean the surface of the pedestal with 3 µl ddH_2O and a Kleenex

A.2.7 CENTRIFUGATION OF DNA-SIP GRADIENTS

The protocol and procedures described here are designed for 5.1 ml polyallomer quick-seal tubes to be spun in a VTI 65.2 vertical rotor. Volumes can be easily down- or upscaled for other tubes and volumes. For DNA, gradients should be prepared at an average BD of 1.71 – 1.72 g ml-1 CsCl before centrifugation. This will ensure an optimal resolution of unlabeled and labeled DNA into 'light' and 'heavy' gradient fractions after centrifugation.

- Per gradient, mix in an 15 ml Falcon tube:
  - 5 ml CsCl (~1.84g/ml)
  - up to 1000 µl of GB containing 5 µg of DNA
  (Nano Drop-quantified, volume must be subtracted from that of GB)
- Mix well, measure refractory index (75 µl aliquot) to control pre-centrifugation average density (RI should be 1.4042 ± 0.0002 = 1.72 g/ml CsCl).
  Adjust by adding 100 µl aliquots of GB or CsCl, when necessary.
- Transfer centrifugation medium into 5.2 ml polyallomer QuickSeal tubes (Beckman) using a 10 ml syringe with a 1.2 mm needle. No air bubbles in tube! Balance opposing tubes for rotor to ± 0.02 g.
- Seal tubes by welding with tube topper, put into VTi vertical rotor, don't close empty slots.
- Spin ~36 h at 20°C and 44.500 rpm (184 000 g_{av}).
- Brake setting to "5" to more gently slow down gradients after run.
- Carefully remove tubes from rotor, minimize any mechanical disruption and proceed immediately with gradient fractionation.

Buffers etc:

RI of pure water at 20°C: 1.3330 (check when using refractometer)

CsCl [~1.84 g/ml]. Add 50 g CsCl (Calbiochem) to 30 ml GB. RI should be at ~1.4164

Gradient Buffer (GB): 0.1 M Tris-HCl (pH 8 = 8,88 g/l Tris-HCL & 5,3 g/l Tris-Base), 0.1 M KCl (7,46 g/l), 1 mM EDTA (0,37 g/l). Prepare with RNase-free reagents in nuclease-free water, filter sterilize (0.2 µm), autoclave in baked glassware.

Fractionation:

- After centrifugation, carry rotor from centrifuge to bench with minimum eruption, remove tube from rotor and adjust within fractionation device.
- Fit sterile 0.4-mm needle to tubing from syringe pump, pump out air
- Carefully poke needle into centrifugation tube at bottom of nozzle. Poke needle minimally into opposing wall for better fixation. Careful: do not puncture opposing wall!
- Poke hole into bottom of tube with a sterile 0.4-mm needle.
- Start syringe pump at rate "7" (1 ml/min), collect 10-11 fractions (~500 µl) in sterile 2-ml cups by manual shifting rack every 30 sec. (for more fractions: 25 sec)

Time steps in 30 sec intervals:

1	2	3	4	5	6	7	8	9	10	11
0:30	1:00	1:30	2:00	2:30	3:00	3:30	4:00	4:30	5:00	5:30

Density measurement:

- Measure refractory index of fractions (75 µl from each fraction), start with lightest (= 13th) fraction. Take care to have refractometer in "RI" mode!
- Due to fractionation, the 11th fraction will contain some water and the refractory index may be much lower than expected. In this case, it should be discarded. If needed, the density can be estimated from the decreasing densities of the other fractions.
- Densities can be calculated from refractory indices by equation in Excel:
  $y = -11{,}293\ x^2 + 42{,}6513\ x - 35{,}9133$

Beware: this standardization has been empirically generated for the above gradient setup! A change in salt batch, concentrations or stocks make re-standardization necessary!

Nucleic acid precipitation:

- Precipitate DNA from fractions with 2 vol (~1000 µl) PEG each.
- Mix thoroughly; spin 30 min at max speed (16000 rpm) and 4°C.
- Take care that all 2-ml cups are orientated similar within the centrifuge, because pellets will not be visible after centrifugation and might be lost by pipetting during washing if localization is not known.
- Remove supernatant with pipet, don't discard or pellet might be lost.
- Wash with 150 µl 70% ice cold EtOH, spin 5 min at max speed
- Remove supernatant with 200-µl pipet
- Elute by placing 25 µl of clean (RNA-grade) EB Buffer on assumed pellet, shake 1 min in Eppendorf Thermomixer at 1400 rpm and 30°C to dissolve uniformly.
- Spin down 1 min at max speed, Place eluted nucleic acids in 8-cup strips with single caps

A.2.8 Amplification of bacterial 16S rRNA gene amplicons for Illumina-Sequencing

- Each student will continue working with one selected set of gradient fractions.
- Each DNA fraction is amplified in only one replicate. Do not forget negative and positive controls!
- Prepare following PCR Master Mix. Pipetting scheme for one PCR reaction (multiply n + 10 %):

2x NEB-Next PCR Kit	25 µl	
50 µM ilu_515f primer	0.3 µl	
50 µM ilu_806rN primer	0.3 µl	
PCR H_2O	22.4 µl	
DNA-template	2 µl	(not include in Master Mix!)

Thermal profile for Cycler:

95°C	3 min	
95°C	30 sec	}
55°C	30 sec	} 30-32 cycles
72°C	60 sec	}
72°C	5 min	
8°C	hold	

Continue with gel electrophoresis and amplicon purification, as adequate.
Amplicons are sent out for sequencing with a company. Further instructions will be provided.

PART B: FLUORESCENCE IN SITU HYBRIDIZATION (FISH)

B.1 OVERVIEW

Fluorescence in situ hybridization (FISH) is a technique developed in the mid 1990s. FISH uses fluorescent DNA oligonucleotide probes to target taxon-specific sequence patterns in ribosomes (ribosomal RNA). This results in fluorescence signals for microbial cells of a certain group, lineage or taxon, that can be detected using a fluorescent microscope.

FISH of microbial cells always consists of four parts:

1) Fixation of the sample containing the target cells. Fixation stabilizes macromolecules and cytoskeletal structures thus preventing lysis of the cells during hybridization. Fixation also permeabilizes the cell walls for the fluorescently-labeled oligonucleotide probe molecules. . Standard fixatives are aldehydes and alcohols.
2) The fixed cells are incubated (hybridized) in a buffer containing the labeled probe at a specified temperature which favours the specific binding of the probe to the target. Ideally, only those probe/rRNA pairs will form which have no mismatches in the hybrid. Consequently, only target cells that contain the full signature sequence on their rRNA will be stained. The subsequent washing step will remove all unbound probe molecules.
3) Finally, the DNA of the hybridized cells is counterstained with DAPI and embedded in antifade mounting medium.
4) The hybridized and counterstained cells are then analyzed with epifluorescence microscopy or super-resolution microscopy.

***Safety*:** Be aware that most nucleic acid stains are believed to be mutagenic. In addition, the fixatives and formamide, which is used in the buffers, have to be handled with great care - use the appropriate gloves, work under the fume hood if necessary and dispose waste according to the waste management system.

Literature

Pernthaler, Jakob; Glöckner, Frank-Oliver; Schönhuber, Wilhelm; Amann, Rudolf. 2001. **Fluorescence in situ hybridization (FISH) with rRNA-targeted oligonucleotide probes**. Methods in Microbiology, Volume 30, Pages 207-210, https://doi.org/10.1016/S0580-9517(01)30046-6

Fuchs, B. M., J. Pernthaler, and R. Amann. 2007. **Single cell identification by fluorescence in situ hybridization**, p. 886-896. In C. A. Reddy, T. J. Beveridge, J. A. Breznak, G. Marzluf, T. M. Schmidt, and L. R. Snyder (ed.), Methods for General and Molecular Microbiology, 3rd ed. ASM Press, Washington, D.C.

Manz, W., R. Amann, W. Ludwig, M. Wagner, and K.-H. Schleifer. 1992. **Phylogenetic oligodeoxynucleotide probes for the major subclasses of proteobacteria: problems and solutions**. Systematic and Applied Microbiology 15:593-600.

Llobet-Brossa, E., R. Rosselló-Mora, and R. Amann. 1998. **Microbial community composition of wadden sea sediments as revealed by fluorescence in situ hybridization**. Applied and Environmental Microbiology 64:2691-2696.

This protocol is a modified version taken from Pernthaler et al. (2001).

https://www.arb-silva.de/fish-probes/fish-protocols/

Table: Properties of the FISH probes. The names, target organisms, sequences, fluorophores and the fluorophore's specific excitation and emission wavelengths (Ex./Em.) of probes and the properties of the 4',6-Diamidin-2-phenylindole (DAPI) counterstain are given.

FISH probes	Target	Sequences (5'→3')	Fluorophore	Ex./Em. [nm]
EUB338	All Bacteria	GCTGCCTCCCGTAGGAGT	Atto-565 or Atto-488	565/590 or 488/520
EPSY549	ε-Proteobacteria	CAGTGATTCCGAGTAACG	Atto-647	644/670
GAM42a	γ-Proteobacteria	GCCTTCCCACATCGTTT	Atto-647	644/670
ARCH915	All Archaea	GTGCTCCCCCGCCAATTCCT	Cy3	550/570
Stains				
DAPI	all DNA	-	-	358/461

Table: Filter combinations of the ÖMIK Fluorescence Microscope

Filter set	Exitation [nm]	Emission [nm]	Dyes
1	365	420	Dapi
2	450-490	515 – 565	Atto 488
3	546	580 – 590	Atto 565
4	Cy3	570 long pass	Cy3

B.2 PRACTICAL PROCEDURES FOR FISH

B.1.1 FIXATION

Fixation of fresh sludge or biofilm samples (Llobet-Brossa et al., 1998)

1) Fix biofilm samples with fresh 4% paraformaldehyde in a 1:1 mixture (2% end conc.). Incubate over night at 4°C.
2) Wash: centrifuge at 16.000 x g for 5 minutes; pour off supernatant and resuspend sample with 1 ml 1 x PBS pH 7.6
3) Repeat step 2 twice.
4) Add 500 µl PBS, resuspend cells well.
5) Add 500 ml cold, absolute ethanol, mix well and store at -20°C or -80°C until further processing OR directly continue with FISH protocol (without EtOH mixing)

Buffers etc.:

PBS Puffer	137 mM NaCl (8 g/L), 2.7 mM KCl (0.2 g/L), 10 mM Na_2HPO_4 (1.44 g/L), and 1.8 mM KH_2PO_4 (0.24 g/L). Autoclave.

B.1.2 SAMPLE PREPARATION

1) Mix fixed sample by flipping and dilute a small aliquot of 5 to 10 µl (depending on cell density and well size) 1:5 or 1:10 in MQ and spot it in wells of a polylysine (no wells) or gelatin (multi-wells) coated microscopy slide
2) Let air-dry at room temperature or 37°C for 30 -60 min
3) Dehydrate slides in a 50%, 80% and 100% ethanol series for 3 minutes each, subsequently air dry at 37°C (or max. 46°C)

B.1.3 HYBRIDIZATION

1) Prepare 2 ml hybridization buffer (**Table 1**) in 2 ml Eppendorf tube

Table1: Hybridization buffer, 35 % stringency		
Reagent	**Volume**	**Final concentration**
5 M NaCl	360 µl	0.9 M
1 M Tris-HCl, pH=8.0	40 µl	20 mM
Formamide, deionized	700 µl	35 %
Sterile H_2O	898 µl	add to 2 ml
10 % SDS (add last)	2 µl	0.01 %

Note: The final formamide concentration depends on the probe used and determines the stringency of hybridization

Details on hybridization probes to be used are given above. Further probes may be selected!
Add 10 µl of probe working solution (10 pmol μl^{-1}) to 90 µl of hybridization buffer in a 0.5-ml microfuge tube; keep probe solutions dark and on ice.

2) Prepare hybridization vessels from 50 ml Falcon tubes: insert a piece of tissue paper into tube and soak it with the remaining hybridization buffer. Wrap in aluminum foil (protection against light) and label tube.

3) Use separate tubes for each concentration of formamide;

4) Add 10 µl of probe mix to the samples in each well and place the slide into the polyethylene tube (in a horizontal position), keep from light.

5) Incubate at 46°C for 2 - 3 hours

B.1.4 WASHING

1) Prepare 50 ml of washing buffer (**Table 2**) in a polyethylene tube and preheat in a 48°C water bath

Table 2: Wash buffer	
Reagent	**Volume**
5 M NaCl	700 µl
1 M Tris / HCl	1 ml
0.5 M EDTA	500 µl
MilliQ	add to 50 ml
10% SDS (added last to avoid precipitation)	50 µl

2) Quickly transfer the slides carefully into preheated washing buffer (**work in fume hood: hot formamide !!**) and incubate for 25 min at 48°C (water bath)

3) Rinse slides carefully with distilled H_2O, let air-dry in the dark
The wells have to be completely dry before embedding, otherwise a fraction of cells will detach during inspection

4) For counterstaining cover each well with 10 µl of DAPI solution (1 µg ml^{-1}), incubate for 3 min; rinse slide with distilled H_2O let air-dry in the dark.

5) Mount samples in a 4:1 mix of 800 µl Citifluor (Citifluor Ltd, London, U.K) and 200 µl Vecta Shield (Vector Laboratories, Inc., Burlingame, CA). Vecta Shield contains a superior antibleaching reagent, but quenches DAPI fluorescence. Prepare in 2 ml tube, add 10 µl per well. Wells have to be completely dry before embedding, otherwise a fraction of cells will detach during inspection.

6) Add a cover slide to the sample. Seal with clear Nagellack. Double stained and air dried slides, as well as mounted slides can be stored in the dark at -20°C for several days without substantial loss of probe fluorescence.

7) Probe-conferred fluorescence fades much more rapidly than DAPI fluorescence in the microscopic image, and UV excitation will also bleach the probe signal. For counting, it is, therefore, safer to first quantify probe stained cells and subsequently all cells from the same field of vision in UV excitation;

PART C: ANALYSIS AND INTERPRETATION OF (META)GENOMIC AND AMPLICON SEQUENCING DATA

Welcome to the sequence analysis section of the practical course for the module Functional Microbiome Research, 2023! What follows is a living document that may change as the course progresses. Please be tolerant of errors therein, and let me know when you find them. I will try to keep the most current version of this document available on the course website and/or the github repository.

C.1 INTRODUCTION

Modern microbial ecology projects typically rely at least in part on culture-free sequencing of DNA or RNA extracted from complex substrates such as soil, water, or animal/plant tissue. In our lectures we have talked extensively about two broadly defined DNA-based methods for doing these culture-free surveys of microbes in the environment: **metabarcoding** (also known as "amplicon sequencing") and **metagenomics**. In this sequence analysis section of the module, we will examine both methods, using some popular software packages. In order to do this, we must also learn about using remote computing resources, and how to work in a Linux command line environment.

Throughout the course, you have been given several theoretical lectures on bioinformatic methods, including some explanations of fundamental algorithms. You have also been given numerous examples of microbiomes in nature. Finally, you are creating data from a microbiome right now, during the course of this practical course. In this section of the module, you will bring your new knowledge and your data together, to conduct a metabarcoding study on your soil samples, and a rudimentary metagenomic analysis of two public data sets.

We will cover three large topics, each of which will have a script associated with it that we as a class will generate together:

1. Scientific computing and Linux – script here
2. Metagenomic methods – script here
3. Metabarcoding methods – script here

Each night I will update the scripts to structure the discussion for the next day, and we will run through them in class together. You will have to adapt them to your own computing environment. Together they will become a record of our activity as a class, as we work through the bioinformatic pipelines. The living versions will be kept in the github repository links above.

C.2 SCIENTIFIC COMPUTING

Many algorithms that we will use require large reference databases and examine large portions of experimental data in real-time. Modern studies generate sequence data that can also be quite large in size, sometimes multiple terabytes in size. We can safely say that bioinformatic analyses are very "memory hungry", sometimes requiring hundreds of gb of RAM to conduct more intensive calculations and alignments. Our data will be relatively modest in size, but even our datasets and databases will sometimes require enough memory to overwhelm the (random-access) memory resources of standard home or office computers.

Scientists address these issues of "big data" in two ways:

(1) Scientists typically minimize the non-essential parts of the programs that they write to implement their algorithms. This means that you will see very few Graphical User Interfaces (GUIs) used for bioinformatic pipelines, though some high-quality bioinformatic GUIs do exist. Instead, software for bioinformatics are usually called from text-based environments called **command-line interfaces** (also called "terminals" or "shells"). Because they are so minimal, these programs are also more flexible - they can be wrapped into other programs, including your own simple programs (**scripts**) that you write. Our first sessions will focus on learning BASH to use these programs, probably the most popular command-line interface for scientific computing.

(2) Scientists do quite a lot of remote computing, using computing clusters and other shared computing resources to conduct memory-intensive operations on their data. In order to conduct modern research, universities, governments, and corporations have invested in shared computing centers with the hardware necessary for this kind of research. For most of these very large computing resources, Linux/Unix-like operating systems are the most efficient and universal work environment. In our case, we will be using Ubuntu Linux virtual machines in the cloud computing resources of the German Network for Bioinformatics Infrastructure (de.NBI), who have kindly offered their support for this course.

C.1.1 GETTING LINUX WORKING LOCALLY

The first step to learning about scientific computing is to get a working Linux environment onto your personal machine. **Please attempt the following installation of Linux before our first day**. This step of getting Linux working on your computer can be complicated, but it is essential for everything else that we do! If it fails for you, we understand, and we will spend much of our first session debugging this step and the following so that everyone can get to their computing resources.

MAC USERS:

You may not know it, but you have been running something very close to Linux every time you opened your computer! So leave your computer at peace, and just open the terminal application: use finder to go to the /Applications/Utilities folder, and double click "terminal".

PC USERS, WINDOWS VERSION >=10 :

Windows has an in-box solution for creating a Linux environment, called **W**indows **S**ubsystem for **L**inux (WSL). If your windows version is 10 or newer, you probably have this available in some form on your computer, but you must activate it. The exact process for activating WSL will depend on your windows version.

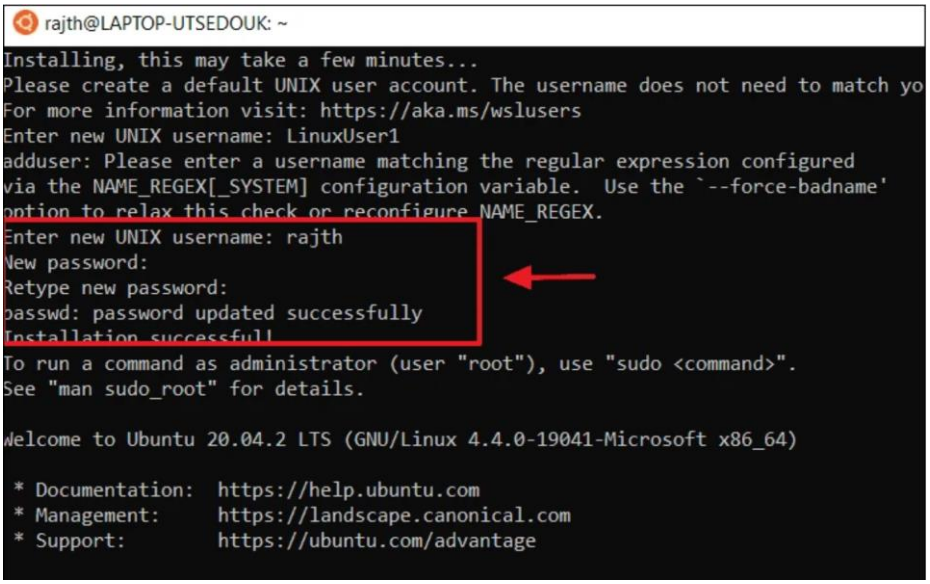
In addition to the official documentation, I found the following website to be very helpful for installation of WSL:

https://www.windowscentral.com/how-install-wsl2-windows-10

This older site was also helpful in the past, though some things may have fallen out of date:

https://www.altisconsulting.com/uk/insights/installing-ubuntu-bash-for-windows-10-wsl2-setup/

Once you have activated WSL on your PC, and have installed Ubuntu, test it out by searching for "Ubuntu" in your start menu. Running this should bring up a terminal, and ask you to create a username and password, something that looks something like:



PC users, Windows version <10 :

Finally, if you are running a really old version of Windows, let me know. We will find another solution.

C.1.2 Getting around a Linux Environment:

Basic shell commands and symbols

These are the universal BASH commands we will be using to get around and to manipulate our files, in addition to specialized bioinformatic programs. We will now try out each in our local BASH shells, so you can get accustomed to them.

pwd – print the current working directory

ls – list the contents of the current directory

cd – change directory

cp – copy a file

mv – move or rename a file

mkdir – make a new directory

echo – print something out to the display

cat – concatenate two or more files

rm [-r] – remove a file. Essentially deletes a file forever. Be careful!!

sudo – execute another command that requires root privileges

apt – access Ubuntu's native software package management system, for updating and upgrading software.

top – opens up a real time display that shows you the "busiest" processes ongoing in your machine. After you start the program, press `i` to see only active processes, press `e` to make memory units more readable, and press `1` then `t` to keep a close eye on your cores.

man – print the reference manual for a utility, command, or program, if available. If not, try the `--help` flag for any given program without a manual.

head – show top several lines from a text file

tail – show last several lines from a text file

less – interactively open a text file for reading

Important symbols for BASH environment

We will also play with the following special characters:

~ home directory (try with cd)

. current directory (try with cd). Also filenames that start with "." are "hidden".

.. parent directory (try with cd)

> direct the output of a process to a file

$ variable expansion

| pipe, connects the output from one command to another

; command separator

/ directory separator (compare to Windows!)

**** escape a character, so that it is no longer a "special character"

& send a process to run in the background

= variable assignment

***** wildcard for multiple characters (try with ls)

? wildcard for single character (try with ls)

comment (place it before a command and see what happens)

Advanced programs

These are more complicated utilities that we will be needing quite frequently:

scp – secure copy. We will use this to move files between our remote server (de.NBI) and our local server.

ssh – secure shell. This opens an encrypted terminal, usually on a remote machine, so you can use that computer as if it were right in front of you.

wget - a file downloader, using standard file transfer protocols.

conda – a software package and environment program that is used to handle the complex installation environments necessary for our kind of work. Conda should activate immediately when you log in to your de.NBI virtual machine. We will learn about the intricacies of Conda as we go.

nohup - a helper utility that we wrap around our big jobs, which tells our remote computer not to stop the process even if the connection breaks.

We will use numerous other programs and features of BASH. I will try to explain them as we use them. Use all the above to explore the terminal, and ask me questions as you wander. Be independent: remember that your computing environment will not look exactly like mine, you will have your own directory names, etc. Part of scientific computing is learning to adapt other's scripts to your own setup!

Finding Windows and Ubuntu filetree systems

In your exploration, you may notice that the directories we have been perusing don't look anything like the folders you see when you look around with the windows file explorer. That's because Windows and Ubuntu Linux do not want to mix their filesystems. You are probably currently wandering around the inside of the Linux filesystem tree. For windows users, now that you have learned how to get around in a linux filesystem, we need to find your old windows file system tree. For most WSL users, while in Linux, your windows file system should be visible here:

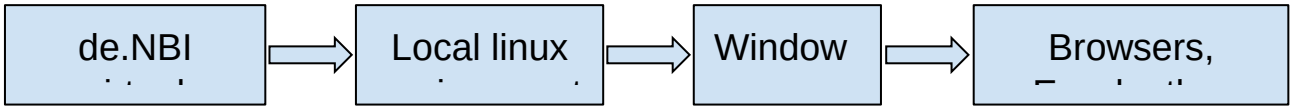
```
cd /mnt/c/
```

From there, can you find your way to some files and folders that you recognize from your Windows environment? To make our products from our Linux-based analyses available to windows, make a directory in that directory/folder, something like:

```
mkdir /mnt/c/PUTLINUXFILESHERE/
```

Can you now find your PUTLINUXFILESHERE folder with the Windows file Explorer program? To find it, look in C:\PUTLINUXFILESHERE\ with Windows file explorer. Pin it to easy-access to make it easy to find next time.

Note – be very careful when using Windows-based programs to edit anything that you want to keep in your Ubuntu environment. Modifying - and sometimes just reading - a Linux-created file with Windows can sometimes corrupt the file for further use by Linux tools. The opposite is usually fine, Linux respects the common file formats of Windows OS. However, to be safe, we will tend to use the following one-way information flow:



C.1.3 REMOTE COMPUTING

Now that your local Linux environment is working, let's use it to talk to your de.NBI virtual machine. We will use SSH and SCP to communicate with our de.NBI machines. To do this, you need to create an SSH keypair.

SSH Key generation

Secure login systems usually rely on an asymmetric, public/private keypair scheme. In such systems, you as a user create both a public key that you can give to servers, and also a complementary private key that only you control. When logging into a server that has your public key, you can prove your identity by providing the only key in the world that "fits" the public key, your very own private key.

We need you to generate your keyset, and following this we need to add your public key to your de.NBI virtual machine. To do this, in your local terminal generate a key pair, we'll use ssh-keygen:

```
ssh-keygen -t rsa -f <nameOfYourKeyHere>
```

-t rsa tells the algorithm which type of key generation algorithm to use.

-f is the name you want to give your key file

You will probably be asked to generate a password associated with your key. Keep this password, you will need it often!

This process will generate two keys. The public key should have a ".pub" extension, and can be shared freely. The private key will not have a file extension – it should be exactly as you named it in the command above. If you are not sure, look at the file with **less** or **cat** or **head**, and it should say -----BEGIN OPENSSH PRIVATE KEY----- in the first line.

Put both keys in your .ssh folder, which should be in your home directory (use "**cd ~**" or just "**cd**" to go to your home directory). If the .ssh folder doesn't exist, make it. Guard the private key with your life, and send me a copy of the public one in an email, with your name clearly in the email somewhere.

Logging into your de.NBI virtual machine

Once we have placed your public key into your assigned de.NBI machine, it should be ready to accept your login.

To get the login for your machine, you will use SSH. The correct SSH login command will be different for each of you, because your virtual machine has its own unique port number through which it is accessed. **As soon as I receive your public key**, I can activate your personal virtual machine (VM) from de.NBI. I can then generate your login information, and I will send you this information in an email. This login information will include your username, ip-address, and port number.

For example, when logging into my de.NBI VM from my home computer, I combine my username, ip-address, location of my private key file, and port number into the following command:

ssh -p 30192 -i </path/to/your/ssh/private/key> ubuntu@129.70.51.6

-p is the port number. Use only this port to interact with on your virtual machine.

-i is where you have stored your private ssh key. If you followed the instructions above, this will be in your .ssh folder in your home directory.

ubuntu is my username, and probably will be yours also (assigned by de.NBI).

If all keys are in place, this should allow you to log into your de.NBI virtual machine. Go ahead and explore it in just the same way you explored your local Linux filetree above, using your new knowledge of BASH commands!

File transfer using SCP:

Think of SCP as a long-distance version of the BASH command "cp", which we learned about above. Both of these commands copy a file from location A to location B. SCP requires a lot of the same information as SSH to do its job (port number, ip address).

To upload a filed called **localFile.txt** that is on the current directory on my office computer, to the directory **farAwayDirectory/** on my de.NBI machine, I do the following:

scp -r -i <...> -P <...> localFile.txt ubuntu@129.70.51.6:/farAwayDirectory/

-i is the path to my private key.

-P is the port on the de.NBI machine (note the **capital -P.** ssh uses a lowercase **-p** flag).

-r (recursive) is not necessary with single files, but it is essential if you want to download an entire folder.

The reverse is also possible. To download **farAwayFile.txt** from my de.NBI virtual machine, to the current directory (**.**) in my office computer, I do the following:

scp -i <….> -P <...> ubuntu@129.70.51.6:/farAwayDirectory/farAwayFile.txt .

Note the dot at the end of the command, indicating the current directory on your local computer.

Now try some file transfers between your local computer and your de.NBI virtual machine!

C.3: METAGENOMIC METHODS

When direct sequencing of environmental DNA and RNA became possible, scientists developed two broadly-defined DNA-based methods for exploring the diversity in a given biological sample: metagenomic and metabarcoding (="amplicon sequencing"). We've talked at length about both in lecture, so we won't dive into intricacies here. Briefly, however, metagenomics involves randomly fragmenting all of the DNA available in a sample, and sequencing as many of these DNA fragments as possible, hopefully without bias. This renders millions of reads that can be considered a statistical sample of the genetic material of all the organisms present in the biological sample, i.e. the metagenome of the sample.

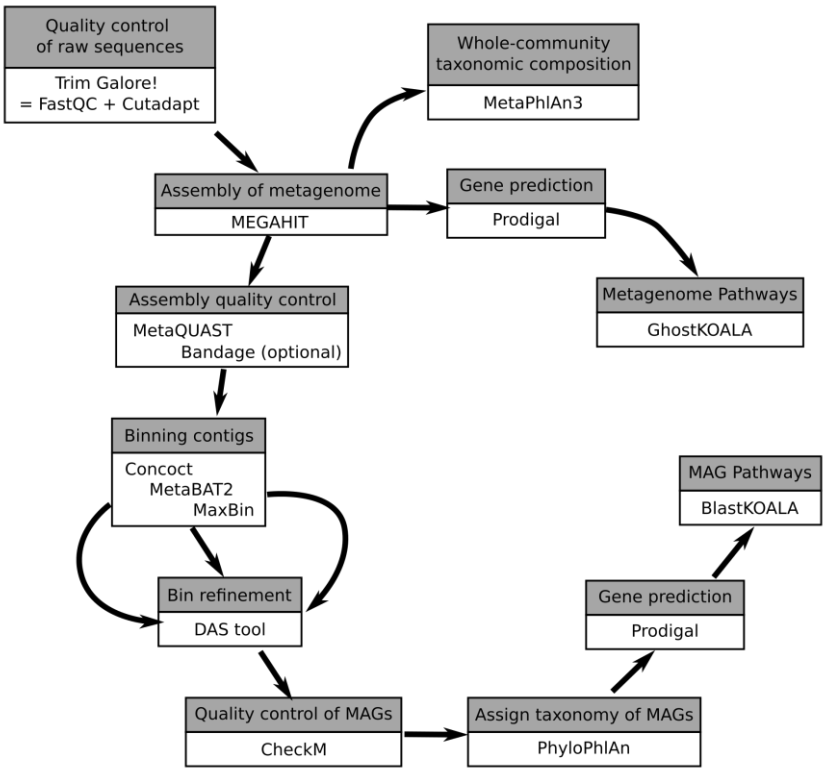
This sequenced sample of the metagenome can be handled in two ways...

Option 1 is to treat a metagenome like a single genome, and observe the metabolic pathways that are present in this "super" genome. We often don't need to know exactly which organisms are responsible for the presence of a metabolic pathway in which we are interested, and often numerous organisms are responsible anyway, due to horizontal gene flow and other forms of ecological redundancy. From this method we can quickly mine our biological samples (soil, water, plant tissue, etc) for interesting genes and pathways such as antibiotic resistance efflux pumps, etc.

Option 2 is to try to recover individual genomes from the soup of our metagenome reads. This approach can give you some very-fine-grain information about the taxonomy and function of the most abundant microbes in your sample. Quality **metagenome-assembled-genomes (MAGs)** are also rapidly increasing our ability to resolve microbial dark matter (see for example Nayfach et al. (2021)). However, this approach requires quite deep sequencing, meaning fewer samples per sequencer run, and more computational steps and computing resources needed. Additionally, very few genomes are typically recovered, especially in very microbially diverse samples or in unusual substrates, and especially without sufficient sequencing depth to allow for de-novo methods of assembly.

Both approaches can yield interesting information about a microbiome, so we will explore both. We will first explore our MAG-creation pipeline with publicly-available mock community dataset, taken from Sereika et al. (2022). We will use read libraries generated from both nanopore and Illumina MiSeq platforms. You will then repeat with a "wild" microbiome from a sample of activated sewage sludge from an anaerobic digester, used in the same paper.

The exact steps of our analysis will be in the code that we produce each day. Here is an overall schematic of the approach, with the particular software packages we will use.



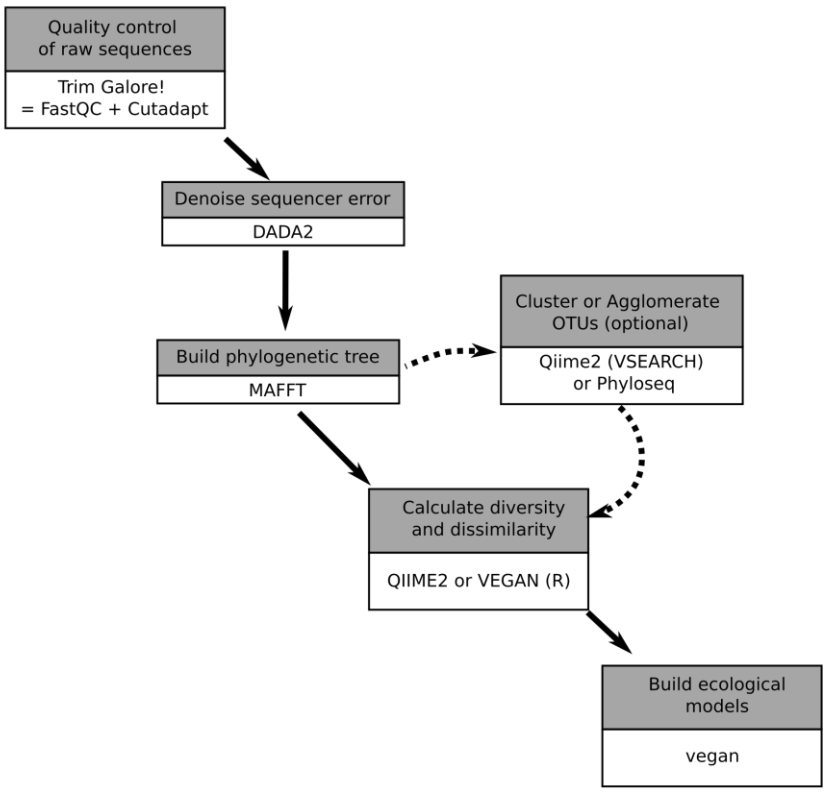
C.4: METABARCODING ("AMPLICON SEQUENCING") METHODS

In contrast to metagenomics, metabarcoding uses selective PCR or bioinformatic methods to target only a particular loci of the genome for all of a group of organisms in a biological sample. If taxonomy of microbes (e.g. prokaryotes and fungi) is of interest, ribosomal genes are typically targeted. These genes are useful barcodes because they are theoretically present in all life, in multiple copies, and differences in their sequences can often predict evolutionary relationships (taxonomy). By targeting these genes, we can draw a general picture of who is present in our sample, and perhaps even relative abundances (careful!!!!). We also begin to model changes in the microbial community structure by ecological predictors such as pH or moisture, if we have this data. In the case of prokaryotes, the 16s small subunit of the rRNA is commonly used. For your soil study with Dr. Lüders, you will be using primers that target the variable region 4 (V4).

Generally, metabarcoding studies do not require the same level of deep-sequencing that metagenome studies require. Many more samples can sequence in a single sequencer run, sometimes as few as several thousand reads is sufficient to saturate diversity curves for a single sample. As such, they are often a good "first step" to understanding your study system.

Other genes are possible, if your question is not taxonomy but function. For example, *Nif* genes can be used as a marker for nitrogen fixation, or secondary metabolic gene clusters can be sampled using different backbone synthase genes.

As above, the exact steps of our analysis will be in the code that we produce each day. Here is an overall schematic of the approach, many of which we will execute within the Qiime2 pipeline:



C.5: WRITING YOUR REPORTS

Once you have run the Sereika reads through the metagenomic pipelines, it's time to think about what you have done, and look for some interesting results. Your report will show me that you had a conceptual understanding of each of the steps that we performed. For your report on the bioinformatic/sequence analysis section of the practical, there is no specific page requirement, but you should address all of the following:

C.5.1 METAGENOMES:

5.1.1 QUALITY CONTROL OF RAW READS:

- Which library (influent or effluent) did you examine?
- How was the quality of the raw read library?
- How long were your reads on average?
- Were these single or paired-end reads?
- Was trimming necessary? And/or enforcing a minimum read length?
- Were 16s primers included in these reads? How do you know?
- Do you think that the authors did a lot/little to their read libraries before releasing them to the public? Why/not?
- Include figures that support your statements.

5.1.2 METAGENOME ASSEMBLY

- Give a general picture of the quality of your assembly:
- How many contigs?
- N50?
- 16S matches to known organisms?
- If so, how well did your metagenome match these reference genomes (genome fraction, and notes of missamblies?
- Are you able to say anything about the general bacterial/archeal community composition of your sample?

5.1.3 BINNING AND REFINEMENT OF BINS

- How did your binning process go? What worked, what did not?
- How many bins did you have before refinement, and how many after? If there is a difference, tell me why that might be.
- How is the quality of your refined bins? Use completeness and contamination metrics to explain your answer.

5.1.4 MAG TAXONOMIC ASSIGNMENTS

- Are your candidate MAGs closely related to anything that has been sequenced before? Do you think it is the same genus/species/strain? Support your answer using Mash average distances.
- If so, what is known about the ecology and metabolism of this organism?
- And if so, does this match your previous 16S-based reference genomes predicted by your analysis of your metagenome (section 1.2 above)?
- If not, what is the closest known MAG or reference genome? And do you think you have discovered an important new genome? You can address this more in your comments about the metabolic pathways of your organism, if you like.

5.1.5 GENES AND PATHWAYS

- How many genes are predicted for your metagenome in its entirety, and how many for your MAG?
- Do you find some (nearly) complete pathways, in either your metagenome or in your MAG?
- Do you find any interesting genes that indicate unique ecological function, or are of human interest (e.g. pathogens, antibiotic resistance, mutalistic/complementary function like essential amino acid synthesis, etc, etc)?.
- Does your MAG have any incomplete pathways, that are perhaps completed by other organisms (i.e. your metagenome)? What might this indicate?
- Feel free to discuss any other interesting data points that popped up in your pathways analysis or in general in the metagenomics pipeline.

C.5.2 METABARCODING

5.2.1 QUALITY CONTROL OF RAW SEQUENCES

- Were these single or paired-end reads?
- Was trimming necessary? And/or enforcing a minimum read length?
- How does the read length compare to Chu et. al? If it is different, why would this be?
- How was the general quality of these reads as compared to the Chu et al. (2018) dataset? If they are very different, speculate on why this might be.

5.2.2 DENOISING

- What were the results of applying a denoising algorithm to your data? More specifically:
- Were many reads lost? If yes, why?
- How many genetically different organisms are predicted in your total data set?
- How deep is your sequencing, by sample, before normalizing (rarefying) your read depths?
- Do you feel good about this depth, or do you think more is needed? Support your answer with figures.

5.2.3 COMMUNITY ANALYSES OF METABARCODING DATA

- After normalization (rarefying), can you compare alpha diversity among your samples?
- Show me ordinations from (1) a simple taxon-based dissimilarity coefficient and (2) a phylogenetically-weighted dissimilarity coefficient.
- Can you find any interesting groupings in these ordinations?
- If so, what explains these groupings - an experimental treatment? Or maybe a co-variate/confounding variable of some kind?
- What effect did changing the dissimilarity coefficient have on your two ordinations?
- Extra credit - what is the algorithm used to make your ordination? Can you find this information?
- Can this kind of community-based analysis pick up the signals we find in SIP with an identification of one-to-several ^{13}C-enriched species?
- Generalize your answer to the previous question a bit: what is the conceptual difference between community analysis methods like ordinations and population-specific methods like the SIP approach? Think in terms of equations/linear models, like we talked about in our lecture on modeling. When is one appropriate, and when is the other useful instead?