LongTREC

Funded by
the European Union

Bioinformatics Summer School

Long-reads Transcriptomics

Carmen Lafuente Sanz

Genoscope, Evry-Courcouronnes,France

LongTREC - The Long-reads TRanscriptome European Consortium
Marie Skłodowska-Curie grant agreement No 101072892

Course Contents

1. **Basic concepts of metatranscriptomics**
2. Experimental design
3. Downstream analysis

4

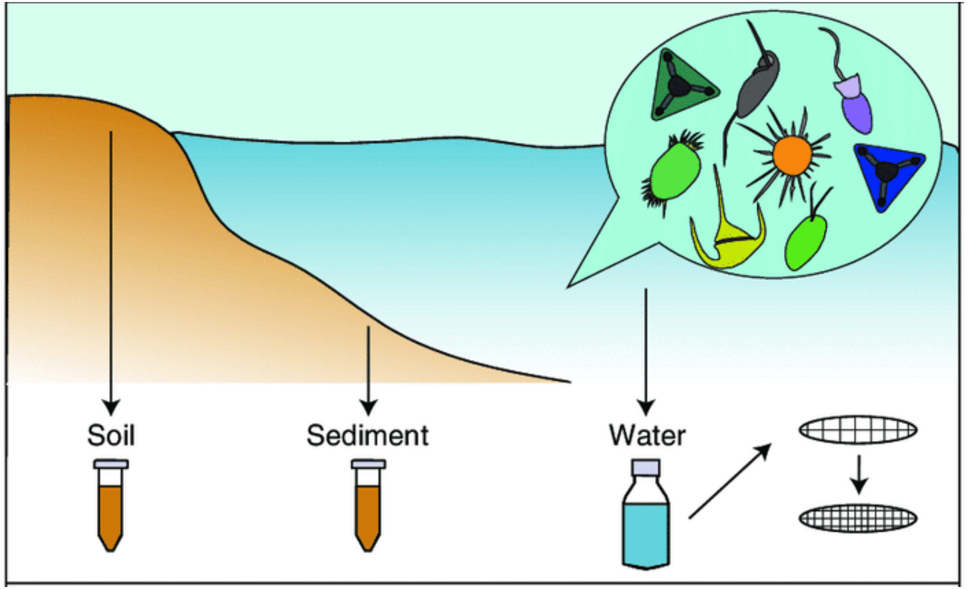
Section 1

Basic concepts of metatranscriptomics

What is metatranscriptomics and why does it matter?

Basic concepts of metatranscriptomics

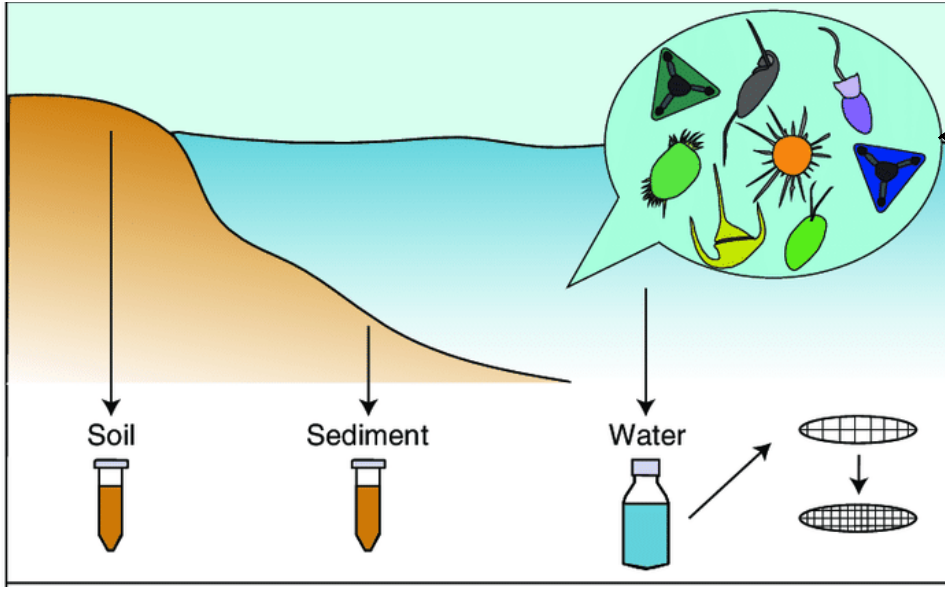
There are different approaches to study microbial communities



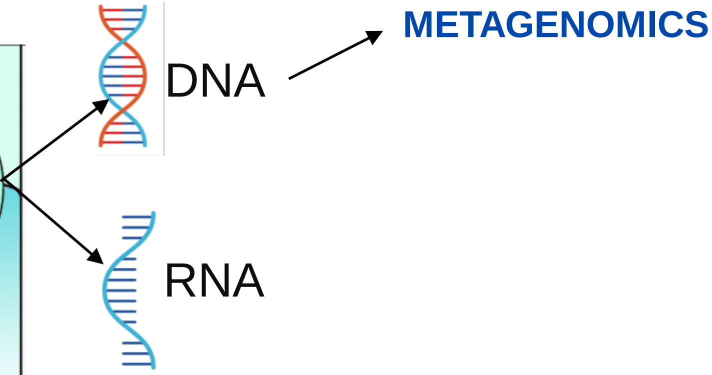
Adapted from: Burki, F., Sandin, M. & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. Current Biology, 31(20), R1233–R1282. https://doi.org/10.1016/j.cub.2021.07.066

Basic concepts of metatranscriptomics

Metagenomics: What genes are present?



Adapted from: Burki, F., Sandin, M. & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. Current Biology, 31(20), R1233–R1282. https://doi.org/10.1016/j.cub.2021.07.066



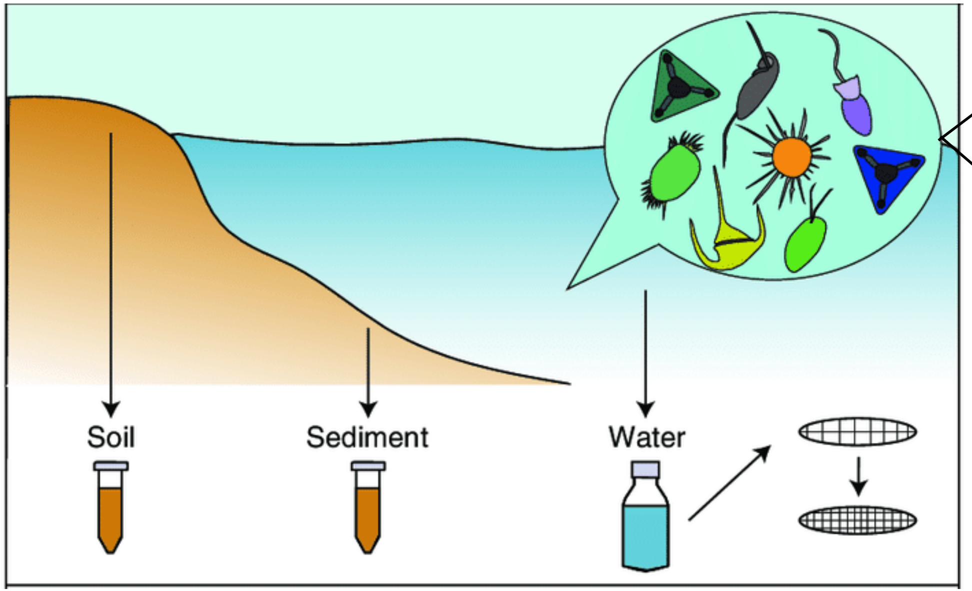
DNA → **METAGENOMICS**

RNA

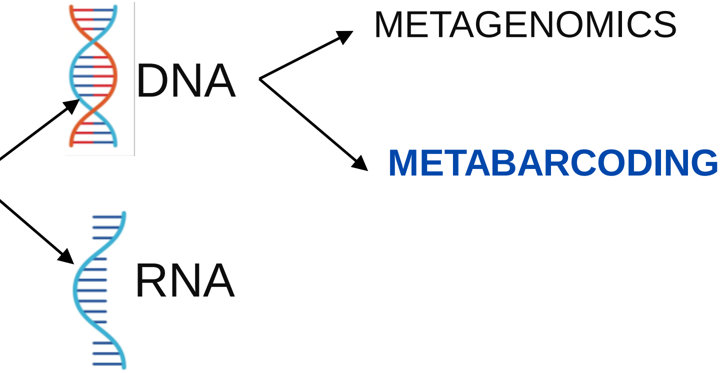
- Sequencing **total community DNA**
- Can reconstruct **genomes**, find **genes** and **functions**

Basic concepts of metatranscriptomics

Metabarcoding: Who is there?



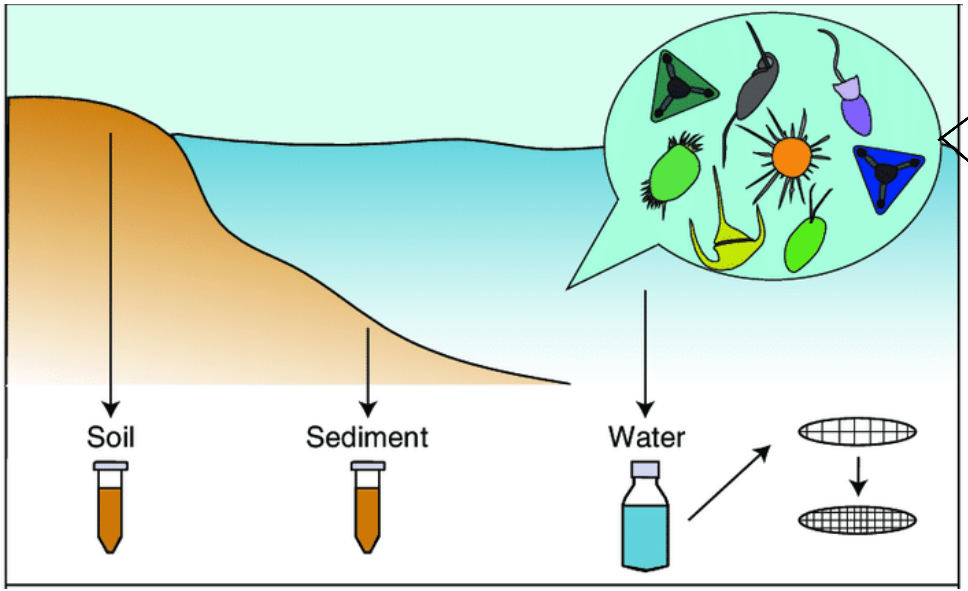
Adapted from: Burki, F., Sandin, M. & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. Current Biology, 31(20), R1233–R1282. https://doi.org/10.1016/j.cub.2021.07.066



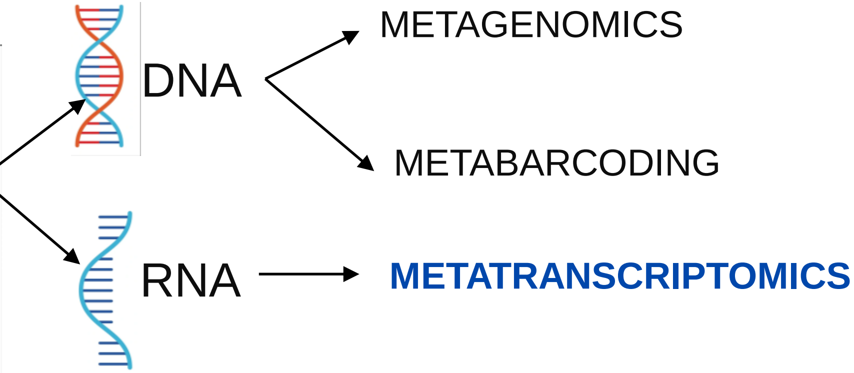
- Targets a **single genetic marker** (e.g., 16S for bacteria, 18S for eukaryotes)
- Good for building **community composition**

Basic concepts of metatranscriptomics

Metatranscriptomics: What genes are active right now?



Adapted from: Burki, F., Sandin, M. & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. Current Biology, 31(20), R1233–R1282. https://doi.org/10.1016/j.cub.2021.07.066



DNA → METAGENOMICS

DNA → METABARCODING

RNA → **METATRANSCRIPTOMICS**

- Assessment of **environmental gene expression**
- Detects **responses to environmental change** (stress, nutrients, pollution)

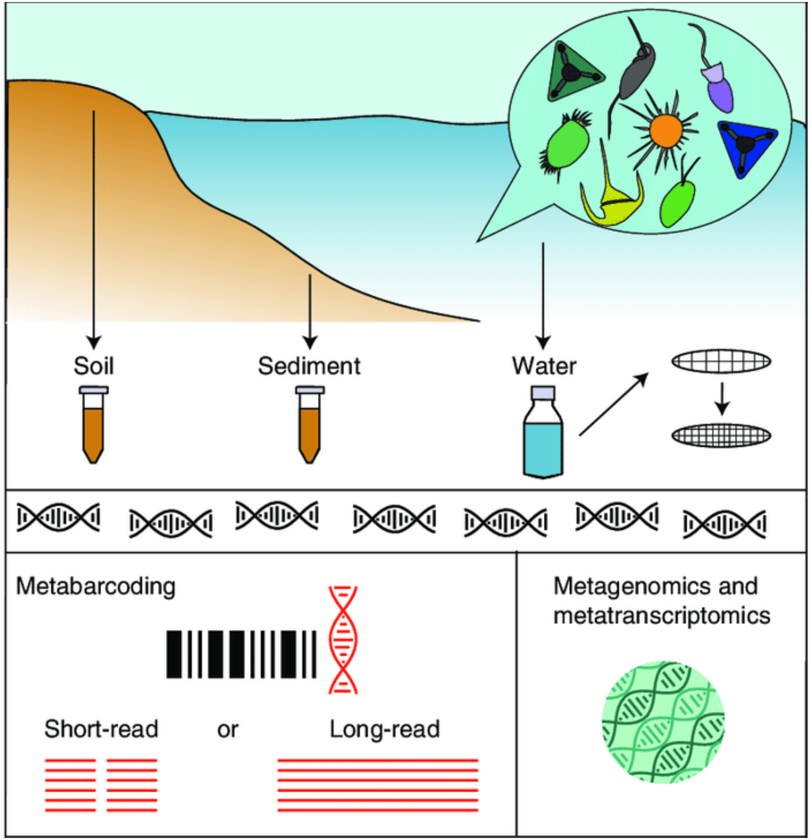
Basic concepts of metatranscriptomics

Metatranscriptomics: Why does it matter?

Identifies which genes are **actively being expressed** by community members, providing insights into functional activity rather than mere taxonomic presence.

Helps understand:

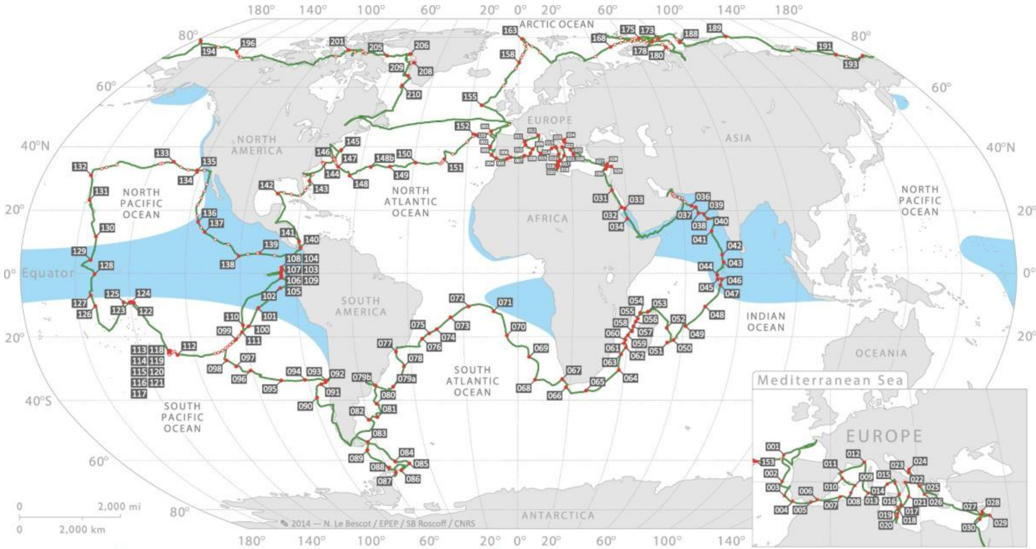
- Ecosystem function
- Microbial behaviour: p.e. human microbiome studies



Adapted from: Burki, F., Sandin, M. & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. Current Biology, 31(20), R1233–R1282. https://doi.org/10.1016/j.cub.2021.07.066

Basic concepts of metatranscriptomics

Practical example: TARA Oceans & marine metatranscriptomics



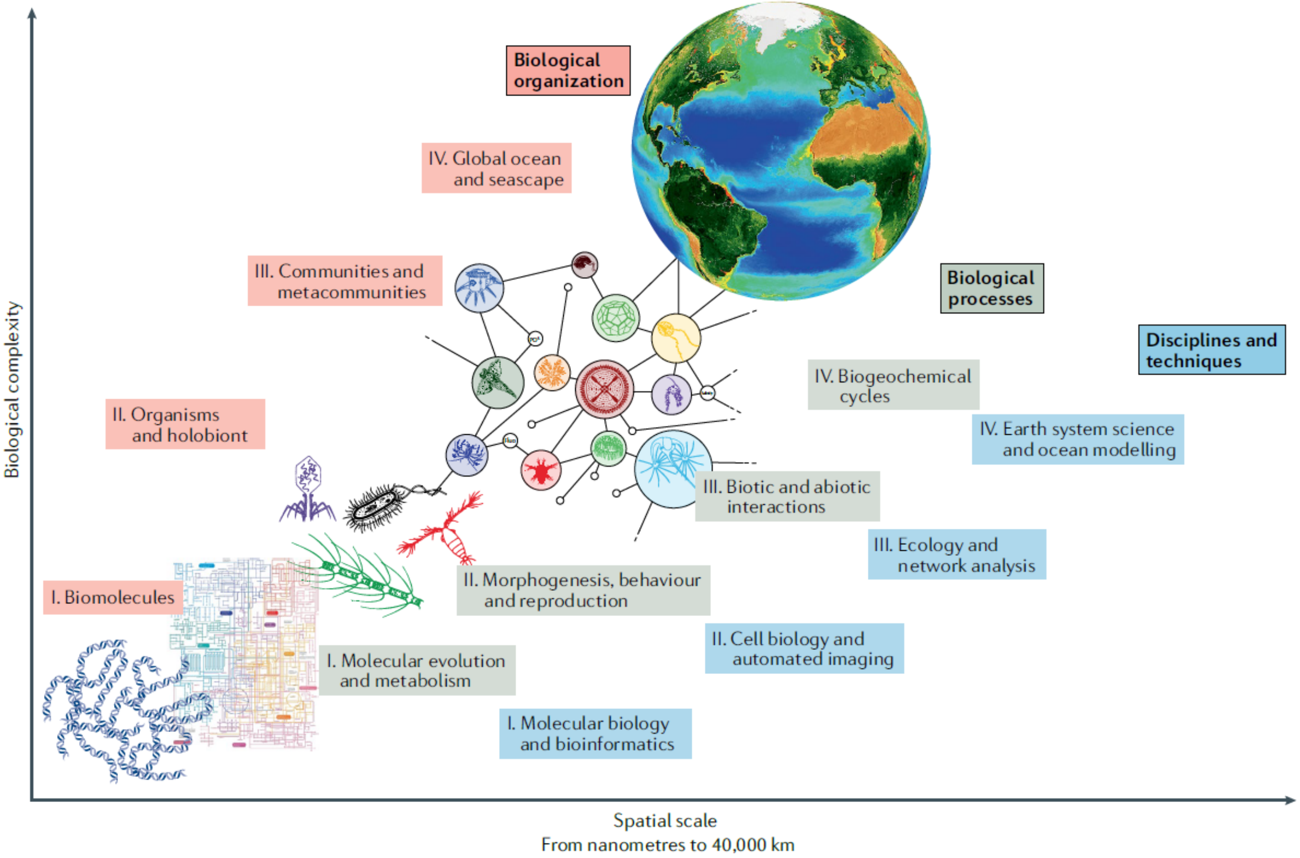
© 2014 — N. Le Bescot / EPEP / SB Roscoff / CNRS

● Environmental/Morphological/Genomic sampling station ○ Environmental sampling station

— Ship's track (2009-2013) **000** Station number 🟦 Annual mean oxygen <2 ml/L at 200 m (WOA09)

- International scientific expedition (2009–2013)
- Sampled ocean water from over 200 global stations
- 40,000 marine samples were collected
- **Goal**: Understand the diversity, function, and structure of planktonic life

Basic concepts of metatranscriptomics



Biological organization
IV. Global ocean and seascape
III. Communities and metacommunities
II. Organisms and holobiont
I. Biomolecules
Biological processes
IV. Biogeochemical cycles
III. Biotic and abiotic interactions
II. Morphogenesis, behaviour and reproduction
I. Molecular evolution and metabolism
Disciplines and techniques
IV. Earth system science and ocean modelling
III. Ecology and network analysis
II. Cell biology and automated imaging
I. Molecular biology and bioinformatics
Biological complexity
Spatial scale
From nanometres to 40,000 km

Challenges and limitations: Ribosomal RNA

More than 95% of RNA is **rRNA** → Difficult mRNA isolation → Wastes sequencing effort

EXPERIMENTAL

Eukaryotes

polyA enrichment

Prokaryotes

mRNA enrichment strategies

- rRNA hybridization/capture
- 5’-3’ exonuclease digestion of processed RNAs
- PolyA tail addition via polyA polymerase (E.coli)
- Antibody capture of mRNAs

COMPUTATIONAL

Computational filtering

Computational filtering

Basic concepts of metatranscriptomics

Challenges and limitations:

RNA stability

RNA degrades rapidly, requiring careful sample handling.

Host contamination

In host-associated samples (p.e. gut) a large fraction of RNA comes from the host.
Can be difficult to separate in complex samples.

Challenges and limitations: Data complexity

- **Mapping reads**: difficult due to the lack of reference genomes
- **Ambigous reads:** Many genes are conserved across species, paralogous genes contain high sequence similarity
- **Normalization**: Transcript abundance is influenced by both gene expression and organism abundance (+ technical artifacts)

Course Contents

1 Basic concepts of metatranscriptomics

2 **Experimental design**

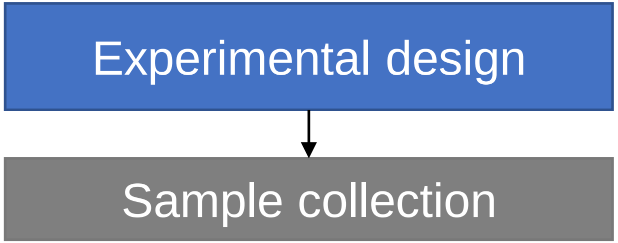
3 Downstream analysis

Section 2.1.

Experimental design

Sampling and RNA extraction

Experimental design: Sample collection



Experimental design

Sample collection

- What type of sample will I collect?
- Does the time or condition of collection matter?
- How can I avoid contamination?
- How much material is needed?

Experimental design: Sample collection

Environmental samples

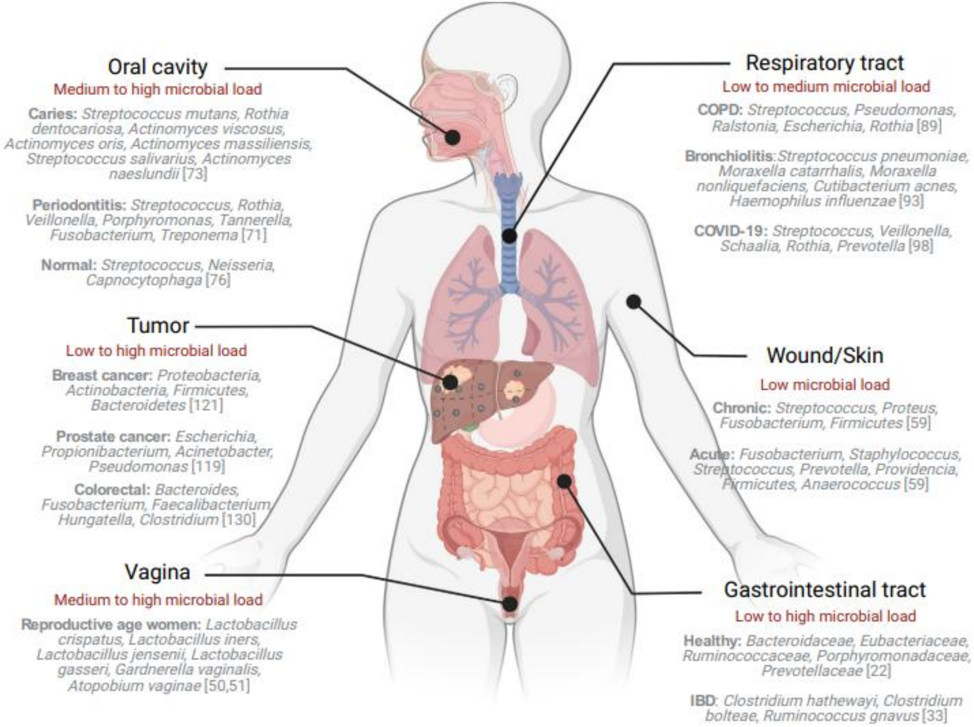
Sediment

Soil

Water

Experimental design: Sample collection

Host-associated biological samples



Oral cavity

Medium to high microbial load

Caries: *Streptococcus mutans, Rothia dentocariosa, Actinomyces viscosus, Actinomyces oris, Actinomyces massiliensis, Streptococcus salivarius, Actinomyces naeslundii* [73]

Periodontitis: *Streptococcus, Rothia, Veillonella, Porphyromonas, Tannerella, Fusobacterium, Treponema* [71]

Normal: *Streptococcus, Neisseria, Capnocytophaga* [76]

Tumor

Low to high microbial load

Breast cancer: *Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes* [121]

Prostate cancer: *Escherichia, Propionibacterium, Acinetobacter, Pseudomonas* [119]

Colorectal: *Bacteroides, Fusobacterium, Faecalibacterium, Hungatella, Clostridium* [130]

Vagina

Medium to high microbial load

Reproductive age women: *Lactobacillus crispatus, Lactobacillus iners, Lactobacillus jensenii, Lactobacillus gasseri, Gardnerella vaginalis, Atopobium vaginae* [50,51]

Respiratory tract

Low to medium microbial load

COPD: *Streptococcus, Pseudomonas, Ralstonia, Escherichia, Rothia* [89]

Bronchiolitis:*Streptococcus pneumoniae, Moraxella catarrhalis, Moraxella nonliquefaciens, Cutibacterium acnes, Haemophilus influenzae* [93]

COVID-19: *Streptococcus, Veillonella, Schaalia, Rothia, Prevotella* [98]

Wound/Skin

Low microbial load

Chronic: *Streptococcus, Proteus, Fusobacterium, Firmicutes* [59]

Acute: *Fusobacterium, Staphylococcus, Streptococcus, Prevotella, Providencia, Firmicutes, Anaerococcus* [59]

Gastrointestinal tract

Low to high microbial load

Healthy: *Bacteroidaceae, Eubacteriaceae, Ruminococcaceae, Porphyromonadaceae, Prevotellaceae* [22]

IBD: *Clostridium hathewayi, Clostridium bolteae, Ruminococcus gnavus* [33]

Ojala, T., Kankuri, E., & Kankainen, M. (2023). Understanding human health through metatranscriptomics. Trends in Molecular Medicine, 29(5), 376–389. https://doi.org/10.1016/j.molmed.2023.02.002

Experimental design: Sample collection

Select the study area for sampling

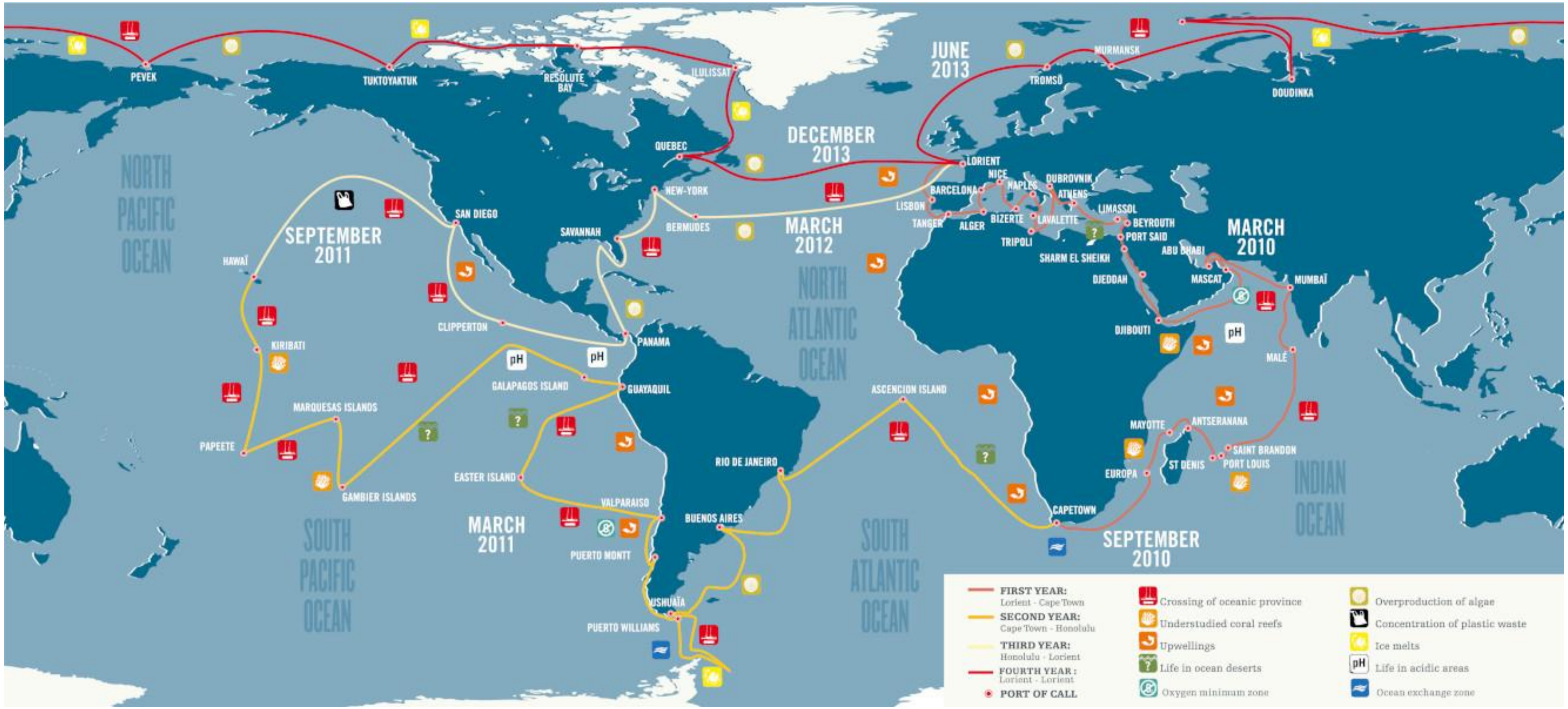
1- In-depth study of the **maps and satellite data** to identify the schooner’s precise station.

2- Certain regions are chosen on the basis of observable **phenomena**, such as phytoplankton blooms.

3- Choose **appropriate tool** for the type of sampling required.

Experimental design: Sample collection

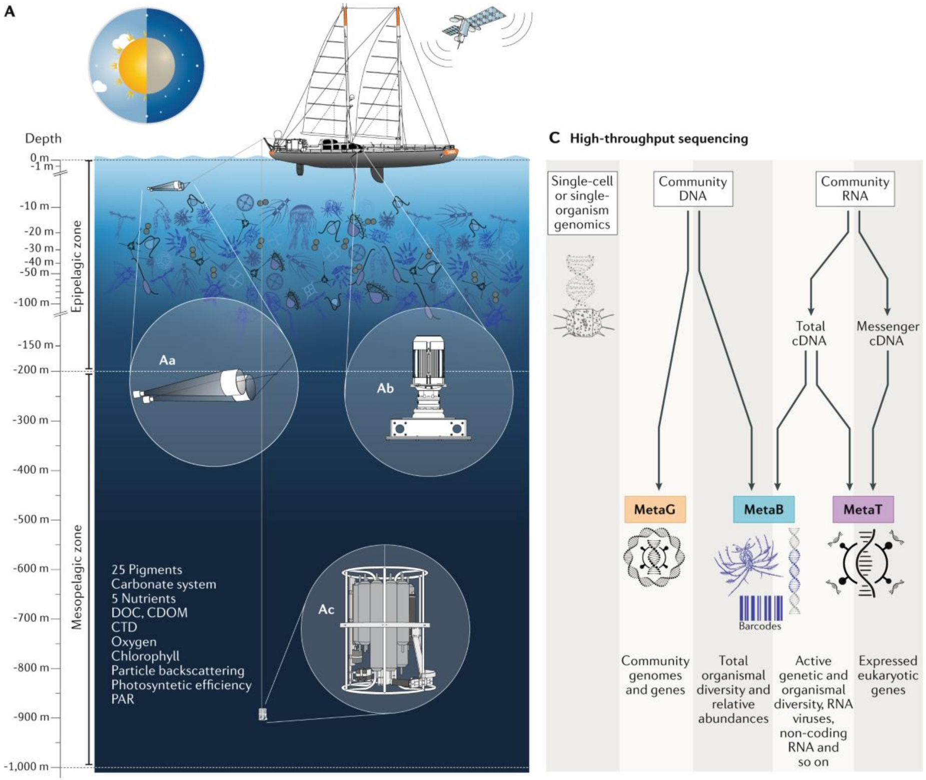
Practical example: TARA Oceans & marine metatranscriptomics



https://www.encyclopedie-environnement.org/app/uploads/2021/02/Tara-Expedition_fig3-trajet-schema.jpg

Experimental design: Sample collection

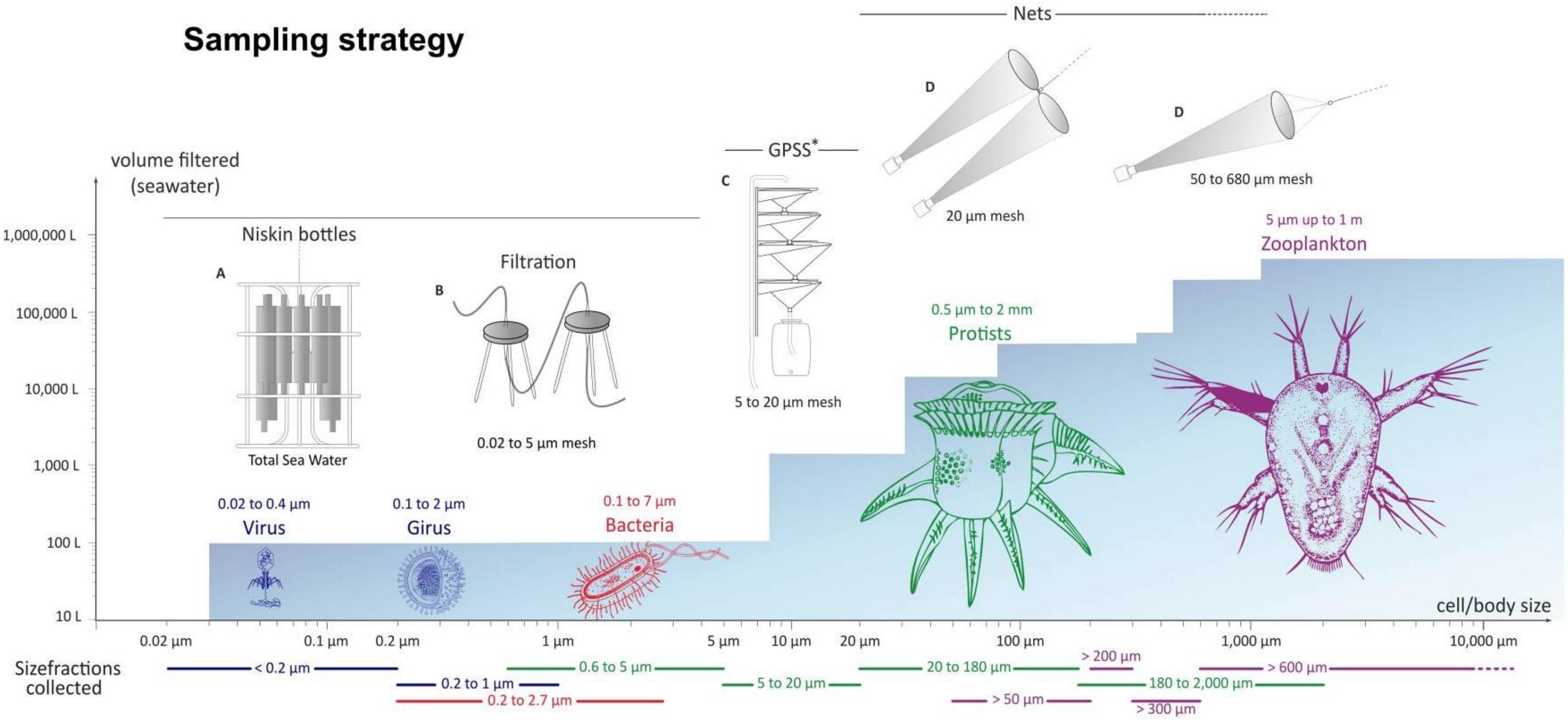
Practical example: TARA Oceans & marine metatranscriptomics



Bringing the Regent net back on board – © Maeva Bardy, Tara Ocean Foundation

Net sampling on board Tara – © Louise Cognard, Tara Ocean Foundation

Experimental design: Sample collection

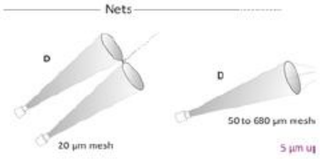


Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., . . . Wincker, P. (2011). A holistic approach to Marine Eco-Systems Biology. PLoS Biology, 9(10), e1001177. https://doi.org/10.1371/journal.pbio.1001177

Experimental design: Sample collection

Niskin bottles
A
Total Sea Water

GPSS*
C
5 to 20 µm mesh

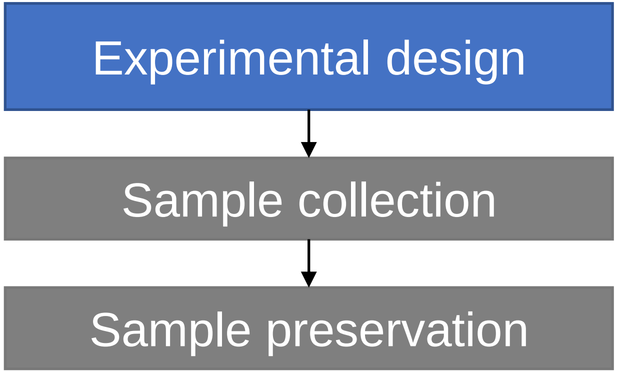


Nets 20-180µm ; >180µm

GPSS
« Gravity Plankton Sieving System »
0.8-5µm ; 5-20µm

Niskin bottles <1µm

Experimental design: Sample preservation



- What type of sample will I collect?
- Does the time or condition of collection matter?
- How can I avoid contamination
- How much material is needed?

- How will I stabilize RNA after collection?

Experimental design: Sample preservation

Preservation methods to maintain RNA integrity

Stabilize against RNA degradation:
- mRNA has a short half-life (minutes)
- Endogenous RNases

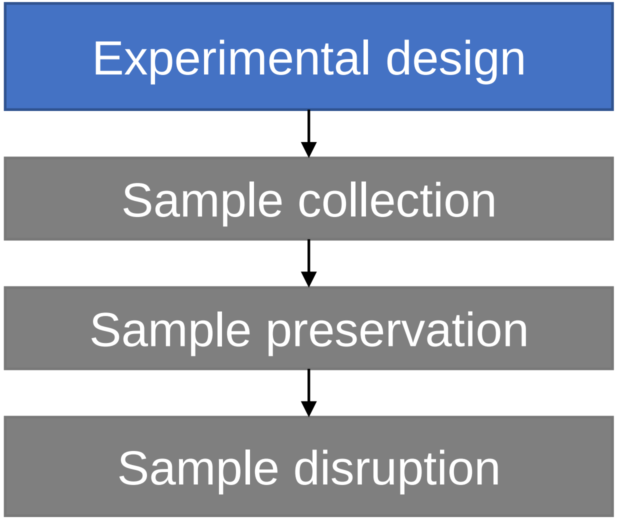
Stabilize gene expression profiles :
- Transport and handling can alter them

- Flash freezing
- RNAlater®
- RNAprotect®
- Phenol-Ethanol
- TRIzol®, QIAzol®, TRI®

Transferring samples to liquid nitrogen – © Leslie Moquin, Tara Ocean Foundation

Flash freezing -> Filters stored at -80ºC -> dry ice for shipping

Experimental design: Sample disruption



- What type of sample will I collect?
- Does the time or condition of collection matter?
- How can I avoid contamination
- How much material is needed?

- How will I stabilize RNA after collection?

- Is my priority RNA yield (qualitative) or reproducibility (quantitative)?

Experimental design: Sample disruption

Main methods:

Mechanical disruption (bead beater)
Enzymatic lysis (lysozyme or lysostaphin)
Proteinase K digestion

Recommendations:

- For **qualitative studies**: Combine all lysis methods to maximize RNA recovery.

As much diversity as possible

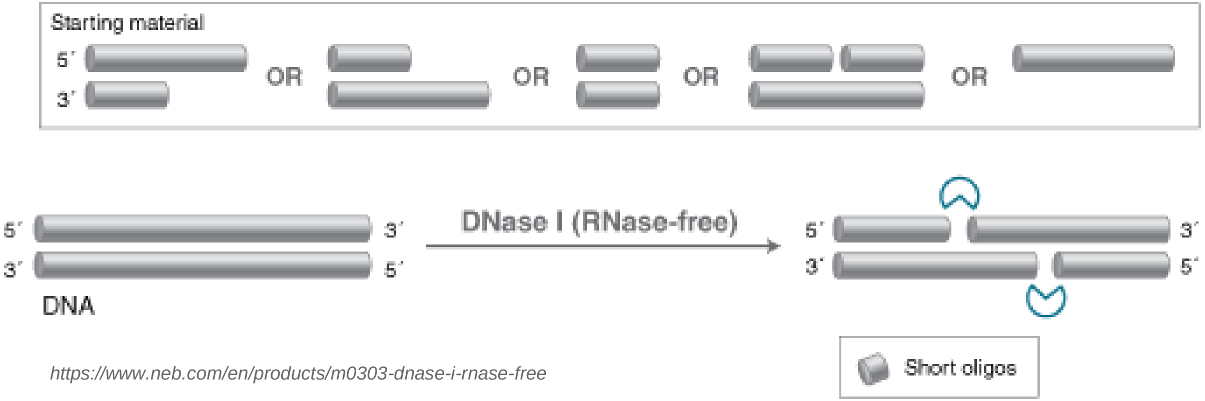
- For **quantitative studies**: Prefer mechanical methods, that are more reproducible.

Reproducibility

Experimental design: DNase treatment

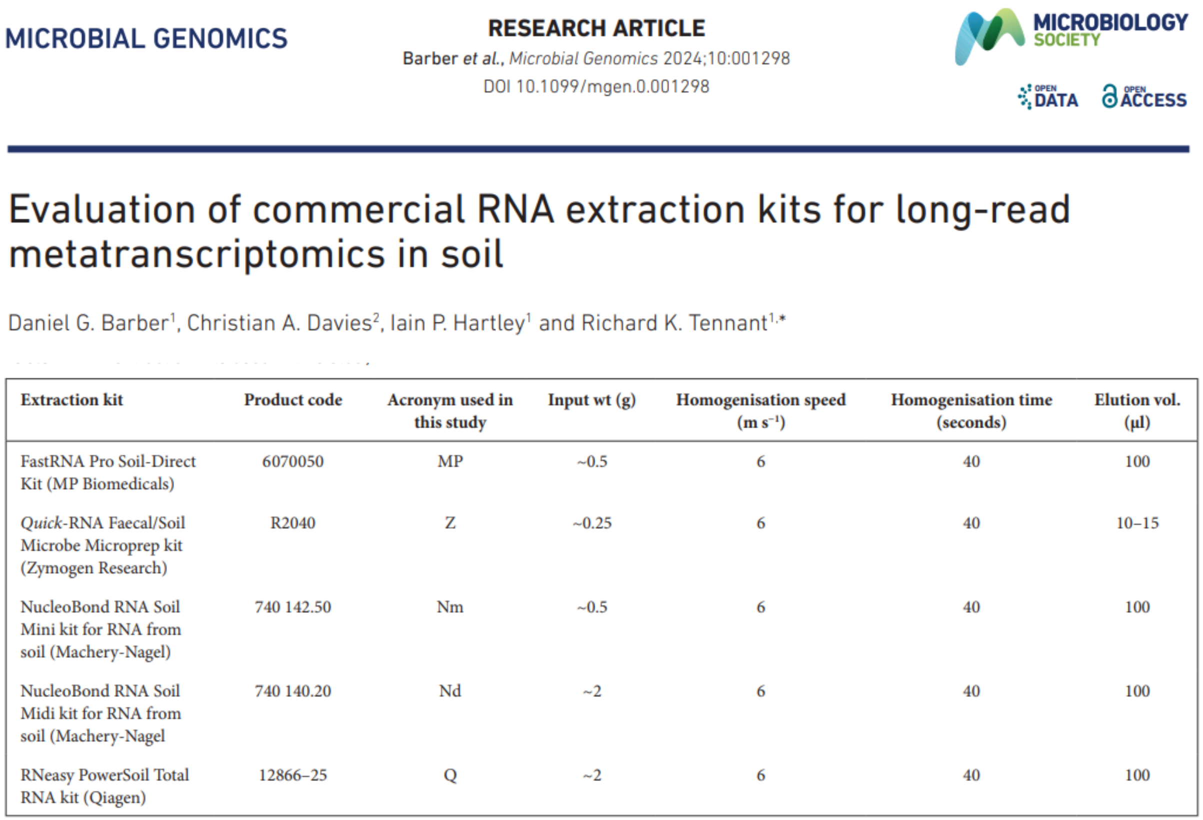
During cell lysis, both DNA and RNA are released simultaneously. Therefore, to obtain high-quality RNA, it is essential to use specific enzymes—such as DNase I—to remove any residual DNA from the sample.

DNase I can degrade double-stranded and single-stranded DNA.



https://www.neb.com/en/products/m0303-dnase-i-rnase-free

Experimental design: RNA Extraction



MICROBIAL GENOMICS

RESEARCH ARTICLE
Barber *et al.*, *Microbial Genomics* 2024;10:001298
DOI 10.1099/mgen.0.001298

MICROBIOLOGY SOCIETY

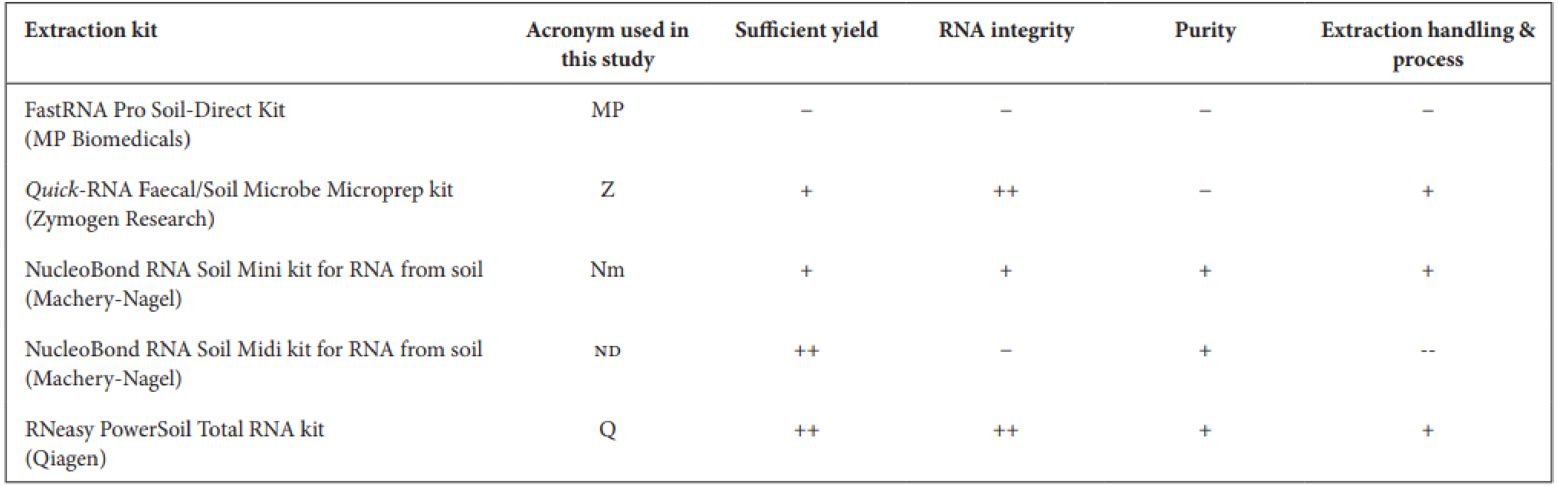
OPEN DATA OPEN ACCESS

Evaluation of commercial RNA extraction kits for long-read metatranscriptomics in soil

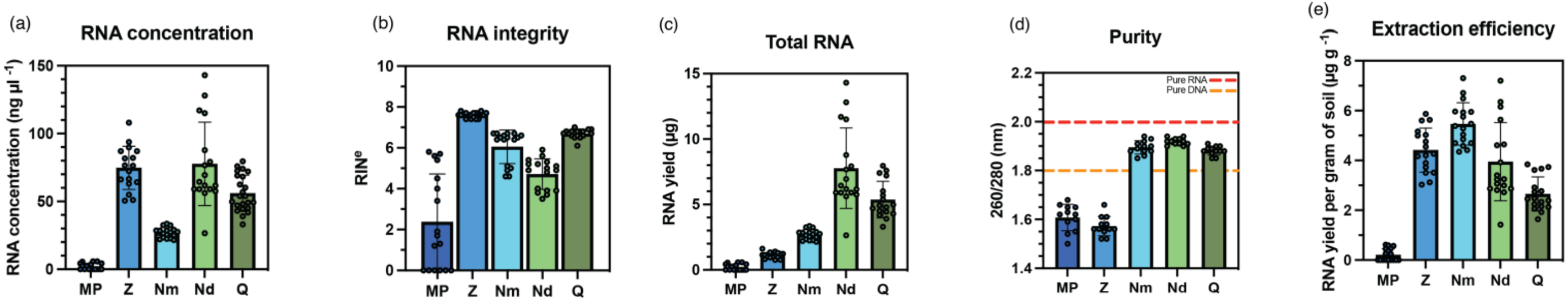
Daniel G. Barber[1], Christian A. Davies[2], Iain P. Hartley[1] and Richard K. Tennant[1,*]

Extraction kit	Product code	Acronym used in this study	Input wt (g)	Homogenisation speed (m s^{-1})	Homogenisation time (seconds)	Elution vol. (µl)
FastRNA Pro Soil-Direct Kit (MP Biomedicals)	6070050	MP	~0.5	6	40	100
Quick-RNA Faecal/Soil Microbe Microprep kit (Zymogen Research)	R2040	Z	~0.25	6	40	10–15
NucleoBond RNA Soil Mini kit for RNA from soil (Machery-Nagel)	740 142.50	Nm	~0.5	6	40	100
NucleoBond RNA Soil Midi kit for RNA from soil (Machery-Nagel	740 140.20	Nd	~2	6	40	100
RNeasy PowerSoil Total RNA kit (Qiagen)	12866–25	Q	~2	6	40	100

Experimental design: RNA Extraction

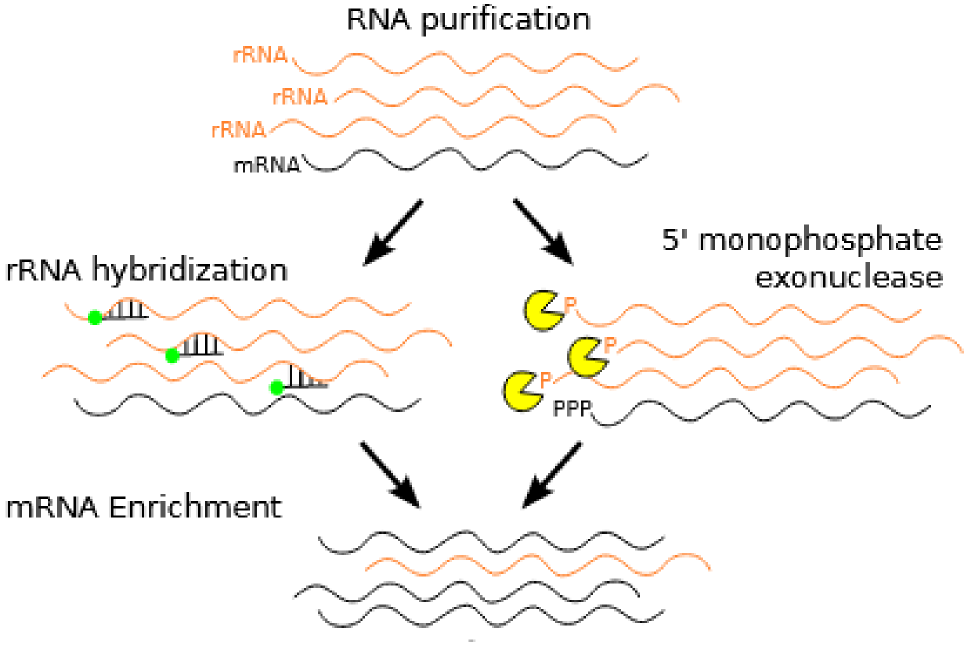


Extraction kit	Acronym used in this study	Sufficient yield	RNA integrity	Purity	Extraction handling & process
FastRNA Pro Soil-Direct Kit (MP Biomedicals)	MP	–	–	–	–
Quick-RNA Faecal/Soil Microbe Microprep kit (Zymogen Research)	Z	+	++	–	+
NucleoBond RNA Soil Mini kit for RNA from soil (Machery-Nagel)	Nm	+	+	+	+
NucleoBond RNA Soil Midi kit for RNA from soil (Machery-Nagel)	ND	++	–	+	--
RNeasy PowerSoil Total RNA kit (Qiagen)	Q	++	++	+	+



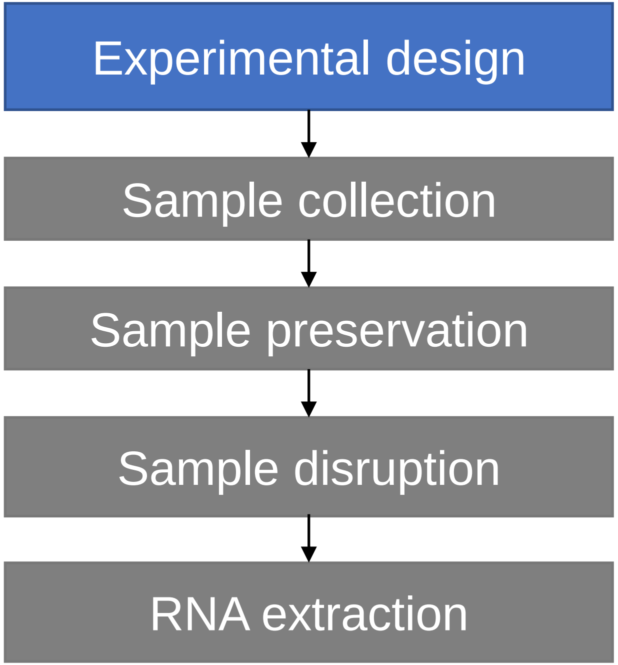
Experimental design: mRNA enrichment

RNA extraction + rRNA depletion



Adapted from Aransay, A. M., & Trueba, J. L. L. (2016). Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing. In Springer eBooks. https://doi.org/10.1007/978-3-319-31350-4

Experimental design: RNA extraction



- What type of sample will I collect?
- Does the time or condition of collection matter?
- How can I avoid contamination?
- How much material is needed?

- How will I stabilize RNA after collection?

- Is my priority RNA yield (qualitative) or reproducibility (quantitative)?

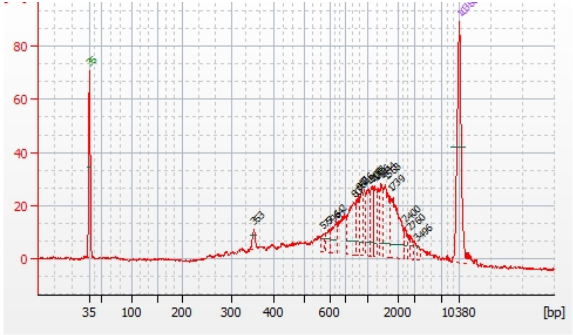
- Should rRNA depletion be performed as part of the RNA extraction process?

Experimental design: RNA quality assesment

RIN and Fragment size distribution

Methods to check RNA quality:

- Femto Pulse: Ultra-sensitive capillary electrophoresis, ultra low input RNA
- Bioanalyzer: Microfluidic electrophoresis
- TapeStation: Automated capillary electrophoresis, high-throughtput



Bioanalyzer

Section 2.2.

Experimental design

Library preparation and RNA-Seq platform selection

RNA-sequencing: What platform is the best for metatranscriptomics?

There is no easy answer. Trade-off between:

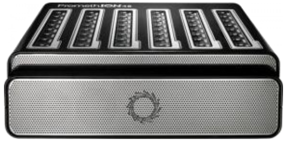
- Overall cost   How much money do you have?

RNA-sequencing: What platform is the best for metatranscriptomics?

There is no easy answer. Trade-off between:

- Overall cost
- Read length

What is your desired read length?



RNA-sequencing: What platform is the best for metatranscriptomics?

There is no easy answer. Trade-off between:

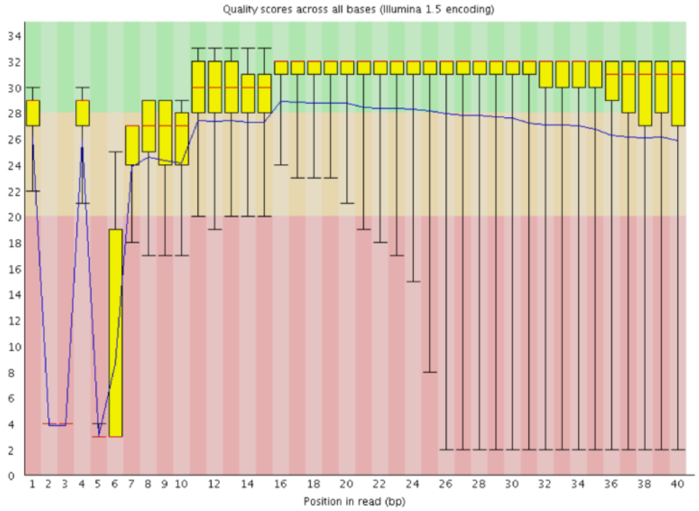
- Overall cost
- Read length
- Sequencing depth

How many reads do you need per sample?

RNA-sequencing: What platform is the best for metatranscriptomics?

There is no easy answer. Trade-off between:

- Overall cost
- Read length
- Sequencing depth
- Quality of the data



RNA-sequencing: What platform is the best for metatranscriptomics?

There is no easy answer. Trade-off between:

- Overall cost
- Read length
- Sequencing depth
- Quality of the data
- Support for the available technology

RNA-sequencing: What platform is the best for metatranscriptomics?

There is no easy answer. Trade-off between:

- Overall cost
- Read length
- Sequencing depth
- Quality of the data
- Support for the available technology

Consider also the mix of two strategies (p.e. ONT and Illumina)

Section 2.3.

Experimental design

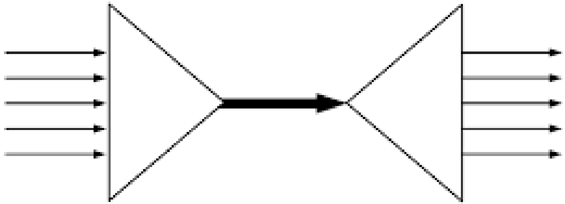
RNA quantity and multiplexing

Experimental design: RNA quantity

Platform	RNA Input	Notes
Illumina	1-1000ng (Illumina Stranded Total RNA Prep)	Can work with low input, Illumina Total RNA Prep with Ribo-Zero Plus
ONT (Nanopore)	300 ng polyA (direct RNA) 1 µg total RNA (direct RNA) 10 ng polyA (cDNA) 500 ng total RNA (cDNA)	Direct RNA requires more input; cDNA kits are more flexible
PacBio	300ng	Less suitable for low-input samples

Experimental design: Multiplexing

Sequencing adapters can include a **barcode** that serves to identify the sample if several samples are mixed in the same run (multiplexing).



RNA-sequencing: Short-reads vs long-reads

Challenges

Short-reads:

- Fragmented reads
- Difficult transcript assembly specially for diverse microbial communities

Long-reads:

- Higher cost, limiting the scale of the experiments
- Sequencing errors are particularly problematic for metatranscriptomic samples
- Lower sequencing depth

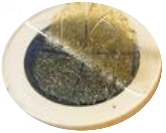
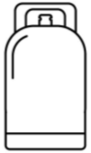
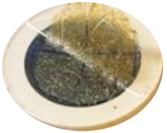
RNA-sequencing: Short-reads vs long-reads

While long-read sequencing platforms generally offer lower sequencing depth compared to short-read platforms, they allow the **direct capture of full-length transcripts**, including those from **low-abundance genes or rare microorganisms**.

In contrast, short-read data requires **assembly or transcript inference**, which often leads to the **loss of rare -** especially problematic in **metatranscriptomics**, where transcriptomes are highly complex.

Therefore, even with fewer reads, **long reads can more faithfully represent transcript diversity.**

Recap: Metatranscriptomics workflow

1- SAMPLING AND STORAGE	2- RNA EXTRACTION	3- METATRANSCRIPTOMIC RAW DATA
Environmental community Sampling   Storage	 DNase I treatment Cell wall lysis RNA Extraction (& mRNA enrichment) RNA QC check	 Library preparation RNA-sequencing

Questions

What is the main advantages of long-read sequencing over short-read sequencing in metatranscriptomics?

In metatranscriptomic studies, what is the key difference between qualitative and quantitative experiments?

Why do you think that sequencing depth is particularly important in metatranscriptomics compared to single-organism transcriptomics?

Course Contents

1. Basic concepts of metatranscriptomics
2. Experimental design
3. **Downstream analysis**
4.

Section 3.1.

Downstream analysis

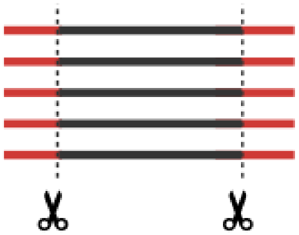
Quality control of raw reads

Downstream analysis: Quality control of raw reads

Trimming

Reads can contain adapter sequences, primers, barcodes.

Trimming help to retain **high-confidence sequences** that will map more accurately and uniquely to references.

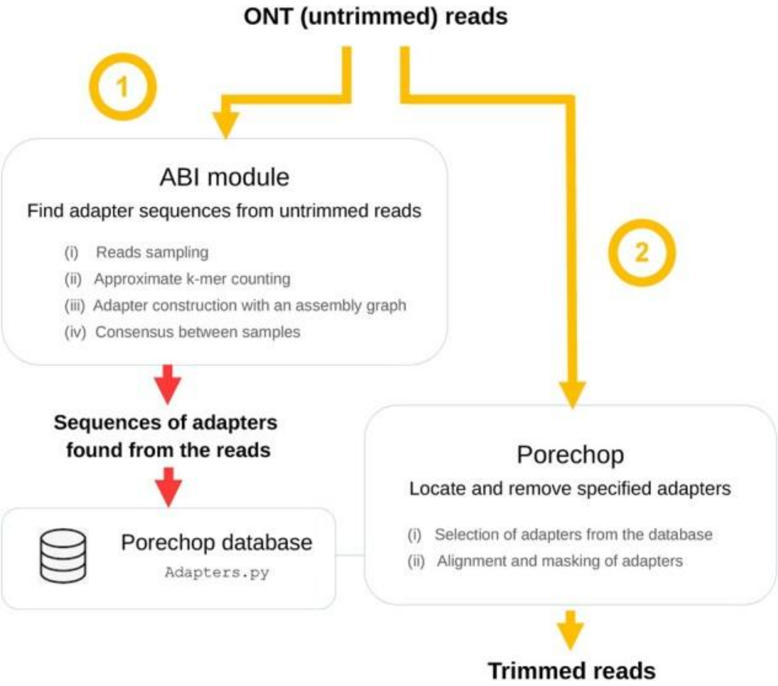


Downstream analysis: Quality control of raw reads

Trimming

Tools:

- Porechop (outdated)
- Pychopper
- Porechop-abi
- Dorado trim



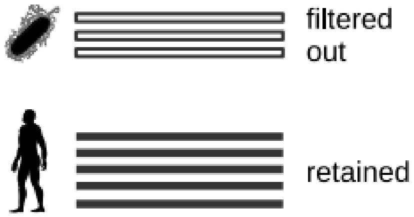
ONT (untrimmed) reads
1
ABI module
Find adapter sequences from untrimmed reads
(i) Reads sampling
(ii) Approximate k-mer counting
(iii) Adapter construction with an assembly graph
(iv) Consensus between samples
2
Sequences of adapters found from the reads
Porechop database
Adapters.py
Porechop
Locate and remove specified adapters
(i) Selection of adapters from the database
(ii) Alignment and masking of adapters
Trimmed reads

Bonenfant, Q., Noé, L., & Touzet, H. (2022). Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. Bioinformatics Advances, 3(1). https://doi.org/10.1093/bioadv/vbac085

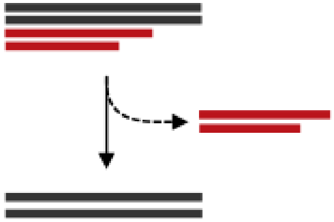
Downstream analysis: Quality control of raw reads

Additional filtering

- Host contamination



- Reads within a length range

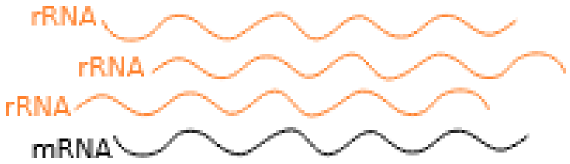


- Others: Mask low complexity regions, remove low quality reads

Downstream analysis: Quality control of raw reads

Removal of rRNA reads computationally

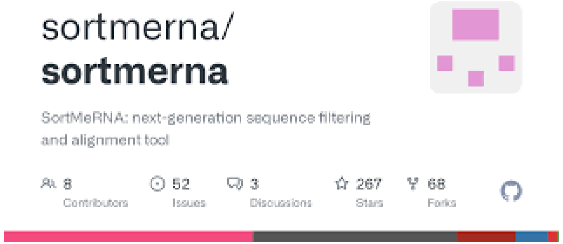
- If we sequence total RNA, most RNA sequences will be ribosomal RNA (rRNA)
- Its important to filter out rRNA before downstream analysis



Downstream analysis: Quality control of raw reads

SortMeRNA: Removal of rRNA reads computationally

- SortMeRNA takes as input files of reads and one or multiple rRNA database file, and sorts apart aligned and rejected reads into two files.
- Can sort a large set of metatranscriptomic reads with high accuracy
- Algorithm implements seeds with errors -> robust to errors of different types of sequencers
- Database can be constructed on any family of sequences provided by the user.



sortmerna/
sortmerna

SortMeRNA: next-generation sequence filtering and alignment tool

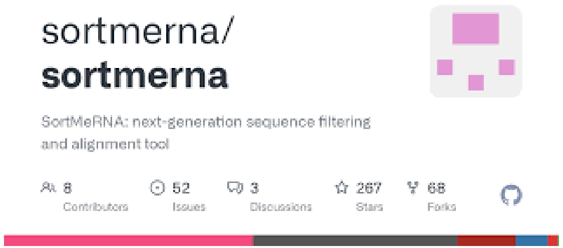
8 Contributors
52 Issues
3 Discussions
267 Stars
68 Forks

Downstream analysis: Quality control of raw reads

LongTREC

SortMeRNA: Removal of rRNA reads computationally

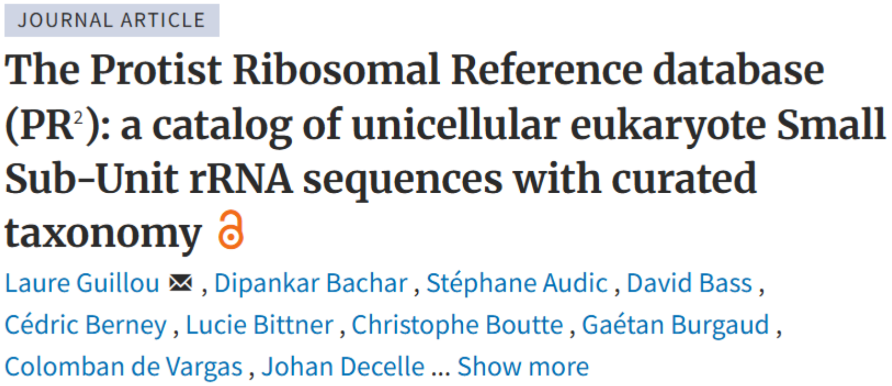
- Database can be constructed on any family of sequences provided by the user.



sortmerna/
sortmerna

SortMeRNA: next-generation sequence filtering and alignment tool

8 Contributors · 52 Issues · 3 Discussions · 267 Stars · 68 Forks



JOURNAL ARTICLE

The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy

Laure Guillou ✉ , Dipankar Bachar , Stéphane Audic , David Bass , Cédric Berney , Lucie Bittner , Christophe Boutte , Gaétan Burgaud , Colomban de Vargas , Johan Decelle ... Show more

Section 3.2.

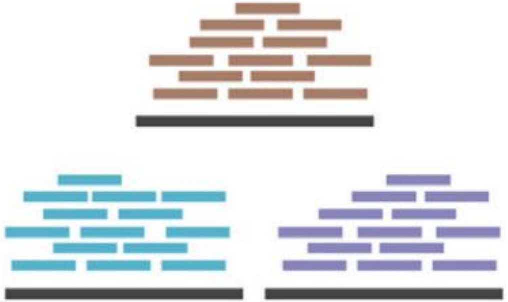
Downstream analysis

Transcript assembly

Downstream analysis: De Novo Assembly

Short reads- **De Novo Assembly**

- Velvet
- SOAPdenovo
- Trinity
- Oases



+ **Cluster** overall results (p.e. cd-hit-est)

Downstream analysis: De Novo Assembly

Long reads- (Not always) De Novo Assembly

Long-read reduce or eliminate the need for de novo assembly

Using raw long reads helps retain a more **faithful representation of the biological sample**, avoiding biases introduced during transcript assembly.

Particularly powerful in complex microbial communities, where assembly can be difficult due to **high diversity and redundancy**.

Section 3.3.

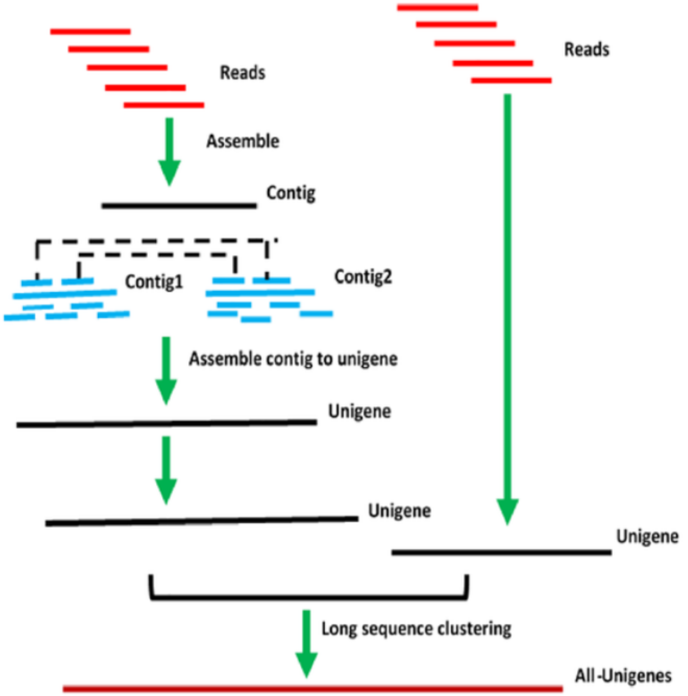
Downstream analysis

Reference sequences for mapping

Mapping on transcripts/proteins catalogs: Unigenes

Non-redundant sets of gene sequences clustered together based on shared sequence similarity

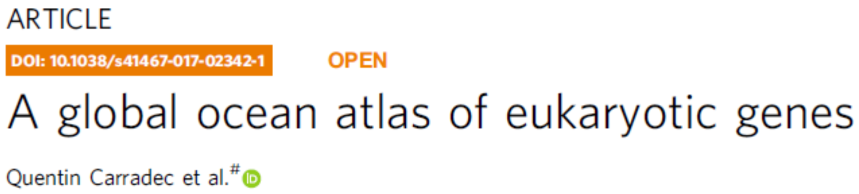
- Functions
- Taxonomy
- Abundance
- Expression



Rasool, K. G., Mehmood, K., Husain, M., Tufail, M., Alwaneen, W. S., & Aldawood, A. S. (2021).

Downstream analysis: Reference sequences for mapping

Mapping on transcripts/proteins catalogs: Unigenes



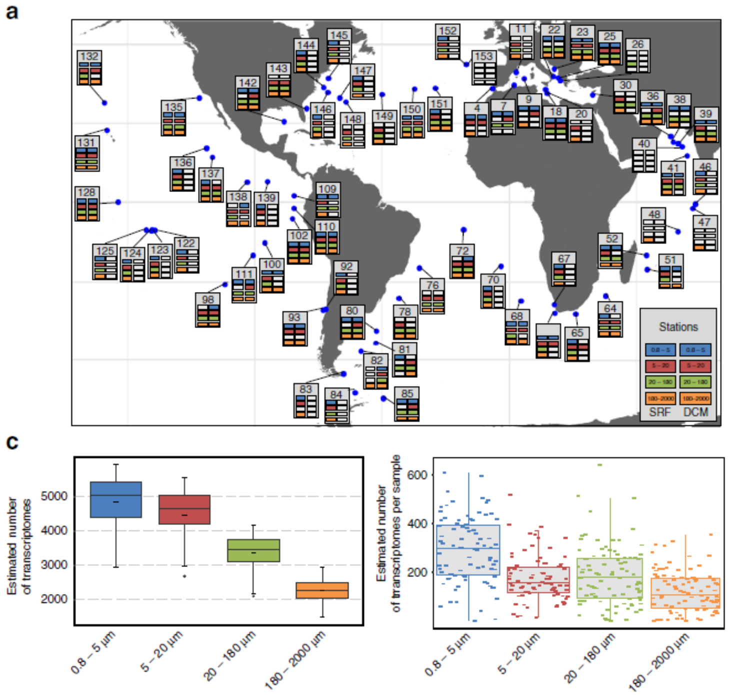
ARTICLE

DOI: 10.1038/s41467-017-02342-1 OPEN

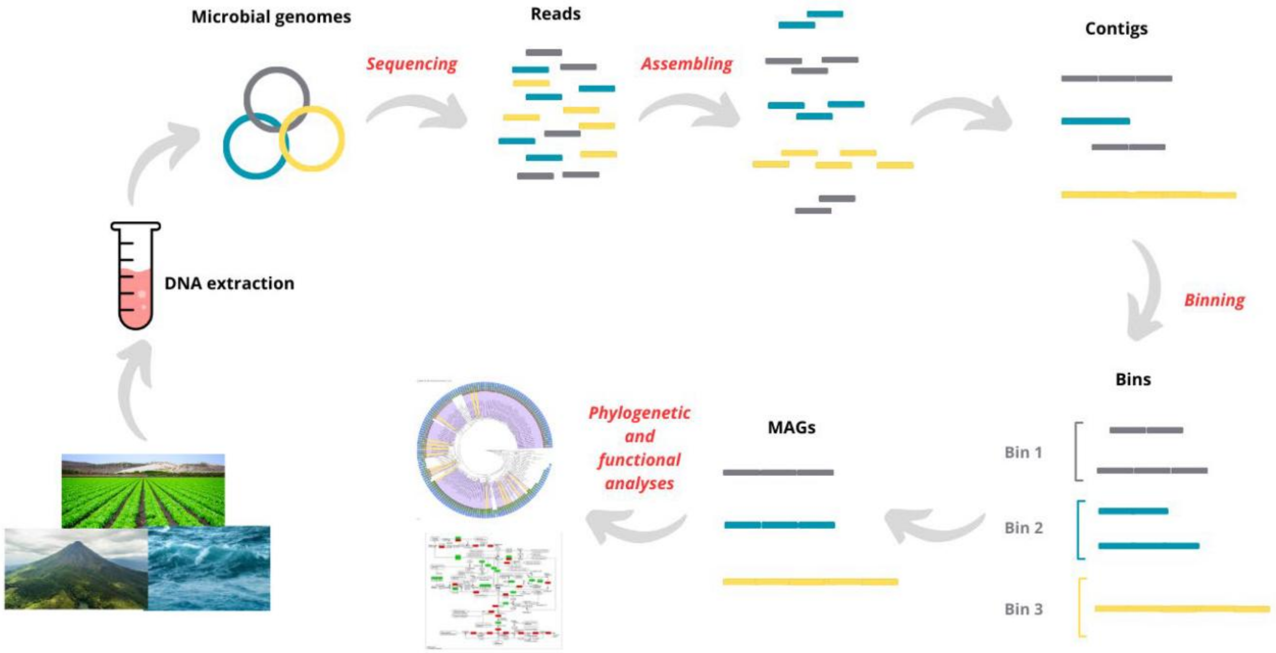
A global ocean atlas of eukaryotic genes

Quentin Carradec et al.#

The individual sequence reads cluster into **116 million unigenes** representing the largest reference collection of eukaryotic transcripts from any single biome.



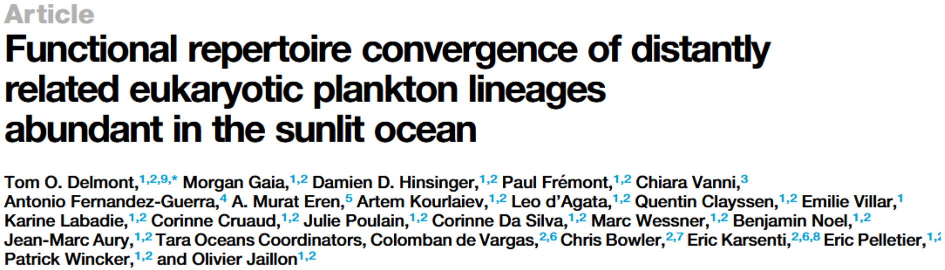
Mapping on « genomes » like Metagenome-Assembled Genomes (MAGs)



Mirete, S., Sánchez-Costa, M., Díaz-Rullo, J., De Figueras, C. G., Martínez-Rodríguez, P., & González-Pastor, J. E. (2025). Metagenome-Assembled Genomes (MAGs): advances, challenges, and ecological insights. Microorganisms, 13(5)

Downstream analysis: Reference sequences for mapping

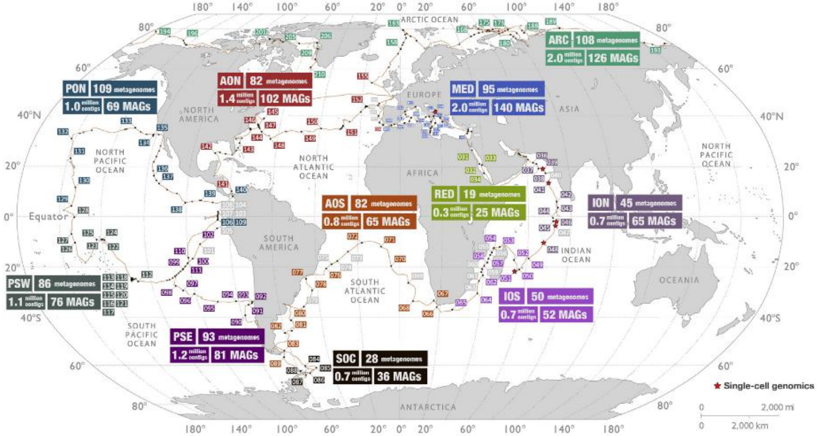
Mapping on « genomes »



Article

Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean

Tom O. Delmont,[1,2,9,*] Morgan Gaia,[1,2] Damien D. Hinsinger,[1,2] Paul Frémont,[1,2] Chiara Vanni,[3] Antonio Fernandez-Guerra,[4] A. Murat Eren,[5] Artem Kourlaiev,[1,2] Leo d'Agata,[1,2] Quentin Clayssen,[1,2] Emilie Villar,[1] Karine Labadie,[1,2] Corinne Cruaud,[1,2] Julie Poulain,[1,2] Corinne Da Silva,[1,2] Marc Wessner,[1,2] Benjamin Noel,[1,2] Jean-Marc Aury,[1,2] Tara Oceans Coordinators, Colomban de Vargas,[2,6] Chris Bowler,[2,7] Eric Karsenti,[2,6,8] Eric Pelletier,[1,2] Patrick Wincker,[1,2] and Olivier Jaillon[1,2]

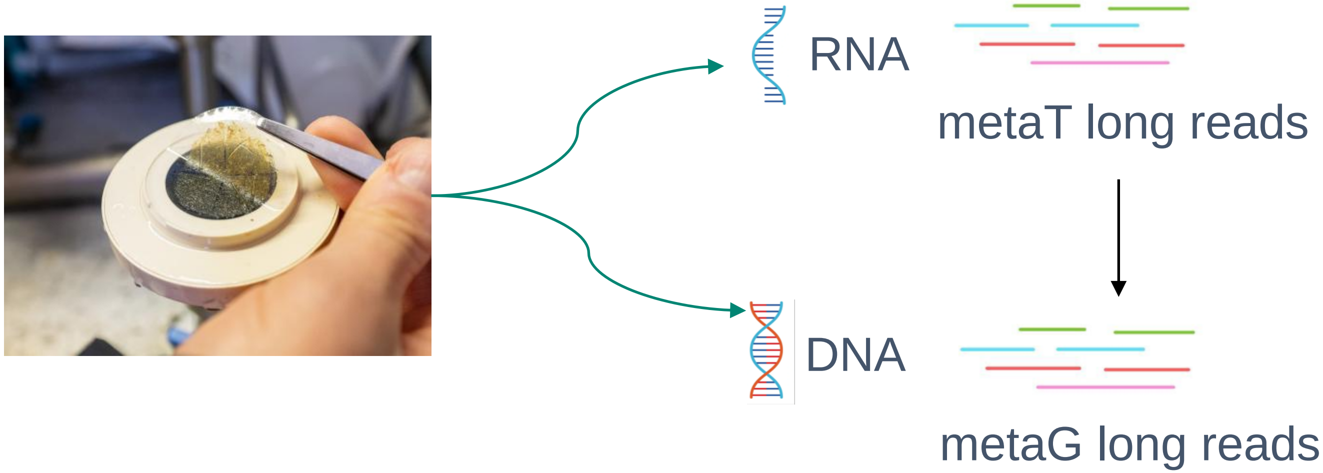

ARCTIC OCEAN
ARC 108 metagenomes 2.0 million contigs 126 MAGs
PON 109 metagenomes 1.0 million contigs 69 MAGs
AON 82 metagenomes 1.4 million contigs 102 MAGs
MED 95 metagenomes 2.0 million contigs 140 MAGs
NORTH AMERICA
EUROPE
ASIA
NORTH PACIFIC OCEAN
NORTH ATLANTIC OCEAN
AFRICA
RED 19 metagenomes 0.3 million contigs 25 MAGs
AOS 82 metagenomes 0.8 million contigs 65 MAGs
ION 45 metagenomes 0.7 million contigs 65 MAGs
Equator
SOUTH AMERICA
INDIAN OCEAN
PSW 86 metagenomes 1.1 million contigs 76 MAGs
SOUTH ATLANTIC OCEAN
OCEANIA
IOS 50 metagenomes 0.7 million contigs 52 MAGs
SOUTH PACIFIC OCEAN
PSE 93 metagenomes 1.2 million contigs 81 MAGs
SOC 28 metagenomes 0.7 million contigs 36 MAGs
★ Single-cell genomics
ANTARCTICA
2,000 mi
2,000 km

683 new eukaryotic genomes of at least 10 Mb in size.

The averaged statistics of the whole dataset were **35.4 Mb in genome size**, about **14,000 genes per MAG**, and a **BUSCO completeness of 40%.**

Downstream analysis: Reference sequences for mapping

Mapping on metagenomic long reads from the same sample



RNA

metaT long reads

DNA

metaG long reads

Metatranscriptomics provides a window into the active fraction of **uncultivable microorganisms**

Section 3.4.

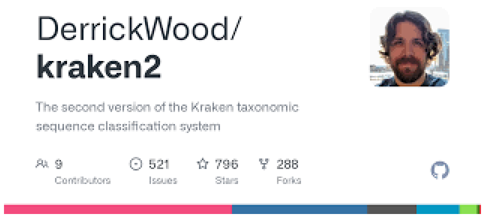
Downstream analysis

Taxonomic and functional characterization

Downstream analysis: Taxonomic and functional analysis

Taxonomic assignment

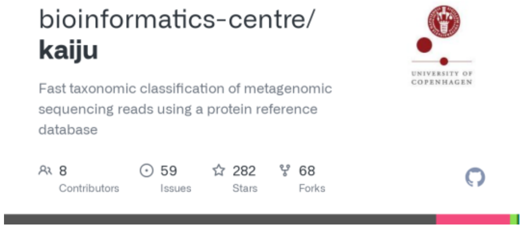
Identify **which organisms** are actively expressing genes in the microbial community.



DerrickWood/
kraken2

The second version of the Kraken taxonomic sequence classification system

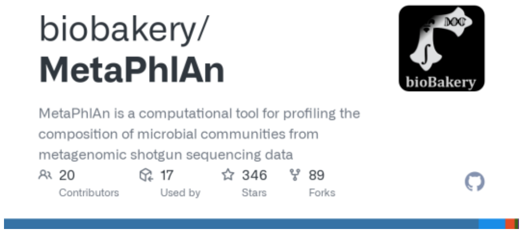
9 Contributors · 521 Issues · 796 Stars · 288 Forks



bioinformatics-centre/
kaiju

Fast taxonomic classification of metagenomic sequencing reads using a protein reference database

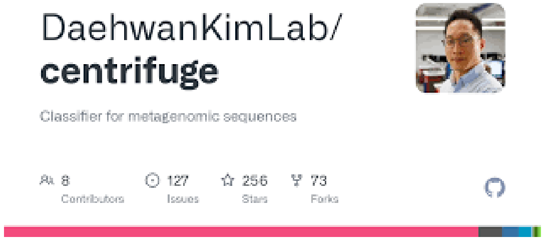
8 Contributors · 59 Issues · 282 Stars · 68 Forks



biobakery/
MetaPhlAn

MetaPhlAn is a computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data

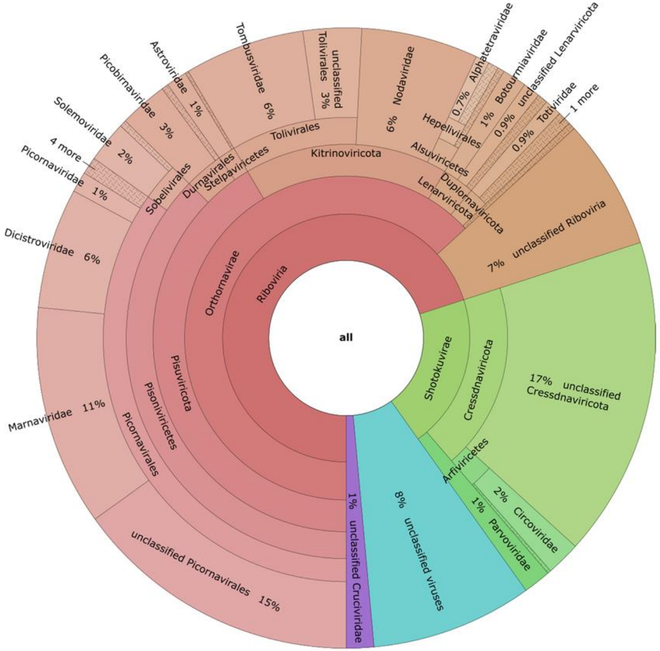
20 Contributors · 17 Used by · 346 Stars · 89 Forks



DaehwanKimLab/
centrifuge

Classifier for metagenomic sequences

8 Contributors · 127 Issues · 256 Stars · 73 Forks

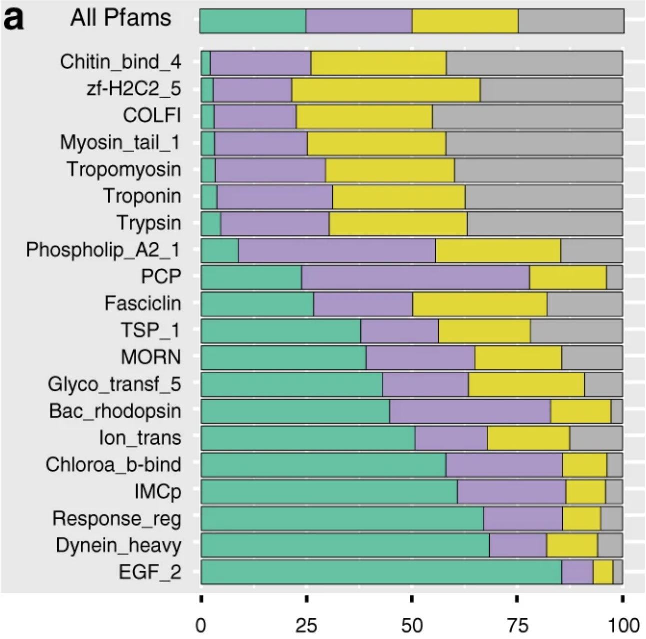


Downstream analysis: Taxonomic and functional characterization

Functional Analysis

Identify the **biological functions** of expressed genes (mRNAs) in the microbial community.

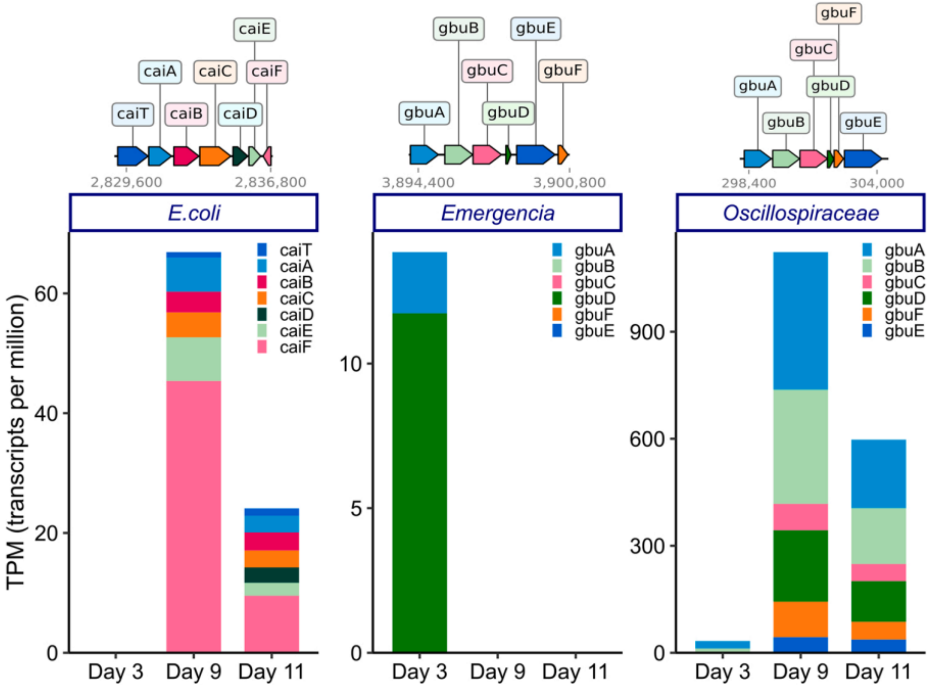
- HUMAnN3
- MEGAN
- DIAMOND
- eggNOG-mapper
- Pfam + HMMER



Downstream analysis

Interpreting metatranscriptomic variation

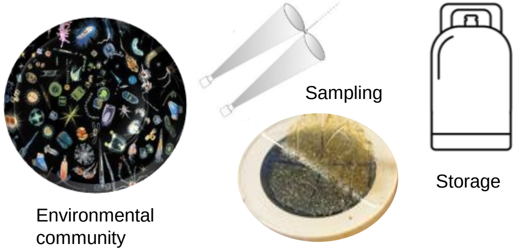
- Alterations in the relative abundance of organisms and their associated genes
- Changes in the expression of genes encoded among the community members



Simó, C., Mamani-Huanca, M., Hernández-Hernández, O., Redondo-Río, Á., Muñoz, S., & García-Cañas, V. (2025).

Metatranscriptomics workflow

1- SAMPLING AND STORAGE



Environmental community

Sampling

Storage

2- RNA EXTRACTION

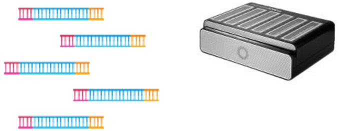
DNase I treatment

Cell wall lysis

RNA Extraction (& mRNA enrichment)

RNA QC check

3- METATRANSCRIPTOMIC RAW DATA

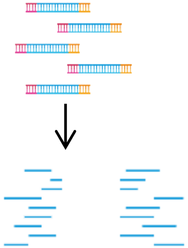


Library preparation

RNA-sequencing

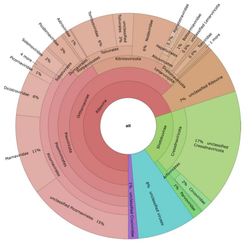
4- PRE-PROCESSING

Trimming
rRNA removal
Host removal
Additional filtering



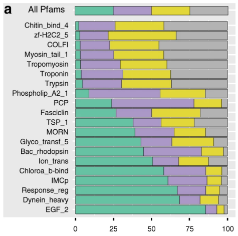
5- TAXONOMIC ASSIGNATION

Active microbial community



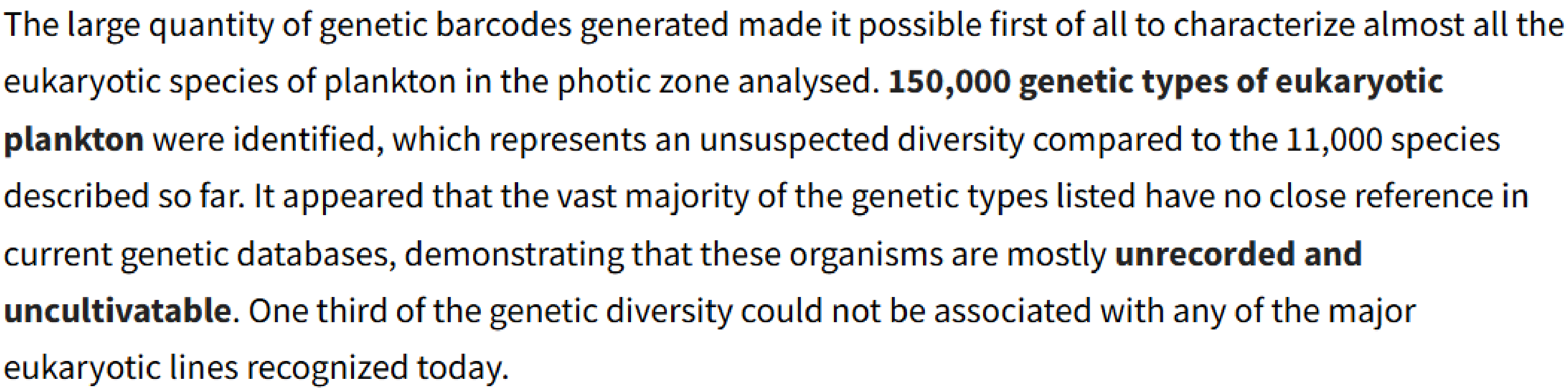
6- FUNCTIONAL CHARACTERIZATION

What is the active microbial community doing



Why do metatranscriptomics important?

Tara Oceans revealed hidden marine diversity

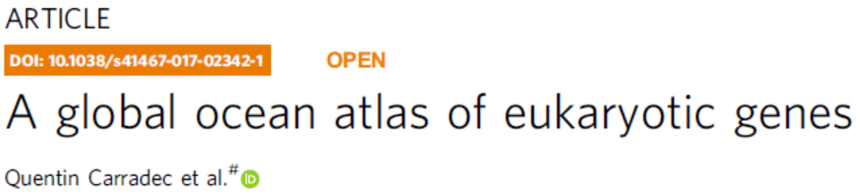


The large quantity of genetic barcodes generated made it possible first of all to characterize almost all the eukaryotic species of plankton in the photic zone analysed. **150,000 genetic types of eukaryotic plankton** were identified, which represents an unsuspected diversity compared to the 11,000 species described so far. It appeared that the vast majority of the genetic types listed have no close reference in current genetic databases, demonstrating that these organisms are mostly **unrecorded and uncultivatable**. One third of the genetic diversity could not be associated with any of the major eukaryotic lines recognized today.

https://www.encyclopedie-environnement.org/en/life/the-tara-oceans-expedition-explores-the-diversity-of-plankton/

Why do metatranscriptomics important?

Tara Oceans revealed hidden marine diversity



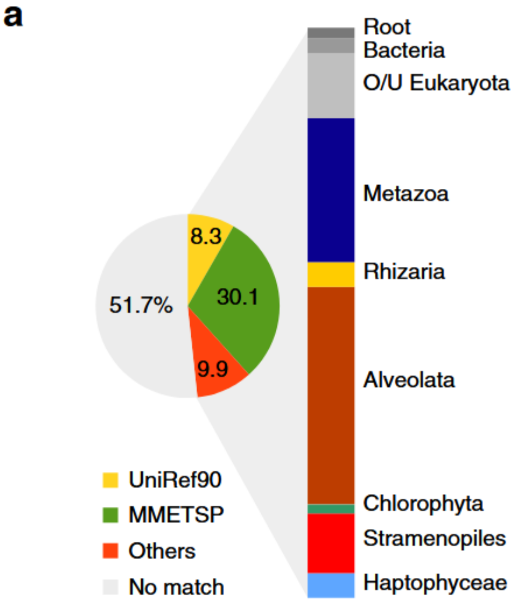
ARTICLE

DOI: 10.1038/s41467-017-02342-1 OPEN

A global ocean atlas of eukaryotic genes

Quentin Carradec et al.#

- 116 million unigenes
- N50 ~ 650bp
- 51.2% of unigenes have no matches in public sequence databases



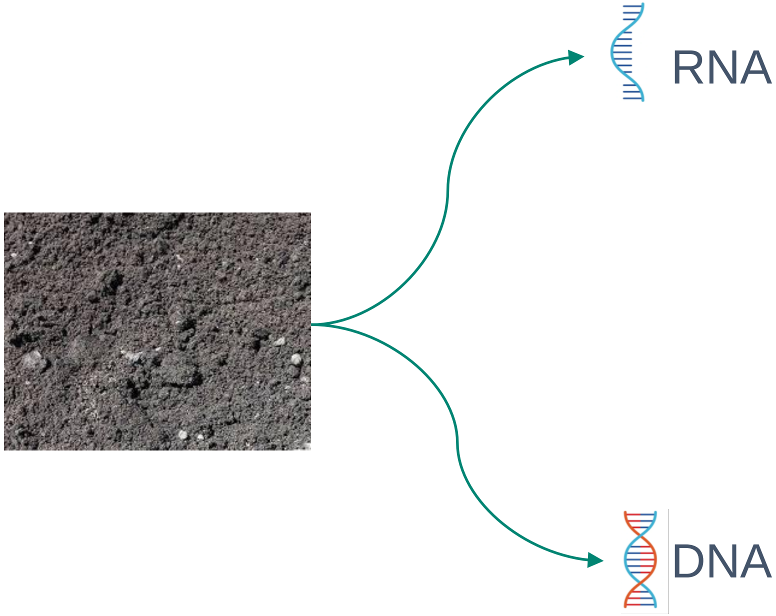
Thank You!

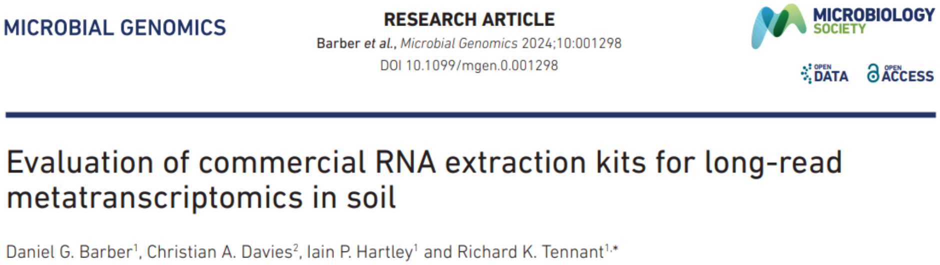
LongTREC

For more information about the LongTREC Summer School:

https://longtrec.eu

Practice: Dataset overview





MICROBIAL GENOMICS

RESEARCH ARTICLE

Barber *et al.*, *Microbial Genomics* 2024;10:001298

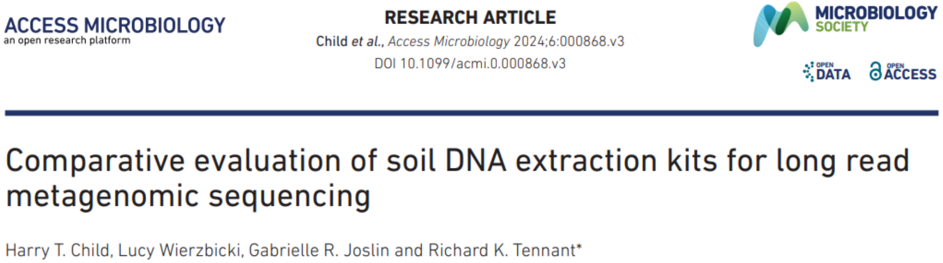
DOI 10.1099/mgen.0.001298

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Evaluation of commercial RNA extraction kits for long-read metatranscriptomics in soil

Daniel G. Barber[1], Christian A. Davies[2], Iain P. Hartley[1] and Richard K. Tennant[1,*]



ACCESS MICROBIOLOGY
an open research platform

RESEARCH ARTICLE

Child *et al.*, *Access Microbiology* 2024;6:000868.v3

DOI 10.1099/acmi.0.000868.v3

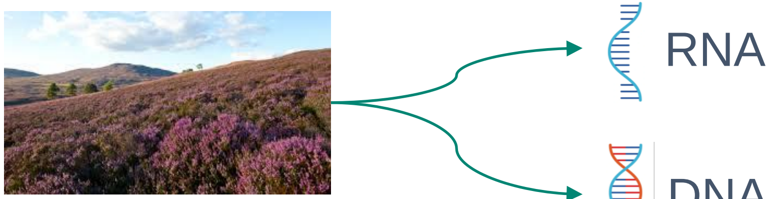
MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Comparative evaluation of soil DNA extraction kits for long read metagenomic sequencing

Harry T. Child, Lucy Wierzbicki, Gabrielle R. Joslin and Richard K. Tennant*

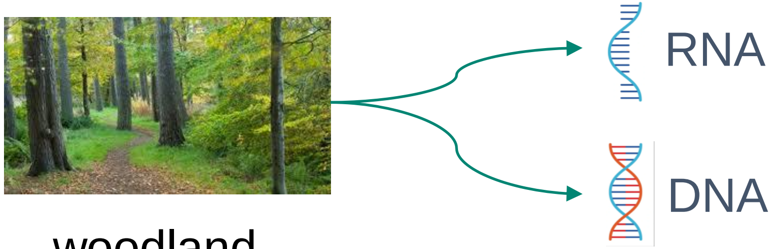
Practice: Dataset overview



heathland

Heath metaT long reads

Heath metaG long reads



woodland

Wood metaT long reads

Wood metaG long reads

Practice

```
conda activate metatranscriptomics

cd /home/train/longTREC/day4/notebooks/metatranscriptomics

jupyter notebook
```