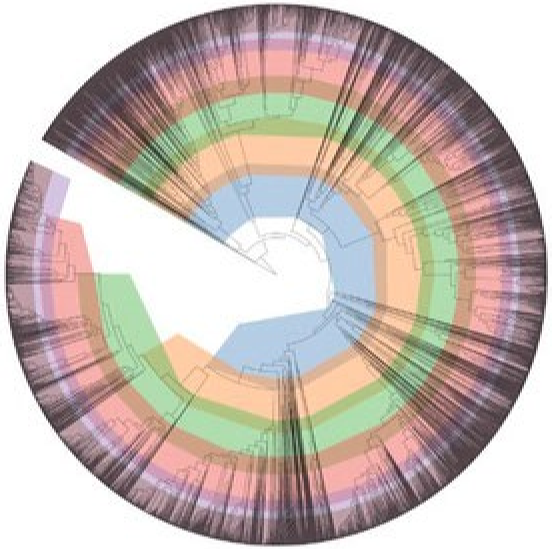
ENVIRONMENTAL METAGENOMICS

Physalia course, online, 13-17 October 2025

Course outline and practical information

Nikolay Oskolkov, Group Leader of Metabolic Research Group at LIOS, Riga, Latvia
Samuel Aroney, Postdoctoral Research Fellow, Queensland University of Technology



Physalia
Courses

NB: original course material courtesy:
Dr. Antti Karkman, University of Helsinki
Dr. Igor Pessi, Finnish Environment Institute (SYKE)

About us

Organizer: Carlo Pecoraro, Physalia courses

info@physalia-courses.org

Instructors:

Dr. Nikolay Oskolkov, Group Leader at LIOS, Riga, Latvia

nikolay.oskolkov@osi.lv

Dr. Samuel Aroney, Queensland University of Technology

samuel.aroney@qut.edu.au

Brief introduction: who am I

@NikolayOskolkov

@oskolkov.bsky.social

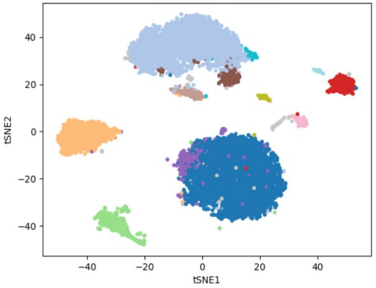
Personal homepage:
https://nikolay-oskolkov.com

2007 PhD in theoretical physics at MSU

2011 medical genetics at Lund University

2016 working at NBIS SciLifeLab, Sweden

Single cell



Biomedical data integration

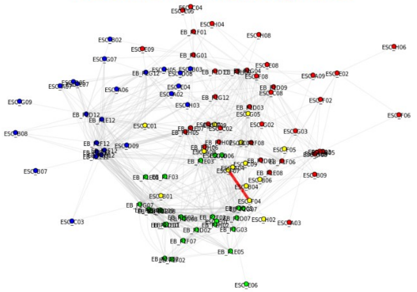
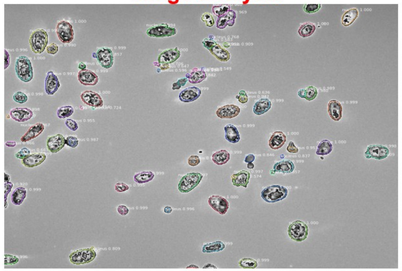
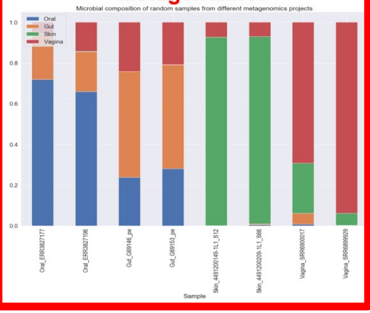


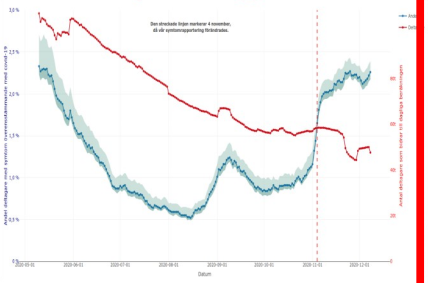
Image analysis



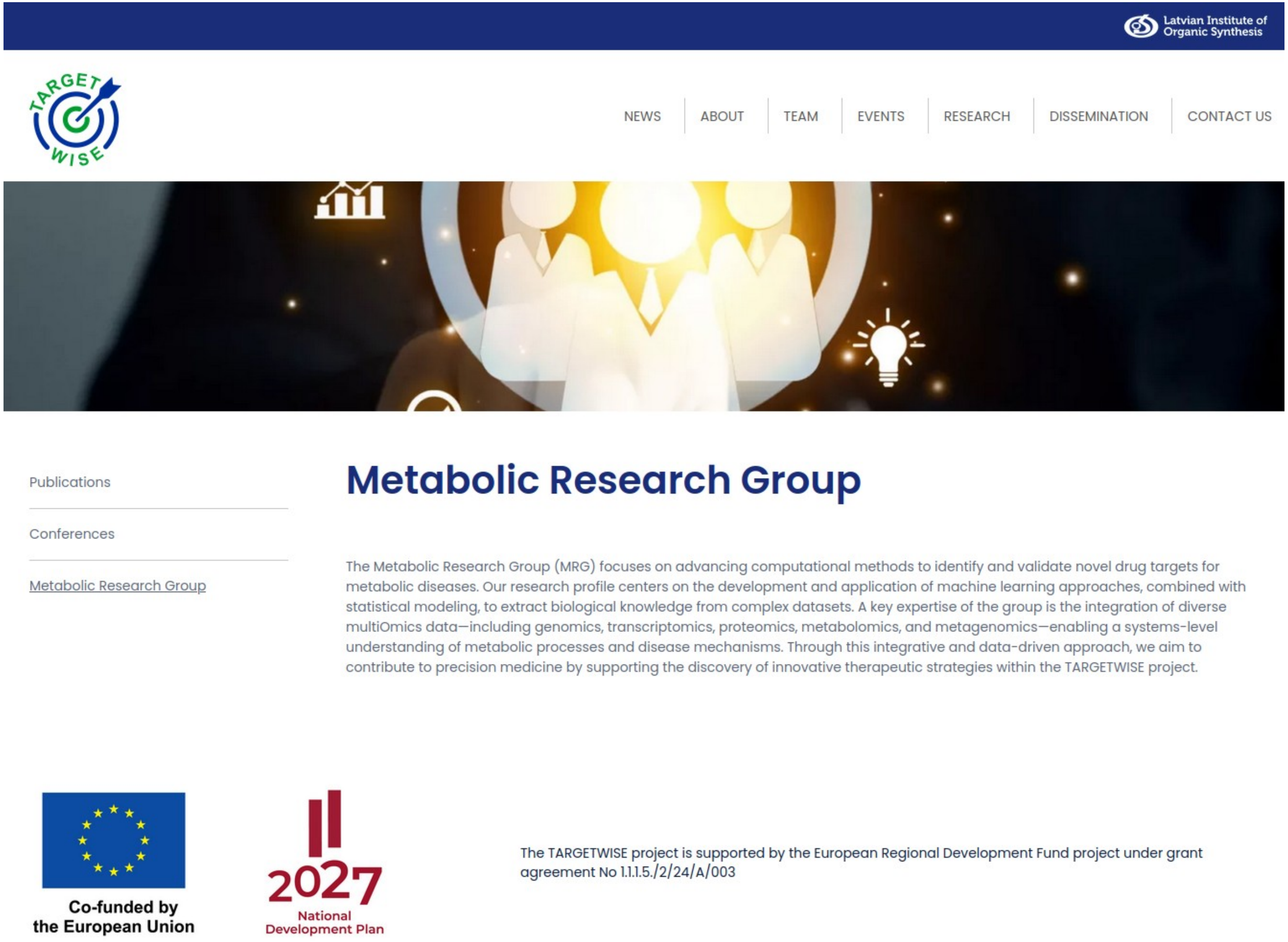
Metagenomics



Epidemiology



Ancient microbiome



Metabolic Research Group

The Metabolic Research Group (MRG) focuses on advancing computational methods to identify and validate novel drug targets for metabolic diseases. Our research profile centers on the development and application of machine learning approaches, combined with statistical modeling, to extract biological knowledge from complex datasets. A key expertise of the group is the integration of diverse multiOmics data—including genomics, transcriptomics, proteomics, metabolomics, and metagenomics—enabling a systems-level understanding of metabolic processes and disease mechanisms. Through this integrative and data-driven approach, we aim to contribute to precision medicine by supporting the discovery of innovative therapeutic strategies within the TARGETWISE project.

The TARGETWISE project is supported by the European Regional Development Fund project under grant agreement No 1.1.1.5./2/24/A/003

PhD. Nikolay Oskolkov

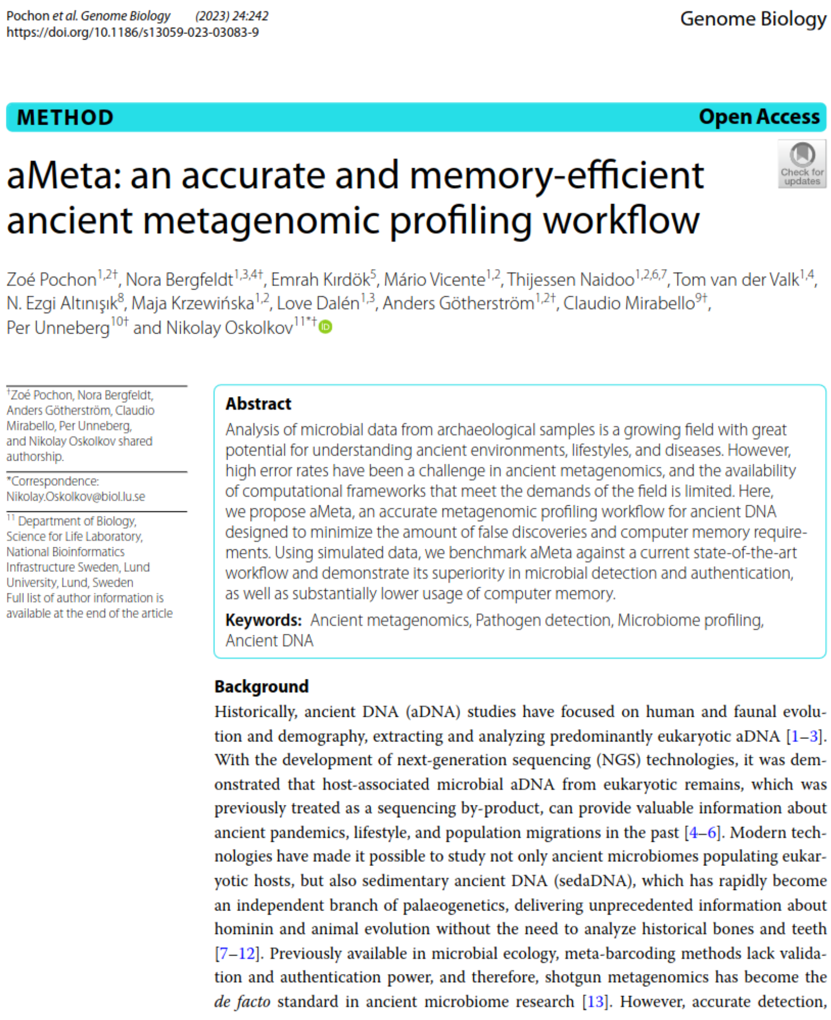
Group Leader (PI) of the Metabolic Research Group

Research interests are primarily focused on applications of mathematical statistics and machine learning to biological and biomedical data.

Daniel Rivas, MD, PhD in AI,
postdoctoral fellow
in Metabolic Research Group

One more postdoctoral fellow and two PhD students to be hired

aMeta: Pochon et al. Genome Biology 2023



Pochon *et al. Genome Biology* (2023) 24:242
https://doi.org/10.1186/s13059-023-03083-9

Genome Biology

METHOD **Open Access**

aMeta: an accurate and memory-efficient ancient metagenomic profiling workflow

Zoé Pochon[1,2†], Nora Bergfeldt[1,3,4†], Emrah Kırdök[5], Mário Vicente[1,2], Thijessen Naidoo[1,2,6,7], Tom van der Valk[1,4], N. Ezgi Altınışık[8], Maja Krzewińska[1,2], Love Dalén[1,3], Anders Götherström[1,2†], Claudio Mirabello[9†], Per Unneberg[10†] and Nikolay Oskolkov[11*†]

[†]Zoé Pochon, Nora Bergfeldt, Anders Götherström, Claudio Mirabello, Per Unneberg, and Nikolay Oskolkov shared authorship.

*Correspondence: Nikolay.Oskolkov@biol.lu.se

[11] Department of Biology, Science for Life Laboratory, National Bioinformatics Infrastructure Sweden, Lund University, Lund, Sweden
Full list of author information is available at the end of the article

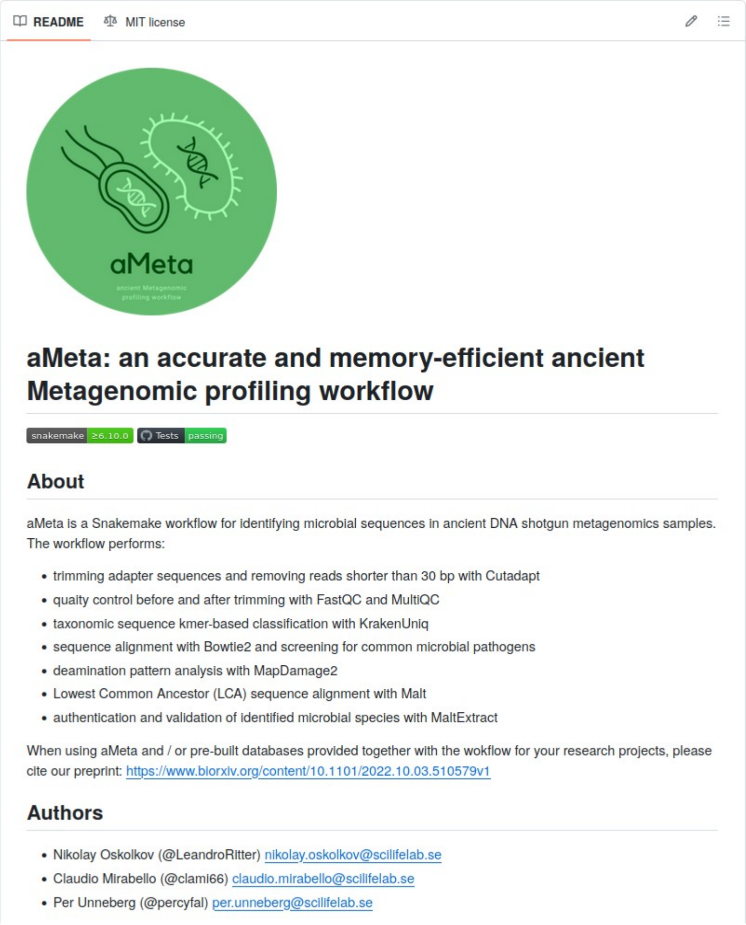
Abstract

Analysis of microbial data from archaeological samples is a growing field with great potential for understanding ancient environments, lifestyles, and diseases. However, high error rates have been a challenge in ancient metagenomics, and the availability of computational frameworks that meet the demands of the field is limited. Here, we propose aMeta, an accurate metagenomic profiling workflow for ancient DNA designed to minimize the amount of false discoveries and computer memory requirements. Using simulated data, we benchmark aMeta against a current state-of-the-art workflow and demonstrate its superiority in microbial detection and authentication, as well as substantially lower usage of computer memory.

Keywords: Ancient metagenomics, Pathogen detection, Microbiome profiling, Ancient DNA

Background

Historically, ancient DNA (aDNA) studies have focused on human and faunal evolution and demography, extracting and analyzing predominantly eukaryotic aDNA [1–3]. With the development of next-generation sequencing (NGS) technologies, it was demonstrated that host-associated microbial aDNA from eukaryotic remains, which was previously treated as a sequencing by-product, can provide valuable information about ancient pandemics, lifestyle, and population migrations in the past [4–6]. Modern technologies have made it possible to study not only ancient microbiomes populating eukaryotic hosts, but also sedimentary ancient DNA (sedaDNA), which has rapidly become an independent branch of palaeogenetics, delivering unprecedented information about hominin and animal evolution without the need to analyze historical bones and teeth [7–12]. Previously available in microbial ecology, meta-barcoding methods lack validation and authentication power, and therefore, shotgun metagenomics has become the *de facto* standard in ancient microbiome research [13]. However, accurate detection,



aMeta: an accurate and memory-efficient ancient Metagenomic profiling workflow

About

aMeta is a Snakemake workflow for identifying microbial sequences in ancient DNA shotgun metagenomics samples. The workflow performs:

- trimming adapter sequences and removing reads shorter than 30 bp with Cutadapt
- quaity control before and after trimming with FastQC and MultiQC
- taxonomic sequence kmer-based classification with KrakenUniq
- sequence alignment with Bowtie2 and screening for common microbial pathogens
- deamination pattern analysis with MapDamage2
- Lowest Common Ancestor (LCA) sequence alignment with Malt
- authentication and validation of identified microbial species with MaltExtract

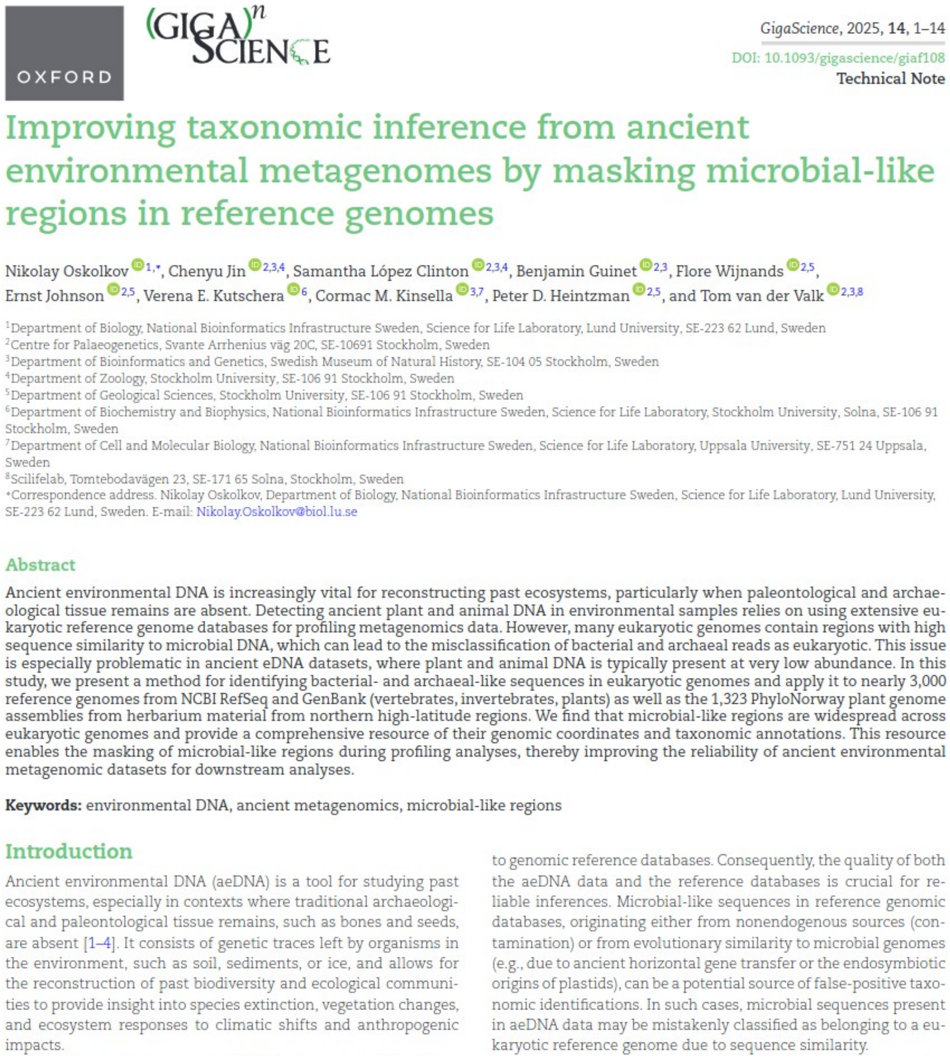
When using aMeta and / or pre-built databases provided together with the wokflow for your research projects, please cite our preprint: https://www.biorxiv.org/content/10.1101/2022.10.03.510579v1

Authors

- Nikolay Oskolkov (@LeandroRitter) nikolay.oskolkov@scilifelab.se
- Claudio Mirabello (@clami66) claudio.mirabello@scilifelab.se
- Per Unneberg (@percyfal) per.unneberg@scilifelab.se

https://github.com/NBISweden/aMeta

GENEX: Oskolkov et al. GigaScience 2025



GigaScience, 2025, **14**, 1–14
DOI: 10.1093/gigascience/giaf108
Technical Note

Improving taxonomic inference from ancient environmental metagenomes by masking microbial-like regions in reference genomes

Nikolay Oskolkov[1,*], Chenyu Jin[2,3,4], Samantha López Clinton[2,3,4], Benjamin Guinet[2,3], Flore Wijnands[2,5], Ernst Johnson[2,5], Verena E. Kutschera[6], Cormac M. Kinsella[3,7], Peter D. Heintzman[2,5], and Tom van der Valk[2,3,8]

[1]Department of Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Lund University, SE-223 62 Lund, Sweden
[2]Centre for Palaeogenetics, Svante Arrhenius väg 20C, SE-10691 Stockholm, Sweden
[3]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, SE-104 05 Stockholm, Sweden
[4]Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden
[5]Department of Geological Sciences, Stockholm University, SE-106 91 Stockholm, Sweden
[6]Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Solna, SE-106 91 Stockholm, Sweden
[7]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, SE-751 24 Uppsala, Sweden
[8]Scilifelab, Tomtebodavägen 23, SE-171 65 Solna, Stockholm, Sweden
*Correspondence address. Nikolay Oskolkov, Department of Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Lund University, SE-223 62 Lund, Sweden. E-mail: Nikolay.Oskolkov@biol.lu.se

Abstract

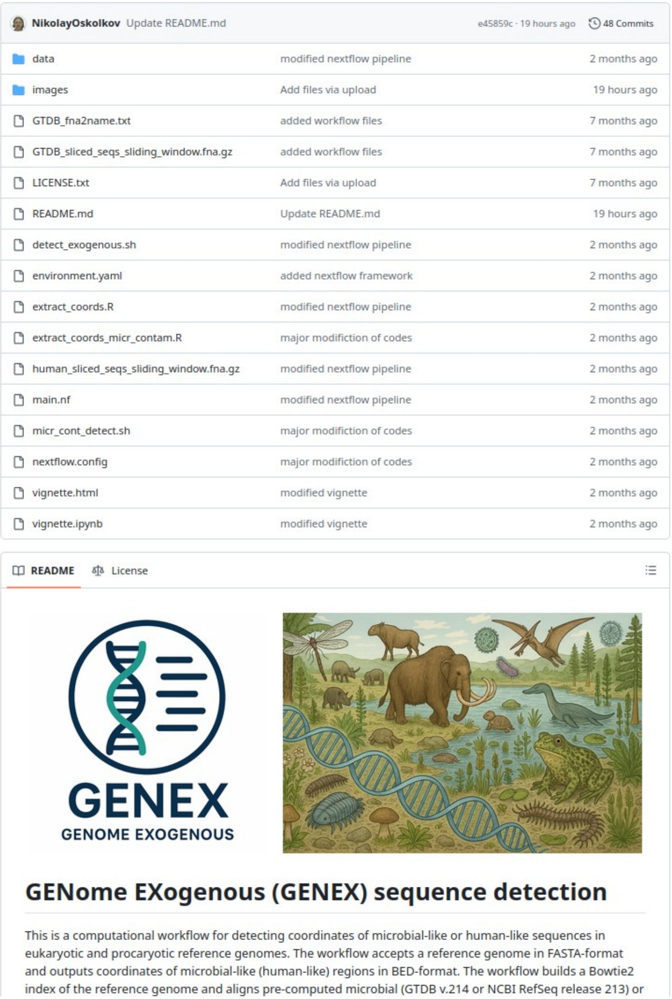
Ancient environmental DNA is increasingly vital for reconstructing past ecosystems, particularly when paleontological and archaeological tissue remains are absent. Detecting ancient plant and animal DNA in environmental samples relies on using extensive eukaryotic reference genome databases for profiling metagenomics data. However, many eukaryotic genomes contain regions with high sequence similarity to microbial DNA, which can lead to the misclassification of bacterial and archaeal reads as eukaryotic. This issue is especially problematic in ancient eDNA datasets, where plant and animal DNA is typically present at very low abundance. In this study, we present a method for identifying bacterial- and archaeal-like sequences in eukaryotic genomes and apply it to nearly 3,000 reference genomes from NCBI RefSeq and GenBank (vertebrates, invertebrates, plants) as well as the 1,323 PhyloNorway plant genome assemblies from herbarium material from northern high-latitude regions. We find that microbial-like regions are widespread across eukaryotic genomes and provide a comprehensive resource of their genomic coordinates and taxonomic annotations. This resource enables the masking of microbial-like regions during profiling analyses, thereby improving the reliability of ancient environmental metagenomic datasets for downstream analyses.

Keywords: environmental DNA, ancient metagenomics, microbial-like regions

Introduction

Ancient environmental DNA (aeDNA) is a tool for studying past ecosystems, especially in contexts where traditional archaeological and paleontological tissue remains, such as bones and seeds, are absent [1–4]. It consists of genetic traces left by organisms in the environment, such as soil, sediments, or ice, and allows for the reconstruction of past biodiversity and ecological communities to provide insight into species extinction, vegetation changes, and ecosystem responses to climatic shifts and anthropogenic impacts.

to genomic reference databases. Consequently, the quality of both the aeDNA data and the reference databases is crucial for reliable inferences. Microbial-like sequences in reference genomic databases, originating either from nonendogenous sources (contamination) or from evolutionary similarity to microbial genomes (e.g., due to ancient horizontal gene transfer or the endosymbiotic origins of plastids), can be a potential source of false-positive taxonomic identifications. In such cases, microbial sequences present in aeDNA data may be mistakenly classified as belonging to a eukaryotic reference genome due to sequence similarity.



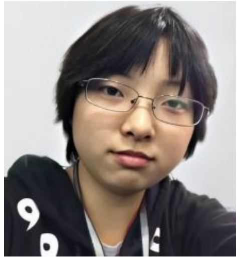
NikolayOskolkov Update README.md e45859c · 19 hours ago 48 Commits

data	modified nextflow pipeline	2 months ago
images	Add files via upload	19 hours ago
GTDB_fna2name.txt	added workflow files	7 months ago
GTDB_sliced_seqs_sliding_window.fna.gz	added workflow files	7 months ago
LICENSE.txt	Add files via upload	7 months ago
README.md	Update README.md	19 hours ago
detect_exogenous.sh	modified nextflow pipeline	2 months ago
environment.yaml	added nextflow framework	2 months ago
extract_coords.R	modified nextflow pipeline	2 months ago
extract_coords_micr_contam.R	major modification of codes	2 months ago
human_sliced_seqs_sliding_window.fna.gz	modified nextflow pipeline	2 months ago
main.nf	modified nextflow pipeline	2 months ago
micr_cont_detect.sh	major modification of codes	2 months ago
nextflow.config	major modification of codes	2 months ago
vignette.html	modified vignette	2 months ago
vignette.ipynb	modified vignette	2 months ago

GENome EXogenous (GENEX) sequence detection

This is a computational workflow for detecting coordinates of microbial-like or human-like sequences in eukaryotic and procaryotic reference genomes. The workflow accepts a reference genome in FASTA-format and outputs coordinates of microbial-like (human-like) regions in BED-format. The workflow builds a Bowtie2 index of the reference genome and aligns pre-computed microbial (GTDB v.214 or NCBI RefSeq release 213) or

https://github.com/NikolayOskolkov/MCWorkflow



About you

- Name
- University/Institute/Company
- Research interest(s)
- Previous experience(s) with microbial ecology, metagenomics, bioinformatics, etc.
- General hopes for this course

Course outline

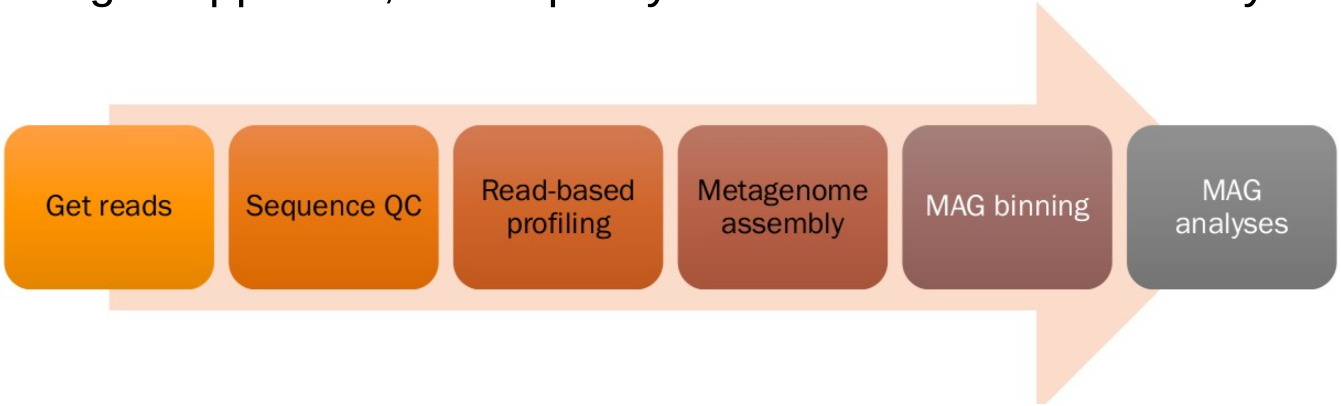
Day 1: introduction, setting up, connecting to server, getting raw data, exploring data

Day2: quality control, adapter and host removal, read-based taxonomic classification

Day3: *de-novo* assembly, taxonomic profiling and abundance quantification of contigs

Day4: assembly of long-read sequencing data, metagenomic binning

Day5: gene catalogue approach, MAG quality control and functional analysis of MAGs



Physalia Courses

Housekeeping rules

- To ask question please raise your hand, unmute yourself and ask. You can also ask questions in zoom chat or slack workspace for this course.

- Please keep your camera on as much as possible for better contact and communication.

- The course includes ~10 lectures (30–60 min each) followed by practical sessions, we will have 15-30 minutes breaks between the sessions depending on how tired the participants are.

- Recordings will be provided after each day, so if you miss a lecture or practical, there will be a chance to catch up

Practical information: GitHub and Zoom

The course will take place in Zoom from 9 am to 1 pm (CET, Berlin time)

Links to the Zoom room will be posted in Slack

The course GitHub repository containing lectures and exercises is:

https://github.com/NikolayOskolkov/Physalia_EnvMetagenomics_2025

Please bookmark this address!

Practical information: Amazon Cloud (AWS EC2)

We will use the Cloud Computing service from Amazon, which we will access via `ssh` (secure shell protocol)

https://github.com/NikolayOskolkov/Physalia_EnvMetagenomics_2025/blob/main/exercises.md#setting-up-the-cloud-computing

See here for information on how to connect, but remember:

- The IP address changes every day
- Everyone is given a username, with a `home` and `shared` folders
    - List of usernames can be found in Slack
    - The `shared` folder is copy-only: do not delete, move, rename, or write

Practical information: conda

conda.io

System for software management (python, R, JavaScript, C++, …)

Allows easy installation of software in dedicated environments, separated from the main environment and other conda environments

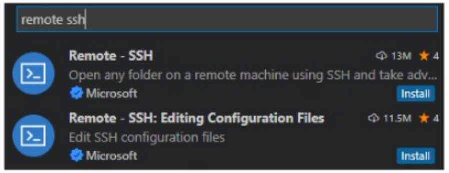
- The environments that we will use have been already set up for everyone

General conda commands

```
> conda env list
> conda activate ambiente
> conda deactivate
```

Practical information: setting up VS Code

- Download and install VS Code: code.visualstudio.com/Download
- Launch VS Code
- Go to View -> Extensions
- Search for and install the extension Remote–SSH
- See here for a step-by-step guide on how to connect to the Amazon Cloud



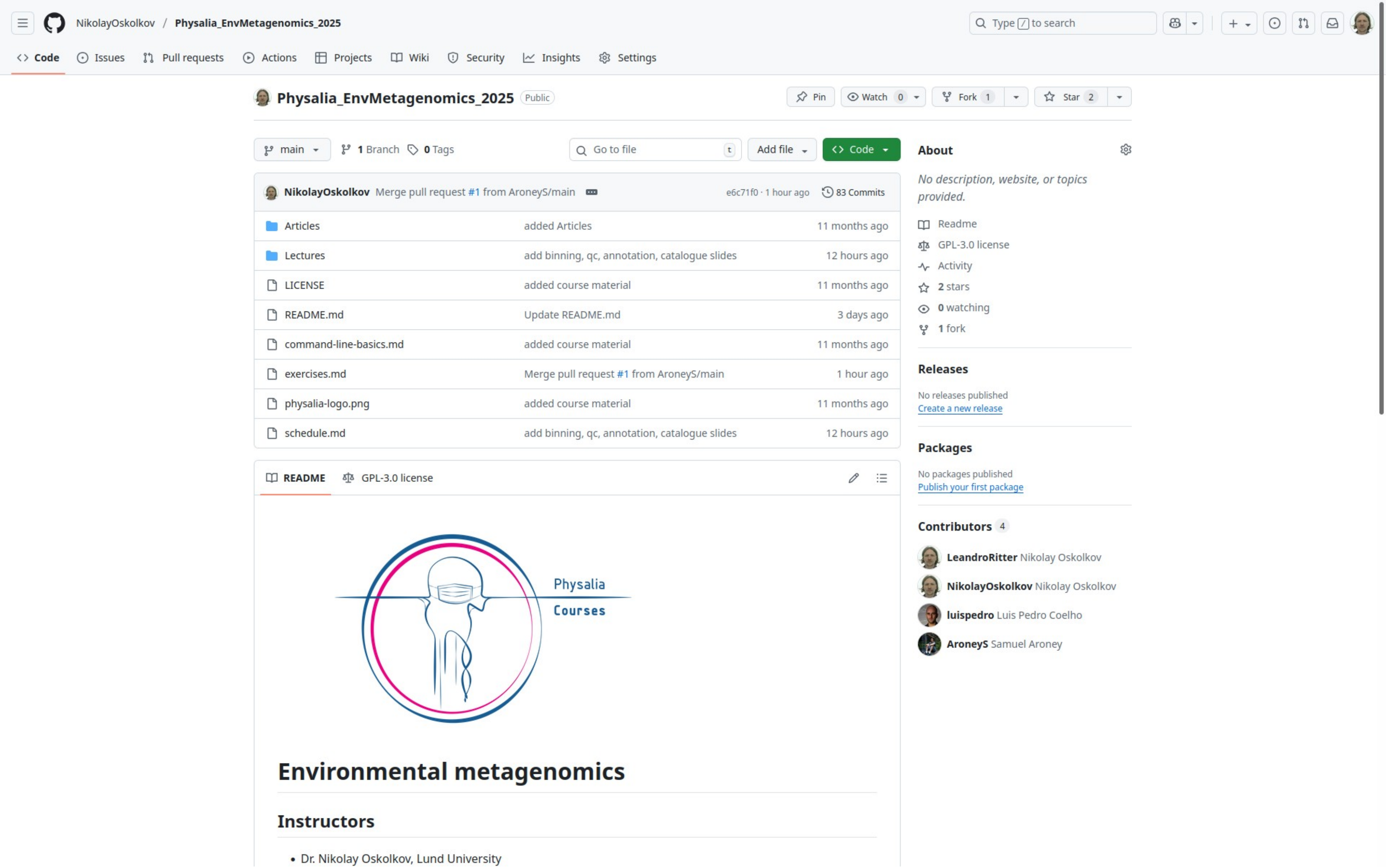
remote ssh

Remote - SSH 13M ★ 4
Open any folder on a remote machine using SSH and take adv...
Microsoft

Remote - SSH: Editing Configuration Files 11.5M ★ 4
Edit SSH configuration files
Microsoft

https://github.com/NikolayOskolkov/Physalia_EnvMetagenomics_2024/blob/main/exercises.md

Physalia
Courses



Physalia_EnvMetagenomics_2025 Public

NikolayOskolkov Merge pull request #1 from AroneyS/main e6c71f0 · 1 hour ago 83 Commits

Articles	added Articles	11 months ago
Lectures	add binning, qc, annotation, catalogue slides	12 hours ago
LICENSE	added course material	11 months ago
README.md	Update README.md	3 days ago
command-line-basics.md	added course material	11 months ago
exercises.md	Merge pull request #1 from AroneyS/main	1 hour ago
physalia-logo.png	added course material	11 months ago
schedule.md	add binning, qc, annotation, catalogue slides	12 hours ago

About

No description, website, or topics provided.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

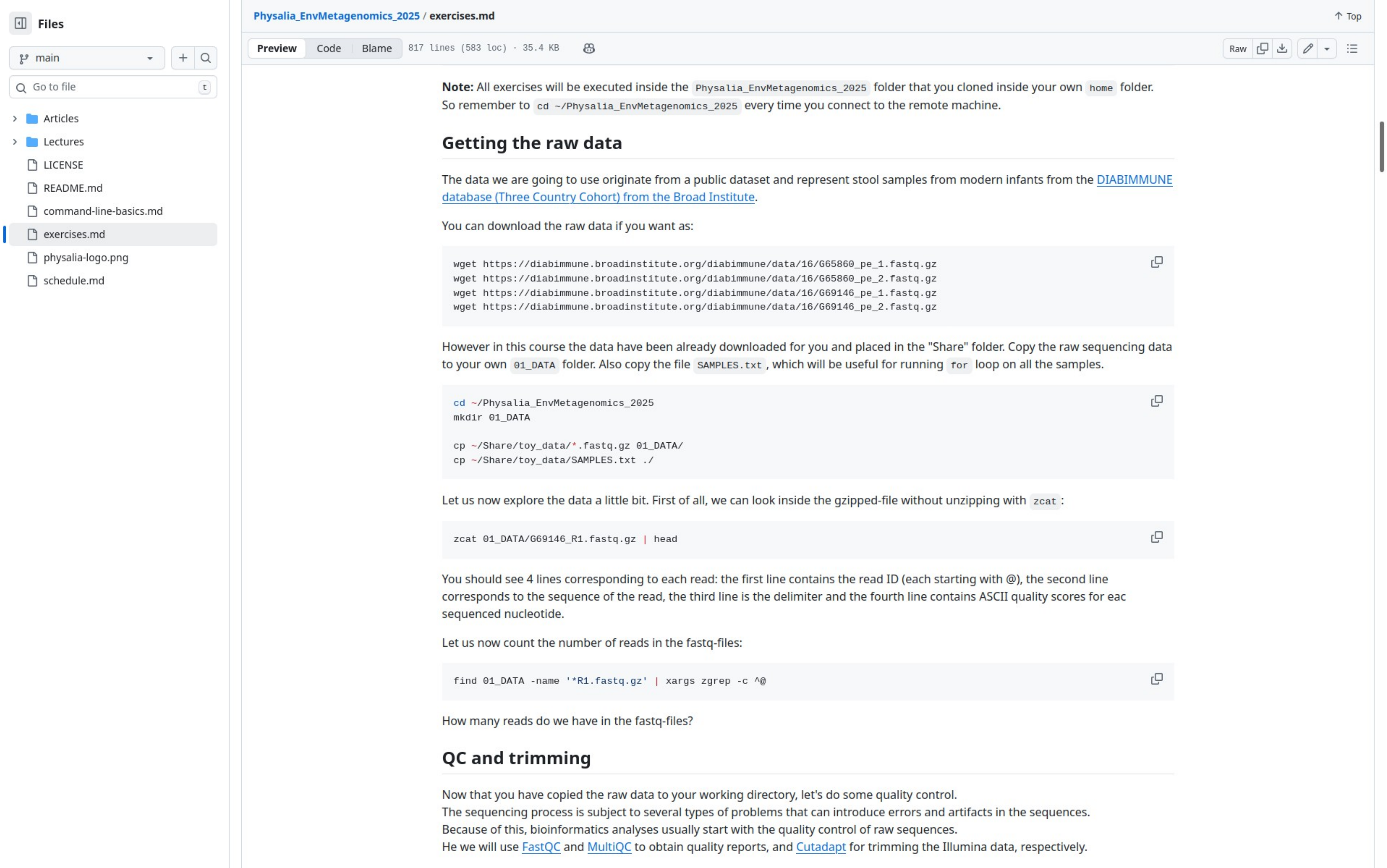
Contributors 4

- **LeandroRitter** Nikolay Oskolkov
- **NikolayOskolkov** Nikolay Oskolkov
- **luispedro** Luis Pedro Coelho
- **AroneyS** Samuel Aroney

Environmental metagenomics

Instructors

- Dr. Nikolay Oskolkov, Lund University



Note: All exercises will be executed inside the `Physalia_EnvMetagenomics_2025` folder that you cloned inside your own `home` folder.
So remember to `cd ~/Physalia_EnvMetagenomics_2025` every time you connect to the remote machine.

Getting the raw data

The data we are going to use originate from a public dataset and represent stool samples from modern infants from the DIABIMMUNE database (Three Country Cohort) from the Broad Institute.

You can download the raw data if you want as:

```
wget https://diabimmune.broadinstitute.org/diabimmune/data/16/G65860_pe_1.fastq.gz
wget https://diabimmune.broadinstitute.org/diabimmune/data/16/G65860_pe_2.fastq.gz
wget https://diabimmune.broadinstitute.org/diabimmune/data/16/G69146_pe_1.fastq.gz
wget https://diabimmune.broadinstitute.org/diabimmune/data/16/G69146_pe_2.fastq.gz
```

However in this course the data have been already downloaded for you and placed in the "Share" folder. Copy the raw sequencing data to your own `01_DATA` folder. Also copy the file `SAMPLES.txt`, which will be useful for running `for` loop on all the samples.

```
cd ~/Physalia_EnvMetagenomics_2025
mkdir 01_DATA

cp ~/Share/toy_data/*.fastq.gz 01_DATA/
cp ~/Share/toy_data/SAMPLES.txt ./
```

Let us now explore the data a little bit. First of all, we can look inside the gzipped-file without unzipping with `zcat`:

```
zcat 01_DATA/G69146_R1.fastq.gz | head
```

You should see 4 lines corresponding to each read: the first line contains the read ID (each starting with @), the second line corresponds to the sequence of the read, the third line is the delimiter and the fourth line contains ASCII quality scores for eac sequenced nucleotide.

Let us now count the number of reads in the fastq-files:

```
find 01_DATA -name '*R1.fastq.gz' | xargs zgrep -c ^@
```

How many reads do we have in the fastq-files?

QC and trimming

Now that you have copied the raw data to your working directory, let's do some quality control.
The sequencing process is subject to several types of problems that can introduce errors and artifacts in the sequences.
Because of this, bioinformatics analyses usually start with the quality control of raw sequences.
He we will use FastQC and MultiQC to obtain quality reports, and Cutadapt for trimming the Illumina data, respectively.