

# aMeta, an accurate and memory-efficient ancient Metagenomic profiling workflow

AUTHOR

Nikolay Oskolkov



## Tip

For this chapter's exercises, if not already performed, you will need to create the [conda environment](#) from the `yaml` file in the following [archive](#), and activate the environment:

```
conda activate aMeta
```


## aMeta: introduction and installation

In this chapter, we will demonstrate an example of using aMeta, an accurate and memory-efficient ancient Metagenomic profiling workflow proposed in [Pochon et al. 2023](#).




THE PREPRINT SERVER FOR BIOLOGY


HOME | SUBMIT | FAQ | BLOG  
| ALERTS / RSS | ABOUT  
| CHANNELS




Advanced Search

New Results


 [Follow this preprint](#)


 Previous


Next 


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


Posted October 05, 2022.


 **Download PDF**













 Email

 **Print/Save Options**

 Share

 Data/Code

 Citation Tools

 Zoé Pochon,  Nora Bergfeldt,  Emrah Kirdök,  Mário Vicente, Thijessen Naidoo,  Tom van der Valk,  N. Ezgi Altınışık,  Maja Krzewińska,  Love Dalen,  Anders Götherström,  Claudio Mirabello,  Per Unneberg,  Nikolay Oskolkov

doi: <https://doi.org/10.1101/2022.10.03.510579>

It can be cloned from NBISweden github repository and installed as follows:

```
git clone https://github.com/NBISweden/aMeta
cd aMeta
mamba env create -f workflow/envs/environment.yaml
conda activate aMeta
```

To ensure that aMeta has been correctly installed, we can run a quick test:

```
cd .test
```

```
./runtest.sh -j 20
```

## Downloading data, databases and indexes

For demonstration purposes we will use 10 simulated with [gargammel](#) ancient metagenomic samples used for benchmarking aMeta. The simulated data can be accessed via <https://doi.org/10.17044/scilifelab.21261405> and downloaded via terminal using following command lines:

```
cd aMeta
mkdir data && cd data
wget https://figshare.scilifelab.se/ndownloader/articles/21261405/versions/1 \
&& export UNZIP_DISABLE_ZIPBOMB_DETECTION=true && unzip 1 && rm 1
```

To run aMeta, we will need a small KrakenUniq database. Here we download a pre-built database based on complete microbial NCBI RefSeq reference genomes:

```
cd aMeta/resources
mkdir KrakenUniq_DB && cd KrakenUniq_DB
wget https://figshare.scilifelab.se/ndownloader/articles/21299541/versions/1 \
&& export UNZIP_DISABLE_ZIPBOMB_DETECTION=true && unzip 1 && rm 1
```

We will also need a Bowtie2 index corresponding to the KrakenUniq reference database:

```
cd aMeta/resources
mkdir Bowtie2_index && cd Bowtie2_index
wget https://figshare.scilifelab.se/ndownloader/articles/21185887/versions/1 \
&& export UNZIP_DISABLE_ZIPBOMB_DETECTION=true && unzip 1 && rm 1
```

The last thing we need to download are a few helping files with useful NCBI taxonomy information:

```
cd aMeta/resources
wget https://figshare.scilifelab.se/ndownloader/files/38201982 && \
mv 38201982 seqid2taxid.map.orig
wget https://figshare.scilifelab.se/ndownloader/files/38201937 && \
mv 38201937 nucl_gb.accession2taxid
wget https://figshare.scilifelab.se/ndownloader/files/37395181 && \
mv 37395181 library.fna.gz && gunzip library.fna.gz
```

## aMeta configuration

Now we need to configure the workflow. First, we need to create a tab-delimited *samples.tsv* file inside *aMeta/config* and provide the names of the input fastq-files:

```
sample  fastq
sample1 data/sample1.fastq.gz
sample2 data/sample2.fastq.gz
sample3 data/sample3.fastq.gz
```

```

sample3 data/sample3.fastq.gz
sample4 data/sample4.fastq.gz
sample5 data/sample5.fastq.gz
sample6 data/sample6.fastq.gz
sample7 data/sample7.fastq.gz
sample8 data/sample8.fastq.gz
sample9 data/sample9.fastq.gz
sample10 data/sample10.fastq.gz

```

Further, we will put details about e.g. databases locations in the *config.yaml* file inside *aMeta/config*. A minimal example *config.yaml* files can look like this:

```

samplesheet: "config/samples.tsv"

krakenuniq_db: resources/KrakenUniq_DB

bowtie2_db: resources/Bowtie2_index/library.pathogen.fna
bowtie2_seqid2taxid_db: resources/Bowtie2_index/seqid2taxid.pathogen.map
pathogenomesFound: resources/Bowtie2_index/pathogensFound.very_inclusive.tab

malt_nt_fasta: resources/library.fna
malt_seqid2taxid_db: resources/seqid2taxid.map.orig
malt_accession2taxid: resources/nuc1_gb.accession2taxid

ncbi_db: resources/ncbi

n_unique_kmers: 1000
n_tax_reads: 200

```

## Prepare and run aMeta

Next, we need to create conda sub-environments of aMeta, then manually tune a few memory related parameters of tools (Krona and Malt) included in aMeta:

```

snakemake --snakefile workflow/Snakefile --use-conda --conda-create-envs-only -j 20

env=$(grep krona .snakemake/conda/*.yaml | awk '{print $1}' | sed -e "s/.yaml://g" \
| head -1)
cd $env/opt/krona/
./updateTaxonomy.sh taxonomy
cd -

cd aMeta
env=$(grep hops .snakemake/conda/*.yaml | awk '{print $1}' | sed -e "s/.yaml://g" \
| head -1)
conda activate $env
version=$(conda list malt --json | grep version | sed -e "s/\\\"//g" | awk '{print $2}')
cd $env/opt/malt-$version
sed -i -e "s/-Xmx64G/-Xmx1000G/" malt-build.vmoptions
sed -i -e "s/-Xmx64G/-Xmx1000G/" malt-run.vmoptions
cd -

```

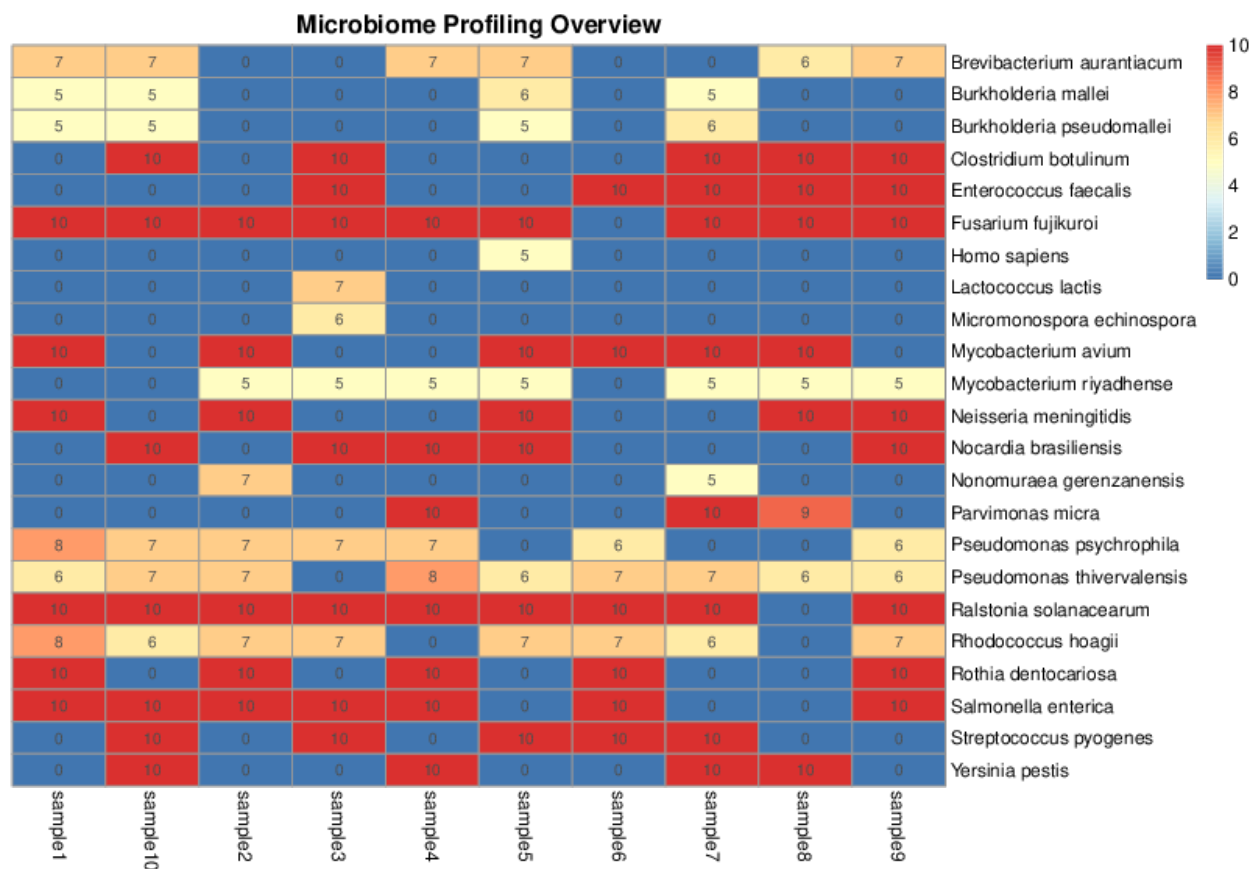
```
conda deactivate
```

And, finally, we are ready to run aMeta:

```
snakemake --snakefile workflow/Snakefile --use-conda -j 20
```

## aMeta output

All output files of the workflow are located in *aMeta/results* directory. To get a quick overview of ancient microbes present in your samples you should check a heatmap in *results/overview\_heatmap\_scores.pdf*.



The heatmap demonstrates microbial species (in rows) authenticated for each sample (in columns). The colors and the numbers in the heatmap represent authentication scores, i.e. numeric quantification of seven quality metrics that provide information about microbial presence and ancient status. The authentication scores can vary from 0 to 10, the higher is the score the more likely that a microbe is present in a sample and is ancient. Typically, scores from 8 to 10 (red color in the heatmap) provide good confidence of ancient microbial presence in a sample. Scores from 5 to 7 (yellow and orange colors in the heatmap) can imply that either: a) a microbe is present but not ancient, i.e. modern contaminant, or b) a microbe is ancient (the reads are damaged) but was perhaps aligned to a wrong reference, i.e. it is not the microbe you think about. The former is a more common case scenario. The latter often happens when an ancient microbe is correctly detected on a genus level but we are not confident about the exact species, and might be aligning the damaged reads to a non-optimal reference which leads to a lot of mismatches or poor evenness of coverage. Scores from 0 to 4 (blue color in the heatmap) typically mean that we have very little

statistical evidence (very few reads) to claim presence of a microbe in a sample.

To visually examine the seven quality metrics

- deamination profile,
- evenness of coverage,
- edit distance (amount of mismatches) for all reads,
- edit distance (amount of mismatches) for damaged reads,
- read length distribution,
- PMD scores distribution,
- number of assigned reads (depth of coverage),

corresponding to the numbers and colors of the heatmap, one can find them in

results/AUTHENTICATION/sampleID/taxID

/authentic\_Sample\_sampleID.trimmed.rma6\_TaxID\_taxID.pdf for each sample sampleID and each authenticated microbe taxID. An example of such quality metrics is shown below:

