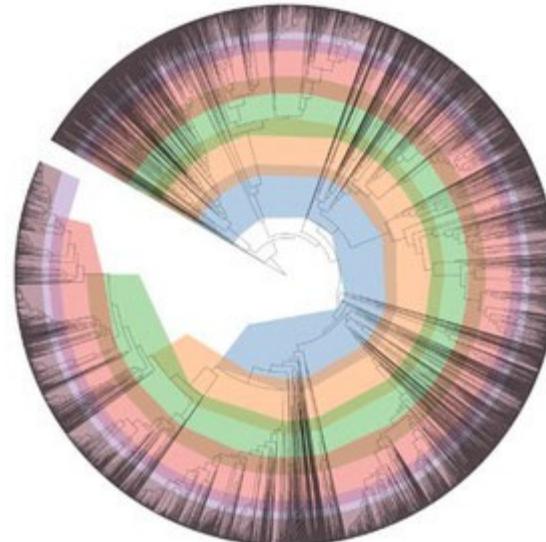


ENVIRONMENTAL METAGENOMICS

Physalia course, online, 13-17 October 2025

Introduction: computational challenges in metagenomics

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)

What is a metagenome? WHY ONLY MICROBIAL?

Marchesi and Ravel *Microbiome* (2015) 3:31
DOI 10.1186/s40168-015-0094-5



Microbiome

EDITORIAL

Open Access

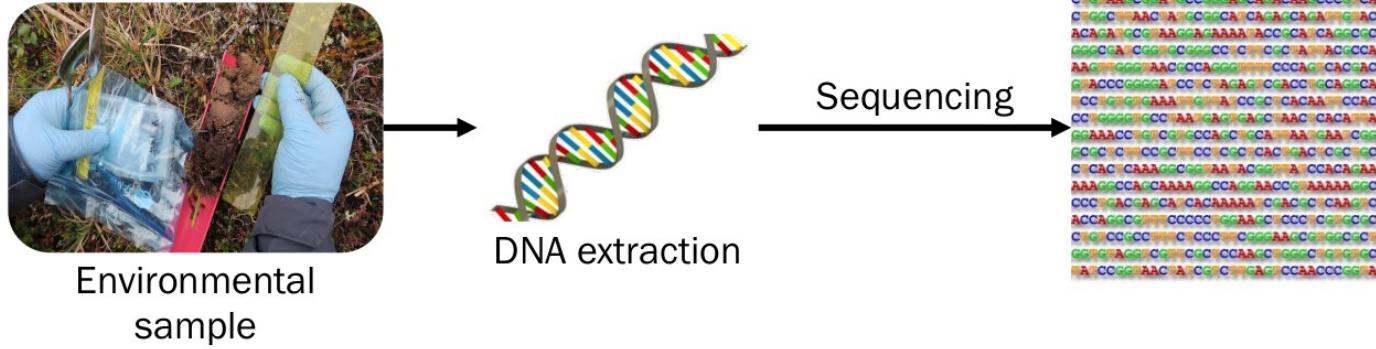
The vocabulary of microbiome research: a proposal

Julian R. Marchesi^{1,2} and Jacques Ravel^{3,4*}

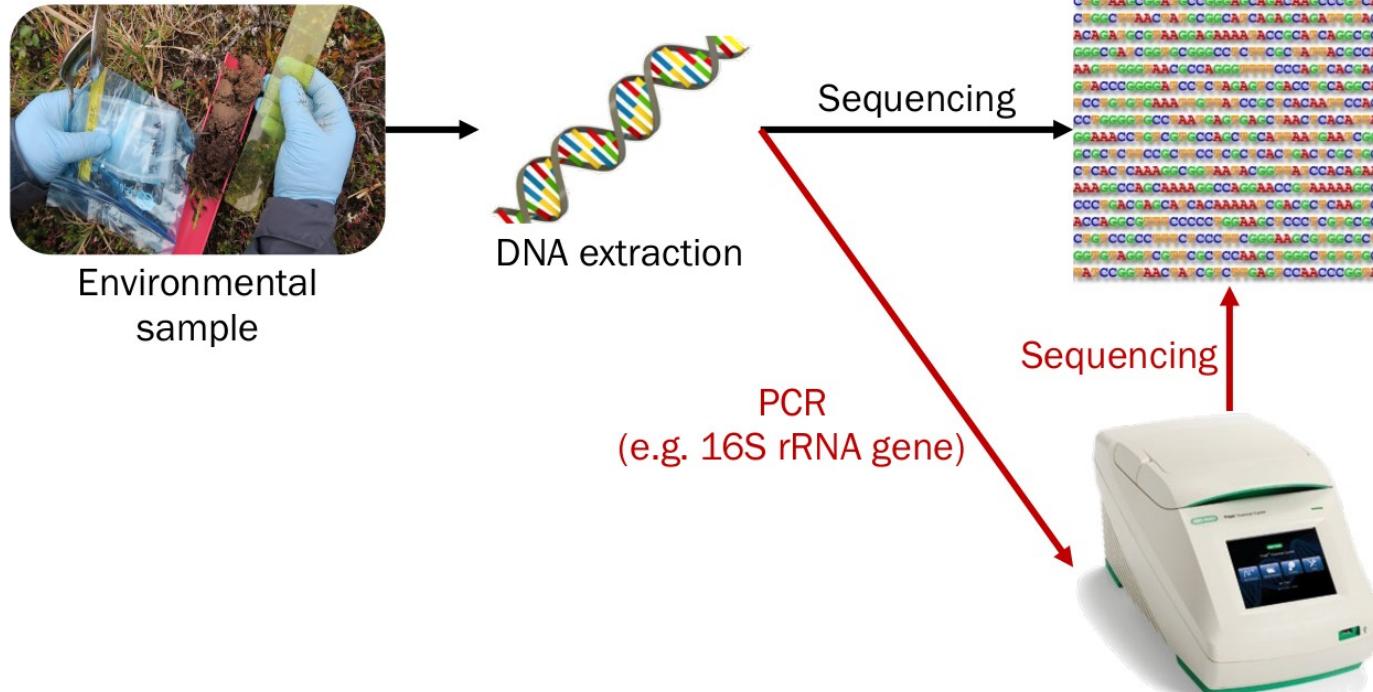


A **metagenome** is a collection of genomes or genes from the members of a microbiota. A **microbiota** is an assemblage of microorganisms present in a defined environment. A **microbiome** refers to an entire habitat, including the microorganisms, their genomes, and the surrounding environmental conditions.

What is metagenomics?



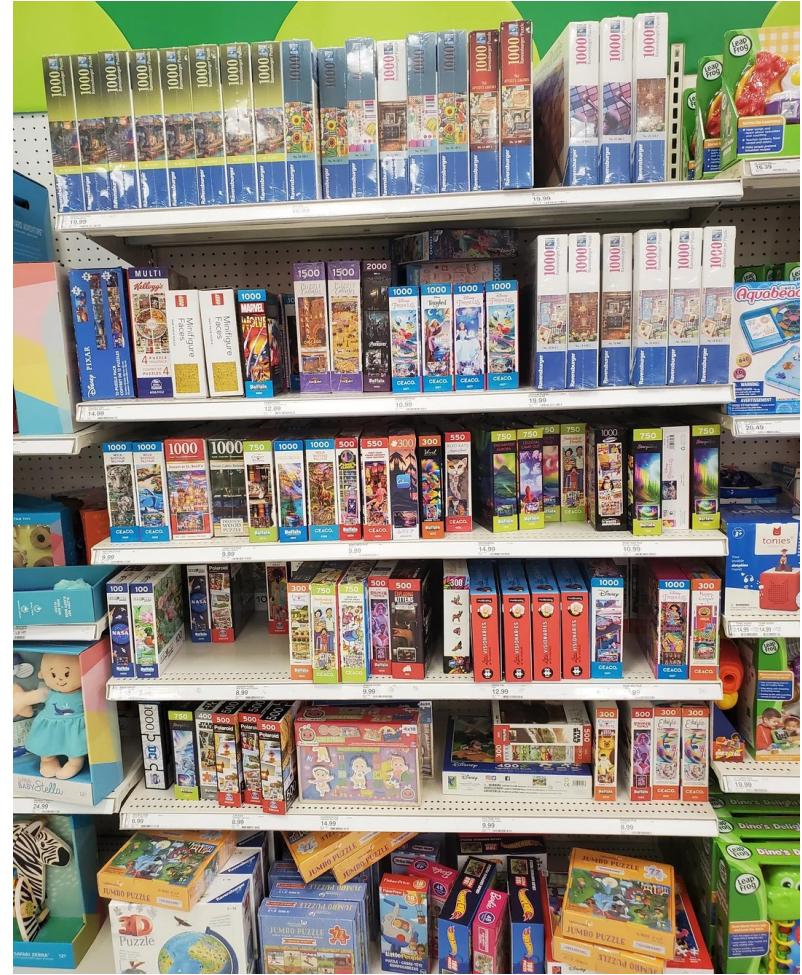
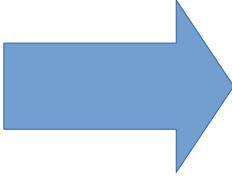
What is NOT metagenomics?



Sequenced genome is a collection of bits of puzzle to be organized



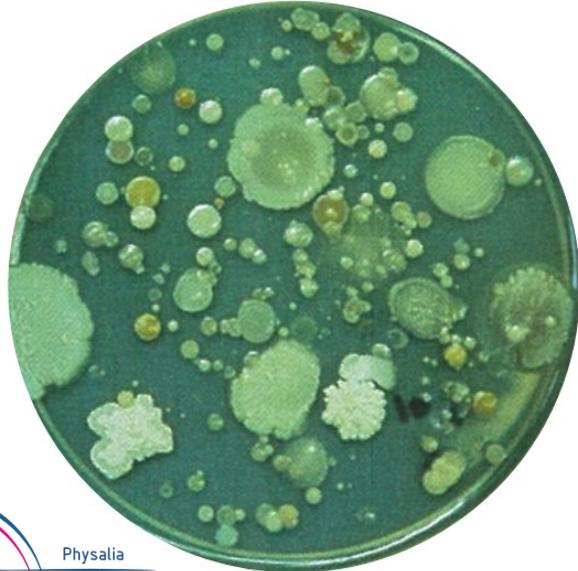
From single genome to metagenome



Who's there?

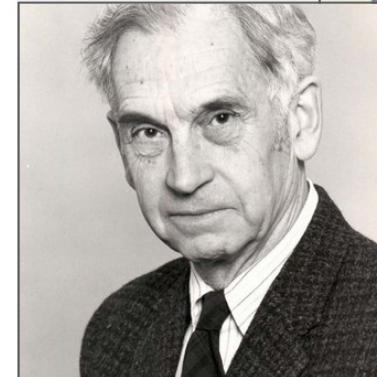
At a most basic level, the first question we usually ask in metagenomics is “Who’s there?”

What is a microbial species?



Physalia
Courses

Ernst Mayr
Biological Species
Concept, 1942



The Birds of Northern Melanesia
Copyrighted Material
SPECIATION, ECOLOGY, AND BIOGEOGRAPHY

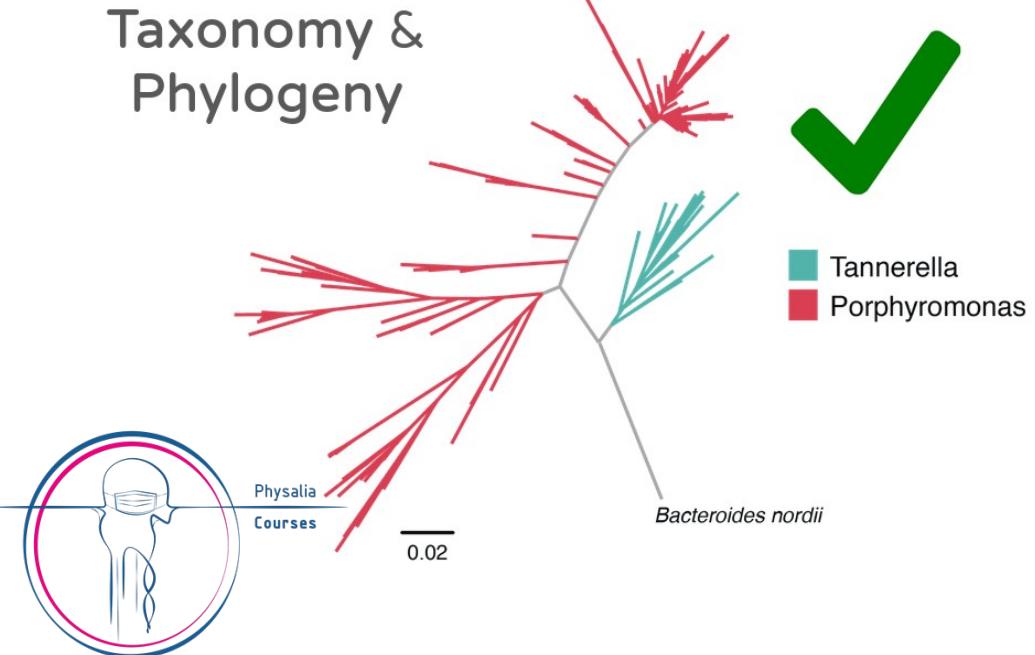


Ernst Mayr ~ Jared Diamond
Copyrighted Material

Who's there?

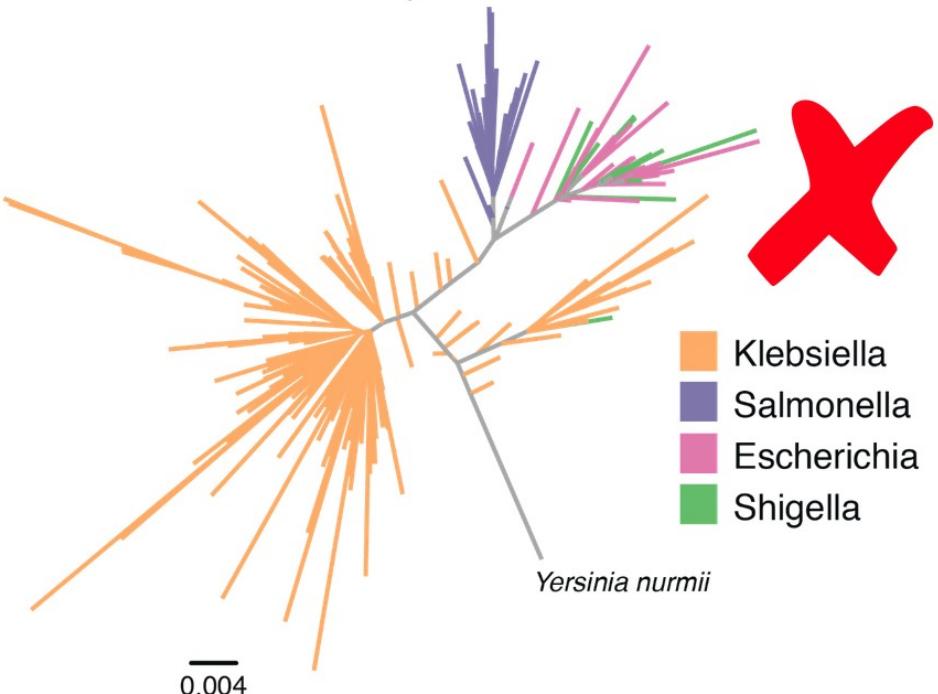
At a most basic level, the first question we usually ask in metagenomics is “Who’s there?”

What is a microbial species?



Taxonomy: classification or categorization of organisms into groups (taxa)

Phylogeny: evolutionary history of a set of taxa

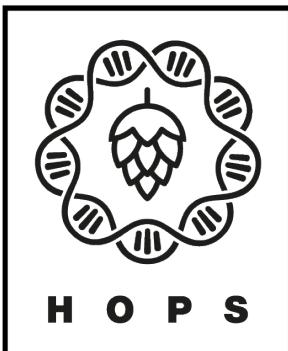


Typical analysis methods used in metagenomics

1) Alignment:



BWA
stands for
Burrows Wheeler Aligner
 Abbreviations.com



2) Classification:



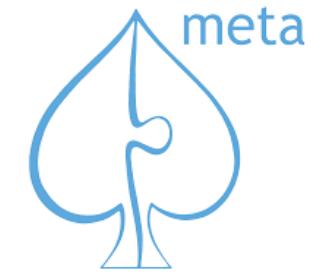
Centrifuge

MetaPhlan

Clark

Reference based:
assume similarity to reference

3) De-novo assembly:



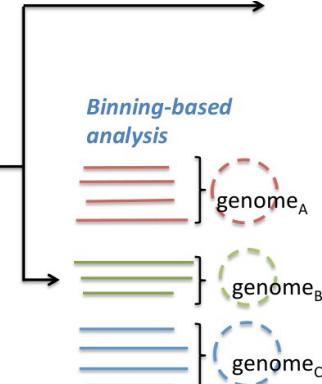
```
>seq1  
GCCGTAGTCC...  
>seq2  
...
```



Assembly

Assembly-based
analysis

gene prediction/
annotation

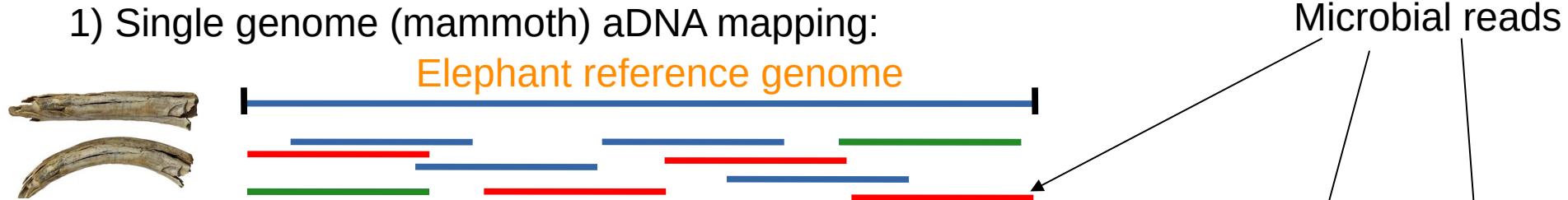


Phylogenetic binning

Reference free:
unbiased but challenging

What is competitive mapping and why you should do it

1) Single genome (mammoth) aDNA mapping:



2) Correcting for human contamination:

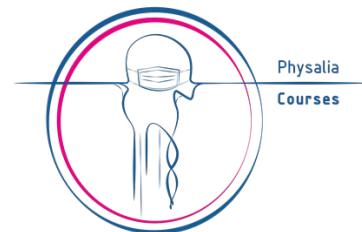
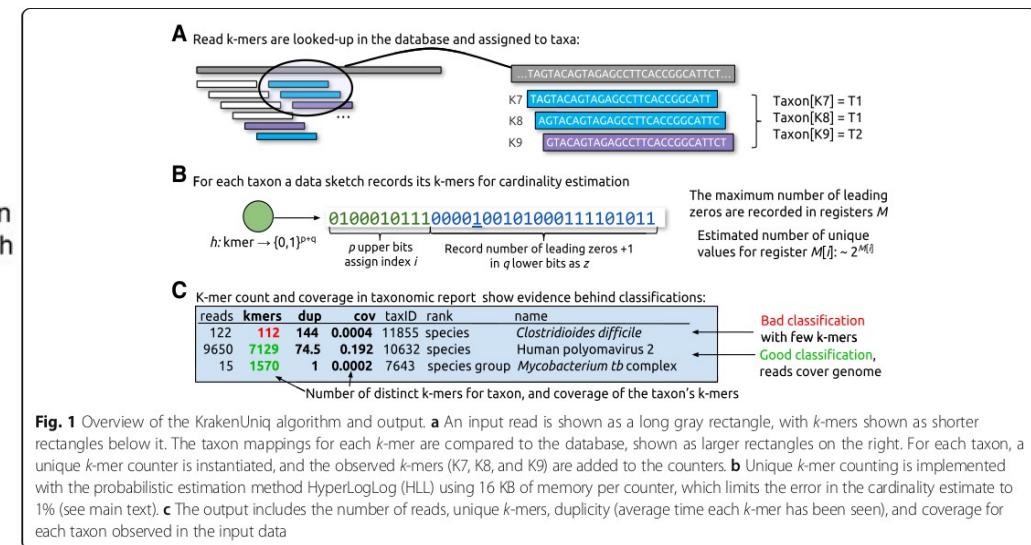
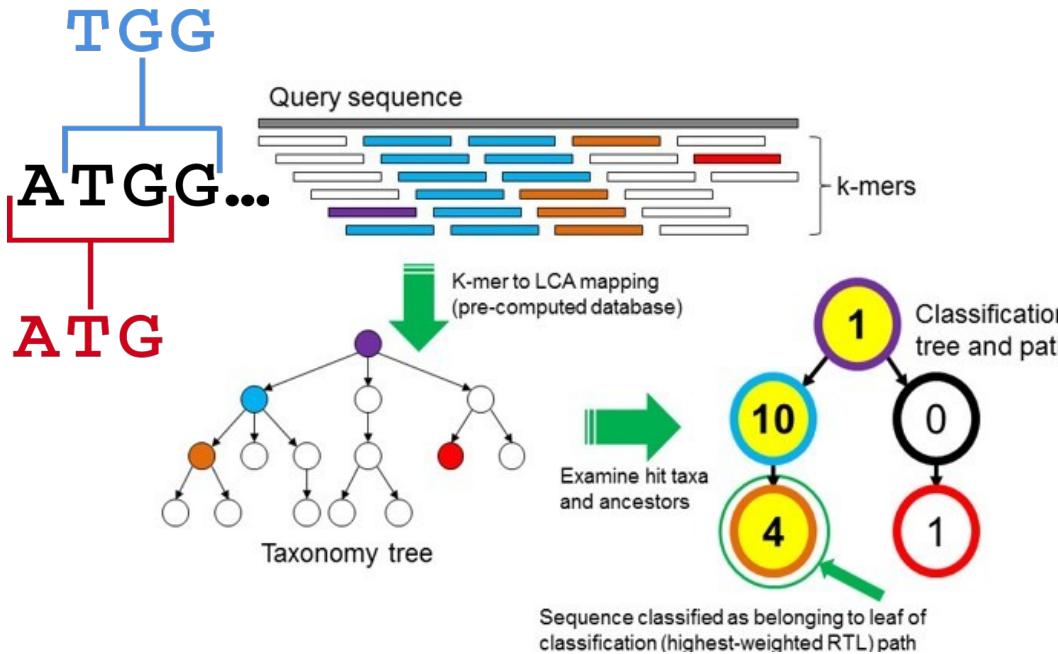


3) Ancient metagenomics mapping:



Competitive mapping is absolutely central for metagenomic analysis!

K-mer based taxonomic profiling: Kraken family of tools



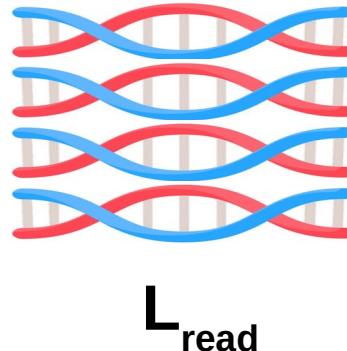
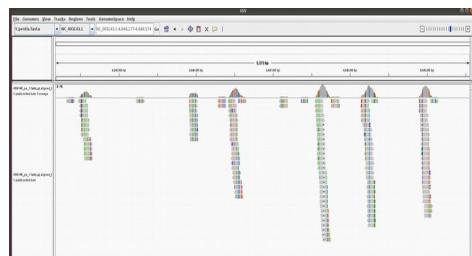
Advantage of classification over alignment: speed, Kraken2 is very fast!

Coverage vs. depth vs. breadth of coverage

Reference genome

A)

GCTACGATCTTAGCTTAGCTGGGATCTGAATTCTCATCTCGGAT



$$L_{\text{genome}} = 4 * L_{\text{read}}$$



Organism
NOT
detected

Reference genome

B)

GCTACGATCTTAGCTTAGCTGGGATCTGAATTCTCATCTCGGAT



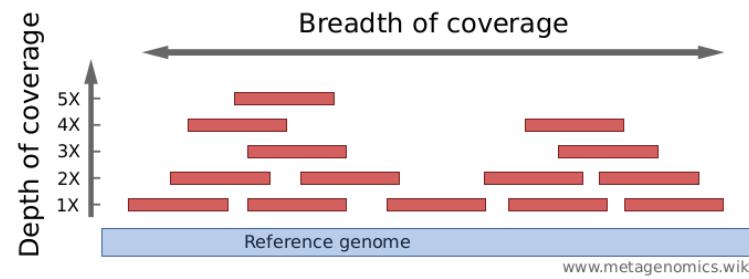
$$L_{\text{genome}} = 4 * L_{\text{read}}$$



Organism
detected

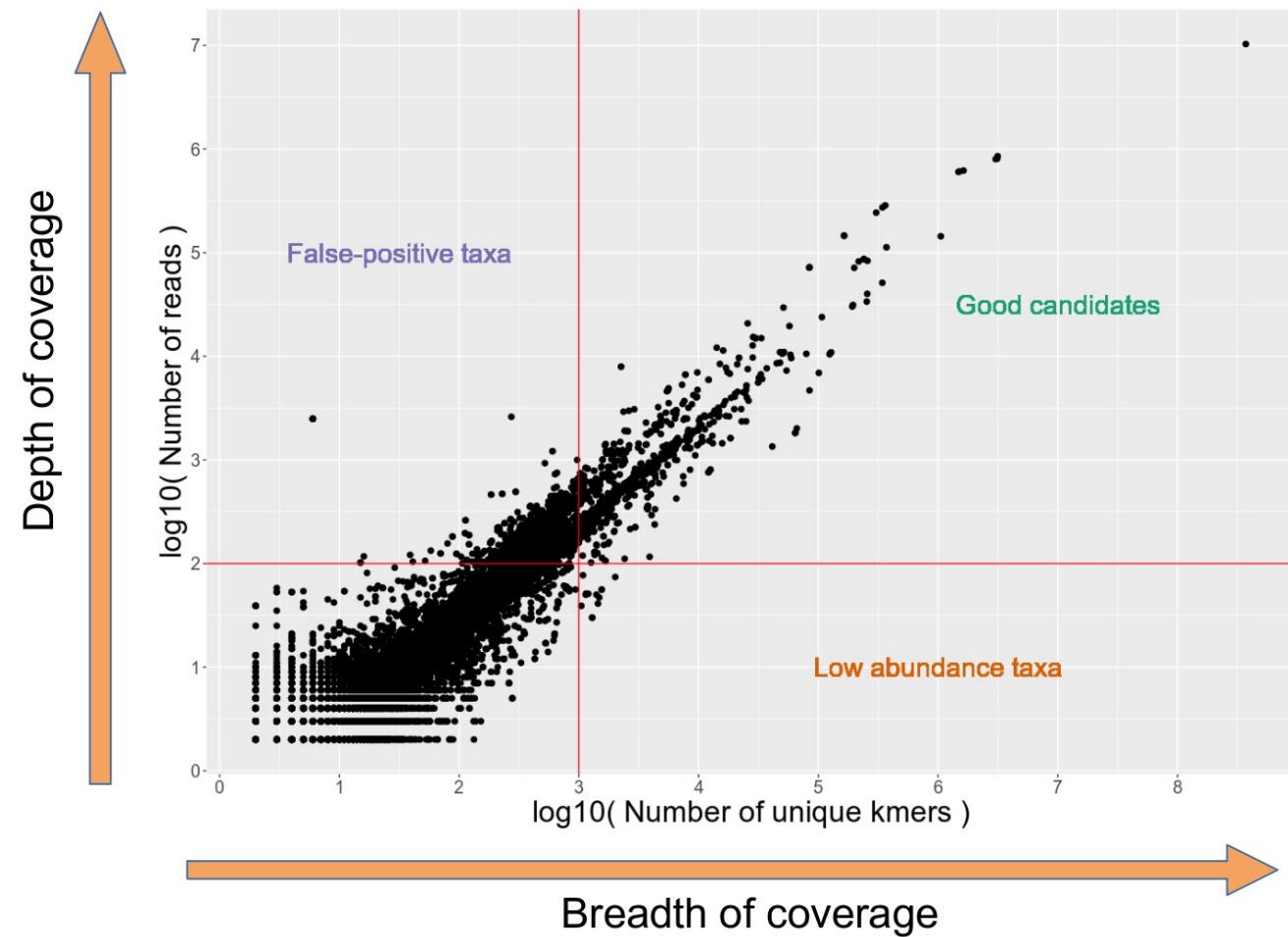


Both A) and B) have identical depth of coverage:
 $\text{Coverage} = (N_{\text{reads}} * L_{\text{read}}) / L_{\text{genome}} = (4 * L) / (4 * L) = 1X$



We can filter KrakenUniq output with respect to both **depth** and **breadth** of coverage.

The **number of unique k-mers** per taxon provided by KrakenUniq is a proxy for breadth of coverage, and can be used for filtering out false positive findings.

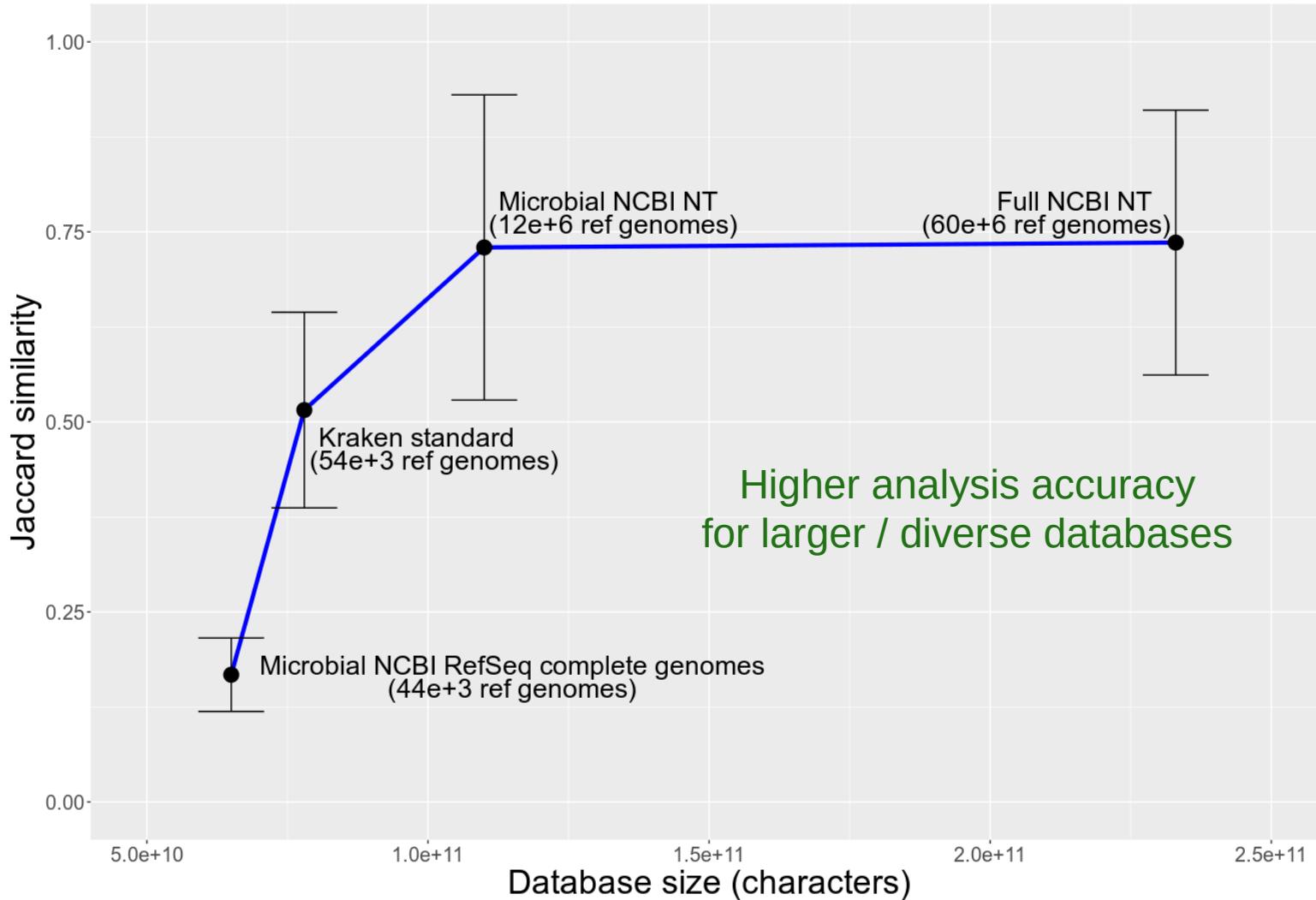


DATABASES!

DATABASES!

DATABASES!

DATABASES!

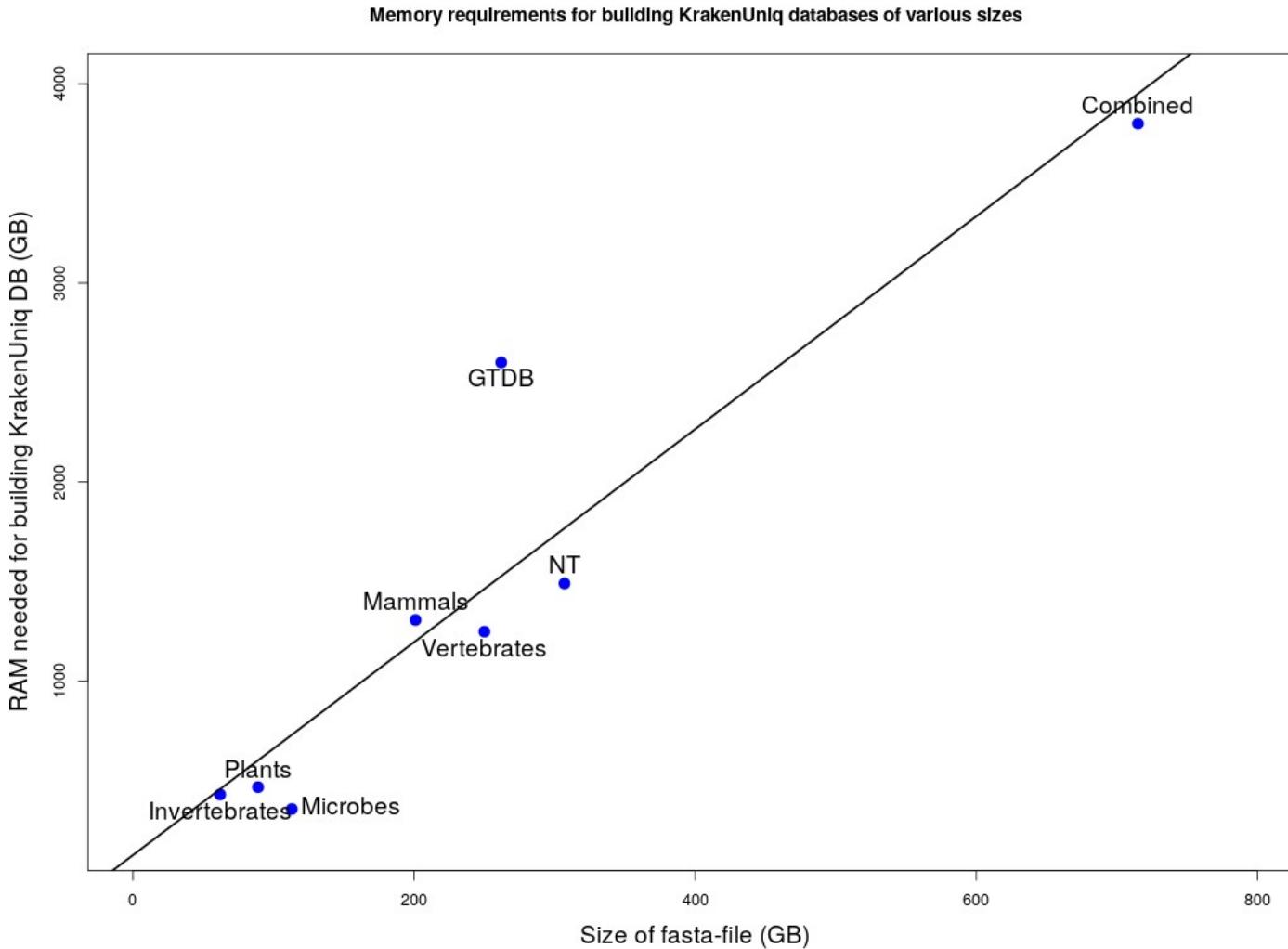




MIXTURE OF READS FROM DIFFERENT ORGANISMS



GRAND DATABASE



Computational challenges in ancient microbial metagenomics

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

METHOD

Genome Biology

Open Access

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

Zoé Pochon^{1,2†}, Nora Bergfeldt^{1,3,4†}, Emrah Kirdök⁵, Mário Vicente^{1,2}, Thijessen Naidoo^{1,2,6,7}, Tom van der Valk^{1,4}, N. Ezgi Altınlıskık⁸, Maja Krzewińska^{1,2}, Love Dalén^{1,3}, Anders Götherström^{1,2†}, Claudio Mirabello^{9†}, Per Unneberg^{10†} and Nikolay Oskolkov^{11†} 

¹Zoé Pochon, Nora Bergfeldt, Anders Götherström, Claudio Mirabello, Per Unneberg, and Nikolay Oskolkov shared authorship.

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

[†]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

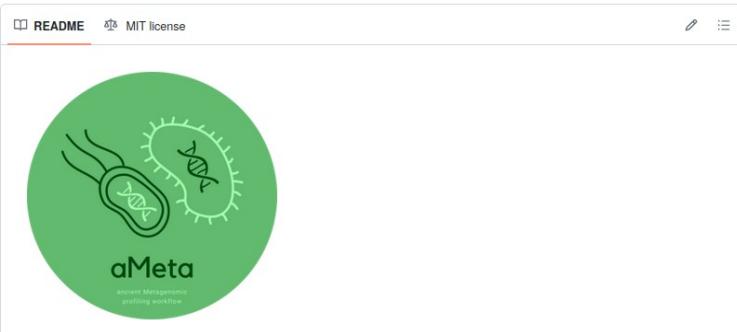
Abstract

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

Snakemake ≥6.10.0 Tests: passing

About

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

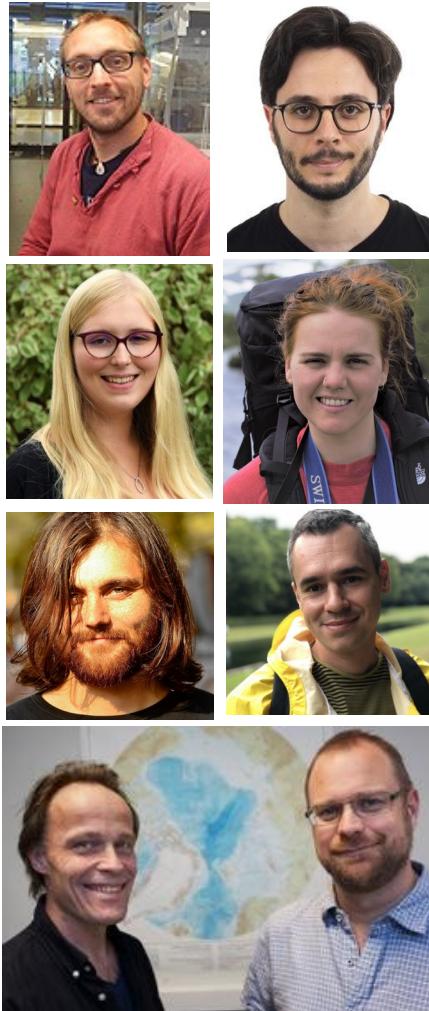
- trimming adapter sequences and removing reads shorter than 30 bp with Cutadapt
- quality control before and after trimming with FastQC and MultiQC
- taxonomic sequence kmer-based classification with KrakenUniq
- sequence alignment with Bowtie2 and screening for common microbial pathogens
- deamination 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 workflow for your research projects, please cite our preprint: <https://www.biorxiv.org/content/10.1101/2022.10.03.510579v1>

Authors

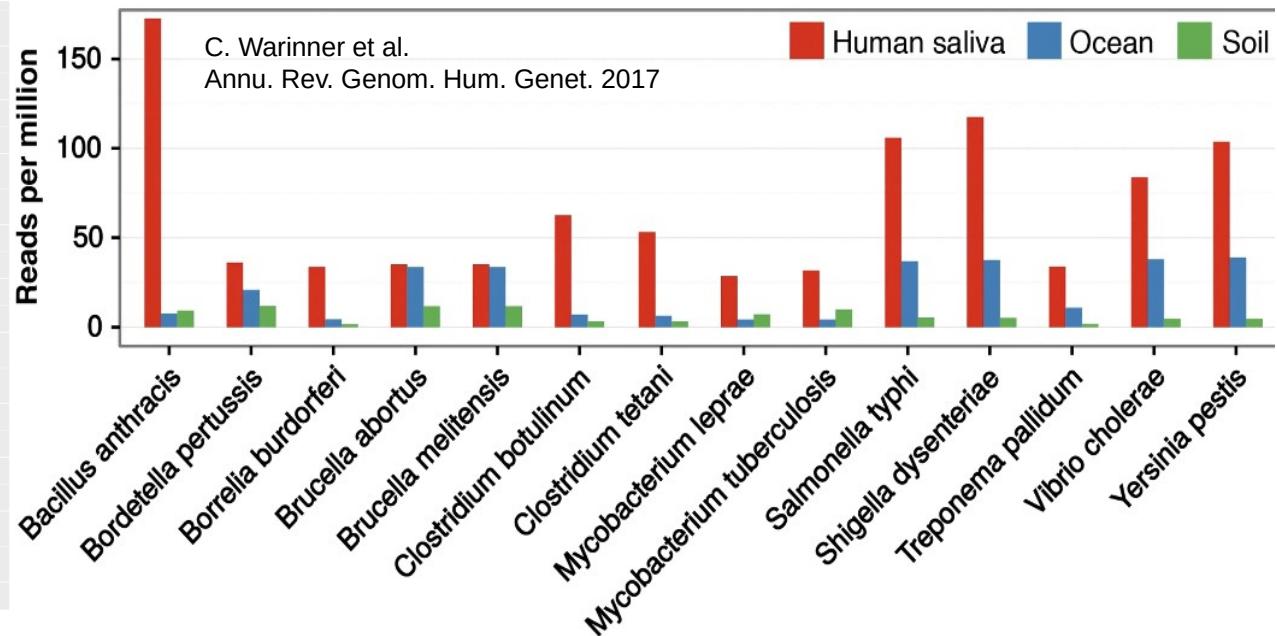
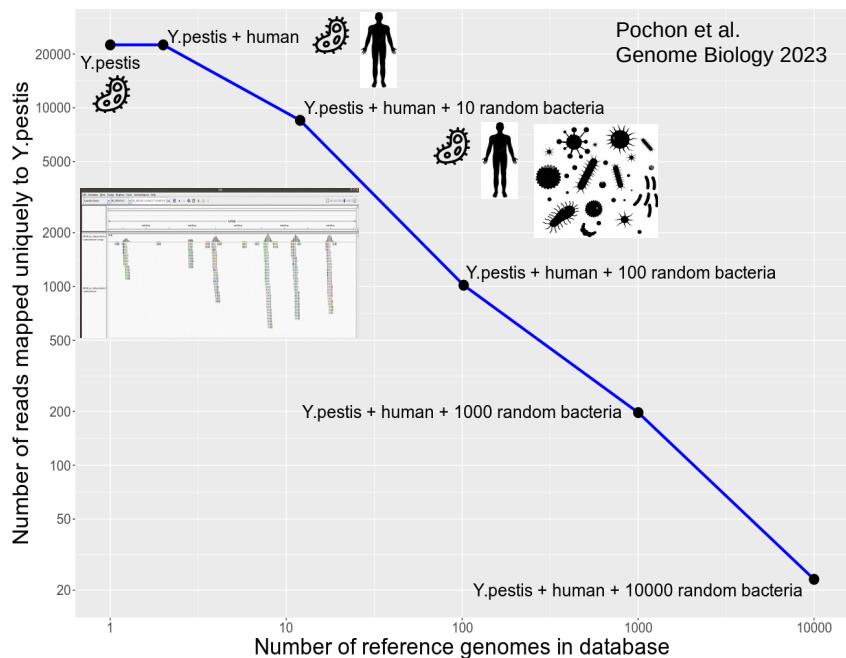
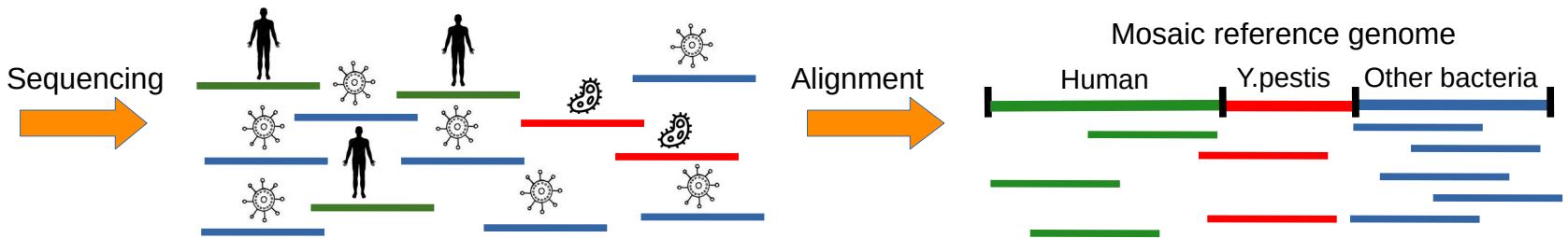
- Nikolay Oskolkov (@LeandroRitter) nikolay.oskolkov@scilifelab.se
- Claudio Mirabello (@clami66) claudio.mirabello@scilifelab.se
- Per Unneberg (@percyfal) per.unneberg@scilifelab.se

<https://github.com/NBISweden/aMeta>





Modern infant stool sample



Support the Guardian
Fund independent journalism with £12 per month

[Support us →](#)

[Print subscriptions](#) [Search jobs](#) [Sign in](#)

News Opinion Sport Culture Lifestyle

World Europe US Americas Asia Australia Middle East Africa Inequality Global development

US news

This article is more than 10 years old

Plague, anthrax and cheese? Scientists map bacteria on New York subway

- Research strongly downplays threat as most of the bacteria is harmless
- More than 1,000 samples were run through a DNA sequencer



Alan Yuhas in New York

Sat 7 Feb 2015 18.46 CET

[Share](#)



Watch

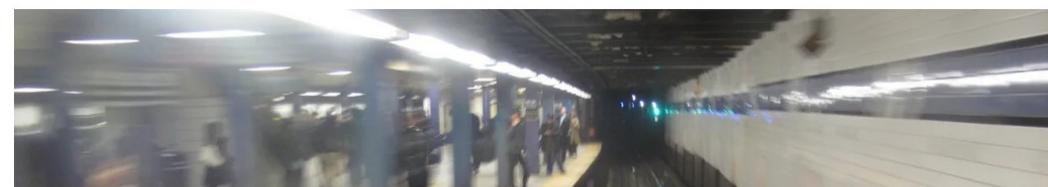
CNN US Crime + Justice

New York subway germ study derailed

By Lorenzo Ferrigno, CNN

2 minute read · Published 1:17 PM EDT, Fri August 14, 2015

[f](#) [X](#) [m](#) [s](#)



Cell Systems

Supports open access

This journal Journals Publish News & events About Cell Press

LETTER • Volume 1, Issue 1, P4-5, July 29, 2015 • Open Archive

[Download Full Issue](#)

Don't Worry, There's Probably No Bubonic Plague on NYC Subways

2 MINUTE READ

Lack of Evidence for Plague or Anthrax on the New York City Subway

Joel Ackelsberg ¹ Jennifer Rakeman ¹ Scott Hughes ¹ Jeannine Petersen ² Paul Mead ² Martin Schriener ² Luke Kingry ² Alex Hoffmaster ³ Jay E. Gee ³ Show less

Affiliations & Notes ▾ Article Info ▾ Linked Articles (2) ▾

[Download PDF](#)

» Main Text

Show Outline In their highly publicized report on the metagenomics of the NYC subway, Afschinnekoo and colleagues display an unfamiliarity with the genetics, microbiology, ecology, and epidemiology of some of the organisms they claim to have identified (Afschinnekoo et al., 2015).

Yersinia pestis, the cause of plague, was first introduced into North America circa 1900 through port cities along the Pacific and Gulf coasts. The organism spread into native rodent populations and became established in the arid western United States. Although a few cases occurred initially in New Orleans, LA, and Pensacola, FL, none were observed along the Atlantic coast, and the organism quickly died out in all areas east of Texas (Kugeler et al., 2015). Rodents, cats, and humans are all exquisitely susceptible to infection with *Y. pestis*, yet naturally occurring infection has never been observed within 1,000 miles of NYC. A plague outbreak in NYC's urban rat population, let alone sporadic human disease, would not go unnoticed. The authors' suggestion that humans and plague bacilli have "interacted (and potentially evolved)" in NYC is unfounded and without scientific merit.

Now the team of scientists, who spent 17 months in New York's sprawling subway system collecting microorganisms, say there is "minimal coverage to the backbone genome of these organisms, and there is no strong evidence to suggest these organisms are in fact present."

[View all journals](#) [Search](#)[Log in](#)[Sign up for alerts](#) [RSS feed](#)[Explore content](#) [About the journal](#) [Publish with us](#)[nature](#) > [articles](#) > [article](#)

Article | Published: 11 March 2020

RETRACTED ARTICLE: Microbiome analyses of blood and tissues suggest cancer diagnostic approach

Gregory D. Poore, Evgenia Kopylova, Oiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, Jad Kanbar, Sandrine Miller-Montgomery, Robert Heaton, Rana McKay, Sandip Pravin Patel, Austin D. Swafford & Rob Knight

[Nature](#) 579, 567–574 (2020) | [Cite this article](#)

107k Accesses | 675 Citations | 979 Altmetric | [Metrics](#)

This article was [retracted](#) on 26 June 2024

This article has been [updated](#)

Abstract

Systematic characterization of the cancer microbiome provides the opportunity to develop techniques that exploit non-human, microorganism-derived molecules in the diagnosis of a major human disease. Following recent demonstrations that some types of cancer show substantial microbial contributions^{1,2,3,4,5,6,7,8,9,10}, we re-examined whole-genome and whole-transcriptome sequencing studies in The Cancer Genome Atlas¹¹ (TCGA) of 33 types of cancer from treatment-naïve patients (a total of 18,116 samples) for microbial reads, and found unique microbial signatures in tissue and blood within and between most major types of cancer. These TCGA blood signatures remained predictive when applied to patients with stage Ia–IIc cancer and cancers lacking any genomic alterations currently measured on two commercial-grade cell-free tumour DNA platforms, despite the use of very stringent decontamination analyses that discarded up to 92.3% of total sequence data. In addition, we could discriminate among samples from healthy, cancer-free individuals ($n=69$) and those from patients with multiple types of cancer (prostate, lung, and melanoma; 100 samples in total) solely using plasma-derived, cell-free microbial nucleic acids. This potential microbiome-based oncology diagnostic tool warrants further exploration.

You have full access to this article via Lund University

[Download PDF](#)

Associated content

RETRACTED ARTICLE: AI finds microbial signatures in tumours and blood across cancer types

Nadim J. Ajami & Jennifer A. Wargo
[Nature](#) | [News & Views](#) | 11 Mar 2020

[Sections](#) [Figures](#) [References](#)

[Abstract](#)

[Main](#)

[TCGA cancer microbiome and its normalization](#)

[Predicting among and within types of cancer](#)

[Biological relevance of microorganism profiles](#)

[Measuring and mitigating contamination](#)

[Predictions using microbial DNA in blood](#)

[Validating microbial signatures in blood](#)

[Discussion](#)

[Methods](#)

[Data availability](#)

[Code availability](#)

[Change history](#)

[References](#)

'All authors agree' to retraction of Nature article linking microbial DNA to cancer

A 2020 paper that claimed to find a link between microbial genomes in tissue and cancer has been retracted following an analysis that called the results into question.

The paper, “[Microbiome analyses of blood and tissues suggest cancer diagnostic approach](#),” was published in March 2020 and has been cited 610 times, according to Clarivate’s Web of Science. It was retracted June 26. The study was also key to the formation of biotech start-up [Micronoma](#), which did not immediately respond to our request for comment.

[Rob Knight](#), corresponding author and researcher at the University of California San Diego, also did not immediately respond to our request for comment.

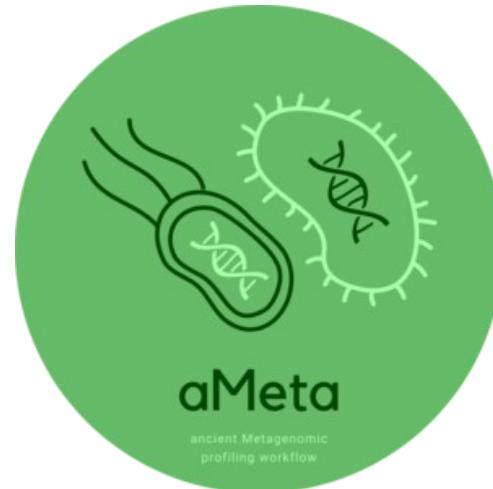
In October 2023, *mBio*, a journal from the American Society for Microbiology, published “[Major data analysis errors invalidate cancer microbiome findings](#).” The paper pointed out several major flaws in the earlier article by Knight’s group.



aMeta goal: minimize amount of false-positives

aMeta key principles

- 1) unbiased (ultra-)competative mapping
- 2) large / diverse reference databases
- 3) controlling evenness of coverage



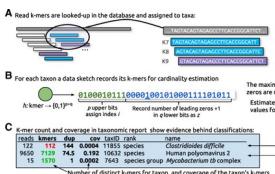
Sample1

Sample2

Sample3

SampleN

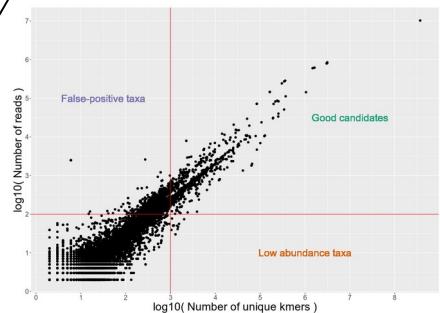
KrakenUniq: pre-screening + filtering



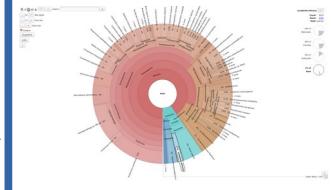
Build Project-Specific MALT Database



Pathogen screening: Bowtie2 + mapDamage



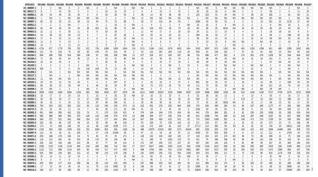
Visualization



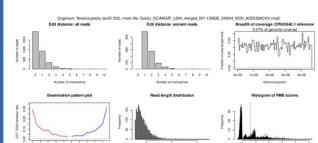
MALT: LCA-alignment + MaltExtract

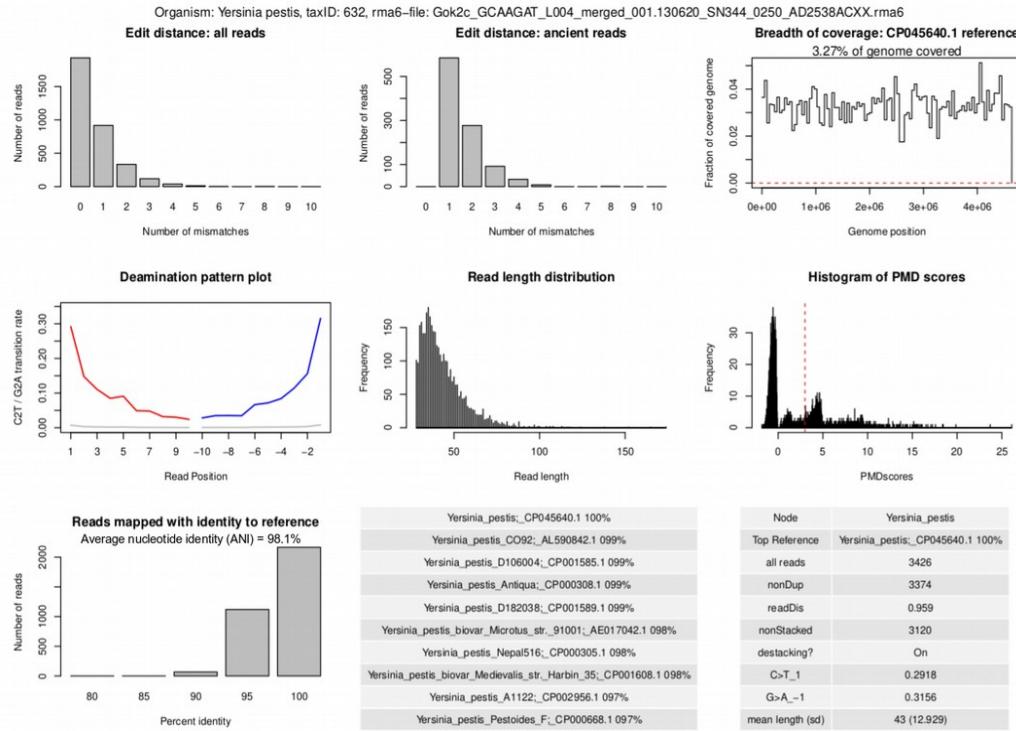


Abundance quantification

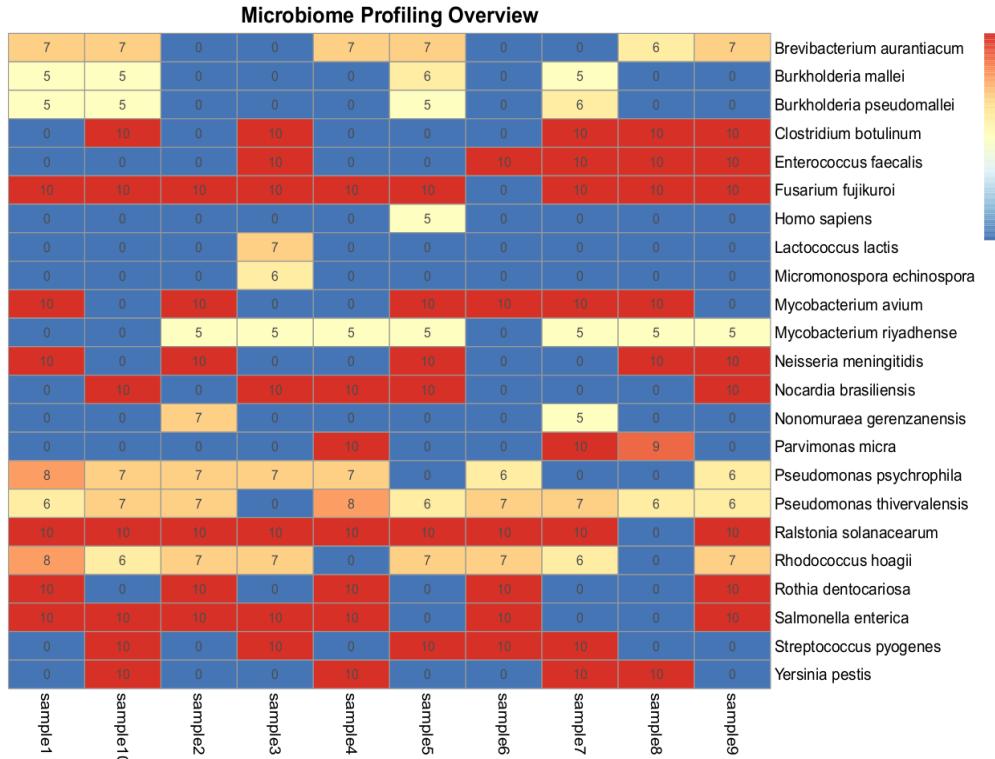


Validation Authentication





Eight quality metrics for assessing presence of microbial organisms and their ancient status



Quantified authentication and validation quality metrics as **scores**

BMC Part of Springer Nature

Search Explore journals Get published About BMC Login

Genome Biology

Home About Articles Submission Guidelines

Method | Open Access | Published: 16 December 2019

HOPS: automated detection and authentication of pathogen DNA in archaeological remains

Ron Hübler, Felix M. Key, Christina Warinner, Kirsten I. Bos, Johannes Krause & Alexander Herbig

Genome Biology 20, Article number: 280 (2019) | Cite this article

5164 Accesses | 35 Citations | 26 Altmetric | Metrics

Abstract

High-throughput DNA sequencing enables large-scale metagenomic analyses of complex biological systems. Such analyses are not restricted to present-day samples and can also be applied to molecular data from archaeological remains. Investigations of ancient microbes can provide valuable information on past bacterial commensals and pathogens, but their molecular detection remains a challenge. Here, we present HOPS (Heuristic Operations for Pathogen Screening), an automated bacterial screening pipeline for ancient DNA sequences that provides detailed information on species identification and authenticity. HOPS is a versatile tool for high-throughput screening of DNA from archaeological material to identify candidates for genome-level analyses.

Download PDF

Collection
Microbiome Biology

Sections Figures References

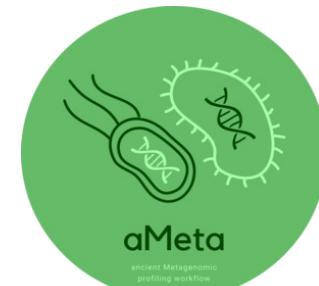
[Abstract](#)
[Background](#)
[Results](#)
[Discussion](#)
[Conclusions](#)
[Methods](#)
[Availability of data and materials](#)
[References](#)
[Acknowledgements](#)



Nice design:
 1) who is there?
 2) how ancient?

SEVERE LIMITATIONS:

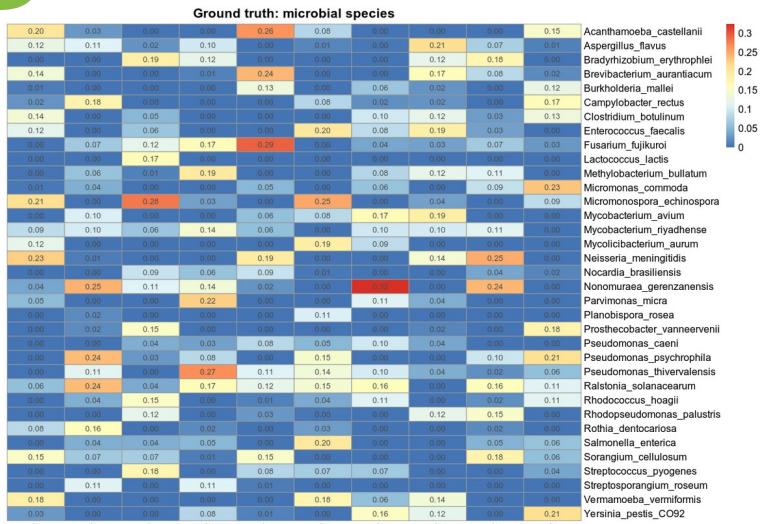
- 1) HOPS / MALT is very resource demanding
- 2) HOPS / MALT only feasible with limited size databases
- 3) HOPS does not use evenness and breadth of coverage



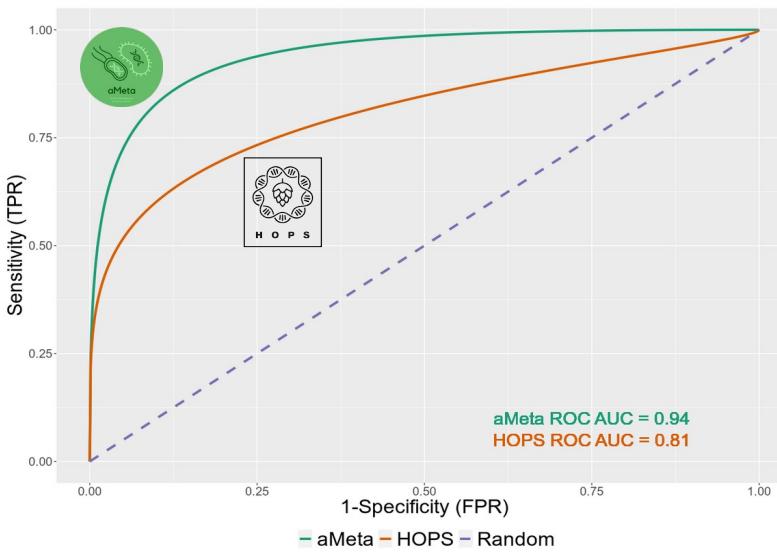
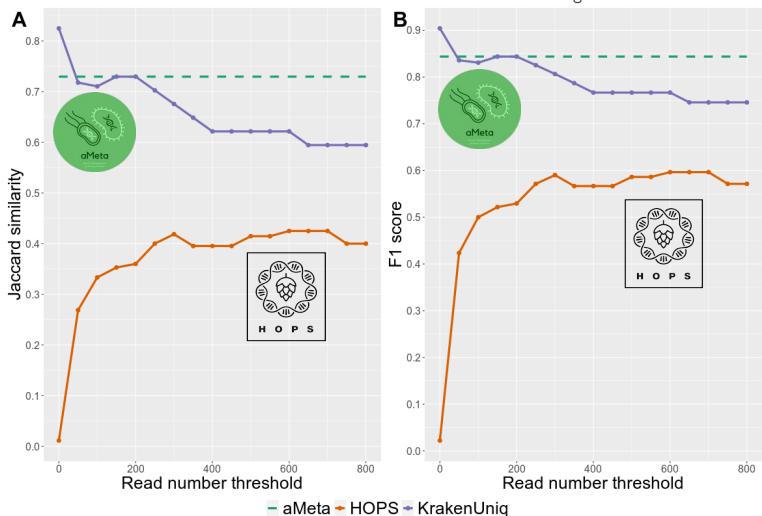
aMeta:
 a way to overcome
 HOPS limitations

aMeta vs HOPS benchmarking

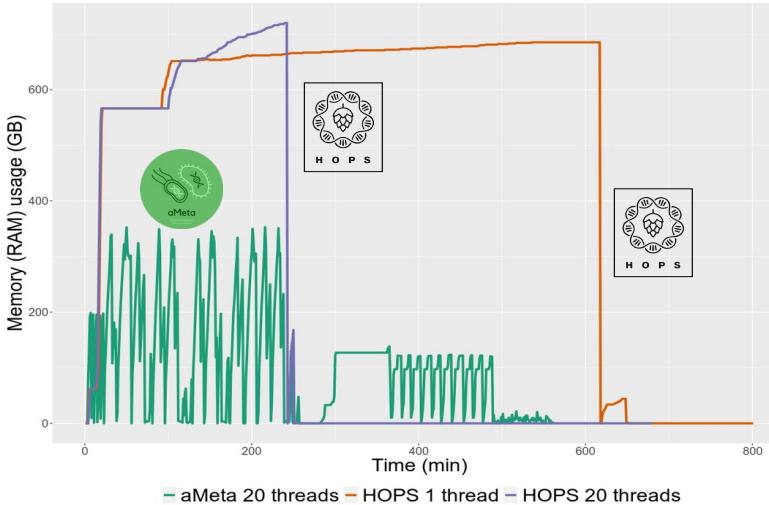
Simulated
ground truth



Detection
error



Authentication
error



Resource
usage

Computational challenges in environmental metagenomics

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

Nikolay Oskolkov  1,*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

¹Department of Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Lund University, SE-223 62 Lund, Sweden

²Centre for Palaeogenetics, Svante Arrhenius väg 20C, SE-10691 Stockholm, Sweden

³Department of Bioinformatics and Genetics, Swedish Museum of Natural History, SE-104 05 Stockholm, Sweden

⁴Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden

⁵Department of Geological Sciences, Stockholm University, SE-106 91 Stockholm, Sweden

⁶Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Solna, SE-106 91 Stockholm, Sweden

⁷Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, SE-751 24 Uppsala, Sweden

⁸Scilifelab, 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.

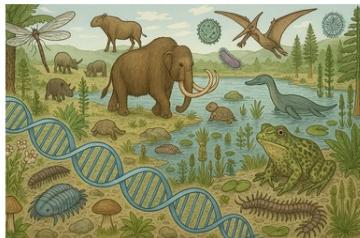
to genomic reference databases. Consequently, the quality of both the aeDNA data and the reference databases is crucial for reliable inferences. Microbial-like sequences in reference genomic databases, originating either from nonendogenous sources (contamination) or from evolutionary similarity to microbial 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.

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

 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	

 README  License

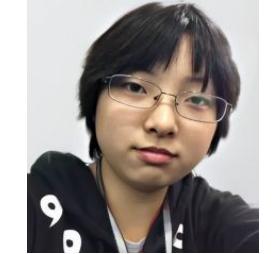

GENEX
GENOME EXOGENOUS



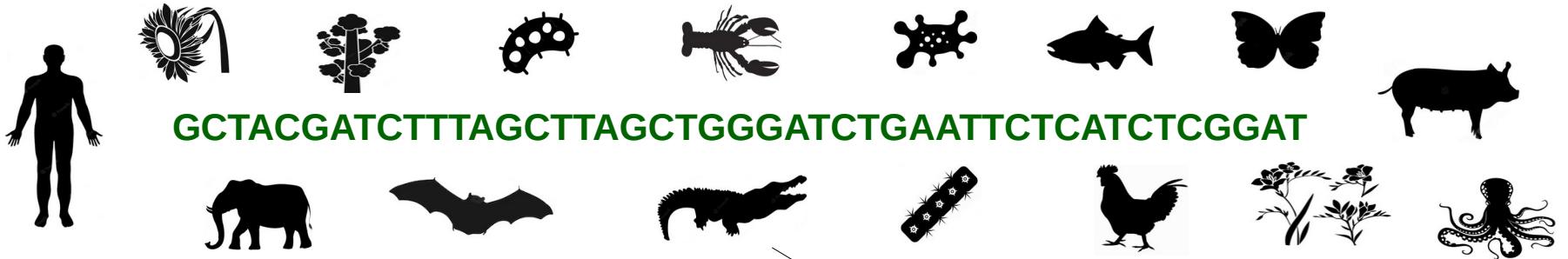
GENome EXogenous (GENEX) sequence detection

This is a computational workflow for detecting coordinates of microbial-like or human-like sequences in eukaryotic and prokaryotic 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>



Does it belong to the organism I am thinking about, or is it contamination?



Elephant

Human

Sunflower

Microbes



Reference genomes: are they OK or contaminated?

Research

Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P. Breitwieser,¹ Mihaela Pertea,^{1,2} Aleksey V. Zimin,^{1,3} and Steven L. Salzberg^{1,2,3,4}

¹Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA; ²Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA; ³Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA; ⁴Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA

Contaminant sequences that appear in published genomes can cause numerous problems for downstream analyses, particularly for evolutionary studies and metagenomics projects. Our large-scale scan of complete and draft bacterial and archaeal genomes in the NCBI RefSeq database reveals that 2250 genomes are contaminated by human sequence. The contaminant sequences derive primarily from high-copy human repeat regions, which themselves are not adequately represented in the current human reference genome, GRCh38. The absence of the sequences from the human assembly offers a likely explanation for their presence in bacterial assemblies. In some cases, the contaminating contigs have been erroneously annotated as containing protein-coding sequences, which over time have propagated to create spurious protein “families” across multiple prokaryotic and eukaryotic genomes. As a result, 3437 spurious protein entries are currently present in the widely used nr and TrEMBL protein databases. We report here an extensive list of contaminant sequences in bacterial genome assemblies and the proteins associated with them. We found that nearly all contaminants occurred in small contigs in draft genomes, which suggests that filtering out small contigs from draft genome assemblies may mitigate the issue of contamination while still keeping nearly all of the genuine genomic sequences.

OPEN ACCESS Freely available online



Abundant Human DNA Contamination Identified in Non-Primate Genome Databases

Mark S. Longo, Michael J. O'Neill, Rachel J. O'Neill*

Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, United States of America

Abstract

During routine screens of the NCBI databases using human repetitive elements we discovered an unlikely level of nucleotide identity across a broad range of phyla. To ascertain whether databases containing DNA sequences, genome assemblies and trace archive reads were contaminated with human sequences, we performed an in depth search for sequences of human origin in non-human species. Using a primate specific SINE, AluY, we screened 2,749 non-primate public databases from NCBI, Ensembl, JGI, and UCSC and have found 492 to be contaminated with human sequence. These represent species ranging from bacteria (*B. cereus*) to plants (*Z. mays*) to fish (*D. rerio*) with examples found from most phyla. The identification of such extensive contamination of human sequence across databases and sequence types warrants caution among the sequencing community in future sequencing efforts, such as human re-sequencing. We discuss issues this may raise as well as present data that gives insight as to how this may be occurring.

Citation: Longo MS, O'Neill MJ, O'Neill RJ (2011) Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. PLoS ONE 6(2): e16410. doi:10.1371/journal.pone.0016410

Editor: Najib El-Sayed, The University of Maryland, United States of America

Received September 1, 2010; Accepted December 23, 2010; Published February 16, 2011

Copyright © 2011 Longo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the NSF (www.nsf.gov). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: racheloneill@uconn.edu

Current Biology



PeerJ

Volume 22, Issue 15, 7 August 2012, Pages R593-R594

Correspondence

Origin of land plants revisited in the light of sequence contamination and missing data

Simon Laurin-Lemay, Henner Brinkmann, Hervé Philippe✉

Science & Society

EMBO
reports

Here, there, and everywhere

From PCRs to next-generation sequencing technologies and sequence databases, DNA contaminants creep in from the most unlikely places

Karl Gruber

No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*

Georgios Koutsopoulos*, Sujaí Kumar*, Dominik R. Laetsch^{a,b}, Lewis Stevens*, Jennifer Daub^c, Claire Conlon^c, Habib Maroon^c, Fran Thomas^c, Aziz Aboobaker^c, and Mark Blaxter^{a,c}

^aInstitute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JL, United Kingdom; ^bThe James Hutton Institute, Dundee DD2 5DA, United Kingdom; and ^cDepartment of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved March 1, 2016 (received for review January 8, 2016)

Tardigrades are meiofaunal ecdysozoans that are key to understanding the origins of Arthropoda. Many species of Tardigrada can survive extreme conditions throughout their life cycle. In a recent paper (Steininger TC, et al., 2015) PeerJ Mat. Asiat. 1:12521:15976–15981), the authors concluded that the tardigrade *Hypsibius dujardini* had an unprecedented proportion (17%) of genes originating through functional horizontal gene transfer (HGT) and suggested that HGT was likely forming a significant proportion of its genome. We independently sequenced the genome of *H. dujardini*. As expected from whole-genome DNA sampling, our raw data contained reads from non-target genomes. Filtering using metagenomic approaches generated a draft *H. dujardini* genome assembly of 132,000,000 bp with 1,000 scaffolds and undergoes further assembly. Additional microbial contaminants likely remain. We found no support for extensive HGT. Among 23,021 gene predictions we identified 0.2% strong candidates for HGT from bacteria and 0.2% strong candidates for HGT from non-arthropod eukaryotes. The vast majority of HGT candidates are from arthropods, suggesting a majority of HGT candidates in the Boothby et al. genome derived from contaminants. We conclude that HGT in *H. dujardini* accounts for at most 1–2% of genes and that the proposal that one sixth of tardigrade genes originate from functional HGT events is an artifact of undetected contamination.

tardigrade | biotools | contamination | metagenomics | horizontal gene transfer

Steininger and Salzberg *Genome Biology* (2020) 21:115
https://doi.org/10.1186/s13059-020-02023-1

Genome Biology

Open Access



METHOD

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

Martin Steineger^{1,2,*} and Steven L. Salzberg^{2,4,5}

*Correspondence:

martin.steineger@rsmu.ac.kr
¹School of Biological Sciences,
Seoul National University, Seoul,
08826, South Korea

²Center for Computational Biology,
Whiting School of Engineering,
Johns Hopkins University, 21218
Baltimore, Maryland, USA
Full list of author information is
available at the end of the article

Abstract

Genomic analyses are sensitive to contamination in public databases caused by incorrectly labeled reference sequences. Here, we describe Conterminator, an efficient method to detect and remove incorrectly labeled sequences by an exhaustive all-against-all sequence comparison. Our analysis reports contamination of 2,161,746, 114,035, and 14,148 sequences in the RefSeq, GenBank, and NR databases, respectively, spanning the whole range from draft to “complete” model organism genomes. Our method scales linearly with input size and can process 3.3 TB in 12 days on a 32-core computer. Conterminator can help ensure the quality of reference databases. Source code (GPLv3): <https://github.com/martin-steineger/conterminator>

Keywords: Genomes, Contamination, Software, RefSeq, GenBank

Unexpected cross-species contamination in genome sequencing projects

Samier Merchant^{1,2}, Derrick E. Wood^{1,3} and Steven L. Salzberg^{1,3,4}

¹Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA

²Department of Computer Science, Brown University, Providence, RI, USA

³Department of Computer Science, Johns Hopkins University, USA

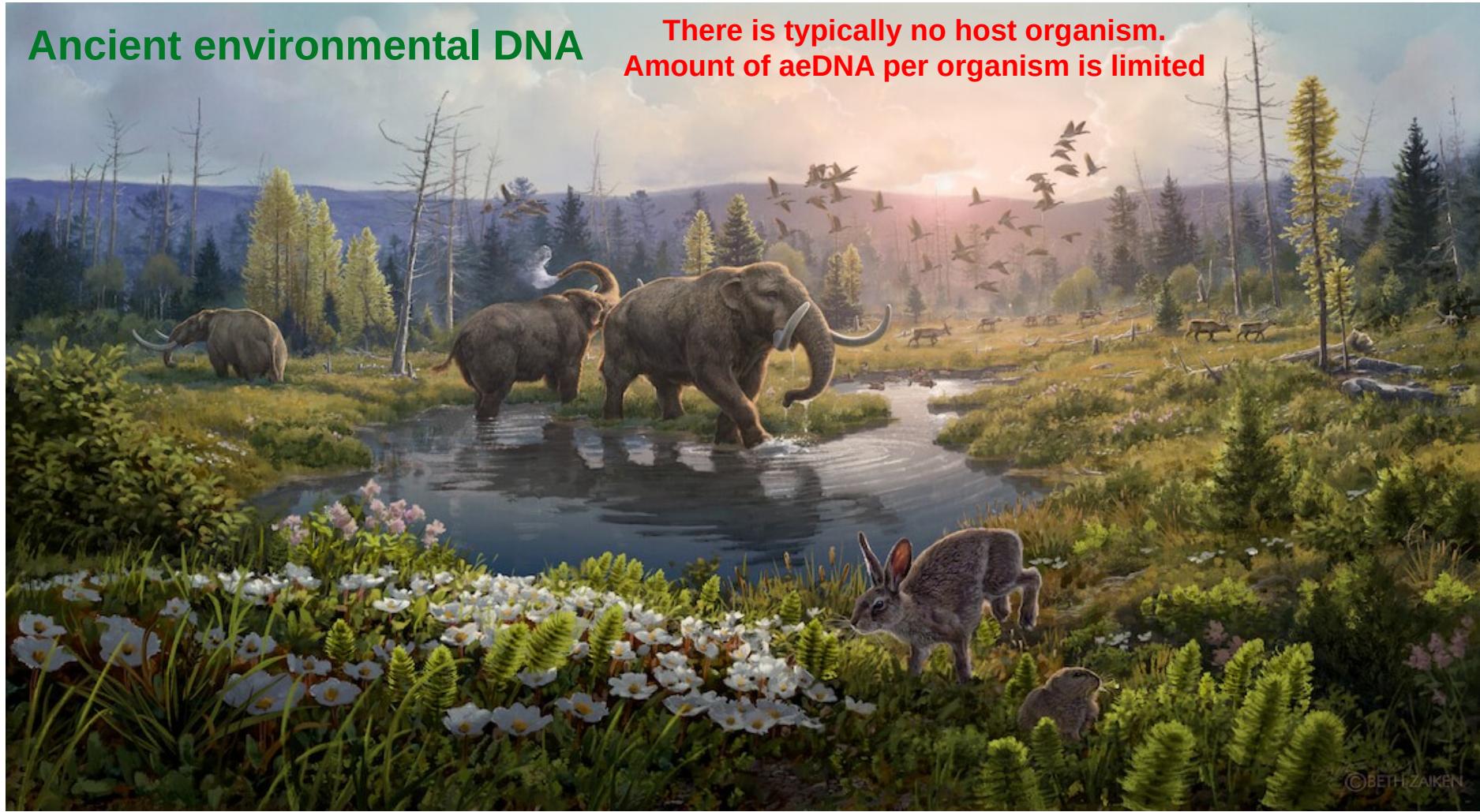
⁴Department of Biomedical Engineering, Johns Hopkins University, USA

ABSTRACT

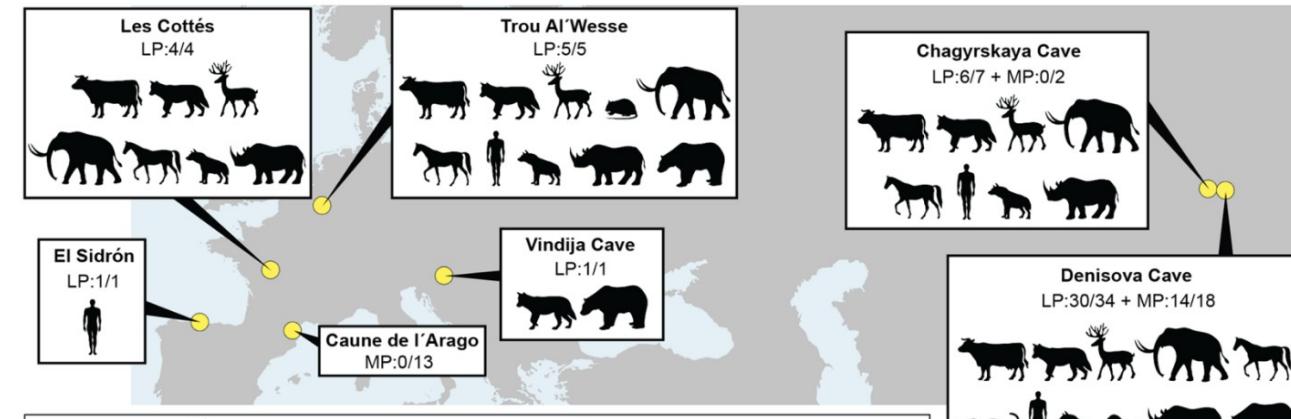
The raw data from a genome sequencing project sometimes contains DNA from contaminating organisms, which may be introduced during sample collection or sequence preparation. In some instances, these contaminants remain in the sequence even after assembly and deposition of the genome into public databases. As a result, searches of these databases may yield erroneous and confusing results. We used efficient microbiome analysis software to scan the draft assembly of domestic cow, *Bos taurus*, and identify 173 small contigs that appeared to derive from microbial contaminants. In the course of verifying these findings, we discovered that one genome, *Neisseria gonorrhoeae* TCDC-NG08107, although putatively a complete genome, contained multiple sequences that apparently derived from the cow and sheep genomes. Our findings illustrate the need to carefully validate findings of anomalous DNA that rely on comparisons to either draft or finished genomes.

Ancient environmental DNA

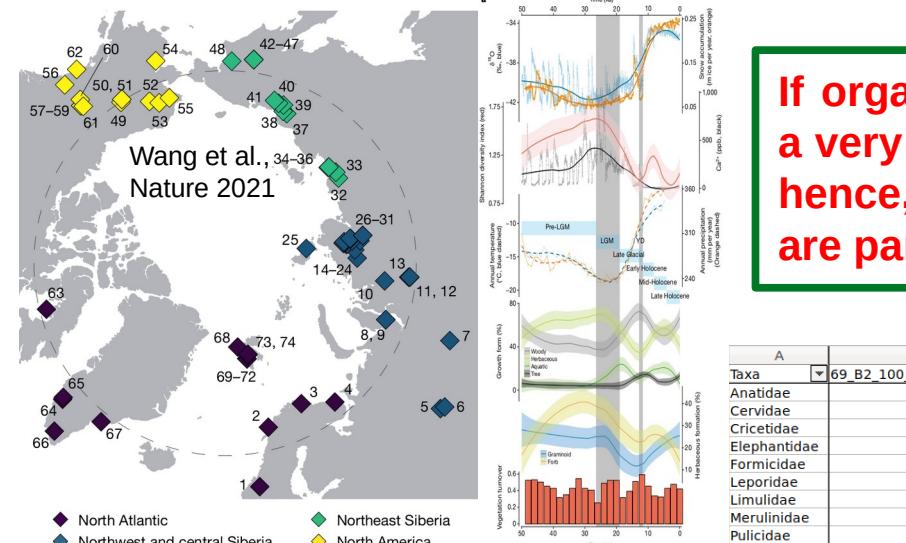
There is typically no host organism.
Amount of aeDNA per organism is limited



An artist's impression of the Kap København formation two-million years ago in a time where the temperature was significantly warmer than northernmost Greenland today. Artist: Beth Zaiken / bethzaiken.com



V. Slon et al., Science 2017



If organism detection relies on a very limited number of reads; hence, high-quality references are particularly important!

A

Taxa	69_B2_100_L0_KapK_12_1_34	69_B2_100_L0_KapK_12_1_34_C	69_B2_100_L0_KapK_12_1_35	69_B2_100_L0_KapK_12_1_35_C	69_B2_100_L0_KapK_12_1_36
Anatidae	372	3	0	0	70
Cervidae	84	0	0	0	0
Cricetidae	1331	17	1890	0	0
Elephantidae	85	0	171	48	118
Formicidae	0	5	0	19	0
Leporidae	542	5	549	0	313
Limulidae	66	0	0	0	23
Merulinidae	18	0	0	0	0
Pulicidae	0	0	0	0	0

B

Sites	119	50	69	74a	74b
Anatidae	Red	Red	Red	Red	Red
Cervidae	Red	Grey	Red	Red	Red
Cricetidae	Red	Grey	Red	Red	Red
Elephantidae	Red	Grey	Red	Red	Red
Leporidae	Red	Red	Red	Red	Red
Formicidae	Red	Grey	Red	Red	Red
Pulicidae	Red	Grey	Red	Red	Red
Limulidae	Red	Grey	Red	Red	Red
Merulinidae	Red	Grey	Red	Red	Red

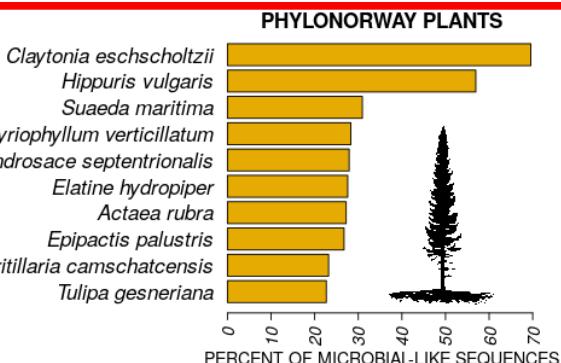
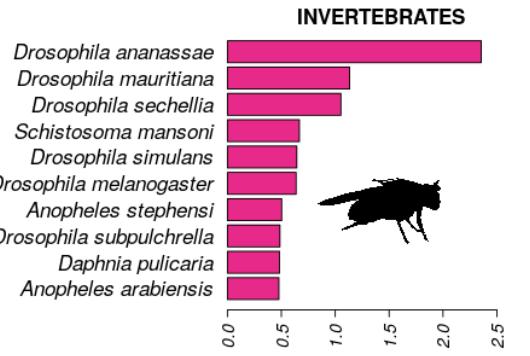
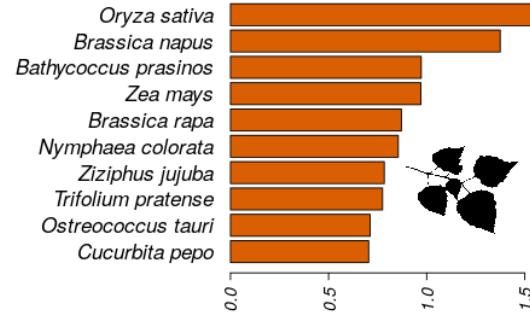
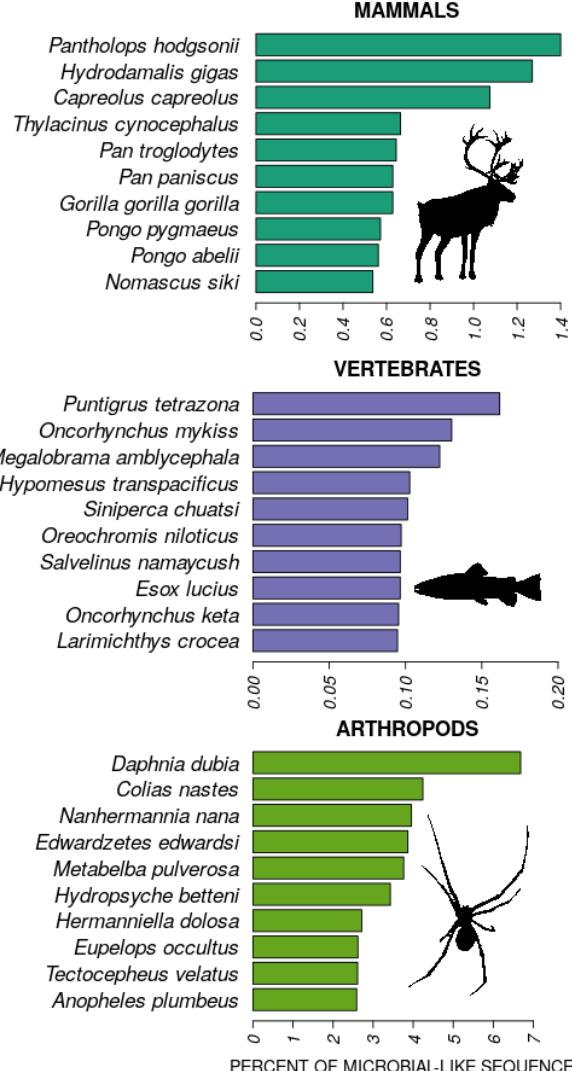
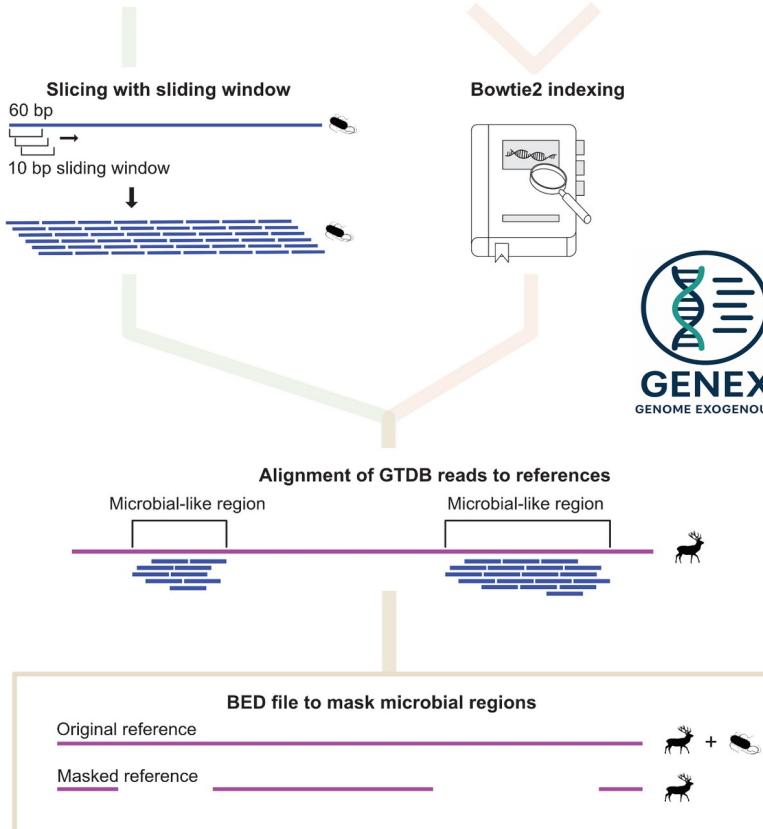
C

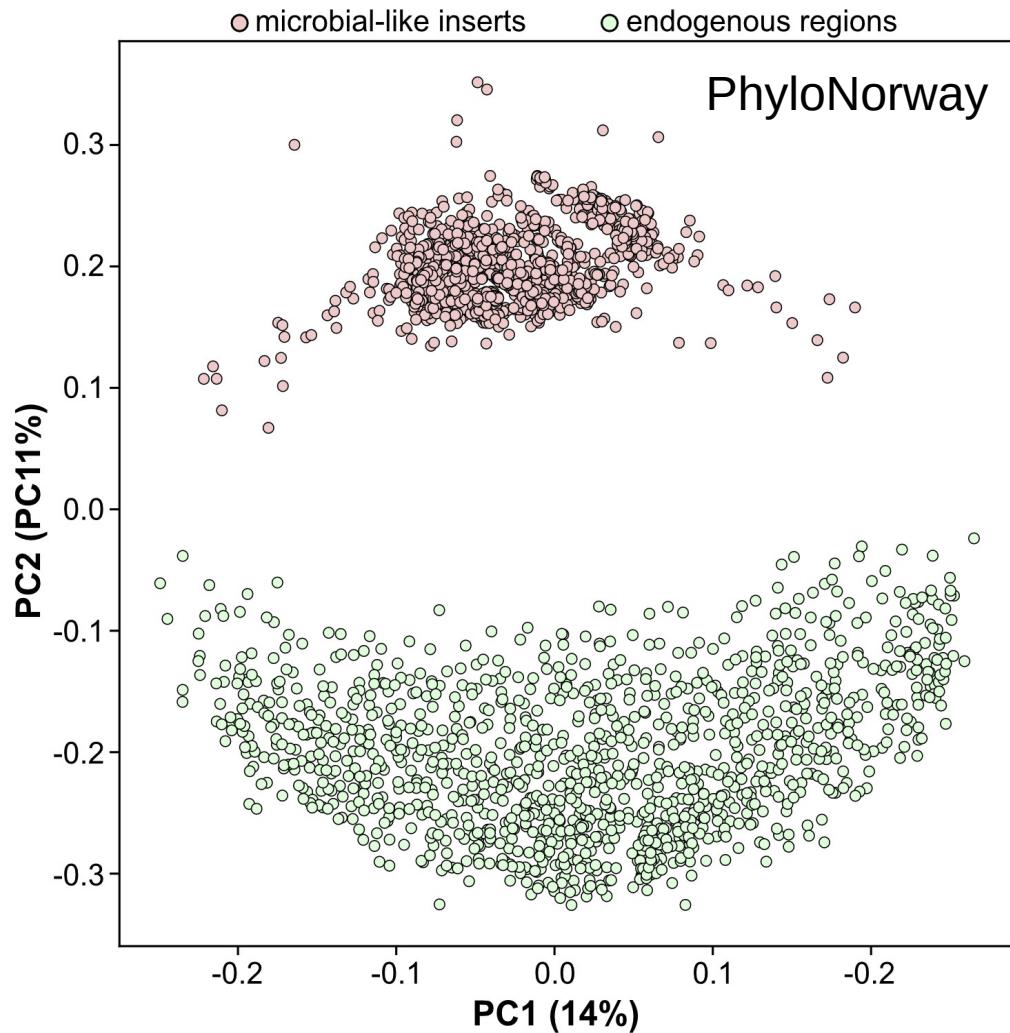
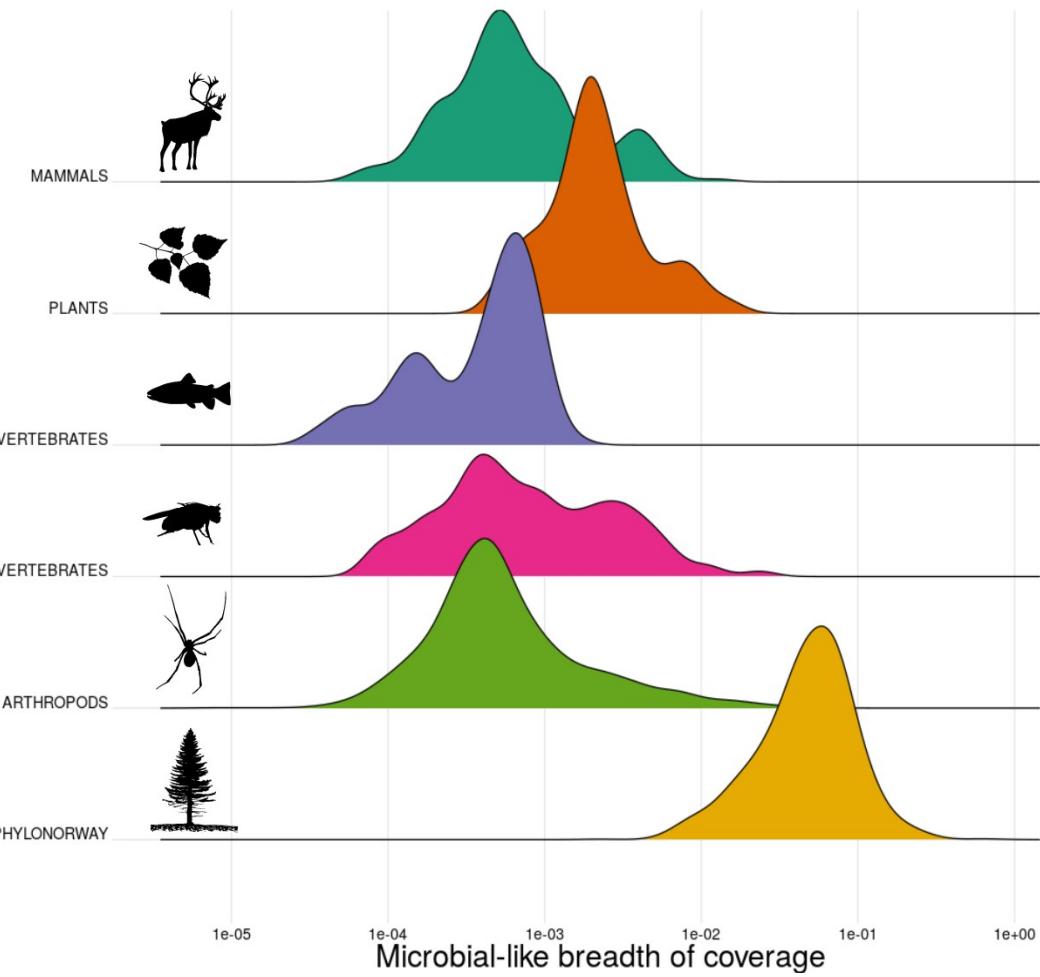
Kjaer et al., Nature 2022

Microbial-like sequences in eukaryotes

GTDB database (release 214.0)	RefSeq reference genomes (chromosome & scaffold level)	PhyloNorway (contig level)
Bacteria (n= 394,932)	Mammals (n=566)	Plants (n=1323)
Archaea (n=7,777)	Other vertebrates (n=162)	Invertebrates (n=114)

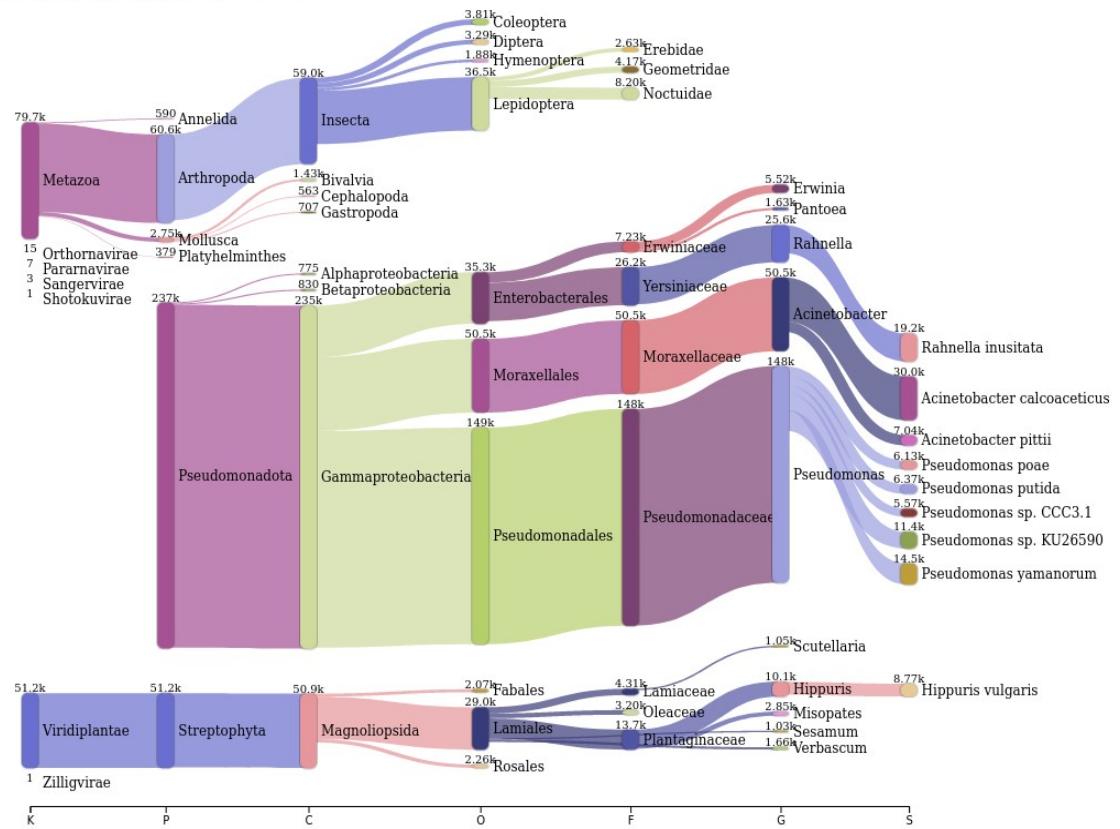
- Mammals (n=566)
- Other vertebrates (n=162)
- Invertebrates (n=114)
- Arthropods (n=2033)
- Plants (n=96)



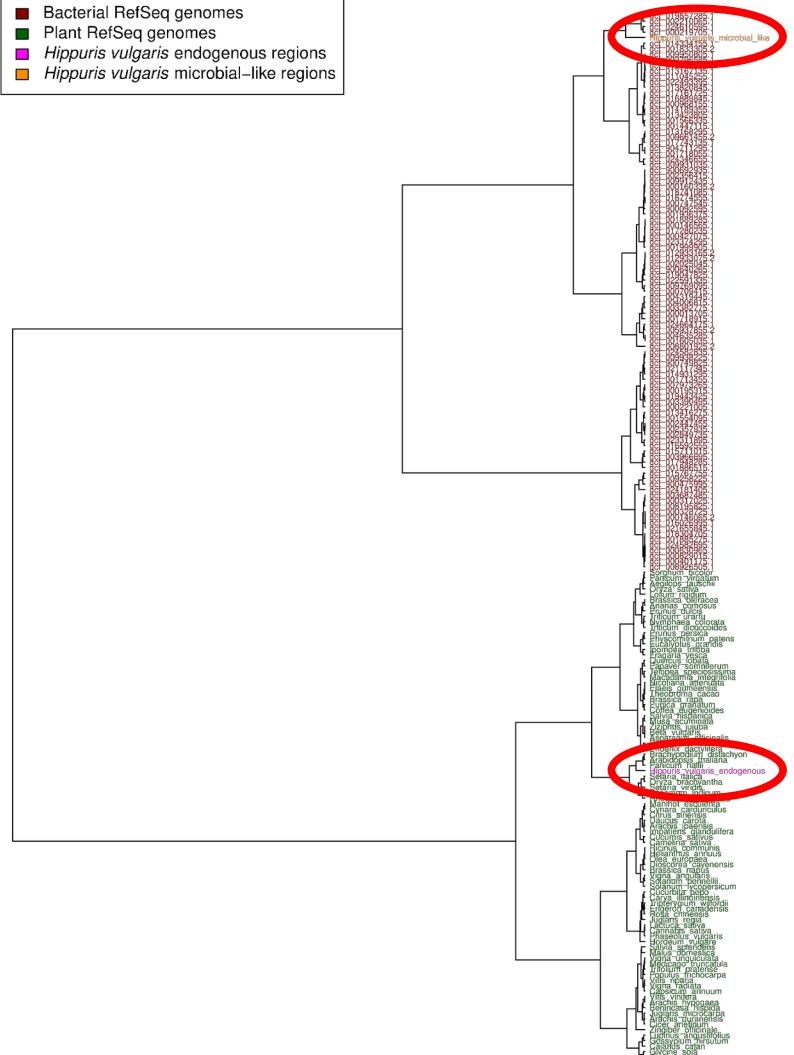




Screening *Hippuris vulgaris* contigs with Kraken2 against NCBI NT DB



- Bacterial RefSeq genomes
- Plant RefSeq genomes
- Hippuris vulgaris* endogenous regions
- Hippuris vulgaris* microbial-like regions



Article

Late Quaternary dynamics of Arctic biota from ancient environmental genomics

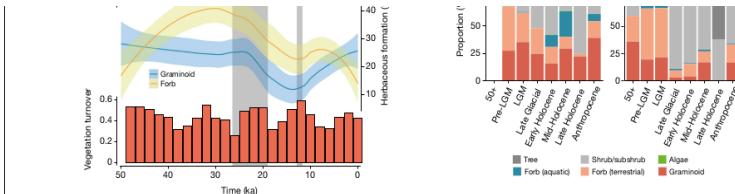


Fig. 2 | Climate and vegetation changes over the past 50 kyr. **a.** Pan-Arctic climate changes and vegetation variations. LGM (26.5–19 ka) and Younger Dryas (YD) (12.9–11.7 ka) are indicated by grey bars. The six time intervals are indicated by light blue bars (Supplementary Information 2). The error bands denotes \pm one standard deviation. **b.** The number of observed genera in different regions. **c.** Regional vegetation turnovers. **d.** Regional vegetation morphological compositions. The samples sizes for each region and time interval are provided in Supplementary Information 2. Calculations are supplied in the Methods.

modelled annual temperature and precipitation for all eDNA sampling sites; the proportion of plant growth forms; and the proportion of herbaceous plant growth forms; and the vegetation turnover rates. **b.** The number of observed genera in different regions. **c.** Regional vegetation turnovers. **d.** Regional vegetation morphological compositions. The sample sizes for each region and time interval are provided in Supplementary Information 2. Calculations are supplied in the Methods.

Regional vegetation dynamics

Underlying the generalized pattern of Holarctic vegetation changes are significant geographical differences. Early in postglacial times, the North Atlantic experienced the sharpest rises in taxonomic richness (Fig. 2b), along with the steepest temperature increase (Extended Data Fig. 2b). The increase in postglacial richness was probably driven by species dispersals coupled with habitat diversification¹⁷, that is, geomorphically dynamic substrates that were exposed by glacial retreat and shaped by meltwater. The resultant vegetation initially had low diversity but was rich in aquatic taxa (Fig. 2b, d). The abundance of aquatic taxa relates in part to the prevalence of samples from lakes in the North Atlantic (Supplementary Information 10), but nonetheless

highlights the ability of aquatic plants to disperse rapidly into newly deglaciated terrain containing abundant streams and lake basins¹⁸. As the postglacial climate continued to warm, the overall proportion of aquatic taxa declined as trees and shrubs (for example, *Betula*, *Salix* and *Vaccinium*) became abundant in this region (Fig. 2d and Extended Data Fig. 3).

Northeast Siberia and North America experienced less radical postglacial changes in vegetation type (Fig. 2c, d). During the Late Glacial, trees and shrubs became more widely distributed, and floristic diversity started to decline – a trend that was especially pronounced in North America (Fig. 2b, d). By about 12 ka, rising sea levels had flooded the Bering Strait, and the vegetation on each side started to diverge (Extended Data Fig. 2a). In northeast Siberia, greater effective moisture

88 | Nature | Vol 600 | 2 December 2021

within the Holocene led to the expansion of aquatic plants (such as *Hippuris* and *Menyanthes*). The previously dominant steppe taxa (for example, *Poa* and *Artemisia*) declined, although sedges, of which many species are hydrophilous, continued to be abundant (Extended Data Fig. 3). The vegetation of this region became a mosaic of steppe and tundra elements. In North America, trees such as *Populus* and *Picea* became more widespread during the Early Holocene and previously widespread steppe species declined (Fig. 2d and Extended Data Fig. 3). Abroad, southern swath of eastern Beringia became boreal forest. In contrast to the changes observed in these regions, vegetation in

and large (horse and mammoth) mammals, indicating that a wetter environment with a high proportion of hydrophilous plants (that is, moisture-loving plants) was a key factor restricting animal distributions. The distribution of mammoths tends to be positively affected by plant NMDS2, which mainly reflects the proportion of woody plants (particularly shrubs and subshrubs) as opposed to herbaceous plants, whereas the reverse is true for horses (Fig. 3). We also found that horses are more sensitive to vegetation composition compared with other herbivores (Supplementary Information 13.3). These findings support the hypothesis that horses were more restricted to a grassland environ-

nature portfolio

<https://doi.org/10.1038/s41586-022-05453-y>

Supplementary information

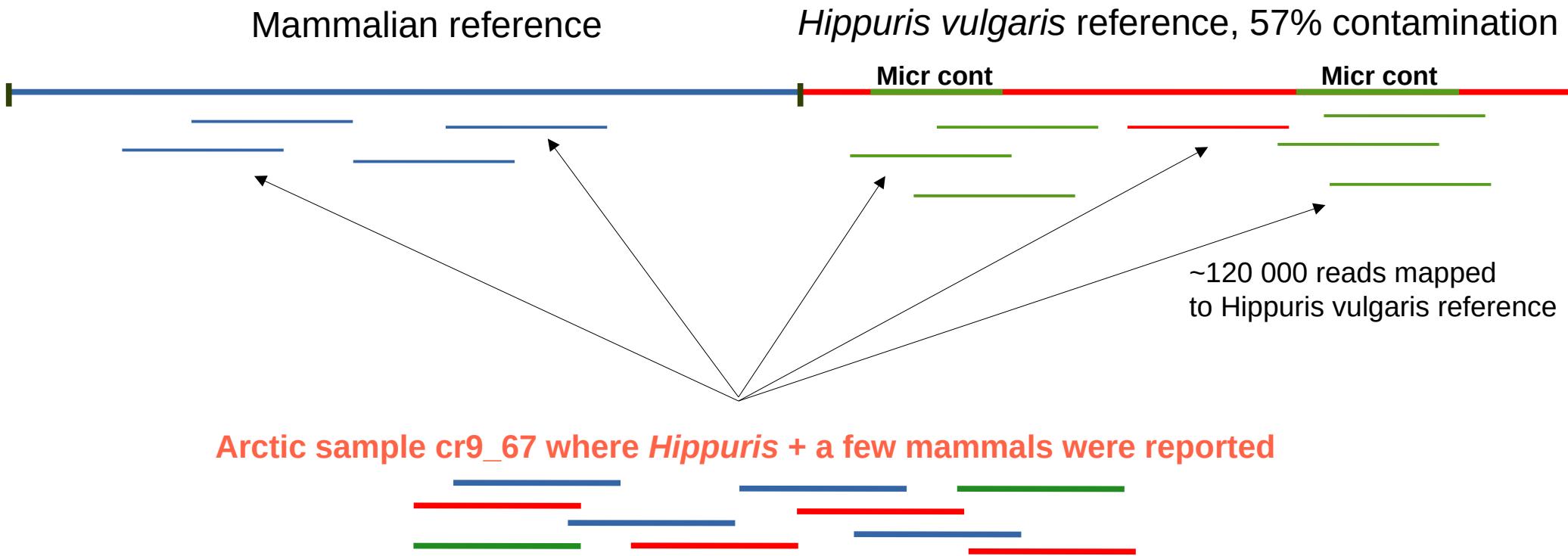
A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA

In the format provided by the authors and unedited

individual samples is consistent with an assemblage from a large low-gradient alluvial catchment, consistent with an estuary with multiple tributaries. All but the two least diverse samples contain both aquatic and terrestrial taxa. The most common terrestrial genera *Andromeda*, *Salix*, *Vaccinium*, *Carex*, *Dryas* and *Equisetum*, dominate the DNA assemblages in almost all samples and are well represented as macrofossils (Source Data S1 - sheet 2). Most of the aquatic taxa are restricted to standing water (e.g., *Hippuris*, *Stuckenia*, *Potamogeton*, *Menyanthes*) or occasionally streams (*Sparganium*, *Callitriches*). This range of genera combined with the presence of ferns and club mosses is consistent with alluvial deposition from drainage containing a range of microhabitats including well-drained slopes, stream valleys, lakes, coastal plains, and beaches. If the requirements and ecological amplitudes of these genera are generally conserved, the occurrence of *Erica*, *Sphagnum*, *Arctostaphylos* and *Kalmia* would suggest the presence of acidic histosols and snow cover; *Alnus* and perhaps *Populus* and *Salix* saturated alluvium / riparian habitats; *Artemisia* and *Astragalus* open, primarily mineral soils and *Dryas* cold, open and windswept areas.

A scenario in which eroded sediment from paleosols or permafrost are redeposited after mixing with younger entrained material could produce an assemblage comprising taxa from a succession of communities adapting over millennia to changing conditions. However, we assert that the biotic assemblages of DNA, macrofossils, and pollen from any discrete sample in the Kap København Formation are largely contemporaneous. Redeposited material is unlikely to contribute substantially to the assemblages for several reasons. The macrofossils were generally well preserved,

Kjaer et al.,
Nature 2021



116 000 reads out of 120 000 mapped reads,
i.e. 97% of reads, intersect with microbial-like regions in *Hippuris* reference