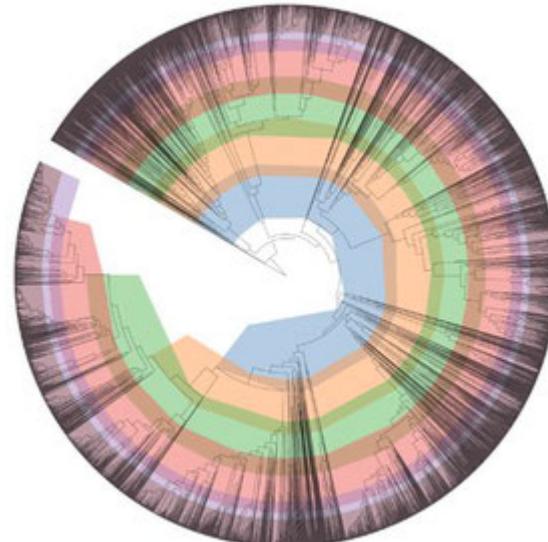


ENVIRONMENTAL METAGENOMICS

Physalia course, online, 11-15 November 2024

Introduction to metagenomics

Nikolay Oskolkov, Lund University, NBIS SciLifeLab
Luis Pedro Coelho, 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?

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



Microbiome

EDITORIAL

Open Access

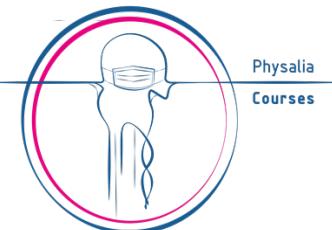


The vocabulary of microbiome research: a proposal

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?

"This collection is obtained through shotgun sequencing of DNA extracted from a sample (**metagenomics**) followed by mapping to a reference database or assembly, followed by annotation."

Marchesi & Ravel 2015, "The vocabulary of microbiome research"

JOURNAL OF BACTERIOLOGY, Feb. 1996, p. 591-599
0021-9193/96/504.00+0
Copyright © 1996, American Society for Microbiology

Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon

JEFFEREY L. STEIN,^{1*} TERENCE L. MARSH,² KE YING WU,³ HIROAKI SHIZUYA,⁴ AND EDWARD F. DELONG^{3*}

Vol. 178, No. 3

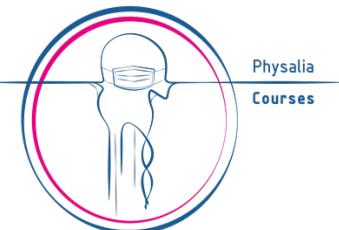
Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman¹, Michelle R Rondon¹, Sean F Brady², Jon Clardy² and Robert M Goodman¹



meta | genome
“beyond the genome”

“[cloning of environmental DNA into *E. coli* for phenotype screening] has been made possible by advances in molecular biology and Eukaryotic genomics, which have laid the groundwork for cloning and functional analysis of the collective genomes of soil microflora, which we term **the metagenome of the soil**”



Metagenomics is the ultimate way to study microbial communities

Community structure

Who is present and in what abundances?

Do communities sampled from different locations share similar composition and structure?

Microbial communities

Interaction and communication

Are organisms competing with one another?

Do they act as partners?

Which proteins and metabolites are signalling molecules?

Diversity and dynamics

How many types of organisms are present?

How are they distributed?

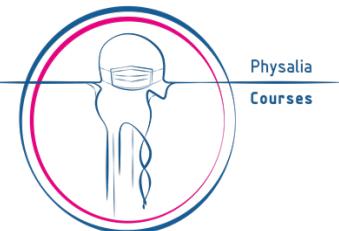
In what ways do they change over time?

Ecosystem function

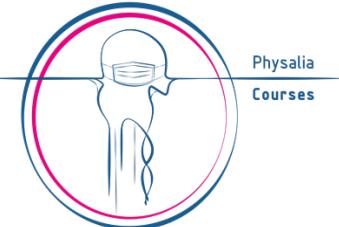
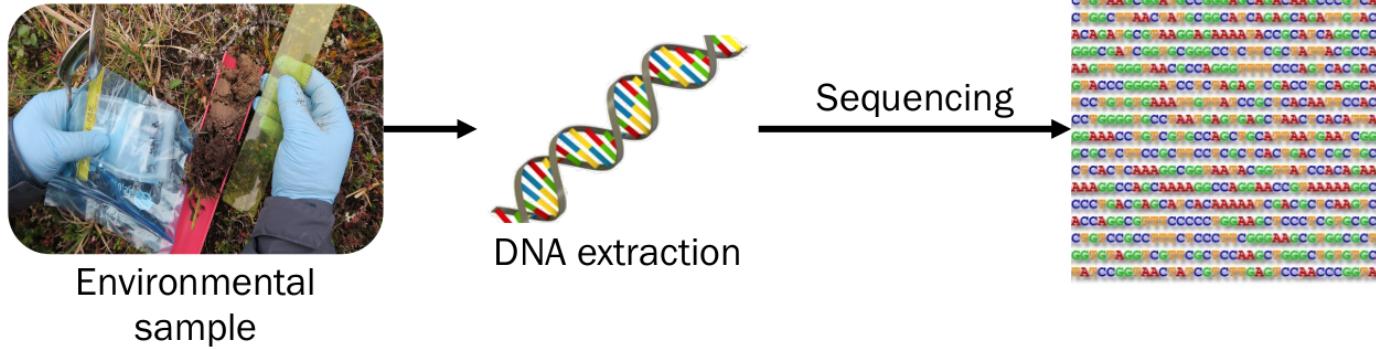
What types of functional genes are present?

Which of these genes are being expressed?

How does this change under different conditions or over time?

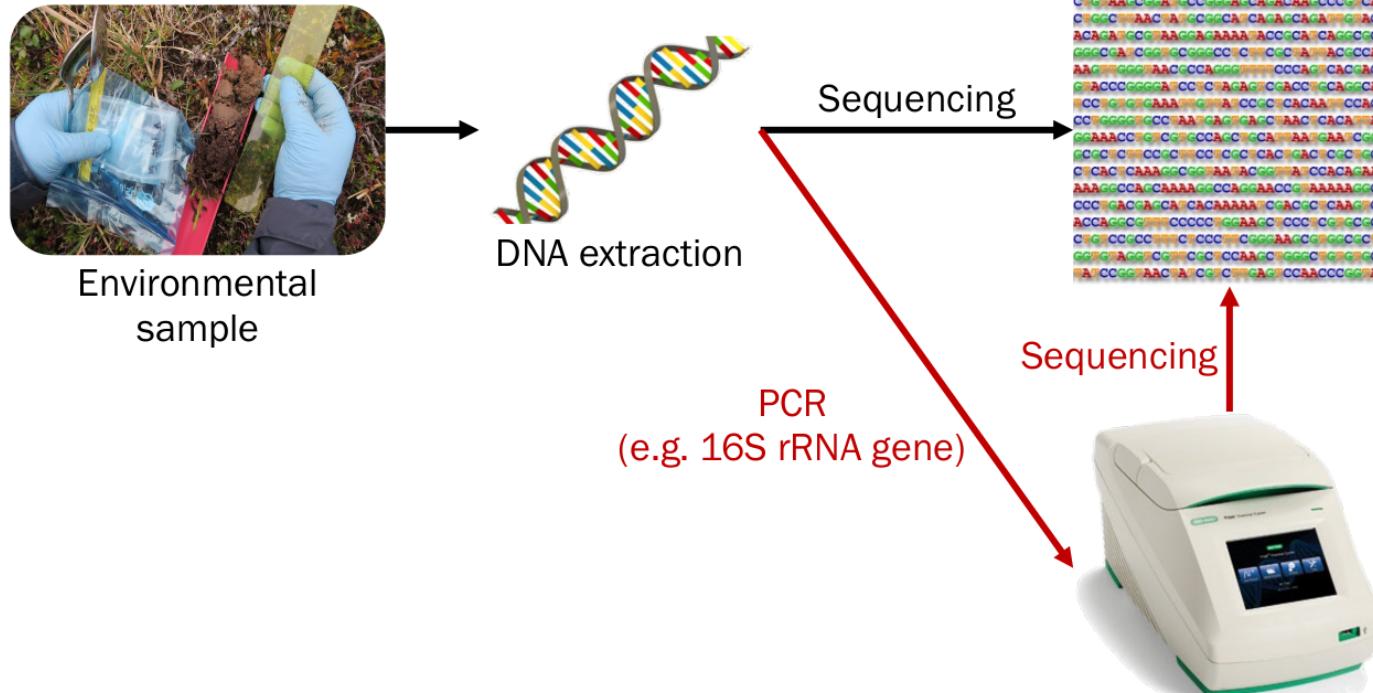


What is metagenomics?

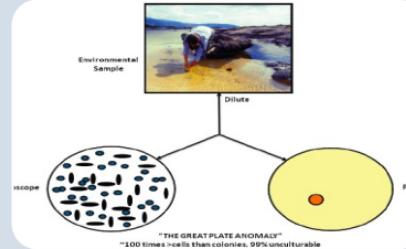
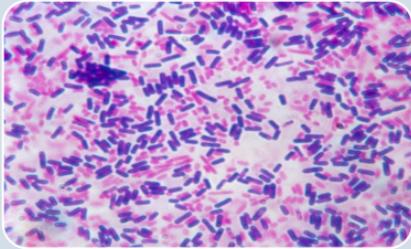
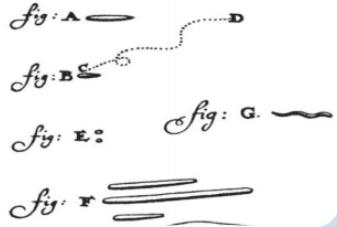


Physalia
Courses

What is NOT metagenomics?



Microbiology and technology go hand in hand



1670's

First observation
of microbes
under the
microscope

1880's

Development of
the Gram staining
method

First isolation of a
bacterium in solid
media

1980's

The Great Plate
Count Anomaly

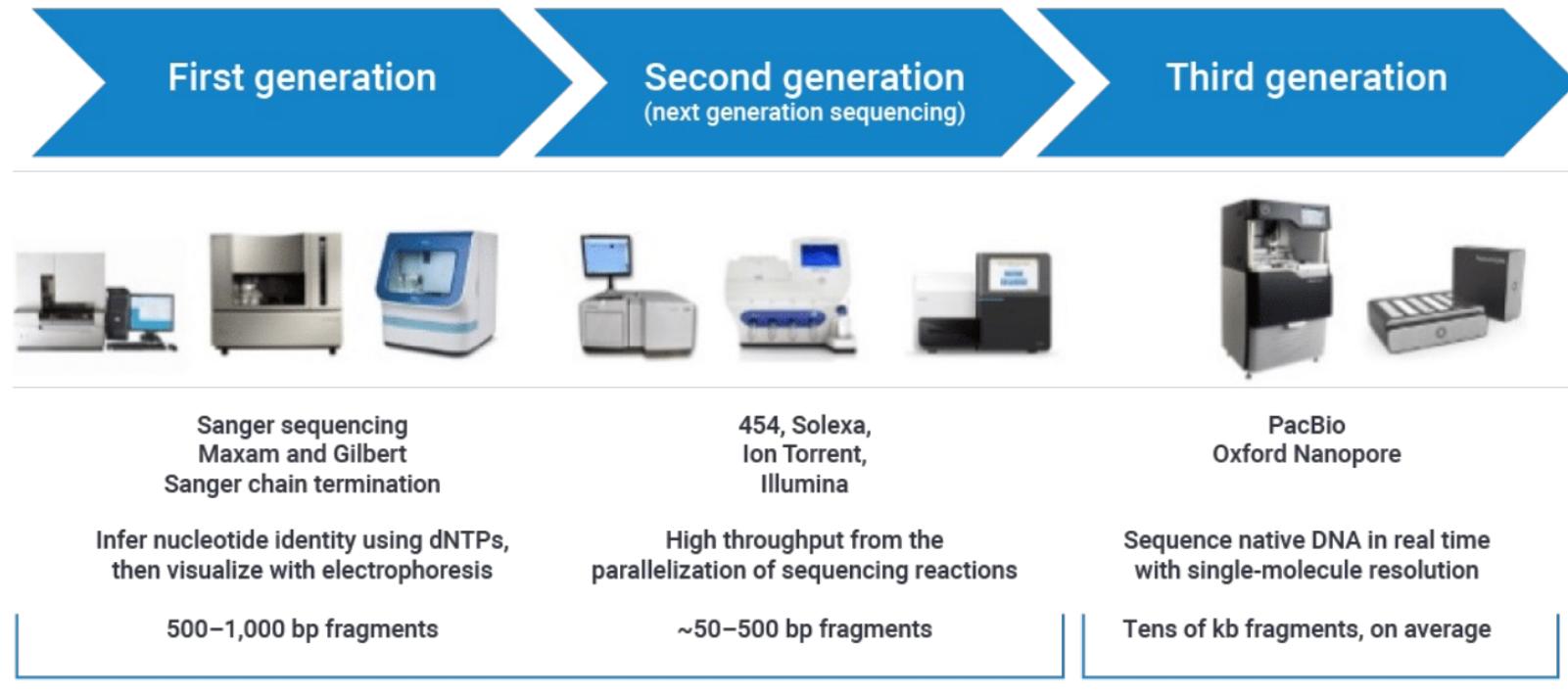
First culture-
independent
studies

2000's

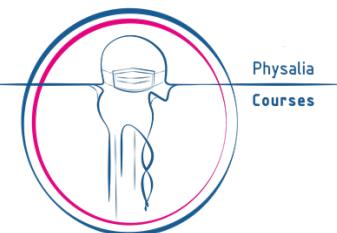
Advent of high-
throughput
sequencing



Metagenomics vs. sequencing technologies



www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

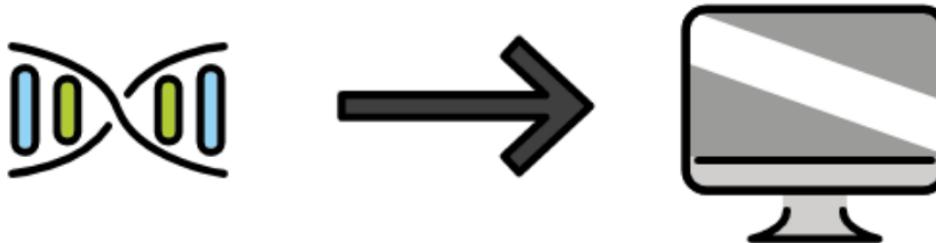


What is Sequencing?

Converting the chemical nucleotides of a DNA molecule

to

ACTG on your computer screen

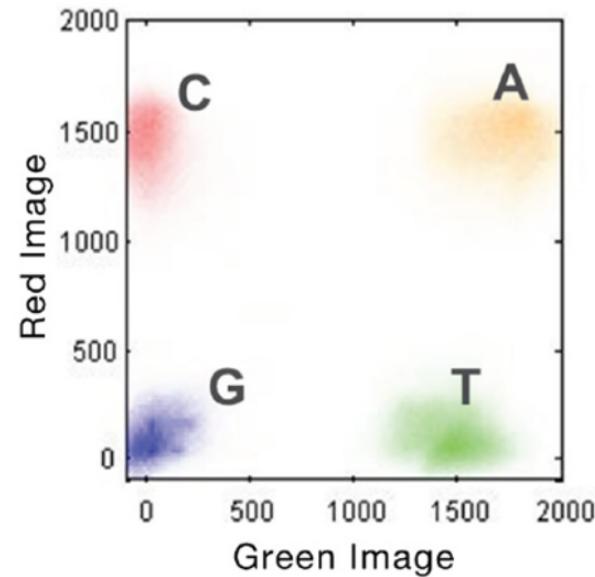


How does it work?

Replicate a strand, but add complementary fluorophore-modified nucleotide, one colour per base

T C A G

Fire mah lazer , and record the colour!
Rinse and repeat!



FASTQ File

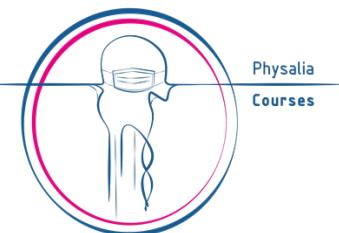
Example (files can be gigabytes in size!)

```
@K00233:37:HGHLYBBXX:3:1101:2646:1121 1:N:0:NACGCATC+NGCTGGTG
NCGCATGAGCCGCCTGTATCAGGCCTGATCGGCCGGCATTGCAGTTGGGATAGATGGGGGAGCACACGTCTG
+
#A7F<<GG<JFJFJJJJFFJJJJJAFFJFJJJJFJAFFFJAJFJJ<FJJJJFFF<FFA--FFFJJJJJ
@K00233:37:HGHLYBBXX:3:1101:4655:1121 1:N:0:NACGCATC+NGCTGGTG
NATGCATGACAGGAGGTGAGGGCATTTCAGATTTCAGGCTGCGACCTTGAGCATTTGCCGCTTCCAGCAC
+
#GG-<FFFF7JFF7JJJJFJJ<JJJJJA7FJJJJJJFF<JFF<J7-<FJJJJFJFFJJGGGGFFJJ--AJAJJ
```

@ <read id, e.g. machine ID, location on flowcell> <extra metadata>
<DNA sequence; Note: N = base couldn't be called!>
+ <a separator>
<base quality scores for each nucleotide in sequence>

Quality score:

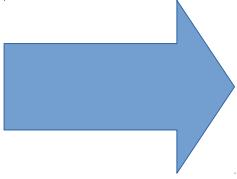
```
!"#$%&' ()*+, -./0123456789:;=>?@ABCDEFGHIJ
0.2.....26...31.....41
```



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



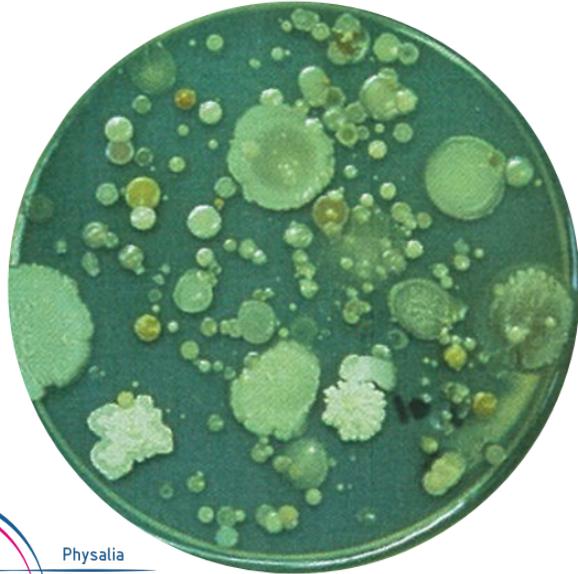
From single genome to metagenome



Who's there?

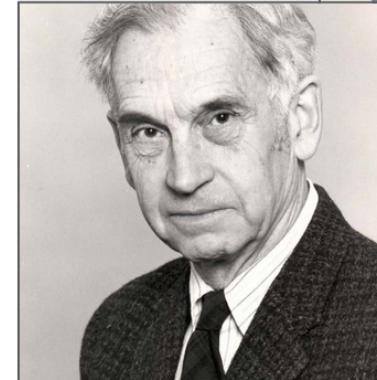
At a most basic level, the first question we usually ask in metagenomics is “Who’s there?”

What is a microbial species?



Physalia
Courses

Ernst Mayr
Biological Species
Concept, 1942



Ernst Mayr ~ Jared Diamond
Copyrighted Material

Who's there?

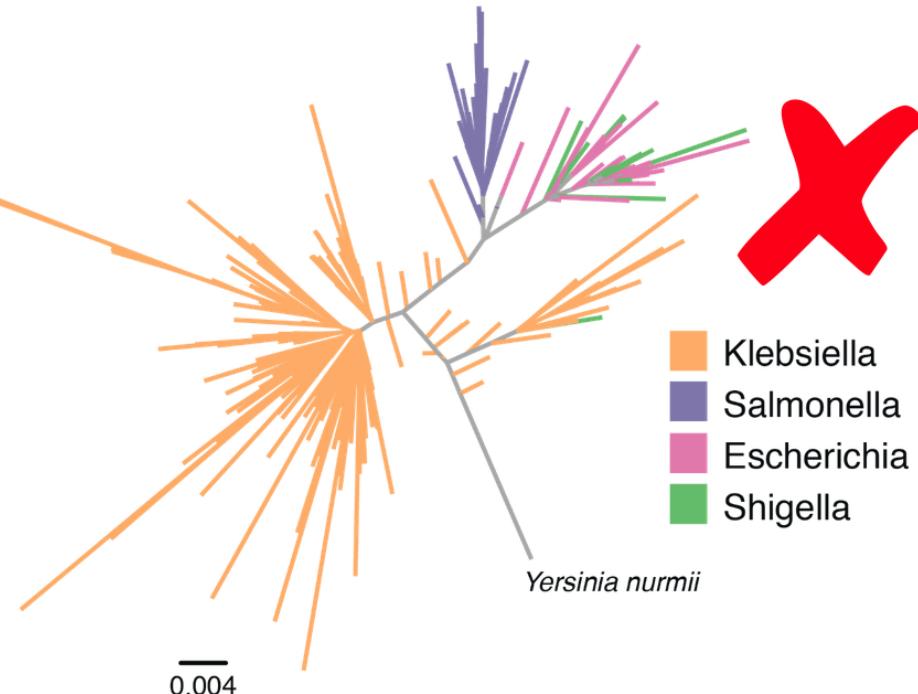
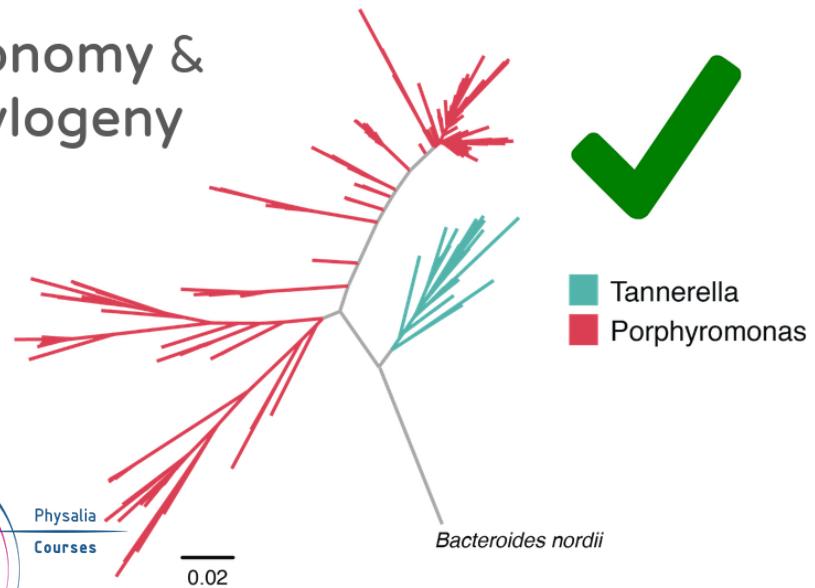
At a most basic level, the first question we usually ask in metagenomics is “Who’s there?”

What is a microbial species?

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

Phylogeny: evolutionary history of a set of taxa

Taxonomy & Phylogeny

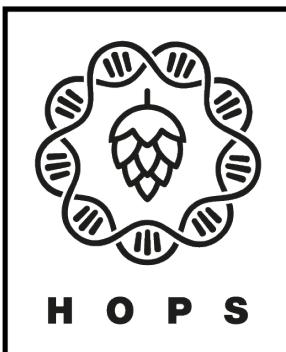


Typical analysis methods used in metagenomics

1) Alignment:



BWA
stands for
Burrows Wheeler Aligner
 Abbreviations.com



2) Classification:



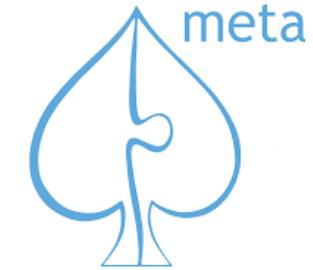
Centrifuge

MetaPhlan

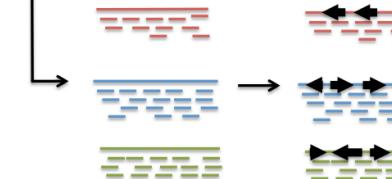
Clark

Reference based:
assume similarity to reference

3) De-novo assembly:



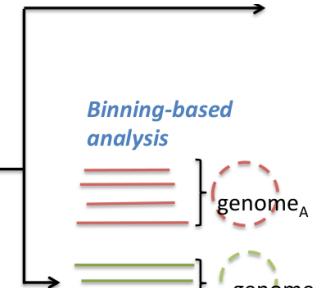
```
>seq1  
GCCGTAGTCC...  
>seq2  
...
```



Assembly

Assembly-based
analysis

gene prediction/
annotation



Binning-based
analysis

genome_A

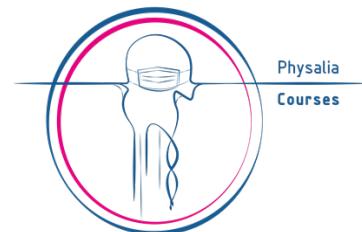
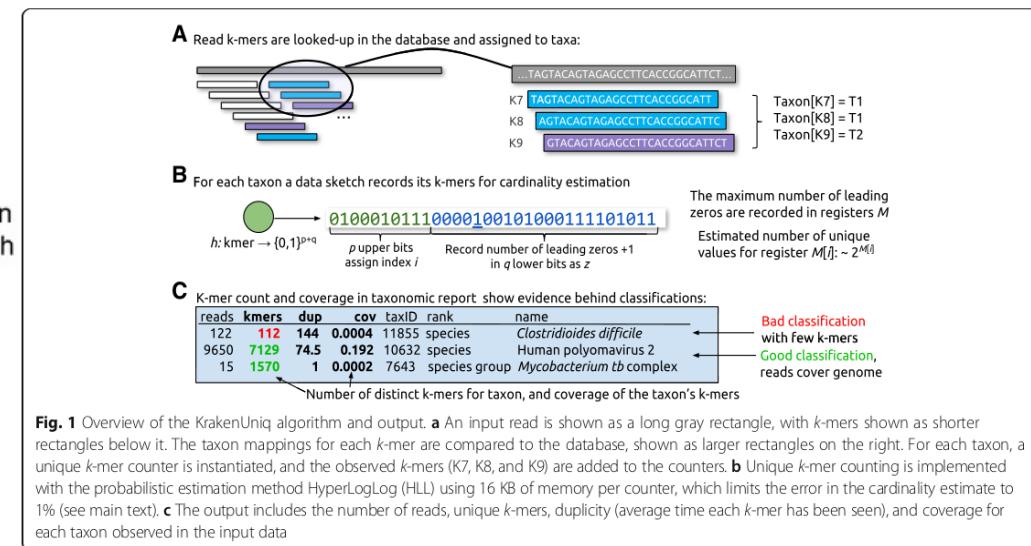
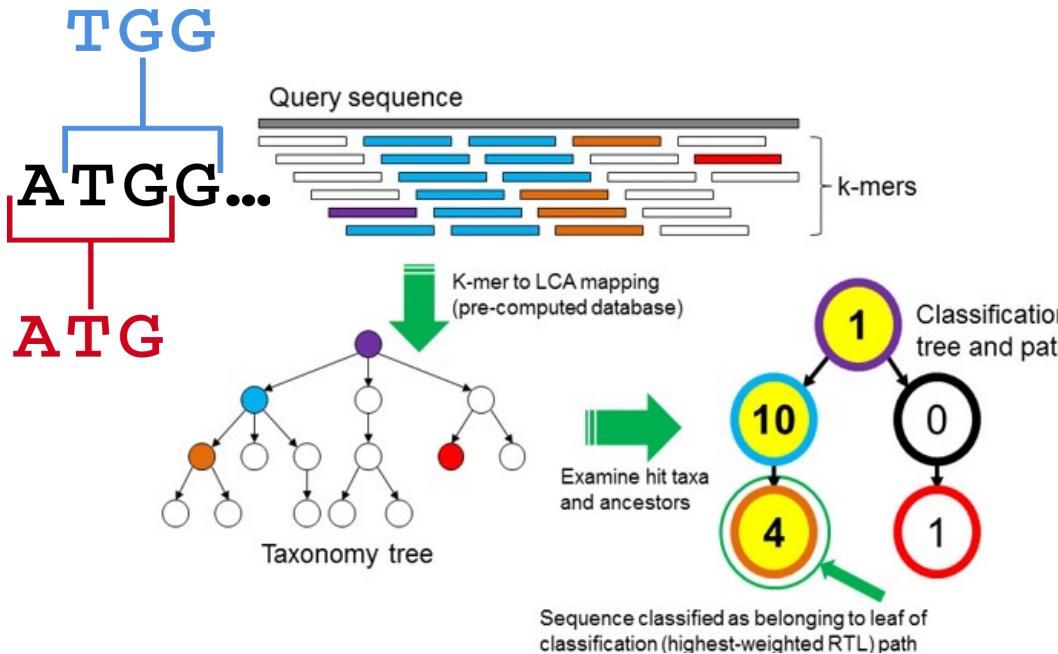
genome_B

genome_C

Phylogenetic binning

Reference free:
unbiased but challenging

K-mer based taxonomic profiling: Kraken family of tools



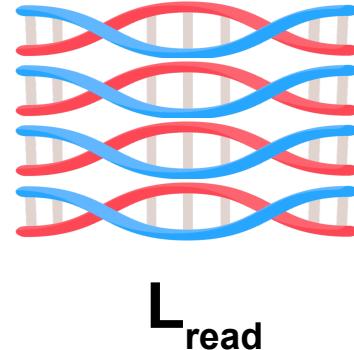
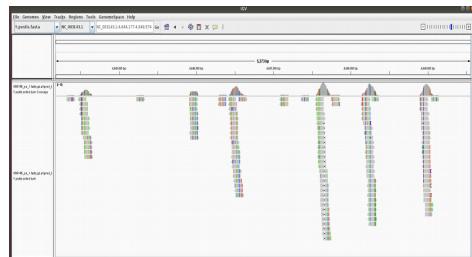
Advantage of classification over alignment: speed, Kraken2 is very fast!

Coverage vs. depth vs. breadth of coverage

Reference genome

A)

GCTACGATCTTAGCTTAGCTGGATCTGAATTCTCATCTCGGAT



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



Organism
NOT
detected

Reference genome

B)

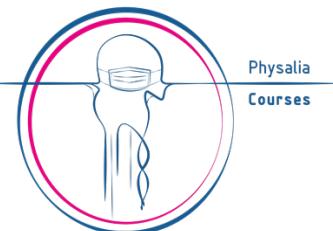
GCTACGATCTTAGCTTAGCTGGATCTGAATTCTCATCTCGGAT



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

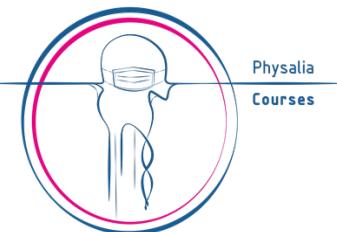
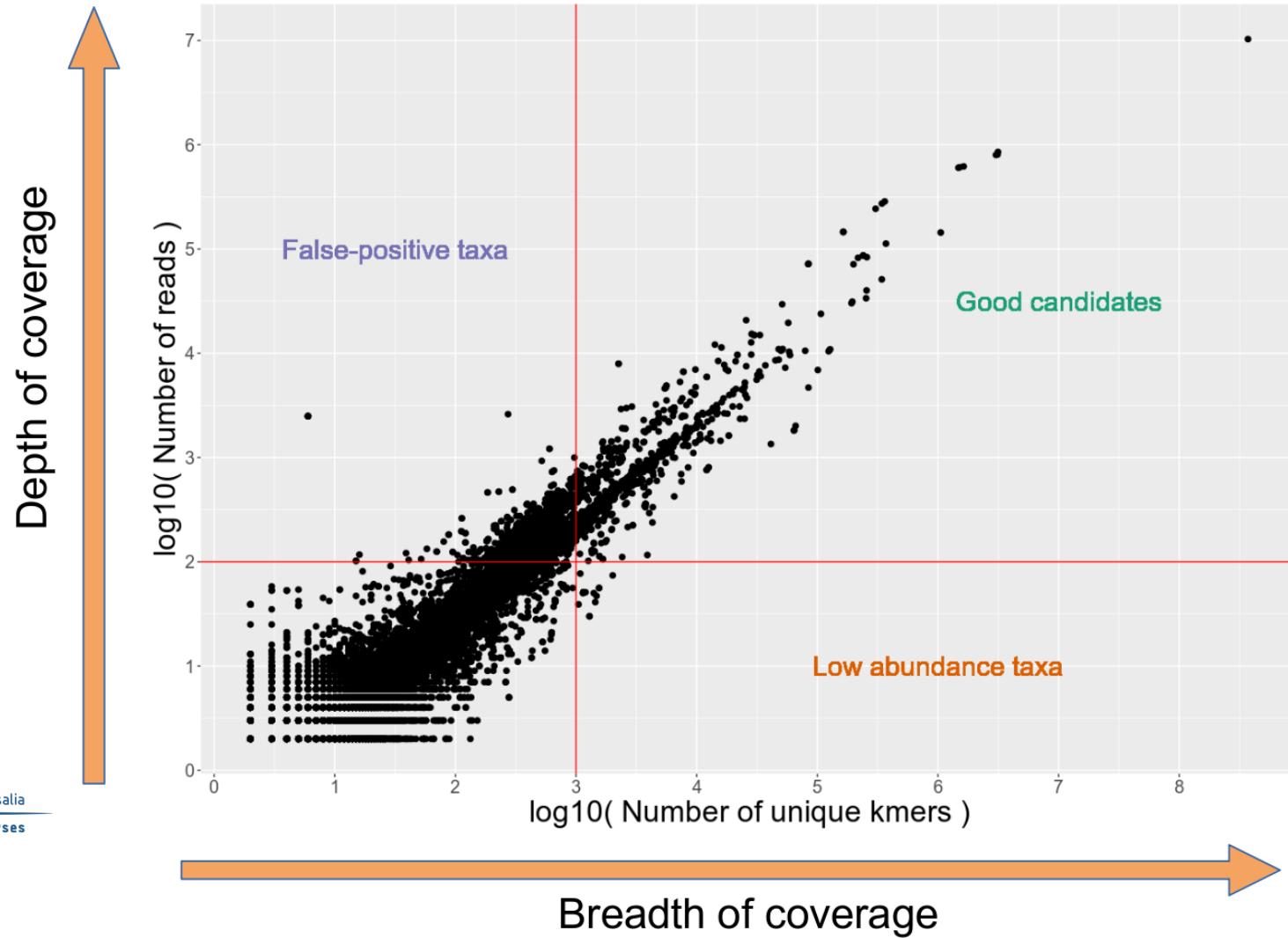


Organism
detected



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

Filtering Kraken output with respect to depth and breadth of coverage

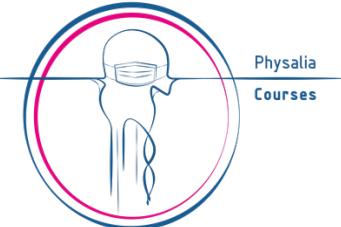


DATABASES!

DATABASES!

DATABASES!

DATABASES!



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 Altıntışık⁸, Małgorzata 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

¹¹Department 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,

README MIT license

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

snakemake ≥6.16.0 Tests passing

About

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

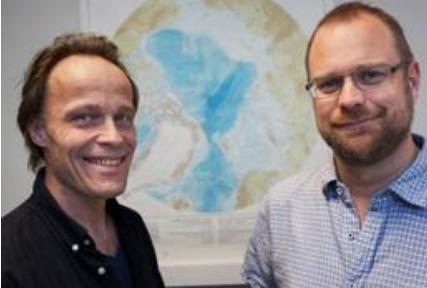
- trimming adapter sequences and removing reads shorter than 30 bp with Cutadapt
- quality control before and after trimming with FastQC and MultiQC
- taxonomic sequence kmer-based classification with KrakenUniq
- sequence alignment with Bowtie2 and screening for common microbial pathogens
- deamination 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>

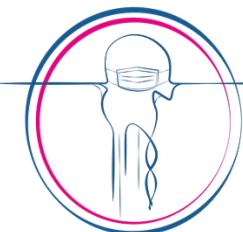
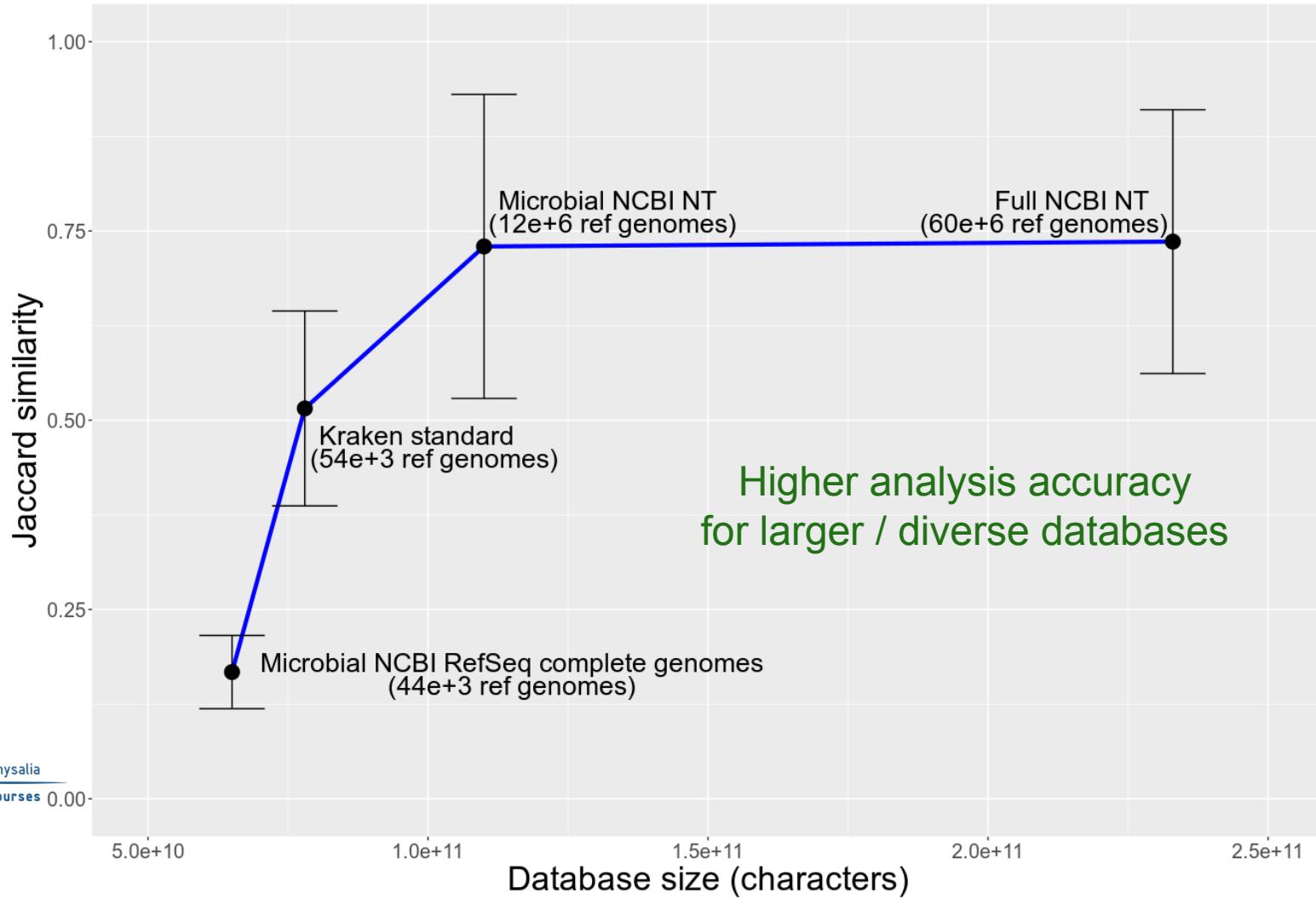
Authors

- 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>



Why large databases are important for accurate taxonomic profiling



Physalia
Courses

Host removal is a good precaution in microbiome research

nature

Explore content ▾ About the Journal ▾ Publish with us ▾

[View all journals](#)

Search

Log in

[Sign up for alerts](#)  [RSS feed](#)

[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, Qiyun 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.



Strengths and weaknesses of read-based metagenomics

Comprehensiveness	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference dbs
Community complexity	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Cannot resolve organisms for which genomes of close relatives are unknown
Computational burden	Can be performed efficiently, enabling large meta-analyses
Genome-resolved metabolism	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes
Expert manual supervision	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision
Integration with microbial genomics	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates

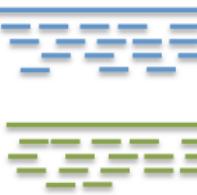
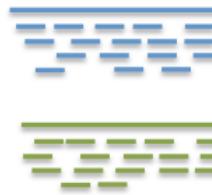
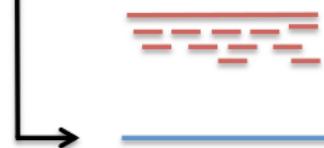


*Read-based
analysis*

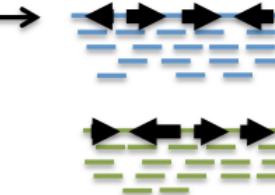
Summary metagenome analysis strategies

Functional/taxonomic annotation
directly on reads (often partial genes)

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



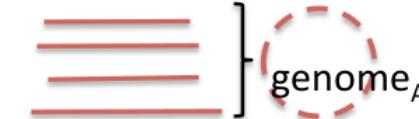
Assembly



gene prediction/
annotation

*Assembly-based
analysis*

*Binning-based
analysis*



Phylogenetic binning



Physalia
Courses

Strengths and weaknesses of assembly-based metagenomics

Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives
Computational burden	Requires computationally costly assembly, mapping and binning
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates



Popular de-novo assemblers for short Illumina reads

OXFORD ACADEMIC Journals Books

Sign in through your institution

Bioinformatics

Issues Advance articles Submit ▾ Alerts About ▾

Bioinformatics Search Advanced Search

Bioinformatics Volume 31, Issue 10 May 2015

JOURNAL ARTICLE

MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph

Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam 
Author Notes

Bioinformatics, Volume 31, Issue 10, May 2015, Pages 1674–1676, <https://doi.org/10.1093/bioinformatics/btv033>

Published: 20 January 2015 Article history ▾

 PDF  Split View  Cite  Permissions  Share ▾

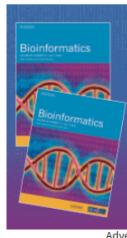
Abstract

Summary: MEGAHIT is a NGS *de novo* assembler for assembling large and complex metagenomics data in a time- and cost-efficient manner. It finished assembling a soil metagenomics dataset with 252 Gbps in 44.1 and 99.6 h on a single computing node with and without a graphics processing unit, respectively. MEGAHIT assembles the data as a whole, i.e. no pre-processing like partitioning and normalization was needed. When compared with previous methods on assembling the soil data, MEGAHIT generated a three-time larger assembly, with longer contig N₅₀ and average contig length; furthermore, 55.8% of the reads were aligned to the assembly, giving a fourfold improvement.

Availability and implementation: The source code of MEGAHIT is freely available at <https://github.com/voutcn/megahit> under GPLv3 license.

< Previous Next >

Physalia Courses



Bioinformatics is now fully open access 

Learn more 

Advertisement

CITATIONS 5.7k  More metrics information

VIEWS 79,539 

ALTMETRIC 129 

Email alerts

Article activity alert
Advance article alerts
New issue alert
In progress issue alert

CSH PRESS GENOME RESEARCH HOME ABOUT ARCHIVE SUBMIT SUBSCRIBE ADVERTISE AUTHOR INFO CONTACT HELP

You'd speed up discovery of drug targets and biomarkers. 

Search for Keyword: Go Advanced Search

« Previous | Next Article » Table of Contents

This Article

Published in Advance March 15, 2017 doi: 10.1101/213959.116  Genome Res. 2017. 27: 824-834 © 2017 Nurk et al.; Published by Cold Spring Harbor Laboratory Press

Abstract  Full Text  Full Text (PDF)  Supplemental Material

All Versions of This Article:  gr.213959.116v1  gr.213959.116v2 27/8/24 most recent

Article Category Method Services Citling Articles Google Scholar PubMed/NCBI ORCID Share Metrics

Total Downloads Abstract: 45,803 Full-text: 8,786 PDF: 16,378 See more details

10X GENOMICS Superior performance for cost-effective small-scale studies

See more details Picked up by 3 news outlets Blogged by 4

Current Issue October 2024, 34 (10) 

From the Cover Alert me to new issues of Genome Research

Advance Online Articles Submit a Manuscript Editorial Board E-mail Alerts & RSS Feeds Recommend to Your Library Job Opportunities

We are going to use Megahit in the exercises