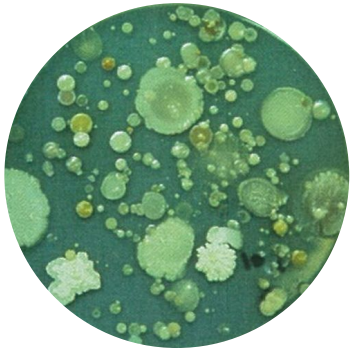
SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

# Lecture 3A: Introduction to Metagenomics

Christina Warinner

# What is a metagenome?



A **metagenome** is the collection of genomes and genes from the members of a microbiota.

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**.

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.

– Marchesi & Ravel 2015, “The vocabulary of microbiome research”


pre-2015
Terminology Wild West

# Why did we need an article about vocabulary?

Because terminology about microbes is a mess!

**Metagenome** was originally coined by Jo Handelsman et al. (1988) and meant something different

**Metagenomics** was occasionally used to refer to 16S rRNA amplification, something we now call **metataxonomics**

**Microbiome** is claimed to have been coined at least twice, each meaning something different - either a **"microbial biome"**, meaning a microbial community (1988); or a **"microbiota -ome"**, meaning a the collective genomes of a microbiota (2001).

…but the term microbiome has actually been in use since at least 1894!

# What is a metagenome?

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

Microbiome

EDITORIAL Open Access

CrossMark

## The vocabulary of microbiome research: a proposal

Julian R. Marchesi[1,2] and Jacques Ravel[3,4*]

# What is ancient metagenomics?

**Ancient metagenomics** is the study of the collection of genes and genomes of the microbiota(s) within a given environment or microbiome, plus all the other DNA mixed in

Basically, **all the DNA in a sample**

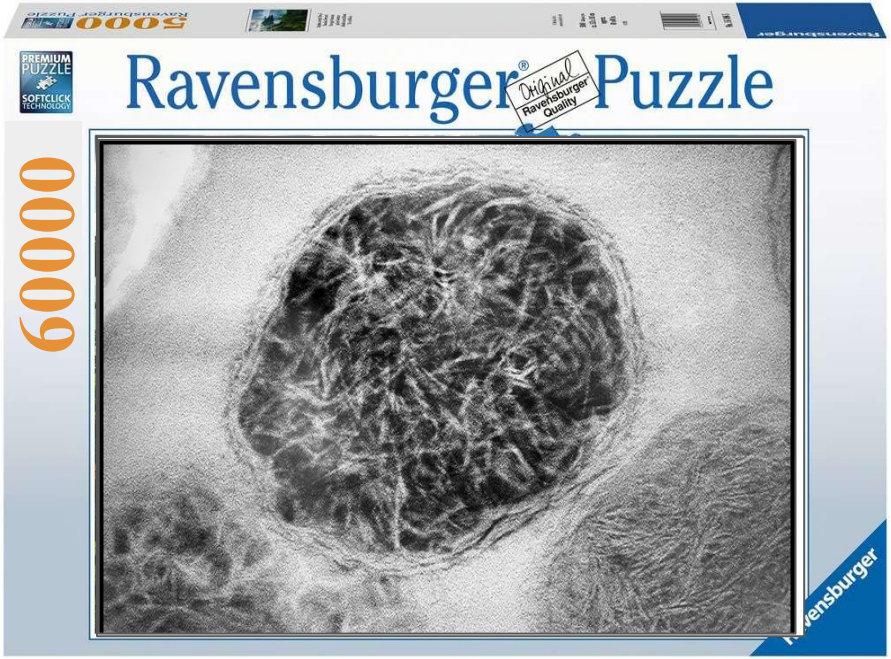
*Key point:* in addition to the **antemortem** genetic material of any microbes present during life, ancient metagenomes almost always contain at least some **postmortem** DNA from the **necrobiome**

Ancient metagenomics is **like regular metagenomics, but harder** because other environmental microbiota of various ages are mixed in and because the DNA is ancient and degraded

# What is ancient metagenomics?

a metaphor…

# Worst puzzle ever metaphor of ancient DNA



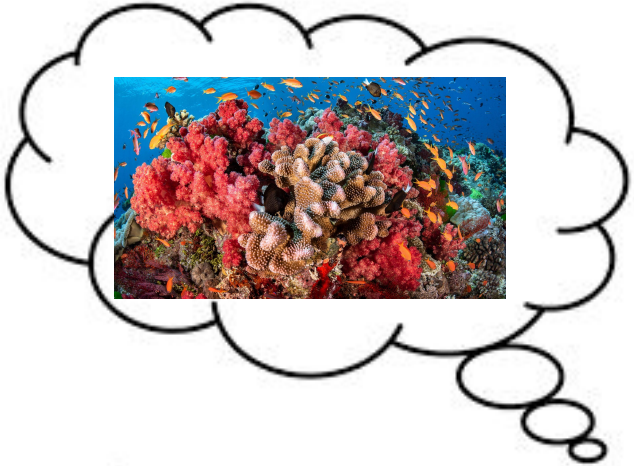

Yersinia pestis
Human genome
Microbiome

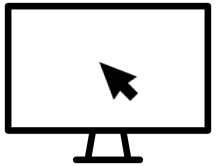
# What is ancient metagenomics?

another metaphor…


everything in
the past
time
aDNA

# Parrotfish metaphor of ancient metagenomics







We Can Do It!
POST FEB. 15 TO FEB. 28
WAR PRODUCTION CO-ORDINATING COMMITTEE

# Starting questions


Who's there?

How preserved is my sample?

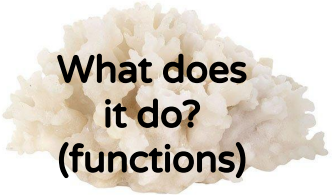
How do I clean up my dataset?

Let's explore its genomes! (MAGs)

Thursday

I'm looking for something very particular (pathogens)

Thursday



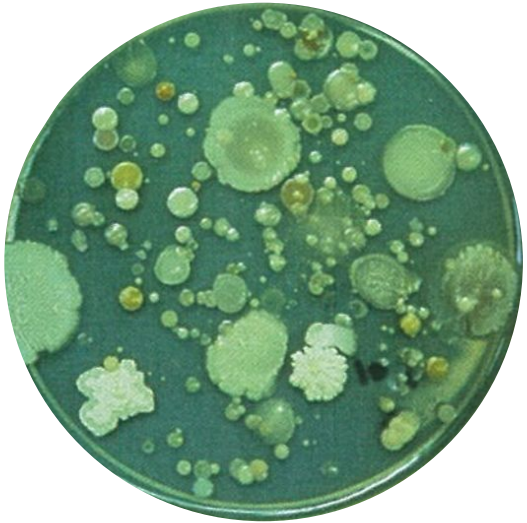
What does it do? (functions)

Friday

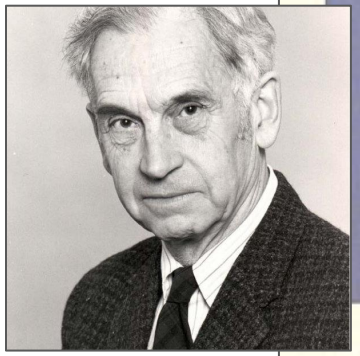
# Who's there?

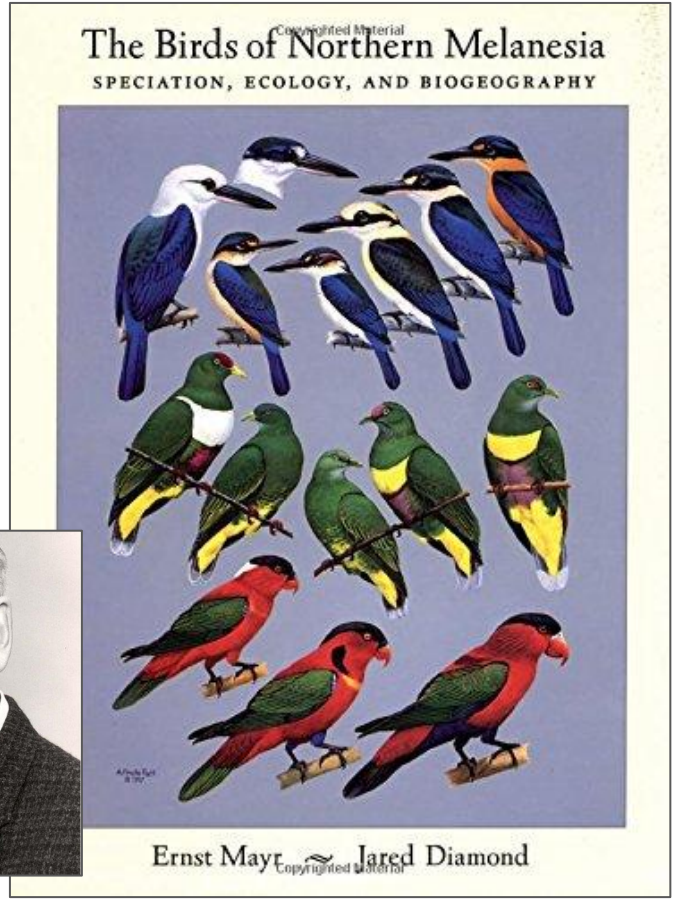
At a most basic level, the first question we usually ask in metagenomics is "**Who's there?**"

What is a microbial species?



**Ernst Mayr**
Biological Species Concept



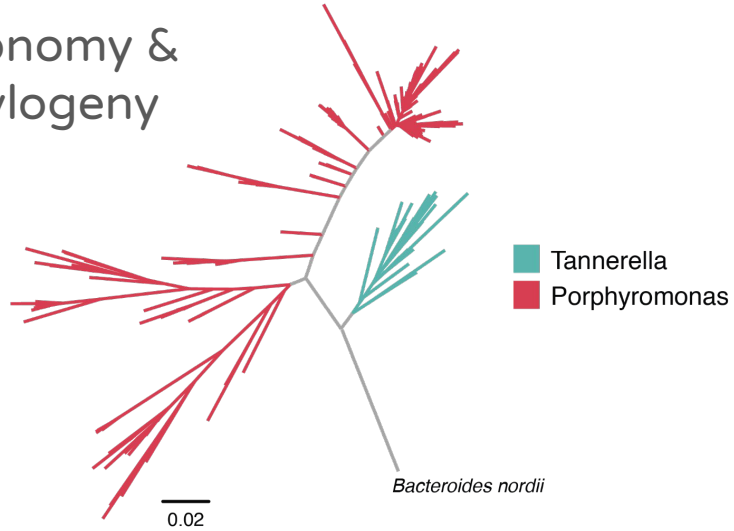


# Who's there?

At a most basic level, the first question we usually ask in metagenomics is "**Who's there?**"

What is a microbial species?

Taxonomy & Phylogeny

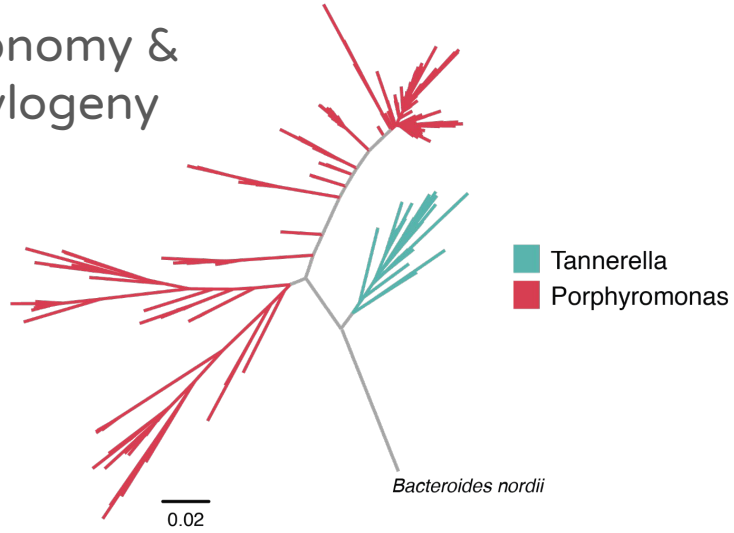


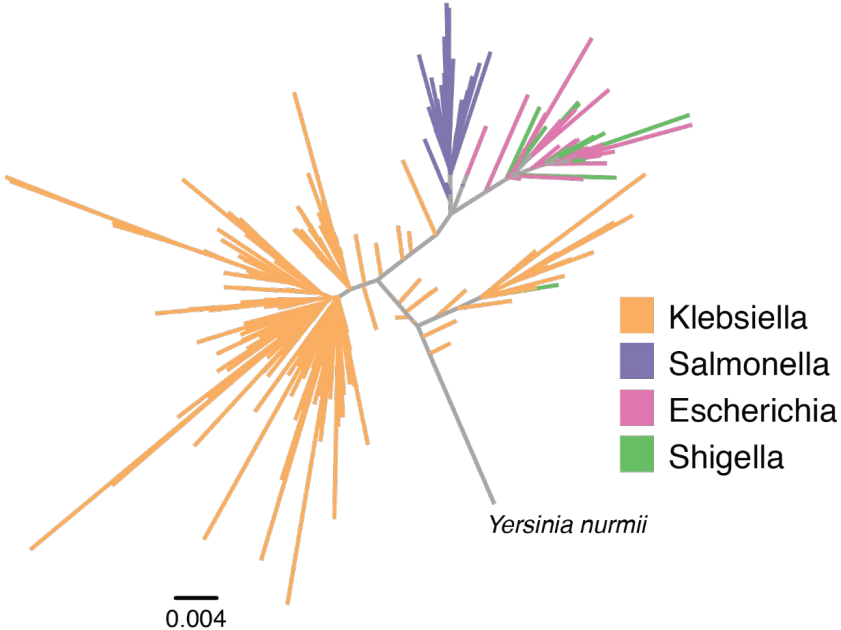
# Who's there?

At a most basic level, the first question we usually ask in metagenomics is "**Who's there?**"

What is a microbial species?

Taxonomy & Phylogeny



# 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?

**species**

# 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?

**species**

| Domain | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|
| **d__Bacteria;** | **p__Bacteroidota;** | **c__Bacteroidia;** | **o__Flavobacteriales;** | **f__Flavobacteriaceae;** | **g__Capnocytophaga;** | **s__Capnocytophaga gingivalis** |

# Who's there?

But how do you go from raw DNA sequences to taxon tables?

We use a **taxonomic profiler**

Several available options:

-> Alignment-based

- QIIME: 16S rRNA marker gene
- MetaPhlAn: marker gene set
- MALT: read alignment and binning

-> Alignment-free

- Kraken: K-mer matching

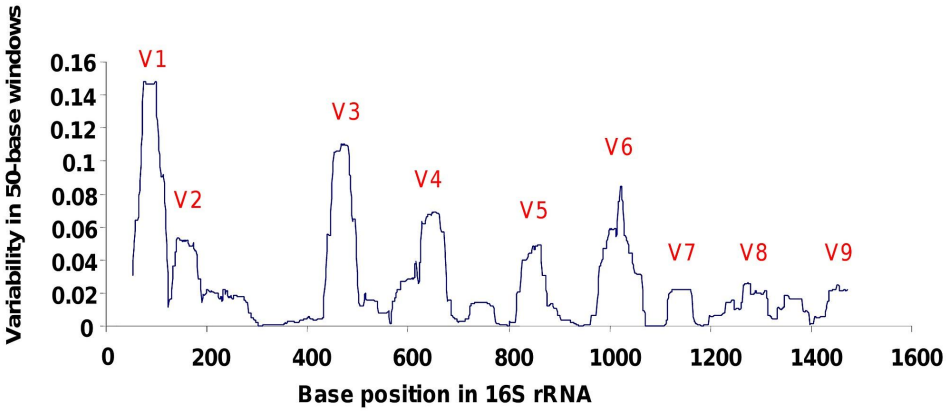
# The Classic

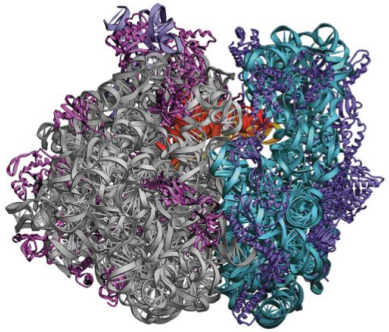
16S rRNA amplicon metataxonomics

# 16S rRNA marker gene

## **Amplicon metataxonomics** of the **16S rRNA** gene

- 16S rRNA gene is ubiquitous among prokaryotes
- Gene is ~1600 bp
- Contains conserved and hypervariable regions



Prokaryotic ribosome (70S)

Small Subunit (30S)

- 16S rRNA (~1540 nt)
- 21 proteins

Large Subunit (50S)

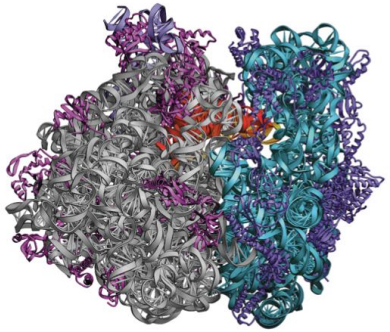
- 5S rRNA (~120 nt)
- 23S rRNA (~2900 nt)
- 31 proteins
- tRNA

# 16S rRNA marker gene

**Amplicon metataxonomics** of the **16S rRNA** marker gene

- **PCR** is used to amplify parts of the 16S rRNA gene
- Gene sequence used as a taxonomic “barcode”
- Also called **metabarcoding**

- Profilers: mothur, RDP classifier, QIIME
- HUGE databases, e.g., SILVA
- Efficient and inexpensive - widely used for modern DNA



Prokaryotic ribosome (70S)

Small Subunit (30S)

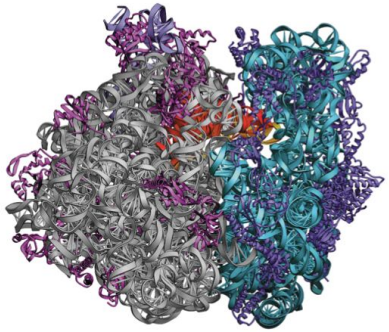
- 16S rRNA (~1540 nt)
- 21 proteins

Large Subunit (50S)

- 5S rRNA (~120 nt)
- 23S rRNA (~2900 nt)
- 31 proteins
- tRNA

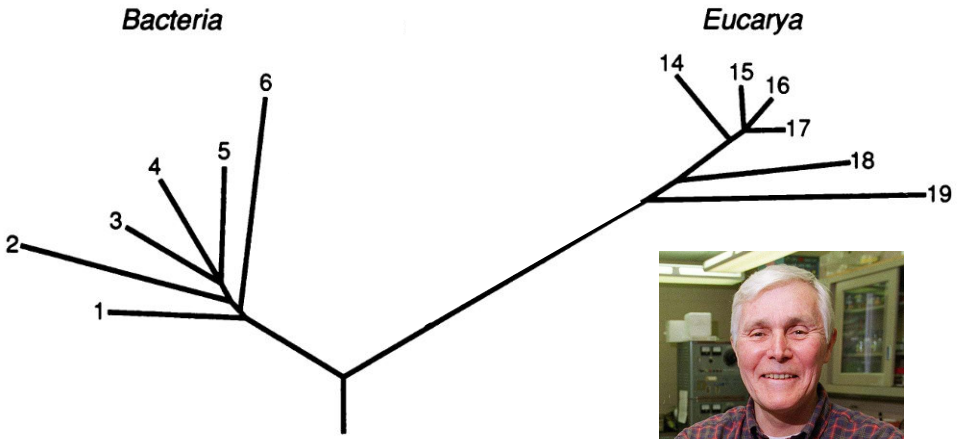
# 16S rRNA marker gene

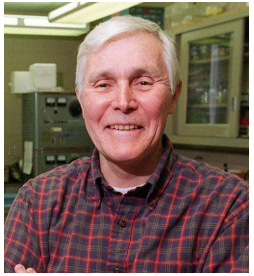
16S rRNA sequences were what led Carl Woese to the 1990 discovery that Archaea are a new domain of life!



Evolution: Woese *et al.*

*Proc. Natl. Acad. Sci. USA 87 (1990)*



Prokaryotic ribosome (70S)

Small Subunit (30S)

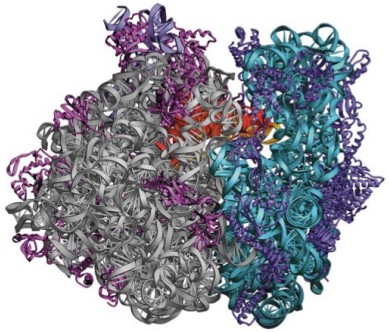
- 16S rRNA (~1540 nt)
- 21 proteins

Large Subunit (50S)

- 5S rRNA (~120 nt)
- 23S rRNA (~2900 nt)
- 31 proteins
- tRNA

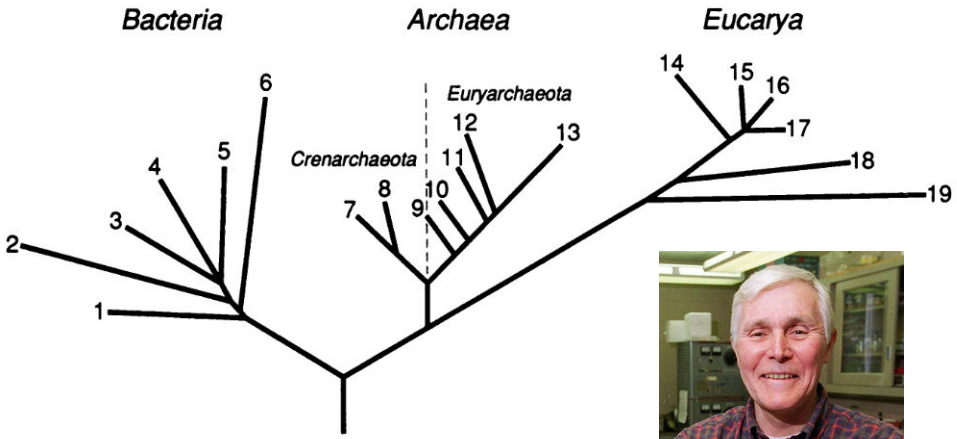
# 16S rRNA marker gene

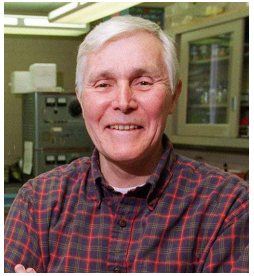
16S rRNA sequences were what led Carl Woese to the 1990 discovery that Archaea are a new domain of life!



Evolution: Woese *et al.*

*Proc. Natl. Acad. Sci. USA 87 (1990)*



Prokaryotic ribosome (70S)

Small Subunit (30S)

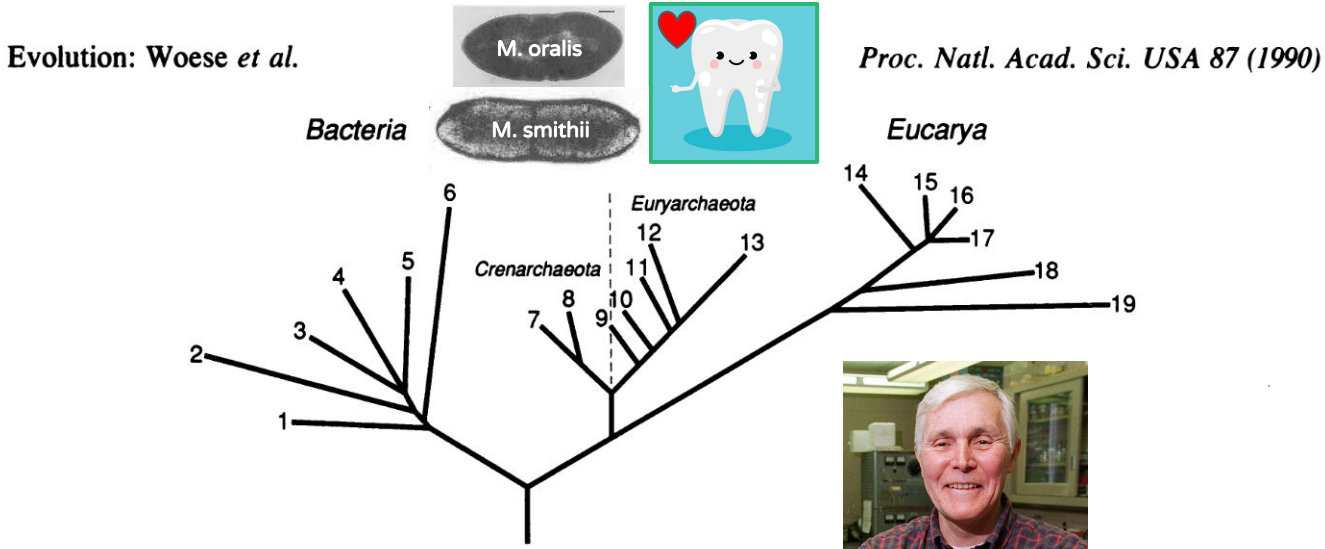
- 16S rRNA (~1540 nt)
- 21 proteins

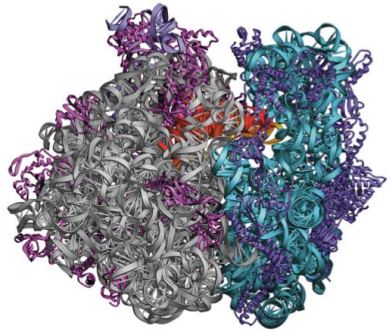
Large Subunit (50S)

- 5S rRNA (~120 nt)
- 23S rRNA (~2900 nt)
- 31 proteins
- tRNA

# 16S rRNA marker gene

16S rRNA sequences were what led Carl Woese to the 1990 discovery that Archaea are a new domain of life!



Prokaryotic ribosome (70S)

Small Subunit (30S)

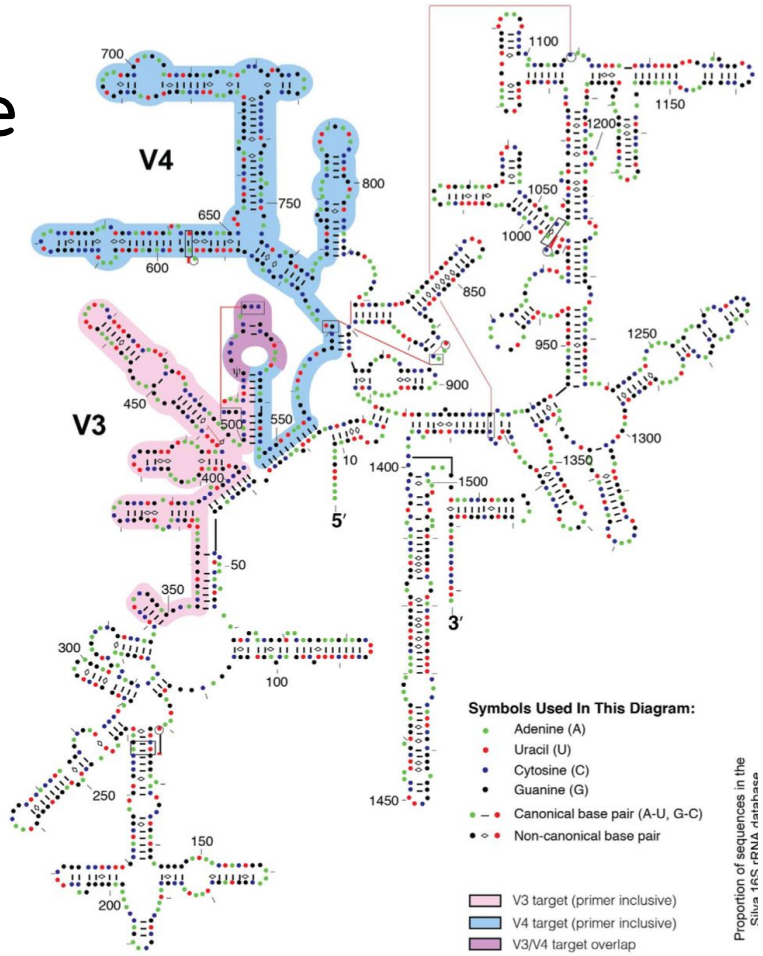
- 16S rRNA (~1540 nt)
- 21 proteins

Large Subunit (50S)

- 5S rRNA (~120 nt)
- 23S rRNA (~2900 nt)
- 31 proteins
- tRNA

# 16S rRNA marker gene

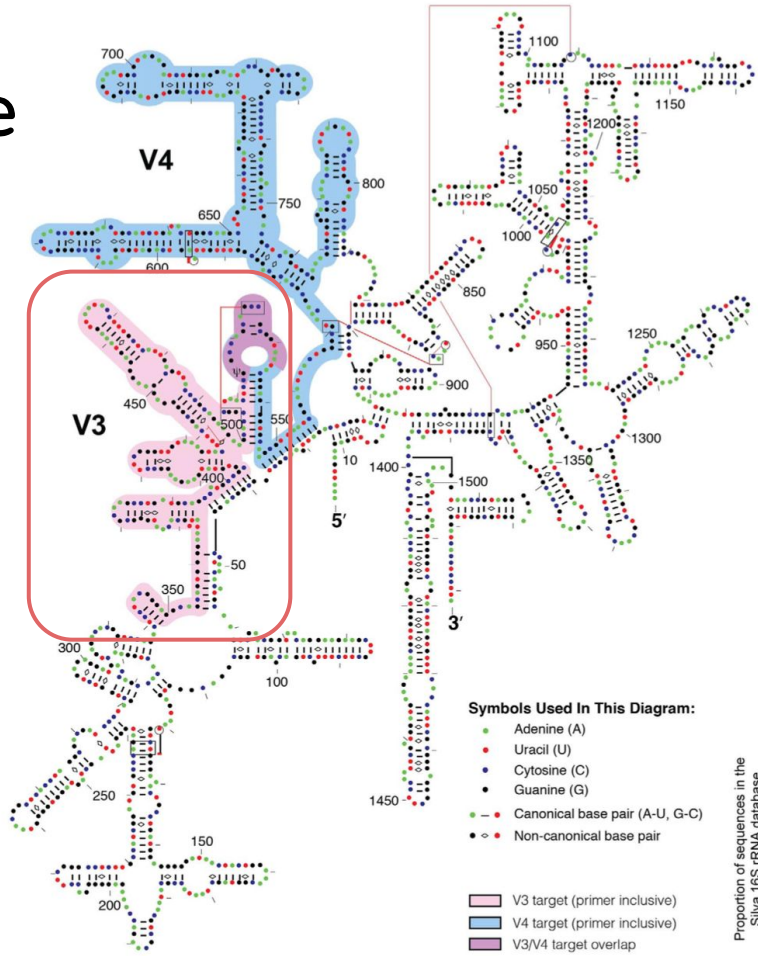
But… there are problems for aDNA



# 16S rRNA marker gene

But… there are problems for aDNA

V3 region is the shortest variable region with good taxonomic discrimination, but it is:

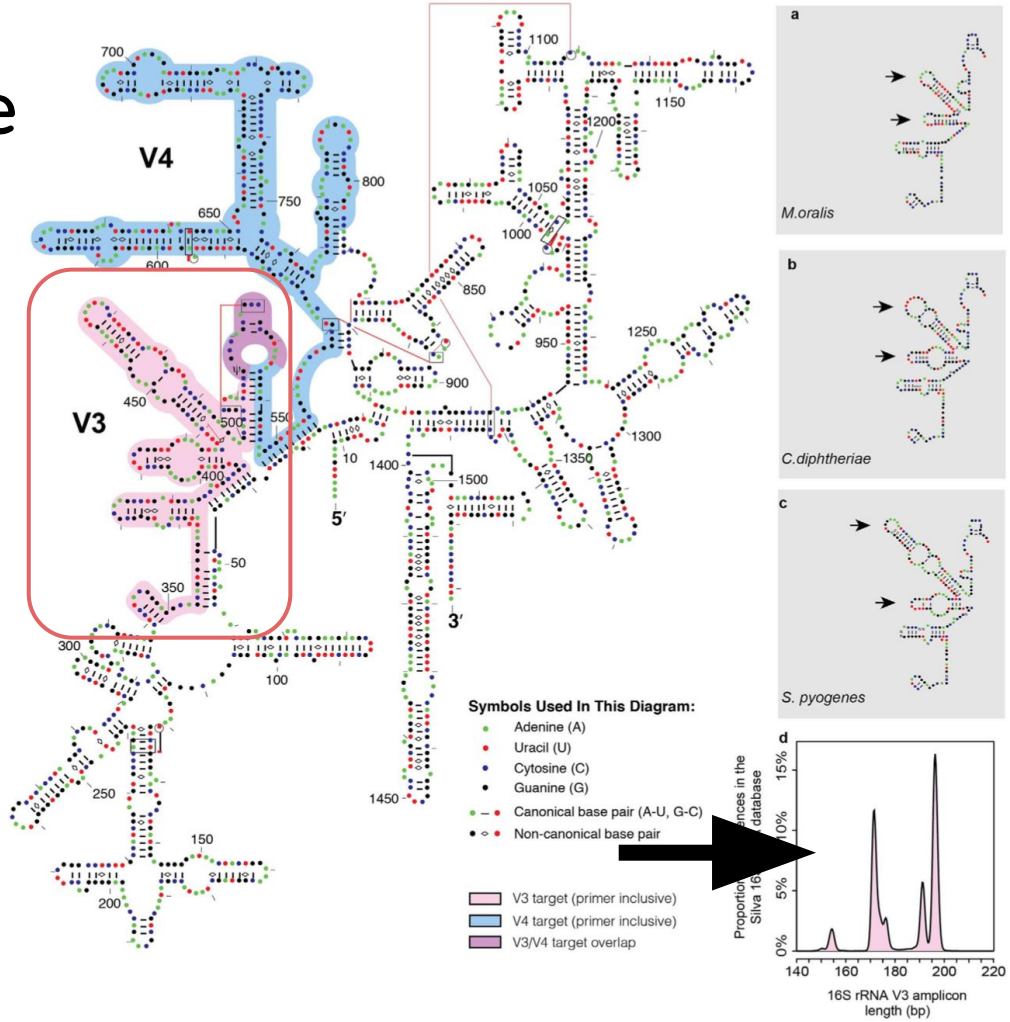


# 16S rRNA marker gene

But… there are problems for aDNA

V3 region is the shortest variable region with good taxonomic discrimination, but it is:

- Length polymorphic

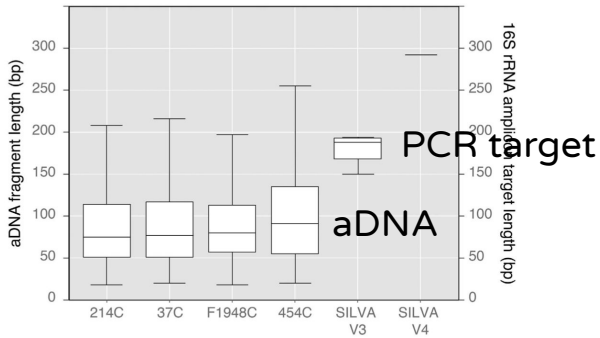


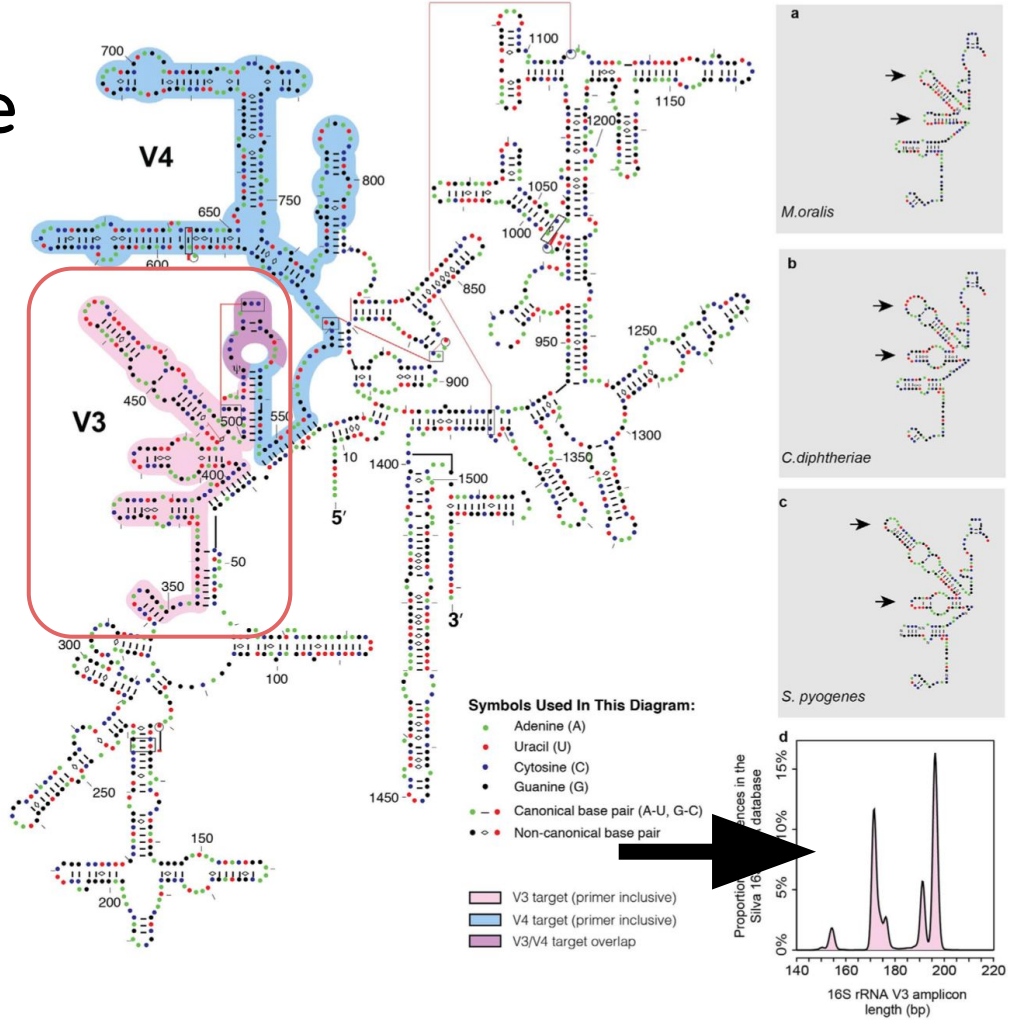
# 16S rRNA marker gene

But… there are problems for aDNA

V3 region is the shortest variable region with good taxonomic discrimination, but it is

- Length polymorphic
- ~180 bp long (too long!)



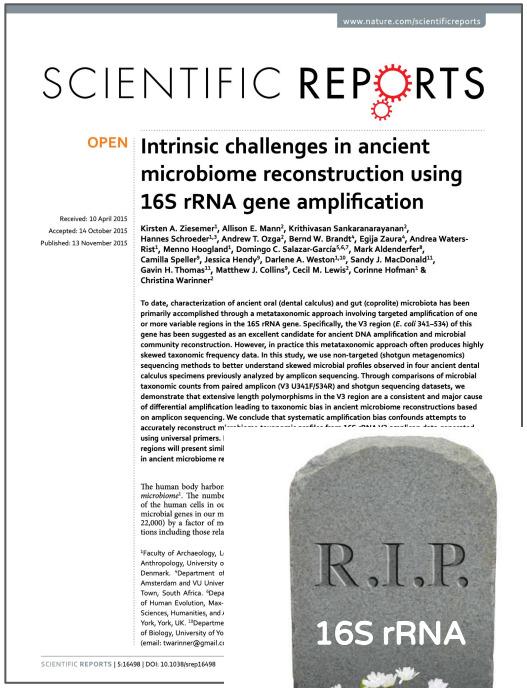
# 16S rRNA marker gene

16S rRNA amplicon **metataxonomics cannot** be used for ancient microbial DNA (Ziesemer 2015)

It is possible to analyze 16S rRNA sequences within **metagenomic** data, but…

- 16S rRNA sequences represent <0.05% of sequences, so it is **inefficient**
- classifying very short 16S rRNA sequences is **error prone**

So we now recommend **alternative approaches** using **metagenomics**


www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

**Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification**

Received: 10 April 2015
Accepted: 14 October 2015
Published: 13 November 2015

Kirsten A. Ziesemer[1], Allison E. Mann[2], Krithivasan Sankaranarayanan[2], Hannes Schroeder[1,3], Andrew T. Ozga[2], Bernd W. Brandt[4], Egija Zaura[4], Andrea Waters-Rist[1], Menno Hoogland[5], Domingo C. Salazar-García[5,6,7], Mark Aldenderfer[8], Camilla Speller[9], Jessica Hendy[9], Darlene A. Weston[1,10], Sandy J. MacDonald[11], Gavin H. Thomas[11], Matthew J. Collins[9], Cecil M. Lewis[2], Corinne Hofman[5] & Christina Warinner[2]

To date, characterization of ancient oral (dental calculus) and gut (coprolite) microbiota has been primarily accomplished through a metataxonomic approach involving targeted amplification of one or more variable regions in the 16S rRNA gene. Specifically, the V3 region (*E. coli* 341–534) of this gene has been suggested as an excellent candidate for ancient DNA amplification and microbial community reconstruction. However, in practice this metataxonomic approach often produces highly skewed taxonomic frequency data. In this study, we use non-targeted (shotgun metagenomics) sequencing methods to better understand skewed microbial profiles observed in four ancient dental calculus specimens previously analyzed by amplicon sequencing. Through comparisons of microbial taxonomic counts from paired amplicon (V3 U341F/534R) and shotgun sequencing datasets, we demonstrate that extensive length polymorphisms in the V3 region are a consistent and major cause of differential amplification leading to taxonomic bias in ancient microbiome reconstructions based on amplicon sequencing. We conclude that systematic amplification bias confounds attempts to accurately reconstruct m
using universal primers.
regions will present simi
in ancient microbiome re

The human body harbor
*microbiome*[1]. The numb
of the human cells in o
microbial genes in our m
22,000) by a factor of m
tions including those rela

[1]Faculty of Archaeology, L
Anthropology, University o
Denmark. [4]Department o
Amsterdam and VU Univer
Town, South Africa. [8]Depa
of Human Evolution, Max-
Sciences, Humanities, and
York, York, UK. [10]Departme
of Biology, University of Yo
(email: twarinner@gmail.c

SCIENTIFIC REPORTS | 5:16498 | DOI: 10.1038/srep16498

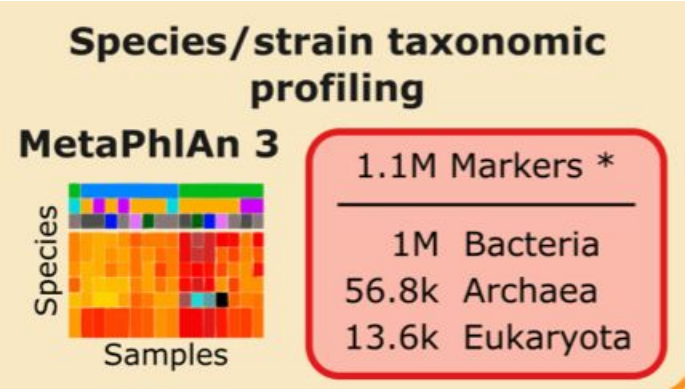
# The Workhorses

MetaPhlAn, MALT, Kraken

# MetaPhlAn: marker gene set

MetaPhlAn is a taxonomic profiler that uses **short read DNA sequence data** and a **database of marker genes** that are highly specific to certain clades

The current marker database contains 1.1 million markers from bacteria, archaea, and microeukaryotes




MetaPhlAn (Segata et al. 2012) and MetaPhlAn2 (Truong et al. 2015) are retired; MetaPhlAn3 (Beghini et al. 2021) is in current use; MetaPhlAn4 is in development and will include MAGs

Available in the bioBakery: https://github.com/biobakery

# MetaPhlAn: marker gene set

Pros:

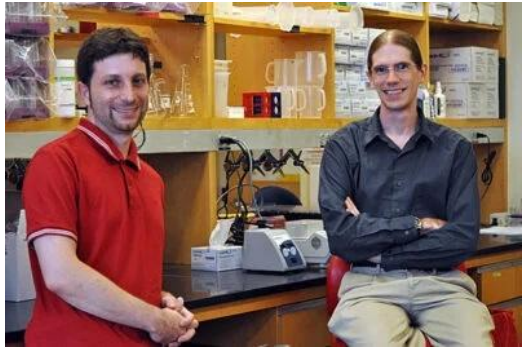
- Uses metagenomic data, and works well with aDNA
- Computationally efficient
- Marker database is good for pathogens and human microbiome

Cons:

- Because it uses a defined marker database, it has low customizability
- Marker database is missing taxa that are relevant for other animal microbiomes or environmental DNA
- Only profiles microbes

Overall, a good option for human-associated ancient microbes and microbiomes

# MetaPhlAn: marker gene set



Developed by Curtis Huttenhower and Nicola Segata, whose team has innovated many microbiome software tools

Other great tools from the same team include:

- PhyloPhlAn - for phylogenetic profiling of genomes and MAGs
- PanPhlAn - for pangenome strain-level analysis
- HUMAnN - for functional profiling (more about this on Friday!)

The team is also vastly expanding available microbial reference genomes through large-scale metagenomic assembly projects (more on Thursday)!

- >150,000 MAGs (Pasolli et al. 2020)
- >200,000 MAGs (Almeida et al. 2021)

# MALT: Read alignment and binning

Developed by Daniel Huson and Alexander Herbig

husonlab/**malt**

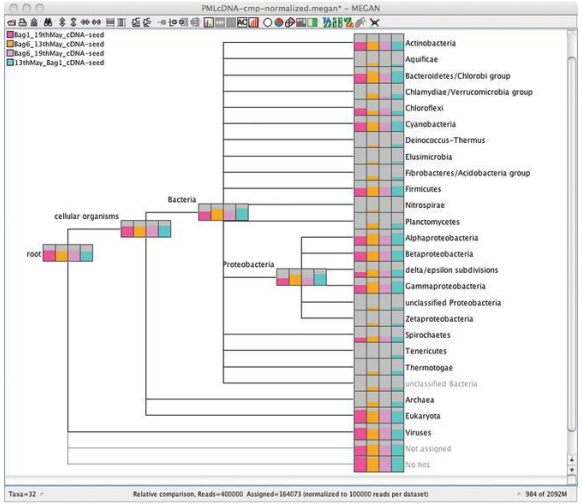
MEGAN alignment tool

**Short read DNA sequence aligner** for metagenomic data (Vågene et al. 2016) integrated into the **MEGAN** (the MEtaGenome ANalyzer) software suite (Huson et al. 2007)

Acronym for **MEGAN Alignment Tool** (MALT)

Works similar to BLAST but much faster

Developed as a DNA alternative to the protein sequence aligner DIAMOND (2015) for use in MEGAN



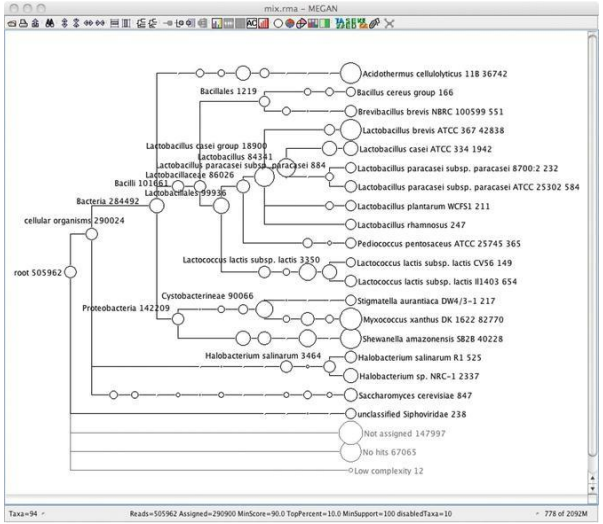
# MALT: Read alignment and binning

MALT uses **all of the DNA** in a dataset to perform taxonomic assignment by aligning to a reference database, such as NCBI nr or RefSeq

This makes it **slow and memory-intensive**, but it maximizes the data available

Database is **customizable**, can be used for all taxa, not just microbes

Uses an **LCA** (lowest common ancestor) algorithm to assign each sequence to a node in the taxonomy



# MALT: Read alignment and binning

Pros:

- Maximizes use of data
- Good database customizability
- Can profile ALL taxa in a sample, not just microbes
- MEGAN interface for quick data inspection
- Integrated into EAGER (Fellows Yates et al. 2021) and compatible with HOPS (Hübler et al. 2019) for pathogen screening
- Because it produces alignments, you can easily create DNA damage profiles

Cons:

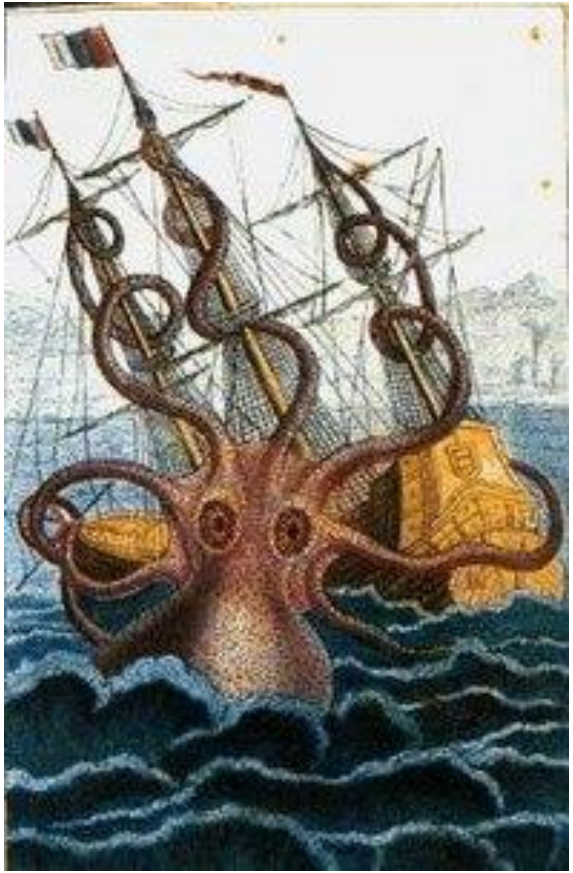
- Very computationally intensive with large databases
- Newest release has a bug in the LCA algorithm that is not yet fixed

# K-mer matching: Kraken

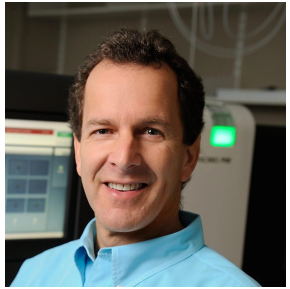
Kraken is a taxonomic profilers that works by **k-mer matching** rather than alignment

This makes Kraken **MUCH faster** and **LESS computationally intensive** that alignment-based profilers

Database is **customizable**, can be used for all taxa, not just microbes



# K-mer matching: Kraken

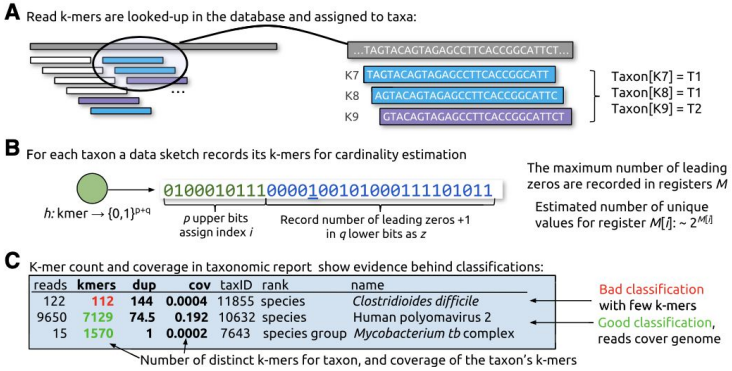


Developed by Derrick Wood and Stephen Salzberg (2014)

Correction developed to account for genome size differences when calculating species abundance with **Bracken** (Lu et al. 2017)

False positives reduced with **KrakenUniq** (Breitwieser et al. 2018)

Made even faster with **Kraken 2** (Wood et al. 2019)



# K-mer matching: Kraken

Pros:

- Fast!
- Can be used for any set of taxa, not just microbes
- Great for quickly seeing what's in your data
- Accuracy good enough for most ancient microbiome studies, but ancient pathogens require more validation

Cons:

- Can be prone to false positives
- Doesn't provide alignment data, so damage analysis must be performed separately

# Comparing taxonomic classifiers

No taxonomic profilers are perfect

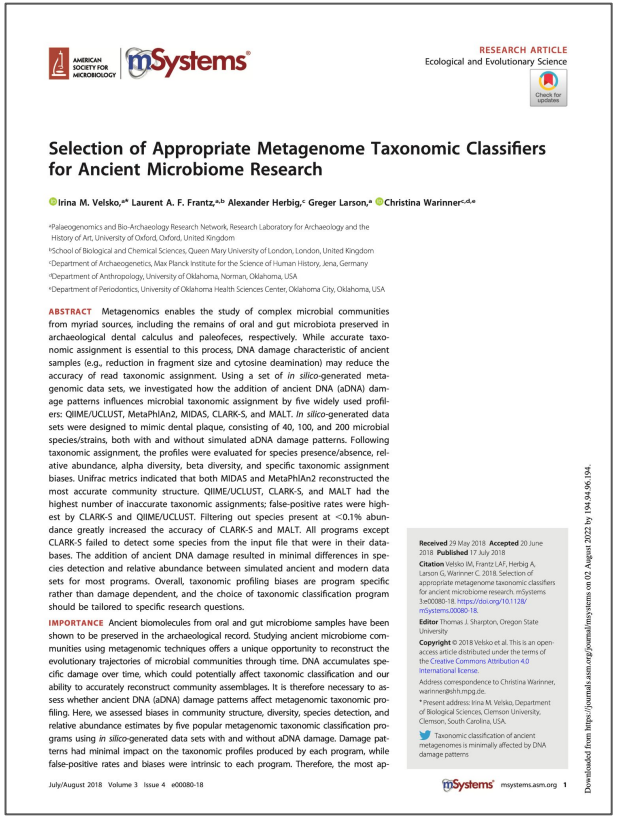
**False positives** tend to be low abundance taxa

Removing **singletons** and **low abundance taxa** helps reduce false positives

Taxonomic profilers generally return broadly **similar results**, but with some **predictable biases**

**Database selection** impacts the precision and accuracy of taxonomic assignment

Select the profiler(s) that will be best for **your study**



AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Ecological and Evolutionary Science

## Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research

Irina M. Velsko,[a*] Laurent A. F. Frantz,[a,b] Alexander Herbig,[c] Greger Larson,[a] Christina Warinner[c,d,e]

[a]Palaeogenomics and Bio-Archaeology Research Network, Research Laboratory for Archaeology and the History of Art, University of Oxford, Oxford, United Kingdom
[b]School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom
[c]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany
[d]Department of Anthropology, University of Oklahoma, Norman, Oklahoma, USA
[e]Department of Periodontics, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA

**ABSTRACT** Metagenomics enables the study of complex microbial communities from myriad sources, including the remains of oral and gut microbiota preserved in archaeological dental calculus and paleofeces, respectively. While accurate taxonomic assignment is essential to this process, DNA damage characteristic of ancient samples (e.g., reduction in fragment size and cytosine deamination) may reduce the accuracy of read taxonomic assignment. Using a set of *in silico*-generated metagenomic data sets, we investigated how the addition of ancient DNA (aDNA) damage patterns influences microbial taxonomic assignment by five widely used profilers: QIIME/UCLUST, MetaPhlAn2, MIDAS, CLARK-S, and MALT. *In silico*-generated data sets were designed to mimic dental plaque, consisting of 40, 100, and 200 microbial species/strains, both with and without simulated aDNA damage patterns. Following taxonomic assignment, the profiles were evaluated for species presence/absence, relative abundance, alpha diversity, beta diversity, and specific taxonomic assignment biases. Unifrac metrics indicated that both MIDAS and MetaPhlAn2 reconstructed the most accurate community structure. QIIME/UCLUST, CLARK-S, and MALT had the highest number of inaccurate taxonomic assignments; false-positive rates were highest by CLARK-S and QIIME/UCLUST. Filtering out species present at <0.1% abundance greatly increased the accuracy of CLARK-S and MALT. All programs except CLARK-S failed to detect some species from the input file that were in their databases. The addition of ancient DNA damage resulted in minimal differences in species detection and relative abundance between simulated ancient and modern data sets for most programs. Overall, taxonomic profiling biases are program specific rather than damage dependent, and the choice of taxonomic classification program should be tailored to specific research questions.

**IMPORTANCE** Ancient biomolecules from oral and gut microbiome samples have been shown to be preserved in the archaeological record. Studying ancient microbiome communities using metagenomic techniques offers a unique opportunity to reconstruct the evolutionary trajectories of microbial communities through time. DNA accumulates specific damage over time, which could potentially affect taxonomic classification and our ability to accurately reconstruct community assemblages. It is therefore necessary to assess whether ancient DNA (aDNA) damage patterns affect metagenomic taxonomic profiling. Here, we assessed biases in community structure, diversity, species detection, and relative abundance estimates by five popular metagenomic taxonomic classification programs using *in silico*-generated data sets with and without aDNA damage. Damage patterns had minimal impact on the taxonomic profiles produced by each program, while false-positive rates and biases were intrinsic to each program. Therefore, the most ap-

**Received** 29 May 2018 **Accepted** 20 June 2018 **Published** 17 July 2018
**Citation** Velsko IM, Frantz LAF, Herbig A, Larson G, Warinner C. 2018. Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. mSystems 3:e00080-18. https://doi.org/10.1128/msystems.00080-18.
**Editor** Thomas J. Sharpton, Oregon State University

Address correspondence to Christina Warinner, warinner@shh.mpg.de.
* Present address: Irina M. Velsko, Department of Biological Sciences, Clemson University, Clemson, South Carolina, USA.

Taxonomic classification of ancient metagenomes is minimally affected by DNA damage patterns

# Databases! Databases! Databases!

Databases matter…a lot

Many databases are incomplete, and you won't find what you can't "see", so always check to make sure your database has your taxon of interest

- ***Example:*** The first MetaPhlAn database lacked *Tannerella forsythia*, so this common oral microbe would "disappear" if you analyzed it with MetaPhlAn. The new MetaPhlAn2 and 3 databases fixed this!

If your taxon is missing a reference genome in the database, your DNA might align to the next best thing, causing a false positive

- ***Example:*** Taxonomic profiling of dental calculus prior to 2012 indicated the skin pathogen *Propionibacterium acnes* was prevalent and abundant. After the genome of the related oral species *Pseudopropionibacterium propionicum* was published in 2012, *P. acnes* "disappeared" from these datasets

# Databases! Databases! Databases!

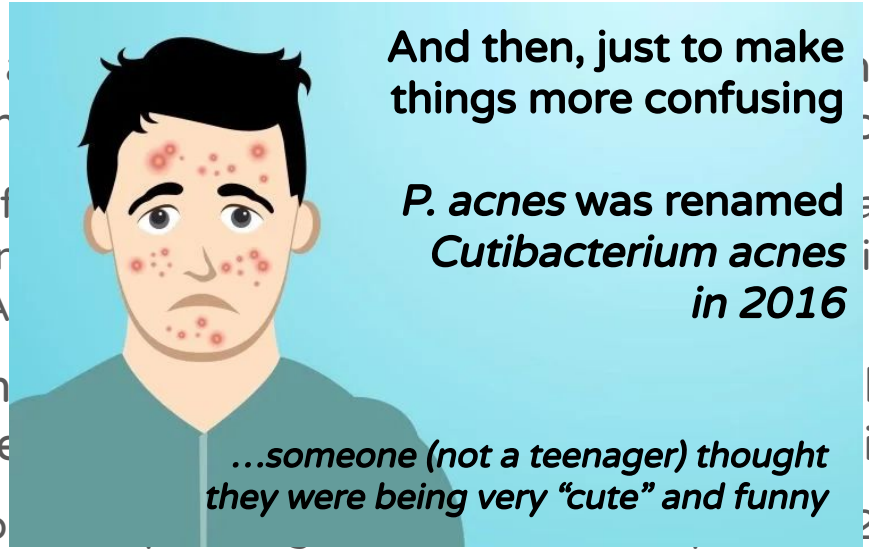
Databases matter…a lot

Many databases at you can't "see", so always check to n on of interest

- ***Example:*** The f a *forsythia*, so this common oral r it with MetaPhlAn. The new MetaPhlA

If your taxon is m base, your DNA might align to the ive

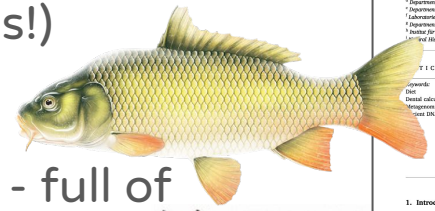
- ***Example:*** Taxo 2012 indicated the skin pathogen *Propionibacterium acnes* was prevalent and abundant. After the genome of the related oral species *Pseudopropionibacterium propionicum* was published in 2012, *P. acnes* "disappeared" from these datasets



# Databases! Databases! Databases!

Databases also contain junk data

Genomes in NCBI (even RefSeq genomes!) contain errors…sometimes BIG errors



ARTICLE IN PRESS

Quaternary International xxx (xxxx) xxx

Contents lists available at ScienceDirect

Quaternary International

journal homepage: www.elsevier.com/locate/quaint

Do I have something in my teeth? The trouble with genetic analyses of diet from archaeological dental calculus

Allison E. Mann [a,*], James A. Fellows Yates [b,c], Zandra Fagernäs [b], Rita M. Austin [d,e,f,i], Elizabeth A. Nelson [b,g,h], Courtney A. Hofman [a,i]

[a] Department of Biological Sciences, Clemson University, Clemson, SC, USA
[b] Department of Archaogenetics, Max Planck Institute for the Science of Human History, Jena, Germany
[c] Institut für Vor- und Frühgeschichtliche Archäologie und Prähistorische Archäologie, Ludwig Maximilian University, Munich, Germany
[d] Department of Anthropology, Smithsonian National Museum of Natural History, Washington, DC, USA
[e] Department of Anthropology, University of Oklahoma, Norman, OK, USA
[f] Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, OK, USA
[g] Department of Anthropology, University of Texas at Arlington, Arlington, TX, USA
[h] Institut für Naturwissenschaftliche Archäologie, Eberhard Karls Universität Tübingen, 72074, Tübingen, Germany
[i] Museum of Natural History Museum Oslo, University of Oslo, Oslo, Norway

ARTICLE INFO

Keywords:
Diet
Dental calculus
Metagenomics
Ancient DNA

ABSTRACT

Dental calculus and other preserved microbiome substrates are an attractive target for dietary reconstruction in past populations through a variety of physical, chemical, and molecular means. Recently, studies have attempted to reconstruct diet from archaeological dental calculus using archaeogenetic techniques. While dental calculus may provide a relatively stable environment for DNA preservation, the detection of plants and animals possibly consumed by an individual through DNA analysis is primarily hindered by microbial richness and incomplete reference databases. Moreover, high genomic similarity within eukaryotic groups - such as mammals - can obfuscate precise taxonomic identification. In the current study we demonstrate the challenges associated with accurate taxonomic identification and authentication of dietary taxa in ancient DNA data using both synthetic and ancient dental calculus datasets. We highlight common errors and sources of contamination across ancient DNA datasets, provide recommendations for dietary DNA validation, and call for caution in the interpretation of diet from dental calculus and other archaeological microbiome substrates.

1. Introduction

Diet is a fundamental component of human culture, biology, and evolution. Shifts in food procurement, production, and processing are inherently linked to shifts in human society and major evolutionary events (Goodman and Redclift, 2002; Larsen, 2003; Bocquet-Appel and Bar-Yosef, 2008; Ma et al., 2016; Andrews and Johnson, 2020). What people choose to eat (or not to eat) provides insight into cultural values and beliefs (Messer, 1984). Archaeological study of the interrelationships between people and foods, such as plants and animals, has revealed complex cultural practices and socio-political structures (Atalay and Hastorf, 2006; Tung et al., 2016; Morehart and Morell-Hart, 2015) organized around food consumption and distribution in which diet is directly related to social status (Coalliar, 2013), ceremonial events, control of food supply, and the establishment of trade (Williams, 2010; Tung and Knudson, 2018). Likewise, the effect of environmental factors such as climate shifts and geologic processes on dietary resources can be investigated through the study of ancient diet (Messner and Stinchcomb, 2014; Nelson et al., 2016). How food was produced and procured throughout human history also provides important historical context for understanding human health in the modern era. Given its importance in understanding the human condition, archaeologists work to reconstruct past diets using a variety of techniques, including analyses of faunal assemblages (e.g., Blasco et al., 2013), paleobotanical remains (e.g., Pearsall, 2018), coprolite analysis (e.g., Reinhard and Bryant, 1992), and dental analyses of wear, development, and disease (e.g., Forshaw, 2014; Molnar et al., 1972; White and Folkens, 2005). The introduction of bulk biomolecular methods, such as stable isotope

* Corresponding author. Clemson University Department of Biological Sciences, 111 Jordan Hall, Clemson, SC, 29634.
E-mail address: amann3@g.clemson.edu (A.E. Mann).

https://doi.org/10.1016/j.quaint.2020.11.019
Received 3 August 2020; Received in revised form 2 November 2020; Accepted 12 November 2020
Available online 19 November 2020
1040-6182/ © 2020 Published by Elsevier Ltd.

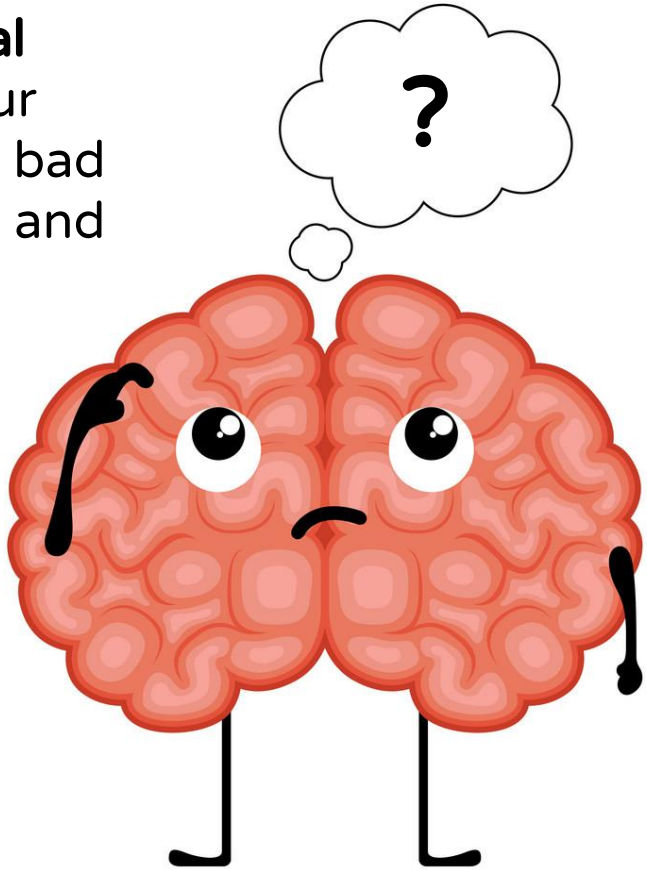


Common carp (*Cyprinus carpio*) genome - full of sequencing adapters!

Tibetan antelope (*Pantholops hodgsonii*) turns up in every metagenomic dataset 🤔

RefSeq genome of the common soil bacterium *Achromobacter denitrificans* contains the entire chicken ovalbumin gene!

Your **brain** and **critical thinking skills** are your best defense against bad databases, bad data, and wrong conclusions




When in doubt, check and double check!

# Starting questions


**Who's there?**

That was a lot of work!

# Starting questions


Who's there?

How preserved is my sample?

# Metagenome composition and quality

Causes of degradation and sources of contamination

- Burial environment (necrobiome)
- Postmortem microbial overgrowth
- Post-excavation handling and storage

Helpful to identify and remove contaminant sequences from your dataset before proceeding to downstream analyses

Software tools can help you characterize your dataset's preservation state and potential contamination

- Source tracking: SourceTracker, Source Predict
- Cleanup: cuperdec, decontam

# Metagenome composition and quality

**Microbial source tracking** can be performed using Bayesian or machine learning methods to estimate to what degree your data derives from a particular microbial source

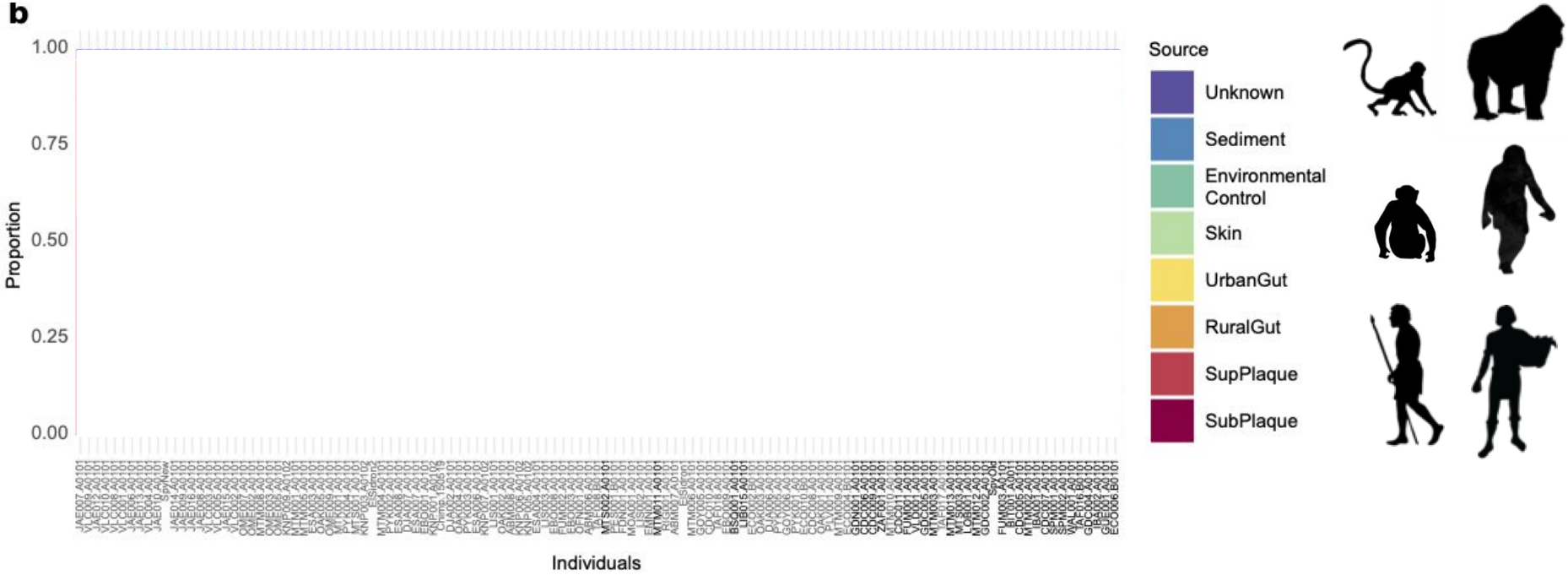
Two main methods:

- SourceTracker2 (Knights et al. 2011)
- Source Predict (Borry 2020)

User provides reference metagenomes (e.g., dental plaque, feces, soil) as sources and the tool estimates the proportion of your dataset that derives from one or more of these sources

# Metagenome composition and quality

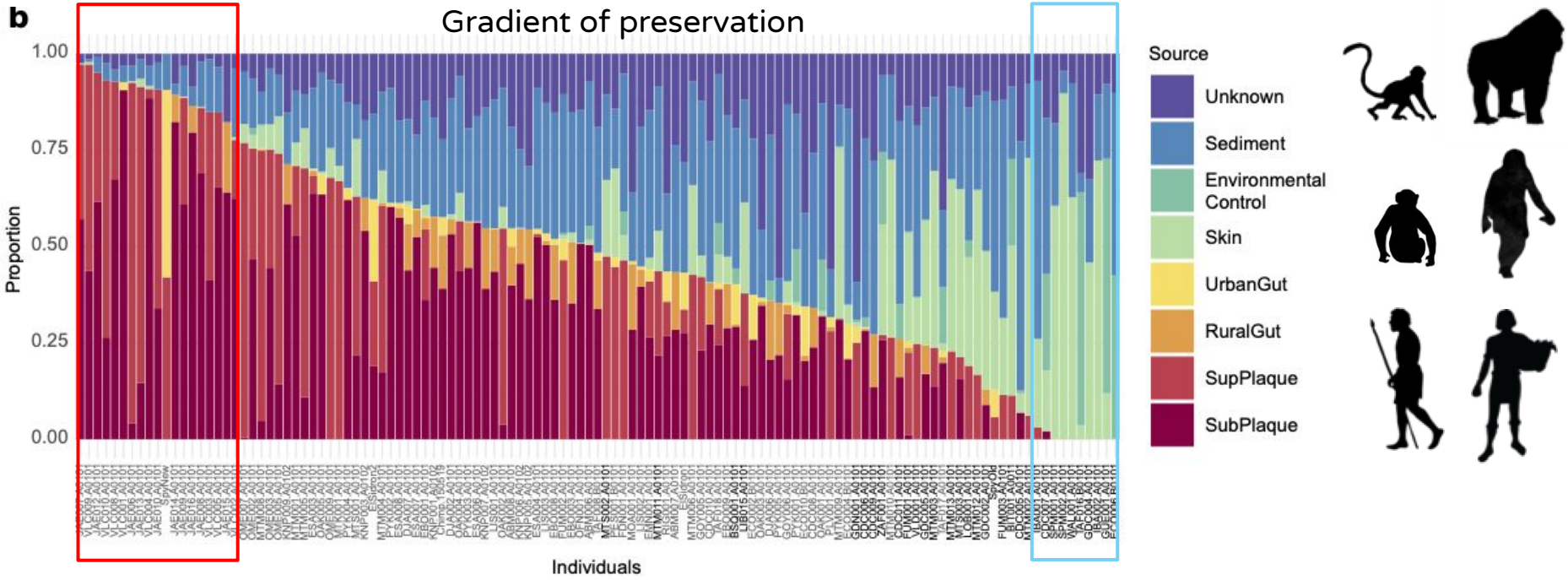
## SourceTracker2




Fellows Yates et al 2021

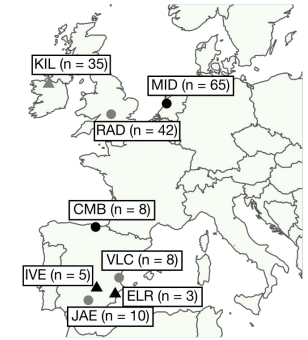
# Metagenome composition and quality

## SourceTracker2

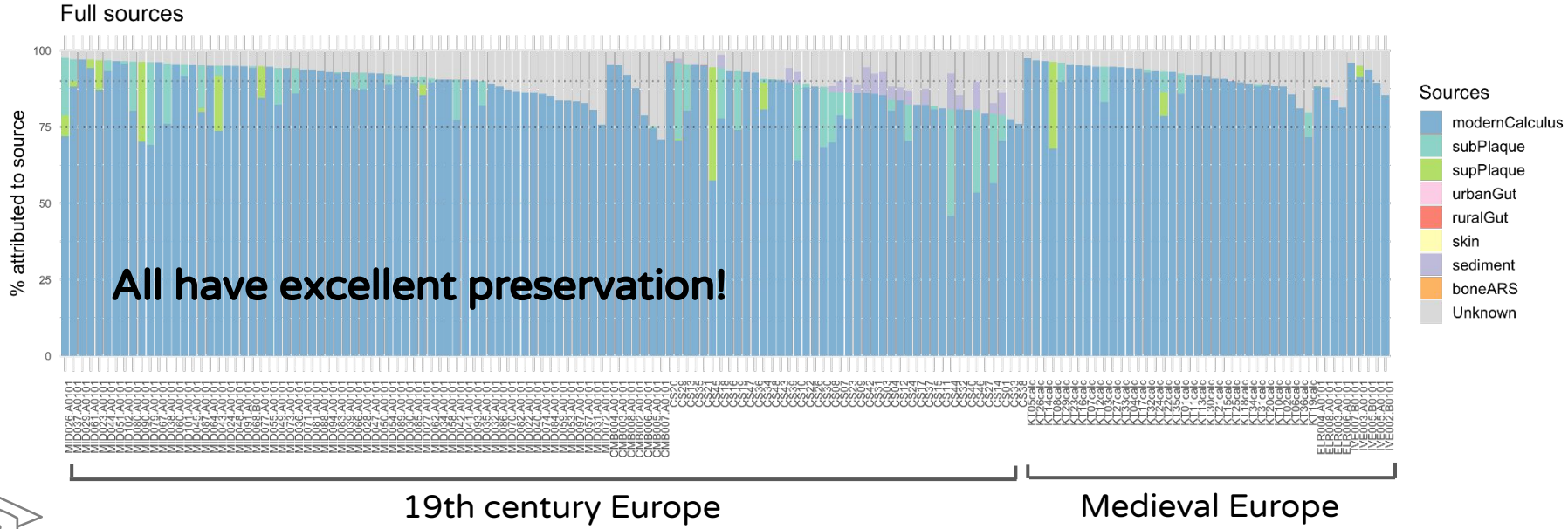



Fellows Yates et al 2021

# Metagenome composition and quality




## SourceTracker2



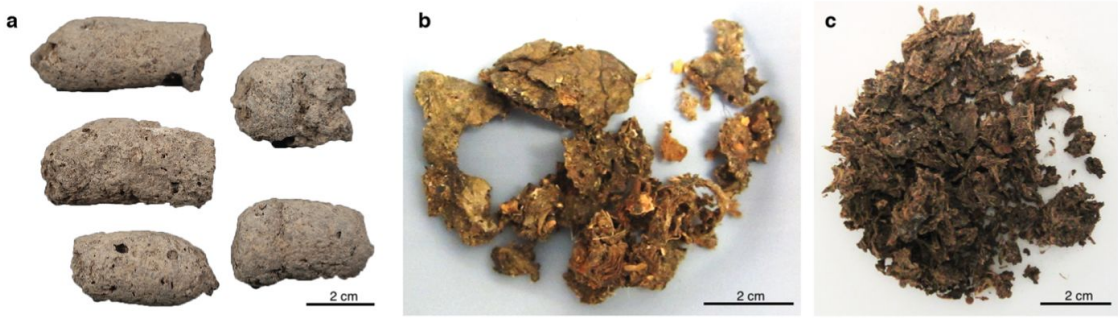

**All have excellent preservation!**

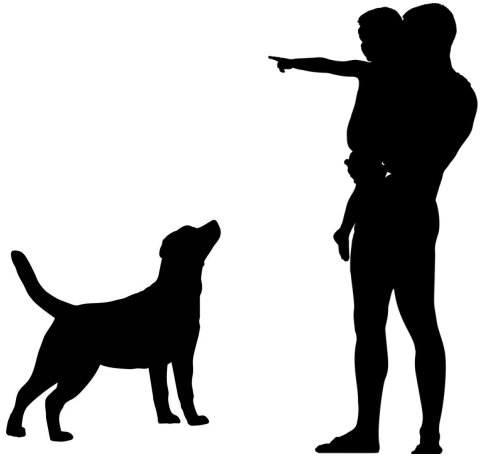
# Metagenome composition and quality

Source Predict

Beyond preservation, you might also what to know, *What is my sample?*

- Human paleofeces
- Dog poop?
- Something else?



# Metagenome composition and quality

*SourceTracker Pro Tips:*

Choose your sources wisely!

- You need at least 10 datasets per source
- Plaque and calculus have similar but distinct profiles
- Archaeological bone is a better proxy for the necrobiome than soil

*Important!* The category “unknown” includes both:

- the proportion of your dataset that cannot be assigned to any source
- the proportion that can be assigned to more than one source

# Starting questions


Who's there?

How preserved is my sample?

How do I clean up my dataset?

# Cleaning up your dataset

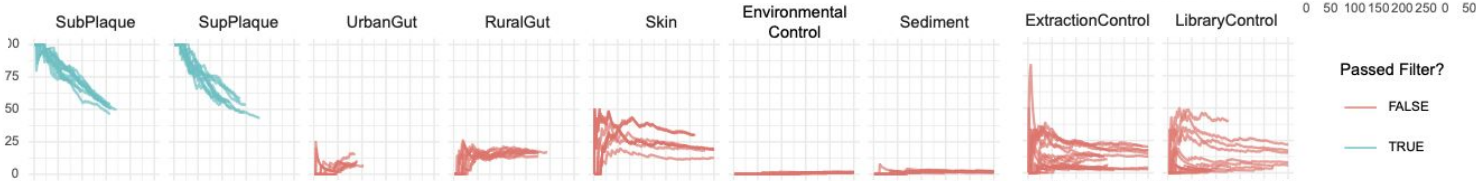
Now that you have a sense of your sample's preservation, you can clean it up for downstream analyses

Two step process:

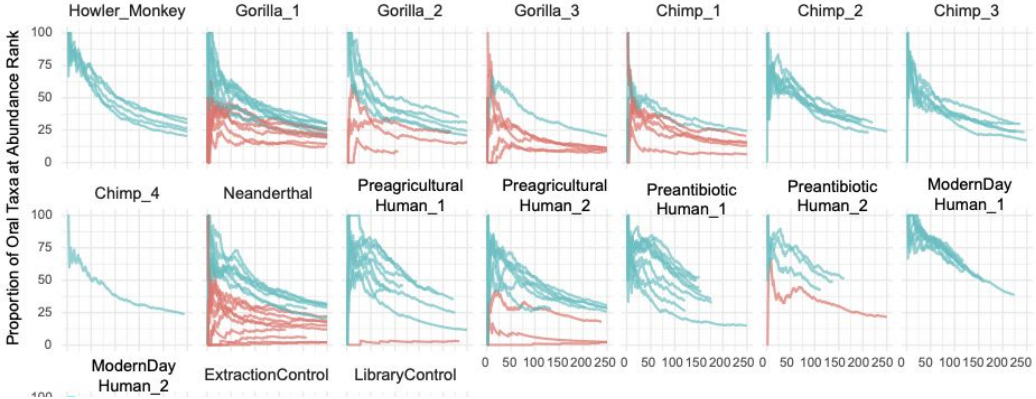
- Identify and remove the very degraded **samples** altogether using **cuperdec** (Fellows Yates et al. 2021)
- Identify and remove low-level laboratory and soil contaminant **taxa** from your datasets using **decontam** (Davis et al. 2018)

# cuperdec - remove the samples beyond hope

References and controls




Samples



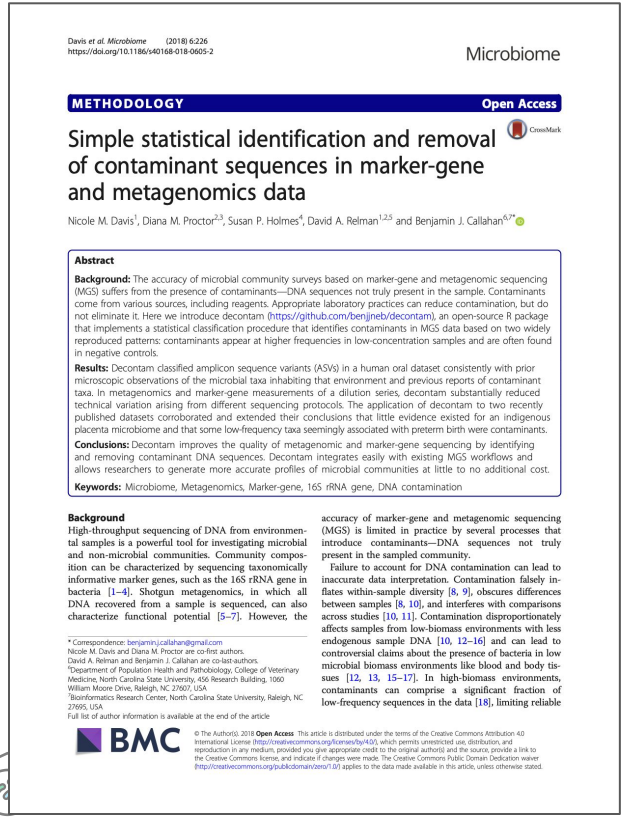

Some **samples** are so degraded and altered postmortem that they aren't worth analyzing

cuperdec can help you identify these so you can remove them from your analyses

cuperdec **removes samples** from your study

# decontam - surgical removal of contaminants



Davis *et al. Microbiome* (2018) 6:226
https://doi.org/10.1186/s40168-018-0605-2

Microbiome

METHODOLOGY Open Access

## Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis[1], Diana M. Proctor[2,3], Susan P. Holmes[4], David A. Relman[1,2,5] and Benjamin J. Callahan[6,7*]

**Abstract**

**Background:** The accuracy of microbial community surveys based on marker-gene and metagenomic sequencing (MGS) suffers from the presence of contaminants—DNA sequences not truly present in the sample. Contaminants come from various sources, including reagents. Appropriate laboratory practices can reduce contamination, but do not eliminate it. Here we introduce decontam (https://github.com/benjjneb/decontam), an open-source R package that implements a statistical classification procedure that identifies contaminants in MGS data based on two widely reproduced patterns: contaminants appear at higher frequencies in low-concentration samples and are often found in negative controls.

**Results:** Decontam classified amplicon sequence variants (ASVs) in a human oral dataset consistently with prior microscopic observations of the microbial taxa inhabiting that environment and previous reports of contaminant taxa. In metagenomics and marker-gene measurements of a dilution series, decontam substantially reduced technical variation arising from different sequencing protocols. The application of decontam to two recently published datasets corroborated and extended their conclusions that little evidence existed for an indigenous placenta microbiome and that some low-frequency taxa seemingly associated with preterm birth were contaminants.

**Conclusions:** Decontam improves the quality of metagenomic and marker-gene sequencing by identifying and removing contaminant DNA sequences. Decontam integrates easily with existing MGS workflows and allows researchers to generate more accurate profiles of microbial communities at little to no additional cost.

**Keywords:** Microbiome, Metagenomics, Marker-gene, 16S rRNA gene, DNA contamination

**Background**

High-throughput sequencing of DNA from environmental samples is a powerful tool for investigating microbial and non-microbial communities. Community composition can be characterized by sequencing taxonomically informative marker genes, such as the 16S rRNA gene in bacteria [1–4]. Shotgun metagenomics, in which all DNA recovered from a sample is sequenced, can also characterize functional potential [5–7]. However, the accuracy of marker-gene and metagenomic sequencing (MGS) is limited in practice by several processes that introduce contaminants—DNA sequences not truly present in the sampled community.

Failure to account for DNA contamination can lead to inaccurate data interpretation. Contamination falsely inflates within-sample diversity [8, 9], obscures differences between samples [8, 10], and interferes with comparisons across studies [10, 11]. Contamination disproportionately affects samples from low-biomass environments with less endogenous sample DNA [10, 12–16] and can lead to controversial claims about the presence of bacteria in low microbial biomass environments like blood and body tissues [12, 13, 15–17]. In high-biomass environments, contaminants can comprise a significant fraction of low-frequency sequences in the data [18], limiting reliable

* Correspondence: benjamin.j.callahan@gmail.com
Nicole M. Davis and Diana M. Proctor are co-first authors.
David A. Relman and Benjamin J. Callahan are co-last-authors.
[2]Department of Population Health and Pathobiology, College of Veterinary Medicine, North Carolina State University, 456 Research Building, 1060 William Moore Drive, Raleigh, NC 27607, USA
[7]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA
Full list of author information is available at the end of the article



Some **samples** are okay, but they have some stubborn **contaminant taxa** you want to remove

If you leave them in, these contaminant taxa could bias or skew your diversity patterns, leading to spurious results and false conclusions

decontam can help you identify the obvious contaminants and remove them

You provide decontam with contaminant sources (e.g., datasets from laboratory blanks, archaeological bone)

decontam **removes contaminating taxa** from your datasets

# What's next?


Who's there?

How preserved is my sample?

How do I clean up my dataset?

# Diversity

Within **ecology** there are many ways to examine the **microbial communities** in your samples in order to better understand them

The two most common ways are to examine and compare their:

- alpha diversity
- beta diversity

# Alpha diversity


Alpha diversity measures the variation within a single sample

**Species richness** (e.g., Chao1 index)

- How many different species are in my microbial community?

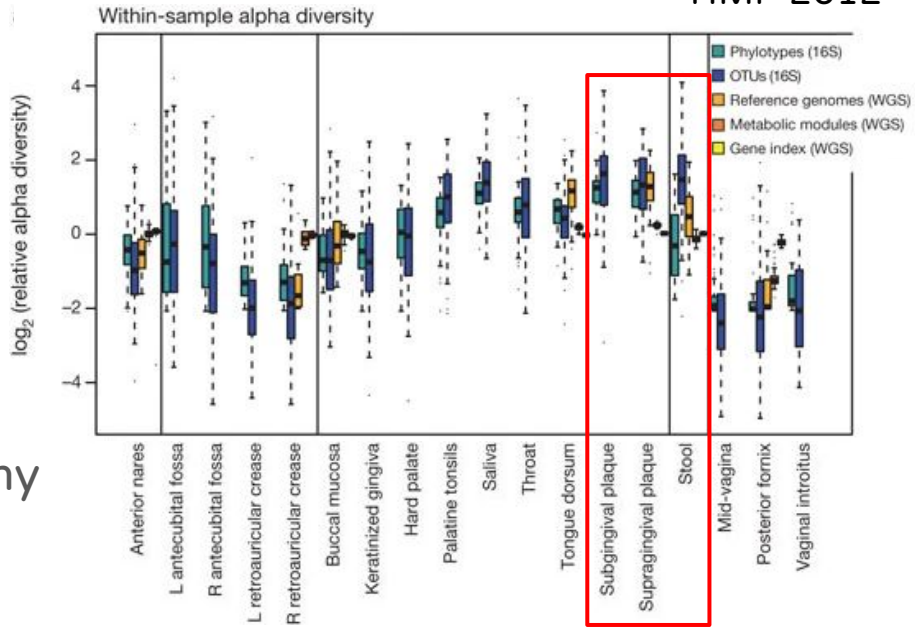
**Species evenness** (e.g., Shannon index)

- How balanced are the species abundances in my community? Do a few taxa dominate the sample or not?

*Pro tip:* alpha diversity is easily skewed in ancient samples by preservation and trace contaminants, so be careful when interpreting ancient alpha diversity!

# Alpha diversity

HMP 2012




Alpha diversity measures the variation within a single sample

**Species richness** (e.g., Chao1 index)

- How many different species are in my microbial community?

**Species evenness** (e.g., Shannon index)

- How balanced are the species abundances in my community? Do a few taxa dominate the sample or not?

*Pro tip:* alpha diversity is easily skewed in ancient samples by preservation and trace contaminants, so be careful when interpreting ancient alpha diversity!

# Beta diversity


Beta diversity measures the variation between samples

**Bray-Curtis dissimilarity**

- To what degree are taxa shared between my samples at same abundances? 0=exactly the same; 1=completely different

**Jaccard distance**

- To what degree are taxa shared between my samples (ignoring abundance)? 0=exact same taxa; 1= completely different taxa

**UniFrac**

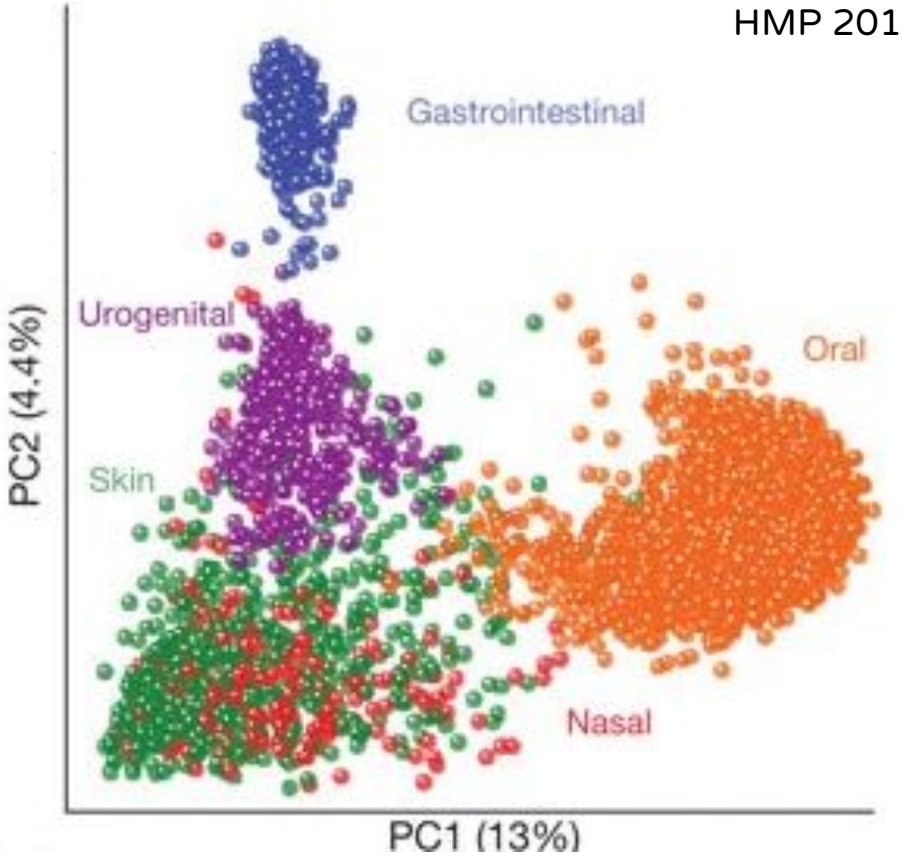
- How phylogentetically similar are the taxa in my samples, taking into account abundance (weighted) or not (unweighted)

# Beta diversity

You can visualize the beta diversity of a given set of samples using **Principal Coordinates Analysis (PCoA)**

Here is an example of a PCoA based on Bray-Curtis distances of the microbial communities present in the **human microbiome**

HMP 2012



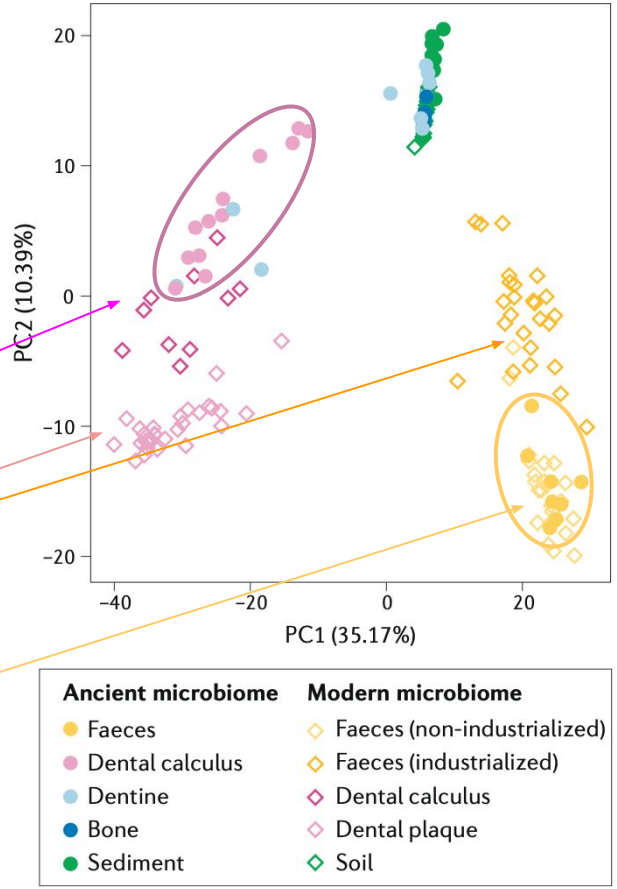
# Beta diversity

Orlando et al. 2021

Here is an example of a PCoA based on Bray-Curtis distances of the microbial communities present in the **archaeological samples**, including **paleofeces** and **dental calculus**

Here you can see compositional differences between modern dental calculus and plaque, and that ancient calculus **overlaps modern calculus**

You can also see that feces from modern industrialized and non-industrialized populations are distinct, and that paleofeces **resembles modern non-industrialized feces**



# PCoA vs PCA

I've never heard of PCoA - what's that?

**PCoA (principle coordinate analysis)** is applied to your distance matrix (Bray-Curtis, Jaccard, UniFrac) in order to visualize your beta diversity in a plot

Alternatively, you can also take an entirely different **compositional approach** by transforming the data in your taxon table using a **centered log-ratio transformation (CLR)**, building a **euclidean distance matrix***, and performing a **PCA (principal components analysis)** to visualize your samples in a plot

*a euclidean distance matrix built from CLR transformed data is also called an **Aitchison distance matrix**; PCAs can only be performed on a euclidean distance matrix

# Standard Model vs Compositional Approach

Which approach is better? It's a bit of a philosophical debate - with **strong feelings** on both sides. Both are valid for metagenomics (with different caveats) and represent your data in slightly different ways. Try both!

*Bottom line:* the two approaches deal with 0 count data and discrepancies in sampling effort differently

Read more about the growing importance of compositional approaches to microbiome analysis in Gloor et al. 2017

**Microbiome Datasets Are Compositional: And This Is Not Optional**

*Gregory B. Gloor[1*], Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]*

*[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada, [2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain, [3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain*

| Operation | Standard approach | Compositional approach |
|---|---|---|
| Normalization | Rarefaction 'DESeq' | CLR ILR ALR |
| Distance | Bray-Curtis UniFrac Jenson-Shannon | Aitchison |
| Ordination | PCoA (Abundance) | PCA (Variance) |
| Multivariate comparison | perManova ANOSIM | perMANOVA ANOSIM |
| Correlation | Pearson Spearman | SparCC SpiecEasi ϕ ρ |
| Differential abundance | metagenomSeq LEfSe DESeq | ALDEx2 ANCOM |

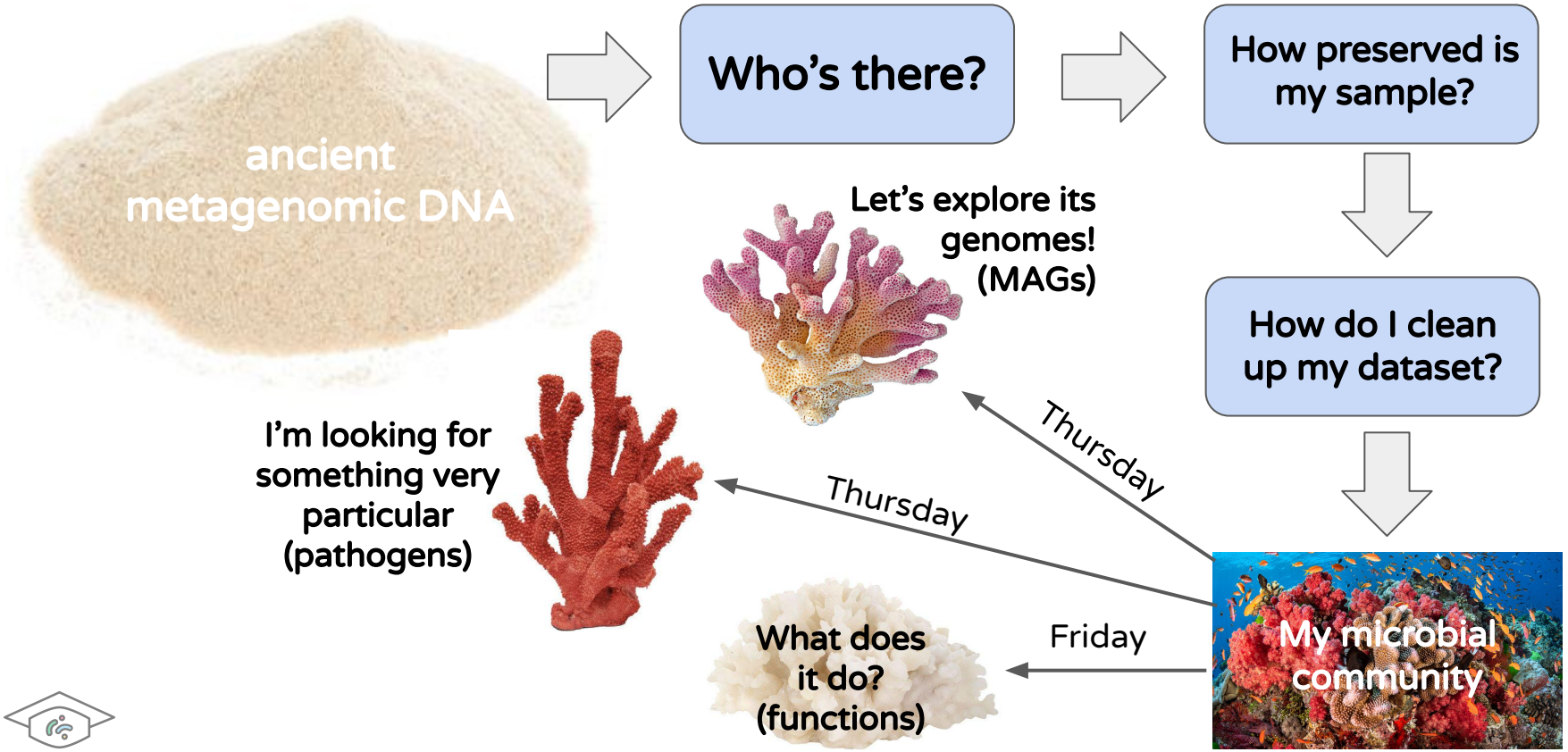
# Intrigued, want to learn more?

**Pat Schloss**, who created mothur, has a series of YouTube videos about ecological analyses and distances, and he explains in detail how to use the R package vegan for microbiome analysis. Check them out!

- Ecological distances in R, https://www.youtube.com/watch?v=xyufizOpc5I
- How to calculating the Aitchison distance in R, https://www.youtube.com/watch?v=ulo7WatBEAo

| Operation | Standard approach | Compositional approach |
|---|---|---|
| Normalization | Rarefaction 'DESeq' | CLR ILR ALR |
| Distance | Bray-Curtis UniFrac Jenson-Shannon | Aitchison |
| Ordination | PCoA (Abundance) | PCA (Variance) |
| Multivariate comparison | perManova ANOSIM | perMANOVA ANOSIM |
| Correlation | Pearson Spearman | SparCC SpiecEasi $\phi$ $\rho$ |
| Differential abundance | metagenomSeq LEfSe DESeq | ALDEx2 ANCOM |

# What's next?



# Want to read more?

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., ... & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature biotechnology, 39(1), 105-114.

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., ... & Segata, N. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nature communications, 11(1), 1-10.

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., ... & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife, 10, e65088.

Borry, M., 2019. Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification. The Journal of Open Source Software.

Breitwieser, F. P., Baker, D. N., & Salzberg, S. L. (2018). KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome biology, 19(1), 1-10.

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome, 6(1), 1-14.

Eisen JA. What does the term microbiome mean? And where did it come from? A bit of a surprise. (2015) microBEnet: Microbiology of the Built Environment Network. Available at http:// www.microbe.net/2015/04/08/what-does-the-term-microbiome-mean-and-where-did-it-come-from-a-bit-of-a-surprise/

# Want to read more?

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology, 5(10), R245-R249.

Hübler, R., Key, F. M., Warinner, C., Bos, K. I., Krause, J., & Herbig, A. (2019). HOPS: automated detection and authentication of pathogen DNA in archaeological remains. Genome Biology, 20(1), 1-13.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome research, 17(3), 377-386.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. and Kelley, S.T., 2011. Bayesian community-wide culture-independent microbial source tracking. Nature methods, 8(9), pp.761-763.

Koslicki, D., & Falush, D. (2016). MetaPalette: a k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. MSystems, 1(3), e00020-16.

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science, 3, e104.

Mann, A.E., Yates, J.A.F., Fagernäs, Z., Austin, R.M., Nelson, E.A. and Hofman, C.A., 2020. Do I have something in my teeth? The trouble with genetic analyses of diet from archaeological dental calculus. Quaternary International.

Marchesi, J. R., & Ravel, J. (2015). The vocabulary of microbiome research: a proposal. Microbiome, 3(1), 1-3.

# Want to read more?

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology, 8, 2224.

Mitreva, M., & Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*, 207-214.

Pasolli, E., De Filippis, F., Mauriello, I. E., Cumbo, F., Walsh, A. M., Leech, J., ... & Ercolini, D. (2020). Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. *Nature communications*, *11*(1), 1-12.

Prescott, S. L. (2017). History of medicine: Origin of the term microbiome and why it matters. *Human Microbiome Journal*, *4*, 24-25.

Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, *4*(1), 1-11.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, *9*(8), 811-814.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., ... & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods*, *12*(10), 902-903.

Vågene, Å. J., Herbig, A., Campana, M. G., Robles García, N. M., Warinner, C., Sabin, S., ... & Krause, J. (2018). Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature ecology & evolution*, *2*(3), 520-528.

# Want to read more?

Velsko, I. M., Frantz, L. A., Herbig, A., Larson, G., & Warinner, C. (2018). Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. *Msystems*, *3*(4), e00080-18.

Warinner, C., Herbig, A., Mann, A., Yates, J. A. F., Weiβ, C. L., Burbano, H. A., ... & Krause, J. (2017). A robust framework for microbial archaeology. Annual review of genomics and human genetics, 18, 321.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, *15*(3), 1-12.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, *20*(1), 1-13.

Yates, J. A. F., Lamnidis, T. C., Borry, M., Valtueña, A. A., Fagernäs, Z., Clayton, S., ... & Peltzer, A. (2021). Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ*, *9*, e10947.