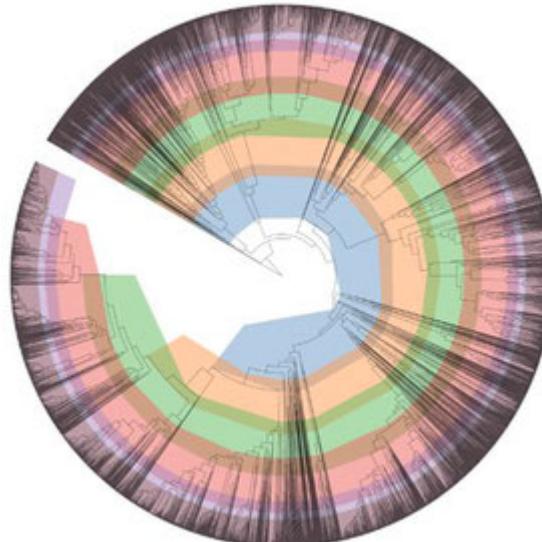


ENVIRONMENTAL METAGENOMICS

Physalia course, online, 11-15 November 2024

Read-based taxonomic profiling

Nikolay Oskolkov, Lund University, NBIS SciLifeLab
Luis Pedro Coelho, Queensland University of Technology



Physalia
Courses

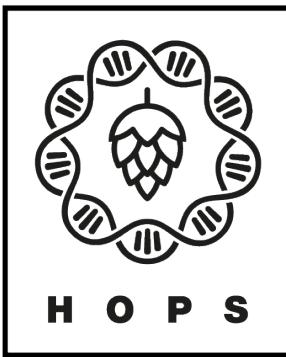
NB: original course material courtesy:
Dr. Antti Karkman, University of Helsinki
Dr. Igor Pessi, Finnish Environment Institute (SYKE)

Typical analysis methods used in metagenomics

1) Alignment:



BWA
stands for
Burrows Wheeler Aligner
 Abbreviations.com



2) Classification:



Centrifuge

MetaPhlan

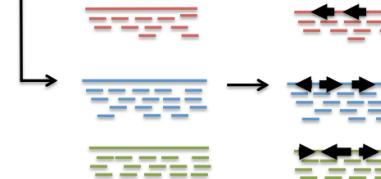
Clark

Reference based:
assume similarity to reference

3) De-novo assembly:



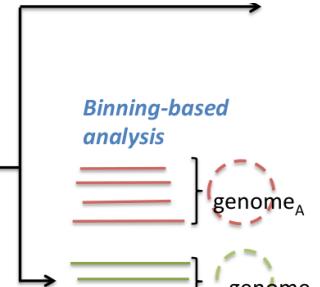
>seq1
GCCGTAGTCC...
>seq2
...



Assembly

Assembly-based
analysis

gene prediction/
annotation



Binning-based
analysis

genome_A
genome_B
genome_C

Phylogenetic binning

Reference free:
unbiased but challenging

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

Wood et al. *Genome Biology* (2019) 20:257
https://doi.org/10.1186/s13059-019-1891-0

Genome Biology

SHORT REPORT

Open Access



Improved metagenomic analysis with Kraken 2

Derrick E. Wood^{1,2}, Jennifer Lu^{2,3} and Ben Langmead^{1,2*}

Abstract

Although Kraken's k-mer-based approach provides a fast taxonomic classification of metagenomic sequence data, its large memory requirements can be limiting for some applications. Kraken 2 improves upon Kraken 1 by reducing memory usage by 85%, allowing greater amounts of reference genomic data to be used, while maintaining high accuracy and speed fivefold. Kraken 2 also introduces a translated search mode, providing increased sensitivity in viral metagenomics analysis.

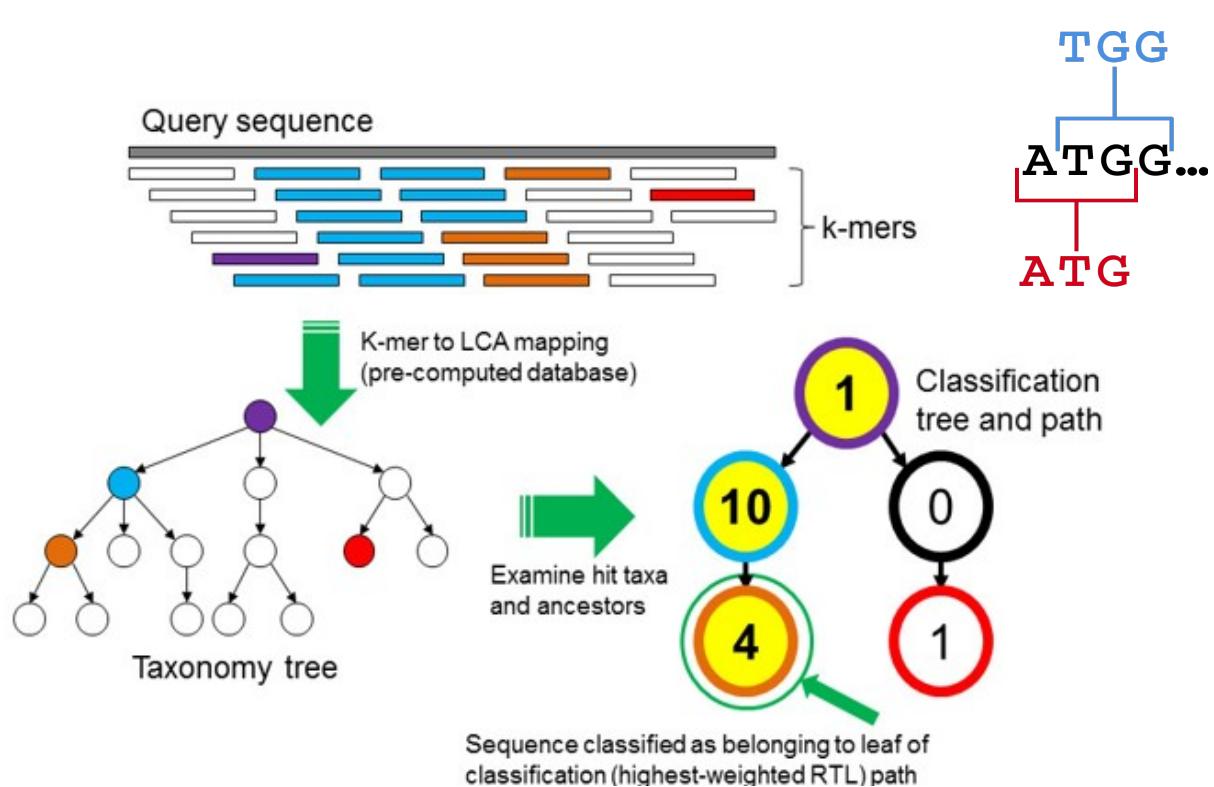
Keywords: Metagenomics, Metagenomics classification, Microbiome, Probabilistic data structures, Alignment-free methods, Minimizers

Assigning taxonomic labels to sequencing reads is an important part of many computational genomics pipelines for metagenomics projects. Recent years have seen several approaches to accomplish this task in a time-efficient manner [1–3]. One such tool, Kraken [4], uses a memory-intensive algorithm that associates short genomic substrings (*k*-mers) with the lowest common ancestor (LCA) taxa. Kraken and related tools like KrakenUniq [5] have proven highly efficient and accurate in independent tool comparisons [6,7]. But Kraken's high memory requirements force many researchers to either use a reduced-sensitivity Minikraken database [8,9] or to build and use many indexes over subsets of the reference sequences [10,11]. Its memory requirements can easily exceed 100 GB [7], especially when the reference data includes large eukaryotic genomes [12,13]. Here, we introduce Kraken 2, which provides a major reduction in memory usage as well as faster classification, a spaced seed searching scheme, a translated search mode for matching in amino acid space, and continued compatibility with the Bracken [14] species-level sequence abundance estimation algorithm.

Kraken 2 addresses the issue of large memory requirements through two changes to Kraken 1's data

structures and algorithms. While Kraken 1 used a sorted list of *k*-mer/LCA pairs indexed by minimizers [15], Kraken 2 introduces a probabilistic, compact hash table to map minimizers to LCAs. This table uses one third of the memory of a standard hash table, at the cost of some specificity and accuracy. Additionally, Kraken 2 only stores minimizers (of length ℓ , $\ell \leq k$) from the reference sequence library in its data structure, whereas Kraken 1 stored all *k*-mers. This change means that, during classification, the minimizer (ℓ -mer) is the substring compared against a reference set in Kraken 2, while Kraken 1 compared *k*-mers (Fig. 1a, b). Kraken 2's index for a specific reference database with 9.1 Gbp of genomic sequences uses 10.6 GB of memory when classifying. Kraken 1's index for the same reference uses 72.4 GB of memory for classification (Fig. 2a, Additional file 1: Table S1). In general, a Kraken 2 database is about 85% smaller than a Kraken 1 database over the same references (Additional file 2: Figure S1).

Kraken 2's approach is faster than Kraken 1's because only distinct minimizers from the query (read) trigger accesses to the hash table. A similar minimizer-based approach has proven useful in accelerating read alignment [16]. Kraken 2 additionally provides a hash-based subsampling approach that reduces the set of minimizer/LCA pairs included in the table, allowing the user to specify a target hash table size; smaller hash tables yield lower memory usage and higher classification



Sequence classified as belonging to leaf of classification (highest-weighted RTL) path

*Correspondence: langmead@jhu.edu

¹Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA

²Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA

Full list of author information is available at the end of the article



Physalia
Courses

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

Same genus

sequence 1 ATGGTCGGCAGGACGTTGCGAGT
sequence 2 CGAGAAGGCAGGACGCCACGTAC

Species 1
Species 2

Ambiguous reads

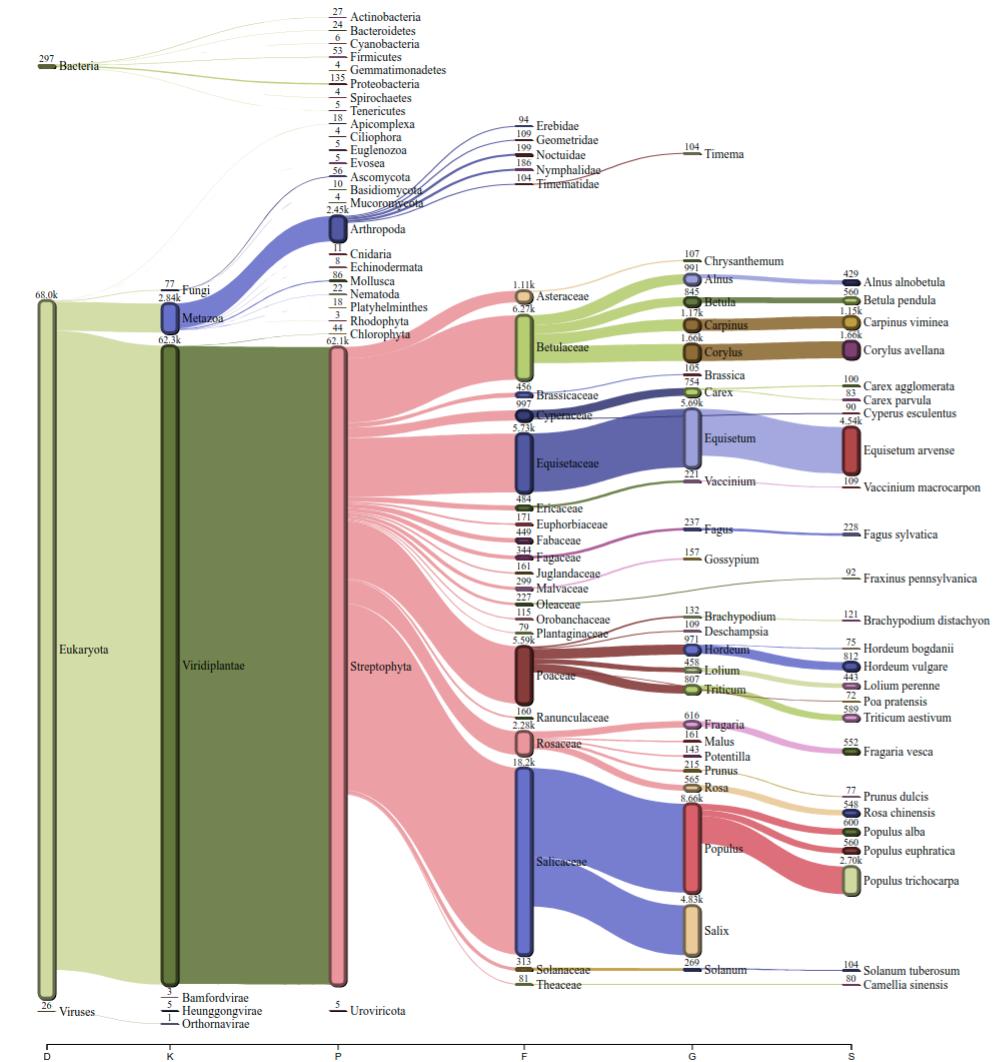
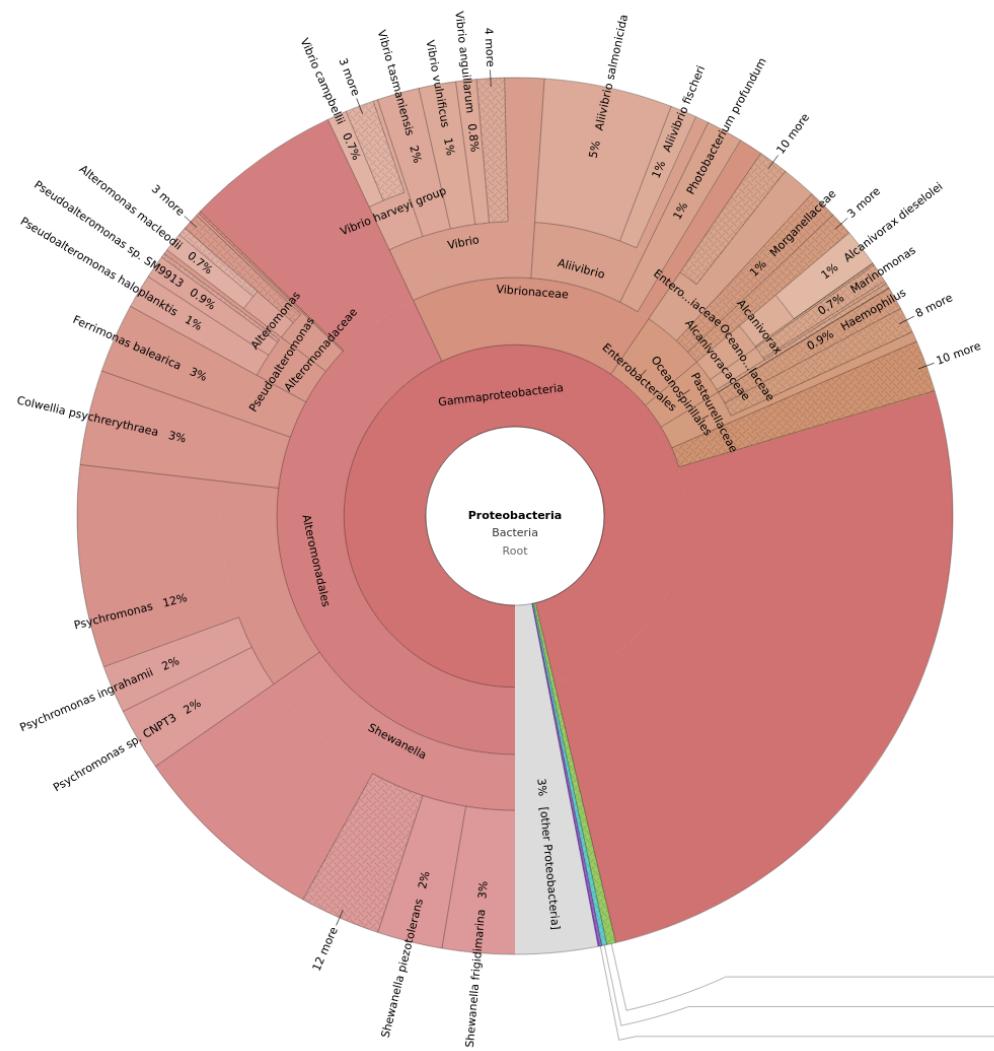
Ignore ambiguous reads
(lose many reads)

Align genus reads
to your species of
interest ref genome
(isn't it too risky?)

Keep them for quantifying
abundance on genus level
(LCA)

Krestovka.merged.trimmed.fastq.gz_kraken2.output - LibreOffice Calc									
File Edit View Insert Format Styles Sheet Data Tools Window Help									
J34									
1	A % of reads	B reads	C taxReads	D n_minimizer	E n_distinct_minimizer	F rank	G taxID	H Name	I J
2	85.76	229988348	229988348	0	0	U	0	unclassified	
3	14.24	38188511	317329	229258315	10956342	R	1	root	
4	14.1	37814928	553677	223347668	10689204	R1	131567	cellular organisms	
5	13.35	35791835	861918	201695602	8612373	D	2	Bacteria	
6	10.84	29060274	1196943	151278879	5085202	P	1224	Proteobacteria	
7	8.22	22056723	407757	106202766	1185104	C	28216	Betaproteobacteria	
8	7.82	20984490	1268537	97499372	1030255	O	80840	Burkholderiales	
9	6	16081257	2224086	68602917	538180	F	80864	Comamonadaceae	
10	2.23	5974652	286945	22944027	90448	G	52972	Polaromonas	
11	1.11	2971020	18648	10967489	38279	S	216465	Polaromonas naphthalenivorans	
12	1.1	2952372	2952372	10917548	38245	S1	365044	Polaromonas naphthalenivorans CJ2	
13	0.95	2560665	1058948	9078851	43621	G1	2638319	unclassified Polaromonas	
14	0.55	1465800	1465800	4741406	23365	S	296591	Polaromonas sp. JS666	
15	0	8327	8327	23317	17	S	1840289	Polaromonas sp. H4N	
16	0	7639	7639	40303	221	S	1840293	Polaromonas sp. H6N	
17	0	7127	7127	35955	154	S	1840301	Polaromonas sp. W10N	
18	0	3429	3429	14648	74	S	1840297	Polaromonas sp. H8N	
19	0	2336	2336	10159	219	S	1840303	Polaromonas sp. W11N	
20	0	1757	1757	7805	272	S	1869339	Polaromonas sp.	
21	0	1251	1251	2288	2	S	1840287	Polaromonas sp. H3N	
22	0	956	956	4451	96	S	1840283	Polaromonas sp. H1N	
23	0	770	770	1317	8	S	416605	Polaromonas sp. A10	
24	0	721	721	1823	1	S	1840281	Polaromonas sp. H12N	
25	0	628	628	712	5	S	2268087	Polaromonas sp. SP1	
26	0	255	255	812	64	S	1840267	Polaromonas sp. E5S	
27	0	245	245	745	49	S	1840265	Polaromonas sp. E3S	
28	0	222	222	1167	24	S	1840239	Polaromonas sp. E10S	
29	0	163	163	405	2	S	480424	Polaromonas sp. GM1	
30	0	20	20	374	1	S	1840257	Polaromonas sp. E19S	
31	0	15	15	51	5	S	642193	Polaromonas sp. RB76	
32	0	15	15	57	20	S	1840275	Polaromonas sp. E9S	
33	0	9	9	28	16	S	1840323	Polaromonas sp. W9N	
34	0	8	8	8	1	S	1705699	Polaromonas sp. 277	

Ways to visualize and interpret taxonomic profilers outputs



How you see false-positives: Kraken2 with NT database

	A	B	C	D	E	F
1	Percent Reads	Reads	TaxReads	Rank	TaxID	Name
2	0.12	6438	6438	S	1009846	Burkholderia cepacia GG4
3	0.09	4683	4683	S	1395570	Burkholderia cepacia JBK9
4	0.09	4685	4685	S	1417228	Paraburkholderia phytofirmans OLGA172
5	0.05	2483	2483	S	1249668	Burkholderia ubonensis MSMB22
6	0.05	2736	2736	S	339670	Burkholderia ambifaria AMMD
7	0.04	2193	2193	S	391038	Paraburkholderia phymatum STM815
8	0.04	2317	2317	S	9615	Canis lupus familiaris
9	0.03	1738	1738	S	754502	Paraburkholderia sprentiae WSM5005
10	0.03	1640	1640	S	1229205	Paraburkholderia phenoliruptrix BR3459a
11	0.03	1664	1664	S	1416914	Pandorea pnomenus 3kgm
12	0.03	1785	1785	S	1112204	Gordonia polysoprenivorans VH2
13	0.02	871	871	S	395019	Burkholderia multivorans ATCC 17616
14	0.02	919	919	S	398577	Burkholderia ambifaria MC40-6
15	0.02	955	955	S	398527	Paraburkholderia phytofirmans PsJN
16	0.02	982	982	S	266265	Paraburkholderia xenovorans LB400
17	0.02	973	973	S	272630	Methylorubrum extorquens AM1
18	0.02	1143	1143	S	223781	Aquila chrysaetos chrysaetos
19	0.02	1192	1192	S	202946	Apteryx mantelli mantelli
20	0.01	755	755	S	406425	Burkholderia cenocepacia MC0-3
21	0.01	623	623	S	32009	Burkholderia gladioli pv. gladioli
22	0.01	581	581	S	999541	Burkholderia gladioli BSR3
23	0.01	772	772	S	1323664	Paraburkholderia caribensis MBA4
24	0.01	536	536	S	264198	Cupriavidus pinatubonensis JMP134
25	0.01	752	752	S	882378	Paraburkholderia rhizoxinica HKI 454
26	0.01	546	546	S	1034807	Flavobacterium branchiophilum FL-15
27	0.01	587	587	S	112262	Ovis canadensis canadensis

Look like reasonable microbes so far ...



This was a blank!

Give me one best metagenomic taxonomic classifier



RESEARCH ARTICLE
Ecological and Evolutionary Science
Check for updates



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,^c Christina Warinner^{c,d,e}

^aPalaeogenomics and Bio-Archaeology Research Network, Research Laboratory for Archaeology and the History of Art, University of Oxford, Oxford, United Kingdom

^bSchool of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom

^cDepartment of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany

^dDepartment of Anthropology, University of Oklahoma, Norman, Oklahoma, USA

^eDepartment 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: QIME/UCLUST, MetaPhAn2, 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. Uniform metrics indicated that both MIDAS and MetaPhAn2 reconstructed the most accurate community structure. QIME/UCLUST, CLARK-S, and MALT had the highest number of inaccurate taxonomic assignments; false-positive rates were highest by CLARK-S and QIME/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 trajectory 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 metagenomic 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

Copyright © 2018 Velsko et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Christina Warinner, warinner@hhmi.org.

* Present address: Irina M. Velsko, Department of Biological Sciences, Clemson University, Clemson, South Carolina, USA.

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

mSystems™ msystems.asm.org 1

Submitted 12 April 2021
Accepted 21 December 2021
Published 24 March 2022

Corresponding authors
Samuel Neuenschwander,
samuel.neuenschwander@unil.ch
Anna-Sapfo Malaspina,
annasapfo.malaspina@unil.ch

Academic editor
Joseph Gillespie

Additional Information and
Declarations can be found on
page 26

DOI 10.7717/peerj.12784

© Copyright
2022 Arizmendi Cárdenas et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study

Yami Ommar Arizmendi Cárdenas^{1,2}, Samuel Neuenschwander^{1,3} and
Anna-Sapfo Malaspina^{1,2}

¹Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

²Swiss Institute of Bioinformatics, Lausanne, Switzerland

³Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland

ABSTRACT

Owing to technological advances in ancient DNA, it is now possible to sequence viruses from the past to track down their origin and evolution. However, ancient DNA data is considerably more degraded and contaminated than modern data making the identification of ancient viral genomes particularly challenging. Several methods to characterise the modern microbiome (and, within this, the virome) have been developed; in particular, tools that assign sequenced reads to specific taxa in order to characterise the organisms present in a sample of interest. While these existing tools are routinely used in modern data, their performance when applied to ancient microbiome data to screen for ancient viruses remains unknown. In this work, we conducted an extensive simulation study using public viral sequences to establish which tool is the most suitable to screen ancient samples for human DNA viruses. We compared the performance of four widely used classifiers, namely Centrifuge, Kraken2, DIAMOND and MetaPhAn2, in correctly assigning sequencing reads to the corresponding viruses. To do so, we simulated reads by adding noise typical of ancient DNA to a set of publicly available human DNA viral sequences and to the human genome. We fragmented the DNA into different lengths, added sequencing error and C to T and G to A deamination substitutions at the read termini. Then we measured the resulting sensitivity and precision for all classifiers. Across most simulations, more than 228 out of the 233 simulated viruses were recovered by Centrifuge, Kraken2 and DIAMOND, in contrast to MetaPhAn2 which recovered only around one third. Overall, Centrifuge and Kraken2 had the best performance with the highest values of sensitivity and precision. We found that deamination damage had little impact on the performance of the classifiers, less than the sequencing error and the length of the reads. Since Centrifuge can handle short reads (in contrast to DIAMOND and Kraken2 with default settings) and since it achieves the highest sensitivity and precision at the species level across all the simulations performed, it is our recommended tool. Regardless of the tool used, our simulations indicate that, for ancient human studies, users should use strict filters to remove all reads of potential human origin. Finally, we recommend that users verify which species are present in the database used, as it might happen that default databases lack sequences for viruses of interest.

How to cite this article Arizmendi Cárdenas YO, Neuenschwander S, Malaspina AS-S. 2022. Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study. *PeerJ* 10:e12784 <https://doi.org/10.7717/peerj.12784>



Article

Benchmarking Metagenomic Classifiers on Simulated Ancient and Modern Metagenomic Data

Vaidehi Pusadkar^{1,2} and Rajeev K. Azad^{1,2,3,*}

¹ Department of Biological Sciences, University of North Texas, Denton, TX 76203, USA;
[vaidehipusadkar@unt.edu](mailto:vaidheipusadkar@unt.edu)

² BioDiscovery Institute, University of North Texas, Denton, TX 76203, USA

³ Department of Mathematics, University of North Texas, Denton, TX 76203, USA

* Correspondence: rajeevazad@unt.edu

Abstract: Taxonomic profiling of ancient metagenomic samples is challenging due to the accumulation of specific damage patterns on DNA over time. Although a number of methods for metagenome profiling have been developed, most of them have been assessed on modern metagenomes or simulated metagenomes mimicking modern metagenomes. Further, a comparative assessment of metagenome profilers on simulated metagenomes representing a spectrum of degradation depth, from the extremity of ancient (most degraded) to current or modern (not degraded) metagenomes, has not yet been performed. To understand the strengths and weaknesses of different metagenome profilers, we performed their comprehensive evaluation on simulated metagenomes representing human dental calculus microbiome, with the level of DNA damage successively raised to mimic modern to ancient metagenomes. All classes of profilers, namely, DNA-to-DNA, DNA-to-protein, and DNA-to-marker comparison-based profilers were evaluated on metagenomes with varying levels of damage simulating deamination, fragmentation, and contamination. Our results revealed that, compared to deamination and fragmentation, human and environmental contamination of ancient DNA (with modern DNA) has the most pronounced effect on the performance of each profiler. Further, the DNA-to-DNA (e.g., Kraken2, Bracken) and DNA-to-marker (e.g., MetaPhAn2) based profilers approached showed complementary strengths, which can be leveraged to elevate the state-of-the-art of ancient metagenome profiling.



Citation Pusadkar V, Azad RK. 2022. Benchmarking Metagenomic Classifiers on Simulated Ancient and Modern Metagenomic Data. *Microorganisms* 11: 2478. <https://doi.org/10.3390/microorganisms11102478>

Academic Editor: Remnath Tan

Received: 18 July 2023

Revised: 26 September 2023

Accepted: 29 September 2023

Published: 2 October 2023



Copyright © 2022 by the author. License MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/>). 4.0).

Microorganisms 2023, 11, 2478. <https://doi.org/10.3390/microorganisms11102478>

<https://www.mdpi.com/journal/microorganisms>

Are we comparing classifiers or filtering strategies of their outputs?



Why is it important to do filtering of Kraken output?

% reads	taxReads	kmers	dup	cov	taxID	rank	taxName
87.090	222111	222111	3577712	1.00	9.593e-03	0	no rank unclassified
12.910	32934	331	329116	1.17	3.026e-06	1	no rank root
12.780	32603	3	324663	1.16	3.032e-06	131567	no rank cellular organisms
7.002	17859	7	111728	1.43	1.388e-06	2759	superkingdom Eukaryota
6.998	17849	0	111664	1.43	1.729e-06	33154	clade Opisthokonta
6.998	17847	0	111664	1.43	1.918e-06	33208	kingdom Metazoa
6.998	17847	0	111664	1.43	1.919e-06	6072	clade Eumetazoa
6.998	17847	0	111664	1.43	1.935e-06	33213	clade Bilateria
6.996	17844	0	111664	1.43	2.365e-06	33511	clade Deuterostomia
6.996	17844	0	111664	1.43	2.392e-06	7711	phylum Chordata
6.996	17844	0	111664	1.43	2.399e-06	89593	subphylum Craniata
6.996	17844	0	111664	1.43	2.399e-06	7742	clade Vertebrata
6.996	17844	1	111664	1.43	2.402e-06	7776	clade Gnathostomata
6.996	17843	0	111640	1.43	2.410e-06	117570	clade Teleostomi
6.996	17843	12	111640	1.43	2.410e-06	117571	clade Euteleostomi
6.973	17785	0	111127	1.43	4.908e-06	8287	superclass Sarcopterygii
6.973	17785	0	111127	1.43	4.920e-06	1338369	clade Dipnotetrapodomorpha
6.973	17785	3	111127	1.43	4.920e-06	32523	clade Tetrapoda
6.972	17781	47	111095	1.43	5.185e-06	32524	clade Amniota
6.945	17714	1	110030	1.44	7.432e-06	40674	class Mammalia

rank = species



% reads	taxReads	kmers	dup	cov	taxID	rank	taxName
4.7660000	12155	5002	72665	1.45	1.342e-03	9785	species Loxodonta africana
0.0019600	5	0	48	1.00	1.719e-03	99490	species Loxodonta cyclotis
0.0337200	86	0	335	1.19	3		Elephas maximus
0.0003921	1	0	8	1.00	3		Mammuthus primigenius
0.3937000	1004	9	3984	1.08	1		Trichechus manatus
0.0172500	44	0	130	1.30	1		Dugong dugon
0.1380000	352	25	1388	1.14	6		Procavia capensis
0.0003921	1	0	4	1.00	3		Dendrohyrax arboreus
0.0200000	51	3	188	1.14	7		Echinops telfairi
0.0047050	12	0	33	1.21	1		Orycteropus afer
0.0019600	5	1	38	1.00	1		Chrysocloris asiatica
0.0003921	1	0	6	1.00	1		Amblysomus hottentotus
0.0011760	3	0	12	1.00	4		Elephantulus edwardii
0.0003921	1	1	2	1.00	3.333e-04	237658	species Petrosaltator rozeti
0.1313000	335	3	1351	1.01	8.105e-07	9612	species Canis lupus
0.0294100	75	2	277	1.06	5.837e-07	9657	species Lutra lutra
0.0007842	2	0	10	1.00	7.836e-07	34882	species Enhydra lutris
0.0007842	2	1	10	1.00	8.168e-07	76717	species Lontra canadensis
0.0019600	5	0	16	1.00	8.997e-07	9715	species Mirounga leonina
0.0011760	3	1	10	1.00	4.989e-07	9720	species Phoca vitulina

kmers = 1000

taxReads = 200



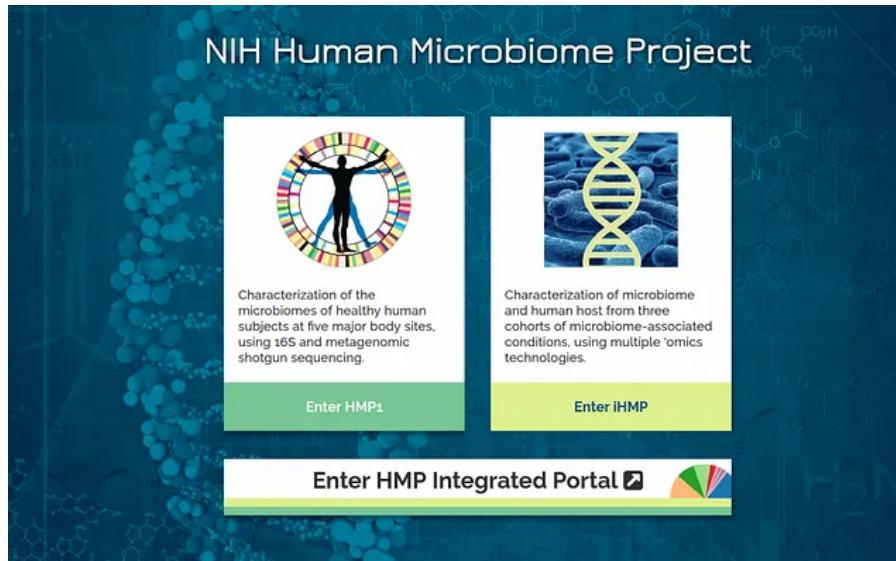
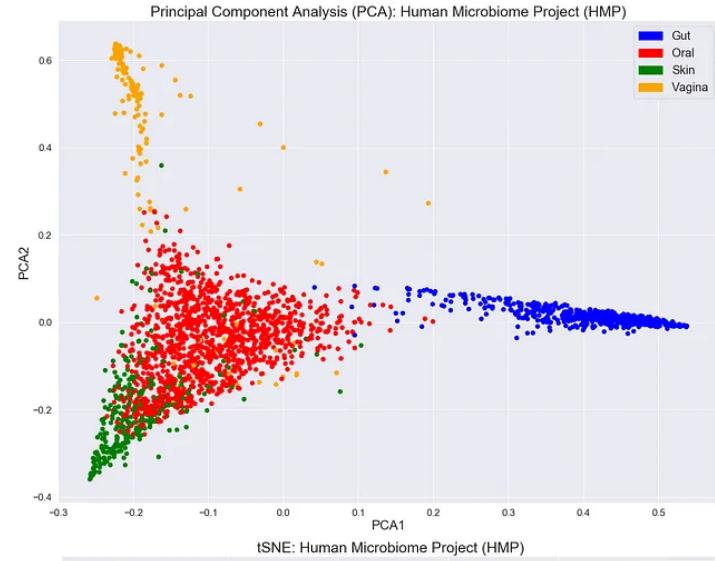
% reads	taxReads	kmers	dup	cov	taxID	rank	taxName
4.7660	12155	5002	72665	1.45	1.342e-03	9785	species Loxodonta africana
0.3937	1004	9	3984	1.08	1.020e-04	9778	species Trichechus manatus
0.1380	352	25	1388	1.14	6.428e-04	9813	species Procavia capensis
0.1313	335	3	1351	1.01	8.105e-07	9612	species Canis lupus
0.1686	430	430	1794	1.00	1.212e-06	9606	species Homo sapiens
5.1680	13180	12862	178949	1.01	5.993e-03	1491	species Clostridium botulinum

Filtering is critical for understanding content of metagenomic samples!

% reads	taxReads	kmers	dup	cov	taxID	rank	taxName
4.7660	12155	5002	72665	1.45	1.342e-03	9785	species Loxodonta africana
0.1686	430	430	1794	1.00	1.212e-06	9606	species Homo sapiens
5.1680	13180	12862	178949	1.01	5.993e-03	1491	species Clostridium botulinum

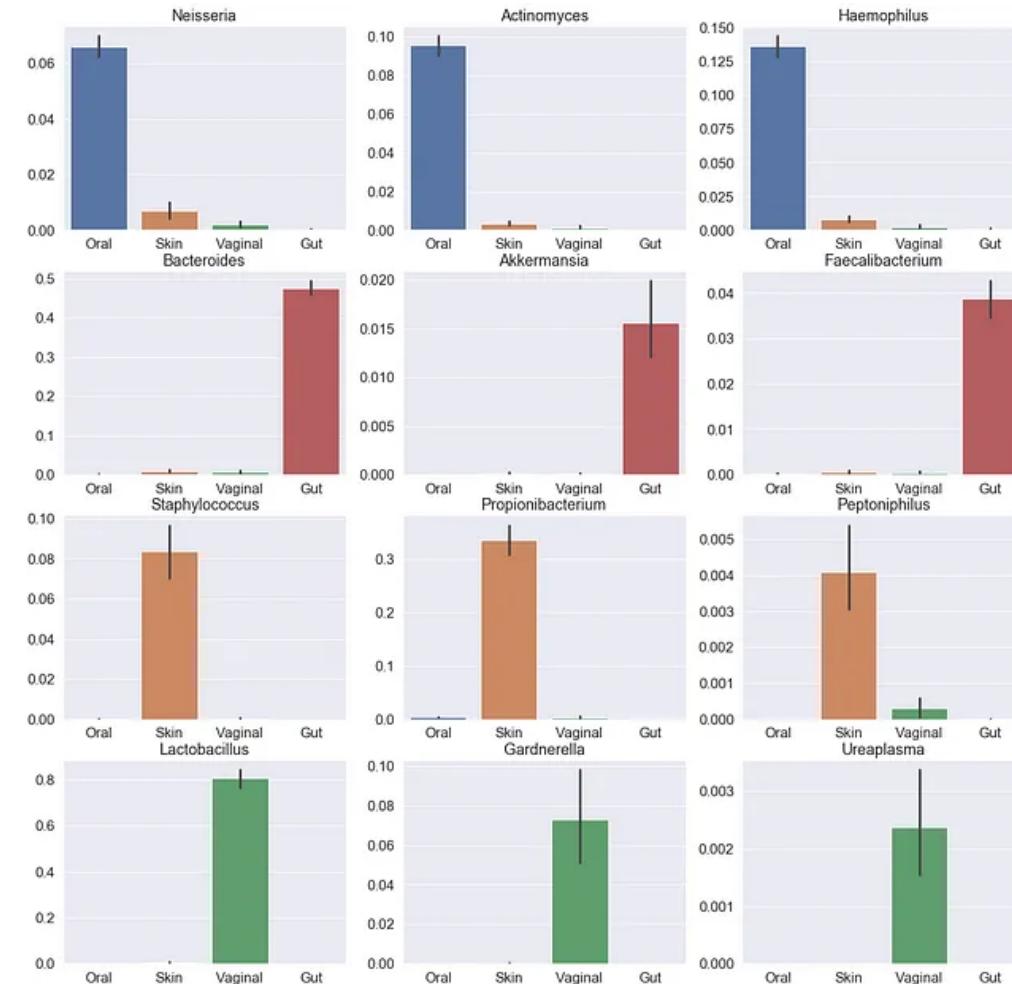
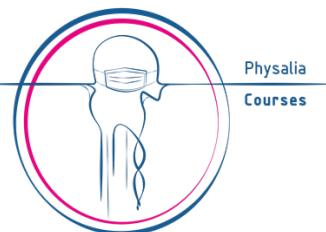
How can we use microbial abundance matrix from Kraken?

	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10
Ralstonia solanacearum	3628	3751	619	1804	1384	1608	1375	0	1112	749
Mycobacterium avium	8236	8546	273	265	3221	4808	7750	6382	0	0
Burkholderia pseudomallei	7095	0	0	0	13082	0	4885	1456	0	7310
Salmonella enterica	4356	4471	4205	3588	0	13854	0	0	1959	2560
Pseudomonas chlororaphis	296	1024	0	977	374	677	276	0	0	294
Neisseria meningitidis	465	502	0	0	7341	0	0	3268	5643	0
Yersinia pestis	0	0	0	7174	957	0	11485	6461	0	11553

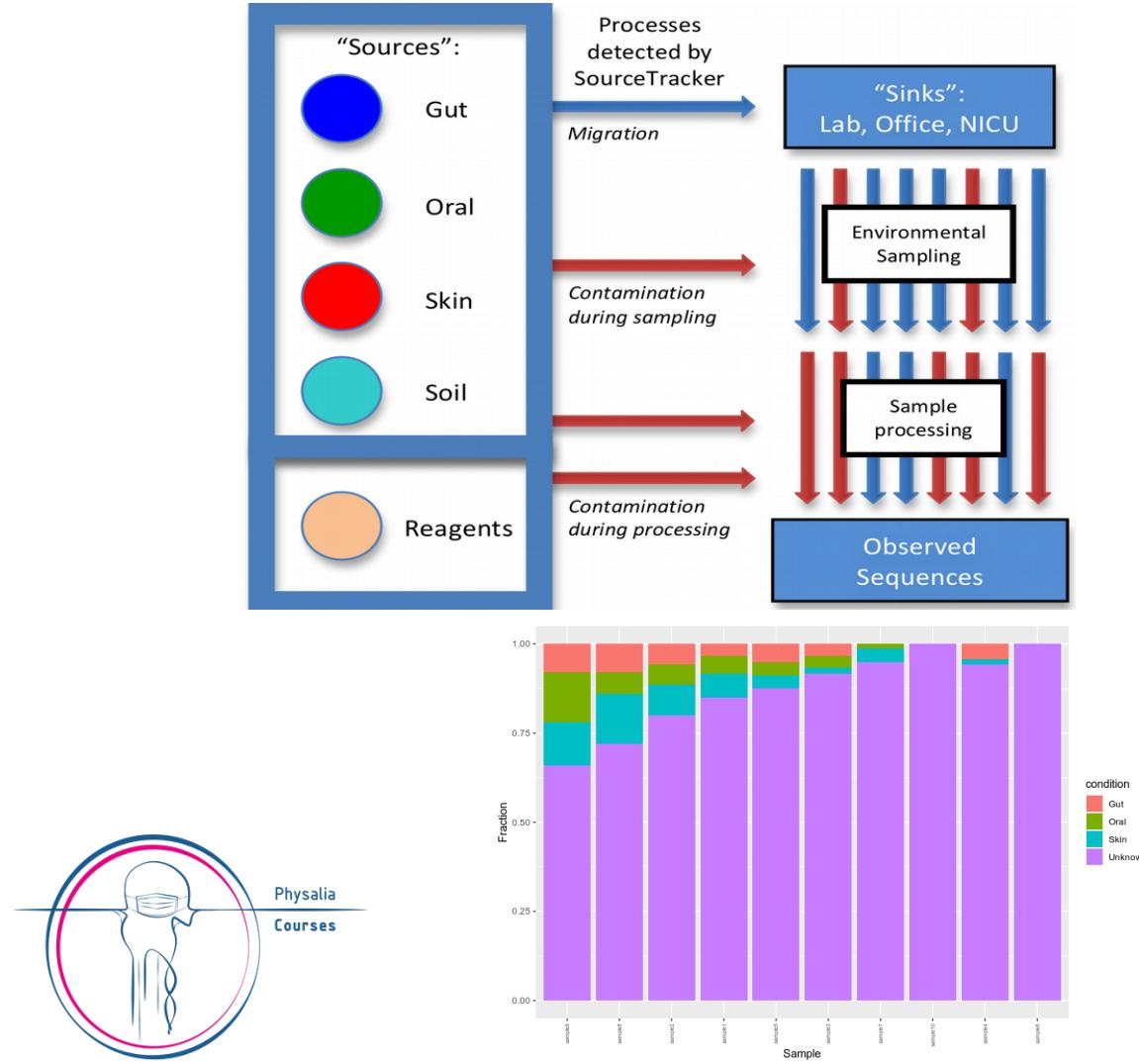


Understanding marker microbes for different environments

	Oral	Gut	Skin	Vagina
0	Neisseria	Blautia	Staphylococcus	Mobiluncus
1	Veillonella	Faecalibacterium	Peptoniphilus	Sphingopyxis
2	Actinomyces	Bacteroides	Citrobacter	Ureaplasma
3	Haemophilus	Dorea	Enhydrobacter	Caulobacter
4	Rothia	Akkermansia	Finegoldia	Gardnerella
5	Leptotrichia	Clostridium	Propionibacterium	Chlamydia
6	Cardiobacterium	Ruminococcus	Acinetobacter	Asticcacaulis
7	Capnocytophaga	Subdoligranulum	Massilia	Mycobacterium
8	Oribacterium	Oxalobacter	Hymenobacter	Herbaspirillum
9	Alloprevotella	Oscillibacter	Corynebacterium	Lactobacillus
10	Gemella	Eubacterium	Bacillus	Achromobacter
11	Fusobacterium	Bilophila	Micrococcus	Atopobium



Microbial source tracking



Duitama González et al. *Microbiome* (2023) 11:243
<https://doi.org/10.1186/s40168-023-01670-3>

Microbiome

Open Access



METHODOLOGY

decOM: similarity-based microbial source tracking of ancient oral samples using k-mer-based methods

Camila Duitama González^{1*}, Riccardo Vicedomini^{1,2}, Téo Lemane², Nicolas Rascovan³, Hugues Richard⁴ and Rayan Chikhi¹

Abstract

Background The analysis of ancient oral metagenomes from archaeological human and animal samples is largely confounded by contaminant DNA sequences from modern and environmental sources. Existing methods for Microbial Source Tracking (MST) estimate the proportions of environmental sources, but do not perform well on ancient metagenomes. We developed a novel method called decOM for Microbial Source Tracking and classification of ancient and modern metagenomic samples using k-mer matrices.

Results We analysed a collection of 360 ancient oral, modern oral, sediment/soil and skin metagenomes, using stratified five-fold cross-validation. decOM estimates the contributions of these source environments in ancient oral metagenomic samples with high accuracy, outperforming two state-of-the-art methods for source tracking, FEAST and mSourceTracker.

Conclusions decOM is a high-accuracy microbial source tracking method, suitable for ancient oral metagenomic data sets. The decOM method is generic and could also be adapted for MST of other ancient and modern types of metagenomes. We anticipate that decOM will be a valuable tool for MST of ancient metagenomic studies.

Keywords Ancient metagenomics, Microbial source tracking, k-mer matrix, Paleogenomics

