Algorithms based on k-mers for ancient oral metagenomics : Tools for contamination removal and assessment in palaeometagenomics

Camila Duitama González

Algorithms based on k-mers for ancient oral metagenomics

Tools for contamination removal and assessment in palaeometagenomics

Camila Duitama González

Thesis submitted for the degree of Philosophiae Doctor

Ecole Doctorale Informatique, Télécommunications et Electronique EDITE (ED130)

Sorbonne Université

Members of the jury :

Dr. Pierre Peterlongo	IRISA, Université de Rennes	President of the jury and reviewer
Dr. Antonio Fernández Guerra	University of Copenhagen	Reviewer
Dr. Prof. Nataliya Sokolovska	Sorbonne Université	Examiner
Dr. Nicolas Rascovan	Institut Pasteur	Invited external member
Dr. Rayan Chikhi	Institut Pasteur	Supervisor
Dr. Hugues Richard	Robert Koch Institute	Supervisor
Dr. Riccardo Vicedomini	Université de Rennes	Supervisor

2024

Series of dissertations submitted to the
Faculté d'Informatique, Sorbonne Université

Print production: Institut Pasteur.

A mi familia y a mis amigas. A quienes no trabajan en ciencia y también contribuyeron a que esta tesis se llevara a cabo.

Preface

This thesis is submitted in partial fulfilment of the requirements for the degree of *Philosophiae Doctor* at the Sorbonne Université. The research presented here was conducted at the Institut Pasteur, under the supervision of Dr. Rayan Chikhi, Dr. Hugues Richard and Dr. Riccardo Vicedomini. This work was supported by the Paris Artificial Intelligence Research Institute (PRAIRIE) Grant ANR-19-P3IA-0001.

The thesis is a collection of two papers, presented in chronological order of writing. I begin with an introduction describing the research output of my PhD, the outreach of my work and the outline of this manuscript. I continue with an introductory chapter that describes some basic biological concepts in the field of palaeomicrobiology. Subsequently, I present the papers I co-wrote during my thesis. Each paper is preceded by a state-of-the-art chapter that provides background and motivation to understand them. The thesis ends with a chapter regarding perspectives and another chapter regarding conclusions.

I am the first author in both of the papers.

Acknowledgements

I would like to thank Rayan for his constant encouragement, despite the issues that came up during these last three years, such as the emotional difficulties that we all faced as the work was done partially during a pandemic. I am sure I could not have completed any of the achievements of this PhD without his supervision, patience and support. I thank Hugues for his continuous motivation, patient explanations and the will to discuss science every other week. I am aware that there are not many women in this field, but I believe that with the persisting support of male supervisors like yourself more (young) women will emerge as leaders in this domain.

I want to express my gratitude towards Riccardo for his advice during our meetings. I am immensely grateful for the opportunity to collaborate with Sam and Téo, working with them was one of the highlights of my PhD and an exceptionally fruitful experience.

I extend my heartfelt thanks to Nataliya Sokolovska, Pierre Peterlongo, Antonio Fernández Guerra and Nicolas Rascovan for their participation as members of the thesis jury. In particular, I am grateful to Nicolas for suggesting the topic of aDNA decontamination and for imparting his passion on the subject.

I also want to acknowledge the support of the SeqBio group. Special thanks to Melanie who consistently and kindly assisted with the bureaucracy necessary to deal with everyday life in Paris. Additionally, I want to thank Yoann for his will to listen, discuss and engage in responsibilities that go beyond his work duties.

Lastly, I express my gratitude to the members of the SPAAM community who attentively listened to my research, provided valuable suggestions, and conducted testing on the methods I developed during my thesis.

Finalmente quiero agradecer a mi familia.

A mis padres, a quienes admiro y amo profundamente, que me enseñaron el valor de cuestionar y cuestionarse, que con su ejemplo incentivaron en mi el deseo de buscar la coherencia, la pasión y la empatía en todo lo que se hace. Mi deseo mas profundo es que mi vida honre la suya.

A mi hermana, a quien tuve que dejar de disfrutar estos años en los que me fui del país, pero que siempre estuvo a mi lado acompañándome y aconsejándome.

Soy nieta de una abuela que no tuvo accesso a la educación secundaria, hija de una madre que trabajó, maternó y estudió toda su vida hasta doctorarse y ahora yo logro, con muchísimas ventajas y ayudas, acceder a un doctorado también a pesar de las dificultades que implica migrar y ser mujer en un campo tan masculino. Sus cuidados amorosos son el ejemplo de que "los logros de las mujeres ejemplares de la historia existen gracias a todas las otras que están por debajo de ellas resolviendo las tareas necesarias para la reproducción de la vida".

No obstante, en mi caso excepcional, es cierto que el apoyo irrestricto de un padre amorosísimo me permitió dedicarme con confianza a esta área del conocimiento, aún si el mundo no fue hecho para que las mujeres disfrutemos y construyamos la ciencia, la ingeniería y la tecnología. Esta tesis es para vos mi pa.

A mis amigues, de tantas latitudes, que me cuidan y hacen de la vida un deleite. Ustedes me mostraron la forma de amor mas libre.

To my friends, from so many latitudes, who take care of me and make my life a delight. You showed me the freest form of love.

Camila Duitama González
Paris, February 2024

List of Papers

Chapter 6

Duitama González, C., Lemane, T., Vicedomini, R., Rascovan, N. ,Richard, H. and Chikhi, R. “decOM: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods”. In: *Microbiome*. Vol. 3, no. 2 (2011), pp. 123–456. DOI: 10.1000/182.

Chapter 8

Duitama González, C., Rangavittal, S., Vicedomini, R., Richard, H. and Chikhi, R. “aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets”. In: *iScience*. Vol. 26, no. 11 (2023), pp. 123–456. DOI: 10.1000/182.

Contents

List of Figures

List of Tables

Chapter 1

Summary

1.1 English summary

Palaeometagenomics is the study of ancient genetic material by using metagenomic sequencing, a process that entails the characterisation of the DNA from all the organisms in a sample. By ancient genetic material we refer to the DNA that comes from a non-living source and that shows signs of molecular degradation. Dental calculus has proven to be an exceptionally rich source of ancient DNA (aDNA) and it has been used to investigate the evolution of the oral microbiome, as well as human oral health and diet. Despite the establishment of rigorous laboratory protocols for aDNA contamination control, aDNA samples are still highly susceptible to contamination from environmental sources, which can drastically alter the microbial composition and lead to erroneous conclusions after downstream analyses. This dissertation proposes two algorithms that rely on k-mers (sub-sequences of DNA) to address two relevant challenges in the field of palaeometagenomics: contamination assessment via Microbial Source Tracking and contamination removal at the read level. The former task resulted in a first-author publication and an open-software called `decOM`, while the latter has also been published as a first-author paper accompanied by an open-software called `aKmerBroom`. Both methods were tested on ancient oral metagenomic data, yet their utility can be extended to samples that do not originate from ancient oral sources. Overall, this thesis has proven that k-mer-based algorithms have an immense potential for contamination removal and contamination assessment of metagenomes, as they leverage the wealth of metagenomic information that has been sequenced and made publicly available throughout the years.

1.2 Résumé en français

La paléométagénomique est l'étude du matériel génétique ancien à l'aide du séquençage métagénomique, un processus qui implique la caractérisation de l'ADN de tous les organismes d'un échantillon. Par matériel génétique ancien, nous entendons l'ADN provenant d'une source non vivante et présentant des signes de dégradation moléculaire. Le tartre dentaire s'est révélé être une source exceptionnellement riche d'ADN ancien et a été utilisé pour étudier l'évolution du microbiome buccal, ainsi que la santé bucco-dentaire et l'alimentation de l'homme. Malgré la mise en place de protocoles de laboratoire rigoureux pour le contrôle de la contamination de l'ADN ancien, les échantillons d'ADN ancien court sont encore très sensibles à la contamination par des sources environnementales, ce qui peut modifier radicalement la composition microbienne et conduire à des conclusions erronées après les analyses en aval. Cette thèse propose deux algorithmes qui s'appuient sur les k-mers (sous-séquences d'ADN) pour relever deux défis importants dans le domaine de la paléométagénomique : l'évaluation de la contamination via le suivi des sources microbiennes et l'élimination de la contamination au niveau des lectures. La première tâche a donné lieu à une publication en première auteure et à un logiciel ouvert appelé `decOM`, tandis que la seconde a également été publiée en tant qu'article du première auteure accompagné d'un logiciel ouvert appelé `aKmerBroom`. Les deux méthodes ont été testées sur des données métagénomiques orales anciennes, mais leur utilité peut être étendue à des échantillons qui ne proviennent pas de sources orales anciennes. Dans l'ensemble, cette thèse a prouvé que les algorithmes basés sur k-mer ont un immense potentiel pour l'élimination de la contamination et l'évaluation de la contamination des métagénomes, car ils tirent parti de la richesse des informations métagénomiques qui ont été séquencées et mises à la disposition du public au fil des ans.

Chapter 2

Introduction

This thesis is the result of my years at the Institut Pasteur, where I started working in October 2020. During my time at the Sequence Bioinformatics group and under the supervision of Dr. Rayan Chikhi and Dr. Hugues Richard I have worked developing bioinformatic tools that use k-mers for the following tasks:

- Contamination assessment via Microbial Source Tracking of ancient oral samples.
- Contamination removal at the read level of ancient oral samples.

2.1 Research output

During this thesis, my work on developing tools for the analysis of ancient metagenomic data resulted in two publications:

- A first author article describing our method for contamination assessment González, C. D. et al. "decOM: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods". In: *Microbiome* vol. 11, no. 1 (2023), pp. 243–243.
- A first author article describing our method for contamination removal González, C. D. et al. "aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets". In: *iScience* vol. 26, no. 11 (2023).

2.2 Presentation and posters

- Learning Meaningful Representations for Life (LMRL) Workshop in NeurIPS 2022.
- Proceedings track paper *"aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets"* in RECOM-Seq 2023.

2.3 Outreach

- decOM was used in the Ancient Metagenomics Summer School 2023 Fellows Yates, J. A. et al. *Introduction to Ancient Metagenomics.* 2022. DOI: 10.5281/zenodo.8027281.
- decOM was cited in Fernandez-Guerra, A. et al. "A 2-million-year-old microbial and viral communities from the Kap København Formation in North Greenland". In: *bioRxiv* (2023), pp. 2023–06.
- decOM was cited in a poster by Maria Lopopolo at the 10^{th} Meeting of the ISBA entitled *"New Horizons in Biomolecular Archaeology"*.
- decOM was cited in the talk presented for SPAAM5 *"Biomolecular perspectives on the uses of birch bark tar in prehistoric Europe"* by Anna White.

2.4 Outline

This thesis is organised as follows:

Chapter 2 An introduction describing the outline of the thesis.

Chapter 3 An introduction of some basic concepts in Palaeomicrobiology and ancient metagenomics.

Chapter 4 An overview to chapters 5 (State of the art: Ancient Microbial Source Tracking) and 7 (State of the art: Ancient reads decontamination), that correspond to the chapters dedicated to explain the concepts required to understand the two papers produced in this thesis. Here I briefly introduce what is the methodological gap that this thesis addresses, with respect to current methods for contamination assessment and contamination removal of ancient oral metagenomes.

Chapter 5 First part of state of the art dedicated to explain the bioinformatics concepts required to understand the paper in Chapter 6.

Chapter 6 `decOM`: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods.

Chapter 7 Second part of state of the art dedicated to explain the bioinformatics concepts required to understand the paper in Chapter 8.

Chapter 8 `aKmerBroom`: Ancient oral DNA decontamination using Bloom filters on k-mer sets.

Chapter 9 Perspectives

Chapter 10 Conclusions

Chapter 3

Palaeomicrobiology

3.1 Microbes, microbiome and the history of palaeomicrobiology

3.1.1 Brief introduction and definitions

Microorganisms, also known as *microbes*, are a diverse range of unicellular organisms that span all three domains of life: bacteria, archaea, and a considerable proportion of eukaryotes [1]. Microbes are estimated to constitute half of the earth's total biomass [2]. Contrary to common beliefs, most microbes are not pathogenic, instead, the vast majority of them are either neutral or beneficial [1]. The earliest known forms of life on Earth were microbes [3]. Fossilised remains of some of these early forms of life were found in Western Australia and are estimated to be older than 3.4 billion years [4].

The *microbiome* refers to the biotic[1] and abiotic[2] factors of an environment, that is to say: the entire habitat, the organisms in it, their genomes and the surrounding environmental conditions [7]. The human microbiome is composed of the microorganisms of the human body, and that includes a massive number of bacteria. Some interesting facts regarding the bacteria in the human microbiome are that the number of bacterial cells (10^{14}) exceeds the number of human cells (10^{13}) [8]. The number of unique bacterial genes in our *"accessory genome"*[3]($\sim 3,300,000$) exceeds the number of our genes ($\sim 22,000$) [11], and the total weight of the bacteria in our microbiome makes up to 1.2 kg which is almost the size of the human brain (1.4 kg) [12].

3.1.2 History of palaeomicrobiology or the study of ancient microbes

3.1.2.1 Pre-Next Generation Sequencing (NGS) era

Throughout the history of palaeontology, chemical and microscopic techniques have been used to study microbial fossils. Microscopic techniques were first used in 1977 to test Gram staining of bacteria coming from fossilised faeces [13]. Concurrently, the first chemical techniques that were used in the field of palaeomicrobiology date to 1978 [14], when they tried to indirectly detect microbes by chasing by-products of their metabolism.

As both detection methods evolved, the field of genetics grew too. The first attempt to detect and identify ancient microbes using genetic techniques occurred in 1984 when the first short DNA fragments were extracted and sequenced from a museum sample coming from an extinct species of zebra [15]. This event is considered the birth of ancient DNA (aDNA) and it was a pivotal moment. It was the first time that humans were able to study genetic information from the past, creating huge expectations about the possibilities researchers had to understand evolution and the origins of the modern world [16]. Later on, the fields of medicine and biology were revolutionised with the appearance of the Polymerase Chain Reaction (PCR), a laboratory technique for rapidly amplifying a fragment of DNA [17]. Shortly after, protocols using PCR were introduced in the field of palaeomicrobiology, when in 1988 mitochondrial DNA sequences from a 7000-year-old brain were amplified by PCR and sequenced [18]. During the upcoming years, subsequent studies reported several prominent species coming from ancient samples using PCR amplification such as *Saccharomyces cerevisiae* [19], *Mycobacterium leprae* [20], *Yersinia pestis* [21], among others. However, problems with reproducibility [22, 23] cast doubt on the PCR technique as a method accurate enough to prove the presence of a species in a given ancient sample. Moreover, it was the first time that issues regarding the ubiquitous contamination and challenges in aDNA validation were raised, questioning the integrity of the conclusions reached in the field [1, 24, 25]. Nonetheless, these first attempts at genetic characterisation of ancient material were a breaking point in the study of palaeomicrobiology [24].

[1] Biotic refers to things related to or involving living organisms [5]

[2] Abiotic or abiological refers to non-biotic [6]

[3] The genes that are common to all strains in a population and are involved in essential functions for survival compose the *core genome*. In contrast to these conserved regions, the *accessory genome* is the portion of the genome that is variably present between individual strains, thus different strains might have different sets of accessory genes [9, 10].

3.1.2.2 Next Generation Sequencing (NGS) era

After the completion of the Human Genome Project (HGP), a landmark global scientific effort to produce the first sequence of the human genome, it became evident that there was a demand for high-throughput sequencing technologies at a reduced cost. This necessity induced the advent of Next Generation Sequencing (NGS) technologies, a term that was coined in the mid-2000s with the remarkable 50,000-fold reduction of the expense of sequencing the human genome since the HGP [26]. NGS allowed to perform "massively parallel" sequencing, producing billions of reads at decreasing costs [27]. Such technologies were ideal for the study of ancient genetic material, as they could handle and recover DNA that was ultrashort and preserve molecular data despite degradation, in a way that could not be done simply by using PCR [16, 28]. The first major breakthrough using this technology in the field was the reconstruction of a complete *Yersinia pestis* genome [29, 30]. Since then, numerous ancient genomes have been sequenced such as *Mycobacterium tuberculosis* (tuberculosis) [31, 32], *Mycobacterium leprae* (leprosy) [33, 34], *Salmonella enterica* (enteric fever) [35, 36, 37], *Saccharomyces cerevesiae* (budding yeast) [38], among many others.

3.2 Metagenomics

Next Generation Sequencing prompted the study of individual ancient microbial species but also enabled the study of entire microbial communities or microbiomes. A *metagenome* is defined as the collection of genomes and genes from the members of a *microbiota*, that is, the DNA recovered from the assemblage of microorganisms existing in a certain environment [7, 39]. Subsequently, *metagenomics* is a term used to describe the process to characterise a data set that includes nucleic acid sequences from all organisms in a sample, in order to gain information on the potential function of the microbiota [39, 40]. This means that, instead of focusing on a single target gene, species or genome, the entire biotic content of a sample (including bacteria, archaea, eukaryotes and viruses) is sequenced by randomly amplifying and sequencing a subset of the total DNA in the sample. This provides a more complete characterisation of the microbiome by analysing all domains simultaneously. Metagenomics as a technology is especially useful, as it overcomes the problems of primer bias and generates whole genome sequencing data; therefore analyses are not limited to questions of taxonomy or phylogeny, but are extended to questions regarding the functionality of the genes present in a sample [41]. Furthermore, shotgun metagenomics allows for the retrieval of microbial genomes without an existing reference and alleviates the challenging task of growing ancient microbes in a laboratory setting. In contrast, some of the disadvantages of metagenomics is that it presents challenges such as the need for computationally intensive algorithms, and requires pipelines that have not yet been widely accepted in the community, part of which may be biased by the use of reference databases [42]. A thorough database for ancient metagenomic studies is the AncientMetagenomeDir (`https://github.com/SPAAM-community/AncientMetagenomeDir`)[43].

3.3 Ancient metagenomics

Genetic material that comes from an organism or tissue that is older than an arbitrary cutoff of 100 years old is also referred to as ancient DNA (aDNA) [44]. Other authors prefer to consider aDNA as any DNA coming from a non-living source that shows signs of molecular degradation [39]. Ancient metagenomics is the use of shotgun metagenomics on ancient genetic material. The genomic information that can be generated using metagenomics is highly suitable for aDNA research, since metagenomic sequencing is not affected by length or sequence variants (unlike PCR) and it is not as sensitive to very short and degraded DNA fragments [1].

3.3.1 Sources of aDNA

As mentioned before, microorganisms are ubiquitous, yet five main sources of ancient microbes that are particularly informative: teeth, bones, palaeofeces, cultural artifact residues, and sediments.

3.3.1.1 Teeth

Teeth are especially valuable in the study of ancient microbes because they reveal information about the oral microbiome present in dental calculus, they have traces of blood-borne pathogens and the communities of bacteria decomposing the body [1].

Dental calculus, also known as mineralized dental plaque, is particularly interesting because it is the richest source of aDNA and it is the only part of the body that fossilises during an individual's lifetime [45]. Besides, the preservation of endogenous DNA in dental calculus is higher as it is less susceptible to decay [1]. For this reason there is aDNA recovered from ancient dental calculus

samples that are up to 100.000 years old [46]. One particularly interesting feature of this source of isolation is the fact that dental calculus is a bio-archive of respiratory pathogens and acts as a trap for microscopic fragments that include human cells, mineralized bacteria, plant micro-fossils and chemical and bio-molecular compounds that pass through the mouth when the person from which the sample was taken was alive [47]. A very recent study proved that archaeological dental calculus samples coming from regions with high temperature and humidity (despite being factors that increase DNA decay rates [48, 49]), still preserve a high proportion of DNA from endogenous oral microbiota, providing more evidence for the relevance of this isolation source in the study of palaeomicrobiology [50].

Furthermore, as teeth are vascularized during life [1], pathogens that infected the individual from which the sample was taken, can flow through the blood and go all the way up to the dental pulp chamber[4], leaving traces of their genetic material. Such traces are highly valuable, as they serve as a footprint for the field of pathogenomics [24, 40]. Due to this, pathogenic microbes such as the *Mycobacterium leprae* [34] or *Klebsiella pneumoniae* [51] have been identified in dental calculus.

3.3.1.2 Bones

Pathological alterations of skeletal tissues are generally caused by long-term chronic infections, and pathogens identified by ancient DNA analysis of pathological bone lesions include not only tuberculosis [33] but also leprosy [52] and syphilis [53]. However, when analysing the microbial composition from archaeological bone, care must be taken to distinguish potential pathogens from close relatives present in the soil that infiltrates the skeleton after death (i.e. soil contamination) [1].

3.3.1.3 Palaeofeces

Palaeofeces, or ancient faecal material, can provide valuable information about the gut microbiome of past populations. By analysing the DNA of microbes present in palaeofeces, also called coprolites, researchers can gain insights into the diets and health of ancient peoples [12, 38, 54]. Ancient human faeces can provide direct evidence of health and diet, sanitation practices, and social organisation in the past, as well as information on the local ecology and environment [55].

3.3.1.4 Cultural artifact residues

Cultural artifacts, such as pottery and dental tools, can also provide information about ancient microbes. By analysing the residues left on these artifacts, researchers can glean insight into both the types of microbes present in the environment and how ancient peoples interacted with them. This source of isolation provides a unique opportunity to better understand the processes of fermentation, culinary practices, and animal domestication and how they have changed through time [1].

3.3.1.5 Sediments

Sediments can provide a wealth of information about ancient microbes, including bacteria, archaea, fungi, microalgae, and phyto and zooplankton [1]. By analysing the DNA of microbes present in sediments, researchers can gain insights into the environmental conditions of the past and the ways in which ancient peoples interacted with their surroundings. Most studied sediments include cave sediments [56], lake sediments [57, 58], ocean sediments [59] and open-air archaeological sites [60].

3.4 aDNA degradation

As mentioned in Section 3.3, the main characteristic of ancient DNA (aDNA) is the fact that it has been degraded due to the passage of time. Degradation refers to the process of biomolecules being broken up and damaged through a variety of chemical and mechanical processes [43].

3.4.1 Damage patterns

The most important DNA decay reactions in fossil material are depurination, nick formation, and cytosine deamination. Interestingly, they both hinder and help the study of palaeometagenomics: they limit the amount of genetic material to sequence, but they are also used for authentication (see Section 3.5 for a more thorough view of this term) [62]. However, it is important to note that post-mortem damage can accumulate with the age of the sample, but the rate at which aDNA

[4]The highly vascularized inner tooth cavity that is contained within the crown and root portions [40]

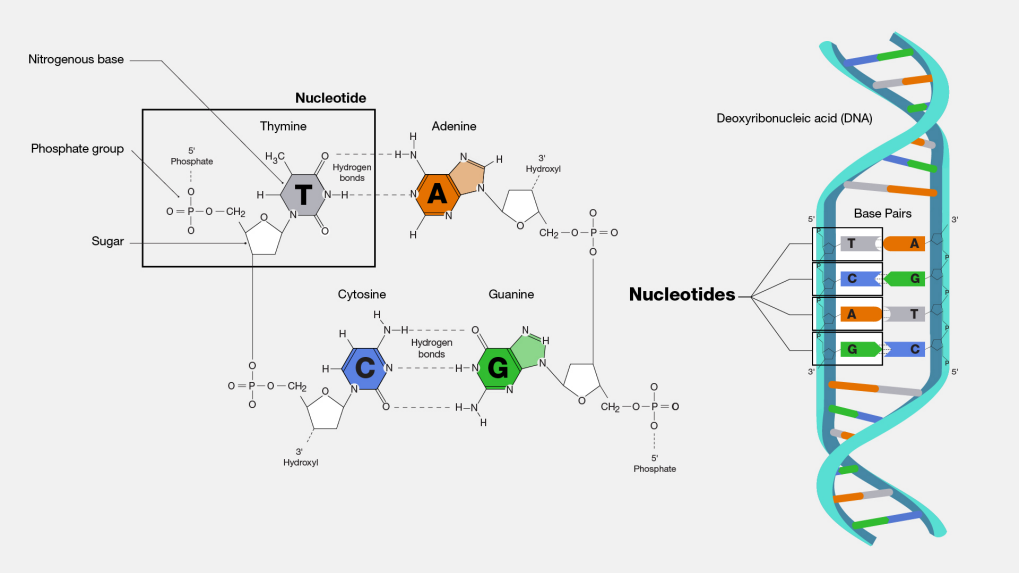


Figure 3.1: **Nucleic acids and nucleotide composition. Nucleic acids and nucleotide composition.** Nucleic acids such as RNA and DNA are composed of nucleotide chains. One single nucleotide is formed by a nitrogenous base, a phosphate group and a sugar. Image taken from [61].

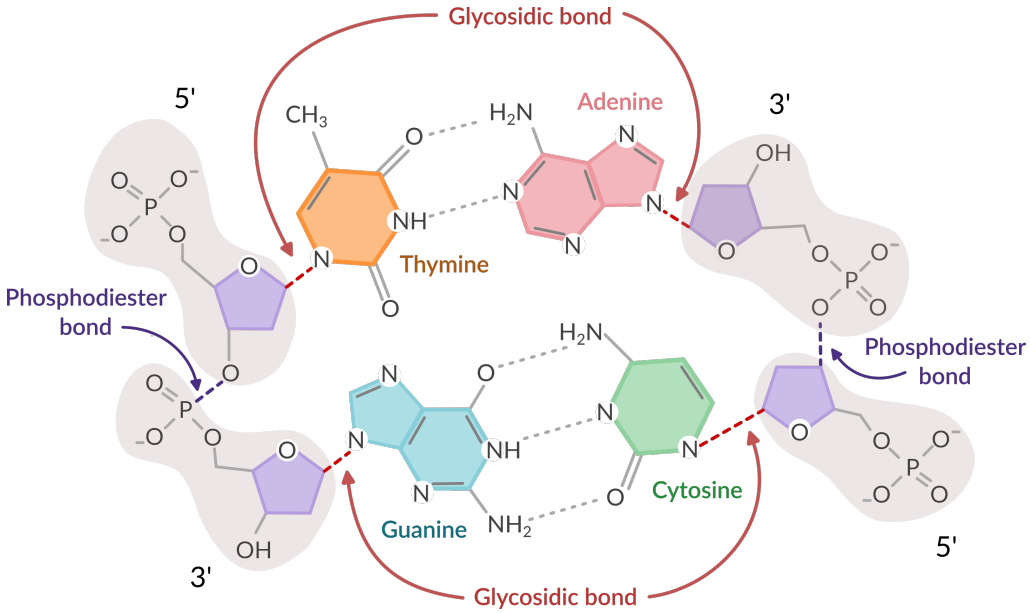


Figure 3.2: **Graphical representation of the *N*-glycosidic bond and the phosphodiester bond.** Graphical representation of the *N*-glycosidic bond between the nitrogenous base and the sugar of a nucleotide (left), and the phosphodiester bond, which links two nucleotides together to form the sugar-phosphate backbone (also called DNA backbone)

degrades is heavily influenced by local environmental conditions, which is why some newer samples can appear more damaged than older ones [16].

To begin, it is important to explain some basic biochemistry regarding the DNA molecule. A nucleotide is the fundamental unit of nucleic acids (DNA or RNA). It consists of a sugar molecule (deoxyribose for DNA and ribose in RNA), a phosphate group, and a nitrogenous base. DNA uses adenine, cytosine, guanine, and thymine as bases, while RNA uses uracil instead of thymine. DNA and RNA are polymers composed of nucleotide chains [61] (See Figure 3.1).

The nitrogenous base is bonded to the sugar by a *N*-glycosidic bond. Whereas the phosphodiester bond is a covalent linkage between the phosphate of one nucleotide and the hydroxyl (-OH) group attached to the 3' carbon of the sugar in the adjacent nucleotide. The latter forms the *"sugar-phosphate backbone"*, also called the *"DNA backbone"* [63] (See Figure 3.2).

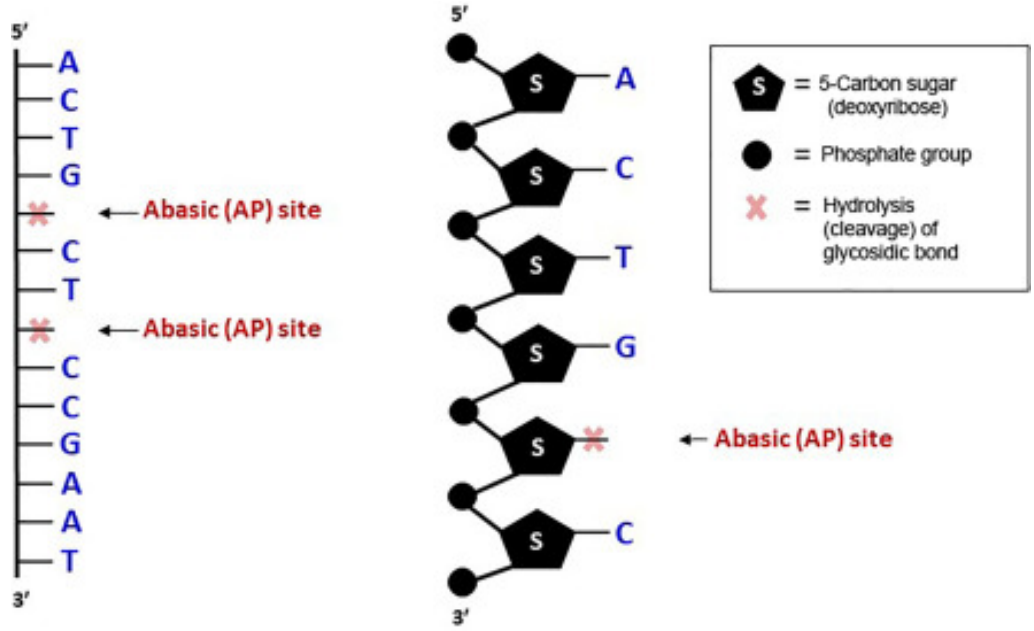


Figure 3.3: **Graphical representation of the formation of an abasic site due to depurination** (loss of a purine base). Image taken from [63]

3.4.1.1 Depurination

Depurination is a form of DNA damage where a purine base is lost (adenine or guanine)[64]. When a purine base is missing, this *apurinic* site is also called *abasic* (without a nitrogenous base). This occurs because the *N*-glycosidic bond between the nitrogenous base and the sugar is broken (see Figure 3.3)

3.4.1.2 Nick formation

Depurination leads to destabilisation of the DNA backbone, which in turn results in nick formations. This type of fragmentation arises as a consequence of the hydrolysis of phosphodiester bonds in the sugar-phosphate backbone of DNA. A Single-Strand Break (SSB) occurs when the phosphodiester bond on one DNA strand is broken, as opposed to a Double-Strand Break (DSB) which involves both DNA strands. SSB are sometimes referred to as *"nicks"* in the DNA backbone [63]. The concrete effect here is that broken DNA results in shorter reads.

3.4.1.3 Cytosine deamination

Deamination is a chemical modification where an amine group (NH2) is removed from a nucleotide through hydrolysis [28]. Cytosine is the most susceptible nucleotide to this type of damage, and after deamination, it is converted into a uracil (see Figure 3.5). It is also the most common miscoding lesion in aDNA, and results in the misreading of cytosine as thymine (T). These C-to-T substitutions occur most often at the ends of sequences (in the single-stranded overhanging termini of aDNA fragments, a deeper explanation of DNA fragmentation will come in the following subsection) [65, 66, 16]. Statistical DNA damage models, like mapDamage [67] or PMDtools [68] allow researchers to explore deamination patterns in their data. The expected deamination profile of a true aDNA sample shows an enrichment of C/T polymorphisms at the ends of the reads. PMDtools also estimates an ancient score (PMD score) for each read. Reads with PMD scores greater than 3 are labelled as reliably ancient, which is why this tool is also useful for separating ancient reads from modern contaminant sequences at the read level [39, 68].

3.4.2 DNA fragmentation

DNA fragmentation is the breakage of the DNA backbone. This is related to depurination (see subsection 3.4.1.1), since depurination destabilises the DNA backbone and results in nick formations (see subsection 3.4.1.2). As time passes, these nicks or Single-Strand Break (SSB) breaks in the DNA backbone become more common, and the resulting aDNA reads shorten up to 30–70 base pairs (bp) [44].

The length distribution of aDNA can be approximated by a lognormal distribution, showing an exponential decline in the tail due to random fragmentation of DNA [62]. There is a fragmentation

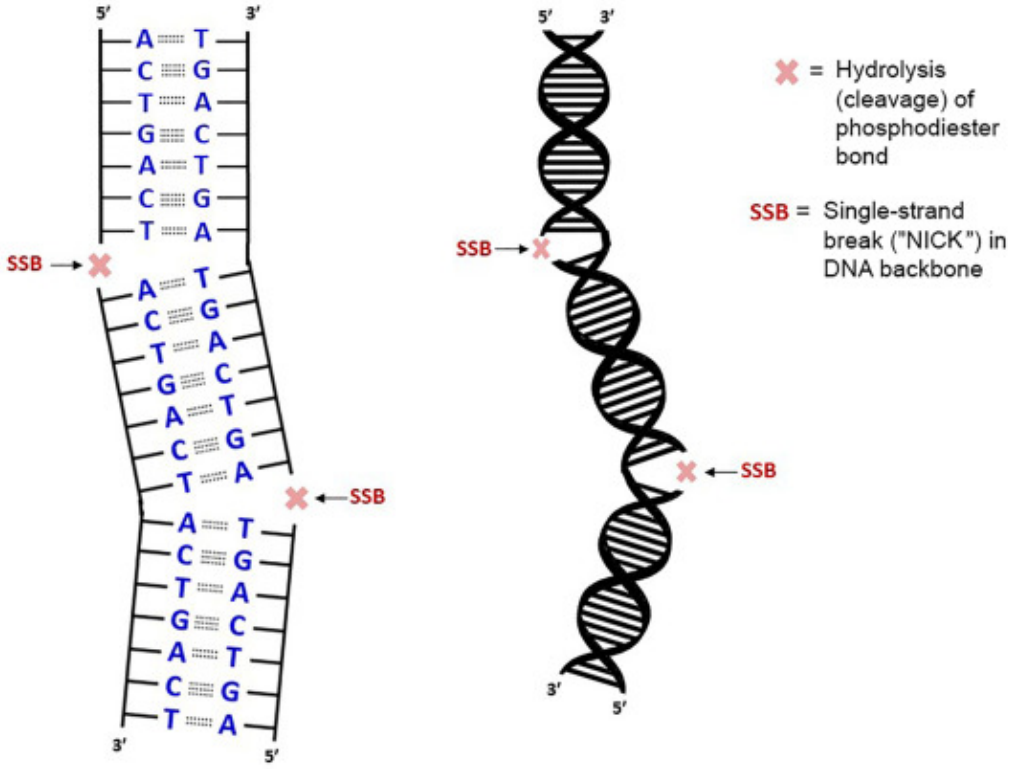


Figure 3.4: **Graphical representation of the Single-Strand Break (SSB)** that occurs in the DNA, leaving one strand broken. This is also called nick formation. Image taken from [63]

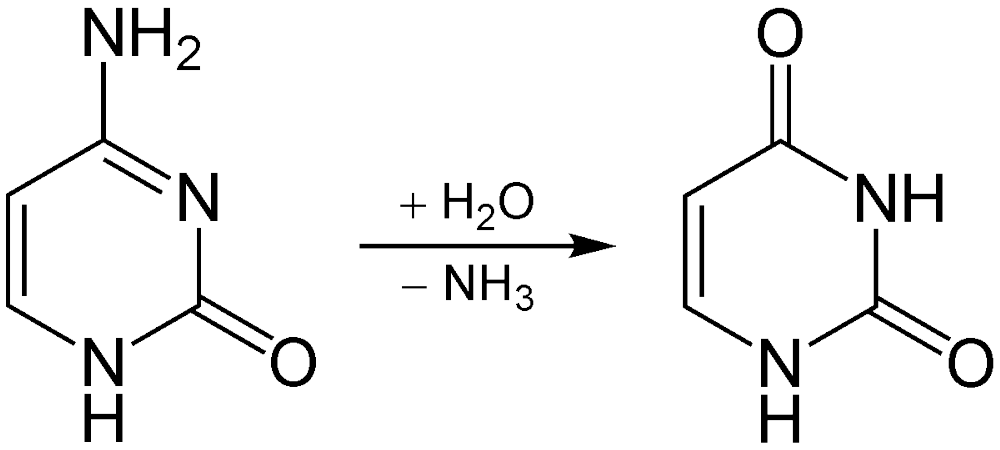


Figure 3.5: **Deamination from cytosine to uracil.** Image taken from [69]

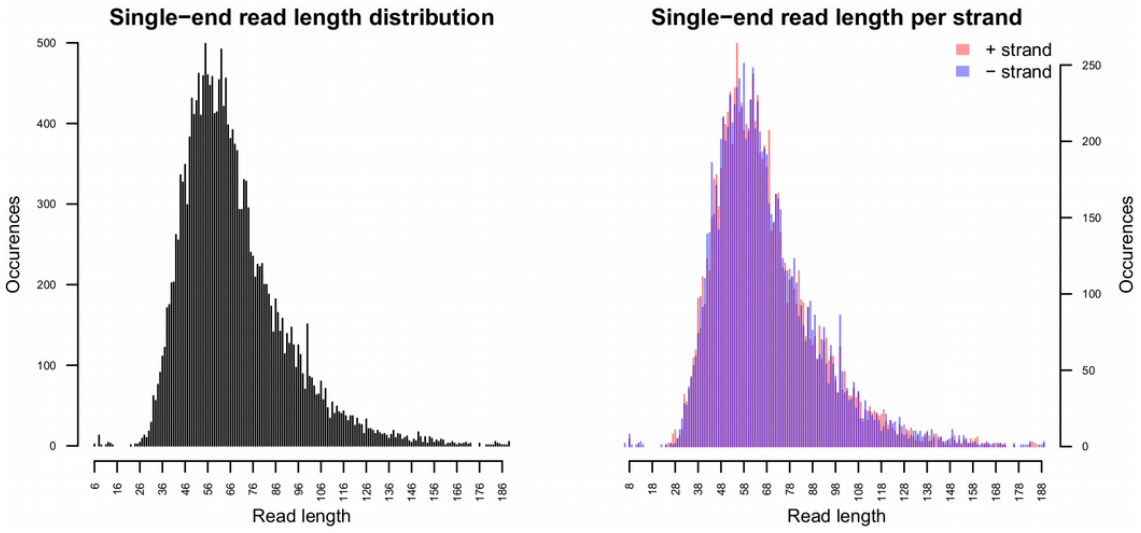
constant (λ) that represents the fraction of bonds broken in the DNA backbone and is an indicator of the magnitude of DNA fragmentation [70, 71]. The distribution of undamaged fragment sizes (x) is modelled with a random Poisson distribution (assuming that the DNA lesions are randomly distributed). The Poisson distribution that represents DNA damage is: [70, 72, 73]:

$$f(x) = \lambda \exp^{-\lambda x} \tag{3.1}$$

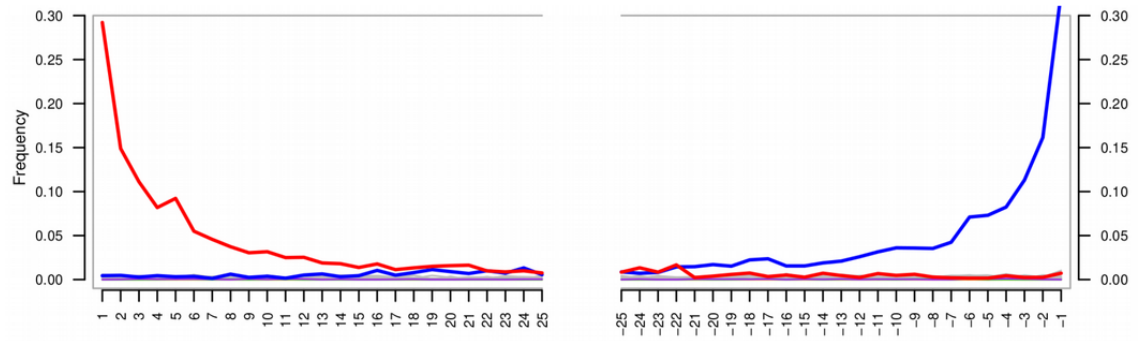
$$\lambda = -\ln \frac{A_D}{A_0} \tag{3.2}$$

Equation 3.1, expresses the distribution of undamaged fragment sizes as an exponential function that depends on λ and x. Lambda is defined in Equation 3.2 as the negative natural logarithm of the ratio between the amount of amplification of the damaged template (A_D), and the amplification product from undamaged DNA (A_0) [70]. This fragmentation constant is estimated after experimental measurements.

DNA fragmentation analysis is important for assessing the authenticity of aDNA, as it undergoes predictable forms of damage and decay. Fragmentation patterns can be visually inspected and are produced by default in software such as mapDamage [67].



(a) Read-length profiles



(b) Deamination profiles

Figure 3.6: **Plots used to assess read-length profiles and deamination profiles as generated by the mapDamage software.** A typical pattern of DNA fragmentation in ancient samples should have read lengths ranging from 30 to 70 base pairs. It is observed that most reads from the ancient sample in Figure 3.6a fall within this range. Reads longer than 100 base pairs are more likely to be the result of modern contamination. In Figure 3.6b, the plots represent the positions of specific nucleotide substitutions along the DNA, with the 5' end on the left-hand side of and the 3' end on the right-hand side of the x-axis of both plots. The red line indicates C-to-T substitutions, which are a result of cytosine deamination, while the blue line represents G-to-A substitutions. Ancient DNA samples are expected to exhibit an excess of C-to-T misincorporations at the 5' ends of sequences, and complementary G-to-A misincorporations at the 3'-termini, due to enhanced cytosine deamination in single-stranded 5'-overhanging ends. The images used in this work are sourced from [74].

3.5 aDNA authentication

One of the main goals of the study of ancient genetic material is to perform aDNA authentication, that is, to determine whether a given set of DNA molecules is truly ancient **and**[5] comes from the sample in question [44, 43]. As mentioned before, authentication can be done by proving that a sample shows signs of DNA damage patterns (e.g., depurination, nick formation, cytosine deamination) and DNA fragmentation. Nevertheless, this is not sufficient to prove aDNA data as being truly ancient, which is why authentication requires additional steps. Some of the reasons why simply studying damage patterns is not enough for authentication include [62]:

- Different molecular tools used during library construction (e.g certain repair enzymes, DNA ligases, polymerases etc) can influence the damage patterns observed in aDNA libraries.
- Genetic data is a mixture of ancient and modern DNA contaminants.
- As the decomposition of the sample begins right after death, there will be microbes involved in the process that also show signs of DNA damage[75]. This is important when researchers

[5]The emphasis is there to clarify that not only proving that a sample is ancient via different techniques is enough, one has to prove as well that the isolation source was indeed the origin of the ancient genetic material.

investigate microorganisms that are closely related to the composition of the soil microbiome [62].

- There are pathogens whose DNA-decay rate is lower, for example, *M. leprae* [33].

This does not mean that DNA fragmentation and damage patterns are not useful signs for aDNA authentication. However, in the case of significant conservation of the genetic material, alternatives such as positioning of the ancient genome in a phylogenetic tree, detection of genomic biomarkers, radiocarbon dating or detection of microbial taxa reflecting the expected sample type (for instance, palaeofeces should contain mostly gut taxa, instead of soil) are effective and necessary alternatives too. Ultimately a deep understanding of the biological context of different taxa may result in falsely attributing the source of an organism of interest [76].

Authentication approaches and guidelines can also involve procedures such as [28, 62]:

- Identifying adequate isolation sources such as dental calculus or petrous bones[6]. That is selecting reliable sources for aDNA collection, for instance, petrous bones and dental calculus have already proven to be rich sources of ancient genetic material [45, 77].
- Verifying damage and fragmentation patterns as described in Section 3.4.
- Molecular procedures such as clean room procedures, as well as the usage and sequencing of negative controls.
- Evaluating non-damage-dependent methods that rely on the estimation of contaminating Mitochondrial DNA (mtDNA) (applicable to samples with eukaryotic DNA) or other population genetics methods.
- Reference-based approaches such as read mapping (for which the user requires a narrow list of potentially interesting genomes) or taxonomic assignment of reads to a database (which, in contrast, demands having such a dataset), are subject to biases that have the potential to influence the obtained results. Additionally, the domain of palaeometagenomics has been challenged by the lack of an abundant collection of ancient reference genomes, unlike there is for other fields.

3.6 aDNA contamination

Contamination is used in this thesis as as proposed by [43]: *"Ancient and modern DNA not deriving from the original organism or sample of interest".* Contamination is one of the major problems in microbial archaeology analyses since fractions of exogenous DNA can lead to false biological and historical conclusions [39]. Reads coming from true aDNA molecules are called *"endogenous"* in contrast to contamination or material that is not of interest that is labelled as *"exogenous"*.

There is an important difference between contamination assessment and authentication in aDNA, which can be illustrated with the following example: One can estimate the presence of different DNA damage patterns of a presumed ancient sample, and this is a positive indication that at least some of the sequences in such sample are ancient. However, this validation does not exclude the presence of other contaminant (exogenous) sequences [78].

3.6.1 Laboratory techniques for aDNA contamination control

Sample and data hygiene standards and precautions are crucial in preventing modern DNA contamination in ancient microbial research. Some of the recommended practices include [62]:

- Use of isolated and dedicated facilities for aDNA manipulation and research.
- Use of High-Efficiency Particulate Arrestance (HEPA) air filtration.
- Use of UV irradiation.
- Use of NaOCl sterilisation.
- Personal protection measures such as full-body suits, double gloving, and eye shields.
- Usage of reagent blanks and negative controls.
- Setting up unidirectional workflows.

[6] Densest and most protected portion of the mammalian skull [28]

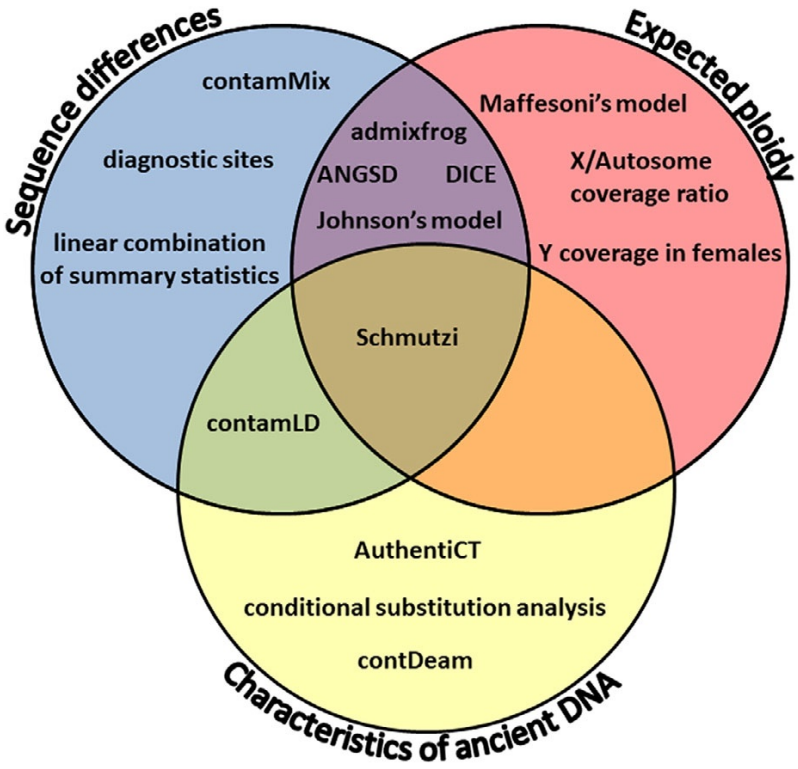


Figure 3.7: **Decontamination tools classified according to contamination signal**: to sequence difference, expected ploidy and characteristics of aDNA. Image taken from [66]

- Testing the reagents for microbial DNA contamination before using them on ancient samples.

It is important to note that despite having the best contamination reduction practices in a laboratory setting, it is impossible to fully remove aDNA contamination from Metagenomic Sequencing (MGS) studies [79, 80].

3.6.2 Computational techniques for aDNA contamination control

Finding solutions to the problem of ancient DNA contamination control remains a primary goal for the field of palaeomicrobiology over the coming years [44], which is why bioinformatic tools applied to aDNA bring a significant contribution to the field.

3.6.2.1 Single species

Methods for contamination assessment of genetic material in single species' genomes from an ancient sample have been classified based on the signals they use to estimate contamination [78].

Sequence differences: One category used for classification uses sequence differences between endogenous and exogenous DNA (see contamMix for instance [81]). All methods in this group require previous knowledge of the relationship between contaminating and ancient individuals, and they are more powerful if the divergence between the contaminating and ancient genome sequences is greater [78].

Expected ploidy: Another category is based on expected ploidy, where deviations from the expected ploidy in sex-chromosomes (applicable to human samples) or regions with large-scale insertion-deletion differences are used to estimate contamination (for an example see Mafessoni's model [82]). In this category, prior information regarding the differences between exogenous and endogenous DNA is not needed, yet, multiple-fold coverage is necessary [78].

Time-dependent characteristics: A third category for classification is based on time-dependent characteristics of aDNA, such as cytosine deamination or DNA length. These methods require few sequences and no previous knowledge of genetic relationships[78] (see AuthentiCT for example [66]).

Some methods rely on multiple signals, combining sequence differences, expected ploidy, and characteristics of ancient DNA to estimate contamination (see Schmutzi [83]).

3.6.2.2 Metagenomics of human samples

As mentioned previously, the advent of NGS technologies brought many advantages to palaeomicrobiology, as it is a non-selective technique whose laboratory protocols allow the rapid and high-throughput processing of genetic material [28]. However, metagenomic studies came with the need for different standards to assess aDNA contamination levels.

Chemical damage is still a useful sign in shotgun palaeometagenomics to computationally remove non-damaged and presumably contaminant sequences from ancient samples, increasing the reliability of posterior analysis after contamination removal [28, 68]. However, DNA damage alone is not enough to authenticate sequences as ancient (for instance when contaminating sequences also show signs of decay, DNA damage is a confusing signal). Two distinct approaches exist to computationally estimate contamination in ancient remains: estimating the percentage of Mitochondrial DNA (mtDNA) sequences originating from a different source or measuring the level of genome-wide contamination using population genetic methods.

The first method, which is reference-based, consists of gathering haploid Mitochondrial DNA (mtDNA) sequences to be assembled into partial or complete mitochondrial genomes. Each fragment is compared against a large database of known contaminants mtDNA [83, 81]. Using this technique mtDNA contamination can be detected, but nuclear contamination cannot be detected.

The second method, based on nuclear data, evaluates the sex of the individual where the sample comes from in various ways [28], such as calculating the ratio of sequences mapping to the Y and X chromosomes [84], estimating the X ratio [7] [85], performing deeper sequencing to study the proportion of heterozygous positions on the haploid X chromosome in males[8] or projecting the genomic data after performing a form of dimensionality reduction such as Principal Component Analysis (PCA) or Multidimensional Scaling (MDS) in order to compare female and male samples from the same site and period [77, 86, 87, 88] .

Metrics such as *edit distance*, that is, the number of mismatches for each read can be useful to determine whether a sample is contaminated or it is truly ancient material. The availability of a reference genome sequence would allow to map potentially ancient reads, one can obtain diagnostic plots to assess the quality of the data (see Fellows Yates, J. A. et al. *Introduction to Ancient Metagenomics.* 2022. DOI: `10.5281/zenodo.8027281`, section 16.1.4 Alignment quality for more details)

Other tools such as decontam [89] and cuperdec [46] have been developed more recently and are widely known in the aDNA community. The former is an R package that is useful when DNA concentration data is available. Even though it was not tailored or specifically tested in aDNA in the original publication, it is still useful for Metagenomic Sequencing studies. It is based on two core hypotheses: sequences from contaminating taxa will have frequencies that inversely correlate with sample DNA concentration and contaminating sequences will have higher prevalence in control samples. For this reason decontam requires the user to have sequenced controls or to provide DNA quantitation data.

The latter, cuperdec, is a R reference-based package for the estimation and visualisation of the endogenous taxonomic content of ancient metagenomes [43]. Cuperdec's main idea is to rank organisms in each sample by their abundance and then compute their enrichment against a reference database that contains a list of microbial organisms specific to a certain tissue/environment. The tool produces Cumulative Percent Decay curves and with the visual help of such curves, the user can identify samples that should be preserved versus samples that should be discarded [39, 43, 90].

References

[1] Warinner, C. "An Archaeology of Microbes". In: *Journal of Anthropological Research* vol. 78, no. 4 (2022), pp. 420–458.

[2] Bunge, J., Willis, A., and Walsh, F. "Estimating the number of species in microbial diversity studies". In: *Annual Review of Statistics and Its Application* vol. 1 (2014), pp. 427–445.

[3] Knoll, A. H. "Paleobiological perspectives on early microbial evolution". In: *Cold Spring Harbor Perspectives in Biology* vol. 7, no. 7 (2015), a018093.

[4] Allwood, A. C., Walter, M. R., Burch, I. W., and Kamber, B. S. "3.43 billion-year-old stromatolite reef from the Pilbara Craton of Western Australia: ecosystem-scale insights to early life on Earth". In: *Precambrian Research* vol. 158, no. 3-4 (2007), pp. 198–227.

[7] Mean coverage on the sex chromosome X normalised by the mean coverage on the autosomes. In females the ratio should be 1.0, in males 0.5

[8] This test is only useful for male samples, as a single allele will exist on the X chromosome and a level of heterozygosity greater than that should be due contamination or sequencing errors

[5] Merriam-Webster. *Abulia*. In: *Merriam-Webster.com dictionary*. URL: https://www.merriam-webster.com/dictionary/biotic (visited on 09/07/2023).

[6] Merriam-Webster. *Abulia*. In: *Merriam-Webster.com dictionary*. URL: https://www.merriam-webster.com/dictionary/abiotic (visited on 09/07/2023).

[7] Marchesi, J. R. and Ravel, J. *The vocabulary of microbiome research: a proposal*. 2015.

[8] Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., et al. "The NIH human microbiome project". In: *Genome research* vol. 19, no. 12 (2009), pp. 2317–2323.

[9] Sim, S. H., Yu, Y., Lin, C. H., Karuturi, R. K. M., Wuthiekanun, V., Tuanyok, A., Chua, H. H., Ong, C., Paramalingam, S. S., Tan, G., et al. "The core and accessory genomes of Burkholderia pseudomallei: implications for human melioidosis". In: *PLoS pathogens* vol. 4, no. 10 (2008), e1000178.

[10] Segerman, B. "The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories". In: *Frontiers in cellular and infection microbiology* vol. 2 (2012), p. 116.

[11] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. "A human gut microbial gene catalogue established by metagenomic sequencing". In: *nature* vol. 464, no. 7285 (2010), pp. 59–65.

[12] Warinner, C., Speller, C., Collins, M. J., and Lewis Jr, C. M. "Ancient human microbiomes". In: *Journal of human evolution* vol. 79 (2015), pp. 125–136.

[13] Fry, G. F. *Analysis of prehistoric coprolites from Utah*. University of Utah Press, 1978.

[14] Lin, D. S., Connor, W. E., Napton, L. K., and Heizer, R. F. "The steroids of 2000-year-old human coprolites". In: *Journal of Lipid Research* vol. 19, no. 2 (1978), pp. 215–221.

[15] Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., and Wilson, A. C. "DNA sequences from the quagga, an extinct member of the horse family". In: *Nature* vol. 312, no. 5991 (1984), pp. 282–284.

[16] Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., et al. "Ancient DNA analysis". In: *Nature Reviews Methods Primers* vol. 1, no. 1 (2021), p. 14.

[17] Mullis, K. B. and Faloona, F. A. "[21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction". In: *Methods in enzymology*. Vol. 155. Elsevier, 1987, pp. 335–350.

[18] Pääbo, S., Gifford, J. A., and Wilson, A. C. "Mitochondrial DNA sequences from a 7000-year old brain". In: *Nucleic acids research* vol. 16, no. 20 (1988), pp. 9775–9787.

[19] Cavalieri, D., McGovern, P. E., Hartl, D. L., Mortimer, R., and Polsinelli, M. "Evidence for S. cerevisiae fermentation in ancient wine". In: *Journal of molecular evolution* vol. 57 (2003), S226–S232.

[20] Rafi, A., Spigelman, M., Stanford, J., Lemma, E., Donoghue, H., and Zias, J. "DNA of Mycobacterium leprae detected by PCR in ancient bone". In: *International journal of Osteoarchaeology* vol. 4, no. 4 (1994), pp. 287–290.

[21] Drancourt, M., Aboudharam, G., Signoli, M., Dutour, O., and Raoult, D. "Detection of 400-year-old Yersinia pestis DNA in human dental pulp: an approach to the diagnosis of ancient septicemia". In: *Proceedings of the National Academy of Sciences* vol. 95, no. 21 (1998), pp. 12637–12640.

[22] Shapiro, B., Rambaut, A., and Gilbert, M. T. P. "No proof that typhoid caused the Plague of Athens (a reply to Papagrigorakis et al.)" In: *International Journal of Infectious Diseases* vol. 10, no. 4 (2006), pp. 334–335.

[23] Gilbert, M. T. P., Cuccui, J., White, W., Lynnerup, N., Titball, R. W., Cooper, A., and Prentice, M. B. "Absence of Yersinia pestis-specific DNA in human teeth from five European excavations of putative plague victims". In: *Microbiology* vol. 150, no. 2 (2004), pp. 341–354.

[24] Bos, K. I., Kühnert, D., Herbig, A., Esquivel-Gomez, L. R., Andrades Valtueña, A., Barquera, R., Giffin, K., Kumar Lankapalli, A., Nelson, E. A., Sabin, S., et al. "Paleomicrobiology: diagnosis and evolution of ancient pathogens". In: *Annual Review of Microbiology* vol. 73 (2019), pp. 639–666.

[25] Willerslev, E. and Cooper, A. "Ancient dna". In: *Proceedings of the Royal Society B: Biological Sciences* vol. 272, no. 1558 (2005), pp. 3–16.

[26] Goodwin, S., McPherson, J. D., and McCombie, W. R. "Coming of age: ten years of next-generation sequencing technologies". In: *Nature Reviews Genetics* vol. 17, no. 6 (2016), pp. 333–351.

[27] Mardis, E. R. "A decade's perspective on DNA sequencing technology". In: *Nature* vol. 470, no. 7333 (2011), pp. 198–203.

[28] Key, F. M., Posth, C., Krause, J., Herbig, A., and Bos, K. I. "Mining metagenomic data sets for ancient DNA: recommended protocols for authentication". In: *Trends in Genetics* vol. 33, no. 8 (2017), pp. 508–520.

[29] Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., McPhee, J. B., DeWitte, S. N., Meyer, M., Schmedes, S., et al. "A draft genome of Yersinia pestis from victims of the Black Death". In: *Nature* vol. 478, no. 7370 (2011), pp. 506–510.

[30] Schuenemann, V. J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., Forrest, S., Coombes, B. K., Wood, J. W., Earn, D. J., et al. "Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death". In: *Proceedings of the National Academy of Sciences* vol. 108, no. 38 (2011), E746–E752.

[31] Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S. R., Schuenemann, V. J., et al. "Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis". In: *Nature* vol. 514, no. 7523 (2014), pp. 494–497.

[32] Sabin, S., Herbig, A., Vågene, Å. J., Ahlström, T., Bozovic, G., Arcini, C., Kühnert, D., and Bos, K. I. "A seventeenth-century Mycobacterium tuberculosis genome supports a Neolithic emergence of the Mycobacterium tuberculosis complex". In: *Genome biology* vol. 21, no. 1 (2020), pp. 1–24.

[33] Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jäger, G., Bos, K. I., Herbig, A., Economou, C., Benjak, A., Busso, P., et al. "Genome-wide comparison of medieval and modern Mycobacterium leprae". In: *Science* vol. 341, no. 6142 (2013), pp. 179–183.

[34] Fotakis, A. K., Denham, S. D., Mackie, M., Orbegozo, M. I., Mylopotamitaki, D., Gopalakrishnan, S., Sicheritz-Pontén, T., Olsen, J. V., Cappellini, E., Zhang, G., et al. "Multi-omic detection of Mycobacterium leprae in archaeological human dental calculus". In: *Philosophical Transactions of the Royal Society B* vol. 375, no. 1812 (2020), p. 20190584.

[35] Key, F. M., Posth, C., Esquivel-Gomez, L. R., Hübler, R., Spyrou, M. A., Neumann, G. U., Furtwängler, A., Sabin, S., Burri, M., Wissgott, A., et al. "Emergence of human-adapted Salmonella enterica is linked to the Neolithization process". In: *Nature ecology & evolution* vol. 4, no. 3 (2020), pp. 324–333.

[36] Vågene, Å. J., Herbig, A., Campana, M. G., Robles García, N. M., Warinner, C., Sabin, S., Spyrou, M. A., Andrades Valtueña, A., Huson, D., Tuross, N., et al. "Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico". In: *Nature ecology & evolution* vol. 2, no. 3 (2018), pp. 520–528.

[37] Zhou, Z., Lundstrøm, I., Tran-Dien, A., Duchêne, S., Alikhan, N.-F., Sergeant, M. J., Langridge, G., Fotakis, A. K., Nair, S., Stenøien, H. K., et al. "Pan-genome analysis of ancient and modern Salmonella enterica demonstrates genomic stability of the invasive para C lineage for millennia". In: *Current Biology* vol. 28, no. 15 (2018), pp. 2420–2428.

[38] Maixner, F., Sarhan, M. S., Huang, K. D., Tett, A., Schoenafinger, A., Zingale, S., Blanco-Míguez, A., Manghi, P., Cemper-Kiesslich, J., Rosendahl, W., et al. "Hallstatt miners consumed blue cheese and beer during the Iron Age and retained a non-Westernized gut microbiome until the Baroque period". In: *Current Biology* vol. 31, no. 23 (2021), pp. 5149–5162.

[39] Fellows Yates, J. A. et al. *Introduction to Ancient Metagenomics.* 2022. DOI: 10.5281/zenodo.8027281.

[40] Spyrou, M. A., Bos, K. I., Herbig, A., and Krause, J. "Ancient pathogen genomics as an emerging tool for infectious disease research". In: *Nature Reviews Genetics* vol. 20, no. 6 (2019), pp. 323–340.

[41] Consortium, T. H. M. P. "Structure, function and diversity of the healthy human microbiome". In: *nature* vol. 486, no. 7402 (2012), pp. 207–214.

[42] Warinner, C., Speller, C., and Collins, M. J. "A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 370, no. 1660 (2015), p. 20130376.

[43] Yates, J. A. F., Valtueña, A. A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-López, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., et al. "Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir". In: *Scientific Data* vol. 8, no. 1 (2021), pp. 1–8.

[44] Der Sarkissian, C., Velsko, I. M., Fotakis, A. K., Vågene, Å. J., Hübner, A., and Fellows Yates, J. A. "Ancient metagenomic studies: Considerations for the wider scientific community". In: *Msystems* vol. 6, no. 6 (2021), e01315–21.

[45] Mann, A. E., Sabin, S., Ziesemer, K., Vågene, Å. J., Schroeder, H., Ozga, A. T., Sankaranarayanan, K., Hofman, C. A., Fellows Yates, J. A., Salazar-García, D. C., et al. "Differential preservation of endogenous human and microbial DNA in dental calculus and dentin". In: *Scientific reports* vol. 8, no. 1 (2018), p. 9822.

[46] Fellows Yates, J. A., Velsko, I. M., Aron, F., Posth, C., Hofman, C. A., Austin, R. M., Parker, C. E., Mann, A. E., Nägele, K., Arthur, K. W., et al. "The evolution and changing ecology of the African hominid oral microbiome". In: *Proceedings of the National Academy of Sciences* vol. 118, no. 20 (2021), e2021655118.

[47] Radini, A., Nikita, E., Buckley, S., Copeland, L., and Hardy, K. "Beyond food: The multiple pathways for inclusion of materials into ancient dental calculus". In: *American journal of physical anthropology* vol. 162 (2017), pp. 71–83.

[48] Kircher, M., Sawyer, S., and Meyer, M. "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform". In: *Nucleic acids research* vol. 40, no. 1 (2012), e3–e3.

[49] Meyer, M., Kircher, M., et al. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing". In: *Cold Spring Harb Protoc* vol. 2010, no. 6 (2010), t5448.

[50] Velsko, I. M. et al. "Exploring archaeogenetic studies of dental calculus to shed light on past human migrations in Oceania". In: *bioRxiv* (2023). DOI: 10.1101/2023.10.18.563027. eprint: https://www.biorxiv.org/content/early/2023/10/19/2023.10.18.563027.full.pdf. URL: https://www.biorxiv.org/content/early/2023/10/19/2023.10.18.563027.

[51] Austin, R. M., Zuckerman, M., Honap, T. P., Lee, H., Ward, G. K., Warinner, C., Sankaranarayanan, K., and Hofman, C. A. "Remembering St. Louis individual—structural violence and acute bacterial infections in a historical anatomical collection". In: *Communications Biology* vol. 5, no. 1 (2022), p. 1050.

[52] Schuenemann, V. J., Avanzi, C., Krause-Kyora, B., Seitz, A., Herbig, A., Inskip, S., Bonazzi, M., Reiter, E., Urban, C., Dangvard Pedersen, D., et al. "Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe". In: *PLoS pathogens* vol. 14, no. 5 (2018), e1006997.

[53] Schuenemann, V. J., Kumar Lankapalli, A., Barquera, R., Nelson, E. A., Iraíz Hernández, D., Acuña Alonzo, V., Bos, K. I., Márquez Morfín, L., Herbig, A., and Krause, J. "Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains". In: *PLoS neglected tropical diseases* vol. 12, no. 6 (2018), e0006447.

[54] Shillito, L.-M., Blong, J. C., Green, E. J., and Asperen, E. N. van. "The what, how and why of archaeological coprolite analysis". In: *Earth-Science Reviews* vol. 207 (2020), p. 103196.

[55] Reinhard, K. J. and Bryant Jr, V. M. "Pathoecology and the future of coprolite studies in bioarchaeology". In: (2008).

[56] Massilani, D., Morley, M. W., Mentzer, S. M., Aldeias, V., Vernot, B., Miller, C., Stahlschmidt, M., Kozlikin, M. B., Shunkov, M. V., Derevianko, A. P., et al. "Microstratigraphic preservation of ancient faunal and hominin DNA in Pleistocene cave sediments". In: *Proceedings of the National Academy of Sciences* vol. 119, no. 1 (2022), e2113666118.

[57] Nwosu, E. C., Brauer, A., Kaiser, J., Horn, F., Wagner, D., and Liebner, S. "Evaluating sedimentary DNA for tracing changes in cyanobacteria dynamics from sediments spanning the last 350 years of Lake Tiefer See, NE Germany". In: *Journal of Paleolimnology* vol. 66, no. 3 (2021), pp. 279–296.

[58] Fernandez-Guerra, A., Borrel, G., Delmont, T. O., Elberling, B., Eren, A. M., Gribaldo, S., Jochheim, A., Henriksen, R. A., Hinrichs, K.-U., Korneliussen, T. S., et al. "A 2-million-year-old microbial and viral communities from the Kap København Formation in North Greenland". In: *bioRxiv* (2023), pp. 2023–06.

[59] Armbrecht, L. H. "The potential of sedimentary ancient DNA to reconstruct past ocean ecosystems". In: *Oceanography* vol. 33, no. 2 (2020), pp. 116–123.

[60] Hudson, S. M., Pears, B., Jacques, D., Fonville, T., Hughes, P., Alsos, I., Snape, L., Lang, A., and Brown, A. “Life before Stonehenge: The hunter-gatherer occupation and environment of Blick Mead revealed by sedaDNA, pollen and spores”. In: *Plos one* vol. 17, no. 4 (2022), e0266789.

[61] Institute), N. (H. G. R. “Talking glossary of genomic and genetic terms”. In: (2022).

[62] Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiß, C. L., Burbano, H. A., Orlando, L., and Krause, J. “A robust framework for microbial archaeology”. In: *Annual Review of Genomics and Human Genetics* vol. 18 (2017), pp. 321–356.

[63] Ambers, A. “Challenges in forensic genetic investigations of decomposed or skeletonized human remains: Environmental exposure, DNA degradation, inhibitors, and low copy number (LCN)”. In: *Forensic Genetic Approaches for Identification of Human Skeletal Remains*. Elsevier, 2023, pp. 15–36.

[64] An, R., Dong, P., Komiyama, M., Pan, X., and Liang, X. “Inhibition of nonenzymatic depurination of nucleic acids by polycations”. In: *FEBS open bio* vol. 7, no. 11 (2017), pp. 1707–1714.

[65] Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., et al. “Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments”. In: *Proceedings of the National Academy of Sciences* vol. 110, no. 39 (2013), pp. 15758–15763.

[66] Peyrégne, S. and Peter, B. M. “AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination”. In: *Genome biology* vol. 21, no. 1 (2020), pp. 1–16.

[67] Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E., and Orlando, L. “mapDamage: testing for damage patterns in ancient DNA sequences”. In: *Bioinformatics* vol. 27, no. 15 (2011), pp. 2153–2155.

[68] Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., and Jakobsson, M. “Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal”. In: *Proceedings of the National Academy of Sciences* vol. 111, no. 6 (2014), pp. 2229–2234.

[69] Wikipedia. *Deamination — Wikipedia, The Free Encyclopedia*. [Online; accessed 04-September-2023]. 2023. URL: `https://en.wikipedia.org/wiki/Deamination#/media/File:DesaminierungCtoU.png`.

[70] Deagle, B. E., Eveson, J. P., and Jarman, S. N. “Quantification of damage in DNA recovered from highly degraded samples–a case study on DNA in faeces”. In: *Frontiers in zoology* vol. 3, no. 1 (2006), pp. 1–10.

[71] Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., et al. “The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils”. In: *Proceedings of the Royal Society B: Biological Sciences* vol. 279, no. 1748 (2012), pp. 4724–4733.

[72] Fernando, L. P., Kurian, P. J., Fidan, M., and Fernandes, D. J. “Quantitation of gene-specific DNA damage by competitive PCR”. In: *Analytical biochemistry* vol. 306, no. 2 (2002), pp. 212–221.

[73] Ayala-Torres, S., Chen, Y., Svoboda, T., Rosenblatt, J., and Van Houten, B. “Analysis of gene-specific DNA damage and repair using quantitative polymerase chain reaction”. In: *Methods* vol. 22, no. 2 (2000), pp. 135–147.

[74] Mikkel Schubert Aurélien Ginolhac, H. J. *mapDamage: tracking and quantifying damage patterns in ancient DNA sequences*. 2021. URL: `https://ginolhac.github.io/mapDamage/#a9` (visited on 12/09/2023).

[75] Metcalf, J. L., Xu, Z. Z., Weiss, S., Lax, S., Van Treuren, W., Hyde, E. R., Song, S. J., Amir, A., Larsen, P., Sangwan, N., et al. “Microbial community assembly and metabolic function during mammalian corpse decomposition”. In: *Science* vol. 351, no. 6269 (2016), pp. 158–162.

[76] Campana, M. G., Robles García, N., Rühli, F. J., and Tuross, N. “False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing”. In: *BMC research notes* vol. 7, no. 1 (2014), pp. 1–15.

[77] Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., Gerritsen, F., Moiseyev, V., Gromov, A., Raczky, P., et al. “Optimal ancient DNA yields from the inner ear part of the human petrous bone”. In: *PloS one* vol. 10, no. 6 (2015), e0129102.

[78] Peyrégne, S. and Prüfer, K. “Present-Day DNA Contamination in Ancient DNA Datasets”. In: *Bioessays* vol. 42, no. 9 (2020), p. 2000081.

[79] Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses”. In: *BMC Biology* vol. 12 (2014), pp. 1–12.

[80] Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. “Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data”. In: *Microbiome* vol. 6, no. 1 (2018), pp. 1–14.

[81] Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., et al. “A revised timescale for human evolution based on ancient mitochondrial genomes”. In: *Current biology* vol. 23, no. 7 (2013), pp. 553–559.

[82] Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., De La Rasilla, M., Lalueza-Fox, C., Rosas, A., Soressi, M., Knul, M. V., Miller, R., et al. “Neandertal and Denisovan DNA from Pleistocene sediments”. In: *Science* vol. 356, no. 6338 (2017), pp. 605–608.

[83] Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. “Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA”. In: *Genome biology* vol. 16, no. 1 (2015), pp. 1–18.

[84] Skoglund, P., Storå, J., Götherström, A., and Jakobsson, M. “Accurate sex identification of ancient human remains using DNA shotgun sequencing”. In: *Journal of archaeological Science* vol. 40, no. 12 (2013), pp. 4477–4482.

[85] Mittnik, A., Wang, C.-C., Svoboda, J., and Krause, J. “A molecular approach to the sexing of the triple burial at the Upper Paleolithic Site of Dolní Věstonice”. In: *PloS one* vol. 11, no. 10 (2016), e0163019.

[86] Malaspinas, A.-S., Tange, O., Moreno-Mayar, J. V., Rasmussen, M., DeGiorgio, M., Wang, Y., Valdiosera, C. E., Politis, G., Willerslev, E., and Nielsen, R. “bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS)”. In: *Bioinformatics* vol. 30, no. 20 (2014), pp. 2962–2964.

[87] Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. “ANGSD: analysis of next generation sequencing data”. In: *BMC bioinformatics* vol. 15, no. 1 (2014), pp. 1–13.

[88] Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., et al. “An Aboriginal Australian genome reveals separate human dispersals into Asia”. In: *Science* vol. 334, no. 6052 (2011), pp. 94–98.

[89] Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. “Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data”. In: *Microbiome* vol. 6 (2018), pp. 1–14.

[90] Yates, J. A. F. *Introduction to cuperdec*. 2021. URL: `https://cran.r-project.org/web/packages/cuperdec/vignettes/cuperdec-intro.html` (visited on 12/09/2021).

Chapter 4

Overview of state of the art (Chapters 5 and 7)

This chapter gives a brief motivation for the development of this thesis and an overview of the State of the art chapters 5 and 7.

Ancient metagenomics is a field that faces several prominent challenges: DNA fragmentation, DNA degradation, and contamination. This thesis focuses on the latter, specifically, on improving the methodological gap that exists in current bioinformatic methods for contamination assessment and contamination removal.

Since removing contamination entirely is impossible, researchers have developed computational tools to account for the contamination of ancient metagenomic samples. Notably, two of the most prominent methods for contamination assessment in palaeometagenomics, namely FEAST [1] and mSourceTracker [2, 3], are probabilistic approaches that require a reference database to construct their input data. However, simply acquiring such a database is challenging because there is a limited proportion of microbial diversity that has been sequenced in the field (thanks to the many challenges faced when sequencing DNA coming from scarce and sensitive fossil records). Not only there is not a clear consensus on what taxa constitutes one microbiome and differentiates it from another, but also both methods require the use of taxonomic classifiers to create their input data (specifically called taxonomic abundance tables or OTU tables), which poses several challenges on its own. For instance, the user needs to select a taxonomic classifier from a large plethora of options, determine various running parameters, and download or even construct their reference database of genomic sequences to create such input tables. These inherent heuristics of reference-based methods open the door to controversies that often deem reproducibility across experiments impossible. Overall, reference-based methods, as their name suggests, depend on a reference sequence that significantly impacts the quality of the downstream analysis and must therefore be as accurate as possible. Unfortunately, many ancient microbial species lack reference genomes of sufficient quality or have no reference genome at all. Particularly in the context of palaeometagenomics, reference methods are ill-suited as metagenomes in this context have not been extensively characterised yet.

In summary, current tools for contamination assessment in paleometagenomics come with the intrinsic biases of reference-based methods, require parameter optimisation for optimal performance, do not guarantee convergence due to their Bayesian nature, and are not deterministic. In light of these limitations, we propose a reference-free method that circumvents the use of taxonomy-based clustering tables called `decOM`. Our method uses k-mer representations that leverage the abundance of metagenomic sequencing data available (at the time) for ancient dental calculus samples. In this thesis, we aimed to deviate from database-dependent methods and instead employ unsupervised approaches that exploit the composition of read-level sequences and the wealth of information encapsulated within previously sequenced metagenomes.

In State of the Art: Ancient Microbial Source Tracking (refer to Chapter 5), I will introduce some bioinformatic concepts that are essential to understanding the paper presented in Chapter 6. To explain the nature of Microbial Source Tracking (MST), I present the minimal input data required for any algorithm of this type: a taxonomic abundance table and a sink/source table. Since the construction of the former is still a source of large controversy and the reason we introduced several ways to produce them in [4], a full subsection regarding the discussion on how to build taxonomic abundance tables was included. Finally, I explain the Bayesian theory underpinning the algorithms for MST that were compared against `decOM` (FEAST and mSourceTracker), and also provide a brief introduction to what k-mers and k-mer matrices are, as they are a foundational component of our contamination assessment method. `decOM`, which is the paper presented in chapter 6, emerged as a solution to the unresolved issue of MST tailored for ancient DNA (aDNA) contamination assessment using a reference-free method. Before delving into the issue of MST, I introduce the concept of taxonomic classification. The rationale behind this is that competing methods to `decOM` require the use of taxonomic classifiers to produce their input data. Moreover, I concisely explain how three algorithms for taxonomic classification work: Kaiju, KrakenUniq, and Centrifuge. Subsequently, I mention briefly other more recent and very prominent pipelines in the field of palaeometagenomics, all of which are reference-based.

Once DNA contamination has been assessed, one could try to remove contaminated (unwanted) reads from the FASTA/FASTQ files, in order to maximize the usage of the scarce biomaterials

from which the samples are derived. Hence, in the part of the State of the art: Ancient reads decontamination (refer to Chapter 7), I present the software DeconSeq and Recentrifuge. They are both competing methods to the algorithm proposed in Chapter 8, a reference-free and novel method that uses a Bloom Filter to perform ancient DNA contamination removal at the read-level called `aKmerBroom`. As explained in Chapter 7, DeconSeq and Recentrifuge require reference databases or sequences negative controls, their performance depends on parameter optimization that has not been benchmarked for aDNA and are methods that were not developed or tested specifically on aDNA data. Furthermore, an explanation is provided regarding the nature of Bloom Filters, which includes hash functions and the variables that affect the false positive rate of such data structure.

The shortcomings that come with state-of-the-art digital methods for contamination removal motivated the development of the paper González, C. D. et al. "aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets". In: *iScience* vol. 26, no. 11 (2023) (see Chapter 8), an alignment-free method that requires no parameter optimization, was developed and tested specifically on aDNA data that also offers the flexibility to be potentially used on other types of metagenomic data.

References

[1] Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., and Halperin, E. "FEAST: fast expectation-maximization for microbial source tracking". In: *Nature Methods* vol. 16, no. 7 (2019), pp. 627–632.

[2] Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. "Bayesian community-wide culture-independent microbial source tracking". In: *Nature Methods* vol. 8, no. 9 (2011), pp. 761–763.

[3] McGhee, J. J., Rawson, N., Bailey, B. A., Fernandez-Guerra, A., Sisk-Hackworth, L., and Kelley, S. T. "Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics". In: *PeerJ* vol. 8 (2020), e8783.

[4] González, C. D., Vicedomini, R., Lemane, T., Rascovan, N., Richard, H., and Chikhi, R. "decOM: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods". In: *Microbiome* vol. 11, no. 1 (2023), pp. 243–243.

[5] González, C. D., Rangavittal, S., Vicedomini, R., Chikhi, R., and Richard, H. "aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets". In: *iScience* vol. 26, no. 11 (2023).

Chapter 5

State of the art: Ancient Microbial Source Tracking

5.1 Taxonomic classifiers

The current production of vast quantities of metagenomic data requires algorithms capable of assessing microbial content in large datasets with reasonable memory and run-time requirements [1]. Metagenomic Sequencing experiments produce large collections of genomic data from a set of species, rather than providing genetic information from a single isolated species. Due to this, algorithms for metagenomic classification are designed to distinguish from a mixture of microbes, create abundance profiles that indicate which species are present in a metagenomic sample and assess the abundance of each of those species in such sample (see Figure 5.1). This task is computationally challenging due to two main reasons: the exponential growth in recent years in the number of sequenced microbial genomes, and the widespread use of NGS technologies that generate millions of short sequences. Well-known algorithms such as Basic Local Alignment and Search Tool (BLAST) [2] have been widely-used by the bioinformatics community, and despite being one of the most sensitive metagenomic alignment methods, it is infeasible to scale it to the millions of raw sequences present in today's metagenomic samples [3].

Tool for taxonomic classification of metagenomic data and estimation of taxon abundance profiles have been divided into two types according to the task performed: taxonomic binning and taxonomic profiling. *Taxonomic binning* refers to the classification of individual sequence reads into reference taxa. *Taxonomic profiling*, on the other hand, refers to the quantitative assessment of relative abundances of taxa within a dataset but not necessarily the classification of individual reads. Taxonomic profiling produces abundance profiles, which are reports of the estimated abundance of each taxa in a metagenomic sample [1, 3].

There are many metagenomic classifiers that have been benchmarked throughout the years [3, 4, 5], yet for the sake of understanding the contribution in Chapter 6 and Chapter 8, only three of them will be briefly introduced: Kaiju [6], KrakenUniq [7] and Centrifuge [8].

5.1.1 Kaiju

Kaiju is a DNA-to-protein classifier[1], that implements a search strategy to find Maximal Exact Matching (MEM) substrings between a query and a database by using a modified version of the backwards search algorithm in the Burrows–Wheeler Transform (BWT)[3, 6]. The BWT is a text compression method that converts a reference sequence database into an easily searchable representation and allows for exact string matching in time proportional to the length of the query. Kaiju uses MEMs to quickly find sequences in the reference database that share the longest possible sub-sequence with the query [6]. Backtracking through the BWT has been made faster by using a lookup table for occurrence counts of each alphabet letter, which was first proposed by [9] and is called FM-index.

Kaiju translates each read into the six possible reading frames, which are then split at stop codons into amino acid fragments. These fragments are sorted by length, and searched for Maximal

[1]Taxonomic classifier that compares DNA sequences against a database of proteins

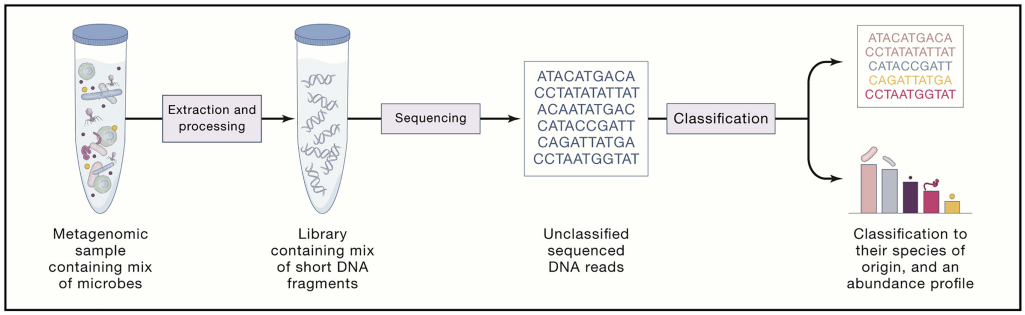


Figure 5.1: **From a metagenomic sample of diverse microbes to an abundance profile.** Image taken from [3]

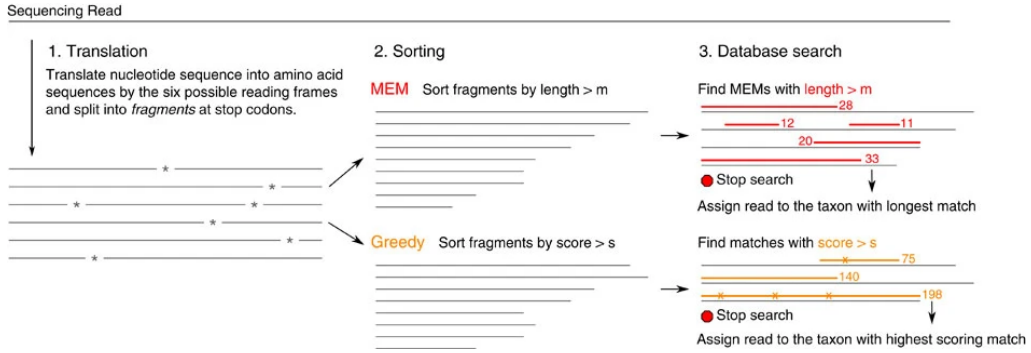


Figure 5.2: **Kaiju's graphical representation of its pipeline.** Image taken from [6].

Exact Matching (MEM)s in the FM-index from the longest to the shortest fragment. Queries are taxonomically assigned to the longest MEM. If equally long matches are found for multiple taxa, Kaiju breaks the tie by determining the Lowest Common Ancestor (LCA)[2] and outputting its taxon identifier. Kaiju also implements a greedy search mode which allows some mismatches at the left end of fragments, by searching backward in the BWT [3]. This algorithm always classifies each read to the lowest possible taxonomic level [6].

5.1.2 KrakenUniq

KrakenUniq is an algorithm based on another classification tool called Kraken [11], that additionally to its predecessor outputs information about the uniqueness of k-mers[3] assigned to each taxa [3, 7].

Kraken, on the other hand, is a classification method that searches for k-mers from a DNA sequence (query) in a pre-computed database that matches such k-mers to the Lowest Common Ancestor taxon of all genomes that contain that taxon (see Figure 5.3). In other words, Kraken represents any genomic sequence query as a k-mer set $K(S)$, and then maps each k-mer to the LCA of all genomes that contain such k-mer. These LCA taxa and their ancestors are represented as a weighted classification tree, where each node has a weight equal to the number of k-mers in the sequence associated with the node's taxon. Then, each Root-To-Leaf (RTL) path in the classification tree is scored by calculating the sum of all node weights along the path. The maximum scoring RTL path in the classification tree is the classification path and the initial DNA sequence (query) is assigned the label corresponding to its leaf. In case there is a tie between paths, the LCA of all those paths' leaves is selected [11].

Kraken was the first taxonomic classification software to introduce exact k-mer matching as a novel classification algorithm. Kraken provides read counts, while KrakenUniq additionally determines the k-mer coverage[4] for each taxonomic classification, a metric that allows the filtering out of false-positive reads. Indeed, there exists reads that are miss-classified to a taxonomic group as present in a sample (also called false-positive reads) when [7] one of the following situations occurs:

- They are contaminating reads. That is, they are reads that belong to contaminant sequences coming from the extraction, handling or sequencing of the samples. The issue of contaminating reads can be even more problematic when there is a scarce amount of input material.
- They are reads that belong to *low-complexity regions*[5] of genomes. For instance, if a certain number of reads match only a portion of a genome that has low-complexity, then the species was probably not present in the sample and the read classification corresponds to a false-positive.
- They are reads classified as hits to a taxon when they are actually the result of contamination in the database of genomes used by the taxonomic classifier.

The authors demonstrate that by reporting the number of unique k-mers, KrakenUniq can efficiently tackle the issue of false-positives and accurately identify species. When reads from a species yield many unique k-mers, it is possible to state more confidently that the taxon is truly present. Conversely, a scarcity of unique k-mers suggests a possible false-positive identification [7].

[2]The LCA of nodes $\boldsymbol{u}$ and $\boldsymbol{v}$ in a tree is the ancestor of $\boldsymbol{u}$ and $\boldsymbol{v}$ that is located farthest from the root [10]

[3]a k-mer is a sub-string of a genomic sequence of length k. More information on this concept will be introduced in upcoming subsections.

[4]Here k-mer coverage refers to the number of unique k-mers per clade divided by genome size [12]

[5]Low-complexity regions are defined as regions of biased composition containing simple sequence repeats [13]

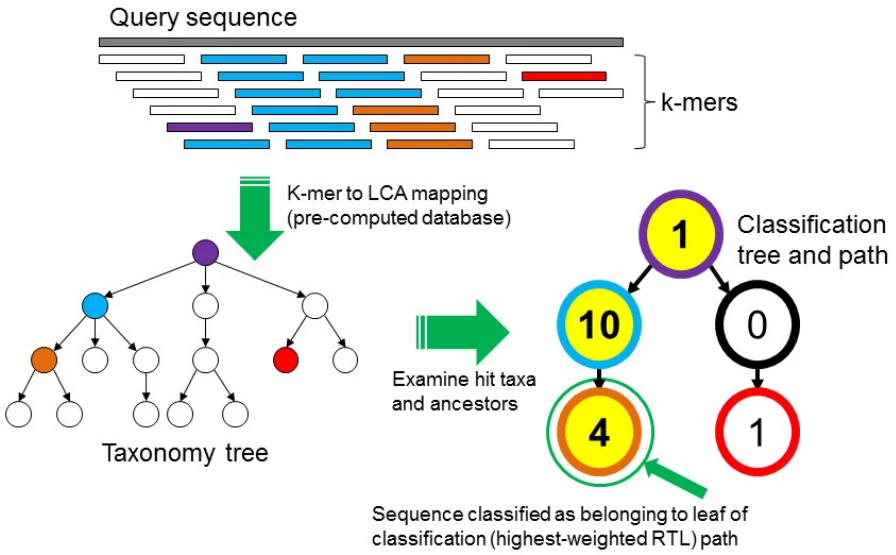


Figure 5.3: **Kraken's graphical representation of its pipeline.** Image taken from [11]

The strategy of unique k-mer counting allows the method to detect reads spread across a genome without alignment.

It is important to note that the authors of both methods (Kraken and KrakenUniq) recommend using KrakenUniq instead of Kraken nowadays [14].

5.1.3 Centrifuge

This metagenomics classifier, created by an team that shares a subset of authors with Kraken, uses the Burrows–Wheeler Transform and FM-index to store and index the genome database [11]. It was released in 2016, and it addresses the memory limitations of Kraken by generating smaller databases based on an FM-index and compression of within-species genomes [15].

Contrasting Kraken, Centrifuge employs a completely different classification algorithm. To begin with, it compresses multiple genomes of the same species by storing near-identical sequences only once, thereby achieving a substantial space reduction for numerous species. Subsequently, an FM-index is constructed from these compressed sequences.

Once the FM-index is built, the sequence classification process begins (see Figure 5.4). Centrifuge takes both the forward and reverse sequences and searches for Maximal Exact Matching (16 bp minimum) in the FM-index, and extends the matches as far as possible. Based on the identified exact matches, the algorithm classifies each read using only those mappings with at least one 22-bp match. Centrifuge scores each species using the formula 5.1, and assigns a sequence to multiple taxonomic categories (5 by default). However, the algorithm reduces the number of assignments by traversing up the taxonomic tree and selecting the genus encompassing the largest number of species. Consequently, if there exists a match for every species from a given genus, the genus is used as the assignment, rather than every matched species that is part of such genus (See Figure 5.4) [8].

$$Score(SpeciesX) = \sum_{hit \in SpeciesX} (length(hit) - 15)^2 \tag{5.1}$$

In comparison to Kraken 2, this taxonomic classifier has several distinctive characteristics [14]:

- Centrifuge uses slightly less memory thanks to the FM-index within species compression.
- Centrifuge can give multiple assignments per read (unlike Kraken 2 that gives just one assignment).

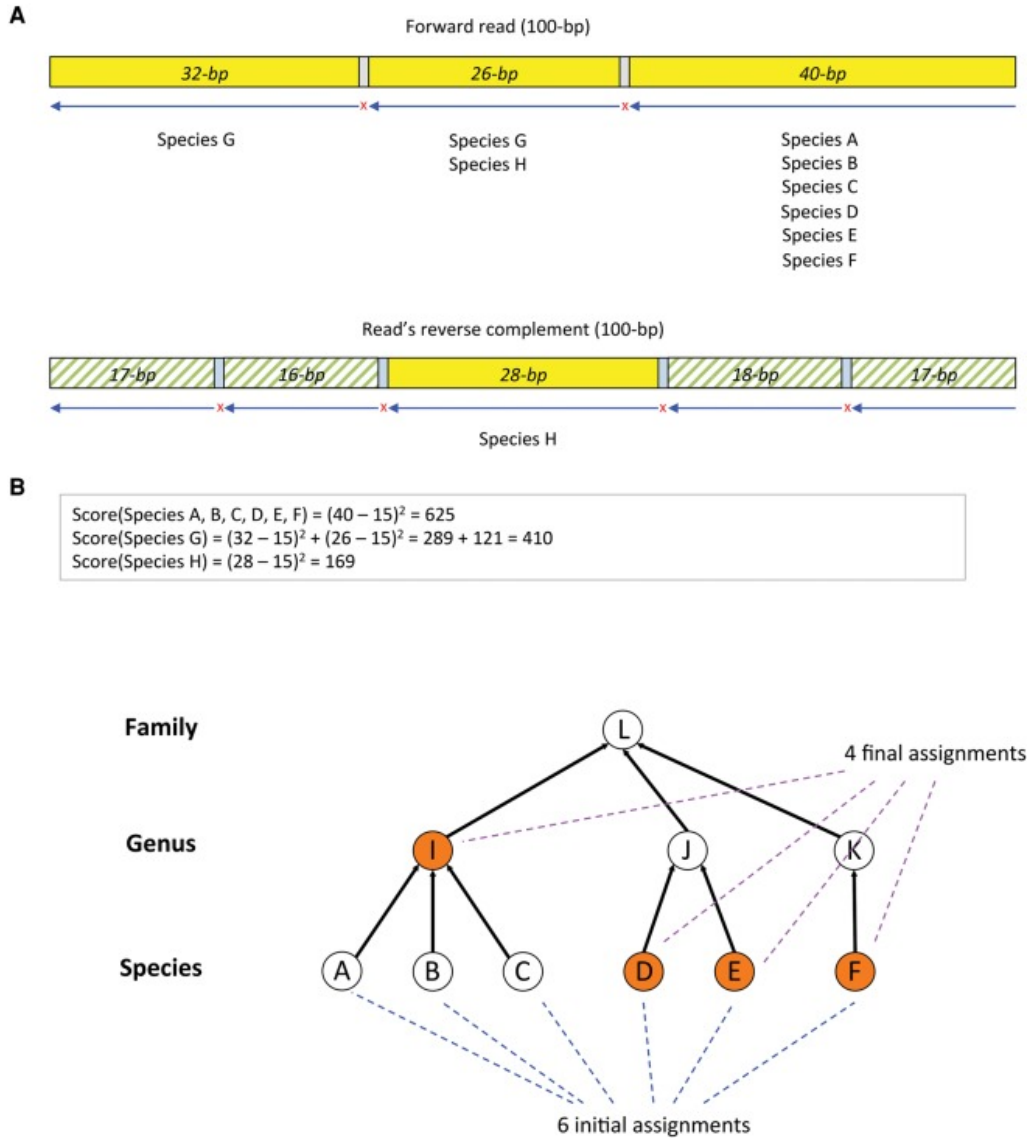


Figure 5.4: **Centrifuge's graphical representation of its pipeline.** Species A,B, and C were found to be a match. Yet, after traversing up the tree, the assignment is attributed to genus I rather than species A, B and C, all of which belong to genus I. Image taken from [8]

- Kraken 2 analyses all the k-mers of the same length in a read, while Centrifuge starts with a 16 bp minimum exact match as a seed, and extends this as far as possible. When a mismatch is found, Centrifuge skips the base and tries to find the next exact match.
- Centrifuge is slower in classification, and requires more time for database building.
- Kraken 2 supports more databases than Centrifuge.

5.2 Ancient metagenomic workflows

Several tools and databases that can be employed for meta-taxonomic analyses of ancient samples [16]. One notable alignment-based method, known as MEGAN ALignment Tool (MALT) [17], was introduced in 2016 and it is one of the most commonly used tools in aDNA analyses. MALT combines alignment and taxonomic binning by first generating an index on a reference database (provided by the user), and subsequently aligning query sequences against said reference database. Once all alignments for a certain read have been estimated, this software finds the Lowest Common Ancestor (LCA) to perform taxonomic binning. The output of MALT can be integrated with the interactive metagenomic analysis software MEGAN[6][18].

[6]Toolbox specifically designed for metagenomic studies, encompassing taxonomic analyses, functional analyses, visualisations, clustering, dimensionality reduction, and various other functionalities. It is important to note that MEGAN does not involve any form of alignment.

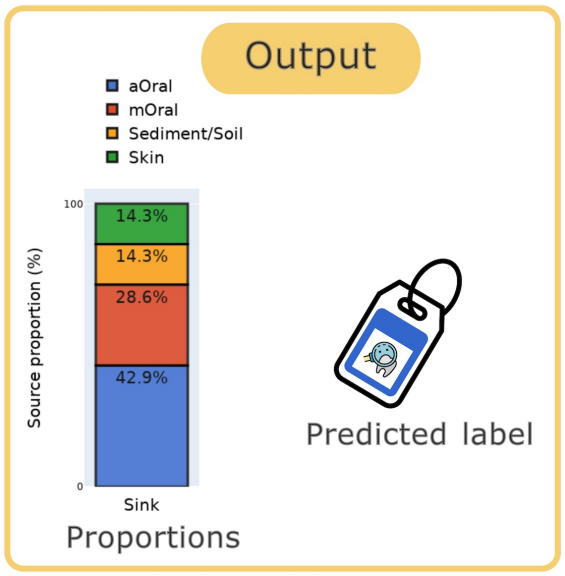


Figure 5.5: **Example of the composition of a metagenomic sample isolated from dental calculus as estimated via Microbial Source Tracking (MST).** The microbial composition of a metagenomic sample of interest (also called sink) isolated from dental calculus can be modelled as a mixture of DNA originating from the following sources: ancient oral (aOral), modern oral (mOral), skin, soil and unknown microbes. This framing of the problem where the composition of a sink is modelled based on the contribution of a set of sources is known as Microbial Source Tracking

Published 3 years after MALT, Heuristic Operations for Pathogen Screening (HOPS)[19] is a pipeline that consists of three steps: MALT alignment with aDNA, usage of MaltExtract to provide statistics for the evaluation of species identification as well as aDNA authenticity, and a post-processing script for visualisation.

Another popular general-purpose aDNA pipeline, nf-core/eager [20], incorporates HOPS as an ancient microbiome profiling module within its framework. In addition to taxonomic profiling, this pipeline enables pathogen screening, microbiome reconstruction, authentication, and analysis of microbial genomes [21].

Recently, in 2023, a metagenomic profiling workflow for aDNA called aMeta [22] was released to minimise false discoveries and reduce computer memory requirements. aMeta demonstrated superior accuracy and memory efficiency when compared to HOPS (hence better than nf-core/eager and MALT). This profiling workflow combines taxonomic classification (via KrakenUniq) and filtering to establish a list of microbial candidates used to build a MALT database, and includes several validation and authentication steps based on the resulting alignments[22]. The purpose of this method is to link the capacity of KrakenUniq to work with large databases with the advantages MALT for result validations via alignment.

It is important to emphasise that all the reference-based methods described in Section 5.1 and Section 5.2, have the inherent limitation of being unable to identify microbial organisms that are not present in the reference database used for alignment. As an additional point, a more comprehensive, yet non-exhaustive, list of computational tools and pipelines for aDNA studies can be found in [21].

5.3 Microbial Source Tracking

Microbial Source Tracking (MST) refers to the modelling of microbial communities in a way that the composition of a metagenome *sink* (sample of interest) is the product of different contributing metagenome *sources* [16, 23]. For example, the microbial composition of archaeological dental calculus (sink), can be modelled as a mixture of DNA originating from dental plaque (source), skin bacteria(source), soil(source), and other unknown sources [24] (see Figure 5.5).

#OTU	SRR1804823	SRR11176636	SRR12557007	SRR988088	ERR687893	SRR2096709
taxa_1	0	5	0	20	100	1
taxa_2	15	5	0	0	44	0
taxa_3	0	13	200	0	3	12
taxa_4	4	5	0	0	0	33

Table 5.1: Example of a taxonomic abundance table.

SampleID	Env	SourceSink
SRR1804823		Sink
SRR11176636	aOral	Source
SRR12557007	aOral	Source
SRR988088	Soil/Sediment	Source
ERR687893	Soil/Sediment	Source
SRR2096709	Skin	Source

Table 5.2: Example of a sink/source table required by both FEAST andmSourceTracker. The environment label for the sink sample (SRR1804823) is unknown and is left to be predicted by the Microbial Source Tracking method.

5.3.1 Input data

In order to use the MST software described in the following subsections, the user has to supply two tables:

- A table that indicates the samples to be used as sources and their origin (i.e. different class environments, for instance, sample A belonging to Soil, sample B belonging to Oral, sample C belonging to Skin) and the sinks (sequenced samples of interest). See the example in Table 5.2.
- A taxonomic abundance table [7]. Such a table can be constructed through the use of taxonomic classifiers such as Kaiju or KrakenUniq. Given a set of unique taxa $O = \{o_1, o_2, o_3..., o_i, ...o_m\}$ of size m, and a set $S = \{s_1, s_2, s_3, ..., s_j, ...s_n\}$ of sample identifiers of size n, a taxonomic abundance table $T(m \times n)$ contains on each element $T(i, j)$ the abundance of the taxa i in the sample j. See example in Table 5.1.

Taxonomic profiling is required for the construction of the taxonomic abundance table used as input by both FEAST and mSourceTracker, and is this characteristic that makes both methods reference-based. The need for a genomic database to run the taxonomic classifiers is limiting to the study of ancient dental calculus: community composition analyses may be missing taxa and underestimating diversity, especially for samples coming from underrepresented locations [26].

5.3.2 Discussion on how to build taxonomic abundance tables

During the process of writing the paper for decOM, numerous discussions took place among the co-authors and reviewers regarding the appropriate construction of a taxonomic abundance table that would serve as input for competing methods such as FEAST and mSourceTracker. As detailed in Section 5.1, there are several benchmarking papers that compare and evaluate algorithms for taxonomy classification in modern data exclusively. Yet, to the best of our knowledge, there is only one benchmarking paper that evaluates the performance of five different metagenomic classifiers (QIIME/UCLUST, MetaPhLAn2, MIDAS, CLARK-S, MALT) on synthetic data specifically designed to emulate ancient dental plaque [27]. This study **does not exhibit unequivocal superiority of any one program over another** but rather highlights that most program biases can be attributed to database construction, which is generally dominated by human-associated bacteria. This further underscores the limitations of using reference-based methods in fields where samples that are underrepresented in public databases play a crucial role, such as the study of ancient microbes.

Unfortunately, by the time we encountered this benchmarking paper of taxonomic classifiers for synthetic ancient dental calculus, we had already constructed the taxonomy-based tables using KrakenUniq and Kaiju. Not only did this process consume a significant amount of time (see subsection 6.5.2), but also there is not a clear consensus on how to build a taxonomic abundance table. In light of this situation, we can't add much more to the analysis of results by overlapping our results with that benchmarking. However, it might be possible that in the future someone

[7] This is often called also Operational Taxonomic Unit (OTU) table. An OTU refers to groups of sequences that are intended to correspond to taxonomic clades or monophyletic groups [25]

migh challenge `decOM`'s capabilities for the multi-class classification task evaluated in our paper, by using alternative ways to produce the input data for FEAST and/or mSourceTracker. As it will be later detailed in Chapter 6, we tried out two taxonomic classifiers to produce such tables and yet showed the superior performance of our method. Additionally, the lack of a clear consensus on the optimal strategy to construct taxonomic abundance tables, supports our assertion on the limitations of using MST methods such as FEAST and mSourceTracker and further encourages the development of methods that do not rely on reference databases.

Lastly, there is an ongoing endeavour by a classifier committee to establish the ultimate benchmarking framework for aDNA meta-taxonomic classification as part of the Standards, Precautions, and Advances in Ancient Metagenomics (SPAAM) initiative. This initiative, which comprises a community of researchers dedicated to ancient metagenomics, aims to advance scientific knowledge, provide training and support, and foster networking opportunities for individuals in the field of palaeometagenomics.

5.4 SourceTracker

SourceTracker[28] is a method published in 2011, and it is perhaps the most popular and widely used tool for Microbial Source Tracking. The core idea of this method is that it considers each sink to be the product of a mixture of known sources and an unknown source, assigning sequences to different source environments based on their conditional distribution. SourceTracker estimates the composition of each sink in an iterative manner. It randomly assigns each taxon to a source environment and then estimates the current proportions of the source environments in the test sample. It then reassigns each sequence based on the conditional distribution until convergence is reached[8]. Let us first introduce a few concepts before digging into the source tracking algorithm used by this method.

5.4.0.1 Latent Dirichlet Allocation (LDA)

The authors postulate that Microbial Source Tracking is analogous to inferring the mixing proportions of conversation topics in a document, a task that has already been tackled with a probabilistic model called Latent Dirichlet Allocation (LDA) [29]. Originally, LDA was used for modelling the problem of having a scientific paper that discusses multiple topics, and how the words that appear in such a paper reflect the particular set of topics it addresses [29]. Latent Dirichlet Allocation was also introduced in parallel during the 2000s in the context of population genetics, to infer population structure and assign individuals to populations. In this case, each population is characterised by a set of allele frequencies at each locus, and individuals in the sample are assigned probabilistically to populations, or jointly to two or more populations if their genotypes indicate that they are admixed [30]. Similarly, a sample (sink) can be the result of the contribution of multiple environments, and the taxa that appear in that sample(sink) reflects the particular set of contributing environments. Viewing samples as mixtures of probabilistic sequences (taxa) makes it possible to formulate the problem of discovering the set of taxa that belong to each environment.

Consider the following notation:

- z_i is the latent variable indicating the environment from which the i^{th} taxon was drawn.
- $P(w_i|z_i = j)$ is the probability of the taxon w_i under the j^{th} environment.
- $P(z_i = j)$ gives the probability of choosing a taxon from environment j in the current sample. This probability will vary across different samples.
- $P(w|z)$ indicates which taxa are important to an environment.
- $P(z)$ is the prevalence of an environment within a sample.

The probability of the i^{th} taxon in a given sink can be written as [31]:

$$P(w_i) = \sum_{v=1}^{V} P(w_i|z_i = j)P(z_i = j) \tag{5.2}$$

Given D samples containing V environments expressed over W unique taxa, we can:

- Represent $P(w|z)$ with a set of V multinomial distributions ϕ over the W taxa, such that $P(w|z = j) = \phi^{w}_{(j)}$.

[8]This concept will be explained in section 5.4.0.2

- Represent $P(z)$ with a set of D multinomial distributions θ over the T environments, such that for a taxon in sample d, $P(z = j) = \theta_j^{(d)}$.

Latent Dirichlet Allocation combines Equation 5.2 with a prior probability distribution on θ to provide a generative model for each sink (as a reminder, the sample from which we want to assess the level of contamination). Sinks are generated by first picking a distribution over environments θ from a Dirichlet distribution, which determines $P(z)$ for taxa in that sink. The taxa distribution in the sink is then obtained by picking an environment j from this distribution and then picking a taxon from that environment according to $P(w|z = j)$, which is determined by a fixed $\phi^{(j)}$. The estimation problem consists of maximising [29]:

$$P(\boldsymbol{w}|\phi, \alpha) = \int P(\boldsymbol{w}|\phi, \theta)P(\theta|\alpha)\delta\theta \tag{5.3}$$

Where $P(\theta)$ is a Dirichlet distribution. Because the integral in this expression is intractable[9], ϕ is estimated via techniques such as Gibbs sampling.

The complete LDA can be written as [29]:

$$\begin{aligned} w_i|z_i, \phi^{(z_i)} &\sim Discrete(\phi^{(z_i)}) \\ \phi &\sim Dirichlet(\beta) \\ Z_i|\theta^{(d_i)} &\sim Discrete(\theta^{(d_i)}) \\ \theta &\sim Dirichlet(\alpha) \end{aligned} \tag{5.4}$$

The Dirichlet parameters, α and β, are prior counts that *"smooth the distributions for low-coverage source and sink samples, respectively"* [28]. In the SourceTracker paper all inferences performed set both α and β to be 0.0001, yet they are both parameters of the model that are left for the user to optimise. For LDA, one takes values that are below 1 to enforce that the posterior discrete distributions are only with a limited number of weights different from 0 (the classical parametrisation is to take 1% of the number of environments and taxa). Note that it is different from traditional estimation with Dirichlet prior where α is greater that 1 to avoid zero counts.

5.4.0.2 Gibbs sampling

Markov Chain Monte Carlo (MCMC) methods are computational techniques designed to generate samples from a given probability distribution $P(\theta)$ (also called target density), and/or to estimate expectations of functions under this distribution. When $P(\theta)$ is not from a simple analytical form, MCMC methods are ideal to overcome the impossibility of evaluating these expectations by exact methods. Gibbs sampling is an instance of MCMC methods that assumes that even though $P(\theta)$ is too complex to draw samples from it directly, its conditional distributions $P(\theta_i|\{\theta_i\}_{i\neq j})$ might be tractable (of a simple analytical form). The application of Gibbs sampling to topic modelling for the LDA model has already been implemented before the application for MST [31]. Gibbs sampling is used in the context of Microbial Source Tracking to explore the distribution of assignments of taxa to source environments within a given sink.

The idea in Gibbs sampling is to generate posterior samples sweeping through each variable and fixing the remaining variables to their current values, this means, the sampling is not done on $P(\theta)$ itself, but samples are simulated by going through all the posterior conditionals, one random variable at a time [32]. Gibbs sampling is used by SourceTracker by assigning each observation of each taxon to a random source environment, and leaving one taxon out for estimation. The initial assumption is that these assignments are correct (even though they are random), and given this assumption it is fairly easy to estimate that the taxon that we left out came from a known or an unknown source. After removing the aforementioned taxon and re-selecting its source environment assignment, SourceTracker updates the tally for the selected source environment, and repeats the process on another randomly selected taxon.

Because the construction of the sampling is done according to an irreducible[10] Markov Chain, Gibbs sampling guarantees to reach *convergence*, or reach a stationary state, where the sample values have the same distribution as if they were sampled from the true posterior joint distribution [32, 34]. Put differently, the algorithm converges on the actual distribution of true assignments from the different source environments.

[9] Solution is computationally prohibitive or the integral has no closed-form solution.

[10] If a chain has any one state that is reachable from anywhere, then the chain is irreducible. In other words, a Markov Chain is irreducible if and only if its graph representation is a strongly connected graph [33].

5.4.1 Main algorithm

SourceTracker considers the following notation:

- Each sink sample $\boldsymbol{x}$ consists of n sequences mapped to taxa.
- Each sequence can be assigned to any of the sources environments $v \in 1 \dots V$, including an unknown source.
- These assignments are treated as hidden variables, denoted as $z_{i=1 \dots n} \in 1 \dots V$.

.

The initial step of the Gibbs sampling process is to initialise z with random source environment assignments and then iteratively reassign each sequence based on the conditional distribution [28]:

$$P(\boldsymbol{z}_i = v | \boldsymbol{z}^{\neg i}, \boldsymbol{x}) \propto P(\boldsymbol{x}_i | v) \times P(v | \boldsymbol{x}^{\neg i}) = \left(\frac{m_{x_i v} + \alpha}{m_v + \alpha m_v} \right) \times \left(\frac{n_v^{\neg i} + \beta}{n - 1 + \beta V} \right) \tag{5.5}$$

In Equation 5.5:

- m_{tv} is the number of training sequences from taxon t in environment v.
- n_v is the number of test sequences currently assigned to environment v.
- $\neg i$ excludes the i^{th} sequence.
- α represents a prior count that smooths the distribution for low-coverage source samples.
- β represents a prior count that smooths the distributions for low-coverage sink samples.

The first fraction gives the posterior distribution over taxa in the source environment; the second gives the posterior distribution over source environments in the test sample.

5.4.2 The contribution of mSourceTracker to SourceTracker

In 2020 metagenomic-SourceTracker (mSourceTracker)[35] appeared as an extension of SourceTracker, by testing the effectiveness of the latter in metagenomic data and adding a diagnostic tool to determine the reliability of the proportion estimates [35]. This option can be used to check if convergence does not take place due to poor taxonomic coverage [36]. Notice that the algorithm that assigns abundance estimates for each source environment **is the same** for SourceTracker and mSourceTracker. The main contributions of the paper for mSourceTracker were:

- The authors proved the effective application of SourceTracker on metagenomic data
- The convergence tests and visualisations implemented via the diagnostic tool helped identify when convergence was not occurring, caused mainly due to poor taxonomic coverage.

5.5 FEAST

mSourceTracker and SourceTracker made an important contribution to the field, yet using Gibbs sampling for parameter estimation is a computationally expensive procedure, only applicable to small datasets with few sources. To address these limitations, other authors developed Fast Expectation-Maximization Microbial Source Tracking (FEAST). This method proved to be efficient on metagenomic datasets and can estimate thousands of source contributions to a sample.

5.5.1 The probabilistic model

The model used by SourceTracker shares many similarities with FEAST, and the main difference between both methods lies in their optimisation procedure (Gibbs Sampling is used by the former, and Expectation-Maximization is used by the latter).

Consider the following notation:

- K is the number of known sources. There are a total of K + 1 sources (including the unknown). Each sink is represented by a vector $\boldsymbol{x}$, where x_j corresponds to the abundance of taxa j, where $1 \leq j \leq N$.
- Every known source is represented by a vector $\boldsymbol{y}_i$, where y_{ij} is the observed abundance of taxa j in source i ($1 \leq i \leq K$).

- $C_i = \sum_{j=1}^{N} y_{ij}$ is the total taxa counts of the known sources.
- $C = \sum_{j=1}^{N} x_j$ is the total taxa counts of the sink.

The model can be described by the following terms [23]:

$$\begin{aligned} \beta_j &= \sum_{j=1}^{K+1} \alpha_i \gamma_{ij} \\ \boldsymbol{y}_i &\sim Multinomial(C_i, (\boldsymbol{\gamma}_{i1}, ..., \boldsymbol{\gamma}_{iN})) \\ \boldsymbol{x} &\sim Multinomial(C, (\beta_1, ..., \beta_N)) \end{aligned} \tag{5.6}$$

Where:

- $\boldsymbol{\alpha}$ is a vector of length K+1, where $\boldsymbol{\alpha_i}$ is to the fraction of source i in the sink. Since the contribution of all sources to the composition of a sink expressed in proportions should add up to 1, $\sum_{i=1}^{K+1} \boldsymbol{\alpha}_i = 1$. This vector is not observed and it is a parameter of the model.
- $\boldsymbol{\gamma}_{ij}$ represents the relative abundance of taxa j in source i. For each source there is a vector $\boldsymbol{\gamma}$, where $\sum_{j=1}^{N} \boldsymbol{\gamma}_{ij}$. This vector is not observed and it is a parameter of the model.

5.5.1.1 Expectation-Maximization (EM) algorithm

This algorithm alternates between the steps of guessing a probability distribution over completions of missing data given the current model (known as the E-step) and then re-estimating the model parameters using these completions (known as the M-step). The EM algorithm attempts to find the parameters that maximise the probability of observing the data, in other words, we want to estimate the model parameters ($\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ in expression 5.6) for which the observed data are most likely [37, 38].

Let $\boldsymbol{X}$ be a random vector that results from a parameterised family. We wish to find θ such that $P(X|\theta)$ is a maximum. This is known as the Maximum Likelihood (ML) estimate for θ. To to estimate θ, it is typical to introduce the log-likelihood function defined as:

$$L(\theta) = \ln P(\boldsymbol{X}|\theta) \tag{5.7}$$

Since $\ln(x)$ is a strictly increasing function, the value of θ which maximises $P(\boldsymbol{X}|\theta)$ also maximises $L(\theta)$. The Expectation-Maximization (EM) algorithm is a two-step iterative procedure for maximising $L(\theta)$. Denote the hidden random vector by $\boldsymbol{Z}$ and a given realisation by $\boldsymbol{z}$, which is used to express the total probability $\mathcal{P}(\boldsymbol{X}|\theta)$[38].

The Expectation-Maximization algorithm consists of iterating the [38]:

1. **E-step:** Determine the conditional expectation $E_{\boldsymbol{Z}|\boldsymbol{X},\theta_n}\{\ln \mathcal{P}(\boldsymbol{X}, z|\theta)\}$. In this step, the missing data are estimated given the observed data and the current estimate of the model parameters.

   The expected complete log-likelihood function $Q(t)$[11] is given by [23]:

$$\begin{aligned} Q(t) = \sum_{i=1}^{K+1} \sum_{j=1}^{N} x_j p(i|j) \dot{\log}(\boldsymbol{\alpha}_i \boldsymbol{\gamma}_{i,j}) + \sum_{i=1}^{K+1} \sum_{j=1}^{N} y_{ij} \log(\boldsymbol{\gamma}_{ij}) + const \\ \text{Where } p(i|j) = \frac{\boldsymbol{\alpha}_i^{(t)} \boldsymbol{\gamma}_{ij}^{(t)}}{\sum_{i=1}^{K} \boldsymbol{\alpha}_i^{(t)} \boldsymbol{\gamma}_{ij}^{(t)}} \end{aligned} \tag{5.8}$$

2. **M-step:** Maximise the conditional expectation with respect to θ. In other words, the likelihood function is maximised under the assumption that the missing data are known. The estimate of the missing data from the E-step is used in place of the missing data.

   The update for the mixing proportions is given by [23]:

$$\boldsymbol{\alpha}_i^{(t+1)} = \sum_{j=1}^{N} \frac{x_j}{C} \frac{\boldsymbol{\alpha}_i^{(t)} \boldsymbol{\gamma}_{ij}^{(t)}}{\sum_{i=1}^{K+1} \boldsymbol{\alpha}_i^{(t)} \boldsymbol{\gamma}_{ij}^{(t)}} \tag{5.9}$$

Convergence of the algorithm is guaranteed only to a local maxima, so FEAST is certainly faster but it might provide different results for different runs. The Expectation and Maximisation steps for FEAST are fully derived in the Methods and Supplementary Material of their paper[23].

[11] Log likelihood function Q at step t

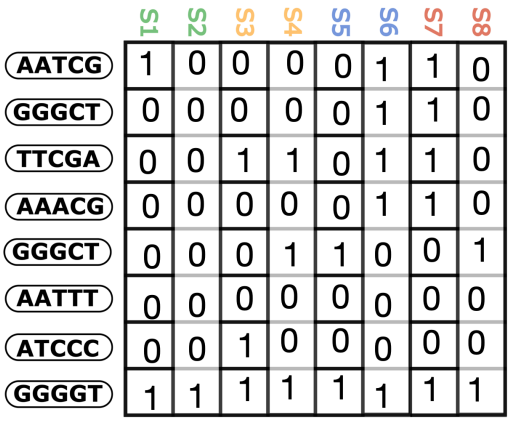


Figure 5.6: **Graphical representation of an example of a binary k-mer matrix.** This was the type of matrix used for `decOM`'s implementation. In the matrix $M_{i,j}$, each row corresponds to a distinct k-mer, and each column to a distinct metagenomic sample. If there is a 1 in the entry $M_{i,j}$, then the k-mer i is present in the sample j, else there is a zero in that position.

5.6 k-mers and k-mer matrices

5.6.1 k-mers

Given a biological sequence S of length L, a k-mer of S is a sub-string of S of length k, usually with $20 \leq k \leq 40$ [39]. One common way to file genomic collections is to store and index data sets as sets of k-mers. Any genomic sequence or genomic dataset (set of reads resulting from the sequencing an individual sample) can be represented as a k-mer set [39, 40].

Algorithms that rely on k-mers are commonly used in bioinformatics to construct an index of all k-mer sets and facilitate basic presence/absence queries. This is achieved by dividing the query sequence into k-mers and determining their presence or absence in the index. A k-mer representation of genomic sequences is a succinct solution, and it is an efficient way of dealing with sequencing errors compared to exact alignment. Unlike aligners that perform inexact pattern matching, k-mer-based methods can examine the matching fraction of k-mers within the query. Consequently, the bioinformatics community has embraced these type of methods, since they enable large comparisons between extensive datasets [40].

5.6.2 *k*-mer matrices

A k-mer matrix is a data structure that allows the representation of sequence content across multiple experiments, via presence/absence or abundance of each sub-sequence of fixed size in all samples. Given a collection S of N samples and size of k-mers of k, a k-mer matrix M contains on each element $M(i, j)$ the abundance or presence/absence (binary) of the k-mer i in the sample j. Since a k-mer matrix represents presence/absence or abundance of k-mers across several samples, its construction relies on k-mer counting [39].

For the construction of the k-mer matrices discussed throughout this thesis, we made use of a tool called `kmtricks` [41]. This software was designed to construct k-mer matrices by relying on disk-based[12] k-mer counting techniques, and it is the first to formalise the concept of k-mer matrices. `kmtricks` consists of a pipeline composed of successive stages that allow step-by-step construction of the matrix of interest [39]. In few words, `kmtricks` takes as input several genomic datasets (that can occupy up to terabytes of memory), and counts k-mers across those samples in a way that is four times faster than state-of-the-art tools [41], producing among other things [13], a k-mer matrix representation of the data. This method performs k-mer counting across multiple metagenomic samples, and creates the base data structure used in `decOM`.

To obtain optimal performance, `kmtricks` uses the concept of partitions. Partitions are sample divisions that are ideal for parallelization, and in `kmtricks` there is a mandatory condition of having all partitions contain roughly equal total number of k-mers. `decOM` relies on `kmtricks` and benefits from several of its features, notably, each input sink is counted using the same partitioning scheme as the sources (represented by a k-mer matrix such as the one in Figure 5.6). The comparison

[12] Disk-based approaches are based on the divide-and-conquer paradigm. The notion behind this is to form groups of k-mers, and process these groups successively or in parallel depending on memory allocation [39].

[13] `kmtricks` can also output a collection of Bloom Filter, one per sample. This concept will be explained later in 7.4

relies on the fact that identical k-mers from the sources and sinks belong to the same partition [39]. Interestingly, we managed to prove that despite considering only one partition, the Microbial Source Tracking results for `decOM` were good enough to beat competing methods (see results the performance of `decOM` when taking a larger number of partitions in Figure A.11, and section 6.6).

Notably, the construction of the k-mer matrix used by `decOM`, with approximately 14 million unique k-mers (the size of one partition), required 2336 MB of memory and 16.4 hours of running time. Larger matrices were tested, yet once the results achieved were optimal, a smaller matrix size was preferred. There were two additional parameters used in the construction of the k-mer matrix of sources: k-mers were kept if they were present in at least 3 samples of the collection (*recurrence_min*), and all k-mers that were seen only once in a sample (*abundance_min*) were removed, as they were most likely sequencing errors.

References

[1] Key, F. M., Posth, C., Krause, J., Herbig, A., and Bos, K. I. "Mining metagenomic data sets for ancient DNA: recommended protocols for authentication". In: *Trends in Genetics* vol. 33, no. 8 (2017), pp. 508–520.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. "Basic local alignment search tool". In: *Journal of molecular biology* vol. 215, no. 3 (1990), pp. 403–410.

[3] Simon, H. Y., Siddle, K. J., Park, D. J., and Sabeti, P. C. "Benchmarking metagenomics tools for taxonomic classification". In: *Cell* vol. 178, no. 4 (2019), pp. 779–794.

[4] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. "Critical assessment of metagenome interpretation—a benchmark of metagenomics software". In: *Nature methods* vol. 14, no. 11 (2017), pp. 1063–1071.

[5] McIntyre, A. B., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., et al. "Comprehensive benchmarking and ensemble approaches for metagenomic classifiers". In: *Genome biology* vol. 18, no. 1 (2017), pp. 1–19.

[6] Menzel, P., Ng, K. L., and Krogh, A. "Fast and sensitive taxonomic classification for metagenomics with Kaiju". In: *Nature communications* vol. 7, no. 1 (2016), pp. 1–9.

[7] Breitwieser, F. P., Baker, D., and Salzberg, S. L. "KrakenUniq: confident and fast metagenomics classification using unique k-mer counts". In: *Genome biology* vol. 19, no. 1 (2018), pp. 1–10.

[8] Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. "Centrifuge: rapid and sensitive classification of metagenomic sequences". In: *Genome Research* vol. 26, no. 12 (2016), pp. 1721–1729.

[9] Ferragina, P. and Manzini, G. "Opportunistic data structures with applications". In: *Proceedings 41st annual symposium on foundations of computer science*. IEEE. 2000, pp. 390–398.

[10] Bender, M. A., Farach-Colton, M., Pemmasani, G., Skiena, S., and Sumazin, P. "Lowest common ancestors in trees and directed acyclic graphs". In: *Journal of Algorithms* vol. 57, no. 2 (2005), pp. 75–94.

[11] Wood, D. E. and Salzberg, S. L. "Kraken: ultrafast metagenomic sequence classification using exact alignments". In: *Genome biology* vol. 15, no. 3 (2014), pp. 1–12.

[12] Breitwieser, F. P. *KrakenUniq: Metagenomics classifier with unique k-mer counting for more specific results*. 2023. URL: `%5Curl%7Bhttps://github.com/fbreitwieser/krakenuniq/blob/cdfabbe3f8d7e5dd09a97101d1738394fd9756ae/src/taxdb.hpp#L1103C1-L1104C1%7D`.

[13] Orlov, Y. L. and Potapov, V. N. "Complexity: an internet resource for analysis of DNA sequence complexity". In: *Nucleic acids research* vol. 32, no. suppl_2 (2004), W628–W633.

[14] Lu, J., Breitwieser, F. P., Wood, D. E., Song, L., Kim, D., Langmead, B., Pockrandt, C., and Salzberg, S. L. *How to Choose Your Metagenomics Classification Tool*. URL: `http://ccb.jhu.edu/software/choosing-a-metagenomics-classifier/`. (accessed: 29.08.2023).

[15] Breitwieser, F. P., Lu, J., and Salzberg, S. L. "A review of methods and databases for metagenomic classification and assembly". In: *Briefings in bioinformatics* vol. 20, no. 4 (2019), pp. 1125–1136.

[16] Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiß, C. L., Burbano, H. A., Orlando, L., and Krause, J. "A robust framework for microbial archaeology". In: *Annual Review of Genomics and Human Genetics* vol. 18 (2017), pp. 321–356.

[17] Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., and Huson, D. H. "MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman". In: *BioRxiv* (2016), p. 050559.

[18] Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. "MEGAN analysis of metagenomic data". In: *Genome research* vol. 17, no. 3 (2007), pp. 377–386.

[19] Hübler, R., Key, F. M., Warinner, C., Bos, K. I., Krause, J., and Herbig, A. "HOPS: automated detection and authentication of pathogen DNA in archaeological remains". In: *Genome Biology* vol. 20, no. 1 (2019), pp. 1–13.

[20] Yates, J. A. F., Lamnidis, T. C., Borry, M., Valtueña, A. A., Fagernäs, Z., Clayton, S., Garcia, M. U., Neukamm, J., and Peltzer, A. "Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager". In: *PeerJ* vol. 9 (2021), e10947.

[21] Der Sarkissian, C., Velsko, I. M., Fotakis, A. K., Vågene, Å. J., Hübner, A., and Fellows Yates, J. A. "Ancient metagenomic studies: Considerations for the wider scientific community". In: *Msystems* vol. 6, no. 6 (2021), e01315–21.

[22] Pochon, Z., Bergfeldt, N., Kırdök, E., Vicente, M., Naidoo, T., Valk, T. van der, Altınışık, N. E., Krzewińska, M., Dalen, L., Götherström, A., et al. "aMeta: an accurate and memory-efficient ancient metagenomic profiling workflow". In: *Genome Biology* vol. 24, no. 1 (2023), p. 242.

[23] Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., and Halperin, E. "FEAST: fast expectation-maximization for microbial source tracking". In: *Nature Methods* vol. 16, no. 7 (2019), pp. 627–632.

[24] Warinner, C. "An Archaeology of Microbes". In: *Journal of Anthropological Research* vol. 78, no. 4 (2022), pp. 420–458.

[25] Edgar, R. C. "UPARSE: highly accurate OTU sequences from microbial amplicon reads". In: *Nature methods* vol. 10, no. 10 (2013), pp. 996–998.

[26] Velsko, I. M. et al. "Exploring archaeogenetic studies of dental calculus to shed light on past human migrations in Oceania". In: *bioRxiv* (2023). DOI: 10.1101/2023.10.18.563027. eprint: https://www.biorxiv.org/content/early/2023/10/19/2023.10.18.563027.full.pdf. URL: https://www.biorxiv.org/content/early/2023/10/19/2023.10.18.563027.

[27] Velsko, I. M., Frantz, L. A., Herbig, A., Larson, G., and Warinner, C. "Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research". In: *Msystems* vol. 3, no. 4 (2018), pp. 10–1128.

[28] Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. "Bayesian community-wide culture-independent microbial source tracking". In: *Nature Methods* vol. 8, no. 9 (2011), pp. 761–763.

[29] Blei, D. M., Ng, A. Y., and Jordan, M. I. "Latent dirichlet allocation". In: *Journal of machine Learning research* vol. 3, no. Jan (2003), pp. 993–1022.

[30] Pritchard, J. K., Stephens, M., and Donnelly, P. "Inference of population structure using multilocus genotype data". In: *Genetics* vol. 155, no. 2 (2000), pp. 945–959.

[31] Griffiths, T. L. and Steyvers, M. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* vol. 101, no. suppl_1 (2004), pp. 5228–5235.

[32] Yildirim, I. "Bayesian inference: Gibbs sampling". In: *Technical Note, University of Rochester* (2012).

[33] Roberts, G. O. and Rosenthal, J. S. "General state space Markov chains and MCMC algorithms". In: (2004).

[34] Resnik, P. and Hardisty, E. *Gibbs sampling for the uninitiated*. Tech. rep. 2010.

[35] McGhee, J. J., Rawson, N., Bailey, B. A., Fernandez-Guerra, A., Sisk-Hackworth, L., and Kelley, S. T. "Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics". In: *PeerJ* vol. 8 (2020), e8783.

[36] Raza, S., Kim, J., Sadowsky, M. J., and Unno, T. "Microbial source tracking using metagenomics and other new technologies". In: *Journal of Microbiology* (2021), pp. 1–11.

[37] Do, C. B. and Batzoglou, S. “What is the expectation maximization algorithm?” In: *Nature biotechnology* vol. 26, no. 8 (2008), pp. 897–899.

[38] Borman, S. “The expectation maximization algorithm-a short tutorial”. In: *Submitted for publication* vol. 41 (2004).

[39] Lemane, T. “Indexing and analysis of large sequencing collections using k-mer matrices”. PhD thesis. Université de Rennes 1, 2022.

[40] Marchet, C., Boucher, C., Puglisi, S. J., Medvedev, P., Salson, M., and Chikhi, R. “Data structures based on k-mers for querying large collections of sequencing data sets”. In: *Genome Research* vol. 31, no. 1 (2021), pp. 1–12.

[41] Lemane, T., Medvedev, P., Chikhi, R., and Peterlongo, P. “kmtricks: Efficient construction of Bloom filters for large sequencing data collections”. In: *bioRxiv* (2021).

Chapter 6

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

Camila Duitama González, Téo Lemane, Riccardo Vicedomini, Nicolas Rascovan, Hugues Richard, Rayan Chikhi

Published in *Microbiome*, November 2023, volume 11, issue 2, pp. 243–243. DOI: 10.1186/s40168-023-01670-3.

6.1 Motivation

This paper emerged as a solution to the unresolved issue of Microbial Source Tracking (MST) tailored for aDNA contamination assessment using a method that is reference-free. The field of palaeometagenomics has long been aware of the recurring problem of contamination with exogenous DNA is recurrent in their genomic data, and the estimation of such contamination might help mitigate bias in downstream bioinformatic analyses. Current reference-based methods to assess contamination levels, which are based of the MST paradigm, employ taxonomic classifiers such as the ones described in Section 5.1. These methods have the fundamental limitation in that they are unable to identify microbial organisms that are not present in the reference database used for alignment. Furthermore, there is a lack of consensus on the optimal approach for constructing taxonomic abundance tables, despite studies that benchmark such taxonomic classifiers on simulated ancient oral data (see section 5.3.2).

Given these circumstances, we have proposed an algorithm based on k-mers and have framed the Microbial Source Tracking problem as a multi-class classification problem. In this approach, we compare the vector representation of a sample of interest (referred to as the "sink") against a matrix of k-mers derived from contaminant and non-contaminant metagenomic samples. This innovative method, named `decOM`, performs MST and classification of ancient and modern metagenomic samples using k-mer matrices. Notably, `decOM` surpasses two state-of-the-art machine learning methods for source tracking, namely FEAST and mSourceTracker. We anticipate that `decOM` will prove to be a valuable tool for studies involving ancient metagenomics.

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.

6.2 Contents

6.3 Background

Ancient metagenomics is the study of multi-species genomic data from samples that have degraded over relatively long time periods[1]. Analysing ancient DNA (aDNA) is particularly challenging due to deterioration and contamination with environmental and modern contaminant DNA sequences. Deterioration refers to DNA damage, which in genetic material from fossil records usually comes in the form of depurination, nick formation and cytosine deamination [2]. Contamination refers to genetic material (ancient or modern) that does not derive from the sample of interest [3]. It can come from the microbes that are present in decaying tissue, from the soil or sediment where the samples were taken, or be an unintended consequence of manipulation during and after excavation [4, 5]. Despite following well-established standards and precautions to prevent modern DNA contamination and reduce the proportion of environmental microbial taxa [5, 6], a certain level of unwanted genetic material in the samples is unavoidable [4]. Under these circumstances, contamination assessment of aDNA samples is crucial not only to avoid misleading results after downstream analysis, but also to decide which samples are worth to be further sequenced [7].

The task of Microbial Source Tracking (MST) is to quantify the proportion of different microbial environments (sources) in a target microbial community (sink) [8]. MST enables quantification of contamination [9] in metagenomics sequencing data and to predict the metadata class of a given microbial sample. That is to say, if a researcher has sequenced their ancient metagenomic sample (sink) and collected a set of sources from environments where the sample might originate, an MST software estimates the contribution of each source to the sink, and optionally report a proportion for unknown sources. For example, if the user has sequenced sink X which is a sample composed of source environments A,B and C, MST should output percentages for the contribution made by source environments A, B and C (and an optional Unknown) that sum up to 100%.

Two of the most widely used methods today for MST in metagenomic data are metagenomic-SourceTracker (mSourceTracker)[10] and Fast Expectation-Maximization Microbial Source Tracking (FEAST) [8], which depend on previously annotated data using taxonomic abundance profiles. mSourceTracker is a metagenomic extension of the popular SourceTracker [9], a method that estimates contamination proportions using a mixture model of taxonomic profiles via Gibbs sampling. It is known that the sensitivity of SourceTracker can be improved through parameter adjustments [11], however more rigorous evaluations are still needed to fully understand the effect of adjusting multiple parameters and hyperparameters on its performance [12]. FEAST, released 8 years after SourceTracker, uses an expectation-maximisation approach that reduced the running time of SourceTracker by a factor of 30 or more. It has been reported to require parameter tuning to achieve optimal performance [13], which is a resource intensive procedure when handling large data sets.

FEAST and mSourceTracker require a reference database which is necessary to build the taxonomy-based clustering tables that both methods use as input. Indeed, in both cases, metagenomic data must be grouped into bins or clusters of sequences sharing the same taxonomic classification, an information that is not only highly dependent on the database used, but also highly biased by the limited proportion of the microbial diversity that has been already sequenced and taxonomically annotated. [14].

Finally, these taxonomy-based clustering tables can also lead to misleading results depending on the sequence similarity metric and the threshold used to define them [15]. To our knowledge, there are no reported reference-free methods for contamination assessment that use MST for large-scale metagenomic analyses [13]. In this work we seek to move away from database-dependent methods and use unsupervised approaches exploiting read-level sequence composition and the wealth of information contained in metagenomes that were previously sequenced.

Over the past years and with the decrease of sequencing costs, large databases of metagenomic collections from all sorts of environments have become available [16, 17, 18]. These metagenomic raw reads collectively require petabases of storage, which prohibits their re-analysis by most labs. This prompted the development of efficient methods for exploring the sequence information contained in these collections, via searching substrings of length k (k-mers) [19]. Such methods build an index of all k-mers and their counts over a collection of samples in the form of a k-mer matrix, where each cell of the matrix represents the abundance (or presence/absence) of a k-mer in a sample. Such matrices are a concise representation of genomic data that deals more efficiently with sequencing

errors and genetic variation [19]. Tools such as `kmtricks` [20] allow the rapid construction of k-mer matrices from massive collections of sequencing data sets.

In this study we developed a novel reference-free and k-mer-based method called `decOM` to perform MST and environmental type prediction of a given microbial sample. `decOM` was evaluated in a collection of ancient oral metagenomes with variable contamination levels. Our results show that `decOM` outperforms two of the most commonly used MST methods in the multi-class classification task of finding the most abundant source environment in a sink. We tested our methodology on a collection of 360 metagenomic data sets of ancient oral samples and its possible contaminants, in an external validation set of 254 ancient oral samples and on a simulated ancient calculus metagenome.

6.4 Implementation

6.4.1 Evaluation setting

Dental calculus or tartar is mineralized dental plaque that contains remnants of microorganisms located in the oral cavity [3], and has been established over the past few years as one of the richest sources of aDNA in the archaeological record [21]. Ancient dental calculus is a great source of biomolecules (including genetic material) that originate from the host, microbes, food and the environment [6]. Dental calculus is an important reservoir of ancient human oral microbiomes, and it offers a unique possibility to examine the links between human health, diet, lifestyle and the environment throughout the course of human evolution [22]. Due to the proven relevance of aOral samples isolated from calculus in the field of ancient paleogenomics, we decided to perform our evaluations on a collection of aOral metagenomic samples and their possible sources of contamination.

The microbial composition of a given aOral sample isolated from dental calculus has been modelled in previous studies as a mixture of DNA originating from dental plaque, skin bacteria, soil and other sources [23, 24]. For this reason, we gathered 360 metagenomic data sets of diverse environment types: ancient oral (aOral), sediment/soil, skin, or modern oral (mOral) (Figure 6.1). We used this collection of real metagenomic data to model the contribution of possible contaminants coming from sediment/soil and skin sources in a group of aOral samples. In addition, we included a set of mOral samples to assess whether our method can tell apart modern and ancient oral environments.

The run accession codes for every aOral sample were retrieved from AncientMetagenomeDir [1](v20.12: Ancient City of Nessebar), a community-curated collection of annotated ancient metagenomic sample lists and standardised metadata. Samples other than aOral were selected either because they had been used by competing MST methods or because they were labelled as aforementioned classes in well-known metagenomic databases such as curatedMetagenomicData [25], the HumanMetagenomeDB [26] or MGnify [27].

We rely on the metadata of each metagenomic sample to assign a true label (i.e. environment type), however, there is no ground truth as to what is the true proportion of aOral, mOral, sediment/soil or skin content in any of them. Several variables accessible through the metadata of each run accession are plotted in the Supplementary File (Figures 1, 2, 3 and 4).

6.4.2 Input data

Both mSourceTracker and FEAST require taxonomy-based clustering tables as input. We built these tables using Kaiju [28] and the reference database NCBI BLAST nr+euk (2021-02-24 release), a non-redundant protein database of bacteria, archaea, viruses, fungi, and microbial eukaryotes (information to download it in Supplementary File, Section 1). To exclude the possibility that the lower performance of competing methods was due to the poor quality of the input taxonomic profiles, we repeated the analyses using KrakenUniq (see Supplementary File Section 2). Also in this case, `decOM` improves over FEAST and mSourceTracker. Moreover, the latter two provide worse results compared to using Kaiju (see ROC and AUC plots in Supplementary Figure 10 and 11, respectively).

On the other hand, `decOM` takes as input a binary k-mer matrix of distinct k-mers across a collection of metagenomic samples. We used `kmtricks` (v1.1.1) to build a presence/absence k-mer matrix from the 360 metagenomic samples in the collection. In order to find patterns that helped us distinguish between samples from different source environments, we kept only k-mers that were present in at least 3 samples in the collection. The k-mer size in `kmtricks` was set to 31. We removed all k-mers seen only once in a sample, which were likely to be sequencing errors. The rest of the parameters of `kmtricks` were set by default.

The complete k-mer matrix contains around 9 billion k-mers, represented by 700 disjoint sets of k-mers called *partitions*. Omitting some technical aspects [29] for clarity, partitions can be seen

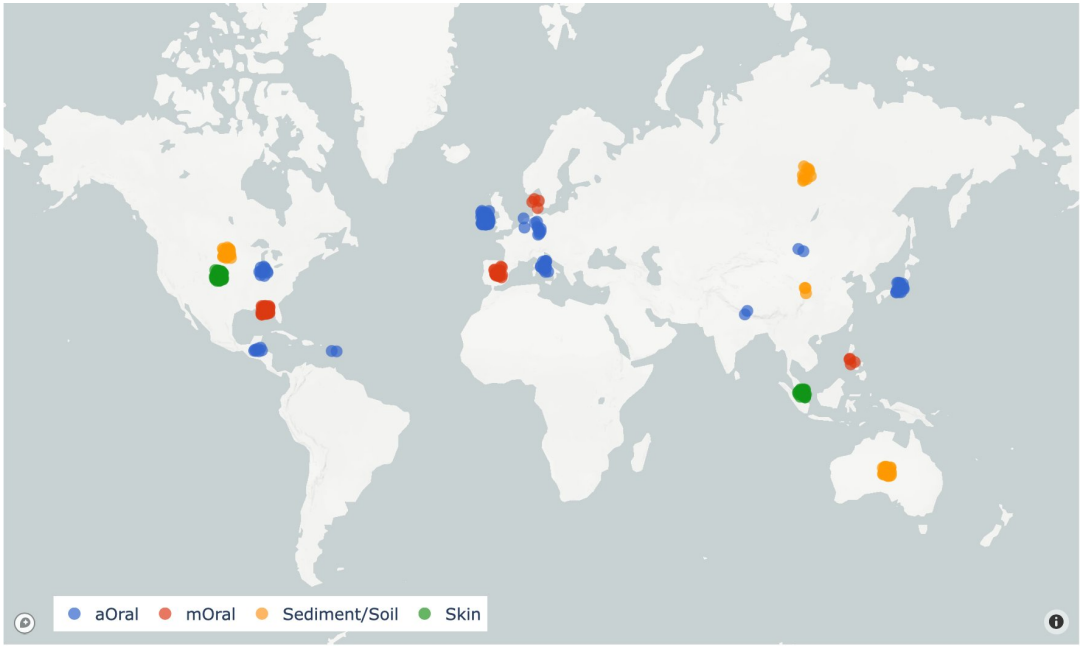


Figure 6.1: **Geographical location of samples coloured by environmental type**. Labels for each sample were retrieved from their metadata. The final collection of metagenomic samples included 116 (32.2%) aOral, 81 skin (22.5%), 79 sediment or soil (21.9%) and 84 mOral (23.3%) samples.

as a random subset of the rows of the k-mer matrix, created to avoid loading the entire matrix in memory [20]. It has been independently shown that partitions enable accurate comparisons between samples [30]. In this work we configure `kmtricks` to only construct a single partition out of the 700, i.e. we consider only a subset of around 14 million k-mers (0.1% of total) for subsequent analysis. We also tested with 7 partitions (Figures 13 and 14 in Supplementary File), and while it improves results marginally, the marked performance improvement when using only 1 partition justifies keeping this regime.

6.4.3 Mathematical formulation

We consider a binary k-mer matrix M (as output by `kmtricks`) that indicates the presence/absence of each k-mer found across several metagenomic data sets, with N number of unique samples (columns) and K number of unique k-mers (rows). Each sample j is represented by a column vector $\mathbf{m}^{(j)} = (m_{1j}, m_{2j}, m_{3j}, ..., m_{Kj})$ where $m_{i,j}$ corresponds to the presence/absence of k-mer i in sample j. We will use the terminology of *sink* and *sources* to respectively denote the sample we want to evaluate the composition of, and the set of samples used as a database.

Consider that the matrix M contains jointly all sources and potential sinks. Let a sample s (where $s \in \{1, 2, \ldots, N\}$) be a sink and $\mathbf{m}^{(s)}$ be its column vector. A source is a collection of $L > 0$ column vectors used to build a matrix of sources M_s of dimensions $K \times (L-1)$. Each column vector in the sources matrix M_s has an associated label that comes from a finite ordered set of environments (classes) $C = \{c_1, c_2, c_3, ..., c_n\}$ determined by the user. In our case $|C| = 4$, as $C = \{aOral, mOral, skin, sediment/soil\}$. The vector of labels for each sample in the sources of length $L-1$ is represented by $\ell = (\ell_1, \ell_2, \ell_3, ..., \ell_{L-1})$, and each entry of the vector can only take one of the values from C as in a multi-class classification problem. The vector of categorical labels ℓ can be further encoded as a highly sparse one-hot binary matrix H of size $(L-1) \times |C|$ where :

$$H_{i,j} = \begin{cases} 1 \text{ if } \ell_i = c_j \\ 0 \text{ otherwise} \end{cases} \tag{6.1}$$

Making an analogy with bins (source environments) and balls (k-mers present in a certain source environment), we are interested in counting the number of balls that fall into each bin. The core idea of `decOM` is that if a k-mer is present in the sink represented by the vector $\mathbf{m}^{(s)}$ *and* in the source vector $\mathbf{m}^{(j)}$ with environment label ℓ_j, then a ball is added to the bin with label ℓ_j. We then compare the sink vector $\mathbf{m}^{(s)}$ against every source vector until all sources are exhausted. The output of this comparison is the vector $\mathbf{w}$ of length $|C|$, where every entry corresponds to the total number of balls in a certain bin, that is, the contribution of each source environment to the sink s.

Counting k-mers of sinks in sources amounts to performing the following matrix vector operation:

$$\mathbf{w} = \mathbf{m}^{(s)^\intercal} \cdot M_s \cdot H \tag{6.2}$$

In order to produce proportions instead of raw counts, we estimate the percentage based on the total number of balls counted per bin (of all known sources) . Such proportions correspond to every element in the vector $\mathbf{p} = \langle w_1, w_2, w_3..., w_{|C|} \rangle$ when multiplied by a scalar, as seen in the following operation:

$$\mathbf{p}' = \frac{\mathbf{p}}{\sum_{i=1}^{|C|} p_i} \tag{6.3}$$

To analyse a new metagenomic sample, one need only compute a presence/absence vector of k-mers for this sample using `kmtricks`, then this new sink is compared against the pre-computed collection of sources. `decOM` incorporates a `kmtricks` module so that the user can give as input a simple FASTQ/FASTA file of their sink of interest, rather than a presence/absence vector. Figure 6.4 provides a graphical representation of our pipeline.

Finally, we are working to include the contribution of an unknown source by characterising it as the number of k-mers that are present in the sink and absent in ***all*** of the sources.

decOM was implemented in Python 3.6 as a conda package and the installation instructions are available in a GitHub repository[31].

6.4.4 Microbial Source Tracking evaluated in four different experimental settings

We perform a metagenomic Microbial Source Tracking to benchmark decOM, mSourceTracker, and FEAST, which all rely on an input matrix. For mSourceTracker and FEAST the input matrix corresponds to a taxonomy-based clustering table, whereas `decOM` takes as input a binary k-mer matrix across metagenomic data sets.

Consider the set $X = \{\mathbf{m}^{(1)}, \mathbf{m}^{(2)}, \mathbf{m}^{(3)}, ...\mathbf{m}^{(N)}\}$, where X contains all the column vectors of the aforementioned k-mer matrix. Let $A = \{\mathbf{m}^{(s)}\}$ be a set of sink vectors, and $B = \{X \setminus \mathbf{m}^{(s)}\}$ a set of sources. In order to estimate the proportion of source environments in each data set in our collection we run our method in a leave-one-out fashion, i.e., every run of our method uses one different sample as sink and leaves the rest of the samples as sources. One run of this experimental setup is described by Algorithm 1.

Algorithm 1 Pseudocode of our method used to estimate proportions of sources in sink s

```
for m_j ∈ B do
    for k = 1, 2, ..., K do
        if m_ks = 1 and m_kj = 1 then
            Add one ball to the bin for class l_j
        end if
    end for
end for
Store proportion of sources in sink s
Store predicted label for sink s as the source environment with highest value
```

Additionally, we performed a 5-fold cross-validation experiment by splitting the collection of metagenomic samples into 5 stratified folds with non-overlapping groups. The groups were defined by the BioProject from which each data set originated. A BioProject is a collection of biological data related to a single initiative originating from a single organisation or from a consortium [32]. The folds were made trying to preserve the percentage of samples for each class, given the constraint that the same group (BioProject) will not appear in two different folds. The idea behind this additional group stratification is to account for the possible bias that might appear when classifying a sink that is very similar to a set of sources simply because they come from the same BioProject and not because there is an underlying sequence similarity between the samples.

For the leave-one-out and cross-validation experiments we evaluated all methods using the Reciever Operating Characteristic (ROC) and Precision-Recall curves, and a hard label was set using as threshold the environment class with the highest contribution to the sink. Performance metrics used were Accuracy, Precision, Recall and F1-score as they are implemented in scikit-learn [33]. Because the framework of evaluation was a multi-class classification task, the performance metrics reported here were estimated for each label and then averaged across classes. Definitions for each performance metric used are specified in Section 5 of the Supplementary File.

Table 6.1: **Environment type prediction performance of `decOM`, FEAST and mSourceTracker**. Accuracy, precision, recall and F1-score for were estimated as an average across all classes in a leave-one-out fashion.

Method	Accuracy	Precision	Recall	F1-score
decOM	**0.8703**	**0.9184**	**0.8703**	**0.8753**
FEAST	0.6816	0.5516	0.7452	0.5479
mSourceTracker	0.8388	0.8388	0.8388	0.8289

We also tested `decOM` on a validation set of 254 aOral samples, none of which belonged to the collection of 360 samples we used to construct the k-mer matrix. For this experiment, the aforementioned matrix is used as sources, whereas the 254 external aOral samples are used as sinks. Because all samples belong to the same class, Precision and F1-score are not well-defined, whereas Recall and Accuracy are equivalent (See Section 5 in Supplementary File), which is why performance is measured using Recall only. Finally we tested `decOM` and its competitors on an uncontaminated simulated ancient oral data set and presented the estimated proportions.

6.5 Results

We created `decOM` as reference-free and open-source Microbial Source Tracking method that is adapted to ancient metagenomic experiments. Our method takes as input a set of source vectors in the form of a presence/absence k-mer matrix (built from a collection of metagenomic data sets ready for the user to download), and one or more FASTA/FASTQ files to be used as sinks. It outputs a set of proportions (percentages) and a predicted metadata class per sink.

6.5.1 decOM robustly predicts metagenome sample labels

6.5.1.1 Leave-one-out experiment

We compared the performance of `decOM` with FEAST [8] and mSourceTracker [9] based on their ability to correctly predict the environmental type of a sample, defined as the highest proportion among the four possible sample types (ancient oral, model oral, skin, soil). For all methods, we used the same collection of 360 metagenomic experiments as sources.

All methods output a set of proportions for each sample. We ran them in a leave-one-out fashion (one sample was used as sink, and the rest were left out as sources). In order to perform a multi-class classification task, we mapped the set of continuous proportions into a hard label, by simply assigning a label to the sample corresponding to the environmental type with the largest proportion among all the predicted sources. The performance metrics presented were calculated using the hard labels.

Table 6.1 shows that `decOM` outperforms both mSourceTracker (+3% Accuracy, +8% Precision, +3% Recall, +5% F1 score) and FEAST (+19% Accuracy, +37% Precision, +12% Recall, +33% F1 score) in the multi-class classification task of predicting source environment with the largest contribution in a sink, when such contribution is estimated using a MST framework. Precision-Recall and ROC curves are shown in the Supplementary File (See Figure 10 and 11).

6.5.1.2 Cross-validation

To further validate that `decOM` does not solely rely on closely related samples for its predictions, we performed a 5-fold cross-validation experiment by dividing the collection into 5 stratified folds with non-overlapping BioProjects. This constraint means that a sink is classified without any other samples from the same BioProject in the sources. This data stratification is relevant because it controls for the possible bias that might come from classifying a sink that is similar to the sources simply because they come from the same sequencing initiative and not because there is some underlying biological similarity between the samples (see Figure 12 in Supplementary File for visualisation of the data splitting).

decOM outperforms mSourceTracker and FEAST in each of the five sink/sources folds for performance metrics such as Accuracy, Precision, Recall and F1 Score (see Figure 6.2) and when metrics are averaged across groups (see Table 2 in Supplementary File). The performance estimates dropped with respect to the leave-one-out MST, which is expected since cross-validation results give a less biased estimate of the model (see also Table 1 and 2 in Supplementary File).

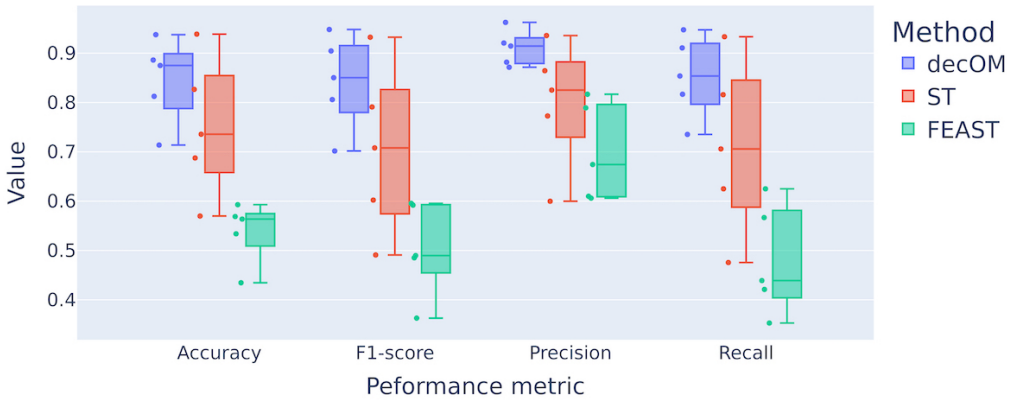


Figure 6.2: **Bioproject stratified 5-fold cross-validation performance of every method**. The performance from every fold was evaluated using accuracy, precision, recall and F1-score

Table 6.2: **Performance of `decOM` in the aOral validation set**. As only one class is present in the validation data set (aOral), performance is measured using precision for this highly imbalanced setting.

Method	Recall
decOM	**0.8654**
FEAST	0.6692
metaSourceTracker	0.6346

6.5.1.3 Validation set

We evaluated `decOM` in an external validation set with 254 aOral samples that were present in the AncientMetagenomeDir [1] but were not part of the matrix of sources previously described. Samples in the validation set belonged to 6 different BioProjects and ranged from 100 to 14800 years old. Furthermore they were isolated from 12 different countries in mostly 2 continents. For more information regarding the metadata of the samples in the validation data set see Supplementary Figures 5 and 6.

Here also `decOM` outperforms mSourceTracker and FEAST by classifying most of the samples as aOral. See Table 6.2 for results in the validations set of only aOral samples.

6.5.1.4 Simulated data set

As a final experiment we tested each of the methods on a simulated ancient dental calculus metagenome generated by other authors [34]. A mock oral microbial community is created using representative genomes of microbes found in the human oral microbiome, further processed to appear similar to an ancient metagenomic sample. As in the validation set, we estimated the source environment contribution of the aOral, mOral, skin and sediment/soil microbial communities by using the samples from the 360 collection as sources. Results for all methods are in Figure 6.3. Given that the synthetic metagenome comes from an uncontaminated mock oral microbial community that has been adapted to appear similar to an ancient calculus sample the expected content is to be 100% oral, `decOM` provides the highest estimation of oral contribution (ancient or modern), followed by mSourceTracker and lastly by FEAST. We encountered reproducibility problems for FEAST that are further explained in the Supplementary Figure 7.

6.5.2 Running times

We measured the running time for `decOM` and mSourceTracker using 250 GB of memory and 10 cores. FEAST did not allow for multithreading. We estimated the time it takes to produce an input matrix for each of the methods (whether it is a taxonomy-based clustering table or k-mer matrix of sources). We also estimated the time it takes to analyse a new sample by splitting the process in two steps: the time it takes to produce a new vector to represent the sample, and the time it takes to perform MST. For the two previously mentioned steps, the average running time was estimated on the 254 samples from the validation set. The consolidated running times can be seen in Table 6.3. `decOM` is considerably faster than the two other methods for creating a source matrix as we selected one partition of the large k-mer matrix produced by `kmtricks` and

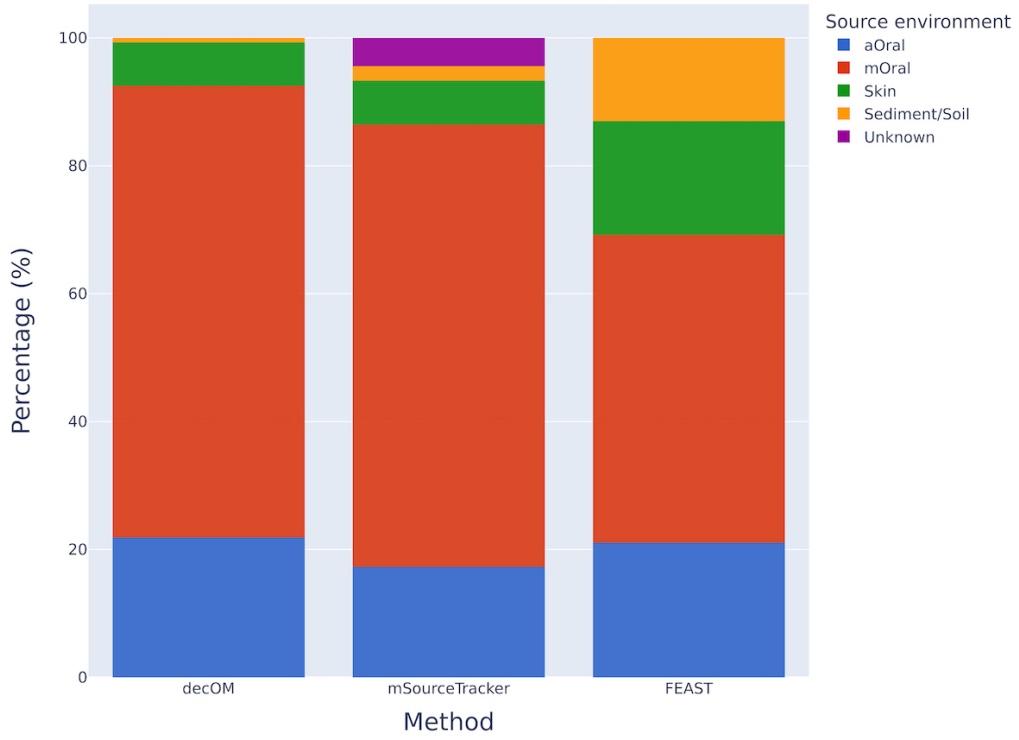


Figure 6.3: **MST on a simulated ancient dental calculus metagenome**. Bar plots for the source environment proportion estimation obtained after evaluating each method using as sources all the samples from the 360 metagenomic collection, and using as sink a synthetic ancient oral data set. The expected content of this synthetic sample is 100% oral

Table 6.3: **Running times of MST**. Wall-clock time was measured in different parts of the pipeline: Time to build the input matrix, time to produce a new vector from an input FASTQ file and time to perform the MST of one sample. Except for the process named "Build source matrix", the average time was estimated on the results from the validation set. MST done by FEAST does not allow for multithreading and was run using 2GB of memory and 1 core, whereas mSourceTracker can not split one sink into multiple jobs, so 1 core and 250GB of memory were allocated for each sink. Every other process was run using 250GB of memory and 10 cores. Results for `decOM` are presented in bold.

Method	Process	Time (h)
decOM	Build source matrix	**6.60**
decOM	Produce new vector	**0.04**
decOM	MST	**0.02**
FEAST	Build source matrix	99
FEAST	Produce new vector	0.28
FEAST	MST	0.07
mSourceTracker	Build source matrix	99
mSourceTracker	Produce new vector	0.28
mSourceTracker	MST	0.01

offered the pre-computed matrix in a Zenodo file for users to implement in their analyses. When producing a new vector, since `decOM` relies on `kmtricks`, it is also considerably faster than FEAST and mSourceTracker. However our evaluation of the time here was based on Kaiju's running times. Optimising the creation of taxonomy-based clustering tables using faster alignment-free methods could improve time performance, potentially at the expense of results quality. Finally, all methods show comparable running times when performing the MST step.

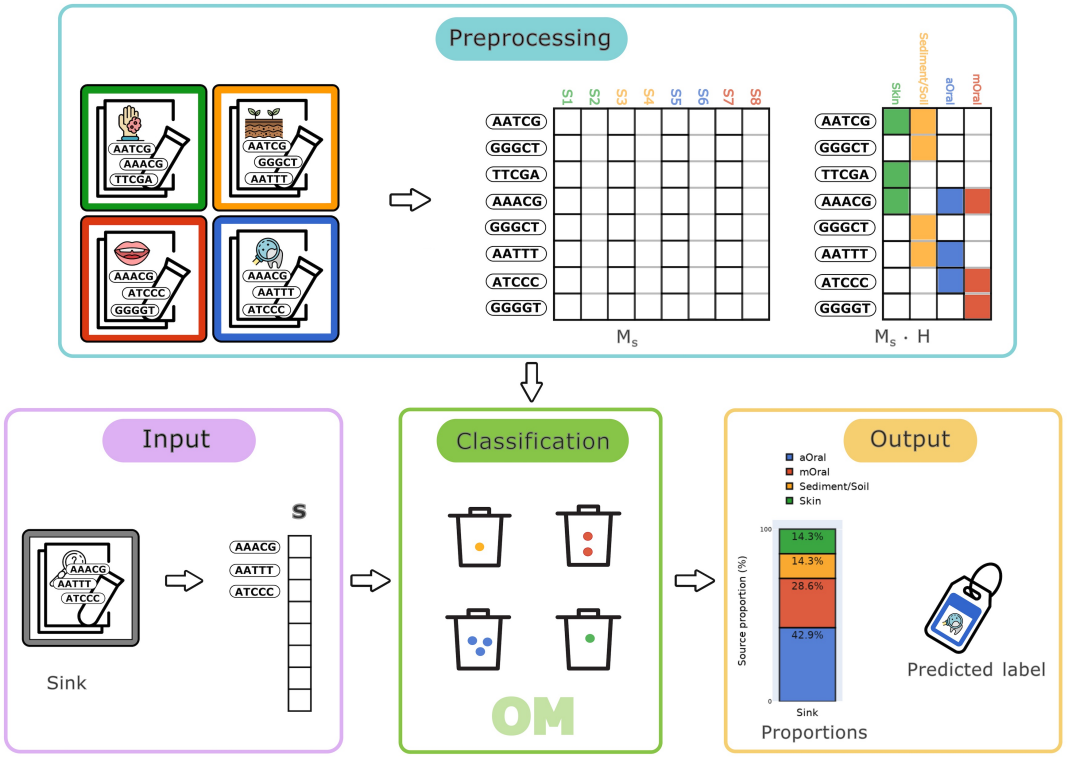


Figure 6.4: **Graphical representation of `decOM`**. Our method preprocesses an input k-mer matrix of aOral metagenomic samples and its possible contaminants, divides it into sinks and sources and then estimates and outputs the proportions of each source environment in the sink. The core idea in the classification step is that if a k-mer is present in the sink s represented by the vector $\mathbf{m}^{(s)}$, and in the source vector $\mathbf{m}^{(j)}$ with environment label l_j, then a ball is added to the bin with label l_j (Ex: k-mer AAACG is present in the input sink S and in source S1 labelled as skin, S5 labelled as aOral and S7 labelled as mOral, hence one ball is added to the bin of skin, aOral and mOral respectively). After every entry in the the sink vector is compared against every entry of every vector in the sources, `decOM` outputs the estimated environment proportions and the hard label assigned to the sink s is that of the environment with the highest contribution.

6.5.3 Ancient oral metagenomic samples come from various environments (multi-source)

After predicting the metadata class of each of the 360 samples in the collection, we also plotted the source proportions according to the estimation done by `decOM`, mSourceTracker and FEAST (Figure 6.5). The proportion bar plots for mSourceTracker and `decOM` are visibly more similar to each other than to FEAST, which seems to output more variable results.

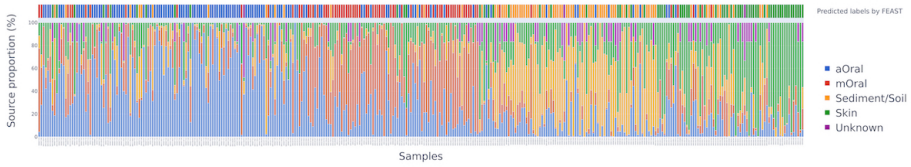
According to the estimation done by `decOM`, there are 4 main predicted groups in the collection with distinct source composition as seen in Figure 6.5a: there is a group of samples that have a higher sediment/soil content, another class of samples with a higher skin content and with a considerable presence of mOral k-mers, a third group that corresponds to the aOral samples and that also share a part of the mOral content. Finally, there is a fourth group of samples in which the contribution of the mOral sequences is considerably higher, however these samples also have some k-mers in common with the skin and aOral metagenomic samples.

Both `decOM` and mSourceTracker find a certain level of skin contamination on mOral samples, as seen in Figure 5a and 5b respectively. We further investigated the issue by plotting a PCA on the k-mer matrix of sources (See Supplementary Figure 18) and saw that effectively some of the mOral samples appear close to the Skin samples. This might be the reason why there was some skin contamination in the mOral samples to begin with.

In additional analyses (see Figure in Supplementary File 15), we divided the samples after `decOM`'s MST estimation into two categories: samples that come mostly from one source environment (mono-source) or samples that come from several environments (multi-source). In addition to the hard label assigned by `decOM`, we further categorised the classification of each sample, qualifying the upper quartile ($> 75\%$) of each class as mostly mono-source samples, and the first and second quartile ($< 75\%$) as samples of diverse origins (more contaminated). According to this threshold, there are 78 mono-source samples (22% of the total collection). These are samples whose recovered label corresponds to the label predicted by `decOM`, and which are not as contaminated by other

(a) decOM

(b) mSourceTracker



(c) FEAST

Figure 6.5: **Bar plots of the source environment contribution on each sink after the leave-one-out experiment as estimated by `decOM`, mSourceTracker and FEAST.** Samples in Figure 6.5a are first sorted by true label, and then sorted by ascending order of the proportion value for such label. Sample order in the x axis for Figures 6.5b and 6.5c is sorted according to the order from 6.5a

.

sources. A collection of low-contaminated and mono-source samples as this could be used as a high-quality multi-class data set of aOral (36%), mOral (27%), sediment/soil (24%) and skin (13%) for benchmarking with a relatively low imbalance (see Figure 16 in Supplementary File). Interestingly, 91% of the samples we call mono-source are also correctly predicted by mSourceTracker and 78% are correctly predicted by FEAST (Figure 17 in Supplementary File). Nearly a quarter of the aOral samples in the collection have contamination levels that are low enough to have them categorised as mono-source, while the remainder of the ancient oral samples, as expected, have varying levels of contamination.

6.6 Discussion

We have proposed and evaluated `decOM` as a tool predict the metadata class of a given metagenomic sample by using a Microbial Source Tracking framework, in order to help paleogeneticists better assess the source content of their ancient samples. Because it was built using a Microbial Source Tracking framework, it can also help determine the composition of any other microbial community (not necessarily ancient or of oral origin), which is a common question in microbiome studies. Let us clarify that our goal is not to define an ancient oral microbial community *per se*, but to give the user an indication on the quality of their sample in terms of ancient genetic material. We leave for immediate future work the extensions of `decOM` to other MST tasks, which could be readily done by creating a k-mer matrix of metagenomic samples of interest with their associated labels and estimating the source proportions using `decOM`.

The utility of `decOM` was established on a collection of aOral metagenomic samples and their possible contamination sources, in a leave-one-out set up experiment where every sample was compared against all others. To control for an overly optimistic performance, we performed a stratified 5-fold cross-validation experiment making sure all the samples from the same BioProject belonged to the same fold. Finally, `decOM` was tested on an external validation data set of 254 aOral samples that were not part of initial collection of metagenomic aOral samples and metagenomes of other contaminants and in a simulated ancient calculus metagenome. We acknowledge that our method would classify the synthetic sample tested on this paper as an mOral sample instead of aOral despite having predicted the largest proportion of aOral source contribution when compared to mSourceTracker or FEAST. However, considering `decOM` has already proven to be useful on real

data, we leave further tuning of the method on synthetic data to be part of the upcoming work. In almost every setting, `decOM` outperformed two of the most widely-used techniques in the field of MST in the multi-class classification task of predicting the label of a metagenomic samples as the source environment with the highest proportion.

Ideally we would test `decOM` on a collection of ancient oral samples with known proportions for each source environment, unfortunately, to our knowledge, such a data set does not exist. The task of creating a synthetic data set with such characteristics poses additional challenges regarding how to avoid overlapping species (originating genomes) between each source environment, and would ultimately not be a good representation of a real sample. For this reason we focused on the evaluation of each method by using the metadata class prediction of a hard label rather than by confirming the proportion predictions were the most accurate.

It could be argued that the lower performance of mSourceTracker and FEAST compared to `decOM` in the multi-class classification task described in this study was due to limitations of the input taxonomy-based clustering table given to the methods. Better results might be achieved by using a larger database or a tool other than Kaiju to estimate taxonomic abundances. To evaluate this, we conducted an additional experiment in which we constructed another taxonomy-based clustering table with KrakenUniq [35] (see Supplementary File, Section 2). Results in this paper are shown only for the taxonomic abundance profile based on Kaiju, which can also be replicated using public data sets and which, in any case, yielded the best results for the competing methods. The results for the taxonomic abundance profile constructed with KrakenUniq are shown in the Supplementary File information (see Figure 10 and 11).

An important hyperparameter of our model is the size of the input k-mer matrix M_s. We explored the effect of using multiple partitions on the performance metrics for the single- and 5-fold cross-validation experiment, but to speed up computations and reduce the memory required, we decided to use only one partition (0.1% of the total k-mer found by `kmtricks`). Remarkably, the performance of `decOM` is still better than that of competing methods (see Figures 13 and 14, Table 1 and 2 in the Supplementary File). In the future it would be interesting to study the impact on the classification performance of varying the hyperparameters for the construction of the k-mers matrix, such as the size of the k-mers, minimum recurrence or minimum abundance.

6.7 Conclusions

We propose a novel and reference-free method to perform Microbial Source Tracking and predict the metadata class of a given (meta)genomic sample. We tested our method on a collection of real metagenomic data sets of aOral origin and its possible contaminants and provided an estimation of the contribution of each source environment on each sample. We anticipate that the incorporation of `decOM` into paleogenomic analyses will prevent erroneous results and help identify contaminated metagenomic samples and ensure their validity.

Acknowledgements

Funding R.C was supported by ANR Full-RNA, SeqDigger, Inception and PRAIRIE grants (ANR-22-CE45-0007, ANR-19-CE45-0008, PIA/ANR16-CONV-0005, ANR-19-P3IA-0001). This work is part of the ALPACA project that has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grants agreements No 956229 and 872539 (PANGAIA).

Availability of data and materials The data sets analysed during the current study are available in the repository for `decOM` [31]. Additional information on the version of the software used, as well as explanation on how to access the run accessions codes for the sources and validation set are present in the Supplementary File.

Ethics approval and consent to participate Not applicable.

Competing interests The authors declare that they have no competing interests.

Consent for publication Not applicable.

Authors' contributions Conceptualisation, C.D, H.R., N.R., R.V. and R.C.; Methodology, H.R, R.C., N.R, R.V and C.D.; Software, C.D., T.L.; Validation, C.D.; Resources, R.C.; Writing – Original Draft, C.D, R.C., R.V., N.R., H.R.; Writing – Review & Editing, C.D, R.C., R.V., N.R, H.R.; Visualisation, C.D.; Supervision, R.C., H.R., R.V; Project Administration, R.C., H.R.; Funding Acquisition, R.C.

Additional Files

6.7.1 Additional file 1 — Appendix I

Supplementary File (Appendix A) with supplementary figures and tables cited throughout the text.

References

[1] Yates, J. A. F., Valtueña, A. A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-López, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., et al. "Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir". In: *Scientific Data* vol. 8, no. 1 (2021), pp. 1–8.

[2] Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., et al. "Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments". In: *Proceedings of the National Academy of Sciences* vol. 110, no. 39 (2013), pp. 15758–15763.

[3] Der Sarkissian, C., Velsko, I. M., Fotakis, A. K., Vågene, Å. J., Hübner, A., and Fellows Yates, J. A. "Ancient metagenomic studies: Considerations for the wider scientific community". In: *Msystems* vol. 6, no. 6 (2021), e01315–21.

[4] Peyrégne, S. and Prüfer, K. "Present-Day DNA Contamination in Ancient DNA Datasets". In: *Bioessays* vol. 42, no. 9 (2020), p. 2000081.

[5] Farrer, A. G., Wright, S. L., Skelly, E., Eisenhofer, R., Dobney, K., and Weyrich, L. S. "Effectiveness of decontamination protocols when analyzing ancient DNA preserved in dental calculus". In: *Scientific reports* vol. 11, no. 1 (2021), pp. 1–14.

[6] Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiß, C. L., Burbano, H. A., Orlando, L., and Krause, J. "A robust framework for microbial archaeology". In: *Annual Review of Genomics and Human Genetics* vol. 18 (2017), pp. 321–356.

[7] Peyrégne, S. and Peter, B. M. "AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination". In: *Genome biology* vol. 21, no. 1 (2020), pp. 1–16.

[8] Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., and Halperin, E. "FEAST: fast expectation-maximization for microbial source tracking". In: *Nature Methods* vol. 16, no. 7 (2019), pp. 627–632.

[9] Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. "Bayesian community-wide culture-independent microbial source tracking". In: *Nature Methods* vol. 8, no. 9 (2011), pp. 761–763.

[10] McGhee, J. J., Rawson, N., Bailey, B. A., Fernandez-Guerra, A., Sisk-Hackworth, L., and Kelley, S. T. "Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics". In: *PeerJ* vol. 8 (2020), e8783.

[11] Henry, R., Schang, C., Coutts, S., Kolotelo, P., Prosser, T., Crosbie, N., Grant, T., Cottam, D., O'Brien, P., Deletic, A., et al. "Into the deep: evaluation of SourceTracker for assessment of faecal contamination of coastal waters". In: *Water research* vol. 93 (2016), pp. 242–253.

[12] Li, L.-G., Huang, Q., Yin, X., and Zhang, T. "Source tracking of antibiotic resistance genes in the environment—Challenges, progress, and prospects". In: *Water Research* vol. 185 (2020), p. 116127.

[13] Raza, S., Kim, J., Sadowsky, M. J., and Unno, T. "Microbial source tracking using metagenomics and other new technologies". In: *Journal of Microbiology* (2021), pp. 1–11.

[14] Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., and Knight, R. "Using QIIME to analyze 16S rRNA gene sequences from microbial communities". In: *Current protocols in bioinformatics* vol. 36, no. 1 (2011), pp. 10–7.

[15] Nguyen, N.-P., Warnow, T., Pop, M., and White, B. “A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity”. In: *NPJ biofilms and microbiomes* vol. 2, no. 1 (2016), pp. 1–8.

[16] Osuolale, O., Mason, C., Consortium, M. I., et al. “The metagenomics and metadesign of the subways and urban biomes (MetaSUB) international consortium inaugural meeting report”. In: *Microbiome* vol. 4, no. 1 (2016), p. 24.

[17] Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, I. N. S. D. “The sequence read archive”. In: *Nucleic acids research* vol. 39, no. suppl_1 (2010), pp. D19–D21.

[18] Cook, C. E., Lopez, R., Stroe, O., Cochrane, G., Brooksbank, C., Birney, E., and Apweiler, R. “The European Bioinformatics Institute in 2018: tools, infrastructure and training”. In: *Nucleic acids research* vol. 47, no. D1 (2019), pp. D15–D22.

[19] Marchet, C., Boucher, C., Puglisi, S. J., Medvedev, P., Salson, M., and Chikhi, R. “Data structures based on k-mers for querying large collections of sequencing data sets”. In: *Genome Research* vol. 31, no. 1 (2021), pp. 1–12.

[20] Lemane, T., Medvedev, P., Chikhi, R., and Peterlongo, P. “kmtricks: Efficient and flexible construction of Bloom filters for large sequencing data collections”. In: *Bioinformatics Advances* (2022).

[21] Ziesemer, K. A., Ramos-Madrigal, J., Mann, A. E., Brandt, B. W., Sankaranarayanan, K., Ozga, A. T., Hoogland, M., Hofman, C. A., Salazar-García, D. C., Frohlich, B., et al. “The efficacy of whole human genome capture on ancient dental calculus and dentin”. In: *American journal of physical anthropology* vol. 168, no. 3 (2019), pp. 496–509.

[22] Warinner, C., Speller, C., and Collins, M. J. “A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 370, no. 1660 (2015), p. 20130376.

[23] Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R. Y., Fiddyment, S., et al. “Pathogens and host immunity in the ancient human oral cavity”. In: *Nature genetics* vol. 46, no. 4 (2014), pp. 336–344.

[24] Ziesemer, K. A., Mann, A. E., Sankaranarayanan, K., Schroeder, H., Ozga, A. T., Brandt, B. W., Zaura, E., Waters-Rist, A., Hoogland, M., Salazar-Garcia, D. C., et al. “Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification”. In: *Scientific Reports* vol. 5, no. 1 (2015), pp. 1–20.

[25] Pasolli, E. et al. “Accessible, curated metagenomic data through ExperimentHub”. en. In: *Nat. Methods* vol. 14, no. 11 (Oct. 2017), pp. 1023–1024. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4468.

[26] Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., Bergen, M. von, Stadler, P. F., Carvalho, A. C. P. d. L. F. d., and Nunes da Rocha, U. “HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes”. In: *Nucleic Acids Research* vol. 49, no. D1 (2021), pp. D743–D750.

[27] Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., et al. “MGnify: the microbiome analysis resource in 2020”. In: *Nucleic acids research* vol. 48, no. D1 (2020), pp. D570–D578.

[28] Menzel, P., Ng, K. L., and Krogh, A. “Fast and sensitive taxonomic classification for metagenomics with Kaiju”. In: *Nature communications* vol. 7, no. 1 (2016), pp. 1–9.

[29] Rizk, G., Lavenier, D., and Chikhi, R. “DSK: k-mer counting with very low memory usage”. In: *Bioinformatics* vol. 29, no. 5 (2013), pp. 652–653.

[30] Irber, L., Brooks, P. T., Reiter, T., Pierce-Ward, N. T., Hera, M. R., Koslicki, D., and Brown, C. T. “Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers”. In: *bioRxiv* (2022). DOI: 10.1101/2022.01.11.475838.

[31] Duitama, C. *decOM*. 2022. URL: https://github.com/CamilaDuitama/decOM (visited on 05/17/2022).

[32] NCBI. *Bioproject FAQ*. Mar. 2018. URL: https://www.ncbi.nlm.nih.gov/bioproject/docs/faq/%5C#what-is-a-bioproject.

[33] Pedregosa, F. et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* vol. 12 (2011), pp. 2825–2830.

[34] Mann, A. E., Yates, J. A. F., Fagernäs, Z., Austin, R. M., Nelson, E. A., and Hofman, C. A. “Do I have something in my teeth? The trouble with genetic analyses of diet from archaeological dental calculus”. In: *Quaternary International* (2020).

[35] Breitwieser, F. P., Baker, D., and Salzberg, S. L. "KrakenUniq: confident and fast metagenomics classification using unique k-mer counts". In: *Genome biology* vol. 19, no. 1 (2018), pp. 1–10.

6.8 Perspectives

The work conducted in this chapter has raised several inquiries for future investigation. One potential avenue for future research involves the construction of a larger and more comprehensive k-mer matrix of sources. It would be intriguing to incorporate metagenomic samples that were not included in this thesis but were subsequently released, particularly those related to aOral and potential contaminants. Moreover, it would be worthwhile to examine the impact of varying hyperparameters on the construction of the k-mer matrix. These hyperparameters include the size of the k-mers used (denoted as k), the minimum recurrence, and the minimum abundance. Additionally, it would be of interest to assess the scalability of `decOM` by expanding the number of sinks and sources used, and to compare its performance against mSourceTracker and FEAST. The potential sources could come from a gold standard set established by the community of users in palaeomicrobiology or from the samples uploaded in the most recent release of the AncientMetagenomeDir.

On the other hand, one interesting aspect that could enhance `decOM`'s capabilities is to explore different forms of count normalisations or assign varying weights to the counts based on the significance of each source environment to the researcher.

In contrast to other existing methods such as FEAST and mSourceTracker, the algorithm developed in this thesis for assessing contamination via MST (`decOM`) has demonstrated the ability to distinguish between mOral and aOral samples. It would be advantageous to further investigate the underlying factors that contribute to `decOM`'s success in achieving this aOral/mOral differentiation.

Finally, it would be valuable to determine the minimum number of k-mers that must be identified in a sink sample in order for `decOM` to yield useful results. Possible enhancements to the software could include introducing this threshold and improving the visual representations, output metrics or taxonomic assignment to the k-mers found in the sink sample.

Chapter 7

State of the art: Ancient reads decontamination

7.1 Contamination removal tools

The study of human aDNA, requires authentication procedures as described in Section 3.5, due to the recurrent challenge in palaeometagenomic field studies, in which contamination from scientists can occur at any stage, from excavation to DNA library preparation [1]. In the study of ancient microbial genomes, it is often very difficult to re-sample from biomaterials that were very difficult to obtain [2, 3], which is why contamination assessment and reduction procedures are fundamental to ensure the best use of available genetic information. Apart from wet-lab based methods for contamination control (see Section 3.6.1), bioinformatic tools such as DeconSeq [4] or Recentrifuge [5] aim to remove genomics sequences that correspond to negative control samples. They are however not specifically tailored for aDNA and require either a database of negative controls or an index of reference genomes to distinguish contamination (that should be removed) from endogenous material.

7.2 DeconSeq

DeconSeq is a framework developed for the rapid and automated identification and removal of sequence contamination in longer-read datasets($\leq$ 150 bp mean read length), available as a standalone tool or a web-based application. This method categorises possible contamination sequences, eliminates redundant hits, and provides graphical visualisations of the alignment results and classifications. In [4] the authors carried out an analysis of 202 metagenomes which revealed significant contamination of non-human associated metagenomes, suggesting its suitability for metagenomic screening. DeconSeq was written in Perl and its last update was released in May 2013.

7.2.1 Pipeline

DeconSeq accepts as input FASTA/FASTQ files containing genomic or metagenomic reads. The workflow of the method is shown in Figure 7.1. The basic idea of the method is to compare an input dataset against a *Remove* database (set of contaminating samples) and identify the sequences with significant similarities to this database. More precisely, sequences are classified as contaminant if they have a match above a (user-defined) threshold against the remove database. DeconSeq uses coverage and identity thresholds to determine if a match is possibly caused by contamination or not. There is an optional *Retain* database that the user might provide, in order to obtain other labels such as sequences that exist in either one of the databases or both of them. This allows to finally classify as contaminants the sequences that are *unique* to the *Remove* database.

This contamination removal method relies on an alignment algorithm called Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW) [6], an algorithm designed to align long sequences (up to 1Mb) against a large sequence database with few gigabytes of memory. In few words, BWA-SW builds FM-indices for reference and query sequences. It uses a prefix trie[1] and a prefix Direct Acyclic Word Graph (DAWG) to represent each of the sequences, respectively. Moreover, it speeds up the process of mapping via dynamic programming and heuristics. Although there exists a faster, and generally more accurate version of this aligner (BWA-MEM [7]), the authors claim that BWA-SW may have better sensitivity when alignment gaps are frequent. Interestingly, several benchmarking and parameter optimisation papers of different mappers have been tested specifically on aDNA data [8, 9, 10]. In [9], authors state that Bowtie2 shows better computational time and increased sensitivity with respect to BWA. On the other hand, authors of [8] authors compare 4 different mapping software to conclude that despite all of them showing some level of reference bias, BWA-aln and NovoAlign and BWA-MEM are recommended as they manage to achieve high mapping precision and reduce bias, particularly after filtering reads with low mapping qualities. Interestingly, NovoAlign requires the use of an International Union of Pure and Applied Chemistry (IUPAC) reference genome. Finally, authors of [10] recommend NovoAlign and BWA-aln

[1]The prefix trie of a string X is a tree with each edge labelled with a symbol such that the concatenation of symbols on the path from a leaf to the roots gives a unique prefix of X (definition from [6])

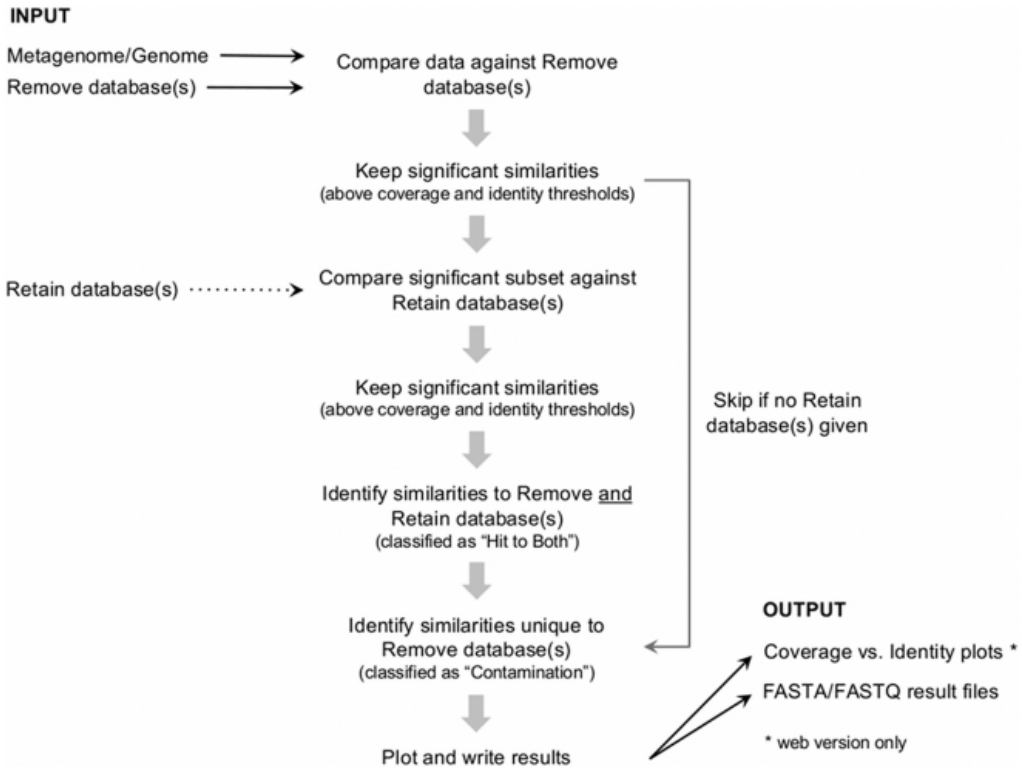


Figure 7.1: **DeconSeq's pipeline**. Image taken from [4].

for research using short reads (as it is the case for ancient DNA research), in particular, they suggest BWA-aln is *"still the gold-standard for aDNA read alignment"*.

The BWA-SW algorithm was modified to fit the needs of DeconSeq, yet the default behaviour of the algorithm does not change.

7.3 Recentrifuge

Recentrifuge [5] is a tool that removes negative-control and crossover taxa, allowing researchers to analyse taxonomic classifications with interactive charts that emphasise confidence levels. It also provides shared and exclusive taxa per sample, enabling contamination removal and comparative analyses in metagenomics. The pipeline of Recentrifuge to remove contaminant reads consists of two steps:

1. Populating and folding logical taxonomic tree.
2. Retrieving set of candidate from control samples.

7.3.1 Populating and folding a logical taxonomic tree

Recentrifuge populates a logical taxonomic tree, and recursively "folds the tree" for any of the samples if the number of assigned reads to a taxon is below a used-defined threshold.

First, for each sample, Recentrifuge populates a taxonomic tree according to the NCBI Taxonomy [11], wherein the terminal nodes correspond to lower taxonomic levels. Subsequently, Recentrifuge performs a recursive process known as "tree folding", wherein the leaves of the tree are aggregated in their parent node until at least one of the following conditions are met:

- The number of assigned reads is below *mintaxa*.
- The corresponding taxonomic tank is below *minrank*.

where *mintaxa* and *minrank* are user-defined parameters.

The new score of parent taxa σ'_p is estimated according to the following equation [5]:

$$\sigma'_p = \frac{1}{n_p + \sum_i^D n_i} \left(\sigma_p n_p + \sum_i^D \sigma_i n_i \right) \forall(\sigma_i, n_i) \tag{7.1}$$

where:

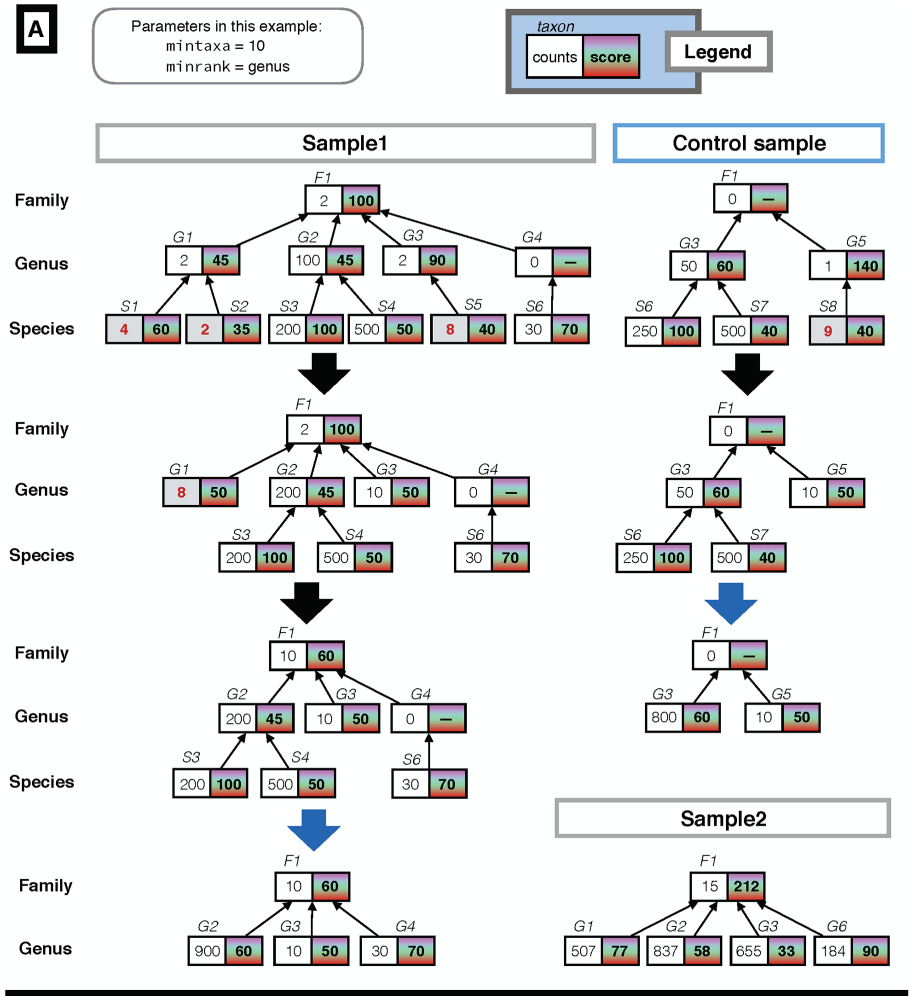


Figure 7.2: **Recentrifuge's first part of the pipeline: populating and folding a logical taxonomic tree.** In this example *mintaxa* is set to 10 and *minrank* is set to genus. The leaves of the tree are recursively accumulated in the parent node until at least one of the necessary conditions is met. The parent score is updated according to Equation 7.1. Image taken from [5].

- n_i are the counts of a taxon i. Where $0 \leq n_i \leq mintaxa$.
- n_p are the counts of parent taxon p.
- σ_p is the score of parent taxon p.
- σ_i is the score of the taxon i.
- σ'_p is the new score of parent taxon p.
- D is the number of descendant taxa that are to be accumulated in the parent taxa.

See Figure 7.2 for an example of this initial part of the pipeline.

7.3.2 Retrieving contaminant sequences

After the "tree folding", Recentrifuge's algorithm retrieves the set of candidates from control samples. Depending on the relative frequency of these taxa in these samples, and if they are also present in other specimens, the algorithm classifies them in one of three contamination levels:

critical, *severe*, *mild* and *other*. All except the members of the *other* group are removed from non-control samples.

To account for crossover contamination, the algorithm removes any taxon in the aforementioned *other* group from a non-control sample unless it passes a crossover check that includes: a statistical test screening for outliers and an order of magnitude test against control samples. More precisely:

$$\begin{aligned} \text{Outliers statistic test}(t_s^k) &: f_{t_s^k} > \text{median}\{f_{t1}^k, ..., f_{t_s^k}\} + \delta Q_n \\ \text{Order of magnitude test}(t_s^k) &: f_{t_s^k} > 10^{\xi} \max\{f_{t_1^k}, f_{t_2^k}, ..., f_{t_N^k}\} \end{aligned} \tag{7.2}$$

In Equation 7.2:

- Q_n is a scale estimator.
- δ is an outliers cutoff factor that ranges from $3 < \delta < 5$.
- ξ is a parameter that sets the difference in order of magnitude between the relative frequency of the candidate to crossover contaminated in a sample s and the greatest of such values among the control samples. Usually $2 < \xi < 3$.

7.4 Bloom Filters

I hereby introduce Bloom Filters, a well-known data structure originally developed in 1970 by Burton H. Bloom. It is at the base of the paper presented in Chapter 8.

A Bloom Filter (BF) [12] is a probabilistic data structure that allows two operations: insertion and lookup. A Bloom Filter is part of a set of data structures called approximate membership query filters: probabilistic data structures designed to address the membership problem in a space-efficient way [13]. Some applications of Bloom Filters in bioinformatics include classification of DNA sequences [14], k-mer counting [15], scaling metagenomic sequence assembly [16], de Bruijn graph compaction [17], k-mer matrix construction [18] among others.

7.4.1 Hash functions

Hash functions are the building block of probabilistic filters such as the Bloom Filter [19].

Suppose you have a dynamic set[2] S in which each element has a key k drawn from a universe U. We could use an array to represent such a dynamic set, in which each slot or position corresponds to an element of U. However, if U is large, storing an array of size $|U|$ becomes computationally prohibitive. By using a *hash function* h to compute the slot from the key k, we introduce a technique called *hashing*, where an element $k \in S$ is stored in a table T at the index corresponding to the hash value $h(k)$. Formally, a hash function is a surjective function $h : U \to \{0, 1, ..., m-1\}$

From the previous definition, an element with key k hashes to slot $h(k)$, and $h(k)$ is the hash value of key k. The point of the hash function is to reduce the range of array indices that need to be handled, going from $|U|$ to m (where $m \ll |U|$). The downside of this, is that two keys may hash to the same slot. This situation is called a *collision* and could be handled with techniques known as chaining or open addressing [20].

7.4.2 Formal definition of Bloom Filters

Definition 7.4.1 (Bloom Filter)**.** A Bloom Filter (BF) for representing a set $S = \{s_1, s_2, s_3, ...s_n\}$ of n elements is described by an array of m bits. Initially, they are all set to 0. Bloom Filters also require the use of k independent hash functions $h_1, h_2, ..., h_k$ with range $\{0, ..., n-1\}$. For each element $s_i \in S$ the bits $h_i(x)$ are set to 1 for $1 \leq i \leq k$ [21]. The following two operations are defined:

- Insertion: $B[h_i(x)] \leftarrow 1, \forall i \in [1...k]$
- Lookup: $\bigwedge_{i=1}^{k} B[h_i(x)]$

See Figure 7.3 for a graphical representation of what a Bloom Filter is.

Notice that a location in the BF can be set multiple times to 1, but only the first change has an effect [19]. Because Bloom Filters "ignore" collisions, the lookup operation may return false

[2]Mathematical sets are unchanging, yet in computer science sets can be manipulated by algorithms (they can grow, shrink, and change over time). Hence they are called *dynamic* sets [20].

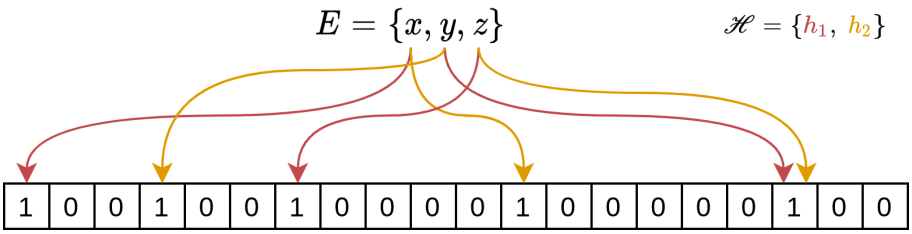


Figure 7.3: **Graphical representation of a Bloom Filter (BF).** Bloom Filter constructed on a set $E = \{x, y, z\}$, where $|E| = 3$. And two hash functions that are part of a set $\mathscr{H} = \{h_1, h_2\}$, where $|\mathscr{H}| = k = 2$. Notice that $h_1(y)$ and $h_2(z)$ cause a collision. Image taken from [13].

positives but not false negatives. That is, a query has always two possible answers: *"possibly in set"* or *"definitely not in set"*.

Other authors have extensively discussed how to estimate the probability to get a false positive after n elements have been added to a Bloom Filter [19]. Such a probability for a presence/absence query for a Bloom Filter built from n elements, with k hash functions and a filter of size m (number of bits in the BF) can be approximated as:

$$p \approx (1 - \frac{1}{e^{\frac{kn}{m}}})^k \tag{7.3}$$

Notice that in Equation 7.3:

- If there are more elements n to store in the BF, then the false positive rate will increase.
- If k increases, there will me more computations and a lower false positive rate as k approaches k_{opt}. Minimising the probability of false positives with respect to k, for a given m and n values gives the following optimal number of hash functions:

$$k_{opt} = \frac{m}{n} \ln 2 \tag{7.4}$$

- The larger m is, the more space in memory is needed and the lower the false positive rate. Assuming the use of the optimal number of hash functions, the number of bits m for a desired number of elements n and a false positive rate p is defined as:

$$m = -\frac{n \ln p}{(\ln 2)^2} \tag{7.5}$$

References

[1] Key, F. M., Posth, C., Krause, J., Herbig, A., and Bos, K. I. "Mining metagenomic data sets for ancient DNA: recommended protocols for authentication". In: *Trends in Genetics* vol. 33, no. 8 (2017), pp. 508–520.

[2] Der Sarkissian, C., Velsko, I. M., Fotakis, A. K., Vågene, Å. J., Hübner, A., and Fellows Yates, J. A. "Ancient metagenomic studies: Considerations for the wider scientific community". In: *Msystems* vol. 6, no. 6 (2021), e01315–21.

[3] Farrer, A. G., Wright, S. L., Skelly, E., Eisenhofer, R., Dobney, K., and Weyrich, L. S. "Effectiveness of decontamination protocols when analyzing ancient DNA preserved in dental calculus". In: *Scientific reports* vol. 11, no. 1 (2021), pp. 1–14.

[4] Schmieder, R. and Edwards, R. "Fast identification and removal of sequence contamination from genomic and metagenomic datasets". In: *PloS one* vol. 6, no. 3 (2011), e17288.

[5] Martí, J. M. "Recentrifuge: Robust comparative analysis and contamination removal for metagenomics". In: *PLoS Computational Biology* vol. 15, no. 4 (2019), e1006967.

[6] Li, H. and Durbin, R. "Fast and accurate long-read alignment with Burrows–Wheeler transform". In: *Bioinformatics* vol. 26, no. 5 (2010), pp. 589–595.

[7] Li, H. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv preprint arXiv:1303.3997* (2013).

[8] Oliva, A., Tobler, R., Cooper, A., Llamas, B., and Souilmi, Y. "Systematic benchmark of ancient DNA read mapping". In: *Briefings in Bioinformatics* vol. 22, no. 5 (2021), bbab076.

[9] Poullet, M. and Orlando, L. “Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes”. In: *Frontiers in Ecology and Evolution* vol. 8 (2020), p. 105.

[10] Oliva, A., Tobler, R., Llamas, B., and Souilmi, Y. “Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment”. In: *Ecology and evolution* vol. 11, no. 24 (2021), pp. 18743–18748.

[11] Federhen, S. “The NCBI taxonomy database”. In: *Nucleic acids research* vol. 40, no. D1 (2012), pp. D136–D143.

[12] Bloom, B. H. “Space/time trade-offs in hash coding with allowable errors”. In: *Communications of the ACM* vol. 13, no. 7 (1970), pp. 422–426.

[13] Lemane, T. “Indexing and analysis of large sequencing collections using k-mer matrices”. PhD thesis. Université de Rennes 1, 2022.

[14] Stranneheim, H., Käller, M., Allander, T., Andersson, B., Arvestad, L., and Lundeberg, J. “Classification of DNA sequences using Bloom filters”. In: *Bioinformatics* vol. 26, no. 13 (2010), pp. 1595–1600.

[15] Melsted, P. and Pritchard, J. K. “Efficient counting of k-mers in DNA sequences using a bloom filter”. In: *BMC bioinformatics* vol. 12, no. 1 (2011), pp. 1–7.

[16] Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., and Brown, C. T. “Scaling metagenome sequence assembly with probabilistic de Bruijn graphs”. In: *Proceedings of the National Academy of Sciences* vol. 109, no. 33 (2012), pp. 13272–13277.

[17] Chikhi, R., Limasset, A., and Medvedev, P. “Compacting de Bruijn graphs from sequencing data quickly and in low memory”. In: *Bioinformatics* vol. 32, no. 12 (2016), pp. i201–i208.

[18] Lemane, T., Medvedev, P., Chikhi, R., and Peterlongo, P. “kmtricks: Efficient and flexible construction of Bloom filters for large sequencing data collections”. In: *Bioinformatics Advances* (2022).

[19] Tarkoma, S., Rothenberg, C. E., and Lagerspetz, E. “Theory and practice of bloom filters for distributed systems”. In: *IEEE Communications Surveys & Tutorials* vol. 14, no. 1 (2011), pp. 131–155.

[20] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to algorithms*. MIT press, 2022.

[21] Broder, A. and Mitzenmacher, M. “Network applications of bloom filters: A survey”. In: *Internet mathematics* vol. 1, no. 4 (2004), pp. 485–509.

Chapter 8

aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets

Camila Duitama González [1], **Samarth Rangavittal, Riccardo Vicedomini, Hugues Richard, Rayan Chikhi**

Published in *iScience*, November 2023, volume 26, issue 11, pp. 123–456. DOI: 10.1000/182.

8.1 Motivation

After having proposed a successful solution for the problem of contamination assessment in ancient oral metagenomic samples using `decOM`, an intuitive consecutive question emerged regarding the process of eliminating contaminant reads once it is established (via MST for contamination assessment, for example) that a sequenced sample contains endogenous DNA. The removal of contamination at the read-level posed an intriguing challenge to overcome, given that the study of ancient microbial genomes involves extracting from limited biomaterials that are exceedingly difficult to resample. Therefore, employing digital contamination removal measures could potentially optimise the utilisation of the available genetic data.

Existing methods for contamination removal such as DeconSeq (see Section 7.2 and Recentrifuge (see Section 7.3) are not tailored for aDNA. Moreover, they require sequenced negative controls or an index of reference genomes to distinguish contaminant from non contaminant material. For these reasons we proposed a novel method called `aKmerBroom`, a reference-free decontamination tool tailored for the removal of contaminant DNA. Our tool performs a two-step lookup in a Bloom Filter (BF) (see section 7.4) constructed from oral k-mers, followed by a lookup in a a set of "anchor reads". This method shows high specificity and sensitivity on real and synthetic data, which is why we anticipate it will be a valuable tool. By avoiding the wastage of useful material that happens to be contaminated and by mitigating the potential for misleading outcomes in subsequent analyses, `aKmerBroom` optimises the processing of ancient microbial metagenomic samples.

Abstract

Dental calculus samples are modelled as a mixture of DNA coming from dental plaque and contaminants. Current computational decontamination methods such as Recentrifuge and DeconSeq require either a reference database or sequenced negative controls, and therefore have limited use cases. We present a reference-free decontamination tool tailored for the removal of contaminant DNA of ancient oral sample called `aKmerBroom`. Our tool builds a Bloom Filter of known ancient and modern oral k-mers, then scans an input set of ancient metagenomic reads using multiple passes to iteratively retain reads likely to be of oral origin. On synthetic data, `aKmerBroom` achieves over 89.53% sensitivity and 94.00% specificity. On real datasets, `aKmerBroom` shows higher read retainment (+60% on average) than other methods. We anticipate `aKmerBroom` will be a valuable tool for the processing of ancient oral samples as it will prevent contaminated datasets from being completely discarded in downstream analyses.

8.2 Contents

[1]Institut Pasteur, 25-28 Rue du Dr Roux, 75015 Paris, France, cduitama@pasteur.fr

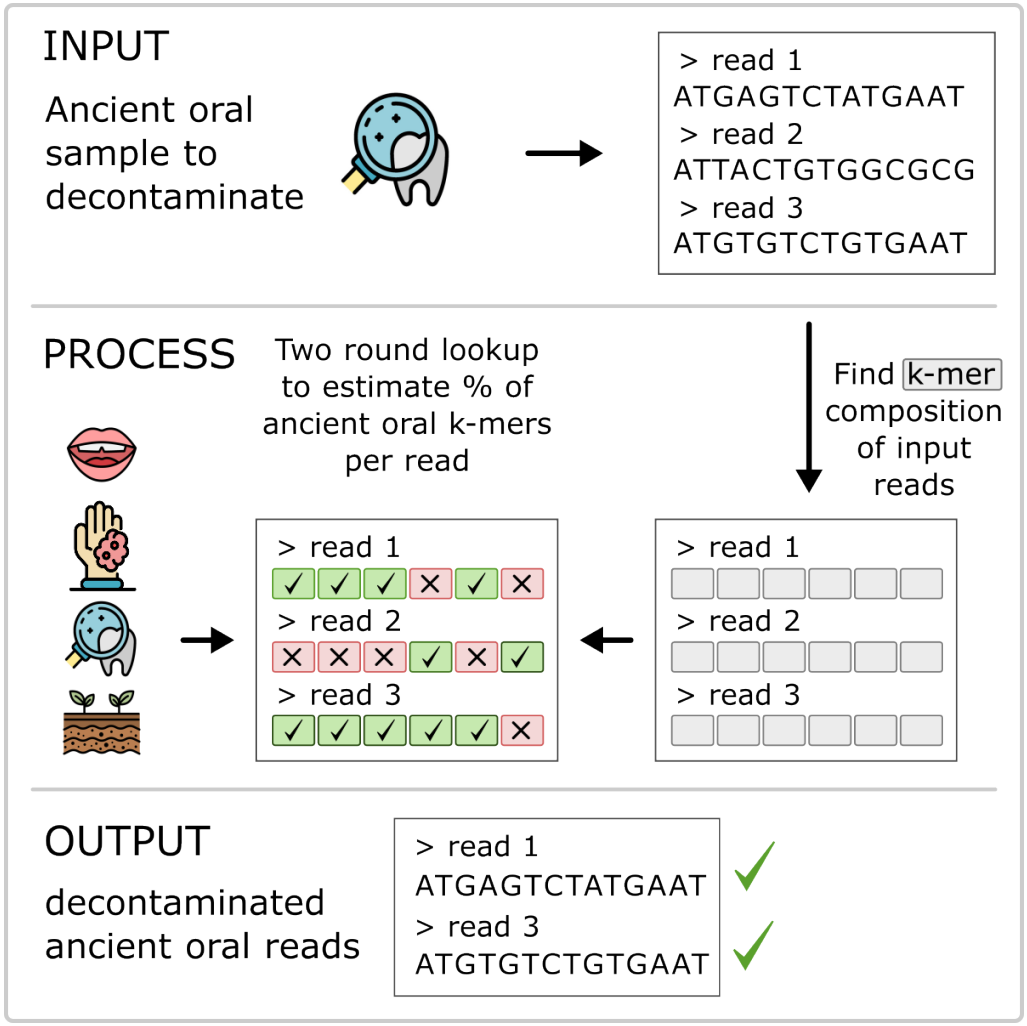


Figure 8.1: Graphical abstract for `aKmerBroom`.

8.3 Introduction

Ancient human dental calculus is a rich source of information on the oral microbial community that allows the study of the oral microbiome evolution, human oral health and diet [1]. It is one of the most relevant sources of isolation in the field of paleometagenomics as it is one of the richest sources of ancient DNA (aDNA) and a crucial reservoir of ancient microbial communities [2, 3]. However, such samples are highly susceptible to contamination from environmental sources, which can drastically alter the microbial composition and lead to erroneous conclusions after downstream analyses [4]. Several studies have shown that contaminant DNA and cross-contamination can confound metagenomic studies, and low microbial biomass samples are particularly vulnerable to contamination[5, 6, 7]. Under these circumstances, contamination estimation and removal are fundamental to avoid the aforementioned risks [8, 5]. In this work, we focus on the removal of contaminated sequences in ancient oral metagenomes.

There are standardised laboratory protocols for the decontamination of ancient DNA (aDNA) samples, guidelines to minimize contamination [1, 9], as well as bioinformatics pipelines for aDNA authentication [10, 11, 12]. For human aDNA, authentication requires to single out genuine ancient human DNA (normally based on characteristic damage patterns and endogenous content [13]), as contamination from field scientists can occur at any stage, from the excavation to the DNA library preparation [10]. In paleometagenomics, it is often very difficult to resample from rare and precious biomaterials[13, 4]. This makes decontamination procedures crucial to ensure the best use of available genetic information, while maintaining low levels of contamination. Apart from wet-lab based methods for contamination control (e.g experimental methods), tools such as DeconSeq[14] or Recentrifuge[15] have digital procedures to remove genomics sequences that correspond to negative control samples. They are however not tailored for aDNA and require either a database of negative controls or an index of reference genomes to distinguish the contamination that should be removed from endogenous material. Moreover, due to the nature of the biosamples processed in aDNA studies, researchers face the challenge of having small sample sizes, a typical feature of the ancient metagenomics field that often leads to underpowered studies [13].

In principle one could also perform read decontamination by read mapping, for instance, to a database of oral microbiota reference genomes while keeping only the reads that align with sufficient identity. However such a reference database does not exist, and the diversity of ancient oral microbiomes is not yet well characterized[16]. Alternatively, one could decontaminate a sample by taking out reads that align to a database of known contaminant reference genomes, such as soil and skin microbes – however also no such database exists and is unlikely to be created given the extensive diversity of these environments[17]. Hence, mapping-based approaches are currently unsuitable for the decontamination of ancient metagenomes, and one must rely on alternative approaches, such as the one presented here.

Terabytes of ancient metagenomic data exist in public repositories, and also petabytes of metagenome data have been produced over diverse environments. As an attempt to globally make this huge amount of data accessible, bioinformaticians have developed efficient algorithmic methods to aggregate substrings of genomic sequences of length k, called k-mers, within these collections. Using tools such as kmtricks[18] one can rapidly construct a matrix of k-mers from large metagenomic collections, allowing to jointly analyze all k-mers present within hundreds to thousands of metagenomic samples. However, such aggregation of k-mer information over hundreds of metagenomes has never been applied to the problem of decontaminating ancient DNA reads yet.

We developed `aKmerBroom`, the first method able to decontaminate ancient oral DNA samples without the need for a control sample nor an extensive set of reference genomes. Our method leverages the wealth of existing ancient oral metagenomes by constructing a database of ancient oral k-mers used to capture reads likely to be of ancient oral origin. In essence, `aKmerBroom` projects the k-mers from an input sample onto a database of reference k-mers and then selects the reads with enough coverage. Technically `aKmerBroom` performs a two-step lookup in a Bloom Filter (BF) of oral k-mers, and then in a set of "anchor" reads. We evaluate `aKmerBroom` on three distinct synthetic datasets and on two real datasets and compare the results with current computational methods for contamination removal. Given its high sensitivity and specificity, `aKmerBroom` is expected to be a useful tool for decontaminating ancient oral samples.

8.3.1 Related work

The advent of large scale metagenomic projects such as the Human Microbiome Project [19, 20], the Earth Microbiome Project [21], Tara Ocean [22] or MetaSub [23] among others, has generated large collections of modern metagenomic sequencing data that has fundamentally changed the study of microbial ecology. Other studies, at a smaller scale, still produced considerable amounts of ancient metagenomic data (approximately 1,000 sequencing runs)[24]. All these sequencing efforts came with increasing amounts of experimental noise, e.g. contamination which plagues both modern and ancient metagenomics. By contamination, we refer here to the observation of sequenced reads in a sample coming from microorganisms that were not originally part of that sample of interest[25].

There are several computational pipelines tailored for the detection of contaminating DNA after sequencing has been performed [26]. Yet we are not aware of tools developed specifically for contamination removal in ancient oral DNA at the read level, despite this sample type being one of the most prevalent source of ancient DNA.

DeconSeq [14], published in 2011, is a method built to detect and identify contamination in microbial metagenomes[27]. It takes as input a set of reads, and compares it against a reference database using a modified version of the BWA-SW algorithm[28]. DeconSeq uses different databases depending on whether the user wants to remove or retain reads. None of the databases were built for ancient oral metagenomic decontamination. The user might create their own ancient index for contaminant screening but this requires having a reference of control samples and increases the running time for contamination removal.

A previous study suggested that the use of negative controls alone is insufficient to inform researchers of measures to minimize contaminants[6]. Tools such as decontam[29] use pre-sequenced quantification data such as Operational Taxonomic Unit (OTU) tables, and remove contaminant taxa from such tables but do not remove contaminants at the read level. On the other hand, microDecon[30] uses proportions of contaminant OTUs from blank samples (negative sequencing controls processed with the same DNA/ PCR amplification kits as the real samples, sequenced on the same run [31]), and also adjust read counts in OTU tables but does not decontaminate the reads themselves. As they are control-based those methods do not account for cross-contamination.

Finally another tool, Recentrifuge [15], identifies cross-contaminations, i.e. DNA exchange between samples within one same study that can create batch effects [29, 32]. It is based on Centrifuge[33], a taxonomic classifier that uses the Burrows–Wheeler Transform (BWT) and a FM-index to store and index a reference database. Recentrifuge reads the score given to the reads by a taxonomic classification software (such as Centrifuge), and uses this information to calculate an average confidence level for each taxon in the taxonomic tree associated with the sample analysed.

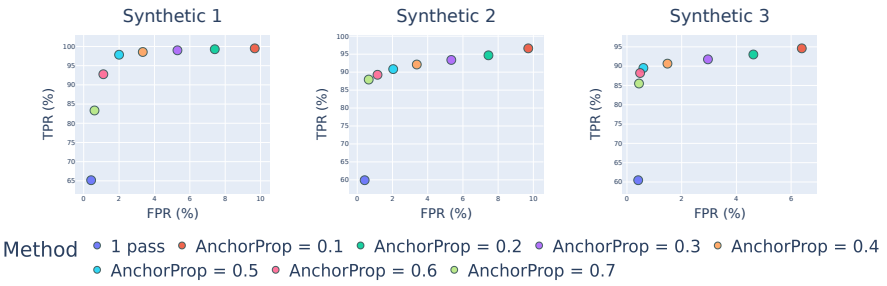


Figure 8.2: **Reciever Operating Characteristic (ROC) curve for selection of anchor proportion threshold**. We optimised the threshold to decide if a read would be classified as ancient or not by running `aKmerBroom` with different values of the parameter τ (the proportion of k-mers found in the anchor k-mers set) and evaluating every run with sample Synthetic 1. As seen on the left panel, the value of τ that has the best trade-off between a high True Positive Rate (TPR) and a low False Positive Rate (FPR) is 0.5. We additionally evaluated an earlier version of `aKmerBroom` that did not include matches against anchor reads and performed only one lookup step in the BF, represented with the blue marker called 1 Pass. Results for samples Synthetic 2 and Synthetic 3 are presented in the middle and right panel respectively.

Tools of this kind rely on sequencing blank samples (controls) to determine baseline contaminant levels of microbes.

To summarise, existing methods are not tailored for the decontamination of ancient oral metagenomic projects, as they are reference-based, and have not been recently updated to scale up to modern dataset sizes (such as DeconSeq) or rely on the sequencing of controls which has limited uses (such as microDecon or Recentrifuge). To remedy this, we propose `aKmerBroom` as a fast, reference-free and precise tool for the decontamination of ancient dental calculus samples.

8.4 Results

8.4.1 Datasets

To evaluate `aKmerBroom` in a controlled setting with known levels of contamination, we constructed three distinct synthetic datasets, corresponding to various scenarios. The Synthetic 1 and Synthetic 2 datasets correspond to the case where all or part of the reads observed in the target sample are from samples used to construct the trusted oral k-mer set, and are thus easier to decontaminate. The Synthetic 3 dataset corresponds to a case where we observe a completely new and unseen sample. Each dataset is built with an equal number of reads belonging to each of the three categories aOral, Sediment/Soil, and Skin, in a 1/3:1/3:1/3 proportion. We also used two real datasets from an ancient oral microbiome study. Table 8.1 presents the datasets.

- **Synthetic 1**: We collected 2 million reads from a source soil dataset, 2 million reads from an aOral sample and 2 million reads from a skin sample. All these source samples were present in the k-mer matrix used to create the trusted k-mers set, hence this is a best-case scenario for decontamination.
- **Synthetic 2**: A second dataset was built by sampling 2 million reads from an external aOral sample that was **not** used to create the trusted k-mers set. We added the 2 million reads from the skin and sediment/soil datasets used for Synthetic 1. Hence this dataset is a semi-artificial best-case scenario.
- **Synthetic 3**: A third and final synthetic dataset was built by sub-sampling reads from aOral, soil and skin datasets that were **not** used to create the trusted k-mers set. For the construction of this dataset, 2 million reads were sub-sampled from am aOral sample, a sediment/soil sample, and a skin sample, respectively.
- **Real data**: Lastly, we evaluated decontamination on two real datasets: First, an aOral sample (accession ERR5670971), isolated from Trentino-South Tyrol, Italy, and dating from the Early Middle Ages (400–1000 CE). Second, a real dataset (accession ERR5670966) isolated from Venosta Valley and dating from the Early Middle Ages too[16]. None of these datasets were used to create the set trusted k-mers. A negative control sequenced and published in the same study was used to run Recentrifuge and DeconSeq, but not `aKmerBroom`.

Dataset	aOral source	Skin source	Sediment/Soil source	nReads (M)	Used to build BF
Synthetic 1	SRR12462946	SRR1620017	ERR671934	6	Entirely
Synthetic 2	SRR13355797	SRR1620017	ERR671934	6	Partially
Synthetic 3	ERR3003655	SRR11426385	ERR3458820	6	No
Real 1	ERR5670971	ERR5670972		64.6	No
Real 2	ERR5670966	ERR5670972		47.8	No

Table 8.1: **Composition of synthetic and real datasets.** For the real dataset, the accession reported for the aOral source corresponds to the sample likely to contain ancient oral microbes, to be decontaminated. The sample reported in the real datasets Sediment/Soil source is a negative control.

8.4.2 Evaluation method

As we know the exact number of reads coming from the aOral sample in each of the three balanced synthetic datasets, we estimated specificity and sensitivity by calculating the True Positive Rate (TPR) and False Positive Rate (FPR). We considered as true ancient oral any read recovered by `aKmerBroom` coming from the aOral samples, and considered as false aOral the reads coming from the soil/skin samples. On the other hand, as we do not know the true number of contaminant reads for the real dataset, we evaluated performance by measuring read retainment, that is the percentage of original reads that were kept after contamination removal.

Competing decontamination methods such as Recentrifuge and DeconSeq were only evaluated on real data since they require negative controls or reference databases which were not available for our 3 synthetic samples. Recentrifuge relies on Centrifuge [33] for taxonomic classification of an input set of reads. We used Centrifuge version 1.0.4-beta on a pre-made index of RefSeq bacteria, archaeal, viral, human sequences [34]. DeconSeq standalone version 0.4.3 was used for performance comparison against `aKmerBroom` on real data. To evaluate the composition of the samples before and after decontamination, we performed a contamination assessment with SourceTracker and using as sources our reference database of 360 metagenomic samples.

8.4.3 Evaluation of decontamination on synthetic data

Dataset	Sensitivity (%)	Specificity (%)
Synthetic 1	97.85	98.00
Synthetic 2	90.84	97.96
Synthetic 3	89.53	94.00

Table 8.2: **Performance of `aKmerBroom` on synthetic samples**. Sensitivity is the percentage of aOral reads that were successfully retained. Specificity is the percentage of non-aOral reads that were successfully removed.

Table 8.2 reports that `aKmerBroom` has excellent performance ($\geq 93\%$ sensitivity and specificity) on synthetic datasets 1 and 2. Synthetic dataset 3 was built by sub-sampling from datasets that were not seen during construction of the trusted k-mers set, hence it is a more realistic case. Here `aKmerBroom` still performs remarkably well with 89.57% sensitivity and 94.00% specificity, albeit shows lower sensitivity than in the first two synthetic datasets. Contamination assessment analyses using SourceTracker (Figure 8.3) show that after decontamination with `aKmerBroom` the final oral composition is above 80% in the three synthetic datasets. This proves that also with alternative metrics to sensitivity and specificity, such as source environment proportions given by MST analyses, our method performs contamination removal effectively.

8.4.4 Evaluation of decontamination on real data

When evaluating `aKmerBroom` on real data, we measured performance with read retainment and compared results with two competing methods: DeconSeq and Recentrifuge. We took two aOral metagenomic samples isolated from the dental calculus microbiome of two people buried in Italy in the Early Middle Ages (400–1000 CE) [16]. Researchers of this study also published a sequenced blank, which we used to create the database for contaminant screening and run DeconSeq with. That same blank was used as negative control when running Recentrifuge, in order to have a reference of taxa that needs to be removed.

The group that collected, sequenced and published those real datasets performed several aDNA authentication analyses to prove their samples were representative of the ancient calculus microbiome. Among others, they ran SourceTracker[35] on the aOral samples and showed that

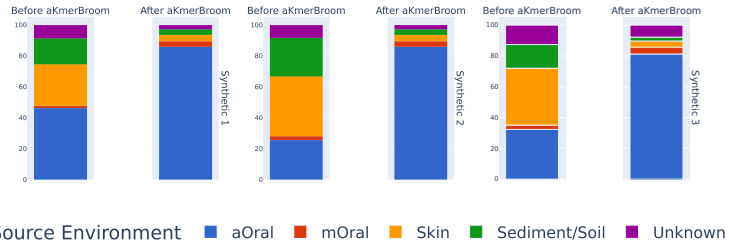


Figure 8.3: **aKmerBroom performance on synthetic data as evaluated by SourceTracker**. We evaluated the source environment composition of each synthetic sample before and after decontamination with aKmerBroom using SourceTracker and our reference collection of 360 metagenomic samples as sources.

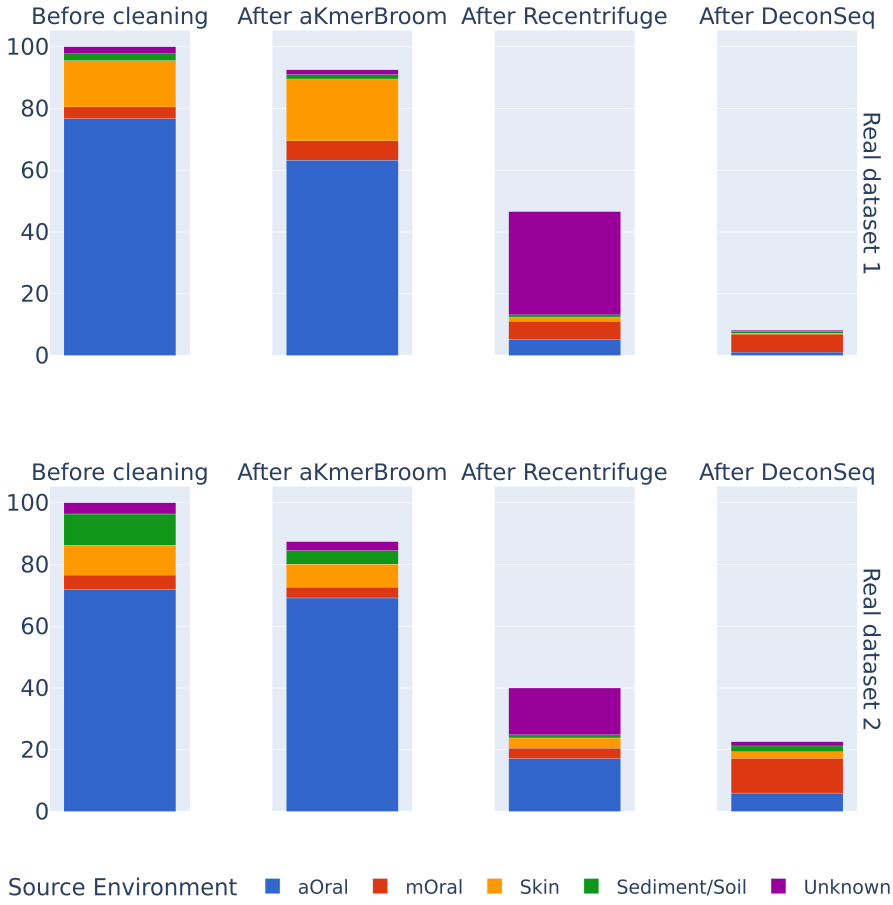


Figure 8.4: **aKmerBroom performance on real data as evaluated by SourceTracker**. Bar plots were scaled according to the percentage of reads retained per decontamination method.

the reads stemming from a known source were predominantly coming from modern calculus and plaque [16], i.e. oral sources. Thus we expect a highly reliable ancient oral content in the real sample evaluated, and a low level of contamination. For this reason we used read retainment and confirmed that aKmerBroom preserves most of the reads of the original aOral sample (92.56%), whereas Recentrifuge and DeconSeq remove most of the sequences (see Table 8.3).

We additionally performed an evaluation of the real samples of ancient oral origin, by running mSourceTracker on each of the samples and against a set of sources represented with an OTU table built from our reference collection of 360 metagenomic samples (sources: ancient oral (aOral), modern oral (mOral), Sediment/Soil and Skin) (further details on the OTU table construction and taxonomic classifier used are detailed in the Supplementary material of decOM [36]). Results are presented in Figure 8.3.

Dataset	Run accession	Method	Reads retained (%)	O.C. after(%)	O.C before(%)
Real dataset 1	ERR5670971	DeconSeq	8.16	84.00	76.69
		Recentrifuge	46.64	23.78	
		aKmerBroom	92.56	75.11	
Real dataset 2	ERR5670966	DeconSeq	22.63	75.89	71.89
		Recentrifuge	40.00	51.19	
		aKmerBroom	87.42	83.09	

Table 8.3: **Decontamination performance on two real datasets**. Four methods were run to decontaminate two samples. For DeconSeq and Recentrifuge, corresponding negative controls were provided as input too. The nReads column shows the total number of reads in case and control samples. The column O.C. (Oral Content) refers to the proportion of oral source environment in the sample *after* and *before* contamination removal with each of the methods, as estimated by SourceTracker.

8.4.5 Computational performance

Using the pre-constructed Bloom Filter of oral k-mers, `aKmerBroom` has a runtime of around 1 hour for a dataset with fewer than 10 million reads, while using approximately 10 Gb of memory. Beyond this input size, the run time and memory requirement scales linearly with the number of unique anchor k-mers in the input reads. Note that if a new Bloom Filter has to be constructed from scratch, this one-time step would take around 6 hours for a file of 1 billion k-mers.

Leaving out the time to build the Bloom Filter or index the control/reference database and evaluating running time on Real dataset 2, a FASTA file of almost 48 million reads, DeconSeq took around 1 day to run, `aKmerBroom` took around 4 hours and Recentrifuge 2 hours. Both DeconSeq and Recentrifuge were run using 2 Gb of memory.

8.5 Discussion

Decontaminating ancient oral metagenomes is a challenging computational problem, currently poorly performed using off-the-shelf tools. This work highlights that current ancient metagenomic studies are hindered by suboptimal decontamination methods. We propose `aKmerBroom`, a tool for contamination removal of ancient oral datasets using a Bloom Filter constructed on a set of trusted oral k-mers, using a large collection of metagenomes.

We evaluated `aKmerBroom` with three distinct synthetic metagenomic datasets subsampling from sample sources that were fully, partially and non included in the construction of the Bloom Filter, and obtained 97.85%, 93.39% and 89.53% sensitivity and 98.00%, 97.96% and 94.00% specificity (respectively). We further measured `aKmerBroom` performance on two real samples and quantified the percentage of ancient oral sample preserved. `aKmerBroom` effectively preserves most of the original sample, and removes contaminant reads as estimated by SourceTracker, whereas other methods (Recentrifuge, DeconSeq) discard over 53% of the sequences and remove true ancient oral content, also as estimated by SourceTracker.

k-mer-based methods such as `aKmerBroom` are relevant to modern day DNA analyses because they are reference-free (e.g. they do not require a database of reference genomes) and they make use of the large corpus of genomic information that has been gathered over the years. Since we use a k-mer-based consensus of samples to decide what to keep, but we do not decide which species specifically are present/absent in the input sample, our method does not suffer from biases coming from using OTU tables or reference databases. Others have reported using k-mers to assess contamination in human whole-genome samples by doing meta analyses across different datasets [37]. Thus a trend emerges on using stored genetic information to tackle the problem of contamination assessment and removal, instead of making it a matter that is unique to each study.

As it simplified the implementation, the second round of lookups was performed using an exact membership data structure (set). Yet as a future step, performance improvements can be made to reduce the memory requirement for large input files with a significant proportion of ancient reads. For example, this second round could also be implemented using a Bloom Filter. This way the memory required can then be independent of the size of the input reads dataset.

The fact that ancient metagenomic samples are rare and have low biomass of ancient remains often translates into underpowered aDNA studies. As more and more ancient metagenomic projects are published, tools such as `aKmerBroom` emerge as a novel and efficient way of incorporating data from previous studies by concisely storing information in the form of a Bloom Filter. Unlike other methods, `aKmerBroom` represents the variety of ancient oral metagenomic material across several BioProjects, while not making specific assumptions about the microbial species that should be expected or ignored. It mitigates the effect of small sample size (as the output of

several metagenomics studies are put together to construct the Bloom Filter) while still making computationally manageable analyses.

`aKmerBroom` brings usability improvements to decontamination methods. Prior to it, users had to make decisions on how to properly carry out analyses. For instance, in the case of Recentrifuge, one needs to estimate whether to run the taxonomic classifier Centrifuge with default or modified parameters, selecting for a pre-made index or building an index with the criteria of the user, which is equivalent to curating a database that ideally would be tailored for ancient oral decontamination. In the case of DeconSeq, users have to select either a "retain" and/or a "remove" database, plus other alignment options that affect BWA-SW results. All these decisions are required even for non-expert users, and they have not been properly benchmarked for aDNA analysis, ultimately leading to sub-optimal results. Although it is out of the scope of this paper to do parameter optimization on all methods to tailor them for ancient oral datasets, we introduce here a method that overcomes much of the parameter selection and database creation burdens that exist in the other decontamination tools.

Some researchers have often emphasised the importance of including negative controls to understand background contamination [38]. While others have focused on implementing the strategy of identifying a "contaminome" profile or list of possible contaminant taxa, to then remove it from the studied sample [31, 37]. The latter, however, rises doubts on whether it can really take into account the possibility that contaminants may come from other samples within the same study[25]. One interesting future work would be to specifically test for this between-sample contamination using `aKmerBroom` and compare performance with methods such as Recentrifuge that are tailored specifically to tackle cross-contamination.

8.6 Limitations of the study

We rely on the metadata of each metagenomic sample to assign a true label (i.e. environment type), however, there is no ground truth as to what is the true proportion of aOral, mOral, Sediment/Soil or skin content in any of them. Overall, one of the biggest challenges in the field of paleometagenomics is that there is not a straightforward (taxonomic) characterisation of an ancient oral metagenome, modern oral metagenome or a contaminant metagenome. Following that line of thought, we acknowledge that our creation of a set of trusted oral k-mers is only an approximation to what a "clean" ancient oral set of k-mers might look like, but there is no way to know for sure that this set of k-mers *only* contains ancient oral DNA. On the other hand, we allow users to input their own set of k-mers for the construction of their own BF, with the hope that experts in the field might be able to come up with their own trusted set of oral k-mers validated by their biological understanding of the problem.

We acknowledge read retainment (percentage of original reads that were kept after contamination removal) is a proxy for how clean a sample is after decontamination, but there might be additional analyses that can be done to authenticate the true content of an ancient oral sample after using `aKmerBroom`, that take into account additional biological information such as deamination or fragmentation patterns.

As most of the k-mers that were used for the construction of the trusted set of oral k-mers come from Illumina HiSeq reads, we know that a limitation of `aKmerBroom` is that it may not show the same performance with higher error rates.

Despite having effectively used a two-round lookup using a Bloom Filter and a set to construct `aKmerBroom` and decontaminate synthetic and real ancient oral data, there might be more memory-efficient data structures to effectively perform the same task that are worth exploring in the future.

8.7 Methods

8.7.1 Key Resources Table

8.7.2 Resource availability

8.7.2.1 Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Camila Duitama González (cduitama@pasteur.fr)

8.7.2.2 Materials availability

This study did not generate new unique reagents.

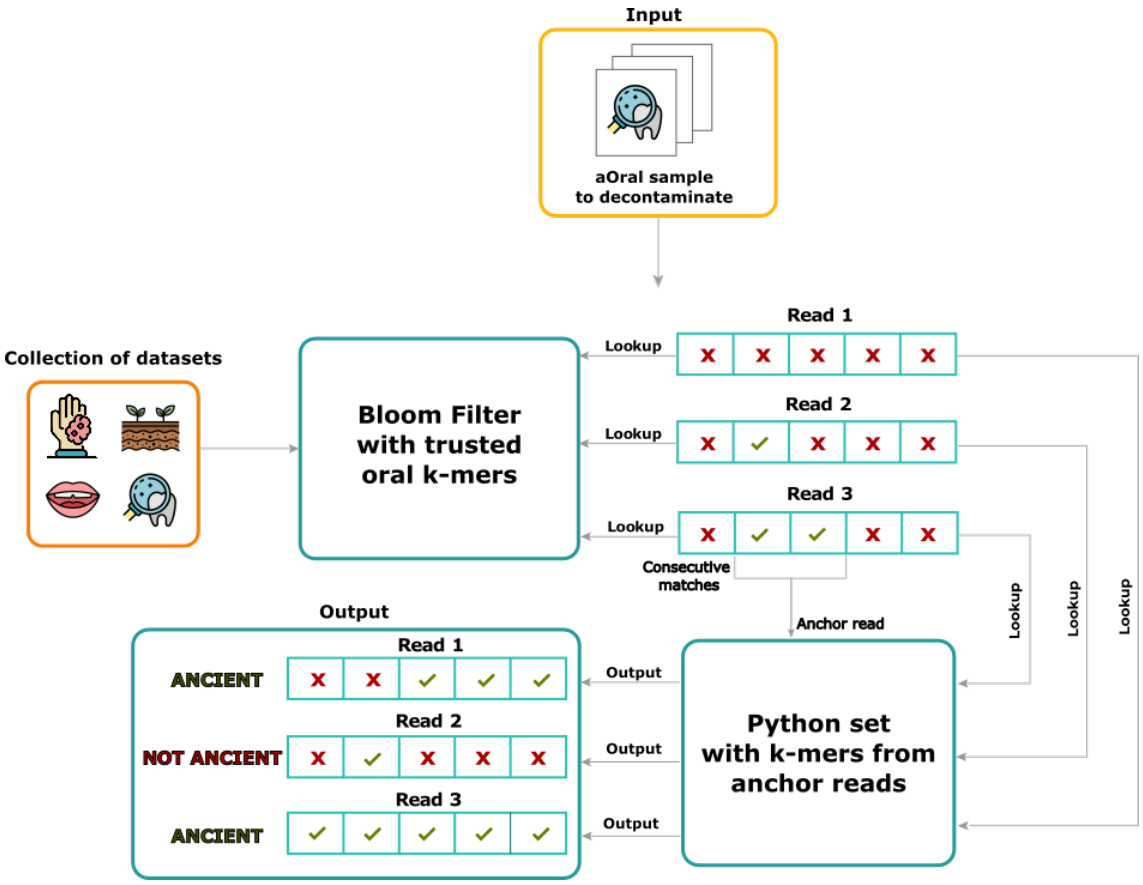


Figure 8.5: **aKmerBroom pipeline**. First, an offline step is performed: a collection of samples representative from diverse sources is used to create a trusted set of oral k-mers. The trusted collection indexes k-mers that appear exclusively in modern and ancient oral samples, but not other samples from contaminant sources (see panel on the left called Collection of datasets). Then this set of oral k-mers is used to decontaminate an input set of reads. The algorithm proceeds by looking up each read k-mer inside the Bloom Filter of trusted oral k-mers, and marking positions of matches. Reads having at least two consecutive matches to the Bloom Filter get passed to the construction of a set containing all k-mers from such reads. Finally, the same input reads are scanned again using the aforementioned set, and reads having a proportion of k-mer matches over a certain threshold are reported to be of ancient oral origin.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
aOral source for Synthetic 1	(Jacobson et al. 2020)	SRR12462946
Skin source for Synthetic 1	(Turnbaugh et al. 2007) HMP	SRR1620017
Sediment/Soil source for Synthetic 1	(Bissett et al. 2016) BASE Project	ERR671934
aOral source for Synthetic 2	(Farrer et al. 2021)	SRR13355797
aOral source for Synthetic 3	(Velsko et al. 2017)	ERR3003655
Skin source for Synthetic 3	(Kim et al. 2022)	SRR11426385
Sediment/Soil source for Synthetic 3	(Cribdon et al. 2020)	ERR3458820
Real 1 aOral sample	(Farrer et al. 2021)	ERR5670971
Negative control for real samples	(Farrer et al. 2021)	ERR5670972
Real 2 aOral sample	(Farrer et al. 2021)	ERR5670966
Deposited Data		
test_1	this paper	doi: 10.5281/zenodo.7590899
test_2	this paper	doi: 10.5281/zenodo.7590899
test_3	this paper	doi: 10.5281/zenodo.7590899
Software and Algorithms		
aKmerBroom	this paper	https://zenodo.org/record/7156306

Table 8.4: Key Resources Table

8.7.2.3 Data and code availability

- Data:

  - This paper analyses existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
  - Synthetic data have been deposited in a Zenodo repository `https://zenodo.org/record/7590899#.Y9lQ_y9w0Us` and are publicly available as of the date of publication. DOIs are listed in the key resources table.

- Code:
  - All original code has been deposited at `https://github.com/CamilaDuitama/aKmerBroom` and is publicly available as of the date of publication. DOIs are listed in the key resources table.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

8.7.3 Method details

We have developed `aKmerBroom`, the first method able to perform read-level decontamination on ancient oral metagenomes. As an input to `aKmerBroom`, the user provides a set of reads to be decontaminated. `aKmerBroom` then scans the input reads against a set of oral k-mers, using two passes to iteratively retain reads likely to be of ancient origin.

The main steps are described below, and a high-level summary is provided here. First a set of high quality oral k-mers is determined from a database of ancient and modern oral samples as well as environmental samples. Then, a Bloom Filter is constructed to represent this set approximately in memory. The tool then scans input reads and retains those that have at least 2 consecutive k-mer matches against the filter. Those reads enable us to enrich the set of ancient k-mers by incorporating new putative ancient k-mers. We refer to those reads as "anchor reads". We then perform another pass over the input reads and identify matching reads against this new subset of k-mers. Reads are finally classified as ancient when $\geq 50\%$ of their k-mers match the set of k-mers generated from anchor reads.

8.7.3.1 Creating a set of trusted oral k-mers

To construct a set of trusted oral k-mers for `aKmerBroom`, we use a resource from `decOM` [36], a method for contamination assessment of ancient oral metagenomic samples. `decOM` constructs a k-mer matrix from 360 metagenomic samples covering a wide range of environments around the world, labelled as ancient oral (aOral), modern oral (mOral), and their possible contaminants (Sediment/Soil and Skin samples). Sample accession numbers are provided in the supplementary material of the `decOM` publication[36]. The `decOM` matrix is built over distinct k-mers of size 31, filtered by retaining k-mers that were present in at least 3 samples in the collection and by removing all k-mers seen only once in a sample, which were likely to be sequencing errors. In order to reduce memory usage, we start by subsampling 10% of the k-mers present in the `decOM` matrix, we select a set of a high quality oral k-mers by filtering each k-mer that satisfies all of the following conditions:

- Present in any of the aOral samples or,
- Present in any of the mOral samples and,
- Absent in all Skin samples and,
- Absent in all Soil samples.

In a boolean formula the conditions could be read as (inAOral or inMOral) and not(inSkin or inSoilSediment).

We obtain over 1.5 billion k-mers, which corresponds to roughly 25% of the subsampled set of k-mers matrix (2.5% of whole set of k-mers of the `decOM` matrix of sources). These k-mers are referred to as *trusted oral k-mers*.

We would like to emphasise that our method is reference-free as it does not require a database or index of reference genomes, however, in the construction of the Bloom Filter, there must be a reference of k-mers considered to be of ancient oral origin. We have defined this set after the conditions previously explained, but the user might come up with their own input set of k-mers too.

8.7.3.2 Constructing a Bloom filter from oral k-mers

A Bloom Filter is a space-efficient probabilistic data structure that enables to query the membership of an element within a set, with false positives but no false negatives[45]. As a preprocessing step, `aKmerBroom` constructs a Bloom Filter (BF) from a set of k-mers (using `pybloomfiltermmap`[46]). In `aKmerBroom`, the user may provide their own set of k-mers, or alternatively use the pre-constructed table of trusted oral k-mers provided with the software and constructed as described in the previous section (see Zenodo file [47]). In the upcoming section, we discuss how we mitigate the issue of false positives.

8.7.3.3 Pass 1: Finding anchor reads

In the first pass, `aKmerBroom` scans each read and looks for k-mer matches in the Bloom Filter. If two consecutive k-mer matches are found, a read is marked as an "anchor" read. These anchors will be used in the next pass to identify reads with ancient origin. Note that requiring only two consecutive k-mer matches has the advantage of being permissive, while also avoiding cases when a single false positive match might result in the read being falsely included as an anchor read.

8.7.3.4 Pass 2: Identifying ancient reads

All anchor reads from the first pass are k-merized and stored into a new anchor k-mer set. The full input dataset is scanned again, and reads having a proportion $\geq 50\%$ of k-mers present in this new anchor k-mer set are retained as likely to be of ancient origin. Note that non-anchor reads may be retained, as some will satisfy this criteria. The final output of `aKmerBroom` consists of the set of retained reads.

8.7.3.5 Parameter selection

The `aKmerBroom` method relies on one main parameter: the anchor proportion threshold τ. In addition, the Bloom Filter implementation requires two other parameters: the capacity and the error rate. There is a trade-off between these two parameters: adding less than *capacity* items ensures that the Bloom Filter will have an error rate less than *error rate*[46]. In `aKmerBroom`, we set the error rate to be 0.001, and set the Bloom Filter capacity to be at least as large as the number of trusted k-mers to be stored. By default, we set it to be 2 billion so that it is larger than the 1.5 billion pre-computed trusted k-mers. One could increase the capacity of the Bloom Filter (or decrease the tolerated error rate), but that would result in a larger Bloom Filter and therefore increase memory requirements. To determine an appropriate value for the anchor k-mer proportion threshold τ, we performed a standard grid search from 10% to 90% over Synthetic dataset 1 (see results). As shown in Figure 8.2, we chose a threshold of 50% because it gives us a suitable trade-off between having a high true positive rate (greater than 85%) while also having a low false positive rate (less than 5%). However, the user can also set τ according to their desired sensitivity/specificity trade-off.

Originally we had tried out only one pass over the initial Bloom Filter built from the set of high quality oral k-mers, and we further improved the results by implementing a second pass based on matches against anchor reads identified from the first pass. Results for this one-step version of `aKmerBroom` are also shown in Figure 8.2 ("1 pass"). Notice that the method with only one pass performs worse than any of the thresholded two-pass methods.

8.7.3.6 Output description

`aKmerBroom` outputs an annotated FASTQ file with 4 fields in the record header:

- **`SeqId`**: sequence identifier
- **`ReadLen`**: length of the read
- **`isConsecutiveMatchFound`**: a binary variable to indicate if 2 consecutive k-mers were found in the first lookup.
- **`AnchorProportion`**: percentage of k-mers that were found in the anchor k-mers set.

Software availability. `https://github.com/CamilaDuitama/aKmerBroom`

8.8 Data availability.

The real datasets as well as the metagenomic samples from which the synthetic samples were built are all available with the corresponding accessions seen in Table 8.2. Synthetic datasets 1, 2 and 3 are available in the following link: `https://zenodo.org/record/7590899#.Y9lQ_y9w0Us`

Acknowledgements. R.C was supported by ANR Full-RNA, SeqDigger, Inception and PRAIRIE grants (ANR-22-CE45-0007, ANR-19-CE45-0008, PIA/ANR16-CONV-0005, ANR-19-P3IA-0001). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grants agreements No. 872539 and 956229.

Declaration of interests. The authors declare no competing interests

Author's contribution. Conceptualisation, C.D, S.R and R.C.; Methodology, S.R. and C.D.; Software, S.R.; Validation, C.D.; Resources, R.C.; Writing – Original Draft, C.D, S.R., R.V., R.C, H.R.; Writing – Review & Editing, C.D, S.R., R.V., R.C, H.R; Visualisation, C.D.; Supervision, R.C., H.R., R.V; Project Administration, R.C., H.R.; Funding Acquisition, R.C.

References

[1] Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R. Y., Fiddyment, S., et al. "Pathogens and host immunity in the ancient human oral cavity". In: *Nature genetics* vol. 46, no. 4 (2014), pp. 336–344.

[2] Ziesemer, K. A., Ramos-Madrigal, J., Mann, A. E., Brandt, B. W., Sankaranarayanan, K., Ozga, A. T., Hoogland, M., Hofman, C. A., Salazar-García, D. C., Frohlich, B., et al. "The efficacy of whole human genome capture on ancient dental calculus and dentin". In: *American journal of physical anthropology* vol. 168, no. 3 (2019), pp. 496–509.

[3] Warinner, C., Speller, C., and Collins, M. J. "A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 370, no. 1660 (2015), p. 20130376.

[4] Farrer, A. G., Wright, S. L., Skelly, E., Eisenhofer, R., Dobney, K., and Weyrich, L. S. "Effectiveness of decontamination protocols when analyzing ancient DNA preserved in dental calculus". In: *Scientific reports* vol. 11, no. 1 (2021), pp. 1–14.

[5] Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. "Contamination in low microbial biomass microbiome studies: issues and recommendations". In: *Trends in Microbiology* vol. 27, no. 2 (2019), pp. 105–117.

[6] Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W. T., Wolfe, A. J., Braun, J., and McWeeney, S. "Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments". In: *MSystems* vol. 4, no. 4 (2019), e00290–19.

[7] Scherz, V., Greub, G., and Bertelli, C. "Building up a clinical microbiota profiling: a quality framework proposal". In: *Critical Reviews in Microbiology* vol. 48, no. 3 (2022), pp. 356–375.

[8] Weiss, S., Amir, A., Hyde, E. R., Metcalf, J. L., Song, S. J., and Knight, R. "Tracking down the sources of experimental contamination in microbiome studies". In: *Genome Biology* vol. 15, no. 12 (2014), pp. 1–3.

[9] Adler, C. J., Dobney, K., Weyrich, L. S., Kaidonis, J., Walker, A. W., Haak, W., Bradshaw, C. J., Townsend, G., Sołtysiak, A., Alt, K. W., et al. "Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions". In: *Nature Genetics* vol. 45, no. 4 (2013), pp. 450–455.

[10] Key, F. M., Posth, C., Krause, J., Herbig, A., and Bos, K. I. "Mining metagenomic data sets for ancient DNA: recommended protocols for authentication". In: *Trends in Genetics* vol. 33, no. 8 (2017), pp. 508–520.

[11] Peyrégne, S. and Prüfer, K. "Present-Day DNA Contamination in Ancient DNA Datasets". In: *Bioessays* vol. 42, no. 9 (2020), p. 2000081.

[12] Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiß, C. L., Burbano, H. A., Orlando, L., and Krause, J. "A robust framework for microbial archaeology". In: *Annual Review of Genomics and Human Genetics* vol. 18 (2017), pp. 321–356.

[13] Der Sarkissian, C., Velsko, I. M., Fotakis, A. K., Vågene, Å. J., Hübner, A., and Fellows Yates, J. A. "Ancient metagenomic studies: Considerations for the wider scientific community". In: *Msystems* vol. 6, no. 6 (2021), e01315–21.

[14] Schmieder, R. and Edwards, R. “Fast identification and removal of sequence contamination from genomic and metagenomic datasets”. In: *PloS one* vol. 6, no. 3 (2011), e17288.

[15] Martí, J. M. “Recentrifuge: Robust comparative analysis and contamination removal for metagenomics”. In: *PLoS Computational Biology* vol. 15, no. 4 (2019), e1006967.

[16] Granehäll, L., Huang, K. D., Tett, A., Manghi, P., Paladin, A., O’Sullivan, N., Rota-Stabelli, O., Segata, N., Zink, A., and Maixner, F. “Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal Methanobrevibacter”. In: *Microbiome* vol. 9, no. 1 (2021), pp. 1–18.

[17] Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., Hirsch, P. R., and Vogel, T. M. “Accessing the soil metagenome for studies of microbial diversity”. In: *Applied and environmental microbiology* vol. 77, no. 4 (2011), pp. 1315–1324.

[18] Lemane, T., Medvedev, P., Chikhi, R., and Peterlongo, P. “kmtricks: Efficient construction of Bloom filters for large sequencing data collections”. In: *bioRxiv* (2021).

[19] The Human Microbiome Project Consortium. “Structure, function and diversity of the healthy human microbiome”. In: *Nature* vol. 486, no. 7402 (2012), pp. 207–214.

[20] The Human Microbiome Project Consortium. “A framework for human microbiome research”. In: *Nature* vol. 486, no. 7402 (2012), pp. 215–221.

[21] Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., et al. “A communal catalogue reveals Earth’s multiscale microbial diversity”. In: *Nature* vol. 551, no. 7681 (2017), pp. 457–463.

[22] Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., et al. “A holistic approach to marine eco-systems biology”. In: *PLoS Biology* vol. 9, no. 10 (2011), e1001177.

[23] Osuolale, O., Mason, C., Consortium, M. I., et al. “The metagenomics and metadesign of the subways and urban biomes (MetaSUB) international consortium inaugural meeting report”. In: *Microbiome* vol. 4, no. 1 (2016), p. 24.

[24] Yates, J. A. F., Valtueña, A. A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-López, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., et al. “Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir”. In: *Scientific Data* vol. 8, no. 1 (2021), pp. 1–8.

[25] Minich, J. J., Sanders, J. G., Amir, A., Humphrey, G., Gilbert, J. A., and Knight, R. “Quantifying and understanding well-to-well contamination in microbiome research”. In: *MSystems* vol. 4, no. 4 (2019), e00186–19.

[26] Renaud, G., Schubert, M., Sawyer, S., and Orlando, L. “Authentication and assessment of contamination in ancient DNA”. In: *Ancient DNA. Humana Press, New York* (2019), pp. 163–194.

[27] Jo, J., Oh, J., and Park, C. “Microbial community analysis using high-throughput sequencing technology: a beginner’s guide for microbiologists”. In: *Journal of Microbiology* vol. 58, no. 3 (2020), pp. 176–192.

[28] Li, H. and Durbin, R. “Fast and accurate long-read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* vol. 26, no. 5 (2010), pp. 589–595.

[29] Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. “Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data”. In: *Microbiome* vol. 6, no. 1 (2018), pp. 1–14.

[30] McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., and Zenger, K. R. “microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies”. In: *Environmental DNA* vol. 1, no. 1 (2019), pp. 14–25.

[31] Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses”. In: *BMC Biology* vol. 12 (2014), pp. 1–12.

[32] Nguyen, N. H., Smith, D., Peay, K., and Kennedy, P. “Parsing ecological signal from noise in next generation amplicon sequencing”. In: *New Phytologist* vol. 205, no. 4 (2015), pp. 1389–1393.

[33] Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. "Centrifuge: rapid and sensitive classification of metagenomic sequences". In: *Genome Research* vol. 26, no. 12 (2016), pp. 1721–1729.

[34] Ben Langmead. *Centrifuge indexex*. URL: `https://benlangmead.github.io/aws-indexes/centrifuge` (visited on 01/28/2023).

[35] Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. "Bayesian community-wide culture-independent microbial source tracking". In: *Nature Methods* vol. 8, no. 9 (2011), pp. 761–763.

[36] Duitama, C. *decOM*. 2022. URL: `https://github.com/CamilaDuitama/decOM` (visited on 05/17/2022).

[37] Chrisman, B., He, C., Jung, J.-Y., Stockham, N., Paskov, K., Washington, P., and Wall, D. P. "The human "contaminome": bacterial, viral, and computational contamination in whole genome sequences from 1000 families". In: *Scientific Reports* vol. 12, no. 1 (2022), p. 9863.

[38] Adams, R. I., Bateman, A. C., Bik, H. M., and Meadow, J. F. "Microbiota of the indoor environment: a meta-analysis". In: *Microbiome* vol. 3 (2015), pp. 1–18.

[39] Jacobson, D., Honap, T., Monroe, C., Lund, J., Houk, B., Novotny, A., Robin, C., Marini, E., and Lewis, C. "Ancient human coprolites and dental calculus demonstrate similar functional diversity as modern gut and oral microbiomes". In: *Philosophical Transactions Royal Society B* (2020), p. 13.

[40] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. "The human microbiome project". In: *Nature* vol. 449, no. 7164 (2007), pp. 804–810.

[41] Bissett, A., Fitzgerald, A., Meintjes, T., Mele, P. M., Reith, F., Dennis, P. G., Breed, M. F., Brown, B., Brown, M. V., Brugger, J., et al. "Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database". In: *GigaScience* vol. 5, no. 1 (2016), s13742–016.

[42] Velsko, I. M., Overmyer, K. A., Speller, C., Klaus, L., Collins, M. J., Loe, L., Frantz, L. A., Sankaranarayanan, K., Lewis, C. M., Martinez, J. B. R., et al. "The dental calculus metabolome in modern and historic samples". In: *Metabolomics* vol. 13 (2017), pp. 1–17.

[43] Kim, H.-J., Oh, H. N., Park, T., Kim, H., Lee, H. G., An, S., and Sul, W. J. "Aged related human skin microbiome and mycobiome in Korean women". In: *Scientific Reports* vol. 12, no. 1 (2022), p. 2351.

[44] Cribdon, B., Ware, R., Smith, O., Gaffney, V., and Allaby, R. G. "PIA: more accurate taxonomic assignment of metagenomic data demonstrated on sedaDNA from the North Sea". In: *Frontiers in Ecology and Evolution* vol. 8 (2020), p. 84.

[45] Bloom, B. H. "Space/time trade-offs in hash coding with allowable errors". In: *Communications of the ACM* vol. 13, no. 7 (1970), pp. 422–426.

[46] Sinha, P. and Mizgiris, V. "pybloomfiltermmap3: a fast implementation of Bloom filter for Python". In: *https://github.com/prashnts/pybloomfiltermmap3* ().

[47] Duitama González, Camila and Vicedomini, Riccardo and Chikhi, Rayan and Richard, Hugues. *aKmerBroom ancient Bloom filter*. Jan. 2023. DOI: `10.5281/zenodo.7587160`.

8.9 Perspectives

The decontamination of ancient oral metagenomes presents a computational challenge, which is currently not properly addressed by bioinformatic methods. As a response, we created `aKmerBroom`, a contamination removal tool that works at the read level and was specifically designed for ancient oral datasets.

In the future, it would be worthwhile to explore the implementation of alternative data structures, potentially ones that are more memory-efficient, consider alternatives for assessing cross-contamination, including additional authentication methods (for example adding deamination and fragmentation patterns), as well as taxonomic labelling for the reads that are removed or retained after each run of `aKmerBroom` on a new sample.

Further enhancements include the creation of a more reliable or biologically informed set of trusted k-mers, as well as demonstrating the advantage of removing contamination at the read-level rather than from a taxonomic table (for example, by comparing `aKmerBroom` with decontam).

Chapter 9

Perspectives

9.1 Contamination assessment via Microbial Source Tracking

The construction of the k-mer matrix used for `decOM` included data from the 2020 release from AncientMetagenomeDir [1] (v20.12: Ancient City of Nessebar), however, as of 2023 there are over 2400 host associated metagenomes and over 500 environmental metagenomic samples from everywhere in the world (see bar plots and location). The most updated release of the AncientMetagenomeDir (v23.09: Historic Centre of Cienfuegos) has 4 times as many host-associated metagenomes and 2 times as many environmental samples as the v20.12: Ancient City of Nessebar release. An interesting future direction would be to incorporate the metagenomic samples included subsequent to the release used in this thesis for aOral and potential contaminants.

Moreover, in the forthcoming work for this thesis, a valuable line of research would be to examine the impact on the multi-class classification performance as described in Chapter 6 when varying the hyperparameters for the construction of the k-mer matrix. These hyperparameters include the size of the the k-mers used (value of k), the minimum recurrence and the minimum abundance[1]. Additionally, it would be interesting to evaluate the scalability of `decOM` with a larger number of sinks and sources, and compare it to mSourceTracker [3] and FEAST [4]. The potential sources could come from either a gold standard set of sources established by the community of users, or from the samples uploaded in the latest release from the AncientMetagenomeDir. Furthermore, one could speculate that an increase in the k-mer size might be problematic considering ancient genetic samples are highly fragmented, and a much smaller k-mer size might end up being not so informative. Nonetheless, I expect that tuning hyperparameters such as minimum recurrence and minimum abundance might end up having an overall positive impact of the performance of `decOM`, by reducing the size of the k-mer matrix of sources, reducing running times and memory requirements, and ultimately having a positive or neutral impact on the classification performance.

Another interesting perspective for `decOM`'s potential improvement would be to test different forms of count normalisations or give different weights to the counts, depending on the importance of each source environment to the researcher. This would broaden the scope of the method but might bring additional usability to users.

The algorithm developed in this thesis for contamination assessment via MST, `decOM`, in contrast to competing methods such as FEAST and mSourceTracker, exhibited the ability to differentiate between mOral samples from aOral samples. During the revision of `decOM`'s paper for the Microbiome journal, several supplementary experiments along those lines were conducted, but further investigation into the reason behind this distinction was not pursued. A prospective area of interest would be to elucidate the underlying factors that enable `decOM` to achieve this aOral/mOral distinction. Recent studies [5] describe the problem there is for distinguishing ancient from modern genetic data as an *"authentication error"*, that is, the error which arises from the ancient status of detected organisms and is caused by modern contamination that is commonly present in archaeological samples. Consequently, modern contaminants are erroneously identified as endogenous samples of ancient origin, and vice versa.

Another compelling discovery is that, despite its status as a highly-cited paper, the FEAST algorithm did not achieve convergence when executed on our data. Furthermore, when repeatedly executing FEAST using the sample data provided in their GitHub repository and the same command line parameters, this approach exhibited issues with reproducibility. Specifically, after multiple iterations with identical input data and command line parameters, the results varied significantly from one another (refer to Figure A.7). Bayesian methods, such as the one employed in FEAST, are anticipated to demonstrate a certain degree of variability in the outcomes. However, the disparity observed in the results was notable. We tried to contact the authors and it would be worthwhile to explore the convergence problems of MST methods that employ a Bayesian framework in the future. We hypothesise that the Expectation-Maximization algorithm used by FEAST to decrease running times is only guaranteed to converge to a local maximum, so although faster, FEAST provides different results for different runs.

[1] The last two hyperparameters are command line options of the software kmtricks [2]

9.2 Contamination removal at the read-level

The decontamination of ancient oral metagenomes poses a complex computational challenge, which is inadequately addressed by the algorithmic tools currently available. Our research emphasises the limitations of current methods employed in ancient metagenomic studies for decontamination purposes. In response, we propose `aKmerBroom`, a contamination removal tool that works at the read level and was specifically designed for ancient oral datasets. This tool uses a two-round lookup by constructing a Bloom Filter from a collection of trusted oral k-mers and draws upon a substantial assortment of metagenomes.

In the future, it would be worthwhile to explore the implementation of alternative data structures, possibly more memory-efficient, for `aKmerBroom` that are not an exact membership data structure like the set used in the second round of look-ups. Additionally, it could be beneficial to consider methods to assess cross-contamination, such as the approach used by Recentrifuge [6], which could be incorporated as an additional module to test for contamination between samples. On the other hand, an improvement to `aKmerBroom` could involve the inclusion of additional authentication methods, such as the analysis of deamination and fragmentation patterns, to further validate the results obtained by the algorithm.

More enhancements include the construction of a more reliable or biologically informed set of trusted k-mers, as well as proving the advantage of taking out the contamination at the read-level versus taking it out from a taxonomic table (for instance by comparing `aKmerBroom` with decontam [7]). Another interesting test would be to compare `aKmerBroom` with PMDtools [8], a method that compute postmortem damage patterns and decontaminates ancient genomes at the read-level which was not tested in Chapter 8.

There are significant advantages to not relying on taxonomic classifiers such as Kaiju or KrakenUniq for contamination assessment and contamination removal, particularly in terms of parameter selection and the bias associated with the use of a reference database. However, one interesting addition to `aKmerBroom` and `decOM` would be to try to map the k-mers that appeared as contaminant/non-contaminant to a specific taxa.

References

[1] Yates, J. A. F., Valtueña, A. A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-López, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., et al. "Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir". In: *Scientific Data* vol. 8, no. 1 (2021), pp. 1–8.

[2] Lemane, T., Medvedev, P., Chikhi, R., and Peterlongo, P. "kmtricks: Efficient construction of Bloom filters for large sequencing data collections". In: *bioRxiv* (2021).

[3] McGhee, J. J., Rawson, N., Bailey, B. A., Fernandez-Guerra, A., Sisk-Hackworth, L., and Kelley, S. T. "Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics". In: *PeerJ* vol. 8 (2020), e8783.

[4] Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., and Halperin, E. "FEAST: fast expectation-maximization for microbial source tracking". In: *Nature Methods* vol. 16, no. 7 (2019), pp. 627–632.

[5] Pochon, Z., Bergfeldt, N., Kırdök, E., Vicente, M., Naidoo, T., Valk, T. van der, Altınışık, N. E., Krzewińska, M., Dalen, L., Götherström, A., et al. "aMeta: an accurate and memory-efficient ancient metagenomic profiling workflow". In: *Genome Biology* vol. 24, no. 1 (2023), p. 242.

[6] Martí, J. M. "Recentrifuge: Robust comparative analysis and contamination removal for metagenomics". In: *PLoS Computational Biology* vol. 15, no. 4 (2019), e1006967.

[7] Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. "Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data". In: *Microbiome* vol. 6, no. 1 (2018), pp. 1–14.

[8] Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., and Jakobsson, M. "Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal". In: *Proceedings of the National Academy of Sciences* vol. 111, no. 6 (2014), pp. 2229–2234.

Chapter 10

Conclusions

This written manuscript is the result of a three year doctoral path that began with a discussion between computational scientists and palaeontologists. The core idea of this meeting was to discuss the challenges faced by paleogeneticists when trying to implement Metagenomic Sequencing techniques to the samples they obtained from fossil records. At the time, contamination assessment of their samples, even though partially solved by several Microbial Source Tracking algorithms, had yet to be explored in a reference-free way.

For this purpose, we suggested a reference-free method to perform MST and predict the metadata class of a given (meta)genomic sample using k-mer matrices. We tested our method on a collection of real metagenomic data sets of aOral origin and its possible contaminants and provided an estimation of the contribution of each source environment on each sample. We performed experiments on real and synthetic data and compared `decOM`'s results with state-of-the-art competing methods to prove that the performance of our algorithm, measured through different quantitative metrics, is considerably better. We believe that the incorporation of `decOM` into paleogenomic analyses will allow to identify contaminated metagenomics samples, to ensure their validity, and possibly exclude them from downstream analyses.

Not only is `decOM` an alignment-free method, but it is also adaptable to other types of metagenomic data. As a result of discussions with potential users of the method, we implemented a module of `decOM` that would allow to create a custom matrix of sources (see decOM-MST). This is the first Microbial Source Tracking method to be developed and tested specifically on aDNA data, more specifically, in metagenomics samples isolated from ancient dental calculus. Moreover there is no need for parameter tuning and it is not a probabilistic method, in contrast to competing methods such as FEAST and mSourceTracker that depend on parameter adjustments and are probabilistic methods (see Table C.1 for a thorough qualitative comparison of the three MST methods described in this thesis). Finally, k-mer-based methods such as `decOM` do not rely on taxonomic profiles (which often leave a large proportion of the reads unclassified). `decOM` therefore is not affected by the incompleteness of databases used to compute the taxonomic abundance tables.

The alignment-free contamination removal method proposed in this thesis, `aKmerBroom`, might also be adjustable to the user needs, if they create their own Bloom Filter using a pre-defined set of trusted aOral k-mers, or any reference k-mers. This flexibility is not unique to our method, yet it has the advantage of not requiring negative controls nor a reference database to index as opposed to Recentrifuge and DeconSeq respectively. Interestingly, `aKmerBroom` was developed and tested specifically for aDNA samples coming from ancient oral fossil records, yet it is adaptable to other types of metagenomic data. For a full qualitative comparison of our method and the competing algorithms see Table C.2.

In the future, it would be exciting to see how palaeogeneticists with more specific research questions and a more knowledgeable approach for constructing a k-mer matrix of sources, would effectively use the adaptability of `decOM` to estimate the extent of contamination in their own ancient samples. Furthermore, they can subsequently integrate this k-mer information with `aKmerBroom` to establish a set of trusted k-mers, use them to construct the Bloom Filter, and decontaminate their own data.

Finding solutions to the issue of ancient DNA authentication remains a primary goal for the field of palaeometagenomics in the upcoming years, which is why bioinformatic tools for the contamination assessment and contamination removal of ancient genetic samples is of great significance. The complexity of this problem is compounded by the challenges associated with the study of aDNA, such as molecular degradation, sequence fragmentation, and inevitable exogenous contamination. The lack of a definitive ground truth for determining the extent of contamination in a given sample or the impossibility to taxonomically characterise in a straightforward manner an ancient oral metagenome, further complicate matters. Consequently, there is still a considerable challenge in getting more reliable data labels before implementing supervised machine learning models with palaeometagenomic data. Nonetheless, both of the methods developed during this thesis, namely `decOM` and `aKmerBroom`, have made valuable contributions to the field by using large collections of of aDNA datasets. These methods have the potential to make better generalisations as more samples are used for the construction of the input data (e.g., k-mer matrix, trusted oral k-mers) and researchers continue to make their data publicly available.

Appendices

Appendix A

A.1 Kaiju taxonomic-based clustering table

Reference database NCBI BLAST nr+euk (2021-02-24 release), is a 64GB non-redundant protein database of bacteria, archaea, viruses, fungi, and microbial eukaryotes can be downloaded at `https://kaiju.binf.ku.dk/database/kaiju_db_nr_euk_2021-02-24.tgz`.

To download and unzip the DB:

```
wget https://kaiju.binf.ku.dk/database/kaiju_db_nr_euk_2021-02-24.tgz
tar -xf kaiju_db_nr_euk_2021-02-24.tgz
```

For every sample analysed (depending whether it single-end or paired-end):

```
kaiju -t nodes.dmp -f kaiju_db_nr_euk.fmi -i reads.fastq
[-j reads2.fastq]

kaiju2table -t nodes.dmp -n names.dmp -r genus -o kaiju.table
input1.tsv [input2.tsv ...]
```

Then simply parse the results from the resulting .tsv files in an OTU table format suitable for FEAST or SourceTracker.

The results presented in the manuscript were obtained with the taxonomy abundance profile built with this reference database.

A.2 KrakenUniq taxonomic-based clustering table

The commands used to produce the database using with KrakenUniq were:

```
krakenuniq-download --db DBDIR taxonomy
krakenuniq-download --db DBDIR --dust refseq/bacteria refseq/archaea
krakenuniq-download --db DBDIR
refseq/vertebrate\_mammalian/Chromosome/species\_taxid=9606
krakenuniq-download --db DBDIR refseq/viral/Any viral-neighbors
krakenuniq-download --db DBDIR --dust microbial-nt
krakenuniq-download --db DBDIR contaminants
```

For every sample analysed:

```
krakenuniq --db DBDIR --threads 10
--report-file report_files/{}_unmapped.tax.report.tsv.gz
--output output_files/{}_unmapped.report.tsv.gz --gzip-compressed
--fastq-input {}
```

The symbol {} in this section corresponds to a run accession (sample name)

A.3 Commands used to perform Microbial Source Tracking

The file *metagenome_OTU.txt* corresponds to the species abundance profile (taxonomic-based clustering table) that results from parsing the results from Section A.1 or A.2.

The file *map.txt* is the additional metadata file used as input by SourceTracker and FEAST with three columns (at least): Sample ID with the unique identifier for each sample, SourceSink with the assignment of the sample to either the category source or sink, and finally a column with the name of environment from which each source comes (NA for sink). See documentation of each specific method for their input formats (SourceTracker : `https://github.com/biota/sourcetracker2` , FEAST : `https://github.com/cozygene/FEAST`)

The symbol {} in this section corresponds to a run accession (sample name)

A.3.1 FEAST

The script *Run_FEAST.R* :

```
#!/usr/bin/env Rscript
args = commandArgs(trailingOnly=TRUE)

#Load libraries
library(FEAST)
library(readr)

#Parse arguments
map <- Load_metadata(metadata_path=args[1])
metagenome_OTU <-  Load_CountMatrix(CountMatrix_path=args[2])
accession<-args[3]
dir_path<-args[4]
setwd(dir_path)

FEAST_output <- FEAST(C = metagenome_OTU, metadata = map,
                      different_sources_flag = 0,
                      outfile=accession,
                      dir_path = dir_path)
```

Can be run from the command line in the following manner:

```
Rscript Run_FEAST.R map.txt metagenome_OTU.txt {} ./
```

A.3.2 mSourceTracker

```
sourcetracker2 -m map.txt -i metagenome_OTU.txt -o output_{}
```

A.3.3 `decOM`

For every sample analysed run the following command:

```
decOM -s {} -p_sources decOM_sources/ -k {}.fof -mem 25GB -t 10
```

See details on the format of the key.fof file in `https://github.com/CamilaDuitama/decOM` depending on whether the sample comes from a single-end or paired-end experiment.

A.4 Software versions and run accession codes of samples used

Versions of the software used:

- Kaiju 1.7.3
- KrakenUniq 0.5.8
- `decOM` 1.0.0
- FEAST 0.1.0
- mSourceTracker 2.0.1-dev

Accession codes are available in `https://github.com/CamilaDuitama/decOM/tree/master/data`, specifically:

- **Collection_accessions.csv:** Accessions for the collection of 360 metagenomic samples that compose the sources of decOM.
- **ValidationSet.csv:** Accessions for the collection of 254 ancient oral metagenomic samples used sinks in validation experiment with an external data set.

A.5 Definition of performance metrics

Consider TP as number of true positives, FN as number of false negatives, $\hat{y}_i$ the predicted value of the i-th sample and y_i the corresponding true value. If we have n samples:

$$\text{Accuracy}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i) \tag{A.1}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{A.2}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{A.3}$$

$$\text{F1-score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{A.4}$$

A.6 Figures

Figure A.1: **Metadata barcharts for collection of 360 metagenomic data sets (sources)**. Additional metadata associated to all of the samples in the collection of data sets used to train the method (this includes aOral, mOral, sediment/soil and Skin samples). Barcharts are presented for metadata features such as Assay Type, BioProject, Platform, Instrument, Library Layout and Center Name.

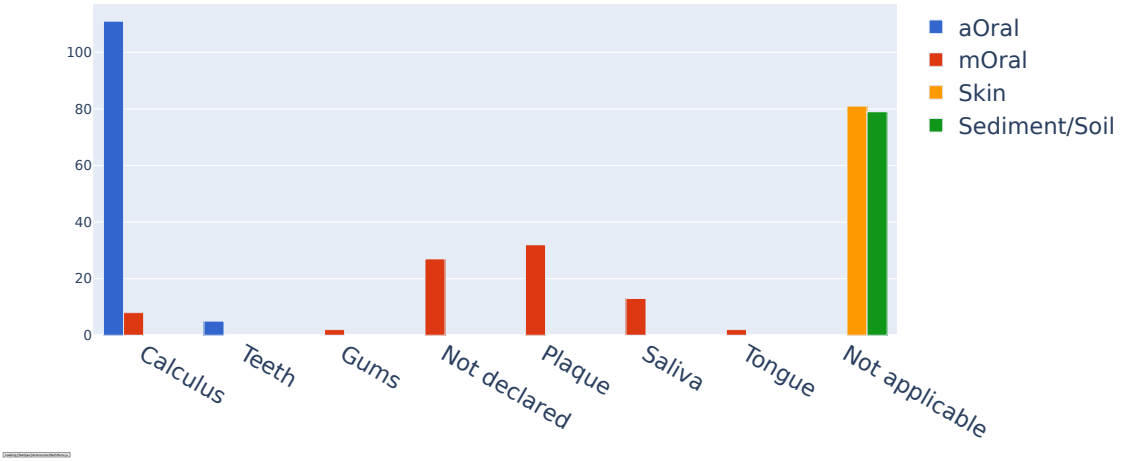


Figure A.2: **Isolation source for samples in the collection of 360 metagenomic data sets (sources)**. Barchart colouring corresponds to the original four source environments studied (aOral, mOral, Skin, sediment/soil).

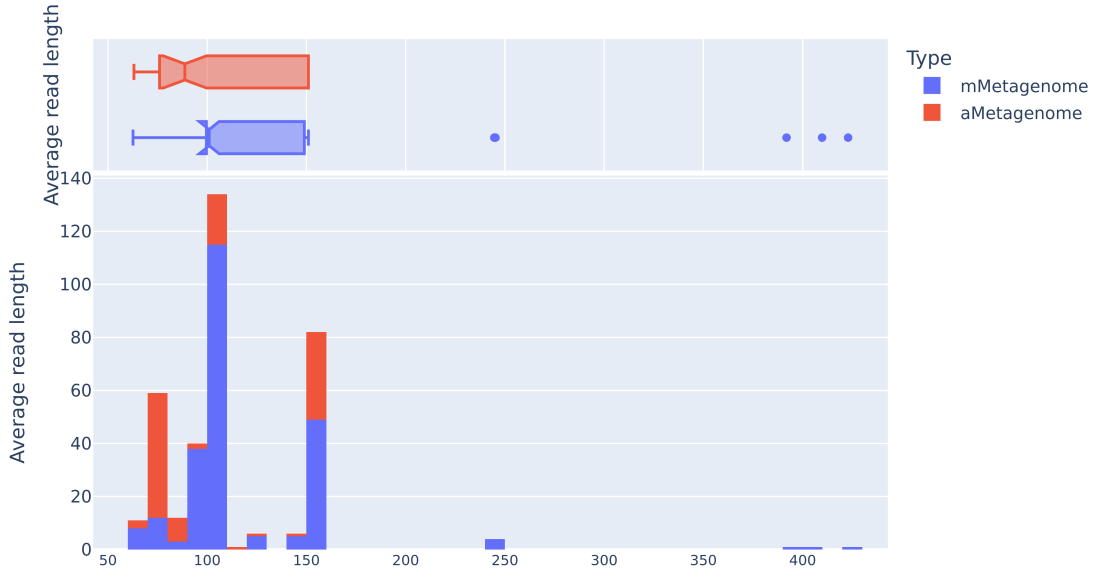


Figure A.3: **Average read length for collection of 360 metagenomic data sets (sources)**. Samples are grouped by ancient metagenomes (aOral) and modern metagenomes. The thinner part of the boxplot on top indicates the median of the distribution.

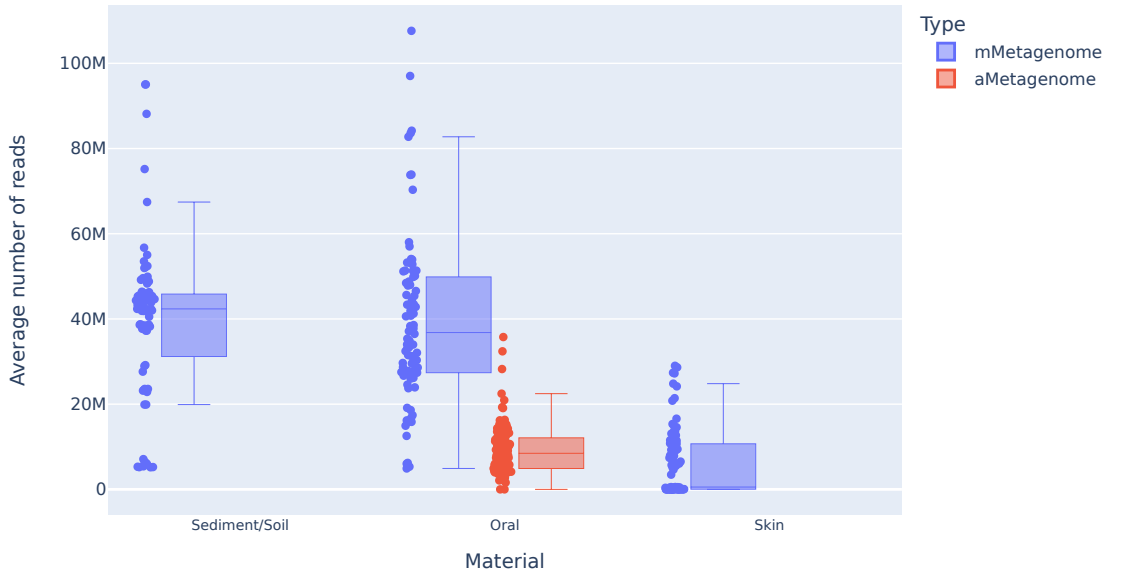


Figure A.4: **Average number of reads for collection of 360 metagenomic data sets (sources)**. Samples are grouped by ancient (aOral) and modern metagenomes (Skin, sediment/soil, mOral)

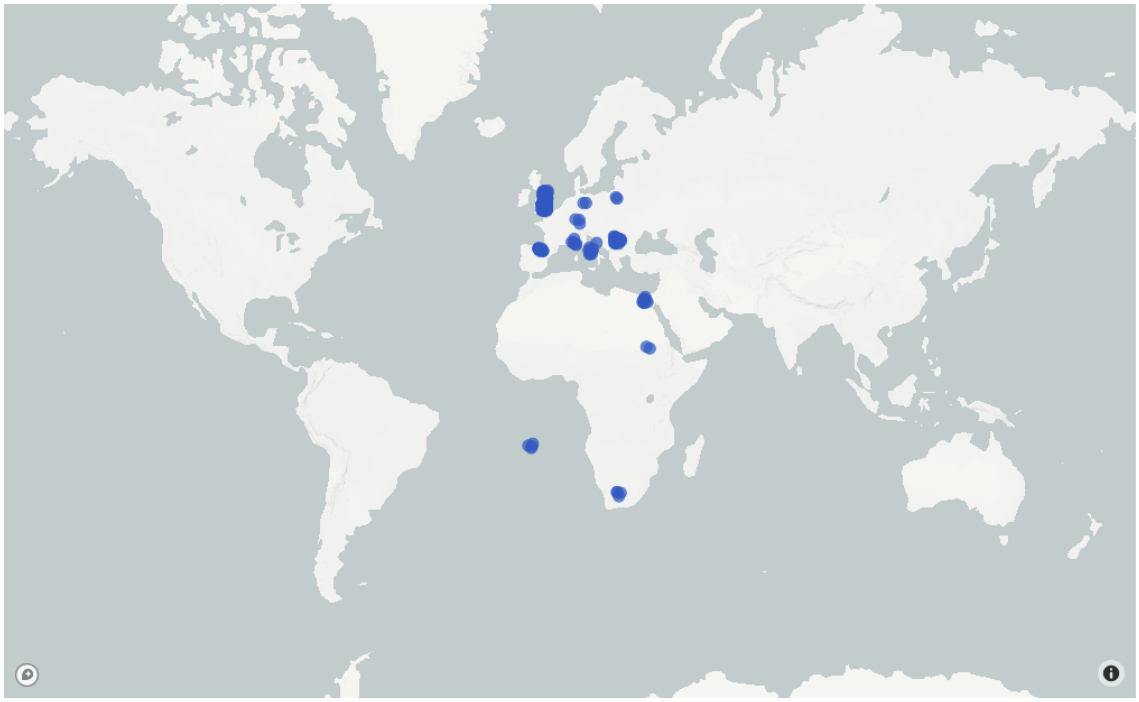


Figure A.5: **Origin of samples in validation data set.** All samples in validation set were ancient oral samples obtained from the AncientMetagenomeDir

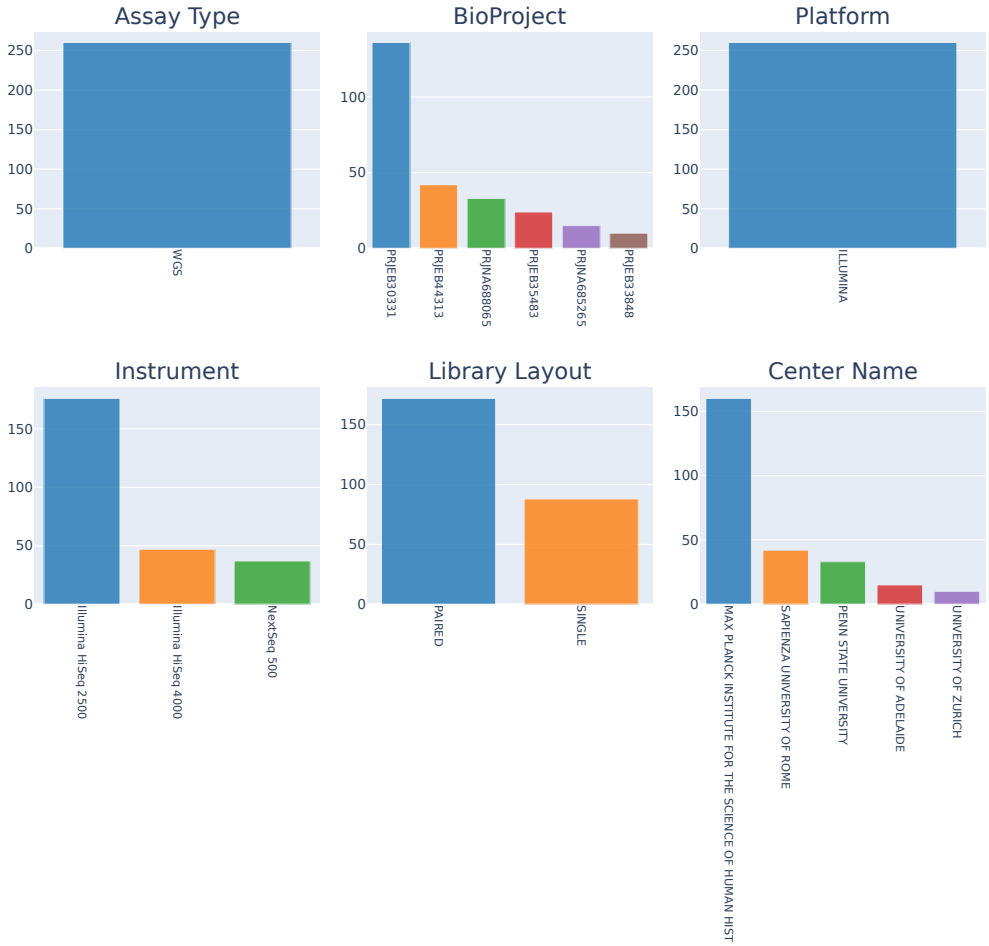


Figure A.6: **Metadata barcharts of validation data set**. All samples in validation set are labelled as ancient oral. Barcharts are presented for metadata features such as Assay Type, BioProject, Platform, Instrument, Library Layout and Center Name.

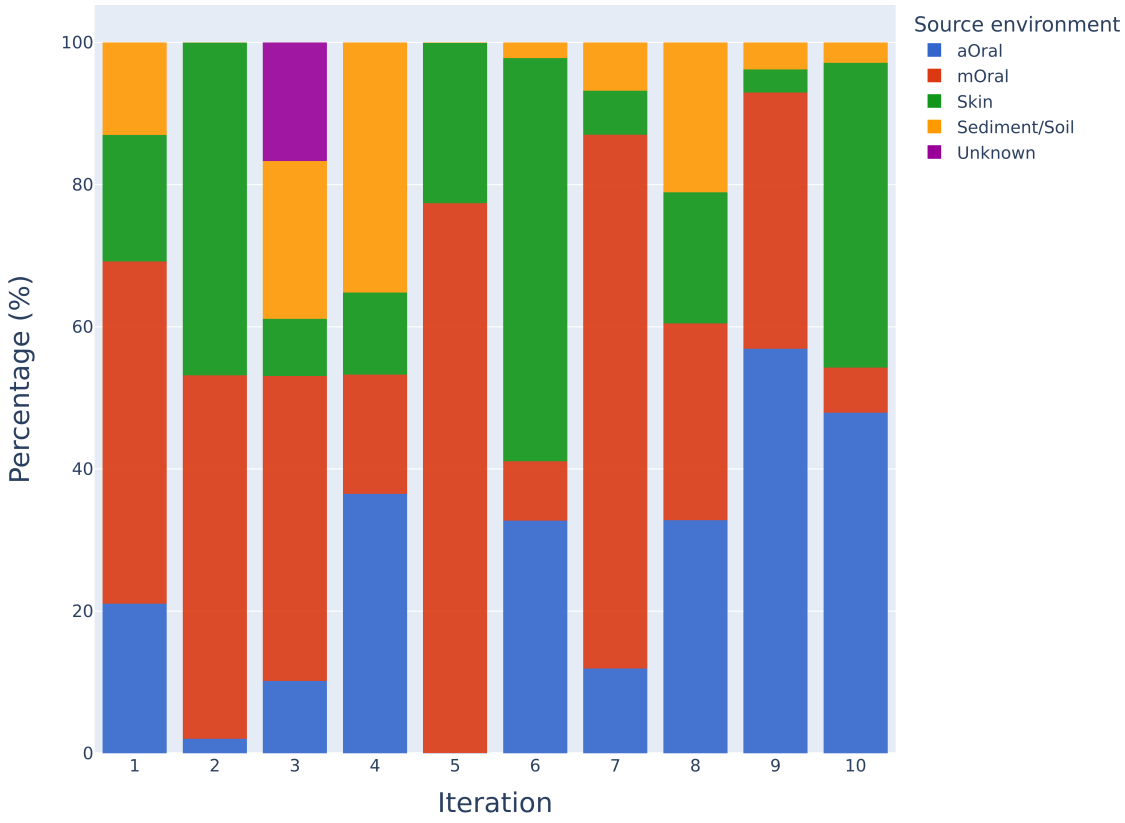


Figure A.7: **Microbial source tracking results by using FEAST on the simulated ancient oral data set**. We ran and plotted the source environment bar plots of every output after using FEAST under the same parameters (10 iterations in total) by using as sources all the samples in the 360 metagenomic collection and as sink a simulated ancient calculus data set. As results for this method vary, we selected the first iteration of the outputs to be part of the main text of the paper.

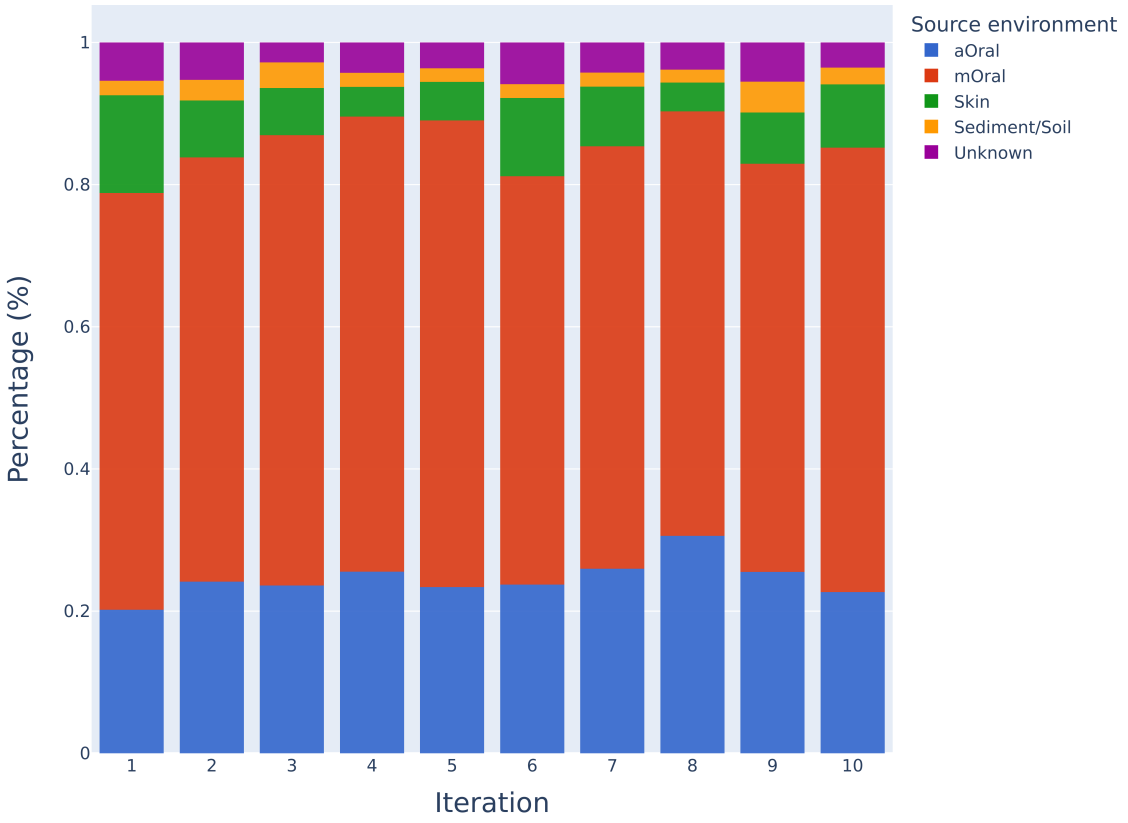


Figure A.8: **Microbial source tracking results by using mSourceTracker on the simulated ancient oral data set**. We ran and plotted the source environment bar plots of every output after using mSourceTracker under the same parameters (10 iterations in total) by using as sources all the samples in the 360 metagenomic collection and as sink a simulated ancient calculus data set. See in comparison with Figure A.7

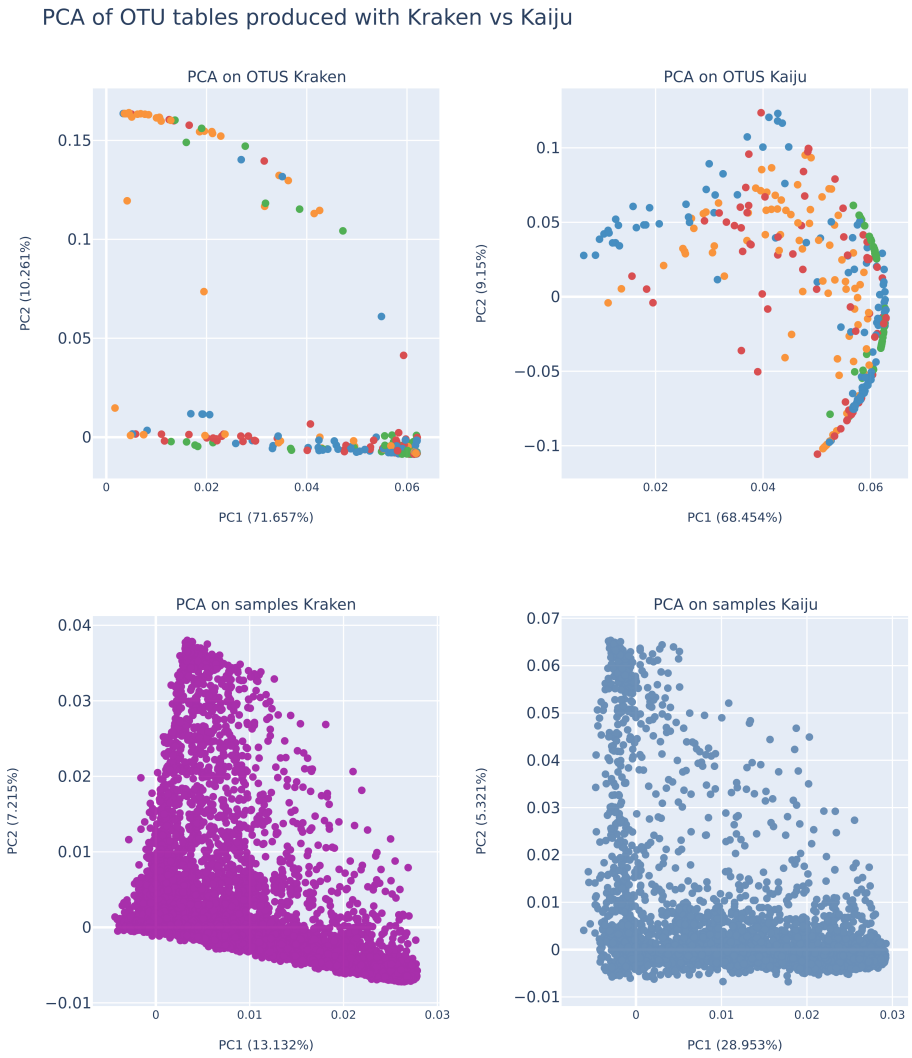


Figure A.9: **PCA of OTU table built with Kaiju vs KrakenUniq**. Colour-coding for upper plots: yellow for sediment/soil, green for Skin, blue for aOral and red for mOral samples. Notice that no clear clusters between source environments are seen for either of the OTU tables.

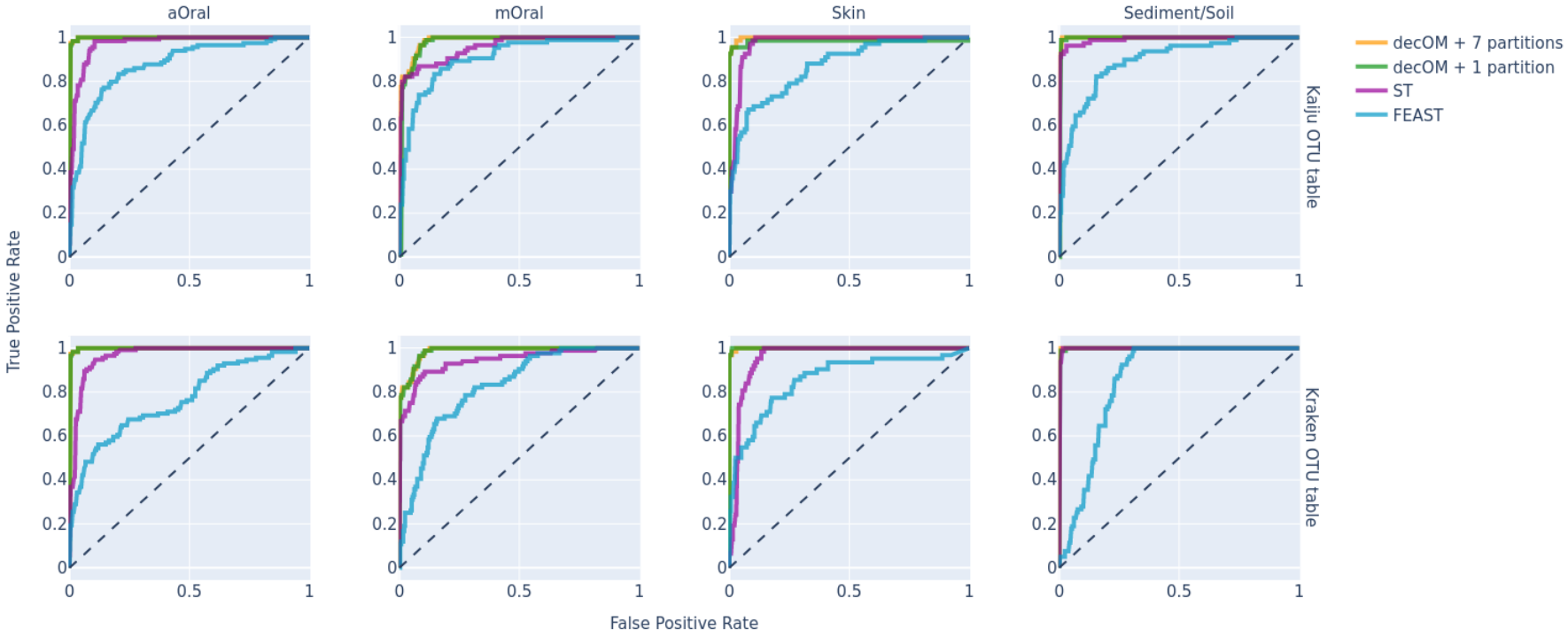


Figure A.10: **ROC per class per method.** Each subplot was built using the true and predicted labels for each class, and it includes the curves for every method evaluated. The ROC curve for `decOM` + 1 partition (yellow) overlaps with the curve for `decOM` + 7 partitions (green), this is, the performance by using less k-mers is almost not affected.

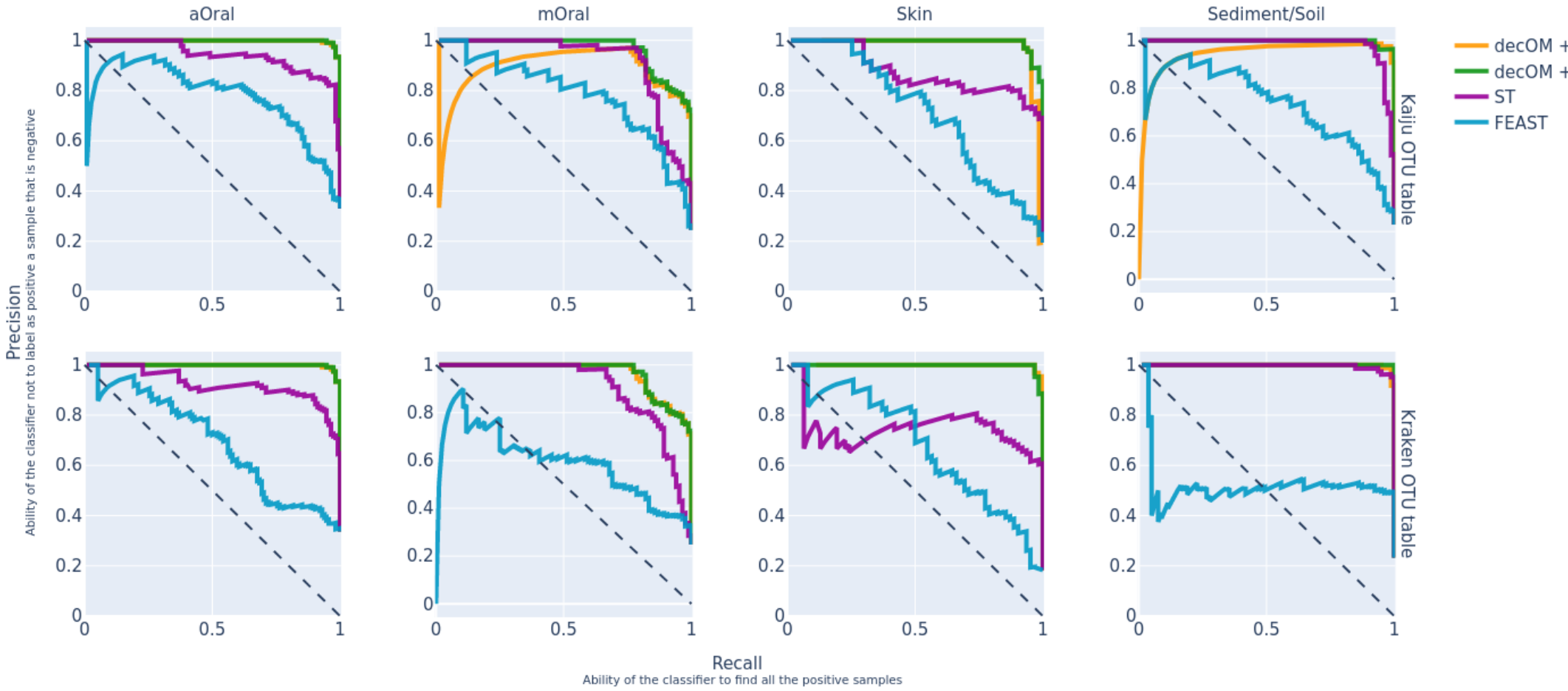


Figure A.11: **Precision-Recall curves per class per method.** Each subplot was built using the true and predicted labels for each class, and each subplot includes the curves for every method evaluated. The PR curve for `decOM` + 1 partition (yellow) overlaps with the curve for `decOM` + 7 partitions (green), this is, the performance by using less k-mers is almost not affected.

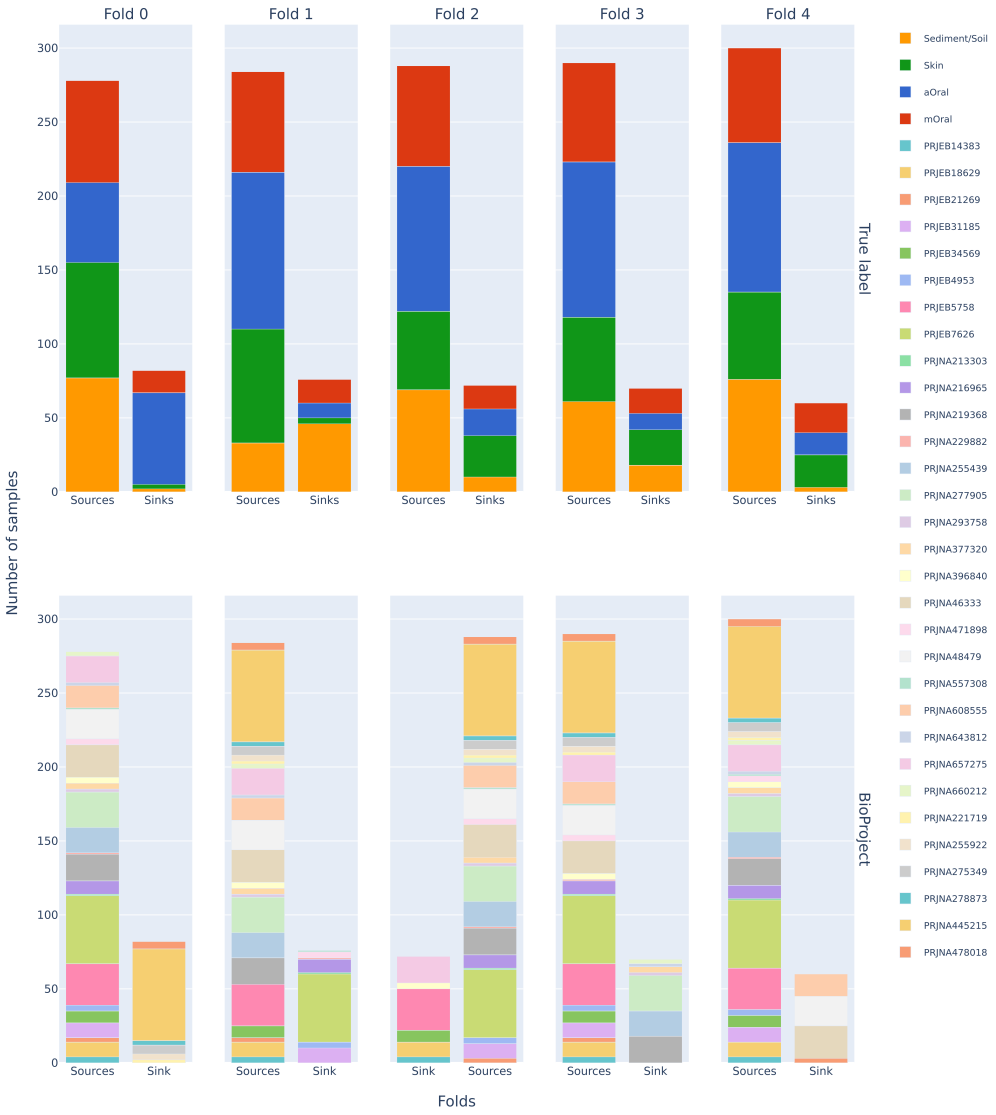


Figure A.12: **5-fold cross-validation data split.** Graphic representation of 5-fold cross-validation data split where all samples belonging to the same BioProject are either part of the training or the test set.

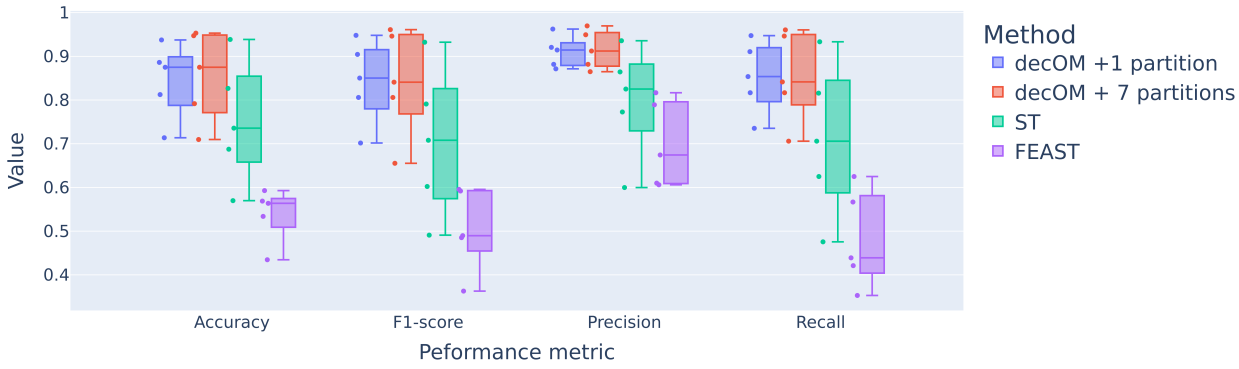


Figure A.13: **Performance in 5-fold cross-validation experiment including `decOM` + 7 partitions.** Box plots for the performance metrics such as Accuracy, F1-Score, Precision and Recall obtained after the the 5-fold cross-validation experiment using `decOM` + 1 partition, `decOM` + 7 partitions, mSourceTracker and FEAST.

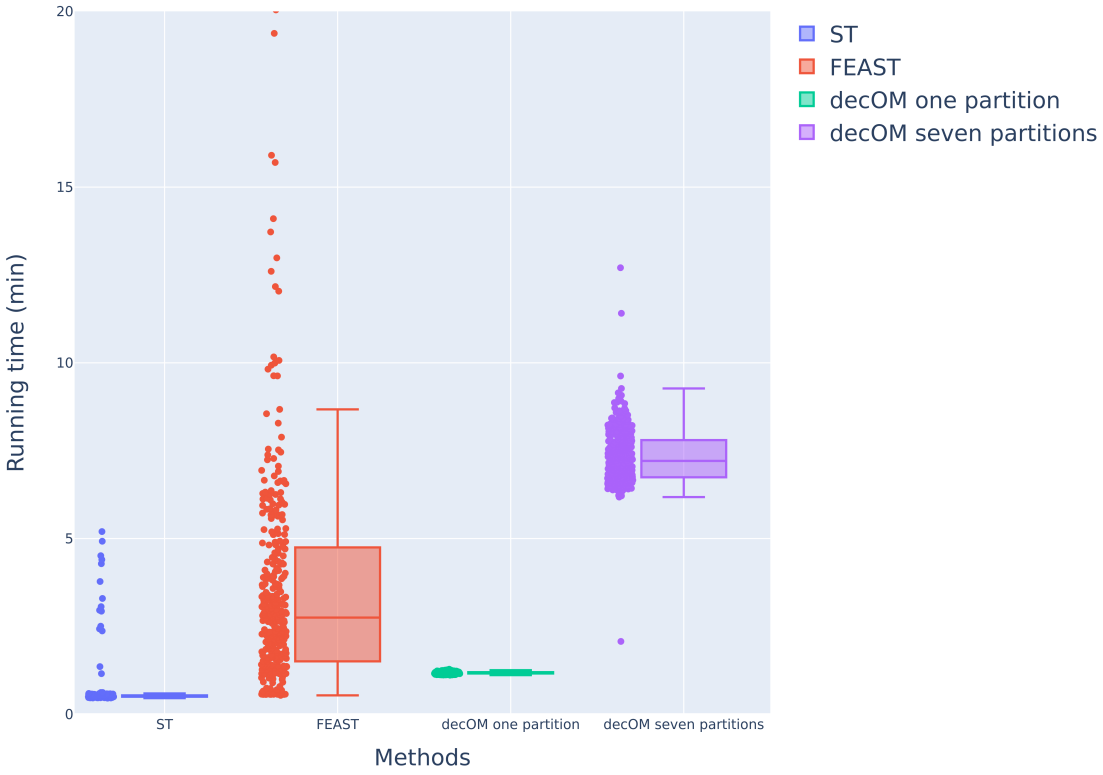


Figure A.14: **Running times in leave-one-out experiment.** Box plots built with the running times for each method when analysing each of the samples from the collection (points to the left of the box plot correspond to each measurement). As seen, using one partition is faster on average than using seven partitions.

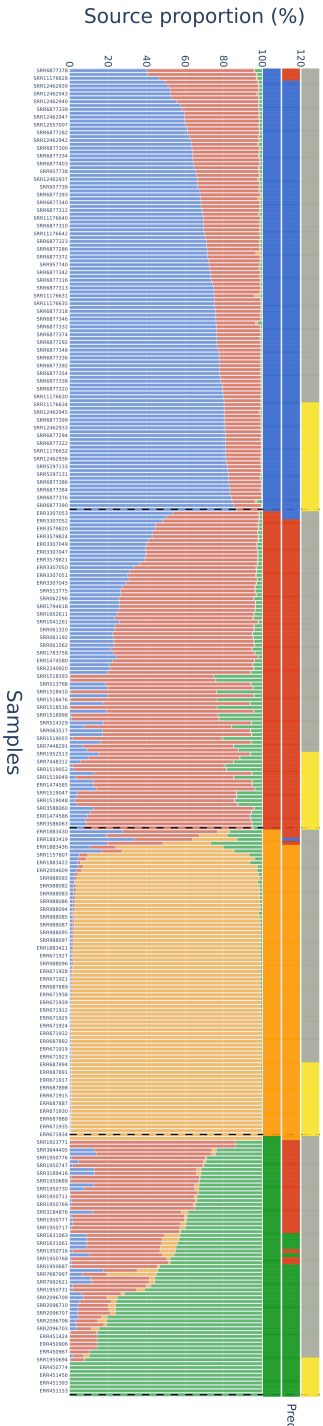


Figure A.15: **Estimations of the source contribution to every samples of the 360 metagenomes collection by `decOM`.** Source estimates considered Skin, Soil, aOral and mOral as possible source environments. The three annotations above the bar chart for each sample from top to bottom correspond to: The legend is light yellow when the sample is considered to be composed of mostly one source for the label assigned (highest 25% of the class, further categorised as mono-source) or grey if it is contaminated and does not come from mostly one source environment (lowest 75% of the class, further categorised as multi-source). The middle annotation corresponds to the predicted label by `decOM`. Finally, the annotation on the bottom corresponds to the true label for each sample. Samples were first sorted according to their true label, and inside each class they were further sorted with respect to the source proportion estimation for each class label.

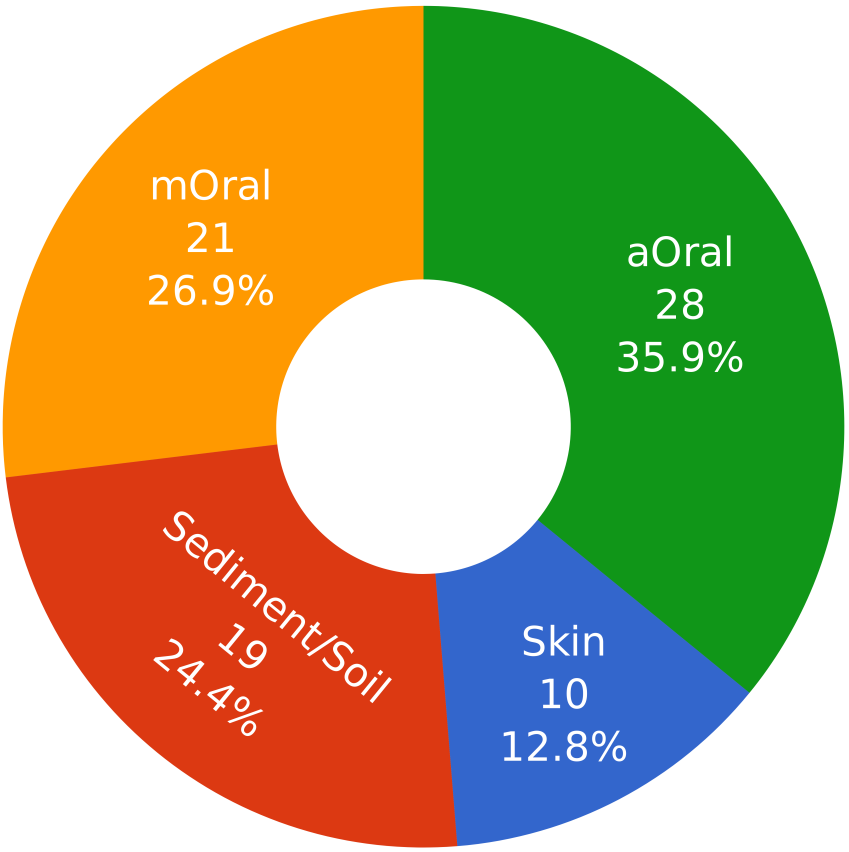


Figure A.16: **Class composition of monosource samples as predicted by `decOM`.** Samples from the collection that we further categorised as monosource samples belong to the classes aOral, mOral, Skin and sediment/soil.

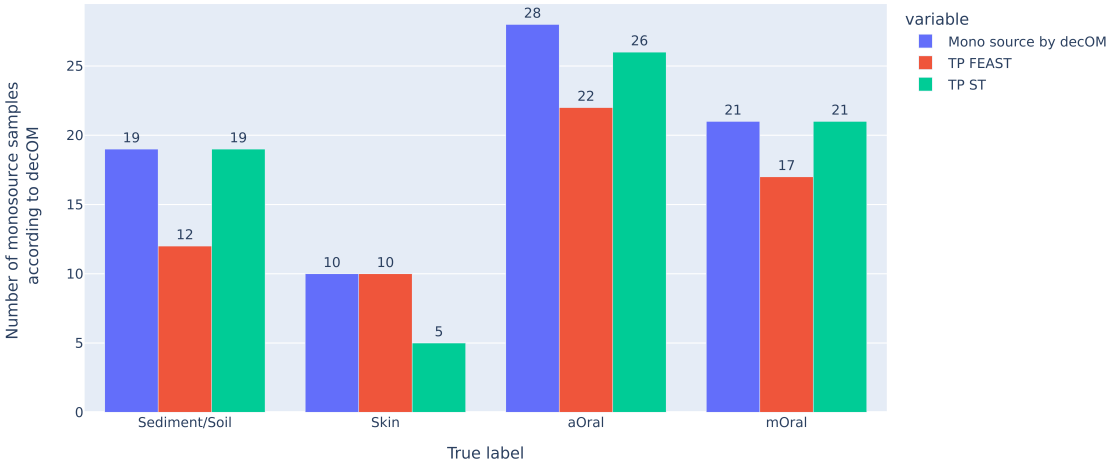


Figure A.17: **Percentage of monosource samples according to `decOM`.** After doing a categorisation of `decOM` predictions we find some samples in the collection to be composed of mostly one source environment. We distinguish them as monosource. Interestingly, from the monosource samples 61/78(78%) are also correctly predicted by FEAST, whereas 71/78(91%) are correctly predicted by mSourceTracker. TP = True positive

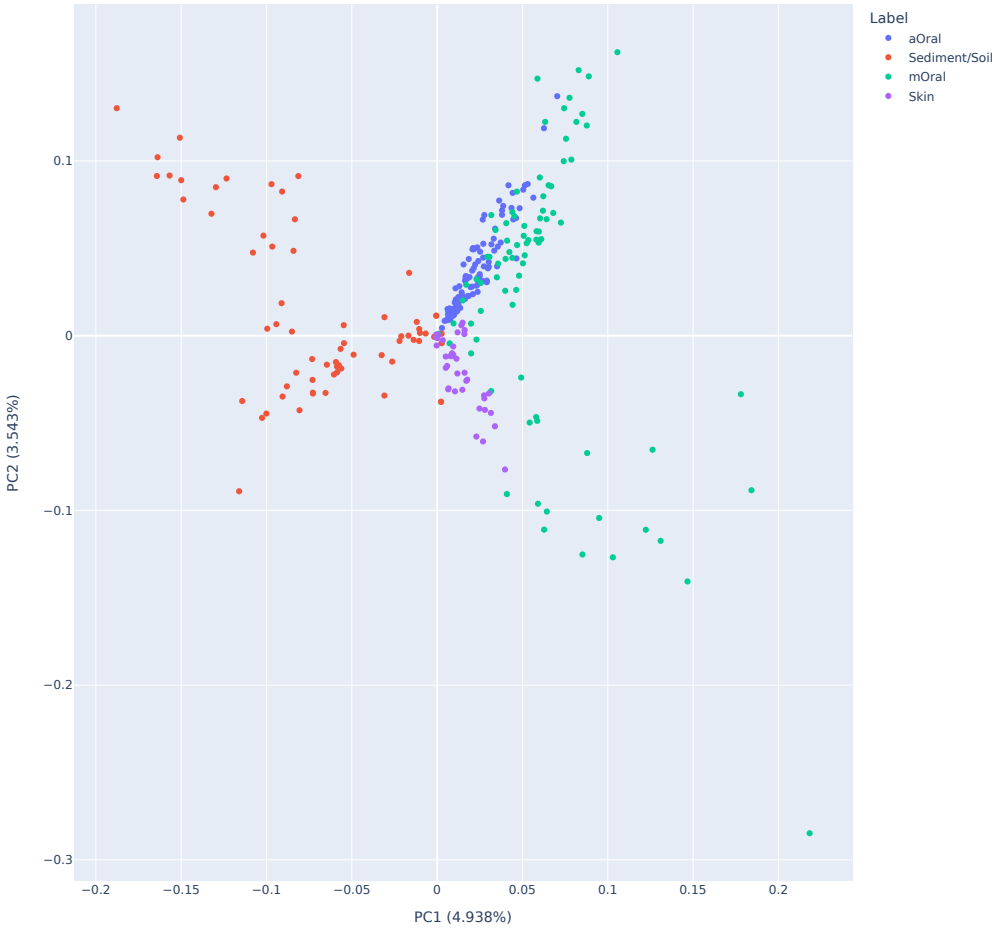
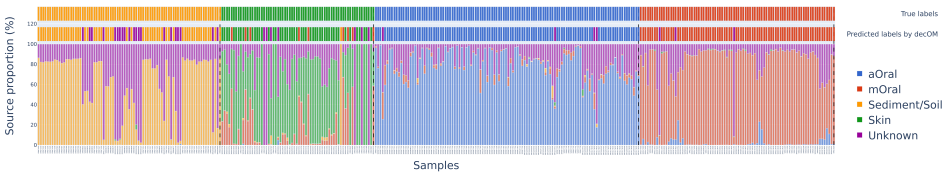


Figure A.18: **PCA of k-mer matrix of sources** We performed a Principal Component Analysis (PCA) on the input k-mer matrix of sources by reducing the number of rows (k-mers) and drawing the 360 samples in a scatterplot built with the two first principal components. Although the variance explained by each of the components is not very large, aOral samples often overlap with mOral samples. Furthermore, some of the mOral samples appear closer to the Skin samples which might explain why some of them seem contaminated and have a higher proportion of Skin contribution, as seen in Figure 5 from the main text.



(a) `decOM`

Figure A.19: **Bar plots of the source environment contribution on each sink after the leave-one-out experiment as estimated by `decOM`** Samples are first sorted by true label. The correction implemented corresponds to counting only the k-mers that are unique to each class, that is, to leave out the k-mers that belong to one or more classes. Notice that for this version of `decOM`, unknowns are estimated as number of k-mers present in sink and absent in all of the sources.

A.7 Tables

Table A.1: **Performance metrics for all three methods compared after leave-one-out experiment.** Performance scores were estimated as the average score across all classes. The result in bold is the one shown in the paper for `decOM`

Method	Accuracy	Precision	Recall	F1-score
`decOM` + 1 partition	**0.8703**	**0.9184**	**0.8703**	**0.8753**
`decOM` + 7 partitions	0.8791	0.9216	0.8791	0.8809
FEAST	0.6816	0.5516	0.7452	0.5479
metaSourceTracker	0.8388	0.8388	0.8388	0.8289

Table A.2: **Performance metrics for all three methods compared after stratified 5-fold cross-validation experiment.** Performance scores were estimated as the average score across all classes. The result in bold is the one shown in the paper for `decOM`

Method	Accuracy	Precision	Recall	F1-score
`decOM` + 1 partition	**0.8450**	**0.9101**	**0.8528**	**0.8421**
`decOM` + 7 partitions	0.8553	0.9156	0.8542	0.8419
FEAST	0.5387	0.6992	0.4809	0.5050
metaSourceTracker	0.7516	0.7996	0.7111	0.7049

Table A.3: **Cardinality of different sets for the k-mer matrix of sources used by `decOM`** Percentages are estimated using the total number of k-mers in the k-mer matrix of sources that is public as a Zenodo file (aprox 14M). If a k-mer is present in at least one sample labelled as class A, then this k-mer is considered to be a member of the set A. If a k-mer is present in at least one sample labelled as class A and is present in at least one sample labelled as class B, then such k-mer is a member of the set called *A AND B*.

k-mer set	**% of k-mers in set**
Sediment/Soil	52.817
Skin	20.044
aOral	16.389
mOral	40.292
Sediment/Soil AND aOral	0.338
Sediment/Soil AND mOral	0.147
Sediment/Soil AND Skin	0.560
aOral AND mOral	11.444
aOral AND Skin	1.314
mOral AND Skin	17.093
aOral AND mOral AND Skin	1.191
aOral AND mOral AND Sediment/Soil	0.072
mOral AND Skin AND Sediment/Soil	0.057
aOral AND Skin AND Sediment/Soil	0.057
aOral AND mOral AND Skin AND Sediment/Soil	0.022

Appendix B

Acronyms

aDNA ancient DNA.

aOral ancient oral.

AUC Area Under the ROC Curve.

BF Bloom Filter.

BLAST Basic Local Alignment and Search Tool.

BWA-SW Burrows-Wheeler Aligner's Smith-Waterman Alignment.

BWT Burrows–Wheeler Transform.

DAWG Direct Acyclic Word Graph.

DSB Double-Strand Break.

EM Expectation-Maximization.

FEAST Fast Expectation-Maximization Microbial Source Tracking.

FPR False Positive Rate.

HEPA High-Efficiency Particulate Arrestance.

HGP Human Genome Project.

HOPS Heuristic Operations for Pathogen Screening.

IUPAC International Union of Pure and Applied Chemistry.

LCA Lowest Common Ancestor.

LDA Latent Dirichlet Allocation.

MALT MEGAN ALignment Tool.

MCMC Markov Chain Monte Carlo.

MDS Multidimensional Scaling.

MEM Maximal Exact Matching.

MGS Metagenomic Sequencing.

ML Maximum Likelihood.

mOral modern oral.

mSourceTracker metagenomic-SourceTracker.

MST Microbial Source Tracking.

mtDNA Mitochondrial DNA.

NGS Next Generation Sequencing.

OTU Operational Taxonomic Unit.

PCA Principal Component Analysis.

PCR Polymerase Chain Reaction.

ROC Reciever Operating Characteristic.

RTL Root-To-Leaf.

SPAAM Standards, Precautions, and Advances in Ancient Metagenomics.

SSB Single-Strand Break.

TPR True Positive Rate.

Appendix C

decOM	FEAST	mSourceTracker
Alignment-free	Needs ref DB	Needs ref DB
Input matrix of sources provided for aOral assessment	Needs input OTU table	Needs input OTU table
Adaptable to other types of metagenomic data		
Evaluated specifically in aDNA	Not so widely used for aDNA	Tested widely on aDNA
No parameter tuning needed	Optimal performance after parameter tuning	Optimal performance after parameter tuning
Method is deterministic	Method is probabilistic	Method is probabilistic

Table C.1: Comparison of contamination assessment methods via Microbial Source Tracking (MST)

DeconSeq	Recentrifuge	**aKmerBroom**
	Needs ref DB	**Alignment-free**
User might create their own index	User needs to provide (-) controls	**User might create their own Bloom Filter**
Parameter optimisation not benchmarked for aDNA. Ex: BWA-SW options and remove/retain DB (DeconSeq), Centrifuge parameters (Recentrifuge)		**No parameter optimisation needed**
Adaptable for other types of metagenomic data		
Not developed specifically for aDNA	Not developed specifically for aDNA	**Evaluated specifically in aDNA**

Table C.2: Comparison of contamination removal methods