

Ancient DNA perspectives on pathogen evolution in domestic animals

Evangelos Antonios Dimopoulos



Thesis submitted for the degree of a Doctor of Philosophy

University of Oxford

Trinity Term 2021

Declaration

I, Evangelos Antonios Dimopoulos, declare that this thesis has never been submitted, either in the same or a different form, to this or any other university, for a degree or any other qualification. All the work presented in this thesis is the result of my own work unless otherwise stated in the author contributions. The work was carried out while I was registered as a graduate student at the Research Laboratory of Archaeology and the History of Art, under the supervision of Prof. Greger Larson.

Abstract

Pathogens, animals (domesticated or not) and humans have intertwined evolutionary trajectories for millennia now, with this relationship intensifying after the Neolithic Demographic Transition and its subsequent epidemiological transition that increased the exposure of animal populations to pathogenic reservoirs. Incorporating the temporal dimension in the study of disease evolution requires the incorporation of ancient pathogen DNA data in phylogenetic and population genetic analyses, which is not a simple task. Limited DNA survival and DNA damage, however, have hindered the search for ancient pathogens. For these reasons the accurate identification of microbial species from archaeological samples is integral and has the potential to provide an unprecedented perspective on the evolutionary history of major diseases that have affected humans and domestic animals. In the first chapter of this thesis I propose a novel analytical method for metagenomic identifications, HAYSTAC, which can also be applied for direct hypothesis testing regarding the presence of pathogens in metagenomic samples. By employing a clear probabilistic model for species identification from shotgun sequencing DNA data, along with incorporating additional algorithmic steps for taxonomic result validation, robust metagenomic identifications from sequencing data generated from archaeological samples become possible, with a significantly lower false positive rate when compared to pre-existing and established methods. In the second chapter of this thesis I used HAYSTAC to screen more than one thousand archaeological and modern chicken samples for Marek's Disease Virus (MDV), a devastatingly virulent virus threatening the modern poultry industry. Following the DNA sequencing of archaeological chicken samples positive for MDV, I reconstructed the genomes of these ancient strains and compared them to modern ones. Through a time dependent phylogenetic reconstruction approach, I identified the

possibility that modern MDV strains are evolving at an accelerated rate compared to ancient ones, with modern highly virulent strains possibly evolving from a potentially non-virulent (ancient) form (prior to 1907) of MDV, a finding that can have a significant impact on improving the health status of domestic chicken population globally. In the third chapter of this thesis, I examine if the process of domestication has had an impact on the disease burden of domestic pigs and dogs, by applying HAYSTAC. By applying hypothesis-driven species identification methods, I focused on screening animal derived metagenomes for zoonotic bacterial pathogens. Since many of the causative agents of these zoonotic diseases became prevalent in a post-industrial context, the results of this approach will be able to offer insights into how human activity has reshaped the landscapes of both the natural environment and animal diseases.

Table of Contents

Declaration	i
Abstract	ii
Table of Contents	iv
Table of Figures	xi
Table of Tables	xii
Acknowledgments	xiii
1. Introduction	1
The motivation behind combining archaeology, ancient DNA and metagenomics	1
Archaeological research, ancient DNA and metagenomics	2
What is Metagenomics and how they are employed in ancient DNA research	4
The challenges of identifying microbial taxa from their DNA sequences	6
Functionality framework of metataxonomic tools	8
What can you do after you identify a species	11
Looking for pathogens in unlikely places	13
Characteristics of ancient DNA for validating metagenomic results	16
Data science and bioinformatics	18
Recapitulation	20
Bibliography	22
2. HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data	33
Statement of Authorship:	33
Authors:	34
Abstract	35
Introduction	37
Results	42
Method Overview	42
Computational Performance	45
Analyses of simulated data sets using RefSeq prokaryotic database	48
Analyses of simulated data sets using genus-level database	57
Case study 1: Ancient Yersinia pestis	61
Case study 2: Historic dental calculus	63

Discussion	67
Species concept, horizontal gene transfer and species identification	67
Issues with databases	69
Conclusions	72
Methods	73
Single Source Identification	73
Likelihood	74
Posterior	75
Multiple Sources Identification	78
Dirichlet prior	80
Posterior mean and Confidence interval	81
Computational architecture	82
A tool for constructing databases	82
Processing, and aligning reads	83
Computing abundance	84
Assessing evenness of coverage	85
Representative species RefSeq database	85
Simulations	86
Analyses of simulated data	88
Case study 1: Ancient Yersinia pestis	89
Case study 2: Historic dental calculus	90
Performance test	90
Acknowledgments	92
Bibliography	93
Supplementary	103
Calculation of false positive, false negative and true positive rates	103
Supplemental Figures	104
Supplemental Tables	109
Permission from co-authors	138
 3. Tracking the evolutionary trajectory of Marek's Disease Virus from the 16th century till the present	 140
Statement of Authorship:	140
Authors:	141
Abstract	142
Introduction	143

Ancient DNA and viruses	143
Alphaherpesviruses and their characteristics	144
Marek's Disease Virus and the vaccines against it	146
Vaccination acting as an evolutionary force for MDV	147
MDV infection overview	148
Motivation	151
Methods	152
DNA extraction	152
Library Preparation & Sequencing	152
HAYSTAC	153
Capture	153
DNA sequencing data alignment	154
SNP calling	155
Ancient genome reconstruction	156
Publicly available (modern) data	156
Data download	156
Sequence preparation for RAxML and BEAST	157
Removing BAC sequences	157
Masking of the modern genomes	157
Multisequence alignment (MSA) with mafft for building phylogenetic trees	158
Extracting and preparing the aligned Open Reading Frames (ORFs) and intergenic regions for BEAST	158
ORFs and intergenic regions with more than 10% overlap	159
Phylogenetic trees	160
BEAST starting tree	160
Maximum Likelihood tree	161
Maximum Likelihood tree based on transversions	161
BEAST	162
Population structure summary statistics	163
Results	165
Metagenomic identifications and MDV genome capture	165
SNP analysis	167
Phylogenetic analysis with Maximum Likelihood trees	169
BEAST analysis	172
Population structure analysis	174
Discussion	176
MDV biology and its implications in infection time course studies	176
Metagenomic identifications and MDV capture	177
SNP analysis	180
Phylogenetic analysis	182

Population structure	189
Human activity shaping the chicken-MDV interaction landscape	190
Conclusions	192
Bibliography	194
Supplementary	210
Permission from co-authors	247
4. Mining for pathogens in ancient pig and dog metagenomes	248
Statement of Authorship:	248
Authors:	248
Abstract	249
Introduction	251
Zoonoses and animal health in domesticated animals	251
What is a domesticated animal and how ancient DNA can elucidate (zoonotic) pathogen transmission	253
Integrating ancient DNA and archaeological evidence can elucidate the dynamics among pathogens, domesticates and humans	254
Examples of zoonotic epidemics affecting animals and humans	256
Motivation for studying (ancient) animal pathogens	259
Swine (zoonotic) pathogens	262
Canine (zoonotic) pathogens	267
Questions	273
Methods	274
DNA extraction	274
Library Preparation & Sequencing	275
Sample pre-processing	275
Pathogen detection	276
Pathogens of Interest	276
Database building	277
Metagenomic identifications	277
aDNA authentication	277
Results	278
Pigs	278
Detected Pathogens	278
Pathogen distribution and species found	287
Domestication burden	289

Dogs	291
Detected Pathogens	291
Pathogen distribution and species found	301
Domestication burden	303
Discussion	305
Factors confounding ancient metagenomic analyses	305
Elaborating on the metagenomic identification results	309
Putting metagenomic results into perspective	311
Bibliography	314
Supplementary	340
Permission from co-authors	374
 5. Conclusions	 375
What was the overall motivation behind studying ancient pathogens and metagenomes ?	375
Laboratory techniques that have advanced the study of ancient pathogens	375
Pathogen discovery in ancient metagenomes	377
How to safely untangle the mysteries of aDNA metagenomics	382
Bioinformatics in metagenomics and data science	386
Challenges for bioinformatic pipelines as genomic data become ‘bigger’	388
What comes after metagenomics and how to use the information they provide us with.	392
Concluding remarks	394
Bibliography	396
 6. Species identification by using bioinformatic methods (Minor Chapter)	 403
Introduction	403
Archaeology and ancient DNA	403
Projects of species identification	403
Introduction to the method	405
Materials and methods	407
Implementation	407
The downgrading system	410
Interpretation of the results	411
Application	411
Results	413

Synthetic dataset: Mice	413
Real data: Chickens	423
Discussion	426
The downgrading system	426
Assessment of the method	427
Synthetic dataset: Mice	427
Real dataset: Chickens	428
Conclusions	430
Bibliography	431
 7. African Bovids (Minor Chapter)	 432
Introduction	432
The problem of species identification	434
Archaeology and ancient DNA	434
Introduction to the method	440
Materials and methods	441
Species Identification Pipeline Method	441
Implementation	442
Notes on the construction of the local database	447
Species identification in metagenomic analysis	448
Interpretation of the results	449
Applications	450
Case Study - The Kalahari Desert debate	452
Motivation and previous studies	452
Results	453
Discussion	457
The limits of aDNA	457
Conclusions	460
Bibliography	461
Supplementary	465
 8. South American Canids (Minor Chapter)	 501
Introduction	501
When and what species of canids appeared in South America	501
Domesticated dogs in South America	503
Hypothesis testing	503
Materials and methods	505
Species Identification Pipeline Method	505
Implementation	506
Notes on the construction of the local database.	511

Interpretation of the results	512
Applications	513
Results	515
Discussion	518
Bibliography	521

9. Demographic Analysis of sheep populations from Asikli, Turkey, across time (Minor Chapter)	523
Asikli population	523
ABC analysis	528
Best Fitting model – Constant NE	535
European Sheep	539
ABC analysis	548
Best Fitting model – Expansion	557
Asikli site and ancient Turkish sheep	571
ABC analysis	577
Best Fitting model – Split	583
Asikli – Assess Bottleneck strength	590
Bibliography	593

Table of Figures

Figure 2.1 HAYSTAC's workflow.	44
Figure 2.2 Computational performance of HAYSTAC and other methods.	47
Figure 2.3 Accuracy of HAYSTAC and other methods for a simple simulation.	52
Figure 2.4 Accuracy of HAYSTAC and other methods for a general microbiome simulation.	54
Figure 2.5 Accuracy of HAYSTAC and other methods for an oral microbiome simulation.	57
Figure 2.6 Accuracy of HAYSTAC and other methods using a reference database restricted to a single genus.	59
Figure 2.7 HAYSTAC inferred posterior abundance levels.	60
Figure 2.8 Posterior abundances of <i>Yersinia</i> species in Case Study 1.	62
Figure 2.9 Posterior abundances of oral microbiome species in Case Study 2.	65
Figure 3.1 Percentage of duplicated reads, after sequencing screening and MDV enriched libraries.	166
Figure 3.2 Histogram of the calculated SFS, across the 19 archaeological and 1 modern newly captured samples (with a log transformed y-axis).	168
Figure 3.3 Rooted ML bootstrap tree of Marek's Disease Virus rooted with Herpesvirus of Turkeys.	171
Figure 3.4 MDV phylogenetic tree under a relaxed clock model.	173
Figure 4.1 Map of pig samples, in which pathogenic species were positively identified in.	280
Figure 4.2 Heatmap of pathogenic species and animal groups that they were identified in.	288
Figure 4.3 Presence of pathogens per time bin of a thousand years.	290

Figure 4.4 Map of canid samples, in which pathogenic species were positively identified in.	292
Figure 4.5 Heatmap of pathogenic species identified in canine samples, grouped according to their domestication status.	302
Figure 4.6 Presence of pathogens per time bin of a thousand years.	304

Table of Tables

Table 2.1 R-squared and p-value metrics from the linear model that was fitted on the computational performance results of each of the four tested methods.	48
Table 2.2 Characteristics of the 6 simulated samples sets of the Simple, Random and Oral microbiome datasets.	49
Table 2.3 False negative and false positive species counts in the ancient and modern Oral Microbiome datasets.	56
Table 3.1 Dxy among the 3 MDV lineages.	174
Table 3.2 Nucleotide diversity within each MDV lineage.	175
Table 4.1 Table with the number of samples in which pathogens from the genera <i>Bordetella</i> , <i>Borrelia</i> , <i>Brucella</i> , <i>Campylobacter</i> , <i>Corynebacterium</i> , <i>Erysipelothrix</i> , <i>Salmonella</i> , <i>Streptococcus</i> and <i>Yersinia</i> were robustly detected in pig and dog samples.	278
Table 4.2 Taxonomic and sample metadata for pig samples in which pathogens were robustly detected.	282
Table 4.3 Taxonomic and sample metadata for dog samples in which pathogens were robustly detected.	294

Acknowledgments

This thesis would not have been possible without the help and support of many people over the four years of my research. First of all, I would like to thank my PhD supervisor, Professor Greger Larson whose enthusiasm for research is infectious and has shown continuous support throughout my DPhil. I would also like to thank Professor Laurent Frantz who has also unofficially acted as a second supervisor, welcomed me into his laboratory group and taught me an incredible amount of bioinformatics and genetics.

My DPhil has been a substantially educational experience. This would not have been possible without the support and patience of a number of colleagues and collaborators. Special thanks go to Anna Linderholm, James Haile, and Ophelie Lebrasseur who have been there for me throughout, from teaching me laboratory techniques right through to providing feedback on my thesis chapters. Special thanks and gratitude goes to Evan K. Irving-Pease for spending innumerable hours teaching me best practices in bioinformatics, supervising the development of my coding and analytical skills, always being present when I needed advice and for being a great friend. I would also like to thank Alex Jamieson and Alberto Carmagnini for their substantial help and presence during my DPhil, and of course for their friendship. As well as these individuals, I wish to thank all of the members of the Palaeobarn as well as the Frantz team, especially our laboratory manager Kristina Tabbada, Sophy Charlton and Aurelie Manin.

I am also grateful to NERC who accepted me onto the DPhil programme and provided the funding to make my research possible. I would like to extend my gratitude to Hertford College and the School of Archaeology who provided funding that made the

trips to present at conferences possible. As well as the funders, I am grateful to all the sample providers and collaborators on each of the projects, especially Steven Fiddaman and Adrian Smith. I have enjoyed working with each and every one of you, and thank you for your invaluable insights.

The final and most important thanks go to my family and close friends who have both been an inspiration and supported me from a young age, helping me along the way. Finally, I would like to dedicate this DPhil to my mother, who has always been supporting me by going above and beyond to ensure my wellbeing and success, constantly believing in me, and reminding me that the only limit is the sky.

Introduction

The motivation behind combining archaeology, ancient DNA and metagenomics

Archaeology has always been a broad ever evolving discipline that incorporates new elements to facilitate its research, focusing on the study of human life in the past. Until the 19th century the practice of Historical Archaeology dominated the discipline, combining research approaches from Anthropology and History and focusing on the study of material remains of societies that had left us written records (Funari et al., 2013). But in the 19th century the concept of Environmental Archaeology started emerging, proposing that the human past cannot be fully comprehended without recognising that past cultures were integrally linked to the environments they were living in. As such, data derived from soils, sediments and various types of organic debris (e.g. bones) could provide great insight into an archaeological site, its history and the people who lived there at any given point in time (Reitz and Shackley, 2012).

Environmental archaeology stresses that interpreting archaeological sites through means of analysing organic material findings and palaeoclimate data allows us to better understand past societies (Reitz and Shackley, 2012). The main goal of historical archaeologists often was to obtain small artifacts worthy of museum display, in effort to obtain enough information and artefacts that would allow them to reconstruct past civilisations. But what historical archaeologists (Orser, 2010; Orser and Jr., 2016) were throwing away while digging, environmental archaeology (Campbell, 2018) took and used it to find out how humans interacted with, and transformed the environments they were inhabiting. Using techniques from physics, analytical chemistry and genetics, scientific archaeology was able to assemble past

ways of living from the biological remains (such as animal or human bone remains, coprolites, skin, hair, soft tissue organs, food items, etc) left behind by past societies.

Most ancient DNA studies that have happened until now have only focused on a small percentage of the DNA generated from archaeological samples, known as endogenous DNA, that is authentic to the species the archaeological specimen under study belonged to (Damgaard et al., 2015), while discarding the mainly microbial exogenous DNA fraction that was sequenced. For this PhD I have decided to follow a strategy similar to the one environmental archaeologists followed a few decades back. The exogenous DNA fraction, which researchers usually discard, has been shown to possess genuine ancient microbial communities. This microbial DNA fraction can be further separated into endogenous (antemortem) DNA, composed of host-related commensal taxa, and exogenous (postmortem) microbial DNA, composed mainly of modern environmental contamination (Warinner et al., 2017). Distinguishing the real archaeological and biological signal from environmental noise from the exogenous fraction yielded from an ancient DNA sample is not an easy task, but if done successfully, querying ancient microbial communities can answer a variety of questions ranging from pathogens and the health status of past populations to the reconstruction of past ecosystems through sediments.

Archaeological research, ancient DNA and metagenomics

Archaeological research on complex microbial samples can be traced almost a century back, with Leigh (1925) morphologically examining the prevalence of dental pathologies in Native American tribes and correlating them with their diets. And while the archaeological record is abundant with materials able to preserve biomolecules,

such as dental calculus, a mineralised bacterial biofilm formed on the surfaces of teeth (Jin and Yip, 2002), their systematic study in archaeology did not start till the 1960s-70s (Warinner et al., 2015). It was even later that the field of ancient DNA microbial genomics took off, after the publication of the first ancient *Yersinia pestis* genome (Bos et al., 2011). Ever since the fields of ancient metagenomics and paleomicrobiology have been constantly expanding by focusing on the taxonomic and/or functional reconstruction of ancient microbial communities and on genome reconstruction of ancient pathogen strains respectively (Velsko and Warinner, 2017; Warinner et al., 2015).

Ancient microbial genomic research though would not have proliferated without two major technological breakthroughs that reduced the cost for generating DNA sequencing data: high-throughput sequencing that combined clonal amplification of DNA molecules through PCR and their concurrent massively parallel sequencing (Loman et al., 2012; Reuter et al., 2015), and targeted genome enrichment protocols that through the use of custom oligonucleotide baits they capture the desired ancient microbial molecules for selective high-throughput sequencing (Burbano et al., 2010; Carpenter et al., 2013; Enk et al., 2014; Gnirke et al., 2009; Hodges et al., 2007). These two techniques, combined or not, have advanced the study of ancient metagenomic samples (Koboldt et al., 2013; Parkhill, 2013).

The total available DNA data obtained from untargeted HTS of microbiome samples are defined as metagenomes (Marchesi and Ravel, 2015), and the analysis of such large datasets is called metagenomics (Thomsen and Willerslev, 2015; Warinner et al., 2017, 2015). Metagenomic DNA samples can be generated from a variety of archaeological material in which microbial DNA could be preserved. For instance,

complex metagenomes can be yielded by sequencing cemented dental plaque also known as calculus (Velsko et al., 2017; Velsko and Warinner, 2017), or the outer layers of excavated bone samples, or even sediments, yielding a mixture of prokaryotic and eukaryotic DNA, also known as sedaDNA (Thomsen and Willerslev, 2015).

What is Metagenomics and how they are employed in ancient DNA research

One of the first approaches to characterising the diversity of microbiome samples was to use amplicon metataxonomics. That was achieved by sequencing marker genes. The usual gene candidate was the 16s unit of the rRNA molecule, a molecule with sufficient interspecific variability, that can be found across all microbes due to its essential functionalities for life (Warinner et al., 2017). This technique, which is also applicable to other various genetic loci, has been known as phylotyping (Tito et al., 2012) and metabarcoding (Andersen et al., 2012; Fantini et al., 2015).

Amplicons of these marker genes have been used for ancient DNA (aDNA) microbiome studies, but with time it has been proven that such an experimental design was not optimal for aDNA molecules. The average length of ancient DNA molecules is far smaller than the size of the hypervariable regions of genes like the 16s rRNA. Additionally, there is a bias towards the amplification of short, degraded DNA fragments (Ziesemer et al., 2015). Instead of using gene amplicon approaches for taxonomic classifications of archaeological microbiome samples, genome-wide data can be generated using shotgun-sequencing, an experimental design suitable for aDNA research (Warinner et al., 2017; Ziesemer et al., 2015). When

characterising a microbiome by analysing DNA data yielded by untargeted HTS, it is called shotgun metagenomics, and these data represent a random subset of the total DNA sequences of all the microbial taxa that are present in a sample (Marchesi and Ravel, 2015). Another experimental design suitable for metagenomic analysis is to target a portion of the total DNA present in a sample prior to sequencing, by using oligonucleotide baits that hybridize with the desired DNA molecules. This way the sequencing library is enriched for the genomes of one or more organisms of interest, a technique also known as sequence capture (Gnirke et al., 2009).

Metagenomics are more well-suited for aDNA microbiome studies than PCR-based methods. PCR based methods require DNA molecules that are long (>100 bp) and well preserved, characteristics that are usually both absent from aDNA molecules (Stoneking and Krause, 2011). Shotgun metagenomics can overcome the effect of length and/or sequence variation that exists in the underlying genome of a taxon, as Next Generation Sequencing does not depend on specific primers, as a PCR protocol otherwise would. Furthermore, HTS platforms can effectively sequence the very short and highly fragmented DNA molecules extracted from an ancient sample, and work around the bias against the amplification of degraded DNA present in PCR amplicon protocols (Warinner et al., 2017). NGS data provide valuable information for taxonomic identifications, deciphering functional relationships in complex microbiome samples, and if shotgun-sequenced a library will be containing genomic data from all four domains of life (Velsko and Warinner, 2017; Warinner et al., 2015).

Non-targeted metagenomic approaches are useful when screening samples for known and/or unknown pathogens as well. Archaeological samples in which

pathogenic DNA reads have been identified through bioinformatic analyses of screening sequencing data are prime candidates for genome capture and resequencing, in order to reconstruct the genomes of ancient pathogens (Bos et al., 2019; Spyrou et al., 2019). But pathogen enrichment techniques are not solely limited to one taxon at a time. If one wishes, targeted genome enrichment can take place for multiple species concurrently, maximising the capture of ancient pathogenic microbial DNA at a time (Velsko et al., 2017; Warinner et al., 2015, 2014).

The challenges of identifying microbial taxa from their DNA sequences

Most frequently, when studying samples that contain DNA from multiple organisms we are interested in discovering the microbial taxa that are contained in it. Here we define microbes as members of the prokaryotic domains Bacteria and Archaea (Warinner et al., 2017).

In an effort to define what a microbial species is, a unitary species concept was given by de Queiroz (2005) and Hey (2006), that is applicable to microbes. According to it, species are metapopulation lineages. Metapopulations can be further considered as sets of connected subpopulations that are maximally inclusive and the limits of which are set by evolutionary cohesive forces (Achtman and Wagner, 2008). And even though bacterial taxonomists have been using sequence similarity as a metric for employing this definition since the 1960s (Goris et al., 2007), so as to delineate species boundaries for microbes, identifying microbial species from their genetic sequences is still a major challenge (Mardis, 2008).

The two major approaches in defining a species through sequence similarity are DNA-DNA hybridisation (DDH) values and the Average Nucleotide Identity (ANI) of two microbial genomes. Two microbes will be assigned to the same species if their DDH value exceeds 70%, which corresponds to a ANI cutoff of 95-96% (Goris et al., 2007; Kim et al., 2014). Interestingly when the sequence similarity comparison is between single marker genes, such as the 16s rRNA gene, then the species cut-off should be more stringent and get recalibrated to 97-99% ANI (Kim et al., 2014; Větrovský and Baldrian, 2013).

If a taxon is only described by the 16s rRNA marker gene then it cannot be described as a species but rather as an Operational Taxonomic Unit (OTU), implying that all microbes assigned to such a unit are related by descent. An OTU, as long as it is clearly defined, can be used as a proxy for any taxonomic level (Hughes and Bohannon, 2004). The ease of use of marker genes, like the 16s rRNA gene, has led to the development of bioinformatic tools that compare sequenced DNA reads to reference databases of either marker genes or whole genomes, when available. Either by comparing DNA reads to whole genomes or just marker genes, these tools (for example MIDAS (Nayfach et al., 2016)) try to apply a taxonomy on the microbial taxa that can be found in a metagenomic sample by analysing sequence similarity (Kim et al., 2014; Richter and Rosselló-Móra, 2009).

But not all metagenomic pipelines are dependent on sequence comparison. For instance, software such as Kraken (Wood et al., 2019; Wood and Salzberg, 2014) rather than focusing on sequence similarity among DNA reads and reference genomes, it infers taxonomic information from reference databases containing

frequency distributions of short sequence fragments (k-mers) across a range of known taxa (Warinner et al., 2017). Such an approach is also known as k-mer binning, and exactly because it uses short DNA fragments for taxonomic identifications, it is well suited for the study of ancient microbial DNA.

Functionality framework of metataxonomic tools

The expansion of metagenomics has been accompanied by the development of a variety of tools to perform metataxonomic assignments with. Depending on the research question that bioinformatic pipelines are called to answer, or the data they need to analyse, different pipelines have different functional approaches to taxonomic characterisation of microbiomes. Thus, metataxonomic pipelines can be broadly categorised based on the way they construct their reference databases and the data they use to populate them, as well as the type of query data that they need to parse in order to perform a taxonomic identification.

The construction and query of a database by a metagenomic classification tool can have a significant impact on the sensitivity and specificity of the overall method but also, they can have an impact on the biases that may rise during the analysis of the data. A reference database can be made of sequences from a single locus. The standard gene in microbial diversity studies has been the 16S rRNA gene. Such databases include many more taxa, compared to databases of multiple loci. Great examples of such databases are SILVA (Quast et al., 2013) and Greengenes (McDonald et al., 2012). Single locus databases do not encapsulate enough

sequence variation to allow distinguishing among closely related taxa on the species, subspecies or strain taxonomic levels.

Alternatively, a database can contain the sequences of multiple genes that regulate functions such as translation, transcription, replication and gene regulation, whose sequences are highly conserved, otherwise known as housekeeping genes (Margos et al., 2008; Segata and Huttenhower, 2011). Usually sequences of housekeeping genes in databases are single copy, so that copy number variation biases during abundance estimation can be avoided. Populating a database with a wider variety of genes also allows for higher amounts of inter- and intra-species sequence variability, compared to just rRNA genes, making the distinction of different strains possible.

Reference databases can also comprise whole genome sequences as well. Whole genomes have higher amounts of sequence diversity making them more fit for finer taxonomic resolution in metagenomic analyses. They are also well suited for aDNA studies, as theoretically, any fragment that originates from a species that is included in the target database can be assigned to it, making the detection of taxa with either low abundance or from short damaged reads possible (Warinner et al., 2017).

Disadvantages of whole genome databases, in the context of metataxonomic characterisation of a microbiome, include biases stemming from horizontal gene transfer and bacterial recombination. Microbial genomes are highly variable in their size, and consequently larger genomes attract more query reads to them, leading to inflated abundance estimations or even false positive hits (Warinner et al., 2017). Far fewer complete genomes are available than 16S rRNA gene sequences making that

sort of database rather incomplete. This can lead to false positive or negative findings when analysing a metagenomic dataset, due to the overrepresentation or underrepresentation of taxa respectively.

The handling of query sequences and the way they query a reference database during taxonomic assignment are equally important in a metagenomic analysis. A classical way of assigning query sequences to reference databases is through traditional sequence alignments. These alignments can be performed with variations of the Needleman-Wunsch (Needleman and Wunsch, 1970) or the Smith-Waterman (Smith and Waterman, 1981) algorithms and they are usually carried out by software like bowtie2 (Langmead and Salzberg, 2012) or bwa (Li, 2013). One major drawback of these methods is that alignments are time consuming, even for optimised versions of the above algorithms.

An alternative way to perform metataxonomic assignments is the use of alignment free methods. In order to remove the computational cost of calculating precise alignments, both all the genomes in a reference database and the query sequences are split in overlapping fragments of length k . Programmes like CLARK-S (Ounit and Lonardi, 2016) and Kraken (Wood and Salzberg, 2014) subsequently use hashing algorithms to match exact kmers from the queries to the reference database. Such a strategy, despite being time efficient, can lead to a large number of false positive hits, especially in cases where samples are being screened for pathogenic taxa. Unfortunately though it is expected that every large metagenomic dataset will generate a number of false positives when such a method is employed (Warinner et al., 2017).

What can you do after you identify a species

Metagenomics entails massively parallel DNA sequencing of clinical or environmental samples (most often containing microorganisms that are uncultured/able) followed by bioinformatic-based analyses to accurately infer species richness.

In addition, sequencing reads obtained by NGS can inform us about the proportion of each microbe in a sample and thus allow us to infer its relative abundance. Relative quantification is crucial, as it can depict the size of microbial communities and can reveal possible ecological relationships among microorganisms (Mardis, 2008), but can also reveal information of epidemiological and potentially clinical interest, like strain type information and type of transmission (Hasman et al., 2014).

With the advent of metagenomics, accurate taxonomic and functional *in silico* assignments became fundamental. Identifying the species a sample contains, from DNA sequences, can also be accompanied by characterising the different metabolic functions that are performed in a microbial community (Lam et al., 2015), leading for instance to the discovery of novel enzymes that allow microbes to adapt to extreme environments (such as alkaline ecosystems) (Mirete et al., 2016). Screening metagenomes, sampled from various ecosystems, for the discovery of novel enzymes overcomes the limitations of traditional culture-based techniques, and apart from the fields of microbiology and ecology, it contributes even to the development of

industrial biotechnological practices (Lam et al., 2015; Mirete et al., 2016; Uchiyama and Miyazaki, 2009).

In biomolecular archaeology, applying metagenomic approaches in ancient DNA WGS data has also enabled the identification of extinct mammalian megafauna, making it possible to recover genomes of ancient *Elephantidae*, such as mammoths, and hominid genomes, such as Neanderthal genomes (Schuster, 2008). This, together with the breakthroughs made for sequencing complex DNA mixtures from most diverse sources will allow for the study of any ecosystem at the sequence level. It will also open a window to the flora and fauna of at least the last 10,000 years, in ways far beyond what would have been deemed possible only with the use of conventional genomic methods (Schuster, 2008).

Further improvement in the study of microbial communities (both environmental and human), in terms of time and biologically relevant questions that can be asked, can be also achieved through metagenomics. Such an advancement can happen as we shift from a classical microbiological approach to a whole genome and/or community sequencing approach (Hasman et al., 2014). Sequencing DNA from purified cultures or DNA extracted straight from the sample itself allows to identify the composition of the community in a much shorter time (approximately 24 hours less when compared to the classical microbiology approach). A direct community sequencing approach could also give a more accurate representation of the gene content of such a community compared even to a whole genome sequencing approach, such as in the case of strains carrying resistance genes that tend to get lost in pure cultures (Hasman et al., 2014).

Looking for pathogens in unlikely places

A significant advantage of the field of metagenomics is that it can contribute significantly to archaeological research, as complex DNA mixtures are yielded every time an archeological sample is shotgun-sequenced. For instance, sediment and soil samples are excellent sources of genetic information for past and modern biodiversity monitoring (Andersen et al., 2012; Thomsen and Willerslev, 2015), permafrost samples can reveal extinct species of the virosphere (Legendre et al., 2014), and archaeological bone samples can give us a deeper insight into the symbiotic microbial flora of past animals and humans and their pathogens (Bos et al., 2011; Spyrou et al., 2019; Velsko and Warinner, 2017; Warinner et al., 2014). As most metagenomic bioinformatic methods, scientific questions and applications do not restrict themselves in the study of modern microbiomes exclusively, they can be easily applied to the study of ancient microbiomes and answer archaeologically related questions.

Having a robust framework for bioinformatic species identification would allow us to expand metagenomic research, not only in terms of the hypotheses that could be tested, but it would also allow us to go back in time and robustly study ancient microbiomes, even in cases where DNA preservation is poor and data is scarce. A hypothesis, with potentially promising results, would be whether it's worthwhile screening nontraditional substrates for pathogen species that might have colonised them opportunistically. A good candidate substrate to test this hypothesis, for instance, could be dental calculus (Warinner et al., 2015). Inside this cemented

material, apart from DNA from the host's own cells, bacterial DNA from the host's oral microbiome can be preserved (Warinner et al., 2014) along with other biomolecules, such as proteins. These properties of dental calculus make it a valuable archaeological material and a suitable substrate for ancient DNA research. Factoring in the proximity of the oral cavity to the tissues of the upper respiratory system, dental calculus becomes the perfect candidate substrate to test if any pathogens of the respiratory system could be detected in it. By screening dental calculus/plaque for respiratory pathogens, we can evaluate whether this substrate can act as a potential reservoir for opportunistic pathogens (Warinner et al., 2014).

Such a robust hypothesis-driven, quantitative species identification method could be useful for ancient DNA research. It can be used as a tool to establish community membership of organisms, whose DNA can be found in the non-endogenous DNA content derived from ancient samples. By screening the metagenomes generated from ancient human or domesticated animal samples for pathogens, we can assess the health status of a past population (Warinner et al., 2015, 2014). It should be noted though, that by detecting the existence of a pathogenic agent in a sample doesn't necessarily mean that this pathogen was the cause of that individual's death. Nevertheless, information can still be obtained about how widespread that microbe was. A confident bioinformatic species identification could also indicate whether any further laboratory work will be conducted, like targeted genome capture, that will pave the way for studies around a microbe's evolution and host interactions.

As more ancient DNA samples are being sequenced from various sources, metagenomics begin to find a new place in the field of palaeopathology, by enabling

the discovery of extinct pathogenic variation (e.g. *Mycobacterium tuberculosis*, *Treponema pallidum*, Hepatitis B) which when combined with modern microbial genomic information can lead to a better understanding of disease evolution (Anastasiou and Mitchell, 2013). Paleovirology too has advanced by utilising metagenomic approaches. Through mining ancient metagenomic data our knowledge of the virosphere has expanded, by recovering genomes of either extinct or not previously sampled viral taxa/strains (Tsangaras and Greenwood, 2019). Characterising previously unsampled viruses, especially from the past, leads to building more robust phylogenies through sampling the taxa that lie on the missing branches among major viral families and orders, especially in the case of divergent RNA viruses (Zhang et al., 2018). Examining bones that have clear signs of pathology left by a disease can yield valuable discoveries of past pathogenic variants (such as *Yersinia pestis*, *Mycobacterium tuberculosis* and *Mycobacterium leprae* (Bos et al., 2011; Stone et al., 2009)), but screening non-traditional source materials can lead to surprising findings as well. Namely, permafrost cores have been proven to be able to preserve extinct species of giant viruses (Legendre et al., 2014) broadening our knowledge of the ancient virosphere.

Metagenomics and robust taxonomic identification can be employed to track the spread of various diseases among animal populations and humans in the past, but more importantly allow to discover the causative agents of past epidemics or mass mortality events. For example, aDNA metagenomic research could identify what microbes caused an increasing trend in mass mortality events of livestock animals during both Roman, and post-Black Death periods, but also during the advent of controlled breeding in the 17th and 18th century (Binois, 2015). In addition, by

creating time series of DNA pathogen data, and constantly expanding the time depth by adding new samples, we could go back as far as the 17th Century Agricultural Revolution (Pretty, 1991; Thomas, 2005) or even as early as the First Transition ten thousand years ago (Stone, 2020), and discover when specific animal pathogens and parasites emerged. By using historical data, we can then trace the transmission routes of identified zoonotic disease agents through commerce, getting a clearer image on how human activity has been shaping the landscape of animal disease. Investigating the evolutionary adaptations of a range of domesticated species (dogs, pigs and chickens) to pathogens of various degrees of virulence, can also offer great understanding in how zoonotic diseases act as an evolutionary pressure that shapes genomes.

Characteristics of ancient DNA for validating metagenomic results

As mentioned before, an analysis of a large metagenomic dataset is highly likely to be plagued with false, either negative or positive identifications. For that reason, it is important to validate the findings of metataxonomic tools as much as possible when studying ancient microbes.

Even though ancient DNA analyses are confounded due to the sequenced DNA reads being short and chemically damaged, DNA damage patterns are used to validate aDNA metagenomic identification results. Enzymatic decomposition of the DNA, caused by endogenous nucleases, and to a lesser extent hydrolysis and oxidation, are the driving factors of molecular lesions characteristic to aDNA (Molak and Ho, 2011). There are two categories of chemical damage occurring in aDNA

molecules, with the first one being lesions that block strand elongation during PCR (e.g. strand breaks) leading to shorter DNA reads (Briggs et al., 2007; Molak and Ho, 2011). The second type of damage is miscoding lesions, that even though they allow for strand elongation during PCR, they cause nucleotide changes that lead to faulty sequencing of these sites (Briggs et al., 2007; Molak and Ho, 2011). As miscoding lesions are expected to be found in higher frequency in aDNA molecules, they are also one of the main criteria used for aDNA sequence authentication (Briggs et al., 2007; Jónsson et al., 2013). The vast majority of post mortem miscoding lesions are transitions (C to T and G to A), with cytosine deamination being mainly responsible for these nucleotide changes (Briggs et al., 2007).

An additional confounding factor, that highlights the need for additional result validation, is that aDNA is normally present in low amounts or absent from many specimens, and traces of modern DNA could possibly be mistaken for ancient endogenous content (Briggs et al., 2007). This challenge could even present itself when the taxa in question are vertebrates, such as detecting human contamination when analysing DNA data from ancient dog samples (Hofreiter et al., 2001; Malmström et al., 2005). WGS metagenomic data generated from archaeological samples are highly complex, including DNA from multiple sources (e.g. endogenous organismal DNA, endogenous ancient microbiota, post mortem colonisers, environmental contamination) that bioinformaticians are required to disentangle (Warinner et al., 2017). When analysing ancient microbiome samples it is necessary to distinguish host-related endogenous microbiota and exogenous microbiota, which includes decomposition-related and recent environmental contaminant microbes (Warinner et al., 2017). Towards that end algorithms, such as SourceTracker

(Knights et al., 2011), are being developed that are able to estimate the sources that have contributed to the composition of a metagenome.

Result validation and sequence authentication in aDNA metagenomic analyses should not only be restricted to estimating base misincorporation frequencies and detecting the presence of damage patterns. When alignment based methods are used for metataxonomic assignments, all valid alignments should have a minimal edit distance (number of mismatches between a read and a reference genome) from a reference genome (Key et al., 2017; Warinner et al., 2017). Concurrently, if a taxonomic identification is authentic, then the DNA reads that are assigned to this taxon should be evenly, almost randomly, distributed across its genome (Warinner et al., 2017). If the reads cluster around specific genomic areas, while leaving large gaps of low coverage throughout the genome, then that hints towards a potential false positive identification (Key et al., 2017).

Data science and bioinformatics

The primary type of data analysed in this PhD was Next Generation Sequencing (NGS) DNA data. In the last decade sequencing platforms have managed to increase their throughput massively (Mardis, 2008), providing us with continuously larger and richer datasets. In order to leverage these DNA sequencing advancements, bioinformatic workflows needed to be combined with techniques used in data science (otherwise known as data stewardship) (Attwood et al., 2019). Thus, bioinformatics have now evolved into a highly integrative field, which needs to be able to combine, manipulate and transform information from multiple data sources

or databases (e.g. DNA sequencing, gene expression experiments, protein sequencing, radiocarbon dating, taxonomy metadata, or chemical damage results) to produce meaningful results (Roos, 2001).

Even for the simplest case of accessioning sample and/or experiment metadata, today's bioinformatician needs to be adept at maintaining efficient databases and enforcing the necessary data hygiene practices. Out of a wide range of tools available to facilitate these tasks, Python (Van Rossum and Drake, 2009) programming language and SQL (Structured Query Language) ("MySQL 8.0 Reference Manual," n.d.; Widenius et al., 2002) are tools highly preferred by bioinformaticians. Both programming languages have additional implementations, biopython and bioSQL respectively (Cock et al., 2009), for handling and manipulating biological data, such as fasta sequence files, fastq sequencing data files, or NCBI sequence data to name a few.

But due to the sheer volume of data produced by DNA sequencing and other experimental procedures, the biological datasets requiring analysis are increasingly getting bigger in terms of informational content and dimensions (e.g. links between genotype and phenotype), obtaining the status of 'big data' (Stephens et al., 2015; Wang et al., 2016). For example genetic analyses, like evolutionary tree building or genome assembling, can suffer from excessive RAM usage, impeding analyses for researchers without access to large computational infrastructure. To tackle such problems, modern bioinformatics need to borrow techniques employed on 'big data', that take advantage of parallel computing to reduce memory requirements, or come

up with heuristic solutions that reduce the complexity of the analysed data, resulting in minimised needs in RAM and/or processing time.

Bioinformatic methods are also reinventing themselves as they adapt to new challenges. Modern bioinformatic tools are now being built on computational frameworks that employ parallel computing for optimal computational resource management (e.g. Snakemake (Köster and Rahmann, 2012)). Bioinformatics have also highlighted the value of reproducibility of analyses and published results, promoting transparency in life sciences. Collaboration and knowledge dissemination are aspects important in modern bioinformatics. To facilitate that, public repositories like bioconda (Grüning et al., 2018) have been set up, where ready-to-use software is hosted and distributed, helping people with not a strong computational background start mining and analysing their data.

Recapitulation

Ancient microbial communities provide an excellent source of information that enables us to ask novel questions about past life. Metagenomic research in aDNA can elucidate pathogen evolution, as in the case of MDV where pathogenicity keeps increasing even in the present day breaking the protective barriers of vaccination, find ancestral strains of pathogens such as *Yersinia pestis* that are basal to modern microbial variants (Bos et al., 2011; Rasmussen et al., 2015), and help in the discovery of extinct viral species (Legendre et al., 2014). Ancient metagenomes can facilitate environmental research (Thomsen and Willerslev, 2015) and can be used to

even reconstruct the species range of extinct megafauna (Schuster, 2008). Most importantly, it provides a toolkit with which we can expand the range of material we can use for bioarchaeological research, such as dental calculus or coprolites, and it therefore allows us to develop cutting edge methods for more efficient aDNA data analysis and mining, as we only grow curiouser.

The overall theme of this PhD thesis is using bioinformatics to answer questions tied to specific taxa or time periods. I first developed a robust and lightweight bioinformatic method, HAYSTAC (Dimopoulos et al., 2020) with a clear probabilistic model to perform metagenomic identifications. I subsequently employed HAYSTAC to screen approximately 1100 modern and ancient chicken samples for *Marek's Disease Virus* (MDV). MDV sequences found in the positive samples were enriched through genome capture (Carpenter et al., 2013), and their reconstructed genomes were combined with modern sequence data in order to study the evolution of MDV's pathogenicity as a response to the intensive agricultural practices introduced by farmers from the second half of the 20th century onwards. For my third main data chapter, I performed an exploratory analysis by screening approximately 1100 and 2300 sequencing libraries of pigs and dogs respectively (from ancient and modern samples) for zoonotic pathogens, so as to understand how zoonotic pathogen prevalence changes throughout time and space and how domestication has shaped the disease landscape of early domesticated animals. This research would not have been possible without the help of the PalaeBARN team, which was responsible for generating sequencing data for thousands of samples from multiple species of domesticated and wild animals.

Bibliography

- Achtman, M., Wagner, M., 2008. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* 6, 431–440.
- Anastasiou, E., Mitchell, P.D., 2013. Palaeopathology and genes: investigating the genetics of infectious diseases in excavated human skeletal remains and mummies from past populations. *Gene* 528, 33–40.
- Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjaer, K.H., Orlando, L., Gilbert, M.T.P., Willerslev, E., 2012. Meta-barcoding of “dirt” DNA from soil reflects vertebrate biodiversity. *Mol. Ecol.* 21, 1966–1979.
- Attwood, T.K., Blackford, S., Brazas, M.D., Davies, A., Schneider, M.V., 2019. A global perspective on evolving bioinformatics and data science training needs. *Brief. Bioinform.* 20, 398–404.
- Binois, A., 2015. Excavating the history of ancient veterinary practices. *Vet. Rec.* 176, 564–569.
- 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., Spyrou, M.A., Krause, J., 2019. Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens. *Annu. Rev. Microbiol.* 73, 639–666.
- 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., Wood, J., Earn, D.J.D., Herring, D.A., Bauer, P., Poinar, H.N., Krause, J., 2011. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* 478, 506–510.
- Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., Pääbo, S., 2007. Patterns of

- damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* 104, 14616–14621.
- Burbano, H.A., Hodges, E., Green, R.E., Briggs, A.W., Krause, J., Meyer, M., Good, J.M., Maricic, T., Johnson, P.L.F., Xuan, Z., Rooks, M., Bhattacharjee, A., Brizuela, L., Albert, F.W., de la Rasilla, M., Fortea, J., Rosas, A., Lachmann, M., Hannon, G.J., Pääbo, S., 2010. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328, 723–725.
- Campbell, G., 2018. Environmental Archaeology. *The Encyclopedia of Archaeological Sciences*. <https://doi.org/10.1002/9781119188230.saseas0218>
- Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y., Wang, J., Gilbert, M.T.P., Willerslev, E., Greenleaf, W.J., Bustamante, C.D., 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* 93, 852–864.
- Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423.
- Damgaard, P.B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., Allentoft, M.E., 2015. Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* 5, 11184.
- de Queiroz, K., 2005. Ernst Mayr and the modern concept of species. *Proc. Natl. Acad. Sci. U. S. A.* 102, 6600–6607.

- Dimopoulos, E.A., Carmagnini, A., Velsko, I.M., Warinner, C., Larson, G., Frantz, L.A.F., Irving-Pease, E.K., 2020. HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2020.12.16.419085>
- Enk, J.M., Devault, A.M., Kuch, M., Murgha, Y.E., Rouillard, J.-M., Poinar, H.N., 2014. Ancient whole genome enrichment using baits built from modern DNA. *Mol. Biol. Evol.* 31, 1292–1294.
- Fantini, E., Gianese, G., Giuliano, G., Fiore, A., 2015. Bacterial metabarcoding by 16S rRNA gene ion torrent amplicon sequencing. *Methods Mol. Biol.* 1231, 77–90.
- Funari, P.P.A., Hall, M., Jones, S., 2013. *Historical Archaeology: Back from the Edge*. Routledge.
- Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., Nusbaum, C., 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27, 182–189.
- Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91.
- Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., Köster, J., Bioconda Team, 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476.
- Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C.A.,

- Frimodt-Møller, N., Aarestrup, F.M., 2014. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J. Clin. Microbiol.* 52, 139–146.
- Hey, J., 2006. On the failure of modern species concepts. *Trends Ecol. Evol.* 21, 447–450.
- Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., McCombie, W.R., 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527.
- Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M., Pääbo, S., 2001. Ancient DNA. *Nat. Rev. Genet.* 2, 353–359.
- Hughes, J.B., Bohannan, B.J.M., 2004. Section 7 update: Application of ecological diversity statistics in microbial ecology, in: *Molecular Microbial Ecology Manual*. Kluwer Academic Publishers.
- Jin, Y., Yip, H.-K., 2002. Supragingival calculus: formation and control. *Crit. Rev. Oral Biol. Med.* 13, 426–441.
- Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., Orlando, L., 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684.
- Key, F.M., Posth, C., Krause, J., Herbig, A., Bos, K.I., 2017. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet.* 33, 508–520.
- Kim, M., Oh, H.-S., Park, S.-C., Chun, J., 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 64, 346–351.

- Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., Kelley, S.T., 2011. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* 8, 761–763.
- Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., Mardis, E.R., 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* 155, 27–38.
- Köster, J., Rahmann, S., 2012. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522.
- Lam, K.N., Cheng, J., Engel, K., Neufeld, J.D., Charles, T.C., 2015. Current and future resources for functional metagenomics. *Front. Microbiol.* 6, 1196.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., Claverie, J.-M., 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U. S. A.* 111, 4274–4279.
- Leigh, R.W., 1925. Dental pathology of Indian tribes of varied environmental and food conditions. *Am. J. Phys. Anthropol.* 8, 179–199.
- Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.
- Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., Pallen, M.J., 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439.
- Malmström, H., Storå, J., Dalén, L., Holmlund, G., Götherström, A., 2005. Extensive

- human DNA contamination in extracts from ancient dog bones and teeth. *Mol. Biol. Evol.* 22, 2040–2047.
- Marchesi, J.R., Ravel, J., 2015. The vocabulary of microbiome research: a proposal. *Microbiome* 3, 31.
- Mardis, E.R., 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402.
- Margos, G., Gatewood, A.G., Aanensen, D.M., Hanincová, K., Terekhova, D., Vollmer, S.A., Cornet, M., Piesman, J., Donaghy, M., Bormane, A., Hurn, M.A., Feil, E.J., Fish, D., Casjens, S., Wormser, G.P., Schwartz, I., Kurtenbach, K., 2008. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. U. S. A.* 105, 8730–8735.
- McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618.
- Mirete, S., Morgante, V., González-Pastor, J.E., 2016. Functional metagenomics of extreme environments. *Curr. Opin. Biotechnol.* 38, 143–149.
- Molak, M., Ho, S.Y.W., 2011. Evaluating the impact of post-mortem damage in ancient DNA: a theoretical approach. *J. Mol. Evol.* 73, 244–255.
- MySQL 8.0 Reference Manual [WWW Document], n.d. URL <https://dev.mysql.com/doc/refman/8.0/en/> (accessed 4.30.21).
- Nayfach, S., Rodriguez-Mueller, B., Garud, N., Pollard, K.S., 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* 26, 1612–1625.

- Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.
- Orser, C.E., 2010. Twenty-First-Century Historical Archaeology. *Journal of Archaeological Research* 18, 111–150.
- Orser, C.E., Jr., 2016. *Historical Archaeology*. Routledge.
- Ounit, R., Lonardi, S., 2016. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics* 32, 3823–3825.
- Parkhill, J., 2013. What has high-throughput sequencing ever done for us? *Nat. Rev. Microbiol.* 11, 664–665.
- Pretty, J.N., 1991. Farmers' extension practice and technology adaptation: Agricultural revolution in 17–19th century Britain. *Agric. Human Values* 8, 132–148.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–6.
- Rasmussen, S., Allentoft, M.E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., Pedersen, A.G., Schubert, M., Van Dam, A., Kapel, C.M.O., Nielsen, H.B., Brunak, S., Avetisyan, P., Epimakhov, A., Khalyapin, M.V., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., Moiseyev, V., Gromov, A., Pokutta, D., Saag, L., Varul, L., Yepiskoposyan, L., Sicheritz-Pontén, T., Foley, R.A., Lahr, M.M., Nielsen, R., Kristiansen, K., Willerslev, E., 2015. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* 163, 571–582.
- Reitz, E.J., Shackley, M., 2012. Introduction to Environmental Archaeology, in: Reitz, E.J., Shackley, M. (Eds.), *Environmental Archaeology*. Springer US, Boston, MA,

pp. 1–39.

- Reuter, J.A., Spacek, D.V., Snyder, M.P., 2015. High-throughput sequencing technologies. *Mol. Cell* 58, 586–597.
- Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131.
- Roos, D.S., 2001. Computational biology. Bioinformatics--trying to swim in a sea of data. *Science* 291, 1260–1261.
- Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18.
- Segata, N., Huttenhower, C., 2011. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One* 6, e24704.
- Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.
- Spyrou, M.A., Bos, K.I., Herbig, A., Krause, J., 2019. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* 20, 323–340.
- Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E., 2015. Big Data: Astronomical or Genomical? *PLoS Biol.* 13, e1002195.
- Stone, A.C., 2020. Getting sick in the Neolithic. *Nat Ecol Evol.*
- Stone, A.C., Wilbur, A.K., Buikstra, J.E., Roberts, C.A., 2009. Tuberculosis and leprosy in perspective. *Am. J. Phys. Anthropol.* 140 Suppl 49, 66–94.
- Stoneking, M., Krause, J., 2011. Learning about human population history from ancient and modern genomes. *Nat. Rev. Genet.* 12, 603–614.
- Thomas, R., 2005. Zooarchaeology, Improvement and the British Agricultural

- Revolution. *Int. J. Hist. Archaeol.* 9, 71–88.
- Thomsen, P.F., Willerslev, E., 2015. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18.
- Tito, R.Y., Knights, D., Metcalf, J., Obregon-Tito, A.J., Cleeland, L., Najar, F., Roe, B., Reinhard, K., Sobolik, K., Belknap, S., Foster, M., Spicer, P., Knight, R., Lewis, C.M., Jr, 2012. Insights from characterizing extinct human gut microbiomes. *PLoS One* 7, e51146.
- Tsangaras, K., Greenwood, A.D., 2019. Paleovirology: Viral Sequences from Historical and Ancient DNA, in: Lindqvist, C., Rajora, O.P. (Eds.), *Paleogenomics: Genome-Scale Analysis of Ancient DNA*. Springer International Publishing, Cham, pp. 139–162.
- Uchiyama, T., Miyazaki, K., 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* 20, 616–622.
- Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual: (Python Documentation Manual Part 2). CreateSpace Independent Publishing Platform.
- Velsko, I.M., Overmyer, K.A., Speller, C., Klaus, L., Collins, M.J., Loe, L., Frantz, L.A.F., Sankaranarayanan, K., Lewis, C.M., Jr, Martinez, J.B.R., Chaves, E., Coon, J.J., Larson, G., Warinner, C., 2017. The dental calculus metabolome in modern and historic samples. *Metabolomics* 13, 134.
- Velsko, I.M., Warinner, C.G., 2017. Bioarchaeology of the human microbiome. *Bioarchaeology International*.
- Větrovský, T., Baldrian, P., 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* 8, e57923.

- Wang, L., Wang, Y., Chang, Q., 2016. Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods* 111, 21–31.
- Warinner, C., Herbig, A., Mann, A., Fellows Yates, J.A., Weiß, C.L., Burbano, H.A., Orlando, L., Krause, J., 2017. A Robust Framework for Microbial Archaeology. *Annu. Rev. Genomics Hum. Genet.* 18, 321–356.
- Warinner, C., Rodrigues, J.F.M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R.Y., Fiddyment, S., Speller, C., Hendy, J., Charlton, S., Luder, H.U., Salazar-García, D.C., Eppler, E., Seiler, R., Hansen, L.H., Castruita, J.A.S., Barkow-Oesterreicher, S., Teoh, K.Y., Kelstrup, C.D., Olsen, J.V., Nanni, P., Kawai, T., Willerslev, E., von Mering, C., Lewis, C.M., Jr, Collins, M.J., Gilbert, M.T.P., Rühli, F., Cappellini, E., 2014. Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* 46, 336–344.
- Warinner, C., Speller, C., Collins, M.J., 2015. A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370, 20130376.
- Widenius, M., Axmark, D., MySQL, A.B., Arno, K., 2002. *MySQL Reference Manual: Documentation from the Source*, 1st ed. “O’Reilly Media, Inc.,” Sebastopol, CA, USA.
- Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257.
- Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46.
- Zhang, Y.-Z., Shi, M., Holmes, E.C., 2018. Using Metagenomics to Characterize an Expanding Virosphere. *Cell* 172, 1168–1172.
- Ziesemer, K.A., Mann, A.E., Sankaranarayanan, K., Schroeder, H., Ozga, A.T.,

Brandt, B.W., Zaura, E., Waters-Rist, A., Hoogland, M., Salazar-García, D.C., Aldenderfer, M., Speller, C., Hendy, J., Weston, D.A., MacDonald, S.J., Thomas, G.H., Collins, M.J., Lewis, C.M., Hofman, C., Warinner, C., 2015. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci. Rep.* 5, 16498.

HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data

HAYSTAC: **H**igh-**A**ccurac**Y** and **S**calable **T**axonomic **A**ssignment of Metagenomi**C** data

Statement of Authorship:

Research design: I designed the research, with input from Greger Larson, Laurent Frantz and Evan K. Irving-Pease. Alberto Carmagnini designed the mathematical framework of HAYSTAC. I coded and designed the software under the guidance and supervision of Evan K. Irving-Pease.

Data generation: I generated the simulated datasets and downloaded the public datasets. Irina M. Velsko generated the dental calculus dataset.

Analysis: I performed all the computational analyses presented in this paper and interpreted the results under the guidance of Laurent Frantz, Christina Warinner, Irina M. Velsko, Alberto Carmagnini, Greger Larson, and Evan K. Irving-Pease.

Manuscript: I wrote the manuscript, with input from all other co-authors.

Authors:

Evangelos A. Dimopoulos^{1*}, Alberto Carmagnini^{2*}, Irina M. Velsko³, Christina Warinner^{3,4}, Greger Larson¹, Laurent A. F. Frantz^{2,5§} and Evan K. Irving-Pease^{1,6§}

*These authors contributed equally to the work

§These authors co-supervised this work

¹ The Palaeogenomics and Bio-archaeology Research Network, Research Laboratory for Archaeology and History of Art, University of Oxford, Oxford, UK.

² School of Biological and Chemical Sciences, Queen Mary University of London, London, UK.

³ Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany.

⁴ Department of Anthropology, Harvard University, Cambridge, USA.

⁵ Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, Munich, Germany.

⁶ Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Abstract

Identification of specific species in metagenomic samples is critical for several key applications, yet many tools available require large computational power and are often prone to false positive identifications. Here we describe **High-AccuracY** and **Scalable Taxonomic Assignment of MetagenomiC** data (HAYSTAC), which can estimate the probability that a specific taxon is present in a metagenome. HAYSTAC provides a user-friendly tool to construct databases, based on publicly available genomes, that are used for competitive reads mapping. It then uses a novel Bayesian framework to infer the abundance and statistical support for each species identification, and provide per-read species classification. Unlike other methods, HAYSTAC is specifically designed to efficiently handle both ancient and modern DNA data, as well as incomplete reference databases, making it possible to run highly accurate hypothesis-driven analyses (i.e., assessing the presence of a specific species) on variably sized reference databases while dramatically improving processing speeds. We tested the performance and accuracy of HAYSTAC using simulated Illumina libraries, both with and without ancient DNA damage, and compared the results to other currently available methods (i.e., Kraken/Braken, MALT/HOPS, and Sigma). HAYSTAC identified fewer false positives than both Kraken/Braken and MALT/HOPS in all simulations, and fewer than Sigma in simulations of ancient data. It runs faster than Sigma for large databases, uses less memory than Kraken/Braken, and is both faster and uses less memory than MALT/HOPS. Lastly, we used HAYSTAC to search for specific pathogens in two

published ancient metagenomic datasets, demonstrating how it can be applied to empirical datasets. HAYSTAC is available from <https://github.com/antonisdim/HAYSTAC>

Introduction

Metagenomics allows high-throughput sequencing of complex microbial communities that may not be possible to grow in a laboratory. Accurate identification of specific species within a complex community is critical for metagenomic applications in a wide variety of fields, including medicine, microbiome sciences, environmental sciences, biosurveillance, and biomolecular archaeology. Confidently assessing the presence of individual species is also crucial for taxonomic profiling (Gonzalez et al., 2016), inferring genomic evolution and phylogenetics (Tett et al., 2019), tracing disease outbreaks (Ahn et al., 2015), generating diagnostics from biofluids (Wilson et al., 2019) and studying ancient pathogens (Spyrou et al., 2019).

Recent studies of ancient pathogens based on high-throughput sequencing technology have provided unique insights into the evolutionary history of major human pathogens. These include the etiologic agents of both the Justinian and Black Death plagues (*Yersinia pestis*; (Bos et al., 2011; Rasmussen et al., 2015; Schuenemann et al., 2011; Spyrou et al., 2016)), tuberculosis (the *Mycobacterium tuberculosis* complex; (Müller et al., 2014; Pósa et al., 2015)), syphilis (*Treponema pallidum*; (Arora et al., 2016; Schuenemann et al., 2018)) and smallpox (*Variola major* and *Variola minor*; (Mühlemann et al., 2020)).

Accurately identifying specific microbial species from ancient metagenomes, however, can be challenging for numerous reasons. Firstly, short read sequences originating from

closely related species may be difficult to distinguish from each other. In addition, the species of interest may have low abundance in the sample, or the genomes in the reference database may not be representative of the strain present in a sample (Warinner et al., 2017). Contamination from the laboratory environment or reagents may also obscure identification (Davis et al., 2018). Identifying specific species within ancient metagenomes is further complicated by issues inherent to ancient DNA, including short molecules (typically between 30-60 bp; (Kistler et al., 2017)), post-mortem miscoding lesions (e.g., cytosine deamination), contamination from soil and sediment bacteria in the burial environment, sample handling, and storage (Mann et al., 2018).

When present in ancient metagenomes, pathogen DNA is usually found in low abundance (Warinner et al., 2017), and researchers often initially screen many samples before deciding which libraries should be sequenced deeper in order to obtain whole genome data, or be further processed using targeted capture techniques (Borówka et al., 2019; Key et al., 2017; Spyrou et al., 2019). Screening for pathogens is typically performed by shallow shotgun sequencing followed by *in silico* analyses of short read data using taxonomic classifiers. Following screening, libraries with a positive identification for a species of interest are then more deeply sequenced and investigated, with or without targeted capture (Spyrou et al., 2019). Confident species identification from metagenomic screening data is therefore critical for studies of ancient pathogens.

Tools for detecting microbial species from metagenomes include metagenomic classifiers, that report taxonomic relative abundances (Ye et al., 2019), such as Kraken2 (Wood et al., 2019; Wood and Salzberg, 2014) and the Megan Alignment Tool (MALT, specifically designed for ancient DNA) (Vågene et al., 2018), as well as species identification pipelines, that classify individual reads, such as Sigma (Ahn et al., 2015), Strain Prediction and Analysis using Representative Sequences (SPARSE) (Zhou et al., 2018), and Heuristic Operations for Pathogen Screening (HOPS) (Hübler et al., 2019). The metagenomic classifiers Kraken2 and MALT provide a good starting point for exploratory studies, especially when it is not known which pathogens may be present (Naccache et al., 2014). These tools, however, can be slow, require a large amount of RAM (Hübler et al., 2019; Vågene et al., 2018), and are prone to false positive identifications of low abundance species (Velsko et al., 2018). The species identification pipelines Sigma (Ahn et al., 2015) and SPARSE (Zhou et al., 2018) are based on more sophisticated statistical models but they require alignment to a large database, which can be slow (Zhou et al., 2018), and they do not provide statistical support for individual species identifications.

To address these issues, we developed **High-Accuracy** and **Scalable Taxonomic Assignment** of **Metagenomic** data (**HAYSTAC** - <https://github.com/antonisdim/HAYSTAC>), a light-weight, fast, and user-friendly species identification tool. HAYSTAC evaluates the presence of a particular species of interest in a metagenomic sample, and provides statistical support for the species assignment.

Our method is designed to estimate the probability that a specific taxon is present in a metagenomic sample given a set of sequencing reads and a database of reference genomes. We first derive the probability that the set of reads has originated from a single species (single source identification). This is useful, for example, for taxonomic identification of unknown source material (e.g. (Prendergast et al., 2017)). We then expand this method to identify multiple species in a metagenomic sample (multiple source identification).

To apply HAYSTAC, the user begins by constructing a database from publicly available genomes via a user-friendly automated process. Sample reads are then mapped to all reference genomes in the database using a Bowtie2 wrapper (Longmead and Salzberg, 2012). HAYSTAC then applies a novel Bayesian framework to infer abundance and statistical support for each species identification, and provides a per-read classification, allowing for both hypothesis-driven and exploratory analyses. HAYSTAC can build and make use of arbitrarily sized databases, with minimal effects on performance accuracy. This means that HAYSTAC can provide reliable species identifications by aligning reads to a handful of reference genomes (instead of thousands), without misidentifying taxa because of the restricted database size. This feature dramatically improves its performance relative to other methods. Altogether, HAYSTAC provides a robust method to assess the presence of a specific species in a metagenomic sample in a single step without the need to run multiple programs to assess assignment accuracy, such as is

required by other methods (i.e., Bracken for Kraken2 (Lu et al., 2017) and HOPS for MALT).

To evaluate HAYSTAC's performance, we applied it to both simulated microbiome datasets and to empirical datasets that were generated from DNA extracts derived from archaeological human bone. We further used HAYSTAC to assess the presence of putative respiratory pathogens in oral microbiome-derived metagenomes from ancient dental calculus. While dental plaque is known to harbor respiratory pathogens (Sumi et al., 2007), few studies to date have reported these species in its mineralized analogue, dental calculus (Eerkens et al., 2018; Warinner et al., 2014), a discrepancy that may reflect identification biases in studies of the latter. Improved identification of ancient respiratory pathobionts (host-associated microbial taxa that may cause disease under specific conditions) has the potential to add to our understanding of pathogen evolutionary dynamics and even aid vaccine development.

Results

Method Overview

HAYSTAC is a metagenomic species identifier that works with arbitrary user defined databases. The database can be built from any combination of an NCBI search query, a user specified list of NCBI accession codes, the RefSeq representative database of prokaryotic species, or a list of user provided reference assemblies. To construct the database, HAYSTAC uses a heuristic method to select a single representative genome per taxon from those available in the input data.

A database-wide index is then built, followed by a competitive alignment (very fast local alignment mode in Bowtie2, with default parameters) of each sample against the full database. This step ensures that reads with no matches in the database are excluded in all subsequent analyses (“dark matter” category; Figure 1).

Reads passing this initial inclusion filter are then aligned against every individual genome in the database (with Bowtie2 in end-to-end mode). For each alignment, we count the number of transitions (Ts) and transversions (Tv), as each type of mismatch contributes differently to our statistical model in order to handle ancient DNA damage. We then compute a posterior probability for each read/taxon alignment pair, using both mismatch count and expected mismatch probability (σ , default of 0.05). Accuracy can

be improved by adjusting the mismatch probability (σ) using prior information on intra- and interspecific variability in the specific set of species analysed.

Reads that possess a posterior probability exceeding a user-defined threshold (default of 0.75, Supplemental Figure 4 and 5) are then assigned to a single taxon. Reads with a lower posterior probability are assigned to the “grey matter” category. For each taxon, the count of reads passing the posterior probability threshold are used to compute the mean posterior abundance (and 95% confidence interval), i.e. the relative abundance of a taxon in a sample, using a Dirichlet distribution of the sample input data as a whole. A final validation step, for low depth sequencing data, using breadth of coverage is then used to assess if the assigned reads represent a random genome sample or cluster around specific genome areas.



Figure 1. HAYSTAC's workflow. HAYSTAC consists of three main modules: 1) DATABASE, which builds a database of reference genomes from various user input sources 2) SAMPLE, which handles downloading sequencing files from the SRA and pre-processing of samples prior to analysis, and 3) ANALYSE, which performs an

analysis of a sample against a database by applying the mathematical model (see methods) for taxonomic abundance estimation.

Computational Performance

We assessed the computational performance of HAYSTAC relative to Sigma, Kraken2/Bracken and MALT/HOPS by measuring peak memory usage and elapsed execution time for each software package on a machine with 16 CPU cores and 258 GB of RAM under controlled conditions.

We first assessed performance when building databases containing 10, 100 and 500 species. For all tested sizes, HAYSTAC ran faster and used less memory than both MALT/HOPS and Kracken2/Bracken, but was slower and used more memory than Sigma, which does not require large indices for its alignments (Figure 2; Supplemental Table S9).

We then assessed the performance of these methods when analysing a one million read dataset against the same three database sizes. HAYSTAC runs faster than Sigma for large databases, uses less memory than Kraken/Braken, and is both faster and uses less memory than MALT/HOPS for all database sizes (Figure 2; Supplemental Table S10).

Lastly, we assessed the performance of these methods when analysing datasets of 10 thousand, 100 thousand and 1 million reads, against a database of 500 species (Supplemental Figure 1; Supplemental Table S10). Memory performance was unchanged from the earlier benchmarks, showing that for all methods, memory usage scales with database size rather than sample size. With the exception of Kraken/Bracken2, all other methods saw only a modest increase in runtime with larger input sizes, showing that database size also has a substantial impact on overall runtime.

To corroborate our interpretation of the computational performance results for each method, we calculated the R^2 values and p-values for each of the fitted lines. All of the p-values were below the 5% threshold of significance, and the majority of the R^2 values also indicated a strong correlation between the number of species present in the database and the elapsed computational time (Table 1). These statistical findings further support our interpretation of the computational performance results.

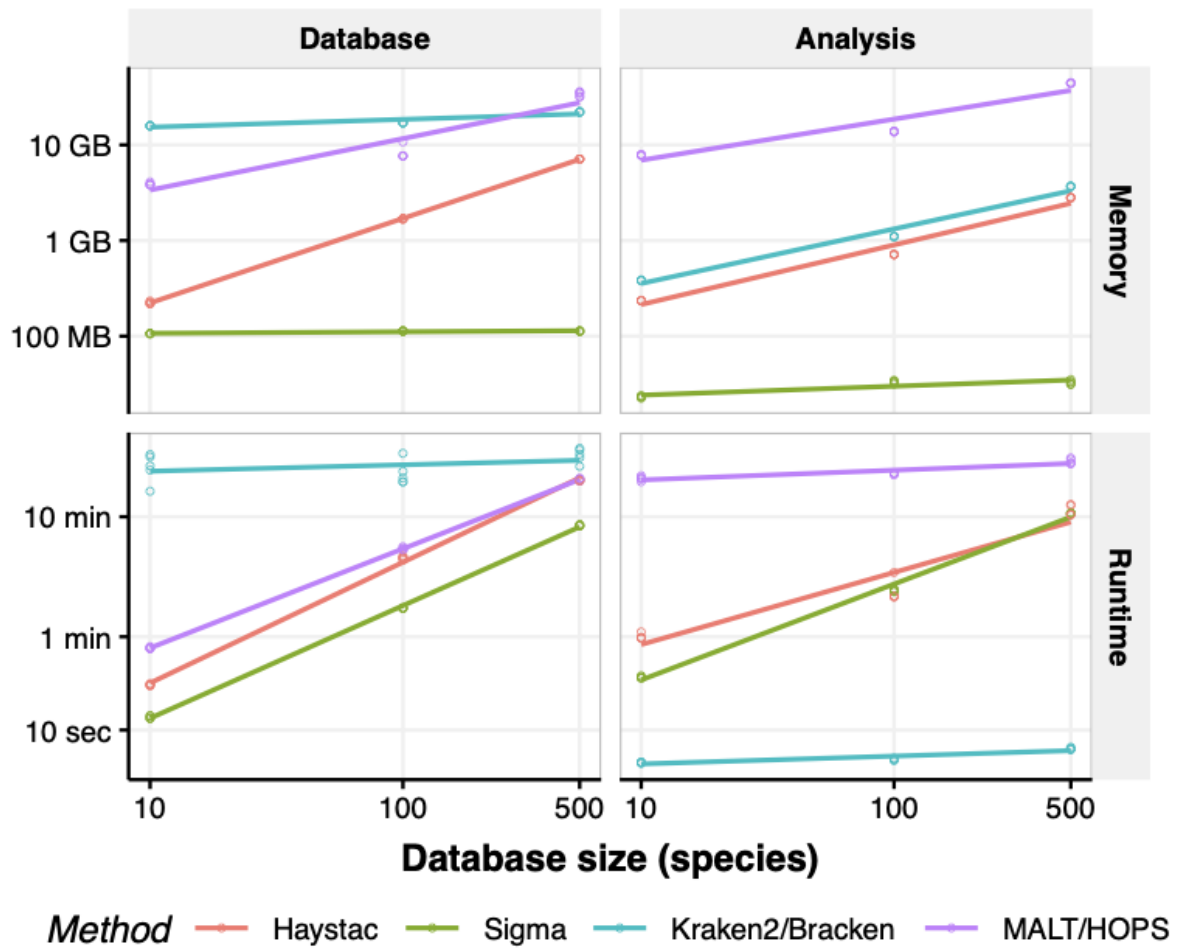


Figure 2. Computational performance of HAYSTAC and other methods. Linear regression of the elapsed time (wall clock) and peak memory usage (maximum resident set size) for a sample of size 1 million reads and reference databases containing either 10, 100 or 500 genomes.

Table 1. R-squared and p-value metrics from the linear model that was fitted on the computational performance results of each of the four tested methods. All the p-values are below 5%, with most of them being highly significant. The majority of the R^2 values also indicate a high correlation between the database size and computational time.

		Database		Analysis	
		Memory	Runtime	Memory	Runtime
Haystac	R^2	0.9991	0.9989	1	0.9768
	p-value	2.20E-16	2.20E-16	2.20E-16	5.26E-12
Sigma	R^2	0.4013	0.9998	0.3635	0.9999
	p-value	0.01123	2.20E-16	0.01736	2.20E-16
Kraken2/ Bracken	R^2	0.9999	0.2833	0.9987	0.9844
	p-value	2.20E-16	0.04113	2.20E-16	3.93E-13
MALT/ HOPS	R^2	0.99	0.9966	0.9994	0.9312
	p-value	2.12E-14	2.20E-16	2.20E-16	6.20E-09

Analyses of simulated data sets using RefSeq prokaryotic database

To characterise the sensitivity and specificity of HAYSTAC, we compared its performance to that of three other widely used metagenomic profiling tools: Sigma, Kraken2/Bracken and MALT/HOPS. We simulated four datasets for these comparisons (Table 2), and then profiled each dataset with all four tools. Each simulated dataset was designed to test a different aspect of performance. For each simulated dataset profiled by each method, we computed: (i) false positive count (total number of species that were reported but not present in the simulated sample); (ii) false negative count (total

number of species present in the simulation that were not identified); and (iii) true positive count (total number of species correctly identified). For a positive identification for HAYSTAC, Sigma and Kraken2/Bracken a relative abundance higher than 0.01% was required, and for the MALT/HOPS pipeline a positive identification required a minimum of 50 reads assigned to a species.

Table 2. Characteristics of the 6 simulated sample sets of the Simple, Random and Oral microbiome datasets.

Simulation set	Dataset	Number of replicates	aDNA damage	Mean number of species
Simple with human DNA	Simple Microbiome	2	No	10
Simple without human DNA	Simple Microbiome	2	No	10
Microbiome 100 species ancient	General Microbiome	2	Yes	100
Microbiome 100 species modern	General Microbiome	2	No	100
Microbiome 500 species	General	2	Yes	500

ancient	Microbiome			
Microbiome 500 species modern	General Microbiome	2	No	500
Oral microbiome aDNA damage	Oral Microbiome	6	Yes	178
Oral microbiome modern	Oral Microbiome	6	No	178

We first generated a dataset of samples comprising 10 species to quantify the ability of each program to accurately detect species in a simple metagenomic community (hereafter “Simple Microbiome” dataset). We generated two additional sets of profiles for this dataset: one that included modern human DNA (“contaminated”) and one that did not (“non-contaminated”). This dataset was created using gargammel (Renaud et al., 2017) by randomly sampling 10 species from a modified RefSeq prokaryote representative genome database (see Methods) and building simulated libraries of one million single-end reads with an equal number of reads per species (100K), both with and without contamination from the human reference genome (25% of the total simulated reads).

All tools correctly identified the 10 species in each simulated sample set (both with and without human contamination; Figure 3). Only two programs reported false positives in

the non-contaminated sample set: Kraken2/Bracken (average of three species) and MALT/HOPS (average of 0.5 species) (see methods for the definition of false positive identifications). False positive rates, however, dramatically increased in the sample set with human contamination. HAYSTAC reported the lowest false positive count (5) relative to Sigma (6), Kraken2/Bracken (12.5), and MALT/HOPS (11.5) (Figure 3). This suggests that HAYSTAC is less prone to false positive identifications caused by human contamination in bacterial reference genomes found in the RefSeq representative prokaryotic database. Across all tested programs for the contaminated simulation set, each false positive identification was to a species whose RefSeq genomes are known to contain human DNA contamination (Breitwieser et al., 2019) (Supplemental Table S7). This result strongly supports the prefiltering and removal of all human reads prior to microbial metagenomic profiling.

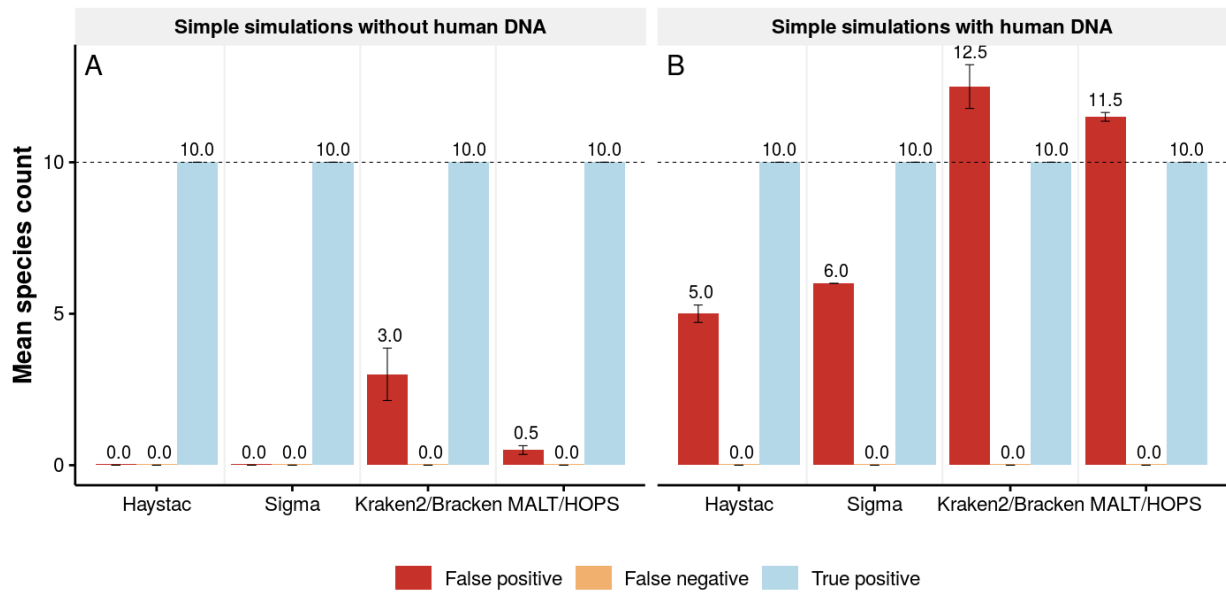


Figure 3. Accuracy of HAYSTAC and other methods for a simple simulation. Bar plot showing the mean count of false positives (red), false negatives (orange), and true detected species (blue) for two versions of the simple simulations dataset, each with two replicates: (A) without human DNA contamination (n=2); and (B) with human DNA contamination (n=2). The dotted line shows the number of simulated species in each set of samples (i.e. the maximum true positive; n=10), and numbers above the error bars indicate the mean species count in each category.

We then tested HAYSTAC's performance when profiling samples with damage patterns characteristic of aDNA (i.e., sequence fragmentation and cytosine deamination). To do so, we generated eight simulated samples consisting of 10 million reads each, both with and without aDNA damage (hereafter "General Microbiome" dataset). Each sample contained simulated reads from either 100 species (four simulated samples; two

random sets of species all with equal abundance) or 500 species (four samples; two random sets of species all with equal abundance) that were randomly sampled from the RefSeq representative prokaryotic database (Table 2).

Overall, the incidence of false negatives was low. HAYSTAC reported 2.5 false negative species on average in the 100 species sample set, and 15.0 in the 500 species sample set. Sigma reported 1.5 and 14.0, Kraken2/Bracken 2.3 and 22.0 and MALT/HOPS 7.8 and 38.0, respectively (Figure 4 and Supplemental Figure 2 panels A&B). False positives, in contrast, were variable between methods. HAYSTAC reported an average of only 0.5 false positives in the 100 species sample set, and an average of 1.0 false positives in the 500 species sample set. Sigma performed similarly, with an average of 0.5 and 1.0 false positives, while Kraken2/Bracken reported 11.5 and 17.0 false positives, and MALT/HOPS reported 27.5 and 25.5 false positives in the 100 and 500 species sample sets, respectively (Figure 4 panels A&B and Supplemental Table S3).

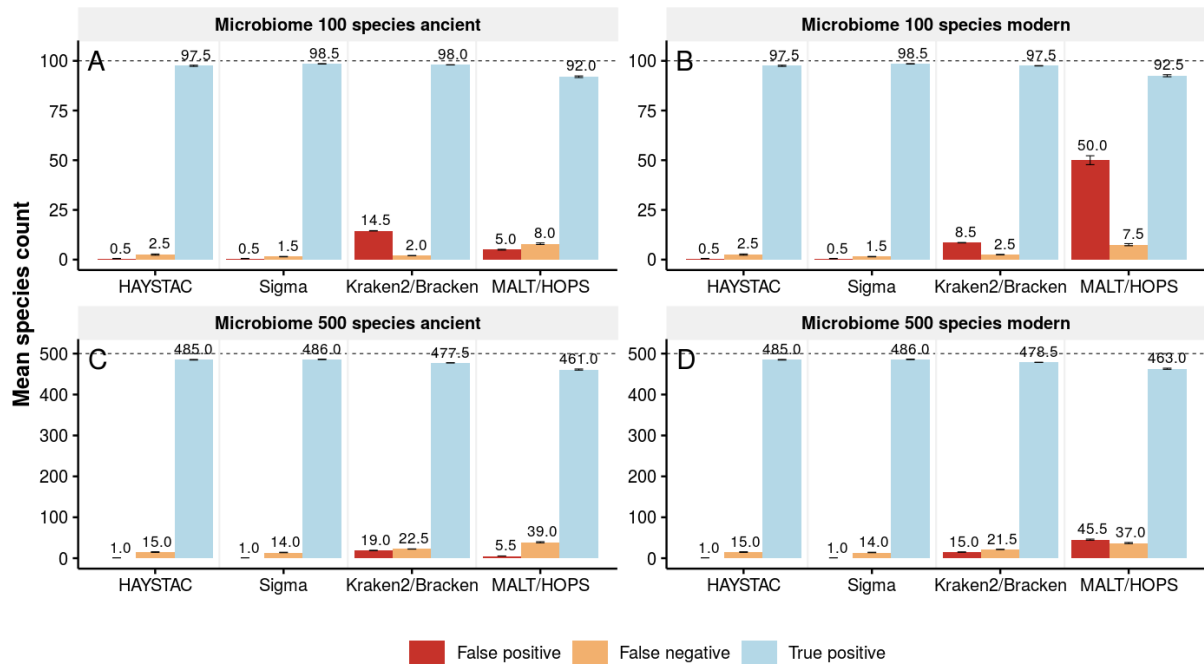


Figure 4. Accuracy of HAYSTAC and other methods for a general microbiome simulation. Mean count of false positive (red), false negative (orange), and true detected species (blue) in the simulated General Microbiome dataset of 100 species simulated as ancient samples (n=2) (A), 100 species simulated as modern samples (n=2) (B), and 500 species simulated as ancient samples (n=2) (C) and 500 species simulated as modern samples (n=2) (D). The dotted line represents the average number of simulated species in each set of samples, and the numbers above the error bars represent the mean species count.

We next performed more realistic simulations based on species compositions that mimic an oral microbial community from a previously analysed subgingival plaque sample

((Velsko et al., 2018); hereafter “Oral Microbiome” dataset). These simulations included 12 simulated samples of 5 million paired-end reads each, with and without aDNA damage, and with 176-180 species in varying or constant species abundance (Table 2). Only 115 of the 196 total Oral Microbiome species had taxonomic representation in the RefSeq representative prokaryotic database, and only four species in the Oral Microbiome dataset had genomes in the modified representative prokaryotic RefSeq database we used (Supplemental Table S6).

As expected, given the incompleteness of the database, false negatives were elevated overall (Table 3). HAYSTAC failed to identify on average 72.5 species in the ancient (damaged) Oral Microbiome dataset, and 70.5 species in the modern Oral Microbiome dataset. Sigma failed to identify on average 67.5 and 67.5 species, Kraken2/Bracken failed to identify 67.3 and 67.2 species, and MALT/HOPS failed to identify 69.0 and 68.3 species in the ancient and modern Oral Microbiome datasets, respectively (Figure 5 panels A&B and Supplemental Figure 3). False positives, however, were highly variable between methods. HAYSTAC reported an average of 23.3 false positives in the ancient Oral Microbiome dataset and 31.8 false positives in the modern Oral Microbiome dataset. Sigma reported 44.2 and 31.5 false positives, Kraken2/Bracken reported 174.2 and 227.8 false positives, and MALT/HOPS reported 115.3 and 325.8 false positives in the ancient and modern datasets, respectively (Figure 5 A&B and Supplemental Table S4). Overall HAYSTAC outperformed the other methods in the ancient Oral Microbiome sample set with the lowest number of false positive identifications (pairwise t-test

p-value between HAYSTAC and Sigma: 0.03, pairwise t-test p-value between HAYSTAC and Kraken2/Bracken: 7.1×10^{-8} , pairwise t-test p-value between HAYSTAC and MALT/HOPS: 0.0, with a p-value significance threshold of 0.05).

Table 3. False negative and false positive species counts in the ancient and modern Oral Microbiome datasets.

	Ancient Oral Microbiome		Modern Oral Microbiome	
	False negative	False positive	False negative	False positive
HAYSTAC	72.5	23.3	70.5	31.8
Sigma	67.5	44.2	67.5	31.5
Kraken2/Bracken	67.3	174.2	67.2	227.8
MALT/HOPS	69.0	115.3	68.3	325.8

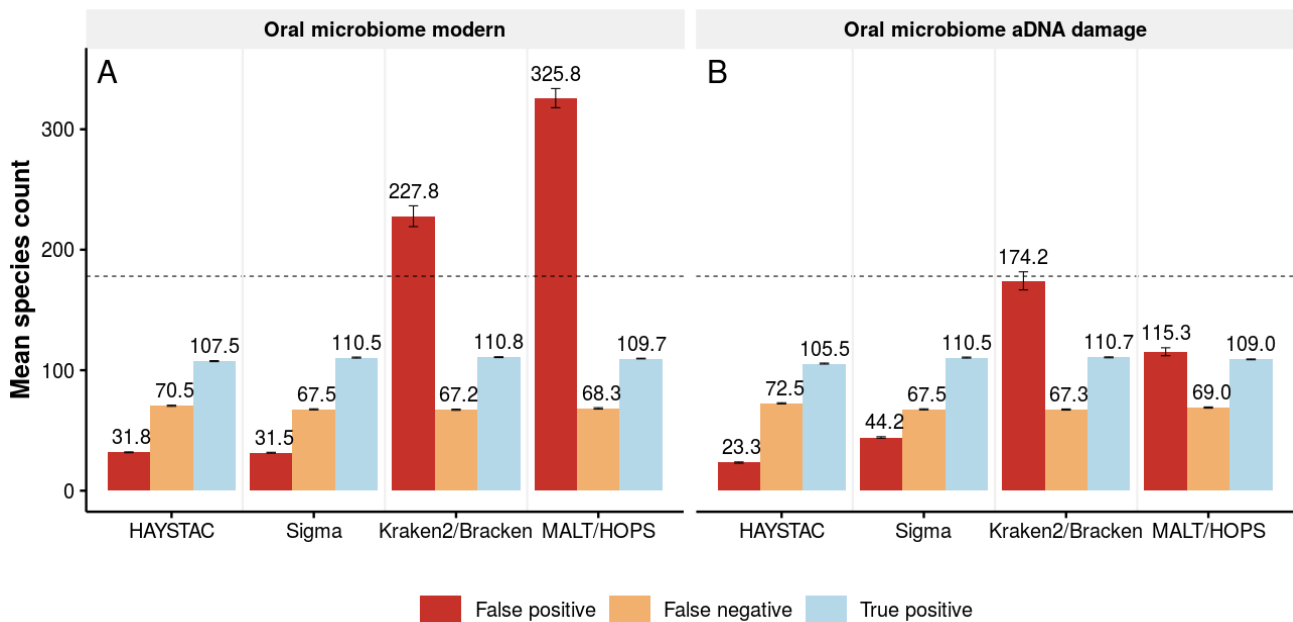


Figure 5. Accuracy of HAYSTAC and other methods for an oral microbiome simulation. Bar plot showing the mean count of false positives (red), false negatives (orange), and true detected species (blue) for two versions of the oral microbiome dataset, each with six replicates: (A) modern simulation, with fixed read lengths (n=6); and (B) ancient simulation, with variable read lengths and post-mortem damage (n=6). The dotted line shows the average number of simulated species in each set of samples (i.e. the maximum true positive; n=178), and numbers above the error bars indicate the mean species count in each category.

Analyses of simulated data sets using genus-level database

We next assessed how accurately HAYSTAC identified specific candidate species using smaller databases. To do so, we re-analysed 20 simulated samples included in the General and Oral Microbiome datasets (General Microbiome 100 species, General

Microbiome 500 species, Oral Microbiome aDNA damage, and Oral Microbiome modern datasets) (Table 2) using nine different genus-specific databases (see Methods). For each randomly selected genus, we constructed a database that included all genomes found in the RefSeq representative prokaryotic database. The nine genera that were selected for this analysis were: *Bacteroides*, *Burkholderia*, *Campylobacter*, *Clostridium*, *Corynebacterium*, *Desulfitobacterium*, *Mycobacterium*, *Solimonas*, and *Streptococcus*. Each database consisted of all RefSeq genomes from all species belonging to the genus instead of the full RefSeq representative prokaryotic database. This resulted in databases with 3-136 species each, as opposed to 5,636 species when using the full RefSeq representative prokaryotic database. For each analysis, we computed false positive, false negative, and true positive species count as before.

Overall, false negatives were low. HAYSTAC reported on average 0.2 species missing in the samples of the General and Oral Microbiome datasets, while Sigma reported 0.1, Kraken2/Bracken 0.1 and MALT/HOPS 0.1 species missing (Figure 6, Supplemental Table S5). False positives, however, were variable between methods. HAYSTAC reported an average of 0.8 false positives in the samples of the General and Oral Microbiome datasets, while Sigma reported 1.1, Kraken2/Bracken 33.1 and MALT/HOPS 3.6 species missing (Figure 6). We also used HAYSTAC to compare the mean posterior abundance computed using the full RefSeq database with abundance computed using the genus-specific databases (Figure 7). We found that while genus-specific abundances were slightly higher than those computed with the full

RefSeq database, there was a very strong correlation between the two ($R^2=0.985$; slope=1.1; p-value < $1-e6$; GLM: $y = 1.1x + 8.2 \cdot 10^{-6}$).

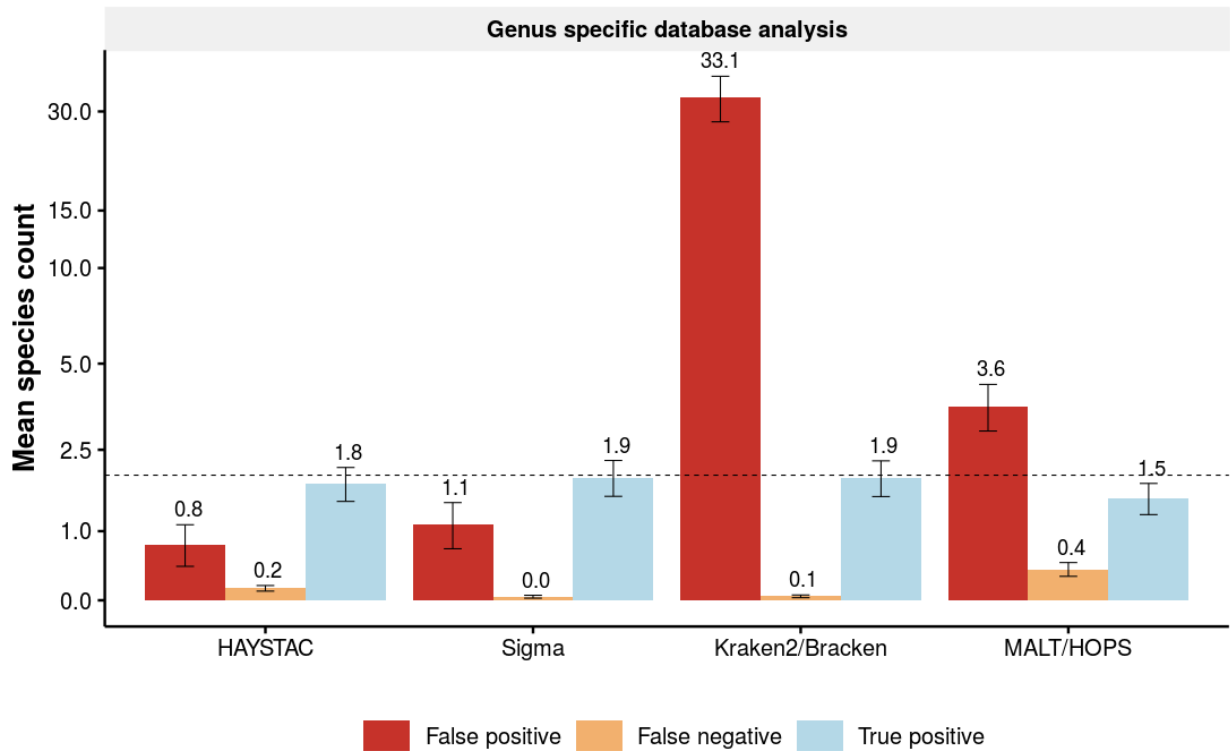


Figure 6. Accuracy of HAYSTAC and other methods using a reference database restricted to a single genus. Bar plot, with a pseudo-log10 transformed y-axis, showing the mean count of false positives (red), false negatives (orange), and true detected species (blue) for 9 different genera (*Bacteroides*, *Burkholderia*, *Campylobacter*, *Clostridium*, *Corynebacterium*, *Desulfitobacterium*, *Mycobacterium*, *Solimonas* and *Streptococcus*), each with 20 samples from the general and oral microbiome datasets. The dotted line shows the average number of simulated species in each set of samples (i.e. the maximum true positive; $n=1.9$), and numbers above the error bars indicate the mean species count in each category.

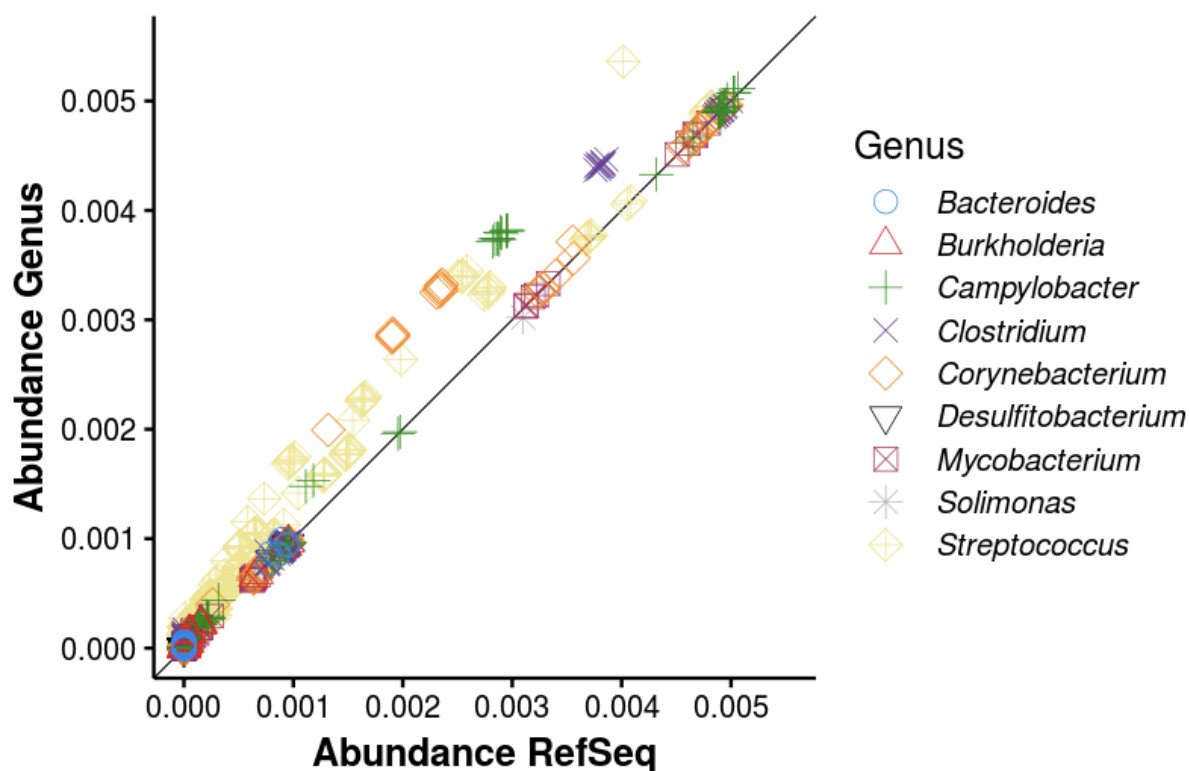


Figure 7. HAYSTAC inferred posterior abundance levels. Scatter plot showing the mean posterior abundances across all taxa ($n=355$) and samples ($n=20$) for either a genus specific database or the entire RefSeq representative database of prokaryotic species (plot truncated at $6e-3$). Using a genus specific database has a small positive bias in mean posterior abundance for taxa within that genus (paired t-test p-value = $2.2e-16$, mean of the differences = $1.4e-05$), nevertheless the overall abundance levels are highly correlated ($R^2=0.985$). Computational runtime for the genus specific analyses are approximately 350 times faster and use 5 times less memory, making genus specific analyses suitable for initial positive screening.

Case study 1: Ancient *Yersinia pestis*

We then tested HAYSTAC's performance to detect specific pathogens in real archaeological samples. We first analysed publicly available 130 plague-positive ancient human libraries (Rasmussen et al., 2015), obtained from seven bone and tooth samples (RISE00, RISE139, RISE386, RISE397, RISE505, RISE509 and RISE511).

HAYSTAC was run against a database consisting of the longest complete genomes for each of the species in the genus *Yersinia* (including plasmid sequences). Our method was able to confidently assign between 3,875 and 550,200 reads to *Y. pestis* across all seven samples (Figure 8). In addition, we found evidence for additional *Yersinia* species in all samples, including *Y. pseudotuberculosis* (311-44,818 reads), *Y. similis* (158-6,263 reads), *Y. canariae* (2-6,029 reads) and *Y. ruckeri* (117-566 reads). Of these, *Y. ruckeri* and *Y. similis* are known to cause infections in animals (specifically fish), but have also been reported to infect humans (De Keukeleire et al., 2014). These identifications, however, are likely surprising, since *Y. ruckeri* and *Y. similis* are phylogenetically basal to other *Yersinia* species (Chen et al., 2010; Reuter et al., 2014) and may be attracting reads belonging diverged sequences.

The posterior abundances for *Yersinia pestis* in the RISE samples ranged from 0.0054% - 0.66%, corresponding to between 3,875 - 550,200 assigned reads. Despite the large number of assigned reads, posterior abundances only exceeded 0.01% for five of the seven samples (RISE00, RISE139, RISE386, RISE505 and RISE511). This

highlights the need to consider both relative and absolute abundance when making positive identifications.

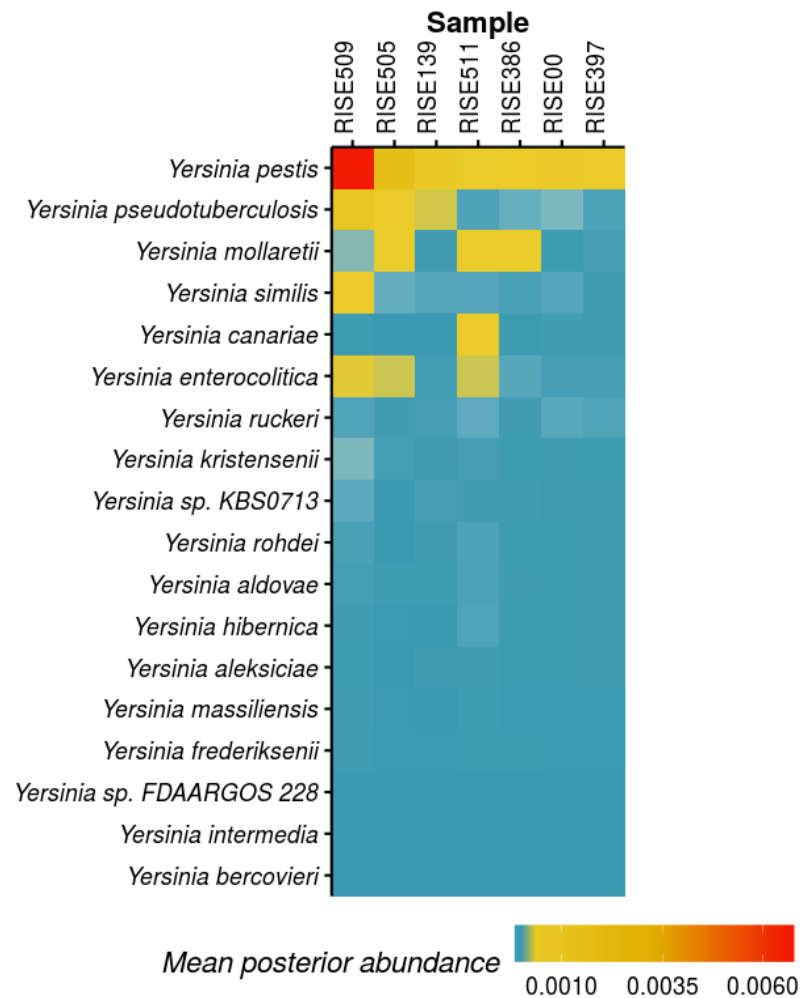


Figure 8. Posterior abundances of *Yersinia* species in Case Study 1. Heatmap showing the mean posterior abundances for the 7 RISE samples, based on a genus specific analysis of 18 *Yersinia* species.

Case study 2: Historic dental calculus

Dental plaque is a biofilm that develops naturally on teeth, and during life it periodically mineralizes to form dental calculus (Warinner et al., 2015), a robust substrate that can preserve biomolecules such as DNA from the biofilm and the host (Warinner et al., 2014). Although this substrate can act as a reservoir for respiratory pathobionts, analysis of ancient dental calculus metagenomic data is often challenging because the low abundance of these species makes it difficult to confidently identify them (Warinner et al., 2014). To test the performance of HAYSTAC on pathobiont aDNA sequences within host-associated ancient microbiomes, we analysed data from 48 dental calculus samples from a 19th century hospital in Oxford (Loe et al., 2014; Velsko et al., 2019).

We screened these libraries for potential upper respiratory pathobionts from the *Corynebacterium*, *Streptococcus*, *Klebsiella* and *Bordetella* genera (Figure 9), in an effort to test if such pathobionts could be preserved and identified in dental calculus. Members of the genera *Corynebacterium*, *Streptococcus* and *Bordetella* had also been included in the Oral Microbiome dataset as well. Within the *Corynebacterium* genus, we identified *Corynebacterium matruchotii* (abundance 0.01-1.63%; 36 samples) and *C. durum* (abundance 0.01-0.52%; 15 samples) both of which are oral commensal taxa. We did not, however, identify *Corynebacterium diphtheriae*, the species that causes diphtheria. We identified multiple species of *Streptococcus*, with the commensal *Streptococcus sanguinis* being identified in 33 samples at relatively high abundance (0.01-2.15%). The pathobiont *Streptococcus pneumoniae* was also positively identified

at low abundance (0.01-0.07%) in seven samples. We did not find any evidence for species belonging to the genera *Klebsiella* and *Bordetella*.

Members of the *Haemophilus* genus were rarely identified, which is expected given that many species from that genus are generally low-abundance in dental plaque compared to other oral surfaces such as the tongue (Mark Welch et al., 2016), and are similarly low-abundance in dental calculus (Velsko et al., 2019). However, *Haemophilus parainfluenzae* was found in 18 samples (abundance between 0.01-0.43%), *H. haemolyticus* was identified in two samples (abundance 0.01-0.03%), and *H. parahaemolyticus* was identified only in sample CS18 (abundance 0.02%). An uncharacterized oral species, *Haemophilus* oral taxon 036, was identified in two samples (abundance 0.01-0.04%) and the commensal *H. sputorum* was also identified in two samples (abundance 0.01-0.03%).

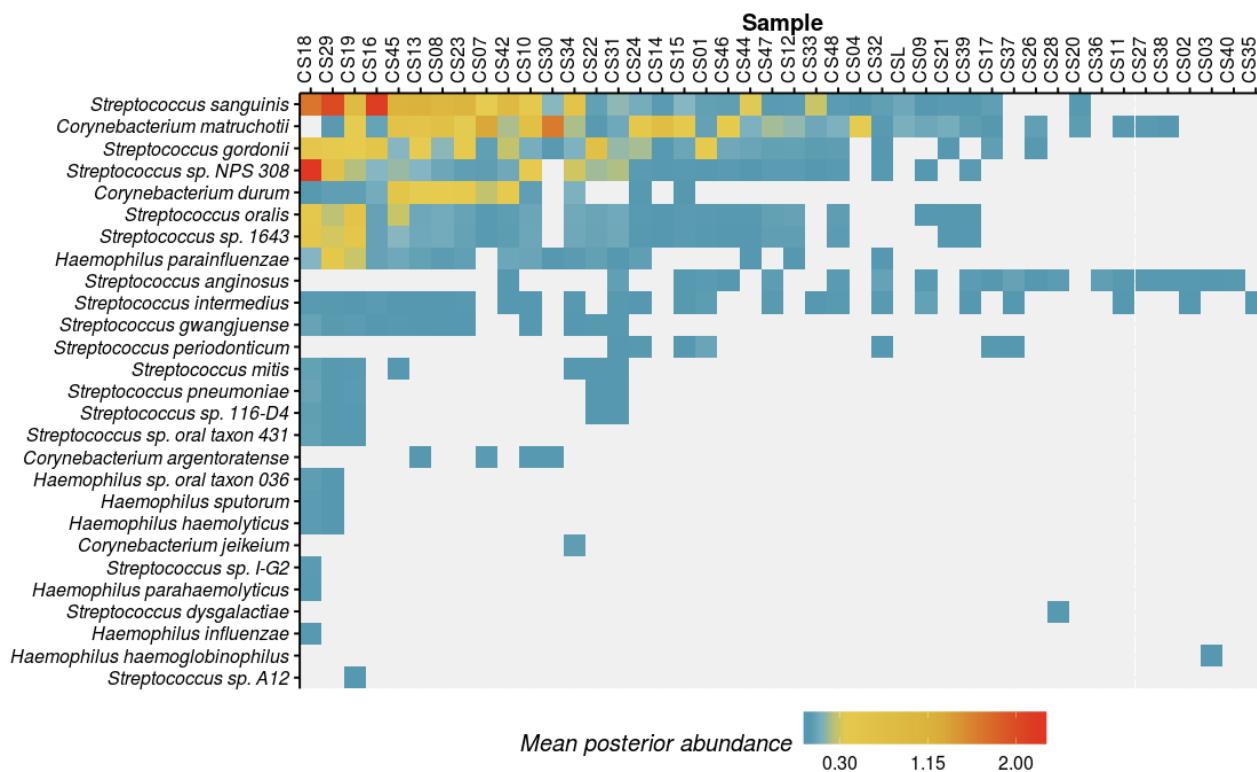


Figure 9. Posterior abundances of oral microbiome species in Case Study 2.

Heatmap showing the mean posterior abundances for the 44 dental calculus samples, based on a custom database that combined the prokaryotic representative RefSeq and pathobionts with complete genomes from the following 5 genera: *Corynebacterium*, *Haemophilus*, *Klebsiella*, *Streptococcus*, and *Bordetella*.

As expected, our analysis indicates that while species naturally occurring in the oral cavity (e.g. *C. matruchotii* and *durum*), were detected in most samples, upper respiratory system pathobionts (e.g. *S. pneumoniae*, *H. parainfluenzae*) were more rare.

Although pathogens were found generally at very low abundance, in only a few individuals, our results indicate that HAYSTAC was able to positively identify some pathogenic species such as *Streptococcus pneumoniae* and *Haemophilus parainfluenzae*, suggesting that dental calculus is a promising substrate to study ancient respiratory pathobionts.

Discussion

Our results demonstrate that HAYSTAC provides a robust and computationally efficient framework to statistically assess the presence of specific species in a metagenomic sample. HAYSTAC identified fewer false positives than both Kraken/Braken and MALT/HOPS in all simulations, and fewer than Sigma in simulations of ancient data. It runs faster than Sigma for large databases, uses less memory than Kraken/Braken, and is both faster and uses less memory than MALT/HOPS. Furthermore, our results show that the use of reduced reference databases (e.g. a few selected genera as opposed to the entire RefSeq representative prokaryotic database), can dramatically reduce computational time (approximately 350 times faster) while still providing highly accurate species identification. Altogether, this makes HAYSTAC a flexible tool to rapidly and accurately screen metagenomic samples, a feature that is attractive for a large range of applications, including archaeogenetics and biosurveillance.

Species concept, horizontal gene transfer and species identification

Although we demonstrated that both HAYSTAC and Sigma provide more robust species assignments than MALT/HOPS or Kraken2/Bracken, their internal computation is based on the assumption that intraspecific polymorphism is lower than interspecific polymorphism. While this is true in most cases, the concept of a species among microorganisms is often not straightforward. This is the case, for example, in the *Mycobacterium tuberculosis* complex in which interspecific variability is very low (Tortoli, 2012), and occasionally lower than the default value of expected mismatch (i.e.,

expected level of intraspecific polymorphism) set in HAYSTAC ($\sigma=0.05$; see eq. 4 below). HAYSTAC deals with this issue by introducing a “grey matter” category, to which reads that are equally likely to match two or more genomes are assigned. This ensures that they do not contribute to posterior abundance computations, which could possibly force false identifications, while also not dismissing them into “dark matter” with reads that cannot be classified. In cases where the levels of intra- and interspecific diversity are known (e.g., within a genus) it is possible to adjust the probability of the expected mismatches between a read and a genome (σ ; see eq. 4), which can increase power while reducing false positive rates when analysing a specific dataset. More specifically, if one is examining a set of closely related species for which a measure of nucleotide diversity can be obtained at the genus level, then the user can set σ to reflect the value of this measure (e.g. by setting $\sigma=\pi$). Concurrently, if a user does not have an estimate of the genetic variation among the taxa included in a database, but they are considering analysing closely related taxa or taxa within a genus with low mutation rate, they might still want to modify the default value of σ , so that the genomes in their database are considered as more similar to each other by HAYSTAC. In practice the user can implement this by setting a more stringent σ value (i.e. set a lower σ value than 0.05) when working with highly similar genomes/closely related species. On the other hand if a user is expecting that the taxa present in their sequencing library are quite divergent from the respective reference genomes in their database, for instance when analysing aDNA data, they can increase the value of σ in order to allow for more mismatches.

This feature means HAYSTAC's performance can be fine-tuned to specific questions and datasets.

Identifying species within a metagenomic sample can also be complicated by horizontal gene transfer (HGT). HGT can lead to surprising, highly supported, species assignments - especially if the sequences originate from an outgroup species that is not represented in the database. HAYSTAC mitigates this issue by inspecting evenness of coverage across the genome for shallow sequencing data. Specifically, reads that belong to an outgroup species and that map to a genome in a reference database because of an HGT or contamination event are expected to map only to parts of the reference genome that resulted from such an event. HAYSTAC assesses whether reads are clustered more than expected by chance across a reference genome using a simple chi square contingency test (see Methods). This provides the user with a quantification of the evenness of spread, and thus a proxy signature of potential HGT or misassignment of reads from a taxon that is not represented in the reference database.

Issues with databases

Another issue facing metagenomics is contaminated sequences incorporated into reference genomes. For example, multiple bacterial RefSeq representative assemblies contain human sequences (Breitwieser et al., 2019). This can lead to spurious, highly supported, species assignments when profiling metagenome samples. Including genome sequences from potential contaminant sources (e.g., the human reference

genome) in the database should alleviate these issues (Feuerborn et al., 2020). Alternatively, it is also possible to filter reads before starting a taxonomic identification/assignment analysis (i.e. by excluding those that map to the contaminant's reference genome). Potential contaminants, however, are not always known, and there are many gaps in reference sequence databases. In fact, inadequate species representation in databases is likely to be an important source of false positives for all species identification methods (Warinner et al., 2017).

Although reads belonging to a species that is not represented in a database are typically assigned to a higher taxonomic node ("grey matter" category in HAYSTAC), incorrect species assignment can occur if the missing taxon is closely related to another species in the database, or if the read derives from a conserved element found within a taxonomic group with sparse representation in the database. Although this is not considered a major problem for general profilers, as closely related species are likely to play a similar function in the ecosystem, such misassignments can greatly complicate the identification of specific pathogen species, especially if they have close environmental relatives.

The "Grey Matter" category is an absorbent taxonomic category, used by HAYSTAC for the assignment of reads that cannot be uniquely assigned to a single taxon. And even though it absorbs all the non-informative reads from a sample, it also includes DNA reads that could have been placed on higher taxonomic levels, such as the genus level.

Unfortunately, that translates into losing information regarding the abundance and composition of a sample on the higher taxonomic nodes. Other methods, e.g. MEGAN (Huson et al., 2007) employ an approach known as the Lowest Common Ancestor (LCA). An LCA algorithm assigns every individual read with a match to a genome in a database, to some taxon in the NCBI taxonomy, regardless of the read's suitability as a phylogenetic marker (Huson et al., 2007). Such an algorithm allows for reads to be assigned to high-order taxa such as genera as well, providing abundance results for each level of the NCBI taxonomy tree. As HAYSTAC's current conservative approach only utilises reads that can be assigned uniquely to a species or subspecies, it disregards the informational content of reads that could be assigned to higher-order taxa. Thus, for a future version of HAYSTAC we will be incorporating an algorithm similar to an LCA approach, to enable the evaluation of the taxonomic composition and estimation of the posterior abundance of higher-order taxa from reads that have been assigned to the "Grey Matter" category.

Conclusions

Metagenomic data are now routinely produced and analysed, providing new insights into microbial communities that were previously unknown. Development of metagenomic analytical tools, however, has mostly focused on frameworks that can profile an entire microbial community. We showed that these methods, however, can produce high levels of false positives when focusing on specific species identification. Here we present a tool that can robustly assess whether a specific species is present in a metagenomic sample in one step, without the need to combine different pipelines to validate the results. HAYSTAC provides a user-friendly and automated method for rapidly constructing databases, and produces robust identifications regardless of the database size, allowing for both rapid hypothesis driven analyses and exhaustive profiling. Because of the mathematical framework it employs, HAYSTAC reliably produces the lowest number of false positive identifications, making it a valuable tool for both ancient and modern DNA microbial research.

Methods

Single Source Identification

We collect a series of observations (reads: $r_1 \dots r_n$) and we denote the full set of observations as R .

$$R = \{r_1 \dots r_n\}$$

We map all the reads (R) to a database of reference genomes and we define the event of sampling from the i -th reference genome in our database as G_i . Assuming that all reads comes from the same source (i.e., a single species represented in the database), the events of sampling from a given genome are exhaustive and mutually exclusive, i.e.

$$G_1 \dots G_n | G_i \cap G_j = \emptyset, \forall i \neq j \text{ and } \bigcup_i^n G_i = \Omega$$

From Bayes Theorem we have:

$$P(G_j|R) = \frac{P(R|G_j)P(G_j)}{P(R)} = \frac{P(R|G_j)P(G_j)}{\sum_k P(R|G_k)P(G_k)} \quad (1)$$

Given that each read is a different realization of a sampling process (i.e., an independent observation), we can express the conditional probability $P(R|G_j)$ as a product of likelihoods:

$$P(R|G_j) = \prod_i P(r_i|G_j) \quad (2)$$

Therefore, by substituting equation 2 in equation 1, we can express the posterior probability of sampling from a given genome G_j as:

$$P(G_j|R) = \frac{P(G_j) \prod_i P(r_i|G_j)}{\sum_k P(G_k) \prod_i P(r_i|G_k)} \quad (3)$$

Likelihood

We next derive a likelihood function to express the probability of a read given a genome $P(r_i|G_j)$. To do so, we use a previously published expression of this likelihood (Ahn et al., 2015) based on the number of mismatches between a read and a reference to which it is aligned:

$$\begin{aligned} P(r_i|G_j) &= \sigma^z (1 - \sigma)^{l-z}; & z \leq U \\ P(r_i|G_j) &= 0; & z > U \end{aligned}$$

where z is the number of mismatches between read r_i and genome G_j , U is the maximum number of mismatches allowed in the alignment and l is the length of read r_i (in bp).

The method assumes a uniform probability of mismatch within a given alignment (σ) to account for population level variability and sequencing errors. We expand this framework to leverage the fact that transition and transversion occur at significantly different rates, particularly in ancient samples due to the effect of deaminations.

For each alignment we can then express the likelihood of generating a read r_i given that it has been sampled from the reference genome G_j based on both number of transition (t) and transversion (v).

Let σ_t be the probability of observing a transition and σ_v the probability of observing a transversion, we can express the likelihood for a single read as:

$$P(r_i|G_j) = \sigma_v^{v_{i|j}} \sigma_t^{t_{i|j}} (1 - \sigma_v - \sigma_t)^{l_i - v_{i|j} - t_{i|j}} \quad (4)$$

where l_i represent the length of the read r_i , while $v_{i|j}$ and $t_{i|j}$ denote the number of transversion and transition between the read r_i and the genome G_j respectively.

Posterior

By substituting the likelihood term (eq.4) into equation 3, we can obtain an explicit form for the posterior probability of sampling a set of reads (R) from a given reference genome:

$$P(G_j|R) = \frac{\sigma_t \sum_i t_{i|j} \sigma_v \sum_i v_{i|j} (1 - \sigma_v - \sigma_t) \sum_i (l_i - t_{i|j} - v_{i|j}) P(G_j)}{\sum_{k=1}^{k=n} P(G_k) \sigma_t \sum_i t_{i|k} \sigma_v \sum_i v_{i|k} (1 - \sigma_v - \sigma_t) \sum_i (l_i - t_{i|k} - v_{i|k})} \quad (5)$$

By rescaling both numerator and denominator by $(1 - \sigma_v - \sigma_t)^{\sum_i l_i}$ we get:

$$P(G_j|R) = \frac{P(G_j) \left(\frac{\sigma_t}{1 - \sigma_v - \sigma_t} \right)^{\sum_i t_{i|j}} \left(\frac{\sigma_v}{1 - \sigma_v - \sigma_t} \right)^{\sum_i v_{i|j}}}{\sum_{k=1}^{k=n} P(G_k) \left(\frac{\sigma_t}{1 - \sigma_v - \sigma_t} \right)^{\sum_i t_{i|k}} \left(\frac{\sigma_v}{1 - \sigma_v - \sigma_t} \right)^{\sum_i v_{i|k}}} \quad (6)$$

Let $V_j = \sum_i v_{i|j}$ denote the total number of observed transversion between all reads and

genome G_j while $T_j = \sum_i t_{i|j}$ represents the total number of transitions. Let $\delta_v = \frac{\sigma_v}{1 - \sigma_v - \sigma_t}$

and $\delta_t = \frac{\sigma_t}{1 - \sigma_v - \sigma_t}$ denote instead the transversion to match ratio and the transition to

match ratio, respectively. By introducing this more compact notation, the previous equation can be re-written as:

$$P(G_j|R) = \frac{P(G_j) \delta_t^{T_j} \delta_v^{V_j}}{\sum_k P(G_k) \delta_t^{T_k} \delta_v^{V_k}} \quad (7)$$

Numerical Representation

Our method assumes a fixed probability of mismatches $(\sigma_v + \sigma_t)$ which can be specified by the user and has a default value of 5%. We estimate the transition transversion ratio

based on the total number of mismatches per categories observed in the entire dataset and we obtain the value for σ_v and σ_t by solving the following linear system:

$$\begin{cases} \sigma_t + \sigma_v = 0.05 \\ \sigma_t = \frac{T}{V} \sigma_v \end{cases} \quad (8)$$

where $V = \sum_j \sum_i v_{i|j} = \sum_j V_j$ and $T = \sum_j \sum_i t_{i|j} = \sum_j T_j$

If we assume a uniform prior in which each genome has a probability of $1/n$ we can simplify eq.7 even further obtaining:

$$P(G_j|R) = \frac{\delta_t^{T_j} \delta_v^{V_j}}{\sum_k^n \delta_t^{T_k} \delta_v^{V_k}} \quad (9)$$

When analysing a large number of reads, evaluating eq.9 might become problematic due to numerical representation issues. To accommodate for this we use an equivalent expression of the posterior probability:

$$P(G_j|R) = \frac{1}{1 + \sum_{k \neq j}^n \delta_t^{T_k - T_j} \delta_v^{V_k - V_j}} \quad (10)$$

Each term in the summation at the denominator of eq.10 represents the likelihood of genome G_k relative to genome G_j . There might be cases in which even the relative likelihood is so close to zero that cannot be represented numerically. We then take a

conservative approach and assign to $\delta_t^{T_k - T_j} \delta_v^{V_k - V_j}$ the smallest number that the machine can represent (for a 64bit system it should be 2.2250738585072014e-308 which we roughly approximate with 10^{-300}).

Multiple Sources Identification

We then expand this method to identify multiple species from metagenomic data (multiple source identification). The multiple-source identification approach involves the following steps:

- Assigning each read to a reference genome
- Obtain the likelihood for the multinomial distribution of abundances
- Obtain the posterior distribution from the Dirichlet distribution
- Calculate the 95% credible interval around each posterior mean
- Assess significance

Operatively we apply eq.9 to each read separately and obtain a posterior probability for a read to originate from a given genome:

$$P(G_j | r_i) = \frac{\delta_t^{t_{i|j}} \delta_v^{v_{i|j}}}{\sum_k^n \delta_t^{t_{i|k}} \delta_v^{v_{i|k}}} \quad (11)$$

Thus, for each read we obtain a vector of posterior probabilities (one value for each genome in our dataset). If a posterior probability value is above a given threshold

(default: 0.75) we consider the read r_i to be informative and assign it to that reference genome. If instead all posterior values are below the threshold, we consider the read to be uninformative and we assign it to the $n + 1$ category where n is the number of reference genomes.

This means that any read r_i that has been successfully aligned, can only be assigned to a single category: either to one reference genome or to the “Grey Matter” (GM, the $n + 1$ category). All reads which belong to species that do not possess a representative reference genome (or a close relative) in the database and therefore haven’t been successfully aligned to any of the reference genomes, will then be assigned to the $n + 2$ category called “Dark Matter” (DM).

Repeating this procedure $\forall r_i \in R$ allows us to populate a matrix $M_{N \times (n+2)}$ in which each row represents a read and each column represents either a reference genome in our database or the GM or the DM category.

Hence, the assignment matrix represents a series of N observations from a multinomial distribution of $n + 2$ variables (where $N = \#R$, the total number of reads). This distribution is parametrized by a probability mass function $q = [q_1, q_2, \dots, q_n, q_{GM}, q_{DM}]$.

The likelihood of N observations has thus the following form:

$$M \sim Multinomial_{n+2}(N, q)$$

$$f(x_1, x_2, \dots, x_{n+2} | N, q) = \frac{N!}{x_1! x_2! \dots x_{n+2}!} \prod_{j=1}^{n+2} q_j^{x_j} \quad (12)$$

where x_j is the sum of the column j of M , i.e. the number of occurrences of genome G_j .

Dirichlet prior

The Dirichlet distribution (*Dir*) is a conjugate prior for the multinomial distribution and it has desirable mathematical properties which make it appropriate to model the composition of metagenomes.

Let π_0 indicate the prior distribution and π^* the posterior.

From Bayes theorem,

$$\pi^*(M|\Theta) \propto L(M|\Theta)\pi_0(\Theta) \quad (13)$$

We employ a uniform prior:

$$\begin{aligned} \pi_0(\Theta) &\sim Dir(\alpha) \\ \alpha_0 &= (\alpha_1, \dots, \alpha_{n+2}), \\ \alpha_i &= 1 \quad \forall i = 1, \dots, n+2 \end{aligned} \quad (14)$$

Because the Dirichlet is a conjugate prior for the multinomial likelihood (eq. 12), the posterior density has the same form, i.e

$$\begin{aligned} \pi_*(M|\Theta) &\sim Dir(\alpha^*) \\ \alpha^* &= (\alpha_1^*, \dots, \alpha_{n+2}^*), \\ \alpha_j^* &= \alpha_j + x_j, \quad j = 1, \dots, n+2 \end{aligned} \quad (15)$$

where, again, x_j is the number of observations (reads) of genome G_j in the assignment matrix M . We can then use this framework to compute the minimum abundance of a species (represented by its closest reference genome) in a metagenome.

Posterior mean and Confidence interval

Let γ_j denote the posterior abundance of genome G_j .

From the property of the Dirichlet we can obtain the posterior mean as follows:

$$E[\gamma_j|X] = \frac{\alpha_j^*}{\alpha^{**}} \quad (16)$$

$$\alpha^{**} = \sum_{k=1}^{n+2} \alpha_k^* = (n+2) + \sum_{k=1}^{n+2} x_k = n+2+N$$

meaning that the denominator of each posterior mean corresponds to the total number of reads plus the number of categories (the number of reference genome + 2).

The marginal distribution of each abundance is:

$$\gamma_j \sim \text{Beta}(\alpha_j^*, \alpha^{**} - \alpha_j^*) \quad (17)$$

and we can use this to obtain the 95% CI numerically.

Computational architecture

HAYSTAC is written in Python and built upon the snakemake workflow engine (Köster and Rahmann, 2012). It operates as a wrapper around custom python scripts and several common bioinformatics tools, including: AdapterRemoval (Schubert et al., 2016), bowtie2 (Longmead and Salzberg, 2012), DeDup (Peltzer et al., 2016), mapDamage (Jónsson et al., 2013), samtools (Li et al., 2009), SeqKit (Shen et al., 2016), seqtk (Li, 2012) and sra-tools (Leinonen et al., 2011). It uses the conda package manager, with bioconda (Grüning et al., 2018), to manage its dependencies.

A tool for constructing databases

HAYSTAC includes a user-friendly tool to download genomes from NCBI and build customisable reference databases. After deciding *a priori* which taxa to include in their database (e.g., anywhere between a single genus to many thousands of taxa), the user constructs a database from any combination of: (i) an NCBI search query; (ii) a user specified list of NCBI accession codes; (iii) the RefSeq representative database of prokaryotic species; and/or (iv) a list of user provided reference assemblies (in FASTA format).

To construct a valid NCBI search query, visit the NCBI Nucleotide database website (<https://www.ncbi.nlm.nih.gov/nucleotide/>) and use the search feature to obtain a correctly formatted query string from the "Search details" box. This search query can

then be used directly with HAYSTAC to automatically download and build a reference database based on the accession codes present in the resultset returned by the query.

We caution that the choice of reference database can dramatically affect species identification for all methods that use reference databases. For example, issues can arise if the reference genomes in the database are not complete, as reads may map uniquely to one reference genome because other incomplete genomes in the database do not possess that specific sequence. To mitigate this issue, HAYSTAC uses a simple heuristic to select the longest genome available per species in the input resultset, which limits issues arising from unequal size of reference genomes in the database.

Processing, and aligning reads

Pre-processed reads can be directly passed onto the alignment step (Figure 1). This is useful when dealing with modern DNA. The user can also decide whether to remove sequencing adapters in the case of raw reads, and for ancient/degraded DNA collapse overlapping mate pairs using a wrapper for AdapterRemoval2 (Schubert et al., 2016). Reads are then aligned to a database using Bowtie2 (Longmead and Salzberg, 2012). We first align to a single Bowtie2 index that contains all the reference genomes, to obtain a BAM file that contains one single best alignment per read. For this alignment step we use the local alignment mode in Bowtie2. This allows us to discard reads that did not map to a single reference genome. This BAM is used by HAYSTAC to compute the expected Ts-Tv ratio from the data, which is used as prior in eq 4. Reads that pass

this first filtering step are then re-aligned to each reference genome separately using Bowtie2 (one index per reference) in end-to-end mode. We set a maximum number of mismatches to calculate the minimum alignment score during the Bowtie alignment. The maximum mismatch allowed during this second alignment step is calculated by the average read length of the input library. Allowing more mismatches results in slower run time as the number of reads included in the analytical step increases.

Computing abundance

Alignment files are then processed to calculate a likelihood score for each read/reference alignment and build a likelihood matrix that possesses $n * r$ entries, where n is the number of reference genomes and r is the number of reads that pass the first filtering stage. This is done by computing T_s and T_v for each alignment, and computing the likelihood using eq 4. Here the user can change the probability of mismatch within a given alignment (σ ; default 0.05) (see eq 4). For each read we find the maximum likelihood, then compute a posterior probability using eq 7. Reads are then filtered based on their posterior probability (default 0.75) and included in the dirichlet assignment step. Reads that have not been aligned to any taxa in the database get assigned to the “Dark Matter” category, a category that absorbs reads whose reference genomes are not included in our database. Reads that have been aligned to multiple taxa in our database are also assigned to another special category called the “Grey Matter”, a category that includes reads that were not informative enough to be either uniquely aligned to a specific taxon or obtain posterior probability below the

threshold that is set for the dirichlet assignment. Both of these categories are included in the mean posterior abundance calculations.

Assessing evenness of coverage

We use a chi square test to assess if the reads assigned to a specific taxon represent a random sample of its genome or whether they are clustering around specific regions. To assess the evenness of coverage, we populate a contingency table with the observed values of the total number of nucleotide positions on a reference genome that are covered at least once by reads uniquely assigned to it, and the total number of the nucleotide bases from all the reads that have been uniquely assigned to that genome. In the case of a true positive identification the reads should not be clustering around specific genome regions (p-value < 5%). A non-significant p-value could indicate a possible HGT event. If a taxon includes more than one reference genomes (e.g. a bacterial chromosome and a plasmid sequence), we then calculate the harmonic mean of the chi-square test p-values across all the different reference genomes for said taxon, so that only one p-value per taxon is obtained in the end. This test is particularly useful for shallow sequencing data that do not allow for more than 1X genome coverage of a given reference genome.

Representative species RefSeq database

For our analyses we build a database out of all the representative species of the prokaryotic RefSeq database. A list of all the species that were included can be found

here:

https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prok_representative_genomes.txt. In cases where the RefSeq representative database contains more than one strain for a given species, the first listed accession was picked. We also added complete genomes from the genera *Klebsiella*, *Streptococcus*, *Corynebacterium*, *Bordetella*, *Haemophilus* and *Yersinia*, for these species that were not present in the prokaryotic representative RefSeq. When possible (i.e. when genome assembly with plasmid data were available) we included plasmid sequences in the database. It should be noted that in the case of composite databases where species from both the representative prokaryotic RefSeq and individual NCBI queries are involved the following priority system decides which genomes will be included in the database: 1) if a species has an accession both in the representative prokaryotic RefSeq and in an NCBI query, the representative prokaryotic RefSeq accession will be preferred, regardless if it the longest genome available, unless the user specifies otherwise and 2) if a species has multiple accessions in the representative prokaryotic RefSeq, only the first accession for that species will be included in the database, unless the user specifies otherwise again.

Simulations

We generated 3 different simulated datasets. The first dataset was generated to determine the accuracy of different methods under simple conditions. We did not apply any chemical damage pattern and the read length was kept constant at 60 bp. This was

done by randomly sampling 10 species from the prokaryotic representative RefSeq database (see above) twice, to get two different sets of 10 species. For each set we then simulated 2 different sequencing library samples of 1 million single-end reads using gargammel (Renaud et al., 2017). One set had no human genome included and all taxa were present at 10% abundance. In the second set 25% of reads were from the human genome, and all remaining taxa were present at 7.5% abundance. This dataset included 4 simulated sample sequencing files.

The second dataset was produced to compare performance of the different tools when analysing ancient and modern DNA, and did not include the human genome. To do so, we first generated two sets of 100 species randomly sampled from the prokaryotic representative RefSeq database, and two sets of 500 randomly sampled species. The species were different between each of the 2 random samplings for both the 100 species and 500 species sets. For each species profile in both the 100 species and 500 species sets, we simulated two different libraries (8 total libraries) using gargammel with 10 million single-end reads of length 125bp from an Illumina HiSeq 2500 run. The species abundances were kept consistent within each library, where libraries of 100 species have all taxa at 1% abundance and libraries of 500 species have all taxa at 0.2% abundance. The first library incorporated aDNA damage patterns based on the profile of reads that mapped to the *Tannerella forsythia* genome from a real ancient dental calculus sample (CS21 from Velsko et al. (2019)) and the second library did not include aDNA damage patterns.

Lastly, to assess our method with more realistic data sets, we analysed simulated data previously published by Velsko et al. (2018) designed to resemble a dental plaque community, which included species that were not in the prokaryotic representative RefSeq database. Twelve datasets of 200 species each were generated: with and without ancient DNA damage, each with even species abundance (0.5%) or with abundance based on values observed in an ancient oral microbiome bacterial community (see Supplemental Table S8). The type of simulation included members of the *Corynebacterium* genus without *Corynebacterium diphtheriae*; the second type included all members of the *Corynebacterium* genus (including *Corynebacterium diphtheriae*); the third type did not include any member of the *Corynebacterium* genus. Each of these twelve samples contained 5 million paired end reads of varying length that were collapsed with AdapterRemoval.

Analyses of simulated data

We trimmed adapters using AdapterRemoval v2 (Schubert et al., 2016) excluding reads that were smaller than 15 bp. Analyses of simulated data with Kraken2/Bracken (Lu et al., 2017; Wood et al., 2019), MALT (Vågene et al., 2018) were conducted using default parameters. The MALT output was subsequently analysed with HOPS (Hübner et al., 2019) with default values but providing a list of candidate species that included all the species in the custom database that we used for the rest of the methods. HAYSTAC and Sigma (Ahn et al., 2015) were run with default parameters. For HAYSTAC,

Kraken2/Bracken and Sigma positive identification abundance threshold was set to 0.01% while we used 50 reads for MALT/HOPS as in Hübler et al. (2019).

In order to assess how HAYSTAC, Sigma, Kraken2/Bracken and MALT/HOPS perform with genus specific database, we analysed 20 simulated samples (based on RefSeq genomes; see above) and aligning reads to all representative genomes in a single genus. Ten genera were picked randomly for these analyses: *Bacteroides* (24 species), *Burkholderia* (3 species), *Campylobacter* (22 species), *Clostridium* (91 species), *Corynebacterium* (83 species), *Desulfitobacterium* (4 species), *Lactobacillus* (139 species), *Mycobacterium* (36 species), *Solimonas* (3 species) and *Streptococcus* (78 species).

Case study 1: Ancient *Yersinia pestis*

We tested HAYSTAC using a published dataset of ancient human bone and tooth samples in which *Yersinia pestis* was identified (Rasmussen et al., 2015), to compare performance against standard techniques used to identify specific pathogens in aDNA samples. We downloaded reads from all seven samples in which the authors detected *Y. pestis* (RISE00, RISE139, RISE386, RISE505, RISE509, RISE511) from the European Nucleotide Archive (NCBI BioProject accession: PRJEB10885). We then used HAYSTAC to build a database that contains the longest complete genome for each species of the genus *Yersinia* and aligned reads using default parameters. HAYSTAC was then used to compute posterior abundance of each species.

Case study 2: Historic dental calculus

We also tested HAYSTAC using an historic dental calculus dataset (Velsko et al., 2019) (PRJEB30331). Raw reads were processed as described in Velsko et al. (2019). We attempted to identify a set of respiratory pathogens in these calculus samples, including: *Corynebacterium diphtheriae* and *Bordetella pertussis*, the causative agents of diphtheria and whooping cough, respectively (Figure 9). We also tested for the presence of species belonging to the *Haemophilus* (ideally *Haemophilus influenzae* or *parainfluezae*), *Klebsiella* (ideally *Klebsiella pneumoniae*) and *Streptococcus* (*Streptococcus pneumoniae*) genera. These species were selected to test if pathobionts of the upper respiratory system could be found in dental calculus.

Performance test

For the performance tests we created a reproducible conda environment, inside which all the different pipelines were installed and run. For each method we used the GNU time command to measure elapsed time and peak memory usage. Scripts to run the performance benchmarks are available from

https://github.com/antonisdim/haystac_paper

We used a total of 500 accessions, one for each species, which were a subset of the accessions used for the analysis of all real and simulated datasets. We built databases with an increasing number of species (10, 100, 500) and the same database was given

as input to all the methods. We only measured the database building time and maximum memory usage for each method, as HAYSTAC was used to efficiently fetch the reference genomes from NCBI.

For sample inputs we used samples from the Human Microbiome Project, with accession SRS078671. We merged the forward mate and the singleton reads in one file that we then treated as single end reads. That file was subsequently subsampled to generate fastq files of 10 K, 100 K, and 1 M SE reads respectively, that were used as inputs for all the tests.

For the sample analysis we performed two tests, one that measured max memory usage and runtime for samples of increasing size against the database of 500 species, and a second where the input file size was kept constant at 1 million reads, while the database size was being varied, again measuring the same variables for memory and elapsed execution time.

Acknowledgments

E.A.D. was supported by the DTP in Environmental Research and funded by NERC award NE/L002612/1. L.A.F.F. and G.L. were supported either by a European Research Council grants (ERC-2013-StG-337574-UNDEAD and ERC-2019-StG-853272-PALAEOFARM) and Natural Environmental Research Council grants (NE/K005243/1, NE/K003259/1, NE/S007067/1, and NE/S00078X/1). E.A.D, L.A.F.F. and A.C. were also supported by the Wellcome Trust (210119/Z/18/Z). The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (<http://dx.doi.org/10.5281/zenodo.22558>) and Queen Mary's Apocrita HPC facility.

Bibliography

- Ahn, T.-H., Chai, J., Pan, C., 2015. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 31, 170–177.
- Arora, N., Schuenemann, V.J., Jäger, G., Peltzer, A., Seitz, A., Herbig, A., Strouhal, M., Grillová, L., Sánchez-Busó, L., Kühnert, D., Bos, K.I., Davis, L.R., Mikalová, L., Bruisten, S., Komericki, P., French, P., Grant, P.R., Pando, M.A., Vaulet, L.G., Fermepin, M.R., Martinez, A., Centurion Lara, A., Giacani, L., Norris, S.J., Šmajš, D., Bosshard, P.P., González-Candelas, F., Nieselt, K., Krause, J., Bagheri, H.C., 2016. Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster. *Nat Microbiol* 2, 16245.
- Borówka, P., Pułaski, Ł., Marciniak, B., Borowska-Strugińska, B., Dziadek, J., Żądzińska, E., Lorkiewicz, W., Strapagiel, D., 2019. Screening methods for detection of ancient *Mycobacterium tuberculosis* complex fingerprints in next-generation sequencing data derived from skeletal samples. *Gigascience* 8. <https://doi.org/10.1093/gigascience/giz065>
- 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., Wood, J., Earn, D.J.D., Herring, D.A., Bauer, P., Poinar, H.N., Krause, J., 2011. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* 478, 506–510.
- Breitwieser, F.P., Perteua, M., Zimin, A.V., Salzberg, S.L., 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* 29, 954–960.

- Chen, P.E., Cook, C., Stewart, A.C., Nagarajan, N., Sommer, D.D., Pop, M., Thomason, B., Thomason, M.P.K., Lentz, S., Nolan, N., Sozhamannan, S., Sulakvelidze, A., Mateczun, A., Du, L., Zwick, M.E., Read, T.D., 2010. Genomic characterization of the *Yersinia* genus. *Genome Biol.* 11, R1.
- Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A., Callahan, B.J., 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6, 226.
- De Keukeleire, S., De Bel, A., Jansen, Y., Janssens, M., Wauters, G., Piérard, D., 2014. *Yersinia ruckeri*, an unusual microorganism isolated from a human wound infection. *New Microbes New Infect* 2, 134–135.
- Eerkens, J.W., Nichols, R.V., Murray, G.G.R., Perez, K., Murga, E., Kaijankoski, P., Rosenthal, J.S., Engbring, L., Shapiro, B., 2018. A probable prehistoric case of meningococcal disease from San Francisco Bay: Next generation sequencing of *Neisseria meningitidis* from dental calculus and osteological evidence. *Int J Paleopathol* 22, 173–180.
- Feuerborn, T.R., Palkopoulou, E., van der Valk, T., von Seth, J., Munters, A.R., Pečnerová, P., Dehasque, M., Ureña, I., Ersmark, E., Lagerholm, V.K., Krzewińska, M., Rodríguez-Varela, R., Götherström, A., Dalén, L., Díez-Del-Molino, D., 2020. Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. *BMC Genomics* 21, 844.
- Gonzalez, A., Vázquez-Baeza, Y., Pettengill, J.B., Ottesen, A., McDonald, D., Knight, R., 2016. Avoiding Pandemic Fears in the Subway and Conquering the Platypus.

mSystems. <https://doi.org/10.1128/mSystems.00050-16>

- Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., Köster, J., Bioconda Team, 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476.
- Hübler, R., Key, F.M., Warinner, C., Bos, K.I., Krause, J., Herbig, A., 2019. HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biol.* 20, 280.
- Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386.
- Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., Orlando, L., 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684.
- Key, F.M., Posth, C., Krause, J., Herbig, A., Bos, K.I., 2017. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet.* 33, 508–520.
- Kistler, L., Ware, R., Smith, O., Collins, M., Allaby, R.G., 2017. A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkx361>
- Köster, J., Rahmann, S., 2012. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522.
- Leinonen, R., Sugawara, H., Shumway, M., International Nucleotide Sequence

- Database Collaboration, 2011. The sequence read archive. *Nucleic Acids Res.* 39, D19–21.
- Li, H., 2012. seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub* 767, 69.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Loe, L., Pollard, M., Gibson, M., Simmonds, A., Cotter, J., Parkin, A., 2014. Radcliffe Infirmary Burial Ground, Oxford. Post-excavation assessment and project design 1–85.
- Longmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie2.
- Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L., 2017. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* 3, e104.
- 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., Frohlich, B., Aldenderfer, M., Hoogland, M., Read, C., Milner, G.R., Stone, A.C., Lewis, C.M., Jr, Krause, J., Hofman, C., Bos, K.I., Warinner, C., 2018. Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Sci. Rep.* 8, 9822.
- Mark Welch, J.L., Rossetti, B.J., Rieken, C.W., Dewhirst, F.E., Borisy, G.G., 2016. Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad.*

Sci. U. S. A. 113, E791–800.

Mühlemann, B., Vinner, L., Margaryan, A., Wilhelmson, H., de la Fuente Castro, C., Allentoft, M.E., de Barros Damgaard, P., Hansen, A.J., Holtsmark Nielsen, S., Strand, L.M., Bill, J., Buzhilova, A., Pushkina, T., Falys, C., Khartanovich, V., Moiseyev, V., Jørkov, M.L.S., Østergaard Sørensen, P., Magnusson, Y., Gustin, I., Schroeder, H., Sutter, G., Smith, G.L., Drosten, C., Fouchier, R.A.M., Smith, D.J., Willerslev, E., Jones, T.C., Sikora, M., 2020. Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science* 369.

<https://doi.org/10.1126/science.aaw8977>

Müller, R., Roberts, C.A., Brown, T.A., 2014. Biomolecular identification of ancient *Mycobacterium tuberculosis* complex DNA in human remains from Britain and continental Europe. *Am. J. Phys. Anthropol.* 153, 178–189.

Naccache, S.N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A.L., Luk, K.-C., Enge, B., Wadford, D.A., Messenger, S.L., Genrich, G.L., Pellegrino, K., Grard, G., Leroy, E., Schneider, B.S., Fair, J.N., Martínez, M.A., Isa, P., Crump, J.A., DeRisi, J.L., Sittler, T., Hackett, J., Jr, Miller, S., Chiu, C.Y., 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* 24, 1180–1192.

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., Nieselt, K., 2016.

EAGER: efficient ancient genome reconstruction. *Genome Biol.* 17, 60.

Pósa, A., Maixner, F., Mende, B.G., Köhler, K., Osztás, A., Sola, C., Dutour, O., Masson,

- M., Molnár, E., Pálfi, G., Zink, A., 2015. Tuberculosis in Late Neolithic-Early Copper Age human skeletal remains from Hungary. *Tuberculosis* 95 Suppl 1, S18–22.
- Prendergast, M.E., Buckley, M., Crowther, A., Frantz, L., Eager, H., Lebrasseur, O., Hutterer, R., Hulme-Beaman, A., Van Neer, W., Douka, K., Veall, M.-A., Quintana Morales, E.M., Schuenemann, V.J., Reiter, E., Allen, R., Dimopoulos, E.A., Helm, R.M., Shipton, C., Mwebi, O., Denys, C., Horton, M., Wynne-Jones, S., Fleisher, J., Radimilahy, C., Wright, H., Searle, J.B., Krause, J., Larson, G., Boivin, N.L., 2017. Reconstructing Asian faunal introductions to eastern Africa from multi-proxy biomolecular and archaeological datasets. *PLoS One* 12, e0182565.
- Rasmussen, S., Allentoft, M.E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., Pedersen, A.G., Schubert, M., Van Dam, A., Kapel, C.M.O., Nielsen, H.B., Brunak, S., Avetisyan, P., Epimakhov, A., Khalyapin, M.V., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., Moiseyev, V., Gromov, A., Pokutta, D., Saag, L., Varul, L., Yepiskoposyan, L., Sicheritz-Pontén, T., Foley, R.A., Lahr, M.M., Nielsen, R., Kristiansen, K., Willerslev, E., 2015. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* 163, 571–582.
- Renaud, G., Hanghøj, K., Willerslev, E., Orlando, L., 2017. gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33, 577–579.
- Reuter, S., Connor, T.R., Barquist, L., Walker, D., Feltwell, T., Harris, S.R., Fookes, M., Hall, M.E., Petty, N.K., Fuchs, T.M., Corander, J., Dufour, M., Ringwood, T., Savin, C., Bouchier, C., Martin, L., Miettinen, M., Shubin, M., Riehm, J.M., Laukkanen-Ninios, R., Sihvonen, L.M., Siitonen, A., Skurnik, M., Falcão, J.P.,

- Fukushima, H., Scholz, H.C., Prentice, M.B., Wren, B.W., Parkhill, J., Carniel, E., Achtman, M., McNally, A., Thomson, N.R., 2014. Parallel independent evolution of pathogenicity within the genus *Yersinia*. *Proc. Natl. Acad. Sci. U. S. A.* 111, 6768–6773.
- Schubert, M., Lindgreen, S., Orlando, L., 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9, 88.
- Schuenemann, V.J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mitnik, A., Forrest, S., Coombes, B.K., Wood, J.W., Earn, D.J.D., White, W., Krause, J., Poinar, H.N., 2011. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proc. Natl. Acad. Sci. U. S. A.* 108, E746–52.
- 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., Krause, J., 2018. Historic *Treponema pallidum* genomes from Colonial Mexico retrieved from archaeological remains. *PLoS Negl. Trop. Dis.* 12, e0006447.
- Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* 11, e0163962.
- Spyrou, M.A., Bos, K.I., Herbig, A., Krause, J., 2019. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature Reviews Genetics*.
<https://doi.org/10.1038/s41576-019-0119-1>
- Spyrou, M.A., Tikhbatova, R.I., Feldman, M., Drath, J., Kacki, S., Beltrán de Heredia, J., Arnold, S., Sitdikov, A.G., Castex, D., Wahl, J., Gazimzyanov, I.R., Nurgaliev,

- D.K., Herbig, A., Bos, K.I., Krause, J., 2016. Historical *Y. pestis* Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. *Cell Host Microbe* 19, 874–881.
- Sumi, Y., Miura, H., Michiwaki, Y., Nagaosa, S., Nagaya, M., 2007. Colonization of dental plaque by respiratory pathogens in dependent elderly. *Arch. Gerontol. Geriatr.* 44, 119–124.
- Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., De Filippis, F., Magnabosco, C., Bonneau, R., Lusingu, J., Amuasi, J., Reinhard, K., Rattei, T., Boulund, F., Engstrand, L., Zink, A., Collado, M.C., Littman, D.R., Eibach, D., Ercolini, D., Rota-Stabelli, O., Huttenhower, C., Maixner, F., Segata, N., 2019. The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* 26, 666–679.e7.
- Tortoli, E., 2012. Phylogeny of the genus *Mycobacterium*: many doubts, few certainties. *Infect. Genet. Evol.* 12, 827–831.
- 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., Bos, K.I., Krause, J., 2018. *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat Ecol Evol* 2, 520–528.
- Velsko, I.M., Fellows Yates, J.A., Aron, F., Hagan, R.W., Frantz, L.A.F., Loe, L., Martinez, J.B.R., Chaves, E., Gosden, C., Larson, G., Warinner, C., 2019. Microbial differences between dental plaque and historic dental calculus are related to oral

- biofilm maturation stage. *Microbiome* 7, 102.
- Velsko, I.M., Frantz, L.A.F., Herbig, A., Larson, G., Warinner, C., 2018. Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research. *mSystems* 3. <https://doi.org/10.1128/mSystems.00080-18>
- Warinner, C., Herbig, A., Mann, A., Fellows Yates, J.A., Weiß, C.L., Burbano, H.A., Orlando, L., Krause, J., 2017. A Robust Framework for Microbial Archaeology. *Annu. Rev. Genomics Hum. Genet.* 18, 321–356.
- Warinner, C., Rodrigues, J.F.M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R.Y., Fiddyment, S., Speller, C., Hendy, J., Charlton, S., Luder, H.U., Salazar-García, D.C., Eppler, E., Seiler, R., Hansen, L.H., Castruita, J.A.S., Barkow-Oesterreicher, S., Teoh, K.Y., Kelstrup, C.D., Olsen, J.V., Nanni, P., Kawai, T., Willerslev, E., von Mering, C., Lewis, C.M., Jr, Collins, M.J., Gilbert, M.T.P., Rühli, F., Cappellini, E., 2014. Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* 46, 336–344.
- Warinner, C., Speller, C., Collins, M.J., Lewis, C.M., Jr, 2015. Ancient human microbiomes. *J. Hum. Evol.* 79, 125–136.
- Wilson, M.R., Sample, H.A., Zorn, K.C., Arevalo, S., Yu, G., Neuhaus, J., Federman, S., Stryke, D., Briggs, B., Langelier, C., Berger, A., Douglas, V., Josephson, S.A., Chow, F.C., Fulton, B.D., DeRisi, J.L., Gelfand, J.M., Naccache, S.N., Bender, J., Dien Bard, J., Murkey, J., Carlson, M., Vespa, P.M., Vijayan, T., Allyn, P.R., Campeau, S., Humphries, R.M., Klausner, J.D., Ganzon, C.D., Memar, F., Ocampo, N.A., Zimmermann, L.L., Cohen, S.H., Polage, C.R., DeBiasi, R.L., Haller, B.,

- Dallas, R., Maron, G., Hayden, R., Messacar, K., Dominguez, S.R., Miller, S., Chiu, C.Y., 2019. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N. Engl. J. Med.* 380, 2327–2340.
- Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257.
- Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46.
- Ye, S.H., Siddle, K.J., Park, D.J., Sabeti, P.C., 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* 178, 779–794.
- Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C., Achtman, M., 2018. Accurate Reconstruction of Microbial Strains from Metagenomic Sequencing Using Representative Reference Genomes, in: *Research in Computational Molecular Biology*. Springer International Publishing, pp. 225–240.

Supplementary

Calculation of false positive, false negative and true positive rates

For the calculation of the false positive, false negative and true positive rates we used the following formulas:

$$FP_r = \frac{FP}{FP+TN}$$

(Eq. 18)

$$FN_r = \frac{FN}{FN+TP}$$

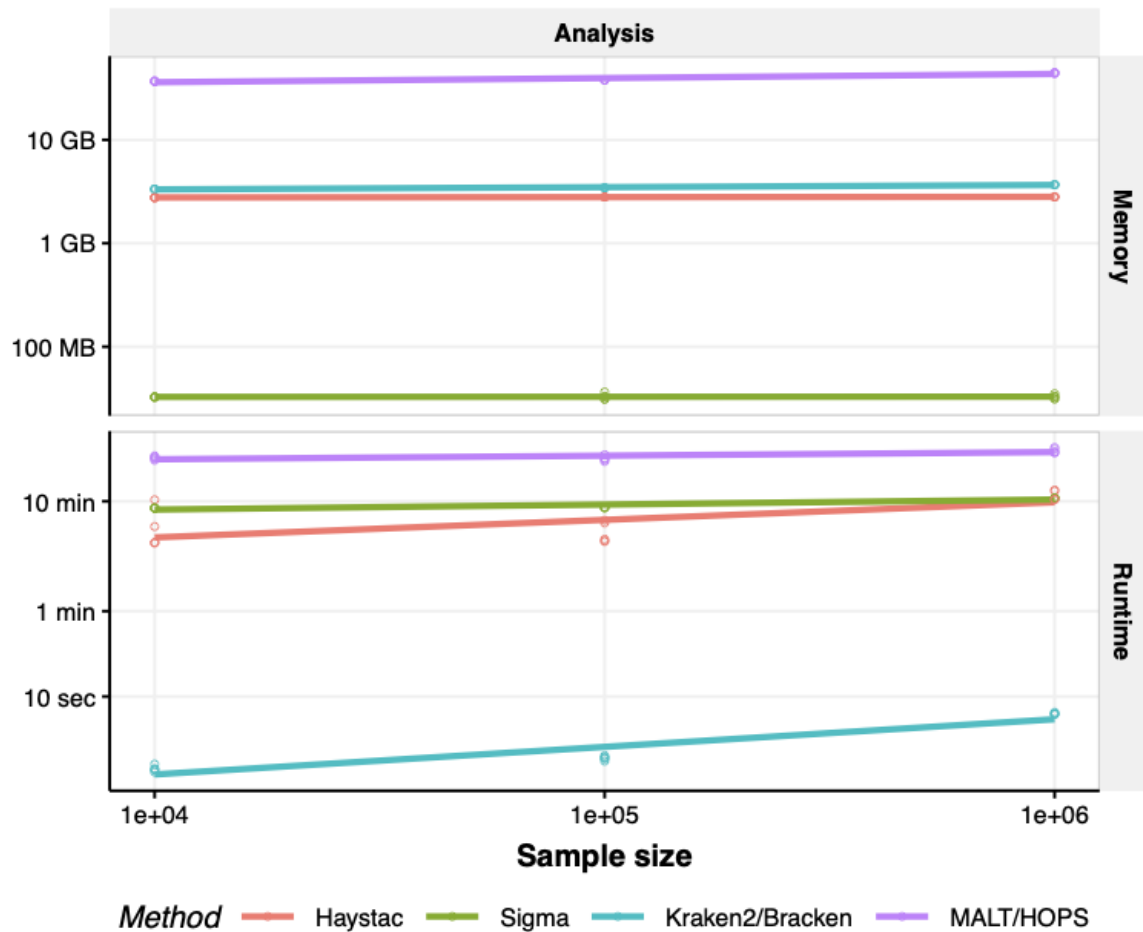
(Eq. 19)

$$TP_r = \frac{TP}{TP+FP}$$

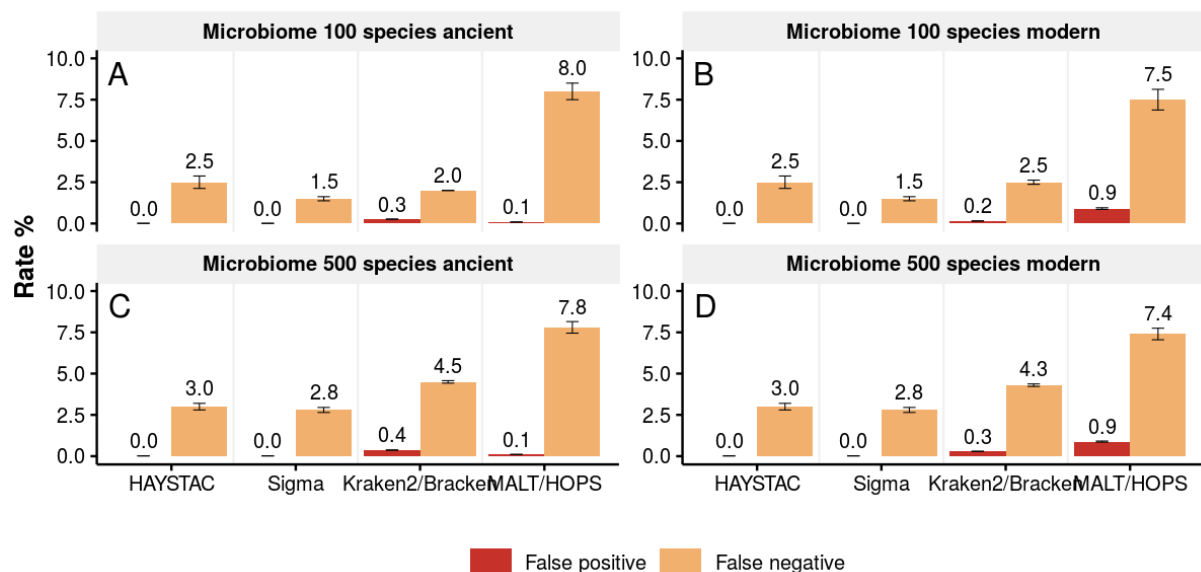
(Eq. 20)

where: TP = number of true positives, FP = number of false positives, TN = number of true negatives, and FN = number of false negatives. For a true positive identification we require an abundance equal or higher than 0.01%.

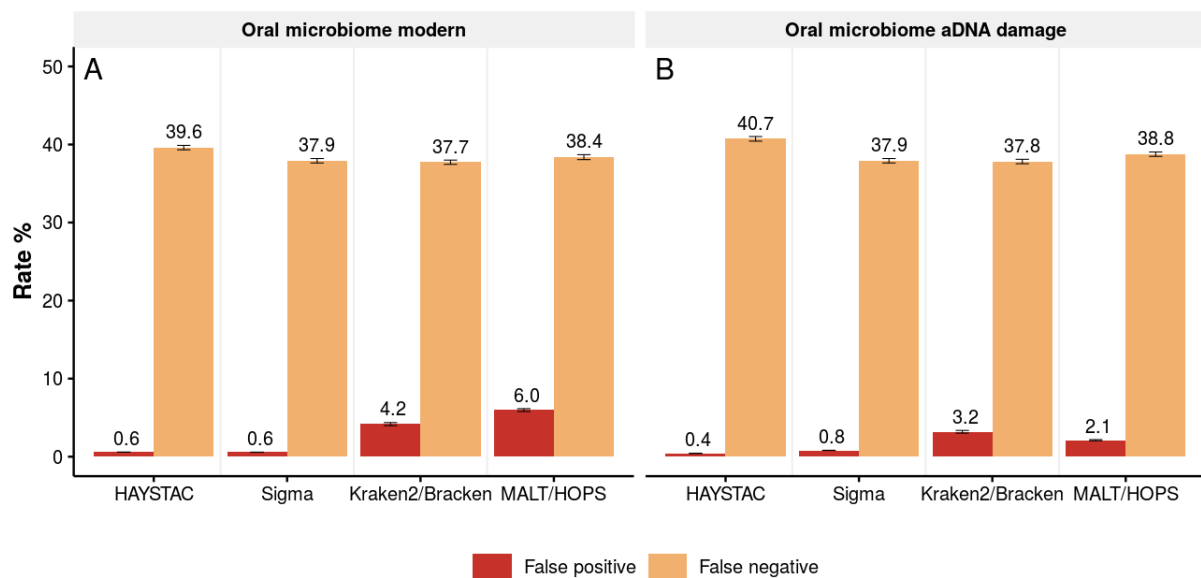
Supplemental Figures



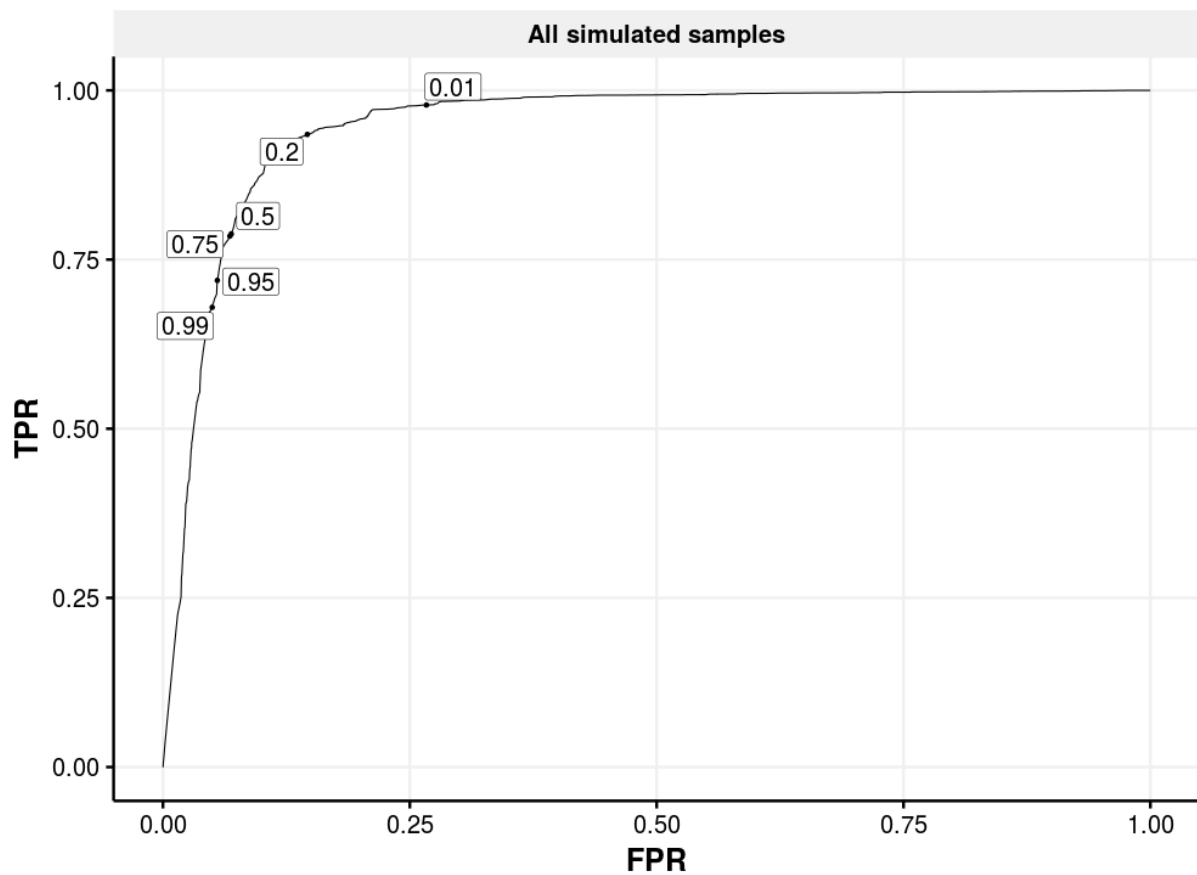
Supplemental Figure 1. Benchmarking for elapsed runtime and memory for HAYSTAC, Sigma, Kraken2/Bracken and MALT/HOPS when analysing samples of 10 K, 100 K and 1 M reads against a database of 500 species. Memory remains constant as sample size increases, and runtime in most methods (other than Kraken2/Bracken) scales with the database size rather than input sample size.



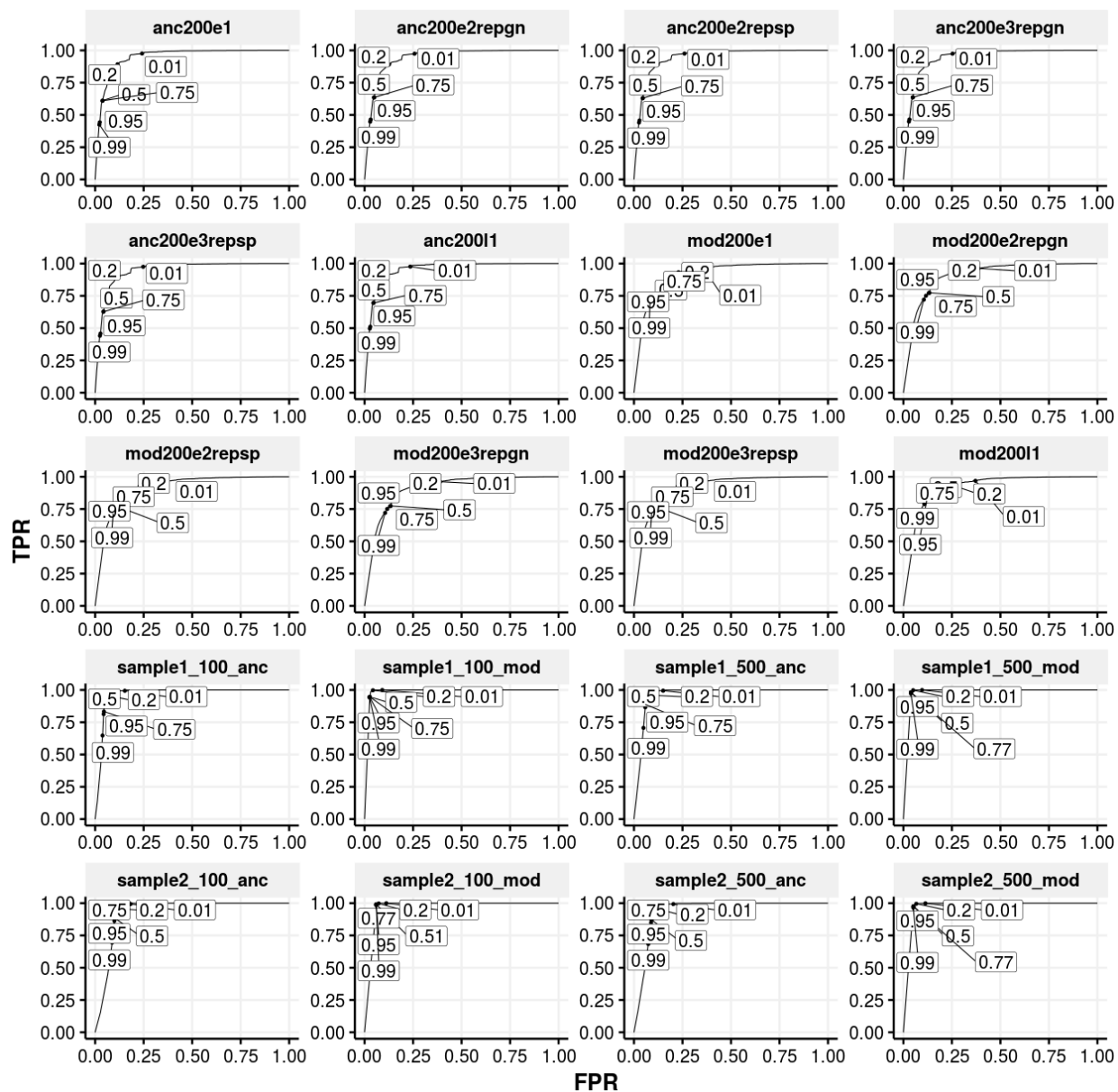
Supplemental Figure 2. False positive (red) and negative (orange) rates per method for the simulated General Microbiome dataset of 100 species simulated as ancient samples (n=2) (A), 100 species simulated as modern samples (n=2) (B), and 500 species simulated as ancient samples (n=2) (C) and 500 species simulated as modern samples (n=2) (D).



Supplemental Figure 3. False positive (red) and negative (orange) rates per method for two versions of the oral microbiome dataset, each with six replicates: (A) modern simulation, with fixed read lengths (n=6); and (B) ancient simulation, with variable read lengths and post-mortem damage (n=6).



Supplemental Figure 4. Receiver operator curve analysis, showing the relationship between the true and false positive ratios (TPR and FPR respectively) to determine the default read posterior probability threshold for the dirichlet assignment across all simulated samples. A rather conservative threshold of 0.75 is appropriate for most types of analysis. If the user requires a more relaxed threshold for discovery purposes the threshold can be lowered to 0.5. Respectively if a user is analysing deep sequencing data or wants to perform a more stringent identification the threshold can also be increased.



Supplemental Figure 5. Receiver operator curve analysis, showing the relationship between the true and false positive ratios (TPR and FPR respectively) to determine the default read posterior probability threshold for the dirichlet assignment per sample. Here we can see how the read length and deamination levels affect the true positive

rate. The user might want to consider these additional factors if they wish to change the default value of 0.75.

Supplemental Tables

Supplementary table S1. Characteristics of the General Microbiome simulated dataset.

Sample	Chemical damage pattern	Abundance	Read length	Number of species
sample1_100_anc	yes	1%	Variable	100
sample1_100_mod	no	1%	Variable	100
sample1_500_anc	yes	0.02%	Variable	500
sample1_500_mod	no	0.02%	Variable	500
sample2_100_anc	yes	1%	Variable	100
sample2_100_mod	no	1%	Variable	100
sample2_500_anc	yes	0.02%	Variable	500
sample2_500_mod	no	0.02%	Variable	500

Supplementary table S2. Characteristics of the Oral Microbiome simulated dataset.

Sample	Chemical damage pattern	Abundance	Read length	Number of species	<i>C. diphtheriae</i> control
anc200e1	yes	0.50%	80 bp	176	positive
anc200e2rep gn	yes	0.50%	80 bp	180	negative - genus
anc200e2rep sp	yes	0.50%	80 bp	178	negative - species
anc200e3rep gn	yes	0.50%	80 bp	180	negative - genus
anc200e3rep sp	yes	0.50%	80 bp	178	negative - species
anc200l1	yes	variable	80 bp	176	positive
mod200e1	no	0.50%	150 bp	176	positive
mod200e2rep gn	no	0.50%	150 bp	180	negative - genus

mod200e2repsp	no	0.50%	150 bp	178	negative - species
mod200e3repgn	no	0.50%	150 bp	180	negative - genus
mod200e3repsp	no	0.50%	150 bp	178	negative - species
mod200l1	no	variable	150 bp	176	positive

Supplemental table S3. Species identification counts and rates for the general microbiome simulated dataset.

Sample	Total Species	False Positive Count	True Detected Species Count	False Negative Count	False Positive Rate %	False Negative Rate %	Method
sample1_100_anc	100	15	98	2	0.27	2	kraken
sample1_100_mod	100	8	97	3	0.14	3	kraken
sample1_500_anc	500	17	476	24	0.33	4.8	kraken
sample1_500_mod	500	13	477	23	0.25	4.6	kraken
sample2_100_anc	100	14	98	2	0.252	2	kraken
sample2_100_mod	100	9	98	2	0.16	2	kraken
sample2_500_anc	500	21	479	21	0.41	4.2	kraken
sample2_500_mod	500	17	480	20	0.33	4	kraken
sample1_100_anc	100	6	94	6	0.11	6	malt
sample1_100_mod	100	39	95	5	0.70	5	malt
sample1_500_anc	500	5	454	46	0.09	9.2	malt
sample1_500_mod	500	52	456	44	1.01	8.8	malt
sample2_100_anc	100	4	90	10	0.07	10	malt
sample2_100_mod	100	61	90	10	1.10	10	malt
sample2_500_anc	500	6	468	32	0.11	6.4	malt
sample2_500_mod	500	39	470	30	0.76	6	malt
sample1_100_anc	100	0	96	4	0	4	haystac
sample1_100_mod	100	0	96	4	0	4	haystac
sample1_500_anc	500	1	481	19	0.02	3.8	haystac

sample1_500_mod	500	1	481	19	0.02	3.8	haystac
sample2_100_anc	100	1	99	1	0.02	1	haystac
sample2_100_mod	100	1	99	1	0.02	1	haystac
sample2_500_anc	500	1	489	11	0.02	2.2	haystac
sample2_500_mod	500	1	489	11	0.02	2.2	haystac
sample1_100_anc	100	0	98	2	0	2	sigma
sample1_100_mod	100	0	98	2	0	2	sigma
sample1_500_anc	500	1	483	17	0.02	3.4	sigma
sample1_500_mod	500	1	483	17	0.02	3.4	sigma
sample2_100_anc	100	1	99	1	0.02	1	sigma
sample2_100_mod	100	1	99	1	0.02	1	sigma
sample2_500_anc	500	1	489	11	0.02	2.2	sigma
sample2_500_mod	500	1	489	11	0.02	2.2	sigma

Supplemental table S4. Species identification counts and rates for the Oral Microbiome simulated dataset.

Sample	Total Species Count	False Positive Count	True Detected Species Count	False Negative Count	False Positive Rate %	False Negative Rate %	Method
anc200e1	176	31	106	70	0.57	39.8	haystac
anc200e2rep gn	180	34	105	75	0.62	41.8	haystac
anc200e2reps p	178	29	108	70	0.53	39.3	haystac
anc200e3rep gn	180	34	105	75	0.62	41.7	haystac
anc200e3reps p	178	28	108	70	0.51	39.3	haystac
anc200l1	176	28	105	71	0.51	40.3	haystac
mod200e1	176	34	106	70	0.62	39.8	haystac
mod200e2rep gn	180	37	107	73	0.68	40.6	haystac
mod200e2reps p	178	32	110	68	0.59	38.2	haystac

mod200e3rep gn	180	36	107	73	0.66	40.6	haystac
mod200e3repsp	178	32	110	68	0.59	38.2	haystac
mod200l1	176	30	105	71	0.55	40.3	haystac
anc200e1	176	46	109	67	0.84	38.1	sigma
anc200e2rep gn	180	45	110	70	0.82	38.9	sigma
anc200e2repsp	178	45	113	65	0.82	36.5	sigma
anc200e3rep gn	180	46	110	70	0.84	38.9	sigma
anc200e3repsp	178	45	113	65	0.82	36.5=	sigma
anc200l1	176	38	108	68	0.67	38.6	sigma
mod200e1	176	33	109	67	0.60	38.1	sigma
mod200e2rep gn	180	33	110	70	0.60	38.9	sigma
mod200e2repsp	178	30	113	65	0.55	36.5	sigma
mod200e3rep gn	180	33	110	70	0.60	38.9	sigma
mod200e3repsp	178	29	113	65	0.53	36.5	sigma
mod200l1	176	31	108	68	0.57	38.6	sigma
anc200e1	176	208	110	66	3.801	37.5	kraken
anc200e2rep gn	180	159	110	70	2.91	38.9	kraken
anc200e2repsp	178	202	113	65	3.70	36.5	kraken
anc200e3rep gn	180	154	110	70	2.82	38.9	kraken
anc200e3repsp	178	205	113	65	3.75	36.5	kraken
anc200l1	176	117	108	68	2.14	38.6	kraken
mod200e1	176	260	110	66	4.76	37.5	kraken
mod200e2rep gn	180	202	110	70	3.70	38.9	kraken
mod200e2repsp	178	266	113	65	4.87	36.5	kraken
mod200e3rep gn	180	208	110	70	3.81	38.9	kraken
mod200e3repsp	178	266	113	65	4.87	36.5	kraken
mod200l1	176	165	109	67	3.02	38.1	kraken
anc200e1	176	131	109	67	2.40	38.1	malt
anc200e2rep gn	180	111	108	72	2.03	40	malt
anc200e2repsp	178	125	111	67	2.30	37.6	malt
anc200e3rep gn	180	107	108	72	1.96	40	malt

anc200e3repsp	178	129	111	67	2.36	37.6	malt
anc200l1	176	89	107	69	1.63	39.2	malt
mod200e1	176	341	110	66	6.24	37.5	malt
mod200e2repgn	180	319	108	72	5.84	40	malt
mod200e2repsp	178	368	111	67	6.73	37.6	malt
mod200e3repgn	180	314	108	72	5.75	40	malt
mod200e3repsp	178	355	111	67	6.50	37.6	malt
mod200l1	176	258	110	66	4.72	37.5	malt

Supplemental table S5. Mean species counts per genus per method for the genus database analysis.

Genus	Method	False Positive Count	False Negative Count	True Detected Species Count
Bacteroides	kraken	23.6	0	0.4
Bacteroides	malt	21.7	0.1	0.3
Bacteroides	haystac	0	0.05	0.35
Bacteroides	sigma	0	0	0.4
Burkholderia	kraken	2.8	0	0.2
Burkholderia	malt	2.8	0	0.2
Burkholderia	haystac	0.5	0	0.2
Burkholderia	sigma	0.8	0	0.2
Campylobacter	kraken	19	0	3
Campylobacter	malt	11.6	0	3
Campylobacter	haystac	0	0.25	2.75
Campylobacter	sigma	0	0.1	2.9
Clostridium	kraken	88.5	0.3	2.5
Clostridium	malt	40.65	0.6	2.2
Clostridium	haystac	0.4	0	2.8
Clostridium	sigma	0.85	0	2.8

Corynebacterium	kraken	53.05	0	3.2
Corynebacterium	malt	49.7	0	3.2
Corynebacterium	haystac	0	0	3.2
Corynebacterium	sigma	0	0	3.2
Desulfitobacterium	kraken	3.9	0	0.1
Desulfitobacterium	malt	3.9	0	0.1
Desulfitobacterium	haystac	0	0.05	0.05
Desulfitobacterium	sigma	0	0	0.1
Mycobacterium	kraken	34.15	0.2	1.25
Mycobacterium	malt	30.85	0.35	1.1
Mycobacterium	haystac	0.45	0.35	1.1
Mycobacterium	sigma	1.45	0.35	1.1
Solimonas	kraken	2.7	0	0.3
Solimonas	malt	2.7	0	0.3
Solimonas	haystac	0	0.1	0.2
Solimonas	sigma	0	0	0.3
Streptococcus	kraken	70.45	0	6.3
Streptococcus	malt	35	0	6.3
Streptococcus	haystac	5.75	0.7	5.6
Streptococcus	sigma	6.8	0	6.3

Supplemental table S6. Oral Microbiome dataset simulated species. A total of 196 species was simulated for the Oral Microbiome dataset. Only 115 species were included in the database, with only 4 of these species sharing a common accession.

Species	Simulated accessions	Database accessions
<i>Abiotrophia defectiva</i>	GCF_000160075.2	ACIN03
<i>Actinomyces cardiffensis</i>	GCF_000364865.1	AQHZ01
<i>Actinomyces dentalis</i>	NZ_AUBL000000000.1	AUBL01
<i>Actinomyces massiliensis</i>	GCF_000269805.1	AKIO01

<i>Actinomyces oris</i>	GCF_001553935.1	NZ_CP014232.1
<i>Aggregatibacter actinomycetemcomitans</i>	CP001733.2, NZ_ADOB000000000.2	NZ_CP012959.1
<i>Alloprevotella tannerae</i>	NZ_ACIJ000000000.2	ACIJ02
<i>Anaeroglobus geminatus</i>	GCF_000239275.1	AGCJ01
<i>Atopobium parvulum</i>	CP001721.1	NC_013203.1
<i>Atopobium rimae</i>	NZ_ACFE000000000.1	ACFE01
<i>Bifidobacterium dentium</i>	NZ_ABIX000000000.2	NZ_AP012326.1
<i>Bordetella bronchiseptica</i>	NC_002927.3	NZ_CP014013.1
<i>Bordetella parapertussis</i>	NC_018828.1	NC_018828.1
<i>Bordetella pertussis</i>	BX470248.1	NZ_CP017405.1
<i>Campylobacter concisus</i>	CP000792.2	NZ_CP012541.1
<i>Campylobacter curvus</i>	CP000767.2	NC_009715.2
<i>Campylobacter gracilis</i>	GCF_001190745.1	NZ_CP012196.1
<i>Campylobacter showae</i>	NZ_AMZQ000000000.1	AMZQ01
<i>Capnocytophaga gingivalis</i>	GCF_000174755.1	ACLQ01
<i>Capnocytophaga granulosa</i>	GCF_000411115.1	ADDA01
<i>Capnocytophaga ochracea</i>	GCF_000023285.1	NC_013162.1
<i>Cardiobacterium valvarum</i>	NZ_AGCM000000000.1	AGCM01
<i>Catonella morbi</i>	GCF_000160035.2	ACIL03
<i>Clostridium botulinum</i>	GCF_000020165.1	NC_010674.1
<i>Clostridium tetani</i>	GCF_000007625.1	NC_004557.1
<i>Corynebacterium diphtheriae</i>	BX248353.1	NZ_LN831026.1
<i>Corynebacterium durum</i>	GCF_000318135.1	AMEM01
<i>Corynebacterium matruchotii</i>	GCF_000175375.1	ACSH02
<i>Corynebacterium urealyticum</i>	AM942444.1	NC_010545.1
<i>Dialister invisus</i>	GCF_000160055.1	ACIM02

<i>Eggerthella lenta</i>	CP001726.1	NC_013204.1
<i>Eikenella corrodens</i>	GCF_000158615.1	ACEA01
<i>Eubacterium brachy</i>	GCF_000488855.1	AXUD01
<i>Eubacterium saphenum</i>	GCF_000161975.1	ACON01
<i>Filifactor alocis</i>	GCF_000163895.2	NC_016630.1
<i>Finegoldia magna</i>	GCF_000010185.1	NC_010376.1
<i>Fusobacterium nucleatum</i>	CP007062.1, AE009951.2	NZ_CP007062.1
<i>Fusobacterium periodonticum</i>	GCF_000158215.3	ACDC03
<i>Gemella sanguinis</i>	NZ_ACRY00000000.2	JNJO01
<i>Granulicatella adiacens</i>	GCF_000160675.1	ACKZ01
<i>Granulicatella elegans</i>	NZ_ACRF00000000.2	ACRF02
<i>Haemophilus parainfluenzae</i>	FQ312002.1	NZ_CP031477.1
<i>Johnsonella ignava</i>	GCF_000235445.1	ACZL01
<i>Kingella denitrificans</i>	NZ_AEWV00000000.1	AEWV01
<i>Kingella oralis</i>	NZ_ACJW00000000.2	ACJW02
<i>Klebsiella pneumoniae</i>	CP000964.1	NZ_WKJE00000000.1
<i>Lachnoanaerobaculum saburreum</i>	GCF_000185385.1	LSDA01
<i>Lactobacillus brevis</i>	CP000416.1	NC_008497.1
<i>Leptotrichia buccalis</i>	NC_013192.1, CP001685.1	NC_013192.1
<i>Methanobrevibacter oralis</i>	GCF_001639275.1	LWMU01
<i>Mogibacterium timidum</i>	GCF_000525775.1	JALU01
<i>Mycobacterium avium</i>	NC_008595.1, NC_021200.1	AGAQ01
<i>Neisseria bacilliformis</i>	GCF_000194925.1	AFAY01
<i>Neisseria elongata</i>	GCF_000818035.1	NZ_CP007726.1
<i>Neisseria lactamica</i>	GCF_000196295.1	NC_014752.1
<i>Neisseria sicca</i>	NZ_ACKO00000000.2	NZ_CP020452.2
<i>Olsenella uli</i>	GCF_000143845.1	NC_014363.1

<i>Parvimonas micra</i>	GCF_000800295.1	NZ_CP009761.1
<i>Peptoniphilus lacrimalis</i>	NZ_ADDO00000000.1	ARKX01
<i>Peptostreptococcus anaerobius</i>	NZ_ADJN00000000.1	ARMA01
<i>Peptostreptococcus stomatis</i>	GCF_000147675.1	ADGQ01
<i>Porphyromonas asaccharolytica</i>	NZ_AENO00000000.1	NC_015501.1
<i>Porphyromonas endodontalis</i>	GCF_000174815.1	ACNN01
<i>Porphyromonas gingivalis</i>	NC_010729.1, CP013131.1, CP012889.1, AP009380.1	NC_010729.1
<i>Prevotella denticola</i>	CP002589.1	NC_015311.1
<i>Prevotella intermedia</i>	NZ_JAEZ00000000.1, GCF_000261025.1	NZ_CP019301.1
<i>Prevotella loescheii</i>	GCF_000378085.1	ARJO01
<i>Prevotella maculosa</i>	NZ_ARNR00000000.1	ARNR01
<i>Prevotella micans</i>	NZ_BAKH00000000.1	BAKH01
<i>Prevotella nigrescens</i>	NZ_AFPX00000000.1	AFPX01
<i>Prevotella oralis</i>	NZ_AEPE00000000.2	AEPE02
<i>Prevotella oris</i>	NZ_BAJC00000000.1	BAJC01
<i>Prevotella oulorum</i>	NZ_ADGI00000000.1	BAJQ01
<i>Prevotella saccharolytica</i>	GCF_000614205.1	BAKN01
<i>Prevotella salivae</i>	AEQO01000001.1	AEQO01
<i>Proteus mirabilis</i>	NZ_ACLE00000000.1	NC_010554.1
<i>Pseudomonas fluorescens</i>	GCF_000237065.1	NC_016830.1
<i>Pseudomonas putida</i>	NC_015733.1	JBOP01
<i>Pseudoramibacter alactolyticus</i>	GCF_000185505.1	AEQN01
<i>Pyramidobacter pisolens</i>	GCF_000177335.1	ADFP01
<i>Rothia dentocariosa</i>	NC_014643.1, CP002280.1	NC_014643.1

<i>Rothia mucilaginosa</i>	NZ_ACVO00000000.1	NC_013715.1
<i>Scardovia wiggsiae</i>	NZ_AKCI00000000.1	AKCI01
<i>Selenomonas sputigena</i>	NZ_ACKP00000000.2	NC_015437.1
<i>Shewanella denitrificans</i>	GCF_000013765.1	NC_007954.1
<i>Slackia exigua</i>	GCF_000162875.1	ACUX02
<i>Solobacterium moorei</i>	NZ_AECQ00000000.1	AUKY01
<i>Staphylococcus saprophyticus</i>	NC_007350.1	JUTO01
<i>Stomatobaculum longum</i>	NZ_AGEL00000000.1	AGEL01
<i>Streptococcus anginosus</i>	NZ_NARD00000000.1	NC_022239.1
<i>Streptococcus anginosus</i>	NZ_NARD00000000.1	NZ_AP013072.1
<i>Streptococcus gordonii</i>	GCF_000017005.1	NC_009785.1
<i>Streptococcus gordonii</i>	GCF_000017005.1	NZ_CP012648.1
<i>Streptococcus intermedius</i>	AP010969.1	NC_022246.1
<i>Streptococcus intermedius</i>	AP010969.1	NZ_AP014880.1
<i>Streptococcus mitis</i>	FN568063.1	NZ_CP014326.1
<i>Streptococcus oralis</i>	GCF_000344275.1	NZ_CP023507.1
<i>Streptococcus parasanguinis</i>	CP002843.1	NC_015678.1
<i>Streptococcus parasanguinis</i>	CP002843.1	NC_017905.1
<i>Streptococcus salivarius</i>	FR873481.1	NZ_CP009913.1
<i>Streptococcus salivarius</i>	FR873481.1	NZ_CP020451.2
<i>Streptococcus sanguinis</i>	CP000387.1	NZ_APLR00000000.1
<i>Streptococcus sobrinus</i>	NZ_AGGP00000000.1	JMLC01
<i>Streptococcus sobrinus</i>	NZ_AGGP00000000.1	NZ_CP029490.1
<i>Streptomyces scopuliridis</i>	GCF_000718095.1	JOEI01
<i>Streptomyces yeochonensis</i>	GCF_000745345.1	JQNR01
<i>Tannerella forsythia</i>	GCF_000238215.1, NZ_MEHX00000000.1, NZ_FMML00000000.1, NZ_FMMM00000000.1, NZ_FMMN00000000.1	NC_016610.1

<i>Treponema lecithinolyticum</i>	NZ_AWVH000000000.1	AWVH01
<i>Treponema maltophilum</i>	GCF_000413055.1	ATFF01
<i>Treponema socranskii</i>	GCF_000413015.1	ATFD01
<i>Treponema vincentii</i>	NZ_ACYH000000000.1	ATFC01
<i>Veillonella parvula</i>	NZ_AEXI000000000.1, GCF_000024945.1, NZ_ADFU000000000.1, ASKE01000001.1	NC_013520.1
<i>Vibrio vulnificus</i>	GCF_000009745.1	NZ_CP014637.1
<i>Yersinia pestis</i>	NC_003143.1, CP000308.1	NZ_CP033699.1
<i>Yersinia pseudotuberculosis</i>	NC_010465.1	NZ_CP033713.1

Supplemental table S7. False positive species, whose RefSeq reference genomes contain human contamination.

False positive species	WGS accession	Accession containing human scaffolds
<i>Paeniglutamicibacter antarcticus</i>	CZJT01	yes
<i>Alcanivorax hongdengensis</i>	AMRJ01	yes
<i>Sanguibacteroides justesenii</i>	JPIU01	yes
<i>Nocardia brevicatena</i>	BAFU01	yes
<i>Thioflexothrix pseukupsi</i>	MSLT01	yes
<i>Gemmata obscuriglobus</i>	ABGO01	yes

Supplemental table S8. Table of species simulated for the Oral Microbiome dataset. Samples containing an 'l' character in their names contain species in various abundances, whereas samples with an 'e' in their name contain species in constant abundance throughout. Genus negative control samples have the whole *Corynebacterium* genus removed, whereas species negative controls do not include *Corynebacterium diphtheriae*.

Species - Positive controls	l samp les abun danc es	e sampl es abun danc es	Species - Negative controls genus	l samp les abun danc es	e sampl es abun danc es	Species - Negative controls species	l samp les abun danc es	e sampl es abun danc es
<i>Abiotrophia defectiva</i>	0.008	0.005	<i>Abiotrophia defectiva</i>	0.008	0.005	<i>Abiotrophia defectiva</i>	0.008	0.005
<i>Actinobaculum sp oral taxon 183</i>	0.006	0.005	<i>Actinobaculum sp oral taxon 183</i>	0.006	0.005	<i>Actinobaculum sp oral taxon 183</i>	0.006	0.005
<i>Actinomyces cardiffensis</i>	0.004	0.005	<i>Actinomyces cardiffensis</i>	0.004	0.005	<i>Actinomyces cardiffensis</i>	0.004	0.005
<i>Actinomyces dentalis DSM 19115</i>	0.03	0.005	<i>Actinomyces dentalis</i>	0.03	0.005	<i>Actinomyces dentalis</i>	0.03	0.005
<i>Actinomyces massiliensis</i>	0.003	0.005	<i>Actinomyces massiliensis</i>	0.003	0.005	<i>Actinomyces massiliensis</i>	0.003	0.005
<i>Actinomyces oris</i>	0.006	0.005	<i>Actinomyces oris</i>	0.006	0.005	<i>Actinomyces oris</i>	0.006	0.005
<i>Actinomyces sp oral taxon 175</i>	0.002	0.005	<i>Actinomyces sp oral taxon 170</i>	0.005	0.005	<i>Actinomyces sp oral taxon 175</i>	0.002	0.005
<i>Actinomyces sp oral taxon 181</i>	0.001	0.005	<i>Actinomyces sp oral taxon 175</i>	0.002	0.005	<i>Actinomyces sp oral taxon 181</i>	0.001	0.005
<i>Actinomyces sp oral taxon 848</i>	0.009	0.005	<i>Actinomyces sp oral taxon 178</i>	0.005	0.005	<i>Actinomyces sp oral taxon 848</i>	0.009	0.005
<i>Aggregatibacter actinomycetemcomitans</i>	0.007	0.01	<i>Actinomyces sp oral taxon 181</i>	0.001	0.005	<i>Aggregatibacter actinomycetemcomitans</i>	0.007	0.01
<i>Aggregatibacter</i>	0.009	0.005	<i>Actinomyces sp</i>	0.009	0.005	<i>Aggregatibacter</i>	0.009	0.005

<i>aphrophilus</i>			<i>oral taxon 848</i>			<i>aphrophilus</i>		
<i>Aggregatibacter segnis</i>	0.001	0.005	<i>Aggregatibacter actinomycetemc omitans</i>	0.007	0.01	<i>Aggregatibacter segnis</i>	0.001	0.005
<i>Alloprevotella tannerae</i>	0.007	0.005	<i>Aggregatibacter aphrophilus</i>	0.009	0.005	<i>Alloprevotella tannerae</i>	0.007	0.005
<i>Anaeroglobus geminatus</i>	0.004	0.005	<i>Aggregatibacter segnis</i>	0.001	0.005	<i>Anaeroglobus geminatus</i>	0.004	0.005
<i>Atopobium parvulum</i>	0.005	0.005	<i>Agrobacterium tumefaciens</i>	0.003	0.005	<i>Atopobium parvulum</i>	0.005	0.005
<i>Atopobium rimae</i>	0.003	0.005	<i>Alloprevotella tannerae</i>	0.007	0.005	<i>Atopobium rimae</i>	0.003	0.005
<i>Atopobium sp oral taxon 810</i>	0.002	0.005	<i>Anaeroglobus geminatus</i>	0.004	0.005	<i>Atopobium sp oral taxon 810</i>	0.002	0.005
<i>Bacillus subtilis</i>	0.001	0.005	<i>Atopobium parvulum</i>	0.005	0.005	<i>Bacillus subtilis</i>	0.001	0.005
<i>Bacteroidetes bacterium oral taxon 272</i>	0.005	0.005	<i>Atopobium rimae</i>	0.003	0.005	<i>Bacteroidetes bacterium oral taxon 272</i>	0.005	0.005
<i>Bacteroidetes oral taxon 274</i>	0.002	0.005	<i>Atopobium sp oral taxon 810</i>	0.002	0.005	<i>Bacteroidetes oral taxon 274</i>	0.002	0.005
<i>Bifidobacterium dentium</i>	0.006	0.005	<i>Bacillus subtilis</i>	0.001	0.005	<i>Bifidobacterium dentium</i>	0.006	0.005
<i>Bifidobacterium longum</i>	0.001	0.005	<i>Bacteroidetes bacterium oral taxon 272</i>	0.005	0.005	<i>Bifidobacterium longum</i>	0.001	0.005
<i>Bordetella bronchiseptica</i>	0.004	0.005	<i>Bacteroidetes oral taxon 274</i>	0.002	0.005	<i>Bordetella bronchiseptica</i>	0.004	0.005
<i>Bordetella parapertussis</i>	0.001	0.005	<i>Bifidobacterium dentium</i>	0.006	0.005	<i>Bordetella parapertussis</i>	0.001	0.005
<i>Bordetella pertussis</i>	0.001	0.005	<i>Bifidobacterium longum</i>	0.001	0.005	<i>Bordetella pertussis</i>	0.001	0.005
<i>Burkholderia mallei</i>	0.001	0.005	<i>Bordetella bronchiseptica</i>	0.004	0.005	<i>Burkholderia mallei</i>	0.001	0.005
<i>Burkholderia pseudomallei</i>	0.001	0.005	<i>Bordetella parapertussis</i>	0.001	0.005	<i>Burkholderia pseudomallei</i>	0.001	0.005

<i>Campylobacter concisus</i>	0.005	0.005	<i>Bordetella pertussis</i>	0.001	0.005	<i>Campylobacter concisus</i>	0.005	0.005
<i>Campylobacter curvus</i>	0.002	0.005	<i>Burkholderia mallei</i>	0.001	0.005	<i>Campylobacter curvus</i>	0.002	0.005
<i>Campylobacter gracilis</i>	0.002	0.005	<i>Burkholderia pseudomallei</i>	0.001	0.005	<i>Campylobacter gracilis</i>	0.002	0.005
<i>Campylobacter rectus</i>	0.004	0.005	<i>Campylobacter concisus</i>	0.005	0.005	<i>Campylobacter rectus</i>	0.004	0.005
<i>Campylobacter showae</i>	0.005	0.005	<i>Campylobacter curvus</i>	0.002	0.005	<i>Campylobacter showae</i>	0.005	0.005
Candidate division SR1 bacterium MGEHA 2517589202 0	0.008	0.005	<i>Campylobacter gracilis</i>	0.002	0.005	Candidate_division SR1 bacterium MGEHA 2517589202 0	0.008	0.005
Candidate division TM7 single cell isolate TM7a	0.002	0.005	<i>Campylobacter rectus</i>	0.004	0.005	Candidate division TM7 single cell isolate TM7a	0.002	0.005
candidate division TM7 single cell isolate TM7b	0.006	0.005	<i>Campylobacter showae</i>	0.005	0.005	candidate division TM7 single cell isolate TM7b	0.006	0.005
candidate division TM7 single cell isolate TM7c	0.003	0.005	Candidate division SR1 bacterium MGEHA 2517589202 0	0.008	0.005	candidate division TM7 single cell isolate TM7c	0.003	0.005
<i>Capnocytophaga gingivalis</i>	0.01	0.005	Candidate division TM7 single cell isolate TM7a	0.002	0.005	<i>Capnocytophaga gingivalis</i>	0.01	0.005
<i>Capnocytophaga granulosa</i>	0.002	0.005	candidate division TM7 single cell isolate TM7b	0.006	0.005	<i>Capnocytophaga granulosa</i>	0.002	0.005
<i>Capnocytophaga ochracea</i>	0.002	0.005	candidate division TM7	0.003	0.005	<i>Capnocytophaga ochracea</i>	0.002	0.005

			single cell isolate TM7c					
<i>Capnocytophaga</i> <i>sp oral taxon</i> 329	0.005	0.005	<i>Capnocytophaga</i> <i>gingivalis</i>	0.01	0.005	<i>Capnocytophaga</i> <i>sp oral taxon</i> 329	0.005	0.005
<i>Capnocytophaga</i> <i>sp oral taxon</i> 332	0.004	0.005	<i>Capnocytophaga</i> <i>granulosa</i>	0.002	0.005	<i>Capnocytophaga</i> <i>sp oral taxon</i> 332	0.004	0.005
<i>Capnocytophaga</i> <i>sputigena</i>	0.004	0.005	<i>Capnocytophaga</i> <i>ochracea</i>	0.002	0.005	<i>Capnocytophaga</i> <i>sputigena</i>	0.004	0.005
<i>Cardiobacterium</i> <i>hominis</i>	0.003	0.005	<i>Capnocytophaga</i> <i>sp oral taxon</i> 329	0.005	0.005	<i>Cardiobacterium</i> <i>hominis</i>	0.003	0.005
<i>Cardiobacterium</i> <i>valvarum</i>	0.009	0.005	<i>Capnocytophaga</i> <i>sp oral taxon</i> 332	0.004	0.005	<i>Cardiobacterium</i> <i>valvarum</i>	0.009	0.005
<i>Catonella morbi</i>	0.01	0.005	<i>Capnocytophaga</i> <i>sputigena</i>	0.004	0.005	<i>Catonella morbi</i>	0.01	0.005
<i>Clostridium</i> <i>botulinum</i>	0.001	0.005	<i>Cardiobacterium</i> <i>hominis</i>	0.003	0.005	<i>Clostridium</i> <i>botulinum</i>	0.001	0.005
<i>Clostridium</i> <i>tetani</i>	0.005	0.005	<i>Cardiobacterium</i> <i>valvarum</i>	0.009	0.005	<i>Clostridium</i> <i>tetani</i>	0.005	0.005
<i>Corynebacteriu</i> <i>m diphtheriae</i>	0.001	0.005	<i>Catonella morbi</i>	0.01	0.005	<i>Klebsiella</i> <i>pneumoniae</i>	0.001	0.005
<i>Corynebacteriu</i> <i>m durum</i>	0.01	0.005	<i>Clostridium</i> <i>botulinum</i>	0.001	0.005	<i>Corynebacteriu</i> <i>m durum</i>	0.01	0.005
<i>Corynebacteriu</i> <i>m matruchotii</i>	0.01	0.005	<i>Clostridium</i> <i>tetani</i>	0.005	0.005	<i>Corynebacteriu</i> <i>m matruchotii</i>	0.01	0.005
<i>Corynebacteriu</i> <i>m urealyticum</i>	0.005	0.005	<i>Desulfobulbus</i> <i>sp oral taxon</i> 041	0.01	0.005	<i>Corynebacteriu</i> <i>m urealyticum</i>	0.005	0.005
<i>Desulfobulbus</i> <i>sp oral taxon</i> 041	0.01	0.005	<i>Dialister invisus</i>	0.009	0.005	<i>Desulfobulbus</i> <i>sp oral taxon</i> 041	0.01	0.005
<i>Dialister invisus</i>	0.009	0.005	<i>Eggerthella lenta</i>	0.002	0.005	<i>Dialister invisus</i>	0.009	0.005
<i>Eggerthella lenta</i>	0.002	0.005	<i>Eikenella</i> <i>corrodens</i>	0.01	0.005	<i>Eggerthella lenta</i>	0.002	0.005

<i>Eikenella</i> <i>corrodens</i>	0.01	0.005	<i>Eubacterium</i> <i>brachy</i>	0.002	0.005	<i>Eikenella</i> <i>corrodens</i>	0.01	0.005
<i>Eubacterium</i> <i>brachy</i>	0.002	0.005	<i>Eubacterium</i> <i>saphenum</i>	0.001	0.005	<i>Eubacterium</i> <i>brachy</i>	0.002	0.005
<i>Eubacterium</i> <i>saphenum</i>	0.001	0.005	<i>Eubacterium</i> <i>yurii</i>	0.005	0.005	<i>Eubacterium</i> <i>saphenum</i>	0.001	0.005
<i>Eubacterium</i> <i>yurii</i>	0.005	0.005	<i>Filifactor</i> <i>alocis</i>	0.01	0.005	<i>Eubacterium</i> <i>yurii</i>	0.005	0.005
<i>Filifactor</i> <i>alocis</i>	0.01	0.005	<i>Finegoldia</i> <i>magna</i>	0.007	0.005	<i>Filifactor</i> <i>alocis</i>	0.01	0.005
<i>Finegoldia</i> <i>magna</i>	0.007	0.005	<i>Fretibacterium</i> <i>fastidiosum</i>	0.007	0.005	<i>Finegoldia</i> <i>magna</i>	0.007	0.005
<i>Fretibacterium</i> <i>fastidiosum</i>	0.007	0.005	<i>Fusobacterium</i> <i>nucleatum</i>	0.035	0.02	<i>Fretibacterium</i> <i>fastidiosum</i>	0.007	0.005
<i>Fusobacterium</i> <i>nucleatum</i>	0.035	0.02	<i>Fusobacterium</i> <i>periodonticum</i>	0.005	0.005	<i>Fusobacterium</i> <i>nucleatum</i>	0.035	0.02
<i>Fusobacterium</i> <i>periodonticum</i>	0.005	0.005	<i>Gemella</i> <i>morbilloorum</i>	0.002	0.005	<i>Fusobacterium</i> <i>periodonticum</i>	0.005	0.005
<i>Gemella</i> <i>morbilloorum</i>	0.002	0.005	<i>Gemella</i> <i>sanguinis</i>	0.009	0.005	<i>Gemella</i> <i>morbilloorum</i>	0.002	0.005
<i>Gemella</i> <i>sanguinis</i>	0.009	0.005	<i>Granulicatella</i> <i>adiacens</i>	0.009	0.005	<i>Gemella</i> <i>sanguinis</i>	0.009	0.005
<i>Granulicatella</i> <i>adiacens</i>	0.009	0.005	<i>Granulicatella</i> <i>elegans</i>	0.002	0.005	<i>Granulicatella</i> <i>adiacens</i>	0.009	0.005
<i>Granulicatella</i> <i>elegans</i>	0.002	0.005	<i>Haemophilus</i> <i>ducreyi</i>	0.001	0.005	<i>Granulicatella</i> <i>elegans</i>	0.002	0.005
<i>Haemophilus</i> <i>parainfluenzae</i>	0.006	0.005	<i>Haemophilus</i> <i>parainfluenzae</i>	0.006	0.005	<i>Haemophilus</i> <i>parainfluenzae</i>	0.006	0.005
<i>Johnsonella</i> <i>ignava</i>	0.001	0.005	<i>Johnsonella</i> <i>ignava</i>	0.001	0.005	<i>Johnsonella</i> <i>ignava</i>	0.001	0.005
<i>Kingella</i> <i>denitrificans</i>	0.004	0.005	<i>Kingella</i> <i>denitrificans</i>	0.004	0.005	<i>Kingella</i> <i>denitrificans</i>	0.004	0.005
<i>Kingella</i> <i>oralis</i>	0.007	0.005	<i>Kingella</i> <i>kingeae</i>	0.005	0.005	<i>Kingella</i> <i>oralis</i>	0.007	0.005
<i>Klebsiella</i> <i>pneumoniae</i>	0.003	0.005	<i>Kingella</i> <i>oralis</i>	0.007	0.005	<i>Klebsiella</i> <i>pneumoniae</i>	0.003	0.005
<i>Lachnoanaeroba</i>	0.01	0.005	<i>Klebsiella</i>	0.004	0.01	<i>Lachnoanaeroba</i>	0.01	0.005

<i>culum saburreum</i>			<i>pneumoniae</i>			<i>culum saburreum</i>		
<i>Lachnospiraceae</i> oral taxon 107	0.005	0.005	<i>Lachnoanaerobaculum saburreum</i>	0.01	0.005	<i>Lachnospiraceae</i> oral taxon 107	0.005	0.005
<i>Lactobacillus brevis</i>	0.002	0.005	<i>Lachnospiraceae</i> oral taxon 107	0.005	0.005	<i>Lactobacillus brevis</i>	0.002	0.005
<i>Lautropia mirabilis</i>	0.01	0.005	<i>Lactobacillus acidophilus</i>	0.001	0.005	<i>Lautropia mirabilis</i>	0.01	0.005
<i>Leptotrichia buccalis</i>	0.006	0.005	<i>Lactobacillus brevis</i>	0.002	0.005	<i>Leptotrichia buccalis</i>	0.006	0.005
<i>Leptotrichia sp</i> oral taxon 215	0.002	0.005	<i>Lactobacillus salivarius</i>	0.01	0.005	<i>Leptotrichia sp</i> oral taxon 215	0.002	0.005
<i>Methanobrevibacter unclassified</i>	0.03	0.005	<i>Lautropia mirabilis</i>	0.01	0.005	<i>Methanobrevibacter unclassified</i>	0.03	0.005
<i>Mogibacterium sp</i> CM50	0.01	0.005	<i>Leptotrichia buccalis</i>	0.006	0.005	<i>Mogibacterium sp</i> CM50	0.01	0.005
<i>Mycobacterium avium</i>	0.009	0.01	<i>Leptotrichia shahii</i>	0.004	0.005	<i>Mycobacterium avium</i>	0.009	0.01
<i>Mycobacterium bovis</i>	0.005	0.005	<i>Leptotrichia sp</i> oral taxon 215	0.002	0.005	<i>Mycobacterium bovis</i>	0.005	0.005
<i>Mycobacterium smegmatis</i>	0.003	0.005	<i>Methanobrevibacter unclassified</i>	0.03	0.005	<i>Mycobacterium smegmatis</i>	0.003	0.005
<i>Mycobacterium tuberculosis</i>	0.001	0.005	<i>Microbacterium sp</i> oral taxon 186	0.001	0.005	<i>Haemophilus ducreyi</i>	0.001	0.005
<i>Neisseria bacilliformis</i>	0.001	0.005	<i>Mogibacterium sp</i> CM50	0.01	0.005	<i>Neisseria bacilliformis</i>	0.001	0.005
<i>Neisseria elongata</i>	0.007	0.005	<i>Neisseria bacilliformis</i>	0.001	0.005	<i>Neisseria elongata</i>	0.007	0.005
<i>Neisseria lactamica</i>	0.02	0.005	<i>Neisseria elongata</i>	0.007	0.005	<i>Neisseria lactamica</i>	0.02	0.005
<i>Neisseria sicca</i>	0.009	0.005	<i>Neisseria gonorrhoeae</i>	0.001	0.005	<i>Neisseria sicca</i>	0.009	0.005
<i>Olsenella uli</i>	0.007	0.005	<i>Neisseria lactamica</i>	0.02	0.005	<i>Olsenella uli</i>	0.007	0.005

<i>Parvimonas micra</i>	0.008	0.005	<i>Neisseria meningitidis</i>	0.01	0.005	<i>Parvimonas micra</i>	0.008	0.005
<i>Parvimonas sp oral taxon 110</i>	0.001	0.005	<i>Neisseria sicca</i>	0.009	0.005	<i>Parvimonas sp oral taxon 110</i>	0.001	0.005
<i>Parvimonas sp oral taxon 393</i>	0.004	0.005	<i>Olsenella uli</i>	0.007	0.005	<i>Parvimonas sp oral taxon 393</i>	0.004	0.005
<i>Peptoniphilus lacrimalis</i>	0.003	0.005	<i>Parvimonas micra</i>	0.008	0.005	<i>Peptoniphilus lacrimalis</i>	0.003	0.005
<i>Peptoniphilus sp oral taxon 386</i>	0.002	0.005	<i>Parvimonas sp oral taxon 110</i>	0.001	0.005	<i>Peptoniphilus sp oral taxon 386</i>	0.002	0.005
<i>Peptostreptococcus anaerobius</i>	0.005	0.005	<i>Parvimonas sp oral taxon 393</i>	0.004	0.005	<i>Peptostreptococcus anaerobius</i>	0.005	0.005
<i>Peptostreptococcus stomatis</i>	0.002	0.005	<i>Peptoniphilus lacrimalis</i> 315 B	0.003	0.005	<i>Peptostreptococcus stomatis</i>	0.002	0.005
<i>Porphyromonas asaccharolytica</i>	0.006	0.005	<i>Peptoniphilus sp oral taxon 386</i>	0.002	0.005	<i>Porphyromonas asaccharolytica</i>	0.006	0.005
<i>Porphyromonas endodontalis</i>	0.008	0.005	<i>Peptostreptococcus anaerobius</i>	0.005	0.005	<i>Porphyromonas endodontalis</i>	0.008	0.005
<i>Porphyromonas gingivalis</i>	0.016	0.02	<i>Peptostreptococcus stomatis</i>	0.002	0.005	<i>Porphyromonas gingivalis</i>	0.016	0.02
<i>Porphyromonas sp oral taxon 278</i>	0.004	0.005	<i>Porphyromonas asaccharolytica</i>	0.006	0.005	<i>Porphyromonas sp oral taxon 278</i>	0.004	0.005
<i>Prevotella denticola</i>	0.004	0.005	<i>Porphyromonas endodontalis</i>	0.008	0.005	<i>Prevotella denticola</i>	0.004	0.005
<i>Prevotella intermedia</i>	0.008	0.01	<i>Porphyromonas gingivalis</i>	0.016	0.02	<i>Prevotella intermedia</i>	0.008	0.01
<i>Prevotella loescheii</i>	0.002	0.005	<i>Porphyromonas sp oral taxon 278</i>	0.004	0.005	<i>Prevotella loescheii</i>	0.002	0.005
<i>Prevotella maculosa</i>	0.001	0.005	<i>Prevotella denticola</i>	0.004	0.005	<i>Prevotella maculosa</i>	0.001	0.005
<i>Prevotella melaninogenica</i>	0.004	0.005	<i>Prevotella intermedia</i>	0.008	0.01	<i>Prevotella melaninogenica</i>	0.004	0.005
<i>Prevotella micans</i>	0.003	0.005	<i>Prevotella loescheii</i>	0.002	0.005	<i>Prevotella micans</i>	0.003	0.005

<i>Prevotella nigrescens</i>	0.007	0.005	<i>Prevotella maculosa</i>	0.001	0.005	<i>Prevotella nigrescens</i>	0.007	0.005
<i>Prevotella oralis</i>	0.02	0.005	<i>Prevotella marshii</i>	0.001	0.005	<i>Prevotella oralis</i>	0.02	0.005
<i>Prevotella oris</i>	0.002	0.005	<i>Prevotella melaninogenica</i>	0.004	0.005	<i>Prevotella oris</i>	0.002	0.005
<i>Prevotella oulorum</i>	0.002	0.005	<i>Prevotella micans</i>	0.003	0.005	<i>Prevotella oulorum</i>	0.002	0.005
<i>Prevotella saccharolytica</i>	0.001	0.005	<i>Prevotella nigrescens</i>	0.007	0.005	<i>Prevotella saccharolytica</i>	0.001	0.005
<i>Prevotella salivae</i>	0.001	0.005	<i>Prevotella oralis</i>	0.02	0.005	<i>Prevotella salivae</i>	0.001	0.005
<i>Prevotella sp oral taxon 299</i>	0.004	0.005	<i>Prevotella oris</i>	0.002	0.005	<i>Prevotella sp oral taxon 299</i>	0.004	0.005
<i>Prevotella sp oral taxon 317</i>	0.004	0.005	<i>Prevotella oulorum</i>	0.002	0.005	<i>Prevotella sp oral taxon 317</i>	0.004	0.005
<i>Prevotella sp oral taxon 473</i>	0.003	0.005	<i>Prevotella saccharolytica</i>	0.001	0.005	<i>Prevotella sp oral taxon 473</i>	0.003	0.005
<i>Propionibacterium propionicum</i>	0.005	0.005	<i>Prevotella salivae</i>	0.001	0.005	<i>Propionibacterium propionicum</i>	0.005	0.005
<i>Proteus mirabilis</i>	0.002	0.005	<i>Prevotella sp oral taxon 299</i>	0.004	0.005	<i>Proteus mirabilis</i>	0.002	0.005
<i>Pseudomonas aeruginosa</i>	0.001	0.005	<i>Prevotella sp oral taxon 317</i>	0.004	0.005	<i>Pseudomonas aeruginosa</i>	0.001	0.005
<i>Pseudomonas fluorescens</i>	0.001	0.005	<i>Prevotella sp oral taxon 473</i>	0.003	0.005	<i>Pseudomonas fluorescens</i>	0.001	0.005
<i>Pseudomonas putida</i>	0.001	0.005	<i>Prevotella veroralis</i>	0.001	0.005	<i>Pseudomonas putida</i>	0.001	0.005
<i>Pseudomonas syringae</i>	0.004	0.005	<i>Propionibacterium avidum</i>	0.001	0.005	<i>Pseudomonas syringae</i>	0.004	0.005
<i>Pseudoramibacter alactolyticus</i>	0.003	0.005	<i>Propionibacterium propionicum</i>	0.005	0.005	<i>Pseudoramibacter alactolyticus</i>	0.003	0.005
<i>Pyramidobacter pisolens</i>	0.007	0.005	<i>Proteus mirabilis</i>	0.002	0.005	<i>Pyramidobacter pisolens</i>	0.007	0.005
<i>Rothia aeria</i>	0.005	0.005	<i>Pseudomonas aeruginosa</i>	0.001	0.005	<i>Rothia aeria</i>	0.005	0.005

<i>Rothia dentocariosa</i>	0.011	0.01	<i>Pseudomonas fluorescens</i>	0.001	0.005	<i>Rothia dentocariosa</i>	0.011	0.01
<i>Rothia mucilaginosa</i>	0.002	0.005	<i>Pseudomonas putida</i>	0.001	0.005	<i>Rothia mucilaginosa</i>	0.002	0.005
<i>Salmonella enterica</i>	0.004	0.02	<i>Pseudomonas syringae</i>	0.004	0.005	<i>Lactobacillus acidophilus</i>	0.001	0.005
<i>Scardovia wiggisiae</i>	0.002	0.005	<i>Pseudoramibacter alactolyticus</i>	0.003	0.005	<i>Salmonella enterica</i>	0.003	0.015
<i>Selenomonas flueggei</i>	0.004	0.005	<i>Pyramidobacter piscolens</i>	0.007	0.005	<i>Scardovia wiggisiae</i>	0.002	0.005
<i>Selenomonas noxia</i>	0.007	0.005	<i>Rothia aerea</i>	0.005	0.005	<i>Selenomonas flueggei</i>	0.004	0.005
<i>Selenomonas sp oral taxon 138</i>	0.003	0.005	<i>Rothia dentocariosa</i>	0.011	0.01	<i>Selenomonas noxia</i>	0.007	0.005
<i>Selenomonas sp oral taxon 149</i>	0.002	0.005	<i>Rothia mucilaginosa</i>	0.002	0.005	<i>Selenomonas sp oral taxon 138</i>	0.003	0.005
<i>Selenomonas sputigena</i>	0.009	0.005	<i>Scardovia wiggisiae</i>	0.002	0.005	<i>Selenomonas sp oral taxon 149</i>	0.002	0.005
<i>Serratia odorifera</i>	0.001	0.005	<i>Selenomonas artemidis</i>	0.004	0.005	<i>Selenomonas sputigena</i>	0.009	0.005
<i>Shewanella denitrificans</i>	0.001	0.005	<i>Selenomonas flueggei</i>	0.004	0.005	<i>Serratia odorifera</i>	0.001	0.005
<i>Slackia exigua</i>	0.004	0.005	<i>Selenomonas noxia</i>	0.007	0.005	<i>Shewanella denitrificans</i>	0.001	0.005
<i>Solobacterium moorei</i>	0.01	0.005	<i>Selenomonas sp oral taxon 138</i>	0.003	0.005	<i>Slackia exigua</i>	0.004	0.005
<i>Staphylococcus aureus</i>	0.003	0.005	<i>Selenomonas sp oral taxon 149</i>	0.002	0.005	<i>Solobacterium moorei</i>	0.01	0.005
<i>Staphylococcus epidermidis</i>	0.002	0.005	<i>Selenomonas sputigena</i>	0.009	0.005	<i>Staphylococcus aureus</i>	0.003	0.005
<i>Staphylococcus saprophyticus</i>	0.001	0.005	<i>Serratia odorifera</i>	0.001	0.005	<i>Staphylococcus epidermidis</i>	0.002	0.005
<i>Stomatobaculum longum</i>	0.009	0.005	<i>Shewanella denitrificans</i>	0.001	0.005	<i>Staphylococcus saprophyticus</i>	0.001	0.005
<i>Streptococcus anginosus</i>	0.004	0.005	<i>Slackia exigua</i>	0.004	0.005	<i>Stomatobaculum longum</i>	0.009	0.005

<i>Streptococcus constellatus</i>	0.001	0.005	<i>Solobacterium moorei</i>	0.01	0.005	<i>Streptococcus anginosus</i>	0.004	0.005
<i>Streptococcus gordonii</i>	0.008	0.005	<i>Staphylococcus aureus</i>	0.003	0.005	<i>Streptococcus constellatus</i>	0.001	0.005
<i>Streptococcus intermedius</i>	0.002	0.005	<i>Staphylococcus epidermidis</i>	0.002	0.005	<i>Streptococcus gordonii</i>	0.008	0.005
<i>Streptococcus mitis</i>	0.006	0.005	<i>Staphylococcus saprophyticus</i>	0.001	0.005	<i>Streptococcus intermedius</i>	0.002	0.005
<i>Streptococcus oligofermentans</i>	0.002	0.005	<i>Stomatobaculum longum</i>	0.009	0.005	<i>Streptococcus mitis</i>	0.006	0.005
<i>Streptococcus oralis</i>	0.03	0.005	<i>Streptococcus anginosus</i>	0.004	0.005	<i>Streptococcus oligofermentans</i>	0.002	0.005
<i>Streptococcus parasanguinis</i>	0.001	0.005	<i>Streptococcus constellatus</i>	0.001	0.005	<i>Streptococcus oralis</i>	0.03	0.005
<i>Streptococcus salivarius</i>	0.004	0.005	<i>Streptococcus gordonii</i>	0.008	0.005	<i>Streptococcus parasanguinis</i>	0.001	0.005
<i>Streptococcus sanguinis</i>	0.005	0.005	<i>Streptococcus intermedius</i>	0.002	0.005	<i>Streptococcus salivarius</i>	0.004	0.005
<i>Streptococcus sobrinus</i>	0.002	0.005	<i>Streptococcus mitis</i>	0.006	0.005	<i>Streptococcus sanguinis</i>	0.005	0.005
<i>Streptococcus sp oral taxon 056</i>	0.006	0.005	<i>Streptococcus oligofermentans</i>	0.002	0.005	<i>Streptococcus sobrinus</i>	0.002	0.005
<i>Streptococcus sp oral taxon 058</i>	0.002	0.005	<i>Streptococcus oralis</i>	0.03	0.005	<i>Streptococcus sp oral taxon 056</i>	0.006	0.005
<i>Streptomyces kanamyceticus</i>	0.001	0.005	<i>Streptococcus parasanguinis</i>	0.001	0.005	<i>Streptococcus sp oral taxon 058</i>	0.002	0.005
<i>Streptomyces scopuliridis</i>	0.004	0.005	<i>Streptococcus salivarius</i>	0.004	0.005	<i>Streptomyces kanamyceticus</i>	0.001	0.005
<i>Streptomyces yeochonensis</i>	0.002	0.005	<i>Streptococcus sanguinis</i>	0.005	0.005	<i>Streptomyces scopuliridis</i>	0.004	0.005
<i>Tannerella forsythia</i>	0.062	0.025	<i>Streptococcus sobrinus</i>	0.002	0.005	<i>Streptomyces yeochonensis</i>	0.002	0.005
<i>Tissierella bacterium</i>	0.015	0.01	<i>Streptococcus sp oral taxon</i>	0.006	0.005	<i>Tannerella forsythia</i>	0.062	0.025

			056					
<i>Treponema denticola</i>	0.018	0.015	<i>Streptococcus</i> sp oral taxon 058	0.002	0.005	<i>Tissierellia bacterium</i>	0.015	0.01
<i>Treponema lecithinolyticum</i>	0.003	0.005	<i>Streptomyces kanamyceticus</i>	0.001	0.005	<i>Treponema denticola</i>	0.018	0.015
<i>Treponema maltophilum</i>	0.004	0.005	<i>Streptomyces scopuliridis</i>	0.004	0.005	<i>Treponema lecithinolyticum</i>	0.003	0.005
<i>Treponema socranskii</i>	0.008	0.005	<i>Streptomyces yeochonensis</i>	0.002	0.005	<i>Treponema maltophilum</i>	0.004	0.005
<i>Treponema vincentii</i>	0.002	0.005	<i>Tannerella forsythia</i>	0.062	0.025	<i>Treponema socranskii</i>	0.008	0.005
<i>Veillonella atypica</i>	0.006	0.005	<i>Tissierellia bacterium</i>	0.015	0.01	<i>Treponema vincentii</i>	0.002	0.005
<i>Veillonella denticariosi</i>	0.002	0.005	<i>Treponema denticola</i>	0.018	0.015	<i>Veillonella atypica</i>	0.006	0.005
<i>Veillonella dispar</i>	0.009	0.005	<i>Treponema lecithinolyticum</i>	0.003	0.005	<i>Veillonella denticariosi</i>	0.002	0.005
<i>Veillonella oral taxon 780</i>	0.001	0.005	<i>Treponema maltophilum</i>	0.004	0.005	<i>Veillonella dispar</i>	0.009	0.005
<i>Veillonella parvula</i>	0.015	0.02	<i>Treponema pallidum</i>	0.001	0.005	<i>Veillonella oral_ axon 780</i>	0.001	0.005
<i>Veillonella sp oral taxon 158</i>	0.003	0.005	<i>Treponema socranskii</i>	0.008	0.005	<i>Veillonella parvula</i>	0.015	0.02
<i>Vibrio cholerae</i>	0.001	0.005	<i>Treponema vincentii</i>	0.002	0.005	<i>Veillonella sp oral taxon 158</i>	0.003	0.005
<i>Vibrio vulnificus</i>	0.002	0.005	<i>Veillonella atypica</i>	0.006	0.005	<i>Vibrio cholerae</i>	0.001	0.005
<i>Xanthomonas axonopodis</i>	0.002	0.005	<i>Veillonella denticariosi</i>	0.002	0.005	<i>Vibrio vulnificus</i>	0.002	0.005
<i>Yersinia enterocolitica</i>	0.001	0.005	<i>Veillonella dispar</i>	0.009	0.005	<i>Xanthomonas axonopodis</i>	0.002	0.005
<i>Yersinia pestis</i>	0.002	0.01	<i>Veillonella oral taxon 780</i>	0.001	0.005	<i>Yersinia enterocolitica</i>	0.001	0.005
<i>Yersinia pseudotuberculo</i>	0.004	0.005	<i>Veillonella parvula</i>	0.015	0.02	<i>Neisseria gonorrhoeae</i>	0.001	0.005

<i>sis YPIII</i>								
			<i>Veillonella sp</i> <i>oral taxon 158</i>	0.003	0.005	<i>Microbacterium_</i> <i>sp oral taxon</i> <i>186</i>	0.001	0.005
			<i>Vibrio cholerae</i>	0.001	0.005	<i>Yersinia</i> <i>pseudotuberculo</i> <i>sis</i>	0.004	0.005
			<i>Vibrio vulnificus</i>	0.002	0.005			
			<i>Xanthomonas</i> <i>axonopodis</i>	0.002	0.005			

Supplemental table S9. Memory and time usage data for database construction across HAYSTAC, Kraken2/Bracken, MALT/HOPS and Sigma.

Database (species)	Method	Max RSS (KB)	Runtime	Runtime s
10	haystac	90091200%	0:23.86	23.86
500	haystac	45068800%	8:38.52	518.52
100	haystac	6825316800.00%	21:01.02	1261.02
10	haystac	6341164800.00%	24:30.02	1470.02
100	haystac	675692800%	4:20.74	260.74
500	haystac	2845032000%	19:52.12	1192.12
100	haystac	3066755200.00%	5:18.10	318.1
100	haystac	45219200.00%	1:44.08	104.08
10	haystac	16373792	0:49.62	49.62
500	haystac	14312580800%	20:31.82	1231.82
500	haystac	8849920000%	31:04.95	1864.95
10	haystac	42566400.00%	0:13.23	13.23
10	haystac	89275200.00%	0:23.38	23.38
500	haystac	45299200%	8:27.07	507.07

100	haystac	6825316800%	23:55.53	1435.53
10	kraken	6341163200.00%	33:00.79	1980.79
500	kraken	2844246400.00%	19:51.89	1191.89
100	kraken	30673264	5:17.67	317.67
100	kraken	676502400%	4:29.17	269.17
100	kraken	45219200%	1:43.43	103.43
10	kraken	1557553600.00%	0:48.22	48.22
500	kraken	12898899200.00%	20:24.62	1224.62
500	kraken	8849918400%	33:09.64	1989.64
10	kraken	42566400%	0:12.47	12.47
10	kraken	86804800.00%	0:23.69	23.69
500	kraken	45068800.00%	8:25.68	505.68
100	kraken	68253168	33:46.71	2026.71
10	kraken	6341164800%	31:35.68	1895.68
100	kraken	675702400%	4:31.19	271.19
500	kraken	2846667200.00%	20:33.56	1233.56
100	malt	4303377600.00%	5:27.63	327.63
100	malt	45219200%	1:43.63	103.63
10	malt	1530721600%	0:47.63	47.63
500	malt	12679528000.00%	19:50.68	1190.68
500	malt	8849920000.00%	37:10.03	2230.03
10	malt	425680	0:12.58	12.58
10	malt	93336000%	0:24.24	24.24
500	malt	45068800%	8:25.27	505.27
100	malt	6825318400.00%	19:31.79	1171.79
10	malt	6341164800.00%	26:33.84	1593.84
100	malt	674892800%	4:38.88	278.88

500	malt	2843795200%	20:02.56	1202.56
100	malt	3071081600.00%	5:37.52	337.52
100	malt	45219200.00%	1:43.65	103.65
10	malt	15381216	0:48.60	48.6
500	sigma	13784364800%	20:42.34	1242.34
500	sigma	8849920000%	26:20.83	1580.83
10	sigma	42566400.00%	0:12.68	12.68
10	sigma	87636800.00%	0:24.26	24.26
500	sigma	45457600%	8:26.42	506.42
100	sigma	6825318400%	19:33.70	1173.7
10	sigma	6341163200.00%	16:19.25	979.25
100	sigma	673260800.00%	4:39.78	279.78
500	sigma	28458112	20:15.54	1215.54
100	sigma	3067256000%	5:19.66	319.66
100	sigma	45219200%	1:44.52	104.52
10	sigma	1527214400.00%	0:48.13	48.13
500	sigma	14329822400.00%	20:55.23	1255.23
500	sigma	8849918400%	35:57.63	2157.63
10	sigma	42566400%	0:13.17	13.17

Supplemental table S10. Memory and time usage data for sample analysis across HAYSTAC, Kraken2/Bracken, MALT/HOPS and Sigma.

Number of Reads	Database (species)	Method	Max RSS (KB)	Runtime	Runtime_s
1000000	10	haystac	31160944	21:08.57	1268.57
1000000	500	haystac	176511408	30:55.78	1855.78
1000000	500	haystac	14708544	0:06.91	6.91
1000000	10	haystac	945040	1:05.84	65.84

1000000	10	haystac	92368	0:27.29	27.29
1000000	100	haystac	2856176	3:26.15	206.15
1000000	10	haystac	1545456	0:05.36	5.36
1000000	500	haystac	140624	10:42.10	642.1
10000	500	haystac	13358864	0:02.17	2.17
100000	500	haystac	150885120	24:36.69	1476.69
10000	500	haystac	130064	8:40.85	520.85
100000	500	haystac	11198576	6:21.41	381.41
100000	500	haystac	13771168	0:02.59	2.59
1000000	500	haystac	11237040	12:25.48	745.48
10000	500	haystac	11084400	10:19.52	619.52
1000000	100	haystac	4432608	0:05.74	5.74
100000	500	haystac	123648	8:47.42	527.42
1000000	100	haystac	54872288	22:55.46	1375.46
10000	500	haystac	148029088	25:55.53	1555.53
1000000	100	haystac	134848	2:29.73	149.73
1000000	10	haystac	31348944	20:30.55	1230.55
1000000	500	haystac	179254192	28:05.71	1685.71
1000000	500	haystac	14770688	0:07.02	7.02
1000000	10	haystac	942624	0:59.40	59.4
1000000	10	haystac	92384	0:27.45	27.45
1000000	100	kraken	2856608	2:11.08	131.08
1000000	10	kraken	1521104	0:05.42	5.42
1000000	500	kraken	125072	10:42.37	642.37
10000	500	kraken	13360288	0:02.21	2.21
100000	500	kraken	151234416	24:38.51	1478.51
10000	500	kraken	131344	8:38.57	518.57
100000	500	kraken	11206944	6:43.13	403.13
100000	500	kraken	13771648	0:02.73	2.73
1000000	500	kraken	11228800	12:40.82	760.82
10000	500	kraken	11073760	5:53.78	353.78

1000000	100	kraken	4388016	0:05.84	5.84
100000	500	kraken	125808	8:47.19	527.19
1000000	100	kraken	55674032	23:08.12	1388.12
10000	500	kraken	148743760	25:27.21	1527.21
1000000	100	kraken	138352	2:22.51	142.51
1000000	10	kraken	31861952	21:01.03	1261.03
1000000	500	kraken	178130624	29:57.20	1797.2
1000000	500	kraken	14717392	0:06.98	6.98
1000000	10	kraken	939936	0:58.62	58.62
1000000	10	kraken	95872	0:27.39	27.39
1000000	100	kraken	2856672	2:10.56	130.56
1000000	10	kraken	1544336	0:05.41	5.41
1000000	500	kraken	132528	10:40.58	640.58
10000	500	kraken	13359152	0:02.16	2.16
100000	500	kraken	151347120	26:35.61	1595.61
10000	500	malt	130080	8:41.15	521.15
100000	500	malt	11190288	4:21.85	261.85
100000	500	malt	13757776	0:02.77	2.77
1000000	500	malt	11228880	10:29.61	629.61
10000	500	malt	11063232	4:11.54	251.54
1000000	100	malt	4413808	0:05.67	5.67
100000	500	malt	131792	8:47.25	527.25
1000000	100	malt	54746736	22:19.20	1339.2
10000	500	malt	146213280	23:40.16	1420.16
1000000	100	malt	133888	2:23.19	143.19
1000000	10	malt	31262576	21:58.04	1318.04
1000000	500	malt	175446688	28:04.90	1684.9
1000000	500	malt	14754496	0:06.89	6.89
1000000	10	malt	934704	0:58.52	58.52
1000000	10	malt	93568	0:27.20	27.2
1000000	100	malt	2865152	2:10.24	130.24

1000000	10	malt	1519952	0:05.29	5.29
1000000	500	malt	132512	10:41.85	641.85
10000	500	malt	13352928	0:02.06	2.06
100000	500	malt	151759312	23:02.28	1382.28
10000	500	malt	131248	8:40.52	520.52
100000	500	malt	11198528	4:31.48	271.48
100000	500	malt	13753792	0:02.80	2.8
1000000	500	malt	11237008	10:22.25	622.25
10000	500	malt	11071440	4:11.26	251.26
1000000	100	sigma	4406256	0:05.74	5.74
100000	500	sigma	145968	8:47.55	527.55
1000000	100	sigma	54947872	23:10.39	1390.39
10000	500	sigma	146476544	24:23.03	1463.03
1000000	100	sigma	126688	2:24.98	144.98
1000000	10	sigma	31051856	19:33.95	1173.95
1000000	500	sigma	176888736	27:34.25	1654.25
1000000	500	sigma	14775680	0:07.20	7.2
1000000	10	sigma	946864	0:58.66	58.66
1000000	10	sigma	89792	0:28.40	28.4
1000000	100	sigma	2856320	2:08.32	128.32
1000000	10	sigma	1535792	0:05.33	5.33
1000000	500	sigma	125984	10:43.43	643.43
10000	500	sigma	13360816	0:02.43	2.43
100000	500	sigma	151830800	23:50.90	1430.9
10000	500	sigma	130064	8:40.56	520.56
100000	500	sigma	11197504	4:18.63	258.63
100000	500	sigma	13771520	0:02.91	2.91
1000000	500	sigma	11229088	10:26.59	626.59
10000	500	sigma	11051504	4:12.86	252.86
1000000	100	sigma	4427824	0:05.58	5.58
100000	500	sigma	132784	8:46.73	526.73

1000000	100	sigma	55682384	23:01.66	1381.66
10000	500	sigma	146914576	24:54.95	1494.95
1000000	100	sigma	129968	2:29.35	149.35

Permission from co-authors

I hereby give permission to Evangelos Antonios Dimopoulos to use our joint work as contribution towards his D. Phil thesis to be submitted for examination at Oxford University.

I confirm that to the best of my knowledge, the author contribution statement below is accurate and Evangelos Antonios Dimopoulos' contribution towards the work is greater or equal than that of any other co-author.

Chapter author contribution statement:

E.A.D., A.C., L.A.F., C.W., I.M.V., G.L., and E.K.I.P. designed the project; E.A.D. generated the simulated datasets and assembled the publicly available datasets; I.M.V. generated the calculus datasets; A.C. devised the mathematical framework of the paper; E.A.D. analysed the simulated and real datasets for the presented case studies; E.K.I.P., L.A.F., C.W. and I.M.V. provided support in interpreting results; E.A.D. and E.K.I.P. developed the computational software; L.A.F. and E.K.I.P. supervised the project; E.A.D. wrote the paper with contributions from all other authors.

Date: 28/05/2021

Name: Alberto Carmagnini

Signature:

A handwritten signature in black ink, reading "Alberto Carmagnini". The signature is written in a cursive style with a long horizontal flourish at the end.

Date: 28-May-2021

Name: Evan K. Irving-Pease

Signature: 

Date: 28 May 2021

Name: Christina Warinner

Signature: 

Date: May 1, 2021

Name: Greger Larson

Signature: 

Date: 31 May 2021

Name: Irina Marie Velsko

Signature: 

Date: 02/06/2021

Name: Laurent Frantz

Signature: 

Tracking the evolutionary trajectory of Marek's Disease Virus from the 16th century till the present

Statement of Authorship:

Research design: I designed the research, with input from Greger Larson, Laurent Frantz and Adrian Smith.

Data generation: Ophelie Lebrasseur generated the screening sequencing datasets and I generated the captured datasets.

Analysis: I performed the computational analyses presented in this paper and interpreted the results under the guidance of Laurent Frantz, Adrian Smith and Greger Larson. Steven Fiddaman aided in the design of the capture probes, selected the modern MDV genome accessions and performed bioinformatic analyses, whose results support the results presented in this paper.

Manuscript: I wrote the manuscript, with input from all other co-authors.

Acknowledgments: I would like to acknowledge Louis du Plessis for his contribution in designing the analyses and interpreting the results presented in this chapter, Bram Vrancken for his contribution in interpreting the results presented in this chapter, and all archaeologists who provided the archaeological material for this study.

Authors:

Evangelos Antonios Dimopoulos¹, Steven Fiddaman², Ophelie Lebrasseur¹, Adrian Smith², Laurent Frantz^{3,4}, Greger Larson¹

¹ The Palaeogenomics and Bio-archaeology Research Network, Research Laboratory for Archaeology and History of Art, University of Oxford, Oxford, UK.

² Department of Zoology, University of Oxford, Oxford, UK.

³ School of Biological and Chemical Sciences, Queen Mary University of London, London, UK.

⁴ Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, Munich, Germany.

Abstract

Since its characterisation in 1907 Marek's Disease Virus (MDV) has become a major chicken pathogen, with its symptoms becoming more acute and lethal. Over the last 50 years, chicken production has increased five hundred percent, chicken growth rate has tripled, and increasingly pathogenic strains of MDV keep emerging every decade, causing outbreaks with high mortality rates.

In order to understand better how avian viral pathogens evolve in relation to the changes in poultry production and increase their pathogenicity, we decided to investigate the evolutionary trajectory of MDV from the earliest possible point in time till the present day. We sequenced DNA yielded by archaeological chicken samples from various assemblages, and screened the generated data for MDV DNA sequences. Comparing the reconstructed genomes of ancient MDV strains to that of modern ones, allowed us to identify accumulated changes which occurred during its evolution from a potentially non-virulent (ancient) form (prior to 1907) to the highly virulent form of today. Our findings can have a significant impact in protecting chickens against MDV, through the design of improved vaccines, therefore improving the health status of poultry flocks globally.

Introduction

Ancient DNA and viruses

The use of ancient DNA (aDNA) has allowed for significant advances in the field of virology, enabling not only discovery of past viral variation, but also the discovery of extinct viral species from permafrost (Legendre et al., 2014), establishing a new field of research called Paleovirology (Emerman and Malik, 2010) that has generated exciting discoveries until now.

For example in the case of Variola Virus, the causative agent of smallpox and the first disease to have been eradicated by vaccination, aDNA research revealed that its lineage diversified only in the 18-19th century, rather than earlier in time (e.g. ancient Egypt or Middle East) as it was hypothesised till recently (Duggan et al., 2016). We have even discovered a three-century old strain of Variola Virus from Siberia (Biagini et al., 2012).

Regarding more recent epidemics, aDNA enabled the discovery that the United States HIV-1 epidemic emerged from a pre-existing epidemic in the Caribbean, placing the introduction of the virus to the USA almost a decade earlier, and even challenging established beliefs about the identity of patient-0 in the USA (Worobey et al., 2016). Especially in the case of retroviruses, by using aDNA, we can get a snapshot of the genomic diversity of simian T-cell leukemia viruses (STLV) infecting primates two hundred years ago, the same group of viruses which through cross-species transmission gave rise to HIV (Calvignac et al., 2008).

Paleovirology has also provided the opportunity to characterise the gene repertoire of other deadly viruses such as the 1918 Spanish Influenza Pandemic Virus

(Taubenberger et al., 2005; Tsangaras and Greenwood, 2019). It even became possible to recreate the 1918 Influenza Virus through reverse genetics and conduct experiments on live animals and human tissues to discover the combination of genes causing its incredible past virulence (Tumpey et al., 2005). Such research highlights that “resurrecting” ancient viruses to study their pathogenicity with the 21st century molecular techniques is paramount for strategising efficient countermeasures for future epidemics to reduce their mortality rates.

Alphaherpesviruses and their characteristics

Viruses, in general, can be defined as intracellular parasites, depending on the host cell for the majority of the functions that support their life cycle (Rickinson and virus Fields, 1996). They constitute a paraphyletic group, whose major lineages have evolved independently from one another, thus making the category of viruses a biological rather than an evolutionary one (Iyer et al., 2001). However, previous studies have shown monophyletic origins for eukaryotic DNA viruses from four viral families: poxviruses, asfaviruses, iridoviruses, phycodnaviruses and mimiviruses (Iyer et al., 2006, 2001). This monophyletic assemblage of large eukaryotic DNA viruses is known as the Nucleo-Cytoplasmic Large DNA Virus (NCLDV) clade, as they either replicate exclusively in the cytoplasm of the host cell or start their life cycle in the host nucleus but complete it in the cytoplasm (Iyer et al., 2001). Interestingly, the divergence of the major NCLDV families occurred at an early stage of evolution, prior to the divergence of the major eukaryotic lineages, and along with other large double stranded DNA (dsDNA) viruses, such as the herpesviruses, they descended from a small group of ancestral dsDNA viruses, that had already reached moderate genome sizes (Iyer et al., 2006).

A particularly interesting group of dsDNA viruses are herpesviruses, a family of widespread animal pathogens that cause diseases of medical and economic impact. Mucocutaneous infections, infections of the central nervous system, and occasionally infections of visceral organs are a few of the diseases they cause (Feierbach et al., 2006). Alphaherpesviruses are a subfamily of Herpesviridae, sharing a common virion structure, several homologous genes, and most importantly the ability to establish a lifelong latent infection (Szpara et al., 2010). They include human pathogens such as the varicella-zoster virus and herpes simplex virus types 1 and 2 (HSV-1 and HSV-2), and animal pathogens like pseudorabies virus (PRV) and bovine herpesvirus 1 (Jane Flint et al., 2020; Szpara et al., 2010).

Most alphaherpesviruses establish a primary infection at a mucosal or epithelial surface and then leave that site by invading the peripheral nervous system, where they establish a latent infection (Szpara et al., 2010). This neuroinvasive ability makes alphaherpesviruses useful not only for the study of their pathogenicity and infection mechanisms, but also for study of the mammalian nervous system (Feierbach et al., 2006). By taking advantage of the broad range of hosts the pseudorabies virus (PRV) can infect, we have been able to uncover chains of functionally connected neurons in the nervous systems of a variety of mammals, but also discover how an alphaherpesvirus infection in neurons disrupts mitochondrial motility and morphology (Kramer and Enquist, 2012; Szpara et al., 2010).

The majority of the alphaherpesviruses have a narrow host range, but due to their monophyletic lineage, they can potentially infect new hosts (Woźniakowski and Samorek-Salamonowicz, 2015). Transmission occurs through the exchange of saliva, nasal discharge between animals, and airborne particles (He et al., 2019).

Preventing cross-species host jumps of alphaherpesviruses is crucial both for livestock health (bovines, pigs) industry (He et al., 2019; Thiry et al., 2006), but also for the health of wildlife species, such as deer, that come into contact with human and domesticated animal populations regularly (Kautto et al., 2012).

Marek's Disease Virus and the vaccines against it

Marek's Disease (MD) is also caused by an alphaherpesvirus (MDV) which commonly infects domestic chickens and rarely other avian species (Trimpert et al., 2017), first described by Jozsef Marek in 1907. MDV is predominantly airborne pathogen, inducing tumours by dissemination of transformed T-cells to skeletal muscles and internal organs, causing asymmetric paralysis to chickens (Osterrieder et al., 2006). MDV can be further classified into two serotypes, with the oncogenic and highly pathogenic strains of MDV belonging to serotype 1, and the nonpathogenic MDV strains belonging to serotype 2 (Lee et al., 2008). Its pathogenicity though has been evolving rapidly throughout the 20th century, an evolutionary process that has been linked to anthropogenic influences (Atkins et al., 2013). Up to the 1950s, MDV's main symptoms were neural tissue lesions with a low mortality rate, but today MDV's mortality rate has reached 100%, in cases, from causing lymphomas (Atkins et al., 2013; Osterrieder et al., 2006; Witter, 1998). To encapsulate the continuously evolving virulence, pathogenic strains can also be categorised in different pathotypes as mild (mMDV), virulent (vMDV), very virulent (vvMDV) and vv+ or hypervirulent MDV (vv+MDV) (Shamblin et al., 2004; Witter, 1997).

Because of its high mortality rate vaccination schemes against the virus were set up, so that every one of the ~23 billion chickens hatched each year by the global poultry

industry is vaccinated against MDV (Trimpert et al., 2017). The vaccines could comprise any of the MDV strains belonging to serotypes 1 or 2, and occasionally they are combined with HVT (Herpesvirus of Turkeys) (Atkins et al., 2013). The HVT live vaccine has been around since the 1970s (Islam et al., 2002), and bivalent live vaccines of HVT and MDV strains have been used since the 1980s (Witter and Lee, 1984). Since the 1990s bivalent vaccines were unable to control symptoms caused by more virulent MDV strains (Witter, 2001) so a new vaccine called Rispens was developed (Lee et al., 2008), from an attenuated form of the MDV-1 strain.

Vaccination acting as an evolutionary force for MDV

Though the existing vaccines are effective at alleviating symptoms, the virus periodically breaks through the protection offered by vaccination by becoming more virulent (Witter, 1998). As a result, there is an acute need for a deeper understanding of the virus' evolutionary trajectory and the factors that result in the emergence of new increasingly pathogenic strains (Witter, 1997). Agricultural advances introduced by humans have affected the evolution of MDV. Firstly, the huge increase of potential hosts (~72 billion chickens were slaughtered in 2019), provided an opportunity for pathogens to evolve faster, which in turn affected both their virulence and transmission rate (Atkins et al., 2013). In addition, new husbandry practices (Kennedy et al., 2018), introduced during the 20th century, such as concentrated animal feeding operations (CAFO), which involve high population densities (e.g. ~0.3m² per chicken) and shorter broiler cohort duration (Atkins et al., 2013), due to improved feeding and selection for faster growth, likely also affected the transmission rate and pathogenicity of the virus (Atkins et al., 2011), as a host's

shorter lifespan can promote the evolution of pathogen virulence (Atkins et al., 2013; Day, 2002).

From previous genomic analyses, it has been observed that MDV virulence appears to have evolved independently in Eurasia and North America, even though these two continents had different vaccination strategies (Trimpert et al., 2017). From the same study it has been estimated that the ancestor of virulent MDV strains should have appeared approximately a decade before the first records of increased virulence. This prediction is a motivation to look further into the past, by using ancient DNA, in an effort to identify ancient MDV strains and create a time series of genome data, allowing us to study its past evolutionary trajectory and possibly also predict its future direction.

MDV infection overview

MDV infects new host chickens through inhalation and subsequently infects the macrophage or dendritic cells in the respiratory tract either directly or after the virus has first replicated in epithelial cells (Osterrieder et al., 2006). After 24 hours the virus has reached the spleen, thymus and the bursa of Fabricius, and in these tissues it begins its first cytolytic replication in the B and later activated CD4⁺ cells that it infects in these organs (Baigent and Davison, 1999; Shek et al., 1983). Infected CD4⁺ cells are not only a viral target for transformation, but they also act as reservoirs of latent MDV and the means the virus employs to spread within an animal until it reaches the skin (Osterrieder et al., 2006). There MDV enters the feather follicle epithelial cells where it assembles its virions and is shed back into the environment, to infect new chicken hosts (Carrozza et al., 1973; Johnson et al., 1975).

Structurally-wise the MDV genome has two repeated regions, one long that has a terminal long (TRL) and an internal long (IRL) copy, and one short region that has a terminal short (TRS) and an internal short copy (IRS), and these repeated regions flank a unique long (UL) and a unique short (US) region respectively (Spatz et al., 2007). The repeated regions are inverted and they harbor virulence related, protein-coding genes (e.g. the Meq gene) (Spatz et al., 2007). A major characteristic of the lytic MDV infection is the downregulation of the Major Histocompatibility Complex (MHC) class I molecules that are being presented on the surface of infected cells (Hunt et al., 2001; Osterrieder et al., 2006). As a result of host-pathogen coevolution, MDV encoded genes sustain the downregulation throughout the virus's lytic cycle, thus avoiding the cytotoxic immune response carried out by CD8⁺ T cells (Hunt et al., 2001). By escaping the cytotoxicity induced by the MHC class I molecules, the virus is able to both establish a latent infection, and also reactivate itself from it, allowing for its dissemination inside a host's organism. Through this process the virus can reach various visceral organs, where it can establish new infections (Calnek and Witter, 1984). It is still not fully understood though if the disseminated virus during this second cytolytic phase is initiated by latently or lytically infected T cells (Osterrieder et al., 2006).

Approaching an MDV infection from the host's perspective, all chicken genotypes are susceptible to MDV but they possess different levels of resistance to it (Kaiser et al., 2003), with resistance being driven by several genetic loci (Calnek, 1985) and the chicken MHC having a strong influence on it (Briles et al., 1977; Longenecker et al., 1977). Host susceptibility and resistance to MDV seems to be inversely proportional to the expression levels of MHC class I molecules. Chicken haplotypes, such as B21, that have the lowest expression levels of MHC class I proteins possess the

highest level of MDV resistance, whereas haplotypes that express the highest levels of class I proteins, like B19, have the highest levels of susceptibility (Kaiser et al., 2003; Kaufman et al., 1995). Additional loci contributing to MDV resistance outside the MHC can be found on chromosome 1 of chickens, in a cluster of genes known as the MDV1 locus (Bumstead, 1998), a region syntenic with the mammalian lectin-like natural killer (NK) cell antigen complex (Lanier and Phillips, 1996; Yabe et al., 1993). Also, another important factor affecting the outcome of an MDV infection is the viral load in the early stages of an infection, with a strong association appearing between high viral load in peripheral blood leukocytes and early onset of clinical MD symptoms (Bumstead et al., 1997; Kaiser et al., 2003).

The poultry industry overall has had a troublesome relationship with trying to control viral infections, by implementing vaccination schemes that prevent the onset of symptoms but does not hinder the transmission of MDV. Even though vaccines against viral diseases protect chicken flocks to an extent, their offered immunity is being breached more frequently since the 1990s, with recurring epidemics of the avian flu (H5N1) (Stephenson and Zambon, 2002) or Newcastle Disease Virus (NDV) (Dimitrov et al., 2017), to name a few, that also have an impact on human health (Katz, 2003). Because of the constant movement of animal and poultry products (Smith and Dunipace, 2011), combined with predation and animal theft (Annapragada et al., 2019), poultry viruses end up finding new available populations to inflict disease and evolve towards a more pathogenic state that escapes vaccine immunity (Escorcia et al., 2008).

Motivation

Considering the high prevalence of MDV in modern flocks of domesticated chickens, I screened approximately 1100 modern and ancient chicken samples to discover the earliest instance of an MDV infected bird in the archaeological assemblages collected by my laboratory. Furthermore, in order to understand better how avian viruses evolve in relation to the changes in poultry production and increase their pathogenicity, I investigated the evolutionary trajectory of MDV from the earliest possible point in time until the present day. I was also interested in detecting signs of selection driven by the human management of chicken flocks across time, that could have left a trace on the underlying demography of the various MDV lineages. By using aDNA methodologies previously applied on bacterial pathogens such as *Yersinia pestis* (Rasmussen et al., 2015), we were able to obtain snapshots of its evolution from different periods in an effort to gain a better understanding of the molecular mechanisms behind it.

For that purpose, I sequenced DNA yielded by archaeological chicken samples from various assemblages, and screened the generated data for MDV sequences. Twenty samples (19 ancient and 1 modern) were subsequently chosen as candidates for targeted genome capture. Comparing the genomes of ancient strains to that of modern ones, allowed me to identify accumulated changes which occurred during its evolution from a potentially non or less virulent (ancient) form (prior to 1907) to the highly virulent form of today. These findings can have a significant impact in protecting chickens against MDV, through the design of improved vaccines, therefore improving the health status of poultry flocks globally.

Methods

DNA extraction

DNA extractions were performed in dedicated ancient DNA (aDNA) facilities at the PalaeoBARN (University of Oxford). DNA was extracted from chicken bone or feather samples (Supplementary table 3) in a dedicated ancient DNA laboratory using the appropriate sterile techniques and equipment. The sample material that was collected for the archaeological chicken samples was bone tissue, more specifically either femur or coracoid bones (Supplementary table 3). All the collected modern samples were feathers, as a means of making the sampling of live chickens as less invasive as possible (Supplementary table 3). Extraction was carried out using the Dabney extraction protocol (Dabney et al., 2013).

Library Preparation & Sequencing

Illumina libraries were built following either the Gansauge and Meyer (2014) or the Carøe et al. (2018) protocol, but with the addition of a six base-pair barcode added to the IS1_adapter.P5 and IS3_adapter.P5+P7 adapter pair. The libraries were then amplified on an Applied Biosystems StepOnePlus Real-Time PCR system to check that library building was successful, and to determine the minimum number of cycles to use during the indexing amplification PCR reaction. A six base-pair barcode was used during the indexing amplification reaction resulting in each library being double-barcoded with an “internal adapter” directing adjacent to the ancient DNA strand and which would be the first bases sequenced, and a traditional external barcode that would be sequenced during Illumina barcode sequencing.

Up to 200 libraries with unique barcode combinations were pooled at equimolar levels (as determined by an Agilent Technologies 2200 TapeStation) and an 80 - 150 bp run was carried out on an Illumina HiSeq 2500/4000/X sequencer.

HAYSTAC

We screened 1111 ancient and modern chicken (*Gallus gallus*) DNA libraries, with the HAYSTAC (Dimopoulos et al., 2020), with default settings, for the Alpha-gallid herpesvirus 2 (MDV), using a custom database of complete herpesvirus genomes found in the NCBI RefSeq database. From this metagenomic screening MDV was robustly identified in 18 samples that were subsequently sent for a targeted genome capture at Arbor BioScience (Michigan).

Capture

Custom DNA probes for a tiled baiting approach were designed and synthesized by Arbor BioScience, based on the Alpha-gallid herpesvirus 2 genome with accession number EF523390.1 on the Nucleotide database of NCBI. A new oligonucleotide bait was designed approximately every 28 nucleotides, yielding 8403 50-mer probe sequences for the whole MDV target genome. The probes allowed for the hybridisation of both modern and ancient strains of MDV, while ensuring that no host (chicken) DNA would be carried over.

The libraries of the samples that were selected for targeted genome enrichment (Supplementary table 3), following the metagenomic screening for MDV DNA reads, were amplified again following the same indexing and PCR amplification strategies as previously described, and were condensed in a final volume of 25 µl. MDV genome capture and paired-end sequencing on Illumina NovaSeq lanes were

performed by Arbor BioScience. One modern chicken sample (OL1099) and two archaeological chicken samples (OL1008, OL1214), which had no MDV assigned reads, were also included, to test if these libraries would also be enriched for MDV DNA molecules after targeted genome capture.

In order to exhaust the informational content of the libraries that had already been enriched for MDV DNA, new sequencing libraries were built from the same DNA extracts of each sample, following the Carøe et al. (2018) protocol, and were amplified by PCR as described previously. The new amplified libraries were condensed at a final volume of 25 µl, and genome capture and paired-end sequencing on Illumina MiSeq lanes were again performed by Arbor BioScience. The number of reads assigned to MDV after screening and after merging the sequencing data from two rounds of genome capture are shown in Supplementary table 2.

DNA sequencing data alignment

The captured data were aligned with the bwa mem algorithm against the EF523390.1 MDV reference genome. Because the MDV genome has terminal long (TRL) and short (TRS) repeat regions (13,903 and 12,184 bp, respectively) (Spatz et al., 2007), one repeat from each of these duplicated regions was masked in order to avoid low alignment scores, stemming from a DNA read aligning to multiple positions, and retain the maximum number of reads.

SNP calling

Aligned reads had 5 base pairs trimmed from both their 5 and 3 prime ends with bamUtil v 1.0.14 (Jun et al., 2015), and read group identifiers were added to the resulting bam files by using picard tools v 2.16.0 ("Picard toolkit," 2019).

The GATK package (Van der Auwera and O'Connor, n.d.) was used for variant calling and sample genotyping. HaplotypeCaller v 4.1.2.0 was employed for SNP (Single Nucleotide Polymorphism) calling with the following parameters: the EF523390.1 masked MDV genome was used as reference, a value of 30 for the minimum base quality (Q) was required, all sites were outputted (even monomorphic ones) in GVCF files, ploidy was set to 1 to reflect the haploid nature of the viral genome, and physical phasing was not allowed (gatk HaplotypeCaller --reference EF523390.1_mask.fasta --min-base-quality-score 30 --output-mode EMIT_ALL_SITES --sample-ploidy 1 -ERC GVCF --do-not-run-physical-phasing true).

The GVCF files, from each individual sample, were aggregated into one with the CombineGVCFs v 4.1.2.0 tool. The aggregated data were inputted into GenotypeGVCFs for joint genotyping of the different samples. The raw variants were further filtered using VariantFiltration, filtering in variants with a minimum base quality (Q) value of 30 and a minimum sequence depth of 4 reads (gatk VariantFiltration --filter-expression "QUAL >= 30.0 && DP >= 4"). VariantsToTable was subsequently employed to convert the resulting VCF files containing the filtered SNPs to tables.

The folded Site Frequency Spectrum (SFS) was calculated using the angsd package v 0.92 (Korneliussen et al., 2014). For the detection of polymorphic sites the GATK genotype likelihood model was employed directly on the bam files containing the

aligned reads for each sample, combined with minimum base and alignment quality values of 30 for filtering purposes. Two folded SFSs were calculated, one that included both transition and transversion nucleotide changes, and one that included only transversions so as to exclude any singleton polymorphisms that could have arisen as a result of post mortem DNA damage (angsd -doSaf 1 -anc EF523390.1_mask.fasta -minMapQ 30 -minQ 30 -gl 2 -fold 1).

Ancient genome reconstruction

For the reconstruction of ancient viral genomes from aligned reads in bam format, we used the htsbox pileup tool (Li, n.d.). The following parameters were used: a minimum read length of 15 base pairs, 5 base pairs were trimmed from each end of an aligned read, a value of 30 for the minimum base quality (Q) and for the minimum read alignment quality (q) was set, for polymorphic sites the majority frequency allele was called, and a minimum sequence depth of 4 reads was also required (htsbox pileup -l15 -T5 -q30 -Q30 -M -s4).

Publicly available (modern) data

Data download

Modern complete MDV genomes were downloaded from NCBI using a custom biopython v 1.78 (Cock et al., 2009) script in fasta formats (42 fasta records in total). A complete modern genome of HVT (Herpesvirus of Turkeys), with accession number NC_002641.1, was also downloaded with the intention of being used as an outgroup in the downstream evolutionary analyses. Data regarding the accessions used, their species taxonomy, genome sequencing method employed and the year when each sample was collected are shown in Supplementary table 4.

Sequence preparation for RAxML and BEAST

In order to study the phylogenetic and evolutionary dynamics of the MDV virus, we combined modern and ancient sequence data.

Removing BAC sequences

MDV reference genomes with accession numbers KT833851.1, KT833852.1, and FJ436097.1 were contaminated with sequences from BAC (Bacterial Artificial Chromosomes) cloning vectors. BAC sequences were removed from these genomes with a custom biopython script, and their non-contaminated respective genomes were outputted in fasta format.

Masking of the modern genomes

Because the consensus MDV sequences in the ancient viral dataset have one terminal repeat of the Meq gene masked, we also masked the same terminal repeat in our modern viral sequence dataset, in an effort to minimise the noise in the downstream phylogenetic analyses.

The modern MDV genomes, now clean from BAC contamination, were split into 500 base pair long k-mers, with a 5 base pair overlap between each k-mer, using the pyfasta split v 0.5.2 tool (Pedersen, n.d.). Each fragmented genome was subsequently aligned against the masked EF523390.1 MDV genome using the bwa mem algorithm v 0.7.5 (Li, 2013) with default settings. The resulting bam files for each modern MDV genome were sorted and passed to htsbox pileup with default settings (htsbox pileup -l15 -M), in order to build new consensus fasta sequences that had the same terminal repeat of the Meq gene masked, as the ancient MDV sequence dataset.

Multisequence alignment (MSA) with mafft for building phylogenetic trees

In order to include the HVT outgroup taxon in our phylogenetic analyses, a multi-sequence alignment needs to be performed. As all the modern MDV genomes were aligned and reconstructed using the same reference genome (the masked EF523390.1 MDV genome), they are already in alignment with each other so the first MSA using the mafft v 7.123 (Katoh and Standley, 2013) algorithm is performed among the modern MDV and HVT genomes. For the incorporation of the ancient MDV genomes a stepwise approach was followed. As the ancient consensus genomes have a high number of unknown bases (Ns), they hinder multi-sequence alignments. To reduce the noise stemming from a high number of N bases each ancient genome is added in a stepwise fashion to the existing fasta alignment of modern MDV and HVT genomes, using again the mafft algorithm (mafft --maxiterate 1000 --thread 5 --nwildcard).

Extracting and preparing the aligned Open Reading Frames (ORFs) and intergenic regions for BEAST

For a BEAST analysis all the MDV genomes will need to have their coding and non coding sequences extracted and partitioned. As all the modern and ancient MDV genomes were aligned and reconstructed using the same reference genome (the masked EF523390.1 MDV genome), they are already in alignment with each other, no further preparation is needed for their ORFs to be extracted.

For the BEAST analysis unique, non-overlapping ORF regions were required, so as to avoid any noise and/or additional complexity from genomic regions where selective pressure is intensified, because they are part of the reading frames of multiple proteins. In order to exclude overlapping coding regions, a bed file

containing the coordinates of each ORF, from the masked EF523390.1 MDV genome, was parsed using a custom biopython script. Overlapping regions were trimmed from every reading frame, while ensuring that the resulting unique open reading frames had complete codons (with 3 base pairs each for the correct subsequent partitioning). Genomic regions that were not part of any unique or overlapping reading frame were also extracted.

The unique open reading frames were then extracted from the multi-sequence alignment using the seqkit subseq tool of the seqkit package v 0.12.1 (Shen et al., 2016). ORFs that were found on the minus (-) strand of the MDV genome had their directionality corrected by having their sequences reverse complemented with the seqtk seq tool v 1.2.95 (Li, 2012).

ORFs and intergenic regions with more than 10% overlap

For each resulting ORF or intergenic region alignment the number of overlapping covered sites was calculated. In order to maximise the informational content of the sequence alignments provided to BEAST, we used a custom biopython script to calculate the percentage of sites that were covered with a base in every sequence, across all modern and ancient MDV consensus fasta sequences. Only ORF or intergenic region alignments with at least 10% of overlapping covered sites across all sequences were selected for a phylogenetic analysis with BEAST, with the unselected alignments being excluded from any downstream phylogenetic analysis with BEAST. All the selected ORF and intergenic region alignments were subsequently concatenated respectively using the seqkit concat tool v 0.12.1.

Phylogenetic trees

Maximum likelihood (ML) phylogenetic trees were built using the RAxML (Stamatakis, 2014) software package. Two trees were built: 1) a starting tree for BEAST and 2) a rooted ML MDV tree.

BEAST starting tree

The percentage of unknown bases (Ns) was calculated for each masked modern and ancient genome with the seqtk comp tool v 1.2.95, and only genomes comprising less than 70% of unknown bases were kept. As a result, 42 modern MDV genomes from previous studies, 1 HVT outgroup, and 5 captured MDV genomes from the present study were aligned with the mafft algorithm, following the approach that was previously described. The 5 captured MDV genomes from the present study included 1 modern genome generated from the feather of a Silkie chicken (OL1099, sampled in 2014) and 4 genomes derived from archaeological bone samples (OL1385, OL1389, OL1986, OL2272).

A rooted tree was subsequently built with RAxML v 8.2.9 (Stamatakis, 2014) with the following parameters: a seed value of 12345 for rapid bootstrapping and parsimony inference, a GTRGAMMA substitution model, 100 bootstrap runs, and using the HVT genome as an outgroup (`raxmlHPC-PTHREADS -f a -T 10 -x 12345 -k -# 100 -p 12345 -m GTRGAMMA`). The best tree topology, as outputted by RAxML, was then parsed with the R package APE v 5.4.1 (Paradis and Schliep, 2019), and had the HVT outgroup pruned. The resulting pruned tree was used as the starting tree for the downstream BEAST analysis.

Maximum Likelihood tree

Again the percentage of unknown bases (Ns) was calculated for each genome with the seqtk comp tool v 1.2.95, and only samples with at least 3564 sites covered (representing more than 2% of the total genome's length) were kept, in an effort to retain and analyse the maximum number of ancient MDV genomes. In total 42 modern MDV genomes from previous studies, 1 HVT outgroup, and 11 captured MDV genomes from the present study were filtered in and aligned with the mafft algorithm, following the approach that was previously described. Out of the 11 captured genomes that were included, 1 genome was from a modern sample (being sampled in 2014) whereas the remaining 10 were yielded from archaeological samples from the 14th till the 20th century (Supplementary table 3).

RAxML was again used to build a rooted tree with the following parameters: a seed value of 12345 for rapid bootstrapping and parsimony inference, a GTRGAMMA substitution model, 100 bootstrap runs, and using the HVT genome as an outgroup (raxmlHPC-PTHREADS -f a -T 10 -x 12345 -k -# 100 -p 12345 -m GTRGAMMA). The best supported tree, along with the bootstrap support values for each branch, was outputted in nexus format and visualised with the program figtree v 1.4.4 (Rambaut, n.d.).

Maximum Likelihood tree based on transversions

In the maximum likelihood tree based only on transversions, we included the same 5 samples that were included in the previous ML phylogenetic tree that was built as an input for BEAST. By using the masked EF523390.1 genome as a reference, if a transition polymorphism was detected at a site of the multisequence alignment, this site was subsequently set as an unknown base (N) in all the genomes in the

alignment with a custom biopython script, so that in the end only the transversions were included. The multi-fasta alignment was subsequently filtered by using the snp-sites tool v 2.5.1 (Page et al., 2016), to exclude sites that included a gap or an undetermined base (N) across every sample or were monomorphic. The downstream strategy and parameters for the phylogenetic tree building were the same as the ones followed in the previous ML tree phylogenetic analyses, with the additional employment of a SNP ascertainment bias likelihood correction model (Lewis, 2001) (raxmlHPC-PTHREADS -f a -T 10 -x 12345 -k -# 100 -p 12345 -m ASC_GTRGAMMA --asc-corr lewis).

BEAST

The BEAST v 1.8.4 (Suchard et al., 2018) package was used for the molecular clock rate estimation and reconstruction of the MDV phylogeny. A total of 45 MDV genomes were included in this analysis: 42 modern, 1 modern captured from the present study (OL1099), and 4 ancient captured genomes from the present study (OL1385, OL1389, OL1986, OL2272). Tip dates were provided to BEAST for both the modern and archaeological samples.

The concatenated ORF alignment, that was previously prepared, was partitioned further into its codon positions with a custom python script. Each partitioned alignment along the concatenated intergenic region alignment, was inputted in BEAST. The ML tree produced from the same 45 genomes was also provided as input, being used as the starting tree for the topology search of the MCMC algorithm.

A Bayesian Skyline tree model was employed with a GTR 4 GAMMA base substitution model, and base frequencies were estimated from the provided sequence dataset. An uncorrelated relaxed clock and a strict clock model were

separately combined with the above set of parameters. Each clock model was run for 100 million MCMC steps 3 times, with a different seed provided in each run . The produced log and tree files were combined for each clock model (3 runs per clock model) using LogCombiner, and a burn-in of 10 million steps was assumed for each MCMC chain. The combined log and tree files were then inspected with Tracer to ensure that the probability space for each parameter had been sufficiently explored and the MCMC algorithm had mixed successfully. The appropriate clock model was chosen based on the comparison of marginal likelihoods of the combined runs and was further used for the molecular clock rate and the tree root age estimation.

Population structure summary statistics

The R package PopGenome v 2.7.5 (Pfeifer et al., 2014) was used to calculate Nei's D_{xy} , i.e. the population divergence defined as the probability of non-identity by descent of two alleles drawn in the two different populations averaged over all loci (Ma et al., 2018; Nei and Miller, 1990), across the different MDV lineages. The same package was also used to calculate the nucleotide diversity within each MDV lineage as well. For this analysis I only used the 47 ancient and modern genomes that were used in the BEAST analysis, as they had the least amount of missing data. The multi-fasta alignment of the dataset used for the BEAST analysis was filtered by using the snp-sites tool v 2.5.1 (Page et al., 2016), to exclude sites that included a gap or an undetermined base (N) across every sample or were monomorphic. The resulting alignment of polymorphic sites was used as an input for PopGenome. Three populations/lineages were considered: Ancient, American and Eurasian. All archaeological samples were assigned to the Ancient population, whereas the rest of

modern samples were assigned to either the American or Eurasian population, depending on the country they were sampled from (Supplementary table 4).

Results

Metagenomic identifications and MDV genome capture

After screening 1111 ancient and modern chicken samples with HAYSTAC by using a herpesvirus specific database, DNA reads assigned to *Gallid alphaherpesvirus 2* were identified in 19 chicken samples (Supplementary table 3), ranging from 1 to 1183 uniquely assigned reads (Supplementary table 1), resulting in a “prevalence” discovery rate of 1.7%. To ensure that the identification results were valid we only included samples whose evenness of coverage ratio of genome coverage to fraction of genome covered was less than 10, and also showed appropriate chemical damage patterns.

Out of all the 116 modern feather samples that were screened, MDV was identified only in sample OL1099. As vaccination against MDV is common practice in modern chicken flock management, one would have expected that it was the vaccine strain that was identified during the metagenomic screening. Interestingly, the bird the feather sample was collected from was exhibiting symptoms of an MDV infection (e.g. cerebral hernia). This identification enables us to be more confident in the rest of the positively identified samples for MDV, as MDV was genetically identified only in this specific modern sample that was exhibiting symptoms.

The 17 archaeological samples and the modern chicken feather sample in which MDV DNA molecules were identified (OL1099), along with the two archaeological

chicken samples in which no MDV DNA was detected (OL1008, OL1214), subsequently underwent an in-solution genome capture. The library enrichment for MDV DNA molecules ranged from 0 to 273.12 fold increase in unique DNA templates, allowed us to increase the genome of coverage of the ancient MDV isolates (ranging from 0.01 to 41.92X), reduce the percentage of missing sequence data and reconstruct consensus genomes (Supplementary table 2). The genome enrichment captured all the DNA templates that were available in the analysed libraries, as shown by the percentage of duplicated reads post capture that ranged from 56.6 to 99.98% (Figure 1, Supplementary table 2).

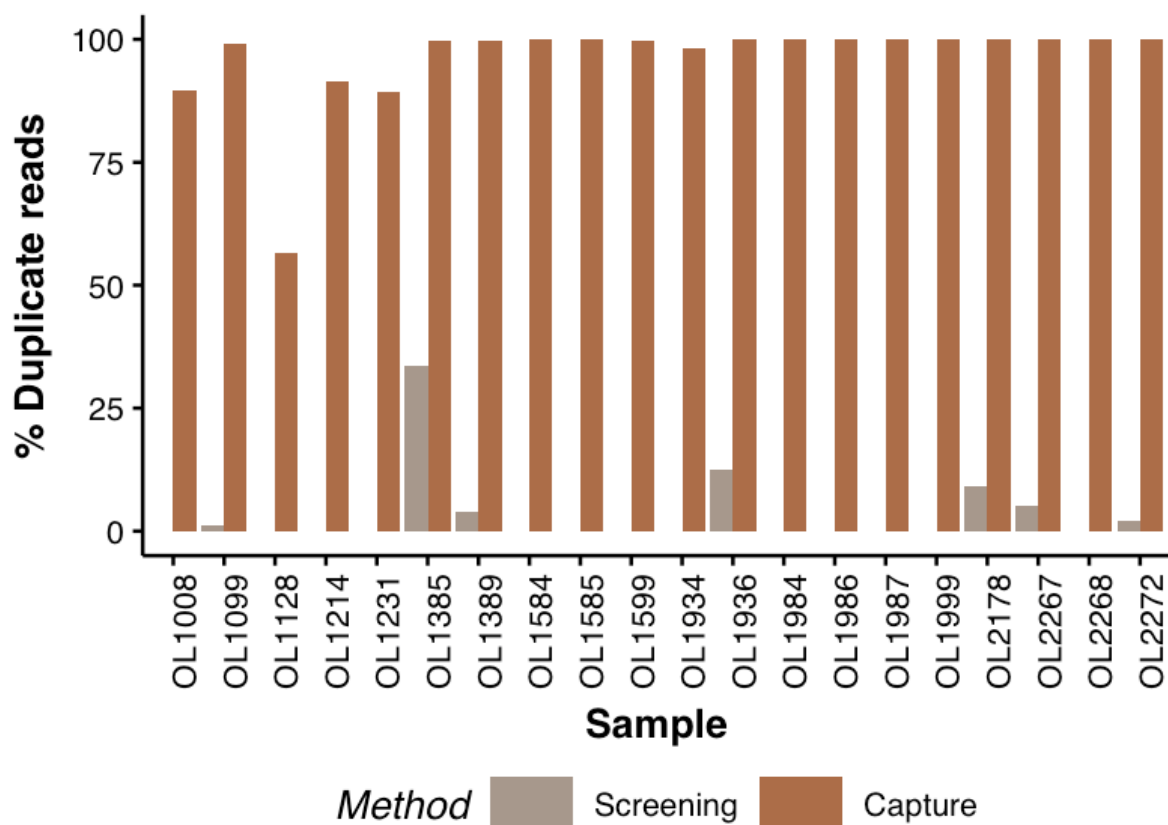


Figure 1. Percentage of duplicated reads, after sequencing screening and MDV enriched libraries.

SNP analysis

Joint SNP calling was performed with the GATK package on the new samples sequenced in this study, and after variant filtration 1993 SNPs and indel polymorphisms were identified in 1950 genomic positions, with a minimum sequencing depth of 4 reads. In order to understand better the distribution of SNPs across the 20 newly sequenced samples for this study, we calculated the folded SFS (Site Frequency Spectrum) for them (Figure 2 and Supplementary table 5). Most sites were found to be monomorphic across the 20 analysed samples. For the majority of the polymorphic sites, as the number of SNPs increased the number of sites observed decreased. In order to discover if the large number of singleton SNPs was a consequence of post mortem DNA deamination damage an additional SFS was calculated in which transition polymorphisms were excluded (Supplementary table 6, Supplementary figure 3). When the transitions are excluded no singleton SNPs are present indicating that these SNPs are the result of DNA damage present in the archaeological chicken samples, and the overall number of SNPs decreases. Furthermore, even with no transitions the distribution of number of SNPs per number of sites for more than 15 SNPs remains highly similar.

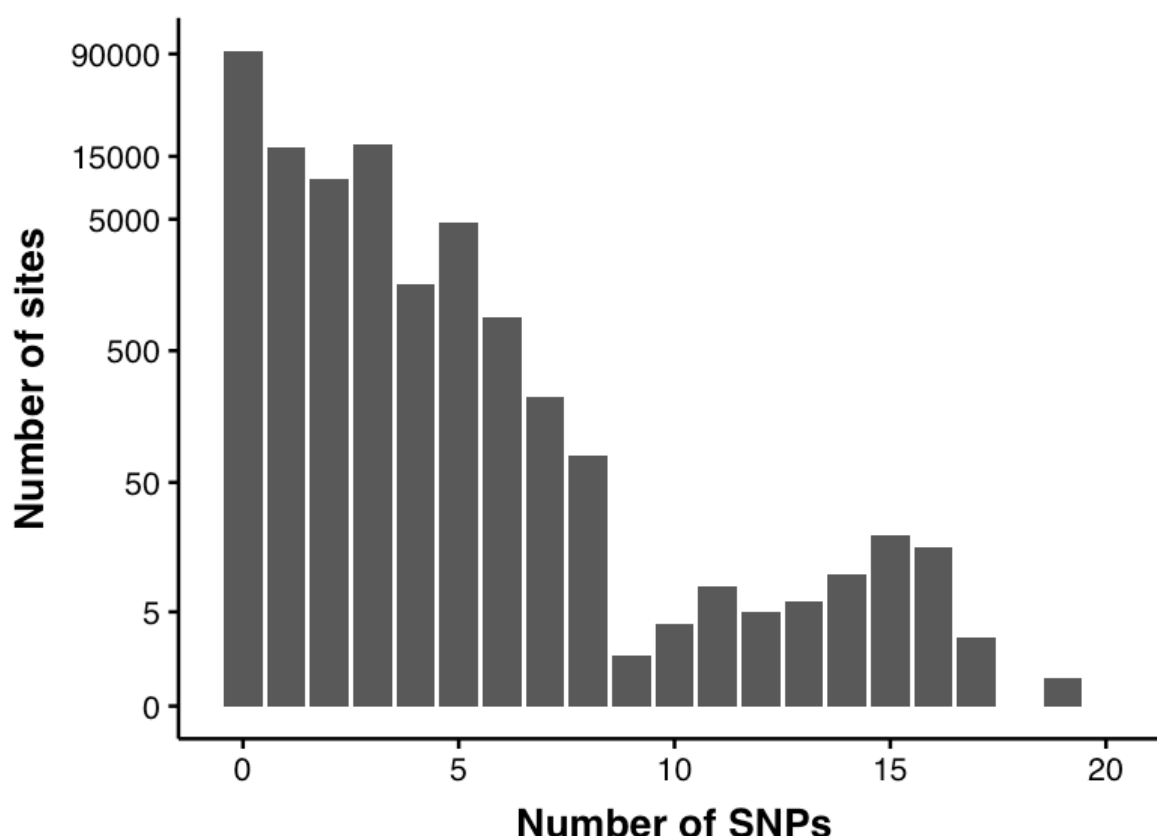


Figure 2. Histogram of the calculated SFS, across the 19 archaeological and 1 modern newly captured samples (with a log transformed y-axis).

Out of the 1950 polymorphic sites identified across the 20 MDV samples, there were 34 sites that had more than 1 alternative allele called. Excluding short indel variants, there were only 11 sites with more than one alternative allele called. As the MDV genome has two repeated regions that harbour virulence related genes, we were motivated to discover how these multiallelic SNP sites were distributed between the unique and repeated genome regions of the MDV chromosome. There was 1 multiallelic site found in unique regions of the MDV genome, whereas there were 10 sites found in duplicated regions of the MDV genome. To analytically test this observed difference, we performed a Fisher exact test (p-value of 1, after rescaling

to account for the 2 copies of the repeated regions), that proved that the frequency of SNPs with multiple alternative alleles is the same throughout MDV's genome, even in the long and short repeated regions of the MDV chromosome, regions where a large proportion of the known pathogenicity related genes are found.

Phylogenetic analysis with Maximum Likelihood trees

The best maximum likelihood tree, with its branch bootstrap support values, is shown in Figure 3. All the ancient MDV genomes always cluster together in a single clade (yellow) that falls basal to the modern Eurasian and American MDV clades (orange and blue), highlighting their genetic divergence from modern pathogenic strains of the virus (also indicated by the relative branch length of that clade). Bootstrap support values vary throughout the different branches and in most cases they exceed 50%, indicating that the majority of the inferred topology is well supported.

Because the majority of the SNPs identified in the ancient MDV genomes were shared by less than 10 individuals, it was important to assess the degree to which polymorphisms, arising from deamination events, could have affected the topology of the phylogenetic tree. In order to assess how robust the previous topology was, we repeated the phylogenetic analysis by removing all the transition mutation polymorphisms, monomorphic sites and sites populated only by gaps and undetermined bases from every genome included in this analysis. Also in order to reduce the noise introduced by the large number of undetermined bases (Ns) present in the sequences of our archaeological samples we only included the same

5 newly sequenced samples (4 ancient and 1 modern) that were included in the BEAST analysis dataset. As in the phylogeny constructed based on all the polymorphic sites (Figure 3), the modern sample OL1099 clustered with the American clade (orange) as it is a strain used in modern vaccines. The 4 archaeological samples clustered together in a clade sister to the American clade (blue) (Supplementary figure 2). Both the American (orange) and the ancient MDV clades were derived from the Eurasian (blue) MDV clade in this phylogeny. Interestingly, the phylogenetic tree built based only on transversion polymorphisms indicates that the majority of the ancient MDV isolates are sufficiently divergent from modern MDV strains (sampled from the second last of the 20th century since the present), thus creating a clade of their own.

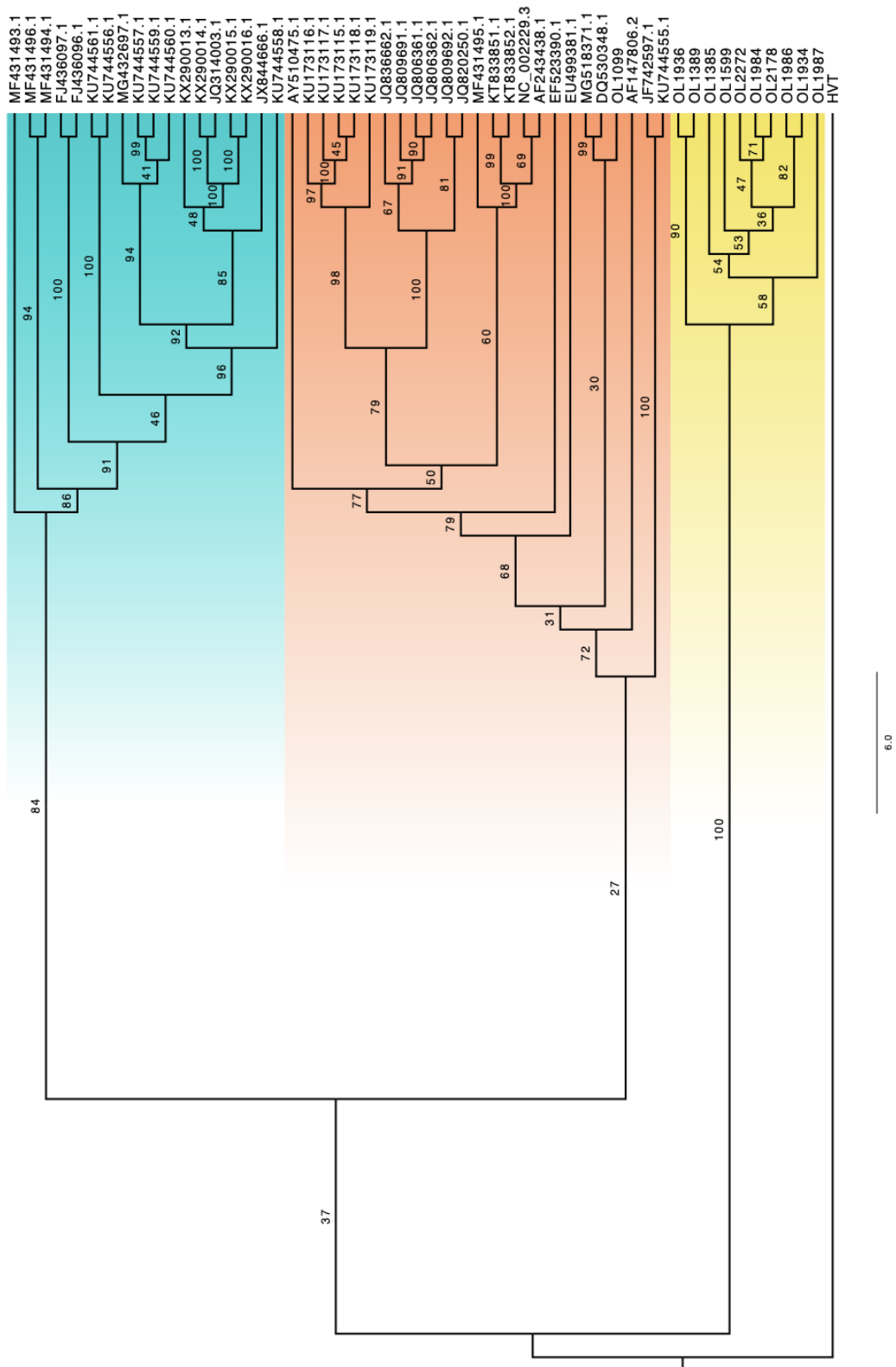


Figure 3. Rooted ML bootstrap tree of Marek's Disease Virus rooted with Herpesvirus of Turkeys. The ancient MDV clade (yellow) is basal to the modern MDV strain clades (blue and orange).

BEAST analysis

For a more detailed phylogenetic analysis, in order to estimate both a robust tree topology and the evolutionary rate of MDV's genome across time, we used the BEAST package. We were only able to include 4 archaeological samples, which had the least amount of unknown nucleotide bases: OL1385, OL1389, OL1986, and OL2272. We also included the modern sample OL1099, that was also sequenced for this study.

A strict and a relaxed clock model were run on our dataset, in an effort to find the best fitting one. Tracer software was used to ensure that the ESS (Effective Sample Size) metric indicated sufficient sampling of the probability space for each parameter of the clock models, and that the MCMC chain had converged successfully after combining the separate runs for each model. Both the strict and the relaxed clock models fit our dataset equally well, with their respective likelihoods being estimated at -82657.2 and -82594.8 respectively. Therefore, we selected the relaxed clock model as the best fitting one, as the marginal likelihood of the relaxed clock model was higher than the one that the strict molecular clock model produced.

The tMRCA of all MDV strains included in this study, as estimated under the relaxed clock model, was ~1477 AD , with a 95% HPD interval ranging from 1388 to 1550. This estimated tMRCA from the uncorrelated relaxed clock model fell approximately 100 years before the date of our oldest ancient sample (mid 15th century). The mean relaxed uncorrelated clock rate was estimated at 9.1E-6 substitutions per site per year, with a 95% HPD interval ranging from 4.7E-6 to 1.4E-5 substitutions per site per year. The coefficient of variation in the current genome dataset was calculated at

0.741 (95% HPD interval 0.49 - 1.01), proving that different branches of the MDV tree are evolving at different speeds. The maximum clade credibility MDV was calculated by TreeAnnotator (Figure 4) and was rooted at its midpoint. In agreement with the Maximum Likelihood tree produced by RAxML, the 4 ancient MDV genomes clustered all together (yellow), and again fell basal to the modern Eurasian and American modern MDV strains (orange and blue).

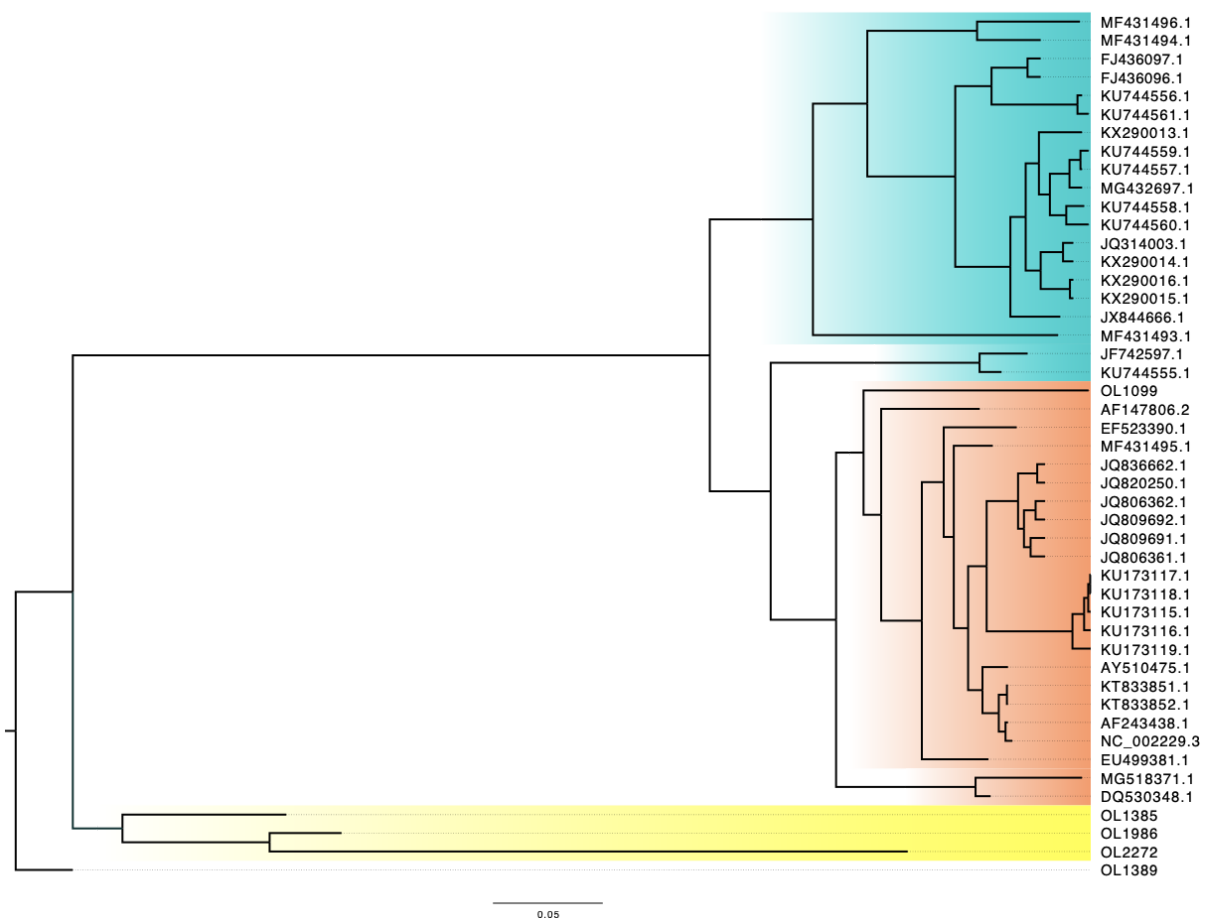


Figure 4. MDV phylogenetic tree under a relaxed clock model. Tip dated maximum clade credibility MDV phylogenetic tree, with a midpoint root, as estimated by BEAST under a relaxed clock model. As seen in the ML tree, all ancient MDV genomes (yellow) cluster together basal to the modern Eurasian and American MDV lineages (blue and orange).

Population structure analysis

The calculated Dxy represents an absolute measure of genetic divergence between the 3 different MDV lineages. The Ancient MDV lineage is estimated to be almost equidistant from both the modern American and Eurasian populations. There is less differentiation observed between the modern American and Eurasian MDV lineages, than when either of them is compared to the Ancient MDV lineage (Table 1). This divergence pattern hints at the presence of an underlying population structure across the ancient MDV lineage and the American and Eurasian MDV lineages that have been sampled during the second half of the 20th century.

Table 1. Dxy among the 3 MDV lineages. Greater genetic divergence can be observed among the Ancient and modern American and Eurasian lineages, whereas the least diverged lineages are the modern Eurasian and American ones.

	Ancient	American	Eurasian
Ancient		136.80	133.99
American			65.95
Eurasian			

Nucleotide diversity within each lineage was also calculated. The ancient MDV lineage had the highest nucleotide diversity, a finding that could partially be attributed to the excess of singleton transition SNPs, likely due to post mortem DNA damage. The American and Eurasian lineages from the second half of the 20th century show similar levels of nucleotide diversity (Table 2).

Table 2. Nucleotide diversity within each MDV lineage. Higher nucleotide diversity levels can be observed in the Ancient MDV lineage, partially due to DNA damage, whereas the American and Eurasian lineages show similar levels of nucleotide diversity.

Lineage	Nucleotide diversity
Ancient	63.50
American	37.33
Eurasian	39.91

Discussion

MDV biology and its implications in infection time course studies

MDV is transmitted via inhalation of contaminated dust particles and once a bird gets infected by MDV it starts shedding the virus approximately after a week. The exact timing of when the shedding begins will depend on its vaccination status (Atkins et al., 2011), and after a bird becomes infectious it continues shedding until its death (Baigent and Davison, 2004; Islam and Walkden-Brown, 2007). As the bird never recovers from an MDV infection (Atkins et al., 2013), there are no time course studies of the disease that span from infection to convalescence. There are though time course studies of MDV infections, which have mainly focused on the time course of viral DNA and antigens being detected in the brain and Central Nervous System (CNS) and its correlation with the time of appearance of brain lesions in the infected birds (Gimeno et al., 2001; Rémy et al., 2013).

After a bird comes into contact with MDV, mainly through inhalation, the virus subsequently travels to the spleen, thymus and bursa of Fabricius, and in these tissues it meets one of its primary targets, the CD4 activated T cells (Hunt et al., 2001; Osterrieder et al., 2006). Infected T cells are not only a primary transformation target of MDV, but they also are the cells in which latent MDV genomes persist and use for transport to the skin, where they enter the feather follicle epithelium (Osterrieder et al., 2006). Due to the virus' ability to cause a persistent latent infection, chickens never recover from an MDV infection and shed MDV virions throughout their lifespan (Atkins et al., 2013). Therefore, detecting MDV in ancient

and historic chickens (prior to the introduction of the vaccination schemes) does indicate that they had been infected with the virus at some point in their life. It does not necessarily prove though that the virus was the cause of their death, as an infected bird could have died of other causes, pathogenic or not, before Marek's Disease (MD) killed it. On the other hand, detecting MDV DNA in modern samples does not always equate to identifying an MDV infection, as most modern chickens are vaccinated against MDV (Lee et al., 2008; Witter, 2001; Witter and Lee, 1984). For a robust identification of an MDV infection in modern chickens, DNA evidence should also be accompanied by a diagnosis based on physiological and neurological symptoms of MD (Witter, 1998) that a bird might be exhibiting.

Metagenomic identifications and MDV capture

Traces of the virus's DNA were only detected in 18 archaeological chicken samples and 1 modern Silkie breed chicken feather sample, leading to an overall discovery rate of 1.7%. Even though no other aDNA study has calculated a pathogen discovery rate from screening a large archaeological sample collection, quite a few species of double stranded DNA viruses have been identified in the human archaeological record. The genomes of two ancient strains of Hepatitis B Virus (HBV) have been sequenced from two mummies dating back to the 16th century, one discovered in Korea (Kahila Bar-Gal et al., 2012) and one discovered in Italy (Patterson Ross et al., 2018) respectively. Ancient Variola Virus genomes have also been reconstructed from 17th century mummies exhumed from Siberia (Biagini et al., 2012) and Lithuania (Duggan et al., 2016), shedding light into the evolution of the virus during

the past 300 years. The archaeological material, these viral genomes are derived from, comes from the tissues of well preserved mummies. It should be noted that it is not common for pathogen DNA to be preserved in archaeological bone samples (Stone and Ozga, 2019), especially when the pathogen is a virus such as MDV. Also archaeological bone samples are posing an additional limitation to the discovery of ancient viruses, as viruses can rarely infect or be isolated from bone tissue (Tsangaras and Greenwood, 2019).

Despite these limitations, the fact that MDV infects the mobile immune system (Osterrieder et al., 2006), along with the fact that an MDV infection lasts for the host's lifetime (unless it kills it) (Atkins et al., 2011), makes it possible to retrieve ancient MDV DNA from bone samples. Concurrently another contributing factor to retrieving ancient MDV is its ability to replicate in high numbers inside host cells (1 million copies every 10 thousand chicken host cells (Baigent et al., 2005; Read et al., 2015)), providing abundant genetic material to be preserved in the archaeological chicken bones. Additionally because MDV is a double stranded DNA virus, its DNA degrades more slowly compared to the genetic material of a single stranded DNA or RNA virus. These underlying biological mechanisms of MDV and traits of its infection increase the chances of retrieving ancient MDV DNA molecules from shotgun-sequencing data of archaeological chicken samples. Subsequently, by employing a targeted genome enrichment approach (Carpenter et al., 2013), the reconstruction of ancient MDV genomes can become possible.

In terms of (archaeological) sample material the main skeletal elements from which DNA was extracted were: modern feathers, archaeological coracoid bones, and

archaeological femur bones. MDV DNA was only identified in libraries that were built from feather, coracoid or femur samples, indicating that pathogen DNA preservation is better in these skeletal elements than in other elements retrieved from archaeological chickens. The majority (16 out of 20) of bone elements in which MDV was identified were coracoid bone samples, suggesting that coracoids could be the optimal skeletal element to be sampled for future avian pathogen (e.g. MDV) or metagenomic studies. It should be highlighted that soft tissues, in which MDV is abundant and replicates (e.g feather follicle epithelium, respiratory and nervous system) (Atkins et al., 2011), do not get preserved in the archaeological record (Tsangaras and Greenwood, 2019). Thus, in the absence of archaeological samples of soft chicken tissue, it is less likely to retrieve MDV DNA frequently enough throughout the archaeological record in order to confidently estimate MDV's prevalence through time and space.

The targeted genome capture approach improved the sequencing efficiency of ancient MDV genomes, in all the samples that HAYSTAC detected DNA reads aligning to MDV. Concurrently no DNA reads were captured from the negative control shotgun chicken library, strongly indicating that the capture probes did not hybridise with non-target DNA. The efficiency of the genome capture varied across the different archaeological samples from a 0 fold increase (for the samples with no MDV reads assigned to them during screening) to a 273.12 fold increase, with the respective final genome coverage ranging from 0 to 41.92x. The genome enrichment successfully exhausted all the DNA templates that were available in the sequencing libraries as the duplication percentage of the captured libraries ranged from 56.6 to 99.98% (Supplementary table 2). It should be highlighted that there were samples in

which thousands of reads were captured after the targeted genome enrichment, but unfortunately their respective consensus genomes vastly consisted of undetermined bases (Ns) (> 99% of the total consensus genome, Supplementary table 2). That can be a consequence of the captured reads clustering tightly around a specific genomic region, and/or of the captured reads aligning poorly to the chosen reference genome, and/or the capture probes having exhausted all the unique MDV DNA templates, thus leading to a high number of sequenced duplicated reads.

SNP analysis

A higher number (p-value of 1) of triallelic SNPs was detected in the repeated regions of the MDV genome, which harbour virulence related genes (such as the Meq gene), than the unique long or short genome regions. As MDV is a virus with a haploid genome, the existence of SNP sites with more than one alternative allele could be justified by the host being infected by multiple viral lineages concurrently (Li et al., 2019; Yu et al., 2020). However, the fact that these polymorphic sites are restricted to the repeated regions of the MDV genome could also indicate within host diversification rather than an infection by multiple strains. This existence of polymorphic sites with multiple alternative alleles in the repeated regions could imply that they have a higher mutation rate, agreeing with findings from previous studies (Spatz et al., 2007; Trimpert et al., 2017). Another biological mechanism that could be contributing towards an abundance of SNPs in these genomic regions could be recombination. It has been previously shown that recombination poses a great risk for chickens, as attenuated live vaccine MDV strains have the potential to enhance

virulence through recombination, by introducing non-synonymous mutations to the MDV genome (He et al., 2020). The same strategy is also followed by *Gallid alphaherpesvirus 1*, which uses recombination to diversify its genome and increase its virulence (Loncoman et al., 2017a). This is not surprising as the viral group of alphaherpesviruses has been using recombination to drive its evolution and pathogenicity (Loncoman et al., 2017b). Many biological features of alphaherpesviruses favour host-cell coinfection and therefore increase the chances of recombination events taking place. Furthermore, there have been indications of genetic drift acting upon the genome of MDV, but its effect has neither been studied thoroughly nor quantified (Pandey et al., 2016).

After calculating the SFS across the 20 samples that were captured for this study, once with transition polymorphisms included and once excluded, it is evident that the majority of the SNPs are shared among a limited number of individuals. The fact that all of the singleton SNPs disappear when transitions are filtered out is indicating that these polymorphisms arise due to the presence of post mortem DNA damage, such as deamination. The same pattern could have also been observed due to the large percentage of missing sequence data, which many ancient samples suffer from. Large amounts of missing data could affect the shape of the SFS by creating a trend of variant frequencies that do not decrease as the number of SNP sites increases. This can lead to an excess of low and/or high frequency variants, as sites contributing to mid frequency variants could be included in the missing data. Concurrently, the shape of the SFS can also be affected by the demographic history of a population (Adams and Hudson, 2004; Lapierre et al., 2017; Marth et al., 2004). More specifically a pattern of an excess in low- and high-frequency variants can also

be observed when selection is acting upon a population (Fay and Wu, 2000; Lapierre et al., 2017). Distinguishing between the two cases can be really complicated, especially in a dataset of aDNA sequences of a virus evolving under selective pressure.

While aDNA damage and missing data could affect the SFS frequencies, both of the calculated spectra (with and without transition SNPs) suggest that ancient MDV strains had a greater level of genetic variation than the sequenced present day MDV strains. To avoid any biases arising from these confounding factors in our downstream analyses, we trimmed the damaged ends of all the DNA reads used for variant calling and phylogenetic analyses, and applied additional quality filters (minimum base and alignment quality values, minimum read length, minimum sequencing depth). Despite our best efforts to work around these caveats, caution is advised when drawing results based on SNPs that are not shared across all the samples.

Phylogenetic analysis

When combining the ancient MDV sequences with MDV genomes from the second half of the 20th century till now in a maximum likelihood tree, the samples are broadly distributed across three different lineages: an ancient, an American and a Eurasian lineage. To ensure that the clustering of the ancient samples was not an artefact of mutations stemming from deamination, we built an ML tree, based on the same set of sequences that were used to build the BEAST input tree, in which we

excluded any site that was monomorphic or where a transition polymorphism was detected. The ancient samples again clustered together into one clade, proving indeed that they are divergent enough from modern MDV strains. This analysis was performed only with the four ancient sequences that had the least amount of missing data. The high number of undetermined bases (Ns) (more than 70%) present in the majority of the ancient samples confounded the phylogenetic analysis when the transition polymorphisms were excluded from the aligned sequences. As the reduced number of polymorphic transversion sites did not provide enough information for a robust phylogeny to be built, the clades of the resulting MDV phylogeny were not well supported. Thus we decided to include only the captured samples with high informational content in the transversion only ML tree analysis.

For the clock rate phylogenetic analysis, we tried to maximise the informational content of the provided alignments by including ancient genomes that contained the least amount of missing data. The BEAST analysis was in agreement with previous studies that analysed complete genome data (Trimpert et al., 2017) and showed that a relaxed clock model can explain the observed accelerated evolution of MDV's virulence since the 1950s (Witter, 1998). Namely, increasingly virulent phenotypes of MDV (e.g. v and vv MDV pathotypes) that cause more severe symptoms (Osterrieder et al., 2006; Shamblin et al., 2004) have been emerging faster since the introduction of the MDV vaccines (Atkins et al., 2013; Shamblin et al., 2004; Witter, 1997). Clock rate analyses focusing only on the Meq gene sequence have previously favored a strict molecular clock model (Deng et al., 2020; Padhi and Parcells, 2016) though. This discrepancy between clock model inferences, depending on whether they were based on complete genome data or single gene sequences, stems from

the fact that pathogenicity is not an evolutionary trait driven by only a single gene (e.g. the Meq gene in the case of MDV). As the virulence of MDV has been linked to multiple genetic loci found in the long and short repeated regions of its genome (Spatz et al., 2007), we stress the importance of analysing multiple gene and intergenic region sequences, so as to obtain clock rate results reflective of the actual evolutionary trajectory of MDV. The various branches of the MDV tree were shown to be evolving at a non uniform rate, with a mean clock rate of $9.1\text{E-}6$, a rate that is lower than what was previously estimated for MDV ($\sim 1.6\text{E-}5$ (Trimpert et al., 2017)). The newly estimated mean clock rate for MDV is on par with clock rates estimated for other species of alpha-herpesviruses, such as the Varicella-Zoster Virus ($\sim 5.5\text{E-}6$) or the Herpes Simplex Virus ($\sim 8.8\text{E-}5$) (Firth et al., 2010). It is also comparable to the clock rates observed in other families of double stranded DNA (dsDNA) viruses, e.g. Variola Virus ($\sim 8.2\text{E-}6$) - a well known member of the *Poxviridae* family. Further supporting the choice of a relaxed clock model is the coefficient of variation (mean of 0.741), whose 95% HPD interval excludes the value of 0 (0.49, 1.01).

For the selection of the appropriate molecular clock model for the MDV phylogeny, our decision was based on the comparison of the marginal likelihoods estimated from the combined runs of the strict and relaxed clock models respectively. The marginal likelihood is commonly used for comparing evolutionary models in Bayesian phylogenetic approaches (Xie et al., 2011), such as BEAST (Suchard et al., 2018). The relaxed clock model, which was also selected, produced the highest marginal likelihood value out of the two tested models. To further corroborate our choice of the relaxed clock model we also draw information from the mean value and 95% HPD

interval of the coefficient of variation. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one dataset to another, while also providing a statistical measure of the dispersion of data points around the mean, in this case the mean value of the clock rate (Brown, 1998). We followed the marginal likelihood approach for the choice of the appropriate molecular clock model, rather than employing an approach such as a Likelihood Ratio Test (LRT), as the marginal likelihood measures the average fit of a model to the data, whereas an LRT approach for model selection bases its decision on the fit of each competing model at its best (i.e. the point in parameter space that maximizes the likelihood) (Xie et al., 2011). Additionally LRTs are biased against strict molecular clock models, when recombination is taking place (Posada, 2001; Schierup and Hein, 2000). It has been found that even small levels of recombination can reject a LRT for a strict molecular clock, falsely providing evidence for clock rate variation caused by selection pressures acting upon the different branches of phylogeny (Schierup and Hein, 2000).

For the BEAST phylogenetic analysis, we employed a tip dating model to calibrate the age of the tree nodes. This leaf-dating method (Shapiro et al., 2011) estimates the age of individual sequences within the Bayesian Markov Chain Monte Carlo (MCMC) framework provided by the software package BEAST (Suchard et al., 2018). The leaf-dating model was further combined with a Bayesian Skyline non-parametric model for the tree prior distribution, in order to infer the past population dynamics of MDV (Drummond et al., 2005). By providing dating data for the analysed sequences, BEAST is able to estimate an evolutionary history timescale for a sample. This is possible as BEAST fixes the dates on the leaf nodes

of a tree while each branch, in units of time, is matched to an expected number of substitutions per site through a molecular clock model (either strict or relaxed) (Shapiro et al., 2011). In the case of a relaxed clock model, no correlation of the rates on adjacent branches of the tree is assumed (Drummond et al., 2006).

When trying to infer tree topologies of datasets that incorporate both modern and ancient DNA data, a relaxed clock model combined with a tip dating method can provide an integrated framework in elucidating ancestral divergence dates and phylogenetic relationships, while accounting for the time-dependent nature of the evolutionary process without assuming a strict molecular clock (Drummond et al., 2006). Concurrently, a relaxed clock model also allows for the estimation of the position of the root of the tree, even in the absence of a non-reversible model of substitutions or a known outgroup (Drummond et al., 2006). On the other hand, a relaxed molecular clock approach can have certain drawbacks, for instance non-accounted rate heterogeneity in the evolutionary process could lead to false estimated tip-dates in sequences with missing dating data (Shapiro et al., 2011). Furthermore, estimating the ages of the leaf nodes depends more on the sequence alignments than the underlying prior distributions the Bayesian model is sampling from, a fact that can be troublesome when analysing ancient DNA sequences that have a high percentage of missing data (Shapiro et al., 2011). Finally, excess variability could be introduced in the analysis when leaf nodes of unknown age attach near the root of the estimated tree (Shapiro et al., 2011).

The estimation of the clock rate from datasets that include both ancient and modern DNA data can be affected by a number of reasons. When a tip-dating model is

employed, the dates provided by the user could be incorrectly reported or estimated, e.g. radiocarbon dates could have been estimated wrongly, thus affecting the calculation of the molecular clock rate and the positioning of the root of the tree (Shapiro et al., 2011). Ancient DNA sequences are also known to be degraded, carrying variable amounts of base misincorporations depending on their age (Hofreiter et al., 2001), thus confounding the estimation of site substitution rates that contribute to the estimation of the molecular clock. Furthermore, BEAST's phylogenetic model assumes that the sequences do not undergo recombination (Shapiro et al., 2011), which could be a potential confounding factor when studying viral genomes across a 500-year evolutionary time scale that are able to recombine. Another factor that could be hindering the calculation of an accurate molecular clock rate is that the coalescent prior distribution may be too constraining or unrepresentative of the data under analysis, especially when aDNA sequences are present.

Even though the results of the present study indicate that MDV lineages have been evolving under a relaxed molecular clock model, the differences in the sampling size of each group should be considered as a source of bias, as the clade of ancient MDV sequences has far fewer individuals compared to both the Eurasian and North American clades. Thus, the possibility of both the molecular clock rate and model being confounded by the uneven number of individuals in each clade should be considered. Although the choice of the relaxed clock model is unlikely to be false, as it is supported by previously published studies on both MDV (Trimpert et al., 2017) and alpha-herpesviruses (Firth et al., 2010) phylogenetics, the clock rate itself could have been affected by the uneven sampling, along with the population dynamics

inferences under the Bayesian Skyline model (Grant, 2015). Sample sizes of less than 50 individuals in total, fail to capture any possible population expansion under a Bayesian Skyline model, as small sample sizes only capture mid-frequency polymorphisms and not the whole extent of mutation diversity in a recently expanded population, a drawback that could also be affecting the calculation of the molecular clock rate (Grant, 2015). Furthermore, aDNA has been shown to inflate the estimation of the clock rate because of base misincorporation and sequencing error due to the degradation of the molecules (Grant, 2015).

Alternatively, the possibility that the molecular clock model is not relaxed enough should be considered, i.e. that the ratio of rates between data partitions, such as different codon positions, is not constant across all lineages (Ho and Lanfear, 2010). Most molecular clock analyses, even relaxed ones, assume that rates are perfectly correlated between the data partitions used for BEAST, so that even though the partitions are allowed to evolve at different rates, the ratio of the rates is fixed across all branches of the tree (Ho and Lanfear, 2010). Such an assumption might not be correct for MDV, whose evolution has been affected by selective pressures introduced by the expanding populations of its host, the chicken, and anthropogenic forces, like the introduction of vaccines.

The time to the most recent common ancestor (tMRCA) was predicted to be at 1477 AD, almost 500 years earlier (1938) than the tMRCA calculated by Trimpert et al (2017), and almost 100 years before the earliest sample included in our dataset. Even though our tMRCA estimation is highly dependent on the age of our oldest sample, it still represents a more accurate estimate of the time that MDV first

emerged in chickens. The tMRCA estimated by Trimpert et al (2017) could not have been the earliest one as the disease was first described in 1907, unless the causative agent for the disease reported by Josef Marek in the early 20th century was a different viral pathogen. Furthermore, the tMRCA by Trimpert et al (2017) was estimated from strains that emerged after the 1960s, and included no information from milder or putatively non-pathogenic strains that could have been circulating among chicken populations prior to that decade. The 4 ancient MDV sequences cluster together in a clade, which falls basal to the Eurasian (blue) and American (orange) clades that consist of sequences sampled from the second half of the 20th century until the present. From the same tree it can be estimated that the Eurasian and American clade diverged approximately 175 years ago, coinciding with the gradual intensification of chicken management in both Europe (Girdland Flink et al., 2014) and North America (Derry, 2015) that took place in the second half of the 19th century. For future work more ancient and modern MDV genomes, for denser temporal sampling, are needed, in order to estimate changes in the effective population size (N_e) of the virus

Population structure

The ancient MDV population is almost equally genetically distant from both modern American and Eurasian MDV populations, with both modern Eurasian and European populations being genetically closer as shown by the pairwise Dxy comparisons, further corroborating the tree topology inferred from our phylogenetic analysis. This lineage divergence effect could be the combined result of the industrialized

management of chickens, since the mid 20th century, along with the introduction of mass vaccination regimes against MDV. It has been proposed that even though the vaccine is effective at fighting the symptoms of Marek's Disease (MD) it is not effective at stopping the transmission or the persistence of the virus (Osterrieder et al., 2006; Witter, 1998). Therefore, while the vaccine protects chicken flocks against the virus, it also acts as an evolutionary pressure that leads MDV to evolve towards a more pathogenic state and break the protection barrier offered by vaccination (Read et al., 2015).

For a more detailed and robust demographic analysis, a greater number of ancient and modern MDV genomes is required. A denser population sampling will allow for populations to be defined both temporally and spatially, instead as per lineage, enabling us to test explicitly different demographic models. Such an approach would allow us to estimate if and when a bottleneck happened, how strongly it affected the effective population size of MDV, and whether it coincided with a time when humans changed and/or intensified their agricultural practices.

Human activity shaping the chicken-MDV interaction landscape

In conclusion, the SNP analysis hints that the repeated genomic regions of the ancient MDV strains might have a higher mutation rate. A comprehensive phylogenetic analysis that combines both a maximum likelihood and a bayesian phylogenetic tree estimation, supports a relaxed molecular clock model for the evolution of MDV. Different branches and lineages are shown to be evolving at

various rates throughout time and space, further supporting the idea that the way humans manage chicken flocks has affected the evolutionary trajectory of MDV.

Our population divergence analysis of the viral MDV lineages show that the Ancient MDV lineage is equidistant from the American and Eurasian lineages that have been circulating among chicken populations since the second half of the 20th century, and also has a higher amount of nucleotide diversity than the extinct viral lineages. It has been empirically observed, especially after 1907 when Marek's Disease was first described, that MDV pathotypes get increasingly virulent in a shorter amount of time (Osterrieder et al., 2006; Witter, 1998). The highly virulent pathotypes circulating today in chicken flocks cause novel and more acute symptoms than the mild strains that inflicted chickens when the disease was first described in 1907 (Osterrieder et al., 2006). More specifically, most virulent MDV pathotypes emerged after the 1970s, when vaccination against MDV was first introduced (Witter, 1998, 1997).

This empirical finding combined with the genetic divergence of the Ancient MDV lineage from the extant American and Eurasian ones hints that MDV's evolutionary trajectory has been shaped by the intensified chicken flock management introduced by humans especially from the 18th century onwards (Derry, 2015; Girdland Flink et al., 2014). Even though humans are not selecting explicitly for MDV's virulence, through shorter broiler cohort times they are probably selecting for increased speed of viral replication and speed of transmission (Atkins et al., 2013, 2011). Additionally after the 1970s, farmers could be inadvertently selecting for the same traits that favour pathogenic phenotypes of MDV, via the introduction of imperfect vaccine programmes introduced against MDV (Atkins et al., 2013; Read et al., 2015),

possibly creating a population bottleneck that is leading to the loss of less virulent strains.

Conclusions

In this study we extracted DNA from bone and feather elements from 1111 modern and archaeological chicken samples, and after performing a metagenomic pathogen screening we identified traces of MDV in 17 archaeological and 1 modern chicken samples. These 18 samples were subsequently enriched for MDV DNA and sequenced, with the capture efficiency ranging from 0.73 to 273.12 fold increase in unique MDV DNA read yield, allowing us to reconstruct consensus genomes with their final genome coverage ranging from 0 to 41.92x.

After analysing the MDV capture sequencing results, we discovered that all the singleton SNPs could possibly be attributed to post mortem DNA damage, and that most polymorphic sites were shared only among a limited number of individuals. A detailed phylogenetic analysis, employing both Maximum Likelihood and Bayesian approaches, revealed that our ancient MDV samples form a lineage that falls basal to the extant American and Eurasian MDV lineages, and this phylogenetic signal is not an artefact of polymorphisms arising from DNA deamination damage in the ancient MDV sequences. Concurrently, we discovered that the MDV lineages have been evolving according to a relaxed molecular clock model, with different branches having different evolutionary rates. After incorporating ancient MDV genomes, we calculated a new tMRCA that places the root of the tree 500 years earlier (1477 AD) than what was previously reported in the literature and we estimated its molecular

clock rate ($9.1\text{E-}6$ substitutions/site/year). Our genomic divergence analysis suggests that the ancient MDV lineage has higher levels of genetic variability compared to modern American and Eurasian lineages. It is also equally divergent from both extant American and Eurasian lineages, further corroborating the tree topology inferred from our phylogenetic analysis.

The estimated divergence time (18th century) between the extant American and Eurasian lineages agrees with historical data describing the intensification of chicken farming beginning approximately the same period of time in North America and Europe. Following the industrialisation of chicken farming in the second half of 20th century empirical data were suggesting an accelerated evolution rate of MDV especially after the emergence of increasingly pathogenic strains after the 1970s, when vaccines against the virus were first introduced. These empirical data further support the results of our study regarding MDV evolving under a relaxed molecular clock model with a faster clock rate than previously estimated. Furthermore, our results underline the effects of human activity (especially after the 1970s) on pathogen evolution, with farmers accidentally selecting for more virulent strains of MDV by industrialising chicken farming and employing inefficient vaccination schemes.

Time series of ancient and modern genomic data and the fact that most viruses have a short generation time have allowed us to observe how pathogens have been evolving as a result of human activity. In the case of MDV, we observed in real time the virus evolving from a putatively less or non virulent state to a lethal pathotype for the chicken host, as a side effect of chicken farming industrialisation. Similar patterns

of evolution have also been discovered in other viral taxa, such as Hepatitis B Virus (HBV) and Variola Virus, whose ancient variants appear to be less virulent than their respective extant and circulating lineages. These findings stress the need of incorporating ancient DNA data in the study of virulence emergence, and the necessity to not only factor in the underlying biology of a virus to explain and predict its pathogenicity, but also the need to consider the effect human activity has been having as an evolutionary pressure.

Bibliography

- Adams, A.M., Hudson, R.R., 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168, 1699–1712.
- Annapragada, A., Borgerson, C., Iams, S., Ravelomanantsoa, M.A., Crawford, G.C., Helin, M., Anjaranirina, E.J.G., Randriamady, H.J., Golden, C.D., 2019. Modeling the Impact of Newcastle Disease Virus Vaccinations on Chicken Production Systems in Northeastern Madagascar. *Front Vet Sci* 6, 305.
- Atkins, K.E., Read, A.F., Savill, N.J., Renz, K.G., Islam, A.F.M.F., Walkden-Brown, S.W., Woolhouse, M.E.J., 2013. Vaccination and reduced cohort duration can drive virulence evolution: Marek's disease virus and industrialized agriculture. *Evolution* 67, 851–860.
- Atkins, K.E., Read, A.F., Savill, N.J., Renz, K.G., Walkden-Brown, S.W., Woolhouse, M.E.J., 2011. Modelling Marek's disease virus (MDV) infection: parameter estimates for mortality rate and infectiousness. *BMC Vet. Res.* 7, 70.
- Baigent, S.J., Davison, F., 2004. 6 - Marek's disease virus: Biology and life cycle, in: Davison, F., Nair, V. (Eds.), *Marek's Disease*. Academic Press, Oxford, pp. 62–ii.
- Baigent, S.J., Davison, T.F., 1999. Development and composition of lymphoid lesions in the spleens of Marek's disease virus-infected chickens: Association with virus spread and the pathogenesis of Marek's disease. *Avian Pathol.* 28, 287–300.
- Baigent, S.J., Petherbridge, L.J., Howes, K., Smith, L.P., Currie, R.J.W., Nair, V.K., 2005. Absolute quantification of Marek's disease virus genome copy number in chicken feather and lymphocyte samples using real-time PCR. *J. Virol. Methods* 123, 53–64.
- Biagini, P., Thèves, C., Balaesque, P., Géraut, A., Cannet, C., Keyser, C., Nikolaeva,

- D., Gérard, P., Duchesne, S., Orlando, L., Willerslev, E., Alekseev, A.N., de Micco, P., Ludes, B., Crubézy, E., 2012. Variola virus in a 300-year-old Siberian mummy. *N. Engl. J. Med.* 367, 2057–2059.
- Briles, W.E., Stone, H.A., Cole, R.K., 1977. Marek's disease: effects of B histocompatibility alloalleles in resistant and susceptible chicken lines. *Science* 195, 193–195.
- Brown, C.E., 1998. Coefficient of Variation, in: Brown, C.E. (Ed.), *Applied Multivariate Statistics in Geohydrology and Related Sciences*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 155–157.
- Bumstead, N., 1998. Genomic mapping of resistance to Marek's disease. *Avian Pathol.* 27, S78–S81.
- Bumstead, N., Sillibourne, J., Rennie, M., Ross, N., Davison, F., 1997. Quantification of Marek's disease virus in chicken lymphocytes using the polymerase chain reaction with fluorescence detection. *J. Virol. Methods* 65, 75–81.
- Calnek, B.W., 1985. Genetic Resistance, in: Payne, L.N. (Ed.), *Marek's Disease: Scientific Basis and Methods of Control*. Springer US, Boston, MA, pp. 293–328.
- Calnek, B.W., Witter, R.L., 1984. Marek's disease.
- Calvignac, S., Terme, J.-M., Hensley, S.M., Jalinot, P., Greenwood, A.D., Hänni, C., 2008. Ancient DNA Identification of Early 20th Century Simian T-Cell Leukemia Virus Type 1. *Mol. Biol. Evol.* 25, 1093–1098.
- Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.H.S., Samaniego, J.A., Wales, N., Sicheritz-Pontén, T., Gilbert, M.T.P., 2018. Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419.
- Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K.,

- Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y., Wang, J., Gilbert, M.T.P., Willerslev, E., Greenleaf, W.J., Bustamante, C.D., 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* 93, 852–864.
- Carrozza, J.H., Fredrickson, T.N., Prince, R.P., Luginbuhl, R.E., 1973. Role of desquamated epithelial cells in transmission of Marek's disease. *Avian Dis.* 17, 767–781.
- Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423.
- Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., Meyer, M., 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* 110, 15758–15763.
- Day, T., 2002. On the evolution of virulence and the relationship between various measures of mortality. *Proc. Biol. Sci.* 269, 1317–1323.
- Deng, Q., Shi, M., Li, Q., Wang, P., Li, M., Wang, W., Gao, Y., Li, H., Lin, L., Huang, T., Wei, P., 2020. Analysis of the evolution and transmission dynamics of the field MDV in China during the years 1995-2020, indicating the emergence of a unique cluster with the molecular characteristics of vv+ MDV that has become endemic in southern China. *Transbound. Emerg. Dis.*
<https://doi.org/10.1111/tbed.13965>

- Derry, M.E., 2015. Chicken Breeding: The Complex Transition from Traditional to Genetic Methods in the USA, in: Phillips, D., Kingsland, S. (Eds.), *New Perspectives on the History of Life Sciences and Agriculture*. Springer International Publishing, Cham, pp. 371–393.
- Dimitrov, K.M., Afonso, C.L., Yu, Q., Miller, P.J., 2017. Newcastle disease vaccines-A solved problem or a continuous challenge? *Vet. Microbiol.* 206, 126–136.
- Dimopoulos, E.A., Carmagnini, A., Velsko, I.M., Warinner, C., Larson, G., Frantz, L.A.F., Irving-Pease, E.K., 2020. HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data. *Cold Spring Harbor Laboratory*. <https://doi.org/10.1101/2020.12.16.419085>
- Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4, e88.
- Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192.
- Duggan, A.T., Perdomo, M.F., Piombino-Mascoli, D., Marciniak, S., Poinar, D., Emery, M.V., Buchmann, J.P., Duchêne, S., Jankauskas, R., Humphreys, M., Golding, G.B., Southon, J., Devault, A., Rouillard, J.-M., Sahl, J.W., Dutour, O., Hedman, K., Sajantila, A., Smith, G.L., Holmes, E.C., Poinar, H.N., 2016. 17th Century Variola Virus Reveals the Recent History of Smallpox. *Curr. Biol.* 26, 3407–3412.
- Emerman, M., Malik, H.S., 2010. Paleovirology--modern consequences of ancient viruses. *PLoS Biol.* 8, e1000301.
- Escorcia, M., Vázquez, L., Méndez, S.T., Rodríguez-Ropón, A., Lucio, E., Nava, G.M., 2008. Avian influenza: genetic evolution under vaccination pressure. *Viol.*

J. 5, 15.

Fay, J.C., Wu, C.I., 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.

Feierbach, B., Piccinotti, S., Bisher, M., Denk, W., Enquist, L.W., 2006.

Alpha-herpesvirus infection induces the formation of nuclear actin filaments.

PLoS Pathog. 2, e85.

Firth, C., Kitchen, A., Shapiro, B., Suchard, M.A., Holmes, E.C., Rambaut, A., 2010.

Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.* 27, 2038–2051.

Gansauge, M.-T., Meyer, M., 2014. Selective enrichment of damaged DNA

molecules for ancient genome sequencing. *Genome Res.* 24, 1543–1549.

Gimeno, I.M., Witter, R.L., Hunt, H.D., Lee, L.F., Reddy, S.M., Neumann, U., 2001.

Marek's disease virus infection in the brain: virus replication, cellular infiltration, and major histocompatibility complex antigen expression. *Vet. Pathol.* 38, 491–503.

Girdland Flink, L., Allen, R., Barnett, R., Malmström, H., Peters, J., Eriksson, J.,

Andersson, L., Dobney, K., Larson, G., 2014. Establishing the validity of domestication genes using DNA from ancient chickens. *Proc. Natl. Acad. Sci. U. S. A.* 111, 6184–6189.

Grant, W.S., 2015. Problems and Cautions With Sequence Mismatch Analysis and

Bayesian Skyline Plots to Infer Historical Demography. *J. Hered.* 106, 333–346.

He, L., Li, J., Peng, P., Nie, J., Luo, J., Cao, Y., Xue, C., 2020. Genomic analysis of a

Chinese MDV strain derived from vaccine strain CVI988 through recombination. *Infect. Genet. Evol.* 78, 104045.

He, W., Auclert, L.Z., Zhai, X., Wong, G., Zhang, C., Zhu, H., Xing, G., Wang, S., He,

- W., Li, K., Wang, L., Han, G.-Z., Veit, M., Zhou, J., Su, S., 2019. Interspecies Transmission, Genetic Diversity, and Evolutionary Dynamics of Pseudorabies Virus. *J. Infect. Dis.* 219, 1705–1715.
- Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M., Pääbo, S., 2001. Ancient DNA. *Nat. Rev. Genet.* 2, 353–359.
- Ho, S.Y.W., Lanfear, R., 2010. Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. *Mitochondrial DNA* 21, 138–146.
- Hunt, H.D., Lupiani, B., Miller, M.M., Gimeno, I., Lee, L.F., Parcels, M.S., 2001. Marek's disease virus down-regulates surface expression of MHC (B Complex) Class I (BF) glycoproteins during active but not latent infection of chicken cells. *Virology* 282, 198–205.
- Islam, A.F.M.F., Wong, C.W., Walkden-Brown, S.W., Colditz, I.G., Arzey, K.E., Groves, P.J., 2002. Immunosuppressive effects of Marek's disease virus (MDV) and herpesvirus of turkeys (HVT) in broiler chickens and the protective effect of HVT vaccination against MDV challenge. *Avian Pathol.* 31, 449–461.
- Islam, A., Walkden-Brown, S.W., 2007. Quantitative profiling of the shedding rate of the three Marek's disease virus (MDV) serotypes reveals that challenge with virulent MDV markedly increases shedding of vaccinal viruses. *J. Gen. Virol.* 88, 2121–2128.
- Iyer, L.M., Aravind, L., Koonin, E.V., 2001. Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* 75, 11720–11734.
- Iyer, L.M., Balaji, S., Koonin, E.V., Aravind, L., 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.* 117, 156–184.
- Jane Flint, S., Racaniello, V.R., Rall, G.F., Hatziioannou, T., Skalka, A.M., 2020.

- Principles of Virology, Volume 2: Pathogenesis and Control. John Wiley & Sons.
- Johnson, E.A., Burke, C.N., Fredrickson, T.N., DiCapua, R.A., 1975. Morphogenesis of marek's disease virus in feather follicle epithelium. *J. Natl. Cancer Inst.* 55, 89–99.
- Jun, G., Wing, M.K., Abecasis, G.R., Kang, H.M., 2015. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* 25, 918–925.
- Kahila Bar-Gal, G., Kim, M.J., Klein, A., Shin, D.H., Oh, C.S., Kim, J.W., Kim, T.-H., Kim, S.B., Grant, P.R., Pappo, O., Spigelman, M., Shouval, D., 2012. Tracing hepatitis B virus to the 16th century in a Korean mummy. *Hepatology* 56, 1671–1680.
- Kaiser, P., Underwood, G., Davison, F., 2003. Differential cytokine responses following Marek's disease virus infection of chickens differing in resistance to Marek's disease. *J. Virol.* 77, 762–768.
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Katz, J.M., 2003. The impact of avian influenza viruses on public health. *Avian Dis.* 47, 914–920.
- Kaufman, J., Völk, H., Wallny, H.J., 1995. A “minimal essential Mhc” and an “unrecognized Mhc”: two extremes in selection for polymorphism. *Immunol. Rev.* 143, 63–88.
- Kautto, A.H., Alenius, S., Mossing, T., Becher, P., Belák, S., Larska, M., 2012. Pestivirus and alphaherpesvirus infections in Swedish reindeer (*Rangifer tarandus tarandus* L.). *Vet. Microbiol.* 156, 64–71.

- Kennedy, D.A., Dunn, P.A., Read, A.F., 2018. Modeling Marek's disease virus transmission: A framework for evaluating the impact of farming practices and evolution. *Epidemics* 23, 85–95.
- Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15, 356.
- Kramer, T., Enquist, L.W., 2012. Alphaherpesvirus infection disrupts mitochondrial transport in neurons. *Cell Host Microbe* 11, 504–514.
- Lanier, L.L., Phillips, J.H., 1996. Inhibitory MHC class I receptors on NK cells and T cells. *Immunol. Today* 17, 86–91.
- Lapierre, M., Lambert, A., Achaz, G., 2017. Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. *Genetics* 206, 439–449.
- Lee, L.F., Lupiani, B., Silva, R.F., Kung, H.-J., Reddy, S.M., 2008. Recombinant Marek's disease virus (MDV) lacking the Meq oncogene confers protection against challenge with a very virulent plus strain of MDV. *Vaccine* 26, 1887–1892.
- Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., Claverie, J.-M., 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U. S. A.* 111, 4274–4279.
- Lewis, P.O., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925.
- Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.

- Li, H., 2012. seqtk Toolkit for processing sequences in FASTA/Q formats. GitHub 767, 69.
- Li, H., n.d. htsbox. Github.
- Li, T., Xie, J., Liang, G., Ren, D., Sun, S., Lv, L., Xie, Q., Shao, H., Gao, W., Qin, A., Ye, J., 2019. Co-infection of vvMDV with multiple subgroups of avian leukosis viruses in indigenous chicken flocks in China. *BMC Vet. Res.* 15, 288.
- Loncoman, C.A., Hartley, C.A., Coppo, M.J.C., Vaz, P.K., Diaz-Méndez, A., Browning, G.F., García, M., Spatz, S., Devlin, J.M., 2017a. Genetic Diversity of Infectious Laryngotracheitis Virus during In Vivo Coinfection Parallels Viral Replication and Arises from Recombination Hot Spots within the Genome. *Appl. Environ. Microbiol.* 83. <https://doi.org/10.1128/AEM.01532-17>
- Loncoman, C.A., Vaz, P.K., Coppo, M.J., Hartley, C.A., Morera, F.J., Browning, G.F., Devlin, J.M., 2017b. Natural recombination in alphaherpesviruses: Insights into viral evolution through full genome sequencing and sequence analysis. *Infect. Genet. Evol.* 49, 174–185.
- Longenecker, B.M., Pazderka, F., Gavora, J.S., Spencer, J.L., Stephens, E.A., Witter, R.L., Ruth, R.F., 1977. Role of the Major Histocompatibility Complex in Resistance to Marek's Disease: Restriction of the Growth of JMV-MD Tumor Cells in Genetically Resistant Birds, in: Benedict, A.A. (Ed.), *Avian Immunology*. Springer US, Boston, MA, pp. 287–298.
- Marth, G.T., Czabarka, E., Murvai, J., Sherry, S.T., 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166, 351–372.
- Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., Feng, J., Jiang, D., Ahani, H., Abbott, R.J., Lascoux, M., Nevo, E., Liu, J., 2018. Ancient polymorphisms and

- divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proc. Natl. Acad. Sci. U. S. A.* 115, E236–E243.
- Nei, M., Miller, J.C., 1990. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics* 125, 873–879.
- Osterrieder, N., Kamil, J.P., Schumacher, D., Tischer, B.K., Trapp, S., 2006. Marek's disease virus: from miasma to model. *Nat. Rev. Microbiol.* 4, 283–294.
- Padhi, A., Parcells, M.S., 2016. Positive Selection Drives Rapid Evolution of the meq Oncogene of Marek's Disease Virus. *PLoS One* 11, e0162180.
- Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., Harris, S.R., 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2, e000056.
- Pandey, U., Bell, A.S., Renner, D.W., Kennedy, D.A., Shreve, J.T., Cairns, C.L., Jones, M.J., Dunn, P.A., Read, A.F., Szpara, M.L., 2016. DNA from Dust: Comparative Genomics of Large DNA Viruses in Field Surveillance Samples. *mSphere* 1. <https://doi.org/10.1128/mSphere.00132-16>
- Paradis, E., Schliep, K., 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528.
- Patterson Ross, Z., Klunk, J., Fornaciari, G., Giuffra, V., Duchêne, S., Duggan, A.T., Poinar, D., Douglas, M.W., Eden, J.-S., Holmes, E.C., Poinar, H.N., 2018. The paradox of HBV evolution as revealed from a 16th century mummy. *PLoS Pathog.* 14, e1006750.
- Pedersen, B., n.d. pyfasta. Github.
- Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E., Lercher, M.J., 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol.*

- Evol. 31, 1929–1936.
- Picard toolkit, 2019. . Broad Institute, GitHub repository.
- Posada, D., 2001. Unveiling the molecular clock in the presence of recombination. Mol. Biol. Evol.
- Rambaut, A., n.d. figtree. Github.
- Rasmussen, S., Allentoft, M.E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., Pedersen, A.G., Schubert, M., Van Dam, A., Kapel, C.M.O., Nielsen, H.B., Brunak, S., Avetisyan, P., Epimakhov, A., Khalyapin, M.V., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., Moiseyev, V., Gromov, A., Pokutta, D., Saag, L., Varul, L., Yepiskoposyan, L., Sicheritz-Pontén, T., Foley, R.A., Lahr, M.M., Nielsen, R., Kristiansen, K., Willerslev, E., 2015. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. Cell 163, 571–582.
- Read, A.F., Baigent, S.J., Powers, C., Kgosa, L.B., Blackwell, L., Smith, L.P., Kennedy, D.A., Walkden-Brown, S.W., Nair, V.K., 2015. Imperfect Vaccination Can Enhance the Transmission of Highly Virulent Pathogens. PLoS Biol. 13, e1002198.
- Rémy, S., Blondeau, C., Le Vern, Y., Lemesle, M., Vautherot, J.-F., Denesvre, C., 2013. Fluorescent tagging of VP22 in N-terminus reveals that VP22 favors Marek's disease virus (MDV) virulence in chickens and allows morphogenesis study in MD tumor cells. Vet. Res. 44, 125.
- Rickinson, A.B., virus Fields, K.E.E.-B., 1996. virology 3rd ed. Fields BN, Knipe DM, Howley PM 2397, 2415.
- Schierup, M.H., Hein, J., 2000. Recombination and the molecular clock. Mol. Biol. Evol. 17, 1578–1579.
- Shamblin, C.E., Greene, N., Arumugaswami, V., Dienglewicz, R.L., Parcels, M.S.,

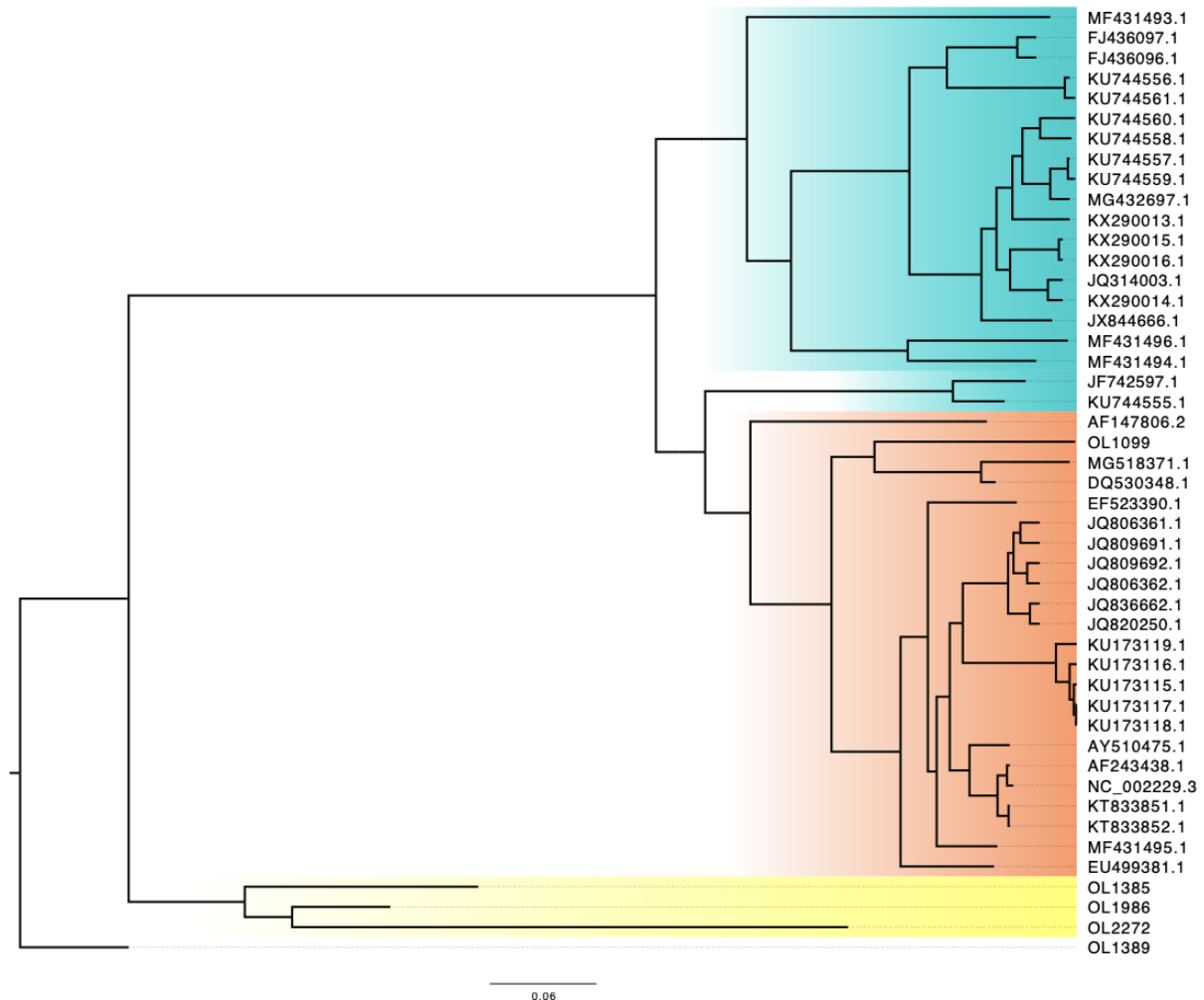
2004. Comparative analysis of Marek's disease virus (MDV) glycoprotein-, lytic antigen pp38- and transformation antigen Meq-encoding genes: association of meq mutations with MDVs of high virulence. *Vet. Microbiol.* 102, 147–167.
- Shapiro, B., Ho, S.Y.W., Drummond, A.J., Suchard, M.A., Pybus, O.G., Rambaut, A., 2011. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* 28, 879–887.
- Shek, W.R., Calnek, B.W., Schat, K.A., Chen, C.H., 1983. Characterization of Marek's disease virus-infected lymphocytes: discrimination between cytolytically and latently infected cells. *J. Natl. Cancer Inst.* 70, 485–491.
- Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* 11, e0163962.
- Smith, G., Dunipace, S., 2011. How backyard poultry flocks influence the effort required to curtail avian influenza epidemics in commercial poultry flocks. *Epidemics* 3, 71–75.
- Spatz, S.J., Zhao, Y., Petherbridge, L., Smith, L.P., Baigent, S.J., Nair, V., 2007. Comparative sequence analysis of a highly oncogenic but horizontal spread-defective clone of Marek's disease virus. *Virus Genes* 35, 753–766.
- Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
- Stephenson, I., Zambon, M., 2002. The epidemiology of influenza. *Occup. Med.* 52, 241–247.
- Stone, A.C., Ozga, A.T., 2019. Chapter 8 - Ancient DNA in the Study of Ancient Disease, in: Buikstra, J.E. (Ed.), *Ortner's Identification of Pathological Conditions in Human Skeletal Remains (Third Edition)*. Academic Press, San Diego, pp. 183–210.

- Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., Rambaut, A., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 4, vey016.
- Szpara, M.L., Kobiler, O., Enquist, L.W., 2010. A common neuronal response to alphaherpesvirus infection. *J. Neuroimmune Pharmacol.* 5, 418–427.
- Taubenberger, J.K., Reid, A.H., Lourens, R.M., Wang, R., Jin, G., Fanning, T.G., 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature* 437, 889–893.
- Thiry, J., Keuser, V., Muylkens, B., Meurens, F., Gogev, S., Vanderplasschen, A., Thiry, E., 2006. Ruminant alphaherpesviruses related to bovine herpesvirus 1. *Vet. Res.* 37, 169–190.
- Trimpert, J., Groenke, N., Jenckel, M., He, S., Kunec, D., Szpara, M.L., Spatz, S.J., Osterrieder, N., McMahon, D.P., 2017. A phylogenomic analysis of Marek's disease virus reveals independent paths to virulence in Eurasia and North America. *Evol. Appl.* 10, 1091–1101.
- Tsangaras, K., Greenwood, A.D., 2019. Paleovirology: Viral Sequences from Historical and Ancient DNA, in: Lindqvist, C., Rajora, O.P. (Eds.), *Paleogenomics: Genome-Scale Analysis of Ancient DNA*. Springer International Publishing, Cham, pp. 139–162.
- Tumpey, T.M., Basler, C.F., Aguilar, P.V., Zeng, H., Solórzano, A., Swayne, D.E., Cox, N.J., Katz, J.M., Taubenberger, J.K., Palese, P., García-Sastre, A., 2005. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* 310, 77–80.
- Van der Auwera, G.A., O'Connor, B.D., n.d. *Genomics in the Cloud*. O'Reilly Media, Inc.

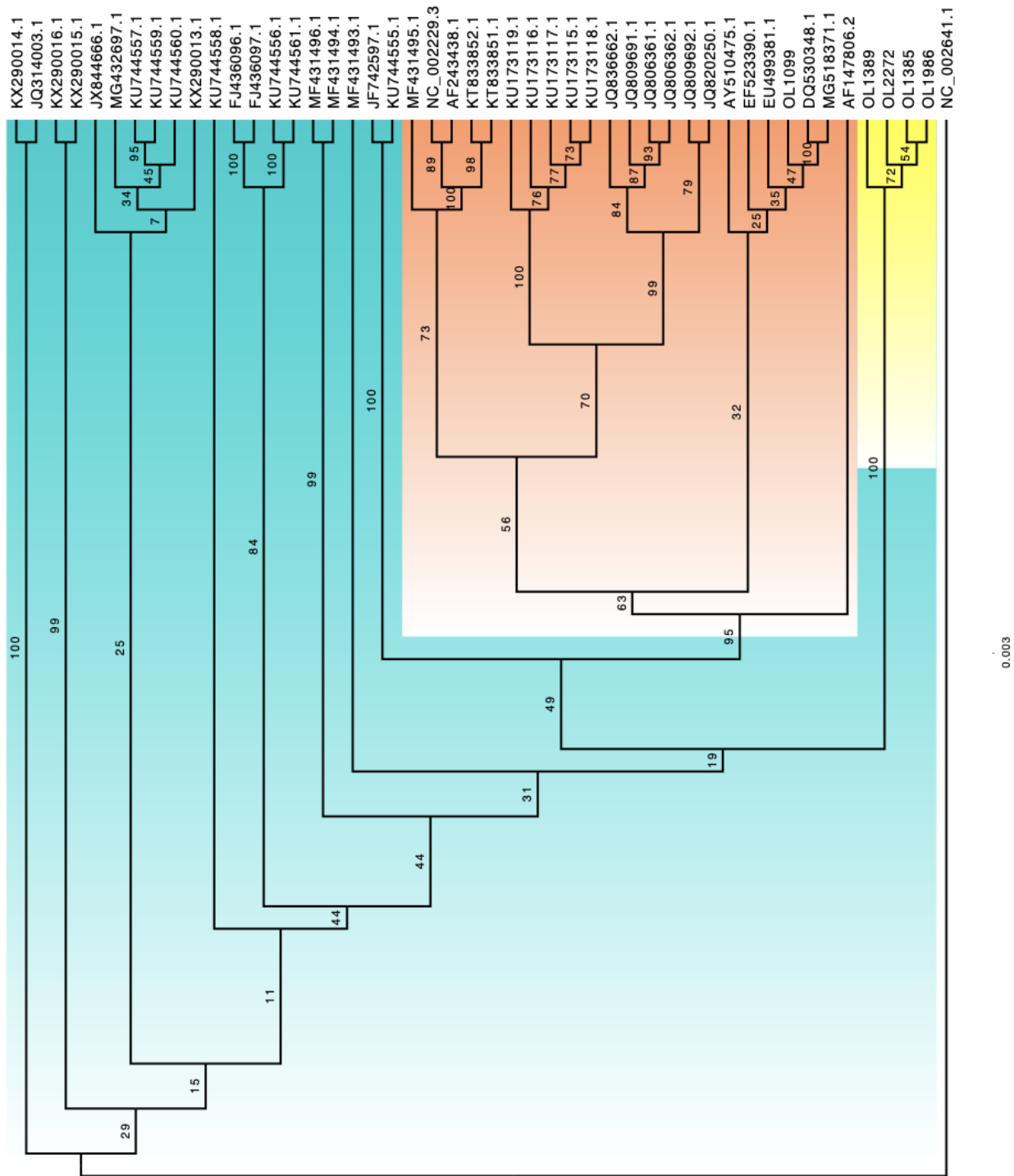
- Witter, R.L., 2001. Protective Efficacy of Marek's Disease Vaccines, in: Hirai, K. (Ed.), Marek's Disease. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 57–90.
- Witter, R.L., 1998. The changing landscape of Marek's disease. *Avian Pathol.* 27, S46–S53.
- Witter, R.L., 1997. Increased virulence of Marek's disease virus field isolates. *Avian Dis.* 41, 149–163.
- Witter, R.L., Lee, L.F., 1984. Polyvalent Marek's disease vaccines: safety, efficacy and protective synergism in chickens with maternal antibodies. *Avian Pathol.* 13, 75–92.
- Worobey, M., Watts, T.D., McKay, R.A., Suchard, M.A., Granade, T., Teuwen, D.E., Koblin, B.A., Heneine, W., Lemey, P., Jaffe, H.W., 2016. 1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* 539, 98–101.
- Woźniakowski, G., Samorek-Salamonowicz, E., 2015. Animal herpesviruses and their zoonotic potential for cross-species infection. *Ann. Agric. Environ. Med.* 22, 191–194.
- Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.-H., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* 60, 150–160.
- Yabe, T., McSherry, C., Bach, F.H., Fisch, P., Schall, R.P., Soudel, P.M., Houchins, J.P., 1993. A multigene family on human chromosome 12 encodes natural killer-cell lectins. *Immunogenetics* 37, 455–460.
- Yu, Z., Zhang, Y., Lan, X., Wang, Y., Zhang, F., Gao, Y., Li, K., Gao, L., Pan, Q., Qi, X., Cui, H., Zhou, L., Sun, G., Wang, X., Liu, C., 2020. Natural co-infection with

two virulent wild strains of Marek's disease virus in a commercial layer flock. Vet. Microbiol. 240, 108501.

Supplementary



Supplementary figure 1. MDV phylogenetic tree under a strict clock model. Tip dated maximum clade credibility MDV phylogenetic tree, with a midpoint root, as estimated by BEAST under a strict clock model. As seen in the ML tree, all ancient MDV genomes (yellow) cluster together basal to the modern Eurasian and American MDV lineages (blue and orange respectively). Its topology is highly similar to the topology that is estimated from the relaxed clock model. To compensate for a lower clock rate across all the branches ($4.7\text{E-}3$ substitutions per site per year), under the strict clock model the age of the root of the tree is pushed back to 1432 AD (with a 95% HPD interval from 1360 to 1490).



Supplementary figure 2. Maximum likelihood phylogenetic tree with no transition mutations allowed. Transition mutations were set as unknown bases (N). The modern OL1099 sample clusters with the American (orange) MDV lineage, as in the ML tree that includes all types of mutations.

Supplementary table 1. Metagenomic identification results. Ancient and modern chicken bone samples that were positive or negative for MDV DNA.

Sample	Taxon	Mean Posterior Abundance	Uniquely Assigned Read #	Evenness of Coverage Ratio	Fraction of Genome Covered	Coverage
OL1008	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.00000104	1	1	3.15E-04	0
OL1099	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.001056777	943	1.23	2.69E-01	0.33
OL1128	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.0000124	1	1	3.54E-04	0
OL1214	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.00000105	0	0	0.00E+00	0
OL1231	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.0000357	5	1	1.61E-03	0
OL1385	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.0000956	1183	1.13	2.02E-01	0.23
OL1389	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.00000319	19	1.67	3.02E-03	0.01
OL1584	<i>Gallid</i> <i>alphaherpesvirus 2</i>	0.0000035	25	2.85	3.19E-03	0.01
OL1585	<i>Gallid</i>	0.00000301	8	1.25	1.65E-03	0

	<i>alphaherpesvirus 2</i>					
	<i>Gallid</i>					
OL1599	<i>alphaherpesvirus 2</i>	0.00000367	4	1	1.33E-03	0
	<i>Gallid</i>					
OL1934	<i>alphaherpesvirus 2</i>	0.00000302	2	1	7.31E-04	0
	<i>Gallid</i>					
OL1936	<i>alphaherpesvirus 2</i>	0.0000175	8	1	2.37E-03	0
	<i>Gallid</i>					
OL1984	<i>alphaherpesvirus 2</i>	0.000015	5	1	1.54E-03	0
	<i>Gallid</i>					
OL1986	<i>alphaherpesvirus 2</i>	0.0000188	25	1	8.29E-03	0.01
	<i>Gallid</i>					
OL1987	<i>alphaherpesvirus 2</i>	0.0000118	14	1.26	3.90E-03	0
	<i>Gallid</i>					
OL1999	<i>alphaherpesvirus 2</i>	0.00000459	4	1.21	1.26E-03	0
	<i>Gallid</i>					
OL2178	<i>alphaherpesvirus 2</i>	0.0000746	87	1	2.41E-02	0.02
	<i>Gallid</i>					
OL2267	<i>alphaherpesvirus 2</i>	0.00000645	20	1.05	4.29E-03	0
	<i>Gallid</i>					
OL2268	<i>alphaherpesvirus 2</i>	0.00000308	15	1.03	4.78E-03	0
OL2272	<i>Gallid</i>	0.000012	47	1.07	1.41E-02	0.01

	<i>alpha</i> herpesvirus 2					

Supplementary table 2. Sample statistics of unknown bases, sequencing depth and coverage after the targeted MDV capture.

	Unique MDV reads from screening libraries	Unique MDV reads from captured libraries	Percentage of duplicate reads in screening libraries	Percentage of duplicate reads in captured libraries	Average sequencing depth (after genome capture)	Average genome coverage (before capture)	Average genome coverage (after capture)	Capture efficiency (X fold increase)	Percentage of unknown bases (Ns) after capture
OL1008	1	59	0	89.72	0	0.00	0.01	8.91	100%
OL1099	931	5423	1.27	99	6.67	0.33	4.04	6.96	58.60%
OL1128	1	24	0	56.6	0	0.00	0	0.73	100%
OL1214	0	451	0	91.3	0	0.00	0.08	0.00	100%
OL1231	5	19	0	89.16	4	0.00	0	0.77	99.99%
OL1385	785	134178	33.64	99.69	38.22	0.23	41.92	57.15	17.08%
OL1389	19	22234	4	99.78	7.94	0.01	6.85	273.12	40.28%

OL1584	24	1536	0	99.9	8.49	0.01	0.65	7.50	99.39%
OL1585	8	949	0	99.87	4.75	0.00	0.26	17.43	99.92%
OL1599	4	12924	0	99.79	5.61	0.00	3.72	153.77	70.33%
OL1934	2	4828	0	98.12	4.89	0.00	1.91	32.44	89.35%
OL1936	7	8613	12.5	99.92	5.96	0.00	3.24	15.79	74.03%
OL1984	5	2868	0	99.97	4.89	0.00	1.12	6.68	96.12%
OL1986	25	14971	0	99.95	6.79	0.01	5.78	12.60	46.16%
OL1987	14	2054	0	99.93	4.87	0.00	0.78	10.06	97.99%
OL1999	4	320	0	99.92	4	0.00	0.13	3.42	99.97%
OL2178	79	6025	9.2	99.98	5.56	0.02	1.84	1.65	89.66%
OL2267	19	683	5	99.98	4.49	0.00	0.26	6.03	99.56%
OL2268	15	732	0	99.98	5.28	0.00	0.27	8.38	99.52%
OL2272	46	16967	2.13	99.92	7.83	0.01	6.81	32.77	35.08%

Supplementary table 3. Sample location and age metadata. Samples were dated either through radiocarbon dating or by archaeological context.

Extract Code	Location	Country	Age	Skeletal element
OL1008	Rudnik	Serbia	14th - 15th century AD	Coracoid
OL1099	Unknown location	Unknown country	21st century AD (20014)	Feather
OL1128	Fishbourne (East)	England	Roman	Coracoid
OL1214	Studenica	Serbia	14th - 15th century AD	Coracoid
OL1231	Vienna Csokorgasse (Avar Cemetery)	Austria	7th century AD	Coracoid
OL1385	Buda Castle, Budapest, Telaki Palace	Hungary	16th - 17th century AD	Coracoid
OL1389	Buda Castle, Budapest, Telaki Palace	Hungary	16th - 17th century AD	Coracoid
OL1584	Kazan City	Russia	16th - 17th century AD	Coracoid
OL1585	Cheboksary City	Russia	16th - 18th century AD	Coracoid
OL1599	Elabuga hillfort	Russia	17th - 18th century AD	Coracoid
OL1934	Andlau	France	10th -12th century AD	Coracoid
OL1936	Strasbourg Rue de Lucerne - Rue du Jeu de Paume	France	16th - 17th century AD	Coracoid
OL1984	Castillo de Montsoriu	Spain	16th century AD	Coracoid

OL1986	Castillo de Montsoriu	Spain	16th century AD	Coracoid
OL1999	Castillo de Montsoriu	Spain	16th century AD	Coracoid
OL2178	Bard-e Mar (Bem), Kurdistan	Iran	Unknown age	Femur
OL2235	Fuerte San Rafael	Argentina	Unknown age	Femur
OL2267	Naderi Tepe	Iran	19th century AD	Coracoid
OL2268	Naderi Tepe	Iran	Unknown age	Coracoid
OL2272	Naderi Tepe	Iran	19th century AD	Femur

Supplementary table 4. Genome metadata for the modern MDV and HVT genomes downloaded from NCBI.

Accession #	Taxon	Country	Sequencing method	Year of collection
AF147806.2	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1964
AF243438.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1977
AY510475.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1977
DQ530348.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	Netherlands	Sanger	1969
EF523390.1	<i>Gallid</i>	USA	Sanger	1981

	<i>alphaherpesvirus 2</i>			
EU499381.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1968
FJ436096.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	United Kingdom	NGS - pyrosequencing	1992
FJ436097.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	United Kingdom	NGS - pyrosequencing	1992
JF742597.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	Sanger	1986
JQ314003.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	Sanger	2007
JQ806361.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1994
JQ806362.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1994
JQ809691.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1994
JQ809692.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1994
JQ820250.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1994
JQ836662.1	<i>Gallid</i>	USA	Sanger	1994

	<i>alphaherpesvirus 2</i>			
JX844666.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	Sanger	2001
KT833851.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	NGS - Illumina	1977
KT833852.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	NGS - Illumina	1977
KU173115.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	NGS - Illumina	2015
KU173116.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	NGS - Illumina	2015
KU173117.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	NGS - Illumina	2015
KU173118.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	NGS - Illumina	2015
KU173119.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	NGS - Illumina	2015
KU744555.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	1974
KU744556.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2011
KU744557.1	<i>Gallid</i>	China	NGS - Illumina	2011

	<i>alphaherpesvirus 2</i>			
KU744558.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2012
KU744559.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2014
KU744560.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2014
KU744561.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2014
KX290013.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2011
KX290014.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2007
KX290015.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2007
KX290016.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2007
MF431493.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	Hungary	NGS - Illumina	2000
MF431494.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	Italy	NGS - Illumina	1992
MF431495.1	<i>Gallid</i>	Hungary	NGS - Illumina	1970

	<i>alphaherpesvirus 2</i>			
MF431496.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	Poland	NGS - Illumina	2010
MG432697.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2011
MG518371.1	<i>Gallid</i> <i>alphaherpesvirus 2</i>	China	NGS - Illumina	2011
NC_002229.3	<i>Gallid</i> <i>alphaherpesvirus 2</i>	USA	Sanger	1979
NC_002641.1	<i>Meleagrid</i> <i>herpesvirus 1</i>	USA	Sanger	1970

Supplementary table 5. Site Frequency Spectrum calculated for the 20 newly sequenced MDV genomes.

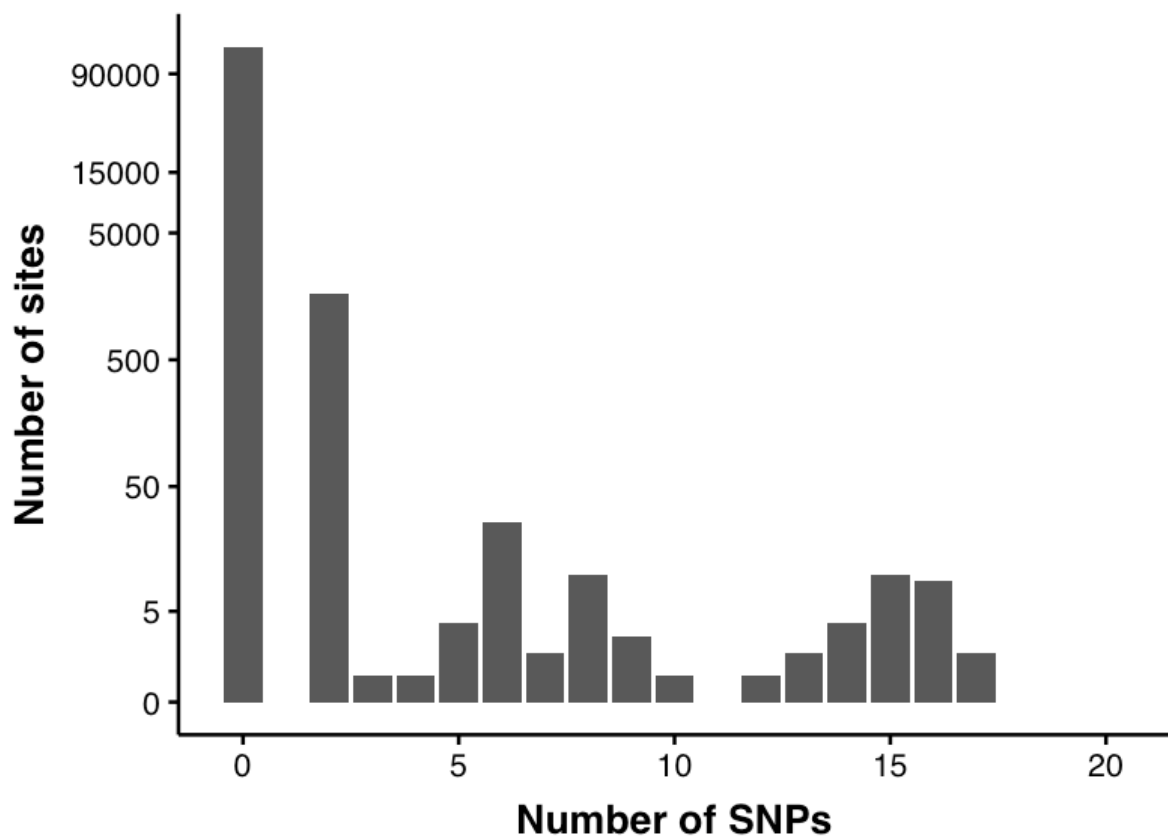
Number of SNPs	Number of sites
0	95407
1	17704
2	10183
3	18456
4	1588

5	4739
6	900
7	222
8	80
9	2
10	4
11	8
12	5
13	6
14	10
15	20
16	16
17	3
18	0
19	1
20	0

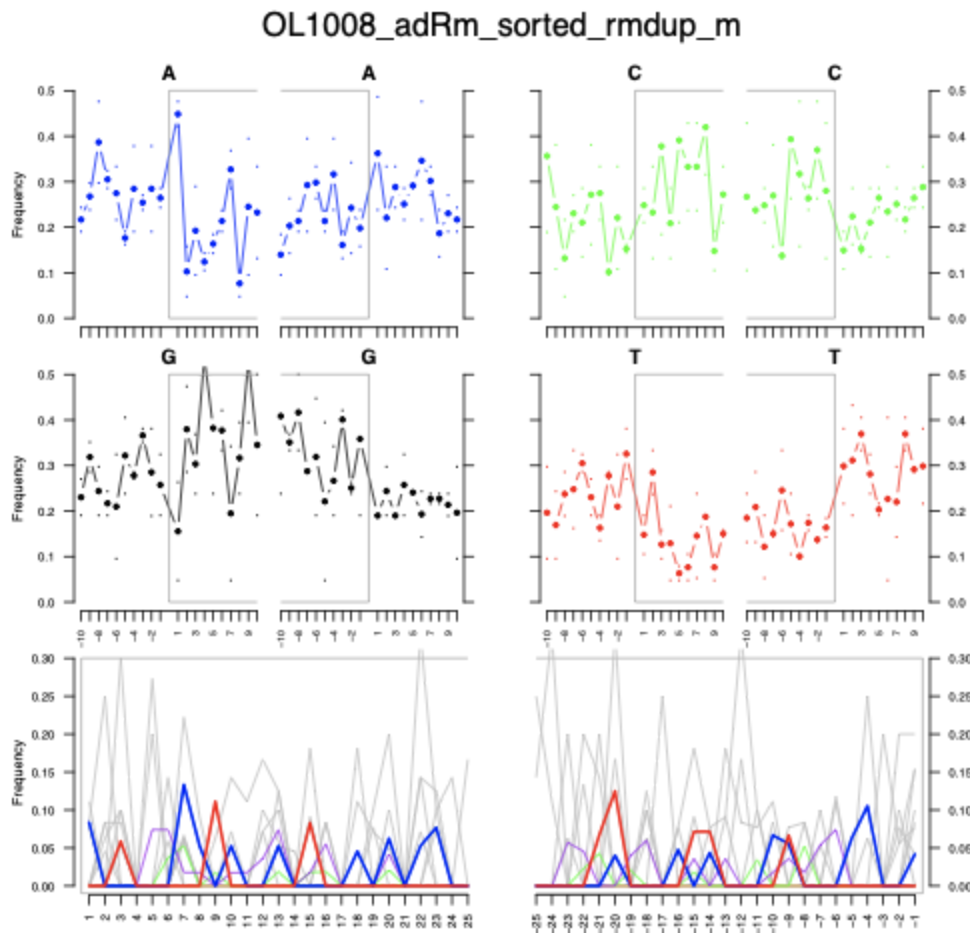
Supplementary table 6. Site Frequency Spectrum calculated for the 20 newly sequenced MDV genomes, excluding all transition mutations.

Number of SNPs	Number of sites
0	147432
1	0
2	1676
3	1
4	1
5	4
6	26
7	2
8	10
9	3
10	1
11	0
12	1
13	2
14	4
15	10

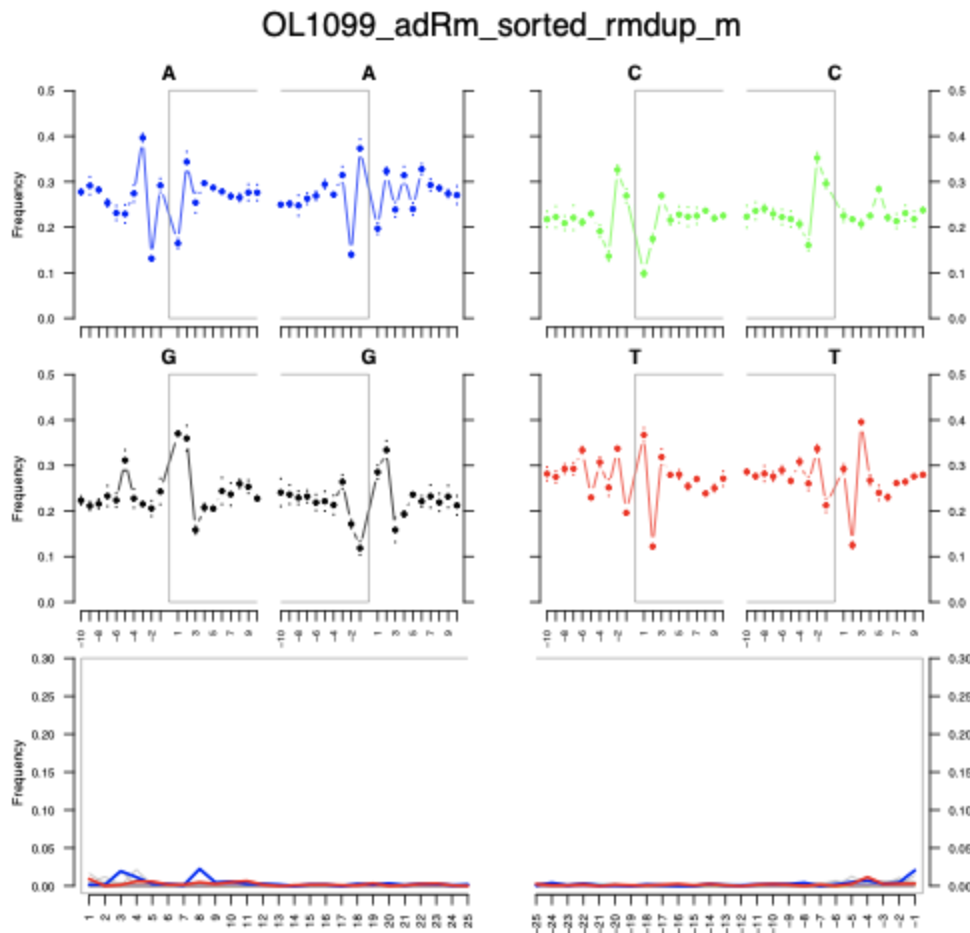
16	9
17	2
18	0
19	0
20	0



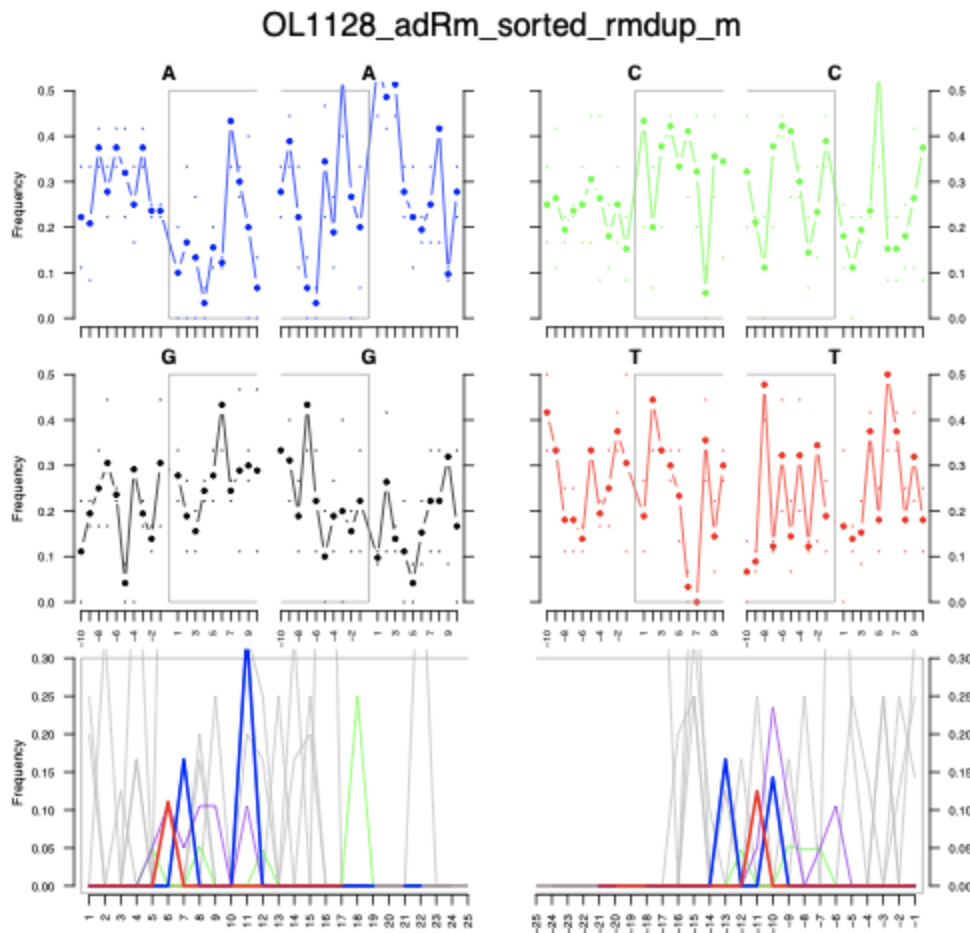
Supplementary figure 3. Histogram of the calculated SFS with no transition polymorphisms, across the 19 archaeological and 1 modern newly captured samples (with a log transformed y-axis).



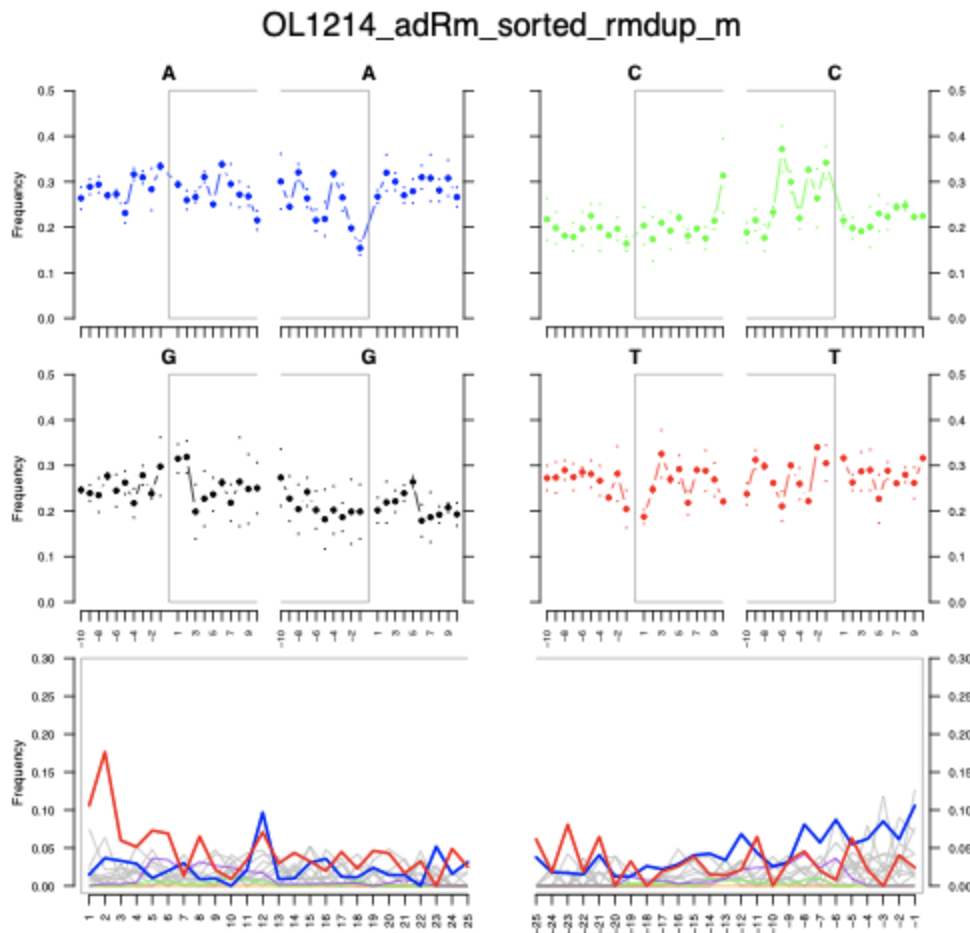
Supplementary figure 4. Deamination damage profile for sample OL1008 for *MDV*, post capture sequencing. No *MDV* DNA was detected in this sample during metagenomic screening, yielding only 59 unique reads after sequencing. This number of reads is too low to give meaningful chemical damage analysis results.



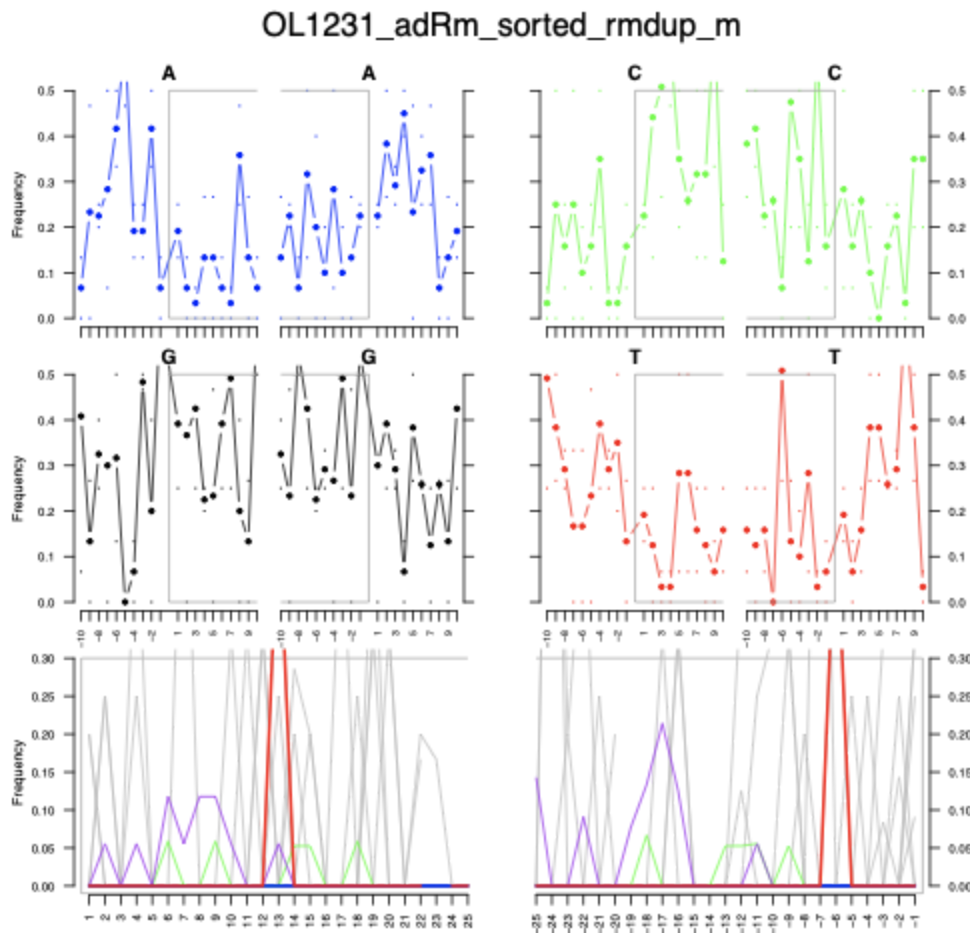
Supplementary figure 5. Deamination damage profile for sample OL1099 for *MDV*, post capture sequencing. This sequencing library was generated from a modern Silkie chicken feather, thus the absence of any DNA damage.



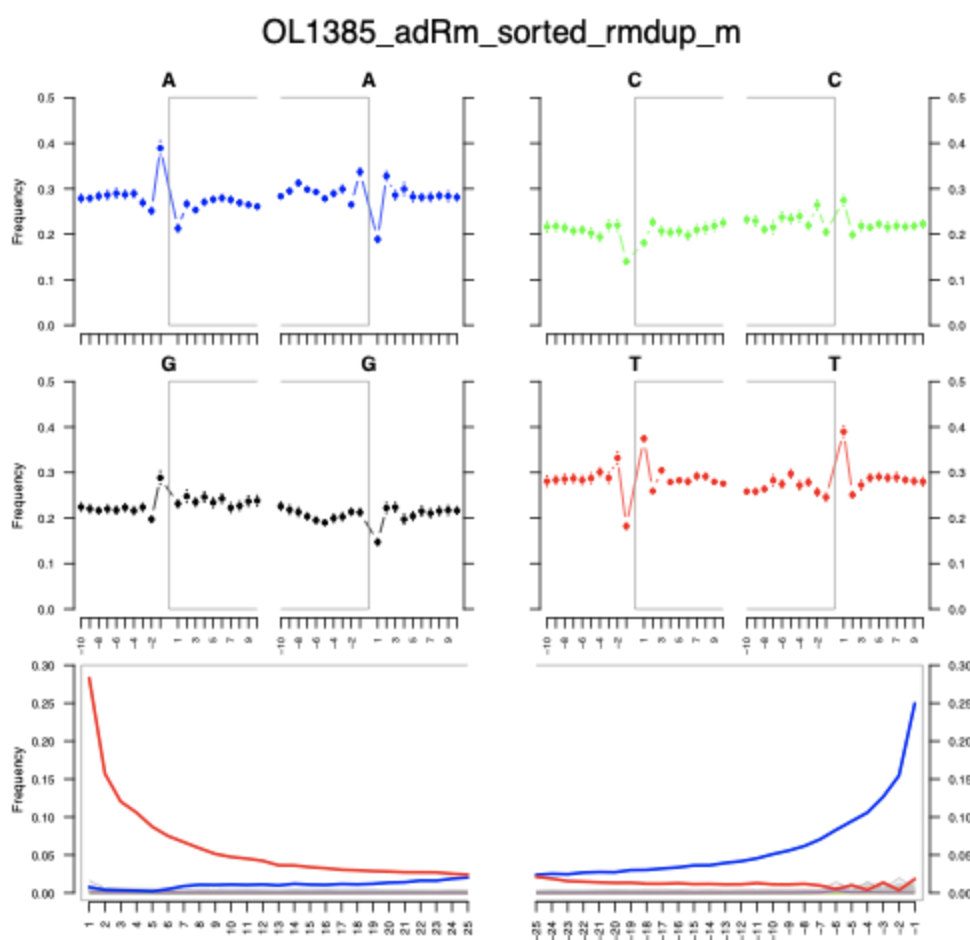
Supplementary figure 6. Deamination damage profile for sample OL1128 for *MDV*, post capture sequencing. Only 1 *MDV* DNA molecule was detected in this sample during metagenomic screening, yielding only 24 unique reads after sequencing. This number of reads is too low to give meaningful chemical damage analysis results.



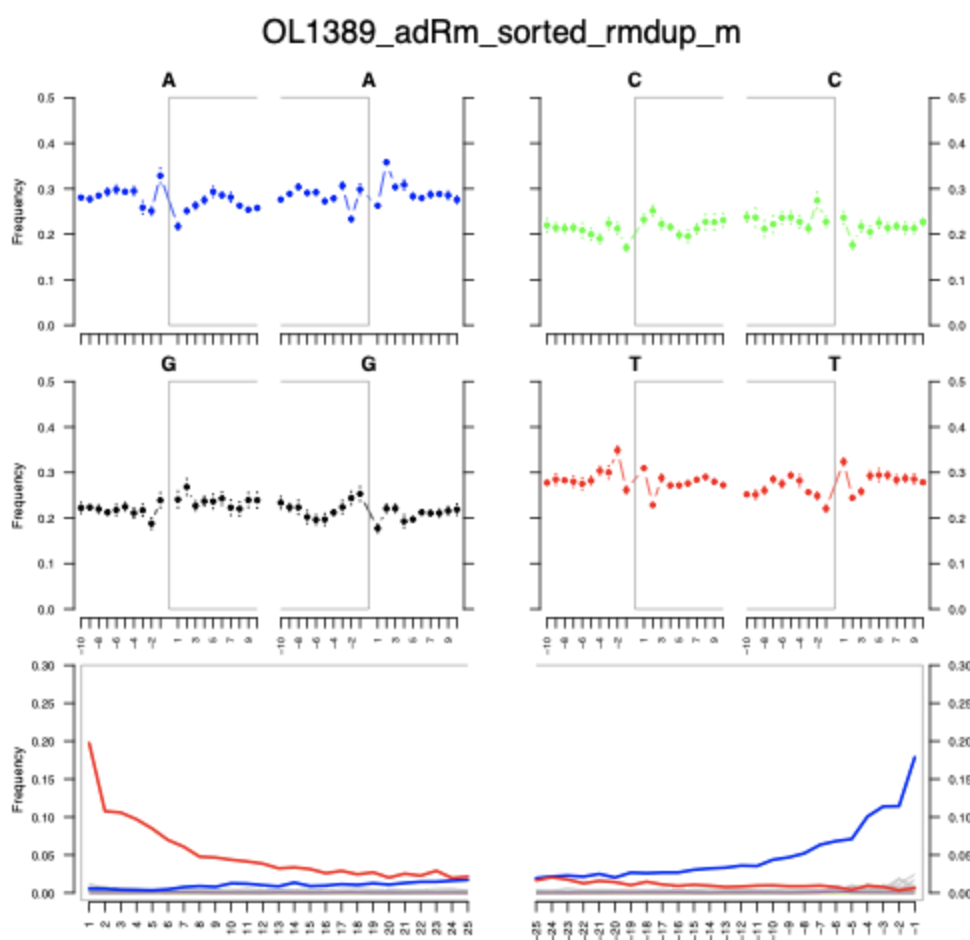
Supplementary figure 7. Deamination damage profile for sample OL1214 for *MDV* post capture sequencing. No *MDV* DNA was detected in this sample during metagenomic screening, yielding 451 unique reads after sequencing the captured library.



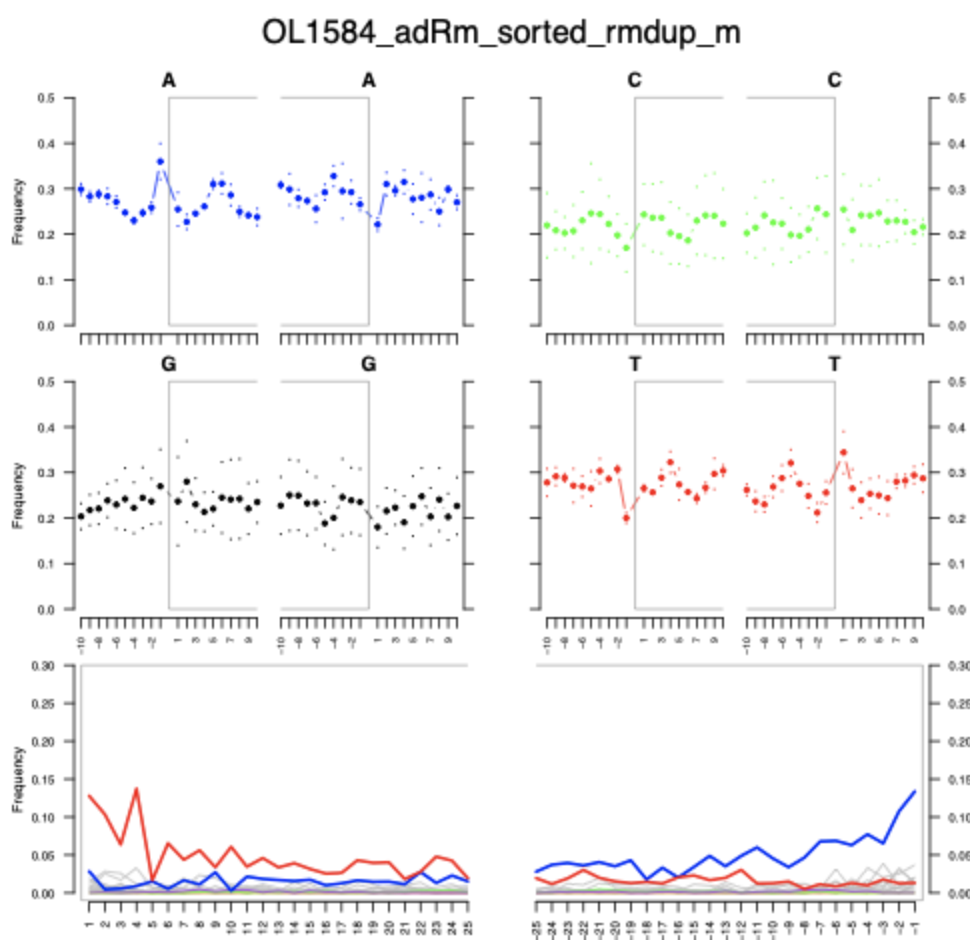
Supplementary figure 8. Deamination damage profile for sample OL1231 for *MDV*, post capture sequencing. Only 5 *MDV* DNA molecules were detected in this sample during metagenomic screening, yielding only 19 unique reads after sequencing.



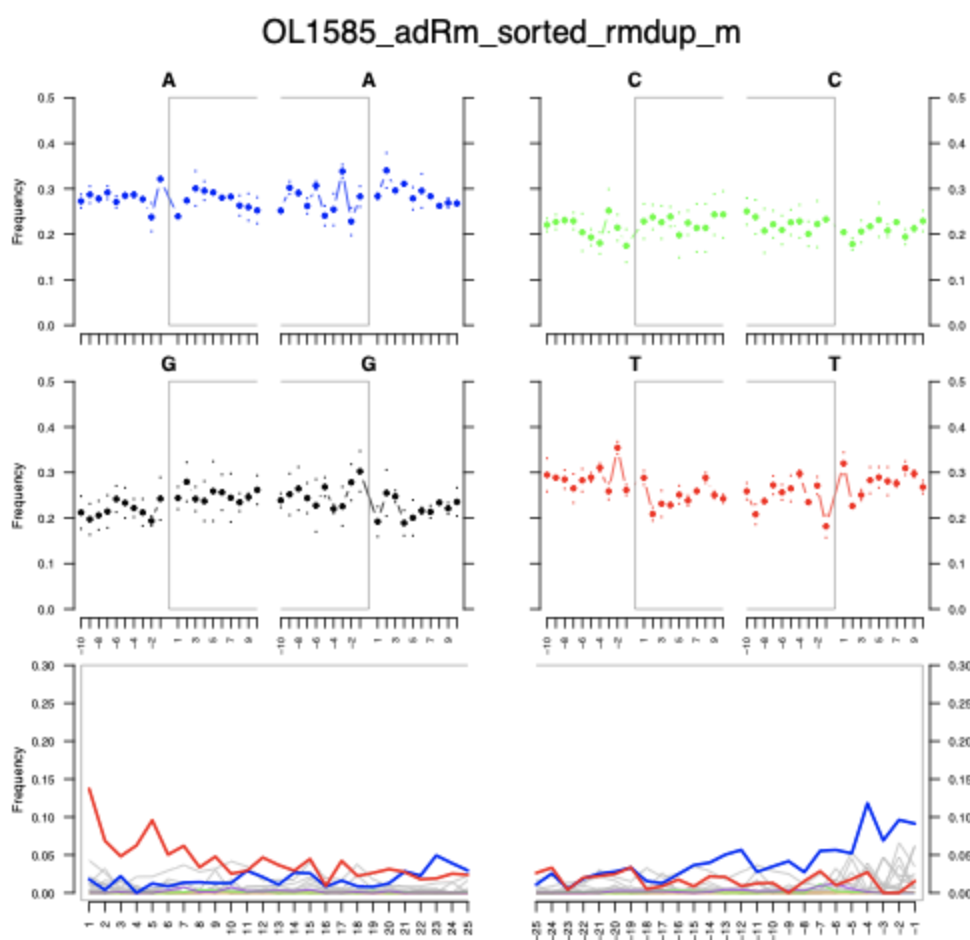
Supplementary figure 9. Deamination damage profile for sample OL1385 for *MDV*, post capture sequencing. More than 130 thousand reads were yielded after sequencing the captured library of this archaeological sample.



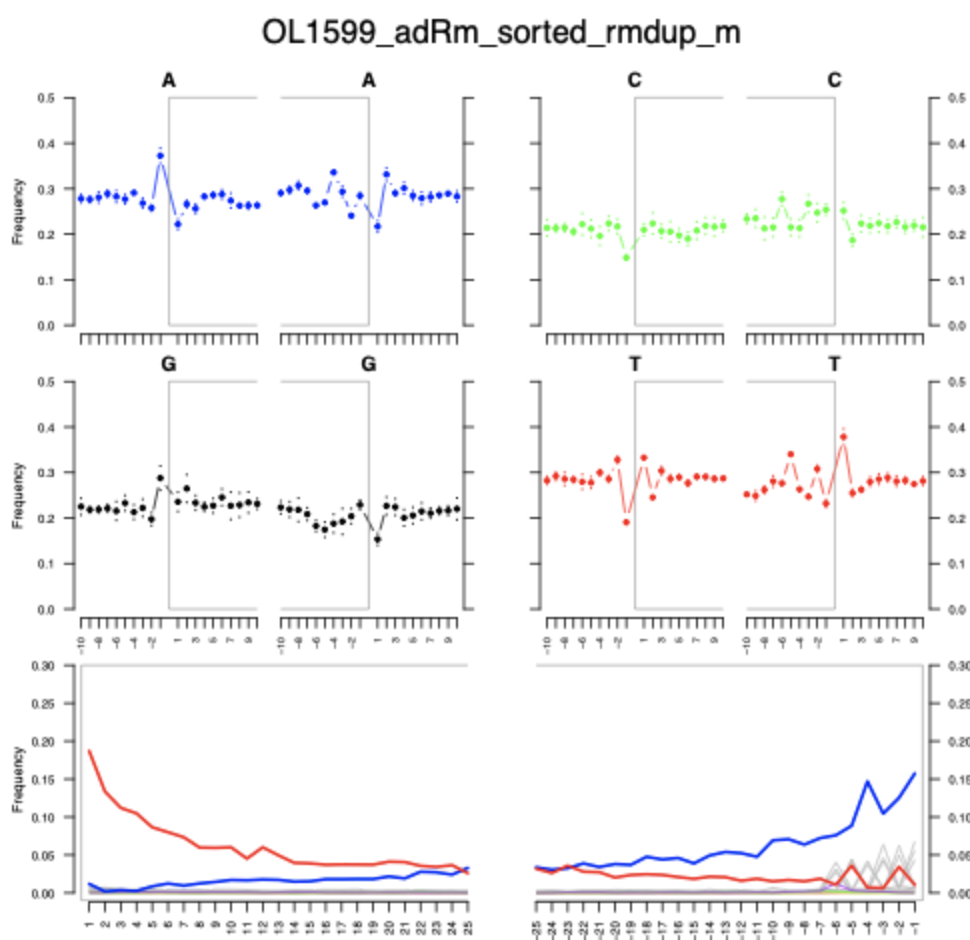
Supplementary figure 10. Deamination damage profile for sample OL1389 for *MDV*, post capture sequencing. More than 22 thousand reads were yielded after sequencing the captured library of this archaeological sample.



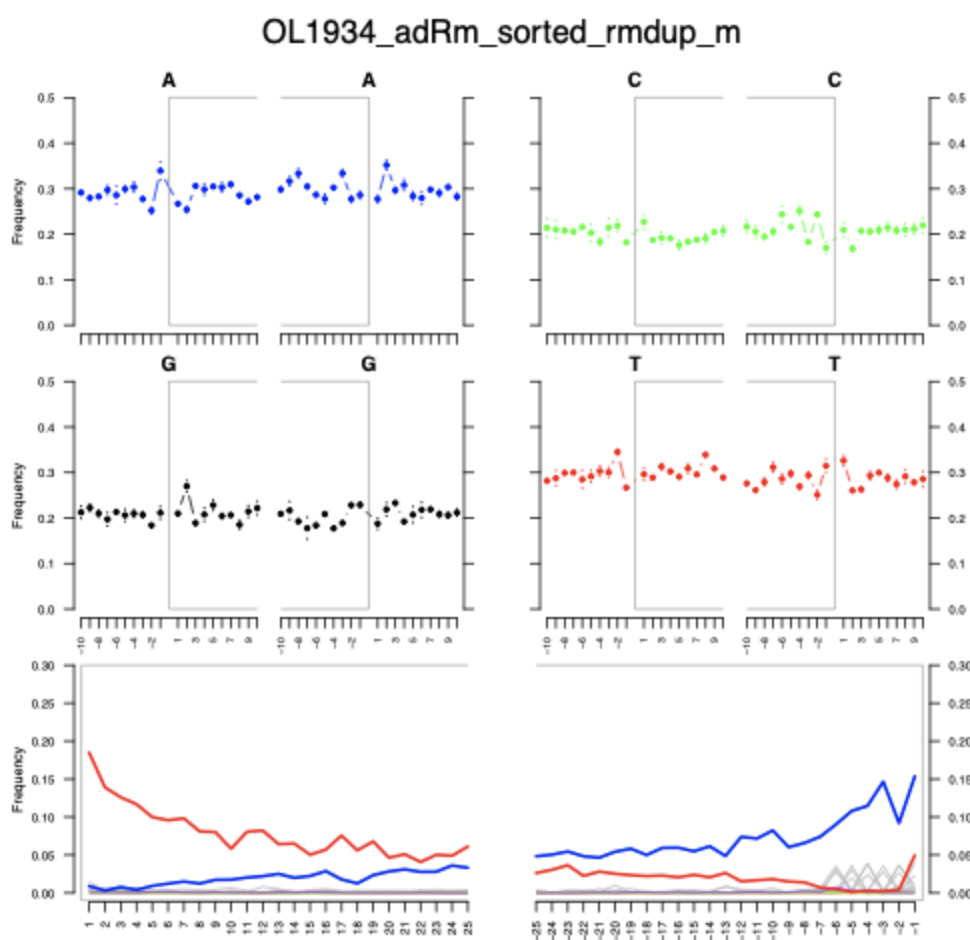
Supplementary figure 11. Deamination damage profile for sample OL1584 for *MDV*, post capture sequencing. More than 1.5 thousand unique reads were yielded after sequencing the captured library.



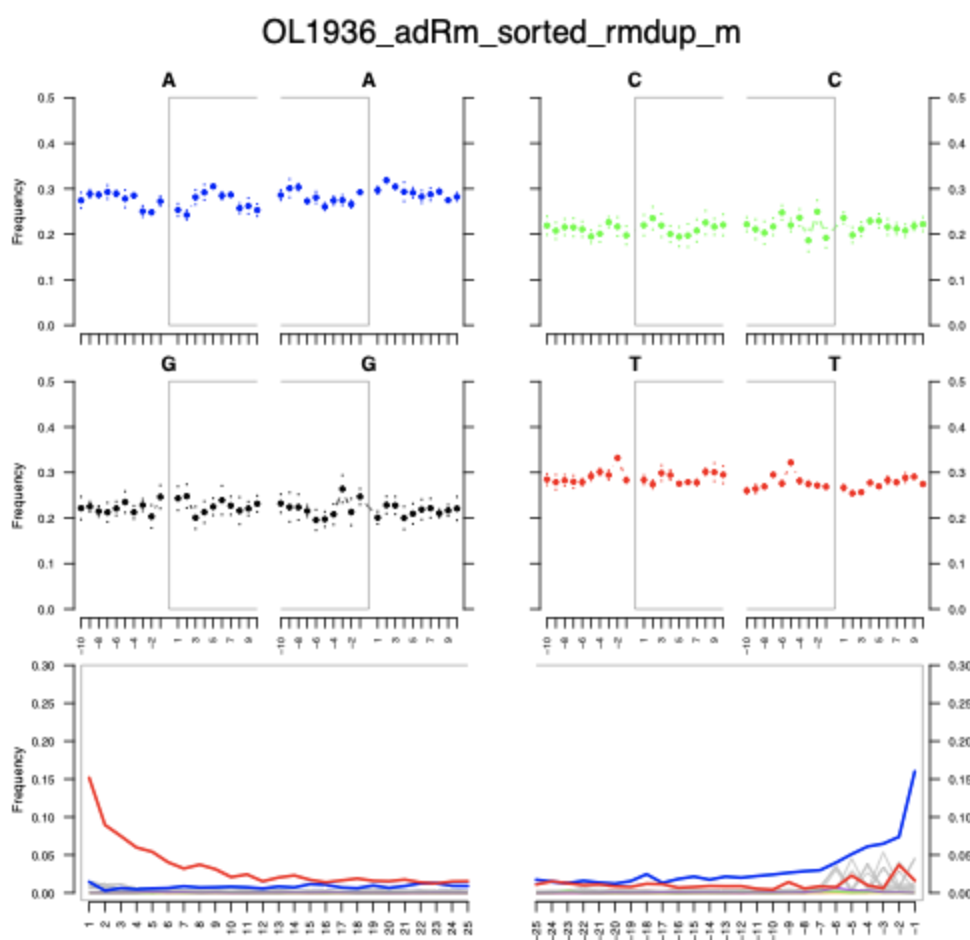
Supplementary figure 12. Deamination damage profile for sample OL1585 for *MDV*, post capture sequencing. Less than 1 thousand unique reads were yielded after sequencing the captured library.



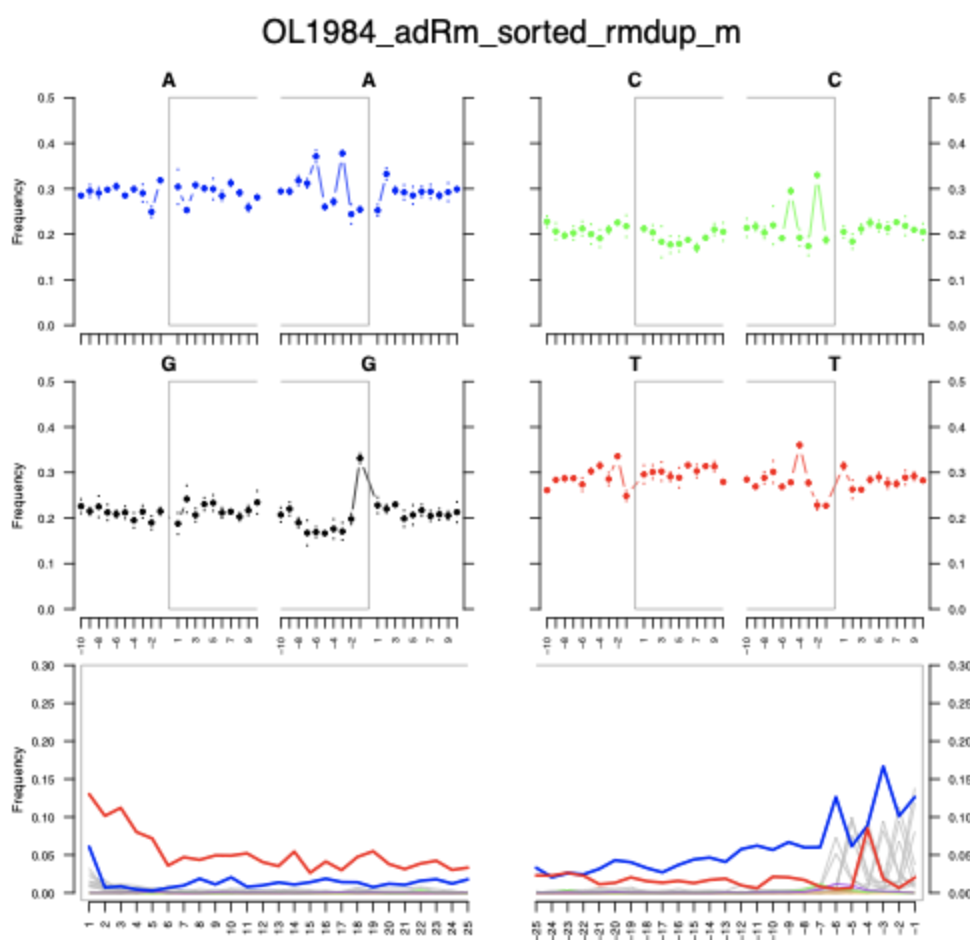
Supplementary figure 13. Deamination damage profile for sample OL1599 for *MDV*, post capture sequencing. Almost 13 thousand unique reads were yielded after sequencing the captured library of this archaeological sample.



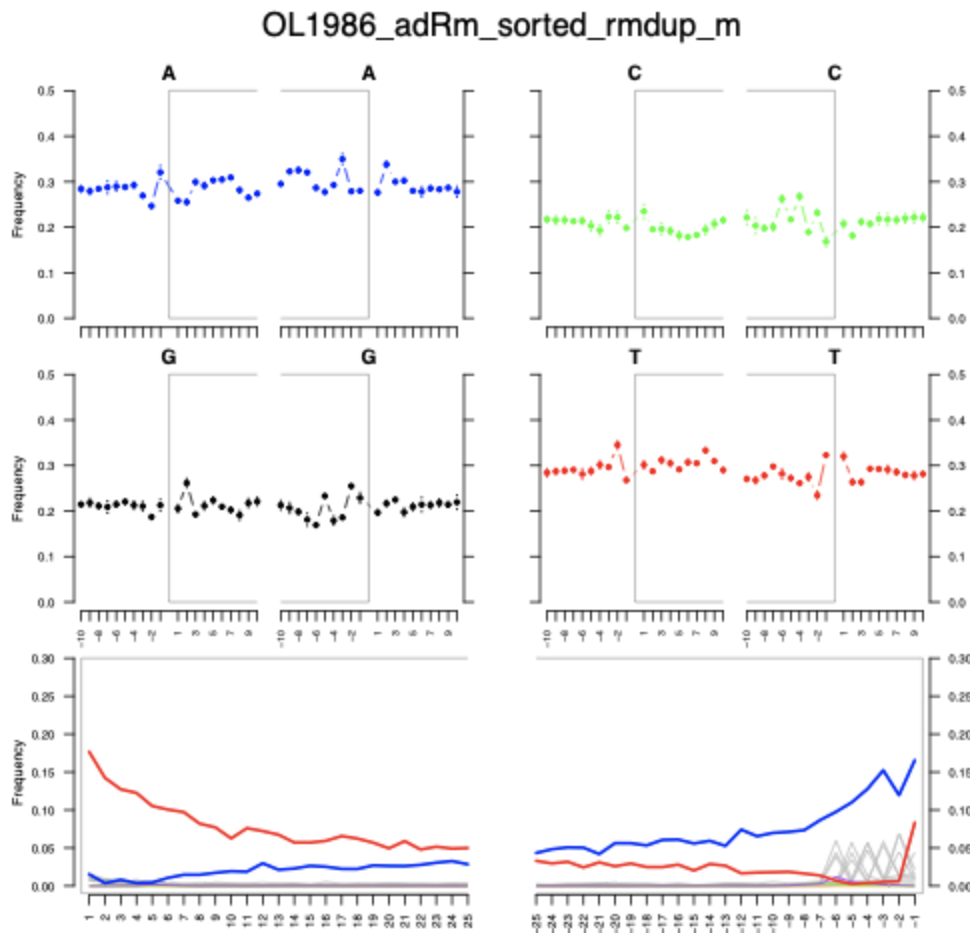
Supplementary figure 14. Deamination damage profile for sample OL1934 for *MDV*, post capture sequencing. Almost 5 thousand unique reads were yielded after sequencing the captured library of this archaeological sample.



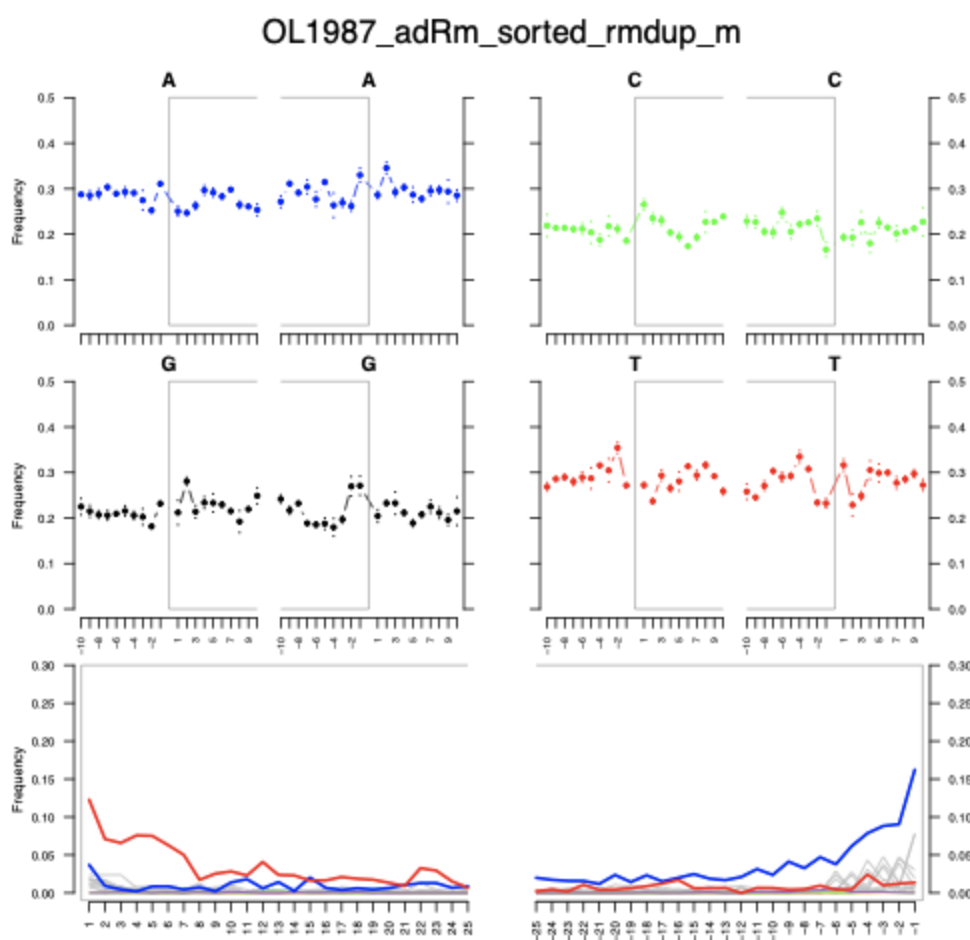
Supplementary figure 15. Deamination damage profile for sample OL1936 for *MDV*, post capture sequencing. Almost 8.5 thousand unique reads were yielded after sequencing the captured library of this archaeological sample.



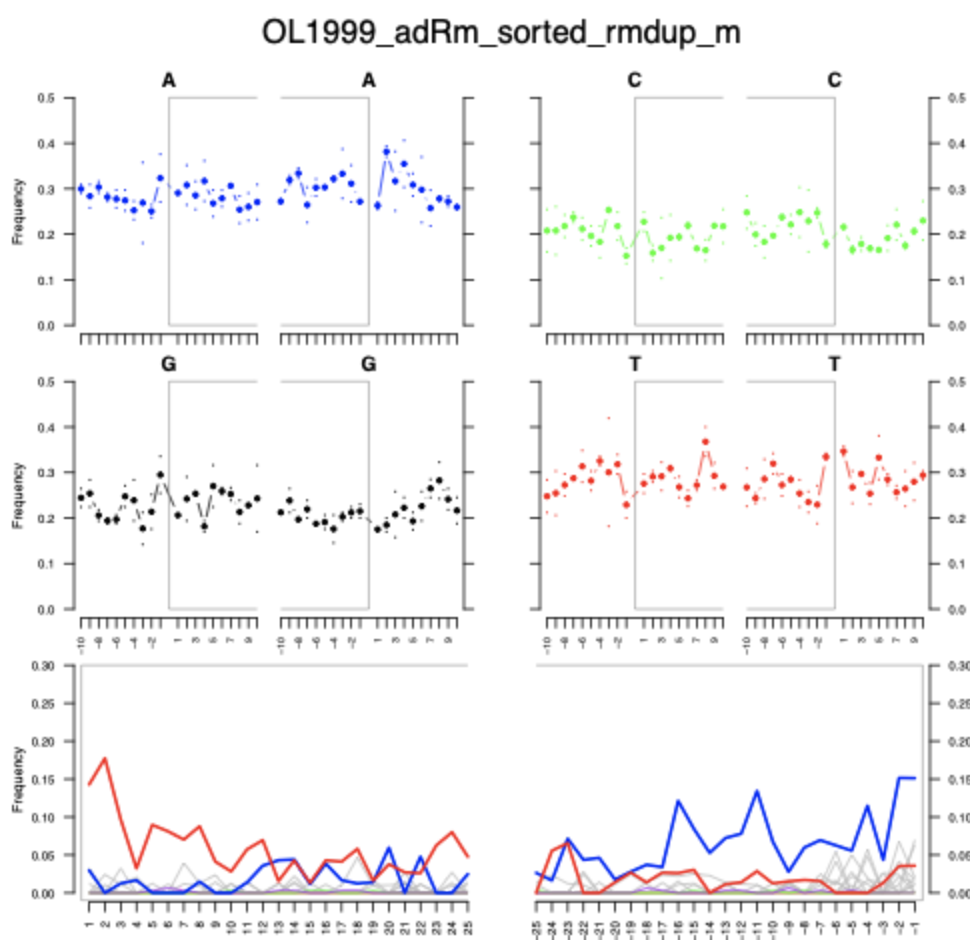
Supplementary figure 16. Deamination damage profile for sample OL1984 for *MDV*, post capture sequencing. Almost 3 thousand unique reads were yielded after sequencing the captured library of this archaeological sample.



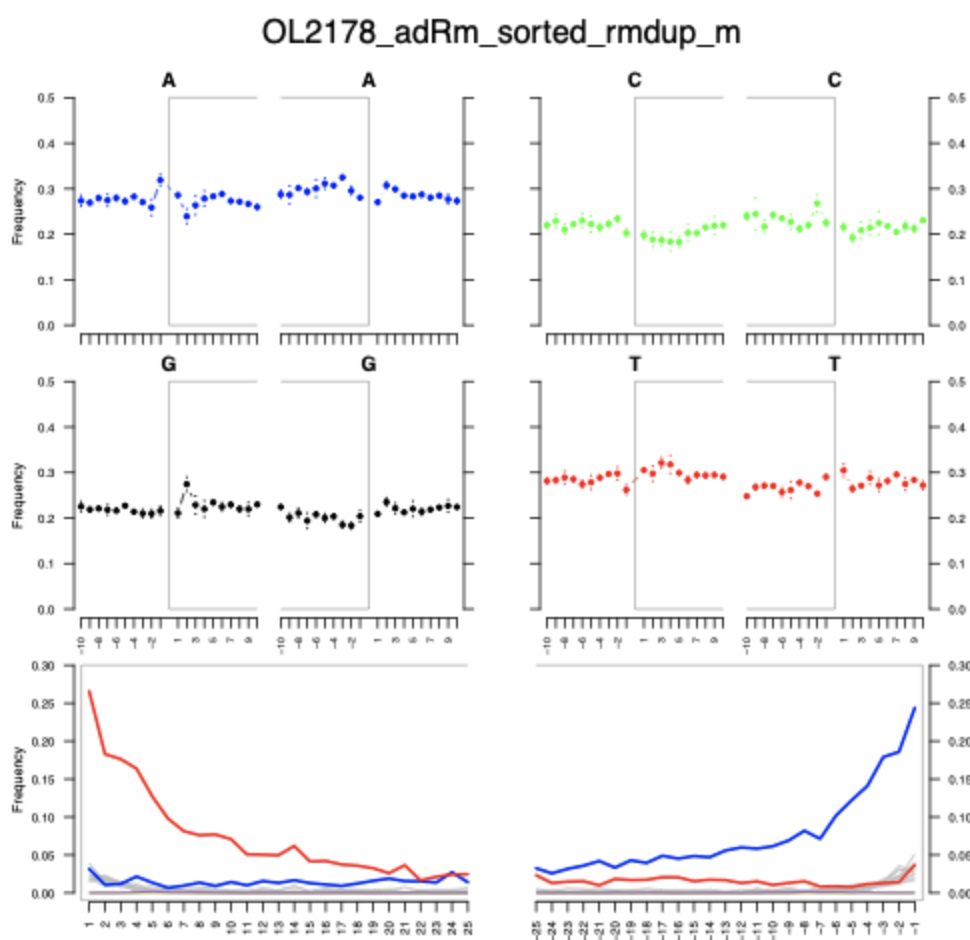
Supplementary figure 17. Deamination damage profile for sample OL1986 for *MDV*, post capture sequencing. Almost 15 thousand unique reads were yielded after sequencing the captured library of this archaeological sample,.



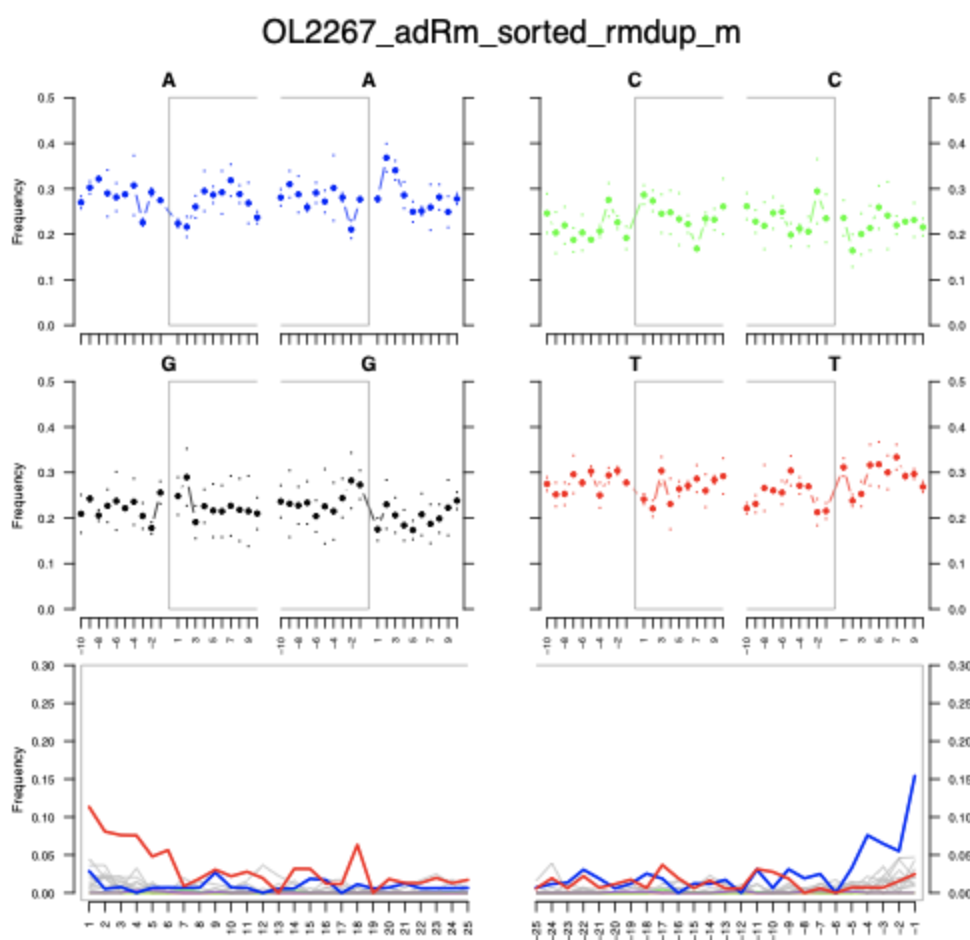
Supplementary figure 18. Deamination damage profile for sample OL1987 for *MDV*, post capture sequencing. Almost 2 thousand unique reads were yielded after sequencing the captured library of this archaeological sample.



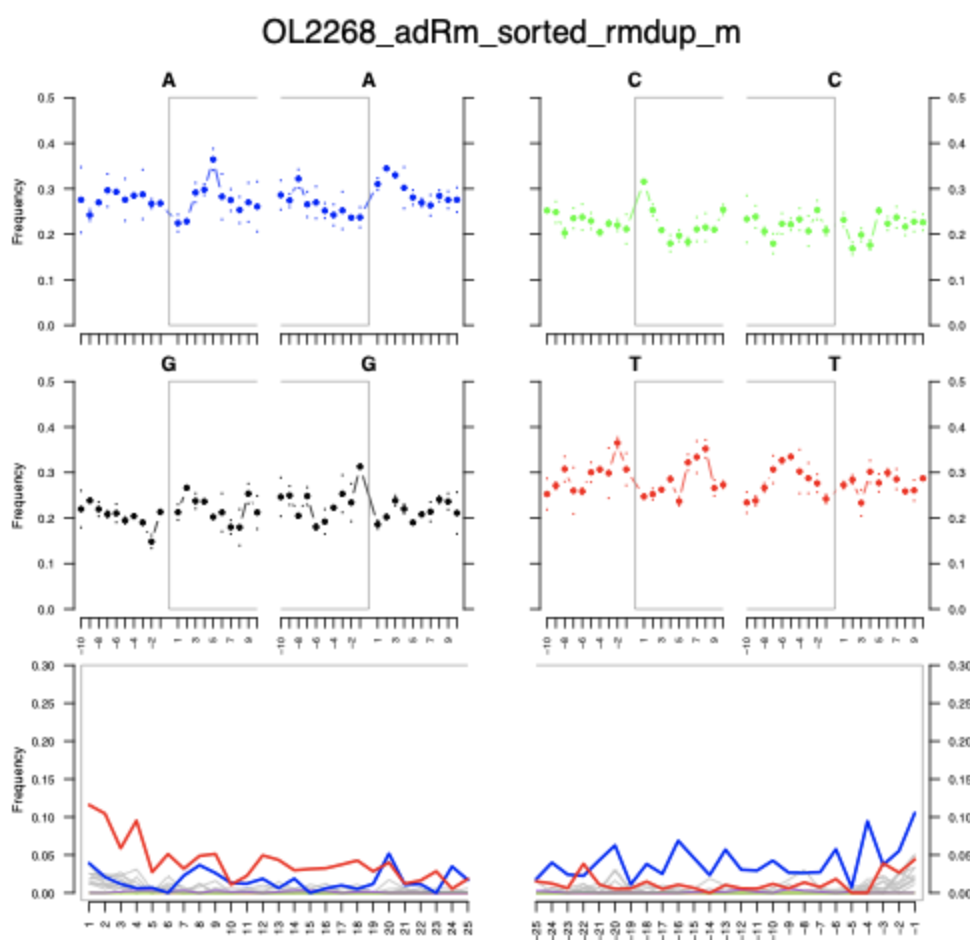
Supplementary figure 19. Deamination damage profile for sample OL1999 for *MDV*, post capture sequencing. Almost 3 hundred unique reads were yielded after sequencing the captured library of this archaeological sample.



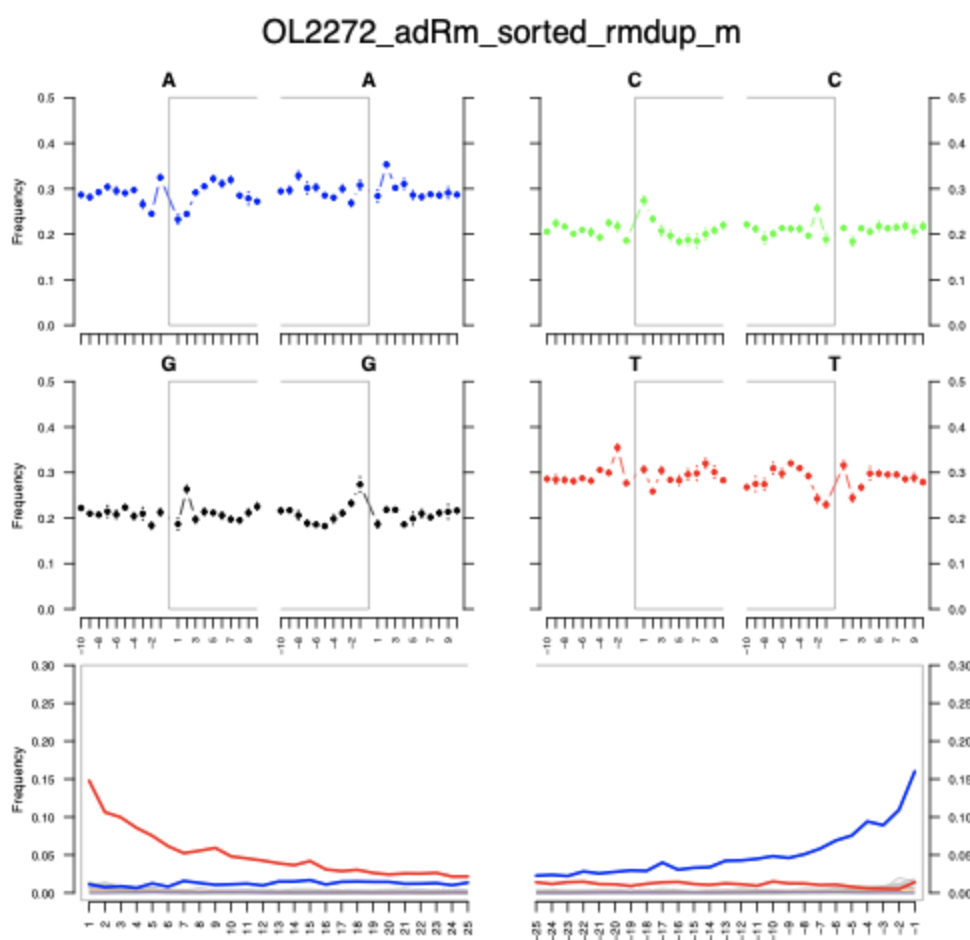
Supplementary figure 20. Deamination damage profile for sample OL2178 for *MDV*, post capture sequencing. Almost 6 thousand unique reads were yielded after sequencing the captured library of this archaeological sample.



Supplementary figure 21. Deamination damage profile for sample OL2267 for *MDV*, post capture sequencing. Almost 7 hundred unique reads were yielded after sequencing the captured library of this archaeological sample.



Supplementary figure 22. Deamination damage profile for sample OL2268 for *MDV*, post capture sequencing. Almost 7 hundred unique reads were yielded after sequencing the captured library of this archaeological sample.



Supplementary figure 23. Deamination damage profile for sample OL2272 for *MDV*, post capture sequencing. Almost 17 thousand unique reads were yielded after sequencing the captured library of this archaeological sample.

Permission from co-authors

I hereby give permission to Evangelos Antonios Dimopoulos to use our joint work as contribution towards his D. Phil thesis to be submitted for examination at Oxford University.

I confirm that to the best of my knowledge, the author contribution statement below is accurate and Evangelos Antonios Dimopoulos' contribution towards the work is greater or equal than that of any other co-author.

Chapter author contribution statement:

E.A.D., S.F., A.S., L.A.F. and G.L. designed the project; E.A.D. generated the captured data; S.F. aided in the design of the capture probes, selected the modern MDV genome accessions, and conducted bioinformatic analyses; O.L. and E.A.D. generated the screening data; E.A.D. analysed the screening and captured data; G.L., L.A.F. and A.S. provided support in interpreting results; E.A.D. wrote the paper with contributions from all other authors.

Date: Name: Signature:

31/05/2021

Ophélie Lebrasseur (OL)



Steven Fiddaman,
31/05/21



Date: May 1, 2021

Name: Greger Larson

Signature:



Laurent Frantz
31/05/2021



Date: 1st June 2021

Name: Adrian L. Smith

Signature:



Mining for pathogens in ancient pig and dog metagenomes

Statement of Authorship:

Research design: I designed the research, with input from Greger Larson and Laurent Frantz.

Data generation: I contributed to the generation of the screening sequencing datasets.

Analysis: I performed all the computational analyses presented in this paper and interpreted the results under the guidance of Laurent Frantz and Greger Larson.

Manuscript: I wrote the manuscript, with input from all other co-authors.

Acknowledgments: I would like to acknowledge James Haile and Anna Linderholm for generating the majority of the screening sequencing suid and canid datasets and all the archaeologists who provided the archaeological material for this study.

Authors:

Evangelos Antonios Dimopoulos¹, Laurent Frantz^{2,3}, Greger Larson¹

¹ The Palaeogenomics and Bio-archaeology Research Network, Research Laboratory for Archaeology and History of Art, University of Oxford, Oxford, UK.

² School of Biological and Chemical Sciences, Queen Mary University of London, London, UK.

³ Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, Munich, Germany.

Abstract

The accurate identification of microbial species from archaeological samples has the potential to provide an unprecedented perspective on the evolutionary history of major diseases that have affected humans and domestic animals. Limited DNA survival and DNA damage, however, have hindered the search for ancient pathogens. Here, we examine if the process of domestication has had an impact on the disease burden of those species, by applying a novel quantitative metagenomic method that complements established metagenomic pipelines.

By applying hypothesis-driven species identification methods to individual libraries constructed on extracts from 938 ancient pig and 2,090 ancient dog samples, we identified the presence of the causative agents for a range of bacterial diseases including, namely Kennel cough and Salmonellosis. We also interrogated the same animal metagenomic datasets and we were able to identify microbes that are capable of infecting both humans and domesticated animal hosts including the causative agents of streptococcal infections and diphtheria. Furthermore, we quantified the differences between microbial identification profiles of domestic pigs and dogs and their wild counterparts.

This method can therefore be employed to test specific hypotheses related to the role of domesticated animal species as pathogen reservoirs. Since many of the causative agents of these zoonotic diseases became prevalent in a post-industrial context, the results of this approach will be able to offer insights into how human activity has reshaped the landscapes of both the natural environment and animal diseases.

Introduction

Zoonoses and animal health in domesticated animals

Animal diseases that are transmissible to humans through contact, food products, water or the environment in general can be defined as zoonoses (Facciola et al., 2018). Domesticated animals and livestock have been experiencing the impacts of non-zoonotic as well as zoonotic (transboundary as well as endemic) diseases, affecting both their and human health (Osofsky, 2019). Foodborne diseases, a category of zoonotic diseases, have been having a significant health and financial impact on human and animal populations, with the CDC estimating that in the United States one in six people contract a foodborne illness each year imposing a burden of 15.5 billion USD (Hoffmann and Scallan, 2017). Foodborne illness occurs when a pathogen is ingested along with food products and infects a human or animal host, or when a toxigenic pathogen multiplies on food products and produces toxins (Heredia and García, 2018), that are subsequently ingested by the animal host. Thus foodborne diseases are broadly classified into foodborne infections or foodborne intoxications (Bintsis, 2017). Inevitably all these different categories of pathogens are putting on a substantial disease burden on modern animal domesticates, which in cases is quite hard to quantify. For instance, estimating the true burden of foodborne diseases is complicated further by the fact that only a few pathogen illnesses can be definitely linked to food consumption (Flint et al., 2005).

Even for well-studied domesticated animals such as dogs, we know that they are responsible for transmitting more than 300 zoonoses to humans, with rabies being

the most prominent one (Ghasemzadeh and Namazi, 2015; Smith et al., 2019), but we lack fundamental knowledge about their ecology and evolution. We still do not fully comprehend what are the exact characteristics that allow zoonotic pathogens to jump among various domesticated species, using them as vehicles for their transmission, or what drives pathogens to evolve as either generalists or specialists (Cleaveland et al., 2001; Woolhouse et al., 2001). Subsequently zoonoses have a detrimental effect on animal and human health (with 60% of human diseases having been identified as zoonotic) that results in significant financial losses, due to damage control and medical care (Hampson et al., 2015; Karesh et al., 2012).

Zoonotic pathogen transmission mechanisms are complex and not always clearly understood (Fong, 2017; Woolhouse et al., 2001). Possible transmission routes for zoonoses include the food chain, for foodborne pathogens, close contact with companion animals, or vertebrate or invertebrate animal vectors (Fong, 2017). Still little is known about the intensity and type of contact that leads to zoonotic pathogen transmission (Klous et al., 2016) from livestock to humans. At the same time not many effective countermeasures are in place to eliminate harmful bacteria from food products (Heredia and García, 2018).

Domestication and animal management led to a larger number of animals living in close quarters and the associated population growth comes with repercussions. The large population sizes of livestock have shifted the evolutionary speed of pathogens (Atkins et al., 2013), as a response to having more individuals available to infect. Multiple times over the past millennia when advances in farming techniques, such as the Neolithic farming transition approximately 9,000 YBP (i.e. the transition from

hunting–gathering to farming) (Fort, 2015; Pearce-Duvet, 2006; Stone, 2020) and the 18th Century Agricultural Revolution, have led to increased domestic animal population sizes, diseases have caused major damage to them. These include the 18th century rinderpest outbreak which killed 80% of infected cattle (Roman-Binois, 2017), and dramatic reductions in stature amongst swine and cattle in England between the Roman Era and the Black Death (Allen, 1999; Thomas, 2005).

What is a domesticated animal and how ancient DNA can elucidate (zoonotic) pathogen transmission

Domestication is a special form of biological mutualism led by the capacity of humans to affect phenotypic and behavioral animal traits (Zeder, 2012). For a definitional approach of domestication to be meaningful, it needs to focus on the evolving relationship between humans and their targeted animal populations, and not the causes or the effects of such an interaction (Zeder, 2012, 2006). Therefore, any resulting morphological and/or genetic changes in domesticated animals should be considered artifacts of this evolutionary process, rather than defining characteristics of it (Zeder, 2006).

Because domestication involves long-term continuous change the use of fixed terms in an effort to define a domesticated animal is arbitrary. Terms such as ‘wild’ and ‘domestic’ are currently being used as categorical, mutually exclusive terms, and not as antonyms that are part of a spectrum. In reality, terms such as ‘wild’ and

‘domestic’ represent previous and later points in time of an evolutionary process and not a binary classification (Larson and Burger, 2013).

Domestication is an evolutionary process with a beginning, but no defined end (Larson and Burger, 2013), thus particularly complicating the robust identification of early domestic animal specimens. For example distinguishing early domesticated dogs from their wolf ancestors is difficult, as key morphological characteristics (e.g. size and position of the teeth, size and proportion of cranial and post cranial elements) established by zooarchaeologists, had not yet been fixed in the early stages of dog domestication (Larson et al., 2012). In an effort to coarsely categorise animals in populations, a wild population can be defined as one that is non-commensal with humans. On the other hand, a domesticated animal population is a population that is commensal with humans or otherwise generally reliant on a human-modified environment for survival (Lord et al., 2020). Although the process of domestication represents a continuum, in this chapter I classify animal samples as wild, domesticated or of uncertain status, for the samples whose zooarchaeological species call was not definitive due to the lack of informational morphological traits. The rationale behind using these dichotomous and oversimplified domestication status calls was to make the objective of observing changes in pathogenic prevalence among wild and domesticated animals a more tractable problem.

Integrating ancient DNA and archaeological evidence can elucidate the dynamics among pathogens, domesticates and humans

Depending on when humans acquired a pathogenic species, a pathogen can be classified as an 'heirloom' or a 'souvenir' species. Pathogens sharing a long co-evolutionary history with human or hominids are categorised as 'heirloom' species, such as pathogens infecting humans in the Palaeolithic (Harkins and Stone, 2015; Harper et al., 2013). Pathogens acquired after the Neolithic demographic transition are characterised as 'souvenir' pathogens, and they are typically acquired after human populations expanded and got exposed to new animal reservoirs or vectors (Harkins and Stone, 2015; Harper et al., 2013). It has been hypothesised that simultaneously with the transition to farming 10000 YBP, an epidemiological transition also occurred (often known as the first epidemiological transition), driven by domesticated animals (Pearce-Duvet, 2006). Poor sanitary conditions due to sedentism and close proximity to domesticated animals resulted in an increased infectious disease burden, which combined with increased fertility rates, led to presumably higher mortality rates (Gage and DeWitte, 2009; Pearce-Duvet, 2006; Stone, 2020). Therefore, under this model livestock is considered to be a disease reservoir, posing a greater risk for transmitting pathogens to agriculturalists than hunter-gatherer populations (Stone, 2020). From modern epidemiological data it has been estimated that more than 500 different species of pathogens are known to infect livestock, out of which approximately 40% are considered to be zoonotic (Cleaveland et al., 2001). Similarly, it has also been estimated that approximately 400 species of pathogens can infect domestic dogs and cats, out of which almost 70% are zoonotic (Cleaveland et al., 2001). These modern epidemiological findings, along with the archaeological evidence supporting the first epidemiological transition, indicate that the close proximity among humans and domesticated animals provides

an opportunity for pathogens to cross the species barrier and infect humans (Woolhouse and Gaunt, 2007).

But a linear relationship of transmission from livestock to humans does not always reflect reality. For example in the case of *Salmonella enterica*, one study suggests that pigs were the intermediate host of strains that later on adapted to humans, speculating that pig strains and early human ones were closely related (Zhou et al., 2018). Another study also suggests that early strains of *S. enterica* were not host specific, had a variable pathogenic potential, and the emergence of human adapted strains was a result of changes in human lifestyle during the process of Neolithisation (e.g. shifting from foraging for food to farming) (Key et al., 2020). In another example, there is evidence that *Mycobacterium tuberculosis* has been transmitted back and forth between humans, cattle (Stone, 2020; Stone et al., 2009) and seals, with the latter being responsible for spreading tuberculosis to the New World (Bos et al., 2014). By contracting the disease from African host species (such as cattle), pinnipeds carried *M. tuberculosis* across the Atlantic, and due to them being commercial commodities in coastal areas of South America a zoonotic transfer of the bacterium back to humans became possible (Bos et al., 2014).

Examples of zoonotic epidemics affecting animals and humans

Because pathogens have been jumping among humans and animals, and have been adapting to new hosts, human and animal health statuses are interconnected. Instances where humans and animals have experienced epidemics caused by the

same pathogens include known recent outbreaks of influenza viruses, such as the avian flu. The term “avian influenza virus” usually refers to the H5N1 flu virus, which is endemic in birds and Asian poultry, first identified in Hong Kong in 1997 (Iwami et al., 2009). The virus rarely gets transmitted from animals to humans, but when it does it causes severe disease (Berns et al., 2012). Transmission among humans has been rare, thus avoiding the occurrence of a human pandemic (Berns et al., 2012; Webster and Hulse, 2005). Even though the H5N1 virus has not been isolated from humans or poultry since 2004 (Webster and Hulse, 2005), it has been evolving continuously ever since. Not only has it expanded its host range, which now includes falcons, herons, tigers and domestic cats, but it has also become more pathogenic, by being able to kill its natural host, the wild waterfowl (Webster and Hulse, 2005).

Another well known example is the H1N1 flu epidemic. The influenza virus H1N1 was first detected in humans in March 2009 and it had caused a pandemic in humans by June of the same year (Schrenzel et al., 2011; van der Meer et al., 2010). Interestingly, the virus was also isolated from a broad range of domestic and wild hosts, such as pigs, dogs, cats, turkeys, skunks and cheetahs. Its ability to infect a diverse range of hosts raised concerns about the possibility of animal reservoirs being created where H1N1 could further persist and evolve (Schrenzel et al., 2011). The strain misleadingly was named as “swine flu” due to its genetic similarity with swine influenza viruses circulating at the time, but the first outbreak on a swine farm did not happen until two months after the first human reports (van der Meer et al., 2010).

It was later concluded that transmission from humans to pigs was the most probable cause of infection (Keenlside, 2013). There are indications that the 2009 H1N1 strain was derived from the 1918 H1N1 human strain that caused the Spanish flu pandemic (Xu et al., 2010). To support that claim it has been hypothesised that in 1918 H1N1 jumped from humans to pigs, evolving into the strains that today cause swine flu (Keenlside, 2013), therefore explaining the high genetic similarity of the 2009 H1N1 virus with the swine flu influenza strains.

But human epidemics because of zoonotic pathogens did not begin in the 20th century, as some of the deadliest epidemics in human history were caused by zoonoses. *Yersinia pestis*, a species that evolved from the zoonotic pathogen *Yersinia pseudotuberculosis* and uses flea rodents as vectors for its transmission (Achtman et al., 1999; Ayyadurai et al., 2010; Eisen et al., 2006) has been infecting humans and animals for millenia. *Yersinia pestis* has so far been implicated in three documented human plague pandemics. The first pandemic was the Justinianic plague (541–544 AD) that persisted until the 8th century AD; the onset of the second pandemic was the Black Death in Europe (1347–1351 AD) and included successive waves, such as the Great Plague (1665–1666 AD), until the 18th century; the third pandemic began in China in the mid-1850s and subsequently spread to Africa, the Americas, Australia, Europe, and other parts of Asia. persisting until the middle of the 20th century (Harbeck et al., 2013; Rasmussen et al., 2015).

Calibrated molecular clock studies have highlighted that the earliest divergence date of *Y. pestis* was 5,783 years ago (95% HPD interval: 5,021–7,022 years ago), with ancestral strains of extant *Y. pestis* being present across Eurasia by the end of the

4th millennium BC (Rasmussen et al., 2015). Further studies have also demonstrated that all *Y. pestis* strains associated with human infections shared a common ancestor sometime between 668 and 729 years ago, indicating that all currently circulating human adapted isolates emerged during the 13th century from East Asia (Bos et al., 2011).

Y. pestis has not always been such a deadly pathogen. Ancient ancestral *Y. pestis* from 5,000 years ago most probably was not pathogenic enough to cause a bubonic plague, but rather was only able to cause pneumonic and septicemic plagues (Rasmussen et al., 2015). Even though plague infections were common during the Bronze Age, crucial virulence genes, which were present in the strains causing the first to third pandemics, were absent from the genomes of the Bronze Age strains (Bos et al., 2011; Rasmussen et al., 2015). This finding suggests that *Y. pestis* had been evolving towards a more pathogenic state, concurrently becoming more deadly and adapting to novel transmission routes, reaching its peak during the Black Death by killing 30–50% of the European population in only a five-year period (Bos et al., 2011).

Motivation for studying (ancient) animal pathogens

Zoonotic pathogens can evolve and adapt to various hosts and environmental conditions imposed by humans (such as antibiotics), giving them the opportunity to even obtain antimicrobial resistance (AMR; ability of a microorganism to not be affected by antimicrobial substances (Verraes et al., 2013)) that can even threaten human health. In pigs, bacterial pathogens such as *Streptococcus suis*, which can

infect and kill both pigs and humans, are rapidly developing antimicrobial resistance (Hadjirin et al., 2020). Additionally, even traditional human pathogens are emerging as zoonotic pathogens with AMR (antimicrobial resistance). For example Methicillin Resistant *Staphylococcus aureus* (MRSA) (Algammal et al., 2020) is now an emerging zoonotic pathogen able to infect multiple host species (dogs, cats, pigs, sheep, cattle, humans), that impacts both the health of domestic animals and food safety. Pigs and dogs are not only two of the first animal species to be domesticated (Larson et al., 2012, 2007), but they have also been in close proximity to human settlements since at least the Neolithic and have moved along with human populations throughout time (Frantz et al., 2019; Perri et al., 2021). Therefore, archaeological and modern samples of these species are useful proxies to study the prevalence of (zoonotic) pathogens in ancient populations, in both time and space.

For the robust temporal study of a zoonotic pathogen's evolution we need to be able to reconstruct and analyse ancestral states of its genome from different time points, a task that can be achieved with the use of ancient DNA data. By incorporating temporal information about the sequences that are used to build a phylogenetic tree, it is possible to calibrate the estimated evolutionary rates to calendar time (Shapiro et al., 2011). Such an approach allows us to estimate the points in time when major pathogen lineages diverged and possibly differentiated to adapt to new hosts (Key et al., 2020; Rasmussen et al., 2015; Spyrou et al., 2016). The value of aDNA in the study of evolution and spread of human pathogens has enabled the discovery of a 5,000-year-old strain of *Yersinia pestis* (Rasmussen et al., 2015), or strains of *Yersinia pestis* that have been hypothesised to have caused even the Justinianic plague (Spyrou et al., 2016), or the discovery of *Salmonella enterica* serotype that

was responsible for an epidemic in Mexico during the 16th century (Vågene et al., 2018). Following a combination of metagenomic analyses (Spyrou et al., 2019) to mine the exogenous DNA fraction yielded from an archaeological samples, along with using targeted genome capture techniques (Carpenter et al., 2013; Enk et al., 2014), ancient pathogenic DNA molecules can be retrieved with great efficiency, providing us with information to track the actual evolutionary course of the pathogen.

Motivated by the recent emergence of zoonotic pathogens as global health threats and advances in the field of ancient metagenomics, we mined the collection of DNA sequencing data, available in the PalaeoBarn group, for zoonotic pathogens. Our goal was to explore the possibility of studying zoonotic pathogens by using ancient DNA data, and discover the earliest instance of each pathogen being detected in the Palaeobarn's sample collection. Confidently identifying traces of a pathogen in whole genome sequence data from archaeological and modern samples provides valuable information about how the distribution of a pathogen changes temporally and spatially, and informs us about its prevalence in its host populations too.

In order to work towards these goals we screened pig and dog WGS data for specific pathogens species. Two lists of pathogens were curated for each domesticated species respectively, which included common zoonotic bacterial pathogens. These species were chosen based on their high prevalence in modern animal populations. Additionally we included bacterial species that are conventionally classified as human pathogens, which have been reported in the literature to be infecting pigs or dogs. The pig samples were screened for the following species: *Brucella suis*, *Campylobacter coli*, *Erysipelothrix rhusiopathiae*, *Salmonella enterica*, *Streptococcus*

suis, *Yersinia pseudotuberculosis*, and *Yersinia enterocolitica*. The dog samples were screened for *Borrelia burgdorferi*, species of the genus *Rickettsia*, members of the genus *Ehrlichia*, *Bordetella bronchiseptica*, *Streptococcus canis*, *Streptococcus pneumoniae*, *Brucella canis*, *Yersinia enterocolitica*, *Salmonella enterica* and *Corynebacterium diphtheriae*

Swine (zoonotic) pathogens

The pig sample collection was screened for *Brucella suis* (Olsen and Tatum, 2017) that causes swine brucellosis, *Campylobacter coli* which causes swine campylobacteriosis (Kempf et al., 2017), *Erysipelothrix rhusiopathiae* (Ogawa et al., 2011; Shimoji et al., 2019; Q. Wang et al., 2010) the causative agent of swine erysipelas in both domestic and wild pigs. We were also interested in *Salmonella enterica* (Martínez-Avilés et al., 2019) which is the pathogen causing salmonellosis in pigs, *Streptococcus suis* (Wertheim et al., 2009) that causes streptococcosis in pigs, and also we looked for *Yersinia pseudotuberculosis* or *enterocolitica* (Laukkanen et al., 2008; Nesbakken et al., 2007), species known to infect pigs. It should be noted I tried to include only genera which have species that specifically infect pigs, while trying to avoid generalist taxa.

Brucella suis is a facultative, intracellular, Gram-negative bacterium, and the main agent of swine brucellosis. It is able to infect both livestock and feral animals, such as domestic pigs, wild boars and European hares (Kreizinger et al., 2014). *Brucella suis* can persist in wildlife reservoirs, allowing for the continuous infection of domestic pig populations (Olsen, 2010). Three main biovars (1-3) can be attributed

to that species with biovars 1 and 3 found in America and Asia, and biovar 2 mainly identified in Europe (Olsen, 2010). The main symptoms of swine brucellosis are abortions, genital lesions and infertility in animals, and it usually enters the body of a host through its mucous membranes (Olsen and Tatum, 2017). It can usually be transmitted venereally or through shedding of mucus and urine, even by male and non-pregnant swine. It has been reported that some pigs are able to recover from a *Brucella suis* infection while other pigs remain permanently infected, with some of them being asymptomatic (B. Lopes et al., 2010).

Campylobacter coli is one of the main species responsible for most *Campylobacter* infections (Alfredson and Korolik, 2007), it causes symptoms of enteric disease (Manser and Dalziel, 1985), and it is also one of the most common food borne pathogens (Qin et al., 2011), especially in processed food products. Its gene repertoire includes toxin (Bang et al., 2003) and antimicrobial resistance genes. Pigs are considered the primary reservoir of *C. coli*, with the primary source of infection for newly born piglets being their mothers and faeces (Alter et al., 2005; Thakur and Gebreyes, 2005).

Erysipelothrix rhusiopathiae is a facultative, non-spore-forming, non-acid-fast, small, Gram-positive bacillus (Q. Wang et al., 2010). It is a ubiquitous species able to persist in all types of environments (Q. Wang et al., 2010). *E. rhusiopathiae* is the causative agent of swine erysipelas in pigs and Erysipeloid in humans. Erysipeloid's clinical manifestation varies in severity, with symptoms ranging from fever and joint aches to lymphadenitis, lymphadenopathy and arthritis (Q. Wang et al., 2010). *E. rhusiopathiae* is also able to cause asymptomatic infections in pigs, as the bacterium

has been found on lymphoid tissues of clinically healthy pigs (Eriksson et al., 2009). Its infection can be treated with antibiotics in both pigs (Eriksson et al., 2009) and humans (Brooke and Riley, 1999; Eamens et al., 2006). Especially in humans transmission can happen through contact with infected animals or their products or wastes and soil (Brooke and Riley, 1999). Swine erysipelas is a disease of great prevalence and economic importance, as financial losses due to the symptoms it causes in infected pigs (mainly chronic arthritis) can be detrimental (WOOD and L, 1999). As a species it has great strain variability and it occasionally manages to break through the protection offered by vaccines, resulting in high mortality rates (Eamens et al., 2006). *E. rhusiopathiae* should not be mistaken for the causative agent of human Erysipelas *Streptococcus pyogenes* (Stevens and Bryant, 2016) which has been established as a human pathogen since the late 19th century.

Salmonella enterica is the most important foodborne pathogen in the swine industry causing multiple cases of poisoning in humans, and its main way of transmission among pigs is fecal shedding (Funk et al., 2001). Studies indicate that transmission is also serovar dependent, with some serovars being transmitted from host to host faster than others (van Winsen et al., 2001). Clinical symptoms of salmonellosis in pigs include fever, diarrhea, vomiting, enterocolitis, and bacteremia (Foster et al., 2003; Rostagno et al., 2011), which are usually treated with antibiotics (Bescucci et al., 2020). Certain serovars (e.g. Typhimurium) after the onset of acute clinical symptoms, such as enterocolitis, can render their hosts into chronic carriers of *S. enterica* (Kim and Isaacson, 2017). The contamination of food products usually takes place during the slaughter process, through the laceration of a contaminated digestive tract, exposing all surrounding swine carcasses to the bacterium (Beloeil et

al., 2004). One of the aspects of its biology, which make it an interesting zoonotic pathogen, is that antimicrobial resistance can be gained through horizontal gene transfer among different serovars (Gebreyes and Thakur, 2005). Additionally *Salmonella enterica* is one of the few microbial species that has been retrieved and had its genome reconstructed from archaeological samples (Key et al., 2020; Vågane et al., 2018). *S. enterica* and its various serotypes are known to be causing outbreaks and/or epidemics in pigs and humans worldwide, spanning from Neolithic Europe to modern-day Australia and United States (Ford et al., 2018; Jackson et al., 2013; Key et al., 2020; Vågane et al., 2018; Zhou et al., 2018).

Streptococcus suis was first reported in 1954 by veterinarians (Wertheim et al., 2009) is the main causative agent of streptococcosis in pigs, causing among others meningitis, septicaemia, arthritis, endocarditis and pneumonia (Wei et al., 2009). *S. suis* is also a zoonotic pathogen, whose reported cases in the literature have been increasing significantly over the recent years (Wertheim et al., 2009), and it can infect humans via exposure to contaminated pigs or pig derived products (Schultsz et al., 2012). *S. suis* infections in pigs are prevalent worldwide, but it can also be isolated from other domesticated mammalian species as well (e.g. cats, dogs, ruminants) (Wertheim et al., 2009). In healthy pigs it can be retrieved from their nasal cavities, tonsils, upper respiratory, genital and alimentary tracts (Marois et al., 2007; Staats et al., 1997). Transmission among pigs usually happens nasally or orally, and it colonises the palatine tonsils of pigs (Srisikandan and Slater, 2006). Among all its different serotypes, a limited number of them are pathogenic, with serotype 2 being the most pathogenic for both pigs and humans alike (Lun et al., 2007). Symptoms in pigs include fever, symptoms of meningitis and encephalitis, and lameness (Smith et

al., 2001). *S. suis* has been known to sporadically cause disease in humans, but since 1998 the number of human cases reported in the literature have been increasing, with most cases originating from South East Asia and China (Tang et al., 2006; Wertheim et al., 2009).

Yersinia enterocolitica is a major foodborne pathogen, prevalent worldwide, that causes yersiniosis in pigs and humans (Gürtler et al., 2005). Symptoms caused by a *Y. enterocolitica* infection range from mild diarrhoea, to acute gastroenteritis and septicaemia (Fredriksson-Ahomaa et al., 2007; Van Damme et al., 2010), while the bacterium is also able to establish chronic infections in pig herds (Jungersen et al., 2006). Human pathogen strains of *Yersinia enterocolitica* are mainly isolated from pigs, and infections in humans have been associated with the consumption of raw pig derived products (Bhaduri et al., 2005). Transmission in pigs happened through the fecal oral route (Bhaduri et al., 2005). Sporadic outbreaks of yersiniosis due to *Y. enterocolitica*, mainly due to the consumption of infected food products, have been reported worldwide since the late 1970s (Ackers et al., 2000; Asakawa et al., 1973; Grahek-Ogden et al., 2007). *Yersinia pseudotuberculosis* on the other hand is primarily an animal pathogen, able to infect pigs with less frequency than *Yersinia enterocolitica* (Bonardi et al., 2016). In pigs, its infection clinically manifests by causing diarrhea or enterocolitis (Harper et al., 1990; Niskanen et al., 2008). It can be rarely transmitted to humans, where it causes yersiniosis, and its epidemiology is still not fully understood (Niskanen et al., 2002).

Canine (zoonotic) pathogens

The dog sample collections as screened for tick borne pathogens such as *Borrelia burgdorferi* (Little et al., 2010; Smith et al., 2012) the causative agent of Lyme's disease, species of the genus *Rickettsia* that are known to cause spotted fever (McQuiston et al., 2011), members of the genus *Ehrlichia* (Beall et al., 2012; Cabezas-Cruz et al., 2016) causative agents of ehrlichiosis. We also screened for zoonotic pathogens that can infect both humans and dogs alike such as *Bordetella bronchiseptica* (Kaczorek et al., 2016) the causative agent of kennel cough, *Streptococcus canis* and *pneumoniae* (Lamm et al., 2010; Tarabichi et al., 2018), and *Brucella canis* (Lucero et al., 2010). Additionally, I wanted to assess if they could act as reservoirs for zoonotic pathogens as *Yersinia enterocolitica* (X. Wang et al., 2010), *Salmonella enterica* (Cummings et al., 2020; Lowden et al., 2015) and the human pathogen *Corynebacterium diphtheriae*.

Borrelia burgdorferi (or *Borrelia burgdorferi*) is the causative agent of Lyme disease, which was first described in Lyme, Connecticut in 1975 (Duncan et al., 2005). Symptoms of canine borreliosis include arthritis, swelling of joints, lameness, neurological abnormalities, nephropathy, lethargy, anorexia, and fever (Skotarczak et al., 2005; Speck et al., 2007), with infections persisting even after antibiotic treatment (Straubinger et al., 1997). In the USA, it is the most frequently reported tick-borne disease and is more common in the Northeast, Midwest and West Coast states (Duncan et al., 2005; Little et al., 2010). *B. burgdorferi* can be transmitted to a variety of mammals (humans, dogs, deer, mice) by ticks of the genus *Ixodes*, with the primary source of transmission for humans being nymphs and for dogs being adult ticks (Little et al., 2010). To control Lyme disease in dogs, vaccination against *B.*

burgdorferi is available, even though its efficiency highly varies (Vogt et al., 2019). Because of its ability to infect multiple mammalian hosts, through tick transmission, it is important to be monitoring the prevalence of *B. burgdorferi* in the environment, something that so far has been done by using dogs as sentinels (proxy populations) for human disease (Bowser and Anderson, 2018). Outbreaks of Lyme disease in dog populations have been sporadically reported since the late 1980s, especially in areas where ticks of the genus *Ixodes* are endemic (Schulze et al., 1987). Furthermore, as the transmission of *B. burgdorferi* is dependent on the ecological niche of its tick vectors, it is predicted that climate change will accelerate the expansion of Lyme borreliosis in the future (Bowser and Anderson, 2018).

Rickettsia is a genus of approximately 27 species, out of which 17 are considered to be zoonotic pathogens. The pathogenic species of that genus are roughly divided into two groups, one causing typhus-like disease and the other causing spotted fever and they are usually transmitted by mites and ticks (Bermúdez and Troyo, 2018). Species of the genus *Rickettsia* are distributed worldwide (Botelho-Nevers et al., 2012), they depend on the density and distribution of their tick vectors (Nicholson et al., 2010) and host reservoirs. *Rickettsia rickettsii*, the causative agent of Rocky Mountain spotted fever, is the most lethal pathogen within that genus (Bermúdez and Troyo, 2018), with dogs infected by it presenting fever, lethargy, vomiting and anorexia (Nicholson et al., 2010). Although both humans and dogs are susceptible to tick-borne *Rickettsia* species, dogs are not the primary animal reservoir for these species (as they can infect multiple mammals) but rather are incidental hosts that can be infected by ticks (Nicholson et al., 2010). In the early 2000s outbreaks of typhus-like and spotted fever rickettsia outbreaks were reported in the United States

and Southern Europe (Campbell et al., 2009; McQuiston et al., 2011; Oteo et al., 2006), with infections occurring in both dogs and humans.

Ehrlichia canis is the primary causative agent of canine ehrlichiosis, and it is endemic in every continent except Australia (Mylonakis et al., 2019). It was first identified in Algeria in 1935 (Harrus and Waner, 2011), and it was first reported in the USA in 1963 (Leiva et al., 2005). It is a zoonotic pathogen, transmitted by the tick species *Rhipicephalus sanguineus* and *Dermacentor variabilis*, unfortunately infecting dogs permanently, even if antibiotic treatment is administered (Leiva et al., 2005). One of the main symptoms of the disease is ocular lesions, with their severity depending on the strain of the species and the host's immune response (Komnenou et al., 2007). Other symptoms of *E. canis* in dogs include fever, myalgia, depression, reduced white cells and reduced platelets that may lead to bleeding disorders (Nicholson et al., 2010). Outbreaks of *E. canis* have been regularly reported since 1989 in the United States and China, infecting both domestic dogs and humans, further underlining its importance as world spread zoonotic pathogen (Hua et al., 2000; Nicholson et al., 2010; Petersen et al., 1989; Wei et al., 2014).

Bordetella bronchiseptica is one of the causative agents of kennel cough, otherwise known as canine infectious respiratory disease (CRID), one of the most infectious diseases in dogs with worldwide prevalence (Schulz et al., 2014). Typical symptoms of the disease include paroxysmal dry cough and nasal discharge (Mochizuki et al., 2008). Because of its ability to infect the respiratory system of canines, even by establishing subclinical infections (Schulz et al., 2014), this disease is a risk factor for both animals and human staff in dog shelters or kennels (Edinboro et al., 2004;

Mochizuki et al., 2008). In an effort to control the disease, vaccination against *B. bronchiseptica* is advised as early as possible in the life of a dog, but vaccine efficacy against the pathogen still varies (Edinboro et al., 2004). Additionally vaccination against *B. bronchiseptica* can pose a threat to immunosuppressed humans, as cases of transmission from vaccinated dogs to immunocompromised patients (Gisel et al., 2010).

Streptococcus canis is a group G pyogenic opportunistic zoonotic pathogen of various animals (dogs, cats, cows, and rats) and humans (Enache et al., 2020; Lam et al., 2007). In healthy dogs and cats it is a commensal to the microflora of the reproductive tract, the tonsils and ear (Lam et al., 2007), but it can be responsible for diseases such as dermatitis, urogenital tract infection, otitis externa, respiratory tract infections, endocarditis, as well as septicemia (Enache et al., 2020). It has also been reported to be able to infect humans. Infected dogs or cats can transmit it through wound bites (Enache et al., 2020) or other types of contact (Tarabichi et al., 2018). Concurrently we were motivated to screen our dog sample DNA data for a streptococcal species that traditionally infects humans (Lynch and Zhanel, 2009), *Streptococcus pneumoniae*, after being reported in the literature to have been isolated from a dog with streptococcal meningoencephalitis (Irwin and Parry, 1999).

Brucella canis is an orally and sexually transmitted bacterium among dogs, which was first recognised in 1966 (Krueger et al., 2014). The primary clinical signs of its infection are late-pregnancy abortions in female dogs and epididymitis and prostatitis in male dogs (Holst et al., 2012). *B. canis* is also able to cause chronic infections in both dogs (Cosford, 2018) and humans (Krueger et al., 2014). There is no canine

vaccine available for its control and antimicrobial treatment is difficult (Krueger et al., 2014). Transmission to humans happens primarily through contact with infected dogs (Lucero et al., 2005), thus increasing the occupational hazard for staff coming to contact with dogs in kennels (Lucero et al., 2010). Infected dogs are able to disseminate the pathogen even months after bacteraemia has ceased, through their secretions (Lucero et al., 2010), but transmission from human to human is thought to be rare (Lucero et al., 2005). It is also thought that the prevalence of human infections is higher than the one indicated by the literature (Lucero et al., 2010). In a period of almost a decade (2007-2016), multiple outbreaks of *B. canis* have been reported in the literature, originating from the northern United States (Brower et al., 2007; Johnson et al., 2018). During the same time span, potentially the first recorded outbreak of *B. canis* in humans was reported in South America (Lucero et al., 2010).

Yersinia enterocolitica is the main cause of yersiniosis in animals and humans (Stamm et al., 2013), causing various enteric syndromes and gastrointestinal disease as well as asymptomatic infections (Fenwick et al., 1994; Stamm et al., 2013; X. Wang et al., 2010), and it is primarily considered to be a foodborne pathogen (X. Wang et al., 2010). Over the recent years its prevalence among other companion animals, such as dogs (Stamm et al., 2013), is being recognised (Bucher et al., 2008), with *Y. enterocolitica* being recovered from both rectal and tonsillar samples of dogs (Murphy et al., 2010). Even though its epidemiology is still not fully understood (Stamm et al., 2013), there are indications that pathogenic strains isolated from domestic dogs are closely related to pathogenic strains infecting humans (X. Wang et al., 2010).

Yersinia pestis, the causative agent of the plague can also be transmitted through flea bites and consumption of infected carcasses to domestic carnivores such as dogs and cats (Gage et al., 1994), turning these companion animals into reservoirs for human infections (Gould et al., 2008). Interestingly dogs can also function as mechanical plague vectors for infected fleas, and have been described as resistant to plague infections, with only mild to moderate symptoms being manifested (Nichols et al., 2014).

Salmonella enterica, a foodborne pathogen, is currently emerging as an important dog pathogen as well, by causing vomiting, diarrhea, and septicemia in infected dogs (Giuliano et al., 2015), or even establish asymptomatic infections (Imanishi et al., 2014). In recent years more dog owners and breeders have been feeding their dogs raw diets (Strohmeyer et al., 2006), that mostly consist of raw meat of low quality not suited for human consumption (Morley et al., 2006). Such raw meat can expose dogs to pathogenic species such as *Salmonella enterica*, who in turn can pass it on to humans, but it poses also a risk to the dog owners handling the meat itself (Strohmeyer et al., 2006). A great example of a chain of transmission of *Salmonella enterica* took place in the United States and Canada, where dog food contaminated with *Salmonella enterica*, produced in a plant in South Carolina, was consumed by dogs, which subsequently infected their owners resulting in an outbreak of human infections (Imanishi et al., 2014).

Corynebacterium diphtheriae was first described in 1884 by Loeffler as the causative agent of diphtheria, and later on in 1893 the species was first reported to cause infective endocarditis (Muttaiyah et al., 2011). Universal vaccination against that

pathogen has reduced and in many countries eliminated fatalities caused by the disease, but since the early 2000s there has been speculation of *C. diphtheriae* reemerging as a zoonotic pathogen (Bonmarin et al., 2009). Diphtheria and diphtheria-like disease can be caused by *C. diphtheriae*, *C. ulcerans* and *C. pseudotuberculosis* that include the diphtheria toxin coding gene in their repertoire (Sing et al., 2016). Even though it is known that *C. ulcerans* and *C. pseudotuberculosis* are zoonotic pathogens (Sing et al., 2016) able to infect various livestock species and domestic animals, such as dogs, instances of *C. diphtheriae* infecting animals have been rarely reported. What previously was thought to be an almost exclusive human pathogen has now been isolated from the ear tracts of domestic cats and even from free roaming red foxes (Hall et al., 2010; Sing et al., 2016), raising the question of whether it could be potentially retrieved from other animal species that are in close contact with humans.

Questions

Since the second half of the 20th century, many of our zoonotic pathogens of interest have been increasingly causing disease outbreaks both in animals and human populations. Whether this increase is a result of more systematic infectious disease diagnosis and reporting, or whether it reflects a genuine change in the biology and pathogenicity of these zoonotic agents is something yet unclear.

In order to find an answer to this problem, we decided to analyse pig and dog samples across a wide temporal and spatial range, and try to robustly identify any of our zoonotic taxa of interest. A presence/absence study of pathogenic agents, that

incorporates different time points, can help us create time series of infection data, allowing us to discover if the frequency of zoonotic infections in domesticated animals has been changing throughout time. If combined with historic and archaeological data relevant to human activity (such as human migrations, cultural changes, commerce, and technological revolutions), we will be able to better understand how human activity has shaped the landscape of disease in domestic pigs and dogs.

The main questions of interest for this paper are: are there pathogens that can be found in modern domesticated animals that can also be identified in their wild counterparts? Also are there any pathogens that can only be found in modern and ancient domestic populations? Is the frequency of pathogens identified higher in the domestic animal compared to the wild populations ? Are there any conclusions we can draw about whether domestication increased the disease burden of the animals?

Methods

DNA extraction

DNA extractions were performed in dedicated ancient DNA (aDNA) facilities at the PalaeoBARN (University of Oxford). DNA was extracted from pig or dog bone or teeth samples (Supplementary table 3) in a dedicated ancient DNA laboratory using the appropriate sterile techniques and equipment. Extraction was carried out using the Dabney extraction protocol (Dabney et al., 2013).

Library Preparation & Sequencing

Illumina libraries were built following either the Gansauge and Meyer (2014) or the Carøe et al (2018) protocol, but with the addition of a six base-pair barcode added to the IS1_adapter.P5 and IS3_adapter.P5+P7 adapter pair. The libraries were then amplified on an Applied Biosystems StepOnePlus Real-Time PCR system to check that library building was successful, and to determine the minimum number of cycles to use during the indexing amplification PCR reaction. A six base-pair barcode was used during the indexing amplification reaction resulting in each library being double-barcoded with an “internal adapter” directing adjacent to the ancient DNA strand and which would be the first bases sequenced, and a traditional external barcode that would be sequenced during Illumina barcode sequencing.

Up to 200 libraries with unique barcode combinations were pooled at equimolar levels (as determined by an Agilent Technologies 2200 TapeStation) and an 80 - 150 bp run was carried out on an Illumina HiSeq 2500/4000/X sequencer.

Sample pre-processing

The previously described wet laboratory and sequencing approach was designed to produce screening data for the estimation of the endogenous DNA fraction that could be yielded from an archaeological sample, so as to determine how best to proceed to generate genome-wide data from these ancient domesticated species. Instead, I mined the non-endogenous DNA fraction from these shotgun sequenced libraries for ancient pathogenic agents. I screened DNA data from 1,121 unique pig libraries

generated from 938 archaeological and modern pig samples, and 2,106 unique dog sequencing libraries generated from 2,090 samples of archaeological and modern dogs.

Adapters and ambiguous bases (Ns) from the ends of DNA reads were removed from the sequenced libraries with AdapterRemoval v2 (Schubert et al., 2016) and the resulting fastq files were compressed (AdapterRemoval --trimns --gzip). For paired-end sequencing data, DNA reads were additionally collapsed (AdapterRemoval --collapse --trimns --gzip), and singleton reads were subsequently removed.

Pathogen detection

Pathogens of Interest

Our pathogens of interest for pigs belonged to the following genera: *Brucella*, *Campylobacter*, *Clostridium*, *Erysipelothrix*, *Leptospira*, *Listeria*, *Salmonella*, *Streptococcus*, and *Yersinia*. The pathogens of focus for the dog screening were: *Brucella*, *Bordetella*, *Borrelia*, *Corynebacterium*, *Ehrlichia*, *Leptospira*, *Rickettsia*, *Salmonella*, *Streptococcus*, and *Yersinia*.

For our analysis we used the HAYSTAC pipeline. We used its different modules for database building, data alignment, and microbial abundance estimation for our screening process.

Database building

A separate database was built for each one of our genera of interest and each genus database contained the longest complete genome for each species. Default parameters were used.

Metagenomic identifications

For the pathogen abundance estimation, each pig or dog DNA library was aligned against the individual database of each genus of interest. Our desired taxonomic resolution is the species level. For a positive identification we required a minimum of 50 reads assigned to a certain species and a χ^2 test p-value < 0.05 , indicating that the assigned reads represent a random genome sample of the identified species.

aDNA authentication

In order to assess if pathogens identified in archaeological samples were ancient, we performed a deamination damage pattern analysis with HAYSTAC, which employs mapDamage v2 (Jónsson et al., 2013).

Results

Pigs

Detected Pathogens

Pathogens were robustly identified in 22 pig samples (Table 2), retrieved mainly from Europe, Middle East and SouthEast Asia (Figure 1). The causative agent of swine brucellosis, *Brucella suis*, was not detected in any of our samples. *Campylobacter coli* was not identified in any of the pig samples analysed. The causative agent of swine erysipelas, *Erysipelothrix rhusiopathiae*, was successfully identified in 10 samples (51 - 9037). One of the causative agents of salmonellosis in pigs, *Salmonella bongori*, was found in 3 samples (105 - 205 uniquely assigned reads), and *Salmonella enterica* in 4 samples (66 - 112 uniquely assigned reads). The causative agent of streptococcosis in pigs, *Streptococcus suis*, was identified in 1 sample (87 reads assigned uniquely). *Yersinia enterocolitica* was detected in 10 samples (58 - 1241 uniquely assigned reads), and the pig production environment specific pathogen *Yersinia hibernica* that was also detected in 2 samples (64 - 89 reads assigned).

Table 1. Table with the number of samples in which pathogens from the genera *Bordetella*, *Borrelia*, *Brucella*, *Campylobacter*, *Corynebacterium*, *Erysipelothrix*, *Salmonella*, *Streptococcus* and *Yersinia* were robustly detected in pig and dog samples.

Pathogen species	Animal species	Number of samples	Number of reads
<i>Brucella suis</i>	Pig - <i>Sus scrofa</i>	0	0
<i>Campylobacter coli</i>	Pig - <i>Sus scrofa</i>	0	0
<i>Erysipelothrix rhusiopathiae</i>	Pig - <i>Sus scrofa</i>	10	51 - 9037
<i>Salmonella bongori</i>	Pig - <i>Sus scrofa</i>	3	105 - 205
<i>Salmonella enterica</i>	Pig - <i>Sus scrofa</i>	4	66 - 112
<i>Streptococcus suis</i>	Pig - <i>Sus scrofa</i>	1	87
<i>Yersinia enterocolitica</i>	Pig - <i>Sus scrofa</i>	10	58 - 1241
<i>Yersinia hibernica</i>	Pig - <i>Sus scrofa</i>	2	64 - 89
<i>Borrelia burgdorferi</i>	Dog - <i>Canis lupus</i>	0	0
<i>Bordetella bronchiseptica</i>	Dog - <i>Canis lupus</i>	0	0
<i>Bordetella parapertussis</i>	Dog - <i>Canis lupus</i>	7	50 - 194
<i>Brucella canis</i>	Dog - <i>Canis lupus</i>	0	0
<i>Corynebacterium diphtheriae</i>	Dog - <i>Canis lupus</i>	5	93 - 462
<i>Ehrlichia canis</i>	Dog - <i>Canis lupus</i>	0	0
<i>Rickettsia rickettsii</i>	Dog - <i>Canis lupus</i>	0	0
<i>Salmonella bongori</i>	Dog - <i>Canis lupus</i>	5	55 - 110

<i>Salmonella enterica</i>	Dog - <i>Canis lupus</i>	3	68 - 93
<i>Streptococcus canis</i>	Dog - <i>Canis lupus</i>	3	71 - 93
<i>Streptococcus pneumoniae</i>	Dog - <i>Canis lupus</i>	1	51
<i>Yersinia enterocolitica</i>	Dog - <i>Canis lupus</i>	20	50 - 6046
<i>Yersinia pestis</i>	Dog - <i>Canis lupus</i>	1	68

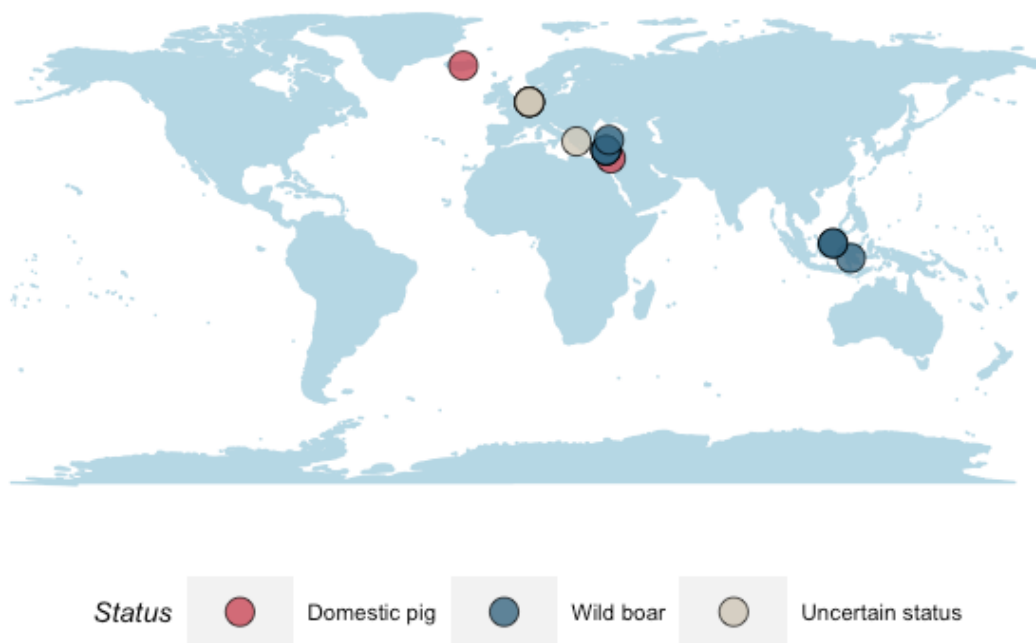


Figure 1. Map of pig samples, in which pathogenic species were positively identified in. The different colours represent the different domestication status calls.

Pathogenic species were identified in 5 domestic pig samples, out of which 3 originated from Europe, 1 from SouthEast Asia and 1 from the Middle East. Metadata regarding the cultural context around these samples were available only for 1 European ancient pig, dating back to the Viking Age, and the 1 sample from the

Middle East, dating back to the Iron Age. There were also 12 wild boar samples positive for various pathogens, 1 from Europe, 1 from the Near East and 7 from SouthEast Asia. Cultural and dating metadata were only present for 5 of these samples, with 3 of them being modern Southeast Asian samples. From the remaining two ancient wild boars, 1 dated back to the pre-pottery Neolithic era in Europe and the 1 dated back to the aceramic Neolithic era in the Near East. Regarding the 5 samples of uncertain domestication status, metadata on their age and cultural context were available for only 3 samples, 2 from Europe and 1 from SouthEast Asia. The two suid European samples were modern and the ancient Southeast Asian pig sample possibly dated back to the Neolithic.

Table 2. Taxonomic and sample metadata for pig samples in which pathogens were robustly detected.

Sample	Taxon	Assigned Read Number	Age (Mean years BP)	Country	Location or Site	Latitude	Longitude	Region	Status	Period	Reason for status call
AA276	<i>Streptococcus suis</i>	87	3240	Israel	Tel Miqne-Ekron	31.78	34.85	Middle East	Domestic pig	Iron Age	Zooarchaeological identification
AA401	<i>Erysipelothrix rhusiopathiae</i>	335	1065	Iceland	Skagafjordur	65.00	-18	Europe	Domestic pig	Viking Age - Medieval	Zooarchaeological identification
AA453	<i>Salmonella bongori</i>	153	0	Netherlands	Wageningen	51.98	5.7	Europe	Uncertain status	Modern	Zooarchaeological identification inconclusive
AA453	<i>Salmonella enterica</i>	112	0	Netherlands	Wageningen	51.98	5.7	Europe	Uncertain status	Modern	Zooarchaeological identification inconclusive
AA453	<i>Yersinia enterocolitica</i>	485	0	Netherlands	Wageningen	51.98	5.7	Europe	Uncertain status	Modern	Zooarchaeological identification inconclusive

AA453	<i>Yersinia hibernica</i>	89	0	Netherlands	Wageningen	51.98	5.7	Europe	Uncertain status	Modern	Zooarchaeological identification inconclusive
AA454	<i>Erysipelothrix rhusiopathiae</i>	1246	0	Netherlands	Wageningen	51.98	5.7	Europe	Uncertain status	Modern	Zooarchaeological identification inconclusive
	<i>Salmonella bongori</i>	205	10800	Cyprus	Klimonas	34.70	33.13	Europe	Wild boar	Pre-Pottery Neolithic	Zooarchaeological identification
AL1031	<i>Salmonella enterica</i>	110	10800	Cyprus	Klimonas	34.70	33.13	Europe	Wild boar	Pre-Pottery Neolithic	Zooarchaeological identification
AL1031	<i>Yersinia enterocolitica</i>	482	10800	Cyprus	Klimonas	34.71	33.13	Europe	Wild boar	Pre-Pottery Neolithic	Zooarchaeological identification
AL1031	<i>Yersinia hibernica</i>	64	10800	Cyprus	Klimonas	34.70	33.13	Europe	Wild boar	Pre-Pottery Neolithic	Zooarchaeological identification
AL476	<i>Yersinia enterocolitica</i>	102	NA	Indonesia	Timor	-3.90	120.86	SE Asia	Wild boar	NA	Sample provider

AL498	<i>Erysipelothrix rhusiopathiae</i>	131	NA	Philippines	Busuanga Is, Palawan Prov	13.61	123.27	SE Asia	Wild boar	NA	Lack of relevant information
AL505	<i>Erysipelothrix rhusiopathiae</i>	286	NA	Malaysia	Sapagaya Forest Reserve	4.26	102.93	SE Asia	Wild boar	NA	Museum information
AL508	<i>Erysipelothrix rhusiopathiae</i>	78	NA	Indonesia	North Sulawesi, Limbe Is	-3.90	120.86	SE Asia	Wild boar	NA	Museum information
AL524	<i>Yersinia enterocolitica</i>	535	20	Indonesia	Bali	-3.90	120.86	SE Asia	Wild boar	Modern	Sample provider
AL542	<i>Yersinia enterocolitica</i>	318	NA	Malaysia	Cave near Pahang River	4.26	102.93	SE Asia	Wild boar	NA	Museum information
AL544	<i>Yersinia enterocolitica</i>	82	NA	Burma	NA	20.89	96.44	SE Asia	Wild boar	NA	Lack of relevant information
AL553	<i>Yersinia enterocolitica</i>	211	NA	Indonesia	Papua Province, Jayapura	-3.90	120.86	SE Asia	Wild boar	NA	Lack of relevant information

AL557	<i>Yersinia enterocolitica</i>	187	0	Borneo	NA	1.40	114.50	SE Asia	Wild boar	Modern	Sample provider
	<i>Salmonella enterica</i>	71	0	Borneo	NA	1.40	114.50	SE Asia	Wild boar	Modern	Sample provider
AL559	<i>Erysipelothrix rhusiopathiae</i>	51	NA	Java	NA	-7.27	110.70	SE Asia	Domestic pig	NA	Sample provider
AL573	<i>Erysipelothrix rhusiopathiae</i>	9037	NA	Corsica	NA	42.19	9.12	Europe	Domestic pig	NA	Sample provider
AL579	<i>Erysipelothrix rhusiopathiae</i>	4393	NA	Hungary	NA	47.23	19.36	Europe	Domestic pig	NA	Sample provider
AL659	<i>Salmonella bongori</i>	105	NA	Vietnam	An Son	14.13	108.10	SE Asia	Uncertain status	Possibly Neolithic	Lack of relevant information
AL659	<i>Salmonella enterica</i>	66	NA	Vietnam	An Son	14.13	108.10	SE Asia	Uncertain status	Possibly Neolithic	Lack of relevant information
AL659	<i>Yersinia enterocolitica</i>	58	NA	Vietnam	An Son	14.13	108.10	SE Asia	Uncertain status	Possibly Neolithic	Lack of relevant information

	<i>Erysipelothrix</i>								Uncertain status	NA	Lack of relevant information
AL701	<i>rhusiopathiae</i>	79	NA	Taiwan	NA	23.73	120.83	SE Asia			
	<i>Erysipelothrix</i>									Aceramic Neolithic	Zooarchaeological identification
AL725	<i>rhusiopathiae</i>	68	10200	Turkey	Aşıklı Höyük	38.40	34.23	Near East	Wild boar		
	<i>Yersinia</i>								Uncertain status		Zooarchaeological identification
OL4351	<i>enterocolitica</i>	1241	2500	Greece	Lehova	37.87	22.58	Europe		NA	inconclusive

Pathogen distribution and species found

I queried the distribution of pathogens among the pig samples for which we had reliable metadata on their domestication status and estimated age (Figure 2). There was one pathogenic bacterial species that was identified both in domestic and wild pig samples: *Erysipelothrix rhusiopathiae*. Interestingly, we also found one species that was identified in domestic pigs: *Streptococcus suis*. Bacterial pathogens, out of our taxa of interest, found only in wild pig samples were: *Salmonella bongori*, *Salmonella enterica*, *Yersinia enterocolitica* and *Yersinia hibernica*.

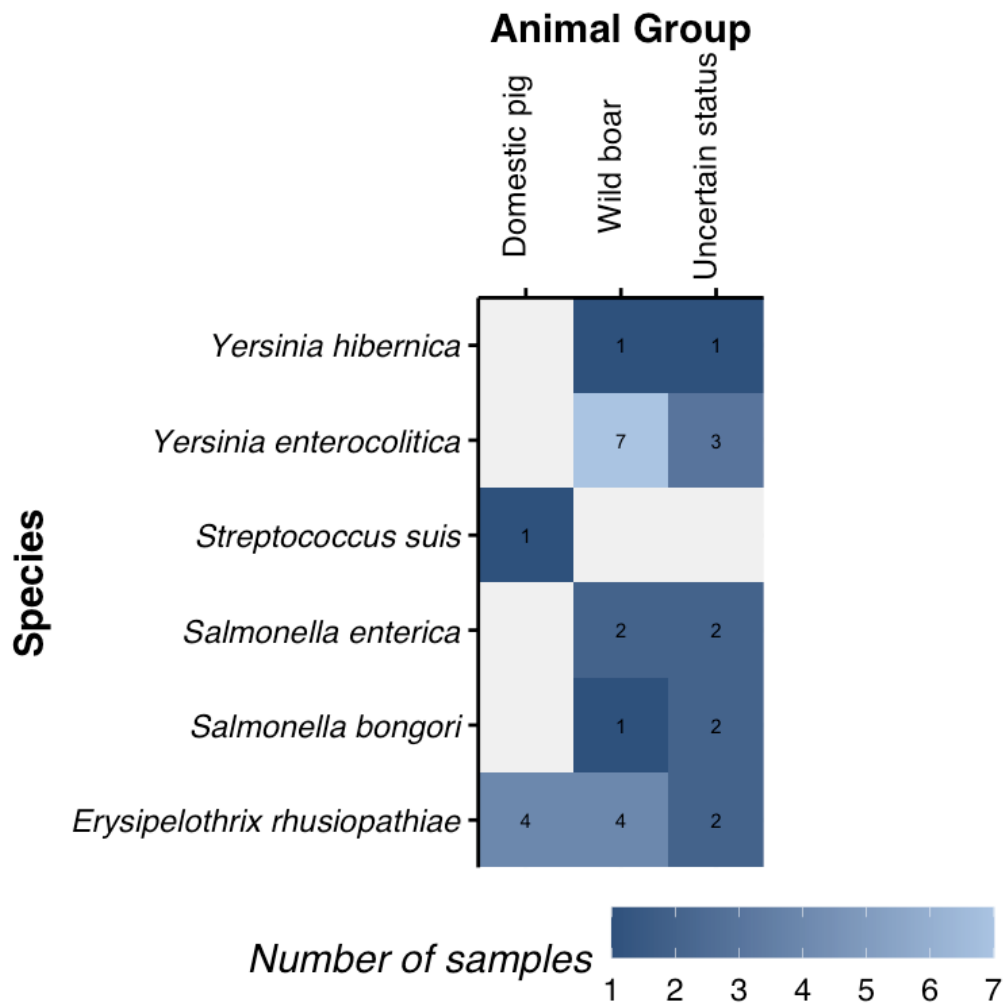


Figure 2. Heatmap of pathogenic species and animal groups that they were identified in. The number of samples, in which a pathogenic species was identified, is included in each square of the heatmap.

When comparing the identified species between modern and ancient domesticated animals, *Streptococcus suis* and *Erysipelothrix rhusiopathiae* were identified in ancient domesticated pigs, whereas no pathogens were confidently identified in modern domesticated pigs.

Domestication burden

Summarising the positive identifications across all the screened pathogenic species in time bins of 1,000 years (with the first time bin representing only the last 100 years), it is clear that pathogen presence discovery rate is low (22 samples out of 938 in total or 2.3%) (Figure 3). Pathogens were identified in two domestic pig samples from the last 4000 years. Both modern and ancient (10000-11000 YBP) wild boars have a higher observed total pathogen count (8 detected pathogens) than both the domestic and the unknown domestication status sample groups. Interestingly the presence of 6 pathogens was also identified in samples with an unclear domestication status, with most of them being identified in samples from the last 100 years.

In an effort to statistically test if the distribution of the identified pathogens differs between the domesticated pigs and wild boars for which we have complete metadata regarding their age and domestication status (pathogens identified in 2 domestic pig samples and 5 wild boar samples respectively), a Fisher's exact test ($p\text{-value} = 0.26$) was performed indicating that the distribution of the identified pathogens does not differ between the domesticated pig and wild boar groups. Of course the power of this statistical test is undermined by the small sample size of suid samples with complete age and domestication metadata.

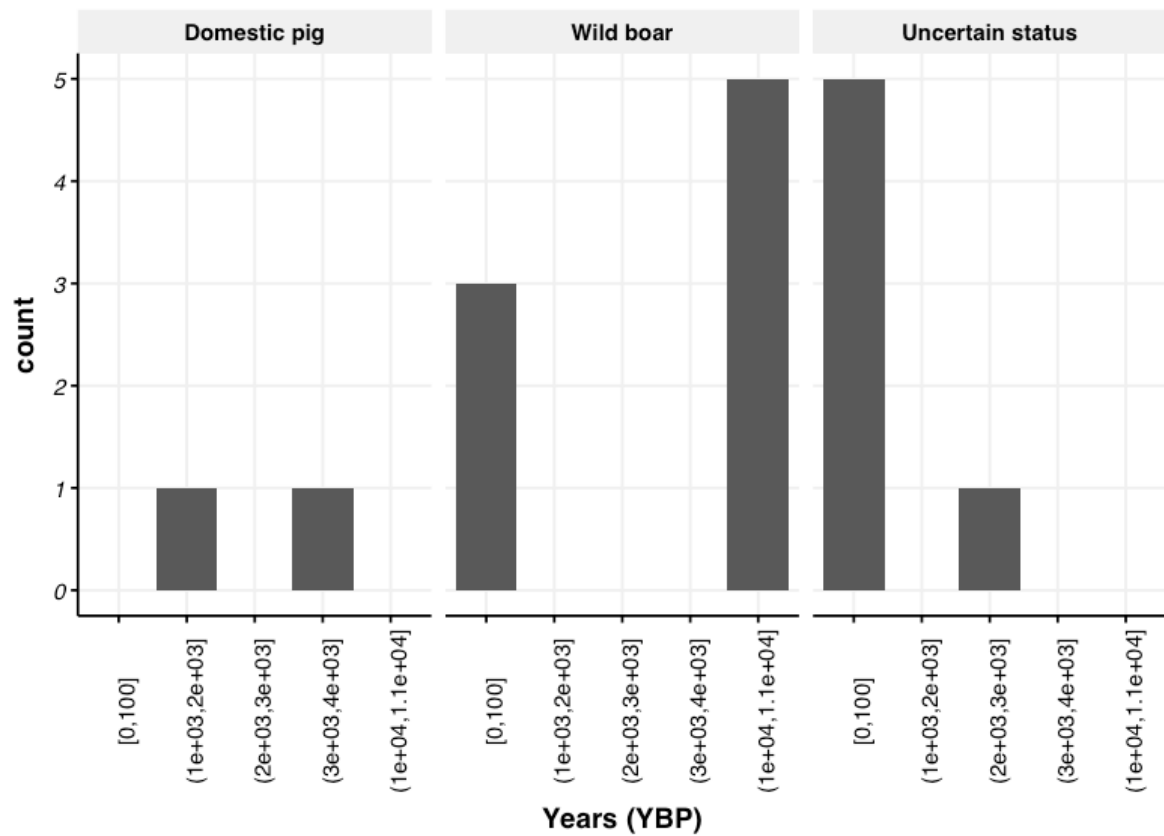


Figure 3. Presence of pathogens per time bin of a thousand years. More pathogens can be identified in wild boar samples than domestic pigs.

Dogs

Detected Pathogens

Pathogens were robustly identified in 34 dog samples (Table 1 and 3), retrieved mainly from Eurasia and North and South America (Figure 4). The causative agent of kennel cough, *Bordetella bronchiseptica*, was not identified in any samples, but *Bordetella parapertussis* was found in 7 samples (50 - 194 uniquely assigned reads). The causative agent of Lyme's disease, *Boreliella burgdorferi*, was not identified in any of the analysed samples. *Brucella canis*, the agent of canine brucellosis, was also not detected in any of the screened samples. The agent of diphtheria, *Corynebacterium diphtheriae*, was identified in 5 samples (93 - 462 uniquely assigned reads). None of the pathogenic agents of the genus *Ehrlichia*, responsible for ehrlichiosis, or any microbes of the genus *Rickettsia*, causing spotted fever or typhus-like disease in dogs, were found in the analysed dog samples. *Salmonella bongori* was detected in 5 samples (55 - 110 uniquely assigned reads), and *Salmonella enterica* was identified in 3 samples (68 - 93 uniquely assigned reads). *Streptococcus pneumoniae* was detected in 1 sample (51 reads uniquely assigned) and *Streptococcus canis*, the causative agent of streptococcosis in dogs, was found in 3 samples (71 - 93 uniquely assigned reads). Also, *Yersinia enterocolitica* was identified in 20 samples (50 - 6046 uniquely assigned reads), as well as *Yersinia pestis* which was detected in 1 sample (68 reads assigned uniquely).

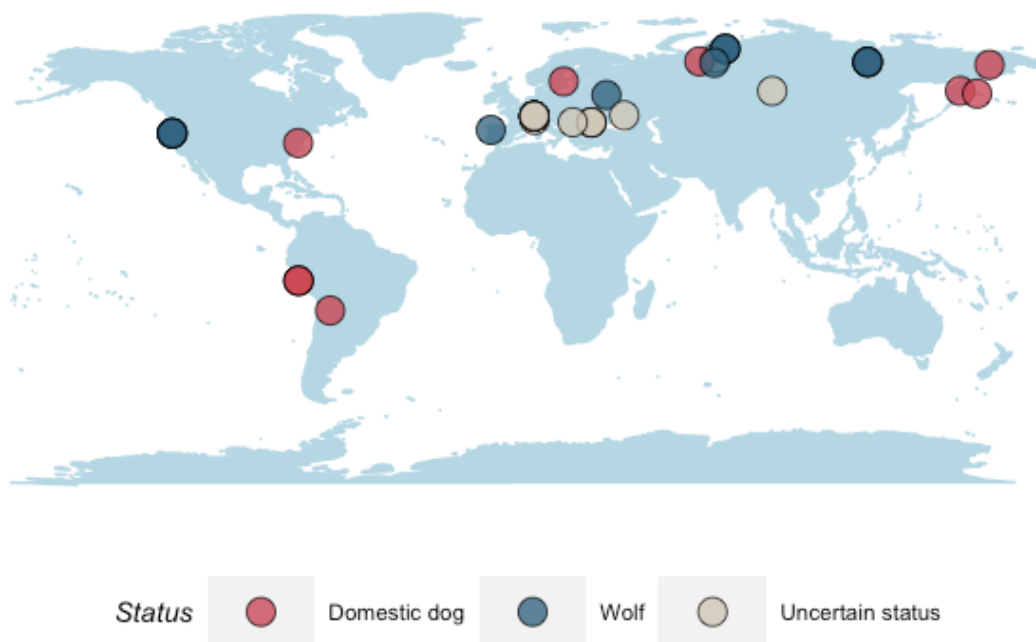


Figure 4. Map of canid samples, in which pathogenic species were positively identified in. The different colours represent the different domestication status calls.

Pathogenic species were identified in 13 domestic dog samples, out of which only 10 had reliable metadata regarding the archaeological context they were retrieved from and their estimated age. Out of these 10 samples 5 were modern dogs from Russian territories in Asia, 1 was modern European, 2 were European dating back to the Neolithic/Bronze Age, 1 of them was from South America dating back to the Pre-Columbian era, and 1 was from North America retrieved from an archaeological context belonging to a Native American culture. Pathogens were also identified in 7 wolf samples. From these 7 samples, 4 of them were modern Russian wolves and 1 was an early modern Russian wolf. Among the positive for pathogen wolf samples there were also 1 sample from Spain dating back to Pleistocene, and 1 sample from

North America dating back to the Rancholabrean period. Additionally there were 14 samples with an uncertain domestication status, in which pathogenic microbes were identified. Reliable archaeological metadata existed for only 12 of them, with 2 modern samples along with 1 dating back to the Pleistocene coming from Eastern Russia, 4 samples coming from Neolithic/Bronze Age Europe, 2 European samples from the Bronze Age, 1 European sample from the Roman Age, 1 Medieval European sample, and 1 modern European sample.

Table 3. Taxonomic and sample metadata for dog samples in which pathogens were robustly detected.

Sample	Taxon	Assigned Read Number	Age (Mean years BP)	Country	Location or Site	Latitude	Longitude	Region	Status	Period	Reason for status call
AC05	<i>Salmonella bongori</i>	69	13750	Russia	Afontova Gora	55.95	92.70	Asia	Uncertain status	Pleistocene	Very early, hard to determine status
AL2066	<i>Yersinia enterocolitica</i>	82	0	Sweden	Stockholm, Sodermanland and Uppland Province	59.36	18.00	Europe	Domestic dog	Modern	Modern / Historic
AL2317	<i>Yersinia enterocolitica</i>	170	12000	Spain	Casa de Moura	41.99	-8.12	Europe	Wolf	Pleistocene	Lack of relevant information
AL2679	<i>Yersinia enterocolitica</i>	50	0	Russia	Novy Port	67.69	72.75	Asia	Uncertain status	Modern	Lack of relevant information

AL2688	<i>Yersinia enterocolitica</i>	214	0	Russia	Nearest urban area (Sasha) Smolensk	54.49	33.30	Asia	Wolf	Modern	Modern
AL3007	<i>Yersinia enterocolitica</i>	423	0	Russia	Petropavlovsk-Kamchatsky, Kamchatka Krai	56.00	160.01	Asia	Domestic dog	Modern	Modern
AL3014	<i>Yersinia enterocolitica</i>	72	0	Russia	Zhigansky District, Sakha Republic	66.30	126.89	Asia	Domestic dog	Modern	Modern
AL3023	<i>Yersinia enterocolitica</i>	94	0	Russia	Anadyr River, Markovo, Chukotka Autonomous Okrug	65.33	170.58	Asia	Domestic dog	Modern	Modern
AL3046	<i>Yersinia enterocolitica</i>	89	0	Russia	Yakutia, Yana River	66.30	126.89	Asia	Wolf	Modern	Modern
AL3052	<i>Yersinia enterocolitica</i>	59	0	Russia	Bering Island	55.00	166.15	Asia	Domestic dog	Modern	Modern

AL 3062	<i>Salmonella bongori</i>	110	0	Russia	Khalmyer Bay, middle reaches of Neite-Yaga River	71.01	75.91	Asia	Wolf	Modern	Modern
AL 3062	<i>Salmonella enterica</i>	93	0	Russia	Khalmyer Bay, middle reaches of Neite-Yaga River	71.01	75.91	Asia	Wolf	Modern	Modern
AL 3062	<i>Yersinia enterocolitica</i>	358	0	Russia	Khalmyer Bay, middle reaches of Neite-Yaga River	71.01	75.91	Asia	Wolf	Modern	Modern
AL 3065	<i>Yersinia enterocolitica</i>	75	0	Russia	Yakutia, Indigirka River	66.30	126.89	Asia	Wolf	Modern	Modern
AL 3065	<i>Yersinia pestis</i>	68	0	Russia	Yakutia, Indigirka River	66.30	126.89	Asia	Wolf	Modern	Modern
AL 3090	<i>Corynebacterium diphtheriae</i>	94	0	Russia	Rostov-on-Don, Sholokhovsky	47.22	39.70	Asia	Uncertain status	Modern	Modern initially

					District, Rostov Province						suggested to be a wolf, but could potentially be a hybrid
	<i>Yersinia enterocolitica</i>	204	1965	Russia	Salekhard, Ust Polni	66.56	66.59	Asia	Domestic dog	NA	Archaeological context
AL3163	<i>Yersinia enterocolitica</i>	204	1965	Russia	Salekhard, Ust Polni	66.56	66.59	Asia	Domestic dog	NA	Archaeological context
	<i>Yersinia enterocolitica</i>	106	350	Russia	Nadym	66.03	72.00	Asia	Wolf	Early modern	morphology
AL3268	<i>Bordetella parapertussis</i>	114	11000	USA	Holden Cave	40.80	-122.30	North America	Wolf	Rancholabrean period	Archaeological context
ANH-1	<i>Corynebacterium diphtheriae</i>	462	11000	USA	Holden Cave	40.80	-122.30	North America	Wolf	Rancholabrean period	Archaeological context
ANH-1	<i>Salmonella bongori</i>	55	11000	USA	Holden Cave	40.80	-122.30	North America	Wolf	Rancholabrean period	Archaeological context
ANH-1	<i>Streptococcus canis</i>	93	11000	USA	Holden Cave	40.80	-122.30	North America	Wolf	Rancholabrean period	Archaeological context

ANH-1	<i>Yersinia enterocolitica</i>	6046	11000	USA	Holden Cave	40.80	-122.30	North America	Wolf	Rancholabrean period	Archaeological context
ANH-10	<i>Corynebacterium diphtheriae</i>	277	1150	Peru	NA	-12.03	-77.04	South America	Domestic dog	Pre-Columbian	Archaeological context
ANH-11	<i>Corynebacterium diphtheriae</i>	93	1150	Peru	NA	-12.03	-77.04	South America	Domestic dog	Pre-Columbian	Archaeological context
ANH-11	<i>Streptococcus canis</i>	71	1150	Peru	NA	-12.03	-77.04	South America	Domestic dog	Pre-Columbian	Archaeological context
ANH-142	<i>Yersinia enterocolitica</i>	109	NA	USA	McPherson's Cave	36.71	-83.13	North America	Domestic dog	NA	Lack of relevant information
ANH-2	<i>Corynebacterium diphtheriae</i>	102	NA	USA	NA	37.55	-77.47	North America	Uncertain status	NA	Lack of relevant information
ANH-6	<i>Streptococcus pneumoniae</i>	51	750	Argentina	Pucara De Tilcara	-22.71	-65.69	South America	Domestic dog	Pre-Columbian	Archaeological context
ANH-79	<i>Yersinia enterocolitica</i>	57	985	USA	Weyanoke	37.30	-77.16	North America	Domestic dog	Native American	Lack of relevant information

CH1581	<i>Bordetella parapertussis</i>	63	3900	Romania	Bordusani	44.68	27.95	Europe	Uncertain status	Bronze Age	Lack of relevant information
CH1665	<i>Bordetella parapertussis</i>	97	3900	Romania	Harsova	44.68	27.95	Europe	Uncertain status	Bronze Age	Lack of relevant information
CH1691	<i>Bordetella parapertussis</i>	50	0	Romania	Harsova	44.68	27.95	Europe	Uncertain status	Modern	Lack of relevant information
OL2186	<i>Yersinia enterocolitica</i>	62	970	Switzerland	NA	45.55	7.58	Europe	Uncertain status	Medieval	Lack of relevant information
OL4153	<i>Streptococcus canis</i>	81	1766	Serbia	Kostolac, Branicevo District	44.74	21.22	Europe	Uncertain status	Roman	Lack of relevant information
OL4167	<i>Bordetella parapertussis</i>	78	6000	Switzerland	NA	46.96	7.43	Europe	Domestic dog	Neolithic/Bronze Age	Lack of relevant information
OL4170	<i>Salmonella bongori</i>	96	6000	Switzerland	Vineltz	46.96	7.43	Europe	Uncertain status	Neolithic/Bronze Age	Lack of relevant information
OL4170	<i>Salmonella enterica</i>	72	6000	Switzerland	Vineltz	46.96	7.43	Europe	Uncertain status	Neolithic/Bronze Age	Lack of relevant information

OL4171	<i>Yersinia enterocolitica</i>	427	6000	Switzerland	Vineltz	46.96	7.43	Europe	Domestic dog	Neolithic/Bronze Age	Lack of relevant information
OL4172	<i>Bordetella pertussis</i>	51	6000	Switzerland	Vineltz	46.96	7.43	Europe	Uncertain status	Neolithic/Bronze Age	Lack of relevant information
OL4173	<i>Bordetella pertussis</i>	194	6000	Switzerland	Vineltz	46.96	7.43	Europe	Uncertain status	Neolithic/Bronze Age	Lack of relevant information
OL4174	<i>Salmonella enterica</i>	68	6000	Switzerland	Vineltz	46.96	7.43	Europe	Uncertain status	Neolithic/Bronze Age	Lack of relevant information
OL4174	<i>Salmonella bongori</i>	65	6000	Switzerland	Vineltz	46.96	7.43	Europe	Uncertain status	Neolithic/Bronze Age	Lack of relevant information
OL4174	<i>Yersinia enterocolitica</i>	218	6000	Switzerland	Vineltz	46.96	7.43	Europe	Uncertain status	Neolithic/Bronze Age	Lack of relevant information
WS17	<i>Yersinia enterocolitica</i>	55	NA	Thailand	Ban Chiang, Udon Thani Province	17.41	103.24	South East Asia	Uncertain status	NA	Lack of relevant information

Pathogen distribution and species found

We decided to further look into the distribution of pathogens among the canine samples for which we had reliable metadata on their domestication status and estimated age (Figure 5). Even though we did not identify all our identified species of interest in our dog sample collection, we decided to look further into the distribution of the (zoonotic) pathogen species we identified. There were four pathogenic bacterial species that were identified both in domestic and wild dog samples: *Bordetella parapertussis*, *Corynebacterium diphtheriae*, *Streptococcus canis* and *Yersinia enterocolitica*. Interestingly, we found one species that was identified in a domestic dog sample: *Streptococcus pneumoniae*. Bacterial pathogens found only in wild dog samples were: *Salmonella bongori*, *Salmonella enterica*, and *Yersinia pestis*.

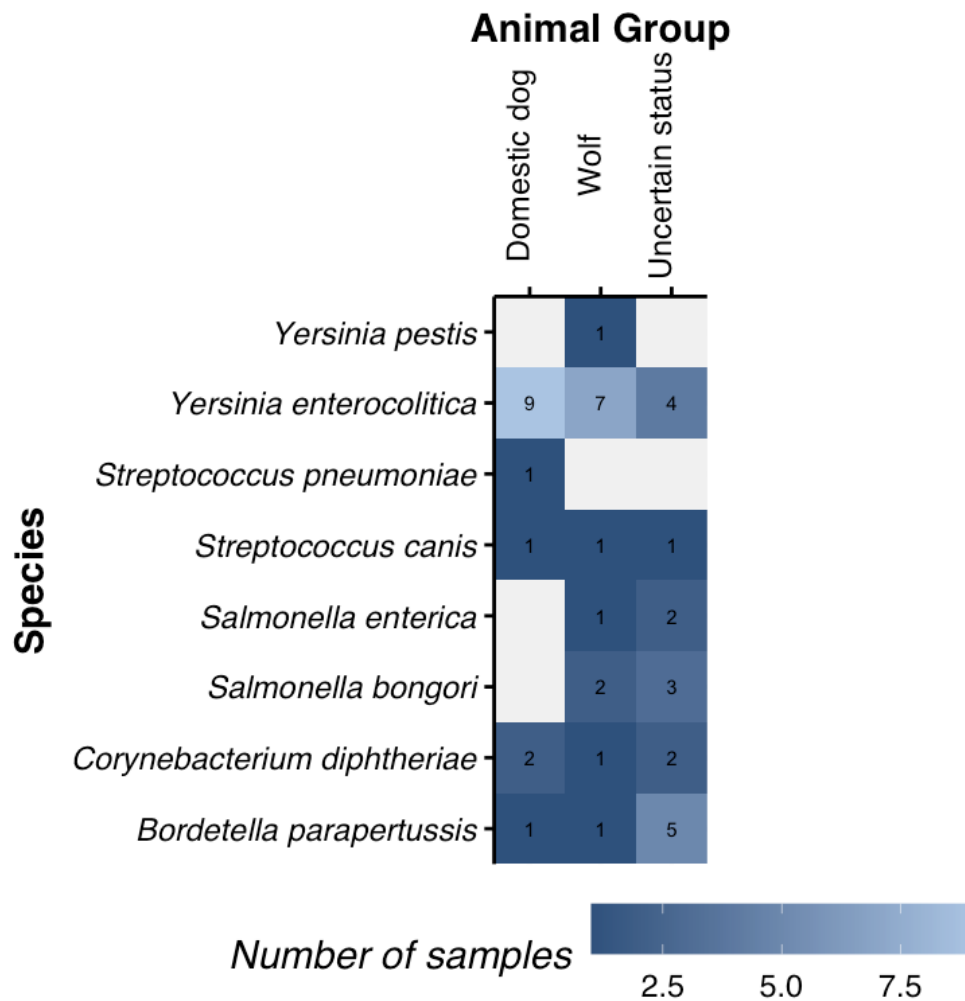


Figure 5. Heatmap of pathogenic species identified in canine samples, grouped according to their domestication status. The number of samples, in which a pathogenic species was identified, is included in each square of the heatmap.

When comparing the species identified in modern and ancient domesticated animals, no pathogens were only identified in modern dog samples. Interestingly *Bordetella parapertussis*, *Corynebacterium diphtheriae*, *Streptococcus canis*, and *Streptococcus pneumoniae* were identified solely in archaeological samples of domesticated dogs. *Yersinia enterocolitica* was the only zoonotic pathogen being identified in both modern and ancient domesticated dogs.

Domestication burden

Summarising the positive identifications across all the screened pathogenic species in time bins of 1000 years (with the first time bin representing only the last 100 years), it is again evident that pathogen presence discovery rate is low (34 positive samples out of 2090 in total or 1.6%) (Figure 6). There were 13 pathogenic species detected in domestic dogs from the last 5000 YPB. More pathogens were identified in wolf samples (14 pathogens), with most of the positive hits coming from wolf samples from the last 1000 years (8 identified species) and from 10-11 thousand years BP (6 detected species). More pathogenic agents were found in canine samples with a non definitive domestication status call (15 detected species in total), with most of these pathobionts being detected in samples dating between five and six thousand years BP.

In an effort to statistically test if the distribution of the identified pathogens differs between the domesticated dogs and wolves for which we have reliable age metadata (pathogens identified in 12 domestic dog samples and 7 wolf samples respectively), a Fisher's exact test ($p\text{-value} = 0.035$) was performed. The statistical test suggests that the distribution of the identified pathogens differed between domesticated dog and wolf groups, with domesticated dogs carrying more pathogens than wolves. This increase, as observed from our data, in pathogen frequency over the last 5000 years, despite the low absolute counts of positive identifications, indicates a possible increase in the disease burden of domestic dog populations across Eurasia and America. Of course this statistical test is again undermined by the small sample size

of canid samples in which the presence of a pathogen has been identified, and have complete age and domestication metadata available.

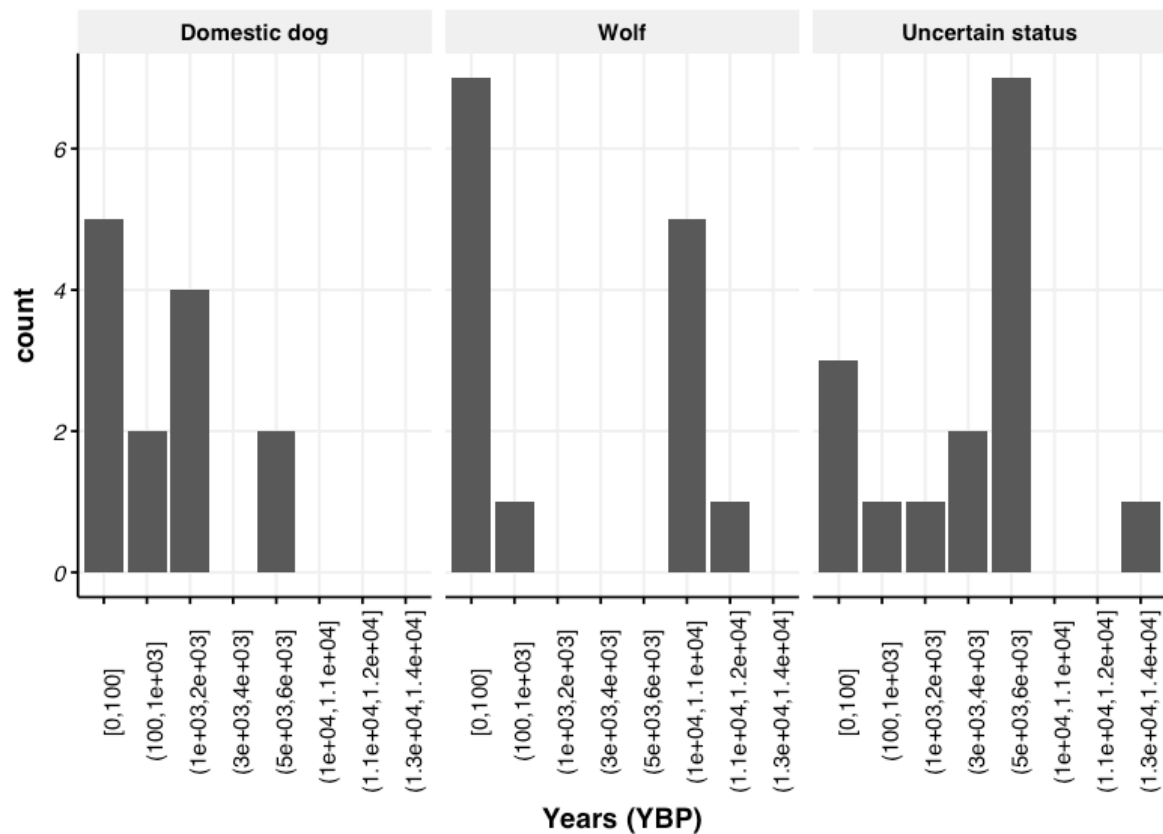


Figure 6. Presence of pathogens per time bin of a thousand years. More pathogens can be identified in domestic and wild samples from the last five thousand years.

Discussion

Factors confounding ancient metagenomic analyses

Pathogen presence discovery rate in archaeological samples is low. In our datasets, the pathogen presence discovery rate in pigs and dogs were 2.3% and 1.6% respectively. This calculation was based on libraries of 938 pig and 2,090 dog samples having been screened with HAYSTAC (Dimopoulos et al., 2020), and the presence of (zoonotic) pathogens was robustly detected in only 22 pig and 34 dog samples. Low pathogen presence discovery rates can be a consequence of a variety of factors. Poor preservation of biological material (such as bone or dental tissue) can lead to poor DNA molecule preservation (Stone and Ozga, 2019). Poorly preserved DNA molecules are not informative, as they are too short or chemically damaged to produce valid alignments, thus hindering metagenomic analyses. Another important factor, that works usually in synergy with poor DNA preservation, is the variable, usually low abundance, of pathogen DNA in a host's tissues (Key et al., 2017). Different pathogens infect and grow optimally in specific tissues in a host's body that might not always be preserved in the archaeological record (e.g. soft tissue). Therefore, any variation in pathogen DNA preservation across different host tissues due to the underlying biology of a microbe, further confounds ancient DNA metagenomic analyses.

The sequencing depth of a aDNA sample is another important factor affecting pathogen discovery. The sensitivity of pathogen detection in NGS metagenomic samples can greatly improve by increasing the sequencing breadth and depth of a sample (Frey et al., 2014; Gu et al., 2021; Schlaberg et al., 2017; Thorburn et al., 2015; Wylezich et al., 2018). The sequencing depth of the microbial fraction of a

sample can be increased either by reducing the number of multiplexed samples per lane, or use platforms with greater throughput, or employing protocols that either deplete host DNA (Hasan et al., 2016) or enrich pathogen DNA (Lee et al., 2017; Thorburn et al., 2015; Wylezich et al., 2018). It needs to be underlined that the even pooling of samples, prior to sequencing, is important for ensuring that each sample yields the same amount of DNA data that maximises the sensitivity of the downstream metagenomic analyses (Schlaberg et al., 2017), and that pathogen enrichment protocols can alter the composition of the sample, possibly leading to the loss of information (e.g. losing DNA reads of off-target pathogens that could be co-infecting a host).

Based on previous studies a threshold level of 5 million reads is required for the robust detection of an organism in a metagenomic sample (Schlaberg et al., 2017). Practically though the number of reads necessary to perform a metagenomic identification is expected to vary depending on the titer of an infection, with high titer infections requiring only hundreds of thousands of reads and low titer infections needing tens of millions of reads (Frey et al., 2014). Moreover, distinguishing between taxa with high sequence similarity, for instance identifying different microbial subspecies, will also require an increase in sequencing depth as well (Schlaberg et al., 2017). Finally, for the characterisation of a pathogen, which follows its identification, (nearly) complete genome representation is required at an increased depth of coverage, creating the need for deeper sequencing (Frey et al., 2014).

In the present study the majority of the analysed NGS data were highly multiplexed when pooled for sequencing and were only sequenced at a shallow depth, yielding

less than 5 million reads per sample in many cases. Therefore, it is highly likely that we failed to identify the presence of pathogens in a large number of suid and canid samples, as the sequencing depth of those samples was not sufficient. Unfortunately, resequencing more than 1000 suid and 2000 canid samples deeper is not an option, despite the correlation between pathogen detection and sequencing depth (Frey et al., 2014; Gu et al., 2021; Schlager et al., 2017; Thorburn et al., 2015; Wylezich et al., 2018). A different downstream analytical strategy could be taken though. A less conservative approach than the one that was followed in this paper could include relaxing the parameters that control HAYSTAC's probabilistic model for species assignment. A combination of a higher base mismatch probability (> 0.05) and a lower read assignment probability threshold ($0.75 <$) would have increased HAYSTAC's sensitivity. Even though not many of our pathogen-positive libraries were deeply sequenced, a standardised approach could also be taken in order to identify the lowest limits of sequencing depth, that is required for robust pathogen detection by HAYSTAC. By subsampling a deeply sequenced suid or canid library at different sequencing depths, followed by a metagenomic analysis with HAYSTAC (leniently parameterised), we could determine provisionally (as other factors, e.g. DNA degradation, are simultaneously affecting aDNA pathogen discovery) the lowest sequencing depth needed for the confident identification of pathogens in ancient metagenomic samples.

Unfortunately, the inability to detect pathogenic DNA in shotgun DNA sequencing data yielded by archaeological samples is not only influenced by DNA preservation or sequencing depth. It is also influenced by the underlying biology of the targeted pathogen itself. Failing to identify a pathogen does not always equate to an animal

being neither a carrier nor infected by a pathogen at the time of its death or earlier, as absence of evidence is not evidence of absence. Genetic traces of bacterial pathogens are hard to find in a host's body that has overcome an infection (unless the bacterium persists). Also, an archaeological sample with a low viral or bacterial load could go undetected under the current sensitivity threshold for a taxonomic call (50 uniquely assigned DNA molecules) in our analysis. Moreover, there's a plethora of undiscovered microbial species and strains whose sequences have never been categorised in a reference database (Ye et al., 2019), further hindering the detection of pathogens in animal samples. This is exacerbated by the expected loss of ancient variants/strains of contemporary zoonotic pathogens, which were circulating among animal populations in the past and are currently extinct, rendering us unable to identify them due to the incompleteness of reference databases. Our ability to detect the presence of pathogens by using aDNA is also affected by the length of time an active infection lasted and whether the pathogen persisted in the body of the animal host after the clinical symptoms of the infection cleared. Thus, the detection of pathogen aDNA is hindered in cases where animals overcome bacterial infections by pathogenic agents that do not persist in their bodies afterwards.

Concurrently detecting pathogen DNA in DNA sequencing data yielded by an archaeological animal sample also does not mean that the animal was undergoing an active infection nor was exhibiting disease symptoms, as it is still possible that the animal was just a carrier of said pathogen. For instance, there is evidence that clinically normal dogs can be chronic carriers of infections caused by species of the genus *Leptospira* (Rojas et al., 2010), and that symptomless pigs can be carriers of *Streptococcus suis* able to infect both other pigs and humans (Lun et al., 2007).

Further complicating the interpretation of identifying the presence of pathogen DNA in suid or canid samples is the lack of time course studies for certain pathogens, such as the species of the genus *Campylobacter* or *Rickettsia* or *Streptococcus canis*. When combined with the fact that most of the zoonotic microbes that our samples were screened for are able to establish chronic or asymptomatic infections, does not allow us to conclude whether the presence of a pathogen's DNA is due to an active infection or asymptomatic carriage. It is important to exercise caution and take the above factors into account when drawing conclusions about the prevalence of a pathogenic agent across time.

Elaborating on the metagenomic identification results

In this paper we screened DNA sequencing data for pathogens belonging to 11 different genera, and for a robust species identification we required a minimum of 50 uniquely assigned reads. Because we built a separate database for each genus we were interested in, reads that could not be uniquely assigned to a species in a genus were assigned to a category called Grey Matter (Dimopoulos et al., 2020). Grey Matter is an absorbent state taxonomic category that encompasses DNA reads of low informational value. This could be owed to these reads representing environmental contamination, genomic regions conserved across a genus, aDNA reads which are too damaged to produce informative alignments, or reads representing ancient microbial strains basal to the genomes of extant members of a genus. In our analyses, there were many cases in which the Grey Matter category had a high relative abundance (but not for each sample for every genus), indicating the presence of a certain genus in the DNA shotgun sequencing data of a sample.

For this paper though, we followed a more conservative approach and only focused on the chosen pathogenic species that could be confidently detected in our samples.

When performing metagenomic identifications, result validation is of great importance, as false positive identification of a pathogen's presence affects the reliability of presence/absence studies. For reliable metagenomic taxonomic identifications, especially in the context of ancient DNA, noise from modern environmental and human contamination needs to be eliminated. Reference genomes that contain human contamination, or genomic regions shared among different prokaryotic species as a result of Horizontal Gene Transfer (HGT) events, can attract DNA reads from multiple taxa (including humans), that give rise to false positive identifications. In an effort to avoid misidentifications from these sources, we ensured that the uniquely assigned reads to each identified species represented a random sample of their genome and did not cluster around specific genetic loci. In the case of ancient samples we also made sure that chemical damage, due to deamination (Briggs et al., 2007), was present. Unfortunately the low number of reads, which were uniquely assigned to our taxa of interest, in many cases were not enough for a deamination frequency 'smiling' effect plot to be produced.

The majority of the pathogens whose presence was identified in this paper had less than 1000 DNA reads assigned to them. Unfortunately this is an insufficient number of DNA molecules to either reconstruct ancient bacterial genomes, or provide robust information for any downstream evolutionary analyses, such as building phylogenetic trees and calculating evolutionary clock rates. For the continuation of this project, we will need to synthesise custom DNA probes that target the DNA molecules of each

pathogen and perform in-solution captures (Enk et al., 2014) of their genomes. Sequencing the bacterial pathogen DNA enriched libraries will allow us to successfully reconstruct the genomes of ancient bacterial microbes and will enable us to study their evolution in detail.

Putting metagenomic results into perspective

Because of the low pathogen presence discovery rate observed in our pig and dog datasets we were not able to analytically test and calculate any increase in the disease burden that comes along with domestication. In the suid dataset, we can qualitatively observe that fewer instances of pathogen presence are identified in domestic pig samples (2 bacterial pathogens in total corresponding to 2 samples) from the last 4,000 YBP than in wild boars (8 bacterial pathogens in total corresponding to 5 samples) from the last 11,000 YBP. A relatively high number of bacterial agents (6 positive hits in total corresponding to 3 samples) were identified in samples that did not have a definite domestication status call unfortunately, hindering our interpretation of how domestication has affected animal health. Unfortunately, statistically there is no significant difference among the number of samples positive for pathogens between domesticated pigs and wild boars.

On the other hand when focusing on samples with reliable domestication and age metadata, a trend can be observed for the canid dataset. The presence of 13 pathogenic agents was detected in domestic dog samples (age range 0 - 6000 YBP), and also the presence of 14 pathogenic agents was detected in wolves (age range 0

- 11000 YBP). The majority of the pathogenic agents present were identified in the unknown domestication status category, with 15 bacterial species in total (corresponding to 12 samples). Although the same number of pathogens present were identified between the domestic dog and wolf samples, in the domestic dog group that corresponded to 12 individual samples whereas the same number corresponded to only 7 individuals in the wolf group. Even though this can be interpreted as an artefact of DNA preservation, it can simultaneously suggest that the disease burden of domestic dogs increased, as humans started to intensify the management of their populations. Unfortunately no sound conclusions can be drawn as the sample size of pathogen-positive samples, with known age and domestication status, is really low. The pattern of increased pathogen presence detection, which we qualitatively observed in the analysed domesticated dog dataset, during the last 3000 years, agrees with what has been reported in the literature regarding an increased frequency and/or emergence of zoonotic diseases during the 20th century (Gibbs, 2005; Shchelkunov, 2013; Wilcox and Gubler, 2005). The increased frequency of pathogen presence detection, and the subsequent increase of the disease burden that it incurs on domesticated animals, could be reflective of how human activity has been reshaping the disease landscape of domesticated animals, though intensive management, breeding, modern agricultural practices and environmental change (Jones et al., 2013).

Unfortunately, the presence of pathogens was robustly identified in only 22 out of the 938 suid samples and 34 out of the 2090 canid samples, leading to a pathogen discovery rate of 2.3% and 1.6% respectively. Out of these pathobiont positive samples, even fewer had complete metadata regarding their age and domestication

status. The low number of individuals that were positive for pathogens did not allow for dense enough sampling throughout time, thus not allowing us to construct time series of pathogen presence/absence and analyse any temporal signal that could have arisen. Furthermore, the low pathogen discovery rates combined with the lack of age and domestication related sample metadata further undermined our ability to statistically test for differences in pathogen prevalence among domesticated and wild animals, as well as modern and ancient animals. For these reasons, we were not able to statistically test and quantify any changes in the disease burden of domesticated or wild animals throughout time and space. Therefore the results presented in this chapter are qualitative observations.

Future work is necessary though in order to gain a clearer insight into the evolutionary trajectory of these pathogens. Additional work will need to be undertaken to generate genomic data of ancient pathogenic strains, through targeted genome enrichment, that can be used to answer questions connecting domestication and the zoonotic disease landscape. Questions surrounding the emergence and transmission of foodborne pathogens are a great example. By constructing robust phylogenies of modern and ancient strains of foodborne pathogens, such as *Yersinia enterocolitica* that was identified in both suid and canid samples, we can trace their transmission routes among various hosts (e.g. pigs, dogs, humans) and gain a better understanding of which domesticated animal they employ as their primary reservoir. The combination of ancient and modern pathogen genetic data can significantly contribute to solving problems pertaining to the disease burden modern animal and human populations carry due to foodborne pathogens.

Bibliography

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., Carniel, E., 1999.

Yersinia pestis, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 14043–14048.

Ackers, M.L., Schoenfeld, S., Markman, J., Smith, M.G., Nicholson, M.A., DeWitt,

W., Cameron, D.N., Griffin, P.M., Slutsker, L., 2000. An outbreak of *Yersinia enterocolitica* O:8 infections associated with pasteurized milk. *J. Infect. Dis.* 181, 1834–1837.

Alfredson, D.A., Korolik, V., 2007. Antibiotic resistance and resistance mechanisms

in *Campylobacter jejuni* and *Campylobacter coli*. *FEMS Microbiol. Lett.* 277, 123–132.

Algammal, A.M., Hetta, H.F., Elkelish, A., Alkhalifah, D.H.H., Hozzein, W.N., Batiha,

G.E.-S., El Nahhas, N., Mabrok, M.A., 2020. Methicillin-Resistant

Staphylococcus aureus (MRSA): One Health Perspective Approach to the Bacterium Epidemiology, Virulence Factors, Antibiotic-Resistance, and Zoonotic Impact. *Infect. Drug Resist.* 13, 3255–3265.

Allen, R.C., 1999. Tracking the Agricultural Revolution in England. *Econ. Hist. Rev.*

52, 209–235.

Alter, T., Gaull, F., Kasimir, S., Gürtler, M., Mielke, H., Linnebur, M., Fehlhaber, K.,

2005. Prevalences and transmission routes of *Campylobacter* spp. strains within multiple pig farms. *Vet. Microbiol.* 108, 251–261.

Asakawa, Y., Akahane, S., Kagata, N., Noguchi, M., Sakazaki, R., 1973. Two

community outbreaks of human infection with *Yersinia enterocolitica*. *J. Hyg.* 71, 715–723.

Atkins, K.E., Read, A.F., Savill, N.J., Renz, K.G., Islam, A.F.M.F., Walkden-Brown,

- S.W., Woolhouse, M.E.J., 2013. Vaccination and reduced cohort duration can drive virulence evolution: Marek's disease virus and industrialized agriculture. *Evolution* 67, 851–860.
- Ayyadurai, S., Sebbane, F., Raoult, D., Drancourt, M., 2010. Body lice, yersinia pestis orientalis, and black death. *Emerg. Infect. Dis.* 16, 892–893.
- Bang, D.D., Nielsen, E.M., Scheutz, F., Pedersen, K., Handberg, K., Madsen, M., 2003. PCR detection of seven virulence and toxin genes of *Campylobacter jejuni* and *Campylobacter coli* isolates from Danish pigs and cattle and cytolethal distending toxin production of the isolates. *J. Appl. Microbiol.* 94, 1003–1014.
- Beall, M.J., Alleman, A.R., Breitschwerdt, E.B., Cohn, L.A., Couto, C.G., Dryden, M.W., Guptill, L.C., Iazbik, C., Kania, S.A., Lathan, P., Little, S.E., Roy, A., Sayler, K.A., Stillman, B.A., Welles, E.G., Wolfson, W., Yabsley, M.J., 2012. Seroprevalence of *Ehrlichia canis*, *Ehrlichia chaffeensis* and *Ehrlichia ewingii* in dogs in North America. *Parasit. Vectors* 5, 29.
- Beloeil, P.-A., Fravalo, P., Fablet, C., Jolly, J.-P., Eveno, E., Hascoet, Y., Chauvin, C., Salvat, G., Madec, F., 2004. Risk factors for *Salmonella enterica* subsp. *enterica* shedding by market-age pigs in French farrow-to-finish herds. *Prev. Vet. Med.* 63, 103–120.
- Bermúdez, C.S.E., Troyo, A., 2018. A review of the genus *Rickettsia* in Central America. *Res. Rep. Trop. Med.* 9, 103–112.
- Berns, K.I., Casadevall, A., Cohen, M.L., Ehrlich, S.A., Enquist, L.W., Fitch, J.P., Franz, D.R., Fraser-Liggett, C.M., Grant, C.M., Imperiale, M.J., Kanabrocki, J., Keim, P.S., Lemon, S.M., Levy, S.B., Lumpkin, J.R., Miller, J.F., Murch, R., Nance, M.E., Osterholm, M.T., Relman, D.A., Roth, J.A., Vidaver, A.K., 2012. Public health and biosecurity. Adaptations of avian flu virus are a cause for

- concern. *Science* 335, 660–661.
- Bescucci, D.M., Moote, P.E., Ortega Polo, R., Uwiera, R.R.E., Inglis, G.D., 2020. *Salmonella enterica* Serovar Typhimurium Temporally Modulates the Enteric Microbiota and Host Responses To Overcome Colonization Resistance in Swine. *Appl. Environ. Microbiol.* 86. <https://doi.org/10.1128/AEM.01569-20>
- Bhaduri, S., Wesley, I.V., Bush, E.J., 2005. Prevalence of pathogenic *Yersinia enterocolitica* strains in pigs in the United States. *Appl. Environ. Microbiol.* 71, 7117–7121.
- Bintsis, T., 2017. Foodborne pathogens. *AIMS Microbiol* 3, 529–563.
- B. Lopes, L., Nicolino, R., P.A. Haddad, J., 2010. Brucellosis - risk factors and prevalence: A review. *Open Vet. Sci. J.* 4, 72–84.
- Bonardi, S., Bruini, I., D’Incau, M., Van Damme, I., Carniel, E., Brémont, S., Cavallini, P., Tagliabue, S., Brindani, F., 2016. Detection, seroprevalence and antimicrobial resistance of *Yersinia enterocolitica* and *Yersinia pseudotuberculosis* in pig tonsils in Northern Italy. *Int. J. Food Microbiol.* 235, 125–132.
- Bonmarin, I., Guiso, N., Le Flèche-Matéos, A., Patey, O., Patrick, A.D.G., Levy-Bruhl, D., 2009. Diphtheria: a zoonotic disease in France? *Vaccine* 27, 4196–4200.
- 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., Campbell, T.J., Majander, K., Wilbur, A.K., Guichon, R.A., Wolfe Steadman, D.L., Cook, D.C., Niemann, S., Behr, M.A., Zumarraga, M., Bastida, R., Huson, D., Nieselt, K., Young, D., Parkhill, J., Buikstra, J.E., Gagneux, S., Stone, A.C., Krause, J., 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 514, 494–497.

- 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., Wood, J., Earn, D.J.D., Herring, D.A., Bauer, P., Poinar, H.N., Krause, J., 2011. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* 478, 506–510.
- Botelho-Nevers, E., Socolovschi, C., Raoult, D., Parola, P., 2012. Treatment of *Rickettsia* spp. infections: a review. *Expert Rev. Anti. Infect. Ther.* 10, 1425–1437.
- Bowser, N.H., Anderson, N.E., 2018. Dogs (*Canis familiaris*) as Sentinels for Human Infectious Disease and Application to Canadian Populations: A Systematic Review. *Vet. Sci. China* 5. <https://doi.org/10.3390/vetsci5040083>
- Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., Pääbo, S., 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* 104, 14616–14621.
- Brooke, C.J., Riley, T.V., 1999. *Erysipelothrix rhusiopathiae*: bacteriology, epidemiology and clinical manifestations of an occupational pathogen. *J. Med. Microbiol.* 48, 789–799.
- Brower, A., Okwumabua, O., Massengill, C., Muenks, Q., Vanderloo, P., Duster, M., Homb, K., Kurth, K., 2007. Investigation of the spread of *Brucella canis* via the U.S. interstate dog trade. *Int. J. Infect. Dis.* 11, 454–458.
- Bucher, M., Meyer, C., Grötzbach, B., Wacheck, S., Stolle, A., Fredriksson-Ahomaa, M., 2008. Epidemiological data on pathogenic *Yersinia enterocolitica* in Southern Germany during 2000-2006. *Foodborne Pathog. Dis.* 5, 273–280.
- Cabezas-Cruz, A., Zweggarth, E., Vancová, M., Broniszewska, M., Grubhoffer, L.,

- Passos, L.M.F., Ribeiro, M.F.B., Alberdi, P., de la Fuente, J., 2016. *Ehrlichia minasensis* sp. nov., isolated from the tick *Rhipicephalus microplus*. *Int. J. Syst. Evol. Microbiol.* 66, 1426–1430.
- Campbell, J., Ereemeeva, M.E., Nicholson, W.L., McQuiston, J., Parks, S., Adjemian, J., McElroy, K., Others, 2009. Outbreak of *Rickettsia typhi* infection-Austin, Texas, 2008. *MMWR Surveill. Summ.* 58, 1267–1270.
- Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.H.S., Samaniego, J.A., Wales, N., Sicheritz-Pontén, T., Gilbert, M.T.P., 2018. Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419.
- Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y., Wang, J., Gilbert, M.T.P., Willerslev, E., Greenleaf, W.J., Bustamante, C.D., 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* 93, 852–864.
- Cleaveland, S., Laurenson, M.K., Taylor, L.H., 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356, 991–999.
- Cosford, K.L., 2018. *Brucella canis*: An update on research and clinical management. *Can. Vet. J.* 59, 74–81.
- Cummings, K.J., Mitchell, P.K., Rodriguez-Rivera, L.D., Goodman, L.B., 2020. Sequence analysis of *Salmonella enterica* isolates obtained from shelter dogs throughout Texas. *Vet Med Sci* 6, 975–979.
- Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B.,

- Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., Meyer, M., 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* 110, 15758–15763.
- Dimopoulos, E.A., Carmagnini, A., Velsko, I.M., Warinner, C., Larson, G., Frantz, L.A.F., Irving-Pease, E.K., 2020. HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2020.12.16.419085>
- Duncan, A.W., Correa, M.T., Levine, J.F., Breitschwerdt, E.B., 2005. The dog as a sentinel for human infection: prevalence of *Borrelia burgdorferi* C6 antibodies in dogs from southeastern and mid-Atlantic States. *Vector Borne Zoonotic Dis.* 5, 101–109.
- Eamens, G.J., Forbes, W.A., Djordjevic, S.P., 2006. Characterisation of *Erysipelothrix rhusiopathiae* isolates from pigs associated with vaccine breakdowns. *Vet. Microbiol.* 115, 329–338.
- Edinboro, C.H., Ward, M.P., Glickman, L.T., 2004. A placebo-controlled trial of two intranasal vaccines to prevent tracheobronchitis (kennel cough) in dogs entering a humane shelter. *Prev. Vet. Med.* 62, 89–99.
- Eisen, R.J., Bearden, S.W., Wilder, A.P., Montenieri, J.A., Antolin, M.F., Gage, K.L., 2006. Early-phase transmission of *Yersinia pestis* by unblocked fleas as a mechanism explaining rapidly spreading plague epizootics. *Proc. Natl. Acad. Sci. U. S. A.* 103, 15380–15385.
- Enache, A.E., Mitchell, C., Kafarnik, C., Waller, A.S., 2020. *Streptococcus canis* multilocus sequence typing in a case series of dogs with ulcerative keratitis. *Vet. Ophthalmol.* 23, 252–258.

- Enk, J.M., Devault, A.M., Kuch, M., Murgha, Y.E., Rouillard, J.-M., Poinar, H.N.,
2014. Ancient whole genome enrichment using baits built from modern DNA.
Mol. Biol. Evol. 31, 1292–1294.
- Eriksson, H., Jansson, D.S., Johansson, K.-E., Båverud, V., Chirico, J., Aspán, A.,
2009. Characterization of *Erysipelothrix rhusiopathiae* isolates from poultry, pigs,
emus, the poultry red mite and other animals. Vet. Microbiol. 137, 98–104.
- Facciola, A., Palamara, M.A.R., D'Andrea, G., Marano, F., Magliarditi, D., Puglisi, G.,
Picerno, I., Di Pietro, A., Visalli, G., 2018. Brucellosis is a public health problem
in southern Italy: Burden and epidemiological trend of human and animal
disease. J. Infect. Public Health 11, 861–866.
- Fenwick, S.G., Madie, P., Wilks, C.R., 1994. Duration of carriage and transmission of
Yersinia enterocolitica biotype 4, serotype O:3 in dogs. Epidemiol. Infect. 113,
471–477.
- Flint, J.A., Van Duynhoven, Y.T., Angulo, F.J., DeLong, S.M., Braun, P., Kirk, M.,
Scallan, E., Fitzgerald, M., Adak, G.K., Sockett, P., Ellis, A., Hall, G., Gargouri,
N., Walke, H., Braam, P., 2005. Estimating the burden of acute gastroenteritis,
foodborne disease, and pathogens commonly transmitted by food: an
international review. Clin. Infect. Dis. 41, 698–704.
- Fong, I.W., 2017. Animals and Mechanisms of Disease Transmission, in: Fong, I.W.
(Ed.), Emerging Zoonoses: A Worldwide Perspective. Springer International
Publishing, Cham, pp. 15–38.
- Ford, L., Moffatt, C.R.M., Fearnley, E., Miller, M., Gregory, J., Sloan-Gardner, T.S.,
Polkinghorne, B.G., Bell, R., Franklin, N., Williamson, D.A., Glass, K., Kirk, M.D.,
2018. The Epidemiology of *Salmonella enterica* Outbreaks in Australia,
2001–2016. Frontiers in Sustainable Food Systems 2, 86.

- Fort, J., 2015. Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe. *J. R. Soc. Interface* 12.
<https://doi.org/10.1098/rsif.2015.0166>
- Foster, N., Lovell, M.A., Marston, K.L., Hulme, S.D., Frost, A.J., Bland, P., Barrow, P.A., 2003. Rapid protection of gnotobiotic pigs against experimental salmonellosis following induction of polymorphonuclear leukocytes by avirulent *Salmonella enterica*. *Infect. Immun.* 71, 2182–2191.
- Frantz, L.A.F., Haile, J., Lin, A.T., Scheu, A., Geörg, C., Benecke, N., Alexander, M., Linderholm, A., Mullin, V.E., Daly, K.G., Others, 2019. Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proceedings of the National Academy of Sciences* 116, 17231–17238.
- Fredriksson-Ahomaa, M., Stolle, A., Stephan, R., 2007. Prevalence of pathogenic *Yersinia enterocolitica* in pigs slaughtered at a Swiss abattoir. *Int. J. Food Microbiol.* 119, 207–212.
- Frey, K.G., Herrera-Galeano, J.E., Redden, C.L., Luu, T.V., Servetas, S.L., Mateczun, A.J., Mokashi, V.P., Bishop-Lilly, K.A., 2014. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genomics* 15, 96.
- Funk, J.A., Davies, P.R., Nichols, M.A., 2001. Longitudinal study of *Salmonella enterica* in growing pigs reared in multiple-site swine production systems. *Vet. Microbiol.* 83, 45–60.
- Gage, K.L., Montenieri, J.A., Thomas, R.E., 1994. THE ROLE OF PREDATORS IN THE ECOLOGY, EPIDEMIOLOGY, AND SURVEILLANCE OF PLAGUE IN THE UNITED STATES. *Proceedings of the Sixteenth Vertebrate Pest Conference* (1994).

- Gage, T.B., DeWitte, S., 2009. What do we know about the agricultural demographic transition? *Curr. Anthropol.* 50, 649–655.
- Gansauge, M.-T., Meyer, M., 2014. Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Res.* 24, 1543–1549.
- Gebreyes, W.A., Thakur, S., 2005. Multidrug-resistant *Salmonella enterica* serovar Muenchen from pigs and humans and potential interserovar transfer of antimicrobial resistance. *Antimicrob. Agents Chemother.* 49, 503–511.
- Ghasemzadeh, I., Namazi, S.H., 2015. Review of bacterial and viral zoonotic infections transmitted by dogs. *J. Med. Life* 8, 1–5.
- Gibbs, E.P.J., 2005. Emerging zoonotic epidemics in the interconnected global community. *Vet. Rec.* 157, 673–679.
- Gisel, J.J., Brumble, L.M., Johnson, M.M., 2010. *Bordetella bronchiseptica* pneumonia in a kidney-pancreas transplant patient after exposure to recently vaccinated dogs. *Transpl. Infect. Dis.* 12, 73–76.
- Giuliano, A., Meiring, T., Grant, A.J., Watson, P.J., 2015. Acute Hepatic Necrosis Caused by *Salmonella enterica* Serotype I 4,5,12:-:1,2 in a Dog. *J. Clin. Microbiol.* 53, 3674–3676.
- Gould, L.H., Pape, J., Ettestad, P., Griffith, K.S., Mead, P.S., 2008. Dog-associated risk factors for human plague. *Zoonoses Public Health* 55, 448–454.
- Grahek-Ogden, D., Schimmer, B., Cudjoe, K.S., Nygård, K., Kapperud, G., 2007. Outbreak of *Yersinia enterocolitica* serogroup O:9 infection and processed pork, Norway. *Emerg. Infect. Dis.* 13, 754–756.
- Gürtler, M., Alter, T., Kasimir, S., Linnebur, M., Fehlhaber, K., 2005. Prevalence of *Yersinia enterocolitica* in fattening pigs. *J. Food Prot.* 68, 850–854.
- Gu, W., Deng, X., Lee, M., Sucu, Y.D., Arevalo, S., Stryke, D., Federman, S., Gopez,

- A., Reyes, K., Zorn, K., Sample, H., Yu, G., Ishpuniani, G., Briggs, B., Chow, E.D., Berger, A., Wilson, M.R., Wang, C., Hsu, E., Miller, S., DeRisi, J.L., Chiu, C.Y., 2021. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* 27, 115–124.
- Hadjirin, N.F., Miller, E.L., Murray, G.G.R., Yen, P.L.K., Phuc, H.D., Wileman, T.M., Hernandez-Garcia, J., Williamson, S.M., Parkhill, J., Maskell, D.J., Zhou, R., Fittipaldi, N., Gottschalk, M., (Dan) Tucker, A.W., Hoa, N.T., Welch, J.J., Weinert, L.A., 2020. Linking phenotype, genotype and ecology: antimicrobial resistance in the zoonotic pathogen *Streptococcus suis*. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2020.05.05.078493>
- Hall, A.J., Cassiday, P.K., Bernard, K.A., Bolt, F., Steigerwalt, A.G., Bixler, D., Pawloski, L.C., Whitney, A.M., Iwaki, M., Baldwin, A., Dowson, C.G., Komiya, T., Takahashi, M., Hinrikson, H.P., Tondella, M.L., 2010. Novel *Corynebacterium diphtheriae* in domestic cats. *Emerg. Infect. Dis.* 16, 688–691.
- Hampson, K., Coudeville, L., Lembo, T., Sambo, M., Kieffer, A., Attlan, M., Barrat, J., Blanton, J.D., Briggs, D.J., Cleaveland, S., Costa, P., Freuling, C.M., Hiby, E., Knopf, L., Leanes, F., Meslin, F.-X., Metlin, A., Miranda, M.E., Müller, T., Nel, L.H., Recuenco, S., Rupprecht, C.E., Schumacher, C., Taylor, L., Vigilato, M.A.N., Zinsstag, J., Dushoff, J., Global Alliance for Rabies Control Partners for Rabies Prevention, 2015. Estimating the global burden of endemic canine rabies. *PLoS Negl. Trop. Dis.* 9, e0003709.
- Harbeck, M., Seifert, L., Hänsch, S., Wagner, D.M., Birdsell, D., Parise, K.L., Wiechmann, I., Grupe, G., Thomas, A., Keim, P., Zöller, L., Bramanti, B., Riehm, J.M., Scholz, H.C., 2013. *Yersinia pestis* DNA from skeletal remains from the 6(th) century AD reveals insights into Justinianic Plague. *PLoS Pathog.* 9,

e1003349.

- Harkins, K.M., Stone, A.C., 2015. Ancient pathogen genomics: insights into timing and adaptation. *J. Hum. Evol.* 79, 137–149.
- Harper, K.N., Zuckerman, M.K., Turner, B.L., Armelagos, G.J., 2013. Primates, Pathogens, and Evolution: A Context for Understanding Emerging Disease, in: Brinkworth, J.F., Pechenkina, K. (Eds.), *Primates, Pathogens, and Evolution*. Springer New York, New York, NY, pp. 389–409.
- Harper, P.A., Hornitzky, M.A., Rayward, D.G., 1990. Enterocolitis in pigs associated with *Yersinia pseudotuberculosis* infection. *Aust. Vet. J.* 67, 418–419.
- Harrus, S., Waner, T., 2011. Diagnosis of canine monocytotropic ehrlichiosis (*Ehrlichia canis*): an overview. *Vet. J.* 187, 292–296.
- Hasan, M.R., Rawat, A., Tang, P., Jithesh, P.V., Thomas, E., Tan, R., Tilley, P., 2016. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. *J. Clin. Microbiol.* 54, 919–927.
- Heredia, N., García, S., 2018. Animals as sources of food-borne pathogens: A review. *Anim Nutr* 4, 250–255.
- Hoffmann, S., Scallan, E., 2017. Chapter 2 - Epidemiology, Cost, and Risk Analysis of Foodborne Disease, in: Dodd, C.E.R., Aldsworth, T., Stein, R.A., Cliver, D.O., Riemann, H.P. (Eds.), *Foodborne Diseases (Third Edition)*. Academic Press, pp. 31–63.
- Holst, B.S., Löfqvist, K., Ernholm, L., Eld, K., Cedersmyg, M., Hallgren, G., 2012. The first case of *Brucella canis* in Sweden: background, case report and recommendations from a northern European perspective. *Acta Vet. Scand.* 54, 18.

- Hua, P., Yuhai, M., Shide, T., Yang, S., Bohai, W., Xiangrui, C., 2000. Canine ehrlichiosis caused simultaneously by *Ehrlichia canis* and *Ehrlichia platys*. *Microbiol. Immunol.* 44, 737–739.
- Imanishi, M., Rotstein, D.S., Reimschuessel, R., Schwensohn, C.A., Woody, D.H., Jr, Davis, S.W., Hunt, A.D., Arends, K.D., Achen, M., Cui, J., Zhang, Y., Denny, L.F., Phan, Q.N., Joseph, L.A., Tuite, C.C., Tataryn, J.R., Behravesh, C.B., 2014. Outbreak of *Salmonella enterica* serotype *Infantis* infection in humans linked to dry dog food in the United States and Canada, 2012. *J. Am. Vet. Med. Assoc.* 244, 545–553.
- Irwin, P.J., Parry, B.W., 1999. Streptococcal meningoencephalitis in a dog. *J. Am. Anim. Hosp. Assoc.* 35, 417–422.
- Iwami, S., Takeuchi, Y., Liu, X., 2009. Avian flu pandemic: Can we prevent it? *J. Theor. Biol.* 257, 181–190.
- Jackson, B.R., Griffin, P.M., Cole, D., Walsh, K.A., Chai, S.J., 2013. Outbreak-associated *Salmonella enterica* serotypes and food Commodities, United States, 1998-2008. *Emerg. Infect. Dis.* 19, 1239–1244.
- Johnson, C.A., Carter, T.D., Dunn, J.R., Baer, S.R., Schalow, M.M., Bellay, Y.M., Guerra, M.A., Frank, N.A., 2018. Investigation and characterization of *Brucella canis* infections in pet-quality dogs and associated human exposures during a 2007-2016 outbreak in Michigan. *J. Am. Vet. Med. Assoc.* 253, 322–336.
- Jones, B.A., Grace, D., Kock, R., Alonso, S., Rushton, J., Said, M.Y., McKeever, D., Mutua, F., Young, J., McDermott, J., Pfeiffer, D.U., 2013. Zoonosis emergence linked to agricultural intensification and environmental change. *Proc. Natl. Acad. Sci. U. S. A.* 110, 8399–8404.
- Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., Orlando, L., 2013.

- mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684.
- Jungersen, G., Sørensen, V., Giese, S.B., Stack, J.A., Riber, U., 2006. Differentiation between serological responses to *Brucella suis* and *Yersinia enterocolitica* serotype O[$\frac{1}{2}$]9 after natural or experimental infection in pigs. *Epidemiology & Infection* 134, 347–357.
- Kaczorek, E., Schulz, P., Małaczewska, J., Wójcik, R., Siwicki, A.K., Stopyra, A., Lew-Kojrys, S., Pomianowski, A., Hryniewicka, K., Mikulska-Skupień, E., 2016. Prevalence of respiratory pathogens detected in dogs with kennel cough in Poland. *Acta Vet. Brno* 85, 329–336.
- Karesh, W.B., Dobson, A., Lloyd-Smith, J.O., Lubroth, J., Dixon, M.A., Bennett, M., Aldrich, S., Harrington, T., Formenty, P., Loh, E.H., Machalaba, C.C., Thomas, M.J., Heymann, D.L., 2012. Ecology of zoonoses: natural and unnatural histories. *Lancet* 380, 1936–1945.
- Keenliside, J., 2013. Pandemic Influenza A H1N1 in Swine and Other Animals, in: Richt, J.A., Webby, R.J. (Eds.), *Swine Influenza*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 259–271.
- Kempf, I., Kerouanton, A., Bougeard, S., Nagard, B., Rose, V., Mourand, G., Osterberg, J., Denis, M., Bengtsson, B.O., 2017. *Campylobacter coli* in Organic and Conventional Pig Production in France and Sweden: Prevalence and Antimicrobial Resistance. *Front. Microbiol.* 8, 955.
- Key, F.M., Posth, C., Esquivel-Gomez, L.R., Hübner, R., Spyrou, M.A., Neumann, G.U., Furtwängler, A., Sabin, S., Burri, M., Wissgott, A., Lankapalli, A.K., Vågane, Å.J., Meyer, M., Nagel, S., Tukhbatova, R., Khokhlov, A., Chizhevsky, A., Hansen, S., Belinsky, A.B., Kalmykov, A., Kantorovich, A.R., Maslov, V.E.,

- Stockhammer, P.W., Vai, S., Zavattaro, M., Riga, A., Caramelli, D., Skeates, R., Beckett, J., Gradoli, M.G., Steuri, N., Hafner, A., Ramstein, M., Siebke, I., Lösch, S., Erdal, Y.S., Alikhan, N.-F., Zhou, Z., Achtman, M., Bos, K., Reinhold, S., Haak, W., Kühnert, D., Herbig, A., Krause, J., 2020. Emergence of human-adapted *Salmonella enterica* is linked to the Neolithization process. *Nat Ecol Evol* 4, 324–333.
- Key, F.M., Posth, C., Krause, J., Herbig, A., Bos, K.I., 2017. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet.* 33, 508–520.
- Kim, H.B., Isaacson, R.E., 2017. *Salmonella* in Swine: Microbiota Interactions. *Annu Rev Anim Biosci* 5, 43–63.
- Klous, G., Huss, A., Heederik, D.J.J., Coutinho, R.A., 2016. Human-livestock contacts and their relationship to transmission of zoonotic pathogens, a systematic review of literature. *One Health* 2, 65–76.
- Komnenou, A.A., Mylonakis, M.E., Kouti, V., Tendoma, L., Leontides, L., Skountzou, E., Dessiris, A., Koutinas, A.F., Ofri, R., 2007. Ocular manifestations of natural canine monocytic ehrlichiosis (*Ehrlichia canis*): a retrospective study of 90 cases. *Vet. Ophthalmol.* 10, 137–142.
- Kreizinger, Z., Foster, J.T., Rónai, Z., Sulyok, K.M., Wehmann, E., János, S., Gyuranecz, M., 2014. Genetic relatedness of *Brucella suis* biovar 2 isolates from hares, wild boars and domestic pigs. *Vet. Microbiol.* 172, 492–498.
- Krueger, W.S., Lucero, N.E., Brower, A., Heil, G.L., Gray, G.C., 2014. Evidence for unapparent *Brucella canis* infections among adults with occupational exposure to dogs. *Zoonoses Public Health* 61, 509–518.
- Lamm, C.G., Ferguson, A.C., Lehenbauer, T.W., Love, B.C., 2010. Streptococcal

- infection in dogs: a retrospective study of 393 cases. *Vet. Pathol.* 47, 387–395.
- Lam, M.M., Clarridge, J.E., 3rd, Young, E.J., Mizuki, S., 2007. The other group G *Streptococcus*: increased detection of *Streptococcus canis* ulcer infections in dog owners. *J. Clin. Microbiol.* 45, 2327–2329.
- Larson, G., Albarella, U., Dobney, K., Rowley-Conwy, P., Schibler, J., Tresset, A., Vigne, J.-D., Edwards, C.J., Schlumbaum, A., Dinu, A., Balaşescu, A., Dolman, G., Tagliacozzo, A., Manaseryan, N., Miracle, P., Van Wijngaarden-Bakker, L., Masseti, M., Bradley, D.G., Cooper, A., 2007. Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc. Natl. Acad. Sci. U. S. A.* 104, 15276–15281.
- Larson, G., Burger, J., 2013. A population genetics view of animal domestication. *Trends Genet.* 29, 197–205.
- Larson, G., Karlsson, E.K., Perri, A., Webster, M.T., Ho, S.Y.W., Peters, J., Stahl, P.W., Piper, P.J., Langaas, F., Fredholm, M., Comstock, K.E., Modiano, J.F., Schelling, C., AgoulNIK, A.I., Leegwater, P.A., Dobney, K., Vigne, J.-D., Vilà, C., Andersson, L., Lindblad-Toh, K., 2012. Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8878–8883.
- Laukkanen, R., Martínez, P.O., Siekkinen, K.-M., Ranta, J., Maijala, R., Korkeala, H., 2008. Transmission of *Yersinia pseudotuberculosis* in the pork production chain from farm to slaughterhouse. *Appl. Environ. Microbiol.* 74, 5444–5450.
- Lee, J.S., Mackie, R.S., Harrison, T., Shariat, B., Kind, T., Kehl, T., Löchelt, M., Boucher, C., VandeWoude, S., 2017. Targeted Enrichment for Pathogen Detection and Characterization in Three Felid Species. *J. Clin. Microbiol.* 55, 1658–1670.

- Leiva, M., Naranjo, C., Peña, M.T., 2005. Ocular signs of canine monocytic ehrlichiosis: a retrospective study in dogs from Barcelona, Spain. *Vet. Ophthalmol.* 8, 387–393.
- Little, S.E., Heise, S.R., Blagburn, B.L., Callister, S.M., Mead, P.S., 2010. Lyme borreliosis in dogs and humans in the USA. *Trends Parasitol.* 26, 213–218.
- Lord, K.A., Larson, G., Coppinger, R.P., Karlsson, E.K., 2020. The History of Farm Foxes Undermines the Animal Domestication Syndrome. *Trends Ecol. Evol.* 35, 125–136.
- Lowden, P., Wallis, C., Gee, N., Hilton, A., 2015. Investigating the prevalence of *Salmonella* in dogs within the Midlands region of the United Kingdom. *BMC Vet. Res.* 11, 239.
- Lucero, N.E., Corazza, R., Almuzara, M.N., Reynes, E., Escobar, G.I., Boeri, E., Ayala, S.M., 2010. Human *Brucella canis* outbreak linked to infection in dogs. *Epidemiol. Infect.* 138, 280–285.
- Lucero, N.E., Escobar, G.I., Ayala, S.M., Jacob, N., 2005. Diagnosis of human brucellosis caused by *Brucella canis*. *J. Med. Microbiol.* 54, 457–461.
- Lun, Z.-R., Wang, Q.-P., Chen, X.-G., Li, A.-X., Zhu, X.-Q., 2007. *Streptococcus suis*: an emerging zoonotic pathogen. *Lancet Infect. Dis.* 7, 201–209.
- Lynch, J.P., 3rd, Zhan, G.G., 2009. *Streptococcus pneumoniae*: epidemiology, risk factors, and strategies for prevention. *Semin. Respir. Crit. Care Med.* 30, 189–209.
- Manser, P.A., Dalziel, R.W., 1985. A survey of *Campylobacter* in animals. *J. Hyg.* 95, 15–21.
- Marois, C., Le Devendec, L., Gottschalk, M., Kobisch, M., 2007. Detection and molecular typing of *Streptococcus suis* in tonsils from live pigs in France. *Can. J.*

Vet. Res. 71, 14–22.

Martínez-Avilés, M., Garrido-Esteba, M., Álvarez, J., de la Torre, A., 2019.

Salmonella Surveillance Systems in Swine and Humans in Spain: A Review.

Vet. Sci. China 6. <https://doi.org/10.3390/vetsci6010020>

McQuiston, J.H., Guerra, M.A., Watts, M.R., Lawaczeck, E., Levy, C., Nicholson,

W.L., Adjemian, J., Swerdlow, D.L., 2011. Evidence of exposure to spotted fever group rickettsiae among Arizona dogs outside a previously documented outbreak area. *Zoonoses Public Health* 58, 85–92.

Mochizuki, M., Yachi, A., Ohshima, T., Ohuchi, A., Ishida, T., 2008. Etiologic study of upper respiratory infections of household dogs. *J. Vet. Med. Sci.* 70, 563–569.

Morley, P.S., Strohmeyer, R.A., Tankson, J.D., Hyatt, D.R., Dargatz, D.A.,

Fedorka-Cray, P.J., 2006. Evaluation of the association between feeding raw meat and *Salmonella enterica* infections at a Greyhound breeding facility. *J. Am. Vet. Med. Assoc.* 228, 1524–1532.

Murphy, B.P., Drummond, N., Ringwood, T., O'Sullivan, E., Buckley, J.F., Whyte, P.,

Prentice, M.B., Fanning, S., 2010. First report: *Yersinia enterocolitica* recovered from canine tonsils. *Vet. Microbiol.* 146, 336–339.

Muttaiyah, S., Best, E.J., Freeman, J.T., Taylor, S.L., Morris, A.J., Roberts, S.A.,

2011. *Corynebacterium diphtheriae* endocarditis: a case series and review of the treatment approach. *Int. J. Infect. Dis.* 15, e584–8.

Mylonakis, M.E., Harrus, S., Breitschwerdt, E.B., 2019. An update on the treatment of canine monocytic ehrlichiosis (*Ehrlichia canis*). *Vet. J.* 246, 45–53.

Nesbakken, T., Iversen, T., Lium, B., 2007. Pig herds free from human pathogenic *Yersinia enterocolitica*. *Emerg. Infect. Dis.* 13, 1860–1864.

Nichols, M.C., Ettestad, P.J., Vinhatton, E.S., Melman, S.D., Onischuk, L., Pierce,

- E.A., Aragon, A.S., 2014. *Yersinia pestis* infection in dogs: 62 cases (2003-2011). *J. Am. Vet. Med. Assoc.* 244, 1176–1180.
- Nicholson, W.L., Allen, K.E., McQuiston, J.H., Breitschwerdt, E.B., Little, S.E., 2010. The increasing recognition of rickettsial pathogens in dogs and people. *Trends Parasitol.* 26, 205–212.
- Niskanen, T., Fredriksson-Ahomaa, M., Korkeala, H., 2002. *Yersinia pseudotuberculosis* with limited genetic diversity is a common finding in tonsils of fattening pigs. *J. Food Prot.* 65, 540–545.
- Niskanen, T., Laukkanen, R., Fredriksson-Ahomaa, M., Korkeala, H., 2008. Distribution of *virF*/*lcrF*-positive *Yersinia pseudotuberculosis* serotype O:3 at farm level. *Zoonoses Public Health* 55, 214–221.
- Ogawa, Y., Ooka, T., Shi, F., Ogura, Y., Nakayama, K., Hayashi, T., Shimoji, Y., 2011. The genome of *Erysipelothrix rhusiopathiae*, the causative agent of swine erysipelas, reveals new insights into the evolution of firmicutes and the organism's intracellular adaptations. *J. Bacteriol.* 193, 2959–2971.
- Olsen, S.C., 2010. Brucellosis in the United States: role and significance of wildlife reservoirs. *Vaccine* 28 Suppl 5, F73–6.
- Olsen, S.C., Tatum, F.M., 2017. Swine brucellosis: current perspectives. *Vet Med (Auckl)* 8, 1–12.
- Osofsky, S.A., 2019. THE GLOBAL BURDEN OF (HOW WE MANAGE) ANIMAL DISEASE: LEARNING LESSONS FROM SOUTHERN AFRICA. *J. Wildl. Dis.* 55, 755–757.
- Oteo, J.A., Portillo, A., Santibáñez, S., Blanco, J.R., Pérez-Martínez, L., Ibarra, V., 2006. Cluster of cases of human *Rickettsia felis* infection from Southern Europe (Spain) diagnosed by PCR. *J. Clin. Microbiol.* 44, 2669–2671.

- Pearce-Duvel, J.M.C., 2006. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. *Biol. Rev. Camb. Philos. Soc.* 81, 369–382.
- Perri, A.R., Feuerborn, T.R., Frantz, L.A.F., Larson, G., Malhi, R.S., Meltzer, D.J., Witt, K.E., 2021. Dog domestication and the dual dispersal of people and dogs into the Americas. *Proc. Natl. Acad. Sci. U. S. A.* 118.
<https://doi.org/10.1073/pnas.2010083118>
- Petersen, L.R., Sawyer, L.A., Fishbein, D.B., Kelley, P.W., Thomas, R.J., Magnarelli, L.A., Redus, M., Dawson, J.E., 1989. An outbreak of ehrlichiosis in members of an Army Reserve unit exposed to ticks. *J. Infect. Dis.* 159, 562–568.
- Qin, S.-S., Wu, C.-M., Wang, Y., Jeon, B., Shen, Z.-Q., Wang, Y., Zhang, Q., Shen, J.-Z., 2011. Antimicrobial resistance in *Campylobacter coli* isolated from pigs in two provinces of China. *Int. J. Food Microbiol.* 146, 94–98.
- Rasmussen, S., Allentoft, M.E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., Pedersen, A.G., Schubert, M., Van Dam, A., Kapel, C.M.O., Nielsen, H.B., Brunak, S., Avetisyan, P., Epimakhov, A., Khalyapin, M.V., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., Moiseyev, V., Gromov, A., Pokutta, D., Saag, L., Varul, L., Yepiskoposyan, L., Sicheritz-Pontén, T., Foley, R.A., Lahr, M.M., Nielsen, R., Kristiansen, K., Willerslev, E., 2015. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* 163, 571–582.
- Rojas, P., Monahan, A.M., Schuller, S., Miller, I.S., Markey, B.K., Nally, J.E., 2010. Detection and quantification of leptospires in urine of dogs: a maintenance host for the zoonotic disease leptospirosis. *Eur. J. Clin. Microbiol. Infect. Dis.* 29, 1305–1309.
- Roman-Binois, A., 2017. L'archéologie des épizooties : mise en évidence et

- diagnostic des crises de mortalité chez les animaux d'élevage, du Néolithique à Pasteur.
- Rostagno, M.H., Eicher, S.D., Lay, D.C., Jr, 2011. Immunological, physiological, and behavioral effects of *Salmonella enterica* carriage and shedding in experimentally infected finishing pigs. *Foodborne Pathog. Dis.* 8, 623–630.
- Schlaberg, R., Chiu, C.Y., Miller, S., Procop, G.W., Weinstock, G., Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology, Microbiology Resource Committee of the College of American Pathologists, 2017. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Arch. Pathol. Lab. Med.* 141, 776–786.
- Schrenzel, M.D., Tucker, T.A., Stalis, I.H., Kagan, R.A., Burns, R.P., Denison, A.M., Drew, C.P., Paddock, C.D., Rideout, B.A., 2011. Pandemic (H1N1) 2009 virus in 3 wildlife species, San Diego, California, USA. *Emerg. Infect. Dis.* 17, 747–749.
- Schubert, M., Lindgreen, S., Orlando, L., 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9, 88.
- Schultsz, C., Jansen, E., Keijzers, W., Rothkamp, A., Duim, B., Wagenaar, J.A., van der Ende, A., 2012. Differences in the population structure of invasive *Streptococcus suis* strains isolated from pigs and from humans in The Netherlands. *PLoS One* 7, e33854.
- Schulz, B.S., Kurz, S., Weber, K., Balzer, H.-J., Hartmann, K., 2014. Detection of respiratory viruses and *Bordetella bronchiseptica* in dogs with acute respiratory tract infections. *Vet. J.* 201, 365–369.
- Schulze, T.L., McDevitt, W.M., Parkin, W.E., Shisler, J.K., 1987. Effectiveness of two insecticides in controlling *Ixodes dammini* (Acari: Ixodidae) following an

- outbreak of Lyme disease in New Jersey. *J. Med. Entomol.* 24, 420–424.
- Shapiro, B., Ho, S.Y.W., Drummond, A.J., Suchard, M.A., Pybus, O.G., Rambaut, A., 2011. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* 28, 879–887.
- Shchelkunov, S.N., 2013. An increasing danger of zoonotic orthopoxvirus infections. *PLoS Pathog.* 9, e1003756.
- Shimoji, Y., Osaki, M., Ogawa, Y., Shiraiwa, K., Nishikawa, S., Eguchi, M., Yamamoto, T., Tsutsui, T., 2019. Wild boars: A potential source of *Erysipelothrix rhusiopathiae* infection in Japan. *Microbiol. Immunol.* 63, 465–468.
- Sing, A., Konrad, R., Meinel, D.M., Mauder, N., Schwabe, I., Sting, R., 2016. *Corynebacterium diphtheriae* in a free-roaming red fox: case report and historical review on diphtheria in animals. *Infection* 44, 441–445.
- Skotarczak, B., Wodecka, B., Rymaszewska, A., Sawczuk, M., Maciejewska, A., Adamska, M., Hermanowska-Szpakowicz, T., Swierzbinska, R., 2005. Prevalence of DNA and antibodies to *Borrelia burgdorferi sensu lato* in dogs suspected of borreliosis. *Ann. Agric. Environ. Med.* 12, 199.
- Smith, F.D., Ballantyne, R., Morgan, E.R., Wall, R., 2012. Estimating Lyme disease risk using pet dogs as sentinels. *Comp. Immunol. Microbiol. Infect. Dis.* 35, 163–167.
- Smith, H.E., Buijs, H., de Vries, R., Wisselink, H.J., Stockhofe-Zurwieden, N., Smits, M.A., 2001. Environmentally regulated genes of *Streptococcus suis*: identification by the use of iron-restricted conditions in vitro and by experimental infection of piglets. *Microbiology* 147, 271–280.
- Smith, L.M., Hartmann, S., Munteanu, A.M., Dalla Villa, P., Quinnell, R.J., Collins, L.M., 2019. The Effectiveness of Dog Population Management: A Systematic

- Review. *Animals (Basel)* 9. <https://doi.org/10.3390/ani9121020>
- Speck, S., Reiner, B., Streich, W.J., Reusch, C., Wittenbrink, M.M., 2007. Canine borreliosis: a laboratory diagnostic trial. *Vet. Microbiol.* 120, 132–141.
- Spyrou, M.A., Bos, K.I., Herbig, A., Krause, J., 2019. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* 20, 323–340.
- Spyrou, M.A., Tukhbatova, R.I., Feldman, M., Drath, J., Kacki, S., Beltrán de Heredia, J., Arnold, S., Sitdikov, A.G., Castex, D., Wahl, J., Gazimzyanov, I.R., Nurgaliev, D.K., Herbig, A., Bos, K.I., Krause, J., 2016. Historical *Y. pestis* Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. *Cell Host Microbe* 19, 874–881.
- Sriskandan, S., Slater, J.D., 2006. Invasive disease and toxic shock due to zoonotic *Streptococcus suis*: an emerging infection in the East? *PLoS Med.*
- Staats, J.J., Feder, I., Okwumabua, O., Chengappa, M.M., 1997. *Streptococcus suis*: past and present. *Vet. Res. Commun.* 21, 381–407.
- Stamm, I., Hailer, M., Depner, B., Kopp, P.A., Rau, J., 2013. *Yersinia enterocolitica* in diagnostic fecal samples from European dogs and cats: identification by fourier transform infrared spectroscopy and matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.* 51, 887–893.
- Stevens, D.L., Bryant, A.E., 2016. Impetigo, Erysipelas and Cellulitis, in: Ferretti, J.J., Stevens, D.L., Fischetti, V.A. (Eds.), *Streptococcus Pyogenes : Basic Biology to Clinical Manifestations*. University of Oklahoma Health Sciences Center, Oklahoma City (OK).
- Stone, A.C., 2020. Getting sick in the Neolithic. *Nat Ecol Evol.*
- Stone, A.C., Ozga, A.T., 2019. Chapter 8 - Ancient DNA in the Study of Ancient Disease, in: Buikstra, J.E. (Ed.), *Ortner's Identification of Pathological*

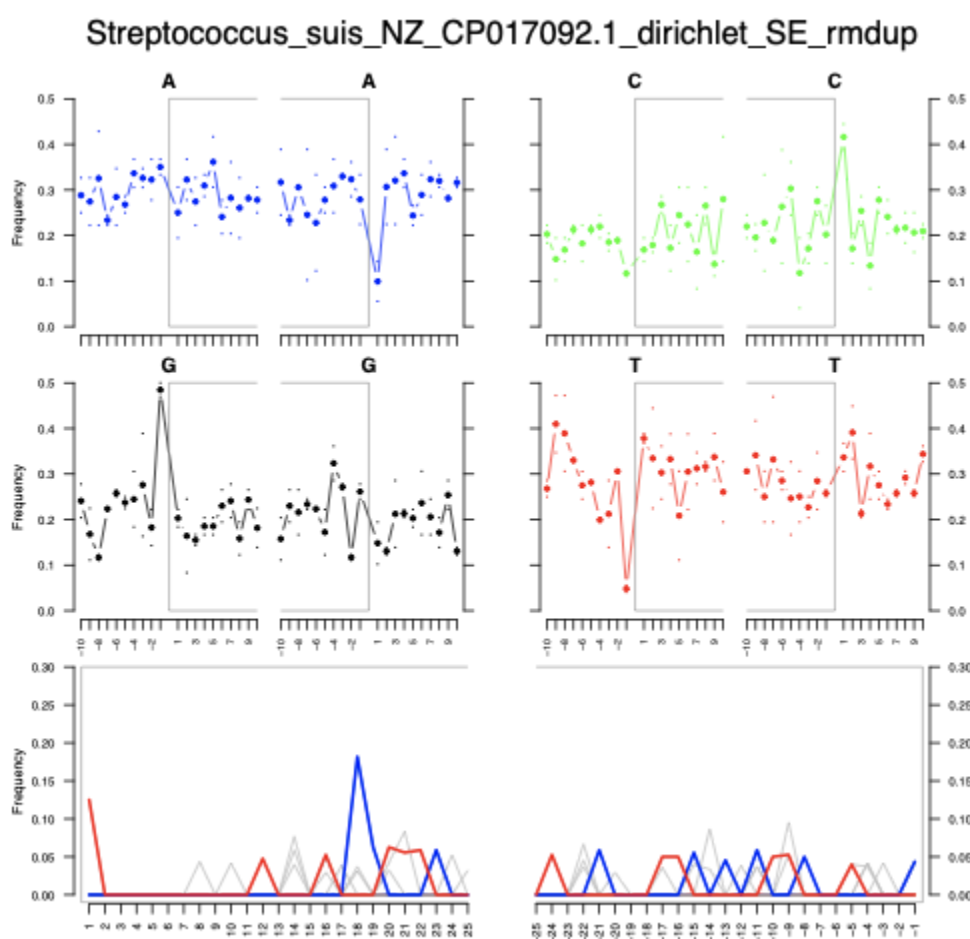
- Conditions in Human Skeletal Remains (Third Edition). Academic Press, San Diego, pp. 183–210.
- Stone, A.C., Wilbur, A.K., Buikstra, J.E., Roberts, C.A., 2009. Tuberculosis and leprosy in perspective. *Am. J. Phys. Anthropol.* 140 Suppl 49, 66–94.
- Straubinger, R.K., Summers, B.A., Chang, Y.F., Appel, M.J., 1997. Persistence of *Borrelia burgdorferi* in experimentally infected dogs after antibiotic treatment. *J. Clin. Microbiol.* 35, 111–116.
- Strohmeyer, R.A., Morley, P.S., Hyatt, D.R., Dargatz, D.A., Scorza, A.V., Lappin, M.R., 2006. Evaluation of bacterial and protozoal contamination of commercially available raw meat diets for dogs. *J. Am. Vet. Med. Assoc.* 228, 537–542.
- Tang, J., Wang, C., Feng, Y., Yang, W., Song, H., Chen, Z., Yu, H., Pan, X., Zhou, X., Wang, H., Wu, B., Wang, H., Zhao, H., Lin, Y., Yue, J., Wu, Z., He, X., Gao, F., Khan, A.H., Wang, J., Zhao, G.-P., Wang, Y., Wang, X., Chen, Z., Gao, G.F., 2006. Streptococcal toxic shock syndrome caused by *Streptococcus suis* serotype 2. *PLoS Med.* 3, e151.
- Tarabichi, M., Alvand, A., Shohat, N., Goswami, K., Parvizi, J., 2018. Diagnosis of *Streptococcus canis* periprosthetic joint infection: the utility of next-generation sequencing. *Arthroplast Today* 4, 20–23.
- Thakur, S., Gebreyes, W.A., 2005. Prevalence and antimicrobial resistance of *Campylobacter* in antimicrobial-free and conventional pig production systems. *J. Food Prot.* 68, 2402–2410.
- Thomas, R., 2005. Zooarchaeology, Improvement and the British Agricultural Revolution. *Int. J. Hist. Archaeol.* 9, 71–88.
- Thorburn, F., Bennett, S., Modha, S., Murdoch, D., Gunson, R., Murcia, P.R., 2015. The use of next generation sequencing in the diagnosis and typing of respiratory

- infections. *J. Clin. Virol.* 69, 96–100.
- 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., Bos, K.I., Krause, J., 2018. *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat Ecol Evol* 2, 520–528.
- Van Damme, I., Habib, I., De Zutter, L., 2010. *Yersinia enterocolitica* in slaughter pig tonsils: enumeration and detection by enrichment versus direct plating culture. *Food Microbiol.* 27, 158–161.
- van der Meer, F.J.U.M., Orsel, K., Barkema, H.W., 2010. The new influenza A H1N1 virus: balancing on the interface of humans and animals. *Can. Vet. J.* 51, 56–62.
- van Winsen, R.L., van Nes, A., Keuzenkamp, D., Urlings, H.A., Lipman, L.J., Biesterveld, S., Snijders, J.M., Verheijden, J.H., van Knapen, F., 2001. Monitoring of transmission of *Salmonella enterica* serovars in pigs using bacteriological and serological detection methods. *Vet. Microbiol.* 80, 267–274.
- Verraes, C., Van Boxtael, S., Van Meervenne, E., Van Coillie, E., Butaye, P., Catry, B., de Schaetzen, M.-A., Van Huffel, X., Imberechts, H., Dierick, K., Daube, G., Saegerman, C., De Block, J., Dewulf, J., Herman, L., 2013. Antimicrobial resistance in the food chain: a review. *Int. J. Environ. Res. Public Health* 10, 2643–2669.
- Vogt, N.A., Sargeant, J.M., MacKinnon, M.C., Versluis, A.M., 2019. Efficacy of *Borrelia burgdorferi* vaccine in dogs in North America: A systematic review and meta-analysis. *J. Vet. Intern. Med.* 33, 23–36.
- Wang, Q., Chang, B.J., Riley, T.V., 2010. *Erysipelothrix rhusiopathiae*. *Vet. Microbiol.* 140, 405–417.
- Wang, X., Cui, Z., Wang, H., Tang, L., Yang, J., Gu, L., Jin, D., Luo, L., Qiu, H., Xiao,

- Y., Xiong, H., Kan, B., Xu, J., Jing, H., 2010. Pathogenic strains of *Yersinia enterocolitica* isolated from domestic dogs (*Canis familiaris*) belonging to farmers are of the same subtype as pathogenic *Y. enterocolitica* strains isolated from humans and may be a source of human infection in Jiangsu Province, China. *J. Clin. Microbiol.* 48, 1604–1610.
- Webster, R., Hulse, D., 2005. Controlling avian flu at the source. *Nature* 435, 415–416.
- Wei, L., Kelly, P., Ackerson, K., El-Mahallawy, H.S., Kaltenboeck, B., Wang, C., 2014. Molecular detection of *Dirofilaria immitis*, *Hepatozoon canis*, *Babesia* spp., *Anaplasma platys* and *Ehrlichia canis* in dogs on Costa Rica. *Acta Parasitol.* 60, 21–25.
- Wei, Z., Li, R., Zhang, A., He, H., Hua, Y., Xia, J., Cai, X., Chen, H., Jin, M., 2009. Characterization of *Streptococcus suis* isolates from the diseased pigs in China between 2003 and 2007. *Vet. Microbiol.* 137, 196–201.
- Wertheim, H.F.L., Nghia, H.D.T., Taylor, W., Schultsz, C., 2009. *Streptococcus suis*: an emerging human pathogen. *Clin. Infect. Dis.* 48, 617–625.
- Wilcox, B.A., Gubler, D.J., 2005. Disease ecology and the global emergence of zoonotic pathogens. *Environ. Health Prev. Med.* 10, 263–272.
- WOOD, L, R., 1999. Erysipelas. *Diseases of Swine* 419–430.
- Woolhouse, M.E., Taylor, L.H., Haydon, D.T., 2001. Population biology of multihost pathogens. *Science* 292, 1109–1112.
- Woolhouse, M., Gaunt, E., 2007. Ecological origins of novel human pathogens. *Crit. Rev. Microbiol.* 33, 231–242.
- Wylezich, C., Papa, A., Beer, M., Höper, D., 2018. A Versatile Sample Processing Workflow for Metagenomic Pathogen Detection. *Sci. Rep.* 8, 13108.

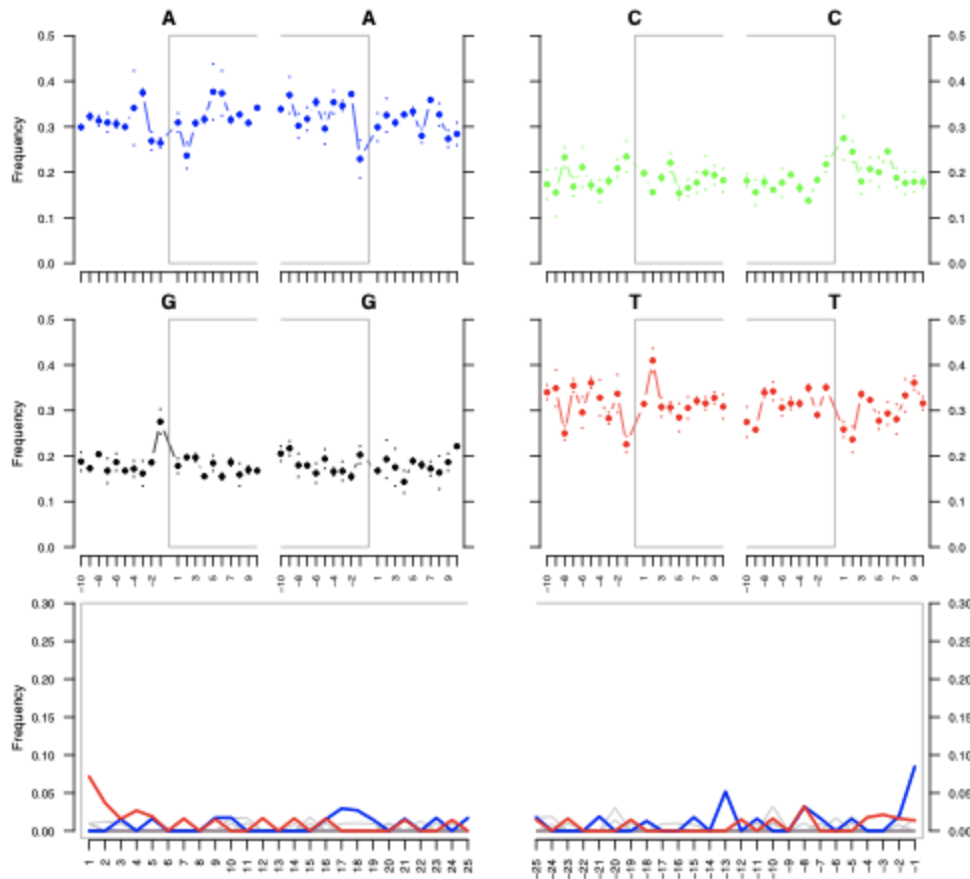
- Xu, R., Ekiert, D.C., Krause, J.C., Hai, R., Crowe, J.E., Jr, Wilson, I.A., 2010. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science* 328, 357–360.
- Ye, S.H., Siddle, K.J., Park, D.J., Sabeti, P.C., 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* 178, 779–794.
- Zeder, M.A., 2012. The Domestication of Animals. *J. Anthropol. Res.* 68, 161–190.
- Zeder, M.A., 2006. Central questions in the domestication of plants and animals. *Evol. Anthropol.* 15, 105–117.
- 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., Hamre, S.S., Casjens, S., Christophersen, A., Quince, C., Thomson, N.R., Weill, F.-X., Ho, S.Y.W., Gilbert, M.T.P., Achtman, M., 2018. Pan-genome Analysis of Ancient and Modern *Salmonella enterica* Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr. Biol.* 28, 2420–2428.e10.

Supplementary



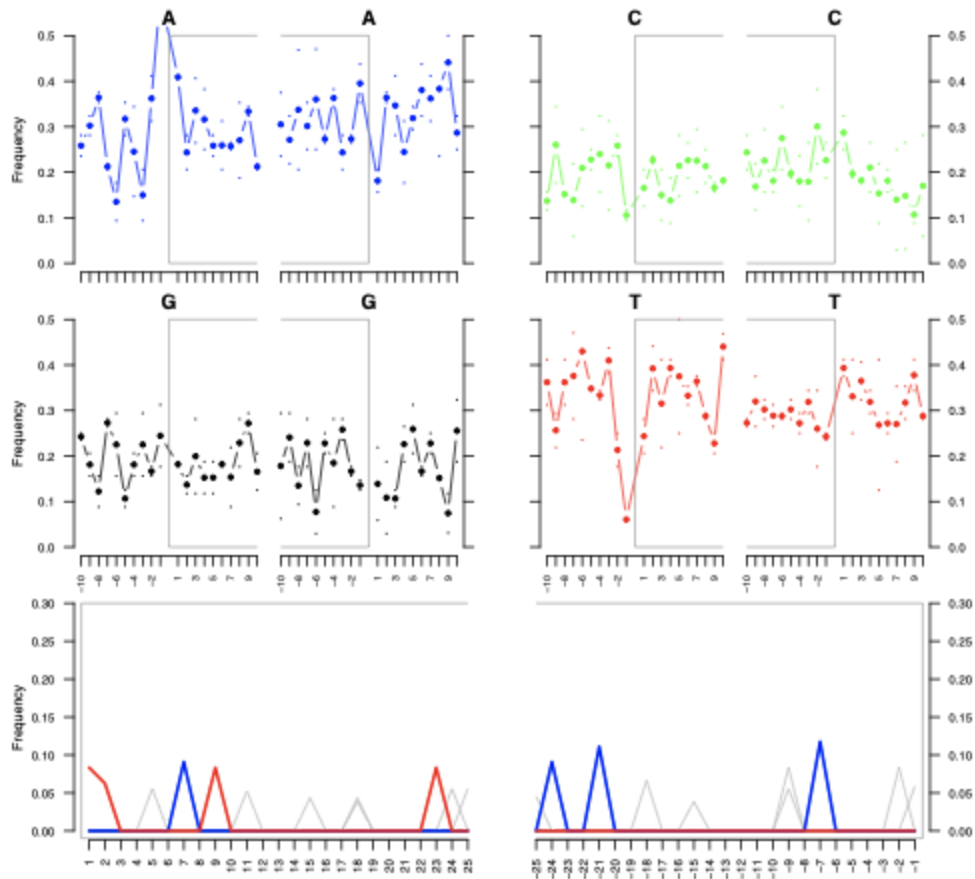
Supplementary figure 1. Deamination damage profile for sample AA276 for *Streptococcus suis*.

Erysipelothrix_rhusiopathiae_NZ_CP041995.1_dirichlet_SE_rmdup

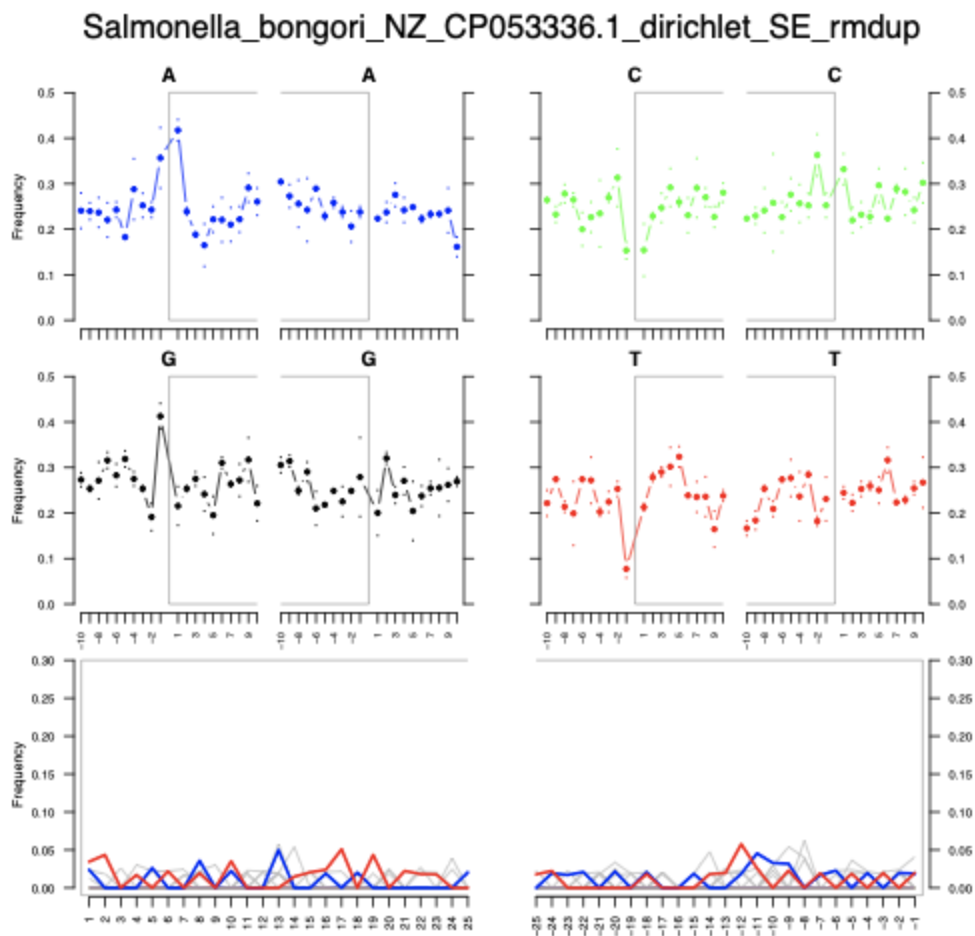


Supplementary figure 2. Deamination damage profile for sample AA401 for *Erysipelothrix rhusiopathiae*.

*Erysipelothrix rhusiopathiae*_NZ_CP041995.1_dirichlet_SE_rmdup

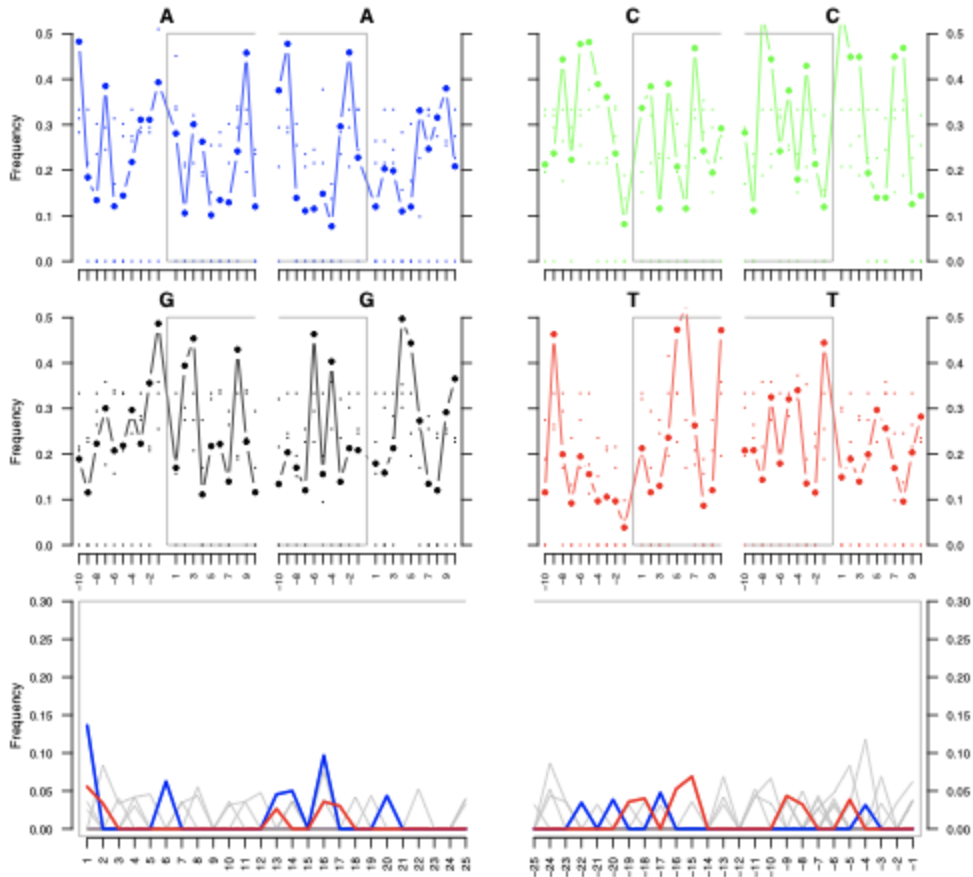


Supplementary figure 3. Deamination damage profile for sample AL725 for *Erysipelothrix rhusiopathiae*.

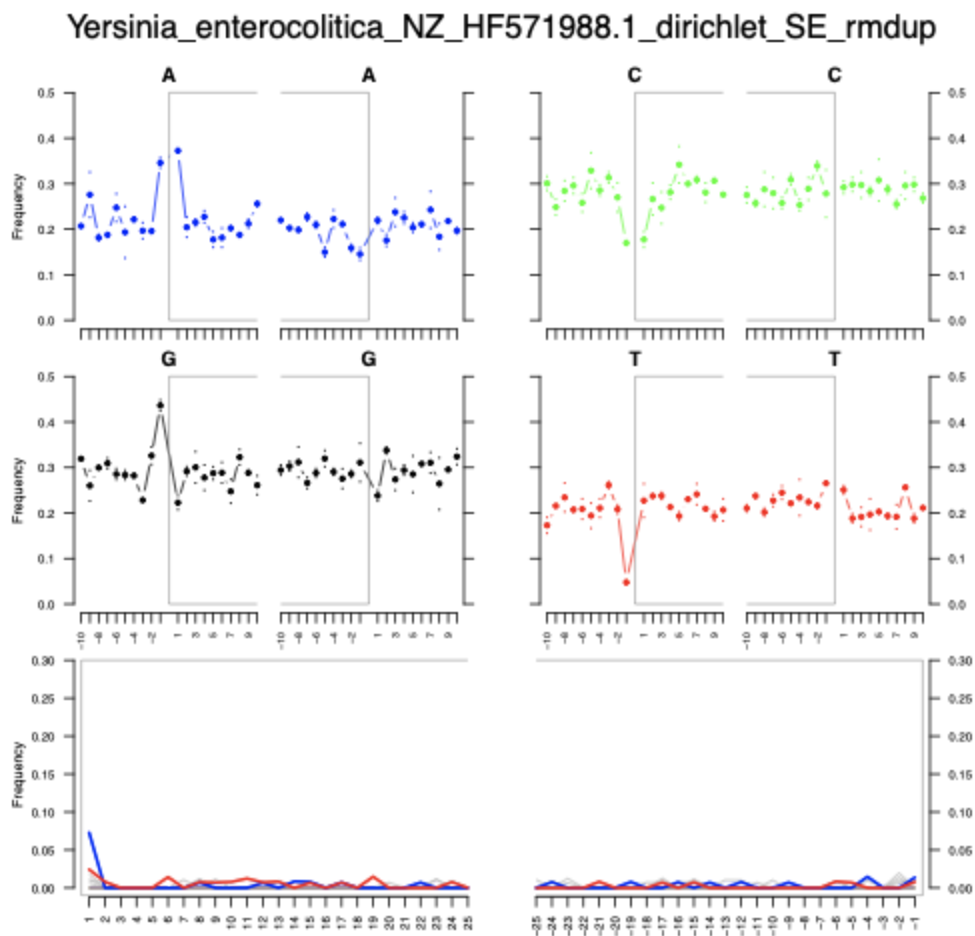


Supplementary figure 4. Deamination damage profile for sample AL1031 for *Salmonella bongori*.

Salmonella_enterica_NZ_CP030219.1_dirichlet_SE_rmdup

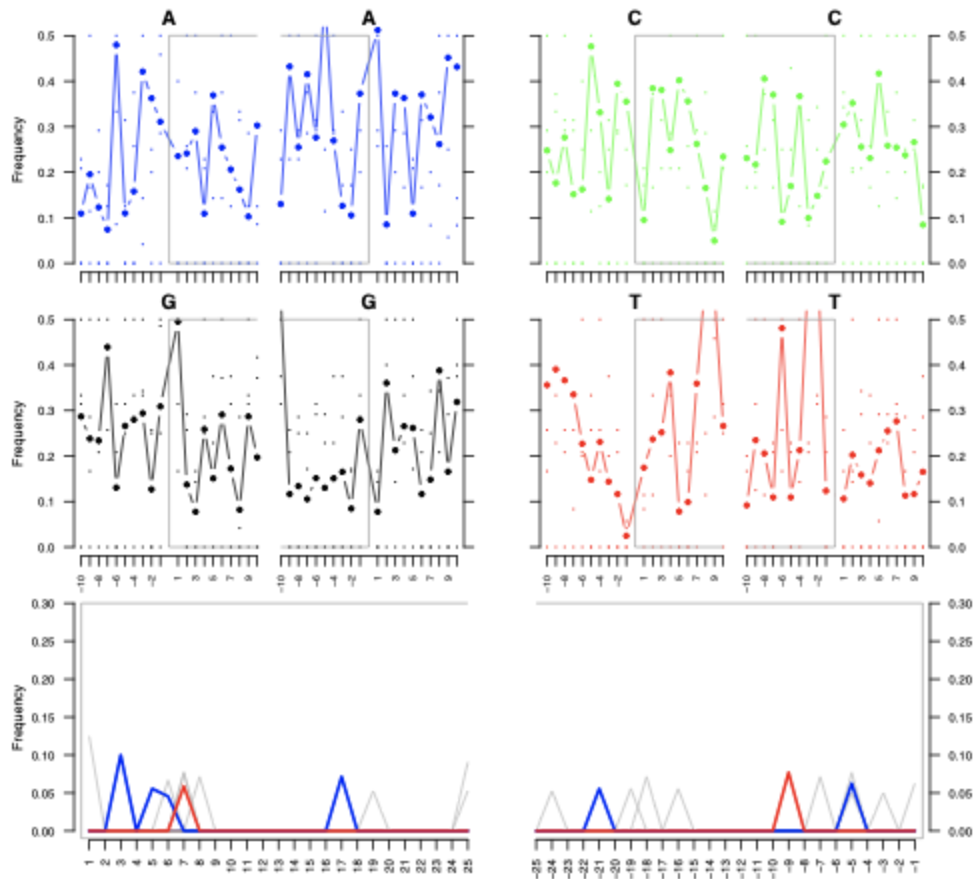


Supplementary figure 5. Deamination damage profile for sample AL1031 for *Salmonella enterica*.



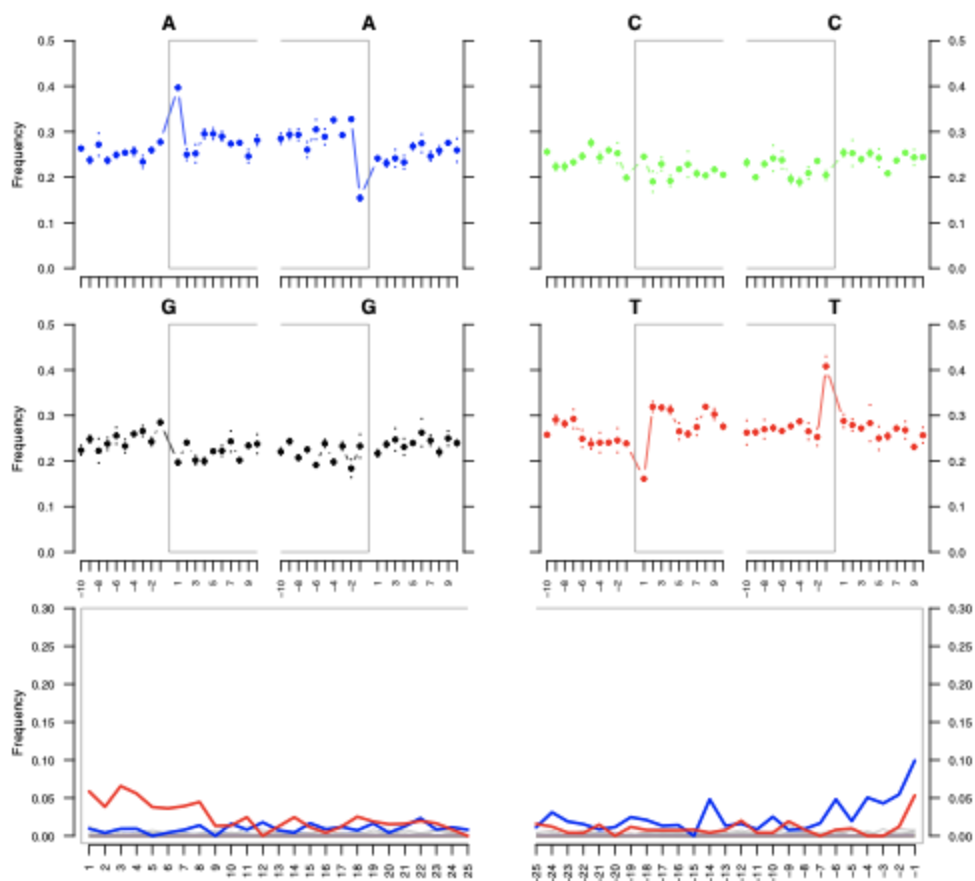
Supplementary figure 6. Deamination damage profile for sample AL1031 for *Yersinia enterocolitica*.

Yersinia_hibernica_NZ_CP032487.1_dirichlet_SE_rmdup

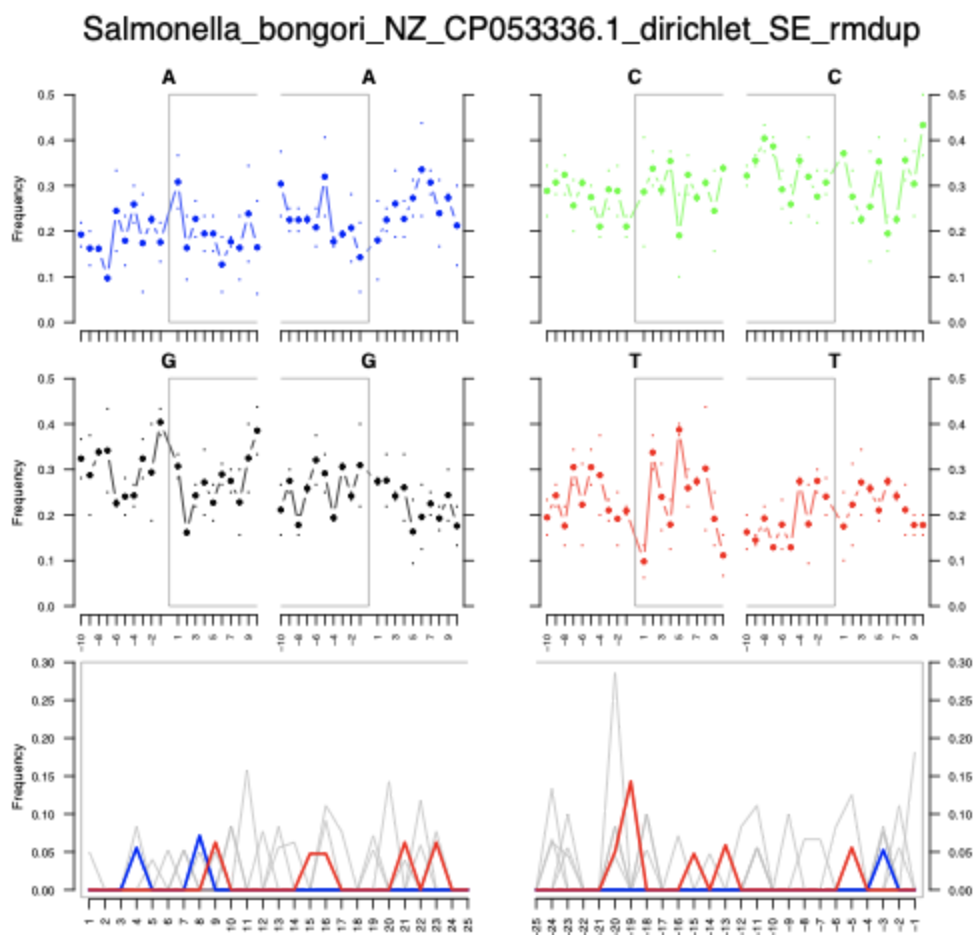


Supplementary figure 7. Deamination damage profile for sample AL1031 for *Yersinia hibernica*.

*Yersinia enterocolitica*_NZ_HF571988.1_dirichlet_SE_rmdup

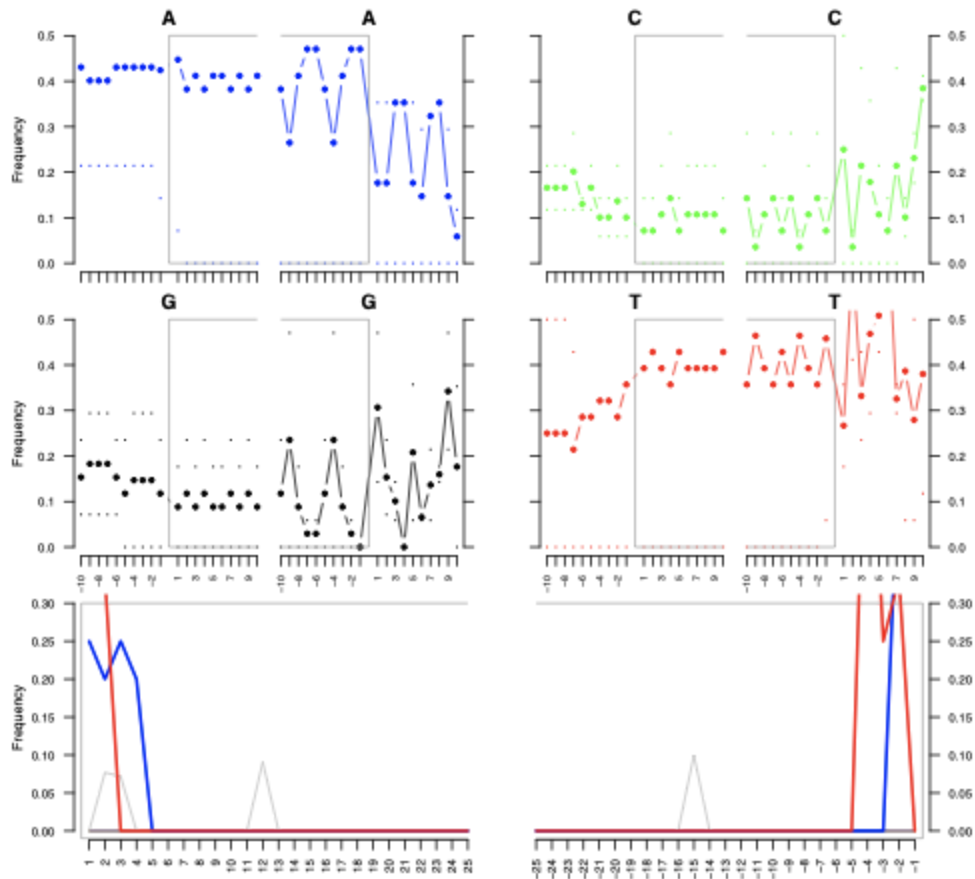


Supplementary figure 8. Deamination damage profile for sample OL4351 for *Yersinia enterocolitica*.



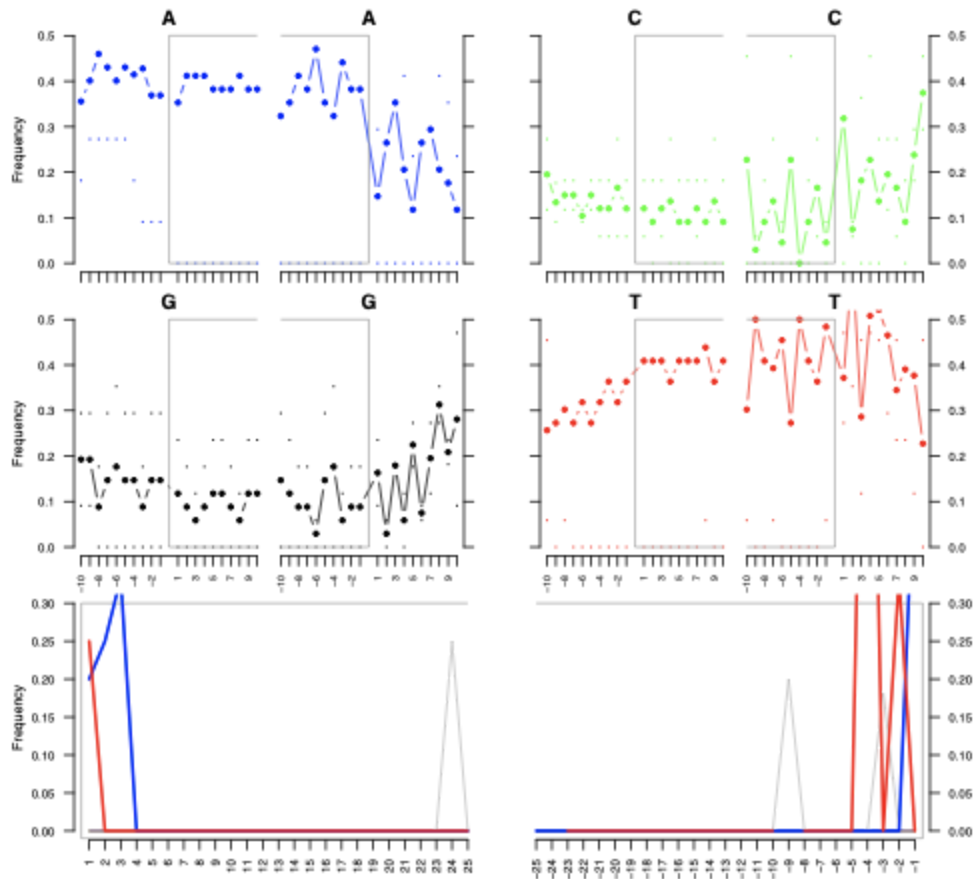
Supplementary figure 9. Deamination damage profile for sample AC05 for *Salmonella bongori*.

*Yersinia enterocolitica*_NZ_HF571988.1_dirichlet_SE_rmdup

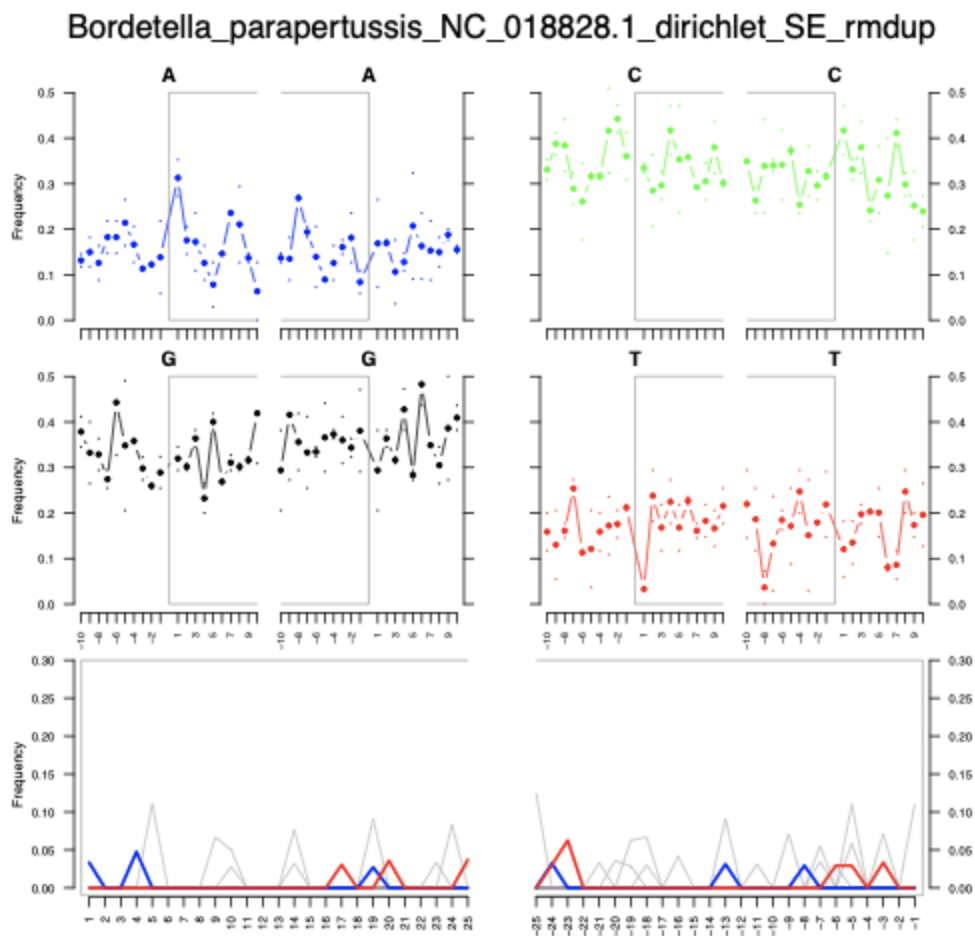


Supplementary figure 10. Deamination damage profile for sample AL3163 for *Yersinia enterocolitica*.

*Yersinia enterocolitica*_NZ_HF571988.1_dirichlet_SE_rmdup

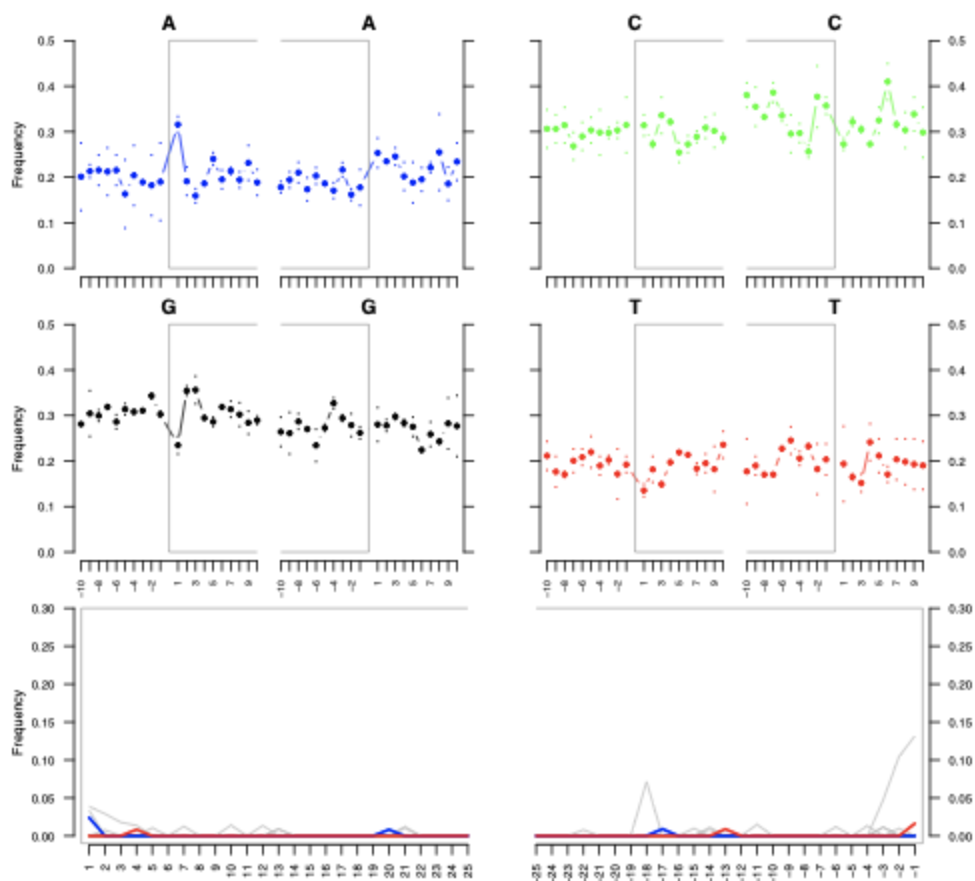


Supplementary figure 11. Deamination damage profile for sample AL3268 for *Yersinia enterocolitica*.



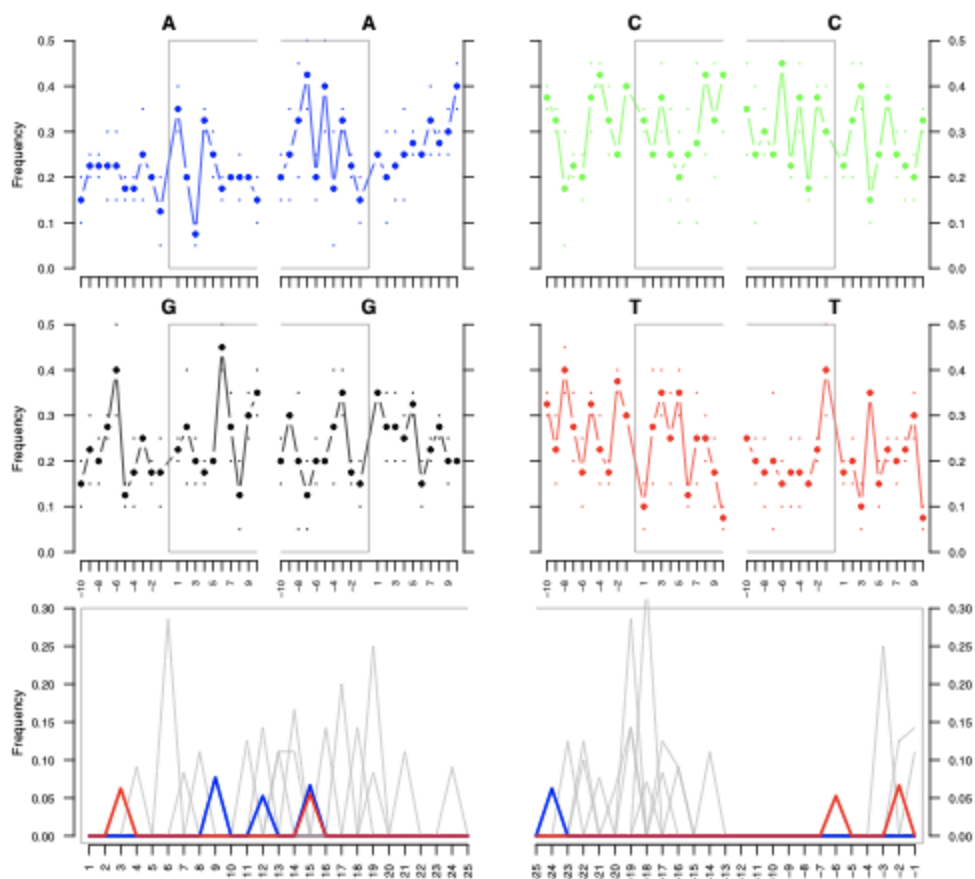
Supplementary figure 12. Deamination damage profile for sample ANH-1 for *Bordetella parapertussis*.

*Corynebacterium diphtheriae*_NZ_CP029644.1_dirichlet_SE_rmdup

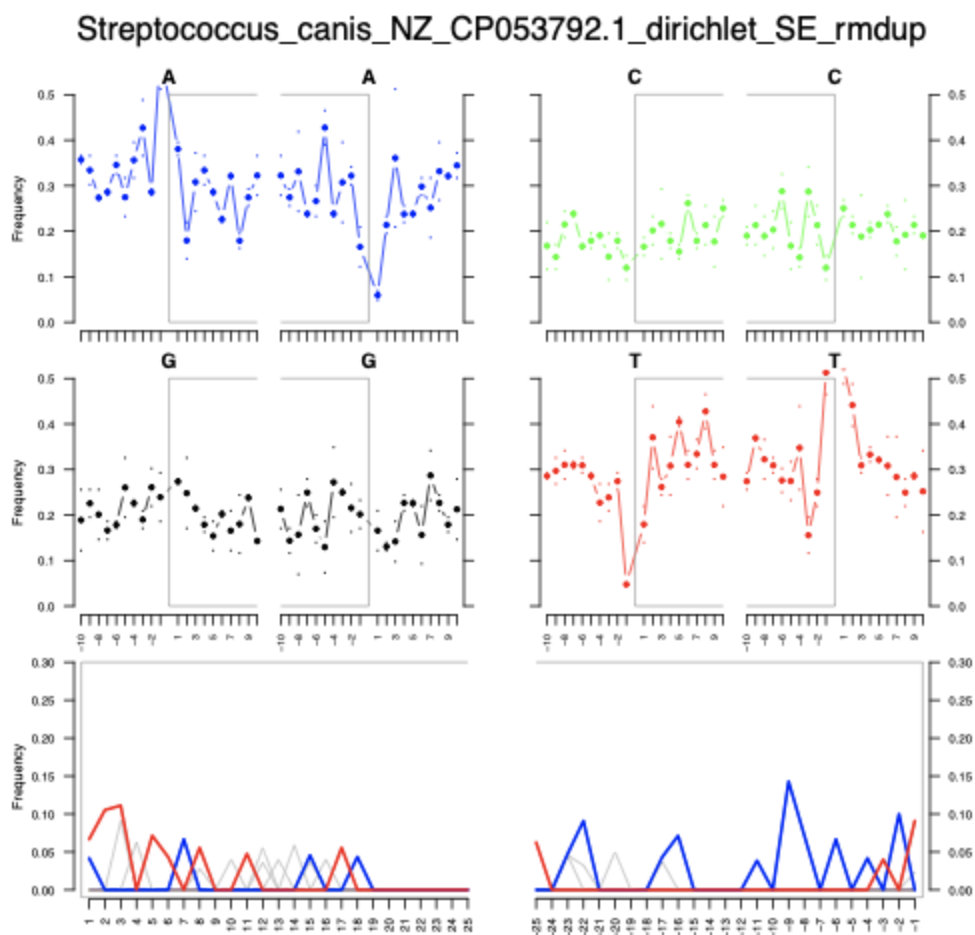


Supplementary figure 12. Deamination damage profile for sample ANH-1 for *Corynebacterium diphtheriae*.

Salmonella_bongori_NZ_CP053336.1_dirichlet_SE_rmdup

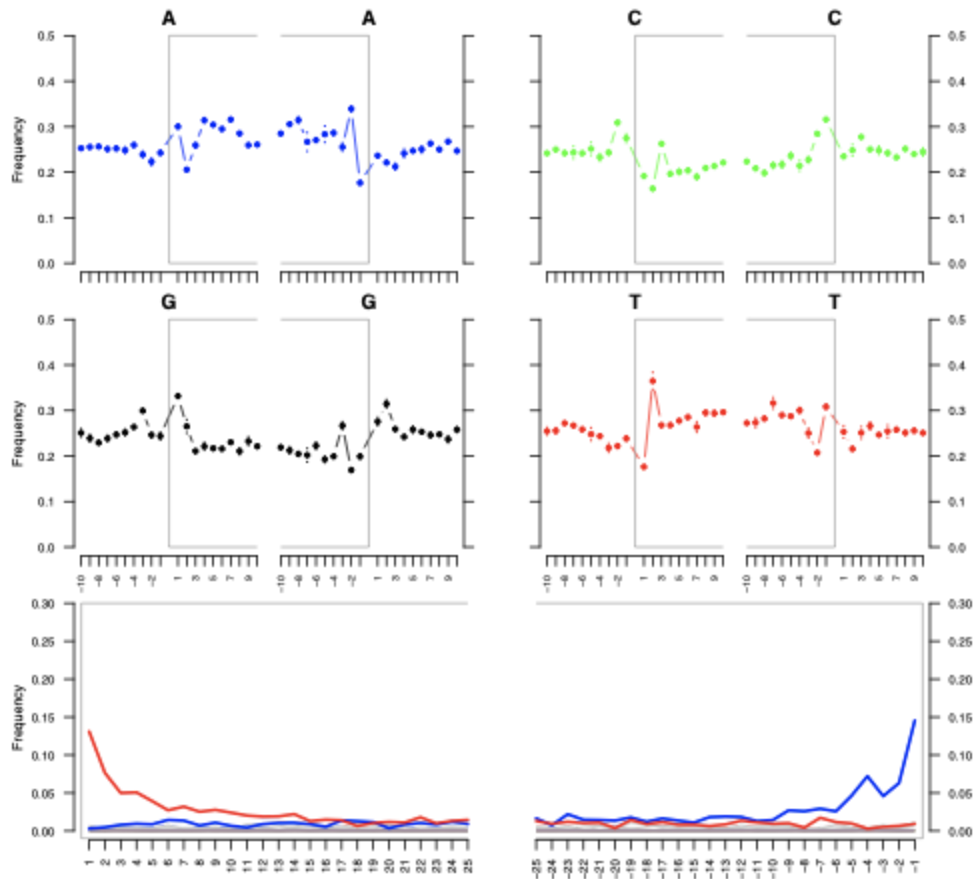


Supplementary figure 13. Deamination damage profile for sample ANH-1 for *Salmonella bongori*.



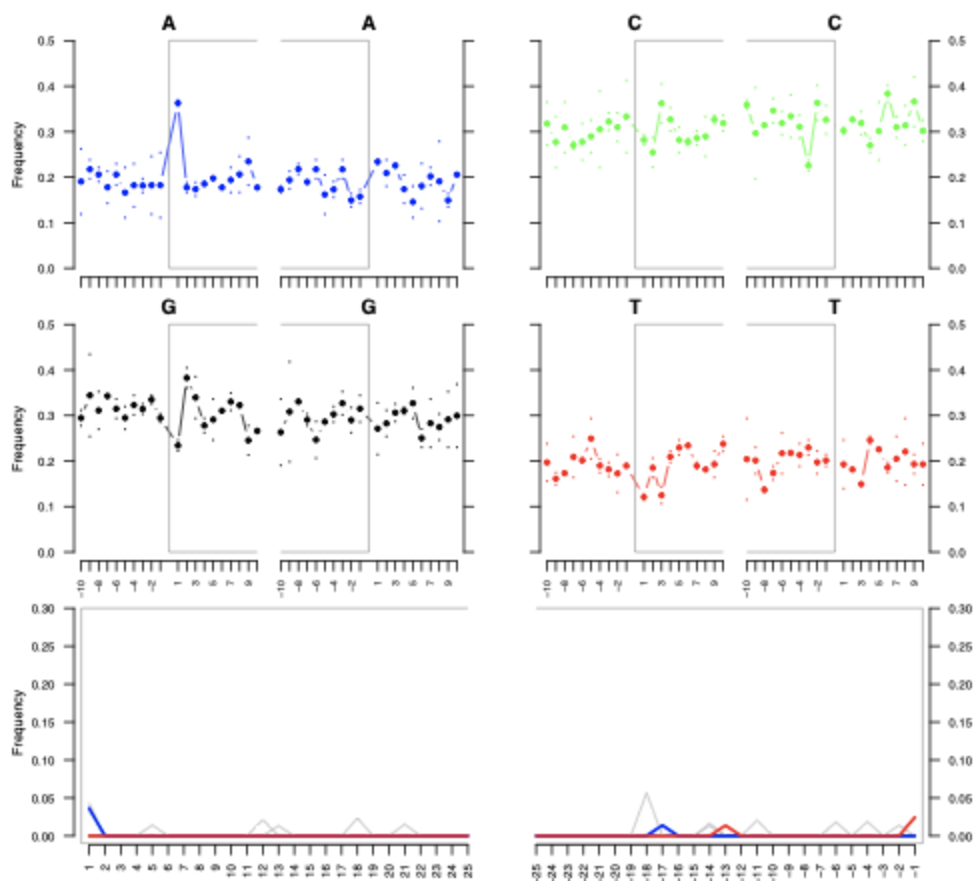
Supplementary figure 14. Deamination damage profile for sample ANH-1 for *Streptococcus canis*.

*Yersinia_enterocolitica*_NZ_HF571988.1_dirichlet_SE_rmdup



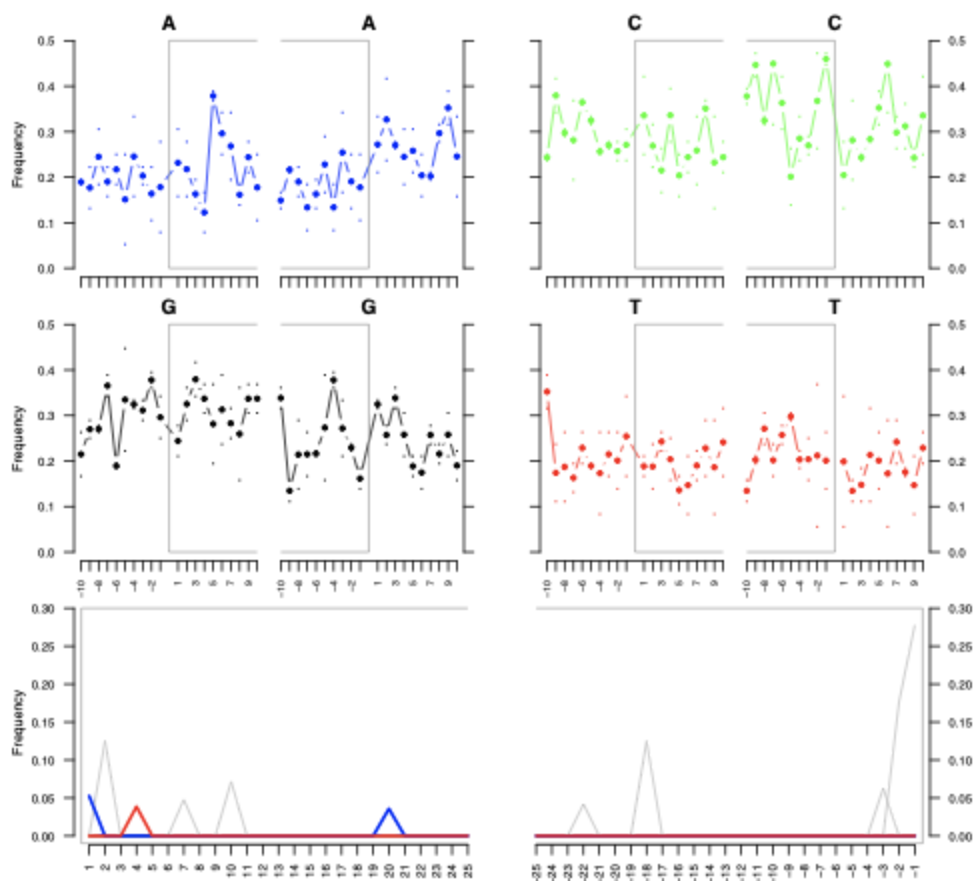
Supplementary figure 15. Deamination damage profile for sample ANH-1 for *Yersinia enterocolitica*.

*Corynebacterium_diphtheriae*_NZ_CP029644.1_dirichlet_SE_rmdup

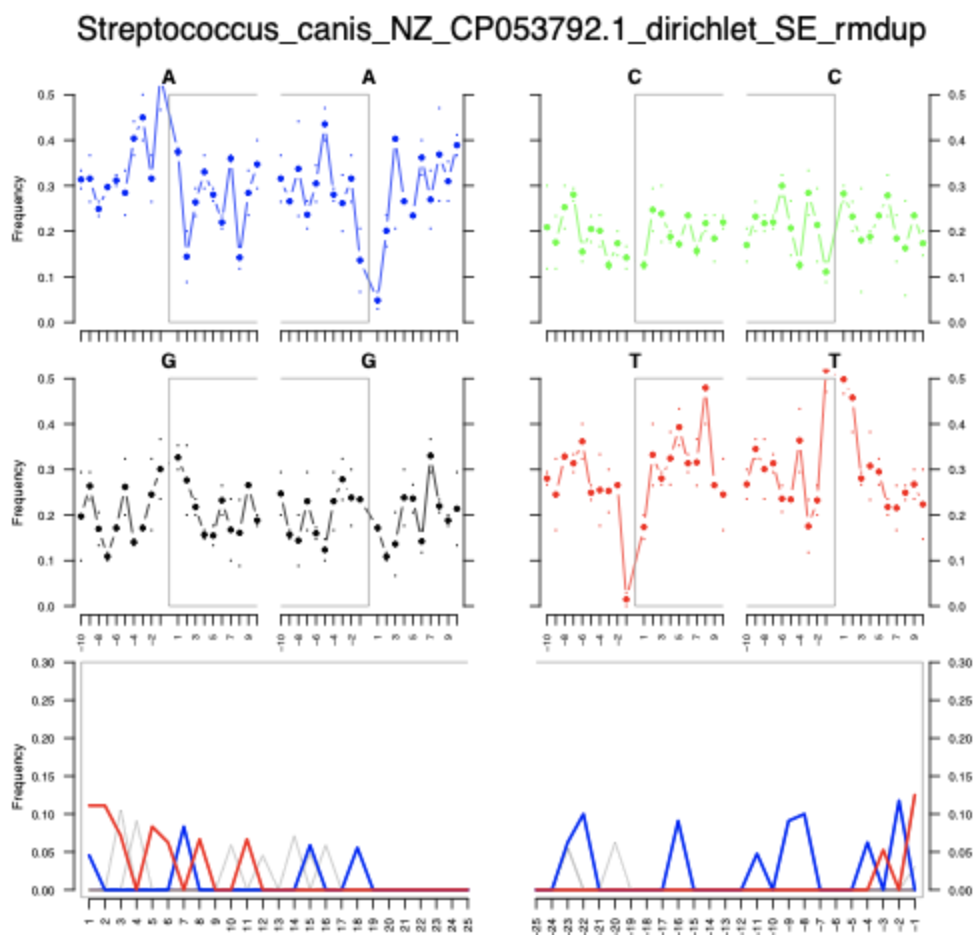


Supplementary figure 16. Deamination damage profile for sample ANH-10 for *Corynebacterium diphtheriae*.

*Corynebacterium diphtheriae*_NZ_CP029644.1_dirichlet_SE_rmdup

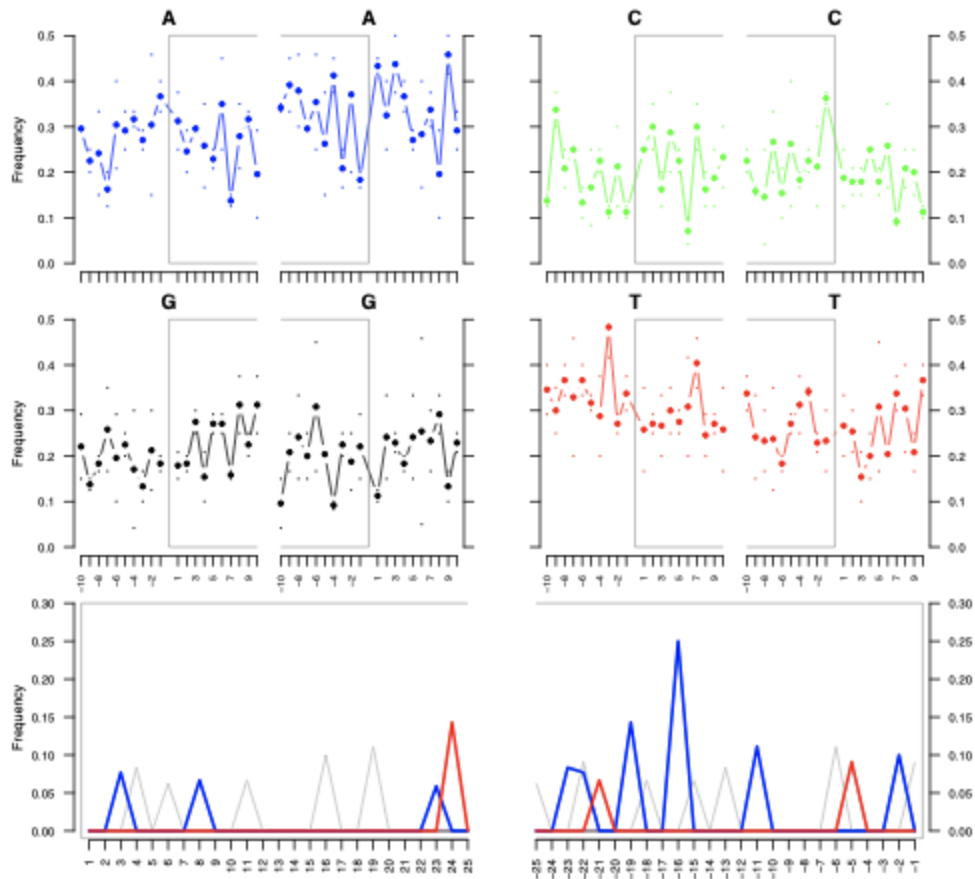


Supplementary figure 17. Deamination damage profile for sample ANH-11 for *Corynebacterium diphtheriae*.

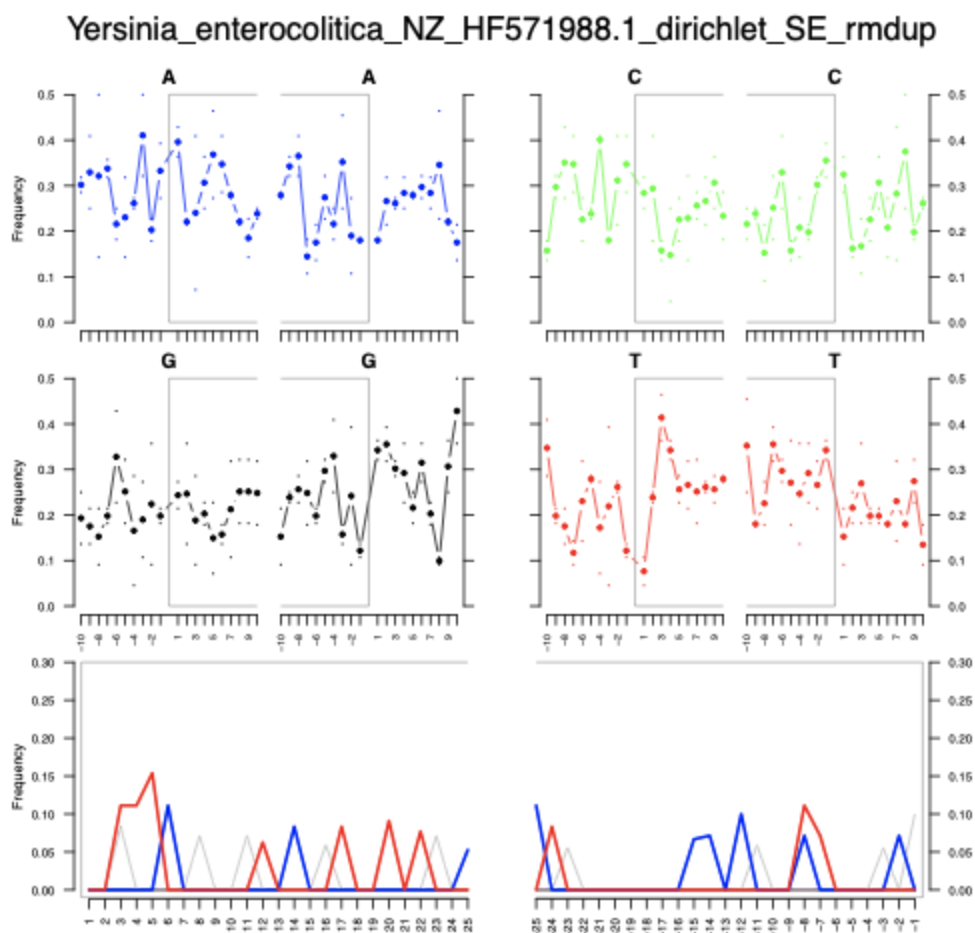


Supplementary figure 18. Deamination damage profile for sample ANH-11 for *Streptococcus canis*.

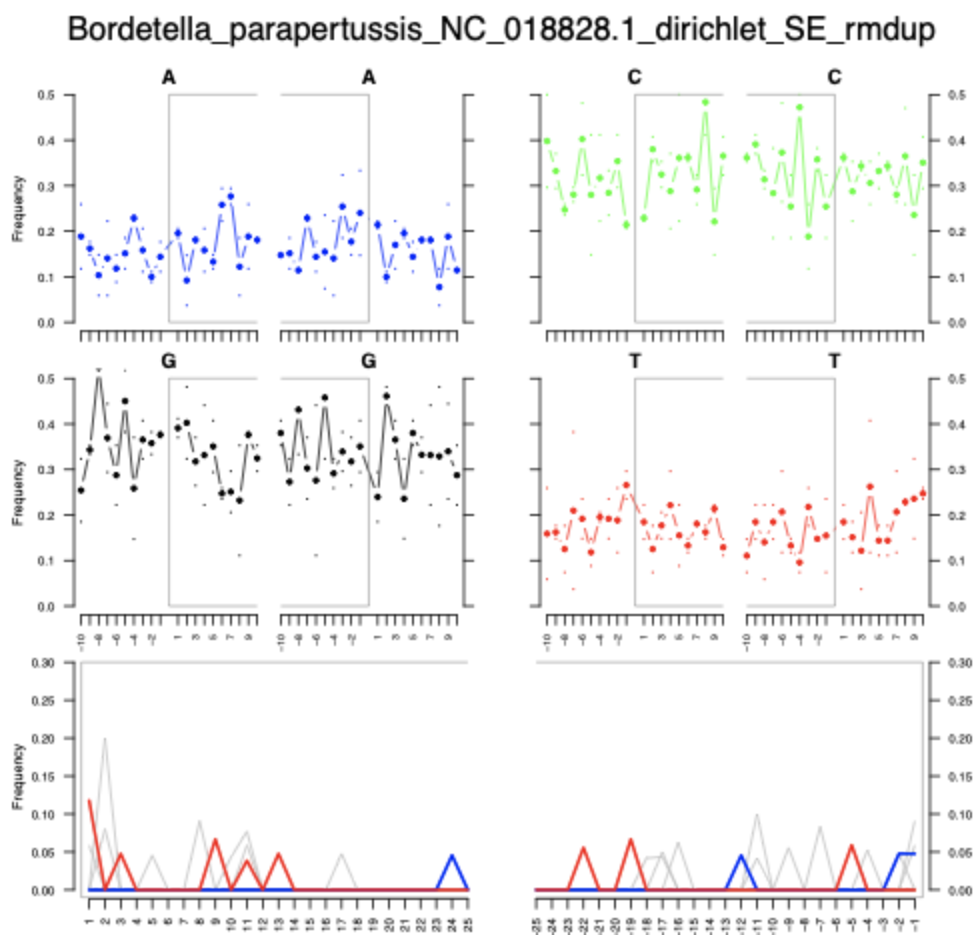
Streptococcus_pneumoniae_NZ_CP007593.1_dirichlet_SE_rmdup



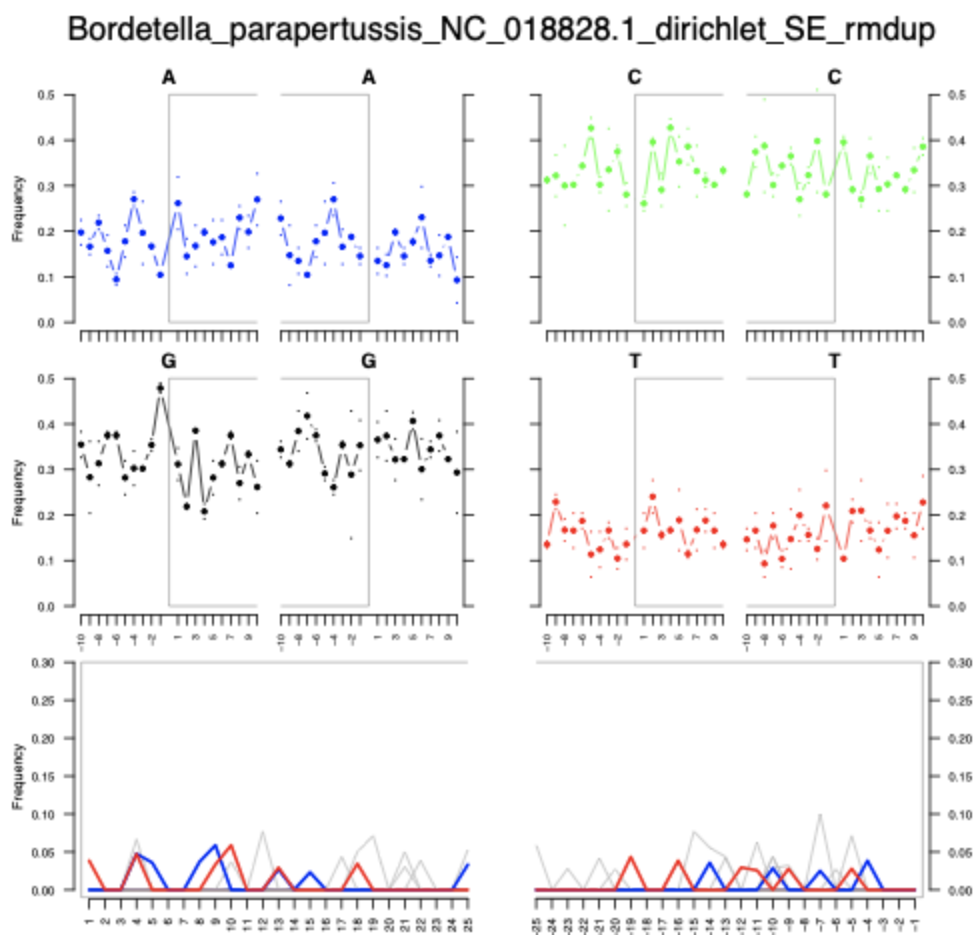
Supplementary figure 19. Deamination damage profile for sample ANH-6 for *Streptococcus pneumoniae*.



Supplementary figure 20. Deamination damage profile for sample ANH-79 for *Yersinia enterocolitica*.

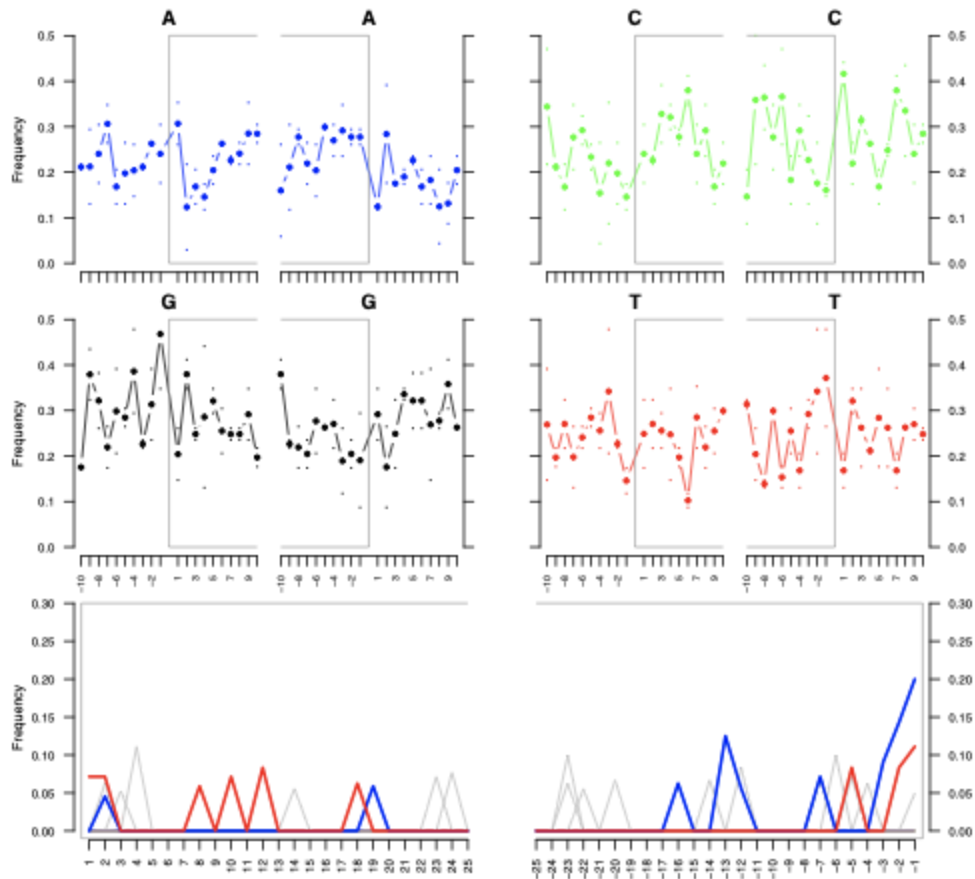


Supplementary figure 21. Deamination damage profile for sample CH1581 for *Bordetella parapertussis*.



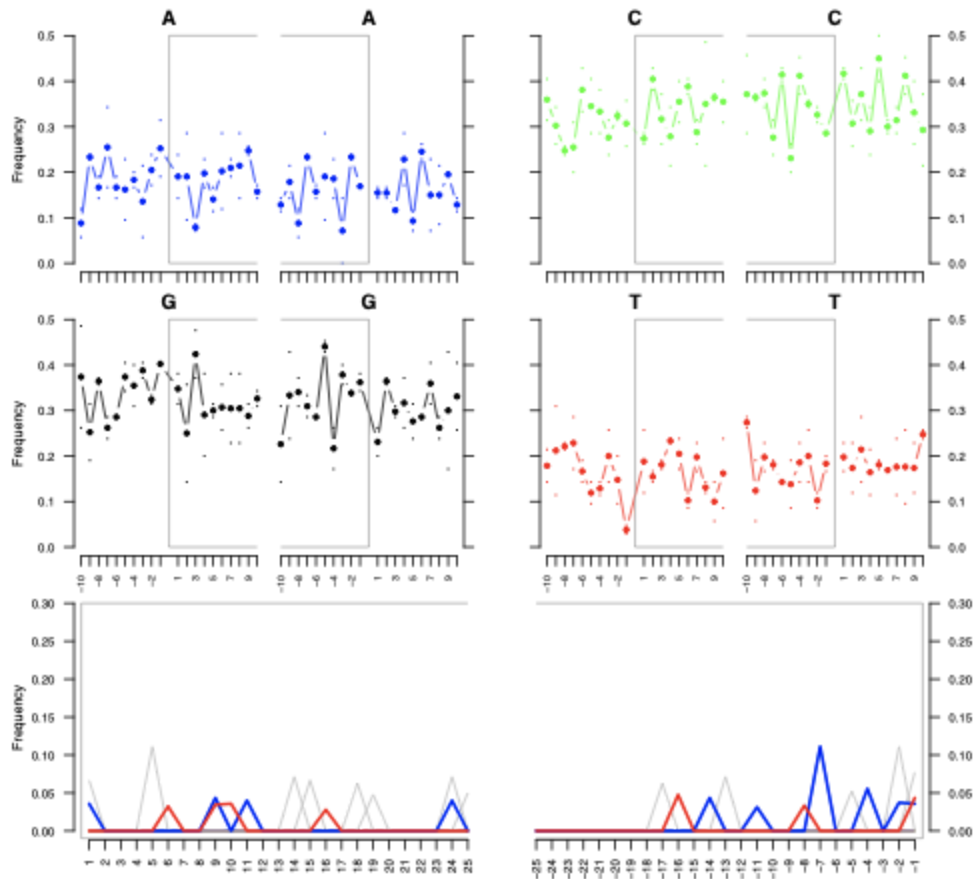
Supplementary figure 22. Deamination damage profile for sample CH1665 for *Bordetella parapertussis*.

Yersinia_enterocolitica_NZ_HF571988.1_dirichlet_SE_rmdup

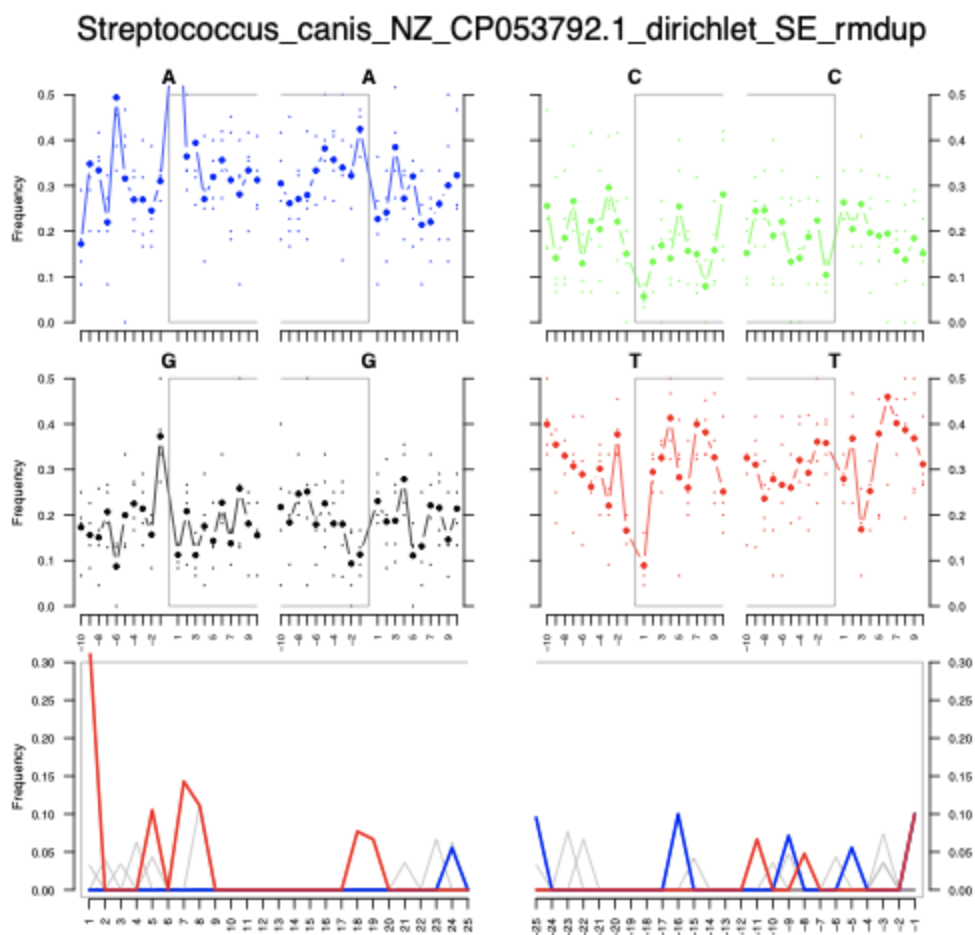


Supplementary figure 23. Deamination damage profile for sample OL2186 for *Streptococcus canis*.

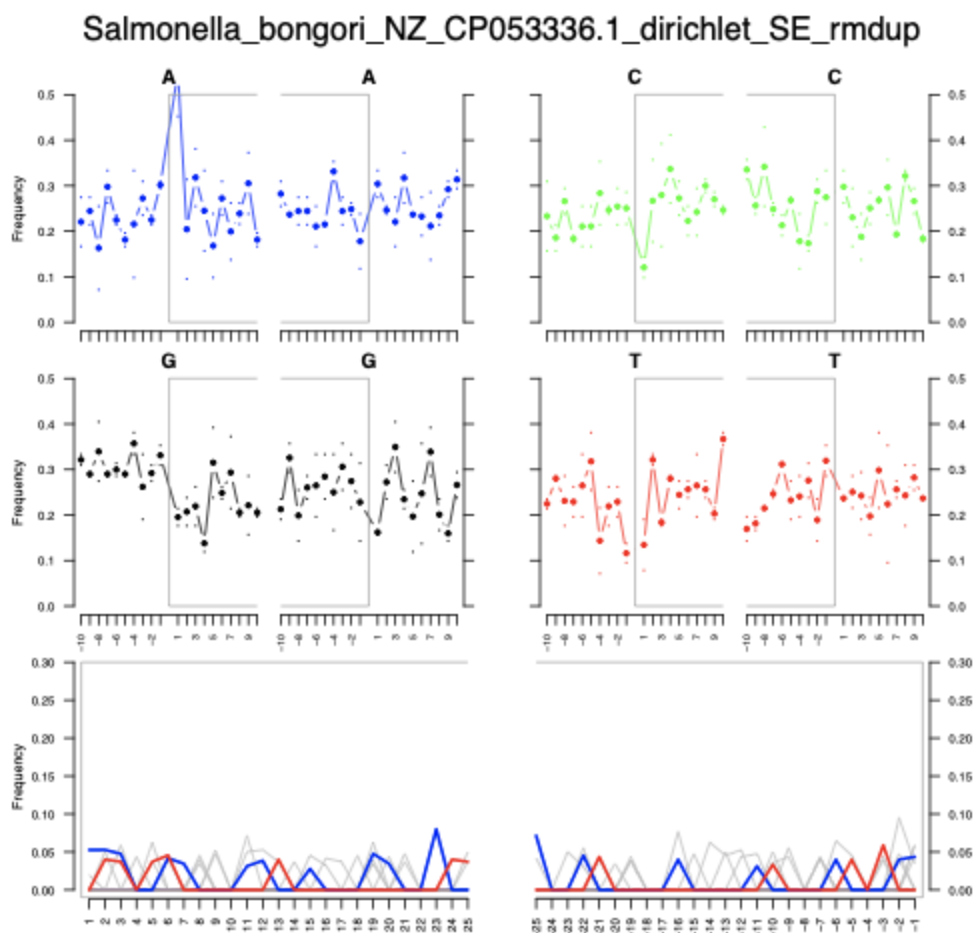
Bordetella_parapertussis_NC_018828.1_dirichlet_SE_rmdup



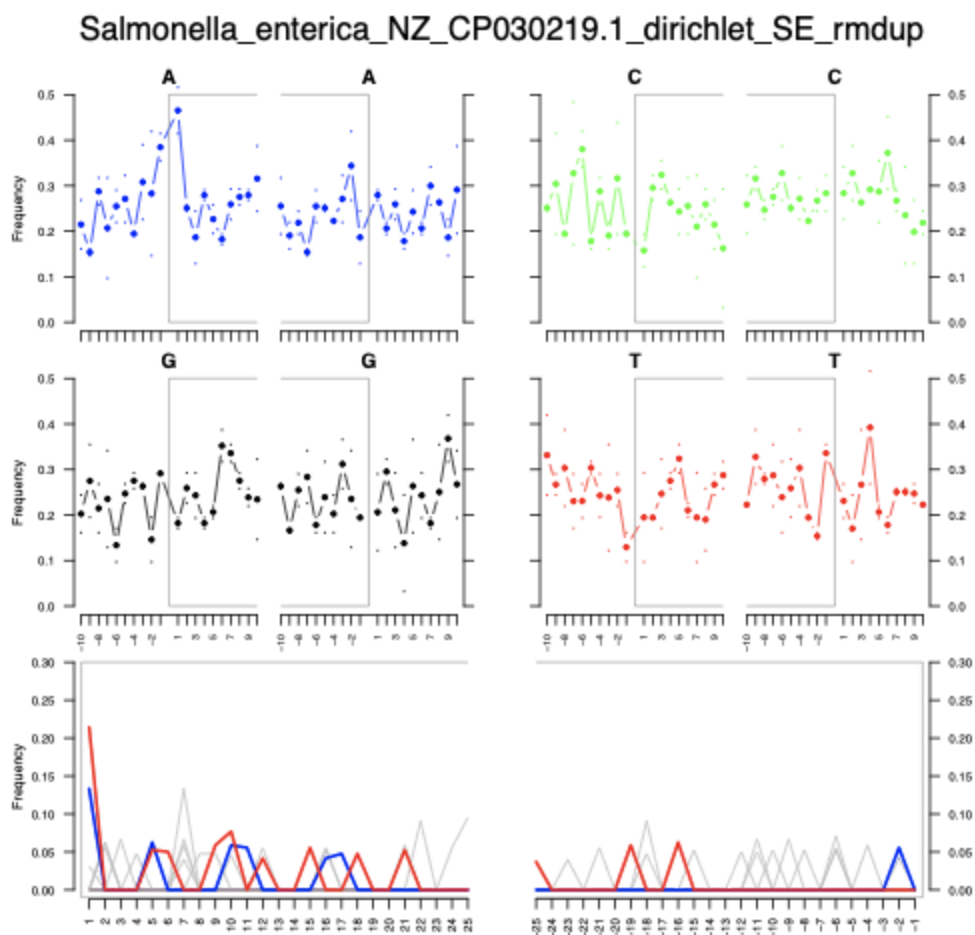
Supplementary figure 24. Deamination damage profile for sample OL4167 for *Bordetella parapertussis*.



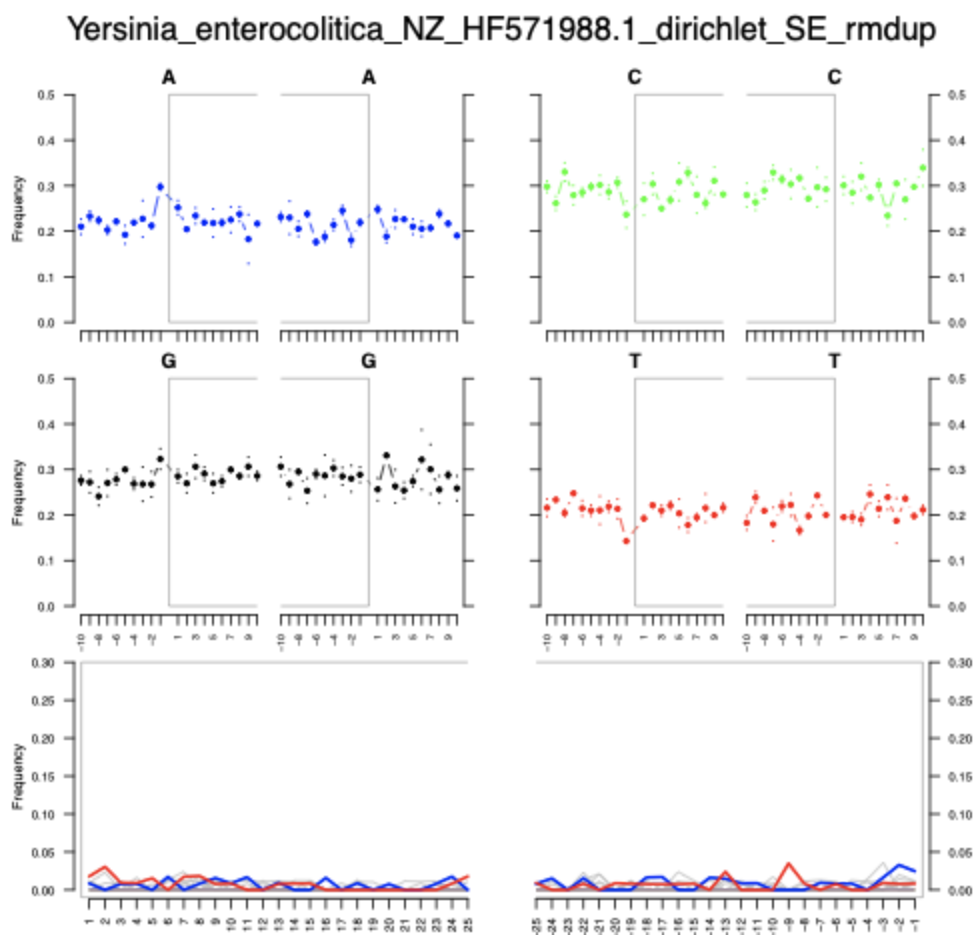
Supplementary figure 25. Deamination damage profile for sample OL4153 for *Streptococcus canis*.



Supplementary figure 26. Deamination damage profile for sample OL4170 for *Salmonella bongori*.

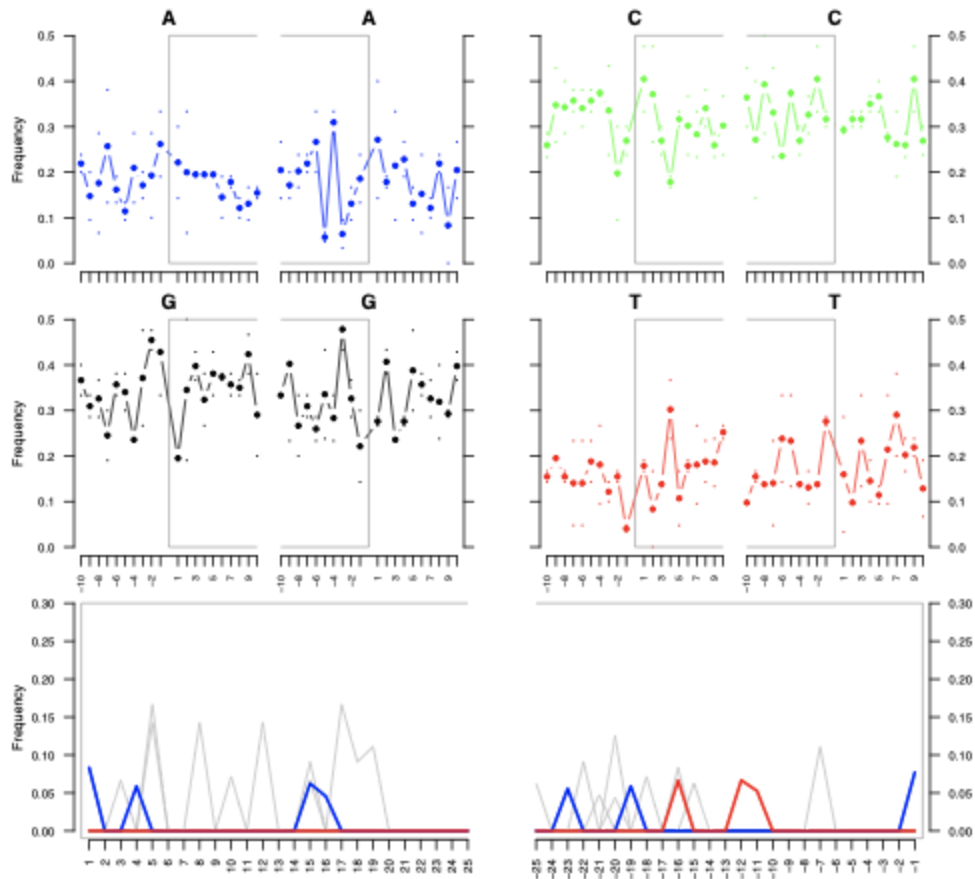


Supplementary figure 27. Deamination damage profile for sample OL4170 for *Salmonella enterica*.

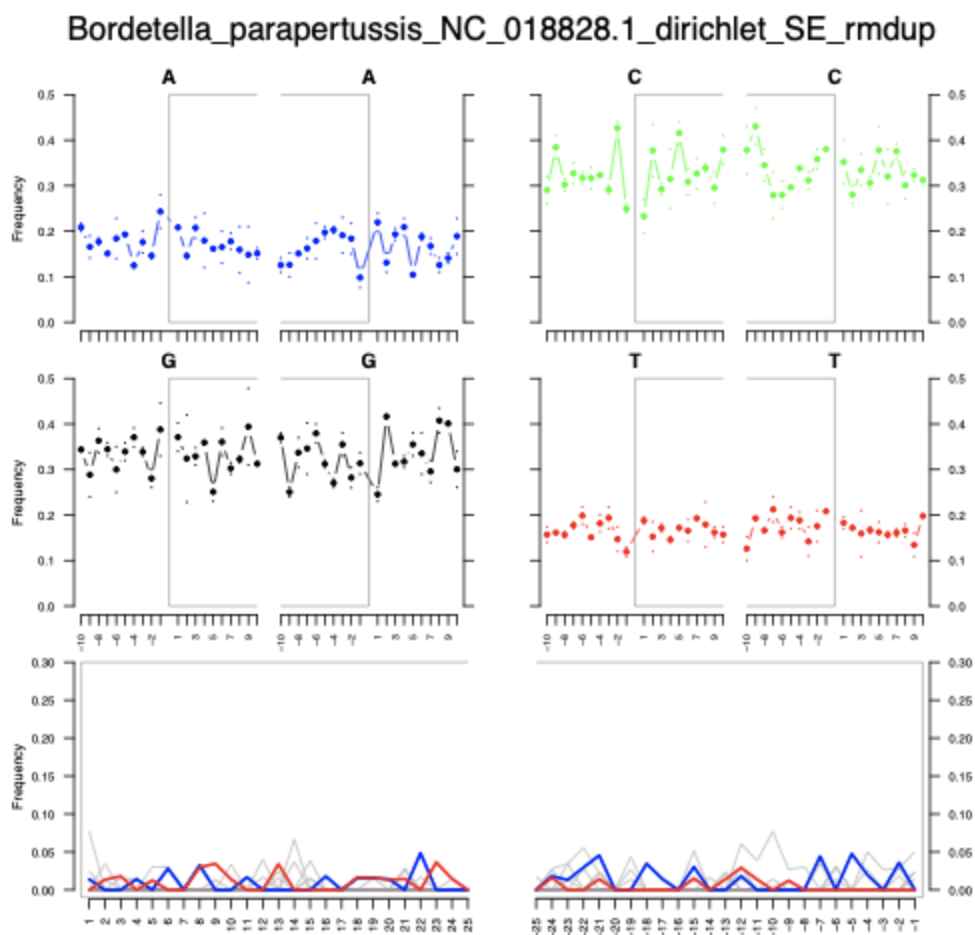


Supplementary figure 28. Deamination damage profile for sample OL4171 for *Yersinia enterocolitica*.

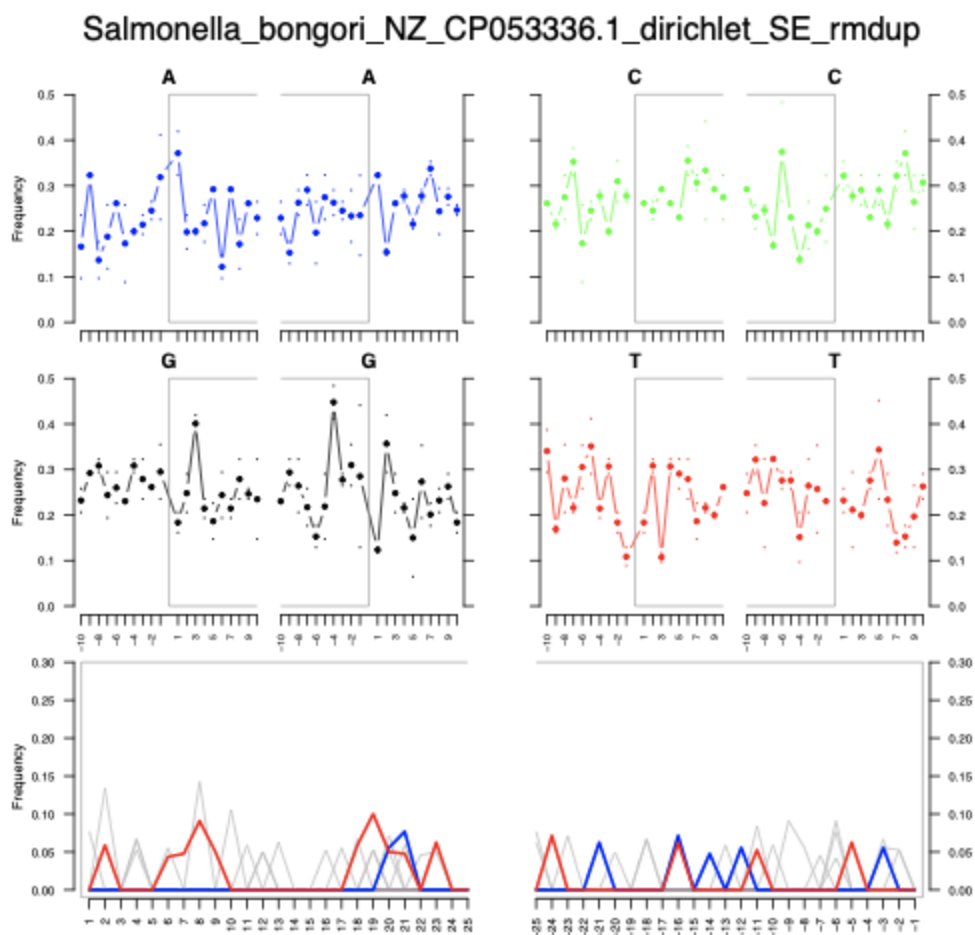
Bordetella_parapertussis_NC_018828.1_dirichlet_SE_rmdup



Supplementary figure 29. Deamination damage profile for sample OL4172 for *Bordetella parapertussis*.

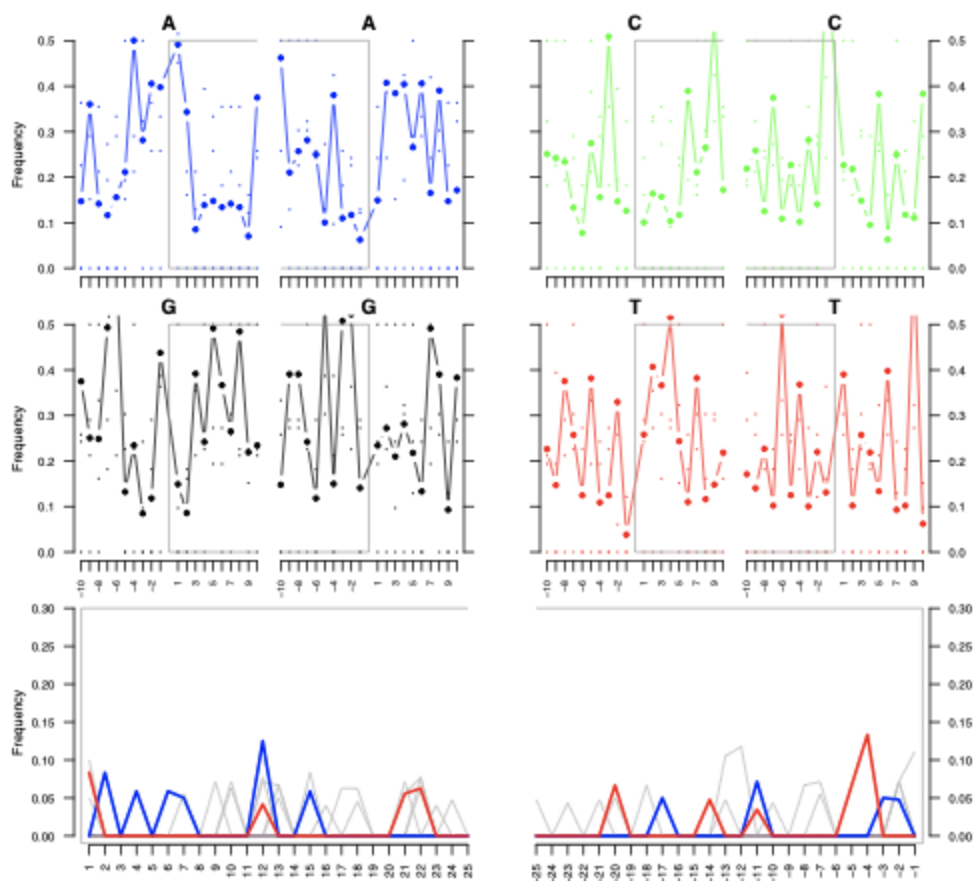


Supplementary figure 30. Deamination damage profile for sample OL4173 for *Bordetella parapertussis*.

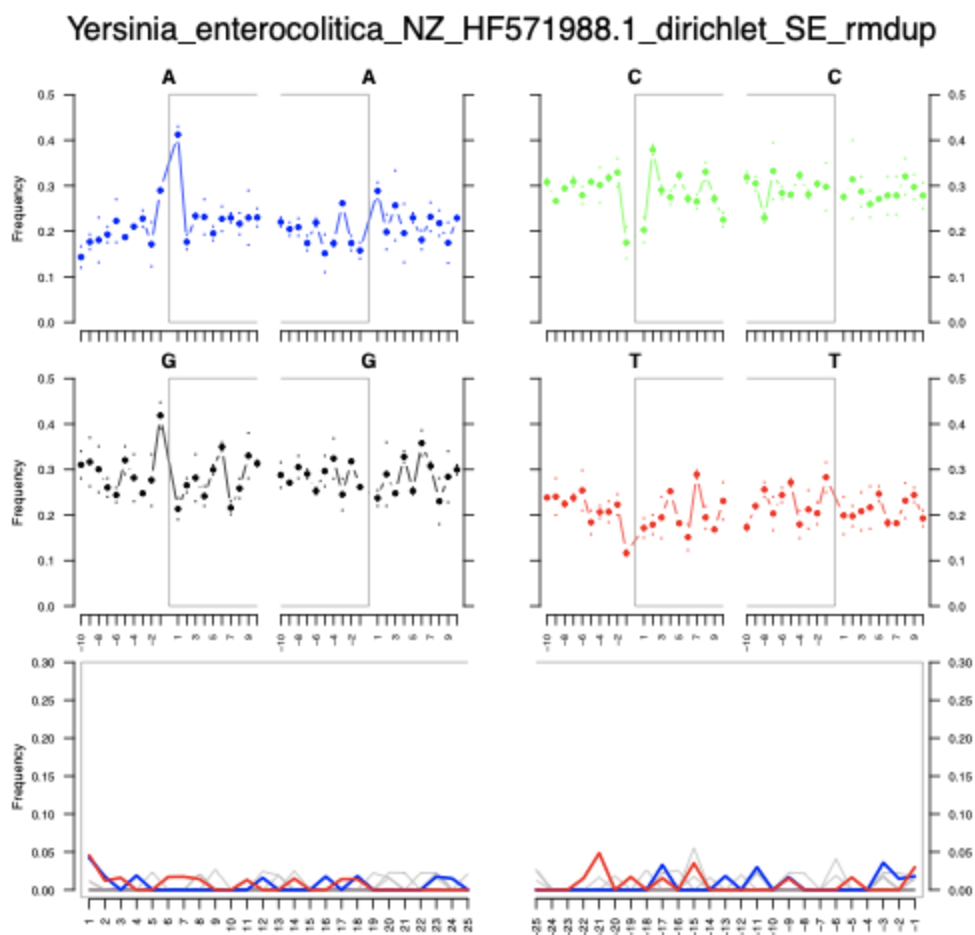


Supplementary figure 31. Deamination damage profile for sample OL4174 for *Salmonella bongori*.

Salmonella_enterica_NZ_CP030219.1_dirichlet_SE_rmdup



Supplementary figure 32. Deamination damage profile for sample OL4174 for *Salmonella enterica*.



Supplementary figure 33. Deamination damage profile for sample OL4174 for *Yersinia enterocolitica*.

Permission from co-authors

I hereby give permission to Evangelos Antonios Dimopoulos to use our joint work as contribution towards his D. Phil thesis to be submitted for examination at Oxford University.

I confirm that to the best of my knowledge, the author contribution statement below is accurate and Evangelos Antonios Dimopoulos' contribution towards the work is greater or equal than that of any other co-author.

Chapter author contribution statement:

E.A.D., L.A.F. and G.L. designed the project; E.A.D. analysed data; G.L. and L.A.F. provided support in interpreting results; E.A.D. wrote the paper with contributions from all other authors.

Date: 29/05/2021

Name: Laurent Frantz

Signature: 

Date: May 1, 2021

Name: Greger Larson

Signature: 

Conclusions

What was the overall motivation behind studying ancient pathogens and metagenomes ?

Ancient pathogen research can contribute greatly to the study of long-term microbial evolution, by enabling us to create well characterised time series of DNA sequence data. Such datasets can be proven insightful, when incorporated into phylogenetic reconstructions, as to how infectious disease agents have coevolved with their hosts (Harkins and Stone, 2015). An even more important argument is that shotgun sequencing of DNA yielded by archaeological samples provides a means of immortalising that precious material (Harkins and Stone, 2015). A single ancient DNA library from an animal or a human specimen, sequenced at sufficient depth, can be used to answer a variety of research questions, from metagenomics to population genetics. The richness of information, which such libraries provide, stresses the importance of a non-targeted approach for sequencing DNA archaeological samples.

Laboratory techniques that have advanced the study of ancient pathogens

Targeted genome capture techniques have significantly increased our ability to maximise the yield of endogenous DNA molecules from archaeological samples (Carpenter et al., 2013; Enk et al., 2014), and have contributed in ancient DNA studies for detecting selection in humans and animals (Jones and Good, 2016; Lippold et al., 2011). Hybridization capture approaches are more efficient at

capturing DNA molecules the shorter they get than any PCR amplification approach, thus making them ideal for capturing endogenous damaged and highly fragmented aDNA reads (Knapp and Hofreiter, 2010). This way most of the sequencing library's original complexity can be retained, while minimising the risk of enriching modern contamination (Krause et al., 2010).

Genome capture has also had a significant impact on the field of ancient pathogens (Jones and Good, 2016). Hybridisation approaches are now routinely being used in microbial paleogenomics as a means to study the emergence of diseases (Bos et al., 2019) and pathogens, like *Yersinia pestis* (Schuenemann et al., 2011) and *Mycobacterium leprae* (Bos et al., 2015). But it is not only our ability to retrieve and sequence ancient pathogen genomes that has been revolutionised. Hybridisation capture has also improved our ability to perform more accurate phylogenetic and metagenomic analyses, by being able to target specific orthologous informative loci shared among various taxa (Jones and Good, 2016).

An in-solution targeted genome capture approach was employed in this PhD, to retrieve ancient *Marek's Disease Virus* genomes, improving genome coverage more than 1,000-fold in some samples. Even though the sequencing coverage results were impressive in many samples, not all of them performed equally well, underlining the importance of having enough DNA template molecules in a sequencing library to begin with (Cruz-Dávalos et al., 2017). Apart from targeted genome enrichment approaches, innovative protocols have also been introduced at the sequencing library preparation stage. Single stranded DNA library preparation protocols are more efficient at converting ancient DNA templates into sequenceable molecules

(Kapp et al., 2021; Wales et al., 2015), yielding sequencing libraries of higher complexity.

As the time and reagent cost of single stranded library preparation protocols decreases (Kapp et al., 2021), I plan on adopting this approach for building more complex DNA libraries. Their maximal endogenous DNA content will be of great help in studying ancient metagenomes, but will also help the specific detection of low abundance pathogens. Concurrently, as the cost of synthesizing custom oligonucleotide baits is decreasing, I also plan on capturing the ancient genomes of the pathogenic species that were identified in various archaeological pig and dog samples, as described in my third chapter, so as to study their temporal evolution in greater detail. Moving forward, I strongly believe that a combination of single stranded DNA libraries and targeted genome capture will be a valuable tool for the study of ancient microbes and disease detection.

Pathogen discovery in ancient metagenomes

The decreasing cost of Next Generation Sequencing (NGS) helped incorporate Whole Genome Sequencing (WGS) in clinical microbiological diagnostic approaches, offering quick result turnaround and high levels of automation compared to traditional culture based methods (Lecuit and Eloit, 2015). A major advantage of WGS for pathogen detection is that it requires little to no prior knowledge of the pathogen that needs to be discovered unlike other clinical assays, making it a valuable diagnostic tool (Frey and Bishop-Lilly, 2015).

Disease outbreaks have been documented throughout human history (e.g. the plague of Athens, the Justinianic Plague and the Black Death), but not all the causative pathogenic agents responsible for them have been identified conclusively. Even though targeted genome enrichment has become the standard in identifying and sequencing ancient pathogens (Carpenter et al., 2013), such an approach requires a prior hypothesis about the targeted pathogenic agents (Harkins and Stone, 2015). Therefore genome capture is not suitable for detecting and/or diagnosing the unknown causative agents of past epidemics (Campana et al., 2014). Theoretically, Whole Genome Sequencing (WGS) DNA data can be used to identify which pathogenic agents were responsible for historical epidemics, but rigorous result validation needs to be in place to avoid false positive identifications, caused by environmental contamination (Campana et al., 2014).

Environmental contamination, DNA preservation, and the tissue a sequenced sample was taken from can affect the discovery rate of pathogens as well as the false positive and negative identification rates. Contamination from soil bacteria can easily lead to false positive identifications. Concurrently failing to detect a pathogen in a DNA sequencing library does not equate to the absence of an infection. For these reasons it is difficult to compare different archaeological collections and datasets in terms of the likelihood of ancient pathogen recovery (Campana et al., 2014). Additionally not all tissues preserve microbial DNA equally well, further complicating the comparison of ancient pathogen discovery rates across different samples. For instance, dental calculus and teeth have greater metagenomic potential, as more microbial taxa are preserved in them, compared to petrous bones,

that despite their high levels of endogenous DNA they have failed to yield any ancient pathogen DNA up to date (Margaryan et al., 2018).

Another important factor affecting pathogen discovery rate in archaeological samples is the biology of the diseases themselves, as not all diseases manifest skeletally or sometimes they can be fatal before they can do so. Identifying a pathogenic species does not also mean we have identified the cause of death (Harkins and Stone, 2015). Microbial variation can emerge and can be lost throughout time, meaning that a microbial taxon could have not been acting as a pathogen throughout its evolutionary history or that a pathogen could have been causing different diseases in different points in time, hindering our efforts to discover the causative agents of past epidemics. All these different factors related to the biology of a disease can affect the discovery of ancient pathogens, and highlight what is called the osteological paradox. It was first defined by (Wood et al., 1992), describing that those who appear healthy in the burial record are the ones that were the sickest. Individuals that have skeletal lesions are individuals who could immunologically fight chronic infections (Harkins and Stone, 2015), a fact that should be taken into account when discovering and analyzing ancient pathogens retrieved from bones.

Furthermore, identifying, even robustly, the presence of pathogen DNA in a metagenome yielded by an archaeological animal sample is not equal to the identification of a pathogenic agent responsible for an active (let alone lethal) infection. Many of the bacterial and viral pathogens our pig, dog and chicken samples were screened for are able to establish chronic or asymptomatic infections using the infected animals as carriers and their respective populations as reservoirs.

Namely, chickens never recover from an MDV infection and keep shedding the virus until they die, even in cases where birds have been vaccinated against MDV and exhibit no symptoms of Marek's Disease (Atkins et al., 2013). Certain bacterial pathogens can establish chronic infections in dogs that can persist even after antibiotic treatment has been administered, e.g. species of the genus *Leptospira* (Rojas et al., 2010) or *Ehrlichia canis* (Leiva et al., 2005). Other bacterial pathogens are able to colonise animal tissues without causing any clinical symptoms to their host using them as carriers, for instance *Streptococcus suis* in pigs (Lun et al., 2007) and *Bordetella bronchiseptica* in dogs (Schulz et al., 2014). But even for well-studied animal pathogens there is little information available regarding the time course of the infection they cause, and the bacterial load that can be retrieved from a host during the different stages of an infection. Worse yet, for other pathogenic species there is a complete lack of time course studies, e.g. species of the genus *Campylobacter* or *Rickettsia*, or *Streptococcus canis*, meaning there is no information about whether these microbes have the ability to infect chronically or asymptotically their hosts, or whether they persist after the host overcomes their infection. For such species we cannot easily make an informed decision, based on ancient DNA evidence alone, about if the presence of pathogen DNA corresponds to an active, chronic or asymptomatic infection. Thus, the lack of detailed time course studies for animal pathogen infections needs to be taken into consideration when drawing conclusions on pathogen prevalence from aDNA metagenomic data.

While mining ancient animal metagenomes for pathogens, in some cases even for zoonotic ones, I tried to work around the aforementioned biological limitations. In all cases WGS DNA data were screened first with the HAYSTAC pipeline and validity of

positive identified species was confirmed by applying a minimum assigned reads threshold, verifying the presence of chemical damage in the assigned DNA molecules, and ensuring that the assigned reads were randomly distributed across a genome rather than clustering around specific genomic areas. Concurrently for all our analyses I chose to look for pathogenic species whose link to a particular disease in a specific animal has been well established. In order to provide a more sound and hypothesis driven background to my exploratory analysis, I also ensured that a mention of each pathogen/animal pair existed in the literature. To further validate my metagenomic identifications, where funding and time allowed, I combined the metagenomic screening of WGS data with a genome capture approach, targeting the positively identified pathogenic species, and further deeper sequencing.

We screened shotgun sequencing libraries generated from 938 pig samples and 2090 dog samples for various pathogens and 1111 chicken samples for MDV. The pathogen discovery rate in pigs was 2.3%, in dogs was 1.6% and in chickens it was 1.7%, corresponding to 22 pig samples, 34 dogs samples and 19 chicken samples respectively. It is evident that the discovery rate of pathogens is extremely low, emphasizing the importance that DNA preservation, tissue compatibility and pathogen biology play in recovering ancient pathogens. Interestingly, across all three taxa and regardless of whether I was screening for one or multiple pathogens, the overall pathogen discovery rates were really similar to each other with an average of 1.9%.

Moving forward a more systematic approach will need to be taken in order to pinpoint the main reason that pathogen discovery rates are so low and how the various inhibiting factors that were previously mentioned exactly contribute to it. For that, I propose a larger scale meta-study that screens the combined sample collections from multiple laboratories, and infer what the driving forces behind pathogen discovery in ancient samples are by factoring in sample and DNA preservation, the biology of the targeted pathogen, the sequencing depth of a DNA library and the overall sample collection size. We need to stop admitting defeat to the stochasticity of pathogen discovery that plagues ancient DNA (aDNA) metagenomics, and develop methodologies that overcome the theorised synergy of aDNA various inhibiting factors so that we no longer have to rely only on luck and large sample sizes for the study of long term microbial evolution.

How to safely untangle the mysteries of aDNA metagenomics

As Whole Genome Sequencing (WGS) is becoming the standard approach for aDNA metagenomics, bioinformatic tools are being adapted to handle more complex genomic datasets (Warinner et al., 2017). DNA data yielded by sequencing archaeological samples typically include molecules from the following sources: endogenous host DNA, endogenous host microbial flora, postmortem microbial colonisers that participate in the process of decomposition, and microbial contamination from the environment the sample was preserved in (Bos et al., 2019; Warinner et al., 2017). Exactly because of the inherent complexity of the composition of these metagenomic samples, additional result validation steps are required both before and after each taxonomic identification analysis.

Quite a few stringent criteria exist in order to authenticate ancient endogenous (microbial) DNA from NGS data bioinformatically and avoid false positive identifications (Tsangaras and Greenwood, 2012). The primary criterion is to validate the existence of postmortem chemical damage caused by deamination at the terminal ends of the DNA reads, that is responsible for C to T or G to A transitions (Bos et al., 2019; Green and Speller, 2017; Key et al., 2017; Warinner et al., 2017). Secondly a true positive match should have its assigned reads relatively evenly distributed across its genome. Peaks around in specific genomic regions, combined with long gaps of no or little coverage, could be hinting at a false positive identification, maybe due to a Horizontal Gene Transfer event (HGT) attracting reads that belong to another species (Key et al., 2017; Warinner et al., 2017). Another important criterion is the edit distance distribution of the reads assigned to a taxon. It is expected that most of the ancient pathogen DNA reads should align to a reference genome without any mismatches if they are genuine, therefore as the number of mismatches in an alignment increases the higher is the probability of a false positive identification (Key et al., 2017).

Metagenomic analyses can be also confounded by both the data and the database a user chooses to use. Ancient, short, deaminated (or generally damaged) DNA reads are highly likely to be misaligned, leading to false positive identifications (Green and Speller, 2017), a problem that is further exacerbated when libraries have not been sequenced deeply enough (Bos et al., 2019). Concurrently, a multitude of methods rely on modern reference genomes for the taxonomic characterisation of ancient metagenomes. Misidentifications can arise from such an approach especially when

trying to either characterise extinct species that have no modern closely related genomes available (Green and Speller, 2017; Hofreiter et al., 2015) or pathogenic species with closely related environmental contaminant taxa (Bos et al., 2019; Hofreiter et al., 2015).

Particular attention should also be paid to the database used in taxonomic characterisations of ancient metagenomes. The genomes that are included in NCBI sometimes are of particularly low quality as there is no robust pipeline in place to validate each entry, resulting in contamination from humans, animals, environmental bacteria or even sequencing adapters to be incorporated in some entries (Green and Speller, 2017). Public genome databases also have neither a complete nor a balanced species representation especially when it comes to environmental microbes (as they overrepresent human pathogenic species (Campana et al., 2014)), further hindering the identification of ancient environmental taxa (Bos et al., 2019).

Since a substantial fraction of my PhD pertains to taxonomic identifications from shallow sequenced ancient metagenomes, I needed to come up with a strategy that yielded credible results. For that reason, I developed the HAYSTAC pipeline (Dimopoulos et al., 2020), a program that incorporates all the aforementioned validation criteria when characterising an ancient metagenomic sample. Reads are assigned to taxa based on their edit distance from a reference genome, thus maximising the informational content of an alignment. For each species that HAYSTAC detects, an evenness of coverage ratio is calculated, to ensure that the DNA reads are randomly distributed and were not assigned to a taxon because they

were attracted by certain genomic regions (e.g. hyper-conserved regions or HGT regions). Meanwhile when analysing DNA from archaeological samples, all the reads that have been uniquely assigned to a taxon are tested for the presence of a 'smiling' damage pattern (more deaminations in the end of aDNA reads), to verify that they do not come from modern microbial contaminants. Finally, in order to try and minimise biases originating from the sequence quality and skewed taxonomic content of public databases, I constructed databases that contained only one complete genome per microbial species that is part of NCBI's RefSeq, an NCBI database whose entries have passed rigorous quality checks.

Having these result validation steps incorporated into the functional framework of one pipeline enabled me to ask specific questions about the taxa present in the analysed samples, and do so in a secure way. This proved useful in my third chapter especially where I screened multiple modern and ancient pig and dog WGS libraries for specific pathogens. As the sequencing coverage of these samples was quite low, the evenness of coverage criterion was informative for assessing the genomic position distribution of the assigned reads. The incorporated chemical damage analysis for ancient samples proved to be really useful, as in many samples various modern bacterial contaminant species, which were passing the minimum read and the evenness of coverage ratio threshold, would have led to false positive identifications as endogenous ancient bacterial species.

It is also important to note that result validation should only be restricted to bioinformatic data processing. Steps for result validation should also be taken during the sampling and sample laboratory processing stages as well. Negative control

samples (such as extraction, library building and PCR blanks or samples containing no pathogens) should always be prepared, sequenced and analysed, so as to be able to detect any contamination and background noise. In the context of using WGS to detect pathogens in archaeological samples additional controls need to be prepared and analysed, which include samples from the surrounding soil and from uninfected archaeologically related individuals (Campana et al., 2014). Only species identified exclusively in the archaeological samples under examination (and not in any of the blank, soil or negative control samples) should be considered as valid ancient pathogen candidates. Such an approach was followed in my second chapter, in which chicken samples negative for *Marek's Disease Virus* (MDV) were also captured (resulting in failure), to ensure that the enriched MDV DNA molecules in the positive samples were not a product of contamination.

Bioinformatics in metagenomics and data science

Modern bioinformatics need to wrangle with biological 'Big Data', highlighting the need for implementation of formalised data science practices in the field (Pal et al., 2020). Exactly because metagenomic datasets are becoming larger, they have started reflecting the actual species richness that characterises microbial communities in nature (Escobar-Zepeda et al., 2015). Current bioinformatic algorithms though struggle to analyse and successfully interpret large amounts of metagenomic data, either because of their sheer volume or their complexity. Therefore more sophisticated and thorough analysis methods are needed so as to mine the information content of shotgun DNA sequencing data.

Metagenomic profiling algorithms should start incorporating machine/deep learning or neural network (Min et al., 2017) approaches for more effective data mining. Such algorithms should be able to analyse sequences from complex and/or large microbial communities, and overcome hurdles stemming from reference bias and database content that heuristic approaches based on k-mer binning or reference genome alignments suffer from (Lan et al., 2018). Apart from alternative algorithmic approaches, data science can contribute data cleaning and integration routines to bioinformatics. Data cleaning in general refers to a process of curating datasets by removing or filling in missing data, identifying outliers, and ensuring validity of the data (Greene et al., 2016; Lan et al., 2018). Data integration broadly refers to data storage, management, and sharing, otherwise known as data stewardship (Attwood et al., 2019; Lan et al., 2018).

But as in any other application of data science, bioinformatic analyses ought to digest big genomic data and output results that facilitate downstream applied research (Greene et al., 2016). Larger metagenomic datasets are being generated as DNA sequencing becomes cheaper, and such an abundance of sequencing data has led to studies with higher statistical power being conducted, showcasing the fact that taxonomic profiling of complex microbial samples is no longer the end result in metagenomics. Rather, taxonomic profiling of metagenomes is becoming a checkpoint after which multiple downstream analyses can take place, especially in the context of disease diagnosis (eg. bacterial vaginosis and irritable bowel syndrome) (van den Bogert et al., 2019), indicating that bioinformatic analysis of big metagenomic data could even innovate modern clinical practice.

Challenges for bioinformatic pipelines as genomic data become 'bigger'

Metagenomics method development was a major point of personal research focus, during which I faced significant challenges around data storage and computational resources limitations. Such a challenge, for example, was coding HAYSTAC's function that calculates the read assignment posterior probabilities to all the different taxa. For that calculation HAYSTAC needs to parse the number of nucleotide mismatches across all possible pairs of DNA reads in a library and taxa in a database, including filling in values of maximum mismatches for missing alignments. For an average screening library of 5 million DNA reads against a database of 5632 species that corresponds to a matrix of 28.16 billion data points of nucleotide mismatches. Parsing that matrix required more than 3 TB of RAM with a required processing time of more than 5 days, making the overall analysis not feasible if a user wanted to profile either larger DNA libraries or use larger databases.

Most of the code libraries that are widely used in bioinformatics unfortunately were designed to neither handle data of that size nor be employed in a parallel computing framework. For example most map/reduce operations in python are treated as monolithic processes handled by only one core, resulting in big requirements for memory and running time. As the data I was trying to analyse starting resemble 'big data' datasets increasingly I started facing challenges regarding management of data and computational resources, as mentioned in (Greene et al., 2016)), and it became apparent that I needed to become more aware of my computing environment (processor, storage, memory) and how I could effectively analyse my metagenomic datasets within it.

By bearing in mind the computational hurdles I had to overcome, I managed to reduce the excessive computational resources my program required by making use of parallel computing and efficient algorithmic design. Thus, I started exploring parallel computing frameworks that could be implemented in python such as Dask (Dask Development Team, 2016) and Apache Spark (Zaharia et al., 2016), but to no avail as again memory and execution time were still prohibitive because of the mathematical operations were still treated as monolithic processes. That led me to rethink if my metagenomic dataset were actually that 'big', and how I could reduce the amount of data, for example the number of mismatches for each DNA read/taxon pair, I needed to parse. I concluded that contemporary bioinformatics programs do not only need to make use of parallel computing to perform a task, but they also ought to be efficient in doing so.

Through this exploratory process I learnt metagenomic analysis does not always need to rely on a hash algorithm, which uses massive amounts of RAM for processing time to be reduced. I discovered that by combining computational frameworks that can parallelise the tasks a program needs to execute, such as Snakemake (Köster and Rahmann, 2012), and by using only sparse matrices that include only information data points, metagenomic analyses can be executed with minimal RAM in a short period of time. As such, it is emphasised that bioinformatic method development has to factor in the scalability and validation of the developed software, so as to deal with the increasing amount of data that are becoming available.

Scalability is a big challenge for bioinformatic software, especially due to the high complexity of the algorithms involved in some analyses, such as in global alignment of multiple sequences (Yang et al., 2017). Efforts to increase the scalability of bioinformatic programs have been focusing on parallelising the computation, but efficiently distributing the data that needs to be analysed has been overlooked (Yang et al., 2017). Currently the focus is shifting from just the mere parallelisation of bioinformatic analyses, to the effective communication among various computational units that are both interconnected and store data. This makes optimal use of innovations in hardware technology such as GPU (Graphics Processing Unit) and Cloud computing and is able to accommodate the varying data related demands end users might have (Yang et al., 2017). It was innovations in Cloud computing and communication among different devices connected to cloud systems that motivated the development of computation frameworks such as Hadoop (Shvachko et al., 2010) and Apache Spark (Zaharia et al., 2016), that facilitated the development of scalable software for bioinformatic analyses.

Software validation ensures that a bioinformatic pipeline produces correct and valid results, a characteristic necessary for translational applications of bioinformatics especially in clinical medicine settings (such as targeted gene panels used in clinical pathology laboratories) (Yang et al., 2017). To ensure that a pipeline produces valid outputs, multiple executions of the software are needed. Every bioinformatics pipeline should be incorporating a form of both/either integration and/or unit testing (Sale, 2014; Taylor and Taylor, 2021) that ensures both that the coding framework of the software is free of bugs and that it actually outputs valid results. For comprehensive integration and/or unit testing, appropriate inputs need to be

designed by method developers and end users (Lundgren and Kanewala, 2016; Srinivasan et al., 2018). Multiple test input types are needed, to encapsulate as much variation as possible that could be present in real data, such as various simulated or ‘gold-standard’ datasets whose ground truth is known (Yang et al., 2017).

Even though bioinformatic analysis pipelines are continuously evolving and enabling us to analyse genomic data in novel ways, our analytical abilities are still restricted. Computationally-wise, the current hardware infrastructure of many institutes struggles to meet the increasing needs for storing ‘big’ genomic data yielded by high throughput sequencing (Papageorgiou et al., 2018). Even though most bioinformatic software is multi-threaded, it is also dominated by memory intensive operations which fail to scale well in modern HPC systems (Al-Ali et al., 2016). For example, when more than eight cores are provided to GATK’s Haplotype caller, the tool ends up using only 25% of the available resources. Consequently, the responsibility for performance optimization falls to the end-user (Al-Ali et al., 2016).

Software scalability issues have also been causing analytical bottlenecks during the ongoing SARS-CoV-2 pandemic. The phylogenetic analysis tools, which are being used to track the evolutionary trajectory of COVID19, were primarily designed to analyse smaller datasets and were not necessarily computationally optimised. But as they are currently being employed for tracking the emergence of new viral variants in almost real time, they struggle to incorporate and analyse new viral genomes, which are becoming available on a daily basis, fast enough (Hodcroft et al., 2021). But the limitations bioinformatic pipelines face are not always computational. Genome

reference databases are commonly used in a wide variety of analyses in metagenomics, such as gene/ORF prediction (Escobar-Zepeda et al., 2015). Mining precise taxonomic information through functional ORF annotation will (always) be challenged by the size and phylogenetic coverage of a database.

What comes after metagenomics and how to use the information they provide us with.

Taxonomic identifications are no longer necessarily the end result of analysing archaeological metagenomic samples. Identifying microorganisms of interest, when followed by deeper sequencing and/or targeted genome capture can allow us to reconstruct genomes of ancient pathogens/microbes, which can act as molecular fossils (Spyrou et al., 2019). Without genomes of ancient pathogens, insights into disease emergence can only be gained through comparative genomics of extant microbial diversity and morphological detection of pathological lesions in archaeological bone remains (Spyrou et al., 2019). By incorporating ancient genomes in phylogenetic analyses, our inferences of the evolutionary trajectories of pathogens become more accurate, even when applying simple Maximum Likelihood (ML) frameworks (Stamatakis, 2014). Additional inferences can be made regarding the biology of a clonal organism, if genomes of archaeological samples that have been dated are used. The age estimates of ancient genomes can be employed for the calibration of molecular clocks, i.e. the nucleotide substitution rate, aiding us to estimate, in a Bayesian context (Suchard et al., 2018), the past demography and lineage divergence of pathogenic species (Bos et al., 2019; Spyrou et al., 2019).

For the second chapter of this thesis, I used metagenomics as a springboard to study the evolutionary trajectory *Marek's Disease Virus* (MDV), and how it was shaped by human activity, such as industrialised chicken flock management and various vaccination schemes. We used HAYSTAC to identify archaeological chicken samples that carried MDV, and subsequently through genome capture we generated ancient MDV genomes. Age estimates were also obtained for all the positive samples either indirectly through their archaeological context or thorough radiocarbon dating. After characterising the SNP variability and distribution that was present in the ancient genome dataset, we combined it with modern MDV genomes in order to reconstruct the diversification of its lineage and estimate the speed at which MDV has been evolving. We estimated a slower clock rate of $9.1\text{E-}6$, and the time to most common recent ancestor (tMRCA) was calculated to be at 1477 AD, almost 500 years earlier (1938) than the tMRCA calculated by studies relying only on modern data (Trimpert et al., 2017), highlighting the value of ancient genomes in evolutionary analyses.

But metagenomics is not only an analytical means of selecting archaeological samples for deeper sequencing. In the third chapter of this thesis I employed metagenomics to study the prevalence of (zoonotic) pathogens in domestic or wild populations of pigs and dogs across time and space. Being able to observe how the pattern of prevalence of a pathogenic species has been shifting spatially and temporally would elucidate how human activity and management of these species have been shaping their disease landscape. Unfortunately pathogen discovery rates were low for both species, giving us low statistical power to estimate changes in patterns of infections in these pigs and dogs. Even though we had to overcome

hindrances caused by DNA preservation, insufficient sequencing depth or even the absence of infected individuals in our collection, we were still able to qualitatively observe the recovery of more pathogenic species from domestic pig, wild boar, domestic dog and wolf samples from the last 2000 years.

Concluding remarks

It should be noted that analysing a metagenomic sample, and especially an ancient one, is not a single-track approach. The analytical route chosen, and respectively the relevant bioinformatic tools, is highly relevant to the research question that one has. If one's question is to find out the general taxonomic makeup of an ancient sample from a WGS library, then even tools such as Kraken2 (Wood et al., 2019) and MALT (Vågene et al., 2018) are appropriate, regardless of the high false positive rates they come with. If one desires to ask specific questions about the taxa that exist in shotgun sequencing in order to select samples for capture enrichment then more conservative methods should be used, such as HAYSTAC (Dimopoulos et al., 2020) and HOPS (Hübler et al., 2019), that apply stringent validation criteria for a species call. There is no right or wrong strategy in choosing an analytical method, as long as a user is aware of the caveats that each pipeline comes with and follows a sound *in vitro* and *in silico* experimental design that yields credible results.

At a point in time when (ancient) genomic data are becoming 'big', there is an increased need for official training in bioinformatics for researchers (Greene et al., 2016). Bioinformatic analysis of large volumes of genetic data is becoming commonplace in ancient DNA genomics, stressing the integral part that

computational skills play in the career path of a researcher. In order to cultivate transparent and correct practices regarding DNA data analysis that yield credible results and promote efficient knowledge dissemination and collaboration among different groups, more thorough training should be provided to early career researchers in ancient DNA, especially to those coming from a pure archaeology background.

The main take home message of this thesis is to show how interdisciplinarity is crucial for studying ancient and modern pathogens. Ancient metagenomics is no longer a standalone set of results, but rather is becoming the starting point of further research into the demography and evolution of microbes throughout time and space. Even taking a step further back it is important to appreciate how genetics, evolutionary biology and archaeology can complement each other, when studying the prevalence and evolution of animal pathogens. The archaeological record is a rich source of information, whose temporal signal can improve our phylogenetic reconstructions and help us trace the first real instances of disease emergence. It is important to understand that no area of research should be a lone island anymore. Archaeological samples give us precious genetic data, which we can sieve through a metagenomic pipeline and discover ancient microbial genomes, which subsequently we can use for population and evolutionary genetic analyses to solve modern-day animal disease related problems.

Bibliography

- Al-Ali, R., Kathiresan, N., Anbari, M.E., Schendel, E.R., Zaid, T.A., 2016. Workflow optimization of performance and quality of service for bioinformatics application in high performance computing. *J. Comput. Sci.* 15, 3–10.
- Atkins, K.E., Read, A.F., Savill, N.J., Renz, K.G., Islam, A.F.M.F., Walkden-Brown, S.W., Woolhouse, M.E.J., 2013. Vaccination and reduced cohort duration can drive virulence evolution: Marek's disease virus and industrialized agriculture. *Evolution* 67, 851–860.
- Attwood, T.K., Blackford, S., Brazas, M.D., Davies, A., Schneider, M.V., 2019. A global perspective on evolving bioinformatics and data science training needs. *Brief. Bioinform.* 20, 398–404.
- Bos, K.I., Jäger, G., Schuenemann, V.J., Vågane, Å.J., Spyrou, M.A., Herbig, A., Nieselt, K., Krause, J., 2015. Parallel detection of ancient pathogens via array-based DNA capture. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370, 20130375.
- 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., Spyrou, M.A., Krause, J., 2019. Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens. *Annu. Rev. Microbiol.* 73, 639–666.
- Campana, M.G., Robles García, N., Rühli, F.J., Tuross, N., 2014. False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing. *BMC Res. Notes* 7, 111.
- Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K.,

- Moreno-Estrada, A., Li, Y., Wang, J., Gilbert, M.T.P., Willerslev, E., Greenleaf, W.J., Bustamante, C.D., 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* 93, 852–864.
- Cruz-Dávalos, D.I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., Librado, P., Seguin-Orlando, A., Pruvost, M., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Scheu, A., Beneke, N., Ludwig, A., Cooper, A., Willerslev, E., Orlando, L., 2017. Experimental conditions improving in-solution target enrichment for ancient DNA. *Mol. Ecol. Resour.* 17, 508–522.
- Dask Development Team, 2016. Dask: Library for dynamic task scheduling.
- Dimopoulos, E.A., Carmagnini, A., Velsko, I.M., Warinner, C., Larson, G., Frantz, L.A.F., Irving-Pease, E.K., 2020. HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2020.12.16.419085>
- Enk, J.M., Devault, A.M., Kuch, M., Murgha, Y.E., Rouillard, J.-M., Poinar, H.N., 2014. Ancient whole genome enrichment using baits built from modern DNA. *Mol. Biol. Evol.* 31, 1292–1294.
- Escobar-Zepeda, A., Vera-Ponce de León, A., Sanchez-Flores, A., 2015. The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Front. Genet.* 6, 348.
- Frey, K.G., Bishop-Lilly, K.A., 2015. Chapter 15 - Next-Generation Sequencing for Pathogen Detection and Identification, in: Sails, A., Tang, Y.-W. (Eds.), *Methods in Microbiology*. Academic Press, pp. 525–554.
- Greene, A.C., Giffin, K.A., Greene, C.S., Moore, J.H., 2016. Adapting bioinformatics curricula for big data. *Brief. Bioinform.* 17, 43–50.

- Green, E.J., Speller, C.F., 2017. Novel Substrates as Sources of Ancient DNA: Prospects and Hurdles. *Genes* 8, 180.
- Harkins, K.M., Stone, A.C., 2015. Ancient pathogen genomics: insights into timing and adaptation. *J. Hum. Evol.* 79, 137–149.
- Hodcroft, E.B., De Maio, N., Lanfear, R., MacCannell, D.R., Minh, B.Q., Schmidt, H.A., Stamatakis, A., Goldman, N., Dessimoz, C., 2021. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* 591, 30–33.
- Hofreiter, M., Paijmans, J.L.A., Goodchild, H., Speller, C.F., Barlow, A., Fortes, G.G., Thomas, J.A., Ludwig, A., Collins, M.J., 2015. The future of ancient DNA: Technical advances and conceptual shifts. *Bioessays* 37, 284–293.
- Hübner, R., Key, F.M., Warinner, C., Bos, K.I., Krause, J., Herbig, A., 2019. HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biol.* 20, 280.
- Jones, M.R., Good, J.M., 2016. Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* 25, 185–202.
- Kapp, J.D., Green, R.E., Shapiro, B., 2021. A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. *J. Hered.* <https://doi.org/10.1093/jhered/esab012>
- Key, F.M., Posth, C., Krause, J., Herbig, A., Bos, K.I., 2017. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet.* 33, 508–520.
- Knapp, M., Hofreiter, M., 2010. Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. *Genes* 1, 227–243.
- Köster, J., Rahmann, S., 2012. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522.

- Krause, J., Briggs, A.W., Kircher, M., Maricic, T., Zwyns, N., Derevianko, A., Pääbo, S., 2010. A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* 20, 231–236.
- Lan, K., Wang, D.-T., Fong, S., Liu, L.-S., Wong, K.K.L., Dey, N., 2018. A Survey of Data Mining and Deep Learning in Bioinformatics. *J. Med. Syst.* 42, 139.
- Lecuit, M., Eloit, M., 2015. The potential of whole genome NGS for infectious disease diagnosis. *Expert Rev. Mol. Diagn.* 15, 1517–1519.
- Leiva, M., Naranjo, C., Peña, M.T., 2005. Ocular signs of canine monocytic ehrlichiosis: a retrospective study in dogs from Barcelona, Spain. *Vet. Ophthalmol.* 8, 387–393.
- Lippold, S., Matzke, N.J., Reissmann, M., Hofreiter, M., 2011. Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. *BMC Evol. Biol.* 11, 328.
- Lundgren, A., Kanewala, U., 2016. Experiences of testing bioinformatics programs for detecting subtle faults, in: *Proceedings of the International Workshop on Software Engineering for Science, SE4Science '16*. Association for Computing Machinery, New York, NY, USA, pp. 16–22.
- Lun, Z.-R., Wang, Q.-P., Chen, X.-G., Li, A.-X., Zhu, X.-Q., 2007. *Streptococcus suis*: an emerging zoonotic pathogen. *Lancet Infect. Dis.* 7, 201–209.
- Margaryan, A., Hansen, H.B., Rasmussen, S., Sikora, M., Moiseyev, V., Khoklov, A., Epimakhov, A., Yepiskoposyan, L., Kriiska, A., Varul, L., Saag, L., Lynnerup, N., Willerslev, E., Allentoft, M.E., 2018. Ancient pathogen DNA in human teeth and petrous bones. *Ecol. Evol.* 8, 3534–3542.
- Min, S., Lee, B., Yoon, S., 2017. Deep learning in bioinformatics. *Brief. Bioinform.* 18, 851–869.

- Pal, S., Mondal, S., Das, G., Khatua, S., Ghosh, Z., 2020. Big data in biology: The hope and present-day challenges in it. *Gene Reports* 21, 100869.
- Papageorgiou, L., Eleni, P., Raftopoulou, S., Mantaïou, M., Megalooikonomou, V., Vlachakis, D., 2018. Genomic big data hitting the storage bottleneck. *EMBnet J* 24.
- Rojas, P., Monahan, A.M., Schuller, S., Miller, I.S., Markey, B.K., Nally, J.E., 2010. Detection and quantification of leptospires in urine of dogs: a maintenance host for the zoonotic disease leptospirosis. *Eur. J. Clin. Microbiol. Infect. Dis.* 29, 1305–1309.
- Sale, D., 2014. *Testing Python: Applying Unit Testing, TDD, BDD and Acceptance Testing*. John Wiley & Sons.
- Schuenemann, V.J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mitnik, A., Forrest, S., Coombes, B.K., Wood, J.W., Earn, D.J.D., White, W., Krause, J., Poinar, H.N., 2011. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proc. Natl. Acad. Sci. U. S. A.* 108, E746–52.
- Schulz, B.S., Kurz, S., Weber, K., Balzer, H.-J., Hartmann, K., 2014. Detection of respiratory viruses and *Bordetella bronchiseptica* in dogs with acute respiratory tract infections. *Vet. J.* 201, 365–369.
- Shvachko, K., Kuang, H., Radia, S., Chansler, R., 2010. The Hadoop Distributed File System, in: *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), MSST '10*. IEEE Computer Society, USA, pp. 1–10.
- Spyrou, M.A., Bos, K.I., Herbig, A., Krause, J., 2019. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* 20, 323–340.

- Srinivasan, M., Shahri, M.P., Kahanda, I., Kanewala, U., 2018. Quality assurance of bioinformatics software: a case study of testing a biomedical text processing tool using metamorphic testing, in: Proceedings of the 3rd International Workshop on Metamorphic Testing, MET '18. Association for Computing Machinery, New York, NY, USA, pp. 26–33.
- Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
- Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., Rambaut, A., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 4, vey016.
- Taylor, J.T., Taylor, W.T., 2021. Unit Testing, in: Taylor, J.T., Taylor, W.T. (Eds.), *Patterns in the Machine: A Software Engineering Guide to Embedded Development*. Apress, Berkeley, CA, pp. 83–96.
- Trimpert, J., Groenke, N., Jenckel, M., He, S., Kunec, D., Szpara, M.L., Spatz, S.J., Osterrieder, N., McMahon, D.P., 2017. A phylogenomic analysis of Marek's disease virus reveals independent paths to virulence in Eurasia and North America. *Evol. Appl.* 10, 1091–1101.
- Tsangaras, K., Greenwood, A.D., 2012. Museums and disease: using tissue archive and museum samples to study pathogens. *Ann. Anat.* 194, 58–73.
- 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., Bos, K.I., Krause, J., 2018. *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat Ecol Evol* 2, 520–528.
- van den Bogert, B., Boekhorst, J., Pirovano, W., May, A., 2019. On the Role of Bioinformatics and Data Science in Industrial Microbiome Applications. *Front.*

Genet. 10, 721.

Wales, N., Carøe, C., Sandoval-Velasco, M., Gamba, C., Barnett, R., Samaniego, J.A., Madrigal, J.R., Orlando, L., Gilbert, M.T.P., 2015. New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. *Biotechniques* 59, 368–371.

Warinner, C., Herbig, A., Mann, A., Fellows Yates, J.A., Weiß, C.L., Burbano, H.A., Orlando, L., Krause, J., 2017. A Robust Framework for Microbial Archaeology. *Annu. Rev. Genomics Hum. Genet.* 18, 321–356.

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257.

Wood, J.W., Milner, G.R., Harpending, H.C., Weiss, K.M., Cohen, M.N., Eisenberg, L.E., Hutchinson, D.L., Jankauskas, R., Cesnys, G., Katzenberg, M.A., Lukacs, J.R., McGrath, J.W., Roth, E.A., Ubelaker, D.H., Wilkinson, R.G., 1992. The Osteological Paradox: Problems of Inferring Prehistoric Health from Skeletal Samples [and Comments and Reply]. *Curr. Anthropol.* 33, 343–370.

Yang, A., Troup, M., Ho, J.W.K., 2017. Scalability and Validation of Big Data Bioinformatics Software. *Comput. Struct. Biotechnol. J.* 15, 379–386.

Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., 2016. Apache Spark: a unified engine for big data processing. *Commun. ACM* 59, 56–65.

Species identification by using bioinformatic methods (Minor Chapter)

Introduction

Archaeology and ancient DNA

aDNA enables us to conduct species identification from environmental samples. For instance identifying bulk bone samples from archaeological excavation sites can be really insightful when it comes to documenting the past biodiversity, understanding the evolution and extinction of species and their needs for conservation (Haile et al., 2009). Along with the use of morphology-based approaches, species identification from DNA sequences can greatly enhance the toolkit of bio-archaeology (Grealy et al., 2016).

But robust experimental design is required to assess the amount of sequencing that is required to obtain enough discriminatory power to perform a species identification from ancient DNA molecules. Thus, I performed this small study on simulated datasets derived from the mitochondrial genome of *Mus musculus domesticus* to assess the informational DNA content that would be needed to perform an identification on the subspecies level, using just mitochondrial reads. Also I performed an additional small study on already published aDNA chicken DNA data, so as to assess what is the lowest taxonomic level at which I could obtain robust taxonomic assignment p-values.

Projects of species identification

- a) Synthetic Datasets: Mice.

For my first project I tried to implement the method suggested by Pretengarst et al and try to distinguish between two subspecies of mice, *Mus musculus musculus* and *Mus musculus domesticus*. Subspecies of domesticated mice differ across different geographic areas, such as Europe and Asia. In excavation sites that are located equally far from these continents it is important to know which domesticated subspecies arrived there first. It is also a good method to assess the diversity between domesticated subspecies of mice and investigate the spatial dispersal of these domesticated mice.

Archaeological remains of bones from many excavation sites can provide us with an insight of the dispersal history of many animals, such as mice, that follow the dispersal of man. Being able to date the onset of house mouse synanthropy and invasion has implications for evolutionary biologists working on the diversification of the commensal subspecies, their phylogeny and their hybridization, and the processes shaping the biodiversity we observe today (Cucchi, Auffray, & Vigne, 2012). Therefore it is crucial to develop a sensitive method that can carry out safe taxonomic assignments even at the subspecies level.

b) Real datasets. Chickens.

The same methodology was also applied to a broader taxonomic group in order to assess its discerning power. Another domesticated animal was chosen, the chicken (*Gallus gallus*) and it was tested if it could be identified from other species. I also wanted to evaluate the power of this method as a preliminary analysis for detecting if there is sufficient amounts of variability in a certain taxon, so that its species and subspecies could be distinguished with certainty by using DNA sequence data.

The introduction of commensal animal species, which hold a certain economic value, is a milestone for the ecology and the history of an area. Being able to pinpoint this introduction event in time, using bones of animal remains that are found in archaeological sites, is of great value. Knowing the approximate time when people first brought chickens in the area of Eastern Africa allows us to infer about which culture may have introduced the rearing of chickens in this area. There are cases however, where it is impossible to identify the species of a bone sample by only using morphological criteria. When cases like that are met, techniques that involve DNA or protein molecules are employed for the identification to happen.

Introduction to the method

The libraries that are produced by shotgun sequencing often contain a large proportion of short reads, which make genomic analysis more difficult and often they cannot be used for identifying the species that they come from. This problem is accentuated when the DNA libraries come from ancient samples, where DNA degradation is substantial and hinders further the analytical methods. For those reasons it seemed fit to find new ways to make use of these short reads for the purpose of species identification.

The DNA data that can be analyzed are either nuclear or mitochondrial. The potential of nuclear DNA data is limited, as the size of the nuclear genomes is too big and there is low interspecific variation. Furthermore, only a few species have full genome data available. Mitochondrial DNA on the other hand is highly variable among different species and there are more full sequence data available. For

instance, currently on the NCBI databases there are approximately 1260 full mitochondrial genomes available, whereas there are only 68 nuclear DNA bird genomes. Additionally, a great advantage of mtDNA, when it comes to ancient samples, is that it is often over-represented in ancient DNA samples, so proportionally more reads of ancient mtDNA are contained in a library than of ancient nuclear DNA.

Materials and methods

In order to identify a robust species I first have to obtain a relative score for each taxon of interest. A means of doing that is using BLAST's e-value score. The e-value is the probability that a query sequence would match the subject sequence that was aligned against by chance, and they are proportional to the size of the BLAST database. So if the e-value score is low then the quality of the alignment is high and the probability of the sequence matching is high.

A library of ancient DNA though contains a high number of short reads, therefore contains many query sequences. Consequently it is of great use to obtain an overall score across all the reads of a library per taxon. For this summarization to happen I make use of the following equations. The overall alignment score for each library/species or library/subspecies combination is given by the formula:

$Ig = \sum_{i=1}^n \log(1 - e_i)$. This is the sum of the decimal logarithms of the probability that a query sequence matches the sequence of a specific taxon.

Implementation

- a) The first step to implement that method is to build a custom BLAST database and include the taxon of interest or a taxon that can be used as a proxy. Such a taxon is one that is phylogenetically closer to the taxon of interest. In order to build a database that actually enables us to produce reliable probabilistic estimations of identity two problems need to be tackled. First, the size of the database needs to be reasonably large, otherwise the e-values will be distorted and unreliable. This is because

when a database is small the probability of observing a match by chance increases. Also the size of the sequences that are contained in the database should be of similar length, otherwise the e-values would be biased. Having target sequences that differ greatly in length could lead to e-values that could be hard to interpret, as larger sequences would allow more alignments to take place than shorter sequences. These two problems can be overcome if the database is built with the sequences of whole mtDNA genomes from many potential species. For example, if the query sequences come from mice then the database will include the full mtDNA sequences of all the species of the *Muridae* family. This approach will give the database the size needed for the e-values to make sense in terms of probability. Furthermore, given that the mtDNA genomes are approximately of the same size, the issue of partial alignments will disappear.

Afterwards the library of the query sequences is blasted against that database.

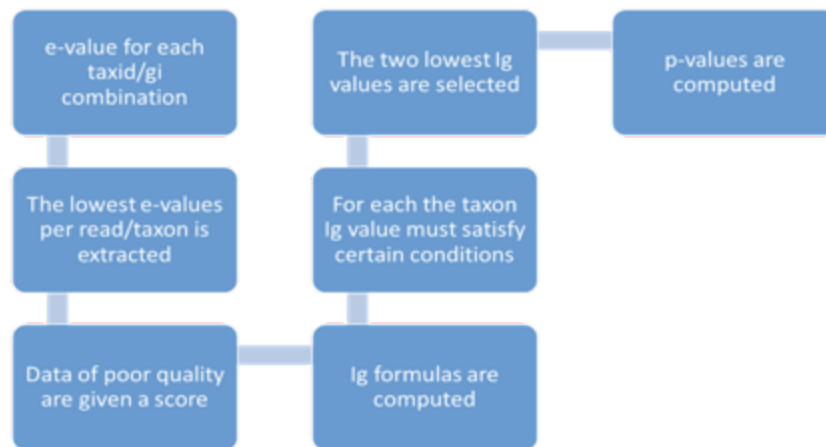
b) Then an algorithm that sorts the BLAST output and computes the scores is employed:

- i) Create a dictionary/database from the BLAST output, and compute the sum of the e-values per gi (gi is a sequence identification number used by the NCBI database), so for each [taxid][gi] = e-value. E-values that are positive or equal to 0 are excluded, as they are hits of low quality.

- ii) The dictionary is sorted according to the values of the g_i and the lowest e-value per read/taxon combination is extracted.
- iii) All the missing data and the data that come from alignments of poor quality are given a certain penalizing value, so that each read/taxon combination has a value.
- iv) Then the I_g formulas for each genus are computed.
- v) In order to filter out bad quality alignments I required that each library's highest I_g value was the result of at least 20 reads alignment with $e < 10e^{-6}$.
- vi) The dictionary of the I_g values of each taxon is sorted again and the two smallest I_g values are chosen.
- vii) Then the relative score for each best and second best taxon, which can be interpreted as a posterior probability, is computed with the equation:

$$p_g = \frac{10^{I_g}}{\sum_{i=1}^n 10^{I_i}}.$$

- c) The algorithm was coded in Python programming language and the plotting of the p-values that are computed was done using the R language and specifically the ggplot library.



The downgrading system

One basic step in this proposed method is to evaluate the quality of the alignments of the query sequences to the custom databases. For that reason, I use the e-value that is produced by BLAST+. The e-value is a score of how many times this alignment/match would be observed due to chance. As long as the e-value is under or equals to one the e-value could be used as a measure of the probability that this match happened due to chance. For alignments of poor quality the e-values exceed 1, and for really bad quality alignments it reaches up to 10 or even higher – it could reach up to the number of sequences that are contained in the custom database. When the alignments are poor the e-value cannot be used by the probabilistic method. In order to obtain an alignment score for every taxon/read pair combination, even for the pairs that haven't aligned, I assigned an e-value of 0.999 for all the missing and poor quality/unusable data, so as to be able to include the mismatches in an analysis. Thus, I was able to downgrade the values of the lg formula according to the number of the mismatches of each taxon and effectively assign samples to taxa.

Interpretation of the results

The p-value that is computed should not be considered as the probability that the sample under examination belongs to the corresponding taxon, but rather to a species that is most closely related to this taxon. It is a rather conservative method of assigning species to samples, as if a sample each sample will produce a high probability for the species that is the most closely related to and really low probabilities for the remaining taxa of the database. A reliable species assignment would take place if the p-values that are produced exceed 0.5. An important limitation of this proposed method is the amount of information that a certain library contains. Libraries with really short reads or low amounts of endogenous DNA can lead to equally low p-values for many taxa.

Application

a) Synthetic dataset: Mice.

For the construction of the synthetic mice dataset initially a mitochondrial genome sequence is provided as the input file. Then the sequence is being fragmented. The length of each fragment, called 'x', comes from a process of random sampling from a normal distribution of mean 50 and standard deviation of 7 and each number that is sampled is unique. Then a random starting point inside the range (0, end of sequence-x) is being selected and then the length x is added, so that the coordinates of each fragment on the initial sequence are [starting point, starting point+x]. This process is repeated for the creation of 4 libraries of 25, 50, 100 and 200 reads respectively. Each type of library is replicated 100 times. Possible overlapping of the fragments is permitted.

For further testing of the method I then provided as an input sequence the hypervariable mitochondrial D-loop region, to see if I could actually detect a stronger signal. The rest of the parameters were kept the same as before (i.e. the size of the libraries, the size of the reads, the number of reads per library type and the replications of each library). Wanting to test the effect of the library size on the produced probabilities I then augmented the library size at 200, 400, 800, 1000 reads respectively.

b) Real data: Chickens.

For this analysis 21 ancient chicken specimens were selected to be shotgun sequenced, so the dataset contains 21 libraries, one for each sample. Each library was subsequently used as an input for the identification pipeline. The samples come from open air archaeological sites of Africa such as Zanzibar (FK, UU), the Comoros (SMA) and Madagascar (MHLK) and their chronology ranges between the 7th and the 13th century.

Results

Synthetic dataset: Mice

The synthetic mice dataset was analyzed and the accompanying simulated libraries are presented below.

First, I used as an input the whole mitochondrial genome of *Mus musculus domesticus*. The results per subspecies are described below.

NCBI's Taxids	Scientific names	Common name
10091	<i>Mus musculus castaneus</i>	southeastern Asian house mouse
10092	<i>Mus musculus domesticus</i>	western European house mouse
39442	<i>Mus musculus musculus</i>	eastern European house mouse
1643390	<i>Mus musculus helgolandicus</i>	

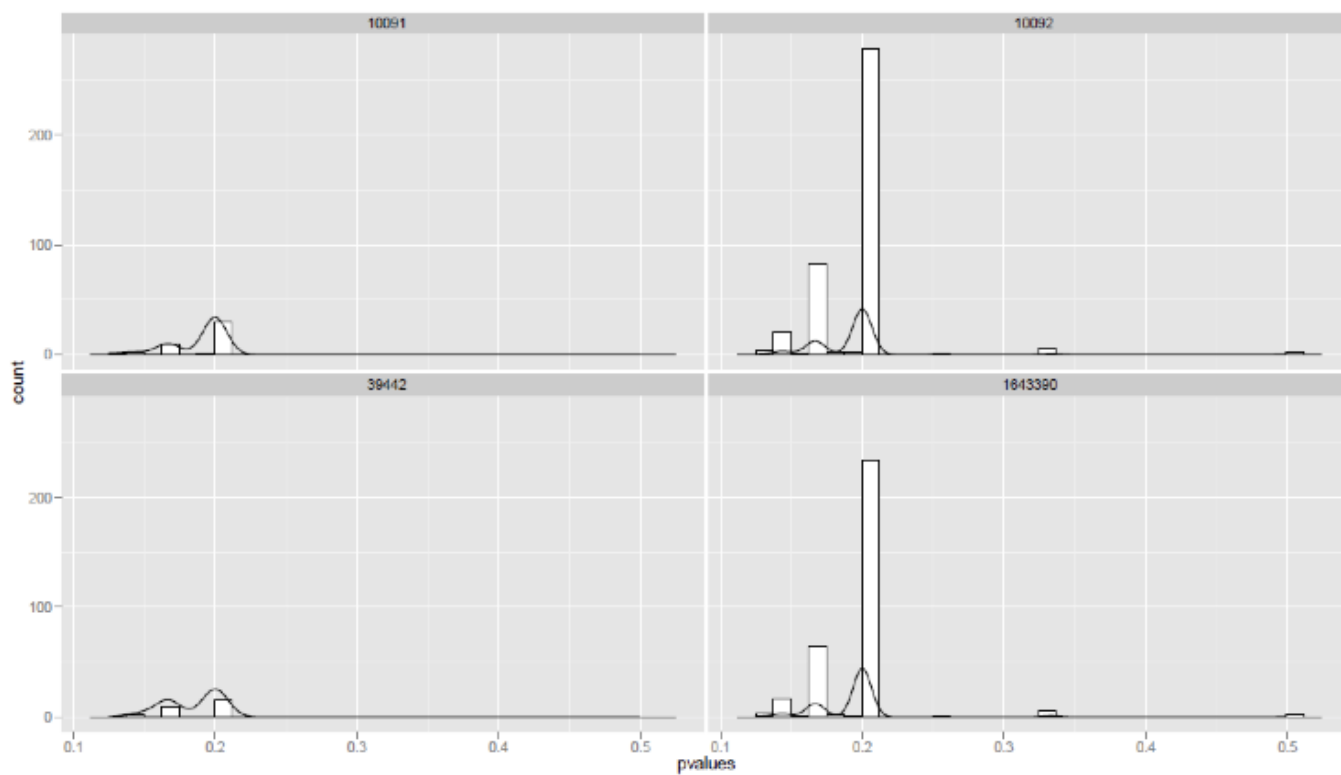


Figure 1. P-values for the whole mitochondrial genome plotted per taxid. No p-value exceeds 0.5 and the majority of them are accumulated around 0.2.

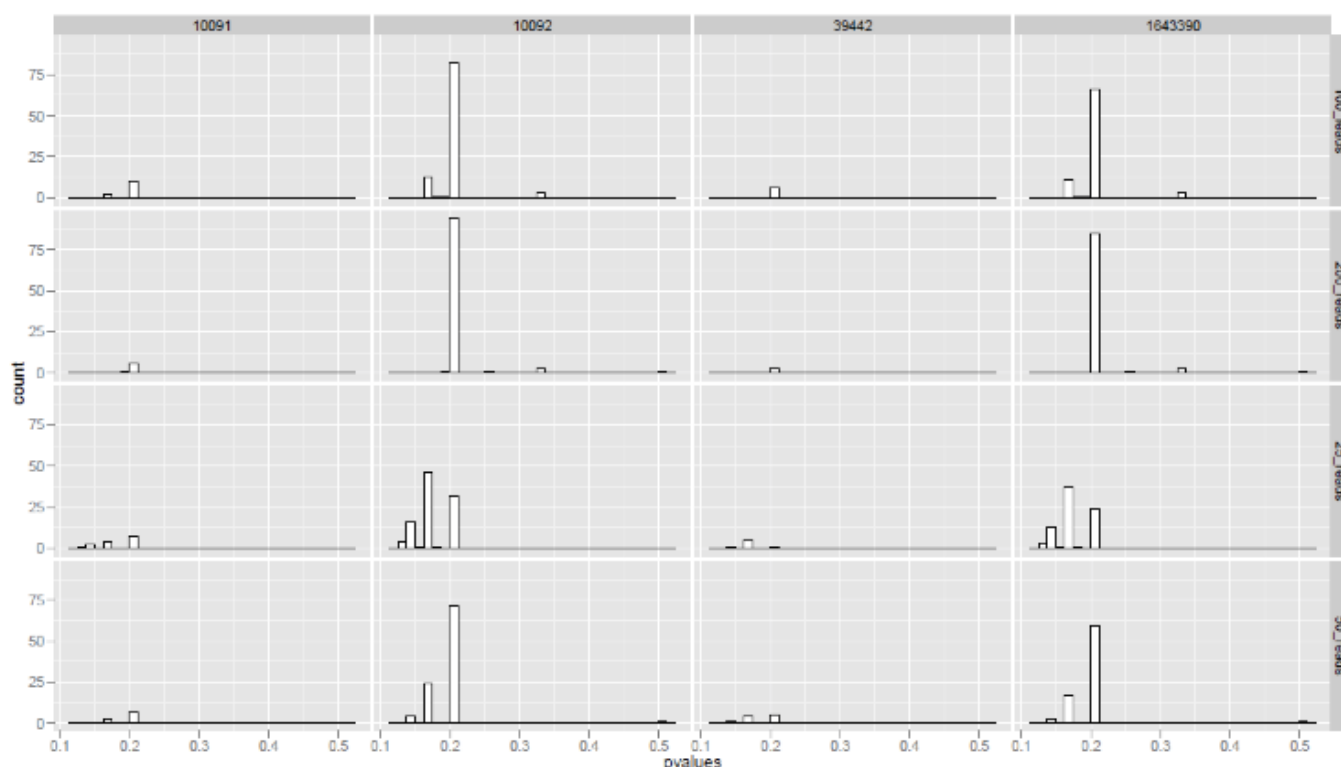


Figure 2. The p-values for the whole mitochondrial genome plotted per taxids and library size. The pattern remains the same as the one in Figure 1.

Initially using ggplot's functions I grouped the produced probabilities per taxid and plotted the frequencies of the p values along with the probability density function of every library. For each library the frequencies of the p-values are usually really low (around 0.2 – Figure 1). Their highest probability is at p-value=0.5. As the plots show most of the values were concentrated around 20%, exceeding it only in the cases of *Mus musculus domesticus* and *Mus musculus helgolandicus*. The probability density function is also generally low.

I then plotted the p-values in a grid of taxids per library size (Figure 2). The pattern remained the same as in the first figure and it was observed that the size of the

library size was slightly affecting the p-values, as it seen from the *Mus musculus domesticus* and *Mus musculus helgolandicus* 200 reads library.

For the next step for the analysis of the synthetic dataset I provided as input sequence the hypervariable mitochondrial D-loop region, to test if I could detect a stronger evolutionary signal. All the other parameters of the analysis remained the same.

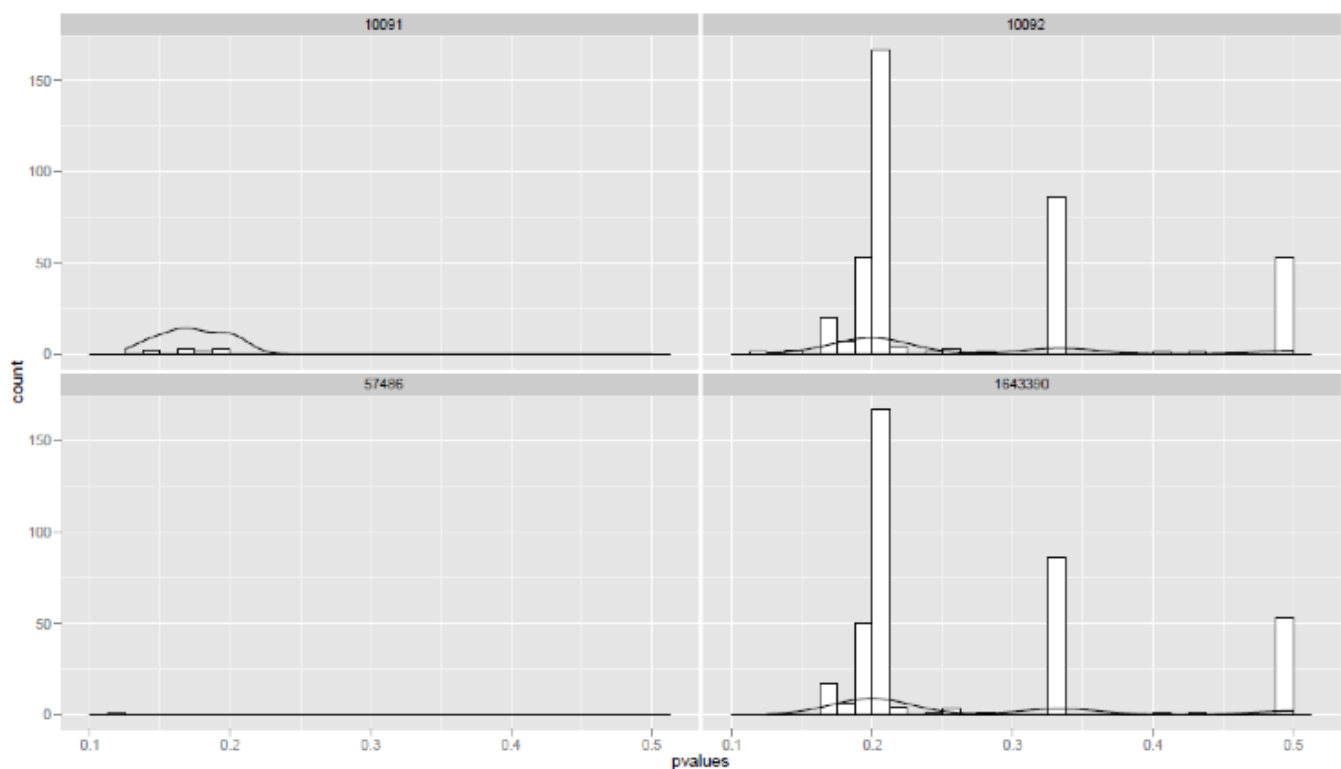


Figure 3. The D-loop region. The subspecies of domesticus and helgolandicus share the majority of the p-values. The p-values do not again exceed 0.5.

The figures of *domesticus* and *helgolandicus* subspecies are highly alike (Figure 3). Searching the NCBI databases I found out that the *domesticus* and *helgolandicus* subspecies share the same mitochondrial genome, due to the very recent evolution of the latter (Babiker & Tautz, 2015). So I redid the analysis, but this time I excluded the *helgolandicus* subspecies from the custom database, so as to include one sequence from each subspecies of the domestic mice. I obtained the p-values and plotted them again using R's ggplot library.

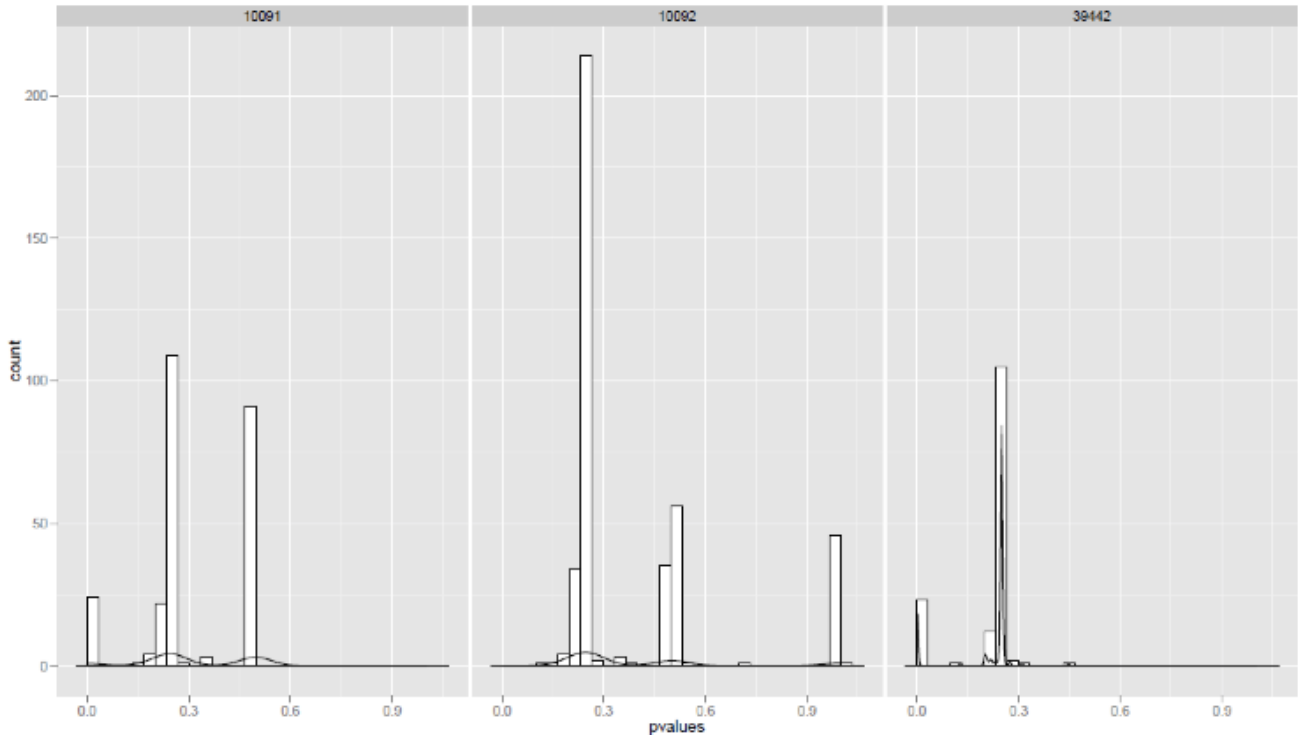


Figure 4. The p-values for the D-loop plotted per taxid. Now the number of subspecies has been reduced to 3. The majority of the p-values accumulate again around 0.2, but the domesticus subspecies has p-values that exceed 0.9. The signal is multiplied.

The number of the major candidate subspecies has been reduced to three, *castaneus*, *domesticus* and *musculus*. The majority of the p-values still accumulate around 0.2, but more p-values start to have values around 0.5. For the *domesticus* subspecies I observe that a few p-values (approximately 50 p-values, as shown in Figure 4) have a value that significantly exceeds 0.9. So by using the D-loop the signal I get multiplies and I get probabilities that allow us to make reliable species assignments.

I then plotted the p-values in a grid of taxids per library size and I observed that as the size of the library increased, the number of p-values that equaled one augmented as well as it is depicted in Figure 5.

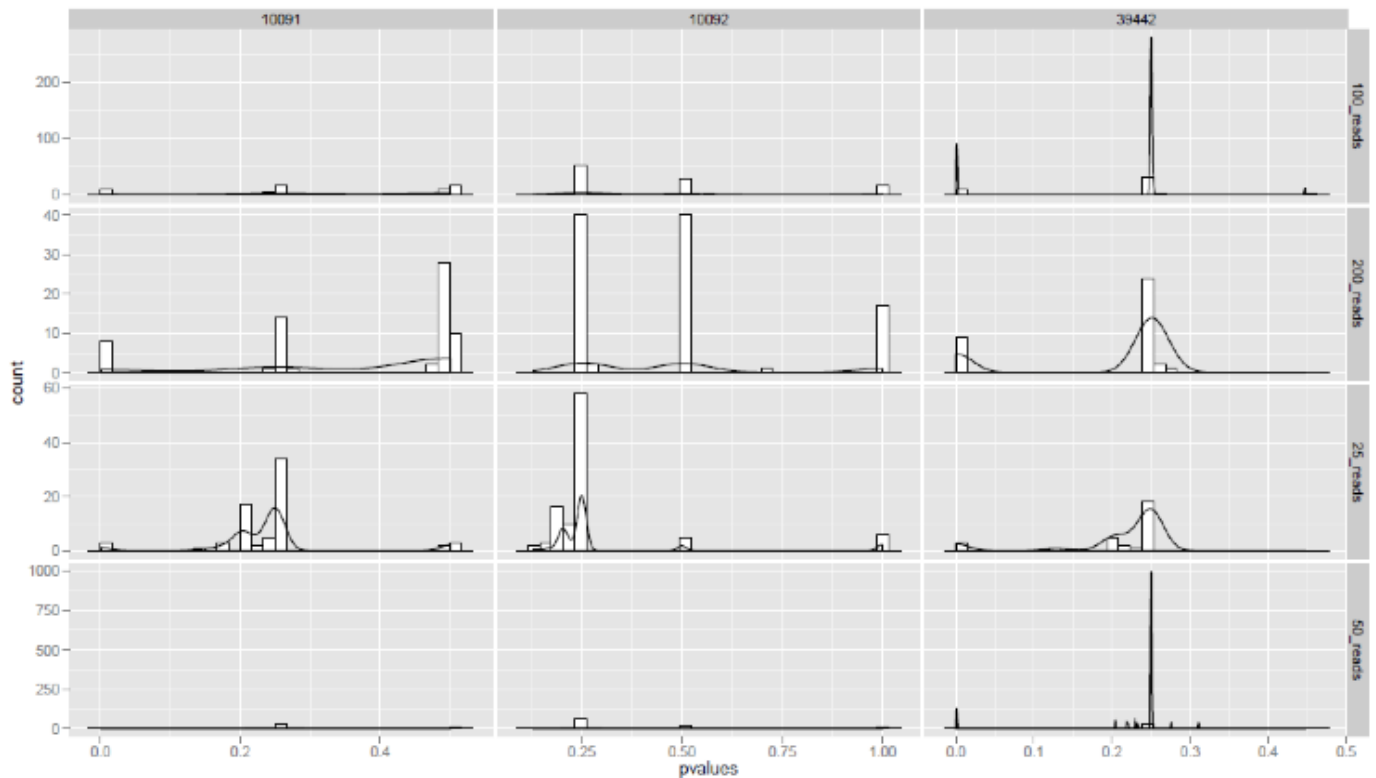


Figure 5. The p-values for the D-loop plotted per taxids per library size. As the size increases the number of p-values that equal 1 increases.

Wanting to test the effect of the library size on the p-values in order, to find the minimum number of reads that yield p-values that enable us to identify subspecies, I rerun the analysis by augmenting the library sizes to 200, 400, 800 and 1000 reads respectively. I plotted the p-values the same way I did for the previous stages.

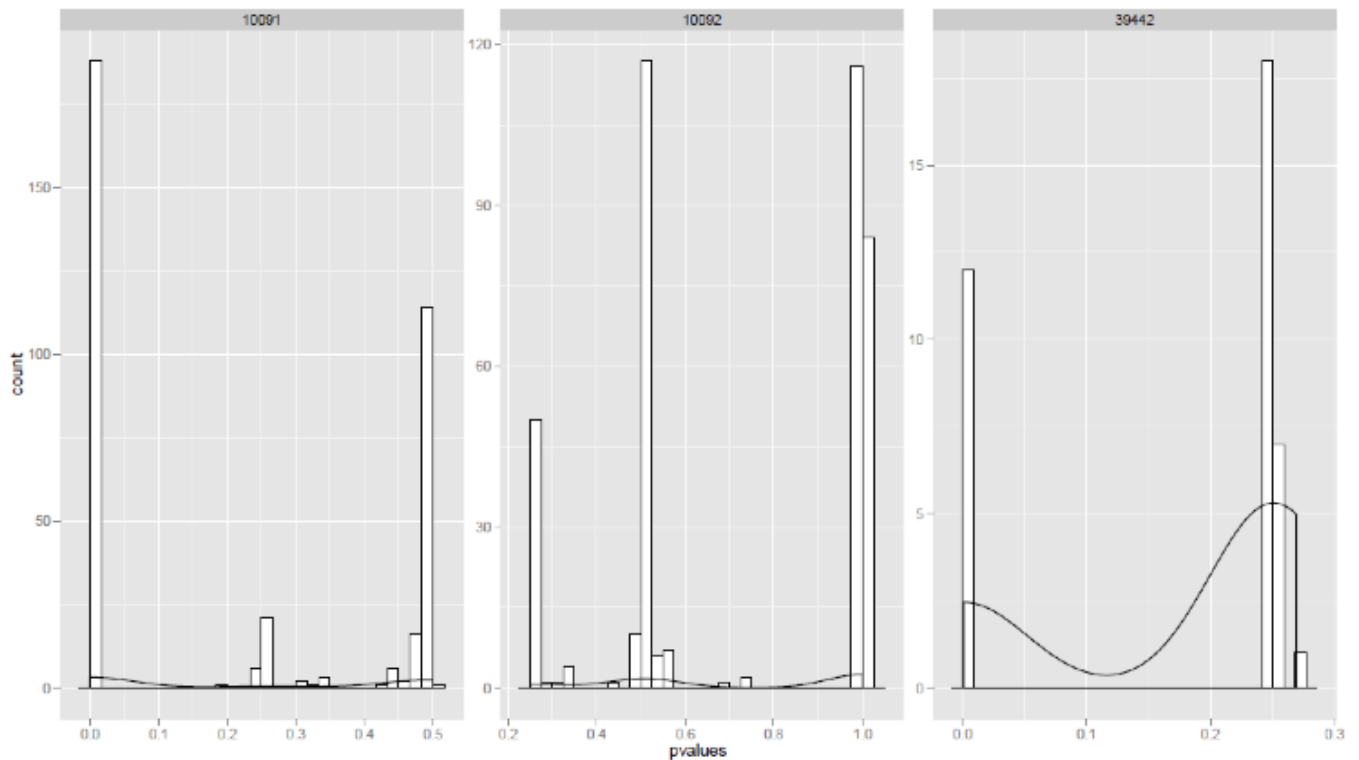


Figure 6. The p-values of the D-loop and increased library size plotted per taxid. The number of p-values equal to one increased.

As the size of the library increased, the number of p-values exceeding 0.5 and being equal to one increased, and especially for the *Mus musculus domesticus* taxid, as Figure 6 portrays. I can see that gradually the signal is multiplied and the discriminatory power of the method augments. Then I plotted the p-values in a grid of taxids per library size.

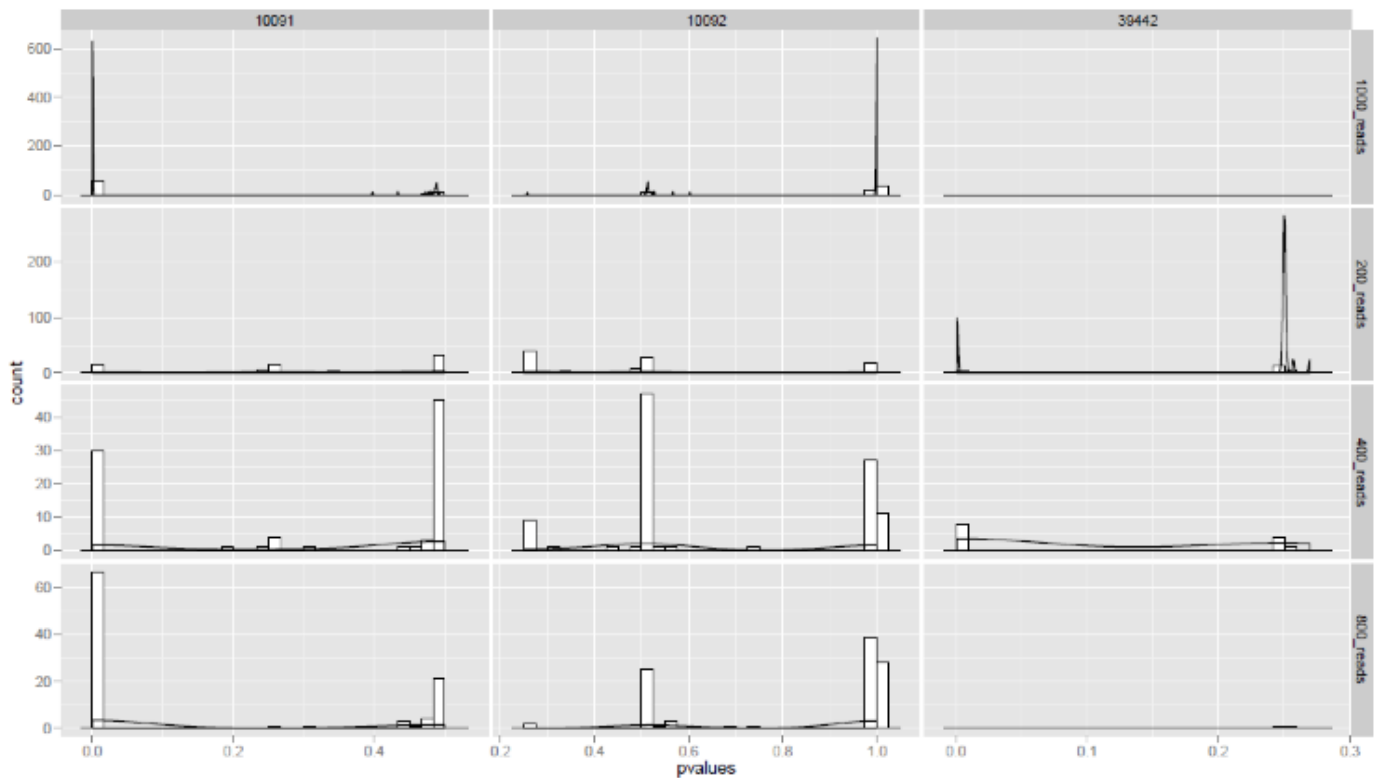


Figure 7. The p-values for the D-loop and increased library size plotted per taxids per read number. As the reads increase the results for the *domesticus* subspecies improve significantly.

As it is evident from the plot (Figure 7) as the number of reads increases, the number of p-values that is equal to one increases as well for the *domesticus* subspecies (10092). For the *castaneus* (10091) subspecies the p-values still don't exceed 0.5 and for the *musculus* subspecies (39442) they don't exceed 0.3.

As we focus on the D-loop and increase the number of the reads each synthetic dataset contains we observe that the p-values start to exceed 0.5 and come closer to one. What also increases significantly is the count of times that p-values are equal to one as the size of the libraries that come from the D-loop increases.

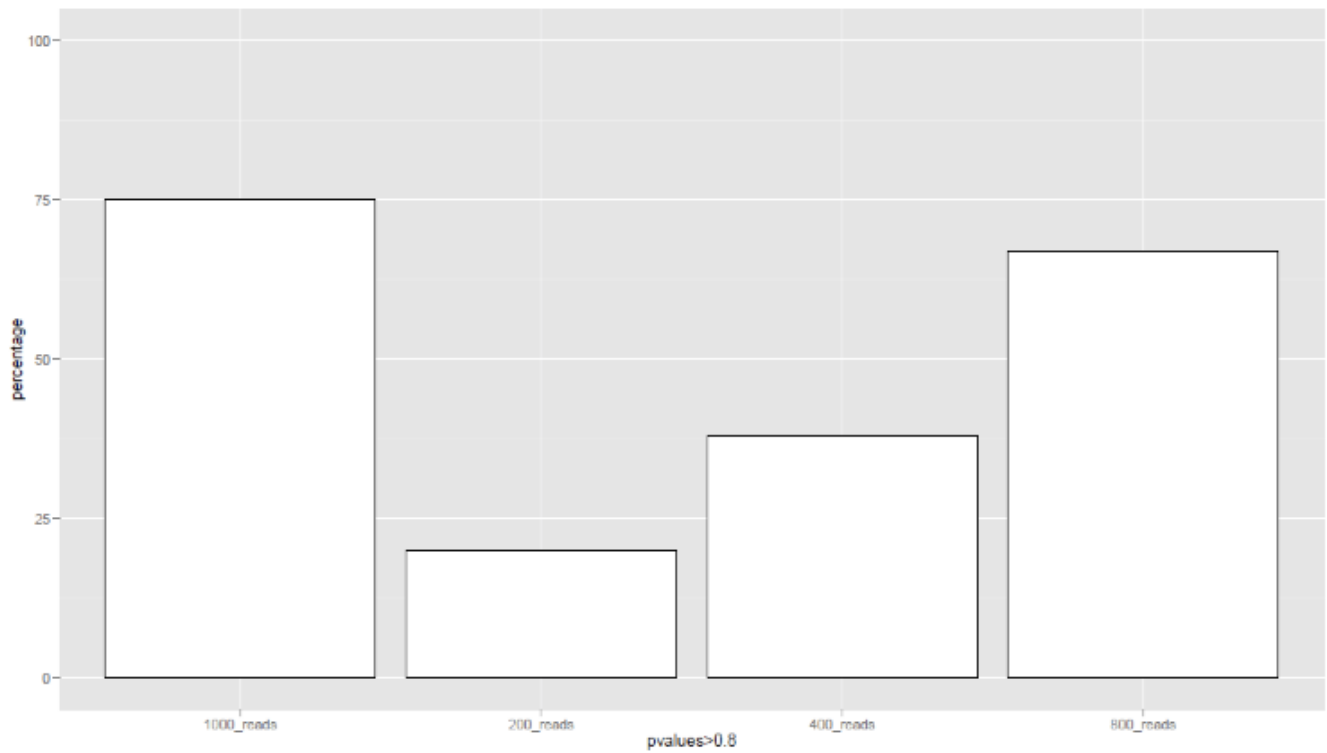


Figure 8. Percentage of p-values that exceed 0.8. All those p-values are for *Mus musculus domesticus*.

Real data: Chickens

Then we moved on to actually test the method on existing chicken data from shotgun sequencing. The failure of some libraries is due to a lack of information, as they might not contain enough reads for the algorithm to compute reliable probabilities or they might contain reads of really short length that produce poor alignments.

Table 1. Table of species assignment using a custom BLAST approach.

Sample code	Regional sample codes	P-value 1	Genus 1	P-value 2	Genus 2
JK2988	UU 11/015	fail	fail	fail	fail
JK2989	UU 10/004	fail	fail	fail	fail
JK2990	KC 1018	fail	fail	fail	fail
JK2991	KC 1015	fail	fail	fail	fail
JK2992	UU 13/002	fail	fail	fail	fail
JK2993	KC 1007	0,997604896	<i>Polyplectron</i>	0,001026137	<i>Numida</i>
JK2994	KC 1003	0,992459168	<i>Syrmaticus</i>	0,002966449	<i>Arborophila</i>
JK2995	UU 1412	1	<i>Gallus</i>	1,12904E-66	<i>Bambusicola</i>
JK2996	SMA 12/13	0,499525561	<i>Gallus</i>	0,498976019	<i>Arborophila</i>

JK2997	UU 11/004	fail	fail	fail	fail
JK2998	UU 11/015	fail	fail	fail	fail
JK2999	UU 13/003	fail	fail	fail	fail
JK3000	FK 12/003	0,535239567	<i>Gallus</i>	0,464760204	<i>Bambusicola</i>
JK3001	KC 1002	1	<i>Numida</i>	7,2382E-55	<i>Acryllium</i>
JK3002	MHLK 2011	fail	fail	fail	fail
JK3003	MHLK 2007	fail	fail	fail	fail
JK3004	MHLK 2005	fail	fail	fail	fail
JK3005	PYS 307	fail	fail	fail	fail
JK3006	KC 1017	fail	fail	fail	fail
JK3007	MHLK 2006	1	<i>Gallus</i>		
JK3008	MHLK 2007	fail	fail	fail	fail

From the 21 samples that were shotgun sequenced and then underwent the custom BLAST analysis 7 (33.3%) of them gave results. Some of those samples, e.g. KC and MHLK samples, gave high p-values that give credibility to this taxonomic assignment. The samples that failed on this analysis failed to produce readable data. A possible reason for that would be degradation of the DNA due to poor sample preservation and low endogenous DNA. Those kinds of libraries gave equivalently low p-values for many genera, due to poor alignments. Moreover, it is possible that

the species that is sequenced is equally related to two species in our custom BLAST database. Thus, we get an equal probability for both related species, which is a value of 0.5.

Discussion

Trying to obtain as much information about the dispersal of animal species that was due to the movement of human populations from archaeological sites is really important for our understanding of animal evolution. Our initial goal was to develop a method that could make optimal use of short DNA reads in order to identify species from ancient samples.

The downgrading system

Our goal was to downgrade the I_g formula by using the penalty value for the missing/poor quality alignment data as much as possible. However, an important limitation that it had to face was that of machine precision, as computers cannot store numbers with infinite precision. Also, another mathematical limitation of the method is that for the p-values to be computed we use products of the power of 10 in our equations. Therefore, we cannot have a denominator that would be rounded to 0, as it would cause division by zero errors. This is a consequence of the insufficient machine precision, as values of the I_g formula that are highly negative will produce a denominator with many decimal digits, thus the denominator will be rounded to 0 and produce a mathematical domain error. Various values were tested as downgrading values, such as the maximum number of 9 decimals (0.99999...) before it was rounded to 1 or the logarithm of the number with the maximum decimal digits ($1e-323$). It was concluded that a value of 0.999 was ideal as it allowed us to have a strict downgrading value, and still to not lose data due to computational restrictions of the p-value formula. However, this value may have to be tuned for different

datasets and for different taxonomic levels. Moreover, the lower the value of the I_g formula, the more conservative the test is.

Assessment of the method

Synthetic dataset: Mice

The results from the synthetic mice dataset showed that our method currently works for distinguishing among subspecies of mice, suggesting they have accumulated enough nucleotide differences. The main problem was the low inter-subspecies variation at mtDNA genome as a whole. The results are by far better when the analysis is focused at the hypervariable D-loop mitochondrial region. The second factor that significantly increases the success of the method and gives p-values of greater value that result in confidently identifying taxa is the size of the initial library. As the number of the reads that a library contains augments the number of p-values that exceed 0.5 augments as well. This is a pattern that can be observed both by the results of the synthetic dataset and the real chicken dataset. Our analysis therefore is not so sensitive at the moment in order to detect this kind of genetic diversity for libraries of low read content. For this kind of analysis a library of 200 reads at least would yield trustworthy results.

From the taxid plots we can see that all the taxa that are recognized are house mice and all are subspecies of *Mus musculus*. The method fails to identify the subspecies of mice when using the whole mitochondrial genome, as the majority of the p-values accumulate around 0.2, (all subspecies are almost equally likely). Nevertheless, the method shows that it is possible to distinguish at the species level (only subspecies

within *Mus musculus* reached high probability). Undoubtedly though there is a certain level of genetic diversity among the different subspecies and species of the Muridae family that allows them to differentiate. The results become more conclusive when we use the D-loop as an input sequence. For the method to succeed when aiming to distinguish samples of organisms among the lower taxonomic levels, sequences with high interspecific variation should be provided as input. Also the contents of the libraries should be reasonably large (generally exceeding 200 reads of 50 bp length). Those recommendations greatly improve the results of this method and yield reliable probabilities that can be interpreted confidently in order to identify the species of each sample. When those guidelines were followed the distinguishing power of the method augments and it works even for the sub-species level, as it happened in our case when we provided the D-loop sequence as an input and increased the number of reads that the libraries contained.

Real dataset: Chickens

As far as the real chicken dataset is concerned the method provided us with better results. The method worked better for distinguishing samples at the genus level, because of the much greater evolutionary depth of this data-set compared to the mice example above. Approximately 33.3% of the chicken samples were taxonomically assigned using that specific methodology. The percentages of the success rate of the method are actually encouraging as they provide better results when compared with other methods that involve the use of aDNA, such as chicken specific primers or 12S gene (Prendergast et al., n.d.). As a method it proved to be

efficient because it provided us with taxonomy results when other standard methods (such as chicken specific primers or 12S fragment) couldn't give an answer to this problem.

The method still has many limitations. The results are not conclusive and lead us to infer that the success of the method highly depends on the variability that exists among the species/subspecies inside a taxonomic group. When compared to another mainstream method used in Zooarchaeology ZooMs, our method has a clear advantage. First, ZooMs use collagen in order to identify the different kinds of animal species. Given the high degree of conservation of the collagen molecules, ZooMs cannot discriminate among subspecies of mice and most likely cannot do it at the species level, or even at the genus level.

More efforts should be made for potential improvement of the method and include components that will increase its sensitivity even when the variability of a taxon is low. Additional ways of extracting information from libraries that contain low amounts of DNA should be developed along with methodologies that make optimal use of really short DNA reads that are contained in DNA libraries of ancient samples. When the samples under examination are ancient or highly degraded, it would be useful to try and enrich our libraries with as many as possible mtDNA molecules. Thus using techniques such as mitochondrial DNA capture, which increases the quantity of mtDNA fragments in our libraries, increases the strength of our test (Slon et al., 2015).

Conclusions

- 1) We have a new method that allows for species identification from small DNA content.
- 2) The method in the end was successful for the mice dataset and it was successful for the chicken dataset as well.
- 3) The mice synthetic dataset mice subspecies test could be considered as the hardest test for this program and the fact that under certain circumstances (high library content, variable mtDNA regions) works is very promising. Thus it is expected to work for most other applications at species/ genus level easily.
- 4) The development of such a method is of great importance because it solves a problem that archaeological researchers have had as it is an unbiased way to compare the likelihood of a bone belonging to one of many potential genera.

Bibliography

- Grealy, A., Douglass, K., Haile, J., Bruwer, C., Gough, C., Bunce, M., 2016. Tropical ancient DNA from bulk archaeological fish bone reveals the subsistence practices of a historic coastal community in southwest Madagascar. *J. Archaeol. Sci.* 75, 82–88.
- Haile, J., Froese, D.G., Macphee, R.D.E., Roberts, R.G., Arnold, L.J., Reyes, A.V., Rasmussen, M., Nielsen, R., Brook, B.W., Robinson, S., Demuro, M., Gilbert, M.T.P., Munch, K., Austin, J.J., Cooper, A., Barnes, I., Möller, P., Willerslev, E., 2009. Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl. Acad. Sci. U. S. A.* 106, 22352–22357.

African Bovids (Minor Chapter)

Introduction

Recent developments in massively parallel sequencing allow for rapid characterization of a large number of sequences in a single run. Massively parallel sequencing is a result of the development of the Next Generation Sequencing methods that have enabled high throughput sequencing. Perhaps more important than the sequencing throughput provided by this technology and its relatively low cost compared with traditional sequencing methods is the type of data it generates. Instead of long reads generated from a PCR-amplified sample, massively parallel sequencing methods provide much shorter reads (~21 to ~400 base pairs), but millions of them. Unlike previous sequencing methods that required DNA amplification (that is, the final sequence was representative of the modal population of DNA templates), sequencing can now be performed from single DNA molecules. The short reads generated in the sequencing of each DNA molecule can be counted and quantified, allowing the identification of mutations in non modal populations of cells (that is, identification of a somatic mutation in a small subpopulation of cells immersed in a modal population with wild-type sequences) and accurate copy number assessment of each genomic region. In addition, with the recent introduction of approaches that allow for the sequences of both ends of a DNA molecule (that is, paired end massively parallel sequencing or mate pair sequencing), it has become possible to detect balanced and unbalanced somatic rearrangements (that is, fusion genes) in a genome-wide fashion.

Not surprisingly, this massive increase in throughput has come at a cost, with the accuracy of each short read being significantly lower than the output generated from

Sanger sequencing. Although this is circumvented by the depth of sequencing (that is, multiple reads of the same region), it is accepted that physical validation using traditional sequencing methods is required. Apart from sequencing whole genomes, massively parallel sequencing can be coupled with DNA capturing methods for focused analysis of specific genomic regions, specific genes or the whole exome (Reis-Filho, 2009).

The major challenge for taxonomy identification based on sequencing data is the selection of marker genes. The 16s rRNA gene in bacteria, the COI mitochondrial gene in animals and the *rbcL* and *matK* genes in plants are a few examples of loci that are informative enough to be used as genetic markers for taxonomic assignments. The unprecedented sequencing depth means that even less abundant species can be detected via their sequences. The taxonomic composition of such samples may be characterized by grouping the sequence fragments (reads) into clusters of operational taxonomic units (OTUs). The number and size of the clusters (OTUs) indicate the diversity of the sample (Bonder et al., 2012).

However those markers in some cases have low discriminatory power, like when the goal is to identify a single species from a mixed sample (Saputra et al., 2015). A possible solution was suggested (Prendergast et al., 2017) to this problem, that employed a combination of shotgun sequencing and the use of customized stand-alone BLAST approach. The libraries of reads produced by the shotgun sequencing were blasted against a specific database and by finding the best possible alignment for each library the taxonomic assignment was then inferred.

The problem of species identification

Taxonomically informative markers (e.g. 16S rRNA gene) have been established as means of identifying species, by using short informative DNA regions (Ratnasingham and Hebert, 2007). The combination of informative markers, next generation sequencing and bioinformatics can constitute a useful toolkit for taxonomy and species identification. Working on species with a well-studied taxonomy promises to reveal the levels and the nature of sequence divergences that typically separate species and advance species discovery in taxonomically understudied groups.

Identifying species from DNA sequences can contribute greatly to many fields of science, such as microbiology, agriculture, taxonomy, metagenomics, archaeology, and consequently have a great economic, scientific and medical/health related impact. By applying metagenomics methods in the field of ancient DNA paved the way towards deciphering ancient elephantid genomes, such as mammoths, and hominid genomes, such as Neanderthal genomes (Schuster, 2008). This, together with the breakthroughs made for sequencing complex DNA mixtures from most diverse sources will allow for the study of any ecosystem of this planet at the sequence level. It will also open a window to the flora and fauna of at least the last 100,000 years, in ways far beyond what would have been deemed possible only with the use of conventional genomic methods (Schuster, 2008).

Archaeology and ancient DNA

The term ancient DNA covers any bulk or trace of DNA from a dead organism or parts of it, as well as extracorporeally encountered DNA of a living organism. Therefore any DNA that has undergone autolytic processes or any kind of fixation is considered to be aDNA. Access to aDNA provides the opportunity to study the

genetic material of past organisms of individual features and evolutionary processes (Hummel and Herrmann, 1994a, 1994b). Examples of such an opportunity include the analysis of DNA recovered from archaeological and historical skeletal material, mummified tissues, archival collections of non-frozen medical specimens, preserved plant remains, ice and permafrost cores. Within living cells, the integrity of DNA molecules is continually maintained by enzymatic repair processes. After the death of an organism, cellular compartments that normally sequester catabolic enzymes break down. As a consequence, the DNA is rapidly degraded by enzymes such as lysosomal nucleases, thus acquiring a certain damage pattern as shown in Figures 1 and 2. The post-mortem instability of nucleic acids is central to the methodological problems inherent in aDNA research. In metabolically active tissues, damage to the DNA molecules is rapidly and efficiently repaired via a host of repair. By contrast, DNA damage accrues over time in inactive (dead or dormant) cells owing to factors such as spontaneous hydrolysis and oxidation, and consequently most ancient specimens do not contain any amplifiable endogenous DNA, while those that do possess only fragments in the 100–500 bp size range. Early research showed that post-mortem DNA decay was characterized by strand breaks, baseless sites, miscoding lesions and crosslinks, and that these caused both sequencing artefacts and the preferential amplification of undamaged contaminant DNA (Willerslev and Cooper, 2005). In addition, the DNA molecule faces an onslaught of bacteria, fungi, and insects that feed on and degrade macromolecules (Pääbo et al., 2004).

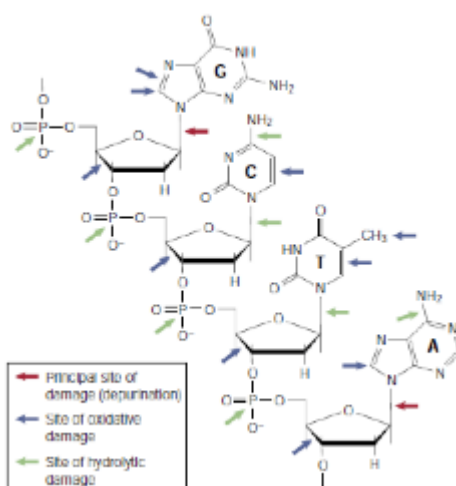


Figure 1. DNA damage shown or likely to affect ancient DNA. A short segment of one strand of the DNA double helix is shown with the four common bases. Principal sites of damage are indicated by red arrows. Sites susceptible to hydrolytic attack are indicated by green arrows and those prone to oxidative damage by blue arrows. G, guanine; C, cytosine; T, thymine; A, adenine (Hofreiter et al., 2001).

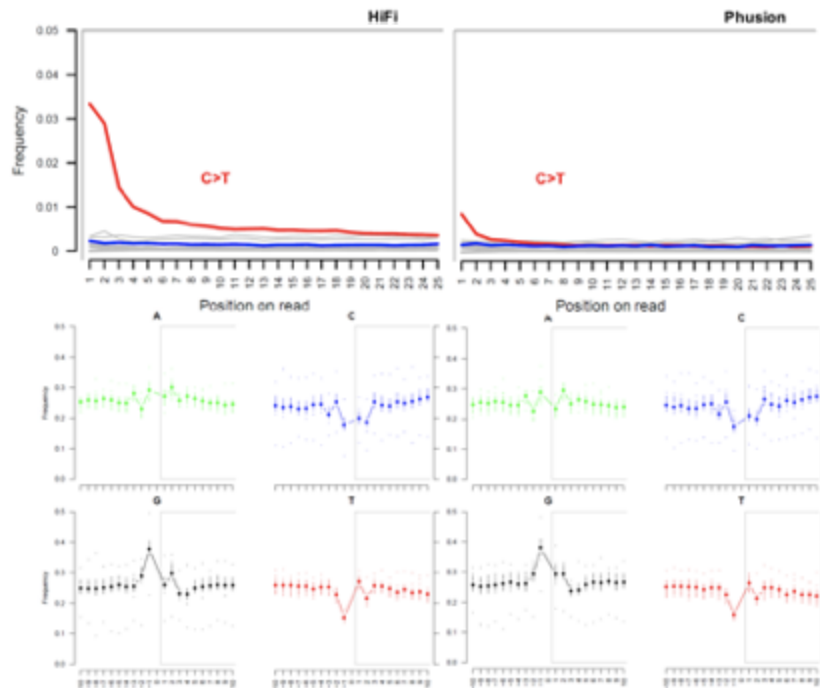


Figure 2. Nucleotide misincorporation at 5'-termini (top) and base composition (bottom) of the paleo-Eskimo individual (Rasmussen et al., 2010). All possible misincorporations are plotted in grey, except for guanine to adenine and cytosine to thymine that are reported in blue and red, respectively. In both graphs, positions are reported with reference to 5'-ends. Nucleotide positions located within reads (top: 1 to 25; bottom: 1 to 10) are reported within a grey frame (Ginolhac et al., 2011).

aDNA enables us to conduct species identification from environmental samples. For instance identifying bulk bone samples from archaeological excavation sites can be really insightful when it comes to documenting the past biodiversity, understanding the evolution and extinction of species and their needs for conservation (Haile et al., 2009). Along with the use of morphology-based approaches, species identification from DNA sequences can greatly enhance the toolkit of bio-archaeology (Grealy et al., 2016).

Identifying species from archaeological samples, such as bones, feathers, fossils, coprolites, can also enhance our knowledge about many aspects pertaining to the population genetics of a certain species. aDNA holds tremendous potential for the study of ancient animal and plant populations. Studies have shown that aDNA can reveal population movements and local extinctions back into the Late Pleistocene. Such studies have considerable power to examine the effects of climate change (e.g. around the Last Glacial Maximum) and to test theories and develop methods used in population genetics and palaeobiology. Studies of ancient populations have repeatedly shown that modern patterns of distribution and diversity are often owing to relatively recent events, and can give very misleading views of past evolutionary processes. Furthermore, studies of several ancient populations of large vertebrates have shown extensive gene flow between taxa recognized as different species or subspecies, suggesting that the morphological plasticity in some fossil groups may have been considerably underestimated. The combination of ancient sequences and coalescent methods has considerable power to reconstruct detailed demographic histories, test models of population genetics and reveal much novel information about microevolutionary processes. These methods can even recover demographic data for taxa that have been through recent population bottlenecks, which would normally remove genetic signals. aDNA also enables us to trace the domestication processes of many animals and plants (chickens, dogs, pigs, bananas) (Lejju et al., 2006) and also the movement of species across distant geographic areas as a result of movement of human populations (Prendergast et al., 2017).

A major trend in species identification is to use PCA based methods of analysis to distinguish between species. The Principal Component Analysis (PCA) is a multivariate technique that allows one to find and plot the major patterns within a

multivariate dataset e.g. multiple loci and multiple samples and the relationship between the distance matrix elements based on their first three principal coordinates. Basically PCA technique concentrates the maximum variance of the data in a reduced number of independent variables, called principal components (PCs). A significant portion of the variance is represented by the first PCs, in other words, higher differentiation among the genetic distances, and the last PCs carry uncorrelated information or noise. PCA plots (2D and 3D) and hierarchical cluster analysis (HCA) dendrograms are convenient representations showing naturally occurring group memberships via these objective classification methods.

Along with other complementary methods, such as amplification with species specific primers or ZooMs analysis, shotgun sequencing in combination with a moderately simple bioinformatic analysis can be of great utility for species identification. The use of a whole genome can be much more informative for inferring taxonomic relationships among organisms than a set of genetic markers. This is because when PCR based methods are used one can only look into one locus or a limited set of them when more loci are multiplexed. The amplicons also have to be quite long so that a robust analysis is conducted, a fact that generally is not the case for aDNA, where the DNA fragments are short and highly fragmented due to post mortem damage. Also it provides the analysis with more discriminatory power when it comes to distinguishing between closely related species, as highly conserved genetic markers wouldn't be differentiated enough. Whole genome sequences are more suitable and informative for genetic structure analyses than a set of short sequence segments, facilitating the study of the evolutionary course of a certain organism or the study of domestication events. However analyzing those fragments requires more sophisticated bioinformatic analysis methods.

Introduction to the method

The libraries that are produced by shotgun sequencing often contain a large proportion of short reads, which make genomic analysis more difficult and often they cannot be used for identifying the species that they come from. This problem is accentuated when the DNA libraries come from ancient samples, where DNA degradation is substantial and hinders further the analytical methods. For those reasons it seemed fit to find new ways to make use of these short reads for the purpose of species identification.

The DNA data that can be analyzed are either nuclear or mitochondrial. The potential of nuclear DNA data is limited, as the size of the nuclear genomes is too big and there is low interspecific variation. Furthermore, only a few species have full genome data available. Mitochondrial DNA on the other hand is highly variable among different species and there are more full sequence data available. For instance, currently on the NCBI databases there are approximately 1260 full mitochondrial genomes available, whereas there are only 68 nuclear DNA bird genomes. Additionally, a great advantage of mtDNA, when it comes to ancient samples, is that it is often over-represented in ancient DNA samples, so proportionally more reads of ancient mtDNA are contained in a library than of ancient nuclear DNA.

Materials and methods

Species Identification Pipeline Method

A common hurdle in many species identification methods via DNA is misidentification, due to the use of the descriptive statistics that are employed by them. As a result, taxonomic identification results from such methods cannot be validated, leading to uncertain sample identifications. This problem becomes prevalent when analysing metagenomic samples. In our proposed method, we employ a hypothesis driven approach by asking if our sample is species X, and therefore we provide a probabilistic estimation of how well supported our taxonomic assignments are. It needs to be underlined that a low assignment probability for species X does not disprove that our sample doesn't belong to that species, but it should be interpreted as our sample's yielded DNA sequences not containing enough information to perform a robust species identification.

In order to identify a certain species robustly, we first have to obtain a relative alignment score for each combination of DNA read and taxon of interest. A means of doing that is using SIGMA's alignment score matrix (Ahn et al., 2015). SIGMA, by using the bowtie2 (Langmead and Salzberg, 2012) aligner, aligns the reads of our DNA library against the genomes of our taxa of interest and produces an alignment score, called a Q score which is an evaluation of the quality of the alignment. The higher the alignment score, the better the quality of the alignment, and the probability of the sequences being a true match is higher.

A library of ancient DNA though contains a high number of short reads, therefore contains many query sequences. Consequently, it is of great use to obtain an overall score across all the reads of a library per taxon of interest. For this summarization to

happen we make use of the following equations. The overall alignment score for each library/species or library/subspecies combination is given by the formula: . This is the sum of the alignment scores (Q scores), that evaluate the quality of an alignment between a DNA read and a sequence of a specific taxon.

Implementation

- a) Initially the questions “Is my sample species X?” or “Does species X exist in my sample?” need to be defined. Therefore, a reference genome of the species/taxon X needs to be downloaded. If species/taxon X is a group of taxa (eg. all mammals), all the genomes of that group can act as a collection of reference genomes.
- b) The pre-processed reads (adapters removed and quality trimmed) included in the DNA library of the sample under study are aligned with SIGMA against the reference genome, by using bowtie2. Only reads which produce an alignment with the reference genome, whose score is above the specified quality threshold, are kept. The rest of the DNA reads are discarded. Sequencing adaptors should have been removed prior to this step.
- c) The selected, potentially positive-hit DNA reads are filtered further by removing duplicate reads. For the duplicate removal, both start and end coordinates of a DNA fragment are taken into account.
- d) After obtaining our filtered unique reads, a damage pattern analysis is performed on them by MapDamage (Ginolhac et al., 2011), to assess if our DNA reads are truly ancient and what is the extent of the damage they have undergone.

e) A crucial step to implement this method is to build a custom RefSeq-like database and include the taxon of interest or a taxon that can be used as a proxy. Such a taxon is one that is phylogenetically closer to the taxon of interest than all other sequenced taxa (Figure 3). In order to build a database that actually enables us to produce reliable probabilistic estimations of identity a few problems need to be tackled. First, the size of the database needs to be reasonably large, otherwise the Q scores will be distorted and unreliable. This is because when a database is small the probability of observing a match by chance increases. Also the size of the sequences that are contained in the database should be of similar length, otherwise the Q scores would be biased. Having target sequences that differ greatly in length could lead to Q scores that could be hard to interpret, as larger sequences would allow more alignments to take place than shorter sequences. Another hurdle for this kind of analysis is the unequal representation of the taxa in the database. This is because many taxa have been sequenced more thoroughly than others, and therefore there are many more publicly available sequences for these taxa. For instance, if we are interested in identifying bones of vertebrate animals the first two problems can be overcome if the database is built with the sequences of whole mtDNA genomes from many potential species. For example, if the query sequences come from mice then the database will include the full mtDNA sequences of all the species of the Muridae family. This approach will give the database the size needed for the Q scores to make sense in terms of probability. Furthermore, given that the mtDNA genomes are approximately of the same size, the issue of partial

alignments will disappear. Likewise, most species of bacteria in the same genus will also have very similarly sized genomes, i.e. all species in the genus *Streptococcus* have a genome size of ~2Mbp. Though, if the sample under study is a metagenome, we might need to readjust the size and the number of taxa included in the custom database, depending on the available data for the taxa of interest. For the third major problem to be tackled, and ensure equal taxonomic representation in the custom-made database, only one sequence per taxonomic level is included. For instance, if we want to perform an identification on the species level, the database is going to include one sequence per species, but if we want to increase our resolution and identify the subspecies of an organism our database will include one sequence per subspecies.

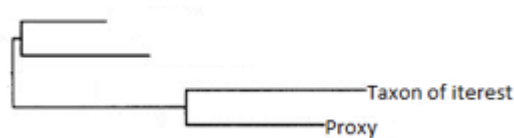


Figure 3. Example of using a taxon as a proxy to infer for closely related taxa of interest.

- f) Afterwards the library of the query sequences is aligned against the custom-made database, using SIGMA. The product of this alignment is a matrix (Q matrix) generated by the SIGMA algorithm, which contains the alignment scores for every read/taxon combination. Alignments, whose quality is lower than a predefined threshold, are given the value of 0 and they do not get reported in the matrix.

g) Then an in-house developed algorithm that sorts the SIGMA Q matrix and computes the probability scores is employed:

- i) Create a database/dataframe from SIGMA's Q-matrix. The rows are representing the DNA reads of our sample and the columns are the different taxa that are included in our custom-made database. After the data frame is constructed, the sum of the Q scores for every column (taxon) is computed.
- ii) The vector is sorted according to the values of the taxa in descending order.
- iii) Then the I_g values (the overall alignment score) for each genus are computed.
- iv) The vector of the I_g values of each taxon is sorted again and the two largest I_g values are chosen.
- v) Then the relative score for each best and second best taxon, which can be interpreted as a posterior probability, is computed with the equation: $p_g = I_g /$

$$\sum_{i=1}^n 1/I_i.$$

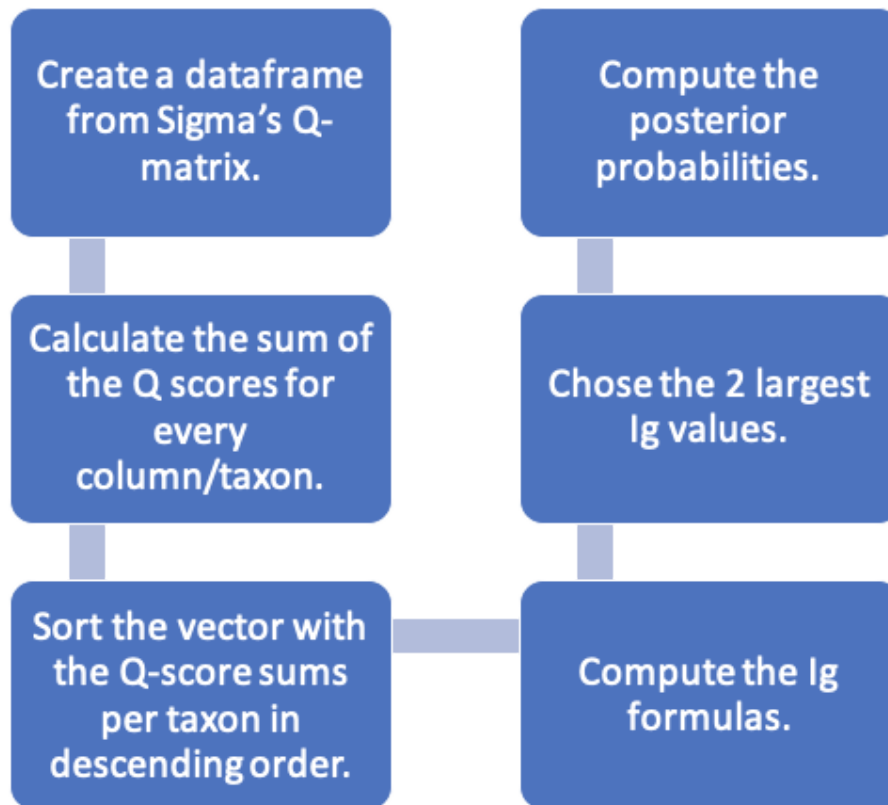


Figure 4 Workflow of the algorithm that parses Sigma's Q matrix.

- h) After the posterior probabilities have been calculated, the 10 first taxa from the vector of I_g values are chosen. These taxa are used to subset the initial dataframe that was constructed, and create a new one only with the columns of interest. The alignment scores of every read per taxon are then plotted, and a nice visualization of the distribution of the alignment scores per taxon is created.
- i) Following the identification of the best taxon, the DNA reads that are aligned specifically to the genome of that taxon can be extracted and be used as input for a second round of damage pattern analysis. This second

step of MapDamage analysis allows us to corroborate that the reads on which the identification of the best species was based are truly ancient.

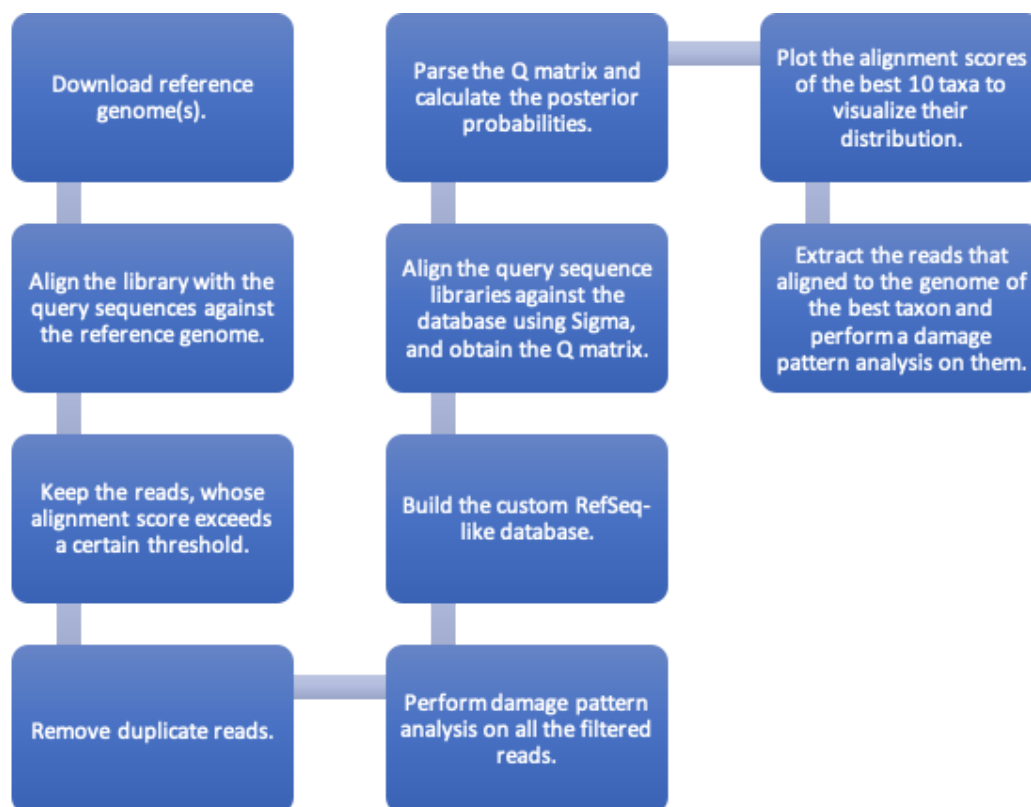


Figure 5 Workflow of the species identification pipeline.

Notes on the construction of the local database

Some additional steps for the construction of the custom-made database need to be highlighted. Initially a search query is posted on NCBI, requesting accession numbers of complete genome sequences (sequence length parameters are often specified). After obtaining a list of accession numbers, their metadata – including information about the sequence records like length, along with various taxonomic information – are downloaded. Depending on the taxonomic level somebody wishes to focus on (e.g. genus, species, subspecies, etc.), only sequence records that have

been taxonomically characterized on that specific level at least are chosen. The rest of the accession numbers are filtered out. To maintain equal species representation in our database, and avoid as much bias as possible at the alignment stage, only one sequence/accession number is selected per taxon, usually the longest one. By the end of this process, a final list of taxa with their respective accession numbers is outputted, and these accession numbers are used to construct the database using a structure based on NCBI Ref-Seq.

Species identification in metagenomic analysis

When our sample is for example a bone, the question that has to be answered is “Is my bone species/taxon X?”. In many cases though, our sample will be containing more than one species – i.e. it is a metagenomic sample. For example, a bone sample could easily be contaminated with soil organisms, microbes left from infection/circulation during life, and in other cases the sample may be a preserved densely-populated microbiome sample such as dental calculus or paleofeces. In this case Identification of a species doesn’t mutually exclude the existence of other species in our sample, so our question can be modified to “Does species/taxon X exist in my sample?”. Therefore, metagenomic samples are analysed in two steps.

Initially the whole of the alignment score matrix (Q matrix) is parsed, and the posterior probabilities are calculated by taking into account all of the reads that were used as input for SIGMA. After this first step, we concentrate on a selected group of closely-related taxa that include the species of interest. We subset the initial Q matrix independently for each of the selected species. Each species subset contains only reads whose best alignment hit is with said species. Then new posterior probabilities

are calculated for only this subset of the matrix, by using the same workflow as described before.

Interpretation of the results

The posterior probability that is computed should not be considered as the probability that the sample under examination belongs to the corresponding taxon, but rather to a species that is most closely related to this taxon. It is a rather conservative method of assigning species to samples, as if a sample each time will produce a high probability for the species that is the most closely related to and really low probabilities for the remaining taxa of our database. A reliable species assignment would take place, as a rule of thumb, if the posteriors that are produced exceed 0.5, depending every time on the taxa that we try to identify and their biology. Subsetting the matrix independently for each species is crucial for confident identification of the species of interest. This is because after obtaining the posteriors from the whole of the matrix we can assess which ones are the major taxa in our metagenomic sample, yet many times the signal that is being picked up is weak, and the resulting posteriors might not be high enough to perform a robust identification. Consequently focusing on one species every time, by analyzing the alignment scores from the reads whose best match was the species/taxon of interest, allows us to mathematically address the question “Is my sample species X?” and therefore increase the sensitivity of our species identification method.

An important limitation of our method is the amount of information that a certain library contains. Libraries with significantly short reads or low amounts of endogenous DNA can lead to equally low posteriors for many taxa. In the case that our sample is or contains a species that doesn't exist in the custom database, then

this library will also get roughly equally low posteriors for all the taxa in our database. Another special case is one where our sample contains a taxon, which is evolutionarily equidistant/equally related to a certain number species/taxa in our database. As a result (relatively high) equal posteriors will be calculated for these certain species/taxa and really low probabilities will be produced for the remaining taxa of our database.

Applications

- a) Single organism sample – one species: When there is a bone assemblage with highly fragmented bones that have lost their diagnostic zones, but it can still be gauged what kind of animal they used to be. By asking the question “Is my sample species X?” a taxonomic identification can be performed. The reference genome for this case of identification is from one species only. The posterior probability produced provides an estimation of how likely this identification is.
- b) Single organism sample – multiple species: The main difference from the previous case is that the bones of that assemblage have lost a great deal of their morphological features, so that an expert can only assign them to a larger taxonomic group, e.g. mammals. In this case, our reference genome consists of the genomes of all the different species in this larger taxonomic group, in our case it’s going to be all the mitochondrial genomes of the mammals. Therefore, the question posed this time is “Is my sample any of these taxa?”.
- c) Metagenomic sample – one species: When a sample contains more than one species in it, like dental calculus – a currently understudied archaeological material, it is of great value to be able to identify the

species content of that sample, but at the same time try and get an estimation for this identification. The best fitting question for these studies would be “Does species X exist in my sample?”. Consequently, the reference genome will be from only one species, as we are trying to identify one specific species.

Case Study - The Kalahari Desert debate

Lively debate surrounds the introduction of non-indigenous domestic livestock to southern Africa (Horsburgh et al., 2016a). More importantly it focuses around two major issues of debate: the processes by which animal husbandry was adopted by formerly foraging people and the nature of the interactions between food-producing and foraging groups occupying the same landscape.

Named the Kalahari Desert debate, because initially it was focusing on the relations of the Bushmen of the Kalahari Desert (Barnard, 1992; Lee, 1979; Lee and Guenther, 1993; Marshall, 2014; Sadr, 1997) with the surrounding populations of agropastoralists. This debate later expanded in the Western Cape Province of South Africa (Schrire, 1992; Whitelaw et al., 1992; Yates and Smith, 1993). A common feature in both these areas is that there is a shortage of robust archaeological evidence (Sadr, 1997; Yellen and Alison, n.d.) regarding the contact between agropastoralist and foraging groups of people.

A meaningful proxy that can act as evidence of contact between foraging and agropastoralist cultures is the existence of domestic animals in places where foraging people live. By identifying species of domestic animals in faunal remains assemblages in areas which were occupied by foragers, like Sehonghong, can act as proof for the advent of domestic animals, animal husbandry and other Neolithic practices in these areas.

Motivation and previous studies

A recent study (Horsburgh et al., 2016a) was published, that focused on identifying the species of various bone samples originating from the area of Sehonghong, using aDNA as a tool. These samples previously had been morphologically identified as

domestic animals, specifically cattle (*Bos taurus*), sheep (*Ovis aries*) and goats (*Capra aegagrus*) (Mitchell, 2009; Mitchell et al., 2008). In said study all but one sample were identified as wild animals, mainly common elands, rheboks and mountain reedbucks, instead of domesticated ones. As a result, the authors of that study (Horsburgh et al., 2016a, 2016b) claimed, by using aDNA as a means of evidence, that the frequency of domestic animals in the faunal assemblages has been significantly overestimated, and that, consequently, the nature of the relationships between highland foragers and agropastoralists has been misconstrued.

Results

The following table is an overview of the results of the study carried out by (Horsburgh et al., 2016a), containing information about the zooarchaeological identifications, the identifications performed via aDNA and information about the bone materials.

Table 1 Different species identifications of the Sehonghong samples using different methods.

Sample code	Zooarchaeologic ally identified species	Identified species via aDNA by Horsburgh et al	Bone Material
SHH_7355	<i>Ovis aries</i>	<i>Redunca fulvorufula</i>	Distal metapodial fragment
SHH_7435	<i>Ovis aries</i>	<i>Pelea capreolus</i>	Right lower molar

SHH_7459	<i>Ovis aries</i>	No DNA recovered	Right proximal lateral radial fragment
SHH_7806	<i>Ovis aries</i>	No DNA recovered	Proximal metacarpal fragment
SHH_7807	<i>Ovis aries</i>	<i>Tragelaphus oryx</i>	Proximal metacarpal fragment
SHH_7246	<i>Ovis/Capra</i>	No DNA recovered	Pelvis ischium/acetabulum
SHH_7315	<i>Ovis/Capra</i>	No DNA recovered	Pelvis spine
SHH_7319	<i>Ovis/Capra</i>	No DNA recovered	Left mandible
SHH_7785	<i>Ovis/Capra</i>	No DNA recovered	Metacarpal fragment
SHH_7416	<i>Ovis/Capra</i>	<i>Pelea capreolus</i>	Left 2nd and 3rd carpals
SHH_7135	<i>Bos taurus</i>	No DNA recovered	Distal 1st phalanx fragment
SHH_7227	<i>Bos taurus</i>	No DNA recovered	Proximal 2nd phalanx fragment
SHH_7297	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	Left upper P3
SHH_7230	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	Distal tibia fragment
SHH_7356	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	Distal 1st phalanx fragment
SHH_7357	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	Complete proximal sesamoid

SHH_7358	<i>Bos taurus</i>	<i>Bos taurus</i>	Distal metapodial fragment
SHH_7449	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	Os centroquartale fragment
SHH_7450	<i>Bos taurus</i>	No DNA recovered	Complete proximal sesamoid

After obtaining the raw DNA sequences from the study that was carried out by Horsburgh et al, the DNA reads were reanalysed by employing the proposed species identification method. The identified species along with their respective posterior probabilities are presented briefly in the following table (for a detailed version of our results please refer to the appendix).

Table 2. Species identification results.

Sample code	Zooarchaeologically identified species	Identified species via aDNA by Horsburgh et al	Identified species via aDNA by the presented method	Posterior probability for the performed identifications	Bone Material
SHH_7355	<i>Ovis aries</i>	<i>Redunca fulvorufula</i>	<i>Redunca fulvorufula</i>	0.64	Distal metapodial fragment
SHH_7435	<i>Ovis aries</i>	<i>Pelea capreolus</i>	<i>Pelea capreolus</i>	0.88	Right lower molar

SHH_7459	<i>Ovis aries</i>	No DNA recovered	No data	NA	Right proximal lateral radial fragment
SHH_7806	<i>Ovis aries</i>	No DNA recovered	No data	NA	Proximal metacarpal fragment
SHH_7807	<i>Ovis aries</i>	<i>Tragelaphus oryx</i>	<i>Tragelaphus oryx</i>	0.53	Proximal metacarpal fragment
SHH_7246	<i>Ovis/Capra</i>	No DNA recovered	No data	NA	Pelvis ischium/acetabulum
SHH_7315	<i>Ovis/Capra</i>	No DNA recovered	<i>Homo sapiens</i>	0.17	Pelvis spine
SHH_7319	<i>Ovis/Capra</i>	No DNA recovered	No data	NA	Left mandible
SHH_7785	<i>Ovis/Capra</i>	No DNA recovered	No data	NA	Metacarpal fragment
SHH_7416	<i>Ovis/Capra</i>	<i>Pelea capreolus</i>	<i>Pelea capreolus</i>	0.87	Left 2nd and 3rd carpals
SHH_7135	<i>Bos taurus</i>	No DNA recovered	<i>Dipus sagitta</i>	0.43	Distal 1st phalanx fragment
SHH_7227	<i>Bos taurus</i>	No DNA recovered	No data	NA	Proximal 2nd phalanx fragment
SHH_7297	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	<i>Tragelaphus oryx</i>	0.58	Left upper P3

SHH_7230	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	<i>Tragelaphus oryx</i>	0.84	Distal tibia fragment
SHH_7356	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	<i>Tragelaphus oryx</i>	0.79	Distal 1st phalanx fragment
SHH_7357	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	<i>Tragelaphus oryx</i>	0.77	Complete proximal sesamoid
SHH_7358	<i>Bos taurus</i>	<i>Bos taurus</i>	<i>Bos taurus</i>	0.44	Distal metapodial fragment
SHH_7449	<i>Bos taurus</i>	<i>Tragelaphus oryx</i>	<i>Tragelaphus oryx</i>	0.52	Os centroquartale fragment
SHH_7450	<i>Bos taurus</i>	No DNA recovered	No data	NA	Complete proximal sesamoid

Discussion

The limits of aDNA

By skimming through the previous table it is easy to observe that the results produced by our method are mostly in agreement with the results produced by (Horsburgh et al., 2016a), and the species identifications are supported by quite robust posterior probabilities.

A major part of our species identification method is to try and make sure that the performed species identifications are true. Therefore, the DNA reads that are used have their damage pattern assessed to ensure that they are ancient, and the

alignment score distribution is visualised in order to observe the quality of the alignments that occurred for the identified species.

Taking into account these quality control steps, the DNA sequences derived from the Seihonghong samples in many cases were not exhibiting damage patterns typical to aDNA. Additionally, the vast majority of the alignments of these reads were of really poor quality. In one or two samples (SHH7315 and SHH7358) contamination is also a possibility. Consequently, it can be concluded that the identification results using aDNA are inconclusive. For results of higher quality DNA resequencing of these samples would be the most helpful.

The described species identification method highly depends on the informational content of the DNA sequences that are yielded from a sample. Highly fragmented and damaged DNA molecules, along with DNA molecules from repetitive or hyper-conserved regions, do not have informative enough sequences to be able to distinguish between different taxa. Having a DNA library rich in polymorphic, diverse and informative sequences increases the power of our analysis by increasing the resolution of our taxonomic identifications, allowing us to go as low as the subspecies level in some cases.

The biology of the taxa that we are trying to identify should be also considered. Genetic diversity the taxa under study is a key factor to a confident species identification. Closely related species for example can be a hindrance to our analysis, as their DNA sequences might not be different enough to allow us to distinguish between them. This can be aggravated by not having insufficient or uninformative DNA reads in the libraries that are being analysed.

For a robust species identification to take place a well curated database needs to be in place. The database needs to include as many related taxa as possible to the chosen candidate species that we are trying to identify. Equal species representation in the database needs to be ensured, as well as making sure that the sequence records have total lengths of similar order of magnitude. Attention needs to be drawn to 2 special cases that relate to the contents of the database, which require special interpretation of the identification results. The case is when our sample comes from a species/taxon whose sequence record isn't included in the database. This will lead to yielding many equally low assignment probabilities for most of the taxa in our database. The second special case is when trying to identify a taxon, which is highly genetically similar with its related taxa. In this case it's possible that our method cannot distinguish among them, therefore yielding equal probabilities for all these taxa.

The described method is quite conservative in assigning a taxonomic identification to a sample. More efforts should be made for potential improvement of the method and include components that will increase its sensitivity even when the variability of a taxon is low. Additional ways of extracting information from libraries that contain low amounts of DNA should be developed along with methodologies that make optimal use of really short DNA reads that are contained in DNA libraries of ancient samples. When the samples under examination are ancient or highly degraded, it would be useful to try and enrich our libraries with as many as possible mtDNA molecules. Thus using techniques such as mitochondrial DNA capture, which increases the quantity of mtDNA fragments in our libraries, increases the strength of our test (Slon et al., 2016). When compared though to another mainstream method used in Zooarchaeology ZooMs, our method has a clear advantage. First, ZooMs use

collagen in order to identify the different kinds of animal species. Given the high degree of conservation of the collagen molecules, ZooMs cannot discriminate among subspecies of mice and most likely cannot do it at the species level, or even at the genus level.

Conclusions

- Our method highly depends on the informational content of our DNA reads and the database we choose.
- The level of the resolution of the identification (e.g. genus, species) varies according to the information our sequences have.
- It is quite conservative in assigning a taxonomic identification to a sample.

Bibliography

- Ahn, T.-H., Chai, J., Pan, C., 2015. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 31, 170–177.
- Barnard, A., 1992. *Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples*. Cambridge University Press.
- Bonder, M.J., Abeln, S., Zaura, E., Brandt, B.W., 2012. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* 28, 2891–2897.
- Ginolhac, A., Rasmussen, M., Gilbert, M.T.P., Willerslev, E., Orlando, L., 2011. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* 27, 2153–2155.
- Grealy, A., Douglass, K., Haile, J., Bruwer, C., Gough, C., Bunce, M., 2016. Tropical ancient DNA from bulk archaeological fish bone reveals the subsistence practices of a historic coastal community in southwest Madagascar. *J. Archaeol. Sci.* 75, 82–88.
- Haile, J., Froese, D.G., Macphee, R.D.E., Roberts, R.G., Arnold, L.J., Reyes, A.V., Rasmussen, M., Nielsen, R., Brook, B.W., Robinson, S., Demuro, M., Gilbert, M.T.P., Munch, K., Austin, J.J., Cooper, A., Barnes, I., Möller, P., Willerslev, E., 2009. Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl. Acad. Sci. U. S. A.* 106, 22352–22357.
- Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M., Pääbo, S., 2001. Ancient DNA. *Nat. Rev. Genet.* 2, 353–359.
- Horsburgh, K.A., Moreno-Mayar, J.V., Gosling, A.L., 2016a. Revisiting the Kalahari debate in the highlands: ancient DNA provides new faunal identifications at Sehonghong, Lesotho. *Azania: Archaeological Research in Africa* 51, 295–306.

- Horsburgh, K.A., Orton, J., Klein, R.G., 2016b. Beware the springbok in sheep's clothing: How secure are the faunal identifications upon which we build our models? *Afr. Archaeol. Rev.* 33, 353–361.
- Hummel, S., Herrmann, B., 1994a. Y-Chromosomal DNA from Ancient Bones, in: Herrmann, B., Hummel, S. (Eds.), *Ancient DNA: Recovery and Analysis of Genetic Material from Paleontological, Archaeological, Museum, Medical, and Forensic Specimens*. Springer New York, New York, NY, pp. 205–210.
- Hummel, S., Herrmann, B., 1994b. General Aspects of Sample Preparation, in: Herrmann, B., Hummel, S. (Eds.), *Ancient DNA: Recovery and Analysis of Genetic Material from Paleontological, Archaeological, Museum, Medical, and Forensic Specimens*. Springer New York, New York, NY, pp. 59–68.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Lee, R.B., 1979. *The! Kung San: men, women and work in a foraging society*. CUP Archive.
- Lee, R.B., Guenther, M., 1993. Problems in Kalahari Historical Ethnography and the Tolerance of Error. *Hist. Afr.* 20, 185–235.
- Lejju, B.J., Robertshaw, P., Taylor, D., 2006. Africa's earliest bananas? *J. Archaeol. Sci.* 33, 102–113.
- Marshall, L., 2014. *The !kung of Nyae Nyae*. Harvard University Press, Cambridge, MA.
- Mitchell, P., 2009. Hunter-gatherers and farmers: Some implications of 1,800 years of interaction in the maloti-Drakensberg region of southern Africa, in: *Interactions between Hunter-Gatherers and Farmers: From Prehistory to Present*. pascal-francis.inist.fr, pp. 15–46.

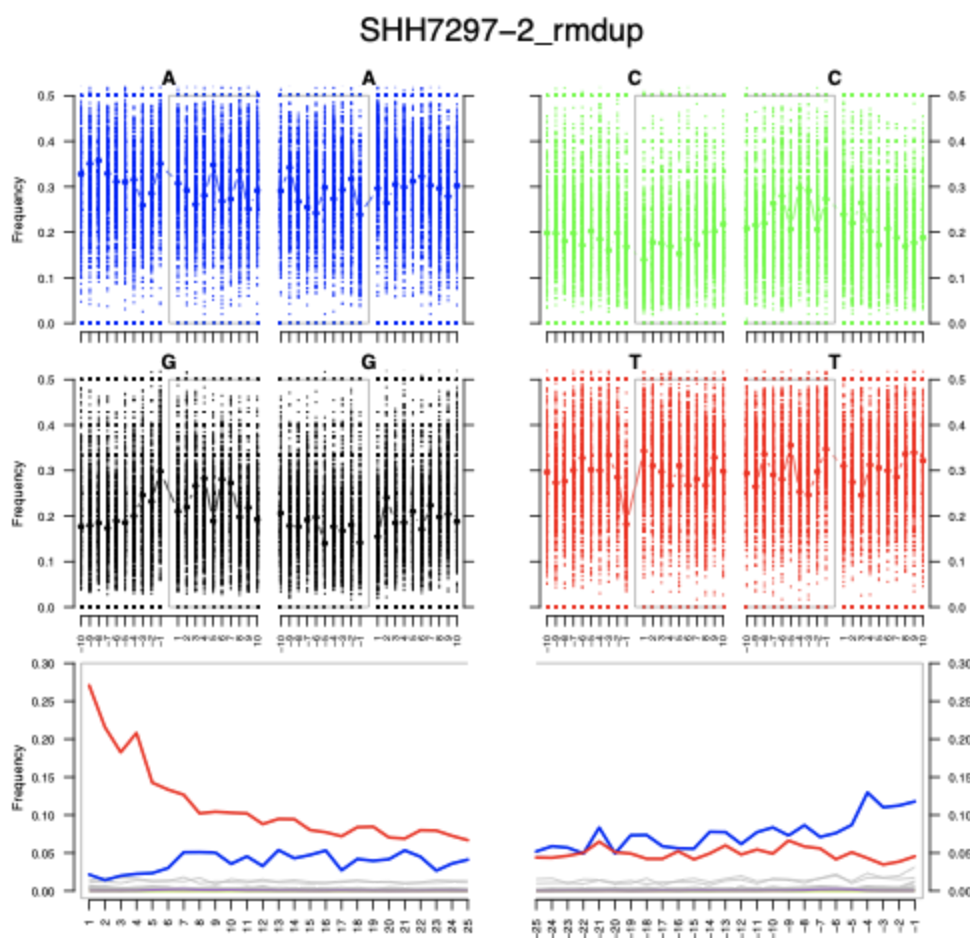
- Mitchell, P., Plug, I., Bailey, G., Woodborne, S., 2008. Bringing the Kalahari debate to the mountains. *Before Farming* 2008, 1–23.
- Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., Hofreiter, M., 2004. Genetic analyses from ancient DNA. *Annu. Rev. Genet.* 38, 645–679.
- Prendergast, M.E., Buckley, M., Crowther, A., Frantz, L., Eager, H., Lebrasseur, O., Hutterer, R., Hulme-Beaman, A., Van Neer, W., Douka, K., Veall, M.-A., Quintana Morales, E.M., Schuenemann, V.J., Reiter, E., Allen, R., Dimopoulos, E.A., Helm, R.M., Shipton, C., Mwebi, O., Denys, C., Horton, M., Wynne-Jones, S., Fleisher, J., Radimilahy, C., Wright, H., Searle, J.B., Krause, J., Larson, G., Boivin, N.L., 2017. Reconstructing Asian faunal introductions to eastern Africa from multi-proxy biomolecular and archaeological datasets. *PLoS One* 12, e0182565.
- Ratnasingham, S., Hebert, P.D.N., 2007. bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* 7, 355–364.
- Reis-Filho, J.S., 2009. Next-generation sequencing. *Breast Cancer Res.* 11 Suppl 3, S12.
- Sadr, K., 1997. Kalahari Archaeology and the Bushman Debate. *Curr. Anthropol.* 38, 104–112.
- Saputra, D., Rasmussen, S., Larsen, M.V., Haddad, N., Sperotto, M.M., Aarestrup, F.M., Lund, O., Sicheritz-Pontén, T., 2015. Reads2Type: a web application for rapid microbial taxonomy identification. *BMC Bioinformatics* 16, 398.
- Schrire, C., 1992. The Archaeological Identity of Hunters and Herders at the Cape over the Last 2000 Years: A Critique. *The South African Archaeological Bulletin* 47, 62–64.

- Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18.
- Slon, V., Glocke, I., Barkai, R., Gopher, A., HersHKovitz, I., Meyer, M., 2016. Mammalian mitochondrial capture, a tool for rapid screening of DNA preservation in faunal and undiagnostic remains, and its application to Middle Pleistocene specimens from Qesem Cave (Israel). *Quat. Int.* 398, 210–218.
- Whitelaw, G., Prins, F., Mazel, A., Maggs, T., 1992. Comment on Schrire: the archaeological identity of hunters and herders at the Cape. *The South African Archaeological Bulletin* 47, 131–132.
- Willerslev, E., Cooper, A., 2005. Ancient DNA. *Proc. Biol. Sci.* 272, 3–16.
- Yates, R., Smith, A.B., 1993. A reevaluation of the chronology of Oudepost: a reply in part to Schrire. *The South African Archaeological Bulletin* 48, 52–53.
- Yellen, J.E., Alison, S.B., n.d. nd The Late Stone Age archaeology of the! Kangwa and/Xai/Xai Valleys, Ngamiland Botswana. *Botsw. Notes Rec.*

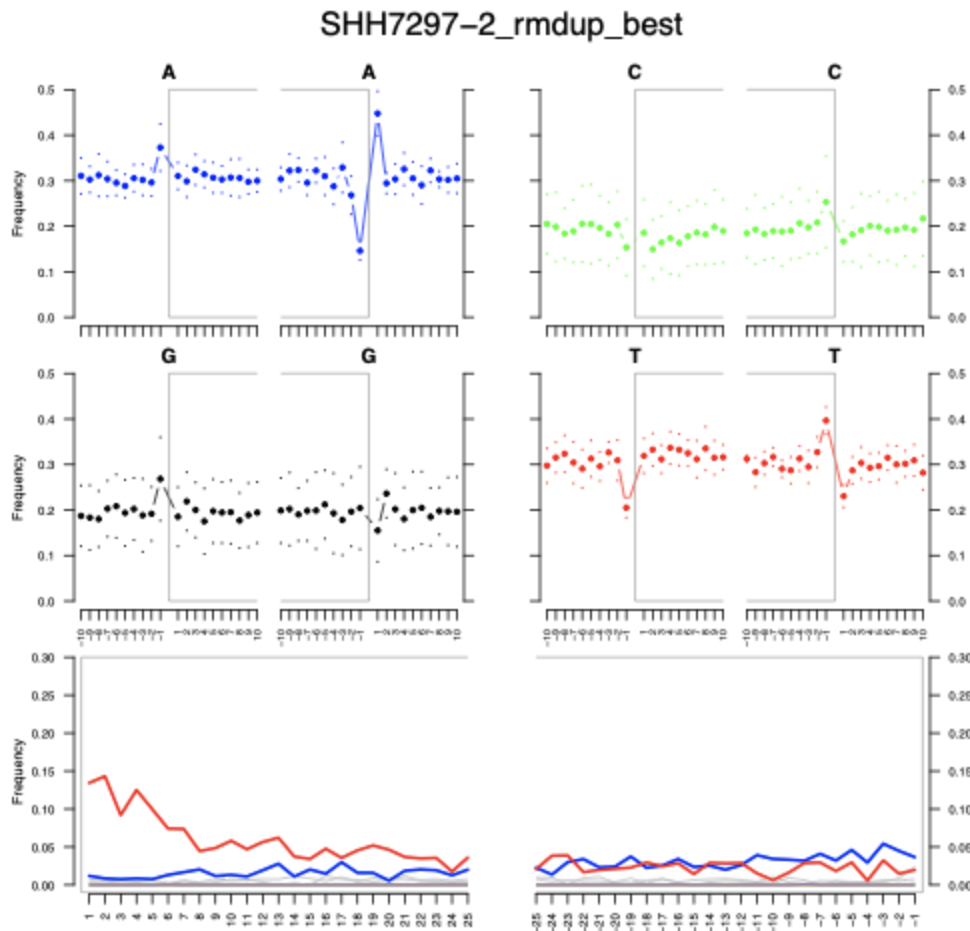
Supplementary

Supplementary table 1. Assignment probabilities for sample SHH7297. The analysis of the alignment score matrix of the sample SHH7297 identified *Tragelaphus oryx* as the best species, with a support probability of 58%. The second best species is *Syncerus caffer*, and has an identification probability of 5.7%.

Sample	SHH7297
1st Assignment Probability	0.58
1st assigned <u>Species</u>	<i>Tragelaphus oryx</i>
2nd Assignment Probability	0.057
2nd Assigned Species	<i>Syncerus caffer</i>
>0	174083
>50	3800
>100	3267



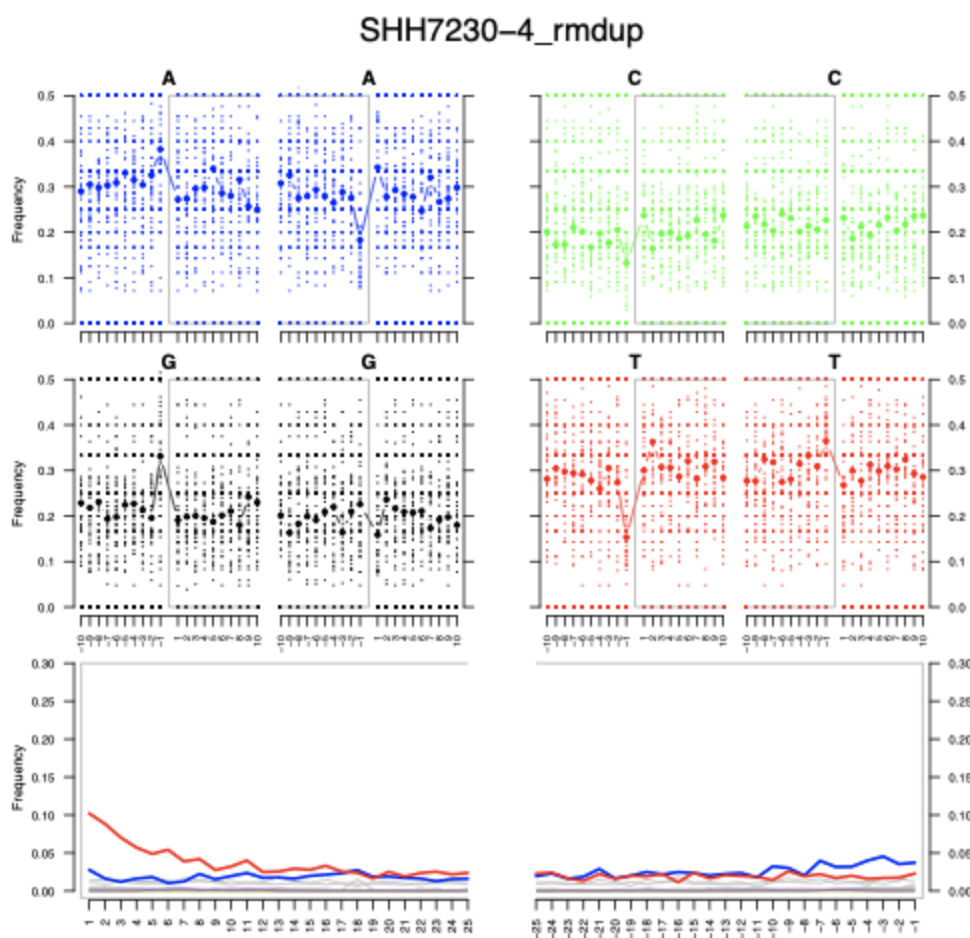
Supplementary figure 1. Damage pattern analysis for all the aligned reads for sample SHH7297.



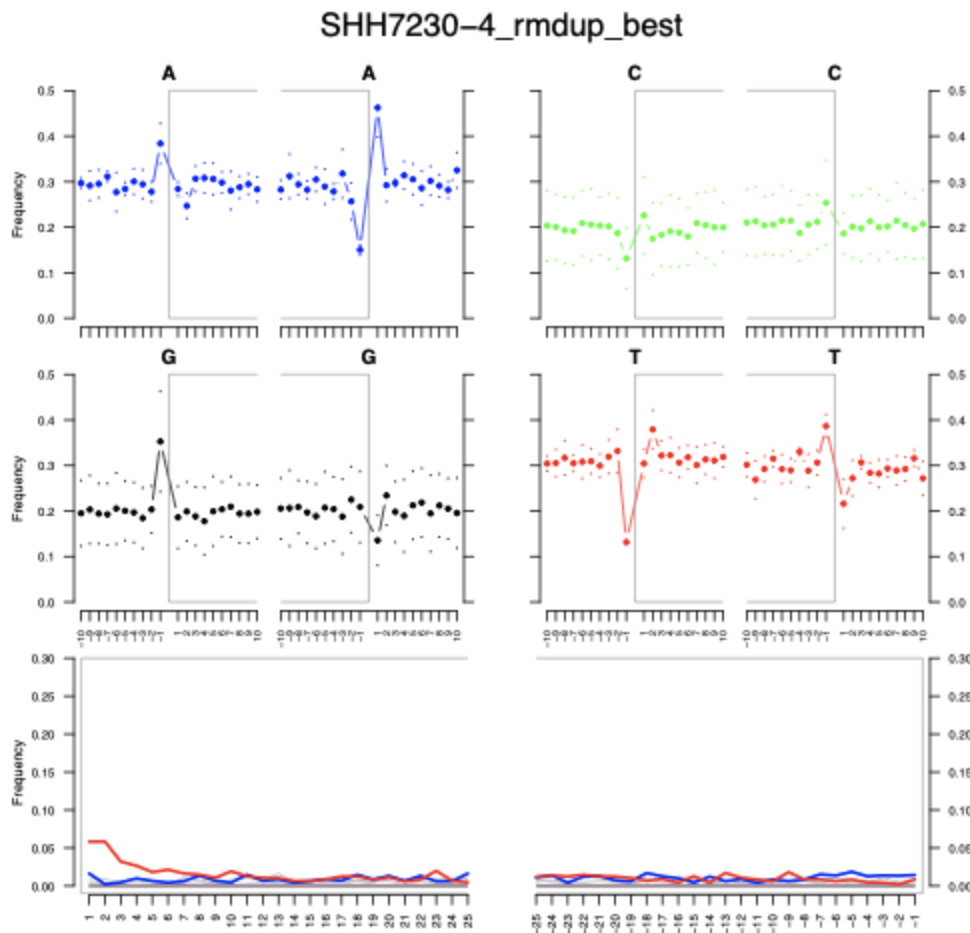
Supplementary figure 2. Damage pattern analysis of the aligned reads for taxon *Tragelaphus oryx*, for sample SHH7297. A high frequency of deaminations can be observed at the 5' end of the reads.

Supplementary table 2. Assignment probabilities for sample SHH7230. From Sigma's alignment score matrix the 2 best species for sample SHH7230 are *Tragelaphus oryx* (Common Eland) with an assignment probability of 84% and *Taurotragus derbianus* (Giant Eland) with an assignment probability of 2.5%.

Sample	SHH7230
1st Assignment Probability	0.84
1st assigned Species	<i>Tragelaphus oryx</i>
2nd Assignment Probability	0.025
2nd Assigned Species	<i>Taurotragus derbianus</i>
>0	154570
>50	3035
>100	2455



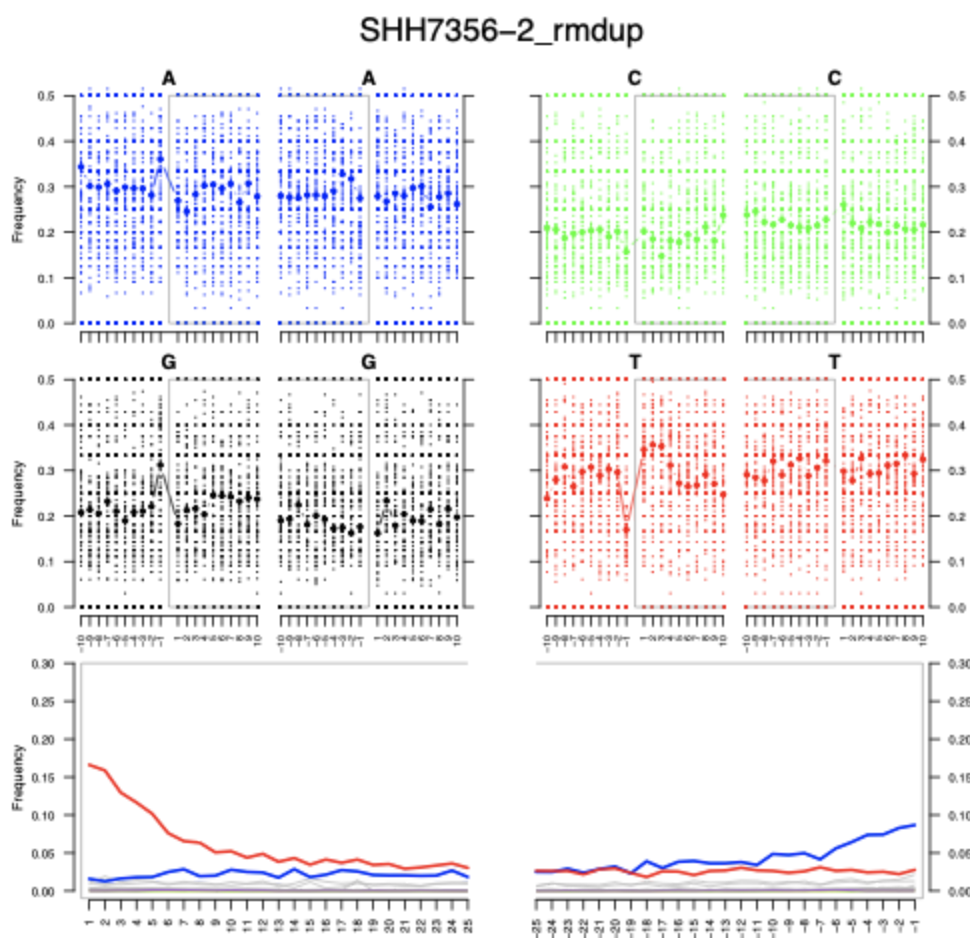
Supplementary figure 3. Damage pattern analysis for all the aligned reads for sample SHH7230.



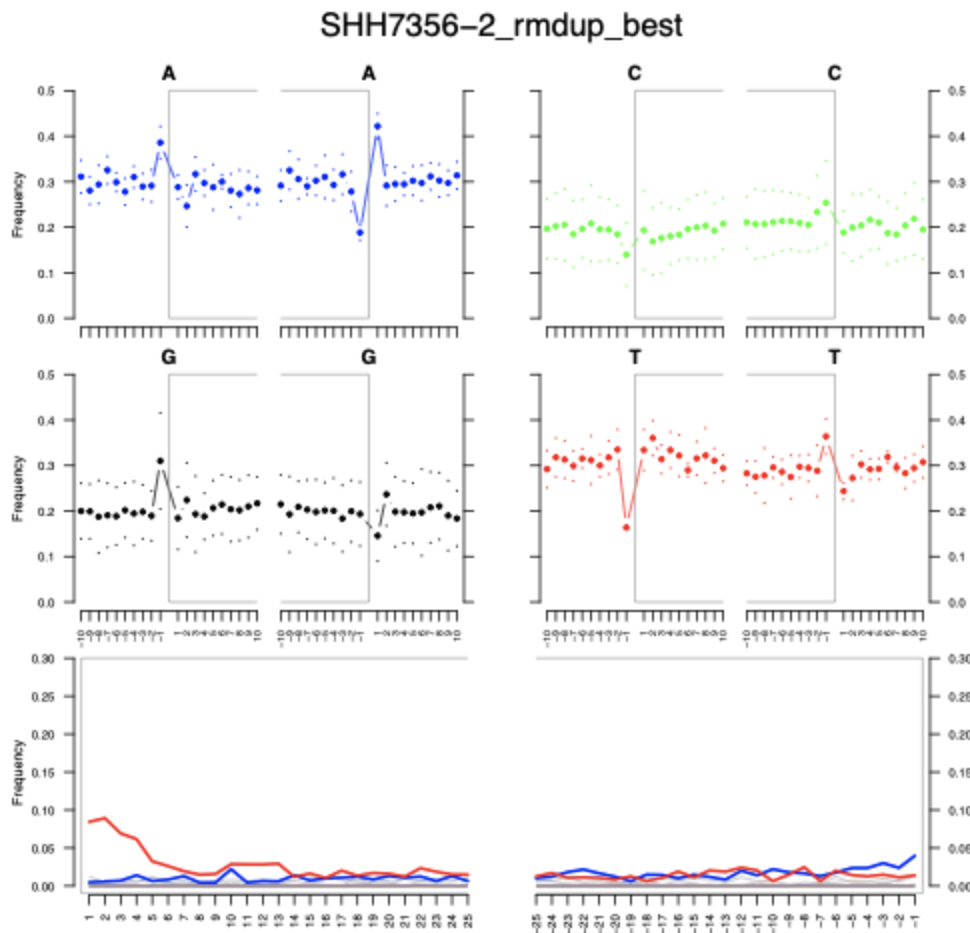
Supplementary figure 4. Damage pattern analysis of the aligned reads for taxon *Tragelaphus oryx*, for sample SHH7230. A high frequency of deaminations can be observed at the 5' end of the reads.

Supplementary table 3. Assignment probabilities for sample SHH7356. From the alignment score matrix produced by Sigma, the best species to assign this sample to was again *Tragelaphus oryx* (the Common Eland) with a probability of 79%. The second best taxon was again *Taurotragus derbianus* (the Giant Eland) with a probability of 2.9%.

Sample	SHH7356
1st Assignment Probability	0.79
1st assigned Species	<i>Tragelaphus oryx</i>
2nd Assignment Probability	0.029
2nd Assigned Species	<i>Taurotragus derbianus</i>
>0	162498
>50	3178
>100	2586



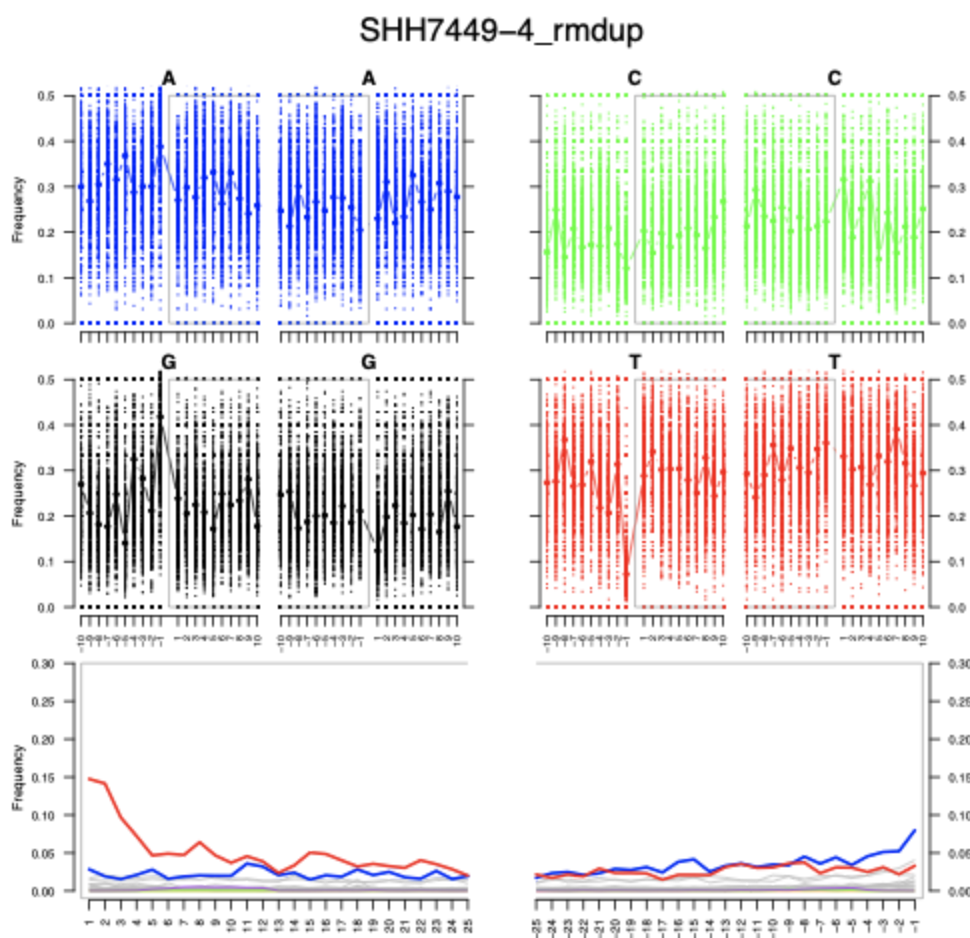
Supplementary figure 5. Damage pattern analysis for all the aligned reads for sample SHH7356.



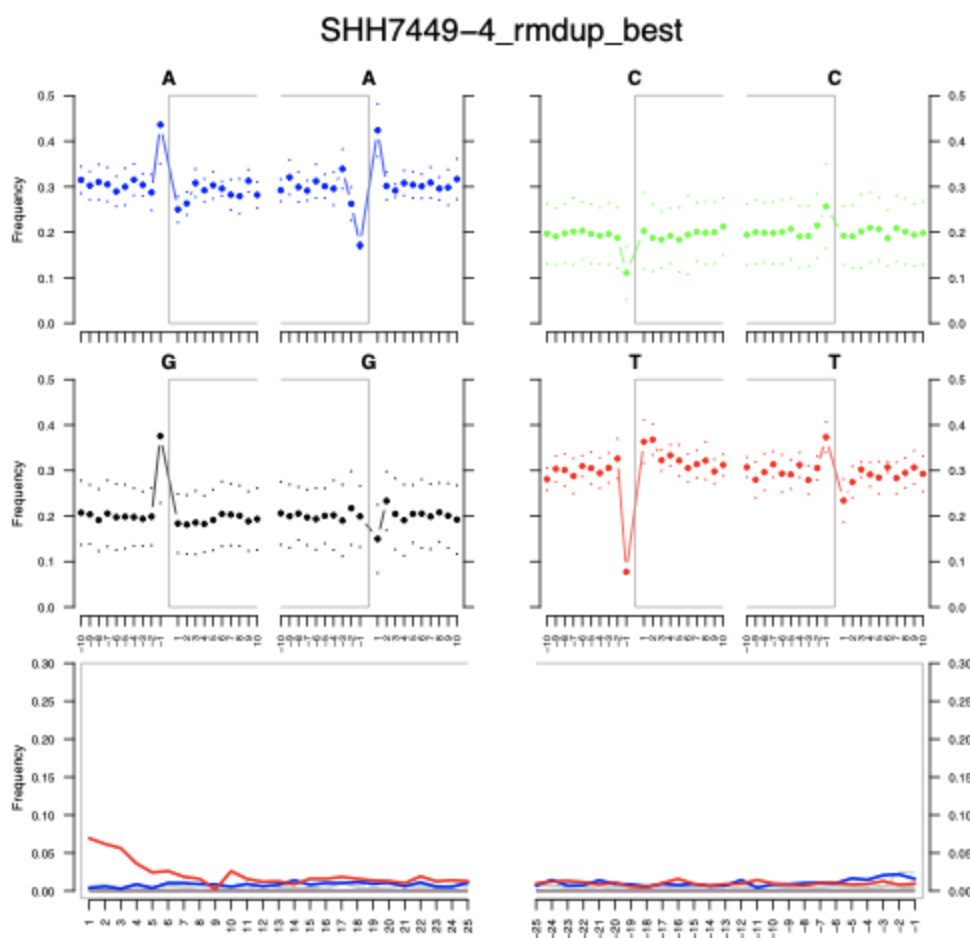
Supplementary figure 6. Damage pattern analysis of the aligned reads for taxon *Tragelaphus oryx*, for sample SHH7356. A high frequency of deaminations can be observed at both the 5' and 3' ends of the reads.

Supplementary table 4. Assignment probabilities for sample SHH7449. After calculating the species assignment probabilities from Sigma's alignment score matrix, *Tragelaphus oryx* came back as the best species, with an assignment probability of 53% and the second best species was *Taurotragus derbianus* with a probability of 4.2%.

Sample	SHH7449
1st Assignment Probability	0.53
1st assigned Species	<i>Tragelaphus oryx</i>
2nd Assignment Probability	0.042
2nd Assigned Species	<i>Taurotragus derbianus</i>
>0	486434
>50	8585
>100	6896



Supplementary figure 7. Damage pattern analysis for all the aligned reads for sample SHH7449.

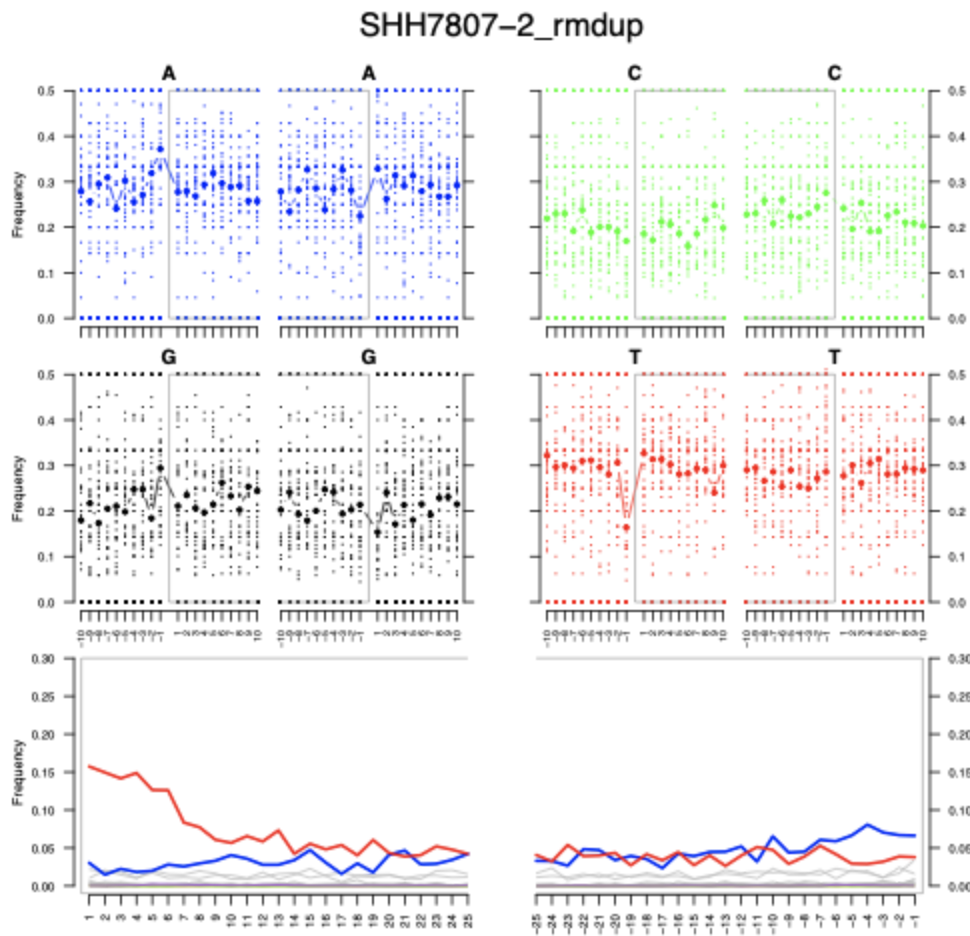


Supplementary figure 8. Damage pattern analysis of the aligned reads for taxon *Tragelaphus oryx*, for sample SHH7449. A high frequency of deaminations can be observed at the 5' end of the reads.

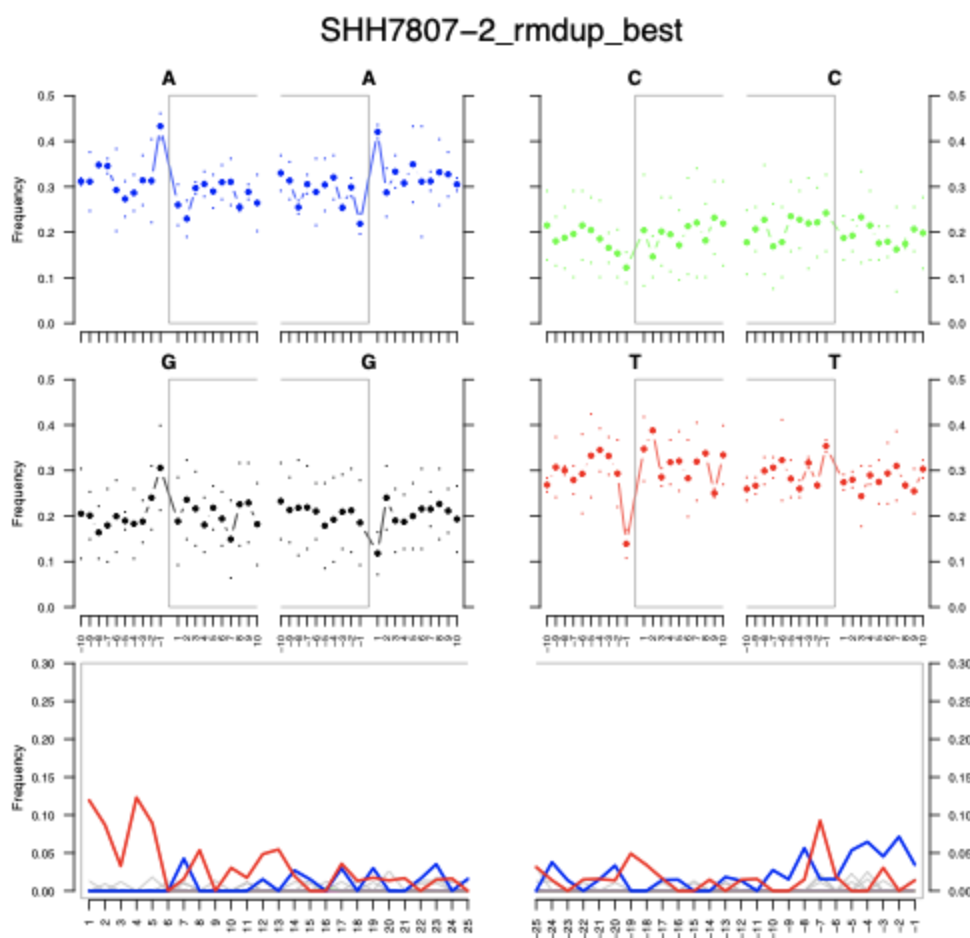
Supplementary table 5. Assignment probabilities for sample SHH7807.

Calculating the species assignment probabilities from the alignment score matrix determines *Tragelaphus oryx* as the best species with a probability of 54%, and the second best species is *Canis lupus* with an assignment probability of 12%.

Sample	SHH7807
1st Assignment Probability	0.54
1st assigned Species	<i>Tragelaphus oryx</i>
2nd Assignment Probability	0.12
2nd Assigned Species	<i>Canis lupus</i>
>0	33609
>50	597
>100	458



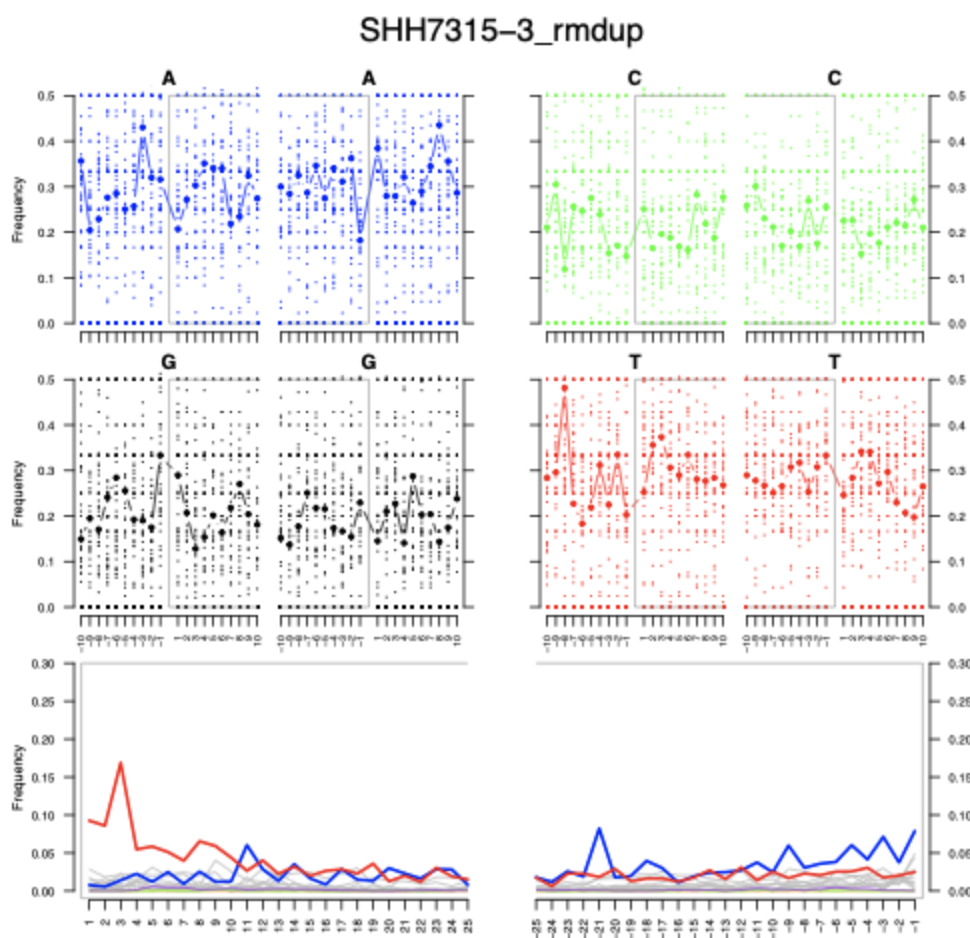
Supplementary figure 9. Damage pattern analysis for all the aligned reads for sample SHH7807.



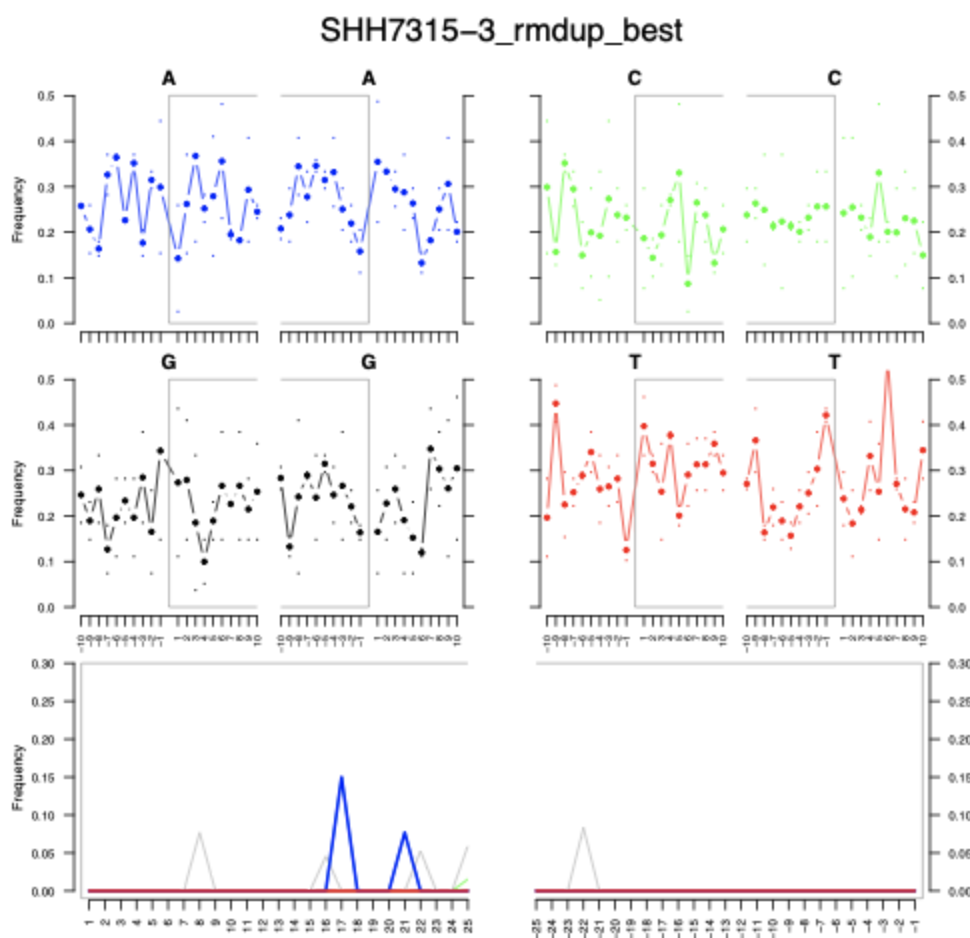
Supplementary figure 10. Damage pattern analysis of the aligned reads for taxon *Tragelaphus oryx*, for sample SHH7807. A high frequency of deaminations can be observed at both the 5' and 3' ends of the reads.

Supplementary table 6. Assignment probabilities for sample SHH7315. After parsing the alignment score matrix produced by Sigma, the two best species identified were *Homo sapiens*, with an assignment probability of 17%, and *Canis lupus*, with an assignment probability of 16%. Both these probabilities are too low to perform any robust and confident species assignment.

Sample	SHH7315
1st Assignment Probability	0.17
1st assigned Species	<i>Homo sapiens</i>
2nd Assignment Probability	0.16
2nd Assigned Species	<i>Canis lupus</i>
>0	11226
>50	316
>100	210



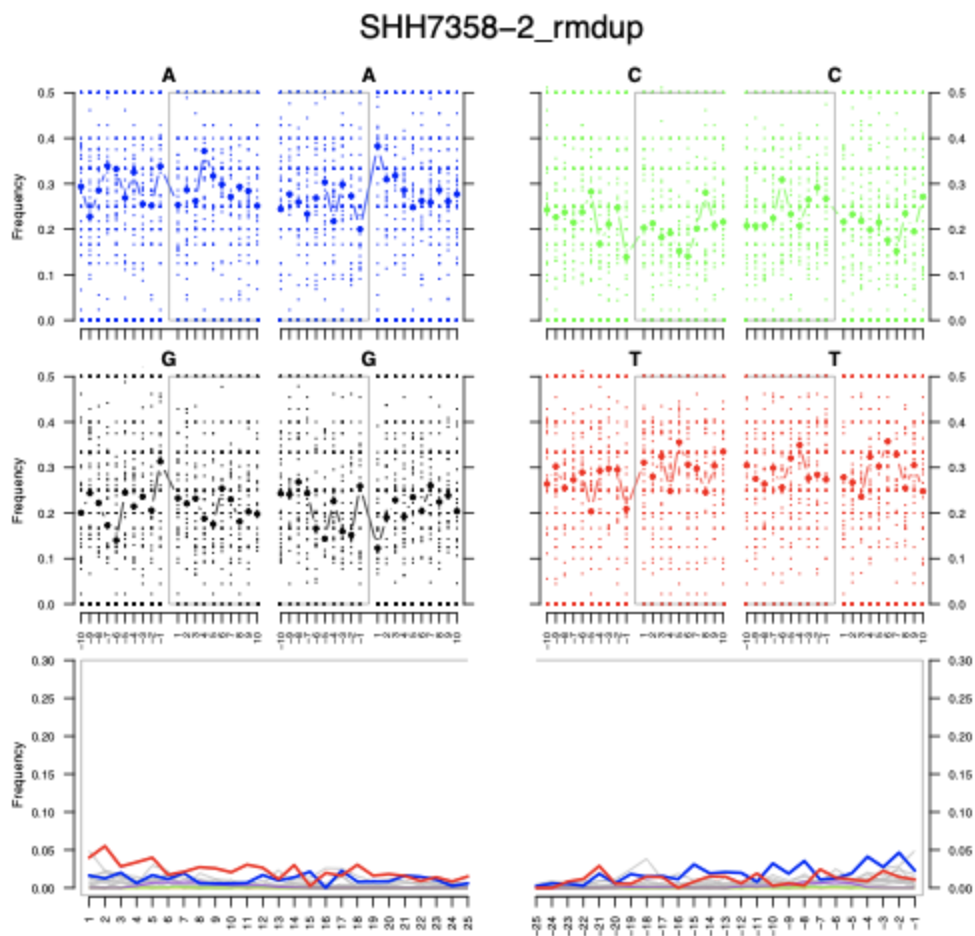
Supplementary figure 11. Damage pattern analysis for all the aligned reads for sample SHH7315.



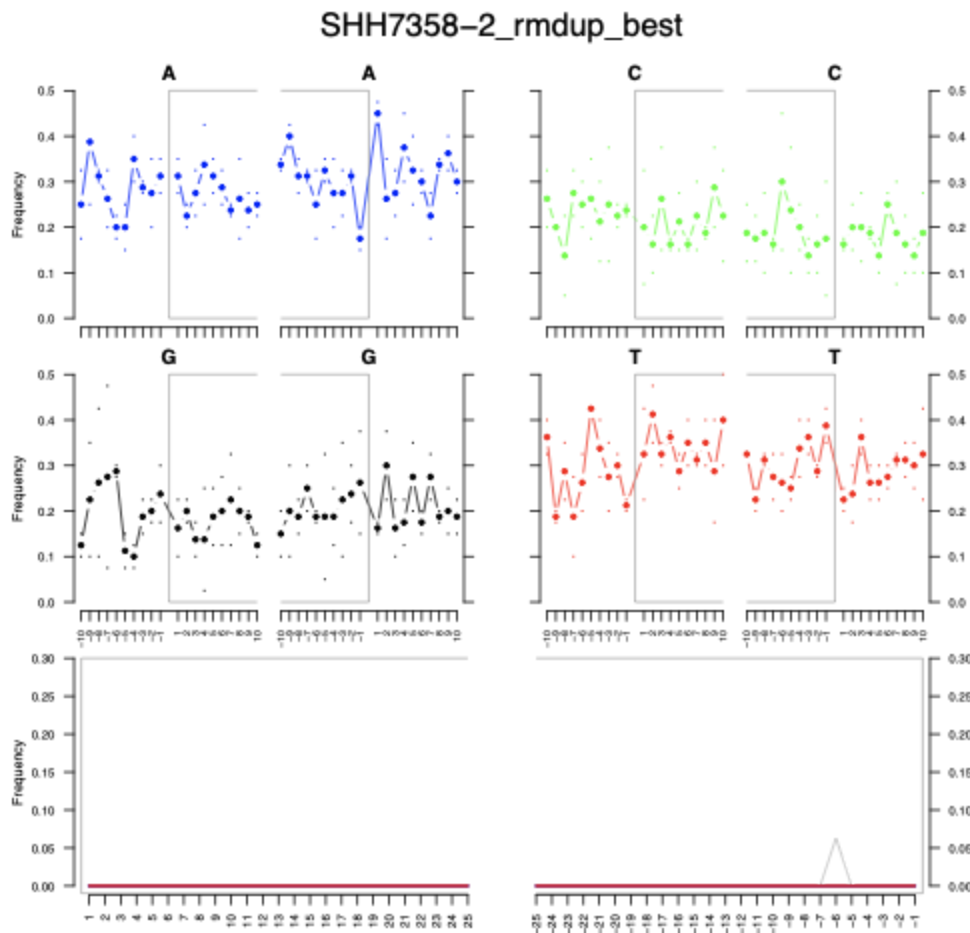
Supplementary figure 12. Damage pattern analysis of the aligned reads for taxon *Homo sapiens*, for sample SHH7315. No deamination pattern can be observed on either end of the DNA reads indicating that these reads are yielded from modern human contamination.

Supplementary table 7. Assignment probabilities for sample SHH7358. By calculating the species assignment probabilities from the alignment score matrix, the two best species that come up are *Bos taurus*, with an assignment probability of 44%, and the second best species is *Homo sapiens*, with an assignment probability of 29%. Neither of these probabilities are high enough to perform a robust species identification.

Sample	SHH7358
1st Assignment Probability	0.44
1st assigned Species	<i>Bos taurus</i>
2nd Assignment Probability	0.30
2nd Assigned Species	<i>Homo sapiens</i>
>0	6308
>50	210
>100	97



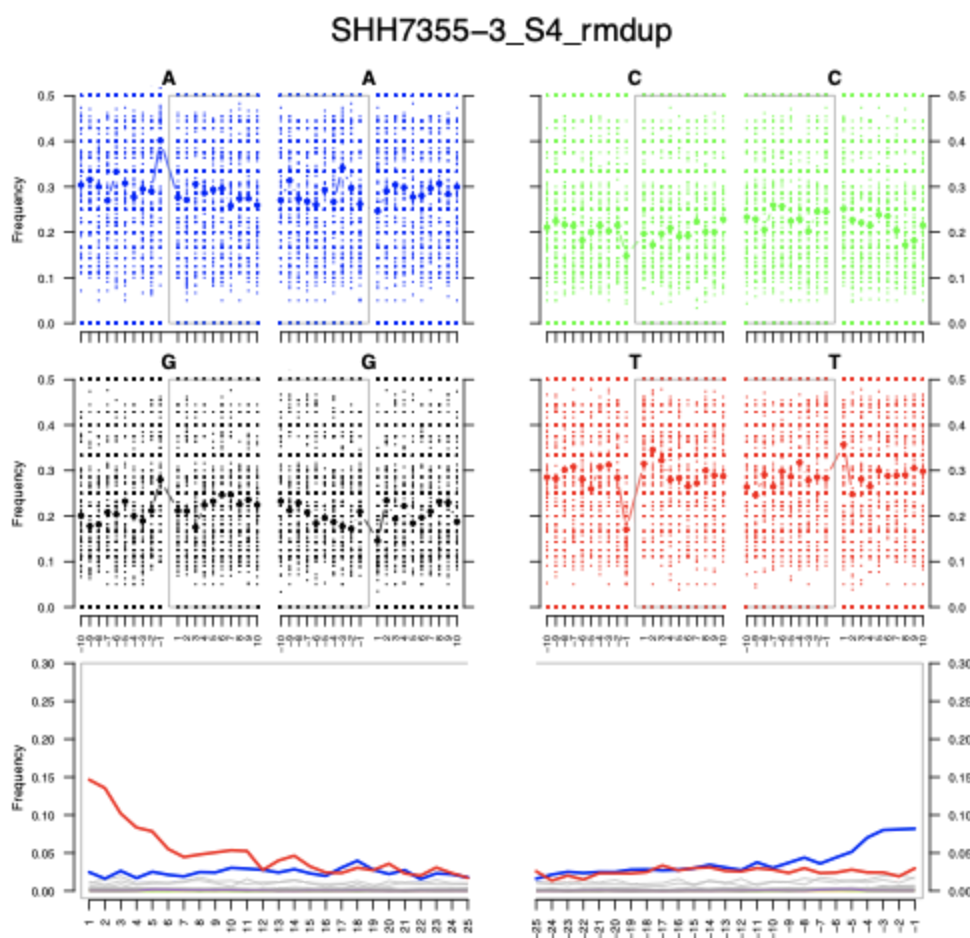
Supplementary figure 13. Damage pattern analysis for all the aligned reads for sample SHH7358.



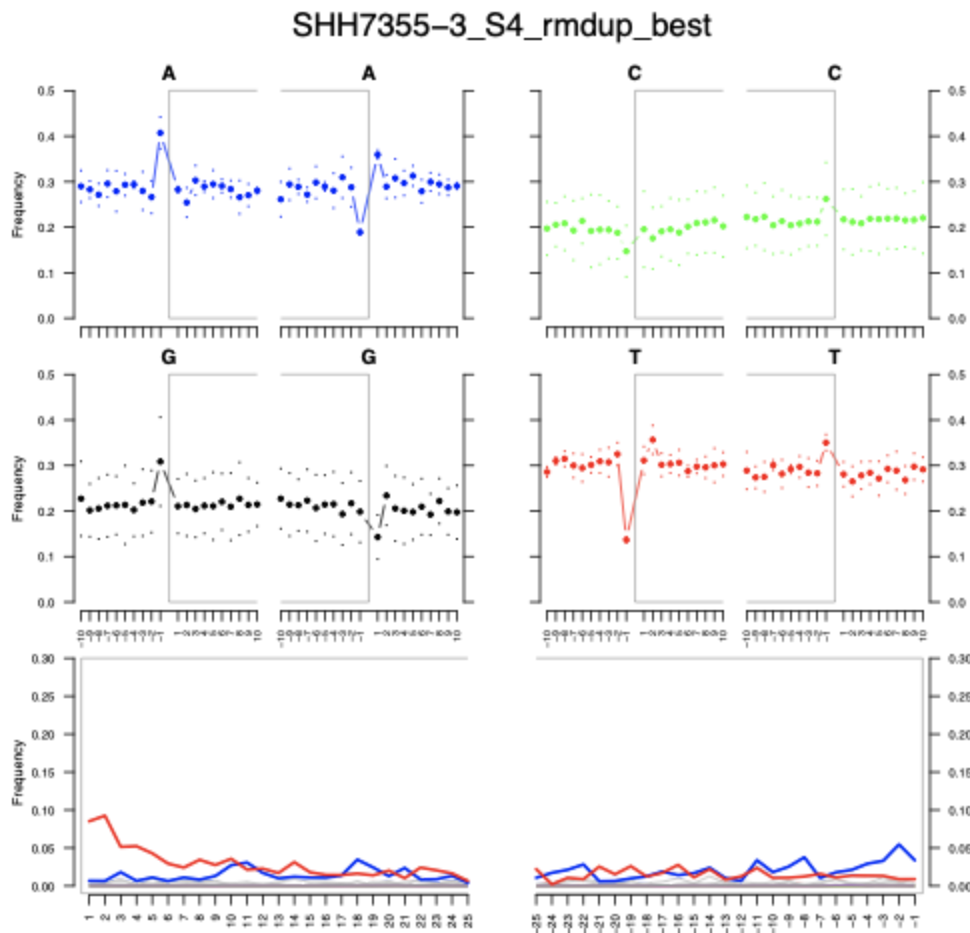
Supplementary figure 14. Damage pattern analysis of the aligned reads for taxon *Bos taurus*, for sample SHH7358. No deamination pattern can be observed in the *Bos taurus* aligned reads, hinting that these reads could be modern contamination.

Supplementary table 8. Assignment probabilities for sample SHH7355. After analysing the alignment score matrix the primary identified species is *Redunca fulvorufula* (Mountain Reedbuck) with an assignment probability of 64%. The second identified species was *Redunca arundinum* (Southern/Common Reedbuck), with an assignment probability of 2%.

Sample	SHH7355
1st Assignment Probability	0.65
1st assigned Species	<i>Redunca fulvorufula</i>
2nd Assignment Probability	0.024
2nd Assigned Species	<i>Redunca arundinum</i>
>0	163161
>50	2873
>100	2430



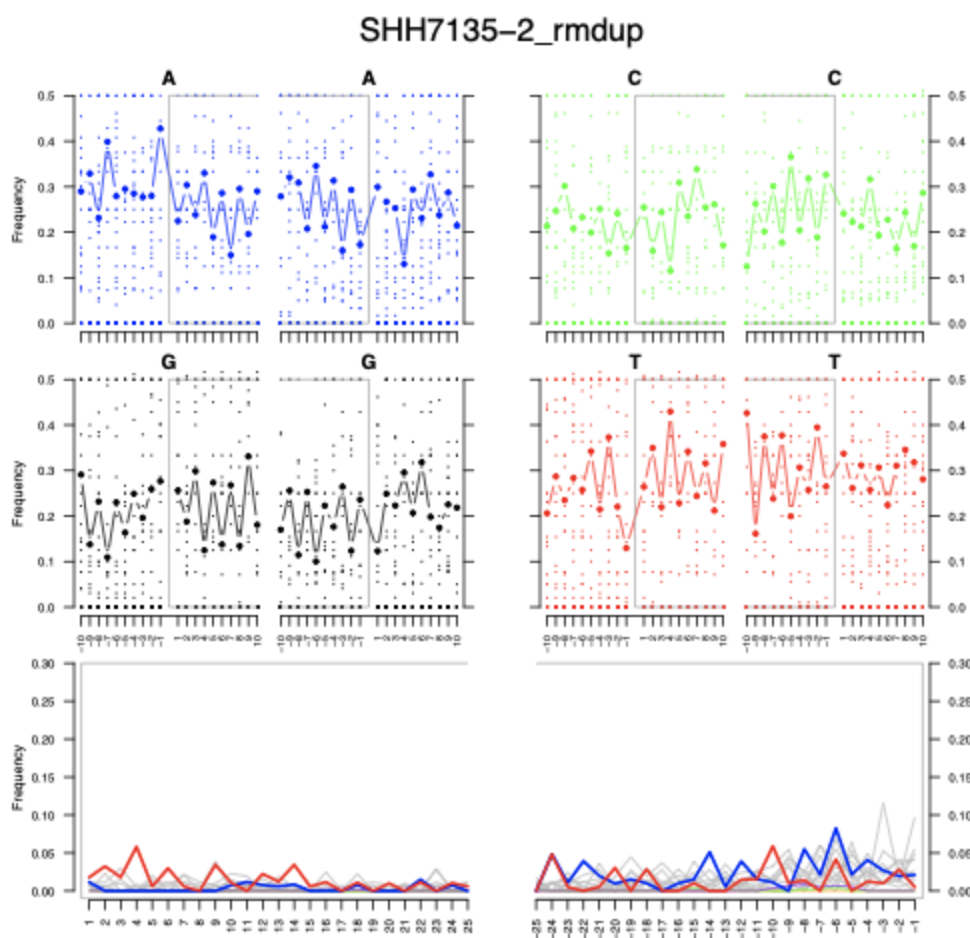
Supplementary figure 15. Damage pattern analysis for all the aligned reads for sample SHH7355.



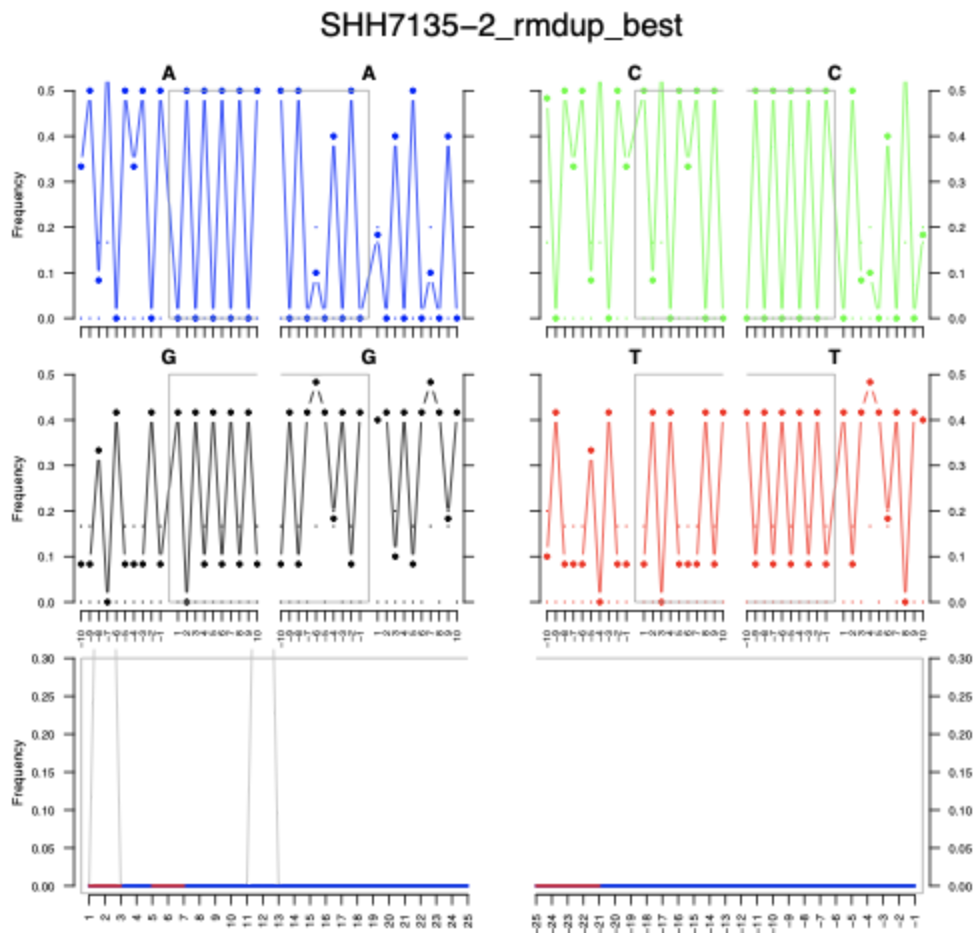
Supplementary figure 16. Damage pattern analysis of the aligned reads for taxon *Redunca fulvorufula*, for sample SHH7355. A high frequency of deaminations can be observed at both the 5' and 3' end of the reads.

Supplementary table 9. Assignment probabilities for sample SHH7135. By analysing the alignment score matrix, the two best identified species were *Dipus sagitta* (Northern three-toed jerboa) with an assignment probability of 43%, and *Panthera pardus* (Leopard) with an assignment probability of 19%. Neither of these probabilities are high enough, to allow us to robustly identify this bone sample.

Sample	SHH7135
1st Assignment Probability	0.43
1st assigned Species	<i>Dipus sagitta</i>
2nd Assignment Probability	0.20
2nd Assigned Species	<i>Panthera pardus</i>
>0	952
>50	22
>100	19



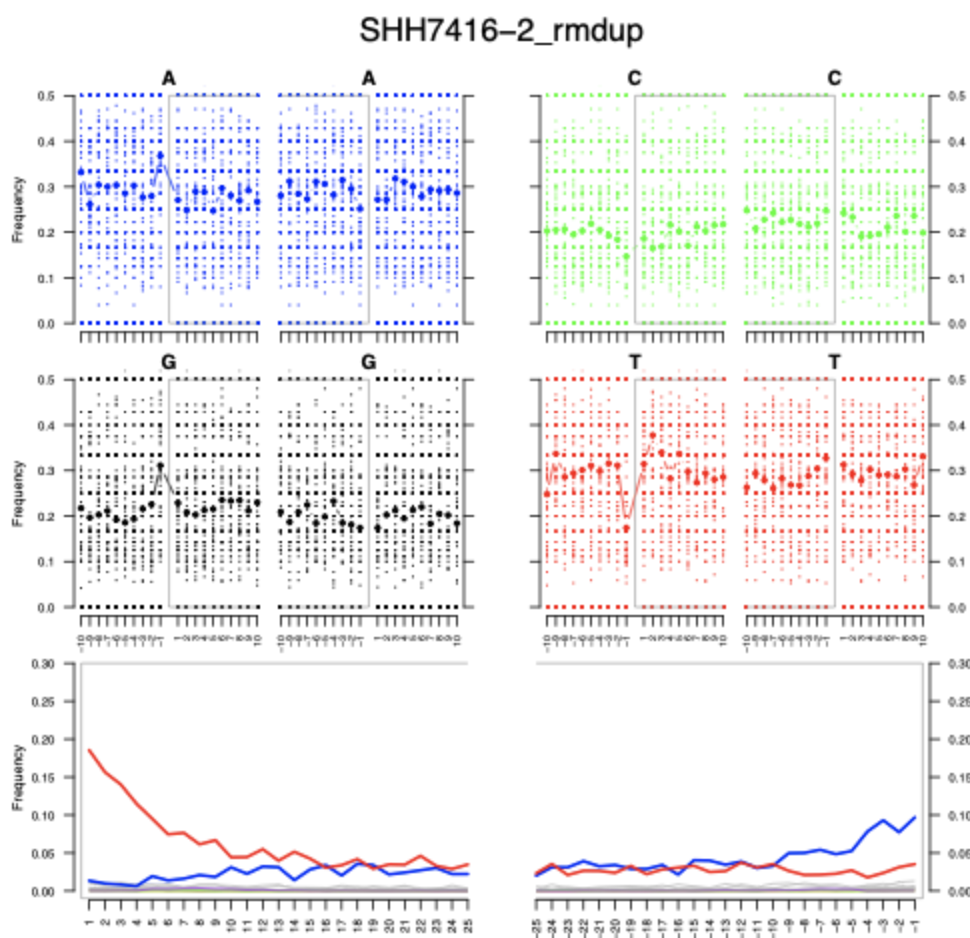
Supplementary figure 17. Damage pattern analysis for all the aligned reads for sample SHH7135.



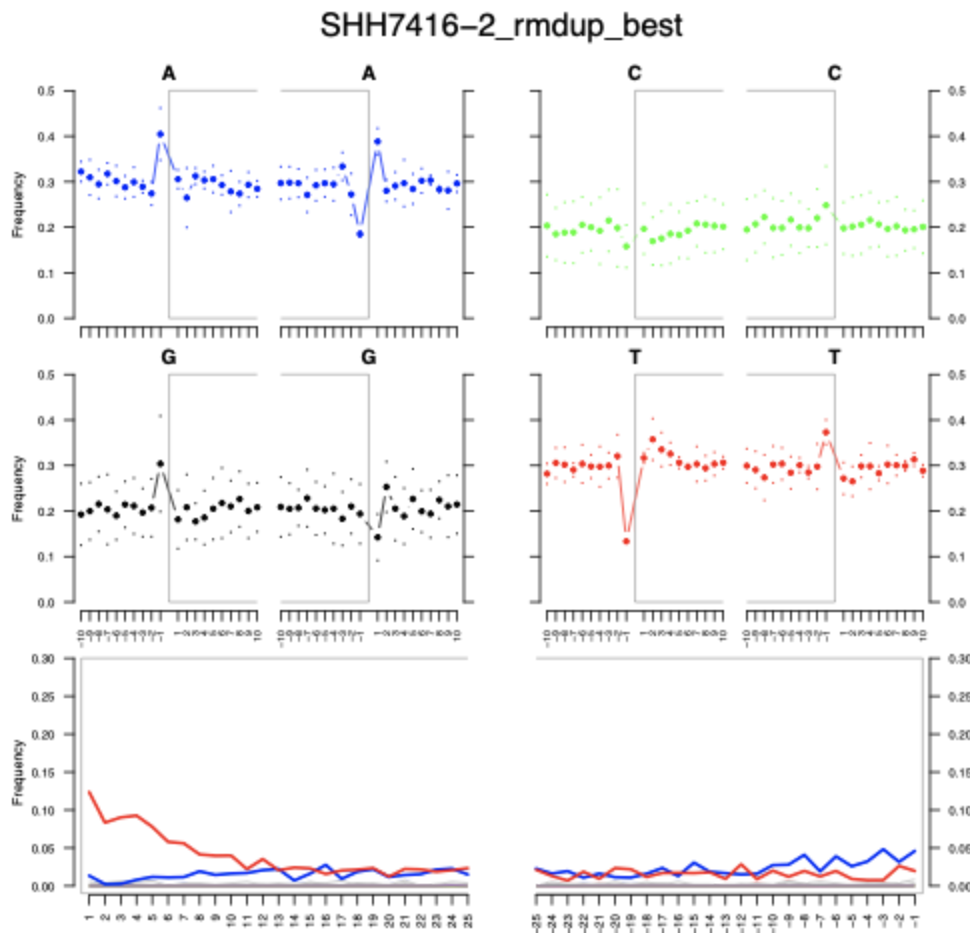
Supplementary figure 18. Damage pattern analysis of the aligned reads for taxon *Dipus sagitta*, for sample SHH7135. A deamination pattern cannot be observed, indicating that modern contamination could be the source of these DNA reads.

Supplementary table 10. Assignment probabilities for sample SHH7416. After analysing the alignment score matrix, the two best species that were identified are *Pelea capreolus* (Grey Rhebok) with an assignment probability of 88% and *Cephalophus ogilbyi* (Ogilby's duiker) with an assignment probability of 1.9%.

Sample	SHH7416
1st Assignment Probability	0.88
1st assigned Species	<i>Pelea capreolus</i>
2nd Assignment Probability	0.019
2nd Assigned Species	<i>Cephalophus ogilbyi</i>
>0	161934
>50	2621
>100	2175



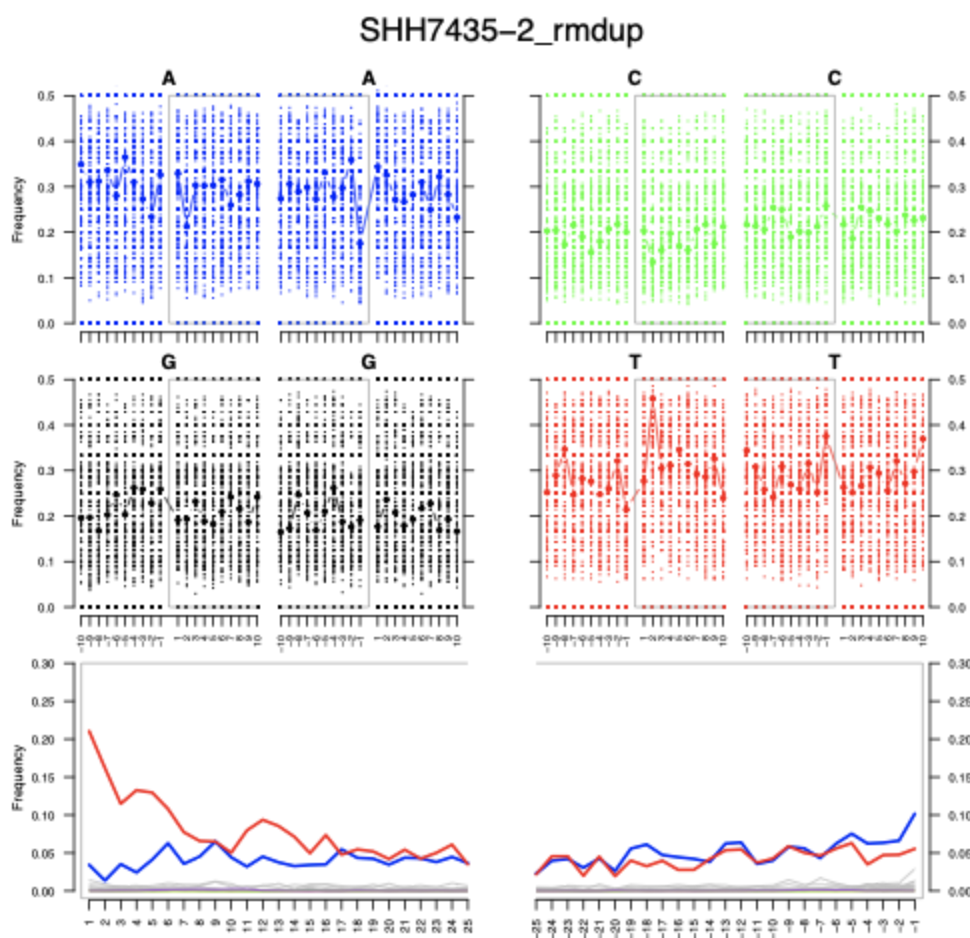
Supplementary figure 19. Damage pattern analysis for all the aligned reads for sample SHH7416.



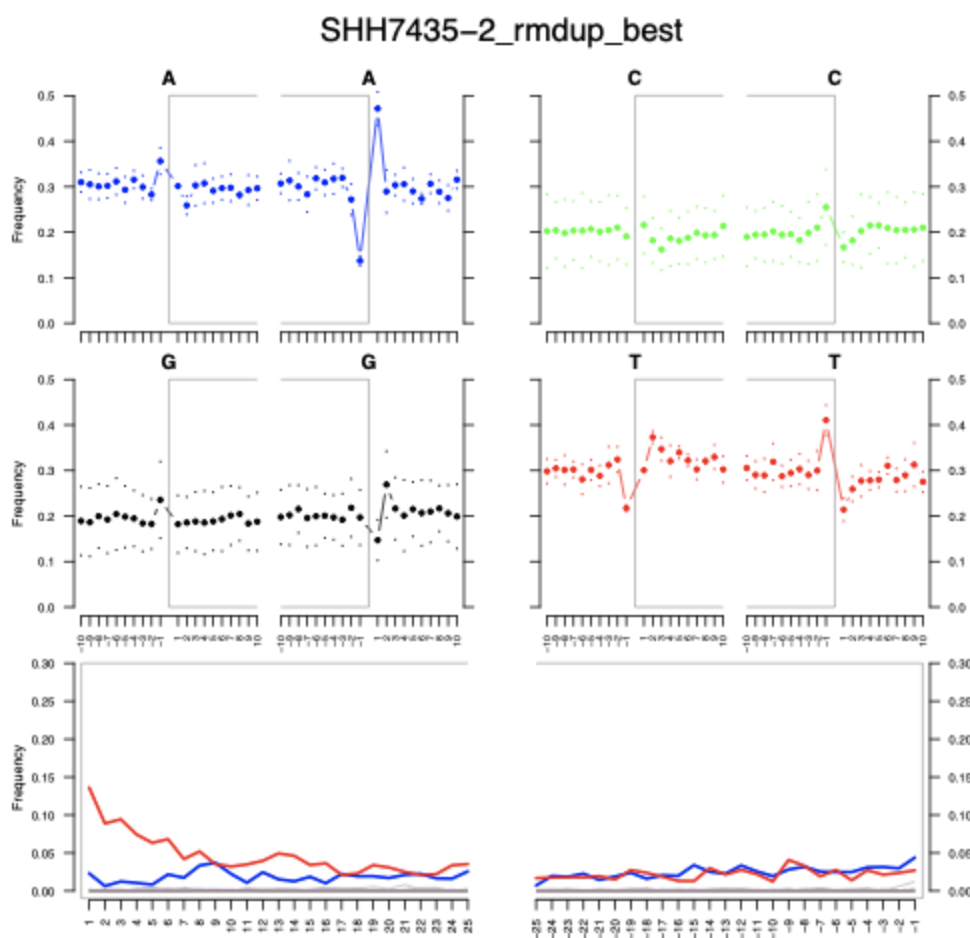
Supplementary figure 20. Damage pattern analysis of the aligned reads for taxon *Pelea capreolus*, for sample SHH7416. A high frequency of deaminations can be observed at both the 5' and 3' end of the reads.

Supplementary table 11. Assignment probabilities for sample SHH7435. The analysis of the alignment score matrix allowed us to identify the two best species for this sample, *Pelea capreolus* (Grey Rhebok) with an assignment probability of 88%, and *Kobus leche* (Lechwe) with an assignment probability of 1%.

Sample	SHH7435
1st Assignment Probability	0.88
1st assigned Species	<i>Pelea capreolus</i>
2nd Assignment Probability	0.01
2nd Assigned Species	<i>Kobus leche</i>
>0	155620
>50	3095
>100	2580



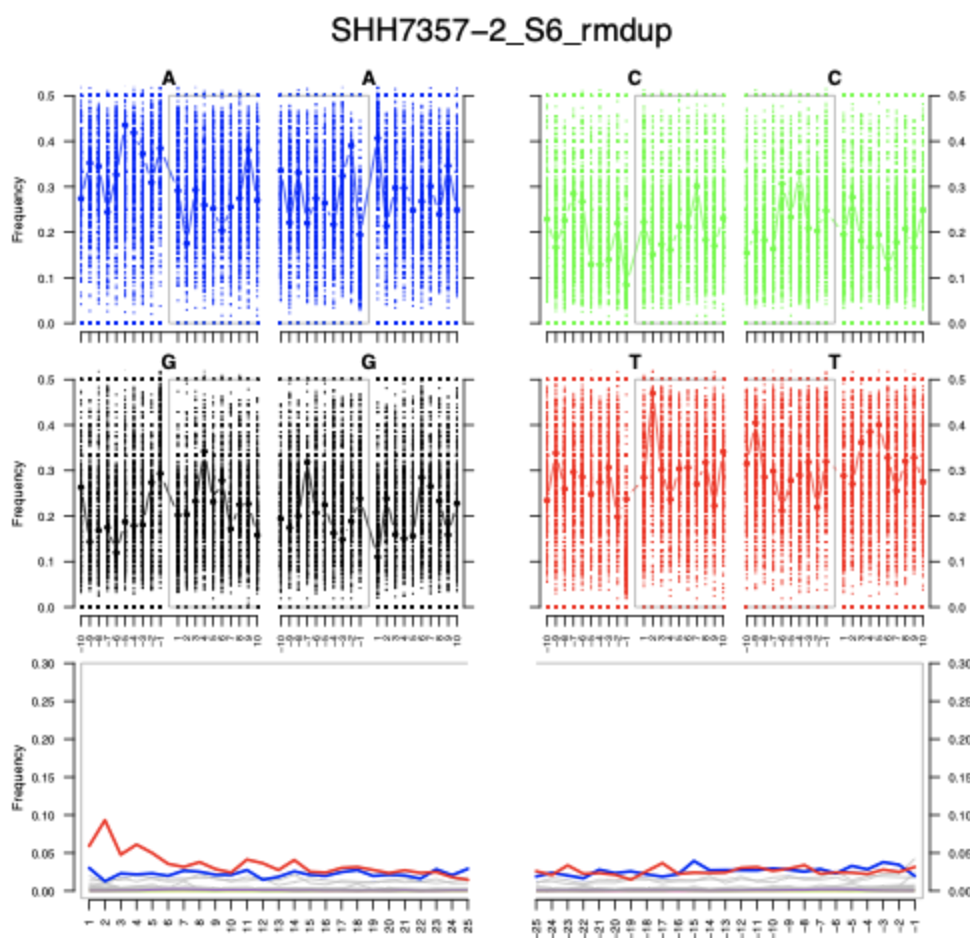
Supplementary figure 21. Damage pattern analysis for all the aligned reads for sample SHH7435.



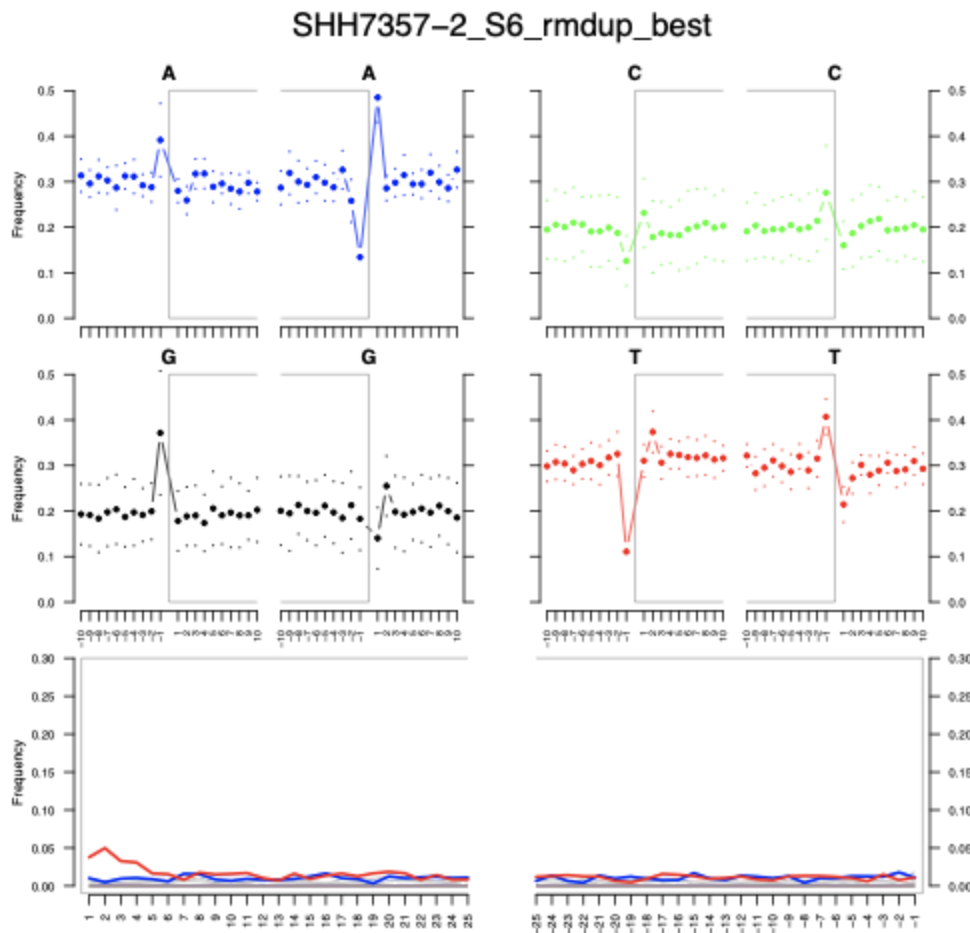
Supplementary figure 22. Damage pattern analysis of the aligned reads for taxon *Pelea capreolus*, for sample SHH7435. A high frequency of deaminations can be observed at the 5' end of the reads.

Supplementary table 12. Assignment probabilities for sample SHH7357. By analysing the alignment score matrix, we calculate the assignment probabilities of the two best species, *Tragelaphus oryx* (Common Eland) with a probability of 79% and *Taurotragus derbianus* (Giant Eland) with a probability of 2.8%.

Sample	SHH7357
1st Assignment Probability	0.79
1st assigned Species	<i>Tragelaphus oryx</i>
2nd Assignment Probability	0.028
2nd Assigned Species	<i>Taurotragus derbianus</i>
>0	290660
>50	5790
>100	4730



Supplementary figure 23. Damage pattern analysis for all the aligned reads for sample SHH7357.



Supplementary figure 24. Damage pattern analysis of the aligned reads for taxon *Pelea capreolus*, for sample SHH7357. A high frequency of deaminations can be observed at the 5' end of the reads.

South American Canids (Minor Chapter)

Introduction

When and what species of canids appeared in South America

Species of family Canidae first appeared in South America in the late Pliocene, with most of its species arriving to the continent later in the Pleistocene (Prevosti et al., 2009). As preserved fox specimens have been found in multiple early archaeological contexts, the archaeological record indicates that humans inhabiting South America during the early Holocene were interacting with foxes endemic to various parts of the continent (Stahl, 2013, 2012). Many of these fox species are still extant and occupy different niches on the continent (Stahl, 2012).

Contemporary South American canids include both wolf- and fox-like species. Wolf-like species include *Chrysocyon brachyurus* and *Speothos venaticus*. The maned wolf *Chrysocyon brachyurus* is a solitary opportunist that forages in the grasslands of central Brazil (Dietz, 1985), whereas *Speothos venaticus* is a carnivorous bushdog that inhabits lowland areas and forests from Panama to Brazil (Beisiegel and Zuercher, 2005).

Fox-like canids in South America also include multiple species from a range of genera. Solitary generalist species such as *Atelocynus microtus* (Leite Pitman and Williams, 2004) and *Cerdocyon thous* (crab-eating fox) (Courtenay and Maffei, 2004) can be found in multiple habitats across Amazonia. The most predominant genus of

native South American fox-like canids though is the genus *Lycalopex*. It includes species such as *Lycalopex culpaeus* (culpeo) (Novaro, 1997), *Lycalopex fulvipes* (Darwin's fox) (Jiménez and McMahon, 2004), *Lycalopex griseus* (grey fox) (González del Solar and Rau, 2004), *Lycalopex gymnocercus* (pampas) (Lucherini, 2004), *Lycalopex sechurae* (Sechuran fox) (Asa and Cossíos, 2004), and *Lycalopex vetulus* (hoary fox) (Dalponte and Courtenay, 2004). All the endemic South American foxes share certain ecological and ethological traits such as broad diets, an opportunistic and solitary behaviour, the ability to inhabit a wide range of ecological niches, and a tendency to habituate to interacting with humans and their respective environments (Stahl, 2012).

The extant South American foxes who have been connected with archaeological findings are *Lycalopex culpaeus* (culpeo), *Lycalopex gymnocercus* (pampas), *Lycalopex sechurae* (Sechuran fox), and *Lycalopex griseus* (grey fox). All of these species are mainly carnivores that occasionally can also function as omnivores, with behavioral characteristics that allow them to live and interact with humans (Stahl, 2012). This relationship between endemic fox species and humans can be corroborated by the historic records, which describe docile South American foxes kept either in captivity or as tamed pets (Lucherini, 2004; Meza, 2004). Similar observations have also been made for *Lycalopex fulvipes* (Darwin's fox) that can be found in the Caribbean Islands (Jiménez and McMahon, 2004), and reportedly was kept as a semi-domesticated pet by indigenous groups of people. Even though these species of foxes were never domesticated in the same way and sense that dogs were, archaeological evidence does suggest a lasting relationship/symbiosis of these species with humans (Stahl, 2012).

Domesticated dogs in South America

Domestic dogs (*Canis lupus familiaris*) are thought to have accompanied early human populations coming into North America from Asia (Fiedel, 2005). At the same time it is hard to provide definitive proof of the domestic dog's presence in South America, with some speculating either its absence from the Amazon basin until recently (Koster, 2009) or a much later introduction during the Holocene (Prevosti et al., 2009). One of the reasons, hindering our ability to argue for an earlier ancient appearance of the domestic dog in South America, is that archaeological canid bones found in old contexts do not have sufficient preserved osteological markers for a robust morphological identification leading to ambiguous calls (Stahl, 2012).

It has been hypothesized that in the absence of a domestic dog early human tribes in South America were interacting with endemic fox species, assuming a symbiotic relationship that encompassed commensalism and mutualism (Stahl, 2013, 2012). Examples of such a relationship include archaeological from the Las Vegas culture evidence of taming of the Sechuran fox (Byrd, 1976), and mute, tamed "dogs" in Panama and Nicaragua (where sometimes they were also eaten in feasts) (Stahl, 2013).

Hypothesis testing

In this chapter I try to identify the species that bones of South American canids belong to. Leveraging developments in High Throughput Sequencing (allowing for

cheaper generation of large volumes of DNA sequence data), completeness of genomic databases (especially for mammalian mitochondrial genomes), and bioinformatic methods (that allow for robust species identification from just a handful of DNA molecules), I aim to taxonomically identify the species of preserved South American Canid samples, belonging to the archaeological sample collection of the Palaeobarn laboratory in the University of Oxford.

Materials and methods

Species Identification Pipeline Method

A common hurdle in many species identification methods via DNA is misidentification, due to the use of the descriptive statistics that are employed by them. As a result, taxonomic identification results from such methods cannot be validated, leading to uncertain sample identifications. This problem becomes prevalent when analysing metagenomic samples. In our proposed method, we employ a hypothesis driven approach by asking if our sample is species X, and therefore we provide a probabilistic estimation of how well supported our taxonomic assignments are. It needs to be underlined that a low assignment probability for species X does not disprove that our sample doesn't belong to that species, but it should be interpreted as our sample's yielded DNA sequences not containing enough information to perform a robust species identification.

In order to identify a certain species robustly, we first have to obtain a relative alignment score for each combination of DNA read and taxon of interest. A means of doing that is using SIGMA's alignment score matrix (Ahn, Chai, & Pan, 2015). SIGMA, by using the bowtie2 (Langmead & Salzberg, 2012) aligner, aligns the reads of our DNA library against the genomes of our taxa of interest and produces an alignment score, called a Q score which is an evaluation of the quality of the alignment. The higher the alignment score, the better the quality of the alignment, and the probability of the sequences being a true match is higher.

A library of ancient DNA though contains a high number of short reads, therefore contains many query sequences. Consequently, it is of great use to obtain an overall

score across all the reads of a library per taxon of interest. For this summarization to happen we make use of the following equations. The overall alignment score for each library/species or library/subspecies combination is given by the formula: . This is the sum of the alignment scores (Q scores), that evaluate the quality of an alignment between a DNA read and a sequence of a specific taxon.

Implementation

- a) Initially the questions “Is my sample species X?” or “Does species X exist in my sample?” need to be defined. Therefore, a reference genome of the species/taxon X needs to be downloaded. If species/taxon X is a group of taxa (eg. all mammals), all the genomes of that group can act as a collection of reference genomes.
- b) The pre-processed reads (adapters removed and quality trimmed) included in the DNA library of the sample under study are aligned with SIGMA against the reference genome, by using bowtie2. Only reads which produce an alignment with the reference genome, whose score is above the specified quality threshold, are kept. The rest of the DNA reads are discarded. Sequencing adaptors should have been removed prior to this step.
- c) The selected, potentially positive-hit DNA reads are filtered further by removing duplicate reads. For the duplicate removal, both start and end coordinates of a DNA fragment are taken into account.
- d) After obtaining our filtered unique reads, a damage pattern analysis is performed on them by MapDamage (Jónsson, Ginolhac, Schubert, Johnson, & Orlando, 2013), to assess if our DNA reads are truly ancient and what is the extent of the damage they have undergone.

e) A crucial step to implement this method is to build a custom RefSeq-like database and include the taxon of interest or a taxon that can be used as a proxy. Such a taxon is one that is phylogenetically closer to the taxon of interest than all other sequenced taxa (Figure 3). In order to build a database that actually enables us to produce reliable probabilistic estimations of identity a few problems need to be tackled. First, the size of the database needs to be reasonably large, otherwise the Q scores will be distorted and unreliable. This is because when a database is small the probability of observing a match by chance increases. Also the size of the sequences that are contained in the database should be of similar length, otherwise the Q scores would be biased. Having target sequences that differ greatly in length could lead to Q scores that could be hard to interpret, as larger sequences would allow more alignments to take place than shorter sequences. Another hurdle for this kind of analysis is the unequal representation of the taxa in the database. This is because many taxa have been sequenced more thoroughly than others, and therefore they are many more publicly available sequences for these taxa. For instance, if we are interested in identifying bones of vertebrate animals the first two problems can be overcome if the database is built with the sequences of whole mtDNA genomes from many potential species. For example, if the query sequences come from mice then the database will include the full mtDNA sequences of all the species of the Muridae family. This approach will give the database the size needed for the Q scores to make sense in terms of probability. Furthermore, given that the mtDNA genomes are approximately of the same size, the issue of partial

alignments will disappear. Likewise, most species of bacteria in the same genus will also have very similarly sized genomes, i.e. all species in the genus *Streptococcus* have a genome size of ~2Mbp. Though, if the sample under study is a metagenome, we might need to readjust the size and the number of taxa included in the custom database, depending on the available data for the taxa of interest. For the third major problem to be tackled, and ensure equal taxonomic representation in the custom-made database, only one sequence per taxonomic level is included. For instance, if we want to perform an identification on the species level, the database is going to include one sequence per species, but if we want to increase our resolution and identify the subspecies of an organism our database will include one sequence per subspecies.



Figure 3. Example of using a taxon as a proxy to infer for closely related taxa of interest.

- f) Afterwards the library of the query sequences is aligned against the custom-made database, using SIGMA. The product of this alignment is a matrix (Q matrix) generated by the SIGMA algorithm, which contains the alignment scores for every read/taxon combination. Alignments, whose quality is lower than a predefined threshold, are given the value of 0 and they do not get reported in the matrix.

g) Then an in-house developed algorithm that sorts the SIGMA Q matrix and computes the probability scores is employed:

- i) Create a database/dataframe from SIGMA's Q-matrix. The rows are representing the DNA reads of our sample and the columns are the different taxa that are included in our custom-made database. After the data frame is constructed, the sum of the Q scores for every column (taxon) is computed.
- ii) The vector is sorted according to the values of the taxa in descending order.
- iii) Then the I_g values (the overall alignment score) for each genus are computed.
- iv) The vector of the I_g values of each taxon is sorted again and the two largest I_g values are chosen.
- v) Then the relative score for each best and second best taxon, which can be interpreted as a posterior probability, is computed with the equation: $p_g = I_g /$

$$\sum_{i=1}^n 1/I_i.$$

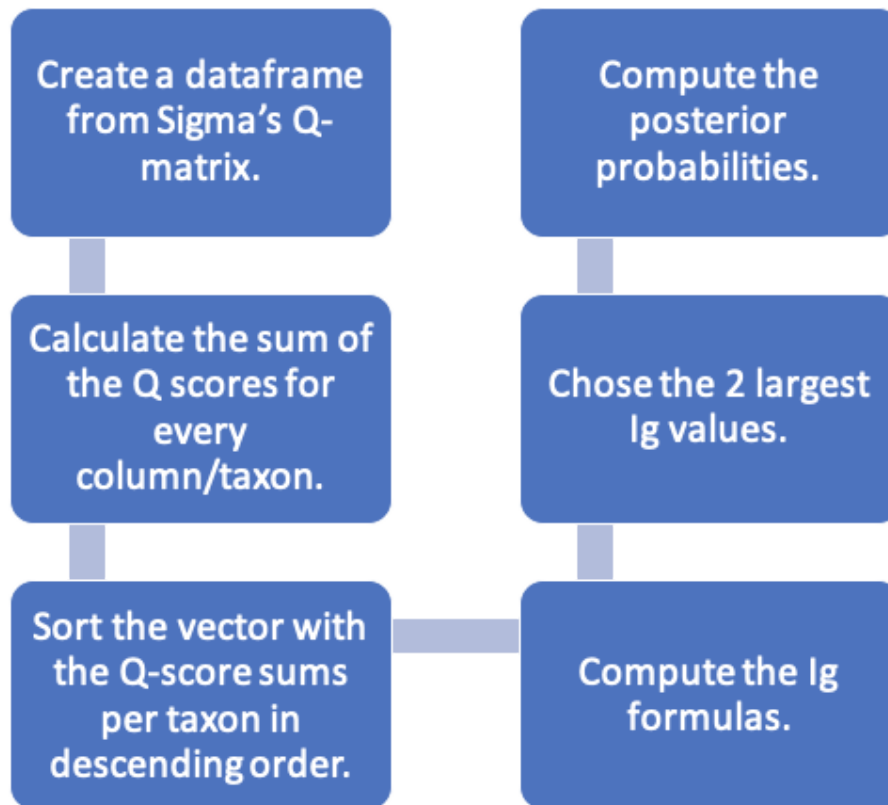


Figure 4 Workflow of the algorithm that parses Sigma's Q matrix.

- h) After the posterior probabilities have been calculated, the 10 first taxa from the vector of I_g values are chosen. These taxa are used to subset the initial dataframe that was constructed, and create a new one only with the columns of interest. The alignment scores of every read per taxon are then plotted, and a nice visualization of the distribution of the alignment scores per taxon is created.
- i) Following the identification of the best taxon, the DNA reads that are aligned specifically to the genome of that taxon can be extracted and be used as input for a second round of damage pattern analysis. This second

step of MapDamage analysis allows us to corroborate that the reads on which the identification of the best species was based on are truly ancient.

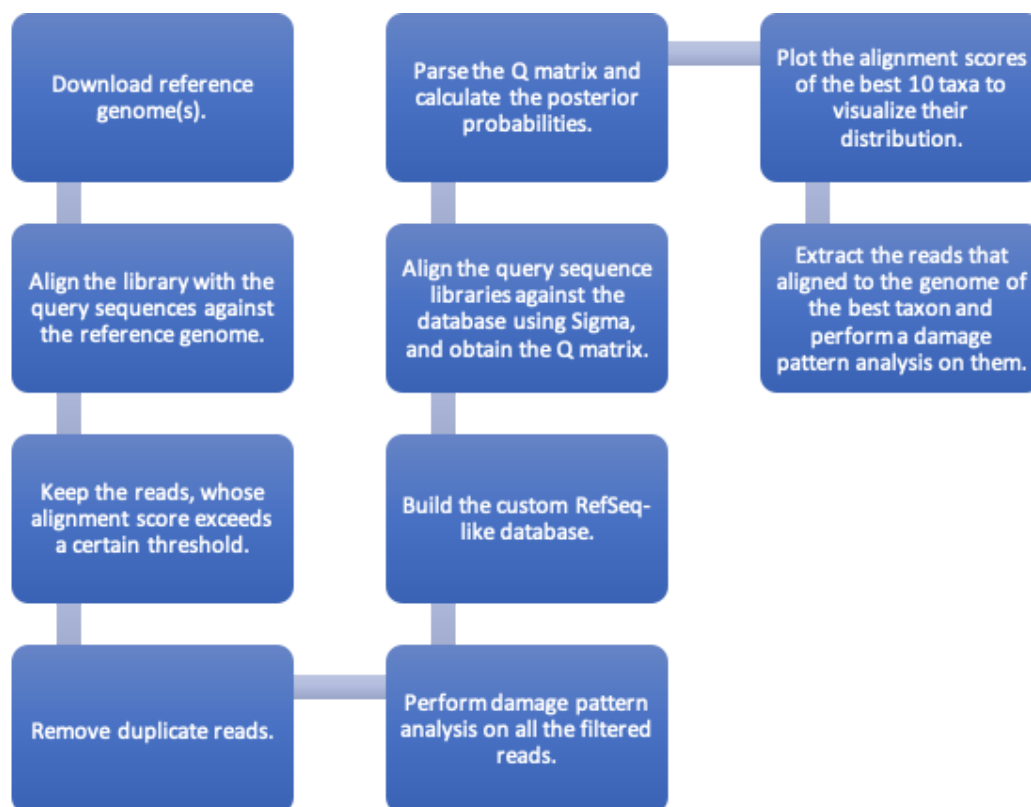


Figure 5 Workflow of the species identification pipeline.

Notes on the construction of the local database.

Some additional steps for the construction of the custom-made database need to be highlighted. Initially a search query is posted on NCBI, requesting accession numbers of complete genome sequences (sequence length parameters are often specified). After obtaining a list of accession numbers, their metadata – including information about the sequence records like length, along with various taxonomic information – are downloaded. Depending on the taxonomic level somebody wishes to focus on (e.g. genus, species, subspecies, etc.), only sequence records that have

been taxonomically characterized on that specific level at least are chosen. The rest of the accession numbers are filtered out. To maintain equal species representation in our database, and avoid as much bias as possible at the alignment stage, only one sequence/accession number is selected per taxon, usually the longest one. By the end of this process, a final list of taxa with their respective accession numbers is outputted, and these accession numbers are used to construct the database using a structure based on NCBI Ref-Seq.

Interpretation of the results

The posterior probability that is computed should not be considered as the probability that the sample under examination belongs to the corresponding taxon, but rather to a species that is most closely related to this taxon. It is a rather conservative method of assigning species to samples, as if a sample each time will produce a high probability for the species that is the most closely related to and really low probabilities for the remaining taxa of our database. A reliable species assignment would take place, as a rule of thumb, if the posteriors that are produced exceed 0.5, depending every time on the taxa that we try to identify and their biology. Subsetting the matrix independently for each species is crucial for confident identification of the species of interest. This is because after obtaining the posteriors from the whole of the matrix we can assess which ones are the major taxa in our metagenomic sample, yet many times the signal that is being picked up is weak, and the resulting posteriors might not be high enough to perform a robust identification. Consequently focusing on one species every time, by analyzing the alignment scores from the reads whose best match was the species/taxon of interest, allows us

to mathematically address the question “Is my sample species X?” and therefore increase the sensitivity of our species identification method.

An important limitation of our method is the amount of information that a certain library contains. Libraries with significantly short reads or low amounts of endogenous DNA can lead to equally low posteriors for many taxa. In the case that our sample is or contains a species that doesn't exist in the custom database, then this library will also get roughly equally low posteriors for all the taxa in our database. Another special case is one where our sample contains a taxon, which is evolutionarily equidistant/equally related to a certain number species/taxa in our database. As a result (relatively high) equal posteriors will be calculated for these certain species/taxa and really low probabilities will be produced for the remaining taxa of our database.

Applications

- a) Single organism sample – one species: When there is a bone assemblage with highly fragmented bones that have lost their diagnostic zones, but it can still be gauged what kind of animal they used to be. By asking the question “Is my sample species X?” a taxonomic identification can be performed. The reference genome for this case of identification is from one species only. The posterior probability produced provides an estimation of how likely this identification is.
- b) Single organism sample – multiple species: The main difference from the previous case is that the bones of that assemblage have lost a great deal of their morphological features, so that an expert can only assign them to a

larger taxonomic group, e.g. mammals. In this case, our reference genome consists of the genomes of all the different species in this larger taxonomic group, in our case it's going to be all the mitochondrial genomes of the mammals. Therefore, the question posed this time is "Is my sample any of these taxa?".

Results

Taxonomic identifications based on DNA, employing the in-house developed method mentioned in the Materials and Methods section, were performed on 173 samples of South American canids. The DNA sequences from each sample were compared against a database of all the complete mammalian mitochondrial genomes, corresponding to 1192 species.

Unfortunately the sequenced libraries of South American canid samples did not contain enough mitochondrial DNA reads to perform a species identification on the species level. In an effort to maximise the informational content of the mitochondrial reads in these libraries we performed a second round of taxonomic identifications, but this time we focused on the genus level. Confidently taxonomically identifying the analysed canid samples on the genus level would still allow us to distinguish domestic dogs from species of South American endemic foxes, as these organisms belong to distinct genera. This way we were able to robustly identify (assignment probability higher than 54%) the genus of 12 South American canids from approximately 14000 years before present (BP) to the present day, based on their mitochondrial DNA.

Table 1. Genus assignment probabilities for the 12 South American canid samples that were positively identified from their mitochondrial DNA sequences.

Sample	Assignment probability	Assigned genus	Number of Reads	Relative Abundance (%)	Filtered Reads	Country	Mean Age BP
AD2010	0.62	<i>Canis</i>	45	95.74	47	Chile	1150
AD2017	0.63	<i>Canis</i>	116	97.48	119	Bolivia	1250
AL2125	0.85	<i>Lycalopex</i>	67	95.71	70	Peru	NA
AL2829	0.54	<i>Canis</i>	2834	86.22	3287	Peru	750
AL2831	0.58	<i>Lycalopex</i>	53	77.94 1176 47	68	Peru	925
AL2838	0.72	<i>Canis</i>	12031	96.10	12519	USA	625
AL2839	0.63	<i>Canis</i>	1947	90.60	2149	Peru	400
AL3206	0.55	<i>Canis</i>	3410	84.01	4059	Peru	0
AL3230	0.57	<i>Canis</i>	6618	92.30 23	7170	Peru	0
AL3231	0.70	<i>Canis</i>	1245	93.19	1336	Mexico	1000
ANH-129	0.63	<i>Canis</i>	218	96.46	226	Mexico	450
OL2107	0.56	<i>Chrysocyon</i>	43	37.39	115	Argentina	NA

Out of the 12 positively identified canid samples 9 belonged to the genus *Canis*, possibly being domestic dogs or any other wild canid, and 3 samples belonged to the South American endemic fox genera *Lycalopex* and *Chrysocyon*. The two samples belonging to the genus *Lycalopex* both were retrieved from Peru, but unfortunately age metadata exist for only one of them, AL2831, that dates to the Pre-Columbian era (925 years BP). One archaeological sample from Argentina was also identified as belonging to the genus *Chrysocyon*, but unfortunately again no age metadata was available. The remaining 9 samples that were identified as members of the genus *Canis* can be traced across a wide geographic range, including Mexico, Peru, Chile, Bolivia, and the USA, with an age range spanning from 1250 years BP (Pre-Columbian era) to the present day.

Discussion

Our results support the presence of endemic South American fox species in the archaeological record of various South American countries. These samples had been previously identified as small domestic or feral dogs, but by using mitochondrial DNA sequences we identified 3 endemic fox samples. Because of the restricted informational content in the sequenced DNA libraries of these canid samples, we used the genus level for our identifications, as there was not enough discriminatory power for a species level identification. Furthermore as domestic dogs and endemic South American foxes belong to different genera, an identification on the genus level would still provide enough taxonomic resolution to assess the prevalence of domestic dogs and endemic foxes in faunal assemblages from South America.

The described species identification method highly depends on the informational content of the DNA sequences that are yielded from a sample. Highly fragmented and damaged DNA molecules, along with DNA molecules from repetitive or hyper-conserved regions, do not have informative enough sequences to be able to distinguish between different taxa, especially on the species level. Having a DNA library rich in polymorphic, diverse and informative sequences increases the power of our analysis by increasing the resolution of our taxonomic identifications, allowing us to go as low as the subspecies level in some cases.

The biology of the taxa that we are trying to identify should be also considered. Genetic diversity the taxa under study is a key factor to a confident species identification. Closely related species for example can be a hindrance to our analysis, as their DNA sequences might not be different enough to allow us to distinguish between them. This can be aggravated by not having insufficient or

uninformative DNA reads in the libraries that are being analysed. An interesting finding from the above analysis was the low number of mitochondrial reads available in the sequenced libraries of the South American foxes (68 - 115 for the genera *Lycalopex* and *Chrysocyon*). Unfortunately the low number and informational content of these mitochondrial reads hindered our ability to identify these samples on the species level, motivating us to focus on the higher taxonomic level of genus.

Even though the class of mammals has a high species representation of complete mitochondrial genomes and nuclear genome assemblies in public genomic databases, there are still species of mammals with not even a complete mitochondrial genome available. Such is the case for the genera *Chrysocyon* and *Lycalopex*, since many of the species they include do not have any complete mitochondrial genomes available, further confounding our taxonomic identification analysis and forcing us to identify these samples on the genus level. Additionally, our ability to successfully identify these canid samples is hindered by the fact that certain species of endemic South American foxes have gone extinct. The extinction of these species is limiting our ability to generate reference mitochondrial genomes that would allow us to identify these extinct species robustly, as now we have to rely on capturing the past genetic variation of those extinct fox taxa through DNA data generated from extant fox species.

This highlights that for a robust species identification to take place a well curated database needs to be in place. The database needs to include as many related taxa as possible to the chosen candidate species that we are trying to identify. Equal species representation in the database needs to be ensured, as well as making sure that the sequence records have total lengths of similar order of magnitude. Attention needs to be drawn to 2 special cases that relate to the contents of the database,

which require special interpretation of the identification results. The case is when our sample comes from a species/taxon whose sequence record isn't included in the database. This will lead to yielding many equally low assignment probabilities for most of the taxa in our database. The second special case is when trying to identify a taxon, which is highly genetically similar with its related taxa. In this case it's possible that our method cannot distinguish among them, therefore yielding equal probabilities for all these taxa.

The described method is quite conservative in assigning a taxonomic identification to a sample. More efforts should be made for potential improvement of the method and include components that will increase its sensitivity even when the variability of a taxon is low. Additional ways of extracting information from libraries that contain low amounts of DNA should be developed along with methodologies that make optimal use of really short DNA reads that are contained in DNA libraries of ancient samples. When the samples under examination are ancient or highly degraded, it would be useful to try and enrich our libraries with as many as possible mtDNA molecules. Thus using techniques such as mitochondrial DNA capture, which increases the quantity of mtDNA fragments in our libraries, increases the strength of our test (Slon et al., 2016).

Bibliography

- Asa, C., Cossíos, E.D., 2004. Sechuran fox (*Pseudalopex sechurae*). Canids: Foxes, Wolves, Jackals, and Dogs. Status Survey and Conservation Action Plan 72–75.
- Beisiegel, B. de M., Zuercher, G.L., 2005. *Speothos venaticus*. Mammalian Species 783, 1–6.
- Byrd, K.M., 1976. Changing animal utilization patterns and their implications: southwest Ecuador (6500 BC-AD 1400). University of Florida.
- Courtenay, O., Maffei, L., 2004. Crab-eating fox *Cerdocyon thous* (Linnaeus, 1766). Canids: foxes, wolves, jackals and dogs. Status survey and conservation action plan 32–38.
- Dalponete, J.C., Courtenay, O., 2004. Hoary fox *Pseudalopex vetulus* (Lund, 1842). Canids: foxes, wolves, jackals and dogs. Status survey and conservation action plan (C. Sillero-Zubiri, M. Hoffmann, and DW Macdonald, eds.). International Union for Conservation of Nature and Natural Resources/Species Survival Commission Canid Specialist Group, Gland, Switzerland 72–76.
- Dietz, J.M., 1985. *Chrysocyon brachyurus*. American Society of Mammalogists: Mammalian Species.
- Fiedel, S.J., 2005. Man's best friend – mammoth's worst enemy? A speculative essay on the role of dogs in Paleoindian colonization and megafaunal extinction. World Archaeol. 37, 11–25.
- González del Solar, R., Rau, J., 2004. Chilla (*Pseudalopex griseus*). Canids: Foxes, Wolves, Jackals, and Dogs. Status Survey and Conservation Action Plan 56–62.
- Jiménez, J., McMahon, E., 2004. Darwin's fox *Pseudalopex fulvipes* (Martin, 1837).

- Canids: foxes, wolves, jackals and dogs. Status survey and conservation action plan. IUCN/SSC Canid Specialist Group. Gland, Switzerland and Cambridge, UK 50–55.
- Koster, J., 2009. Hunting Dogs in the Lowland Neotropics. *J. Anthropol. Res.* 65, 575–610.
- Leite Pitman, M.R.P., Williams, R.S.R., 2004. The short-eared dog (*Atelocynus microtis*). Canids: species status and conservation. Action plan (eds Sillero-Zubiri C, Gingsberg JR, Macdonald D.) 26–31.
- Lucherini, M., 2004. Pampas fox (*Pseudalopex gymnocercus*). Canids: foxes, wolves, jackals, and dogs. Status survey and conservation action plan.
- Meza, C., 2004. Relaciones entre el zorro de Sechura, *Pseudalopex sechurae* (Thomas), y el hombre en el Perú. *Ecología Aplicada* 3, 134–138.
- Novaro, A.J., 1997. *Pseudalopex culpaeus*. *Mammalian Species* 1–8.
- Prevosti, F.J., Tonni, E.P., Bidegain, J.C., 2009. Stratigraphic range of the large canids (Carnivora, Canidae) in South America, and its relevance to quaternary biostratigraphy. *Quat. Int.* 210, 76–81.
- Slon, V., Glocke, I., Barkai, R., Gopher, A., Hershkovitz, I., Meyer, M., 2016. Mammalian mitochondrial capture, a tool for rapid screening of DNA preservation in faunal and undiagnostic remains, and its application to Middle Pleistocene specimens from Qesem Cave (Israel). *Quat. Int.* 398, 210–218.
- Stahl, P.W., 2013. Early Dogs and Endemic South American Canids of the Spanish Main. *J. Anthropol. Res.* 69, 515–533.
- Stahl, P.W., 2012. Interactions Between Humans and Endemic Canids in Holocene South America. *etbi* 32, 108–127.

Demographic Analysis of sheep populations from Asikli, Turkey, across time (Minor Chapter)

Asikli population

In order to study the demographic dynamics of the possibly first domesticated sheep in the Asikli site in Turkey, we used mitochondrial DNA (mtDNA) sequences of archaeological samples from the that site to test whether we could detect any noticeable human management affecting the matrilineal lineage. The mean calendar year age of the samples was converted in sheep generation times for population demographic modelling purposes, assuming a generation step of 3.6 years (Leroy et al., 2013).

The data from the site was divided into four-time bins. For each time bin, we computed summary statistics: K (the number of segregating haplotypes), S (the number of segregating sites), Tajima's D, Fu's FS, and π (the average pairwise difference), as well as global and pairwise F-statistics. These summary statistics were used for the ABC analysis below.

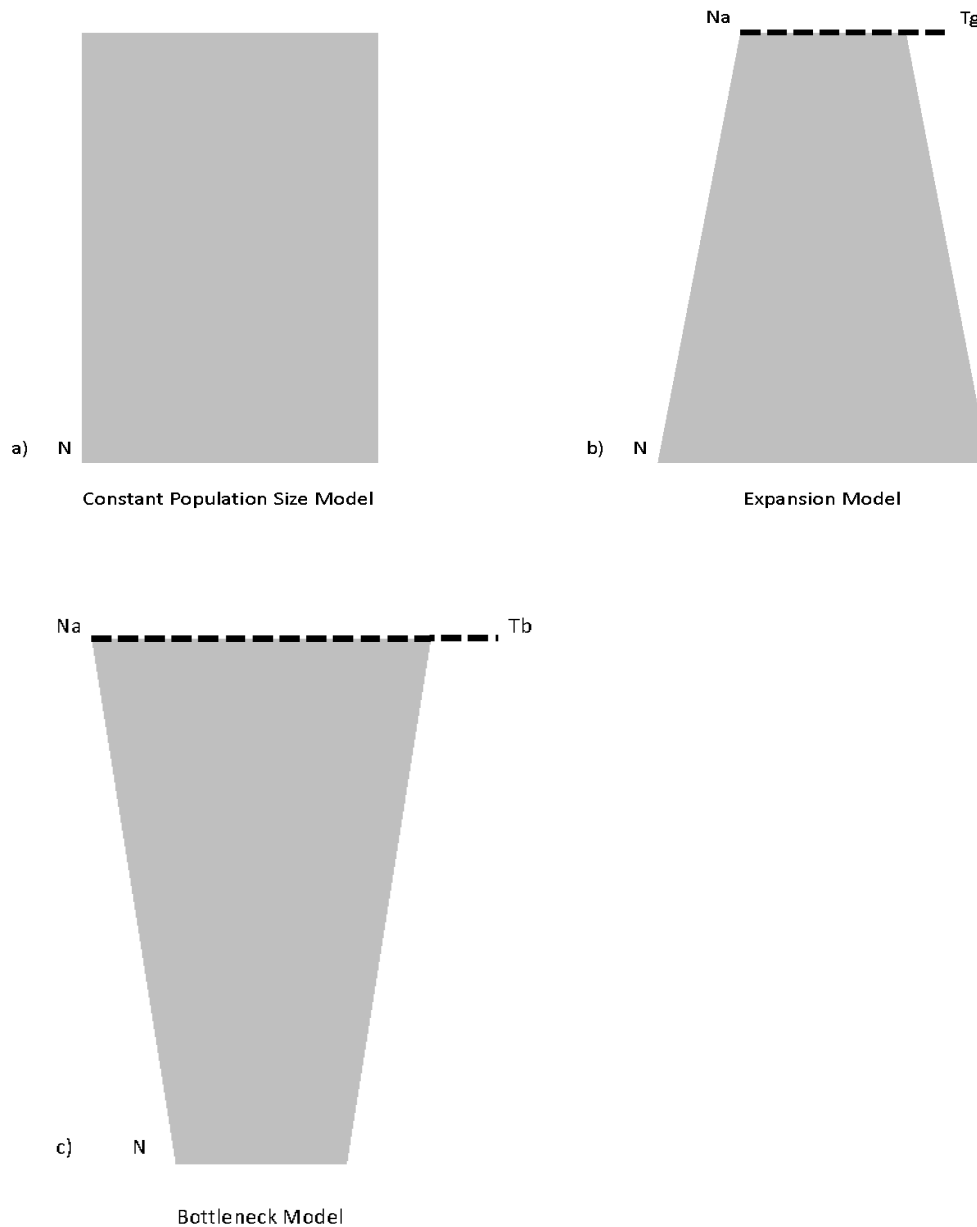


Figure 1: Demographic models of the Asikli archaeological samples. 1a) constant population model. N : number of individuals. 1b) Expansion model. N_a : number of individuals before the expansion event. N : number of individuals at the time of sampling after the expansion. T_g : time of growth.

The calculated observed summary statistics for each time bin were plotted. Also, we summarised pairwise statistics in diagonal matrices.

Table 1. Estimated values of K, S, π and Tajima's D per time bin/population, calculated from observed data.

Pop_Name	K	S	Pi	D
2618_gen_ago	11	258	89.5455	0.0242488
2673_gen_ago	12	281	76.9429	-0.576253
2736_gen_ago	11	288	115.538	-0.112519
2819_gen_ago	19	324	92.3714	0.0559447

We also summarised the population pairwise pi and F statistics.

Table 2. Pairwise F statistics for every pairwise combination of time bins/populations.

	2618_gen_ago	2673_gen_ago	2736_gen_ago	2819_gen_ago
2618_gen_ago		0.00898575	-0.00587792	-0.0351923
2673_gen_ago			-0.00281394	0.0354673
2736_gen_ago				-0.0249707
2819_gen_ago				

Table 3. Pairwise Π statistics for every pairwise combination of time bins/populations.

	2618_gen_ago	2673_gen_ago	2736_gen_ago	2819_gen_ago
2618_gen_ago		83.8303	102.126	87.9134
2673_gen_ago			95.7692	87.9714
2736_gen_ago				101.143
2819_gen_ago				

The statistics from Table 1 were also plotted using R and ggplot2.

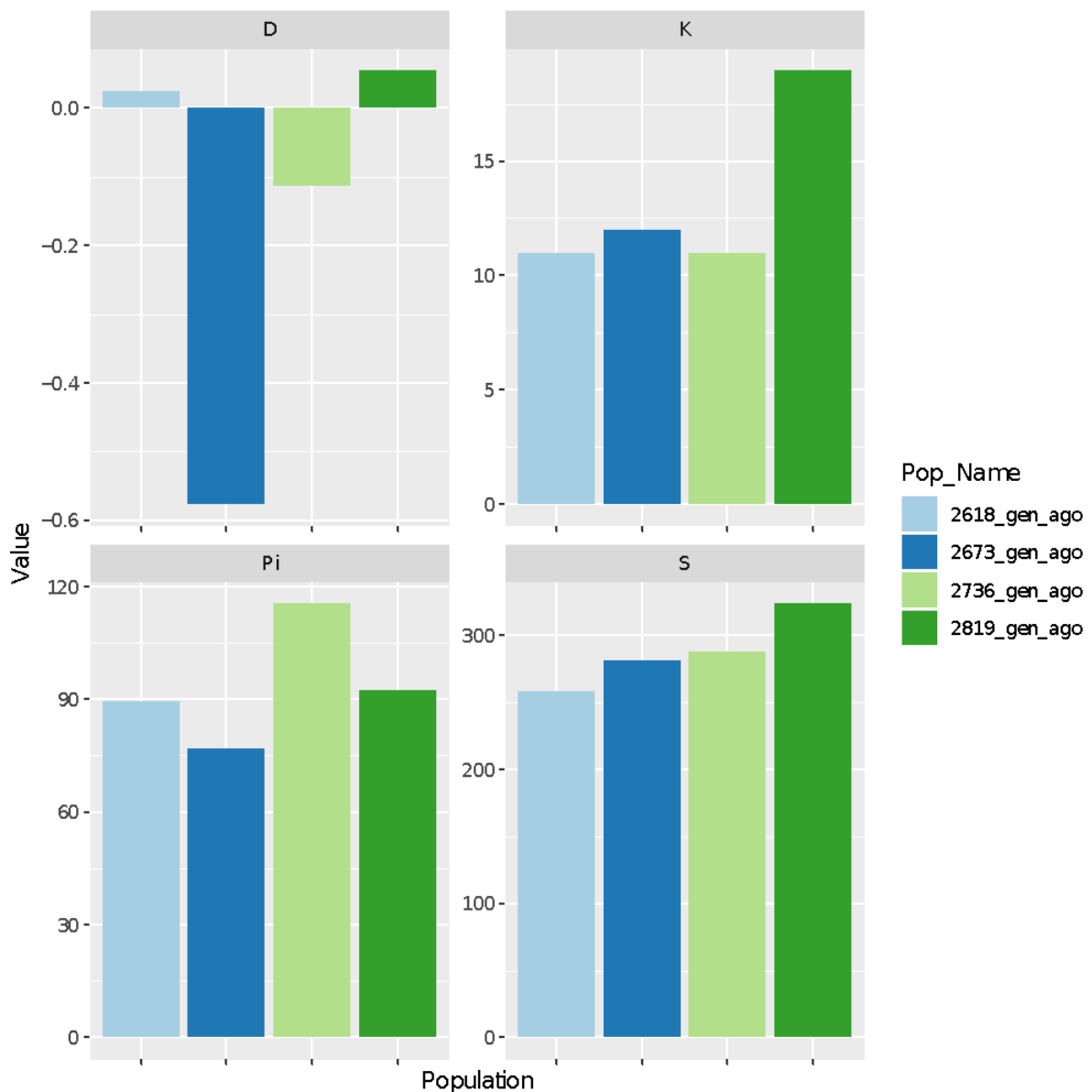


Figure 2. Plots of summary statistics presented in Table 1. These samples were used as the observed data in the varying NE population history models.

The increase of K and S, in the different populations, might be an artefact of the fact that the populations don't have the same number of individuals, but instead the sample size increases the further we go back in time. Tajima's D on the other hand is much more informative for each population in the different time bins. Populations at

2673 and 2736 generations ago have negative Tajima D's values indicating that there might have been a population expansion after a recent bottleneck event. Populations at 2618 and 2819 generations ago have slightly positive D values hinting at maybe a slight population contraction.

ABC analysis

We used fastsimcoal (Excoffier et al., 2013; Excoffier and Foll, 2011) to simulate data under three demographic models: a constant Ne model (one parameter: Ne), a bottleneck model (three parameters: Ne, T: time of bottleneck and Nrel: severity of the bottleneck) and an expansion model (three parameters: Ne, T: time of expansion, and Nrel: severity of the expansion). Parameter values were drawn from uniform distributions presented in Table 4, time units were in sheep generation steps and the population times were recalibrated so that the most recent population had an age of 0 generation steps. Historical samples were simulated as suggested in the fastsimcoal manual. For model testing purposes, we performed 100,000 simulations, for each model, for which we calculated the same summary statistics that were calculated from our observed data of the Asikli site sheep population.

Table 4. Summary of parameters for each demographic model and the distribution characteristics from which each parameter was drawn during the ABC simulations.

	<i>Parameter</i>	<i>Prior</i>	<i>Prior_min</i>	<i>Prior_max</i>
<i>Constant NE</i>	N	Log Uniform	3	5.7
	Mutrate	Uniform	1E-09	0.000001
<i>Bottleneck</i>	N	Log Uniform	3	5.7

	Nanc	Log Uniform	3	5.2
	Mutrate	Uniform	1E-09	0.000001
	Tbot	Uniform	0	2778
<i>Expansion</i>	N	Log Uniform	3	5.7
	Nanc	Log Uniform	3	5.7
	Mutrate	Uniform	1E-09	0.000001
	Texp	Uniform	0	2778

We chose a set of informative summary statistics to apply a partial least-squares discriminant analysis on (Frantz et al., 2015; Peter et al., 2012) using the `plsda` function in R (Lê Cao et al., 2009). After retaining a total of 1,500 simulations closest to the observed data, we compared all models (computing marginal likelihood and posterior probability) simultaneously using a standard ABC generalized linear model (GLM) approach as implemented in ABCtoolbox (Wegmann et al., 2010). The best fitting model was the constant NE size throughout time.

We also computed the average Root Mean Square Error (RMSE) for each parameter using pseudo-observed data to assess our power to infer each parameter in the model.

Table 5. Model fitting on the observed data. The constant NE model fits the observed data the best.

	<i>obsPvalue</i>	<i>Marginal Density</i>	<i>Bayes Factor</i>	<i>Posterior Probability</i>	<i>Chosen Model</i>
<i>Constant NE</i>	0.789	0.00117951	1.16675	0.538479	TRUE
<i>Bottleneck</i>	0.799	0.000581417	0.361345	0.265432	
<i>Expansion</i>	0.634	0.000429524	0.243919	0.196089	

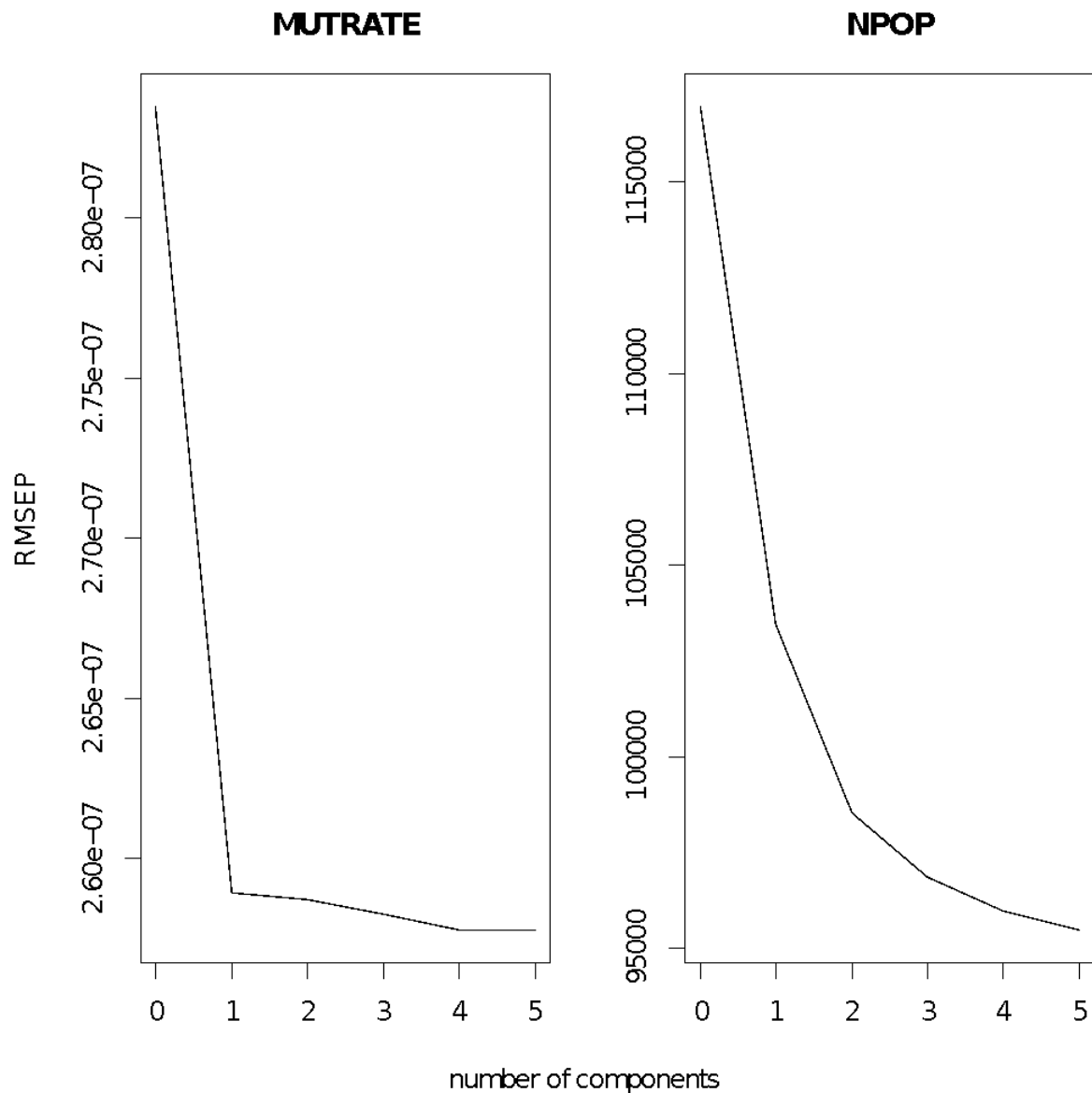


Figure 3. Combined RMSEP. Proportion of the variation observed in the simulated data, explained by the mutation rate and effective population size from the 3 models combined.

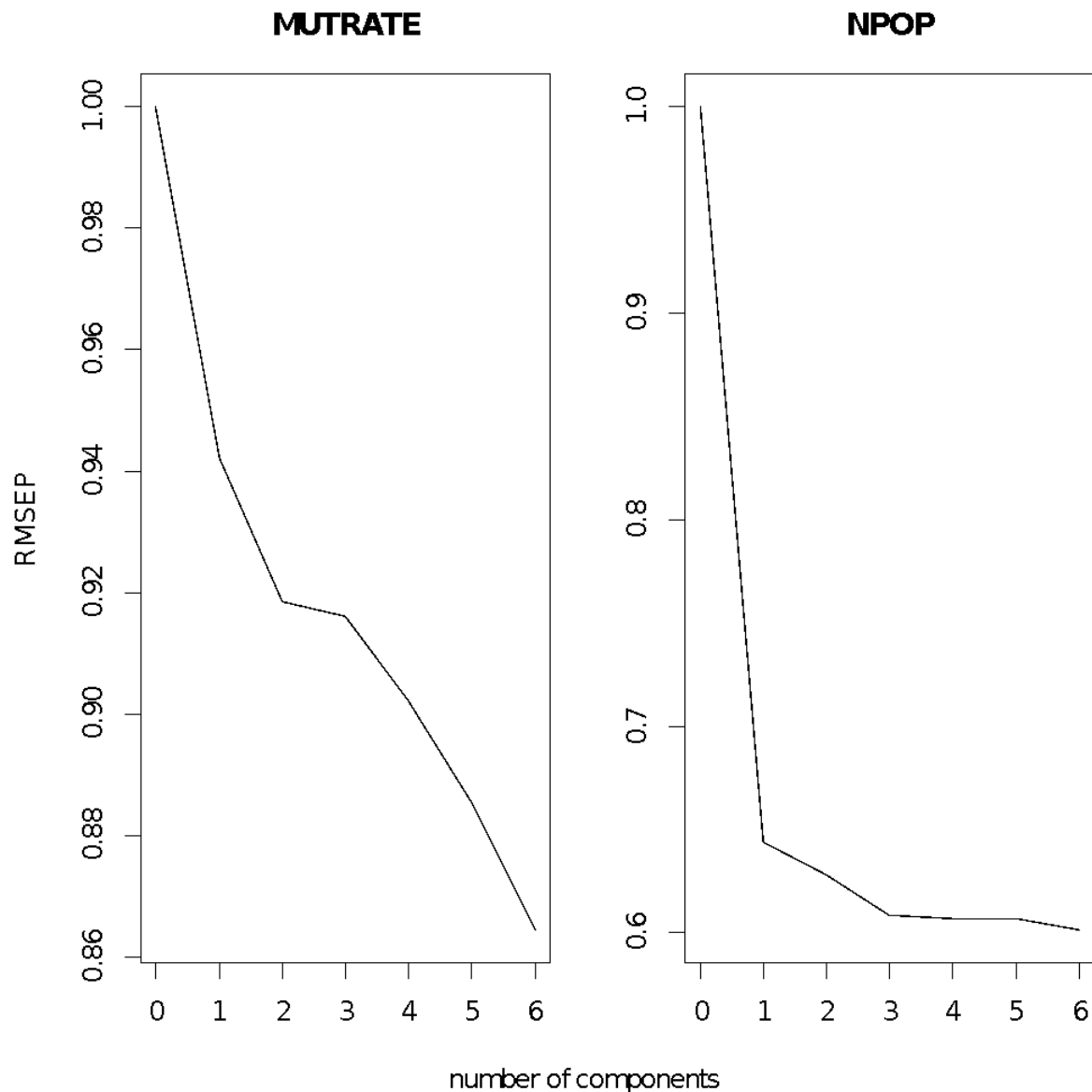


Figure 4. Constant NE RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size.

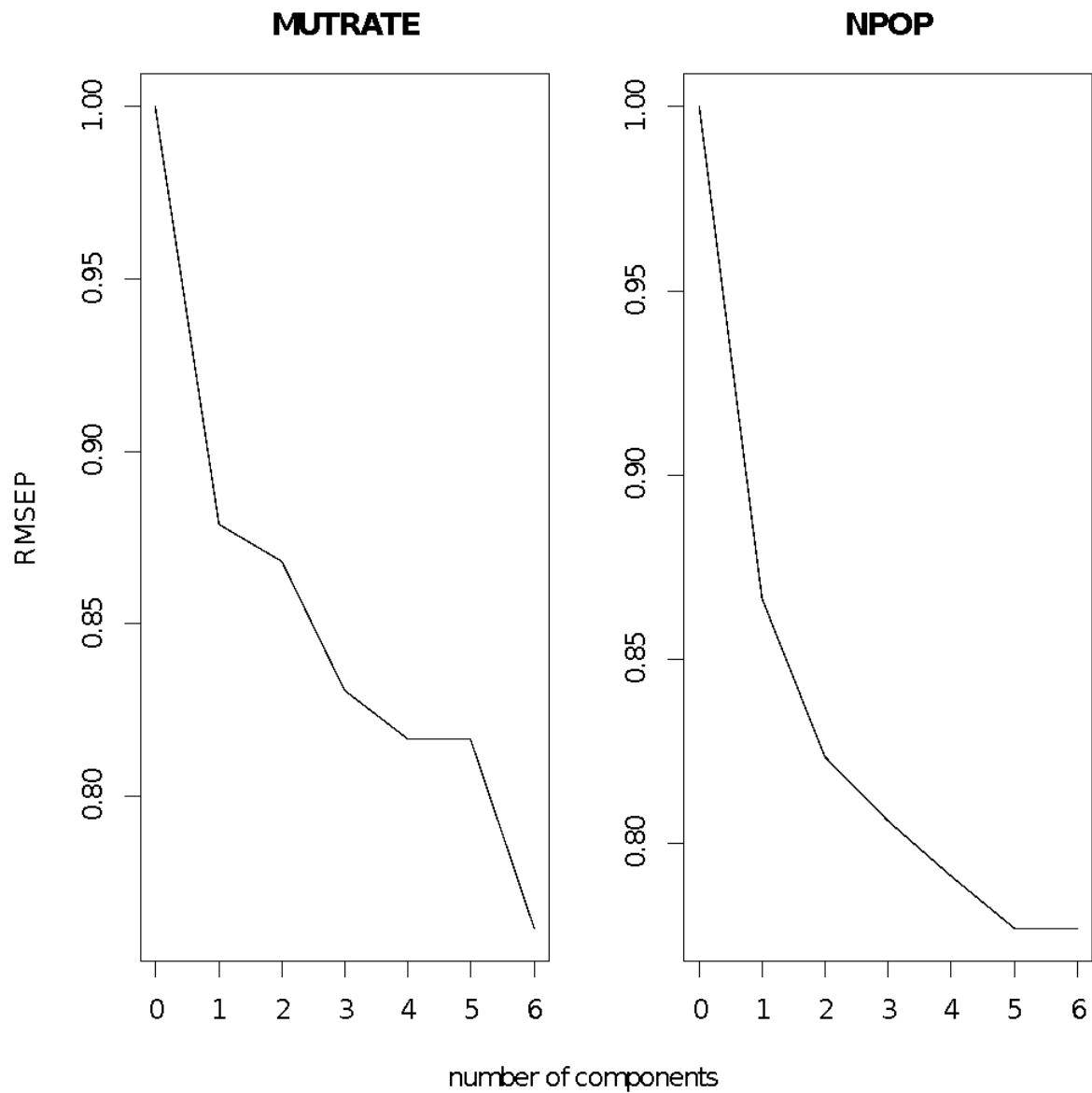


Figure 5. Bottleneck RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size.

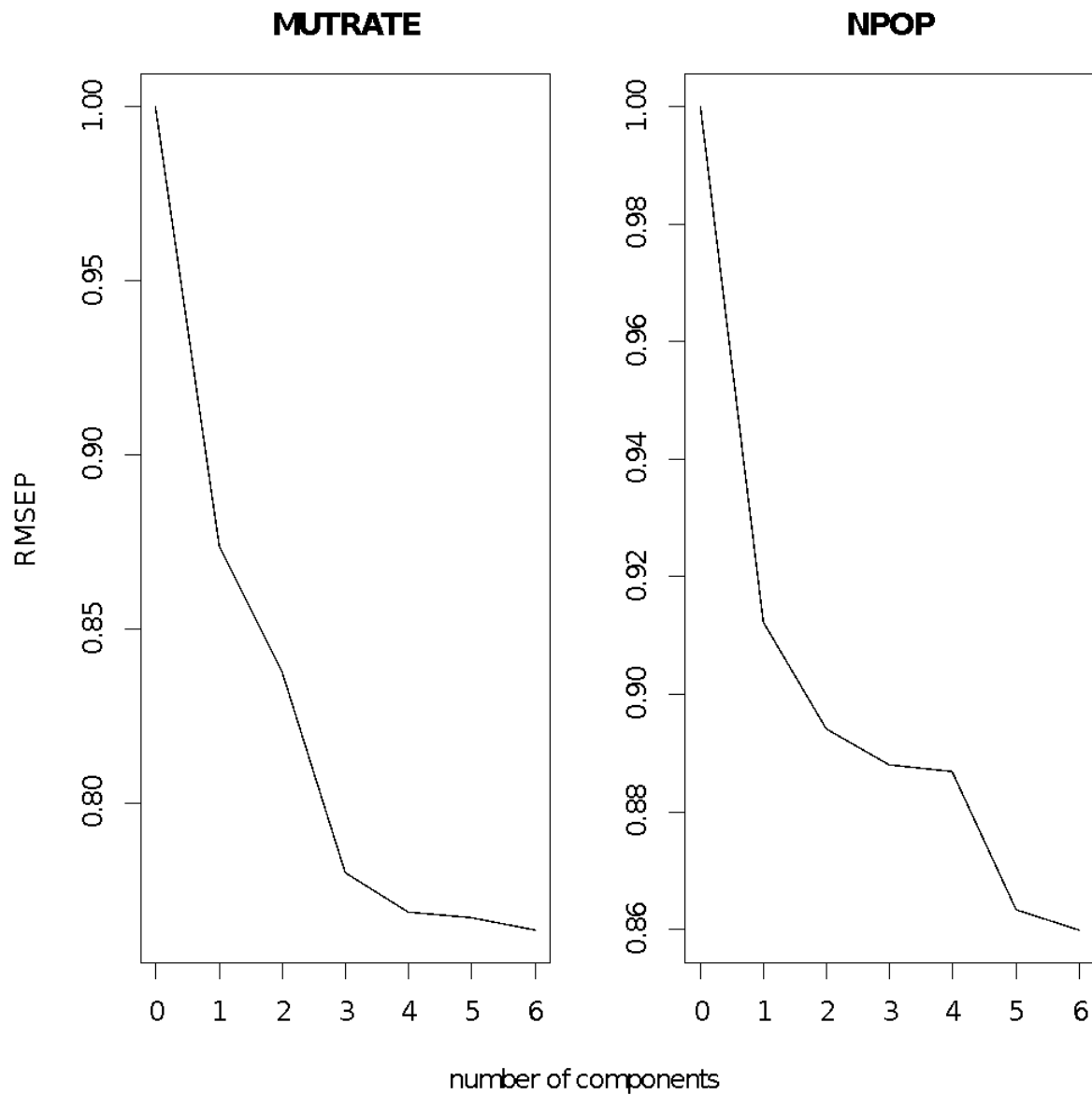


Figure 6. Expansion RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size.

Best Fitting model – Constant NE

After identifying the constant NE size model as the one that fits our observed data the best, model parameter estimation was performed. We ran a total of 2 million simulations under the constant NE model, and for each simulation summary statistics were calculated. These summary statistics were subsequently used as an input for a PLS analysis. The computed PLS were used to transform the summary stats to find the 5 best linear combinations.

The 5 best linear combinations were used as an input for the ABCtoolbox estimator in order to calculate the posterior distributions of NPOP and mutation rate parameters. For this purpose, we retained a total of 10,000 simulations closest to the observed data and applied a standard ABC-GLM. The characteristics of the posterior distribution for each parameter of the constant NE model are summarised in the following table.

Table 6. The characteristics of marginal posterior distributions of the parameters of the constant NE mo. del after 2 million simulations.

Parameter	Prior	Prior min	Prior max	RMSE	Mode	HPDI 95 lower	HPDI 95 upper
N	Log Uniform	3	5.7		3.68182	3.06818	4.61085
Mutrate	Uniform	1.00E-09	0.000001		7.28E-07	1.04E-07	9.95E-07

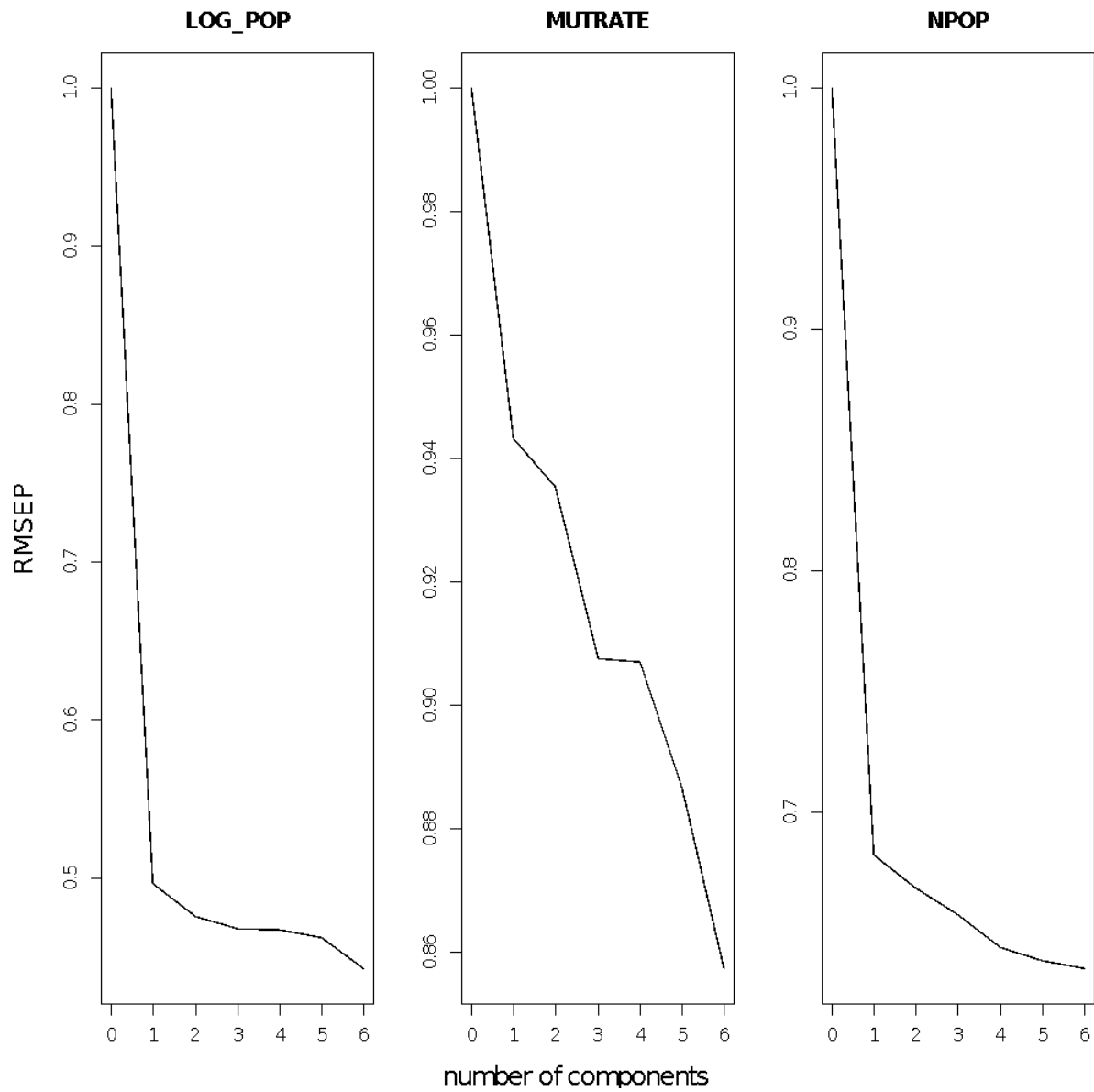


Figure 7. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population (in log and non-log scales) size in the constant NE model.

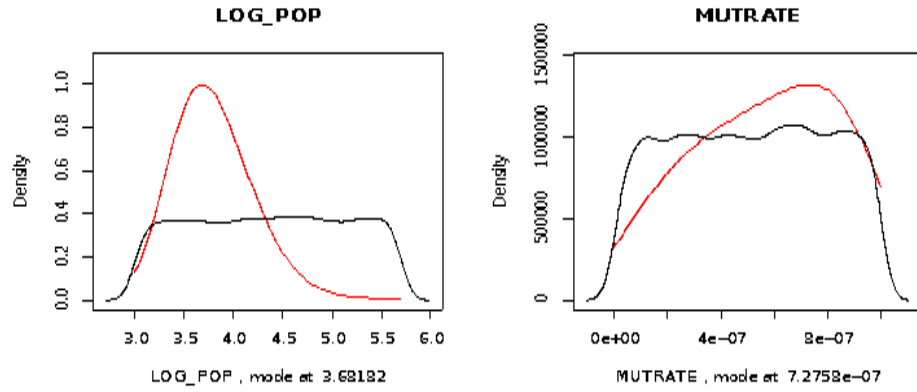


Figure 8. Plots of the uniform posterior distributions of the effective population size and the mutation rate. Blue lines represent the simulated values that have been retained, and the black lines represent the prior distributions of these values.

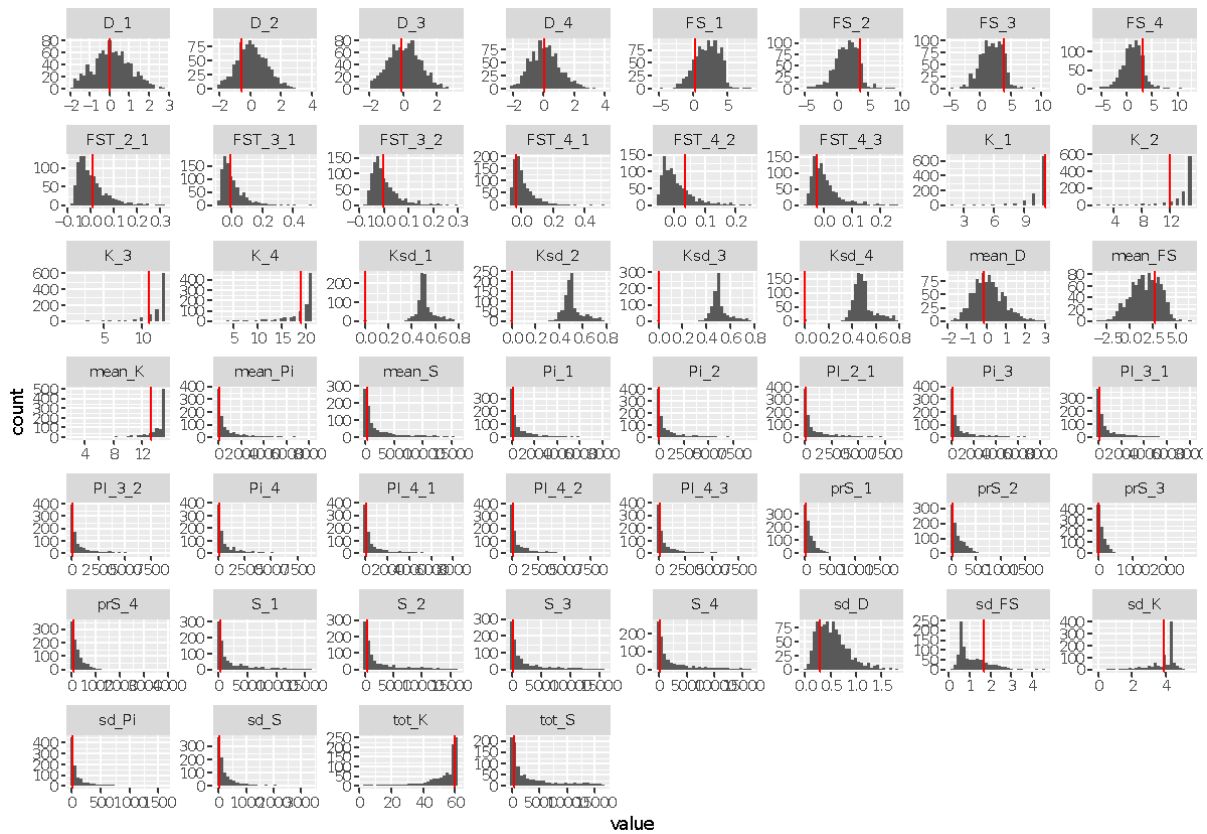


Figure 9. Histogram distribution of the summary statistics of the simulated data. The red bars represent the observed values of the chosen summary statistics.

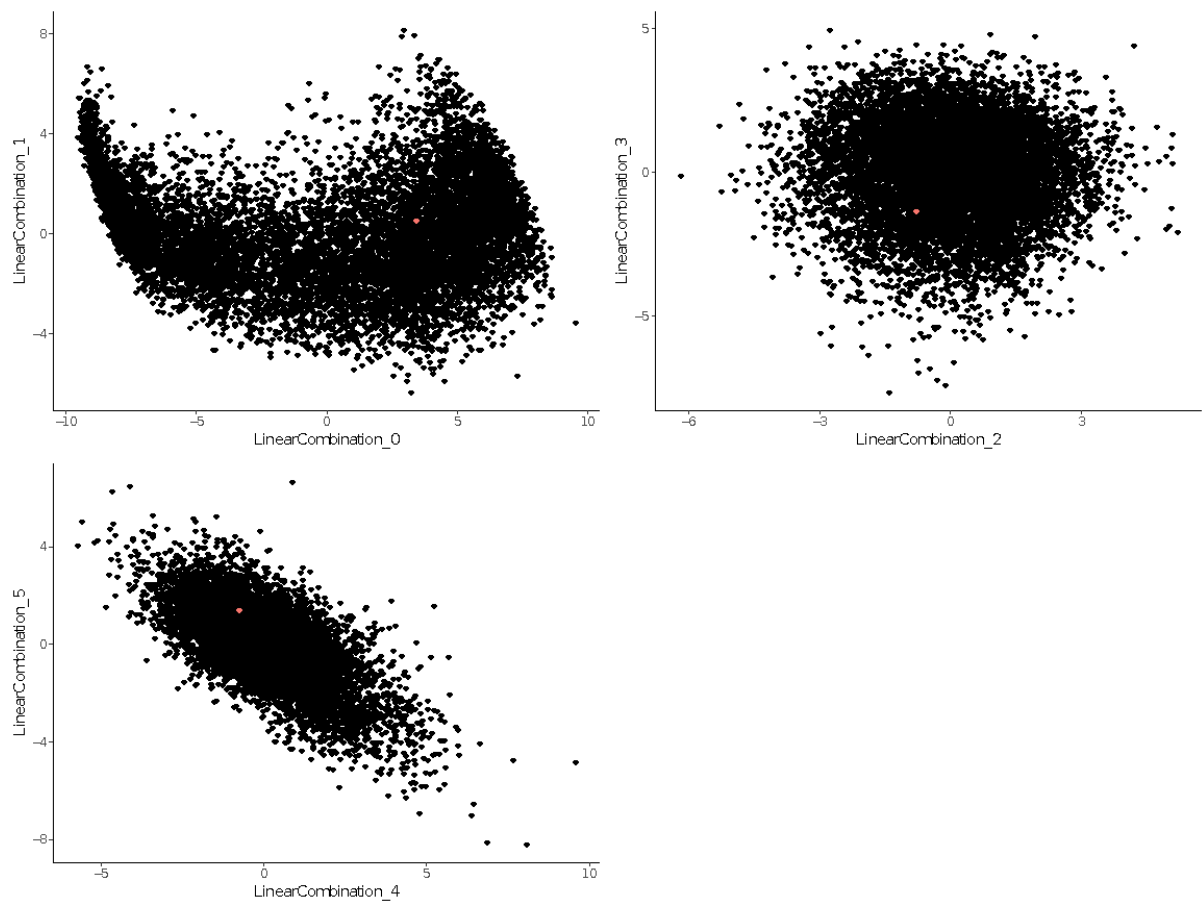


Figure 10. PLS distribution of 10,000 (out of 2,000,000) retained simulations and observed data under the full model. Simulations are shown in black, and observed data are shown in red.

European Sheep

In order to study the demographic dynamics of the European sheep from the Neolithic till the modern day, we used mitochondrial DNA sequences of both archaeological and modern samples collected from a large number of European countries. Our demographic modelling was focused on trying to detect how human management has affected the European sheep populations. The mean calendar year age of the ancient samples was converted in sheep generation times for the population demographic modelling purposes, assuming a generation step of 3.6 years (Leroy et al., 2013).

The data from the site was divided into twenty-one-time bins. For each time bin, we computed summary statistics: K (the number of segregating haplotypes), S (the number of segregating sites), Tajima's D , Fu's FS , and π (the average pairwise difference), as well as global and pairwise F -statistics. These summary statistics were used for the ABC analysis below.

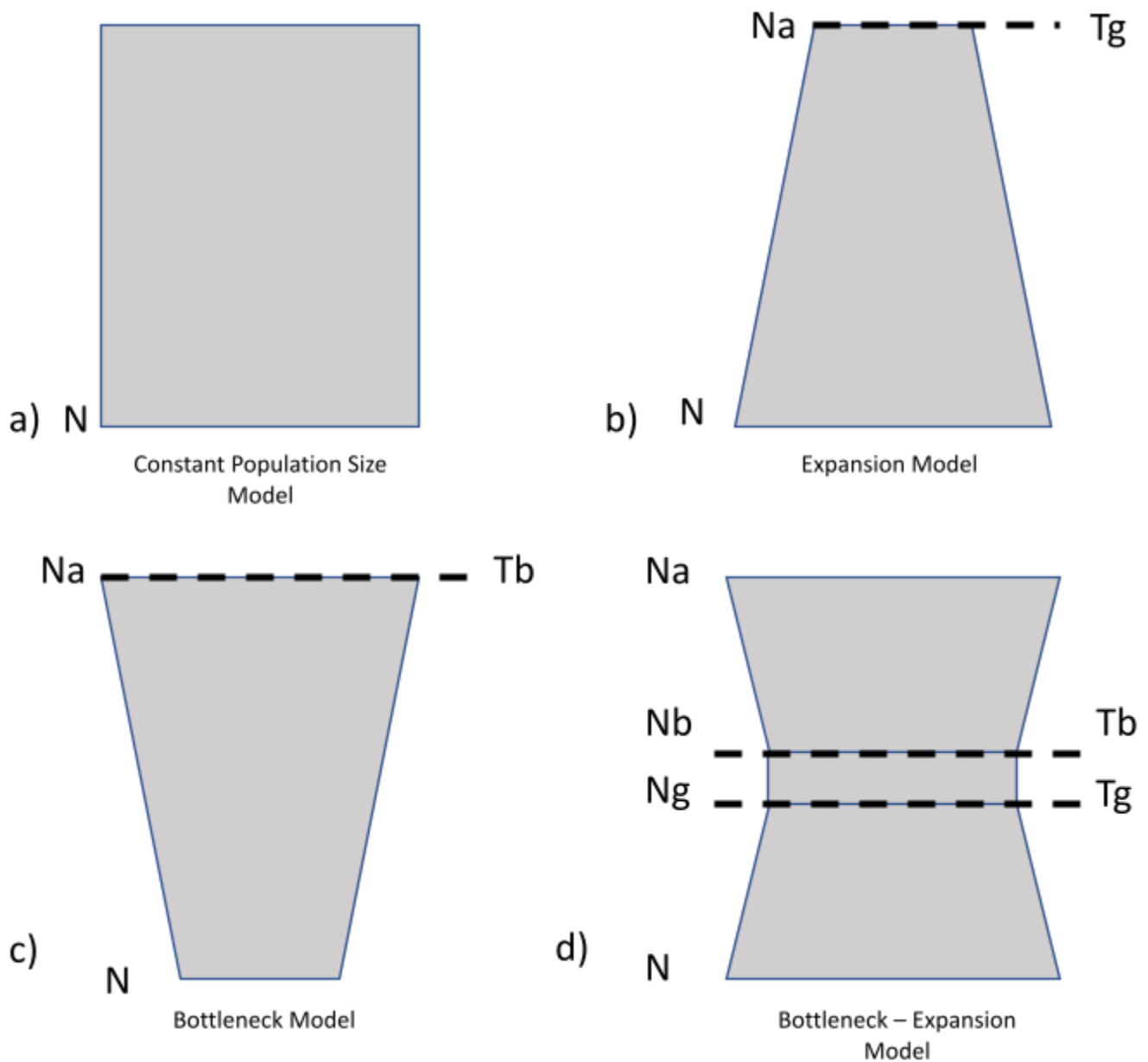


Figure 11: Demographic models of the European samples. 1a) constant population model. N : number of individuals. 1b) Expansion model. N_a : number of individuals before the expansion event. N : number of individuals at the time of sampling after the expansion. T_g : time of growth. 1c) Bottleneck model. N_a : number of individuals before the bottleneck event. N : number of individuals at the time of sampling after the expansion. T_b : time of bottleneck. 1d) Bottleneck-Expansion model. N_a : number of individuals before the bottleneck event. N_b : number of individuals at the time of

sampling after the bottleneck. Ng: number of individuals before the expansion. N: number of individuals at the time of sampling after the expansion. Tb: time of bottleneck. Tg: time of growth.

The calculated summary statistics for each time-bin/population were summarised in tables and subsequently plotted.

Table 7. Estimated values of K, S, π and Tajima's D per time bin/population, calculated from observed data.

POP NAME	K	S	Pi	D
0_GEN_AGO	194	829	39.5016	-2.3032
300_GEN_AGO	1	0	0	0
800_GEN_AGO	1	0	0	0
967_GEN_AGO	1	0	0	0
1000_GEN_AGO	15	258	62.0667	-0.965211
1150_GEN_AGO	2	23	23	0
1167_GEN_AGO	3	24	16	0
1400_GEN_AGO	1	0	0	0
1700_GEN_AGO	2	95	95	0
1800_GEN_AGO	1	0	0	0
1850_GEN_AGO	1	0	0	0
1900_GEN_AGO	2	16	16	0
2325_GEN_AGO	3	18	12	0
2940_GEN_AGO	5	29	12.4	-0.815789

3000_GEN_AGO	1	0	0	0
4400_GEN_AGO	2	1	1	0
4500_GEN_AGO	1	0	0	0
6160_GEN_AGO	1	0	0	0
7000_GEN_AGO	13	14	2.74359	-1.62597
7100_GEN_AGO	2	3	3	0
7125_GEN_AGO	2	13	13	0

We also summarised the population pairwise pi and F statistic

Table 8. Pairwise Pi statistics for every pairwise combination of time bins/populations.

	0	300	800	967	1000	1150	1167	1400	1700	1800	1850	1900	2325	2940	3000	4400	4500	6160	7000	7100	7125
	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN	GEN
	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO	AGO
0 GEN AGO		28.40	34.30	32.50	54.30	34.20	32.90	25.90	58.40	34.80	27.40	31.60	29.70	30.20	26.00	26.80	20.90	27.40	25.30	21.30	30.20
300 GEN AGO			7.00	7.00	47.53	16.50	12.67	9.00	50.50	8.00	11.00	14.00	9.33	11.60	11.00	9.50	6.00	10.00	6.85	6.00	12.50
800 GEN AGO				12.00	52.33	20.50	17.67	14.00	55.50	13.00	16.00	19.00	14.33	16.60	16.00	14.50	11.00	15.00	11.85	11.00	17.50
967 GEN AGO					50.80	20.50	15.00	14.00	53.50	13.00	14.00	19.00	14.33	16.40	15.00	14.50	8.00	15.00	11.69	10.00	17.50
1000 GEN AGO						50.10	49.67	42.00	60.33	51.93	43.20	48.10	46.56	46.89	43.80	42.90	39.27	43.80	42.38	36.20	46.90
1150 GEN AGO							16.83	13.50	57.50	22.50	14.50	19.50	15.17	17.30	15.50	14.00	12.00	14.50	12.31	10.00	17.50
1167 GEN AGO								11.67	56.17	18.67	13.67	17.00	11.33	15.60	13.67	12.17	10.33	12.67	10.92	7.83	15.83
1400 GEN AGO									50.50	15.00	4.00	9.00	7.00	7.80	6.00	4.50	5.00	5.00	4.00	3.50	8.50
1700 GEN AGO										55.00	51.50	56.00	53.17	54.30	52.50	52.00	44.50	52.50	50.12	43.50	55.25
1800 GEN AGO											17.00	20.00	15.33	17.60	17.00	15.50	10.00	16.00	12.85	11.50	18.50
1850 GEN AGO												11.00	9.67	10.20	8.00	6.50	7.00	7.00	5.85	5.50	10.50

Table 9. Pairwise F statistics for every pairwise combination of time bins/populations.

0	GEN 300	800	967	1000	1150	1167	1400	1700	1800	1850	1900	2325	2940	3000	4400	4500	6160	7000	7100	7125
AG	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO	GEN AGO
0 GEN AGO	-0.39	-0.15	-0.21	0.09	-0.03	0.03	-0.52	0.11	-0.14	-0.44	-0.05	-0.01	0.04	-0.52	-0.08	-0.89	-0.44	0.07	-0.30	-0.07
300 GEN AGO		1.00	1.00	-0.31	-0.39	-0.26	1.00	-0.88	1.00	1.00	-0.14	-0.29	-0.07	1.00	0.89	1.00	1.00	0.60	0.50	-0.04
800 GEN AGO			1.00	-0.19	-0.12	0.09	1.00	-0.71	1.00	1.00	0.16	0.16	0.25	1.00	0.93	1.00	1.00	0.77	0.73	0.26
967 GEN AGO				-0.22	-0.12	-0.07	1.00	-0.78	1.00	1.00	0.16	0.16	0.24	1.00	0.93	1.00	1.00	0.77	0.70	0.26
1000 GEN AGO					-0.03	0.07	-0.48	-0.20	-0.20	-0.44	-0.03	0.04	0.11	-0.42	-0.05	-0.58	-0.42	0.22	-0.21	-0.03
1150 GEN AGO						-0.11	-0.70	-0.03	-0.02	-0.59	0.00	-0.08	0.09	-0.48	0.14	-0.92	-0.59	0.48	-0.30	-0.03
1167 GEN AGO							-0.37	0.18	0.14	-0.17	0.06	-0.24	0.11	-0.17	0.15	-0.55	-0.26	0.42	-0.36	0.06
1400 GEN AGO								-0.88	1.00	1.00	-0.78	-0.71	-0.59	1.00	0.78	1.00	1.00	0.31	0.14	-0.53
1700 GEN AGO									-0.73	-0.84	0.01	0.18	0.35	-0.81	0.08	-1.13	-0.81	0.69	-0.13	0.02
1800 GEN AGO										1.00	0.20	0.22	0.30	1.00	0.94	1.00	1.00	0.79	0.74	0.30
1850 GEN AGO											-0.45	-0.24	-0.22	1.00	0.85	1.00	1.00	0.53	0.45	-0.24

The statistics from Table 6 were also plotted using R and ggplot2.

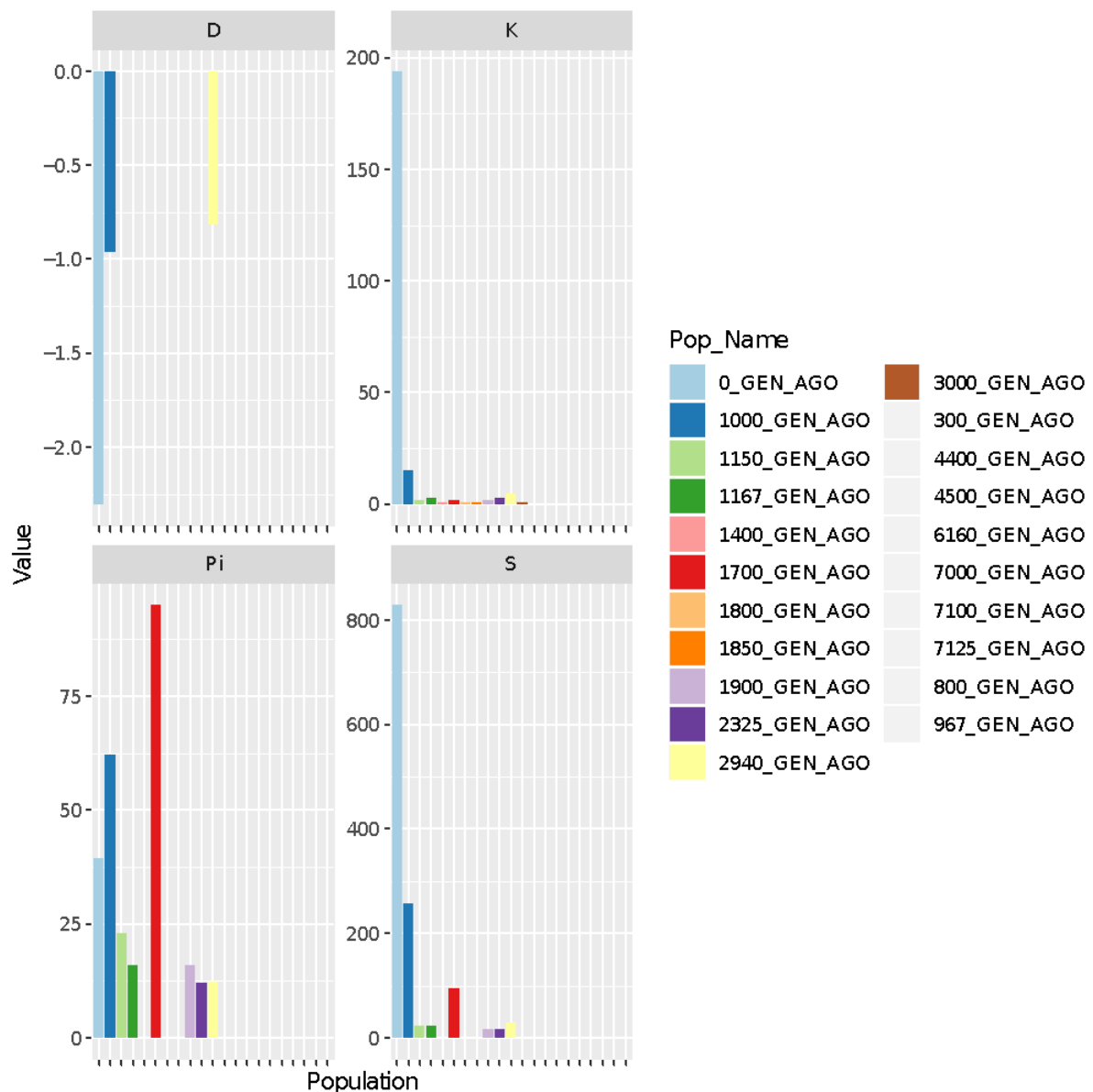


Figure 12. Plots of summary statistics presented in Table 6. These samples were used as the observed data in the ABC modelling population history models.

The increase of K and S, in the different populations, is an artefact of the fact that the populations don't have the same number of individuals, but instead the sample size varies as we go back in time. In 9 out of the 21 populations the sample size is only 1 individual. For that reason, summary statistics like the number of segregating site (S)s and haplotypes (K) cannot be calculated, making the calculation of more

complex statistics, like π and Tajima's D , impossible for these populations. Populations at 0, 1000 and 2940 generations ago show a negative Tajima's D indicating a (strong) population expansion (potentially after a bottleneck). Populations at 0, 1000 and 1700 generations ago show increased values of π as a result of an excess of segregating sites in these populations.

ABC analysis

We used fastsimcoal (Excoffier et al., 2013; Excoffier and Foll, 2011) to simulate data under three demographic models: a constant N_e model (one parameter: N_e), a bottleneck model (three parameters: N_e , T : time of bottleneck and N_{rel} : severity of the bottleneck), an expansion model (three parameters: N_e , T : time of expansion, and N_{rel} : severity of the expansion), and a fourth model of a bottleneck and followed by an expansion (four parameters N_e , T_g : time of expansion, T_b : time of bottleneck, N_{rel_g} : severity of the expansion and N_{rel_b} : severity of the bottleneck). Detailed parameter values were drawn from uniform distributions presented in Table 9, time units were in sheep generation steps and the population times were calibrated in such a way so that the modern sheep populations had an age of 0 generation steps. Historical samples were simulated as suggested in the fastsimcoal manual. For model testing purposes, we performed 100,000 simulations, for each model respectively, for which we calculated the same summary statistics that were calculated from our observed data of the European sheep populations.

Table 10. Summary of parameters for each demographic model and the distribution characteristics from which each parameter was drawn during the ABC simulations.

	<i>Parameter</i>	<i>Prior</i>	<i>Prior_min</i>	<i>Prior_max</i>
<i>Constant NE</i>	N	Log Uniform	3	5.7
	Mutrate	Uniform	1.00E-09	0.000001
<i>Bottleneck</i>	N	Log Uniform	3	6
	Nanc	Log Uniform	3	6
	Mutrate	Uniform	1.00E-09	0.000001
	Tbot	Uniform	0	10000
<i>Expansion</i>	N	Log Uniform	3	6
	Nanc	Log Uniform	1	3
	Mutrate	Uniform	1.00E-09	0.000001
	Texp	Uniform	0	7200
<i>Bottleneck - Expansion</i>	N	Log Uniform	3	6
	Nbot	Log Uniform	3	6
	Nexp	Log Uniform	1	3
	Mutrate	Uniform	1.00E-09	0.000001
	Tbot	Uniform	0.00E+00	10000
	Texp	Uniform	0	7200

We chose a set of informative summary statistics to apply a partial least-squares discriminant analysis on (Frantz et al., 2015; Peter et al., 2012) using the plsda function in R (Lê Cao et al., 2009). After retaining a total of 10,000 simulations closest to the observed data, we compared all models (computing marginal

likelihood and posterior probability) simultaneously using a standard ABC generalized linear model (GLM) approach as implemented in ABCtoolbox (Wegmann et al., 2010). We also performed two rounds of model fitting, one with 4 models, which included both an expansion and a combination of a bottleneck and expansion, and a second round that included only a simple expansion scenario. The best fitting model in both cases was the expansion model.

We also computed the average Root Mean Square Error (RMSE) for each parameter using pseudo-observed data to assess our power to infer each parameter in the model.

Table 11. Model fitting on the observed data with 4 models. The expansion model fits the observed data the best.

	<i>obsPvalue</i>	<i>Marginal Density</i>	<i>Bayes Factor</i>	<i>Posterior Probability</i>	<i>Chosen Model</i>
<i>Constant NE</i>	0	3.09E-48	5.12E-42	5.12E-42	
<i>Bottleneck</i>	0	2.06E-46	3.43E-40	3.43E-40	
<i>Expansion</i>	0.01	3.11E-07	1.06811	0.516466	TRUE
<i>Bottleneck - Expansion</i>	0.011	2.91E-07	0.936236	0.483534	

Table 12. Model fitting on the observed data with 3 models. The expansion model fits the observed data the best.

	<i>obsPvalue</i>	<i>Marginal Density</i>	<i>Bayes Factor</i>	<i>Posterior Probability</i>	<i>Chosen Model</i>
<i>Constant NE</i>	0	3.09E-48	9.92E-42	9.92E-42	
<i>Bottleneck</i>	0	2.06E-46	6.63E-40	6.63E-40	
<i>Expansion</i>	0.01	3.11E-07	inf	1	TRUE

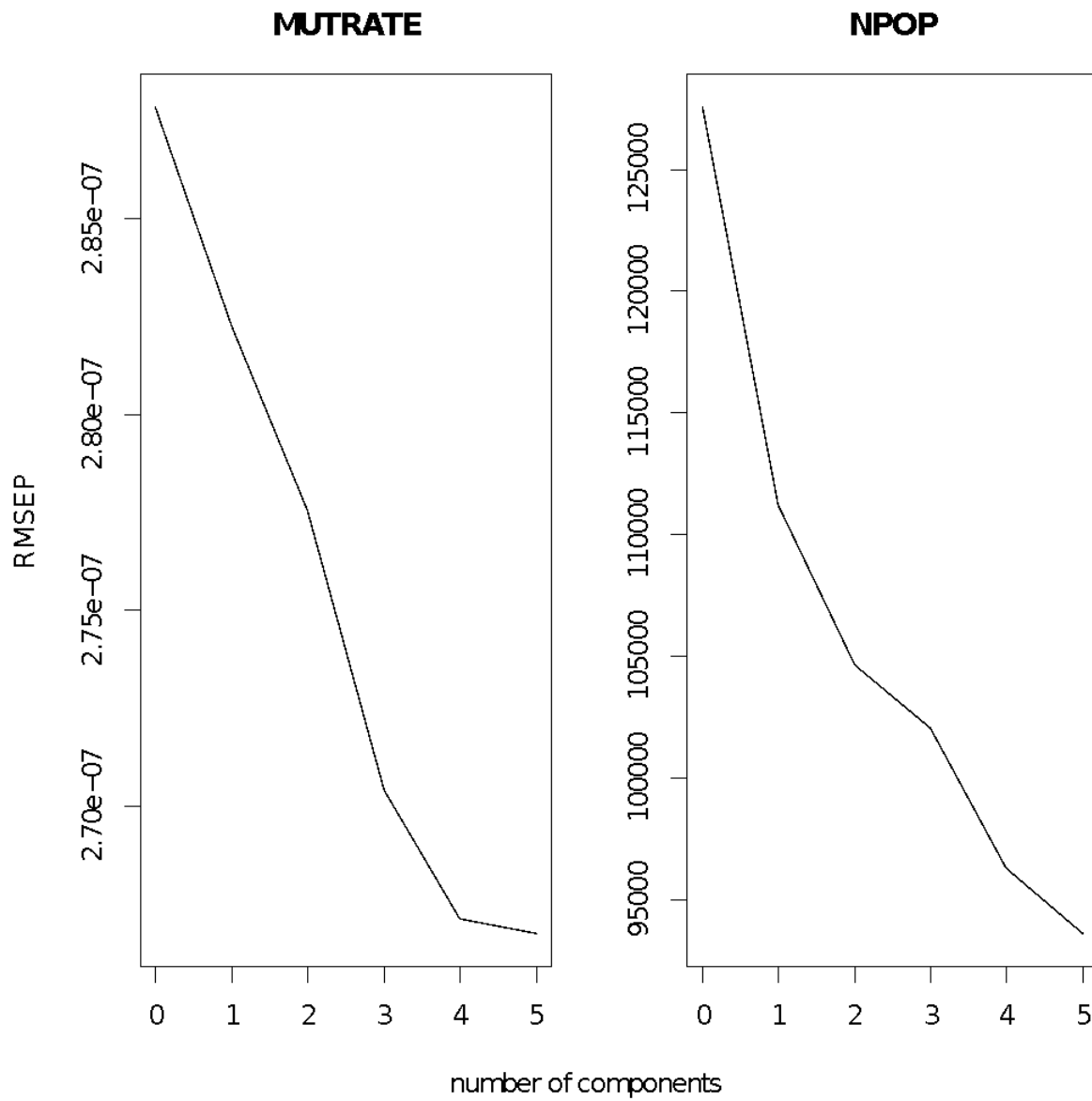


Figure 13. Combined RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size from the 4 models combined.

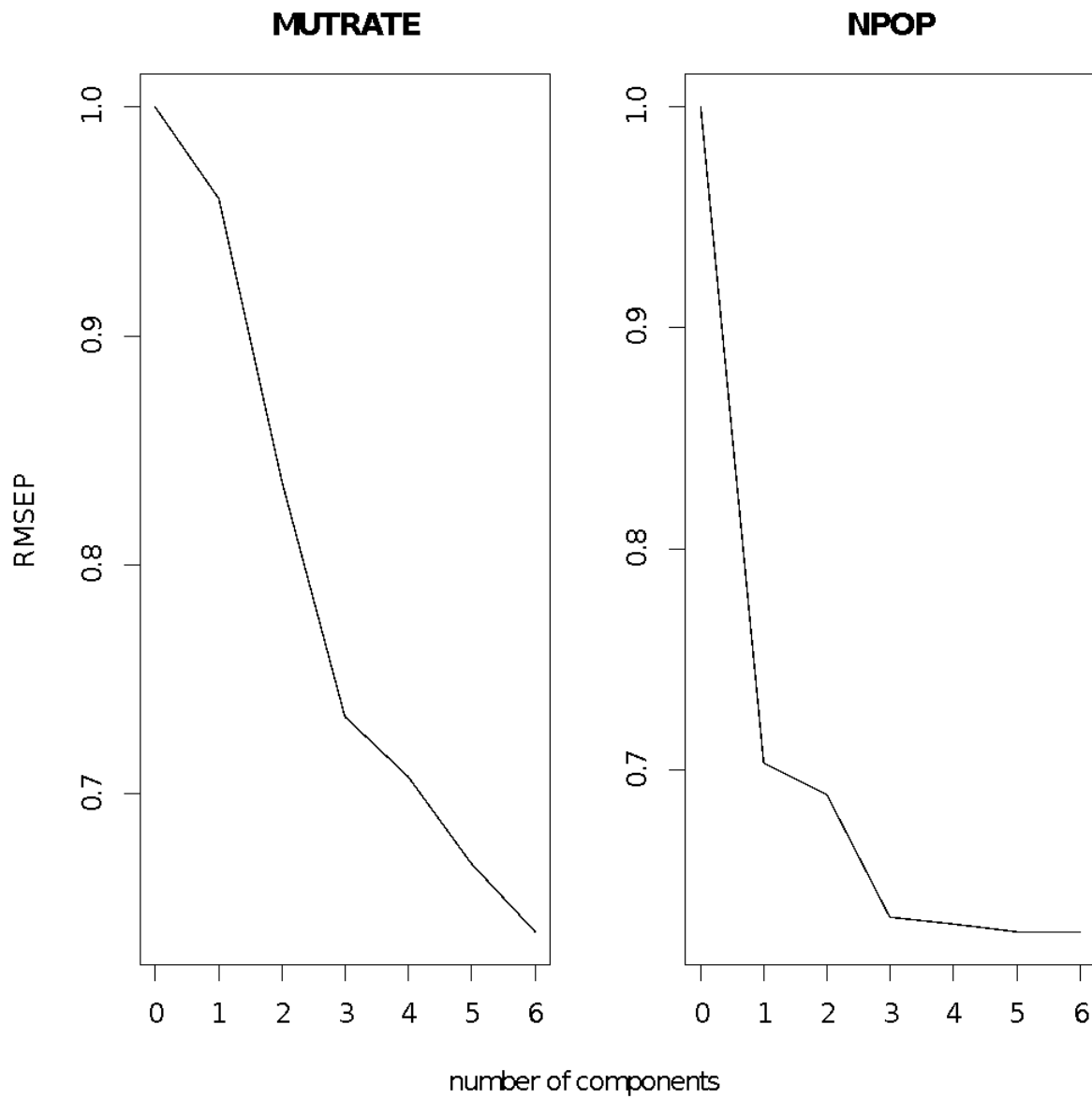


Figure 14. Constant NE RMSE. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size.

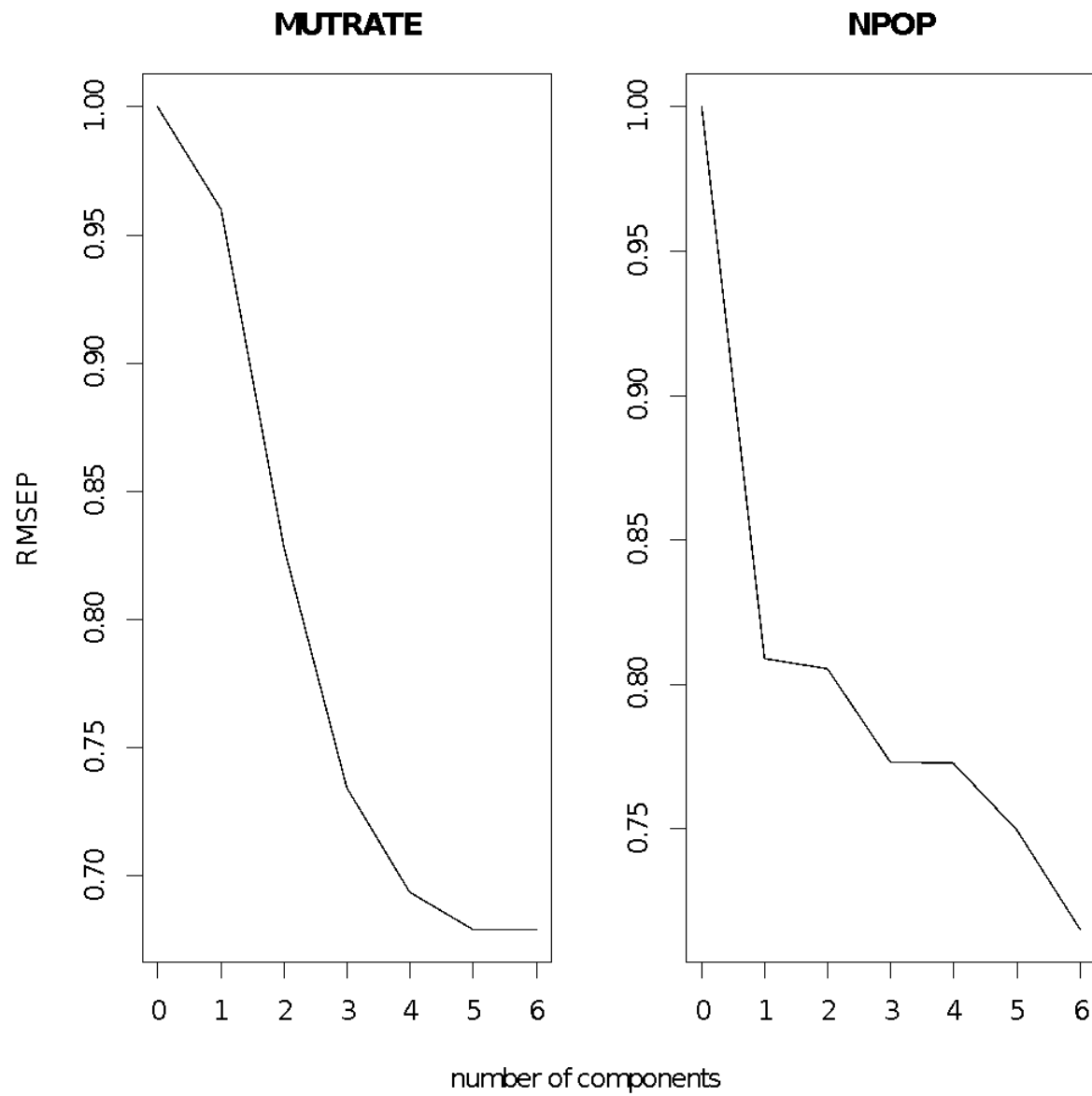


Figure 15. Bottleneck RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size.

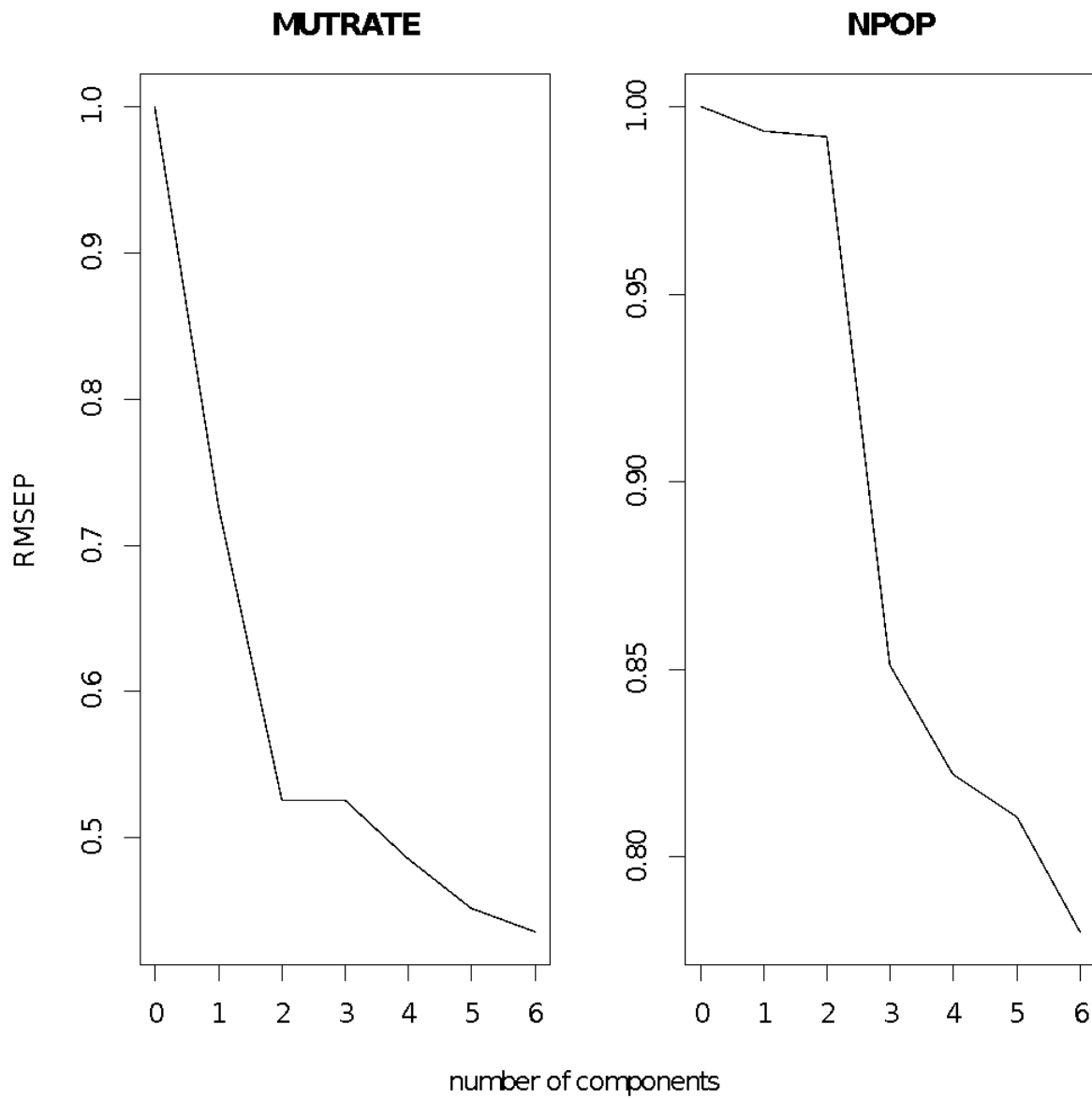


Figure 16. Expansion RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size.

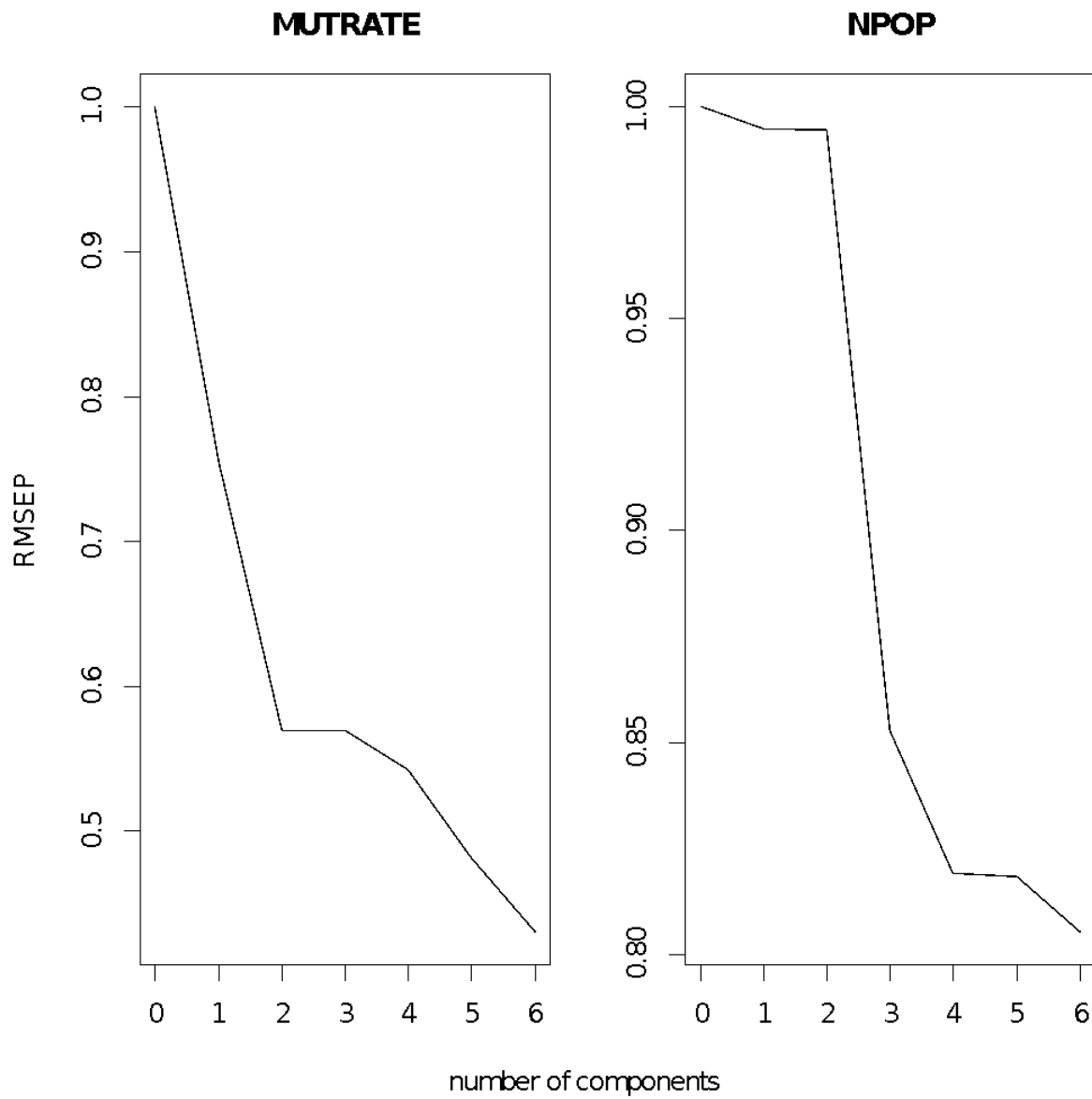


Figure 17. Bottleneck – Expansion EMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size.

Best Fitting model – Expansion

After identifying the expansion model as the one that fits our observed data the best, model parameter estimation was performed. We ran a total of 2 million simulations under the expansion model, and for each simulation summary statistics were calculated. These summary statistics were subsequently used as an input for a PLS analysis. The computed PLS were used to transform the summary stats to find the 5 best linear combinations.

The 5 best linear combinations were used as an input for the ABCtoolbox estimator in order to calculate the posterior distributions of N_e before and after the expansion event, the time of expansion, the magnitude of the expansion event and mutation rate parameters. For this purpose, we retained a total of 10,000 simulations closest to the observed data and applied a standard ABC-GLM. The characteristics of the posterior distribution for each parameter of the expansion model are summarised in the following table.

Table 13. The characteristics of marginal posterior distributions of the parameters of the Expansion model after 2 million simulations.

<i>Parameter</i>	<i>Prior</i>	<i>Prior min</i>	<i>Prior max</i>	<i>RMSE</i>	<i>Mode</i>	<i>HPDI 95 lower</i>	<i>HPDI 95 upper</i>
<i>N Now</i>	Log Uniform	3	6		4.90	3.77	5.72

<i>N Exp</i>	Log Uniform	1	3		2.23	1.17	2.94
<i>Time of Expansion</i>	Uniform	0	7200		1964.36	397.705	3527.78
<i>Mutrate</i>	Uniform	1E-09	0.0000 01		3.84E-07	2.38E-07	5.49E-07

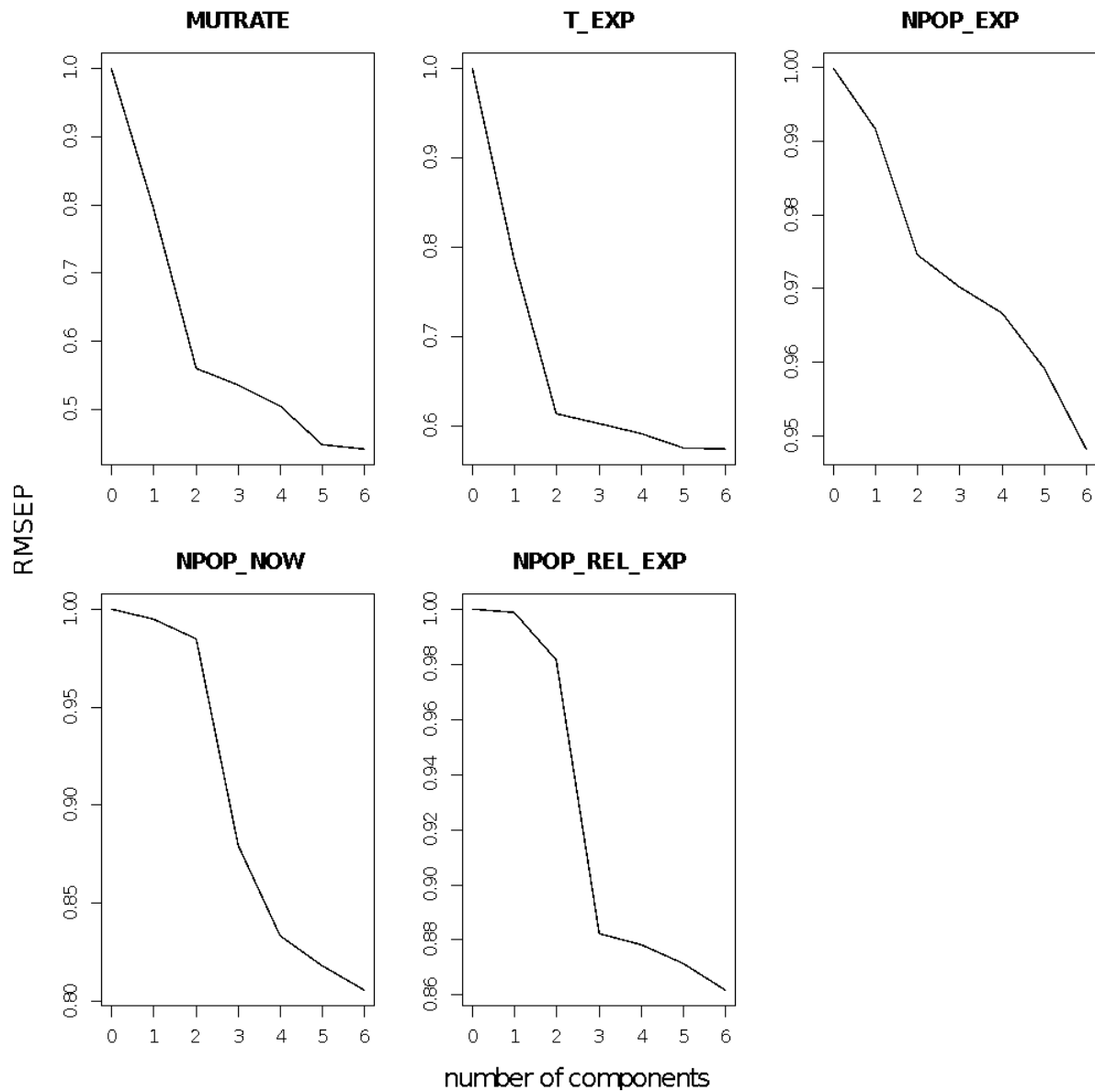


Figure 18. Percentage of the variation observed in the simulated data, explained by the mutation rate, effective population of the European sheep before and after the expansion event, the magnitude and time of the expansion event.

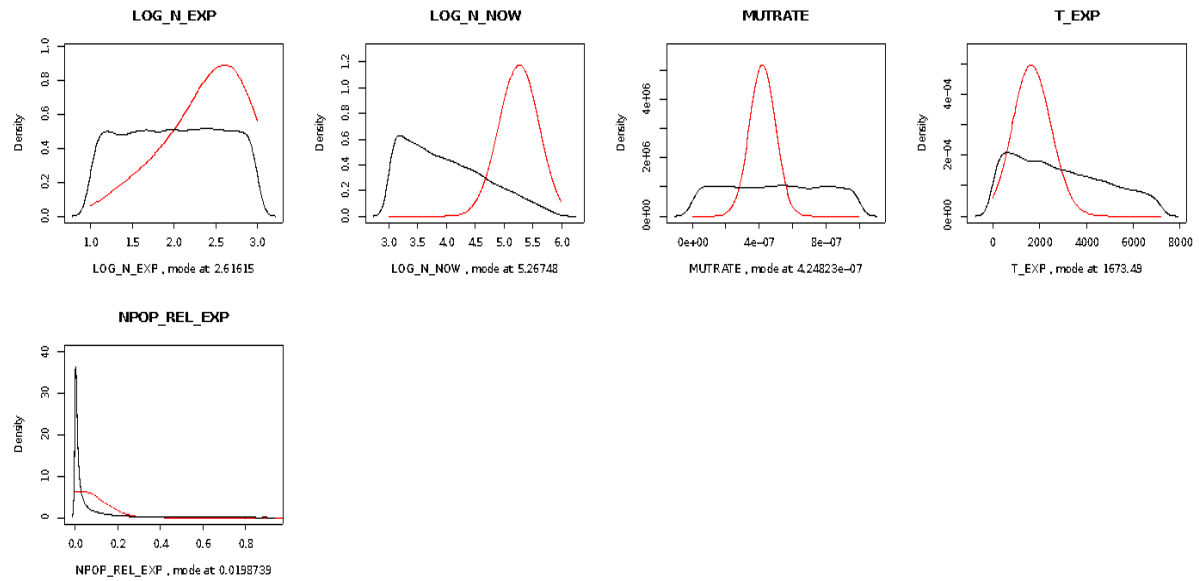
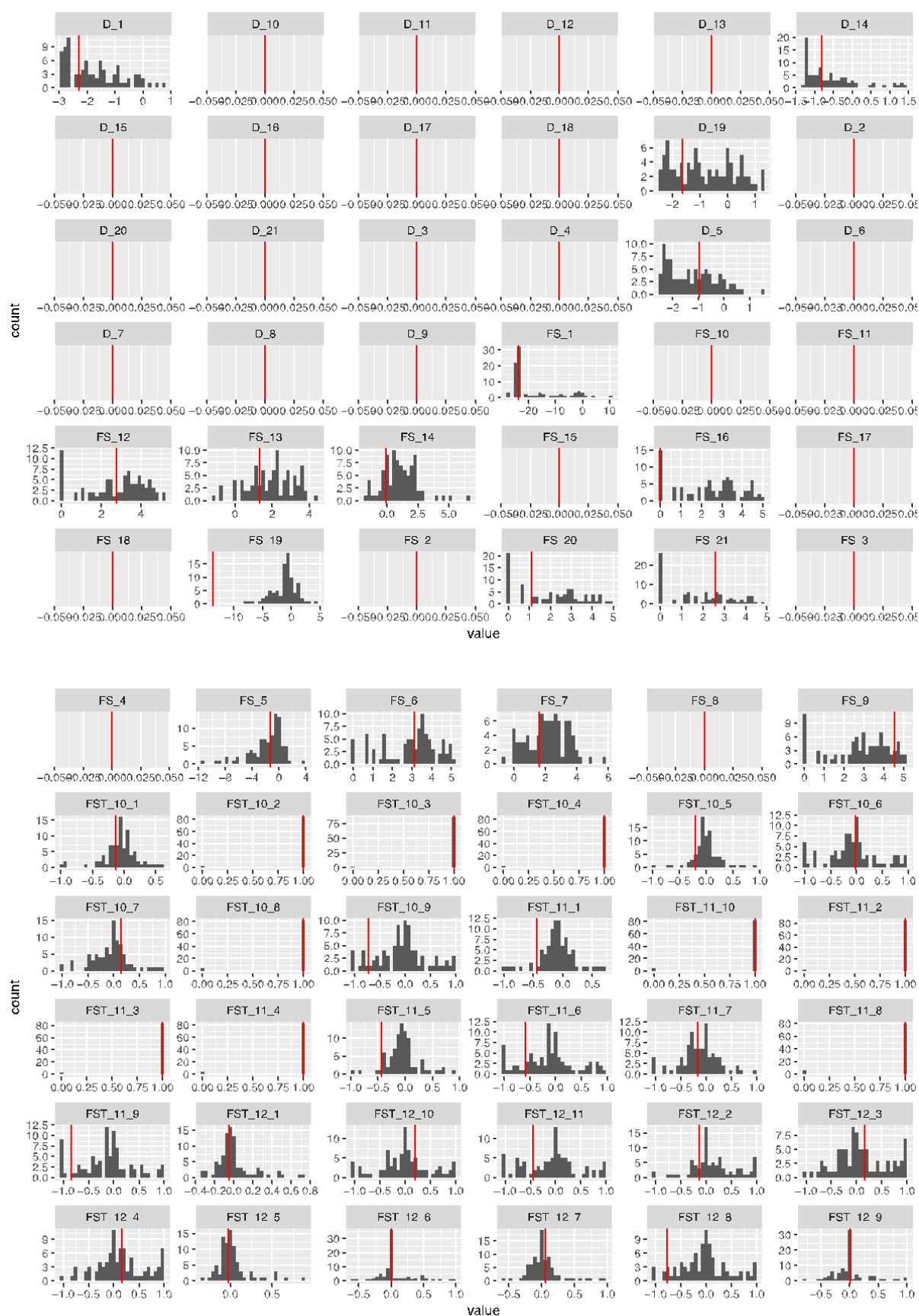
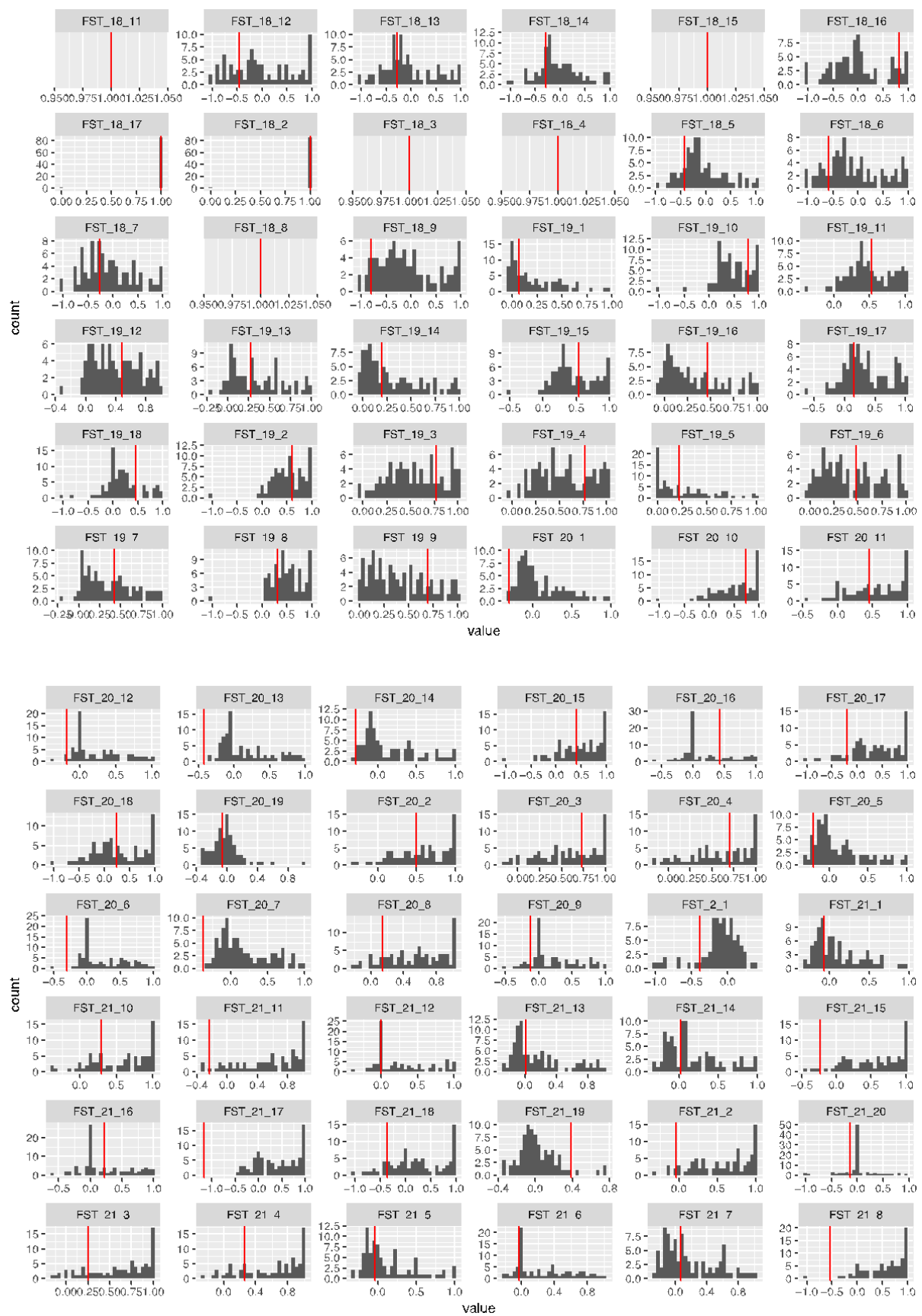
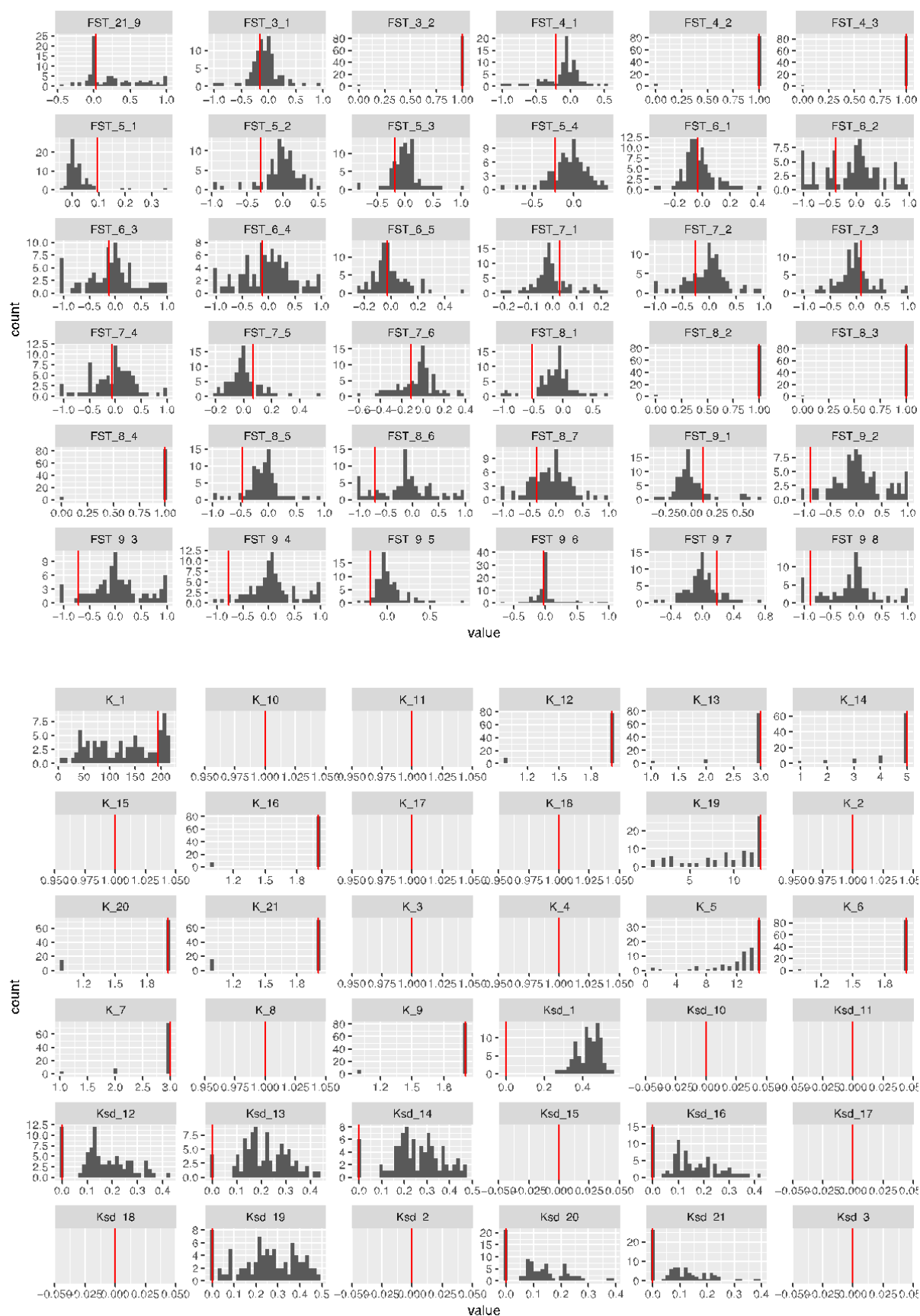


Figure 19. Plots of the uniform posterior distributions of the effective population size before and after the expansion of the European sheep populations, the mutation rate and magnitude and time of the expansion event. Blue lines represent the simulated values that have been retained, and the black lines represent the prior distributions of these values.

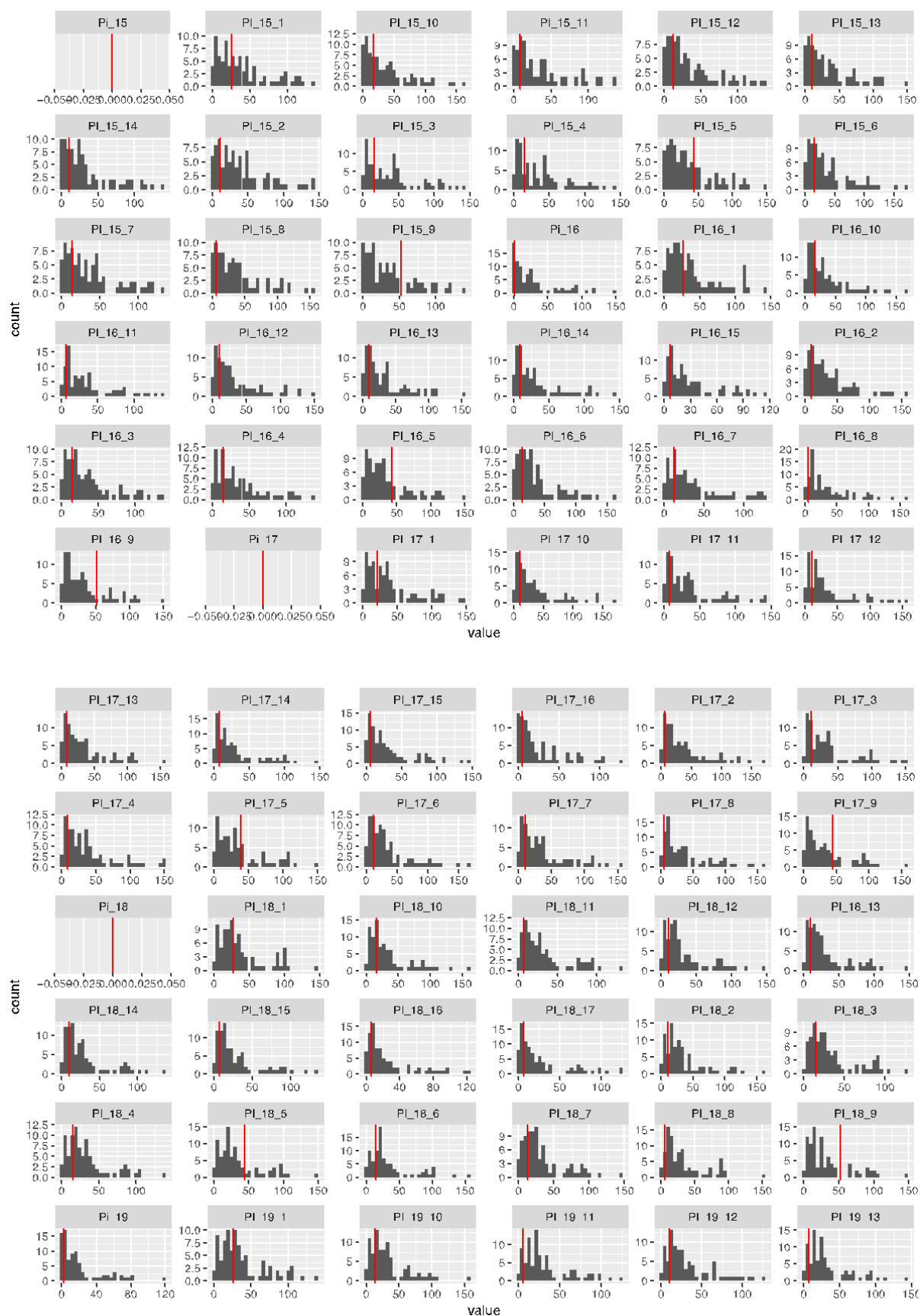


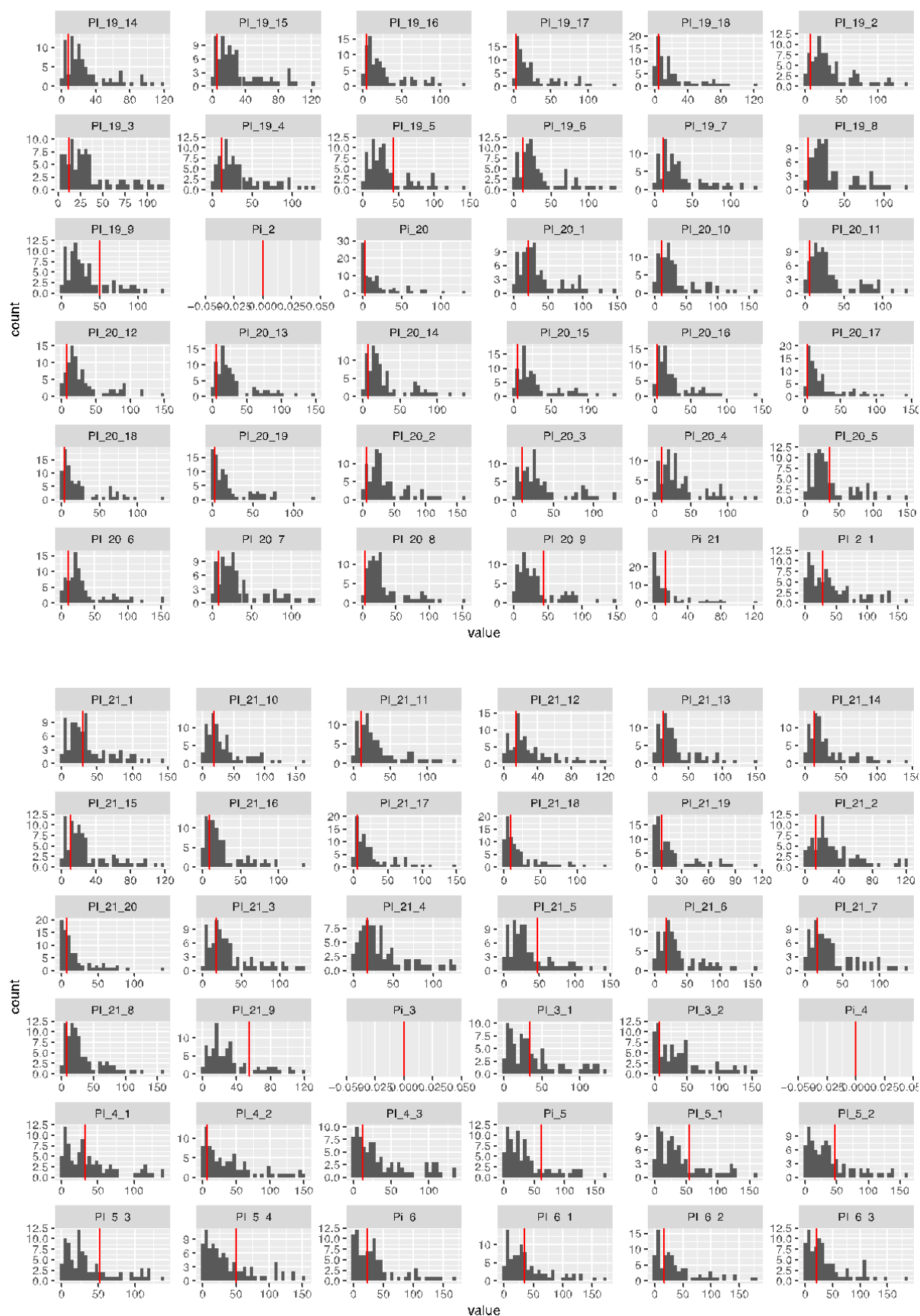


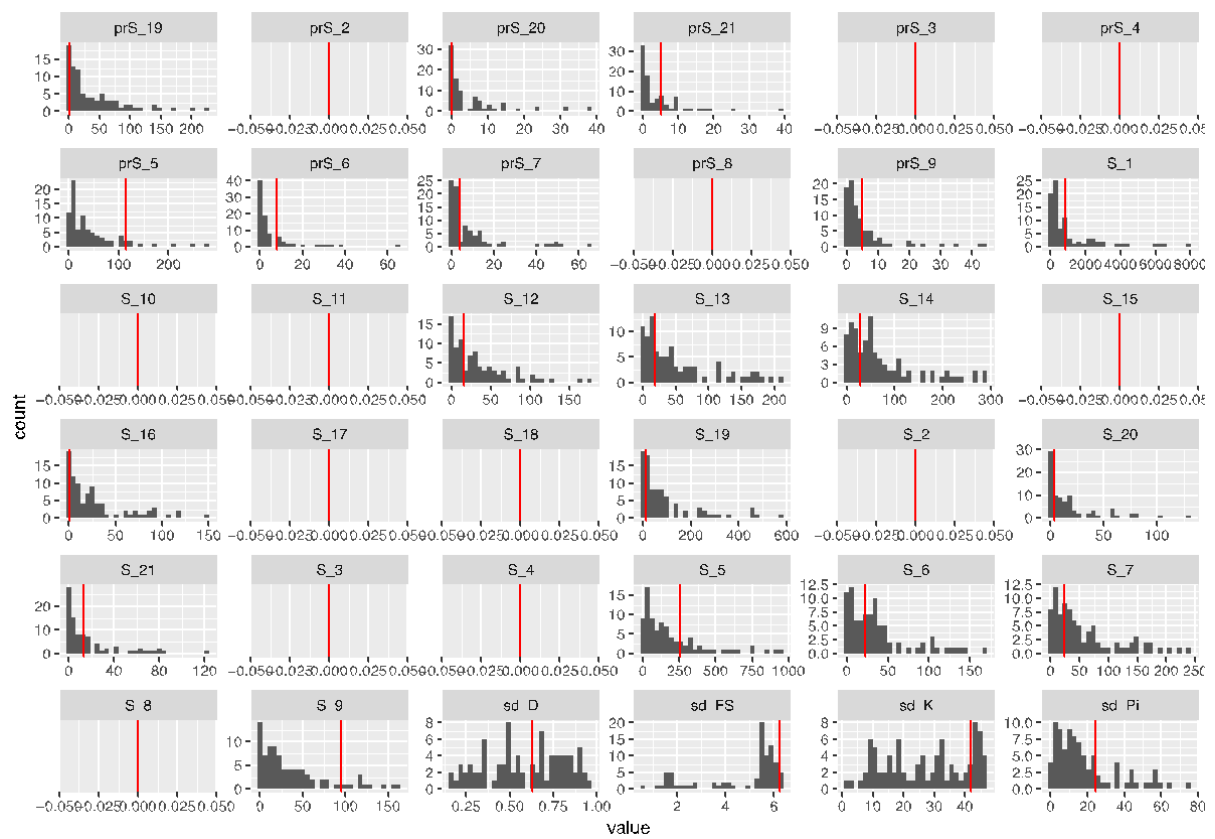












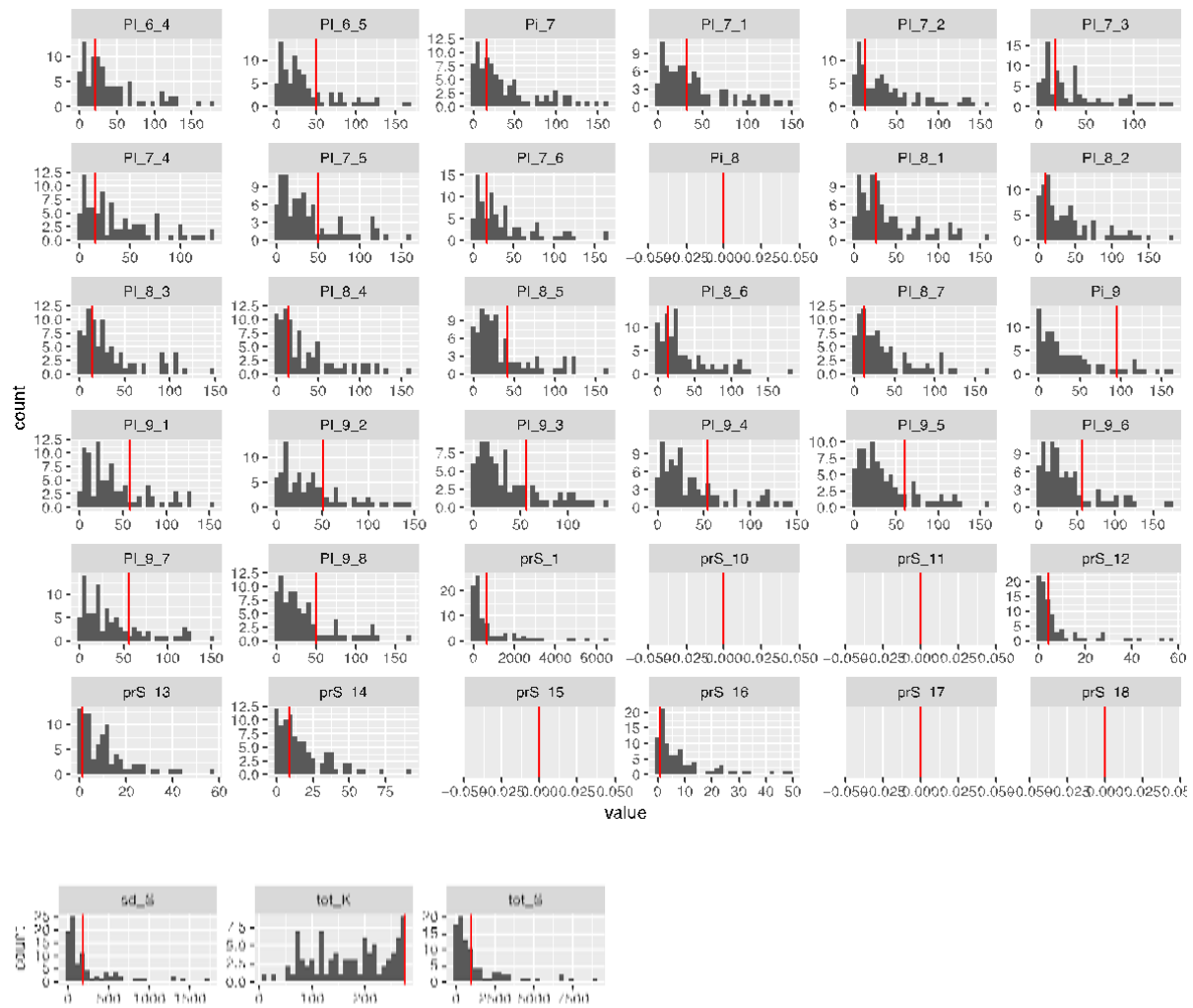


Figure 20. Histogram distribution of the summary statistics of the simulated data. The red bars represent the observed values of the chosen summary statistics.

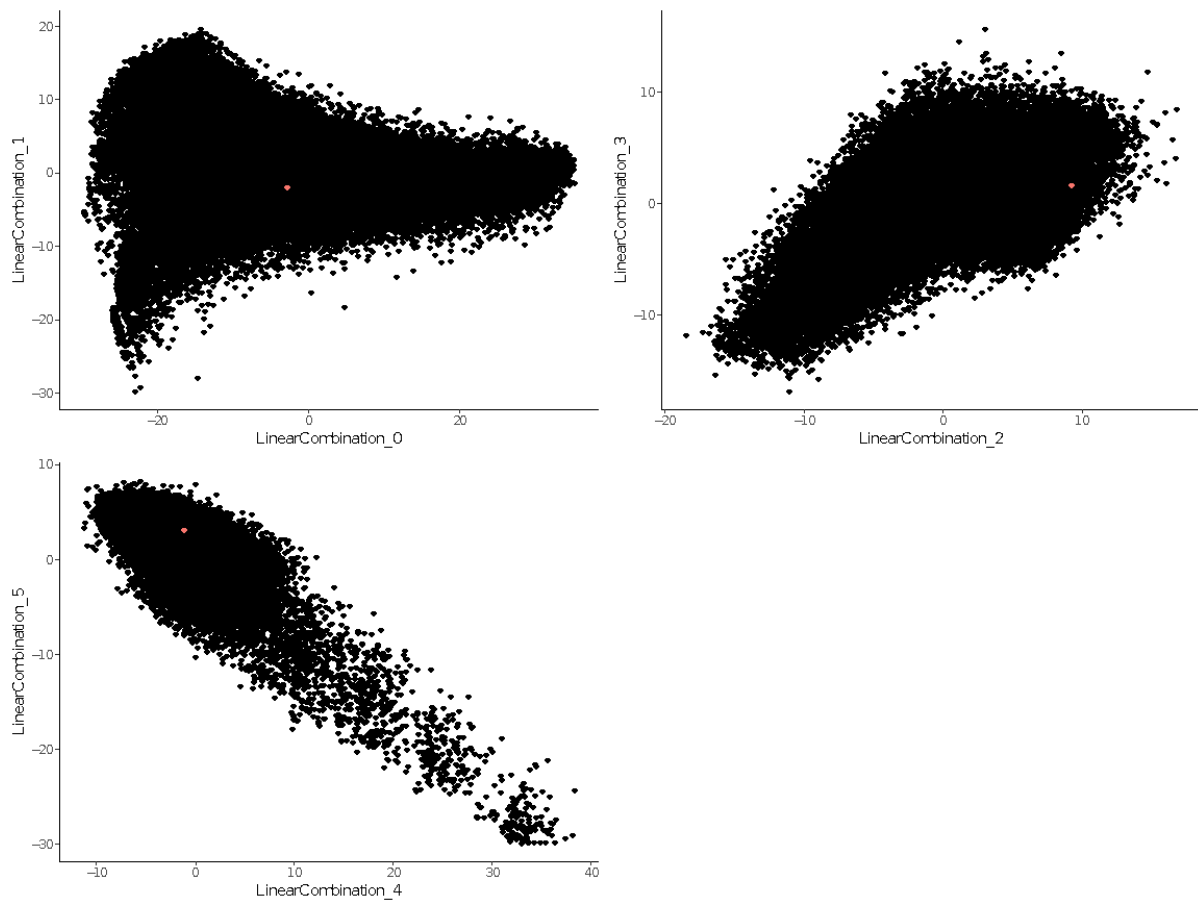


Figure 21. PLS distribution of 10,000 (out of 2,000,000) retained simulations and observed data under the full model. Simulations are shown in black, and observed data are shown in red.

Asikli site and ancient Turkish sheep

In order to study the demographic dynamics of the sheep populations in Asikli and ancient Turkey, we once again used mitochondrial DNA sequences from ancient sheep samples from Asikli, but we also included sequences from Neolithic sheep from various other archaeological Turkish sites. Our demographic modelling focused on the split between the lineages of the Asikli and Ancient Turkish sheep populations, also trying to detect whether a genetic bottleneck happened in the Ancient Turkish sheep population following the split. Testing these split models could also allow us to establish whether there has been continuity between the sheep populations in Asikli and the ancient Turkish ones. The mean calendar year age of the samples was converted in sheep generation times for population demographic modelling purposes, assuming a generation step of 3.6 years (Leroy et al., 2013).

The ancient data from all the sites were divided into 9-time bins/populations, 5 time-bins for the Ancient Turkish sheep populations and 4 time-bins for the ancient Asikli populations. For each time bin, we computed summary statistics: K (the number of segregating haplotypes), S (the number of segregating sites), Tajima's D , Fu's FS , and π (the average pairwise difference), as well as global and pairwise F -statistics. These summary statistics were used for the ABC analysis below.

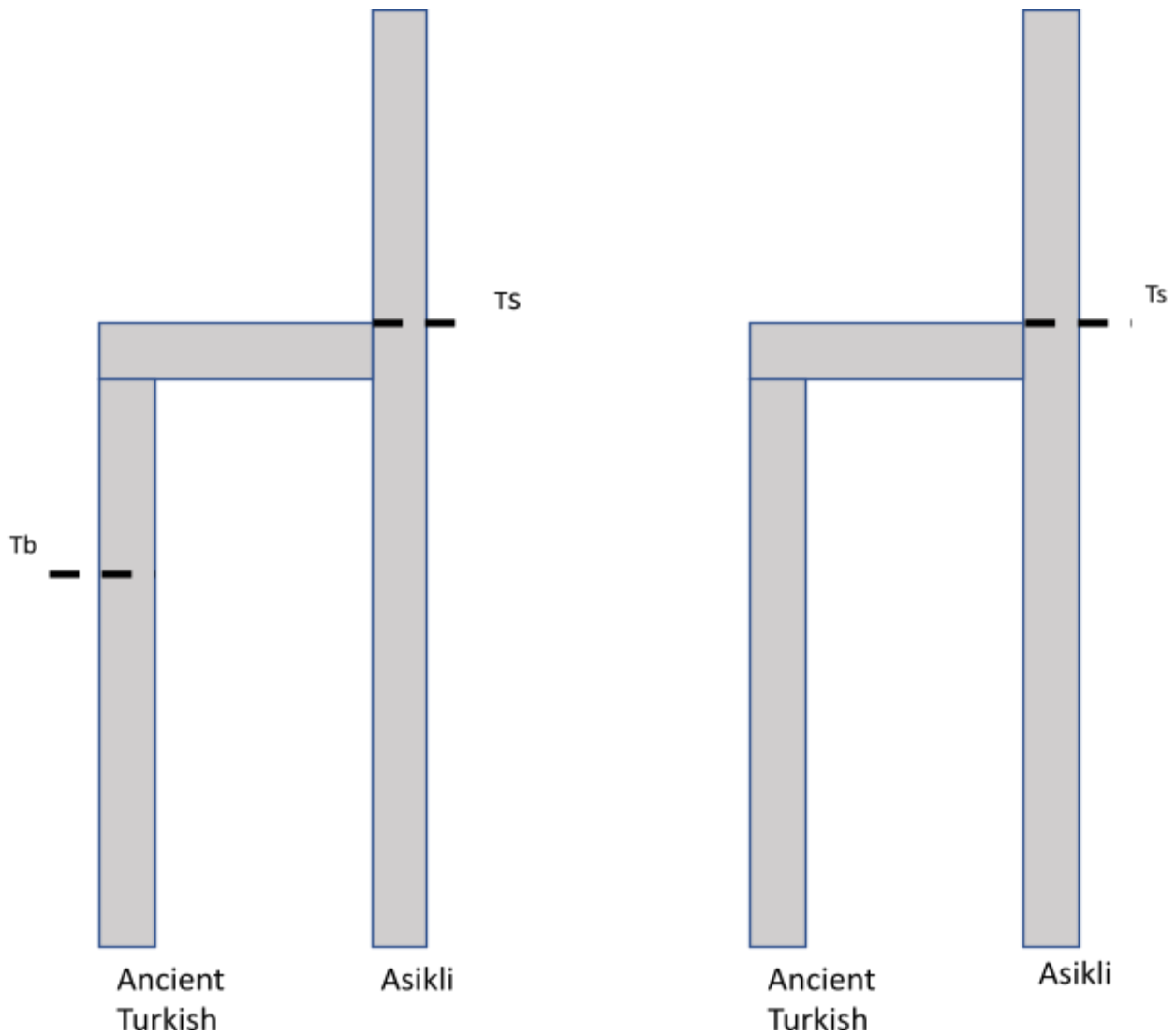


Figure 22: Demographic models of the ancient Asikli and ancient Turkish samples. 1a) Split model between the Asikli and ancient Turkish lineages. T_s : Time of the split. T_b : Time of the bottleneck. 1b)) Split model between the Asikli and ancient Turkish lineages. T_s : Time of the split. The calculated summary statistics for each time-bin/population were summarised in tables and subsequently plotted.

Table 14. Estimated values of K, S, π and Tajima's D per time bin and population, calculated from observed data.

Pop_Name	K	S	Pi	D
Turkish_7000_GEN_AGO	5	110	45.7	-1.0802
Turkish_7062_GEN_AGO	5	8	3.4	-0.807344
Turkish_7159_GEN_AGO	7	22	8.21429	-0.166822
Turkish_7200_GEN_AGO	2	5	5	0
Turkish_7247_GEN_AGO	2	4	4	0
Asikli_9425_GEN_AGO	11	258	89.5455	0.0242488
Asikli_9625_GEN_AGO	12	281	76.9429	-0.576253
Asikli_9850_GEN_AGO	11	288	92.4615	-0.112519
Asikli_10150_GEN_AGO	19	324	92.3714	0.0559447

We also summarised the population pairwise pi and F statistics.

Table 15. Pairwise Pi statistics for every pairwise combination of time bins/populations.

	Turkish 7000 GEN AGO	Turkish 7062 GEN AGO	Turkish 7159 GEN AGO	Turkish 7200 GEN AGO	Turkish 7247 GEN AGO	Asikli 9425 GEN AGO	Asikli 9625 GEN AGO	Asikli 9850 GEN AGO	Asikli 10150 GEN AGO
Turkish 7000 GEN AGO		27.68	30.33	28.30	27.30	86.58	63.85	79.69	90.58
Turkish 7062 GEN AGO			6.83	3.90	3.20	80.24	54.76	79.03	88.89
Turkish 7159 GEN AGO				7.50	6.69	82.58	57.73	81.34	91.02
Turkish 7200 GEN AGO					4.00	81.14	55.63	79.96	89.93
Turkish 7247 GEN AGO						80.55	54.83	79.27	89.31
Asikli 9425 GEN AGO							83.83	90.59	87.91
Asikli 9625 GEN AGO								84.23	87.97
Asikli 9850 GEN AGO									89.60
Asikli 10150 GEN AGO									

Table 16. Pairwise F statistics for every pairwise combination of time bins/populations.

	Turkish 7000 GEN AGO	Turkish 7062 GEN AGO	Turkish 7159 GEN AGO	Turkish 7200 GEN AGO	Turkish 7247 GEN AGO	Asikli 9425 GEN AGO	Asikli 9625 GEN AGO	Asikli 9850 GEN AGO	Asikli 10150 GEN AGO
Turkish 7000 GEN AGO		0.11	0.18	-0.13	-0.16	0.18	0.00	0.09	0.17
Turkish 7062 GEN AGO			0.11	0.00	-0.13	0.31	0.13	0.26	0.30
Turkish 7159 GEN AGO				0.03	-0.04	0.37	0.19	0.32	0.35
Turkish 7200 GEN AGO					-0.13	0.15	-0.03	0.12	0.19
Turkish 7247 GEN AGO						0.15	-0.04	0.11	0.19
Asikli 9425 GEN AGO							0.01	0.00	-0.04
Asikli 9625 GEN AGO								0.00	0.04
Asikli 9850 GEN AGO									-0.03
Asikli 10150 GEN AGO									

The statistics from Table 6 were also plotted using R and ggplot2.

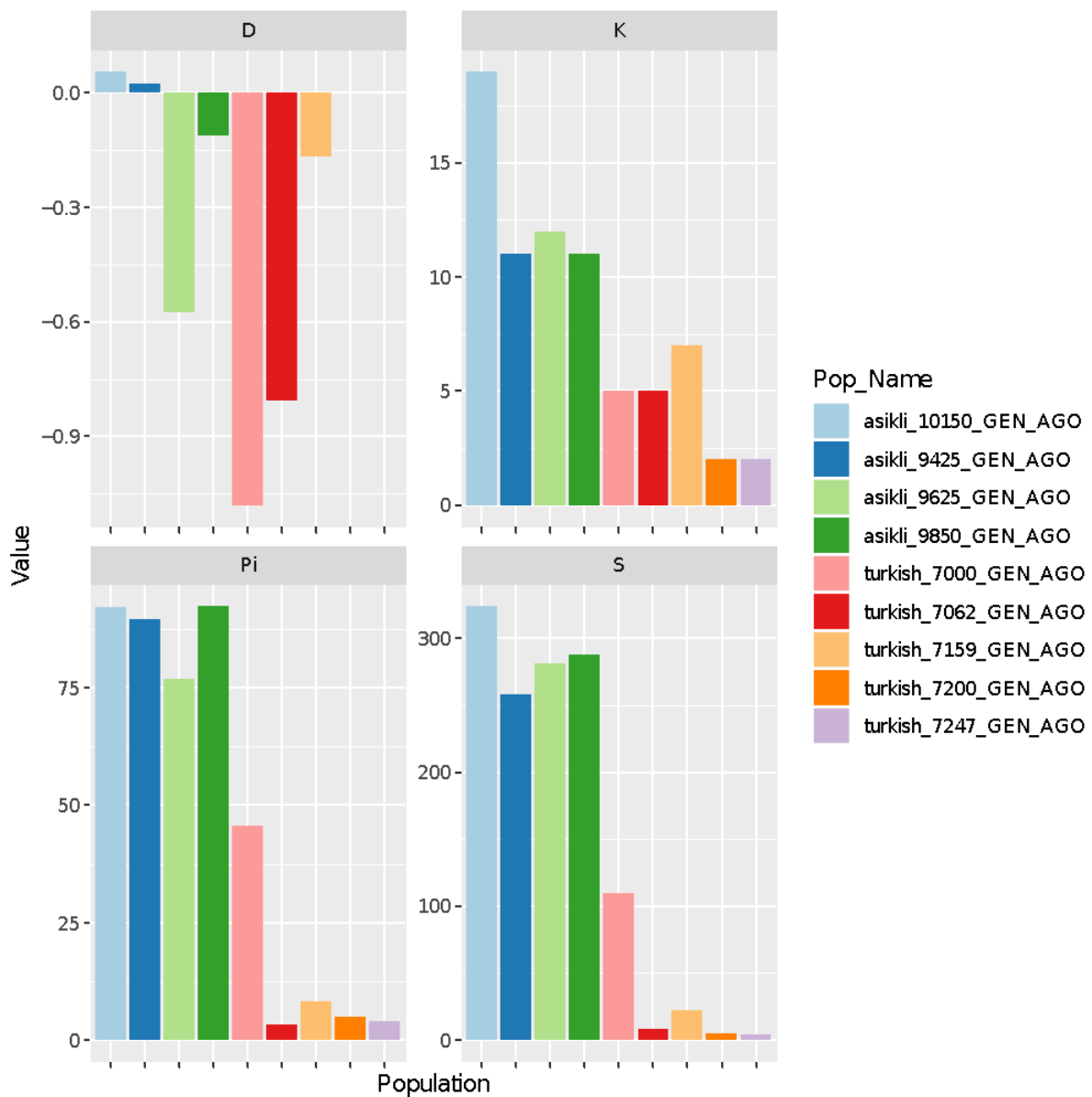


Figure 23. Plots of summary statistics presented in Table 1. These samples were used as the observed data in the ABC modelling population history models that followed.

The increase of K, S and Pi, in the different Asikli populations, is an artefact of the fact that these populations almost have double the number of individuals compared to the ancient Turkish ones. Once again though in both the Asikli and the Turkish populations the sample size varies as we go back in time. Asikli populations at 9625

and 9850 generations ago and Turkish populations at 7000 7062 7159 generations ago show a negative Tajima's D indicating a (strong) population expansion (potentially after a bottleneck). All the Asikli populations along with the Turkish population at 7000 generations ago show increased values of nucleotide diversity (π) compared to the other ancient Turkish populations.

ABC analysis

We used fastsimcoal (Excoffier et al., 2013; Excoffier and Foll, 2011) to simulate data under 2 demographic models: 1) a split model between the ancient Asikli and the Ancient Turkish sheep populations (4 parameters: N Asikli, N Turkish, Mutation rate and Time of the split), 2) a split model between the ancient Asikli and the Ancient Turkish sheep populations, followed by a bottleneck in the Ancient Turkish lineage (5 parameters: N Asikli, N Turkish, Mutation rate, Time of the split and Time of the Bottleneck). Detailed parameter values were drawn from uniform distributions presented in Table 4, time units were in sheep generation steps and the population times were calibrated in such a way so that the modern sheep populations had an age of 0 generation steps. The time of the split was allowed to happen while the ancient Asikli lineage was still active, and the time of the bottleneck was allowed to be anywhere in-between the first and the last sample from the ancient Turkish lineage.

Historical samples were simulated as suggested in the fastsimcoal manual. For model testing purposes, we performed 100,000 simulations, for each model respectively, for

which we calculated the same summary statistics that were calculated from our observed data of the European sheep populations.

Table 17. Summary of parameters for each demographic model and the distribution characteristics from which each parameter was drawn during the ABC simulations.

	<i>Parameter</i>	<i>Prior</i>	<i>Prior_min</i>	<i>Prior_max</i>
<i>Split</i>	N Asikli	Log	3	5.7
		Uniform		
	N Turkish	Log	3	5.7
		Uniform		
	MUTRAT E	Uniform	1E-09	0.000001
	T split	Uniform	2013.05	2819.44
<i>Split with Bottlenck</i>	N Asikli	Log	3	5.7
		Uniform		
	N Turkish	Log	3	5.7
		Uniform		
	MUTRAT E	Uniform	1E-09	0.000001
	T split	Uniform	2013.05	2819.44
	T bot	Uniform	1944.44	2013.05

We chose a set of informative summary statistics to apply a partial least-squares discriminant analysis on (Frantz et al., 2015; Peter et al., 2012) using the *plsda* function in R (Lê Cao et al., 2009). After retaining a total of 10,000 simulations

closest to the observed data, we compared all models (computing marginal likelihood and posterior probability) simultaneously using a standard ABC generalized linear model (GLM) approach as implemented in ABCtoolbox (Wegmann et al., 2010). The best fitting model was the constant NE size throughout time.

We also computed the average Root Mean Square Error (RMSE) for each parameter using pseudo-observed data to assess our power to infer each parameter in the model.

Table 18. Model fitting on the observed data. The split model fits the observed data the best.

	<i>obsPvalue</i>	<i>Marginal Density</i>	<i>Bayes Factor</i>	<i>Posterior Probability</i>	<i>Chosen Model</i>
<i>Split</i>	0.004	1.15E-08	5.79272	0.852784	TRUE
<i>Split with Bottleneck</i>	0.002	1.98E-09	0.17263	0.147216	

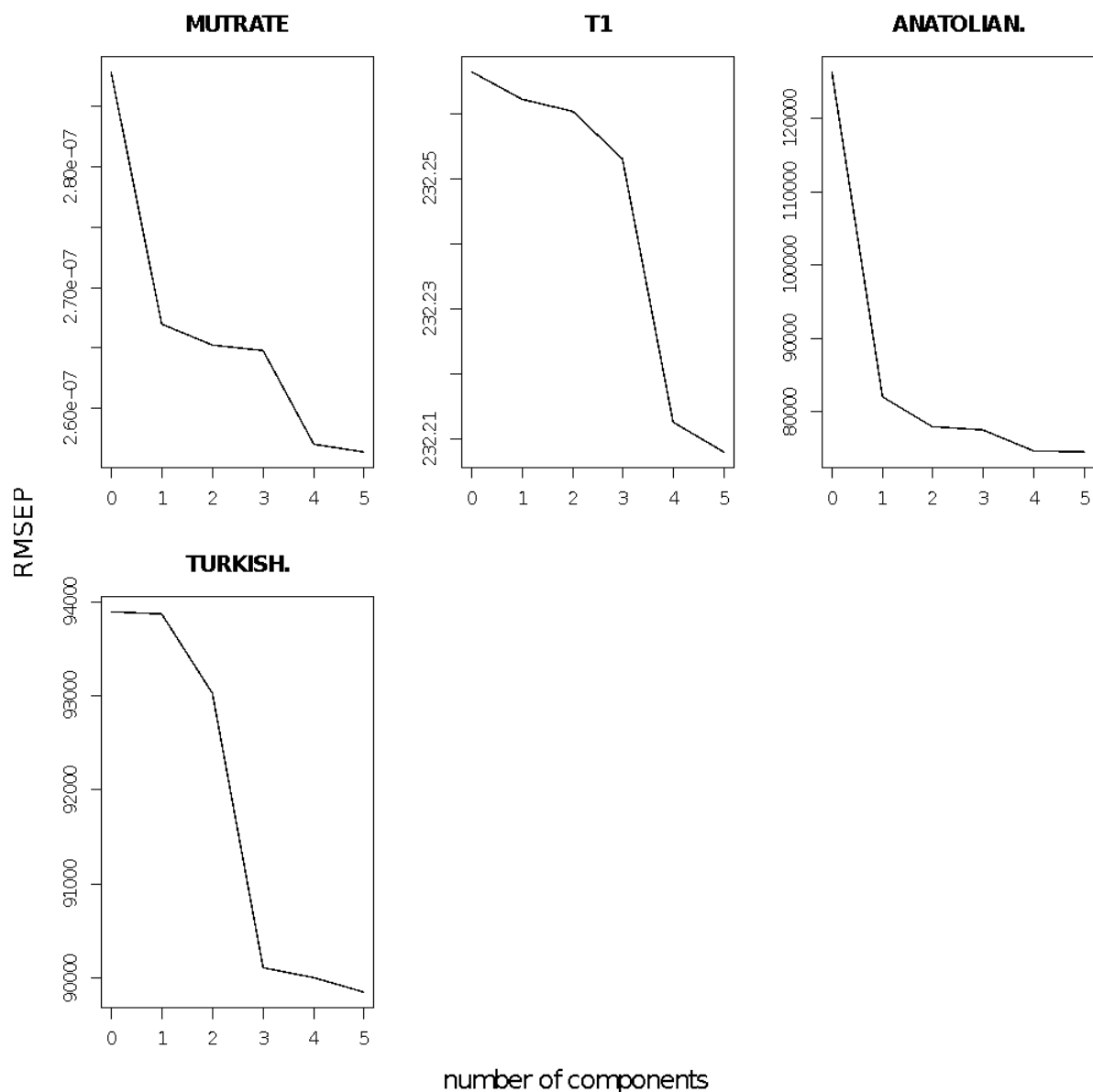


Figure 24. Combined RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate and effective population size of Asikli and Turkish sheep and the time of the split.

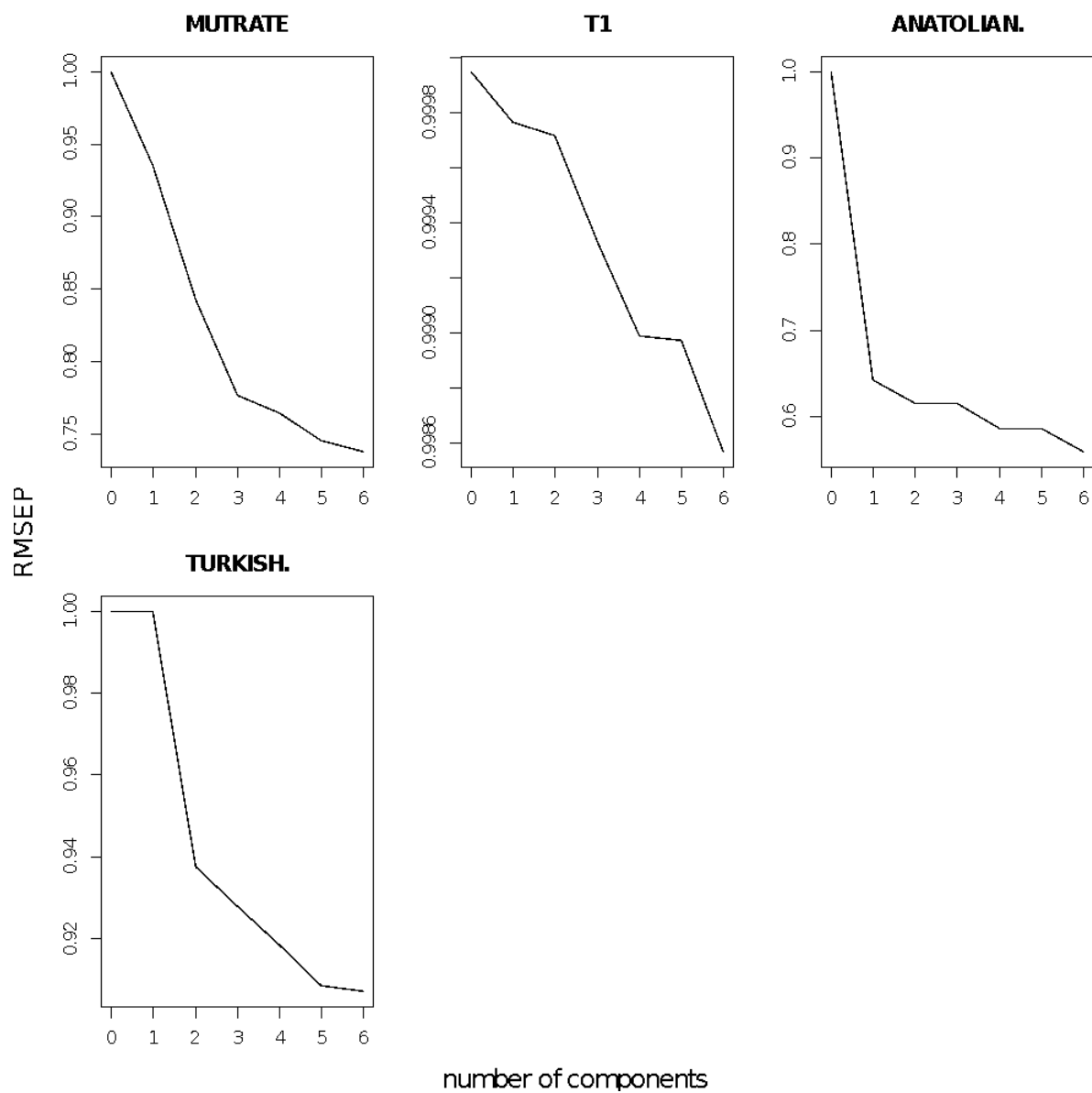


Figure 25. Split RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate, effective population size of Asikli and Turkish sheep and the time of the split.

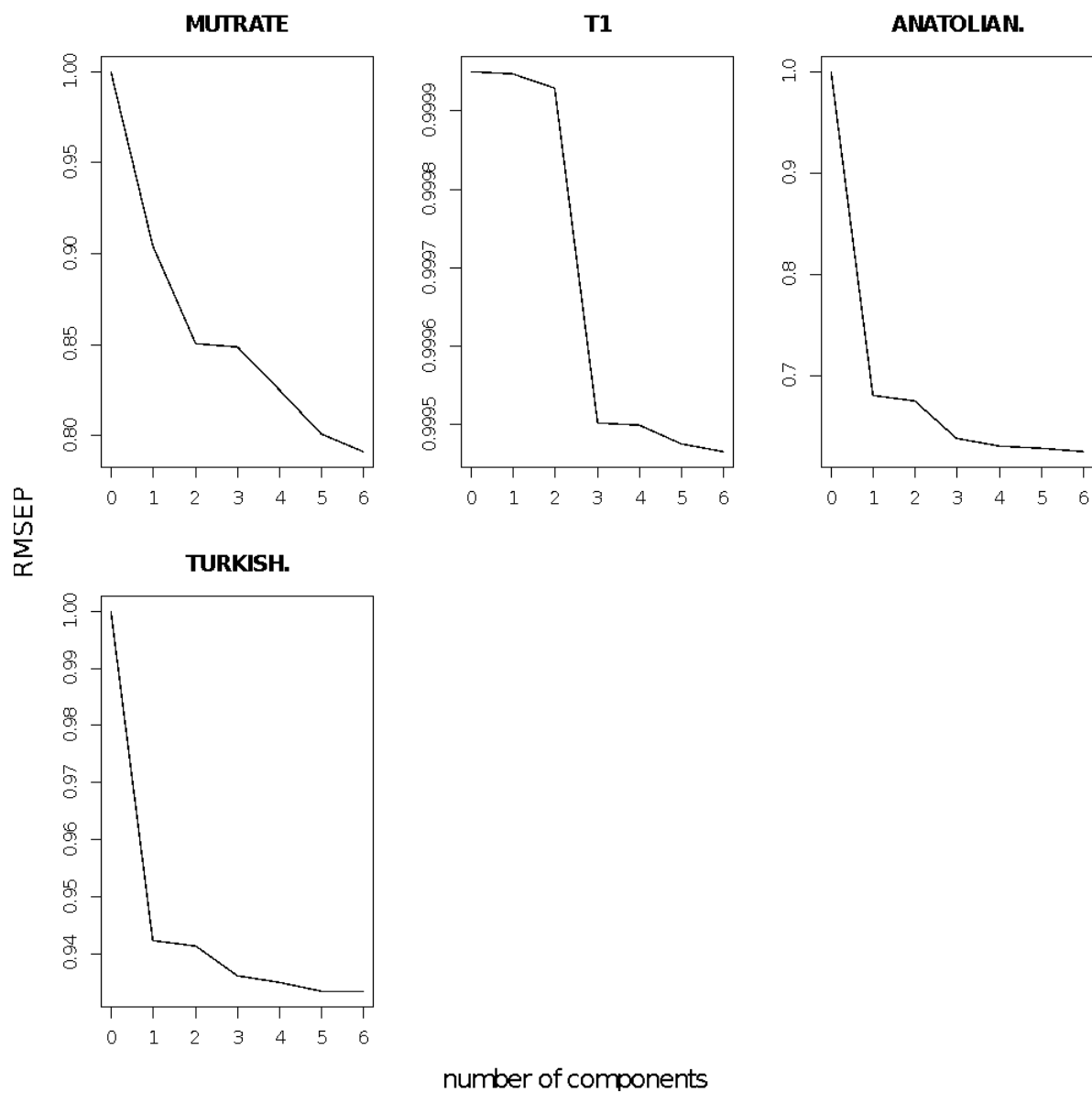


Figure 26. Split with Bottleneck RMSEP. Percentage of the variation observed in the simulated data, explained by the mutation rate, effective population size of Asikli and Turkish sheep and the time of the split.

Best Fitting model – Split

After identifying the split model as the one that fits our observed data the best, model parameter estimation was performed. We ran a total of 2 million simulations under the split model, and for each simulation summary statistics were calculated. These summary statistics were subsequently used as an input for a PLS analysis. The computed PLS were used to transform the summary stats to find the 5 best linear combinations.

The 5 best linear combinations were used as an input for the ABCtoolbox estimator in order to calculate the posterior distributions of N_e of Asikli and Turksih populations, mutation rate and Time of the split parameters. For this purpose, we retained a total of 10,000 simulations closest to the observed data and applied a standard ABC-GLM. The characteristics of the posterior distribution for each parameter of the split model are summarised in the following table.

Table 19. The characteristics of marginal posterior distributions of the parameters of the split model after 2 million simulations.

Parameter	Prior	Prior min	Prior max	RMSE	Mode	HPDI 95 lower	HPDI 95 upper
N Asikli	<i>Log</i> <i>uniform</i>	3	5.7		3.78	3.54	5.07
N Turkish	<i>Log</i> <i>uniform</i>	3	5.7		4.88	3.93	5.66

MUTRATE	<i>Uniform</i>	<i>1.00E-09</i>	<i>0.000001</i>		<i>7.07E-07</i>	<i>2.21E-07</i>	<i>9.95E-07</i>
Time of split	<i>Uniform</i>	<i>2013.05</i>	<i>2819.44</i>		<i>2786.43</i>	<i>2219.97</i>	<i>2814.93</i>

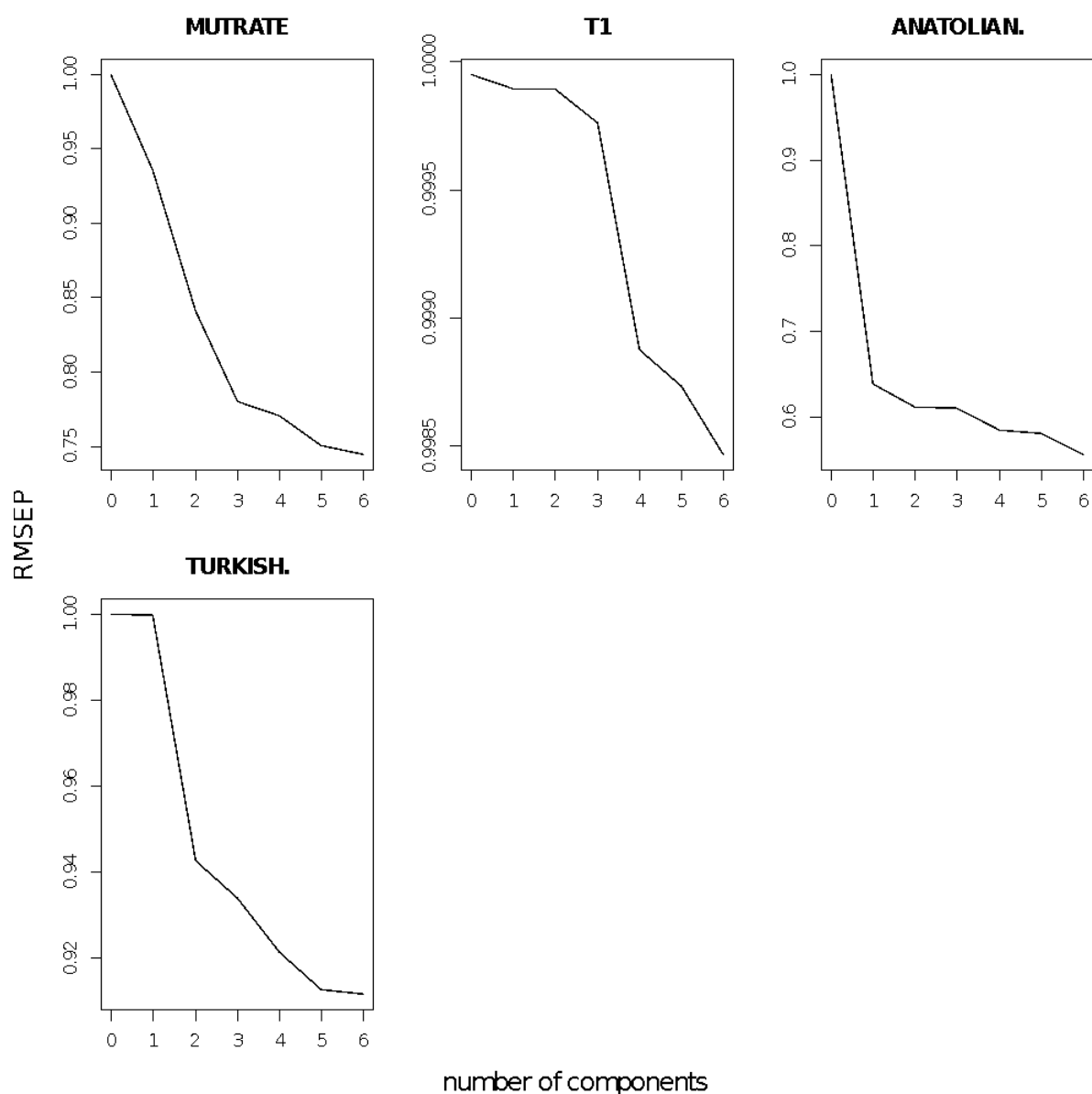


Figure 27. Percentage of the variation observed in the simulated data, explained by the mutation rate, effective population (in log and non-log scales) sizes of the Asikli (Anatolian) and Turkish population and Time of the Split (T1) in split model.

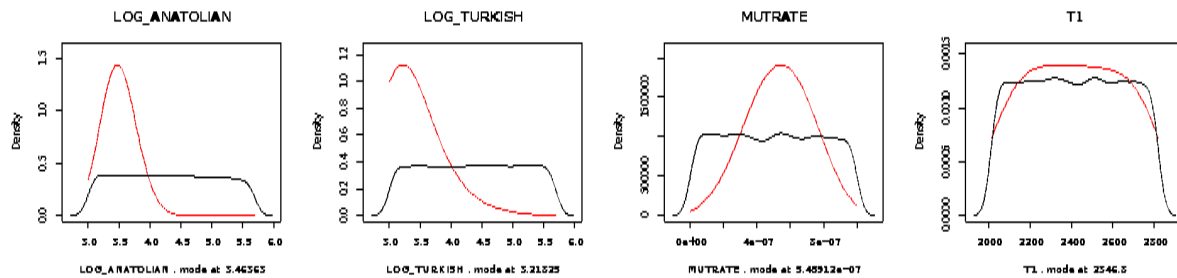
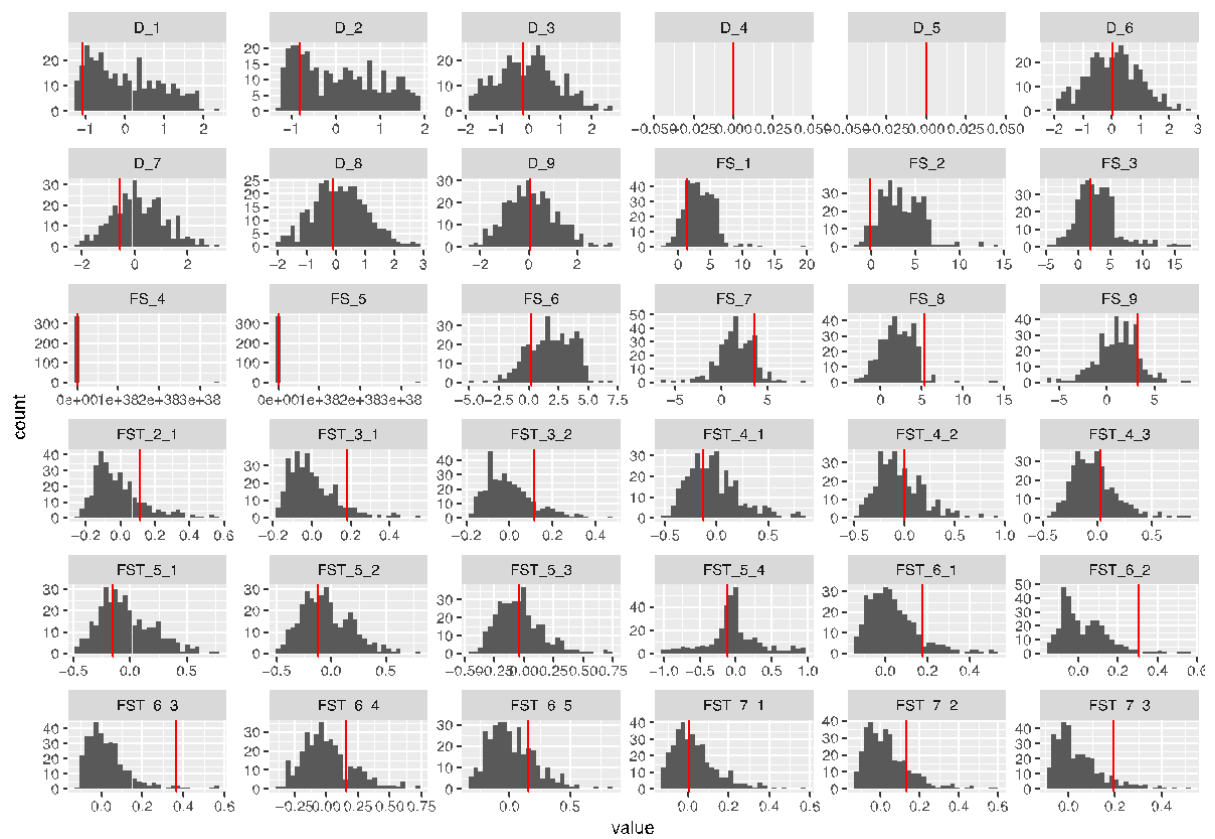


Figure 28. Plots of the uniform posterior distributions of the effective population size of the Asikli and Turkish populations, the mutation rate and time of the split. Blue lines represent the simulated values that have been retained, and the black lines represent the prior distributions of these values.





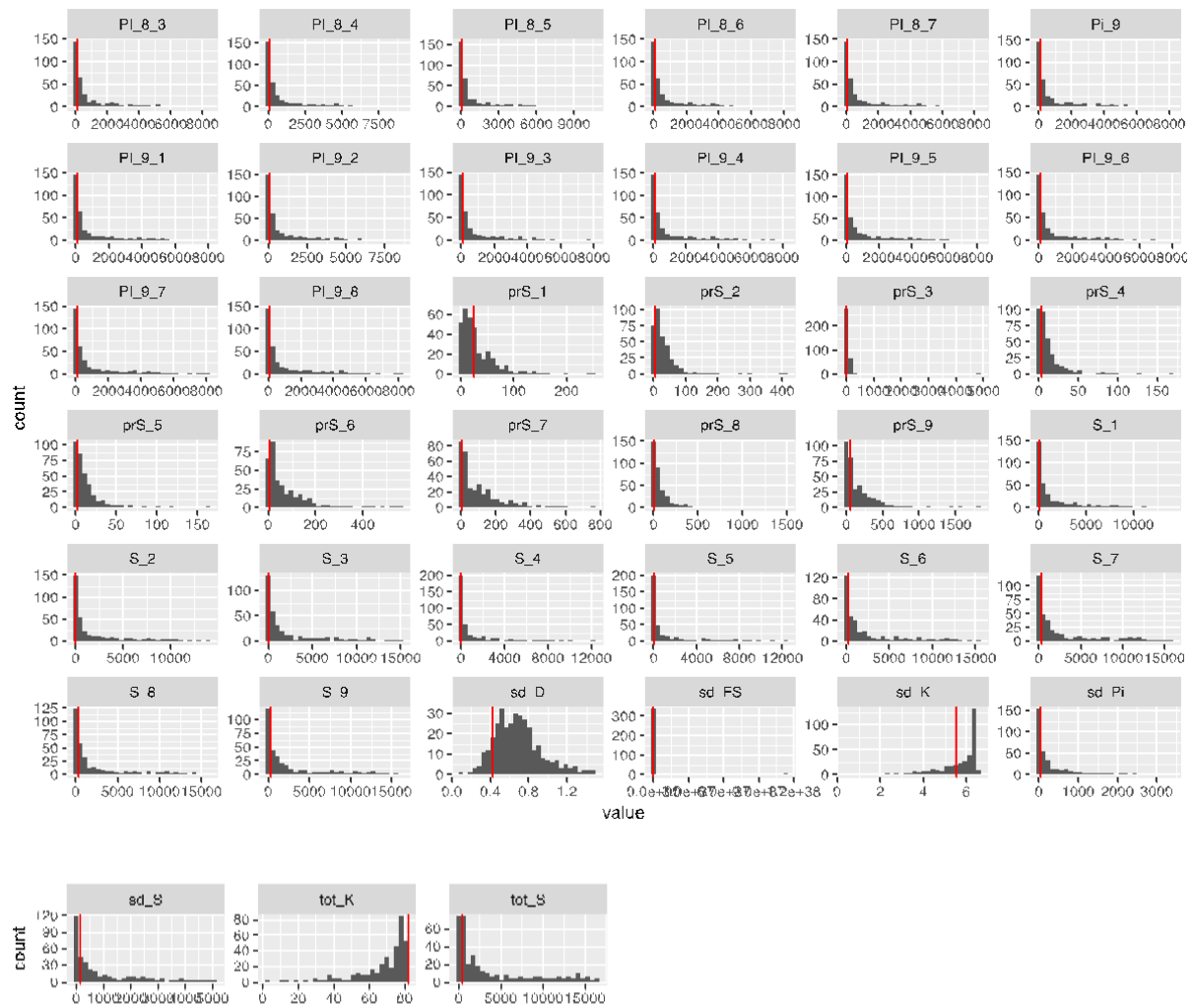


Figure 29. Histogram distribution of the summary statistics of the simulated data. The red bars represent the observed values of the chosen summary statistics.

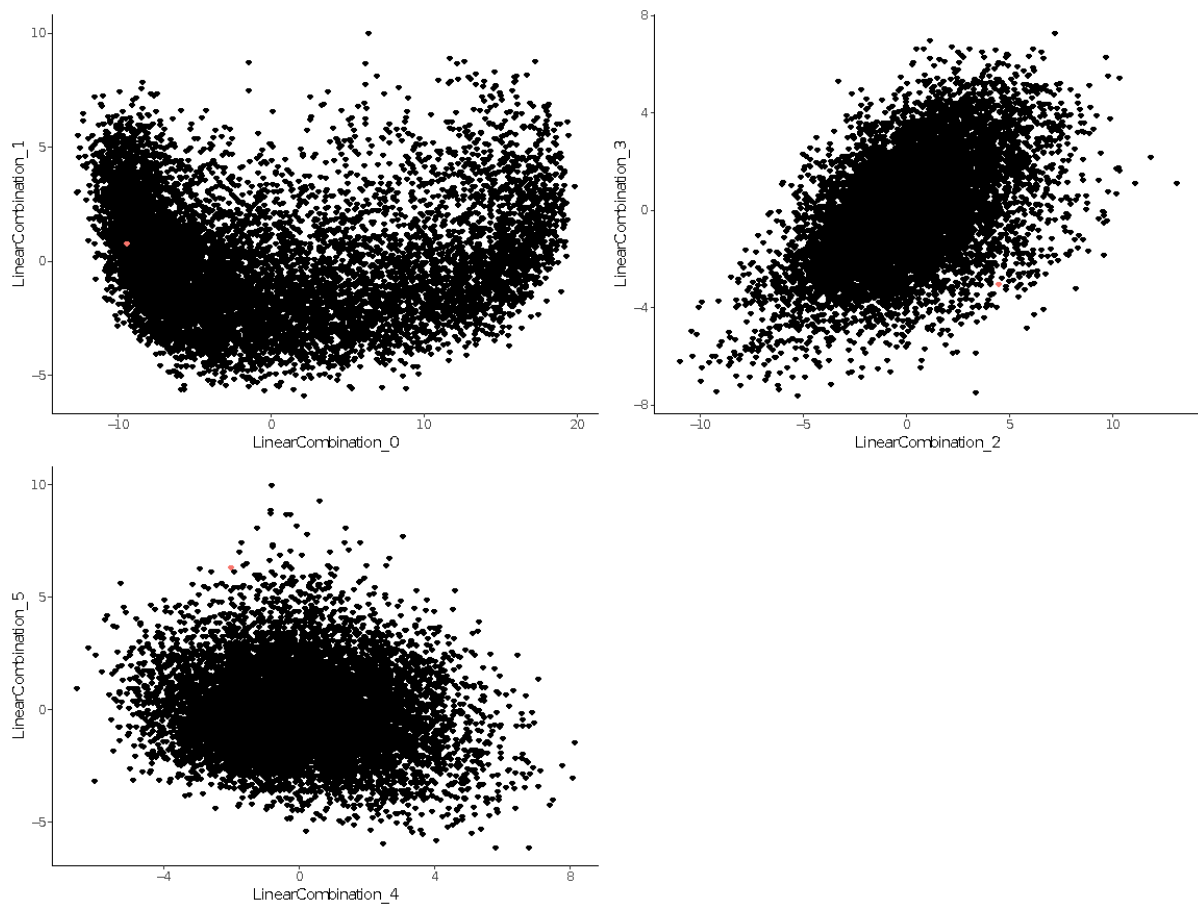


Figure 30. PLS distribution of 10,000 (out of 2,000,000) retained simulations and observed data under the full model. Simulations are shown in black, and observed data are shown in red.

Asikli – Assess Bottleneck strength

Another goal was to estimate the strength of a bottleneck required to leave a detectable trace on the mitochondrial lineage of the Asikli sheep population.

In order to do that we used the summary statistics that were calculated from the 100,000 simulations that were performed for the Asikli population bottleneck model. Each time a set of summary statistics from the simulations were considered as summary statistics derived from observed data. The rest of the summary statistics, from the simulated data of the bottleneck model, were used to perform the model fitting on the pseudo observed summary statistics.

A set of informative summary statistics was chosen to apply a partial least-squares discriminant analysis on (Frantz et al., 2015; Peter et al., 2012) using the `plsda` function in R (Lê Cao et al., 2009). After retaining a total of 5,000 simulations closest to the pseudo observed data, we assessed the fit of model (computing marginal likelihood and posterior probability) using a standard ABC generalized linear model (GLM) approach as implemented in ABCtoolbox (Wegmann et al., 2010). Each time we stored the model fit statistics.

After repeating that process 100,000 times, we summarised the resulting marginal densities, posterior probabilities, observed p values and Bayes factors in a table and plotted them in heatmap, showing the correlation between the time of the bottleneck (T Growth) and the ratio of reduction in the effective population size (NPOP_REL).

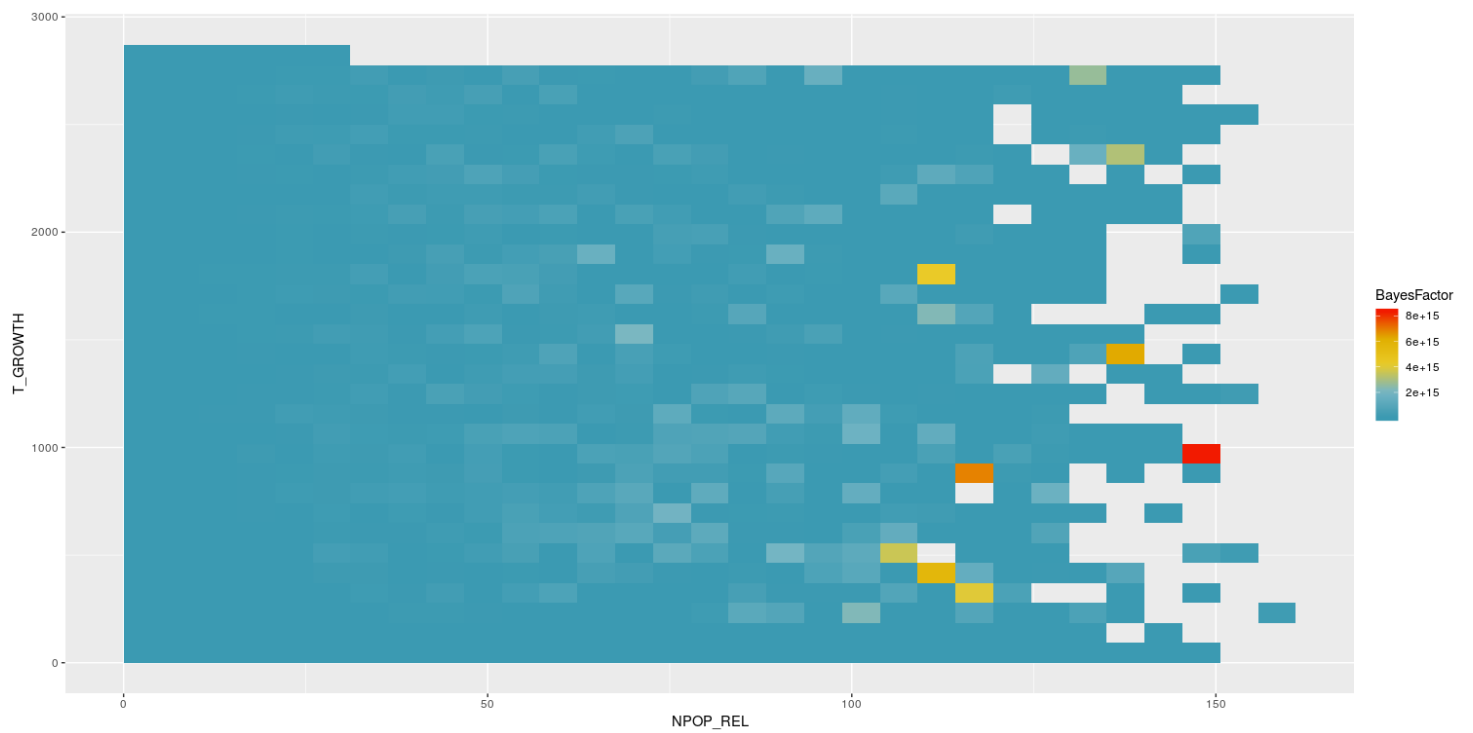


Figure 31. Heatmap between the time of the bottleneck and the ratio of the reduction in effective population size of the Asikli sheep population, the fill is the Bayes Factor from the ABC GLM model fitting. In red are the combinations of time and N_e reduction ratios that gave a high Bayes Factor.

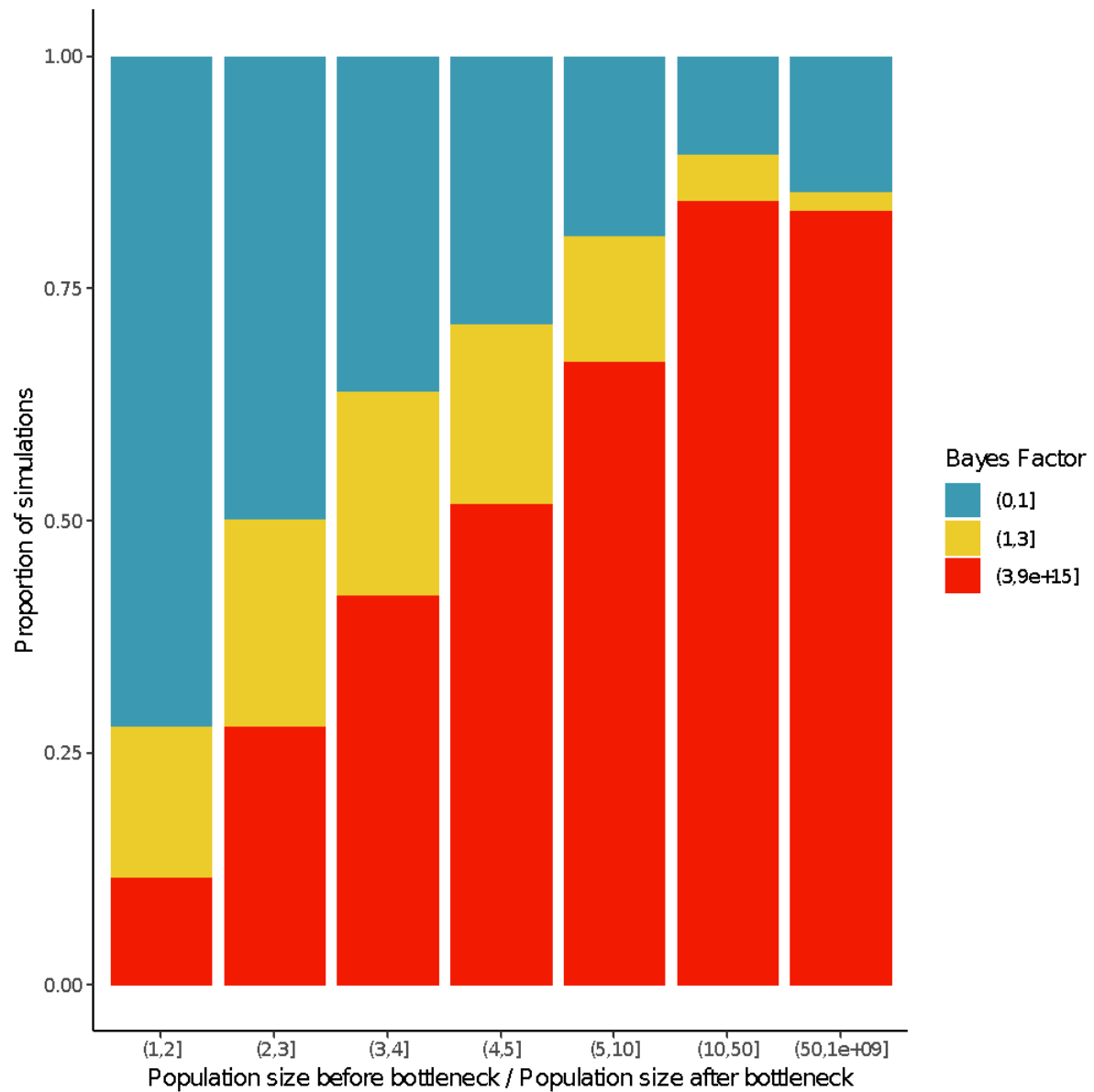


Figure 32. Bar plot frequency of different Bayes Factors for different simulated bottleneck severities. Bayes Factors divided in bins showing the different support for the null hypothesis, among (0-1] evidence for no bottleneck, (1-3] anecdotal evidence for a bottleneck, and (3, 9e+15] strong support for the bottleneck. Ne ratios divided in bins showing the different severity of bottlenecks.

Bibliography

- Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., Foll, M., 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9, e1003905.
- Excoffier, L., Foll, M., 2011. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27, 1332–1334.
- Frantz, L.A.F., Schraiber, J.G., Madsen, O., Megens, H.-J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R.P.M.A., Larson, G., Groenen, M.A.M., 2015. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* 47, 1141–1148.
- Lê Cao, K.-A., González, I., Déjean, S., 2009. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 25, 2855–2856.
- Leroy, G., Mary-Huard, T., Verrier, E., Danvy, S., Charvolin, E., Danchin-Burge, C., 2013. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. *Genet. Sel. Evol.* 45, 1.
- Peter, B.M., Huerta-Sanchez, E., Nielsen, R., 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet.* 8, e1003011.
- Wegmann, D., Leuenberger, C., Neuenschwander, S., Excoffier, L., 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* 11, 116.