Epigenomic study on the domestication of the horse using ancient DNA

Marine Poullet

▶ **To cite this version:**

Marine Poullet. Epigenomic study on the domestication of the horse using ancient DNA. Animal production studies. Université Paul Sabatier - Toulouse III, 2022. English. NNT : 2022TOU30040 . tel-03714400

HAL Id: tel-03714400
https://theses.hal.science/tel-03714400

Submitted on 5 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Marine POULLET

Le 9 mars 2022

Etude épigénomique sur la domestication du cheval à l'aide d'ADN ancien

---

Ecole doctorale : **BSB - Biologie, Santé, Biotechnologies**

Spécialité : **ANTHROPOBIOLOGIE**

Unité de recherche :
CAGT - Centre d'Anthropobiologie et de Génomique de Toulouse

Thèse dirigée par
Ludovic ORLANDO

Jury

M. Gianni LITI, Rapporteur
M. Pierre PONTAROTTI, Rapporteur
M. Gabriel MARAIS, Examinateur
M. Eric CRUBEZY, Examinateur
Mme Clio DER SARKISSIAN, Examinatrice
M. Ludovic ORLANDO, Directeur de thèse

Acknowledgment

First and foremost, I would sincerely like to thank Ludovic Orlando for all the time and energy he dedicated to me. This three-year-long journey was not always easy, but you have always been accessible and patient in helping and guiding me until the very end. You also allow me to be part of an exciting research project and research group.

I would also like to thank all the members of my yearly PhD committee, Etienne Danchin and Laurent Frantz, for supporting me and giving me important advice. I would also like to thank all the members of my jury to evaluate my work.

This PhD would not have been possible without the substantial participation of my work colleagues. In particular, I would like to thank Pablo for sharing his experience and valuable guidance throughout the past three years. I am also thankful to Andaine for the morning motivation routine when I returned to the office and started writing this manuscript. It really cheered me up. Many thanks to Yvette and Evelyn for proofreading and editing this manuscript. I would definitely want to thank Ophélie for her joie de vivre, infectious enthusiasm, and motivation "pom-pom"! It was really amazing every time, and I will, for sure, miss these little moments. Extra thanks to Dirk for sharing his office with me for my first year. I certainly need to thank also my exceptional colleagues and friends, Andreita and Duhita, for sharing the best office ever with me and supporting me until the last moment. Finally, I also want to extend my thanks to the other's members of the team "Archaeology, Genomics, Evolution and Societies" (AGES) for their warm-hearted welcome, which allowed me to work in a pleasant environment: Xuexue, Stephie, Clio, Loreleï, Pierre, Oscar and many more.

I want to address my warm thanks to all my friends at home who have been supporting me throughout these past three years, for dealing with my stressful times, and for supporting me in less happy times, in particular the Bananananation: Nanas, Thomas, Yujin and specifically David, Daz, Rahim and Soub who precisely understand what I am talking about.

Lastly, this three-year-long journey would not have been possible without the support of my family. I want to thank my parents, who have inspired me so much and have always believed in me, and I want to say a special thanks to my grandmother which I am very proud of: 102 years old! You did it! Now, we will have to find another goal together. Finally, it is impossible not to thank you, Marin, my love. I will just say that you were the best to support me on this adventure, and it would have been so much difficult to survive these years without you. Now, since you are following the same path, I will also do my best to help you.

Again, thanks to everybody here.

Abstract

The emergence of Next Generation Sequencing (NGS) in the early 2000s was one of the milestones in the recent history of research. This technology revolutionised not only the world of molecular biology but also that of science in general, allowing advances in both archaeology and cancerology. Using ultra-high-throughput DNA sequencing technology, researchers have characterised thousands of ancient genomes to date, including some now extinct lines. This major technological breakthrough has also revealed that beyond the genome sequences themselves, it is possible to recover ancient epigenetic marks from DNA traces preserved in archaeological material. This made it possible to consider studying the potential evolution of epigenetic modifications carried by individuals who have gone through significant changes in their living conditions, such as domestication for animals. As part of this PhD, we wanted to assess and help resolve some of the various technological and methodological obstacles that make producing reliable estimates of ancient epigenetic marks particularly difficult. Among the various epigenetic marks accessible by ancient DNA, we have only focused on the methylation of DNA present at the level of CpG dinucleotides. The methods developed have enabled us to tackle the question of the epigenetic modifications that have accompanied the domestication of the horse from its origins, around 5,500 years ago, to the present day, a period during which several hundred distinct breeds have been developed. While helping to improve certain aspects of epigenetic inference from ancient data, our work nonetheless revealed significant experimental and analytical limitations. Current procedures for analysing modern epigenetic data are unfortunately not always compatible with the use of ancient data. In addition, this thesis work has provided several avenues for future improvements concerning our ability to manipulate ancient DNA molecules and obtain more reliable alignments on the reference genomes.

Keywords: Ancient DNA, Epigenetics, DNA Methylation, Horse Domestication

Résumé

L'émergence du séquençage de nouvelle génération (NGS) au début des années 2000 a été l'une des étapes marquantes dans l'histoire récente de la recherche. Cela a révolutionné non seulement le monde de la biologie moléculaire mais aussi celui des sciences en général, en permettant des avancées tant en archéologie qu'en cancérologie. Muni d'une technologie de séquençage d'ADN à très haut débit, les chercheurs ont pu caractériser des milliers de génomes anciens à ce jour, y compris parmi des lignées aujourd'hui éteintes. Cette avancée technologique majeure a également révélé qu'au-delà des séquences des génomes elles-mêmes, il est possible de recouvrer des marques épigénétiques anciennes à partir des traces ADN préservée dans le matériel archéologique. Cela a permis ainsi d'envisager étudier l'évolution possible des modifications épigénétiques portées par des individus ayant traversé de grands changements dans leurs conditions de vie, comme la domestication pour les animaux. Dans le cadre de ce doctorat, nous avons voulu évaluer et contribuer à résoudre certains des différents verrous technologiques et méthodologiques qui rendent la production d'estimations fiables des marques épigénétiques anciennes particulièrement difficile. Parmi les différentes marques épigénétiques accessibles par l'ADN ancien, nous ne nous sommes concentrés que sur la méthylation de l'ADN présente au niveau des dinucléotides CpG. Les méthodes développées nous ont permis d'aborder la question des modifications épigénétiques qui ont accompagné la domestication du cheval depuis ses origines, il y a environ 5 500 ans, jusqu'à nos jours, période au cours de laquelle plusieurs centaines de races distinctes ont été développées. S'ils ont contribué à améliorer certains aspects de l'inférence épigénétique de données anciennes, nos travaux ont néanmoins révélé des limites à la fois expérimentales et analytiques importantes. Les procédures actuelles permettant l'analyse de données épigénétiques modernes ne se révèlent malheureusement pas toujours compatibles avec l'utilisation de données anciennes. Par ailleurs, ce travail de thèse a donné plusieurs pistes d'améliorations futures concernant notre capacité à manipuler les molécules d'ADN ancien et obtenir des alignements plus fiables sur les génomes de référence.

Mots clés : ADN ancien, Épigénétique, Méthylation de l'ADN, Domestication du cheval

List of abbreviations

aDNA	Ancient DNA
BCE	Before Common Era
bp	Base pairs
BS	Bisulfite treatment
BWA	Burrows-Wheeler Aligner
CE	Common Era
CGI	CpG Island
cpDNA	Chloroplast DNA
DNA	Desoxyribonucleic Acid
ENA	European Nucleotide Archive
FAANG	Functional Annotation of Animal Genomes
HTS	High-Throughput Sequencing
Kya	Thousand years ago
LIMS	Laboratory information management systems
Mb	Mega bases
MBD	Methylated Binding Domains
mtDNA	Mitochondrial DNA
NGS	Next-generation sequencing
nuDNA	Nuclear DNA
PCA	Principal component analyses
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
RRBS	Reduced representation bisulfite sequencing
SBS	Sequencing-By-Synthesis
SNP	Single-Nucleotide Polymorphisms
TSS	Transcriptional start site
Vg	Variation graph
^{5m}C	5-methyl-cytosines

Table of contents

I. Introduction: from ancient DNA molecules to epigenomics 11

1. **A brief introduction to ancient DNA research** 12
   - 1.a 1984, 1985: the first ancient DNA studies 12
   - 1.b From DNA cloning to PCR amplification 13
   - 1.c The first technical game changer: PCR 13
   - 1.d The second game changer: using osseous remains 14
   - 1.e A race to find the oldest and most spectacular almost compromises the future of ancient DNA research 14
   - 1.f The third game changer: setting up strict guidelines 15
   - 1.g Characterisation of post-mortem damages 16
   - 1.h Almost no studies successfully characterised nuclear loci until the mid 2000s 18
   - 1.i The fourth game changer: introduction of Next-Generation Sequencing 19
   - 1.j The fifth game changer: finding true aDNA reservoirs 20
   - 1.k Most recent game changers 22
   - 1.l Bioinformatic Challenges 23

2. **Ancient epigenetics** 27
   - 2.a Epigenetics and its definition 27
   - 2.b Molecular mechanisms underlying epigenetic traits 28
   - 2.c Characterising methylomes 31

3. **How ancient (epi)genomic studies can advance our understanding of animal domestication?** 35
   - 3.a Introduction to animal domestication 35
   - 3.b State of the art of domestication studies 37
   - 3.c Limitations encountered with the study of ancient domesticated genomes 39
   - 3.d Temporal and geographic origins of horse domestication 41
   - 3.e Impact of horse domestication on the dynamics of human history 44
   - 3.f Assessing ancient DNA methylation profiles underlying horse domestication 46

4. **Objectives of the PhD project** 49

II. Methodological improvement towards the analysis of ancient methylomes 51

1. Context 53
2. Rationale and Hypothesis 55
3. Article 1: Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes 56
4. Discussion 71

III. Unravelling technological and methodological obstacles for characterising epigenetic marks in ancient equine specimens 73

1. Context **75**
2. Article 2: Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNAs **76**

IV. Experimental design for the reconstruction of ancient DNA methylation maps 113

1. Introduction **113**
2. Material and Methods **114**
3. Results **120**
4. Discussion **127**

V. Perspectives 131

1. Improving ancient DNA read mapping **131**
2. Detecting epigenetic marks **132**
3. Accessing ancient metagenomes **133**

VI. Appendix 137

1. Appendix 1: Supplementary information - Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes **137**
2. Appendix 2: Supplementary information - Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA **147**
3. Appendix 3: Supplementary information - Experimental design for the reconstruction of ancient DNA methylation maps in horses **169**

VII. References 177

I. Introduction: from ancient DNA molecules to epigenomics

This manuscript comprises two main research articles, and is complemented with an additional section describing unpublished analyses.

The first article helps resolve certain aspects of epigenetic inference from ancient data. It highlights the importance of using a dedicated mapping strategy when aligning ancient DNA sequences against modern reference genomes and identifies an alignment procedure that improves downstream reconstruction of DNA methylation maps. The second article revealed significant experimental and analytical limitations of one of the main enzymatic treatments that is used in ancient DNA analyses. While this treatment can improve ancient DNA sequence data quality and is instrumental for the prediction of ancient DNA methylation marks, it was found to also affect negatively some experimental procedures, such target-enrichment capture, that are aimed at improving local sequence coverage and facilitate the characterisation of ancient genomes.

The following introduction contains three chapters. The first chapter is a brief introduction to ancient DNA, emphasising major technological milestones that have allowed studying many high-quality ancient genomes, including horses. The second chapter contains an introduction to ancient epigenetics with a particular focus on the methodology. The last section focuses on horse domestication.

These chapters are not intended to provide a comprehensive guide to ancient DNA research but to provide sufficient background information for the two research articles, and the methods part included in this PhD thesis.

1. A brief introduction to ancient DNA research

1.a 1984, 1985: the first ancient DNA studies

Ancient DNA (aDNA) research began in 1984 with the extraction and sequencing of the first short DNA fragments from the dried muscle of a Quagga zebra museum specimen. This species is a member of the horse genus (*Equus*) that went extinct in South Africa at the beginning of the 20th century (Higuchi et al., 1984, Figure 1). The following year, molecular cloning was used to amplify small sequences retrieved from the skins of Ancient Egyptian mummies (Pääbo et al., 1985). The fraction of DNA molecules extracted from those ancient animal and human remains (endogenous DNA) were limited to very low concentrations of short, damaged fragments of multi-copy loci. However, it was the first time that molecular biology techniques offered researchers the opportunity to travel back in time and directly obtain DNA sequence information from a long-dead individual. The ability to study DNA from museum specimens and archaeological specimens opened up the possibility to answer an entirely new array of questions relating to anthropology, evolutionary biology, and the environmental and archaeological sciences. The purpose of this chapter is to give a brief introduction to the main milestones that have marked the history of aDNA research.

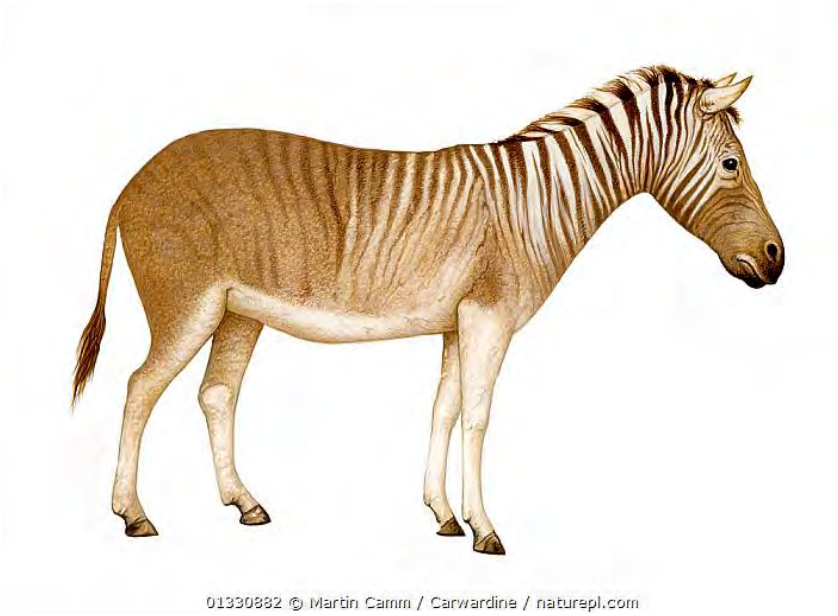


Figure 1: Quagga zebra specimen (*Equus quagga quagga*)

Illustration from Martin Camm at the Wildlife Art Company. The Quagga zebra went extinct in the wild by end of the 19th century. Ancient DNA research began in 1984 with the analysis of the first short DNA fragments from a museum specimen.

1.b From DNA cloning to PCR amplification

Molecular cloning in living bacteria made it possible to extract and sequence DNA from ancient organic remained preserved in different contexts (Higuchi et al., 1984, Pääbo et al., 1985) and suggested that nuclear DNA, as well as mitochondrial DNA (mtDNA), may persist for millennia. However, this cloning procedure was time-consuming, could be easily contaminated and proved sub-optimal when applied to short, damaged DNA fragments. Since then, a series of technological and methodological advances have enabled the successful sequencing of complete genomes from ancient organisms such as ancient humans, dogs, mammoths and equids (Marciniak and Perry, 2017). The advent of the Polymerase Chain Reaction (PCR) is probably the first on the list of such instrumental technical (Mullis and Faloona, 1987), and was used for the first time to recover sequence data from a 7,000-year-old ancient human brain (Pääbo et al., 1988).

1.c The first technical game changer: PCR

The development of the PCR technique made it possible to routinely amplify a large quantity of a specific DNA sequence even when the number of starting templates was limited (Higuchi et al., 1988). Thus, in addition to being less time consuming than molecular cloning in living bacteria, this technique enabled the replication of undamaged sequences of interest (Pääbo and Wilson, 1988).

Various analyses have shortly followed the development of PCR to allow further experimental manipulation (Höss et al., 1996a, Höss et al., 1996b, Thomas et al., 1989). For example, one study has inferred the phylogenetic history of the extinct marsupial wolf (*Thylacinus cynocephalus*) with mtDNA from museum specimens using direct PCR sequencing (Thomas et al., 1989). Another one recovered 1,100 base pairs (bp) of mitochondrial ribosomal DNA of the extinct ground sloth (*Mylodon darwinii*) and inferred phylogenetically that the arboreal lifestyle had evolved at least twice among sloths (Höss et al., 1996a). Those examples further established the extraction and sequencing of DNA from museum specimens and archaeologists by PCR as a viable approach to the genetic characterisation of extinct animals. In addition, they opened up access to other types of remains.

1.d The second game changer: using osseous remains

The first aDNA studies focused only on soft, wet materials preserved in peat bogs or materials preserved by dehydration, mummification or freezing (Higuchi et al., 1984, Pääbo et al., 1985). Five years later, a new significant technical advancement had been performed with the discovery of experimental procedures capable of recovering DNA from osseous remains (Hagelberg et al., 1989), such as bones and teeth. In 1989, utilisation of PCR techniques allowed for aDNA extraction and amplification from human's bones that were more than 5,000-year-old (Hagelberg et al., 1989). This latter study showed that the aDNA field could now tap into a considerably larger resource of archaeological and palaeontological remains to provide biological insights into past populations.

1.e A race to find the oldest and most spectacular almost compromises the future of ancient DNA research

Unfortunately, the PCR technology suffers from some limitations that can affect the authenticity of the amplified sequences, especially when DNA damage is present (Pääbo et al., 1990). For example, a phenomenon called "jumping PCR", recombining different molecules together, can generate erroneous sequences during amplification (Pääbo et al., 1990). In addition, other limitations to this PCR process are the small size of the ancient DNA templates, the fact that it is only possible to target a limited subset of the genome at once, and the propensity for contamination (Poinar et al., 1998, Dabney et al., 2013). Indeed, due to endogenous ancient DNA molecules being highly fragmented and preserved only in small amounts, PCR is sensitive to many sources of contamination: the environmental DNA surrounding the fossil (Höss et al., 1996a), modern human contamination caused by the handling of the ancient samples (Zischler et al., 1995), contamination of the reagents (Höss et al., 1996b) and also hydrolytic and oxidative DNA damage that both accumulate naturally after death (Pääbo et al., 1989).

Due to the amplification power of PCR, any source of intra-laboratory contamination is increased. This resulted, in the nineties, in the publication of numerous false-positive claims that have since been disproved. For example, the sequencing of DNA from 80 million years old dinosaur bones found in eastern Utah in the roof of an underground mine (Woodward et al., 1994), later turned out to be a human mitochondrial gene insertion in the nuclear genome

(Zischler et al., 1995). Additionally, supposedly weevil DNA from amber inclusions that was claimed to be 120–135 million years old (Cano et al., 1993) was instead modern fungal contamination (Gutierrez and Marin, 1998). Contamination has also resulted in incorrect conclusions being drawn from studies on plants (Golenberg et al., 1990, Soltis et al., 1992, Goloubinoff et al., 1993) and other amber inclusions (Cano et al., 1995, DeSalle et al., 1992, 1993, Poinar et al., 1993). During this decade, however, the most contentious area was human DNA due to the difficulty of recognising contamination from modern humans (Pääbo et al., 1989, Zischler et al., 1995). Moreover, as many human museum specimens have been handled without strict procedures for a long time, contaminant molecules could also appear ancient and exhibit appropriate molecular behaviour (Handt et al., 1994). Thus, questions about the authenticity of the amplified sequences have often been raised and require a proper methodology to be mitigated or alleviated.

1.f The third game changer: setting up strict guidelines

Fortunately, these formidable obstacles within human ancient DNA studies led aDNA researchers to develop strict guidelines, methodological procedures and clean facility protocols to keep the contamination of ancient samples under control and to allow reproducibility of their studies (Handt et al., 1994, Cooper and Poinar, 2000, Pääbo et al., 2004, Briggs et al., 2007, Fortea et al., 2008). This new methodology is expensive and time-consuming but necessary since one aerosol droplet can easily contain a thousand times the amount of amplifiable mtDNA present in most ancient human remains (Handt et al., 1994). Besides, the analysis of aDNA requires the destruction of irreplaceable, finite subfossil material that is part of our common bio-cultural heritage. Therefore, it is essential to guarantee that results are authentic, even if they can never totally exclude traces of modern contamination caused by the minute amounts and degraded nature of surviving DNA (Cooper and Poinar, 2001, Hofreiter et al., 2001).

Furthermore, it seems that if the contamination occurs a long time before aDNA extraction, through multiple handling or washing during the excavation process (Richards et al., 1995), many of the techniques used to decontaminate contemporary human DNA on the outer surface of the material (with the application of bleach, exposure to UV irradiation, grinding and shot-blasting) are almost inefficient on bone and teeth samples (Gilbert et al., 2005). The natural shield made by enamel around the teeth offers more resilient protection against contamination

than bones but, it is not enough since both samples are quickly and almost permanently contaminated over time (Gilbert et al., 2006). In this way, knowing the history of how the samples were handled before the analysis can prove critical.

1.g Characterisation of post-mortem damages

The next breakthrough in the characterisation of aDNA pertains to improved understanding of post-mortem damage. Quickly after death, some enzymatic and microbial factors can lead to irreversible damage, characterised by a dramatic reduction in size down to an average of 30 to 50bp (Briggs et al., 2007, Figure 2a). However, under favourable environmental conditions, such as low temperatures, dry or frozen conditions, these processes may slow down and even become inhibited before destroying all endogenous DNA (Dabney et al., 2013). As a consequence, some older samples can appear less damaged than recent ones depending upon their local environmental conditions. These damages are due to different mechanisms that degrade the DNA. In living cells, the chemical DNA degradation process is an ongoing process that is constantly restored and almost fully balanced by an entire DNA repair machinery (Lindahl, 1993). However, when the organism is no longer alive, such enzymatic processes stop, and considerable damage accumulates over time, leading to shorter DNA molecules. In addition, several chemical mechanisms participate in the degradation of DNA shortly after death, including free radicals, DNA oxidation and a non-enzymatic hydrolytic cleavage. This latter degradation begins with a depurination event which leaves an abasic site in the middle of the double-stranded DNA (Briggs et al., 2007, Figure 2b). This step is followed by a beta-elimination reaction that generates a single-stranded nick following the abasic site (Lindahl and Andersson 1972). The resulting unrepaired single-strand break has then more chance to become a double-strand DNA break. This process is considered as one of the main contributors to DNA fragmentation (Glocke and Meyer, 2017).

Nucleotidic bases are also the target of important post-mortem modifications, particularly cytosines, which can rapidly undergo a hydrolytic deamination process. Such damages are significant as they can inhibit the activity of PCR (Poinar et al., 1998). Cytosine deaminations are primarily found in the single strand overhangs that are formed following fragmentation by depurination (Figure 2b). Cytosines are first converted into uracils by a hydrolytic deamination process, and are then mis-identified as thymines during the PCR amplification

step resulting in the accumulation of CG→TA substitutions in the sequence data (Hansen et al., 2001).

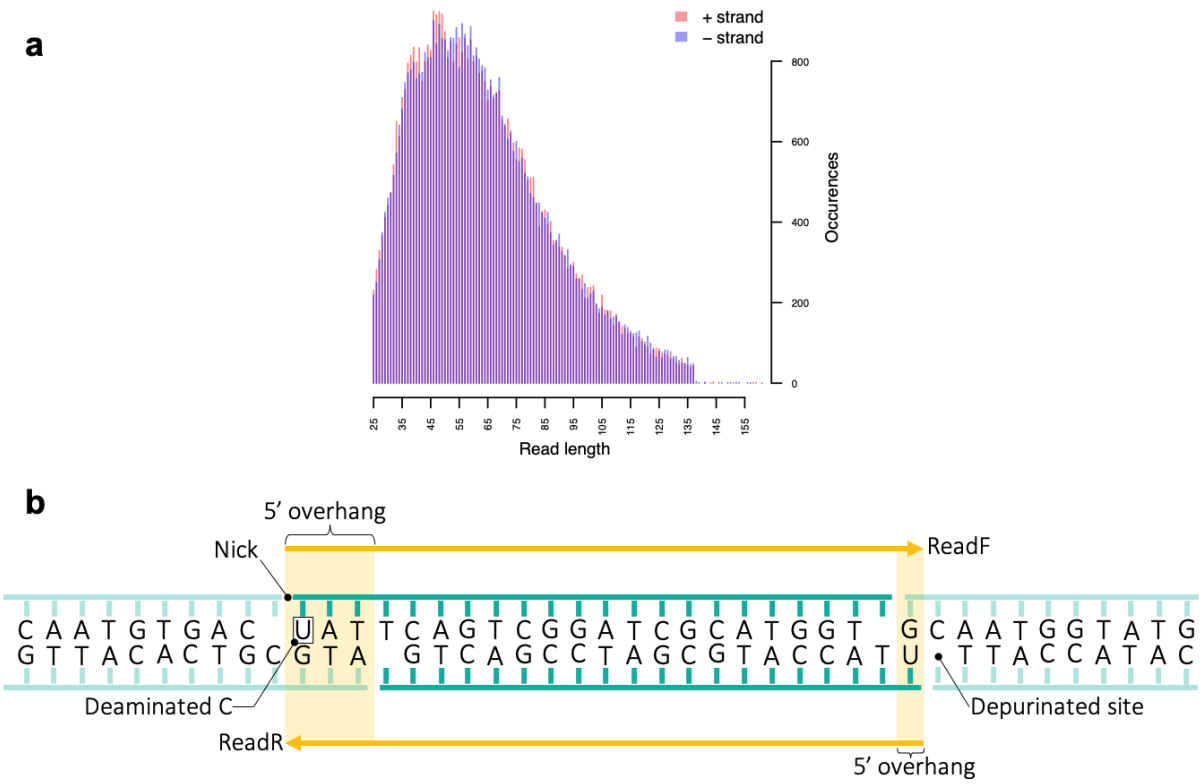


Figure 2: Characterisation of port-mortem DNA damages

Modified from Orlando 2021. **(a) Single-end read length per strand.** mapDamage plot of an ancient horse specimen. The distribution of the read length shows a small average size which is specific of ancient DNA reads for both strands. **(b) Process of DNA fragmentation in an ancient DNA molecule.** Degradation begins with a depurination event resulting in the formation of overhanging ends and single-stranded nicks. The resulting unrepaired break has more chance to become a double-strand DNA break.

DNA chemical decay and high sensitivity of polymerases to modern contamination limited the success of PCR on aDNA molecules and restricted most studies to phylogenetic inference from small sequence alignments during most of the first two decades of aDNA research. Next, a new methodological approach emerged to assemble long, continuous DNA sequences using minimal amounts of fragmented ancient DNA as a template (Römpler et al., 2006a, Römpler et al., 2006b). It consisted of a two-rounds of multiplex PCR amplification that bypasses the limitations and the speed of the standard PCR. In comparison to simplex PCRs, this multiplex

protocol has the advantages of extracting several kilobases of sequence from limited template amounts of DNA using simultaneously multiple primer pairs in only one reaction and finally amplifying rapidly a very high number of loci (Römpler et al., 2006a). However, despite opening access to new investigations without exhausting bone resources, the two-round multiplex still suffered from some limitations typical of PCR-based approaches. These limitations include the necessity to know in advance the sequence of at least a closely related species to enable primer design and the relative scarcity of DNA templates of sufficient size and the presence of PCR inhibitors (Kemp et al., 2014). Indeed, in a sufficient amount, numerous substances co-extracted with aDNA have the potential for inhibiting the PCR.

1.h Almost no studies successfully characterised nuclear loci until the mid 2000s

Until the revolutionary advent of next-generation DNA sequencing (Poinar et al., 2006), the scope of aDNA studies was often limited to mitochondrial or chloroplast DNA (cpDNA). Little was known about DNA sequence variation at nuclear loci at the beginning of the 21st century because of the degraded nature of aDNA extracts resulting in very short fragments (Pääbo et al., 1989). However, it is also essential to sequence several thousand base pairs from many individuals to observe enough variable positions to assess variation accurately. Mitochondrial DNA has proven to be the main target for genetically characterising ancient samples and is present in several hundreds to thousands of copies per cell, in contrast to the single-copy nuclear genome. Thus, target sequences of mtDNA are more likely to be present in any single extract and accessible for amplification than single-copy nuclear locus (Ljungman and Hanawalt, 1992). However, even if a sample of ancient DNA contains more mtDNA copies, the number of copies well preserved enough to be analysed is limited at most. In some well-preserved samples, low- and single-copy nuclear DNA (nuDNA) sequences have been reported from plants preserved in dry environments (Jaenicke-Despres et al., 2003), museum specimens (Bunce et al., 2003, Huynen et al., 2003), as well as several short nuclear loci carrying Single-Nucleotide Polymorphisms (SNP) of functional relevance (Römpler et al., 2006b, Krause et al., 2007). In addition to these studies, another strategy was implemented in 2005 to recover no less than 27 kilobases (kb) of a cave bear nuclear genome (Noonan et al., 2005). This methodology relied on the shotgun sequencing of DNA libraries multiplied in bacterial vectors and revealed the exogenous origin

of most sequences present in aDNA extracts, a significant discovery that triggered research in ancient metagenomics.

1.i The fourth game changer: the introduction of next-generation sequencing

Next-generation sequencing (NGS) has been the most transformative step in the history of ancient DNA research. It has profoundly affected wet-laboratory and dry-laboratory activities and outstripped the state-of-the-art Sanger sequencing. These approaches could generate millions of short DNA reads per run (Metzker, 2010), which perfectly fit the ultra-short nature of ancient DNA fragments that are too short for conventional PCR amplification. Instead, they rely on the development of sequencing libraries by the ligation of DNA adapters. These adapters then serve as primers both for amplification and for sequencing. Besides, these High Throughput DNA Sequencing technologies (HTS) grant access to entire genomes instead of only target regions, as is the case with Sanger sequencing. In addition, HTS reduces the cost and time of DNA sequencing.

The first application of those second-generation platforms to ancient remains was in 2006 with the large-scale sequencing of an extinct animal (Poinar et al., 2006). Using the 454 Life Sciences platform (Margulies et al., 2005), in no more than a few days, 28 Megabases (Mb) of metagenomic DNA were sequenced, from which 13 Mb (45.4%) of the sequencing reads originated from a 27k years-old woolly mammoth preserved in a dedicated permafrost museum. For the first time, thanks to the high percentage of endogenous DNA recovered from this single mammoth remain, it was demonstrated that sequencing of whole mammalian genomes could probably be afforded and performed in a reasonable time frame and price range, despite the relatively limited amounts of DNA usually preserved in fossil remains. Furthermore, this pyrosequencing technology took no more than two years to provide the first near-complete nuclear genome of a 20k years-old mammoth (Miller et al., 2008). Ancient DNA was thus not limited to analysing a handful of genes anymore but could complete whole-genome sequencing from ancient individuals and extinct species. These analyses will answer historical questions in molecular evolution and tackle the molecular basis of speciation as to whether or not analyses are correctly done. Up until this time, fossil remains yielded little genetic insight into evolutionary processes. This was due to poor preservation of their DNA

and the limited ability to retrieve a suitable amount of endogenous DNA within the mixture of bacterial, fungal and often human contaminants present.

1.j The fifth game changer: finding true aDNA reservoirs

Bones and teeth were usually considered the most sustainable physical evidence of human or animal presence at an archaeological site, and they were also the most widely used samples for aDNA studies. Furthermore, sampling teeth is less demanding, and teeth can be unequivocally assigned to an individual skull (Hummel and Hermann, 1994). Past analyses have been done to confirm their suitability as a source of aDNA. Their long conservation has been explained by the reasonably low water and enzyme content of hard tissues to avoid diagenetic effect (Hummel and Hermann, 1994) and by a mechanism reminiscent of mummification (Bell et al., 1996, 2001). Despite these assets, an awareness of sample handling as a source of contamination motivated further research, resulting in another crucial development in the history of aDNA. It occurred only two years after the release of NGS approaches to identify keratinous appendices such as hair samples acting as actual aDNA reservoirs (Gilbert et al., 2008). Indeed, ancient hair samples showed a lower susceptibility to contamination with modern DNA, unlike osseous remains.

Thanks to a new throughput sequencing technology, the Sequencing-By-Synthesis (SBS) Illumina technology, the mitochondrial genome of a Paleo-Eskimo human was sequenced from a 3,400- to 4,500-year-old frozen hair excavated in an early Greenlandic Saqqaq settlement. These findings revealed that the Saqqaq are not descendent of Native Americans, but instead, they are likely derived from a population in the Bering Sea area (Gilbert et al., 2008). These results were dwarfed two years later by the publication of the first ancient hominin genome, obtained from the same 4 thousand years ago (kya) permafrost-preserved-hair (Rasmussen et al., 2010). These results showed that hair samples could genuinely be used to uncover genetic signatures (Figure 3), and paved the way to studies that ended up dramatically improve our knowledge of humans' evolutionary history and archaic hominins.

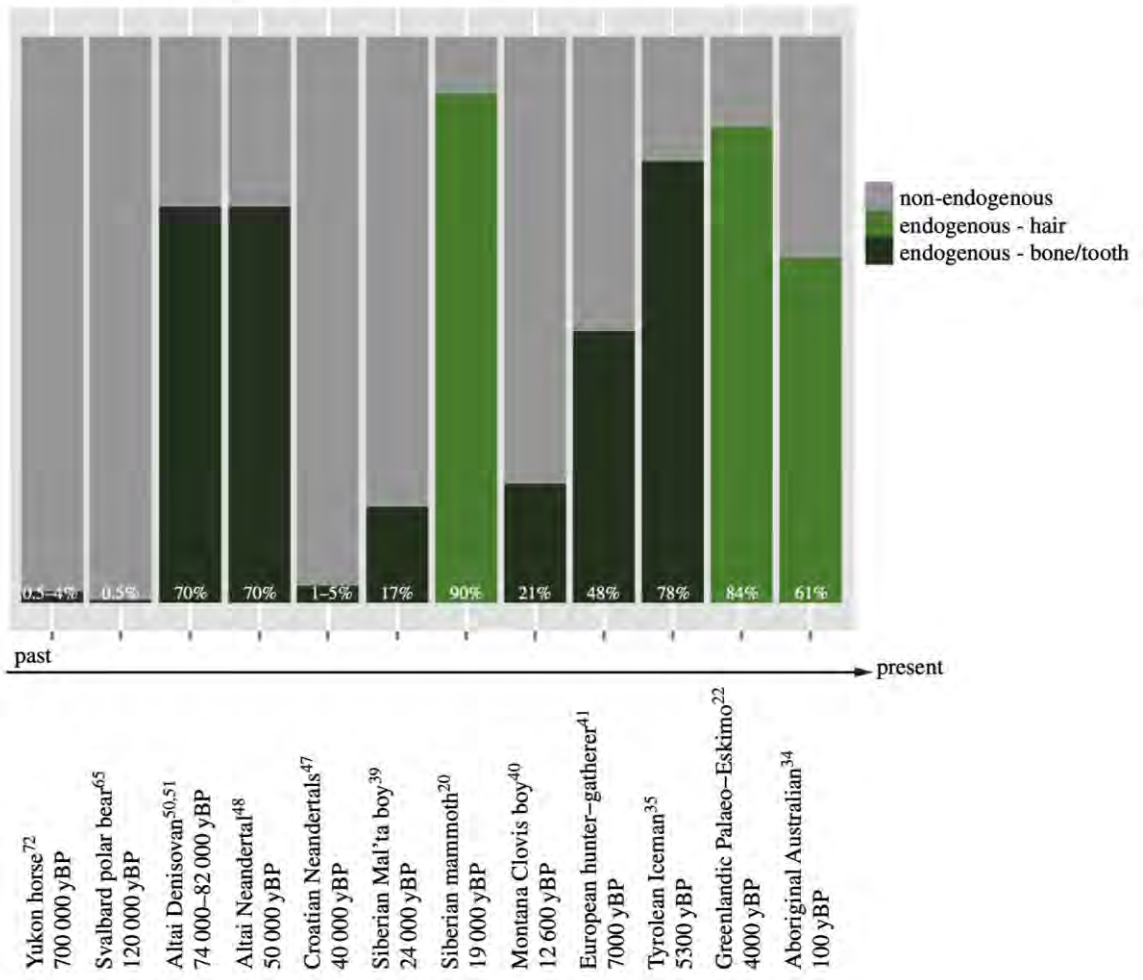


Figure 3: Endogenous content of ancient genomic extracts.

Modified from Der Sarkissian et al., 2015. Datasets are ordered from the most ancient sample to the most recent. 'yBP', years before present. Hair samples show a higher endogenous in comparison to other ancient genomic extracts.

Despite the field's slow start, developments in aDNA research enabled the application of whole-genome sequencing approaches to a variety of aDNA sources such as soft tissues, bones, teeth (Margaryan et al., 2018), hair (Gilbert et al., 2008), ear ossicles (Sirak et al., 2020), petrous bone (Gamba et al., 2014), plants (Goloubinoff et al.,1993, Suyama et al., 1996, Jaenicke-Despres et al., 2003, Li et al., 2011, Ramos-Madrigal et al., 2019), or more surprisingly sediments (Willerslev et al., 2003, Willerslev et al., 2004, Zavala et al., 2021), calcified dental plaque (Adler et al., 2013, Warinner et al., 2014), and coprolites (Kohn and Wayne, 1997, Poinar et al., 2003, Bon et al., 2012, Slon et al., 2017, Qvarnström et al., 2021). Identifying body parts that provide optimal contexts for DNA preservation, namely hair,

tooth cementum, and the temporal bone's petrous portion, boosted the efficiency of shotgun sequencing from palaeontological and archaeological material (Damgaard et al., 2015, Gamba et al., 2014). These findings opened up for future studies at the population scale.

1.k Most recent game changers

With the growing amount of ancient DNA data available, our current prospect of genetic variation is not anymore limited to a glimpse of the diversity of present-day populations worldwide. Instead, temporal genomic information is continually enriched and allows us to track changes in the ancestries of human, animal and plant populations (Orlando et al., 2021). More than three decades later, the time depth of ancient DNA analyses has long surpassed the limit of 100,000 years. With the advent of HTS platforms, the recovery of specimens that broke the Middle Pleistocene time barrier (125-781 kya) is now possible. For example, with the recovery of hominin individuals dated to approximately 430,000 years ago (Meyer et al., 2016), permafrost bones from a 780,000 years old horse (Orlando et al., 2013) and three mammoth molars from Siberia, two of which are more than one million years old (Van der Valk et al., 2021). Analysing ancient genomes from the Early and Middle Pleistocene epochs would help emphasise the potential of deep-time Paleogenomics and understand the complex nature of evolutionary change and speciation events.

The last crucial developments in aDNA research occurred throughout the previous decade mostly, combining molecular and in silico improvements. First, with the elaboration of new extraction and library preparation in pressurised laboratories tailored to ancient DNA (Gamba et al., 2016, Boessenkool et al., 2016, Wales et al., 2015, Gansauge et al., 2017, 2020, Rohland et al., 2015, 2018). Second, enzymatic treatment of chemical DNA damage to retrieve endogenous sequences even from highly degraded material (Briggs et al., 2010, Rohland et al., 2015). Third, target enrichment methods (e.g. hyRAD capture) based on DNA or RNA probes designed to capture the whole mitochondrial genome and millions of informative SNPs along the genome for minimal cost and sequencing effort (Fu et al., 2013, 2016, Cruz-Dávalos et al., 2017, Suchan et al. 2021). Fourth is the development of a new sequencing capacity to sequence the successfully extracted DNA from subfossil and museum specimens (Poinar et al., 2006, Rasmussen et al., 2010). Finally, computational authentication of aDNA through patterns of DNA decay and removal of damage-related sequencing errors improved the

performance of computational analyses of extensive genomic datasets (Jónsson et al., 2013, Renaud et al., 2015, Schubert et al., 2014).

1.l Bioinformatic Challenges

Molecular and technological advances have significantly increased the yield of aDNA sequence data allowing for large-scale palaeogenomics projects. As a result, it is now expected that such projects involve processing and analysis of an extensive number of samples. Consequently, metadata documentation becomes essential to track those samples throughout the whole aDNA data production process, optimise work, and ensure quality control (Orlando et al., 2021). Specific laboratory information management systems (LIMS), such as CASCADE (Dolle et al., 2020), have been created to meet this need.

To analyse aDNA sequences, many softwares and methods are available. Depending on the nature of the research plan's objectives, the choice of the methods used can be different. Nevertheless, in most studies, multiple steps in the workflow are in common (Figure 4). These include handling raw sequence data, aligning against reference genomes, assessing authenticity, and finally, investigating error rates (Orlando et al., 2021).

The processing of raw sequencing reads into alignments is the first step in many comparative genomics analyses. This process can be automated using complete bioinformatic pipelines such as PALEOMIX (Schubert et al., 2014) and EAGER (Peltzer et al., 2016) to handle large-scale genomic datasets or conduct separately (Figure 4). For the latter, the resulting single or paired-end reads are efficiently trimmed to remove adapters or damaged positions at reads termini ("AdapterRemoval v2", Schubert et al., 2016). Then, they are collapsed into consensus sequences, filtered, and aligned using different software with different parameters depending on the type of study. Reconstructing a genome from sequencing reads can be achieved either by reference-guide mapping, consisting of the alignment of reads to a known reference genome or by "de novo" assembly, which is the assembly of short overlapped reads into longer fragments without the use of a reference genome. The most common approach in aDNA studies is the reference-guide mapping such as BWA (Li and Durbin, 2010), Bowtie2 (Langmead and Salzberg, 2012), and to a lower extent, MIA (Green et al., 2008), as it is computationally intensive. However, due to the inherent characteristics of aDNA, the application of Bowtie2 and BWA aligners with default parameters is not usually entirely

adapted, as these methods were initially developed to align short read sequences obtained from fresh DNA extracts. Therefore, specialised strategies and mapping recommendations tailored for each type of study have been identified with aDNA data or using aDNA simulators ("Gargammel", Renaud et al., 2017). These strategies can be used, for example, to improve efficiency and sensitivity (Schubert et al., 2012, Cahill et al., 2018), to mitigate the extent of reference bias in the data (Günther and Nettelblad, 2019, "Vg", Martiniano et al., 2020) or to improve the accuracy of DNA methylation maps (Poullet and Orlando., 2020, article 1).

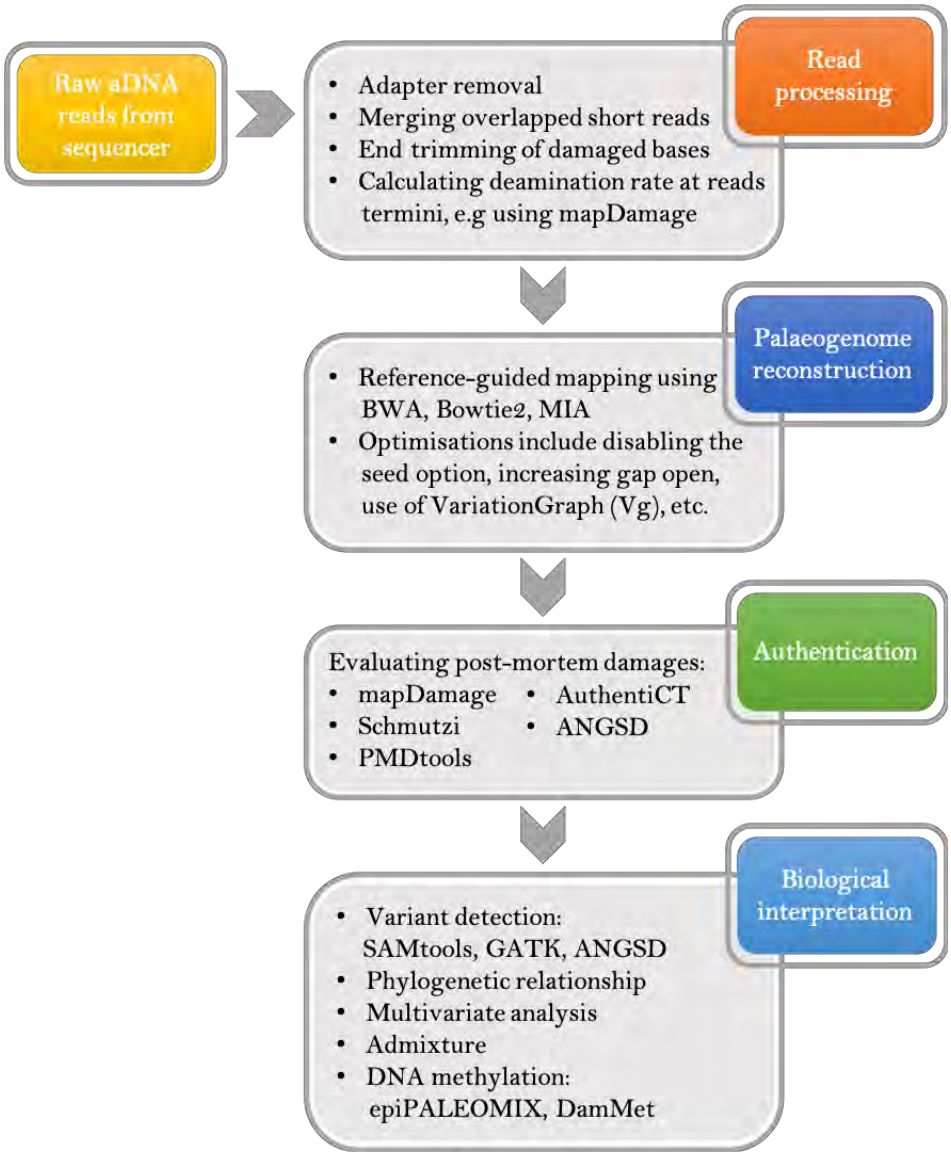


Figure 4: Bioinformatic pipeline in ancient DNA studies

Modified from Lan and Lindqvist, 2018. This pipeline shows all the different steps recommended for analysing ancient DNA data. These include handling raw sequence data from sequencer, aligning against reference genomes with specific tools, assessing authenticity, investigating error rates and, finally, interpret the data (Orlando et al., 2021).

After the reconstruction of ancient paleogenomes, the authentication of genomic DNA is required. This authentication requires evaluating post-mortem damages, mostly deaminated cytosines misincorporated toward read termini, and contamination levels found in haploid chromosomes (Krause et al., 2010, Orlando et al., 2021). Following this, softwares have been developed to calculate overall cytosine deamination frequencies at each sequencing position ("mapDamage", Jónsson et al., 2013, "PMDtools", Skoglund et al., 2014). In addition, the estimation of contaminating DNA has been calculated in haploid chromosomes by quantifying deamination patterns and fragment length distributions on the mitochondrial genome ("Schmutzi", Renaud et al., 2015) and nuclear genomes, using the single X chromosome in males ("ANGSD", Korneliussen et al., 2014). Finally, new methods are using the model of cytosine deamination, instead of the ploidy-based approaches, to evaluate the proportion of present-day DNA contamination in ancient DNA datasets generated from single-stranded DNA libraries ("AuthentiCT", Peyrègne and Peter, 2020). This method is handy for analysing any poorly preserved specimens, as long as their reference genome is available for alignment.

Various downstream bioinformatic tools are used for interpreting the vast number of datasets. These tools could be used for the detection of variants, genotypes or methylation signatures. Then, offering unprecedented opportunities to better understand the chronology and tempo of evolution at the genomic level. Unfortunately, with some exceptions, ancient genomes are sequenced at a limited depth of coverage and include an overrepresentation of post-mortem damage of aDNA sequences. This precludes the determination of accurate genotypes from standard analysis toolboxes such as SAMtools (Li et al., 2009), bedtools (Quinlan and Hall, 2010), and the Genome Analysis Toolkit ("GATK", McKenna et al., 2010). Fortunately, some recently developed programs have been optimised for substantially removing those damages and handling low and medium coverage data ("aDNA_GenoCaller", Botigué et al., 2017, "ANGSD"). Thus, making it possible to perform downstream analyses including admixture events, demographic fluctuations, genetic kinship and principal component analyses (PCA) without great effort.

Furthermore, armed with time-stamped full genomes or genome-scale sequence data, ancient DNA research can now address key evolutionary questions, considerably improving the inference power of past genomic and population evolutionary histories (Malaspinas et al., 2012, Joseph and Pe'er, 2019, Dehasque et al., 2020).

Ancient DNA is a powerful research tool for studying the past and capable of addressing diverse questions. Ancient DNA represents also the only approach for the characterisation of the genetic diversity predating domestication in some species. In addition, although the growth of the field of palaeogenomics during the past decades has been close to exponential, at present, large datasets of ancient genomes are only available for some systems, including, for example, horses (Fages et al., 2019). Indeed, the history of the horse had considerably benefited from recent aDNA developments. Starting in 1998 with the publication of a 90bp fragment from three horse subfossils (Lister et al., 1998), Y-chromosome and ancient nuclear analyses unveiled the substantial diversity that existed before domestication and explained the diffusion and origin of phenotypic traits in modern breeds (Lippold et al., 2011, Wutke et al., 2018). Furthermore, novel sequencing technologies have provided unprecedented amounts of information and have deeply revised current models of horse domestication (Schubert et al., 2014, Librado et al., 2021). As a result, horses are now the organisms in which the most extensive time series of ancient genomes have been characterised (Fages et al., 2019). It is also the organism where one of the oldest genomes has been sequenced (Orlando et al., 2013).

This first section has provided a snapshot of historical concepts, challenges and state-of-the-art methods illustrating the diversity of ancient DNA. This research area has moved forward over almost 40 years, in parallel with technological and computational advances in the life sciences, leading to tremendous breakthroughs. Such advances have also revealed that beyond genomic sequences, ancient genome-scale epigenetic information can be retrieved and characterised from subfossil material (Orlando and Willerslev, 2014). Although the identification of several epigenetic layers of gene regulation is in its early stages in modern horses, new statistical methods have started to unveil the first ancient epigenetic maps (Orlando, 2020). The second section will then focus on ancient epigenomics.

2. Ancient epigenetics

2.a Epigenetics and its definition

The field of epigenetics emerged for the first time in 1942 when Conrad Waddington explained that some phenotypic changes might not be due to simple genetic variation but from complex and dynamic interactions between the developmental environment and the genome (Waddington, 1942). Interest for epigenetics has increased rapidly over the last decade, in a diverse range of research, including cancer research (Daniel et al., 2011), developmental biology (Tobi et al., 2009), host-pathogen interactions in plants (Boyko and Kovalchuk, 2011), ecology (Burris and Baccarelli, 2014) and psychology (Zhou et al., 2014). Numerous biological and otherwise unexplained phenomena have been attributed to epigenetics, including paramutation in maize (Chandler, 2007), mental disorders (Rutten and Mill, 2009), silencing of a large fraction of transposable elements present within our genome (Slotkin and Martienssen, 2007) and imprinting of specific paternal and maternal loci in mammals (Li and Sasaki, 2011). However, despite several decades of research, epigenetics still lacks a consistent definition and supporting empirical data (Deans and Maggert, 2015). The simplest definition of epigenetic is a limited interpretation requiring the epigenetic trait to be stable, mitotically and/or meiotically heritable and to occur without changes in the DNA sequence (Berger et al., 2009). However, this definition is supposed to be used to describe phenomena that utilise molecular mechanisms to mediate a phenotype change, including histone modifications (Bell et al., 2011), DNA methylation (Plongthongkum et al., 2014), nucleosome positioning (Struhl and Segal, 2013), and non-coding RNA interactions (Chen et al., 2016). A more comprehensive explanation has been proposed, defining epigenetic modifications as: "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" (Bird 2007). This latter definition does not automatically require a feature to be heritable and comprises both temporary and permanent reprogramming of the DNA such that the resulting phenotype changes are not simply a by-product of the genome. This definition accounts for the subtle phenotypic changes in monozygotic twins during their lifetime (Li et al., 2020). It also includes environmental-related factors that can influence epigenomic regulation, such as dietary factors (Andersen et al., 2019), smoking (Kaur et al., 2019), alcohol (Portales-Casamar et al., 2016), current habitat and lifestyle (Fagny et al., 2015) and past environmental, social and climatic crises such as the Dutch Hunger Winter (Heijmans et al., 2008, Tobi et al., 2018).

Such findings have demonstrated that epigenetic analyses can deliver critical biological insights underlying divergence, speciation and extinction. Thus, there was a growing interest in the potential evolutionary role of epigenetic variation in ancient DNA studies, and the number of research mentioning epigenetic marks increased over the past few years (Orlando and Willerslev, 2014, Pedersen et al., 2014, Hanghøj et al., 2016, 2019, Gokhman et al., 2014, 2017, 2020, Wagner et al., 2020). Such a research has paved the way to look for possible epigenetic modifications in ancient individuals through massive evolutionary transitions, such as the Neolithic agricultural revolution, the Industrial Revolution or the domestication of plants and animals (Orlando and Willerslev, 2014, Ding and Chen, 2018, Seguin-Orlando et al., 2021). However, a direct test of this evolutionary role would require analysing epigenetic marks over timescales and measuring past environmental and social conditions with precision. This would be done for example by exploiting the natural degradation processes of cytosines to reconstruct ancient DNA methylation maps of extinct individuals (Gokhman et al., 2014) and then to compare loci whose DNA methylation level is environmentally responsive to infer about ancient daily life (Gokhman et al., 2017). This would also be done by comparing the epigenotype of specimens that lived before with those that lived after some major environmental phenomenon (such as domestication process, the Last Glacial Maximum or the transition from hunting and gathering to farming). To this day, lots remain to be addressed, both at the technical (some of which was covered by the two papers included within this manuscript) and the domestication level. Nonetheless, the capacity to gain knowledge about the past makes ancient epigenomes worthy of analysis.

2.b Molecular mechanisms underlying epigenetic traits

In most eukaryotes, epigenetic reprogramming can be carried out by a multitude of biological mechanisms (e.g. DNA methylation on cytosines, chromatin remodelling from a condensed state to a transcriptionally active state, long non-coding RNAs, microRNA, histone tail modifications). Within some preserved ancient DNA extracts, it is also possible to recover certain chromatin features and epigenetic regulatory marks present at the time of death (Gaffney et al., 2012, Pedersen et al., 2014, Hanghøj et al., 2016, 2019, Gokhman et al., 2017, 2020, Wagner et al., 2020, Figure 5). These include histone modifications (particularly acetylation and methylation (Kistler et al., 2017)), nucleosome positioning (Snyder et al., 2016, Hanghøj et al., 2016) and DNA methylation, with cytosine modifications (Llamas et al., 2012, Hanghøj et al., 2016).

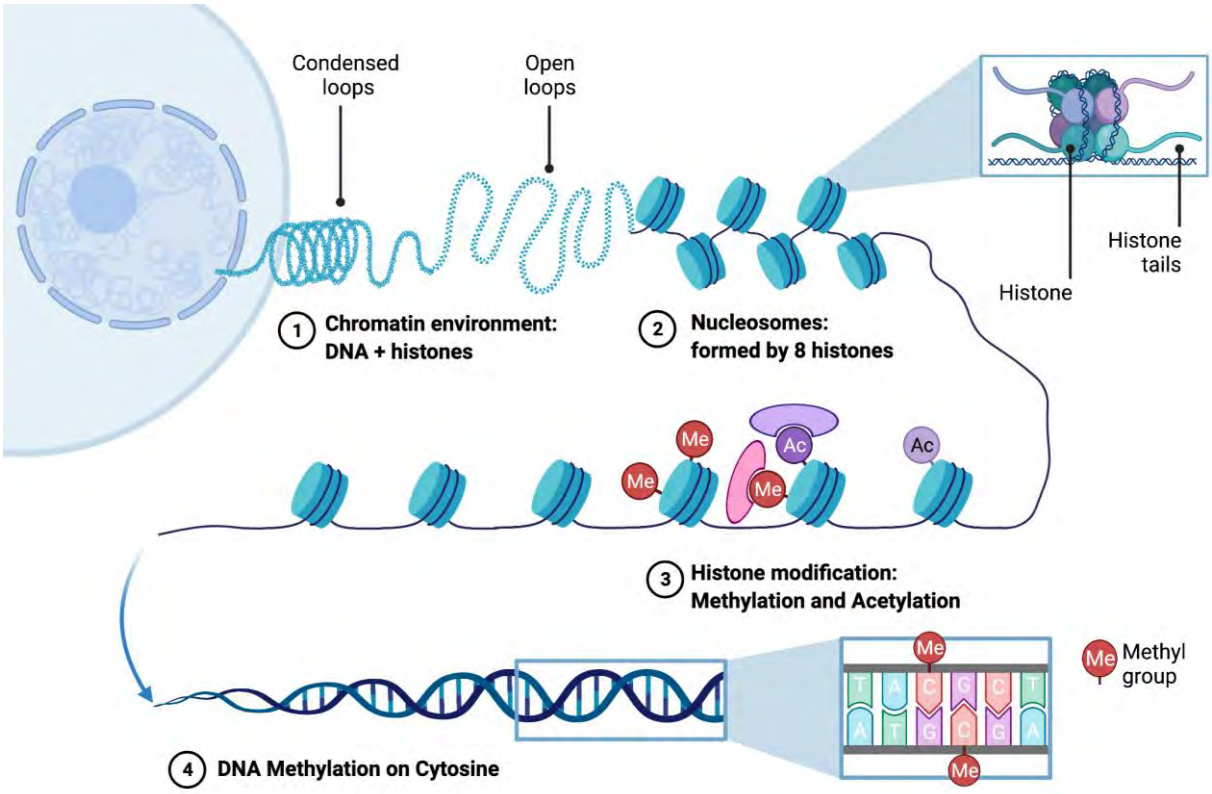


Figure 5: Epigenetic marks

Created with BioRender.com. Simplified schematic drawing of chromatin structure showing the possible mechanisms related to epigenetic factors. In ancient DNA research, nucleosome positioning (2) and DNA methylation (4) are the most studied epigenetic processes.

Among the above-mentioned list of epigenetic marks, the most studied and best understood epigenetic phenomenon in ancient DNA is DNA methylation (Figure 5) because this is the one that ancient DNA research, given its nature, can directly and easily study.

DNA methylation is a fundamental regulator of gene expression that is widespread in eukaryotes and prokaryotes. DNA methylation consists of methyl groups covalently added to the fifth carbon position of adenine or cytosine bases (Plongthongkum et al., 2014, Figure 5). These methyl groups can be added or removed dynamically but can remain stable throughout multiple cell divisions. Cytosine methylation is the most common eukaryotic DNA modification, even though the rate of cytosine DNA methylation can differ significantly between species, starting from almost none in *Caenorhabditis* to around 4% in human genomes and about 14% from leaf tissue of *Arabidopsis thaliana* (Hu et al., 2015, Capuano et al., 2014, Matsuda et al., 2018).

In mammals, DNA methylation can occur at cytosines in any genome region but is not randomly distributed (Lee et al., 2014). The vast majority of 5-methyl-cytosines (^{5m}C) residues are predominantly found in CpG dinucleotide sites in somatic cells (Weber et al., 2005). In contrast to other nucleotidic bases and dinucleotide combinations, CpG dinucleotide sites are found 4-times less than expected given the average base composition of the human genome. This lack of CpG dinucleotides is due to an excess of methylation targeting mostly cytosines. The resulting ^{5m}C is then prone to deamination and C to T mutations compromising the genomic stability of these regions in evolutionary time scales (Bird, 2002, Panchin et al., 2016). CpG dinucleotides are often found within high-density regions, such as CpG islands (CGIs), where they are often repeated (Smith et al., 2015). But, they can be found in different genomic regions, either inside or outside of CGI contexts, for example in coding regions, regulatory elements and repeat sequences. More than half of the genes, including housekeeping genes (Zhu et al., 2008, Hong et al., 2013), in vertebrate genomes contain short (around 1kb) CGIs at their promoters (Pachano et al., 2021). However, most CGIs remain unmethylated in somatic cells even though, in general, repetitive DNA sequences including transposable elements, satellite DNA and coding regions (especially exons) come heavily methylated (Li and Zhang, 2014). These methylation-resistant CGIs are ideal for housekeeping genes that are frequently expressed in several types of tissues (Hong et al., 2013). In contrast, the 5' and 3' gene-flanking regions are relatively hypomethylated when compared to intragenic regions (Lee et al., 2014). Since high DNA levels are often associated with low gene expression, promoter regions with high levels of ^{5m}C correlate inversely with gene expression and may prevent the binding of methylation-sensitive transcription factors (Lai et al., 2005). However, this model does not always apply partly due to the position of the methylated site on the transcriptional unit or also the state of the chromatin compaction and the positioning of other regulatory elements that condition genomic expression landscapes. For example, methylation near transcriptional start sites (TSS) blocks initiation, but methylation in the gene body does not stop the activity and sometimes even stimulates transcription (Jones, 2012, Liu et al., 2020).

2.c Characterising methylomes

Many technological advances have driven epigenetic science over the past decades, which allowed for the recovery of DNA methylation information based on several techniques that can distinguish ^{5m}C from cytosines (Petropoulos et al., 2016). As a result, a few initiatives,

such as the ENCODE consortium (Encyclopedia Of DNA Elements, ENCODE Project Consortium, 2012), the NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010) and the Database of CpG islands as well as the first online methylation analytical tool such as DBCAT (Kuo et al., 2011), have started to record and map functional elements such as epigenetic marks in humans. Such maps are also being characterised for other species, including mice, plants and domesticated animals (the FAANG consortium, standing for Functional Annotation of Animal Genomes, Andersson et al., 2015). Furthermore, methylation profiles can be used to estimate the cellular composition of more recent samples based on the deconvolution of cell types (Schmidt et al., 2020, 2021), to estimate with a relatively high precision the sample age with the "epigenetic clocks" procedure or to inform us about the medical situation of the sample. This would also allow us to elucidate important parameters for a multitude of palaeogenomic and anthropological investigations. This method consists of measuring the accumulation of methyl groups on a set of CpG markers (Horvath et al., 2013, Weidner et al., 2014, Pedersen et al., 2014).

To characterise the epigenomic variation present in ancient individuals and generate accurate methylation data from high-throughput assays, it is essential to become familiar with the characteristics and limits of each available approach. Thus, knowing the amount of DNA required, and the existing amount of DNA damage is important as this analysis generally requires significant amounts of starting material and minimal DNA decay (Hattori and Ushijima, 2017, Seguin-Orlando et al., 2021). Furthermore, it is also important to know the impact of technical batch effect, sex-specific signatures on DNA methylation inference and the flexibility in selecting CpG sites that we want to analyse as the local genome instability (and related hypermutability) makes the problem of confounding polymorphisms even more acute when it comes to predict ancient DNA methylation patterns. To map DNA methylation features on ancient genomes, two main experimental approaches have been applied.

The first relies on approaches frequently used on fresh DNA, such as sodium bisulfite conversion (Llamas et al., 2012, Smith et al., 2015) and immunoprecipitation (Seguin-Orlando et al., 2015a). Sodium bisulfite treatment converts unmethylated cytosines into uracil rapidly by deamination, whereas ^{5m}C is resistant to bisulfite-induced deamination (Hayatsu et al., 1970). Under normal conditions, a difference in the methylation status of CpG sites can be converted into a difference in the sequence, TpG or CpG. In such cases, the unmethylated cytosine is sequenced as thymine analogues, whereas the methylated cytosine remains

unaffected. Combined with HTS, bisulfite sequencing (BS) can give quantitative information on the ratio of methylated and unmethylated DNA molecules, and it can provide methylation maps at the single-base level or in a given region of the genome (Gehring, 2016). By using a BS protocol of a handful of loci from late Pleistocene Bison remains, a study has described the single nucleotide resolution of cytosine methylation patterns in an ancient mammalian genome (Llamas et al., 2012), providing the first evidence that ancient methylation epialleles can be identified. Using a similar methodology, Smith and colleagues have analysed 30 ancient Native Americans ranging in age from approximately 230 to 4,500 years before present (Smith et al., 2015). They managed to recover the methylation values of a single CpG within a repetitive element. They showed again that DNA preservation levels strongly impacted the accuracy of the methodology, as unmethylated cytosines are also degraded by post-mortem DNA degradation reactions in the absence of BS treatment although methylated cytosine residues deaminate at a higher rate than unmethylated cytosine residues (Smith et al., 2014). To overcome some of the limitations of the methods then available, Methylated Binding Domains (MBD) immunoprecipitation have been used to recover DNA fragments from the remains of a Palaeo-Eskimo Saqqaq individual, some mammoths, polar bears and two equine species (Seguin-Orlando et al., 2015a). Unfortunately, the authors could not retrieve a large amount of the methylome, suggesting that such procedures are inappropriate for analysing ancient specimens affected by high DNA fragmentation and deamination.

The second experimental approach to analyse ancient DNA methylation relies on patterns of DNA misincorporations. Such post-mortem patterns are omnipresent in animals and have been exploited to map ancient methylomes (Gokhman et al., 2014, Pedersen et al., 2014, Figure 6). An alternative computational proxy method has been tested with next-generation sequencing data produced following a specific enzymatic treatment and implemented in different open-source software such as epiPALEOMIX (Hanghøj et al., 2016) and DamMet (Hanghøj et al., 2019). The method leverages the fact that cytosines residues deaminate to uracil residues, whereas methylated cytosines residues deaminate to thymine residues (Figure 6). Mixed with a robust library procedure, this redesigned methodology open access to ancient epigenetic information (Seguin-Orlando et al., 2015b).

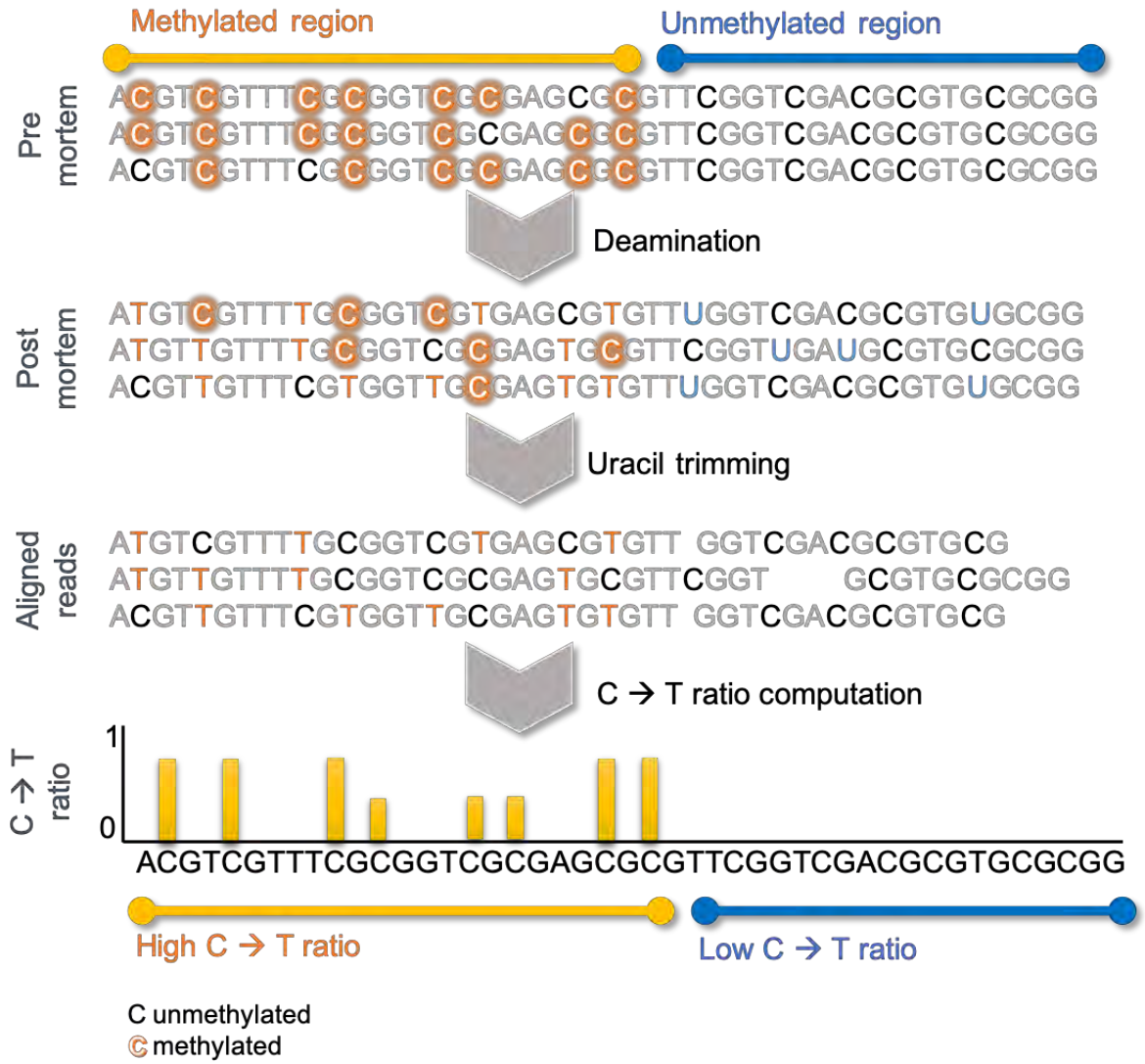


Figure 6: Reconstructing ancient methylomes

Modified from Gokhman et al., 2014. Post-mortem patterns of DNA misincorporations affect all tissues and have been used to reconstruct ancient methylation maps. The C to T ratio was measured by dividing the number of TpG, which result for the post-mortem deamination of ^{5m}CpG, by the total number of CpG dinucleotides, deaminated or not.

Among the post-mortem modifications discussed within the ancient DNA section, the most critical for the efficiency of ancient epigenetic study is DNA misincorporation of cytosine residues. Indeed, cytosine residues can be subject to a rapid hydrolytic deamination process resulting in C to T misincorporations (Briggs et al., 2007). When the deamination of cytosines is extreme, both methylated and unmethylated epialleles can give CpG to TpG conversions in the sequence, be extracts treated with sodium bisulfite or not. It then becomes practically

impossible to computationally distinguish thymine residues from deamination events, true variants and true epialleles (Jónsson et al., 2013, Gokhman et al., 2014, Hanghøj et al., 2019). Then, to remove the misincorporations of unmethylated cytosine, a wet lab protocol should be done upstream of the library construction (Briggs et al., 2007, 2010, Rasmussen et al., 2010, Seguin-Orlando et al., 2015b). The most popular protocol is the USER reagent treatment of ancient DNA extracts. This treatment removes uracils and cleaves the resulting abasic sites, thereby cutting out damage (Figure 6). Since the deamination of cytosine residues into uracils represents the most frequent post-mortem DNA damage degradation reaction (Hansen et al., 2001), this treatment drastically reduces the fraction of C to T misincorporations observed (Figure 7). Furthermore, this USER treatment is equally interesting for our subject since it does not remove methylated deaminated cytosine residues as they are converted directly to thymine residues. This property was initially used to reconstruct DNA methylation maps of both anatomically modern and hominins (Rasmussen et al., 2010, Gokhman et al., 2014) and applied to a comprehensive series of ancient samples. Additionally, procedures have been developed to display the deamination pattern of methylated cytosine only. By amplifying ancient DNA libraries, following the original Illumina sample preparation protocol with an enzyme that cannot bypass uracils, patterns of C to T transitions reflect the deamination of 5mC and can be used to infer methylation and genome-wide nucleosome occupancy maps (Seguin-Orlando et al., 2015b). Their application, however, has remained more limited.

Within this second section, we have delivered a snapshot of historical concepts, challenges and specific methods surrounding the analysis of ancient epigenetic signatures. The third section will then focus on animal domestication, one of the greatest transitions in recent human evolution. Additionally, it will also review how the field of paleo-(epi)genomics has revolutionised our understanding of animal domestication.

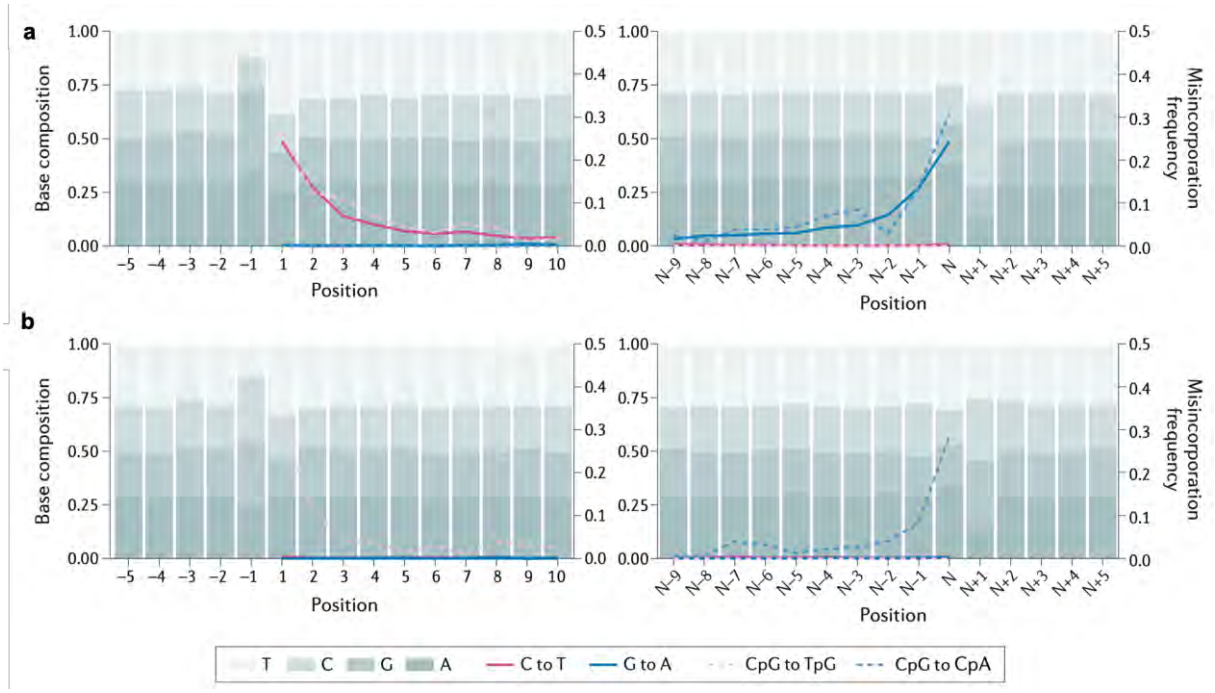


Figure 7: Deamination profiles of ancient DNA libraries

Modified from Orlando et al., 2021. Base composition and misincorporations frequency for ancient DNA libraries generated in the absence **(a)** or in the presence **(b)** of the USER reagent. Left panels correspond to read starts; right panels correspond to read ends. In the absence of USER treatment, the sequence data show an increasing excess of C to T misincorporations at read starts. In the presence of USER treatment, the excess of C to T disappears.

3. How ancient (epi)genomic studies can advance our understanding of animal domestication?

3.a Introduction to animal domestication

Over the past 15,000 years, humans have brought a wide variety of animals under domestication. Animal domestication was one of the most significant and complex transitions in human history in many aspects. It has long been an area of unresolved debate about the timing, location and number of domestication sources (Vigne, 2011, Larson et al., 2014, Figure 8). Domestic animals belong to all classes, such as mammals, birds, reptiles, fish and insects. Domestication led, first, to a shift in human subsistence patterns with a stable source of primary products such as protein-rich nutrition with milk, eggs, and meat. In a later stage, domesticated animals have gone through an extensive range of positions. Some opportunistic

ones started to feed on garbage around houses and become tolerated as household pets. Some others were venerated within rituals or religious practices or provided humans secondary products such as haulage, clothing and even warfare (Frantz et al., 2020). Between 11,000 and 4,000 years ago, following the advent of mixed-crop farming societies, massive changes have been decisive for spreading human language and culture (Zeder, 2012). Because of its importance for a wide range of disciplines, including archaeology, genetics, ecology and physical sciences (Larson et al., 2014, MacHugh et al., 2017, McHugo et al., 2019, Lahtinen et al., 2021), domestication has been studied extensively. Such studies have looked at the impact of domestication on human societies and how animal genomes and epigenomes have been reciprocally shaped by changes in human culture and technology (Larson et al., 2014, McHugo et al., 2019). In addition, the alteration and diversification of phenotypes, evident in multiple domesticated taxa, has also provided a convenient critical model to study the various pathways humans and their animal partners have followed into domestication (MacHugh et al., 2017, Figure 8). However, unlike most other common domesticated, some species, such as horses, do not seem to develop morphological characters, such as altered anatomy, to earmark domestication until long after the process began (Olsen et al., 2006, Librado et al., 2017).

The pathways that different animal species followed into domestication are varied, shaped by the biological constraints of the animals brought into domestication and the diverse cultural contexts of their human partners (MacHugh et al., 2017, Zeder, 2012). However, these domestication scenarios could be divided into three main approaches that a progressive intensification of animal-human relationships could characterise: The commensal pathway, the prey pathway and the directed pathway.

First, the commensal pathway describes a relationship initiated by wild animals, as is the case for dogs, chickens, cats and pigs to some extent. These wild animals have started to feed on household wastes or small prey around houses, and they have developed closer social bonds with their host over time (Coppinger and Coppinger, 2001, Larson et al., 2005).

The second pathway is the prey pathway. Rather than initiating a mutual relationship, these animals were first hunted for their meat. In this model, to avoid the extinction of the food supply, herd management replaced hunting. Animals such as sheep, cattle, pigs and horses are suggested to have been domesticated following this pathway (Olsen et al., 2006, Conolly et al., 2011, Larson et al., 2014). Although other pathways have also been suggested (Zeder, 2012), horses seem to best fit the prey pathway and would have been first hunted and then

progressively kept in captivity for their meat and secondary products, such as milk and, later, transport (Outram et al., 2009, Gaunitz et al., 2018).

Finally, the last pathway to domestication is a more deliberate and directed process. The domestication process following this pathway often results in direct human interest for food resources or transportation purposes. The rabbit provides a textbook example of such a pathway, although the donkey was also suggested to have been domesticated using this pathway (Rossel et al., 2008).

3.b State of the art of domestication studies

The first species likely to have followed the domestication pathway was the dog (*Canis familiaris*), and it was also the only ones domesticated by mobile hunter-gatherers at least 15,000 years ago in Eurasia (Larson et al. 2014, MacHugh et al., 2017, Lahtinen et al., 2021, Figure 8). Earlier canid remains, dating back to over 30,000 years ago (Germonpré et al., 2009) from sites in Belgium, Ukraine and Russia, were first described as putative domesticated dogs. Nevertheless, their status remains highly controversial (Perri, 2016) and thus the date and location of their first domestication. It is, however, now extensively accepted that all dogs were domesticated from an extant wolf ancestor, the grey wolf (*Canis lupus*).

For the domestication process of the dogs, two hypotheses have been proposed. First, it has occurred only in one time from a single ancient now-extinct wolf population, or possibly multiple closely related wolf populations within Europe, Central Asia or East Asia (Germonpré et al., 2009, Thalmann et al., 2013, MacHugh et al., 2017, Bergström et al., 2020). Second, it has happened in two different locations and times, as suggested by other recent studies (Frantz et al., 2016). Interestingly, the Southwest Asia region seems to be the cradle of the domestication of many animals since many studies have shown exciting findings related to this place. Sheep, cattle, goats and pigs were some of the first livestock to be domesticated in that region 10,000 to 11,000 years ago (Zeder, 2011, Larson et al., 2014, Figure 8). Even with the harsh conditions for preserving archaeological DNA in such an important region, it is essential to note that genomic information has been successfully retrieved from material held in non-Artic environments like the Levant, the Arabian Peninsula or Iran (Orlando et al., 2006).

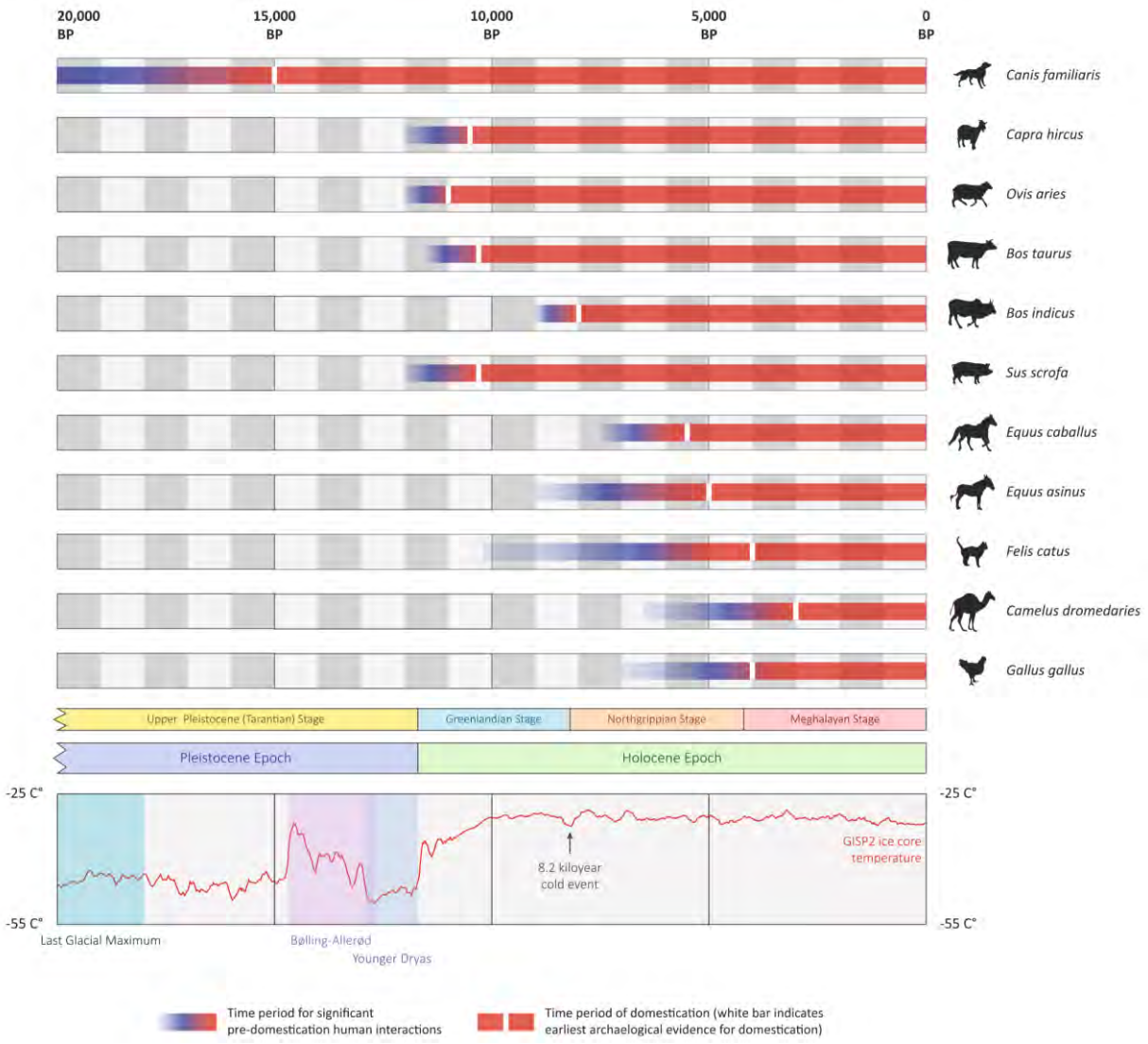


Figure 8: Chronological chart of domesticated animals

Modified from McHugo et al., 2019. Domestication timelines for 11 animal species with stratigraphy information and quaternary temperatures. For each species, the time period of domestication and pre-domestication is indicated. The first species likely to have followed the domestication pathway was the dog at least 15,000 years ago in Eurasia. Then, it was followed by livestock animals.

Following human migrations, domesticated animals then travel to Europe, involving patterns of introgression and admixture for some of them. Introgression has been seen firstly within indigenous American dogs. Later European dogs have completely replaced them with some introgression from Inuit dogs (Ameen et al., 2019). This was not the only case, since large-scale population replacement also occurred in wolves (Skoglund et al., 2015, Frantz et al., 2016), cattle (Bailey et al., 1996, Verdugo et al., 2019), horses (Fages et al., 2019) and pigs

(Frantz et al., 2019). The domestication of these animals significantly impacted human history by providing a permanent high-protein food source and other secondary products all year long. Later domesticated animals, such as equids, chickens, rabbits and camels (Figure 8), also provided communities with many resources and features, ranging from food sources, such as meat and milk, to workforce and transportation (McHugo et al., 2019).

3.c Limitations encountered with the study of ancient domesticated genomes

Paleogenomics has immense capacity to discuss any questions regarding the understanding of the genetic consequences of domestication over time. Unfortunately, there are still lots of unknowns in this area, and some major sets of questions tend to reoccur in the domestication literature: When did the domestication process start? Where? How? As the number of aDNA studies increases over the past few years, the geographic range and temporal resolution of each dataset will allow every study to investigate in more detail the early phases of domestication to understood the process from the beginning. In particular, since 2018, the number of studies using ancient DNA material with whole-genome sequence data from domesticated species exploded (Figure 9, Frantz et al., 2020).

In addition to archaeological records, one approach used to document the domestication process was genetic data obtained from modern domestic animals (Vilà et al., 2001, Larson et al., 2005). The objective is both the reconstitution of the geographical and temporal origins of the animal and the discovery of the genetic basis underlying domestic traits. However, genomic information obtained from living animals provides only a limited understanding of the long-term evolutionary process of domestication and makes contentious the validity of inferences about the past that are exclusively based on analyses of modern populations (Larson et al., 2014, Schuenemann et al., 2017, Dehasque et al., 2020, Loog, 2021).

Modern DNA can be blind, for example, to population replacement because the extinct populations made little contribution to modern genetic variation (Haak et al., 2015, Gaunitz et al., 2018). Furthermore, if the ancestral state of the species is unknown, it is more difficult for modern DNA to effectively detect admixture from an unsampled extinct species (Prüfer et al., 2014, Park et al., 2015). Fortunately, in the last decade, the study of animal domestication has been transformed by the advent of modern and ancient next-generation sequencing

platforms (Larson and Bradley, 2014, Gerbault et al., 2014), opening access to genetic information from past populations and offering the opportunity to combine the morpho-anatomical findings in archaeology through space and time with the resolution of genetic data.

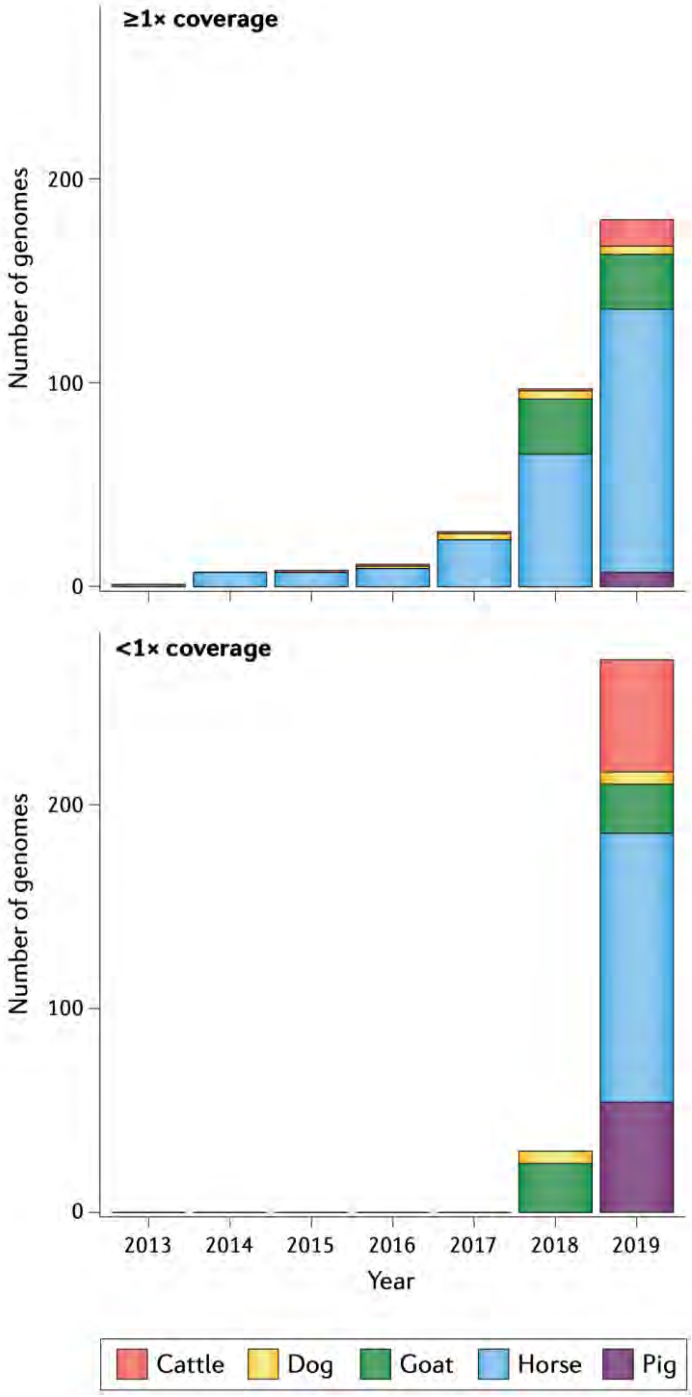


Figure 9: Ancient samples published with whole-genome sequence data

Modified from Frantz et al., 2020. Cumulative number of genomes (≥1-fold coverage) or genome-wide data (<1-fold coverage) that have been published each year since 2013 for five major domesticated species. In 2020, genomics data for 451 specimens have been reported.

Another essential but limited approach to characterise genes involved during the early phases of animal domestication is to use a large genome-wide dataset. Unfortunately, those types of datasets are more present in modern DNA than in ancient DNA, as they benefit from the relative ease of sampling and low cost of data generation compared to aDNA.

Thus, the use of aDNA data or the use of methods mixing both is recommended. Although the number of genomes sequenced and published has increased rapidly since the first publication of an ancient horse genome in 2013 (Orlando et al., 2013), many species have not yet benefitted from extensive sequencing of archaeological remains, including chickens and sheep. Thus, the number of available genomes with enough coverage is not substantial in comparison to modern species. Genomics data for 451 specimens were reported last year, including whole-genome (≥ 1-fold coverage) sequences for 180 specimens and genome-scale (< 1-fold coverage) data for 270 specimens (Frantz et al., 2020, Figure 9). There is a disparate sequencing effort among these genomics data between ancient domesticates, and a large proportion of these genomes belong to horses. With nearly 300 ancient genomes sequenced at above onefold coverage, it is indeed the species with the most complete ancient genome dataset after humans (Brunson and Reich, 2019). It led to several exciting findings, with profound implications for both standard evolutionary models of horse domestication and archaeological scenarios of past human migrations. To this day, one in-depth palaeogenomics study has unveiled convincing evidence in support of a loss of genetic diversity in modern horses and provided an example of swift changes in the horse gene pool. This study also includes recent breeding methods changes (Fages et al., 2019, Orlando, 2020) and adaptation to different environments (Librado et al., 2015). Nevertheless, the genetic, geographic and temporal origins of modern domestic horses remained unknown. Despite the availability of such high-density datasets, most of the other questions remain open and unsolved for many species.

The following section will give an overview on current understanding of horse domestication, its impact on human history and how to study this animal from an epigenetic perspective.

3.d Temporal and geographic origins of horse domestication

Among all domestic animals, horses are undoubtedly the one that has most influenced the history of humans and human settlements. The relationship between humans and equids began millennia before the start of horse husbandry. The first evidence of this link was unearthed in the site of Shönningen, in Germany, about 300,000 years ago (Lang et al., 2015). Several thousand bones, primarily horses, were excavated together with artefacts of human origin and wooden spears. Since the signs of degradation observed on the bones, such as specific cut marks, seem to be related to the human tools used at the time, they have been described as putative remains of hunted animals (Lang et al., 2015). In addition to this, many other excavations have uncovered objects dating from the Paleolithic era long before its domestication, such as fossils, portable figurines, engravings and parietal art in which the horse is the most commonly featured animal (Figure 10). Within the Vogelherd cave in Germany, researchers have found an ivory figurine representing a horse dating back to 32,000 years ago (Figure 10a), same in the Basque country where a horse statuette has been found dating back to 15,000 years ago. In others sites in Portugal (Foz Côa site, 24,000-12,000 years ago), Spain (Altamira cave, 36,000-22,000 years ago) or France (Chauvet cave, 37,000 - 28,000 years ago, Figure 10b), some horse parietal representations have been found alongside human handprints and other animals, such as aurochs, bison, deer and lions (Gonzaga et al., 2003). These paintings are widespread across western Eurasian caves, and importantly, predating the earliest evidence of horse domestication, around 5,500 years ago, by a considerable margin (Gaunitz et al., 2018). The influence of the horse did not stop there and probably strengthened throughout the Neolithic, as revealed by the collection of horse remains in different sites. Archaeologists have identified multiple sites from the Samara culture, dated to 6,000-7,000 years old, on the Pontic-Caspian steppe that included horse bones (Anthony, 2007). And yet, the temporal and geographic origin of horse domestication remains very contentious. However, based on the available number of samples at that time, several candidates for the domestication have been suggested, ranging in time from 5,500 years to 4,500 years ago and geographically spanning the steppes from central to western Eurasia (Gonzaga et al., 2004, Outram et al., 2009, 2014, Warmuth et al., 2011, Leonardi et al., 2018).

For a long time, Przewalski's horses, which was discovered in the Asian steppes, was considered as the only extant wild horse (Der Sarkissian et al., 2015). However, a genome-wide aDNA study of ancient domestic showed that Przewalski's horses are not truly wild but

the feral descendants of the first domestic horses ~5,500 years ago (Gaunitz et al. 2018). They, thus, belong to a lineage that was once successfully domesticated but further returned feral. Also, those horses suffered a complete genetic turnover by ~4,000 years ago, coinciding with the dramatic population expansion associated with the Yamnaya culture during the Early Bronze Age (Allentoft et al., 2015, Gaunitz et al., 2018). However, the exact timing of this turnover, and the geographic origin of the population, which gave rise to all modern domestic horses, remains unknown (Figure 11).

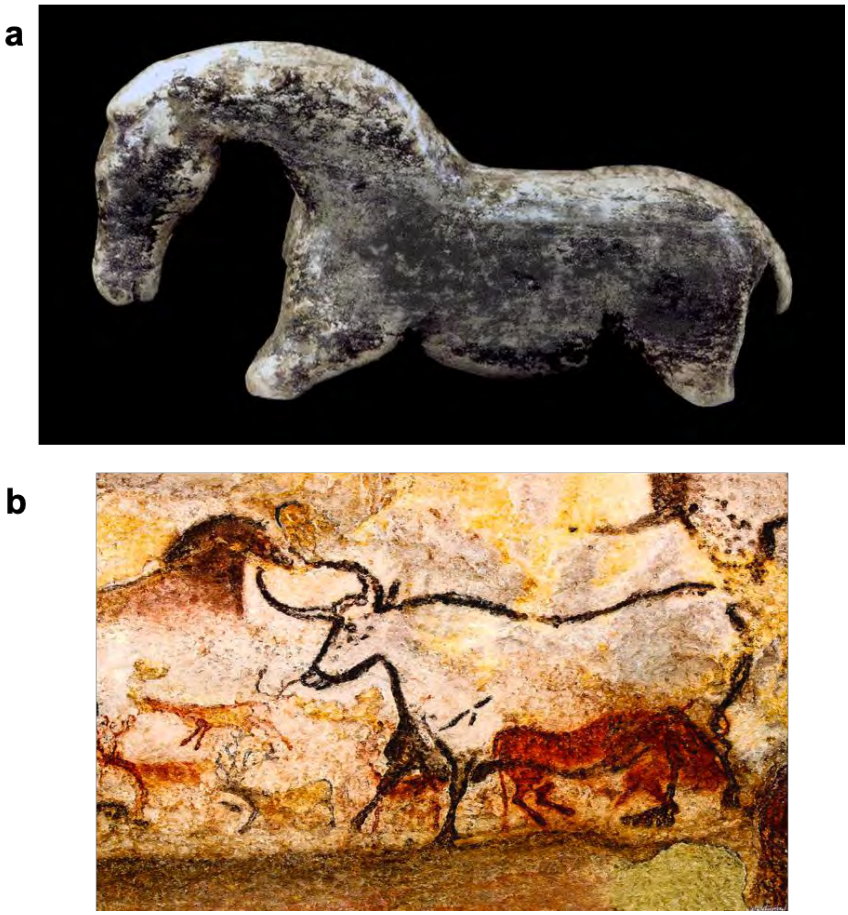


Figure 10: Statuette and parietal art representing horses

Modified from Gonzaga et al., 2003. **(a)** Sculpture of a horse carved in mammoth ivory from the Volgelherd Cave, Baden-Würtemberg, south Germany dated to 32,000-28,000 BC. **(b)** Horses, bulls and other animals painted in black. From the 'Rotunda' cave at Lascaux, France.

Potential evidence of horse rituals that belong to the 6th millennium before common era (BCE) was found in the steppes surrounding central to western Eurasia and Pontic-Caspian steppes (Outram et al., 2009, Figure 11). They correspond to horse skulls and carved bone

figurines of horses buried with humans and domestic livestock. These elements cannot be used to determine the precise date and place of horse domestication since it is impossible to know whether they represent wild or domestic individuals. However, they certainly attest to the deeply symbolic value of horses given by specific communities during the Late Neolithic, and they have been used to suggest that horses have been domesticated in this geographical area first (Outram et al., 2009).

Other potential centres of domestication include Anatolia and Iberia (Figure 11). In Anatolia, early domesticated horses were reported in the early third millennium BCE (Sherratt, 1983, Arbuckle, 2009). For Iberia, paleoclimatic reconstructions have indicated that the peninsula was covered by open landscapes instead of the rest of Europe throughout the early Holocene and thus could have been used as a refugium by wild horses (Warmuth et al., 2011, Leonardi et al., 2018). Furthermore, a study found an Iberian horse lineage and a Siberian horse that both existed during early domestication but became extinct. However, both of them did not genetically contribute to modern domestic horses (Fages et al., 2019).

Finally, another convincing evidence for horse domestication lay in the Pontic-Caspian steppes some ~5,500 years ago, in the Eneolithic culture of Botai, in present-day northern Kazakhstan (Kelekna, 2009, Figure 11). The Botai culture would have gradually transitioned from nomadic hunting to the borders of sedentary states with an economic system relying on horses. This critical change is illustrated by the large settlement with huts and comprising many animal bones where the horse represents >99% of the bone assemblages (Outram et al., 2009). These bones are not only the remains of their hunts since traces corresponding to enclosures have been found within the Botai site in addition to horse dung, skull fractures likely caused by pole-axing, tooth pathologies indicating horses were bridled, and lipid residues in ceramics, suggesting that the horses were slaughtered, possibly ridden and milked, at Botai (Outram et al., 2009, 2014). It has, however, been recently recognised that modern domestic horses do not descend from this Eneolithic Botai culture (Gaunitz et al. 2018, Fages et al., 2019). Recent studies are also questioning the legitimacy of early horse domestication at Botai (Taylor and Barrón-Ortiz, 2021) and pointing toward Western Eurasian steppes as a new centre (Librado et al., 2021).

3.e Impact of horse domestication on the dynamics of human history

The different stages of the domestication of the horse remain poorly understood from an archaeological point of view and are complex to trace from genetic data collected on current breeds. Both the scarcity of archaeological horse remains in the millennia preceding domestication, and the temporal coalescence patterns of mitochondrial haplotypes have suggested that the demographic trajectory of the horse was declining at the time of their first domestication (Gaunitz et al., 2018). It is thus likely that the human groups have developed a strategy to effectively maintain access to a perennial resource, both in the form of meat, milk and dairy products (Outram et al., 2009). Then, even though domestication probably saved the horse from extinction during the Neolithic and resulted in the emergence of a new diversity of forms and traits for domestic breeds, it also brought wild horses to extinction. The relatively limited number of equine species living today is in sharp contrast to the large number of Pleistocene species described in the fossil record (Orlando, 2020). Same for the genomic diversity available nowadays in domestic species, which has increasingly decreased following the growing influence of Oriental and Arabian bloodlines during the last thousand years (Fages et al., 2019).

Throughout the five to six millennia of horse-human domestic interactions, the domestic horse has provided many civilisations with essential primary resources and was instrumental at many levels (Kelekna, 2009). Used for its physical strength, it became the fastest distance-running quadruped on earth and was used to transport and communicate between past communities (Forrest, 2016). It has had profound consequences on the movement of people and goods and cultural exchanges, such as disseminating ideas, diffusing religions, science, and languages across Eurasia (Kelekna, 2009). Major migrations are known to document early nomadic movement across the Eurasian steppes, in addition to later military invasions of sedentary centres thanks to horse domestication (Gazagnadou, 2013). Later, the first millennium BCE was marked by the spread of cavalry and iron weaponry in both the eastern and western steppe, and with it the emergence of communities with legendary fierce mounted armies, including the Sassanids and the Assyrians in the Near East (Potts, 2007), the Scythians in Central Asia (Gnecchi-Ruscone et al., 2021), and Xiongnu horse riders in the steppes of Mongolia and China (McMiken, 1990). The first millennium BCE also saw the rise of selective breeding to produce larger horses fully capable of carrying an armed warrior with combat weapons. By the end of this millennium, war chariots and cavalry had been introduced to all

centres of civilisation across Eurasia. There were gradually considered one of the primary drivers of an army and even empires (Sidnell, 2006). Emblematic examples of equestrian conquests include the Roman Empire, which controlled the Mediterranean littoral during most of the Antiquity, Genghis Khan's Great Mongol Empire, which ruled the largest land empire in world history in the 13th century CE (Common Era), the conquest of horseless Amerindian societies with the second voyage of Columbus in 1493, and the Napoleonic wars which tore Europe apart during the 19th century CE.

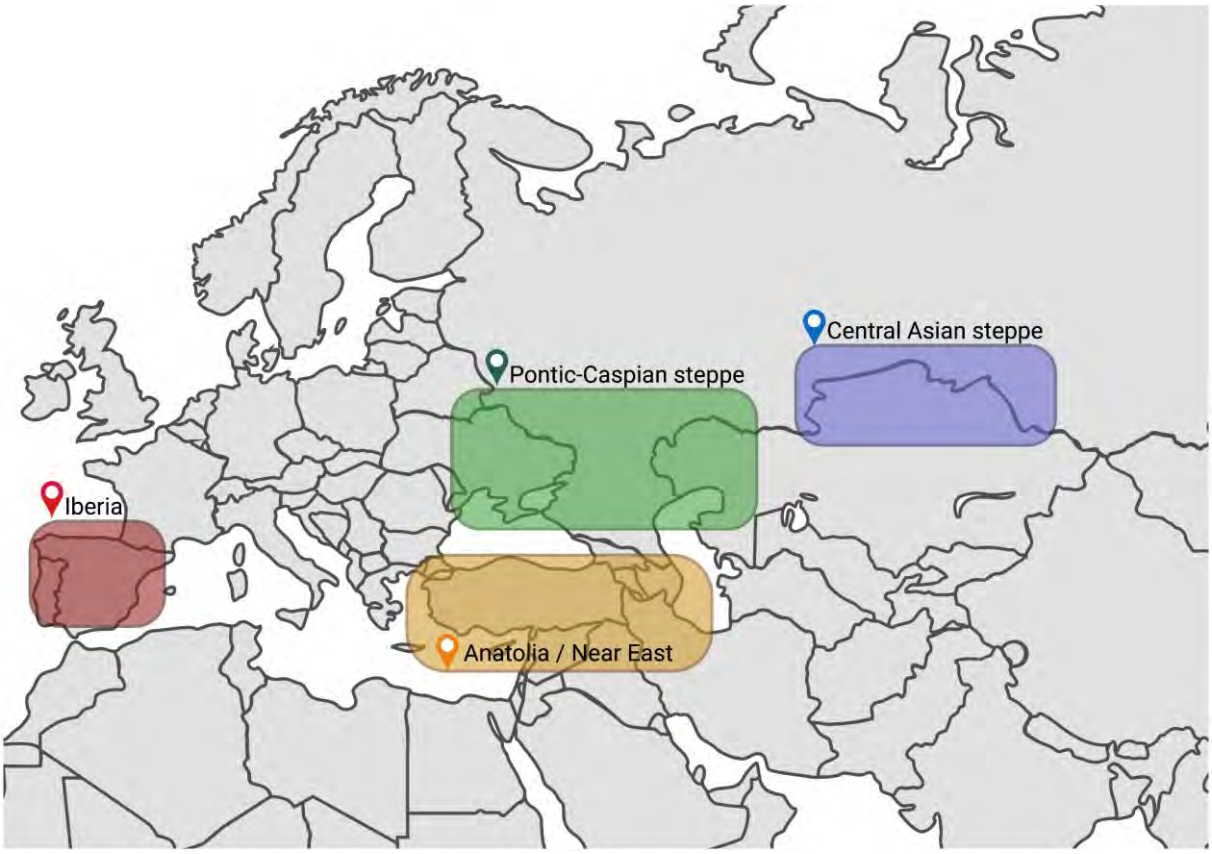


Figure 11: Putative centres of horse domestication

The exact timing of the turnover, and the geographic origin of the horse population, which gave rise to all modern domestic horses, remains unknown. Four different centres of domestication are studied.

In recent times, the influence of the horse on human societies has reduced in most parts of the world due to mechanisation and the industrial revolution. Nevertheless, it remains strong through leisure and sports with the horse-racing and breeding industry in Western Europe and Northern America (Goodwin, 2007). Moreover, horses still fulfil their traditional roles in developing countries where they provide a workforce in agriculture, transport, meat, milk and leather.

3.f Assessing ancient DNA methylation profiles underlying horse domestication

Animal domestication is a process that occurs in a short evolutionary time and leads to considerable phenotypic alterations (Zeder, 2012). However, at present, no well-known morphological markers can be used to identify wild from domestic horses, and demographic profiles are also not very useful in distinguishing which management strategies has been used. This is especially the case when dealing with an archaeological record of a mix of domesticated and wild hunted horses, such as assemblages found at Botai (Outram et al., 2009). Instead, researchers tend to find suitable indirect evidence to monitor this process: the presence of quantities of horse dung in human settlements, evidence of enclosure, lipid signatures of horse milk on pots, tooth pathologies that could be related to clamping jaw or specific cut marks on the skull (Outram et al., 2009). However, accumulating evidence shows that the phenotypic variations within domesticated animals can respond to selection and be influenced by epigenetic factors and genetic factors (Jensen, 2015). For example, acquiring an advantageous alteration for the species and which remains even in the offspring can take a very long time, about thousands of years (Jensen, 2015). This is the opposite of the rapid rate at which domestication occurs (Zeder, 2012). Additionally, variation in DNA methylation has been shown to affect various phenotypes such as coat colour and social behaviour (Gokhman et al., 2017). Furthermore, a recent study connects CpG-related mutations with speciation events in chickens (Pértille et al., 2019). As a result, it is not unexpected to see epigenetic changes proposed as an evolutionary driver of domestication (Jensen, 2015, Jablonka, 2017, Pértille et al., 2019).

The role of epigenetics in the domestication process is a promising new research area but is still in its infancy. Genome-wide methylation maps start to be reconstructed for ancient hominins (Gokhman et al., 2014, Pedersen et al., 2014) and also outside humans with equids and aurochs (Hanghøj et al., 2016). As our understanding of epigenetics improves, the ability to retrieve epigenetic from ancient DNA also become increasingly important. Thanks to the significant efforts of the FAANG consortium (Andersson et al., 2015), the characterisation of various epigenetic layers in modern horses has started and has been growing since then (Kingsley et al., 2021). However, while the DNA sequence of every somatic cell from a specific individual is nearly identical, the epigenetic layers can be pretty distinct and hard to characterise. This latter distinction is even more pronounced between two differentiated cells

of different individuals. Therefore, comparing the methylomes of ancient specimens, as recovered from two differential tissues (mostly osseous and hair material for instance), to those reconstructed from other modern tissues is tedious. Such comparison with modern tissues is necessary to reveal reliable ancient DNA methylation profiles and show differential methylation patterns in the bones of modern versus archaic horses associated with domestication.

4. Objectives of the PhD project

The first aDNA molecules were extracted more than three decades ago from a species of a sister genus to that of the horse (Higuchi et al., 1984). Since then, new milestones have been reached such as the first application of NGS to ancient remains, improving our capacity to understand the changes underlying the genetic history of the horse and its domestication. These include the characterisation of new extinct horse lineages (Fages et al., 2019), the finding of relevant insights into the divergence and admixture between pre-domestication horse populations (Librado et al., 2021) and, the detection of some early targets of selection (Librado et al., 2017). In addition, other essential game-changers such as the recovery of genomes now exceeding the symbolic one million years old barrier and, the reconstruction of the first ancient epigenome, contributed to the development of the aDNA field (Van der Valk et al., 2021, Pedersen et al., 2014). However, several research topics remain either controversial, such as the temporal and geographical origin for horse domestication or not yet studied such as the epigenomic impact of the domestication of the horse. This is partly due to sparse and fragmented fossil records belonging to the suggested time and localisation of domestication, and to the difficulty of identifying the molecular markers used such as methylated cytosines. During the three years of my PhD, I have attempted to address and some of these questions, resulting in two research papers.

The research project carried out during my PhD is of fundamental interest in understanding the importance of epigenetic mechanisms in ancient domesticated animals and was built on the latest advances in aDNA and HTS technologies. It benefited from recent developments in horse genomics, expertise and work from my colleagues that have generated the data at the Centre for Anthropobiology and Genomics of Toulouse.

This PhD project was developed around two main objectives:
The first objective was to assess the optimal sequence alignment approaches for the evaluation of ancient genomes and methylomes. This would allow the production of ancient genomic and epigenomic data of better quality for further analyses, such as the construction of DNA methylation maps. Furthermore, to construct such DNA methylation maps, USER treatment of ancient DNA extracts is necessary. However, the work carried out in this PhD revealed

some experimental and analytical limitations of this treatment, especially when combined with target-enrichment, which affect negatively some procedures.

The second objective of this project was to produce reliable estimates of ancient epigenetic marks, specifically DNA methylation maps, from ancient horses. A large dataset of ancient horse genomes was available within the laboratory. The production of such ancient methylation maps would have allowed to reveal specific epigenetic modifications associated with domestication within the horse's bone tissues. However, while helping to improve certain aspects of epigenetic inference from ancient data, this second part of my PhD project revealed significant experimental and analytical limitations that make the production of these ancient epigenetic marks particularly difficult. A number of procedures have been designed to tackle these technological and methodological obstacles.

II. Methodological improvement towards the analysis of ancient methylomes

Article 1:

Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes

Poullet*, Orlando, Frontiers, 2020

---

1. Context

A few ancient studies have used bisulfite-like and methylated binding domains methods to recover methylation data (Llamas et al., 2012, Smith et al., 2015, Seguin-Orlando et al., 2015a). However, these molecular approaches pose several challenges and limitations when used with ancient samples (Smith et al., 2014, Seguin-Orlando et al., 2015a). Alternative methods relying on statistical proxies were then developed and used on different ancient samples with relatively high coverage. They exploit primarily the transformation of cytosine residues into thymine analogues (Pedersen et al., 2014, Gokhman et al., 2014, Hanghøj et al., 2016, 2019, Wagner et al., 2020, Article 1). One of these approaches was used, for the first time, to reconstruct an ancient methylome (Pedersen et al., 2014), by using the DNA from an ancient human hair tuft generated in 2010 and sequenced to an average-fold coverage of 16X (Rasmussen et al. 2010). Furthermore, to try to increase the endogenous DNA content of the library, statistical protocols targeting post-mortem patterns of DNA degradation along the genome (Pedersen et al., 2014, Gokhman et al., 2017) and mammal-like DNA methylation signatures enriched for methylated CpG sites prior to library construction (Seguin-Orlando et al., 2015a) have been elaborated. With these protocols, access to important ancient epigenetic information, such as chromatin features and regulatory epigenetic marking, became possible (Pedersen et al., 2014, Gokhman et al., 2014, Seguin-Orlando et al., 2015b, Hanghøj et al., 2016, 2019, Wagner et al., 2020, Article 1). Such information may contribute to learning more about our evolutionary past and would allow us to elucidate some of the palaeogenomic and anthropological questions related to ancient environmental changes and cultural transitions. However, when these first methylomes were characterised, only a low number of high-quality ancient genomes had been sequenced, limiting the availability of dense and rich high-quality comparative panels outside of humans (Der Sarkissian et al., 2015). Following this, more recent studies have looked for ways to circumvent the main limitation of these methods, the lack of high-quality ancient genomes.

The identification of methylated cytosines poses a first limitation, since it is highly dependent on the level of deamination rates (Pedersen et al., 2014, Hanghøj et al., 2018). It is then, essential to improve our ability to detect these methylated cytosines accurately. Then, for example, a particular attention should be given to single nucleotide polymorphisms at CpG dinucleotides. Indeed, it would be problematic for the accuracy of the prediction of ancient DNA methylation patterns to count as DNA methylation signatures what is in fact a product

of the substitution of cytosines into thymines. Additionally, cytosine residues can be also subject to a rapid hydrolytic and post-mortem deamination process resulting in C to T misincorporations both at methylated and unmethylated CpGs (Briggs et al., 2007, 2010). However, the latter are removed by enzymatic treatment, hence, leaving C to T signatures as marks of past DNA methylation and showing post-mortem cytosine deamination rates declining faster from sequencing ends (Briggs et al., 2010, Seguin-Orlando et al., 2015a, Smith et al., 2015). Therefore, focusing on sequencing starts, where post-mortem DNA deamination is maximised (Briggs et al., 2007), should improve the ability to retrieve deamination events, thus, detecting C to T mismatches and increasing the sequence quality that is usually lower toward read ends (Pedersen et al., 2014). Another limitation is due to the limited efficacy of USER-treatment at the end of the template, which would, if not accounted for, inflates the false-positive rate of DNA methylation inference (Article 2). Furthermore, the quality of sequencing depth and the accuracy of underlying alignments play an essential role in detecting epigenetic data (Cahill et al., 2018, Article 1, Nunn et al., 2021), it is then essential to find a suitable read-mapping strategy for each type of study.

Recent computational software and methods have been released in the past few years to meet these previous needs (Gokhman et al., 2014, Hanghøj et al., 2016, 2019, Article 1). In addition, the method used to infer the methylation state is mainly based on the number of CpG to TpG misincorporations carried at CpG dinucleotide sites. In particular, this statistical method measures the number of C to T misincorporations observed in a given genomic region in an ancient genome and compared this count for regional DNA methylation levels relative to a reference genome. Therefore, this method misses several essential elements that can impact the inference of DNA methylation levels. Indeed, all observed C to T substitutions in ancient genomes, including sequencing and mapping errors, single nucleotide polymorphisms at CpG sites and deamination of unmethylated cytosine residues does not reflect true signals of methylation and should be considered (Hanghøj et al., 2016). More recently, however, a software program has been released and tackle all the previous limitations (DamMet, Hanghøj et al., 2019). Unlike the latter studies focusing only on human methylation data, DamMet can handle any read alignment file against any reference genome. This software can recover DNA methylation signatures in the presence of USER treatment, it can recover the post-mortem deamination rates of methylated and unmethylated CpG dinucleotides when sequencing data where non-USER treated and it accounts for the presence of polymorphisms avoiding the problem of recording false signals of methylation (Hanghøj et al., 2019). Furthermore, since

it has specifically been designed for ancient samples, we have used the DamMet package to carry out methylation analyses for the first part of this PhD project.

2. Rationale and Hypothesis

The aim of this study was to identify optimal mapping parameters for CpG dinucleotides, as these represent the loci for which alignments provide DNA methylation information. For this study, the read alignment tools most commonly used in ancient DNA research have been tested on sequencing data generated following USER treatment. It was at that time unknown whether the BWA and Bowtie2 aligners show similar performance in the presence or the absence of USER treatment. However, if these two aligners show different sensitivity and specificity when using this reagent, they are both likely to impact ancient DNA methylation inference. Then, rather than comparing the performance of both aligners in a range of environmental conditions, we needed instead to compare matched sequence datasets coming from the same preservation conditions and generated in the presence or in the absence of USER treatment. Furthermore, there is no specific aligner designed for the mapping of ancient DNA templates aiming at identifying epigenetic information. Then, benchmarking the performance of the read alignment tools most commonly used in ancient DNA research is essential and should help to limit both sequencing costs and sample destruction while optimising the amount of recoverable epigenetic information. This has been done on data simulated with Gargammel (Renaud et al., 2017, Jónsson et al., 2013) and real ancient DNA sequence data of Sunghir Upper Paleolithic (Gamba et al., 2014, Sikora et al., 2017), USER treated or not. These sequence data were processed with Paleomix (Schubert et al., 2014) and the two aligners were tested alternatively on different parameters. The Burrows-Wheeler Aligner has been developed to handle long and undamaged sequencing reads generated from modern extracts and using platforms with low rates of indels and sequencing errors (Li and Durbin, 2010). However, as mentioned by Schubert et al., 2012, this read alignment tools can be employed for ancient DNA samples by following some mapping guidelines which improve the identification of true ancient DNA sequences. On the contrary, Bowtie2 is developed as an ultra-fast and memory-efficient tool that can align reads between 50 to 100bp in a competitive manner but remained much less used than BWA in ancient DNA studies (Langmead et al. 2012).

Since the mapping step is mandatory in every study, combining this first methodology with other software such as DamMet and epiPALEOMIX for the following steps will have significant consequences for the nascent field of ancient epigenomics.

3. **Article 1:** Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fevo.2020.00105

Assessing DNA Sequence Alignment Methods for Characterizing Ancient Genomes and Methylomes

Marine Poullet[1] *and Ludovic Orlando*[1,2]*

[1] *Laboratoire d'Anthropobiologie et d'Imagerie de Synthèse, CNRS UMR 5288, Faculté de Médecine de Purpan, Toulouse, France,* [2] *GLOBE Institute, University of Copenhagen, Copenhagen, Denmark*

Applying high-throughput DNA sequencing technologies to the ancient DNA molecules preserved in subfossil material can provide genetic information from past individuals, populations, and communities at the genomic scale. The combination of dedicated statistical techniques and specific molecular tools aimed at reducing the impact of post-mortem DNA damage can also help recover epigenetic data from ancient individuals. However, the capacity of different sequence aligners to identify ultrashort and deaminated ancient DNA templates and their impact on the characterization of ancient methylomes remain overlooked. In this study, we use both simulated and real ancient DNA sequence data to benchmark the performance of the read alignment tools most commonly used in ancient DNA research. We identify a read alignment strategy making use of the Bowtie2 aligner that substantially reduce computational times but shows increased sensitivity relative to previous recommendations based on the BWA aligner. This strategy significantly improves the genome coverage especially when DNA templates are shorter than 90 bp, as is typically the case for ancient DNA. It also impacts on ancient DNA methylation estimates as it maximizes coverage improvement within CpG dinucleotide contexts, which hold the vast majority of DNA methylation marks in mammals. Our work contributes to improve the accuracy of DNA methylation maps and to maximize the amount of recoverable genetic information from archeological and subfossil material. As the molecular complexity of ancient DNA libraries is generally limited, the mapping strategy recommended here is essential to limit both sequencing costs and sample destruction.

Keywords: ancient DNA, DNA methylation, DNA damage, alignment, mapping, coverage, genome, methylome

OPEN ACCESS

Edited by:
Michael Knapp,
University of Otago, New Zealand

Reviewed by:
Kieren James Mitchell,
University of Adelaide, Australia
Katharina Dulias,
University of York, United Kingdom
Peter D. Heintzman,
UiT The Arctic University of Norway, Norway

***Correspondence:**
Ludovic Orlando
ludovic.orlando@univ-tlse3.fr

Specialty section:
This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution

Received: *28 November 2019*
Accepted: *31 March 2020*
Published: *06 May 2020*

Citation:
Poullet M and Orlando L (2020) Assessing DNA Sequence Alignment Methods for Characterizing Ancient Genomes and Methylomes. Front. Ecol. Evol. 8:105. doi: 10.3389/fevo.2020.00105

INTRODUCTION

The first genome from an ancient human individual was sequenced in 2010 (Rasmussen et al., 2010) and was immediately followed by the genome sequencing of a Neanderthal (Green et al., 2010) and Denisovan (Reich et al., 2010) individual, two extinct archaic hominins. Since then, hundreds of ancient genomes have been characterized across many branches of the tree of life, including humans, horses, dogs, pigs, cattle, goats, wooly mammoths, but also many human pathogens and crops such as maize, sorghum, and barley (see Marciniak and Perry, 2017 and Brunson and Reich, 2019 for reviews). Ancient genome time series have made it possible to chart migration,

admixture, and selection through space and time at unprecedented resolution. They have provided many opportunities to revisit evolutionary scenarios developed from patterns of cultural variation among archeological sites (e.g., the spread of steppe-related ancestry during the Eneolithic and early Bronze Age, see Allentoft et al., 2015; Haak et al., 2015; Damgaard et al., 2018a; Narasimhan et al., 2019; Wang et al., 2019) and from patterns of genetic variation in present-day populations (e.g., the temporal and geographic rise of lactose tolerance in western Eurasia, Mathieson et al., 2015; Ségurel and Bon, 2017).

The variation present in ancient DNA sequences does not only inform us about the genetic affinities of past individuals, populations, and species. It can also provide insights into ancient epigenetic landscapes, which play a crucial role in the regulation of gene expression (Lea et al., 2018) in response to infection (Smith et al., 2014; Pacis et al., 2015) as well as social (Laubach et al., 2019; Santos et al., 2019; Sanz et al., 2019; Snyder-Mackler et al., 2019) and environmental (Fagny et al., 2015) cues. It can, thus, help predict individual phenotypes in the past (see Pedersen et al., 2014 and Hanghøj et al., 2016 for age predictions on ancient individuals, or Gokhman et al., 2019 for morphological predictions).

Although methods have been developed to infer nucleosome maps in ancient tissues (Pedersen et al., 2014; Hanghøj et al., 2016), most of ancient epigenetic work thus far has focused on detecting DNA methylation within CpG dinucleotide (CpG) contexts. While molecular tools such as bisulfite sequencing (Llamas et al., 2012; Smith et al., 2015) or immunoprecipitation (Seguin-Orlando et al., 2015) have been used, genome-wide DNA methylation maps have been mostly produced through statistical inference leveraging the differential sequence footprint of post-mortem DNA damage at methylated and unmethylated sites, in particular at CpGs (see Hanghøj and Orlando, 2018 for a review). When molecular tools are used to prevent the sequencing of those unmethylated CpGs that have been degraded into UpGs, ancient methylated CpGs can indeed be revealed through CpG→TpG mis-incorporations in the sequence data (see "Materials and Methods"). Most recent methodologies have been proposed to mitigate the impact of evolutionary divergence and/or sequence variation at CpG sites on the calculation of DNA methylation scores (Hanghøj et al., 2019).

High-quality DNA sequence alignments against a reference genome are essential to make accurate predictions of past genetic and epigenetic variation. Yet, the vast majority of ancient DNA studies make use of read aligner software that were developed for mapping short read sequences produced from rather long DNA molecules extracted from fresh tissues. They are, thus, not optimized for the ultra-short and degraded nature of ancient DNA templates (Dabney et al., 2013). Several studies have contrasted a range of mapping conditions to identify those most specific and most sensitive (e.g., Schubert et al., 2012; Cahill et al., 2018) or to mitigate the extent of reference bias (Günther and Nettelblad, 2019; Martiniano et al., 2019). Yet, the sensitivity and specificity of other read aligners for ancient DNA data, such as Bowtie2 (Langmead and Salzberg, 2012), as well as the impact of different read alignment strategies on ancient methylation inference, remain untested. However, since the latter relies on patterns of CpG→TpG mis-incorporations introduced in the genome sequence data by post-mortem DNA damage, the read-to-reference edit distance is expected to be increased at methylated sites. This may affect the alignment sensitivity at such sites and, in turn, impact the accuracy of DNA methylation inference for ancient individuals. It is, thus, essential to investigate the possible sensitivity of read alignment methods at CpG dinucleotides so as to not underestimate DNA methylation levels along the genome and to accurately identify differentially methylated regions between individuals showing different levels of post-mortem DNA damage.

In this study, we assess the performance of 11 read alignment strategies for mapping ancient DNA sequence data against reference genomes and their impact on the inference of ancient DNA methylation. Our main purpose is not to carry out an exhaustive investigation about the impact of mapping parameters on an entire array of sequence data reflecting various post-mortem DNA decay conditions (for such studies, please refer to, e.g., Schubert et al., 2012; Cahill et al., 2018; Renaud et al., 2018). We instead focus on identifying those parameters and factors with potential impact on DNA methylation, using publicly available sequence data carefully selected from the literature to have been generated both in the absence and presence of USER treatment of the same ancient DNA extracts. The latter currently provide the best source of information to estimate CpG methylation level from patterns of CpG→TpG mis-incorporations (see Hanghøj and Orlando, 2018 for a review). Overall, we uncover that the end-to-end alignment mode of Bowtie2 shows better performance than all other commonly used alternatives. Simulated and real ancient human DNA sequence data reveal that the coverage can be increased by up to 2.1–9.4% for a given sequencing effort. The gain in recovered read alignments is particularly important within CpGs and significantly impacts the inference of regional DNA methylation levels. Applying such alignment procedures thus, improves both the quality of the genome and epigenetic data produced from ancient individuals and extinct species.

MATERIALS AND METHODS

Ancient DNA Sequence Datasets

Previously published raw sequence data from four ancient human individuals were downloaded from the European Nucleotide Archive (**Table 1**). For three of the four ancient humans (SI, SIII, and SIV, Sikora et al., 2017), Illumina DNA sequences were generated for libraries prepared with and without treatment with the USER enzymatic mix (Rohland et al., 2015). The USER treatment makes use of a first enzymatic activity, the uracil DNA glycosylase, to eliminate uracil residues (U) accumulated in ancient DNA templates due to post-mortem deamination of cytosine residues (C). This leaves abasic sites as targets for a second enzymatic reaction in which the Endonuclease VIII cleaves the DNA backbone 3′ of the abasic site. As a result, the fraction of DNA library templates containing U residues is reduced, which limits the number of C→T

TABLE 1 | Sample and sequence information.

Name	Experimental conditions	Age (years ago)	Location	Bone	Mean fragment length	Read pairs	Libraries	ENA accession number	Publication
Sunghir SI	USER+	33,875–31,770	Russia	Molar root	50.636	15,069,820	SI_388_USER_14_CGTATA	PRJEB22592	Sikora et al., 2017
Sunghir SI	USER−	33,875–31,770	Russia	Molar root	55.162	20,168,236	SI_388_NOT_USER_13_CTATCA	PRJEB22592	Sikora et al., 2017
Sunghir SIII	USER+	35,154–33,031	Russia	Molar root	64.873	12,419,661	SIII_386_USER_39_CGACCT	PRJEB22592	Sikora et al., 2017
Sunghir SIII	USER−	35,154–33,031	Russia	Molar root	70.630	13,336,119	SIII_386_NOT_USER_2_CGATGT	PRJEB22592	Sikora et al., 2017
Sunghir SIV	USER+	34,485–33,499	Russia	Femur	60.055	12,029,755	SIV_392_USER_23_AGCATG	PRJEB22592	Sikora et al., 2017
Sunghir SIV	USER−	34,485–33,499	Russia	Femur	64.448	10,100,614	SIV_392_NOT_USER_10_TAGCTT	PRJEB22592	Sikora et al., 2017
NE1	USER−	5,070–5,310	Hungary	Petrous bone	69.762	63,774,886	NE1_SRR1186790	PRJNA240906	Gamba et al., 2014

The name, ages, and location of each ancient DNA specimen considered in this study are provided with respect to the original publication reporting the DNA sequence data. The numbers of read sequencing pairs considered in the analyses are provided and represent only a subset of the overall data available for download at the European Nucleotide Archive (ENA). The experimental conditions indicate whether raw ancient DNA extracts where treated (USER+) or not (USER−) with the USER enzymatic mix prior to DNA library construction. Ages are calibrated radiocarbon ages. The mean fragment length corresponds to BWA ds.

nucleotide mis-incorporations introduced during sequencing (Briggs et al., 2010). USER treatment is, however, inefficient for those C residues that were methylated but deaminated post-mortem, as the uracil DNA glycosylase shows no activity on the resulting thymine (T) residues. Therefore, C→T nucleotide mis-incorporations are mostly restricted to methylated loci in the presence of USER treatment (Pedersen et al., 2014). In these conditions, the read-to-genome edit distance can be expected to be inflated at such sites, which may affect the performance of read alignment software. In the absence of USER treatment, this effect is, however, expected to impact all C residues that were deaminated post-mortem, be methylated or not. Contrasting sequence data generated from raw or USER-treated ancient DNA extracts provided, thus, an opportunity to assess the performance of read alignment software at methylated loci. The sequence data underlying the ancient genomes of Sunghir Upper Paleolithic individuals originate from similar preservation conditions and were generated both in the presence and in the absence of USER treatment. These data thus provided us with an opportunity to assess whether mapping conditions could affect regional methylation prediction.

Data Simulation

Quantifying the sensitivity and predicted positive value of DNA alignment software requires the identification of the fraction of reads correctly mapped (true positives), those not correctly mapped (false positives), and those not mapped at all (false negatives). In order to assess those performance statistics, we simulated DNA sequence data using the human (hg19, The Genome Sequencing Consortium, 2001) reference genome and Gargammel (Renaud et al., 2017). This software returns DNA sequences of a selected size and can include sequencing errors typical of Illumina DNA sequencing instruments and, optionally, DNA mis-incorporations reflecting post-mortem DNA damage. A total of 3.3 million read pairs were simulated both in the presence and in the absence of ancient DNA damage for an entire size range of DNA templates overlapping typical ancient DNA size distributions. This included 100,000 read pairs for each size increment of one nucleotide within the 25–45 bp range, as well as 100,000 read pairs for each size increment of five nucleotides within the 45–90 bp range, and finally 100,000 read pairs for each size increment of 10 nucleotides within the 90–120 bp range. DNA damage was simulated using the DNA mis-incorporation of sample SIII produced by mapDamage2 (Jónsson et al., 2013). The alignment file that was used as input for mapDamage2 was generated by Paleomix (version 1.2.13.2, Schubert et al., 2014) using the same human reference genome as above and the default end-to-end alignment mode of Bowtie2, sensitive.

Read Processing and Alignment

Both simulated and real ancient DNA sequence data were processed using Paleomix. This automated computational pipeline carries out a number of read processing steps, including adapter trimming, pair collapsing, mapping, quality/size filtering, duplicate removal, and local realignment. Mapping was performed using both BWA (Li and Durbin, 2009) and Bowtie2 (Langmead and Salzberg, 2012), which represent the

two most commonly used read alignment software in ancient DNA research. BWA version 0.7.17 was used in this study, together with two main alignment modes (backtrack and mem). The backtrack algorithm was applied both using seed or disabling seeding with default parameters (-n 0.04), as recommended by Schubert et al. (2012) for ancient DNA data. Version 2.3.5.1 of the Bowtie2 read mapper was used, applying both the local and end-to-end alignment modes and the four sensitivity options provided (very fast, fast, sensitive, and very sensitive). Combined, this represented a total of 11 read alignment conditions. Read pairs were automatically collapsed as single reads when showing sufficient sequence overlap and the base quality was recalculated according to sequence match at those overlapping positions, following the default procedure implemented in the AdapterRemoval2 software (Schubert et al., 2016). Reads shorter than 25 bp post-trimming and/or collapsing were disregarded except for BWA mem where reads shorter than 30 bp were disregarded. Computational running times were recorded using the time bash command.

Coverage and DNA Methylation Calculations

Binary Alignment Map (BAM) read alignment files and summary files obtained from Paleomix were processed for a number of analyses. First, average depth-of-coverage was calculated disregarding alignments showing quality scores strictly lower than 30. This corresponded to the estimated endogenous coverage provided in the Paleomix summary file. Second, average depth-of-coverage estimates were calculated at CpG, CpA, CpC, and CpT dinucleotides using the coverage option of Bedtools (Quinlan and Hall, 2010), conditioning on the bed coordinates of each dinucleotide type present in the human reference genome (-d option). The coordinates were obtained using Seqkit Version 0.3.1.1 (Shen et al., 2016). Third, we repeated the previous calculations after soft-clipped bases present in the read alignments were masked using the Jvarkit Biostar84452 tool (Lindenbaum, 2015). All previous analyses were carried out on both the read and simulated DNA sequence data. For simulated data, we also estimated the sensitivity and positive predicted value of each alignment condition. The alignment sensitivity was measured by dividing the number of true-positive alignments by their sum with the number of false-negative alignments [i.e., true positives/(true positives + false negatives] (Schubert et al., 2016). The alignment positive predictive value was estimated as the fraction of all simulated reads that were correctly mapped [i.e., true positives/(true positives + false positives)] (Schubert et al., 2016). Reads were considered as true positives if they showed a minimum of 80% of their length overlapping the known genomic coordinates used for simulation. Reads were considered as false negative when not mapping and false positives otherwise. These three categories were identified using python version 2.7.5 and the pysam library (Li et al., 2009). Additionally, DNA methylation analyses were carried out using the recently developed DamMet package (version 1.0.1, Hanghøj et al., 2019), in which the fraction of DNA methylation, f, can be estimated for a given genomic region including a pre-selected number of CpG dinucleotides. In this study, we selected 22,845 regions showing a total of 100 CpG dinucleotides in the human reference genome as the amount of sequence data considered was not sufficient to retrieve genuine estimates in regions of smaller sizes (data not shown). The corresponding genomic coordinates were provided to DamMet in the form of a BED coordinate file using the -B option. The f DNA methylation values were directly retrieved for each genomic window from the DamMet output. For consistency, coverage estimates within each window were calculated using the approach described above. All plots were generated using RStudio Version 1.1.463 (RStudio Team, 2016) and the ggplot2 library (Wickham, 2016).

RESULTS

Overall Alignment Performance

BWA (Li and Durbin, 2009) represents the most common software for aligning ancient DNA data against a reference genome. Previous work has established that disabling seeding in BWA increased mapping sensitivity for ancient DNA data, owing to the presence of inflated mis-incorporation rates at read ends (Schubert et al., 2012). Additional work investigated the specificity and sensitivity of Bowtie2 for ancient DNA data (Cahill et al., 2018). The performance of both aligners has, however, not been benchmarked on ancient DNA data with the specific aim to assess their possible impact on the inference of ancient DNA methylation. We, thus, compared their overall alignment performance on previously published ancient DNA data from four ancient humans, consisting of three Upper Paleolithic individuals excavated at Sunghir (SI, SIII, and SIV) and one Neolithic individual from Hungary (NE1) (**Table 1**). This represented a total of 11 mapping conditions, including 3 for BWA (with/out seeding, and mem) and 8 for Bowtie2 (very fast, fast, sensitive, and very sensitive options for both the local and end-to-end alignment modes). Alignment performance was calculated by normalizing the genome coverage obtained in one mapping condition relative to that obtained when disabling seeding in BWA, after quality filtering and duplicate removal (**Figure 1**).

We first confirmed previous work reporting reduced BWA performance when seeding, corresponding to a loss of 0.19–0.51% coverage across all four ancient DNA sequence datasets investigated in the absence of USER treatment (**Figure 1**, USER−). BWA mem was found to show increased performance in all three Sunghir individuals, in which a gain of 1.66–3.63% coverage was obtained. However, the performance was reduced (1.25%) for the NE1 individual. This indicates that the individual features of ancient DNA datasets, which reflect different post-mortem DNA preservation conditions, can have both a positive and a negative impact on the performance of the mem alignment procedure. The same was found for the four sensitivity options (very fast, fast, sensitive, and very sensitive) of the Bowtie2 local alignment mode, in which up to 2.63% coverage could be gained and up to 1.75% could be lost depending on the procedure considered. In these conditions, the very fast sensitivity option was the only one associated with a performance drop in all four

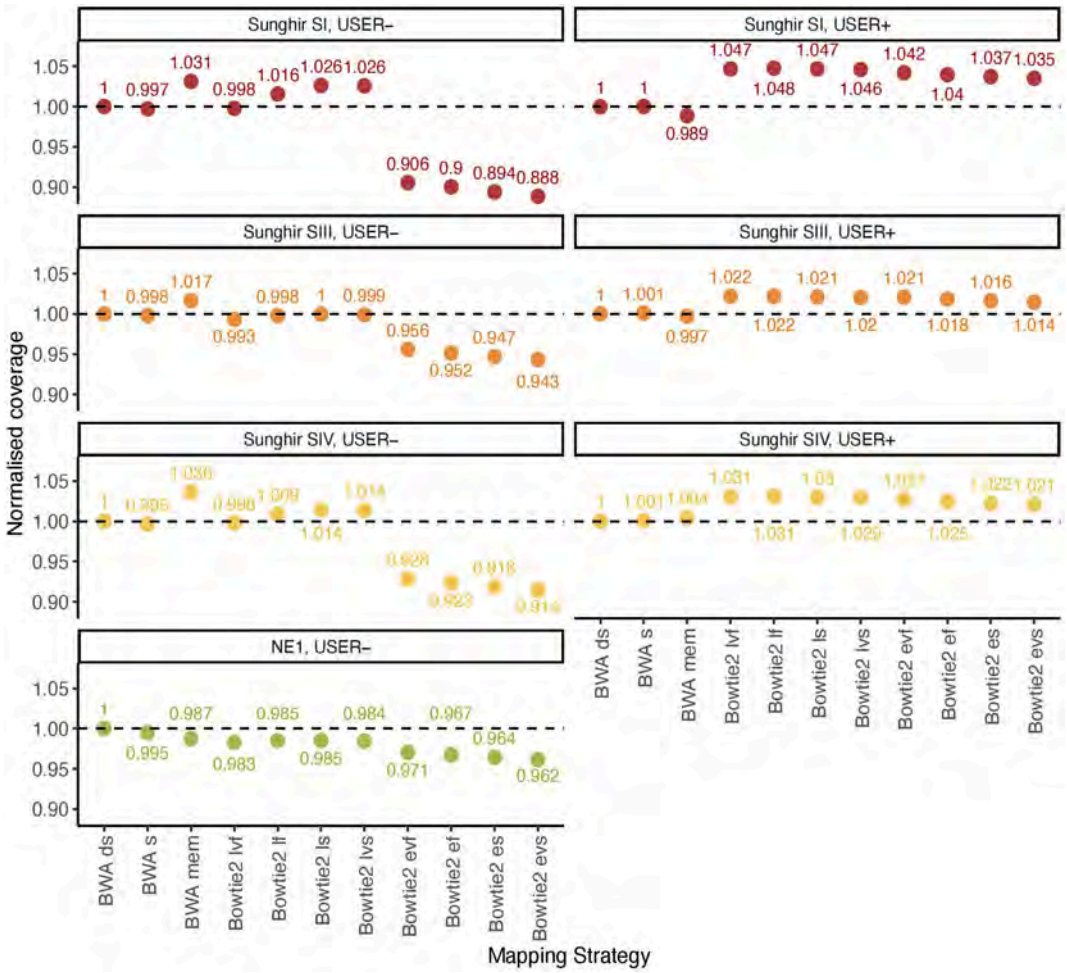


FIGURE 1 | Normalized coverage across all 11 mapping conditions investigated (real data). The average depth of coverage was estimated by filtering alignments for minimal mapping quality scores of 30 (MQ ≥ 30) and removing PCR duplicates. All coverage estimates are reported relative to those obtained with the BWA read aligner when disabling seeding (BWA ds). The other mapping conditions investigated in BWA correspond to seeding (BWA s) and mem (BWA mem). Two alignment modes (local, l; and end-to-end, e) were tested in Bowtie2, together with four options (fast, f; very fast, vf; sensitive, s; and, very sensitive, vs). The mean fragment length corresponding to BWA ds is indicated in **Table 1**.

samples tested (0.20–1.75%). In contrast to what was observed for the local alignment mode, the end-to-end mode in Bowtie2 was consistently found to show reduced performance, representing a loss of 2.93%–11.15% coverage.

We next assessed the mapping performance of ancient DNA data generated following USER treatment of raw DNA extracts (USER+). This treatment was developed to reduce the amount of DNA mis-incorporations resulting from post-mortem Cytosine deamination, which represents the most common DNA degradation reaction taking place after death (Briggs et al., 2010; Dabney et al., 2013). USER-treated ancient DNA data were available for the three Sunghir individuals but not for the NE1 individual. We found marginal coverage gain (0.01–0.11%) when disabling seeding in BWA, and different performance for the mem alignment procedure, in which a fraction of coverage could be gained (0.43%) or lost (1.12%) (**Figure 1**, USER+). Interestingly, all eight alignment modes tested for Bowtie2 were associated with increased performance, corresponding to a gain of 1.45–4.76% coverage relative to what was obtained when disabling seeding in BWA. This is in striking contrast with the reduced performance observed for the end-to-end alignment mode in the absence of USER treatment and indicates that the USER treatment modified the properties of ancient DNA data sufficiently enough to positively impact on the alignment performance.

To further gain insights into the alignment consequences of USER treatment, we simulated ancient DNA sequence data of increasing size (25–120 bp) and assessed the fraction of true positives, false positives, and false negatives obtained for each of the 11 alignment procedures tested (**Figure 2** and **Supplementary Figures S1, S2**). We found that the fraction of false-negative alignments was minimal when using the BWA aligner, except for the mem alignment mode and DNA templates of sizes inferior to 35 bp. The end-to-end alignment mode

in Bowtie2 also led to virtually no false-negative alignments across all size categories investigated, including for DNA templates of 25–26 bp in which a detectable proportion of false-negative alignments was obtained when using the local alignment mode (albeit more limited than that observed in BWA mem for larger sizes, **Figure 2A**). Applying strict mapping quality thresholds of 30 was found appropriate for eliminating all false-positive alignments obtained in all mapping conditions investigated (**Figure 2B**). Interestingly, the fraction of true-positive alignments showing mapping quality scores strictly inferior to 30 increased in BWA rather than in Bowtie2 for DNA templates of limited sizes (25–70 bp) (**Figure 2C**). This fraction increased for larger sizes when using the Bowtie2 local alignment mode (70 bp) or the Bowtie2 end-to-end alignment mode (90 bp). This indicates that the mapping quality scores returned by BWA for short size categories such as those generally observed with ancient DNA data are more conservative than those returned by Bowtie2. Moreover, when applying a strict mapping quality threshold of 30 (as commonly practiced, e.g., Sikora et al., 2017), the sensitivity (i.e., the fraction of true positives relative to both true positives and false negatives) of BWA was more limited in the mem alignment mode than when seeding or disabling seeding for short DNA templates. It, however, returned to ∼100% for size categories superior or equal to 40 bp (**Supplementary Figure S3**). Maximal sensitivity was observed in Bowtie2 using the end-to-end alignment mode (Damgaard et al., 2018b), resulting in a significant loss of true-positive alignments in BWA compared to Bowtie2 (**Supplementary Figure S3**). This effect is reversed for size categories larger than 70 bp (local mode) or 90 bp (end-to-end mode), but this is expected to minimally impact ancient DNA datasets due to the generally extensive DNA fragmentation that takes place post-mortem (**Figure 2C**).

We next tested this prediction by measuring the overall alignment performance of the 11 procedures investigated by calculating the total coverage achieved after applying a strict mapping quality threshold of 30 and removing PCR duplicates (**Figure 3**). We confirmed that Bowtie2 showed an increased performance relative to BWA for DNA templates of size inferior to 70 bp when running with the local mode and for templates of size inferior to 90 when running the end-to-end mode.

Altogether, our USER-treated read simulations revealed that across all size categories. The sensitivity of the local alignment mode was reduced for DNA templates of size strictly inferior to 38 bp, but was generally larger than that observed with BWA mem. The quality scores returned by BWA in the short size range were found to be conservative, leading to the loss of a significant fraction of true positives (9.4–10.0%) when applying strict quality thresholds (**Figure 2C**).

Alignment Performance at CpG Sites and DNA Methylation Inference

The most commonly used strategy available for estimating ancient methylation maps leverages patterns of C→T mis-incorporations at CpG dinucleotide sites as identified from BAM alignment files providing ancient DNA sequence alignment against a reference genome. We next investigated if the different mapping conditions investigated above showed different performance at CpG dinucleotide sites and could lead to different estimates of ancient DNA methylation levels. We first calculated the coverage achieved at each CpN dinucleotide context (i.e., CpA, CpC, CpG, and CpT) when applying the 11 mapping conditions to the sequencing data available for the three Sunghir individuals (**Figure 4** and **Supplementary Figure S4**). This revealed results largely consistent with those obtained when measuring coverage genome-wide, in which the Bowtie2 end-to-end mode showed the poorest performance when considering data generated in the absence of USER treatment (**Figure 4**, USER−). The performance drop was more pronounced at CpG dinucleotides. This is most likely due to the faster cytosine deamination rates reported at such sites when methylated (Seguin-Orlando et al., 2015; Smith et al., 2015), which increases the read-to-reference edit distance and, thus, limits the alignment sensitivity.

In striking contrast, Bowtie2 showed increased performance for all eight alignment conditions investigated when considering data generated following USER treatment (**Figure 4**, USER+). The performance gain was generally found to be especially pronounced within CpG dinucleotide contexts. This indicates that the USER treatment restored a fraction of reads that could not be previously aligned by reducing the read-to-reference edit distance. Since USER treatment is inefficient on those CpG dinucleotides that are methylated (Briggs et al., 2010; Pedersen et al., 2014; Hanghøj et al., 2016), we deduce that the Bowtie2 alignment conditions tested are more prone to result in gain of coverage at unmethylated CpG dinucleotides, which could have important consequences when deriving estimates of ancient DNA methylation levels.

The sensitive option of the end-to-end Bowtie2 alignment mode was found to show favorable running performance speed (**Supplementary Figure S5**). It also returned maximal sensitivity using simulated sequence data (**Supplementary Figure S3**) and maximal coverage gain at CpG dinucleotides on ancient DNA sequence data generated following USER treatment (**Figure 4**). We, thus, next compared the impact of this mapping condition on DNA methylation estimates relative to that used in all previous ancient DNA data work and consisting of disabling seeding in BWA (Gokhman et al., 2014, 2019; Pedersen et al., 2014; Hanghøj et al., 2016). To achieve this, we divided the human reference genome in windows comprising a total of 100 CpG dinucleotides and counted the number of such windows covered by at least one sequencing read in both alignment conditions. Although both alignment conditions identified read alignments in the vast majority of such genomic windows, we found that the total number of windows returning non-null coverage was larger for Bowtie2 than for BWA (**Figure 5A**), in line with the increased coverage observed with both simulated and real data with this mapper. This demonstrates that Bowtie2 retrieves data within regions for which no sequencing data could be aligned with BWA, thereby extending the genomic contexts into which DNA methylation can be estimated. Additionally, the distribution of sequencing depth obtained across genomic windows of 100 CpG dinucleotides was shifted toward larger values when applying

FIGURE 2 | Alignment performance of simulated data. A total of 100,000 reads were simulated for each size category considered in the presence of typical Illumina sequencing errors as well as nucleotide mis-incorporations remaining following USER treatment. MQ refers to the mapping quality scores of the read alignments. **(A)** Fractions of true-positive, false-positive, and false-negative alignments. **(B)** Mapping quality scores of false-positive alignments. **(C)** Mapping quality scores of true-positive alignments.

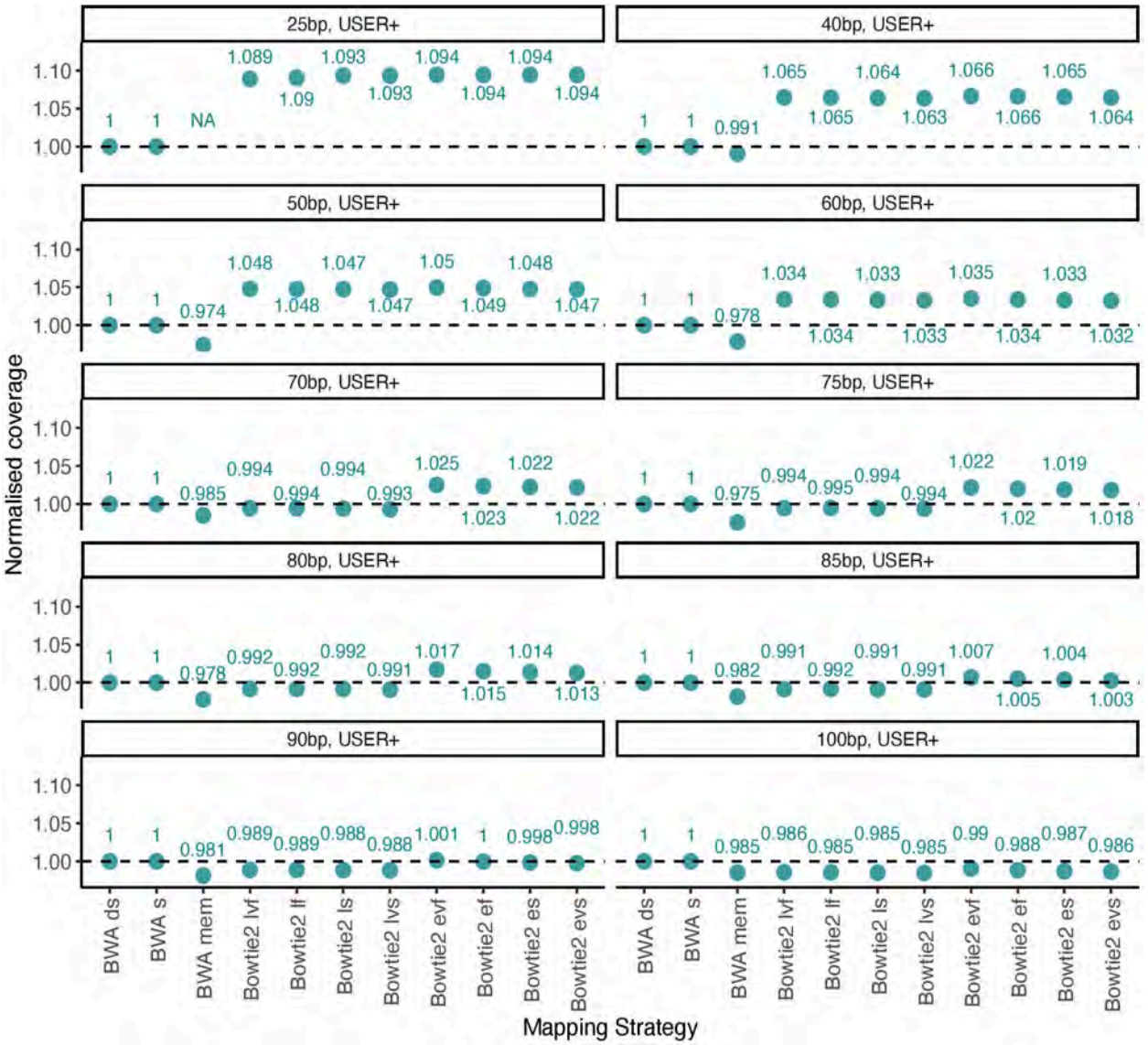


FIGURE 3 | Normalized coverage across all 11 mapping conditions investigated (simulated data). The average depth of coverage was estimated by filtering alignments for minimal mapping quality scores of 30 (MQ ≥ 30) and removing PCR duplicates. All coverage estimates are reported relative to those obtained with the BWA read aligner when disabling seeding (BWA ds). The other mapping conditions investigated in BWA correspond to seeding (BWA s) and mem (BWA mem). Two alignment modes (local, l and end-to-end, e) were tested in Bowtie2, together with four options (fast, f; very fast, vf; sensitive, s, and; very sensitive, vs). A total of 100,000 reads were simulated for each size category considered in the presence of typical Illumina sequencing errors as well as nucleotide mis-incorporations remaining following USER treatment. Each panel provides the coverage for each size category (25, 40, 50, 60, 70, 75, 80, 85, 90, and 100 bp).

Bowtie2 instead of BWA (**Figure 5B**, in which dashed lines indicate mean depth of coverage). This indicates that regional DNA methylation inference based on Bowtie2 alignments can build on more data than when based on BWA. This is important as the inference accuracy for ancient DNA methylation levels was previously shown to improve with sequencing depth (Hanghøj et al., 2019).

We next used DamMet (Hanghøj et al., 2019) to calculate in both alignment conditions the DNA methylation levels, f, for those genomic windows encompassing 100 CpG dinucleotides (**Figure 5C**). We found that the distributions of differences between the *f* values returned from Bowtie2 and BWA read alignments were centered around zero, indicating that the two mapping conditions resulted in similar regional methylation estimates. However, a fraction of the windows considered returned *f* values of one (i.e., full methylation) when using the sequence data aligned with BWA and values of zero (i.e., full demethylation) when using Bowtie2 alignments. This represented a fraction of 0.040–0.103% of the windows across the three ancient individuals investigated. Reciprocally, a fraction of the windows considered returned *f* values of zero when using the sequence data aligned with BWA and

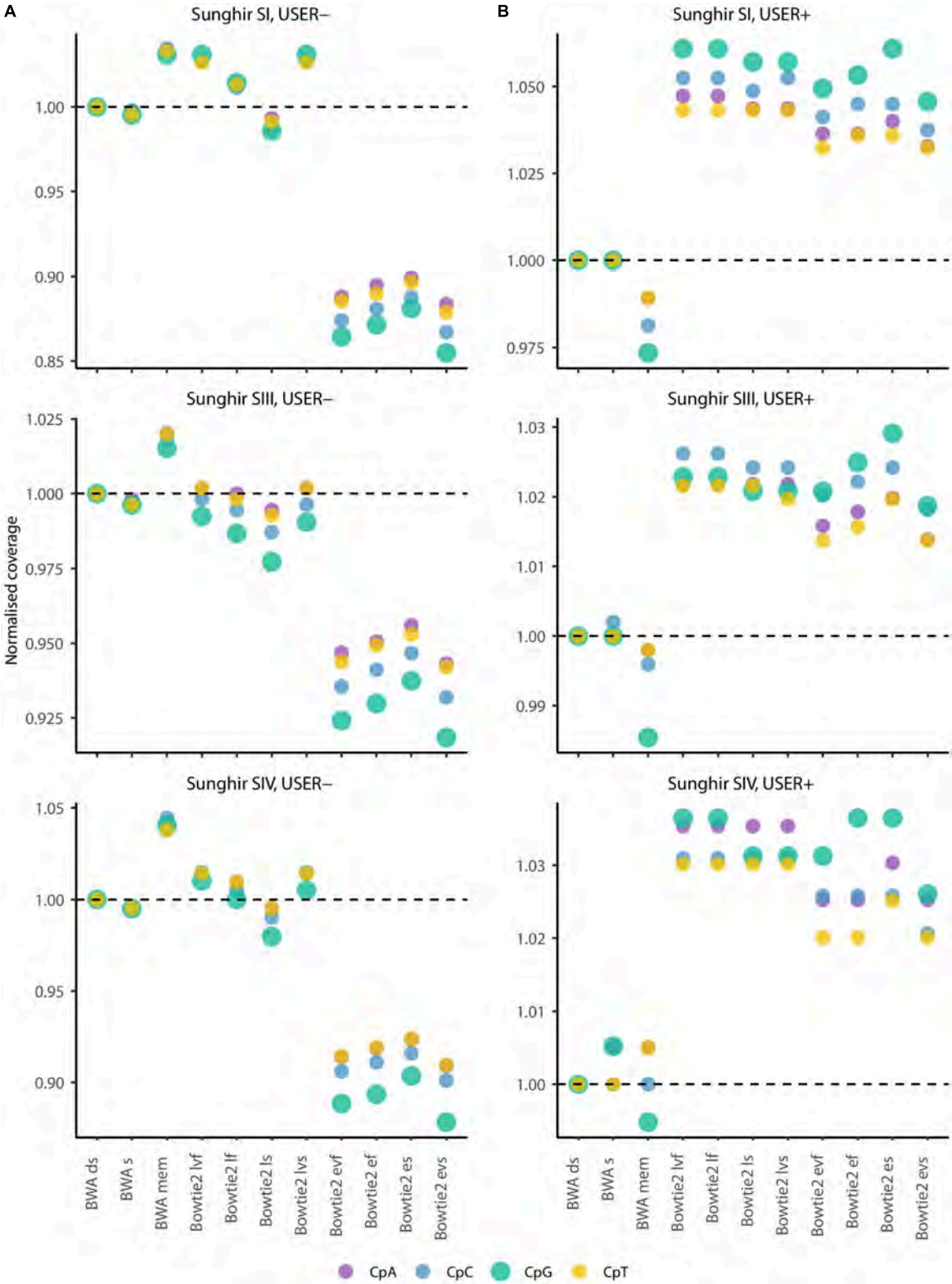


FIGURE 4 | Average depth of coverage in four dinucleotide contexts (real data). **(A)** Average depth of coverage when real data are generated in the absence of USER treatment. **(B)** Average depth of coverage when real data are generated following USER treatment. The average depth of coverage was estimated by filtering alignments for minimal mapping quality scores of 30 (MQ ≥ 30) and removing PCR duplicates. Coverage values are calculated in the dinucleotide sequence context most affected by DNA methylation (CpG), as well as the three other dinucleotides potentially affected by post-mortem cytosine deamination at the same position (i.e., CpA, CpC, and CpT). The differences observed are not due to soft-clipped bases as the values returned in the presence or not of soft-clipping masking are identical (**Supplementary Figure S4**).

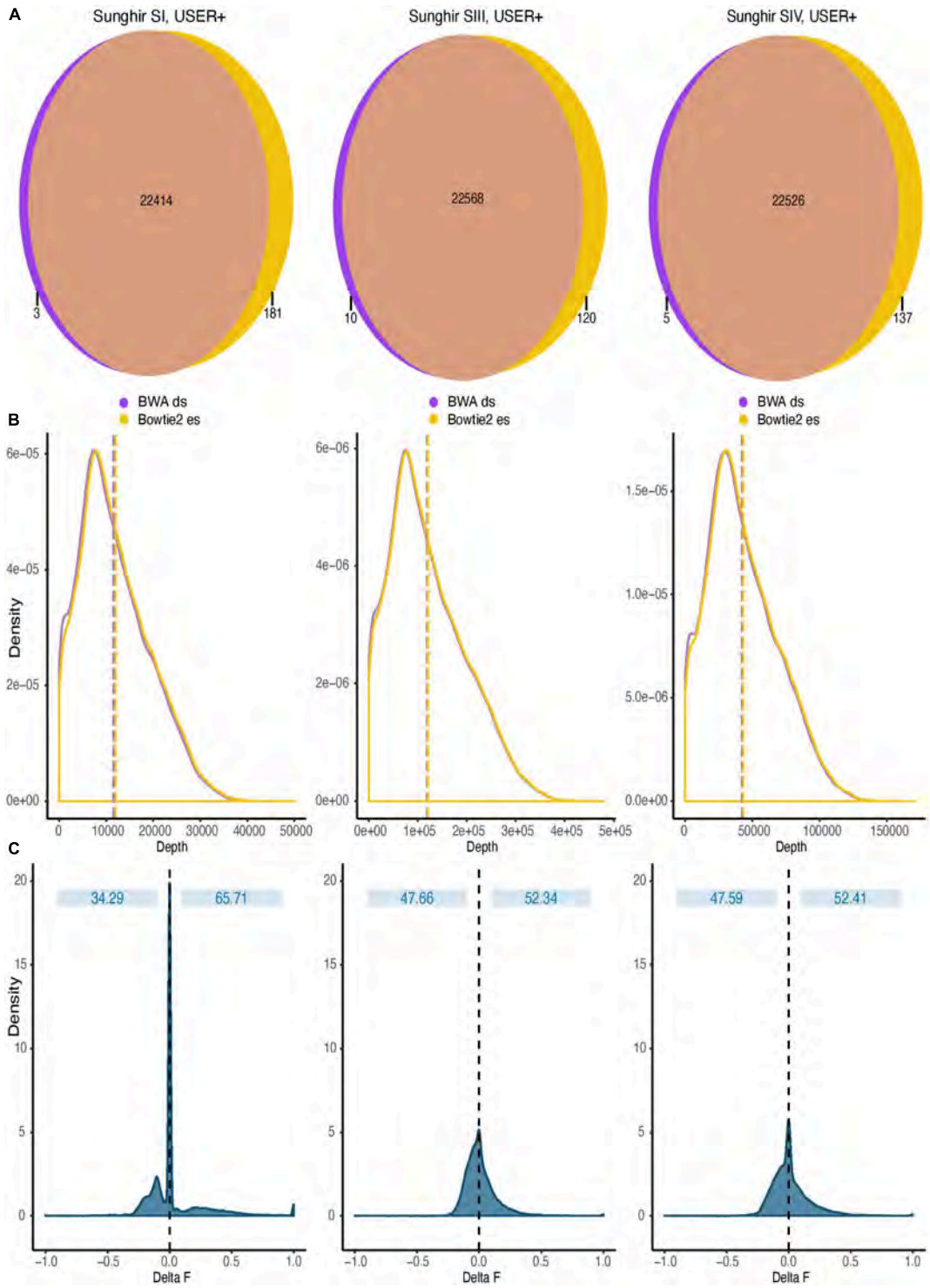


FIGURE 5 | Impact of read alignment conditions on ancient DNA methylation inference. The analyses were carried out using the sequence data generated following USER treatment of the raw DNA extracts of the three ancient Sunghir individuals (SI, left; SIII, center, and; SIV, right). The consequences of two mapping conditions on DNA methylation inference are investigated (BWA disabling seeding, BWA ds versus Bowtie2 end-to-end sensitive, Bowtie2 es). **(A)** Venn diagram of genomic windows showing non-null sequence coverage. Numbers indicate the total of genomic windows comprising 100 CpG dinucleotides and for which non-null sequence coverage was observed. Most windows are covered in both mapping conditions, but a fraction was exclusively identified by only one read aligner. Each circle is not scaled proportionally to increase readability. **(B)** Coverage distribution of genomic windows containing 100 CpG dinucleotides. Dashed lines represent the mean depth, respectively. **(C)** Distribution of the difference observed between the DNA methylation values inferred by two mapping conditions (Delta F). The difference (Delta F) reported corresponds to the difference between the *f* values estimated from Bowtie2 es alignments and those estimated from BWA ds alignments (f.Bowtie2 es - f.BWA ds).

values of one when using Bowtie2 alignments. This represented a larger fraction of the windows across the three ancient individuals investigated (0.316–1.392%), which demonstrates the increased sensitivity of the Bowtie2 aligner for those reads carrying CpG→TpG substitutions and informing on regional methylation levels. This demonstrates that the alignment procedures can significantly impact on the inference of regional DNA methylation levels.

DISCUSSION

In this study, we report that Bowtie2 shows a higher performance than BWA when aligning ancient DNA data generated following USER treatment. This effect is especially pronounced within the shorter size range (25–70 bp), due to the combined effects of a higher sensitivity for the Bowtie2 read aligner, and more conservative mapping quality scores for the BWA aligner. Moreover, in the absence of USER treatment, BWA mem was found to impact positively the coverage estimates for some samples, but negatively for others. This may be related to the respective representation of ultrashort templates among the sequencing data, as the most positive impact is found for those libraries showing the shortest average sizes. Although this remains to be tested systematically, it suggests that individual post-mortem DNA preservation conditions will significantly affect the performance of this alignment procedure. Nonetheless, the vast majority of DNA fragments retrieved from archeological and paleontological remains are of limited sizes; the mapping conditions investigated with Bowtie2 can be expected to substantially improve genome coverage estimates and, hence, data quality. Using the sequence data obtained from three Upper Paleolithic Sunghir individuals, we found that Bowtie2 could improve the genome average depth of coverage by up to 1.62–3.72%. This improvement may appear modest at first glance but represents a significant improvement for ancient DNA research as the material available for destructive DNA extraction is finite and not replaceable. Additionally, improving sequencing depth is not always possible due to the limited molecular complexity of the DNA libraries available for sequencing. Importantly, read alignment conditions were found to impact not only depth of coverage but also the inference of regional ancient DNA methylation levels.

While ancient DNA methylation has received increasing scholar attention over the last 5 years, and while several statistical inference methods have been developed (e.g., BindDB, Livyatan et al., 2015; epiPaleomix, Hanghøj et al., 2016; and DamMet, Hanghøj et al., 2019), how much different read alignment methods could impact on methylation predictions had not been investigated. This study reveals that the Bowtie2 mapping conditions recommended (sensitive option, end-to-end mode) returns with larger numbers of read alignments that increase the number of genomic windows available for inference as well as their sequence coverage, which improves accuracy of the predictions. This has important consequences for the nascent field of ancient epigenomics, which, to the best of our knowledge, based all previous predictions on BWA DNA alignments. The fact that different read alignment conditions significantly impact the inference of ancient DNA methylation levels also implies that strictly identical alignment procedures are used when comparing DNA methylation levels in different ancient remains, including when groups showing different evolutionary distances to the reference genome used for alignments are considered (e.g., archaic hominins and anatomically modern humans, Gokhman et al., 2019).

Recent work has revealed that mapping ancient, ultrashort, and damaged ancient DNA reads against a linear reference genome can introduce substantial reference bias in the data, with possible impact on downstream population genetics inference (Günther and Nettelblad, 2019). Alignment procedures including a variation graph recapitulating the known genetic variation within a panel of modern individuals further mitigated this bias and helped effectively recover non-reference variants (Martiniano et al., 2019). Future work should focus on assessing the impact of such alignment procedures on ancient DNA methylation inference.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the ENA Accession number Sunghir: PRJEB22592, ENA Accession number NE1: PRJNA240906.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LO conceived the study and provided material and infrastructure, and wrote the manuscript. MP carried out the analyses, with significant input from LO and plotted the figures.

FUNDING

This work was supported by the Initiative d'Excellence Chaires d'attractivité, Université de Toulouse (OURASI), AMADEUS (CNRS LIA), and the Villum Fonden miGENEPI research project. LO has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (grant agreement no. 681605-PEGASUS).

ACKNOWLEDGMENTS

We are grateful to all members of the AGES research group (Archaeology, Genomics, Evolution and Societies) for discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00105/full#supplementary-material

REFERENCES

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population genomics of Bronze Age Eurasia. *Nature* 522:167. doi: 10.1038/nature14507

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* 38:e87. doi: 10.1093/nar/gkp1163

Brunson, K., and Reich, D. (2019). The promise of paleogenomics beyond our own species. *Trends Genet.* 35, 319–329. doi: 10.1016/j.tig.2019.02.006

Cahill, J. A., Heintzman, P. D., Harris, K., Teasdale, M. D., Kapp, J., Soares, A. E. R., et al. (2018). Genomic evidence of widespread admixture from polar bears into brown bears during the last ice age. *Mol. Biol. Evol.* 35, 1120–1129. doi: 10.1093/molbev/msy018

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., et al. (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. doi: 10.1073/pnas.1314445110

Damgaard, P. D. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018a). 137 ancient human genomes from across the Eurasian steppes. *Nature* 557, 369–374. doi: 10.1038/s41586-018-0488-1

Damgaard, P. D. B., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., Peyrot, M., et al. (2018b). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360:eaar7711. doi: 10.1126/science.aar7711

Fagny, M., Patin, E., MacIsaac, J. L., Rotival, M., Flutre, T., Jones, M. J., et al. (2015). The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nat. Commun.* 6:10047. doi: 10.1038/ncomms10047

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* 5:5257. doi: 10.1038/ncomms6257

Gokhman, D., Lavi, E., Prüfer, K., Fraga, M. F., Riancho, J. A., Kelso, J., et al. (2014). Reconstructing the DNA Methylation maps of the neandertal and the denisovan. *Science* 344, 523–527. doi: 10.1126/science.1250368

Gokhman, D., Mishol, N., de Manuel, M., de Juan, D., Shuqrun, J., Meshorer, E., et al. (2019). Reconstructing denisovan anatomy using DNA Methylation maps. *Cell* 179, 180–192. doi: 10.1016/j.cell.2020.01.020

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the neandertal genome. *Science* 328, 710–722.

Günther, T., and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* 15:e1008302. doi: 10.1371/journal.pgen.1008302

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207. doi: 10.1038/nature14317

Hanghøj, K., and Orlando, L. (2018). "Ancient epigenomics," in *Population Genomics*, ed. O. P. Rajora (Cham: Springer).

Hanghøj, K., Renaud, G., Albrechtsen, A., and Orlando, L. (2019). DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *Gigascience* 8:giz02. doi: 10.1093/gigascience/giz025

Hanghøj, K., Seguin-Orlando, A., Schubert, M., Madsen, T., Pedersen, J. S., Willerslev, E., et al. (2016). Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Mol. Biol. Evol.* 33, 3284–3298. doi: 10.1093/molbev/msw184

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). MapDamage2.0: fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Laubach, Z. M., Faulk, C. D., Dolinoy, D. C., Montrose, L., Jones, T. R., Ray, D., et al. (2019). Early life social and ecological determinants of global DNA methylation in wild spotted hyenas. *Mol. Ecol.* 28, 3799–3812. doi: 10.1111/mec.15174

Lea, A. J., Vockley, C. M., Johnston, R. A., Del Carpio, C. A., Barreiro, L. B., Reddy, T. E., et al. (2018). Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife* 7:e37513. doi: 10.7554/eLife.37513

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lindenbaum, P. (2015). *JVarkit: Java-Based Utilities for Bioinformatics*. Available online at: https://github.com/lindenb/jvarkit (accessed July 22, 2019).

Livyatan, I., Aaronson, Y., Gokhman, D., Ashkenazi, R., and Meshorer, E. (2015). BindDB: an integrated database and webtool platform for "reverse-ChIP" epigenomic analysis. *Cell Stem Cell* 17, 647–648. doi: 10.1016/j.stem.2015.11.015

Llamas, B., Holland, M. L., Chen, K., Cropley, J. E., Cooper, A., and Suter, C. M. (2012). High-resolution analysis of cytosine Methylation in ancient DNA. *PLoS One* 7:e30226. doi: 10.1371/journal.pone.0030226

Marciniak, S., and Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* 18:659. doi: 10.1038/nrg.2017.65

Martiniano, R., Garrison, E., Jones, E. R., Manica, A., and Durbin, R. (2019). Removing reference bias in ancient DNA data analysis by mapping to a sequence variation graph. *BioRxiv* [Preprint], doi: 10.1101/782755

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528:499. doi: 10.1038/nature16152

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., et al. (2019). The formation of human populations in South and Central Asia. *Science* 365:eaat7487. doi: 10.1126/science.aat7487

Pacis, A., Tailleux, L., Morin, A. M., Lambourne, J., MacIsaac, J. L., Yotova, V., et al. (2015). Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res.* 25, 1801–1811. doi: 10.1101/gr.192005.115

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., et al. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* 24, 454–466. doi: 10.1101/gr.163592.113

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463, 757–762. doi: 10.1038/nature08835

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from denisova cave in Siberia. *Nature* 468, 1053–1060. doi: 10.1038/nature09710

Renaud, G., Hanghøj, K., Willerslev, E., and Orlando, L. (2017). Gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33, 577–579. doi: 10.1093/bioinformatics/btw670

Renaud, G., Petersen, B., Seguin-Orlando, A., Bertelsen, M. F., Waller, A., Newton, R., et al. (2018). Improved de novo genomic assembly for the domestic donkey. *Sci. Adv.* 4:eaaq0392. doi: 10.1126/sciadv.aaq0392

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil – DNA – glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20130624. doi: 10.1098/rstb.2013.0624

RStudio Team (2016). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Santos, H. P., Bhattacharya, A., Martin, E. M., Addo, K., Psioda, M., Smeester, L., et al. (2019). Epigenome-wide DNA methylation in placentas from preterm infants: association with maternal socioeconomic status. *Epigenetics* 14, 751–765. doi: 10.1080/15592294.2019.1614743

Sanz, J., Maurizio, P. L., Snyder-Mackler, N., Simons, N. D., Voyles, T., Kohn, J., et al. (2019). Social history and exposure to pathogen signals modulate social status effects on gene regulation in rhesus macaques. *Proc. Natl. Acad. Sci. U.S.A.* 201820846. doi: 10.1073/pnas.1820846116

Schubert, M., Ermini, L., Sarkissian, C. Der, Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9:1056. doi: 10.1038/nprot.2014.063

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A. S., Willerslev, E., et al. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* 13:178. doi: 10.1186/1471-2164-13-178

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Seguin-Orlando, A., Gamba, C., Sarkissian, C. Der, Ermini, L., Louvel, G., Boulygina, E., et al. (2015). Pros and cons of methylation-based enrichment methods for ancient DNA. *Sci. Rep.* 5:11826. doi: 10.1038/srep11826

Ségurel, L., and Bon, C. (2017). On the evolution of lactase persistence in humans. *Annu. Rev. Genomics Hum. Genet.* 18, 297–319.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 11:e0163962. doi: 10.1371/journal.pone.0163962

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A., et al. (2017). Ancient genomes show social and reproductive behavior of early upper paleolithic foragers. *Science* 358, 659–662. doi: 10.3389/fpsyg.2017.02247

Smith, O., Clapham, A. J., Rose, P., Liu, Y., Wang, J., and Allaby, R. G. (2014). Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. *Sci. Rep.* 4:5559. doi: 10.1038/srep05559

Smith, R. W. A., Monroe, C., and Bolnick, D. A. (2015). Detection of cytosine methylation in ancient DNA from five native american populations using bisulfite sequencing. *PLoS One* 10:e0125344. doi: 10.1371/journal.pone.0125344

Snyder-Mackler, N., Sanz, J., Kohn, J. N., Voyles, T., Pique-Regi, R., Wilson, M. E., et al. (2019). Social status alters chromatin accessibility and the gene regulatory response to glucocorticoid stimulation in rhesus macaques. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1219–1228. doi: 10.1073/pnas.1811758115

The Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., et al. (2019). Ancient human genome-wide data from a 3000-year interval in the caucasus corresponds with eco-geographic regions. *Nat. Commun.* 10:590. doi: 10.1038/s41467-018-08220-8

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, PH, declared a past collaboration with one of the author, LO, to the handling Editor.

Copyright © 2020 Poullet and Orlando. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

4. Discussion

In the first article of this thesis, some guidelines are given to characterise ancient methylomes by using the appropriate reference-guide aligner between BWA and Bowtie2. The primary purpose of this study was not to perform an extensive analysis of the impact of all the mapping parameters on a full range of ancient sequence data, showing the specific computational challenges posed by aDNA, such as extensive fragmentation and deamination features. Previous studies have arguably addressed such objectives, leading to several recommended tools and methods tailored to each type of ancient DNA study (e.g. Schubert et al., 2012, Cahill et al., 2018, Renaud et al., 2018, Martiniano et al., 2020).

More recently, different studies have also explored several mapping strategies across different read mapping software (Martiniano et al., 2020 Oliva et al., 2021) to examine the impact of reference bias and mapping artefacts in downstream population genetics inference. Due to the inherent characteristics of aDNA and the absence of specific reference-guide aligners for ancient DNA, such artefacts are inevitable. Alignment procedures including a variation graph recapitulating the known genetic variation within a panel of modern individuals (Martiniano et al., 2019) or an IUPAC-based parameterisation with the positions of all the variants available for each reference (Oliva et al., 2021) would offer high precision and low levels of bias for non-USER treated, and USER treated reads. However, some difficulties could arise when mapping ancient DNA if only a distantly related reference is available, if the reference is not available or with the presence of contaminant sequences with high sequence similarity to the target species. Most reference-guide aligners available (BWA, Bowtie2) would allow the modifications of their parameters, but this will probably decrease the sensitivity by allowing more mismatches (Taron et al., 2018).

In contrast, reference-free alignment procedures could address these limitations and even show that alignment to a reference limits biological discovery (Wang, 2021). Reference-free alignment procedures does not use or produce alignment against a reference genome but instead it is directly working from the raw fastq file. These approaches are computationally less expensive, suitable for whole genome comparisons and they do not rely on estimates regarding the evolutionary trajectories of sequence modifications (Zielezinski et al., 2017). The main mapping-free approaches are applied in many fields including, for example, phylogenetic studies (Bromberg et al., 2016), metagenomics (Meinicke, 2015), sequence assembly (Zerbino and Birney, 2008) or epigenomics (Pinello et al., 2014). For the latter,

reference-free alignment procedures have been used to investigate specific sequence features associated with epigenetic patterns in modern DNA, such as nucleosome positioning (Kaplan et al., 2009), histone modifications (Pinello et al., 2011) and DNA methylation (Zheng et al., 2013). Among the list of alignment-free methods that can define these epigenetics features, the most common method is based on k-mer or words frequencies (Pinello et al., 2014), i.e. any substring of length k that are taken in the nucleotide sequence. Future benchmarking studies should focus on characterising the impact of such alignment procedures on ancient DNA methylation inference.

Lastly, future work should also focus on assessing the impact of different read sizes on ancient DNA studies. Indeed, the most popular protocol used in ancient DNA studies (USER reagent) reduces the size of the templates by removing nucleotide misincorporations. Then, it would be interesting to wonder whether this USER treatment would decrease the alignment quality. The next article of this thesis has started to implement this.

III. Unravelling technological and methodological obstacles for characterising epigenetic marks in ancient equine specimens

Article 2:

Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA, (submitted).

Suchan, Chauvey, Poullet*, Tonasso-Calvière, Schiavinato, Clavel, Lepetz, Seguin-Orlando, Orlando.

---

1. Context

Throughout the previous decade, diverse molecular and in silico improvements have enriched the ancient DNA research with the aim to characterise the present-day biodiversity at both the community and the population levels (Dabney et al. 2013, Gamba et al. 2016, Boessenkool et al. 2016, Slon et al., 2017, Gansauge et al., 2020, Suchan et al., 2021). With the advent of high-throughput DNA sequencing (HTS) combined with new extraction methods tailored to ancient DNA, it has been possible to unlock genetic information at the genome scale from various past, and even extinct, specimens (Leonardi et al., 2017). Furthermore, the cost of such study has substantially declined since HTS have become even more effective (Goodwin et al., 2016). Among the target-enrichment method, the hybridisation method (hyRAD) based on DNA or RNA probes is normally designed to capture only a specific fraction of the genome for minimal cost and sequencing effort (Suchan et al., 2016). This method is particularly useful when analysing aDNA material, since it is often contaminated by diverse environmental sources (Suchan et al., 2021). The hyRAD methodology works in two steps. First, it uses a restriction enzyme which cut fresh DNA fragments from a closely related species. Then, the resulting DNA fragments are immortalised or captured to generate target-enrichment RNA probes (Suchan et al., 2016). This technique offers to researchers the opportunity to synthesise the probes within their own laboratories, eliminating one of the most common limitation of target-enrichment methods which is the lack of specific probes designed for their species or project. Furthermore, it helps to identify DNA present even in minute amounts which is often the case for ancient DNA materials (Slon et al., 2017), and it can characterise past molecules from calcified material, such as bones and teeth (Suchan et al., 2021). However, this hyRAD technique was only applied to a limited number of cases and once on ancient osseous material (Suchan et al., 2021).

Actually, this methodology was developed in our laboratory and was then available to assess some of the objectives of this PhD. Since it is cost effective, we used it to assess the impact of USER and capture on data recovery, with the aim of testing impact on DNA methylation inference. To reduce nucleotide misincorporations coming from post-mortem DNA damage, ancient DNA templates are treated with the USER enzymatic mix. In the first instance, this USER reagent makes use of the uracil DNA glycosylase to eliminate uracil residues that are coming from the deamination of the unmethylated cytosines (Briggs et al., 2010). Then, with a second enzyme, the Endonuclease VIII, the resulting abasic sites are cleaved. As a result, the number of post-mortem damages is reduced as the size of the ancient DNA template. This

reduction is size could be then problematic for the capture efficacy and reliability since capture technique are limited by a number of factors including the probe molecular features (Cruz-Davalos et al., 2017), the DNA adapter length (Rohland et al., 2015) and the size of the ancient DNA templates (Suchan et al., 2021). However, this enzymatic mix is used to improve ancient DNA sequence quality and is also fundamental for the prediction of ancient DNA methylation marks. Indeed, the methylated deaminated cytosine are not eliminated by the USER treatment as they are converted directly into thymine residues. Then, data produced via the USER reagent would allow us to further conduct methylation analyses.

2. **Article 2:** Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA

MOLECULAR ECOLOGY
RESOURCES

Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA

Journal:	*Molecular Ecology Resources*
Manuscript ID	Draft
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Suchan, Tomasz; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CGAT); W Szafer Institute of Botany Polish Academy of Sciences Chauvey, Loreleï; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CGAT) Poullet, Marine; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CGAT) Tonasso-Calvière, Laure; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CGAT) Schiavinato, Stéphanie; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CGAT) Clavel, Pierre; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CGAT) Clavel, Benoit; Muséum National d'Histoire Naturelle, Archéozoologie, Archéobotanique: sociétés, pratiques et environnements (AASPE) Lepetz, Sebastien; Museum National d'Histoire Naturelle, Archéozoologie, Archéobotanique: sociétés, pratiques et environnements (AASPE) Seguin-Orlando, Andaine; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT) Orlando, Ludovic; Universite Toulouse III Paul Sabatier, Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT)
Keywords:	Ancient DNA, Equids, DNA damage, DNA capture, DNA methylation, Reference bias

SCHOLARONE™
Manuscripts

Title

Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA

Tomasz Suchan[1,2*], Lorelei Chauvey[1*], Marine Poullet[1], Laure Tonasso-Calvière[1], Stéphanie Schiavinato[1], Pierre Clavel[1], Benoit Clavel[3], Sébastien Lepetz[3], Andaine Seguin-Orlando[1], Ludovic Orlando[1$]

*These authors equally contributed to this work

$Correspondence should be sent to Ludovic Orlando, ludovic.orlando@univ-tlse3.fr

[1]Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT), Université Paul Sabatier, Faculté de Médecine Purpan, Bâtiment A, 37 allées Jules Guesde, 31000 Toulouse, France

[2]W. Szafer Institute of Botany, Polish Academy of Sciences, Lubicz 46, 31-512 Kraków, Poland

[3]Archéozoologie, Archéobotanique: sociétés, pratiques et environnements (AASPE), Muséum national d'histoire naturelle, CNRS, CP 56, 55 rue Buffon 75005 Paris, France

For Review Only

Abstract

Ancient DNA preservation in subfossil specimens provides a unique opportunity to retrieve genetic information across the last 1.6 million year time range. As ancient DNA extracts are generally dominated by molecules originating from environmental microbes, capture techniques are often used to economically retrieve orthologous sequence data at the population scale. Post-mortem DNA damage, especially the deamination of cytosine residues into uraciles, also considerably inflates sequence error rates unless ancient DNA extracts are treated with the USER enzymatic cocktail prior to library construction. While both approaches have recently gained popularity in ancient DNA research, the impact of USER-treatment on capture efficacy still remains untested. In this study, we applied hyRAD capture

to eight ancient equine subfossil specimens from France (1st-17th century CE), including horses, donkeys and their first-generation mule hybrids. We found that the USER-treatment could reduce capture efficacy and introduce significant bias toward the long, least damaged and least methylated ancient DNA templates. We also recovered unbalanced proportions of donkey-specific and horse-specific alleles in mule capture sequence data, due to the combined effects of reference bias and USER-treatment. Our work demonstrates that while USER-treatment can improve the quality of ancient DNA sequence data, it can also significantly affect capture outcomes, introducing bias in the sequence data, which may affect cross-sample comparisons and estimates of ancestry contributions from highly-divergent lineages in individual samples.

Introduction

The survival of ancient DNA molecules within subfossil specimens provides a unique opportunity to gather genetic information from past, and even extinct, organisms within the last 1.6 million year time range (van der Walk et al. 2021). During the last decade, the experimental costs underlying the sequencing of ancient genomes have dramatically declined as high-throughput DNA sequencing instruments have become increasingly performant (Goodwin et al. 2016). The ultra-fragmented and degraded nature of ancient DNA, which is more difficult to manipulate than fresh DNA contaminants, represents a major technical limitation to ancient genome characterization (Dabney et al. 2013a). In fact, the vast majority of subfossil DNA extracts are dominated by large proportions of exogenous DNA molecules, mostly derived from various environmental microbial sources (Green et al. 2009). Therefore, shotgun sequencing does not generally provide a cost-effective solution, except for material such as petrosal bones (Gamba et al. 2014), ossicles (Sirak et al. 2020) and tooth cementum (Damgaard et al. 2015), which can show better DNA preservation and limited microbial DNA content.

For Review Only

Over the last years, an increasing number of molecular techniques have enriched the ancient DNA toolkit with the aim to improve DNA extraction (Dabney et al. 2013b, Gamba et al. 2016, Boessenkool et al. 2016, Korlević et al. 2019) as well as DNA library conversion (Meyer et al. 2012, Gansauge et al. 2017, Gansauge et al. 2020, Kapp et al. 2021) and immortalization (Dabney et al. 2012). Other approaches have been developed to focus sequencing efforts on endogenous fragments of interest (Carpenter et al. 2013, Enk et al. 2014, Mathieson et al. 2015), while reducing the proportion of nucleotide mis-incorporations resulting from post-mortem DNA damage (Briggs et al. 2010, Rohland et al. 2015). The latter can be achieved by treating ancient DNA extracts with the USER enzymatic cocktail, which removes those cytosine residues that were deaminated after death and would otherwise be sequenced as thymines.

Focusing sequencing efforts on target regions of interest can be achieved using various in-solution capture techniques, including both fully commercial and home-made solutions (Maricic et al. 2010, Carpenter et al. 2013, Suchan et al. 2021). The myBaits Expert Human Affinities kit from Arbor Bioscience® provides an example of commercial capture reagents that target approximately 2 million loci spread across the human genome. hyRAD capture provides an example of procedures reducing experimental costs related to probe synthesis and allowing users to prepare RNA probes covering a defined fraction of the genome using fresh DNA material obtained from extant species of interest (Suchan et al. 2016). This methodology is scalable and has been successfully used to improve the characterization of the DNA variation that was present in 4.8-7.2% of the horse genome and shared across multiple individuals dating back to beyong the radiocarbon range (Suchan et al. 2021).

Capture techniques can achieve on-target enrichment folds of several orders of magnitude (Ávila-Arcos et al. 2011, Furtwängler et al. 2020) but are limited by a number of factors, including the probe molecular features (Cruz-Davalos et al. 2017), the DNA adapter length (Rohland et al. 2015) as well as the size of the ancient DNA templates (Suchan et al.

For Review Only

2021). While USER treatment reduces the impact of post-mortem damage on the sequencing data (Rohland et al. 2015), it does not repair those cytosines that were deaminated after death but cleaves instead the DNA strand 3' of their location. By reducing the length of ancient DNA templates and therefore potentially the probe-template heterodimer stability, this treatment may, thus, impede the efficacy of target-enrichment techniques. Conversely, the removal of post-mortem damage limits the probe-template edit distance, which may facilitate probe-template annealing and, ultimately, capture. How ancient DNA capture is affected by DNA methylation, which protects from USER treatment (Pedersen et al. 2014, Hanghøj et al. 2016), and thus can increase both read edit distances and sequence length, remains unclear. Surprisingly, the impact of USER treatment on the capture efficacy has not been tested, despite both techniques gaining popularity in ancient DNA research.

In this study, we applied hyRAD capture to eight ancient equine bone specimens from present-day France dating to the 1st-17th centuries CE (Common Era). We compared the capture efficacy on DNA libraries that were constructed in the absence of or following USER treatment of DNA extracts. The specimens analyzed were selected to include horses, donkeys as well as their first-generation mule hybrids, and hyRAD probes were constructed from both horse and donkey fresh DNA extracts. This design allowed us to detect possible capture bias toward one parental species as mules are expected to contain balanced proportions of donkey and horse DNA. We found that USER treatment impacts on capture efficacy, leading to an over-representation of those longer and least damaged templates. These included templates showing more limited DNA methylation. Therefore, following USER-treatment, capture sequencing data can reduce the power to access to hypermethylated regions and to detect post-mortem DNA damage signatures that are yet important for data authentication (Rohland et al. 2015, Orlando et al. 2021). Our work also reveals the presence of a slight reference bias affecting the balance of donkey-specific or horse-specific alleles in mules. Such bias was not fully mitigated and even enhanced following capture with probes from one parental species only. Combined, our results indicate that the probability to detect

For Review Only

highly divergent alleles in an ancient individual, thus, the ancestry contribution from potential divergent ghost-lineages, is not only a function of the sequences represented in the probes utilized for capture but also their size.

Material and Methods

Data generation

In a previous study, we used shallow DNA sequencing to screen a total of 874 equine subfossil specimens for DNA preservation levels (Lepetz et al. Submitted). The sequence data collected were processed within the Zonkey pipeline (Schubert et al. 2017) to identify horses, donkeys and mules showing high levels of endogenous DNA (Table S1). We selected 2 horses, 3 donkeys and 3 mules with 37.9-78.1% endogenous DNA to carry out our experiments as this ensured that sufficient sequence data could be obtained to characterize the consequences of USER-treatment and/or hyRAD capture from minimal sequencing efforts.

Probes production followed protocol by Suchan et al. (2021), starting from 500 ng of horse or donkey DNA, digested for 3 h at 37 °C with 3 U of MseI and 6 U of PstI (New England Biolabs – NEB) in 10 µl reactions. The reaction was then supplemented with ATP (final concentration = 1nM) and each adapter to the final concentration of 0.5 µM (P1 and P2, see Suchan et al. 2021) in 20 µl reactions and incubated 3 h at 16 °C. The reactions were then purified with AMPure with a bead-to-buffer ratio of 1.5:1 and eluted in 30 µl of Tris. The samples were then size-selected for 190-390 bp size range using the Blue Pippin instrument and 2% agarose cassettes with external marker (Sage Science). Fragments with P2 adapters present were then separated using biotinylated MyOne C1 Dynabeads (Invitrogen) as in Suchan et al (2021) and eluted in 30 µl of Tris. The resulting fragments were amplified in three PCR reactions each, using KAPA HiFi HotStart ReadyMix (Roche), 0.6 µM of IS4 primer and indexing primer (Meyer & Kircher, 2010), and 5 µl of the bead solution obtained in the

For Review Only

previous step. The PCR consisted of denaturation for 5 min at 95°C, followed by 15 cycles (denaturation for 20 sec at 98°C, annealing for 15s at 60°C, and; elongation for 30 s at 72°C), followed by elongation step for 5 min at 72°C. The triplicates were pooled, purified with AMPure with a bead-to-buffer ratio of 1:1 and eluted in 60 µl of Tris. The resulting products were digested, using 50 µl of the purified product, 7.5 U of MseI (NEB) in 75 µl reactions. The RNA probes were prepared using HiScribe T7 Quick High Yield RNA Synthesis Kit (NEB), according to the manufacturer's instructions, using 1 µl of the template, incubated overnight, and the DNA template was removed using 2U of TURBO DNAse in 30 min reaction at 37 °C. RNA probes were purified using RNEasy Mini kit (Qiagen®) with the standard protocol, except using 665 µl ethanol to the RNA+RTL mix, The probes were then diluted to around 100 ng/µl. Blocking RNAs were prepared as described in Suchan et al. (2021).

Ancient DNA extraction, USER-treament and library construction procedures followed the work from Fages and colleagues (2019). For each ancient DNA extract, one aliquot was prepared and subjected to USER-treatment before preparing a triple-indexed DNA library while a second aliquot was directly used for library construction. A volume of 14.9µL of raw DNA extract was used for library construction (final volume = 25µL), together with adapters including unique 7 bp indices located at the ends undergoing ligation with ancient DNA templates (Rohland et al. 2015). Two aliquots containing 3µL of DNA libraries were amplified in two parallel 25µL reactions, containing 0.4µL AccuPrime *Pfx* polymerase (ThermoFisher Scientific), 1mg/ml BSA, 200mM of the inPE1 (5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT) and external index (5'-CAA GCA GAA GAC GGC ATA CGA GAT NNN NNN GTG ACT GGA GTT CAG ACG TGT), where NNN NNN represent the external 6 bp library index. The amplification conditions consisted of a first denaturation step for 5 min at 95°C, and were followed by 12 cycles (denaturation for 15 sec at 95°C, annealing for 30s at 60°C, and; elongation for 30 s at 68°C) and a final elongation step for 5 min at 68°C. PCR amplifications were purified using AMPure XP beads (Beckman Coulter), with a bead-to-buffer ratio of 1.4:1, and eluted in 25 µL EB. Two aliquots of each library amplification (7.5 µL

For Review Only

each) were then reamplified in parallel reactions using the same conditions as above, except that the reaction volume was 50 µL and that 1µM of the primers IS5 (5'-AATGATACGGCGACCACCGA) and IS6 (5'-CAAGCAGAAGACGGCATACGA) were used (Meyer & Kircher 2010). The required number of PCR cycles were determined using qPCR (7 to 10 cycles). The two parallel reactions were purified on MinElute columns (Qiagen®) and eluted in 11 µL EB before being mixed to a single tube, thus, providing 22 µL of amplified library. Library concentrations were estimated using Qubit (ThermoFisher Scientific) between 180 and 216 ng/µL.

Capture conditions followed those described by Suchan and colleagues (Suchan et al. 2021), and included a volume of 7µL of each amplified library and 550 ng hyRAD RNA probes. The hybridization reaction was carried out for 40 hours at 55°C in a 30 µL volume. Dynabeads® M-280 Streptavidin (ThermoFisher Scientific) were prepared following three successive washes with 200 µL V1 TEN buffer, and eluted in 70 µL V2 TEN buffer. Hybridization reactions were then purified on an Opentrons OT2 liquid-handling instrument using the automated procedure described by Suchan and colleagues (Suchan et al. 2021), with the slight modifications provided in the Supplemental Material. The final elution step was carried out in 30 µL of Tris 10mM. Two 8 µL aliquots of this first hybridization round were amplified in parallel 50 µL reactions using the KAPA HiFi HotStart ReadyMix (Kapa Biosystems) and the IS5 and IS6 PCR primers (500 nM each). The amplification conditions consisted of a first denaturation step for 3 min at 95°C, and were followed by 9-12 cycles (denaturation for 20 sec at 98°C, annealing for 15s at 60°C, and; elongation for 30 sec at 72°C) and a final elongation step for 1 min at 65°C. Parallel amplification reactions were pooled and purified on MinElute columns (Qiagen®) and eluted in 10 µL EB. A total of 7 µL of this purified amplification was used in a second hybridization reaction, followed by Dynabeads purification and amplification using the same conditions as above, except that the hybridization lasted for 18 hours. The capture product was amplified in PCR as above, for 3 cycles, and purified using AMPure XP beads (Beckman Coulter), with a bead-to-buffer ratio

For Review Only

of 1:1, and eluted in 30 µL EB tween 0.05%. Library profiles and concentrations were estimated using the TapeStation 4200 instrument (Agilent) and Qubit HS dsDNA assay (Invitrogen), before pooling at equimolar concentration for sequencing.

Sequencing was carried out on the Illumina MiniSeq instrument available at CAGT using the paired-end sequence mode (2x81bp, except for the horse probe library, which was sequenced using 2x156bp reactions). Raw DNA sequences were deposited to the European Nucleotide Archive and are available for download under the XXX-TO-BE-UPLOADED-FOLLOWING-ACCEPTANCE-XXX project.

Data analysis

Probe DNA sequences were demultiplexed using CutAdapt v2.10 (Martin 2011; --discard-untrimmed -g "TAATACGACTCACTATAGGGCGG;max_error_rate=0.15;min_overlap=20" -G "^TAA;max_error_rate=0;min_overlap=3"). Raw sequence fastq files generated from ancient DNA libraires were demultiplexed on the basis of the two internal indices using AdapterRemoval2 (Schubert et al. 2016; --barcode-mm-r1 1 --barcode-mm-r2 1 --minlength 25 --trimns --trimqualities --minadapteroverlap 3 --mm 5). This pipeline also collapses paired-end reads showing significant sequence overlap into single end reads, creating so-called '*collapsed*' reads or '*collapsed truncated*' reads, if ends show low base quality scores and require further trimming. Both collapsed, collapsed truncated and those paired-end reads showing no significant overlap were then aligned against reference genomes using Bowtie2 v2.3.5.1 (Langmead et al. 2012) and the Paleomix v1.2.13.2 pipeline (Schubert et al. 2014). Read alignments were carried out following the optimized parameters from Poullet & Orlando (2020) and repeated against the horse reference genome (EquCab3, (Kalbfleish et al. 2019)), supplemented by the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), and the donkey reference genome (Wang et al. 2021). Read alignments were also carried out against the horse (Accession Number = NC_001640) and the donkey

For Review Only

(JADWZW010000001.1) mitochondrial genomes. PCR duplicates were removed using Paleomix and alignments showing mapping qualities strictly inferior to 25 were disregarded. Similar mapping parameters were applied to the demultiplexed set of probe paired-end sequences and ON-target regions were identified applying Bedtools v2.27.1 bamtobed to the alignments obtained from the probe sequence libraries. As daisy chaining reactions during capture can lead to the recovery of sequence data in those regions immediately flanking probes (Cruz-Davalos et al. 2017), we defined OFF-target regions as located at least 250 bp away from ON-target regions. ON-target and OFF-target read alignments were identified using samtools view v1.11, considering a minimum of 1bp sequence overlap.

In silico digestion of the horse and the donkey nuclear genomes were carried out using the script fragmatic.pl (Chafin et al. 2018), with -r "CTGCA^G T^TAA". This provided the expected fraction of each genome represented in each size class (Supplementary Fig. S2). These size distributions were then compared to those measured empirically from the sequence data generated from the probe libraries, using ON-target regions bed coordinates. The probe library non-redundant sequence content was estimated using the Preseq v2_0 lc_extrap command (Daley & Smith 2014) and the limited sequence data generated from the probe DNA libraries. Size, %GC and %CpG distributions were estimated from the fraction of collapsed reads to ensure that ancient DNA templates were characterized across their full sequence length. The size of each DNA template was obtained from the samtools view command, while %GC and %CpG distributions of those genomic regions were obtained by applying seqtk v1.3-r117-dirty comp (Li 2013), with default parameters, to the reference genome of interest, considering genomic windows of 100 bp. Post-mortem DNA damage signatures, especially C→T mis-incorporations within and outside CpG dinucleotides, were obtained from PMDtools v0.60 (Skoglund et al. 2014). Calculations and statistical tests were carried out using standard functions in R (R Core Team 2014), and plots were generated using the R ggplot2 package (Wickham 2009). F-methylation scores were calculated using DamMet v1 (Hanghøj et al. 2019), with standard parameters.

For Review Only

Horse-specific and donkey-specific alleles were identified using published genome sequence data for a total of 27 horses and 27 donkeys, respectively, which encompassed a whole range of breeds in both species (for horses: Akhal-Teke (Jagannathan et al. 2019), Duelmener (Schrimpf et al. 2016), Franches Montagnes (Jagannathan et al. 2019), Haflinger (Jagannathan et al. 2019), Hanoverian (Schrimpf et al. 2016), Holsteiner (Jagannathan et al. 2019), Icelandic (Andersson et al. 2012), Jeju (Lee et al. 2018), Lipizzan (Wallner et al. 2017), Marwari (Jun et al. 2014), Mixed UK Warmblood (Jagannathan et al. 2019), Mongolian (Do et al. 2014), Noriker (Jagannathan et al. 2019), Painted (Jagannathan et al. 2019), Quarter (Jagannathan et al. 2019), Reit (Jagannathan et al. 2019), Shetland (Jagannathan et al. 2019), Sorraia (Metzger et al. 2015), Standardbred (Cosgrove et al. 2020), Swiss Warmblood (Jagannathan et al. 2019), Thoroughbred (Cosgrove et al. 2020), Trakener (Jagannathan et al. 2019), Egyptian Arabian (Cosgrove et al. 2020), Welsh (Jagannathan et al. 2019), Westfale (Jagannathan et al. 2019), and Yakutian (Librado et al. 2015); for donkeys: Au-1, Ch-by1, Ch-by3, Ch-gl1, Ch-ht1, Ch-jm1, Ch-kl2, Ch-qy2, Ch-tlf1, Ch-xj1, Ch-yn3, Eg-1, Eg-3, Et-4, Et-8, Ir-7, Ke-16, Ky-6, Ni-6, Sp-10, Sp-12, Sp-15, Sp-17, Sp-2, Sp-5, Sp-7, Ti-4; Wang et al. 2021). Sequence data were aligned following the same procedure as described above, and resulting BAM alignment files were processed in ANGSD (-minQ 30, -minMapQ 25 -baq 0 -rmTriallelic 1e-4 -SNP_pval 1e-6 -C 50; Korneliussen et al. 2014) to identify minor and major alleles at sites that were polymorphic and covered in at least 40 individuals. ANGSD output files were next restricted to those sites showing minimal genotype quality scores of 0.99 in all individuals covered to identify positions where allelic frequency differences between horses and donkeys were at least 95%. This provided a provisional list of genome positions in which different variants neared fixation in donkeys and horses.

Results

For Review Only

Previous work (Lepetz et al. Submitted) allowed us to identify a total of eight ancient equine subfossils from France (1st-17th centuries CE) showing excellent DNA preservation, including 2 horses, 3 donkeys and 3 mules (Table S1). Their high endogenous DNA content (37.9%-78.1%) guaranteed that a large number of equine DNA templates could be identified, and thus, the effect of capture following USER treatment assessed, even from limited sequencing efforts (289,599-1,997,334 reads; Table S1). We used fresh DNA extracts from modern horses and donkeys to prepare two hyRAD probe libraries following the methodology from Suchan and colleagues (Suchan et al. 2021). We produced a total of 6.4 and 12.7 million paired-end sequences to characterize the horse and donkey probe library content. The sequence data generated from the horse probe library covered approximately 118.99 Mb (~4.73%) of the horse reference genome. The donkey sequence probes obtained covered 110.50 Mb (~4.54%) of the donkey reference genome. In the following, these sequences were used to identify those ancient DNA read alignments located ON-target. Since daisy-chaining reactions can extend capture in the regions immediately flanking probes (Cruz-Davalos et al. 2017), we considered that the sequences located at least 250 bp-away from probes as a conservative set of OFF-target regions. It is important to notice that further sequencing of DNA probe libraries would have extended the ON-target range (Fig. S1). *In silico* digestion indeed indicates that probe regions could encompass 15.01% and 14.83% of the horse and donkey genomes, respectively. Therefore, we can estimate that the fraction of genomic regions considered OFF-target but including probes that have not yet been sequenced is approximately of 10.28 and 10.29% for both genomes. Therefore, the calculations presented in the following sections, including enrichment folds and the differences detected between ON- and OFF-target regions, are conservative.

Interestingly, Preseq (Li 2013) predicted that an average 25-30 million unique templates were present in the horse probe library versus 15-25 million for the donkey library (Fig. S1). This is in line with previous work using the same combination of enzymes for hyRAD

For Review Only

probe preparation (HspI-MseI, (Suchan et al. 2021)), and suggests 25-30 million read pairs as a conservative sequencing effort to characterize the full mappable probe library content.

The ratio of the total number of read alignments overlapping ON-target regions following hyRAD capture or shotgun sequencing showed enrichment-folds of 2.08-3.08 for the horse probes (Fig. 1AB). Enrichment-folds were more limited when using the donkey probes (1.29-2.33), possibly due to their more limited size (Supplementary Fig. S2). While the enrichment-folds obtained are conservative, they may seem limited. It is, however, important to consider that only samples characterized by high endogenous content were considered in our experiments. The probability that shotgun data overlap a probe genomic region by chance is considerably increased in such specimens relative to standard subfossil DNA extracts, which are dominated by environmental microbial DNA. Therefore, higher enrichment-folds are expected when working on material with a more limited DNA preservation, in line with the results from Suchan and colleagues (2021).

Interestingly, higher enrichment-folds were obtained when capturing DNA libraries constructed on raw DNA extracts with horse probes than when capturing DNA libraries prepared from USER-treated DNA extracts (Wilcoxon signed rank test, *p*-values = 0.014-0.036; Fig.1 AB). The same was true five (seven) out of the eight samples analyzed here, when donkey probes were used for capture and when the sequence data were mapped against the horse (donkey) genome (Wilcoxon signed rank test, *p*-values = 0.039 and 0.151; Fig.1 AB).

Our hyRAD probe sequences represented a non-random fraction of the genome and were characterized by both a higher %GC and %CpG dinucleotide content (Supplementary Fig. S3; Kolmogorov-Smirnov test, *p*-values < $2.2.10^{-16}$). The %GC and %CpG dinucleotide content of the ancient DNA sequence data were also significantly increased following capture, relative to shotgun sequencing (Fig. 2BC, Supplementary Fig. S4AB; Kolmogorov-Smirnov test, *p*-values < $2.2.10^{-16}$). This effect was more pronounced in ON-target regions than in OFF-target regions, in line with the latter including only a limited fraction of yet uncharacterized probes. Ancient DNA templates sequenced post-capture were also

For Review Only

significantly longer post-capture (Fig. 2A; Kolmogorov-Smirnov test, *p*-values < $2.2.10^{-16}$), especially in ON-target regions, in line with previous reports. That the sequence data (1) are more often located in ON-target regions following capture and (2) comprise a larger fraction of longer templates while (3) mirror base compositional features characteristics of the probes demonstrates that the hyRAD capture procedure was applied successfully. This suggested that the resulting sequence data could be used to assess the effect of USER-treatment on capture efficacy.

Interestingly, when capture was carried out with horse probes, the increment in size distribution of the ON-target sequence data generated was generally more pronounced following USER-treatment than in the absence of USER treatment (Fig. 2A, Supplementary Fig. S5A; Kolmogorov-Smirnov test, *p*-values < $2.2.10^{-16}$). This was unexpected since USER-treatment cleaves off un-methylated cytosine residues that were deaminated into uraciles post-mortem (Briggs et al. 2010). This effect was not apparent, and in fact reversed (Kolmogorov-Smirnov test, *p*-values < $2.2.10^{-16}$), when donkey probes were used for capture (Fig. 2A, Supplementary Fig. S5A), suggesting probe-specific features impacting on the capture outcome.

Several phenomena are known to limit the efficacy of USER-treatment. The first is DNA methylation as post-mortem deamination at methylated cytosine leads to the formation of thymine residues, instead of uracile residues at unmethylated sites (Hanghøj et al. 2016). Since thymine residues cannot be cleaved by USER enzymes, the observed increase in the size distribution of ON-target reads following capture with horse hyRAD probes may, thus, indicate that the probes encompassed genomic regions with higher-than-average DNA methylation levels. Additionally, molecular breathing at both ends of ancient DNA fragments is known to limit the efficacy of USER-treatment as the first and last positions within reads generally show relatively high nucleotide mis-incorporation rates driven by post-mortem cytosine deamination (Rohland et al. 2015). Therefore, the observed increase in the size distribution of ON-target reads following capture with horse hyRAD probes may also reflect the more limited efficacy of USER-treatment at the fragment termini.

For Review Only

To test each of these two non-mutually exclusive hypotheses, we compared C→T mis-incorporation rates at the first read position, where post-mortem DNA damage is maximal, and within the following 24 read positions (Fig. 3AB, Supplementary Fig. S6-8). We first looked at non-CpG dinucleotides as sites mostly unaffected by DNA methylation (Bird 2002) (Supplementary Fig. S6-S7). As expected, such C→T mis-incorporation rates were higher in the absence of USER-treatment than following USER-treatment in all experimental conditions and read positions considered (Wilcoxon signed rank test, *p*-values = $1.554.10^{-4}$-0.014). This indicated that the USER treatment was applied successfully. Next, we looked at both site categories for C→T mis-incorporation rates at non-CpG dinucleotides. We found that such rates were markedly reduced in ON-target regions relative to OFF-target regions, following capture both in the absence and presence of USER-treatment. This was true regardless of whether horse or donkey probes were used (Wilcoxon signed rank test, *p*-values = $3.052.10^{-5}$-0.016), which illustrates the general capture propensity to retrieve those least damaged DNA templates.

Conditioning the same analyses on CpG dinucleotides revealed reduced ON-target CpG→TpG mis-incorporation rates post-capture both in the presence and the absence of USER-treatment (Wilcoxon signed rank test, *p*-values = $3.052.10^{-5}$). This was equally true for the first read position as well as the following 24 nucleotides (Fig. 3AB, Supplementary Fig. S8), which rules out molecular breathing at the fragment ends as a main factor. The fact that ON-target CpG→TpG mis-incorporation rates were lower and not higher post-capture also rules out DNA methylation protecting from USER-treatment as a possible cause for the incremental size shift observed following capture with horse hyRAD probes. CpG→TpG mis-incorporation rates post-USER treatment are indeed indicative of genomic regions showing more limited DNA methylation levels (Hanghøj et al. 2016), which could be confirmed here using DamMet (Hanghøj et al. 2019) for predicting F-methylation scores of ON-target and OFF-target regions (Fig. 4). ON-target regions were found to be associated with lower DNA methylation values than OFF-target regions as the fraction of CpG dinucleotides characterized by at least 50%

For Review Only

DNA methylation levels was consistently reduced at such regions, even in the absence of capture. This was true for both horse and donkey probes, and regardless of whether capture or shotgun DNA sequencing were performed (Wilcoxon signed rank test, *p*-values =$1.397.10^{-9}$-$5.122.10^{-8}$). Therefore, the probe regions were characterized by DNA methylation levels lower than the remaining genomic fraction. Combined with previous observations, this indicates that neither molecular breathing nor DNA hyper-methylation were responsible for the observed increase in the size distribution of ON-target reads following capture with horse hyRAD probes.

Our data, however, suggest that, following USER-treatment, capture with horse probes provided preferential access to a sub-population of ancient DNA templates characterized by longer sizes, lower DNA methylation levels, and more limited post-mortem DNA damage. Capture with donkey probes provided access to ancient DNA molecules showing similar molecular features, except for their size which remained longer following USER-treatment than in the absence of USER-treatment. As this could not be explained by molecular breathing or differences in DNA methylation, the difference observed in the DNA templates captured is likely due to the size difference of the donkey and horse probes themselves. The former were significantly shorter (median = 78 bp) than the latter (median = 168 bp), despite similar experimental production procedures were applied (Supplementary Fig. S2; Kolmogorov-Smirnov test, *p*-value < $2.2.10^{-16}$). As horse probes were longer, they likely disproportionally favored the annealing of the longest ancient DNA templates, while the shorter donkey probes have allowed annealing to those more fragmented ancient DNA templates.

We next leveraged our experimental design to assess whether horse-specific alleles and donkey-specific alleles had equal chances to be detected following capture with horse or donkey hyRAD probes, respectively. Species-specific alleles for horses and donkeys were identified as those alleles nearing fixation in a worldwide panel representing 27 genomes for both species. Regardless of the reference genome considered, we found that the sequence

For Review Only

data generated from the two horse specimens only showed horse-specific alleles (Fig. 5AB) The reverse was true for the sequence data generated from the three donkey specimens, which only showed donkey-specific alleles. This supports previous taxonomic identification of these specimens (Lepetz et al. 2021). Interestingly, the three mule specimens showed almost balanced proportions of horse- and donkey-specific alleles. However, mule sequence alignments against the horse reference genome generally slightly over-represented horse-specific alleles. Reciprocally, mule sequence alignments against the donkey reference genome were slightly enriched in donkey-specific alleles. This is characteristic of limited but consistent reference bias in our analyses. Importantly, applying hyRAD capture with horse probes slightly inflated the proportion of horse-specific alleles identified in mule specimens, hence, the reference bias (Fig. 5A). Conversely, the reference bias against horse alleles detected when aligning mule data against the donkey reference genome was partly mitigated when using horse hyRAD probes libraries (Fig. 5B). This was true both in the presence and the absence of USER-treatment. In contrast, the reference bias against donkey alleles detected when aligning mule data against the horse reference genome was over-compensated when using donkey hyRAD probes libraries (Fig. 5A). The effect was more pronounced after USER-treatment and was not present when aligning the same sequence data against the donkey reference genome (Fig. 5A). This was observed regardless whether transition SNPs were filtered or not (Fig. 5, Supplementary Fig. S8), which rules out DNA damage as a potential driver.

Discussion

In this study, we extracted ancient DNA from eight subfossils of horses, donkeys and mules and constructed double-stranded DNA libraries using both raw DNA extracts and DNA extracts that were subjected to USER-treatment. Amplified DNA libraries were then shotgun sequenced or captured using the hyRAD technology and probes that were prepared from

For Review Only

horse or donkey fresh DNA extracts. Our experimental design allowed us to assess, for the first time, whether USER-treatment impacted on the capture performance and provided equal chances to identify both parental alleles in hybrid individuals.

We found that the sequence data generated following capture were more often located on probe regions than when performing shotgun sequencing. Additionally, ON-target sequence data recapitulated the molecular features of the probes, including increased %GC and %CpG dinucleotide content relative to the remaining fraction of the genome, and decreased DNA methylation levels. This ensured that the hyRAD procedures implemented here succeeded in facilitating the sequence characterization of a pre-selected genome fraction. Interestingly, the size distribution of the ancient DNA fragments sequenced over their full size was increased in all capture conditions. However, USER-treatment led to the recovery of ancient DNA fragments that were on average longer than those retrieved in the absence of USER-treatment when horse probes were used for capture. This was not the case when capturing with donkey probes that encompassed a shorter size range (Supplementary Fig. S2). The shorter size of the donkey probes used in our experiments may have favored the formation of short probe-template heterodimers, thus, providing no further advantage to the longest templates during capture. This resulted in donkey-specific alleles being more efficiently enriched than horse-specific alleles when capturing mule specimens (Fig. 5). Our findings, thus, highlight the size selection as a critical step in the hyRAD probe preparation procedure. In this study, we used the same methodology for preparing both probe sets and yet recovered probe spanning different size ranges. Improving the reproducibility of this step will be critical before hyRAD capture can be applied across large sample panels without risking to introduce important technical batch effects. No such difficulties are expected with capture protocols using synthetic oligonucleotides of controlled sizes (e.g. 52-mers in (Slon et al. 2017), 60-mers in (Cruz-Davalos et al. 2017) and 120-mers in (Ramos-Madrigal et al. 2019).

While mules carry even numbers of donkey and horse autosomes, horse-specific alleles were slightly over-represented when using the horse reference genome for sequence

For Review Only

458 alignments. Conversely, donkey-specific alleles were slightly over-represented when aligning
459 the mule sequence data against the donkey reference genome. This is typical of a reference
460 bias (Günther and Nettelblad, 2019), favoring the parental species used as reference. We
461 expect such effects to be commensurate with the sequence divergence when species
462 others than the horse and the donkey are investigated. Importantly, using horse probes
463 increased this bias further when aligning mule sequence data against the horse reference
464 genome. Using horse probes, however, reduced this bias when aligning mule sequence data
465 against the donkey reference genome and led to the recovery of almost balanced
466 proportions of horse- and donkey-specific alleles. This may have suggested mixing horse and
467 donkey probes as the most optimal strategy for limiting the extent of reference bias
468 characterized here. However, the donkey probes disproportionally favored the detection of
469 donkey alleles when aligning mule sequences against the horse genome, an effect that was
470 magnified following USER-treatment and that was most likely driven by the shorter probe size
471 range. We, thus, recommend to merge probe panels prepared from divergent lineages only
472 after testing that similar size ranges were obtained during probe preparation as any
473 significant size difference can significantly impact the accuracy of the ancestry proportions
474 estimated. The reciprocal mapping strategy implemented in our study, in which the
475 sequencing data were aligned against multiple reference genomes from closely related
476 species, provides an easy control for assessing the quality of ancestry proprotions inferred. In
477 the light of our results, performing such controls appears particularly important when
478 combining data obtained on from extracts that were or were not subjected to USER
479 treatment, as the over-representation of donkey alleles in the mule sequence data obtained
480 following capture with hyRAD donkey probes was even more pronounced when considered
481 USER-treated DNA extracts.

482 Post-mortem DNA damage signatures at CpG sites also indicated that the ancient
483 DNA fragments retrieved post-USER treatment carried fewer deaminated cytosines and
484 encompassed genomic regions showing lower levels of DNA methylation. We interpret these

For Review Only

findings in terms of increased probe-template heterodimer stability as deaminated cytosines increase the number of probe-template sequence mismatches. It is interesting to notice that DNA methylation offers a protection against USER-treatment as methylated cytosine residues are deaminated in thymines. The average size of those ancient DNA fragments including methylated cytosines is, thus, expected to be longer following USER treatment than those DNA fragments containing no methylated cytosine residues. This increased size should improve probe-template stability. However, these fragments also include higher CpG→TpG mis-incorporations (Hanghøj et al. 2016), which reduce the probe-template stability. In the experimental conditions investigated here, we found that the ON-target CpG→TpG conversion rates sharply decreased following capture and USER-treatment. Therefore, those ancient DNA templates sequenced post-capture were characterized by lower-than-average DNA methylation levels. This indicates that DNA methylation protection against USER-treatment offered no particular advantage for capture. Whether targeting hypermethylated regions would lead to similar results remains to be determined.

Regardless, our data showed that shotgun sequencing and hyRAD capture can yield different cytosine deamination rate estimates on data yet collected from the same individual (Fig. 3, Supplementary Fig. 5). This suggests that the capture procedures do not provide a random subset of the pool of ancient DNA library templates covering the regions of interest but can instead over-represent those showing molecular features increasing probe-template affinities, including those least fragmented, deaminated and methylated. This has important ramifications for data authentication, which partly rely on the presence of particular post-mortem DNA damage signatures.

Acknowledgements

We thank all members of the AGES research group at CAGT for fruitful discussions and Dr Albano Beja-Pereira for providing modern DNA extracts. This project has received funding

For Review Only

from: the University Paul Sabatier IDEX Chaire d'Excellence (OURASI); the CNRS Programme de Recherche Conjoint (PRC); the CNRS International Research Project (IRP AMADEUS); the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 797449; the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 681605).

References

Andersson, L., Larhammar, M., Memic, F., Wootz, H., Schwochow, D., & Rubin, C. et al. (2012). Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature*, *488*(7413), 642-646. https://doi.org/10.1038/nature11399

Ávila-Arcos, M., Cappellini, E., Romero-Navarro, J., Wales, N., Moreno-Mayar, J., & Rasmussen, M. et al. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports*, *1*(1). https://doi.org/10.1038/srep00074

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, *16*(1), 6-21. https://doi.org/10.1101/gad.947102

Boessenkool, S., Hanghøj, K., Nistelberger, H., Der Sarkissian, C., Gondek, A., & Orlando, L. et al. (2016). Combining bleach and mild predigestion improves ancient DNA recovery from bones. *Molecular Ecology Resources*, *17*(4), 742-751. https://doi.org/10.1111/1755-0998.12623

Briggs, A., Stenzel, U., Meyer, M., Krause, J., Kircher, M., & Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, *38*(6), e87-e87. https://doi.org/10.1093/nar/gkp1163

Carpenter, M., Buenrostro, J., Valdiosera, C., Schroeder, H., Allentoft, M., & Sikora, M. et al. (2013). Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *The American Journal Of Human Genetics*, *93*(5), 852-864. https://doi.org/10.1016/j.ajhg.2013.10.002

Chafin, T. K., Douglas, M. R., & Douglas, M. E. (2018). MrBait: universal identification and design of targeted-enrichment capture probes. *Bioinformatics (Oxford, England)*, *34*(24), 4293–4296. https://doi.org/10.1093/bioinformatics/bty548

Clavel, P., Dumoncel, J., Der Sarkissian, C., Seguin-Orlando, A., Calvière-Tonasso, L., & Schiavinato, S. et al. (2021). Assessing the predictive taxonomic power of the bony labyrinth 3D shape in horses, donkeys and their F1-hybrids. *Journal Of Archaeological Science*, *131*, 105383. https://doi.org/10.1016/j.jas.2021.105383

Cosgrove, E.J., Sadeghi, R., Schlamp, F. *et al.* Genome Diversity and the Origin of the Arabian Horse. *Sci Rep* 10, 9702 (2020). https://doi.org/10.1038/s41598-020-66232-1

Cruz-Dávalos, D., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., & Soubrier, J. et al. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Molecular Ecology Resources*, *17*(3), 508-522. https://doi.org/10.1111/1755-0998.12595

Dabney, J., Knapp, M., Glocke, I., Gansauge, M., Weihmann, A., & Nickel, B. et al. (2013a). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings Of The National Academy Of Sciences*, *110*(39), 15758-15763. https://doi.org/10.1073/pnas.1314445110

Dabney, J., Meyer, M., & Paabo, S. (2013b). Ancient DNA Damage. *Cold Spring Harbor Perspectives In Biology*, *5*(7), a012567-a012567. https://doi.org/10.1101/cshperspect.a012567

Dabney, J., & Meyer, M. (2012). Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques, Feb;52(2):87-94. https://doi.org/10.2144/000113809. PMID: 22313406.

Daley, T., & Smith, A. (2014). Modeling genome coverage in single-cell sequencing. *Bioinformatics*, *30*(22), 3159-3165. https://doi.org/10.1093/bioinformatics/btu540

Damgaard, P., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., & Allentoft, M. (2015). Improving access to endogenous DNA in ancient bones and teeth. *Scientific Reports*, *5*(1). https://doi.org/10.1038/srep11184

Do, K., Kong, H., Lee, J., Lee, H., Cho, B., & Kim, H. et al. (2014). Genomic characterization of the Przewalski's horse inhabiting Mongolian steppe by whole genome re-sequencing. *Livestock Science*, *167*, 86-91. https://doi.org/10.1016/j.livsci.2014.06.020

Enk, J., Devault, A., Kuch, M., Murgha, Y., Rouillard, J., & Poinar, H. (2014). Ancient Whole Genome Enrichment Using Baits Built from Modern DNA. *Molecular Biology And Evolution*, *31*(5), 1292-1294. https://doi.org/10.1093/molbev/msu074

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., & McCrory Constantz, C., et al. (2019). Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. *Cell*, *177*(6), 1419–1435.e31. https://doi.org/10.1016/j.cell.2019.03.049

Felkel, S., Vogl, C., Rigler, D., Dobretsberger, V., Chowdhary, B., & Distl, O. et al. (2019). The horse Y chromosome as an informative marker for tracing sire lines. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-42640-w

Furtwängler, A., Neukamm, J., Böhme, L., Reiter, E., Vollstedt, M., & Arora, N. et al. (2020). Comparison of target enrichment strategies for ancient pathogen DNA. *Biotechniques*, *69*(6), 455-459. https://doi.org/10.2144/btn-2020-0100

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A., Alquraishi, S., & Al-Rasheid, K. et al. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources*, *16*(2), 459-469. https://doi.org/10.1111/1755-0998.12470

Gamba, C., Jones, E., Teasdale, M., McLaughlin, R., Gonzalez-Fortes, G., & Mattiangeli, V. et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, *5*(1). https://doi.org/10.1038/ncomms6257

For Review Only

Gansauge, M., Aximu-Petri, A., Nagel, S., & Meyer, M. (2020). Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA. *Nature Protocols*, *15*(8), 2279-2300. https://doi.org/10.1038/s41596-020-0338-0

Gansauge, M., Gerber, T., Glocke, I., Korlević, P., Lippik, L., & Nagel, S. et al. (2017). Single-stranded DNA library preparation from highly degraded DNA usingT4DNA ligase. *Nucleic Acids Research*, gkx033. https://doi.org/10.1093/nar/gkx033

Goodwin, S., McPherson, J., & McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333-351. https://doi.org/10.1038/nrg.2016.49

Green, R., Briggs, A., Krause, J., Prüfer, K., Burbano, H., & Siebauer, M. et al. (2009). The Neandertal genome and ancient DNA authenticity. *The EMBO Journal*, *28*(17), 2494-2502. https://doi.org/10.1038/emboj.2009.222

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, *15*(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

Hanghøj, K., Seguin-Orlando, A., Schubert, M., Madsen, T., Pedersen, J., Willerslev, E., & Orlando, L. (2016). Fast, Accurate and Automatic Ancient Nucleosome and Methylation Maps with epiPALEOMIX. *Molecular Biology And Evolution*, *33*(12), 3284-3298. https://doi.org/10.1093/molbev/msw184

Hanghøj, K., Renaud, G., Albrechtsen, A., & Orlando, L. (2019). DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *Gigascience*, *8*(4). https://doi.org/10.1093/gigascience/giz025

Jagannathan, V., Gerber, V., Rieder, S., Tetens, J., Thaller, G., Drögemüller, C., & Leeb, T. (2019). Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Animal Genetics*, *50*(1), 74-77. https://doi.org/10.1111/age.12753

Jun, J., Cho, Y., Hu, H., Kim, H., Jho, S., & Gadhvi, P. et al. (2014). Whole genome sequence and analysis of the Marwari horse breed and its genetic origin. *BMC Genomics*, *15*(S9). https://doi.org/10.1186/1471-2164-15-s9-s4

Kalbfleisch, T., Rice, E., DePriest, M., Walenz, B., Hestand, M., & Vermeesch, J. et al. (2019). Author Correction: Improved reference genome for the domestic horse increases assembly contiguity and composition. *Communications Biology*, *2*(1). https://doi.org/10.1038/s42003-019-0591-3

Kapp, J., Green, R., & Shapiro, B. (2021). A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. *Journal Of Heredity*, *112*(3), 241-249. https://doi.org/10.1093/jhered/esab012

Korlević, P., & Meyer, M. (2019). Pretreatment: Removing DNA Contamination from Ancient Bones and Teeth Using Sodium Hypochlorite and Phosphate. *Methods in molecular biology (Clifton, N.J.)*, *1963*, 15–19. https://doi.org/10.1007/978-1-4939-9176-1_2

Korneliussen, T., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/s12859-014-0356-4

Lee, W., Park, K., Taye, M., Lee, C., Kim, H., Lee, H., & Shin, D. (2018). Analysis of cross-population differentiation between Thoroughbred and Jeju horses. *Asian-Australasian Journal Of Animal Sciences*, *31*(8), 1110-1118. https://doi.org/10.5713/ajas.17.0460

Lepetz, S. (2021). Animals in funeral practices in Belgic Gaul between the end of the 1st century BC and the beginning of the 5th century AD: From gallic practices to Gallo-Roman practices. *Hyper Article en Ligne - Sciences de l'Homme et de la Société*, ID : 10670/1.ycqmbj.

Li, H. (2013). Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences. Retrieved from https://github.com/lh3/seqtk

Librado, P., Der Sarkissian, C., Ermini, L., Schubert, M., Jónsson, H., & Albrechtsen, A. et al. (2015). Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proceedings Of The National Academy Of Sciences*, *112*(50), E6889-E6897. https://doi.org/10.1073/pnas.1513696112

Langmead, B., & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Maricic, T., Whitten, M., & Pääbo, S. (2010). Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *Plos ONE*, *5*(11), e14004. https://doi.org/10.1371/journal.pone.0014004

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., & Roodenberg, S. et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499-503. https://doi.org/10.1038/nature16152

Metzger, J., Karwath, M., Tonda, R., Beltran, S., Águeda, L., & Gut, M. et al. (2015). Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC Genomics*, *16*(1). https://doi.org/10.1186/s12864-015-1977-3

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448-pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Meyer, M., Kircher, M., Gansauge, M., Li, H., Racimo, F., & Mallick, S. et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, *338*(6104), 222-226. https://doi.org/10.1126/science.1224344

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P.W., Ávila-Arcos, M.C., et al. (2021). Ancient DNA analysis. *Nat Rev Methods Primers* 1, 15. https://doi.org/10.1038/s43586-021-00016-3

Pedersen, J., Valen, E., Velazquez, A., Parker, B., Rasmussen, M., & Lindgreen, S. et al. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Research*, *24*(3), 454-466. https://doi.org/10.1101/gr.163592.113

Poullet, M., & Orlando, L. (2020). Assessing DNA Sequence Alignment Methods for Characterizing Ancient Genomes and Methylomes. *Frontiers In Ecology and Evolution*, *8*. https://doi.org/10.3389/fevo.2020.00105

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Ramos-Madrigal J, Wiborg Runge AK, Bouby L, Lacombe T, Samaniego Castruita JA, Adam-Blondon A.-F., et al. (2019). Palaeogenomic insights into the origins of French grapevine diversity. *Nat Plants* 5(6), 595-603. https://doi.org/10.1038/s41477-019-0437-5

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., & Reich, D. (2015). Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *370*(1660), 20130624. https://doi.org/10.1098/rstb.2013.0624

Schrimpf, R., Gottschalk, M., Metzger, J., Martinsson, G., Sieme, H., & Distl, O. (2016). Screening of whole genome sequences identified high-impact variants for stallion fertility. *BMC Genomics*, *17*(1). https://doi.org/10.1186/s12864-016-2608-3

Schubert, M., Mashkour, M., Gaunitz, C., Fages, A., Seguin-Orlando, A., & Sheikhi, S. et al. (2017). Zonkey: A simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages. *Journal Of Archaeological Science*, *78*, 147-157. https://doi.org/10.1016/j.jas.2016.12.005

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1). https://doi.org/10.1186/s13104-016-1900-2

Schubert, M., Ermini, L., Sarkissian, C., Jónsson, H., Ginolhac, A., & Schaefer, R. et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, *9*(5), 1056-1082. https://doi.org/10.1038/nprot.2014.063

Sirak, K., Fernandes, D., Cheronet, O., Harney, E., Mah, M., & Mallick, S. et al. (2020). Human auditory ossicles as an alternative optimal source of ancient DNA. *Genome Research*, *30*(3), 427-436. https://doi.org/10.1101/gr.260141.119

Skoglund, P., Northoff, B., Shunkov, M., Derevianko, A., Pääbo, S., Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings Of The National Academy Of Sciences*, *111*(6), 2229-2234. https://doi.org/10.1073/pnas.1318934111

Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., de la Rasilla, M., Lalueza-Fox, C., Rosas, A., Soressi, M., Knul, M. V., Miller, R., Stewart, J. R., Derevianko, A. P., Jacobs, Z., Li, B., Roberts, R. G., Shunkov, M. V., de Lumley, H., Perrenoud, C., Gušić, I., Kućan, Ž., … Meyer, M. (2017). Neandertal and Denisovan DNA from Pleistocene sediments. *Science (New York, N.Y.)*, *356*(6338), 605–608. https://doi.org/10.1126/science.aam9695

Suchan, T., Kusliy, M., Khan, N., Chauvey, L., Tonasso-Calvière, L., & Schiavinato, S. et al. (2021). Performance and automation of ancient DNA capture with RNA hyRAD probes. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13518

Suchan, T., Pitteloud, C., Gerasimova, N., Kostikova, A., Schmid, S., & Arrigo, N. et al. (2016). Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. *PLOS ONE*, *11*(3), e0151651. https://doi.org/10.1371/journal.pone.0151651

van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., & Hartmann, S. et al. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*(7849), 265-269. https://doi.org/10.1038/s41586-021-03224-9

Wallner, B., Palmieri, N., Vogl, C., Rigler, D., Bozlak, E., & Druml, T. et al. (2017). Y Chromosome Uncovers the Recent Oriental Origin of Modern Stallions. *Current Biology*, *27*(13), 2029-2035.e5. https://doi.org/10.1016/j.cub.2017.05.086

Wang, C., Li, H., Guo, Y., Huang, J., Sun, Y., & Min, J. et al. (2021). Author Correction: Donkey genomes provide new insights into domestication and selection for coat color. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21014-9

Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*. Springer New York.

Data accessibility

Raw DNA sequences are deposited to the European Nucleotide Archive and are available for download under the XXX-TO-BE-UPLOADED-FOLLOWING-ACCEPTANCE-XXX project.

Author contributions

Designed and conceived the study: TS and LO. Performed wet-lab work: TS, LC, LTC, SS, ASO. Provided samples, material and reagents: BC, SL and LO. Analyzed data: LO. Interpreted the data: TS, LO. Wrote the article: LO, with input from all co-authors.

Figures.

Figure 1. ON-target enrichment-folds.
Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were subjected to capture with hyRAD horse (H-capture) or donkey (D-capture) probes. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Panel A: Reads were aligned against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), and filtered for PCR duplicates and minimal mapping quality of 25. Panel B: Same as panel A, except that the sequence alignments against the donkey reference genome (Wang et al. 2021) were used.

Figure 2. Size and base composition of ancient DNA templates following shotgun sequencing and hyRAD capture.

For Review Only

Panel A: Size distribution of endogenous DNA templates. Panel B: %GC content of the genomic regions covered by endogenous DNA templates. Panel C: Proportions of CpG dinucleotides present in the genomic regions covered by endogenous DNA templates. Read pairs that could not be collapsed into full-length templates were not used for mapping and only unique, high-quality alignments were considered. The horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), was used for read alignment. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

Figure 3. CpG→TpG conversion rates.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG→TpG conversions report the relative fraction of read alignments in which a CpG dinucleotide is found in the horse reference genome considered but a TpG dinucleotide is found in the sequence data. Calculations were conditioned on those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

Figure 4. Proportion of CpG dinucleotides showing F-methylation levels above 50%.
F-methylation levels were calculated using DamMet (Hanghøj et al. 2019) with default parameters. Calculations were conditioned on those read alignments against the reference genome that were located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

Figure 5. Proportions of horse-specific and donkey-specific alleles identified following shotgun and hyRAD-capture.
Panel A: Alignments against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-contigs from Felkel and colleagues (Felkel et al. 2019). Panel B: Alignments against the donkey reference genome (Wang et al. 2021). Calculations indicate the difference between the number of reads carrying horse-specific and donkey-specific alleles, respectively. Only transversion alleles were considered. Positive (negative) fractions indicate an over-representation of horse-specific (donkey-specific) alleles in the sequence data. Calculations were conditioned on those read alignments against the reference genome that were located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

For Review Only

Supplementary Tables

Supplementary Table 1. Sample description and sequencing and hyRAD capture statistics. DNA extracts from 2 ancient horses (light green), 3 ancient donkeys (deep blue) and 3 ancient mules (light blue) were converted into double-stranded DNA libraries, following treatment (+) with the USER enzymatic mix, or in the absence of USER treatment (-). DNA libraries were either subjected to shotgun sequencing or hyRAD capture using horse (Horse Probes) or donkey probes (Donkey Probes). The sex of each specimen was determined by Clavel and colleagues (Clavel et al., Submitted) following analysis of shallow DNA sequencing data with Zonkey (Schubert et al. 2017). Raw paired-end reads were processed using AdapterRemoval2 (Schubert et al. 2016), providing collapsed, collapsed truncated and non-collapsed pairs that were mapped against the horse (HR, (Kalbfleish et al. 2019)) and donkey (DR, Wang et al. 2021) reference genomes. The horse genome was supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019). Endogenous DNA content is approximated from the total fraction of aligned reads. The number of unique (U), high-quality (MQ25) read alignments against both reference genomes and mitochondrial genomes are also provided. The total number of ON- and OFF-target read alignments are conservative estimates based on limited sequencing of probe libraries and following PCR duplicate removal (U) and mapping quality filtering (MQ25). ON-target fold enrichments are estimated following hyRAD capture and relative to shotgun sequencing. ON/OFF = ratios of ON- and OFF-target read alignments. CE = Common Era. Colors follow those from Clavel and colleagues (Clavel et al. 2021).

Supplementary Figures

Supplementary Figure 1. PreSeq calculation of the probe library content.
Panel A: The probe DNA library that was prepared from modern horse DNA extracts was subjected to shallow sequencing (red dashed line). Reads were then aligned against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019) (left), or against the donkey reference genome (Wang et al. 2021) (right). Panel B: Same as panel A, except that the probe DNA library was prepared from modern donkey DNA extracts. The PreSeq statistical model indicated that the horse probe DNA library contained an average of ~25-30 million genomic fragments versus 15-25 million for the donkey probe DNA library. Solid lines report the average number of unique genomic fragments that can be expected to be discovered with increasing sequencing efforts. The prediction confidence range is also provided.

Supplementary Figure 2. Probe size distributions.
Panel A: Size distribution of expected DNA templates following in Silico double restriction of the donkey (top) and horse (bottom) reference genomes. Panel B: Size distribution of the probe genomic fragments. Paired-end sequence data generated from horse and donkey hyRAD probe libraries were mapped against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019). Panel C: Same as panel B, except that mapping was performed against the donkey reference genome (Wang et al. 2021).

Supplementary Figure 3. Probe base composition.
Panel A: %GC content of the genomic regions encompassing (Probes+) or not (Probes-) the hyRAD probes used for capture. Read alignment was carried out against both the horse

For Review Only

reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), and the donkey reference genome (Wang et al. 2021). Panel B: Same as panel A, except that the fractions of CpG dinucleotides within the genomic regions encompassing (Probes+) or not (Probes-) are provided.

Supplementary Figure 4. Ratios of ON-target and OFF-target read alignments.
Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were either shotgun sequenced (yellow) or subjected to capture with hyRAD horse (H-capture, blue) or donkey (D-capture, red) probes. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Panel A: Reads were aligned against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), and filtered for PCR duplicates and minimal mapping quality of 25. Panel B: Same as panel A, except that the sequence alignments against the donkey reference genome (Wang et al. 2021) were used. Resulting ON-target enrichment folds are provided on Figure 1 by dividing the number of read alignments located ON-target following capture relative to shotgun sequencing.

Supplementary Figure 5. Size and base composition of ancient DNA templates following shotgun sequencing and hyRAD capture.
Panel A: Size distribution of endogenous DNA templates. Panel B: %GC content of the genomic regions covered by endogenous DNA templates. Panel C: Proportions of CpG dinucleotides present in the genomic regions covered by endogenous DNA templates. Read pairs that could not be collapsed into full-length templates were not used for mapping and only unique, high-quality alignments were considered. The donkey reference genome (Wang et al. 2021) was used for read alignment. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

Supplementary Figure 6. CpG→TpG conversion rates.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG→TpG conversions report the relative fraction of read alignments in which a CpG dinucleotide is found in the donkey reference genome considered but a TpG dinucleotide is found in the sequence data. Calculations were conditioned on those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

Supplementary Figure 7. C→T conversion rates outside CpG dinucleotides.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG→TpG conversions report the relative fraction of read alignments in which a C is found in the horse reference genome considered but a T is found in the sequence data. Calculations were conditioned on non-CpG dinucleotide sites and those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes

For Peer Review Only

were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

Supplementary Figure 8. C→T conversion rates outside CpG dinucleotides.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG→TpG conversions report the relative fraction of read alignments in which a C is found in the donkey reference genome considered but a T is found in the sequence data. Calculations were conditioned on non-CpG dinucleotide sites and those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

Supplementary Figure 9. Proportions of horse-specific and donkey-specific alleles identified following shotgun and hyRAD-capture.
Panel A: Alignments against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-contigs from Felkel and colleagues (Felkel et al. 2019). Panel B: Alignments against the donkey reference genome (Wang et al. 2021). Calculations indicate the difference between the number of reads carrying horse-specific and donkey-specific alleles, respectively. Both transversion and transition alleles were considered. Positive (negative) fractions indicate an over-representation of horse-specific (donkey-specific) alleles in the sequence data. Calculations were conditioned on those read alignments against the reference genome that were located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

For Peer Review Only

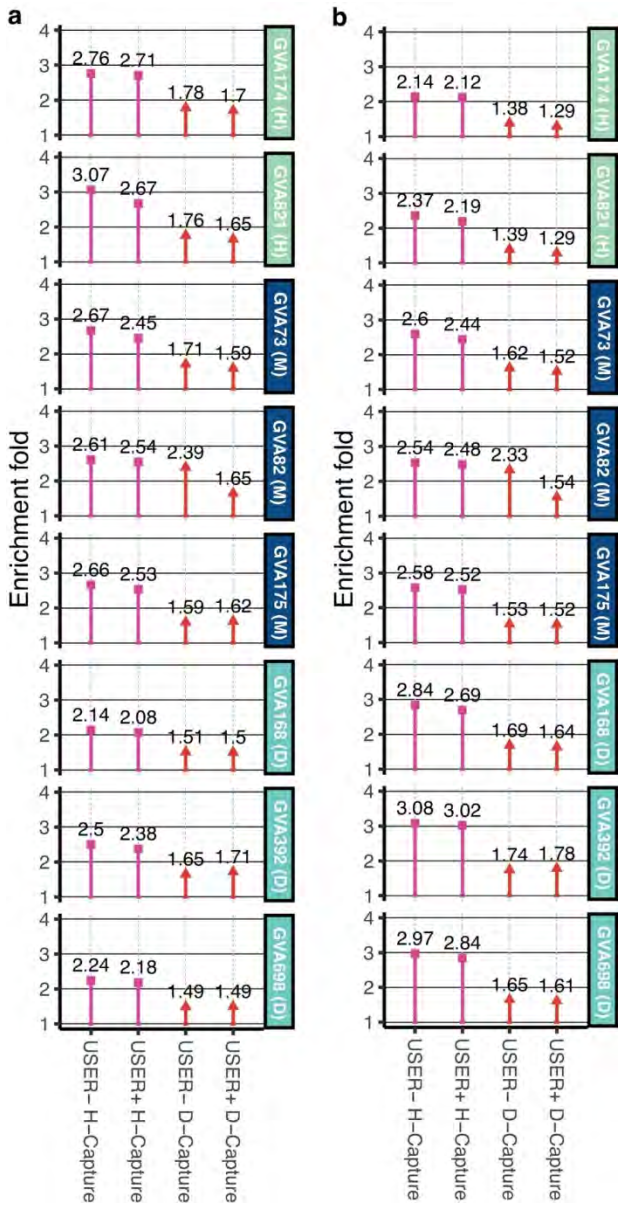


Figure 1. ON-target enrichment-folds.
Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were subjected to capture with hyRAD horse (H-capture) or donkey (D-capture) probes. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Panel A: Reads were aligned against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), and filtered for PCR duplicates and minimal mapping quality of 25. Panel B: Same as panel A, except that the sequence alignments against the donkey reference genome (Wang et al. 2021) were used.

75x148mm (300 x 300 DPI)

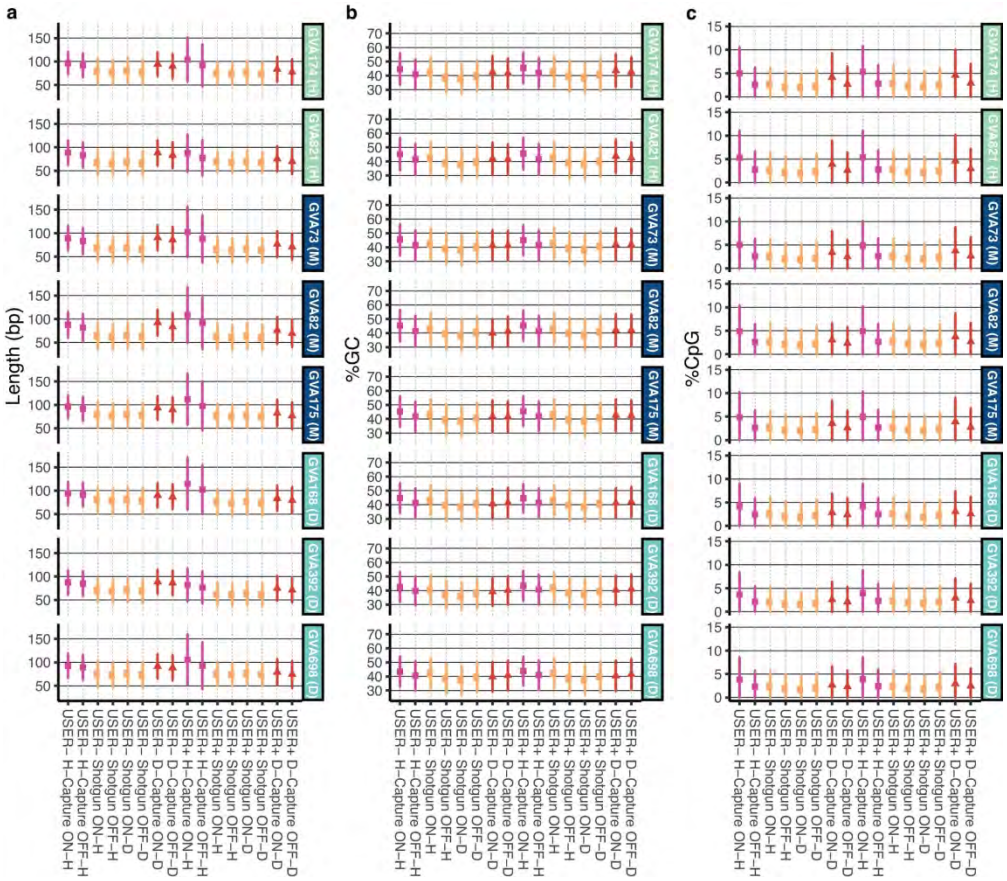


Figure 2. Size and base composition of ancient DNA templates following shotgun sequencing and hyRAD capture.

Panel A: Size distribution of endogenous DNA templates. Panel B: %GC content of the genomic regions covered by endogenous DNA templates. Panel C: Proportions of CpG dinucleotides present in the genomic regions covered by endogenous DNA templates. Read pairs that could not be collapsed into full-length templates were not used for mapping and only unique, high-quality alignments were considered. The horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), was used for read alignment. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

165x144mm (300 x 300 DPI)

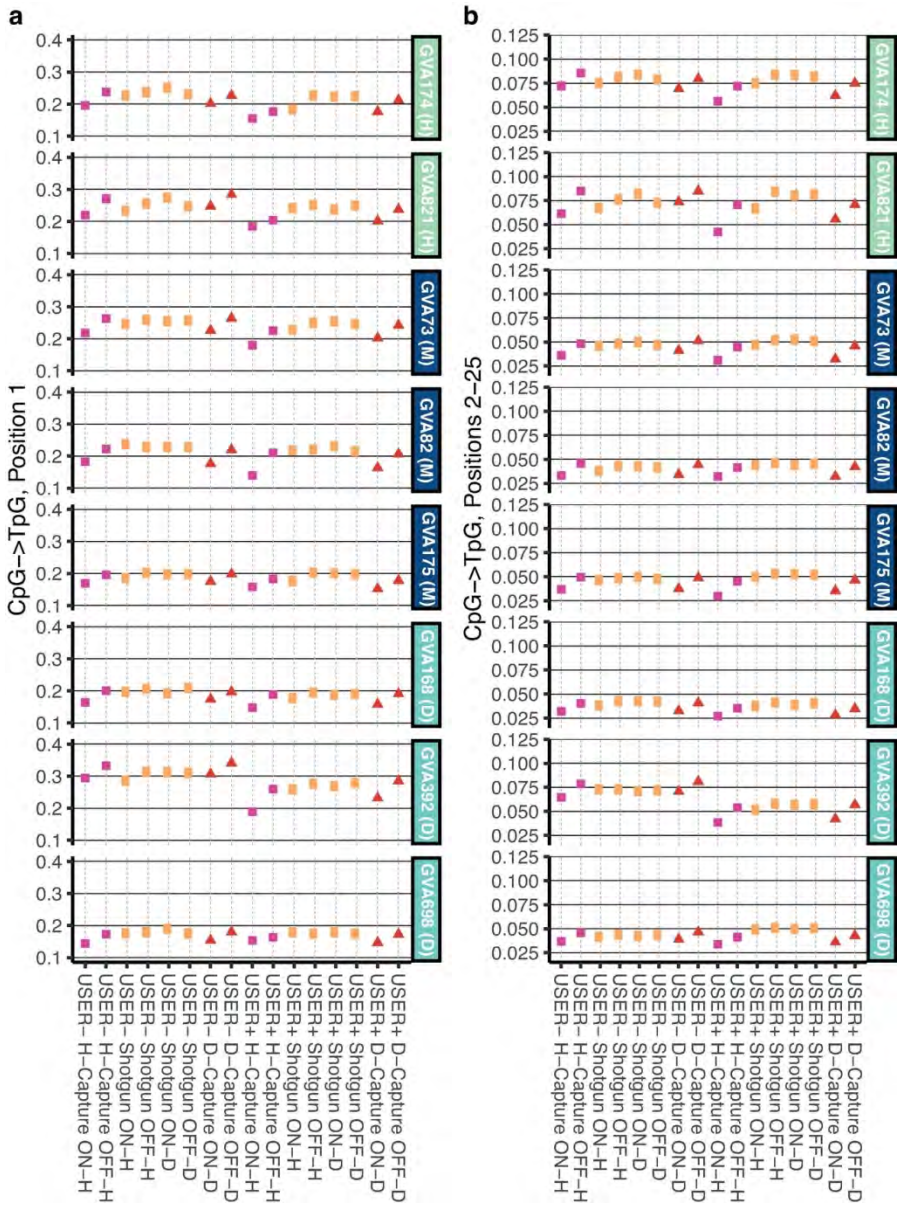


Figure 3. CpG□TpG conversion rates.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG□TpG conversions report the relative fraction of read alignments in which a CpG dinucleotide is found in the horse reference genome considered but a TpG dinucleotide is found in the sequence data. Calculations were conditioned on those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

109x147mm (300 x 300 DPI)

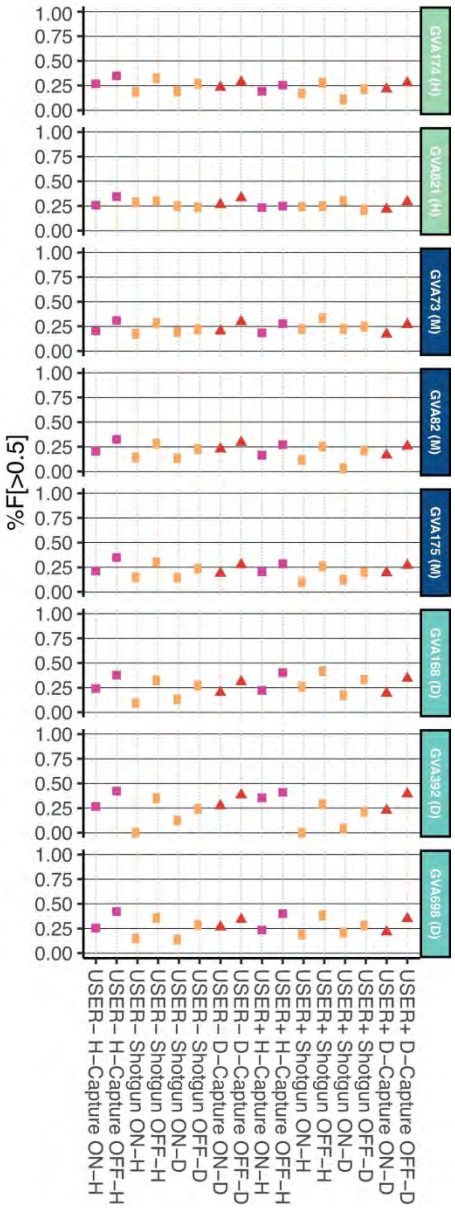


Figure 4. Proportion of CpG dinucleotides showing F-methylation levels above 50%. F-methylation levels were calculated using DamMet (Hanghøj et al. 2019) with default parameters. Calculations were conditioned on those read alignments against the reference genome that were located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

78x211mm (300 x 300 DPI)

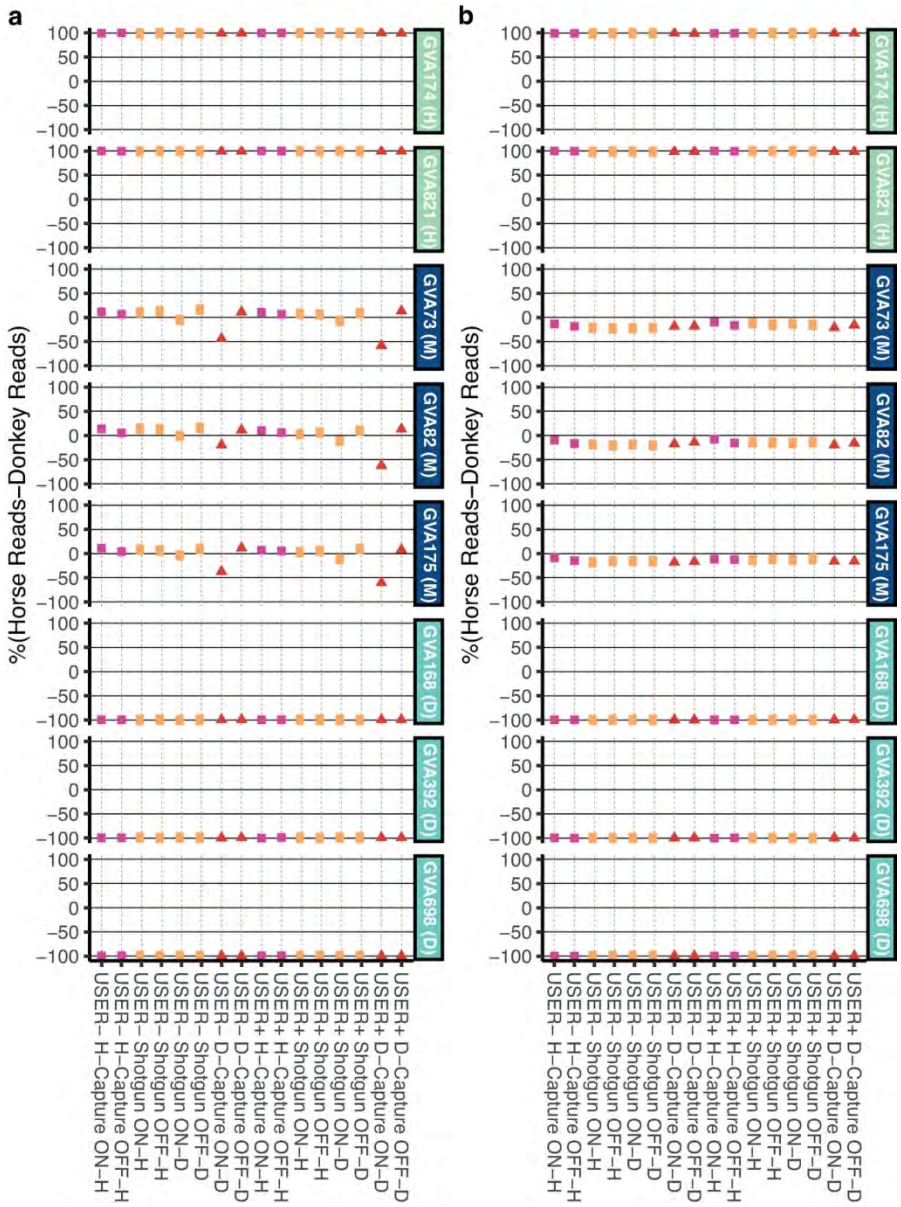


Figure 5. Proportions of horse-specific and donkey-specific alleles identified following shotgun and hyRAD-capture.

Panel A: Alignments against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-contigs from Felkel and colleagues (Felkel et al. 2019). Panel B: Alignments against the donkey reference genome (Wang et al. 2021). Calculations indicate the difference between the number of reads carrying horse-specific and donkey-specific alleles, respectively. Only transversion alleles were considered. Positive (negative) fractions indicate an over-representation of horse-specific (donkey-specific) alleles in the sequence data. Calculations were conditioned on those read alignments against the reference genome that were located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.

IV. Experimental design for the reconstruction of ancient DNA methylation maps in horses

1. Introduction

With the advent of technological and molecular progress in ancient DNA (aDNA) research, the number of studies mentioning epigenetic marks have increased over the past few years (Orlando and Willerslev, 2014, Zhenilo et al., 2016, Seguin-Orlando et al., 2021). Among the several ancient epigenetic marks that have been studied from DNA traces preserved in archaeological and paleontological specimens (Gaffney et al., 2012, Pedersen et al., 2014, Gokhman et al., 2014, 2020, Hanghøj et al., 2016, 2019), one of the best understood is the DNA methylation. This epigenetic modification can be affected by inherited genetic variation, nutrition and environmental changes (Andersen et al., 2019, Portales-Casamar et al., 2016, Fagny et al., 2015, Tobi et al., 2018). In mammals, DNA methylation is found mainly at cytosines within CG dinucleotides (CpG sites). Being a covalent modification of DNA, cytosine methylation can be inferred by a variety of methods. Four popular protocols are using bisulfite treatment (BS) to detect methylation marks: reduced representation bisulfite sequencing (RRBS), whole-genome bisulfite sequencing (WGBS), targeted bisulfite sequencing (TBS), and Infinium BeadChip distributed by Illumina (Hanghøj and Orlando, 2018). The set of techniques declared above makes use of the differential sensitivity of cytosine and 5-methylcytosine towards chemical modification. This methodology is challenging and burdensome when applied to the whole genome but mostly at the level of bioinformatics analysis and cost per sample (Smith et al., 2015). Combined with High Throughput DNA Sequencing technologies (HTS), bisulfite sequencing can gives quantitative information on the ratio of methylated and un-methylated DNA molecules, and it can provide methylation maps at the single base level or in a given region of the genome if RRBS or TBS techniques are used. Otherwise by using WGBS method, all the available cytosines residues will be retrieved. These protocols can be used for mammals but also in other species where methylated alleles can be found in different sequence contexts.

With this study we wanted to investigate epigenetic markers in ancient horse populations following domestication processes by computationally design ancient horse DNA methylation

maps. We hypothesise that modifications in DNA methylation can play a role in the rapid adaptation of large vertebrate populations to environmental changes (Liu et al., 2020). This hypothesis is corroborated by differences that have been found in the methylome from different populations that survived major environmental shifts or that have been influenced by their habitat and lifestyle (Fagny et al., 2015, Tobi et al., 2018, Andersen et al., 2019, Liu et al., 2020).

For this study, we used a set of publicly available modern horse sequence data selected from the literature and produced using RRBS and, ancient genomic data produced within the laboratory following the USER treatment. With this, we focused on identifying genomic sites with potential impact on DNA methylation inference on modern and ancient horse datasets. We wanted to see whether the estimates as recovered from an ancient bone, correlate to those retrieved in similar fresh tissues. Following the literature, the procedures selected to study each of the two datasets currently provide the best source of information to obtain CpG methylation level on ancient DNA data (Smith et al., 2015, Hanghøj and Orlando, 2018 for a review). Such procedures could allow the characterisation of high-quality genome-wide methylation maps and then, it could later help reconstructing extinct horse methylomes. However, instead of characterising the epigenetic markers in ancient horse populations following domestication processes, this work faced multiple experimental and analytical limitations that make difficult the production of consistent estimates of ancient methylation marks.

2. Material and Methods

DNA sequence datasets

Previously published and unpublished raw sequence data from twelve modern horse individuals were downloaded from the European Nucleotide Archive (Table 1). The modern sample set consists of a mix of soft and osseous tissues extracted from different young mares (Heart, Liver, Lung, Sesamoid, Metacarpe and Blood), each of the six samples have one replica. Except for the white blood cells samples (Blood_1, Blood_2, Ząbek et al., 2019), the other samples are unpublished and reported by the Functional Annotation of Equine Genome consortium (FAANG). For all the modern samples, isolated Illumina DNA sequences were generated for RRBS library production with methylation-insensitive restriction enzyme MspI

to generate short fragments since only a fraction of the genome is covered with the RRBS procedure. As a result, the unmethylated cytosines are sequenced as thymine analogues, whereas the methylated cytosines stay identical. Bisulfite treatment followed by PCR amplification gives quantitative information on the ratio of methylated and unmethylated DNA molecules and provides the opportunity to assess methylation maps at the single-base level.

In addition to this modern dataset, one ancient sample has been added within this study for later comparison ("RusxNx53"). It consists of a petrous bone extracted from a female horse radiocarbon-dated to the early bronze age (4195 y BP) and excavated in Turganik, Russia. This sample has been selected due to its high coverage (~21X) as this ensured that sufficient sequence data could characterise methylation inference (Hanghøj et al., 2019). Indeed, in a recent paper the author suggests at least 20X coverage to estimates methylation inference with confidence (Hanghøj et al., 2019). Furthermore, recent genomic analysis within the laboratory has suggest that this sample is carrying more DOM2 (modern horses of the second domestic clade) genetic affinity than it should be for a wild horse originating from this site. Then, this strongly suggests than this sample has undergone early horse management and herding practices, which is instrumental to study differentiated methylation marks (Librado et al., 2021). Illumina DNA sequences coming from this sample were generated for libraries prepared following the USER enzymatic mix (Rohland et al., 2015). As a result, the number of post-mortem damages accumulated in ancient DNA templates due to deamination of cytosine residues through time is reduced. These data allow us to further conduct methylation analyses.

Read processing and alignment methods

The processing of raw modern DNA sequence data was handled to generate high-quality sequence reads before the alignment. The trimming of low-quality bases on sequence ends, and sequence adapter (MspI, 5-CCGG-3') has been managed by AdapterRemoval2 (Schubert et al., 2016, "--basename output_single --adapter1 AGATCGGAAGAGC --trimns --trimqualities --gzip --threads 16 --minlength 30 –qualitybase 33"). Then, a further 2 bp removal from the 3' and 5' ends has been done to avoid using artificial methylation calls from the filled-in cytosine positions close to the MspI site in sequenced fragments. This step has been recommended by the trim_galore script (Krueger, 2015).

Name	Tissue	ENA study accession	ENA run accession	Assay type	Instrument model	Library layout	Sex
Heart_1	Heart Left Ventricle 1	PRJEB32645	ERR3329199	RRBS	Illumina HiScanSQ	Single	F
Heart_2	Heart Left Ventricle 2	PRJEB32645	ERR33291200	RRBS	Illumina HiScanSQ	Single	F
Liver_1	Liver left lateral lobe 1	PRJEB32645	ERR33291201	RRBS	Illumina HiScanSQ	Single	F
Liver_2	Liver left lateral lobe 2	PRJEB32645	ERR33291202	RRBS	Illumina HiScanSQ	Single	F
Lung_1	Left lung 1	PRJEB32645	ERR33291205	RRBS	Illumina HiScanSQ	Single	F
Lung_2	Left lung 2	PRJEB32645	ERR33291206	RRBS	Illumina HiScanSQ	Single	F
Sesamoid_1	Forelimb sesamoid bone 1	PRJEB32645	ERR33291211	RRBS	Illumina NextSeq 500	Single	F
Sesamoid_2	Forelimb sesamoid bone 2	PRJEB32645	ERR33291212	RRBS	Illumina NextSeq 500	Single	F
Metacarpe_1	Cannon bone diaphysis forelimb 1	PRJEB32645	ERR33291213	RRBS	Illumina NextSeq 500	Single	F
Metacarpe_2	Cannon bone diaphysis forelimb 2	PRJEB32645	ERR33291214	RRBS	Illumina NextSeq 500	Single	F
Blood_1	White blood cells 1	PRJNA517684	SRR8502735	RRBS	Illumina HiScanSQ	Single	F
Blood_2	White blood cells 2	PRJNA517684	SRR8502736	RRBS	Illumina HiScanSQ	Single	F

Table 1. Modern samples and sequences information

All the information registered within this table are provided with respect to the European Nucleotide Archive (ENA) reporting the DNA sequence data. The names, type of tissues, ENA study accession and instrument model of each modern DNA specimen considered within this study are indicated. With the exception of the white blood cells samples, the other samples are unpublished and reported by the Functional Annotation of Equine Genome consortium (FAANG).

The next step is the mapping of modern genomic DNA treated with bisulfite agent. A specific method capable of handling potential C to T conversions is required to estimate methylation levels since simply aligning these BS-seq by using standard aligners results in poor mapping efficiency (Krueger and Andrews, 2011, Tran, 2018). Many bisulfite sequencing alignment methods exist today and rely primarily on two types of strategies. The first method relies on an in-silico C to T conversion strategy with, for example, Bismark (Krueger and Andrews, 2011), MethylCoder (Pedersen et al., 2011), Bison (Ryan and Ehninger, 2014), BRAT-BW (Harris et al., 2012) and BS-seeker2 (Guo et al., 2013). In contrast, other methods replace the C found within the reference genome with a wildcard letter and tolerate both Cs and Ts in the sequencing reads. Such approaches are used with BSMAP (Xi and Li, 2009), Last (Frith et al., 2012) and RRBSMAP (Xi et al., 2012). A few comprehensive benchmarking studies of these softwares exist for mapping bisulfite data (Tran et al., 2014, Kunde-Ramamoorthy et al., 2014, Tsuji and Weng, 2016, Sun et al., 2018, Grehl et al., 2020, Nunn et al., 2021). They compare

different features such as the effects of the bisulfite conversion rate, the sequencing error rate, the run time, the memory consumption, the maximum number of mismatches allowed, the number of uniquely mapped reads and the types of output. None of these studies was decisive in defining the best aligner for general bisulfite sequencing. For our work, we decided then to choose the aligner according to the run time, the memory, the sensitivity, the number of parameters available and the possibility to use the output easily to conduct further analyses. Based on this, the BSMAP algorithm v2.89 (Xi and Li, 2009) has been selected to align bisulfite sequences against the horse reference genome (EquCab3, Kalbfleish et al. 2019, "-s 12 -D C-CGG -p 8 -q 30 -n 1 –v 0.08"). This algorithm was also used, with different parameters, for the white blood cells samples that we are analysing (Ząbek et al., 2019), which would give a way to cross validate results. We have selected specific RRBS parameters, -D, which activates the RRBS mode and sets the restriction enzyme, as well as -s which is the seed size in a RRBS mode and -q for the quality threshold. Finally, the -n option corresponds to the mapping strand information (1 is for paired end reads) and -v is giving the maximum number of mismatches allowed. We followed the software's recommendation to choose the -v 0.08 which depends on the read length. An additional quality filtering step was carried out to remove alignments showing mapping qualities inferior to 25 with Samtools.

The processing of raw ancient DNA sequences data corresponding to RusxNx53 was handled using Paleomix v1.2.13.2 (Schubert et al., 2014). This bioinformatic pipeline can efficiently perform several read processing steps starting from the trimming of specified adapter sequences from single-ended and paired-ended sequences with AdapterRemoval2. This pipeline can also eliminate low-quality base calls, ambiguous bases and collapse overlapping read pairs creating a single "collapsed" sequence showing a lower error rate. After these preliminary steps, very short sequences have been eliminated (<25bp by default) to reduce the fraction of wrong alignments. The resulting sequences were then aligned against the horse reference genomes (EquCab3, Kalbfleish et al. 2019) using Bowtie2 v2.3.5.1 end-to-end sensitive (Langmead et al. 2012) and Paleomix. Read alignments were carried out following the optimised parameters from Poullet & Orlando (2020). An additional quality filtering step was carried out to remove PCR duplicates with Picard MarkDuplicates and alignments showing mapping qualities inferior to 25 were discarded. Finally, nucleotide misincorporation and DNA fragmentation patterns are quantified with mapDamage2.0 (Jónsson et al., 2013) and used for local realignment. The files obtained from Paleomix were then processed.

Validation of RRBS results

After the alignment of the modern data, the output BAM read files were processed for several analyses. First, base-by-base methylation calling was calculated from BSMAP mapping results with at least twenty reads of coverage. The BED coordinates of each CpG dinucleotide were obtained using Seqkit v0.3.1.1 (Shen et al., 2016) conditioned on the horse reference genome (EquCab3, Kalbfleish et al. 2019). For each CpG site, the methylation ratio was estimated as the percentage of cytosines aligned to a given genome location with an actual C base in the reads. Due to bisulfite treatment, methylated Cs are intact. The percentage will then provide a methylation score on that particular base. The corresponding 95% confidence interval for the ratio has been calculated by Wilson score for a binomial proportion. Validation and statistical tests were carried out using the Methylkit package available in R (Akalin et al., 2012, R core Team 2014), and plots were generated using the R ggplot2 package (Wickam, 2009). The Methylkit package can analyse and annotate DNA methylation information obtained by bisulfite sequencing, in particular RRBS. This package was used to assess the quality of NGS read alignment. Percent methylation distribution and read coverage per base were calculated for each CpG. The latter helps ensure that PCR duplicates bias is correctly removed. Bases with more than the 99.9^{th} percentile of coverage distribution or less than 20 reads were discarded. Following default parameters, coverage values were also normalised. Based on the similarity of their methylation profiles, a hierarchical clustering of the samples has been done using Canberra distance, which sorts each sample into groups according to how closely or distantly related they are to each other. The cluster tree was drawn using the ward method, which used a bottom-up approach. This method merges the pairs of data points with the smallest variance compared to other possible combinations.

DNA methylation calculations and filtering process

The processing of ancient BAM read alignment files obtained from Paleomix was carried out using the DamMet package v1.0.1 (Hanghøj et al., 2019). However, this computational approach can be sensitive to errors introduced during NGS read alignment due to very low coverage and post-mortem DNA damage. Additionally, it is known that sequencing errors can exceed nucleotide misincorporations in case of a base quality inferior to 25 (Schubert et al., 2012). In this way, to work with the least noisy data possible, we need to perform several

filtering steps and quality tests with DamMet. This package allows for DNA methylation inference, called f, for a specific coverage, within a given genomic region and including a pre-selected number of CpG dinucleotides. Thus, we tried two different set of tests. First, we down-sampled to a selection of different coverage with an increment of five within the 10-30 coverage range, then we conducted DamMet in windows of different sizes (10, 25, 50, 75, 85 and 100 base pairs) following recommendations of the authors to choose the interval (Hanghøj et al., 2019). For each sliding window, the corresponding genomic coordinates were provided to DamMet in the form of BED coordinate files. These files were obtained using Seqkit v0.3.1.1 (Shen et al., 2016) and a custom python script availing the itertools module and conditioned on the horse reference genome (EquCab3, Kalbfleish et al. 2019). Different comparisons were carried out between the RRBS data and the ancient sample based on logistic regression and the use of Fisher's test to find the best filter to see whether the estimates as recovered from an ancient bone, correlate to those retrieved in similar fresh tissues. Calculations and statistical tests were carried out using standard functions in R (R Core Team 2014), and plots were generated using the R ggplot2 package (Wickham 2009). Finally, a last filter has been implemented to remove the noise that could arise from analysing different types of RRBS tissues (osseous, blood and soft). The purpose of this filter is to recover a set of unique and non-overlapping RRBS positions. We considered as a high-quality set of DNA methylation values in fresh samples, those consistently estimated to similar levels across all tissues. Therefore, the ancient tissue should be expected to show similar values, providing an easy test for assessing the inference accuracy. To do this, methylated RRBS sites shared by at least one tissue in every pair of at least 4 or 5 out of 6 pairs have been extracted. For each selected site, we have calculated a methylation score (Ms), its respective confidence interval boundaries (lower: CI_low and upper: CI_up) and the difference between the low and upper bound (Diff = CI_up - CI_low). Then, the cumulative sum of the delta of the methylation score confidence interval (Σ_Diff_siteN = Diff_tissue1 + Diff_tissue2 + Diff_tissue3 +… Diff_tissue12) for each site N was calculated. In addition, the sites whose cumulative sum is less than the median of all cumulative sums were filtered. This thus makes it possible to keep the positions with the most reliable methylation score. Furthermore, overlapping positions have been filtered out with awk to inforce data independency. These specific genomic positions related to RRBS were retrieved from the ancient individual with a custom python script. DamMet was run on these sites with different parameters, including coverage (10, 15, 20, 25 and 30X) or window size around the site (5, 10, 20 or 50 bp). Finally, comparisons between moderns and ancient samples were carried out using the standard functions available in R.

3. Results

Overall alignment performance and DNA methylation inference for RRBS data

To help reconstruct extinct horse methylomes, use was made of modern horse data produced through Reduced Representation Bisulfite Sequencing (Table 1). This represented six types of tissues extracted from different young mares, including osseous and soft tissues. These samples were aligned against the horse reference genome following the specific BSMAP recommendations about RRBS sequencing (version 2.89, Xi and Li, 2009). This includes trimming RRBS adapters and low-quality sequences from the 3' end of reads and the number of mismatches allowed. After the alignment, different statistics about the methylation data, such as percent methylation (Figure 11) and coverage (Figure 12) have been calculated with a 95% confidence interval.

We first summarise previous findings with the observation of a U-shape typical of a bimodal distribution within the percent methylation histogram, e.g. two peaks on both sides (Figure 11, Zhang et al., 2017) showing the two states of cytosines, either methylated or not. All of the tissues investigated show this pattern (Figure 11), with a higher peak for fully methylated bases (17.2-24.8%) and a lower peak for fully unmethylated bases (10.3-15.6). However, the results are slightly different for the osseous tissues. Indeed, the difference in the methylation state seems more pronounced for them (sesamoid 1-2, metacarpe1-2), as well as the distribution of the methylation histogram, which is inverted (21.3-29.2% for fully methylated bases against 21.4-36.1% for fully unmethylated bases). Thus, more cytosines appeared unmethylated than methylated, which is the opposite for all the soft tissues, including the blood, often used as a proxy when comparing ancient and modern data.

We next assessed the quality of the RRBS data with the calculation of read coverage per base for each CpG (Figure 12). This measurement helps ensure that PCR duplicates bias has correctly been removed in every tissue. Bases with less than 20X coverage have been discarded by the methylkit R package (Akalin et al., 2012). Across all the modern samples, the distribution of the data stays identical, showing that these samples are not carrying any PCR duplication bias. Otherwise, the frequency distribution would show another peak for higher read coverage per base. The median coverage per base varies between 25X and 32X among the different tissues. Additionally, approximately half of the bases are located within the first bin (39.6-66.7%), meaning that these bases have between 20X and 25X read coverage per base.

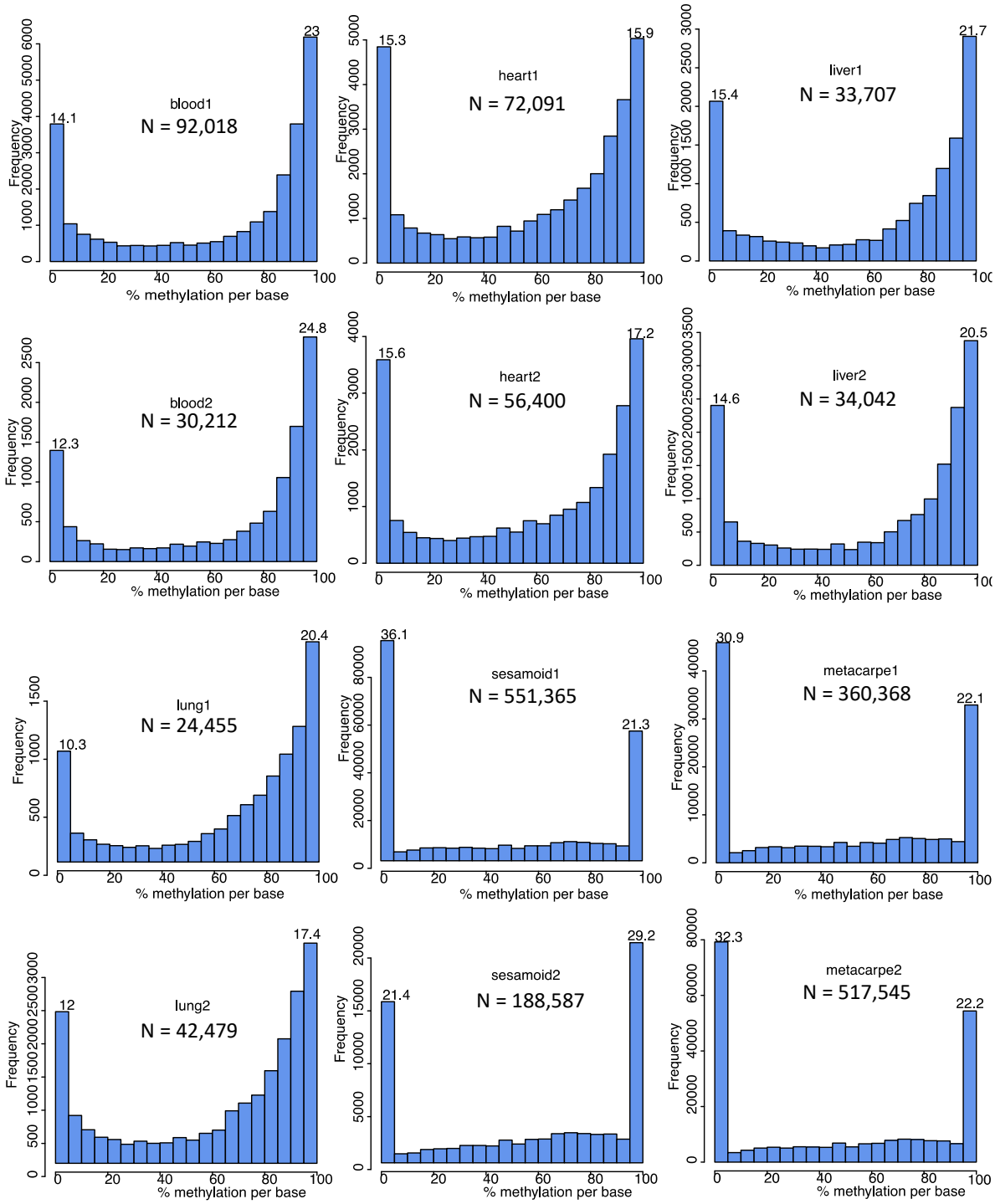


Figure 12: Percent methylation distribution across all 12 modern samples

Base-by-base methylation calling was calculated from BSMAP mapping results with at least twenty reads of coverage. The percent methylation score was estimated for each CpG sites as the percentage of cytosines aligned to a given genome location with an actual C base in the reads. The corresponding 95% confidence interval for the methylation ratio has been calculated by Wilson score for a binomial proportion. The numbers on bars denote what

percentage of sites are contained in that bin. N represents the total number of CpG sites available for this study.

Based on the similarity of their methylation profiles, a clustering of the samples has been done. Many different algorithms can achieve a clustering. Within this study we used a hierarchical clustering based on Canberra distance and Ward's methods (Figure 13). Interestingly, with our data, the first dendrogram shows expected results except for the osseous tissues (Figure 13a). Indeed, all the soft tissues are clustered with their respective replicate on one branch, which is not the case for the osseous tissues separated on another branch. Instead of being merged by tissue, the osseous samples clustered by method and with an equivalent height. Furthermore, the blood tissue is also the closest to osseous tissues. When adding the ancient sample, RusxNx53, to the comparison, the clusters are different (Figure 13b). As expected, the ancient bone sample is clustered with the majority of the modern osseous tissues, which seems to confirm our prediction on methylation inference for ancient and modern DNA data. However, the location of all the other modern samples is completely modified and does not fit anymore with a biological signal. It is then necessary to adjust the method used to design and filter the methylation maps.

Accounting for limitations to reconstruct reliable methylation maps

Different filters (coverage, removal of outliers, sliding windows) more or less strict have been applied to our data to remove outliers that could arise from analysing different types of RRBS tissues (Figure 14). The purpose is the extraction of a set of unique, non-overlapping positions that allow all pairs of somatic samples to make the same contribution to methylation. We have, first, compared the distribution of the coverage (Figure 14a). Indeed, the higher the coverage values for every test, the better the comparison between each RRBS tissue (Supplementary Figure1). Furthermore, for an identical coverage (Supplementary Figure 2), depending on the number of outlier tissues removed, the correlation value between each RRBS tissue is modified. With only two outliers discarded (Blood2, Sesamoid2, mean coverage = 0.901) the correlation value is lower than when three outliers (Blood2, Sesamoid2, Lung1, mean coverage = 0.961) are discarded (Supplementary Figure 2, Figure 14b). Then, to decide which coverage to use, we looked first the number of resulting sites (Figure 15a). For a coverage of 30, the number of positions available is 972. For 25X and 20X, respectively, 1302 and 1925 sites are available. Unfortunately, after filtering for the highest coverage, the number of remaining sites is not enough to pursue the analysis. When comparing the 20X

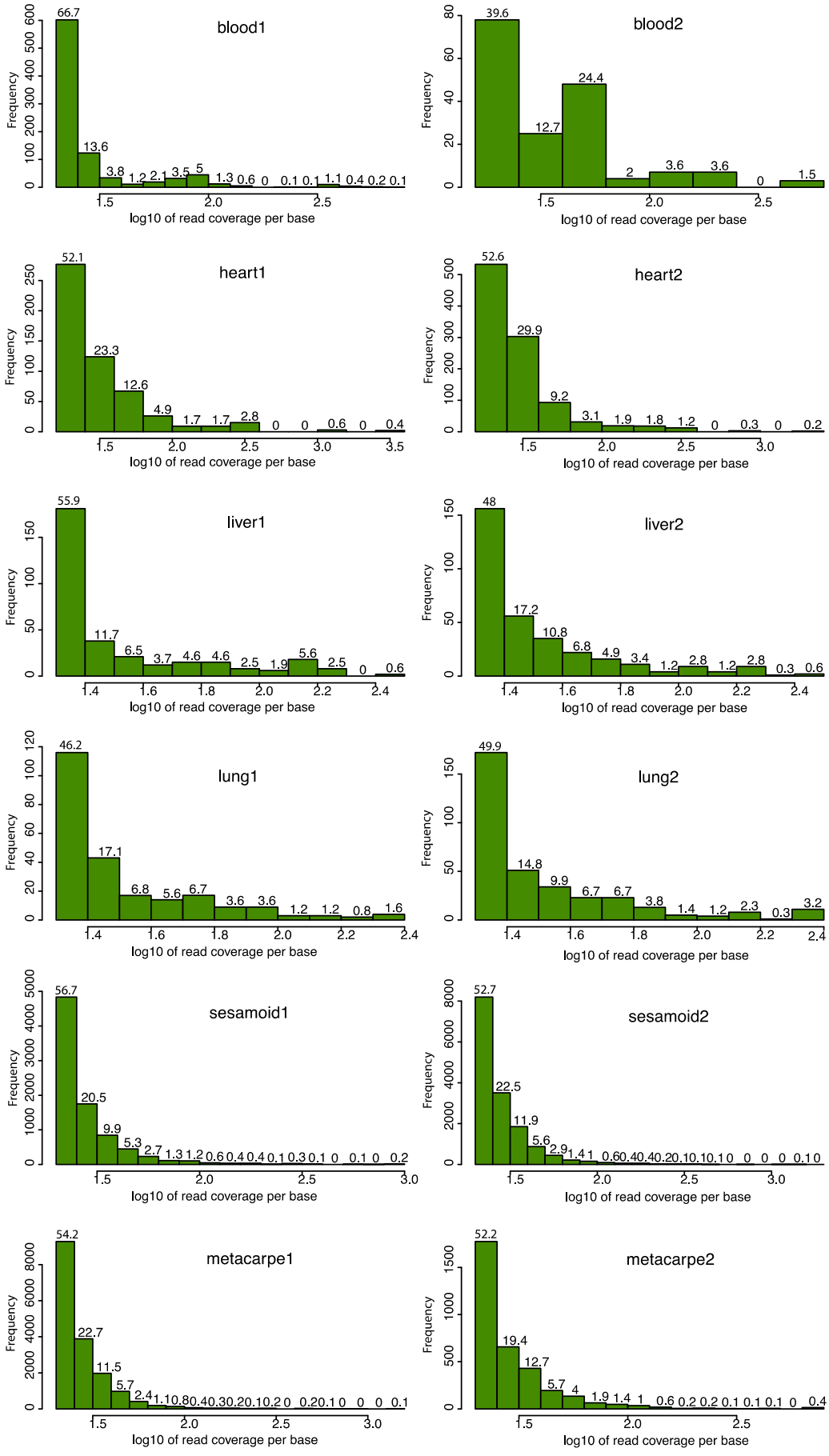

blood1
Frequency
log10 of read coverage per base
66.7
13.6
3.8
1.2
2.1
3.5
5
1.3
0.6
0
0.1
0.1
1.1
0.4
0.2
0.1
blood2
39.6
12.7
24.4
2
3.6
3.6
0
1.5
heart1
52.1
23.3
12.6
4.9
1.7
1.7
2.8
0
0
0.6
0
0.4
heart2
52.6
29.9
9.2
3.1
1.9
1.8
1.2
0
0.3
0
0.2
liver1
55.9
11.7
6.5
3.7
4.6
4.6
2.5
1.9
5.6
2.5
0
0.6
liver2
48
17.2
10.8
6.8
4.9
3.4
1.2
2.8
1.2
2.8
0.3
0.6
lung1
46.2
17.1
6.8
5.6
6.7
3.6
3.6
1.2
1.2
0.8
1.6
lung2
49.9
14.8
9.9
6.7
6.7
3.8
1.4
1.2
2.3
0.3
3.2
sesamoid1
56.7
20.5
9.9
5.3
2.7
1.3
1.2
0.6
0.4
0.4
0.1
0.3
0.1
0
0.1
0
0.2
sesamoid2
52.7
22.5
11.9
5.6
2.9
1.4
1
0.6
0.4
0.4
0.2
0.1
0.1
0.1
0
0
0
0
0.1
0
metacarpe1
54.2
22.7
11.5
5.7
2.4
1.1
0.8
0.4
0.3
0.2
0.1
0.2
0
0.2
0.1
0
0
0
0.1
metacarpe2
52.2
19.4
12.7
5.7
4
1.9
1.4
1
0.6
0.2
0.2
0.1
0.1
0.1
0
0.4

Figure 13: Read coverage per base across all 12 modern samples

Bases with less than 20X coverage have been discarded by the methylkit R package. All the RRBS tissues are represented. This measurement helps ensure that PCR duplicates bias has correctly been removed in every tissue. The median coverage per base varies between 25X and 32X among the different tissues.

and 25X coverage distribution (Figure 14a), the 25X coverage curve (mean = 1.0303, standard deviation = 0.3024) appears to follow a more Gaussian distribution than 20X (mean = 1.1757, standard deviation = 0.3770), with a more centred curve and less dispersed values. Thus, we could hypothesise that, the methylation inference of sites estimated from a 25X coverage filtering step are more reliable than the others since their confidence interval is smaller. Furthermore, to decide how many outliers we should remove, we also compared the distribution of the cumulative sum of the delta of the methylation score confidence interval (Σ_Diff_siteN, Figure 14b). When removing three outliers (mean = 1.1757, standard deviation = 0.3770) instead of only two (mean = 1.5310, standard deviation = 0.6255), the curve is more centred and reduced, indicating that we need to remove three outliers. We then extracted the sites with a coverage of 25X and without the three tissues outliers (Blood2, Sesamoid2, Lung1). This gives 1302 CpG positions (Figure 15a). Additionally, each sample has been compared with all the sites (Figure 14c) and only with the 1302 CpG extracted sites (Figure 14d). In both cases, we can observe that our three samples (Blood2, Sesamoid2, Lung1) behave like outliers by having higher median values than the other samples (non-outliers group: median = 0.112, standard deviation = 0.047, outliers group: median = 0.141, standard deviation = 0.060). A correlation matrix between all the remaining sites and samples has been processed (Figure 15b). All the coefficient correlation values are superior to 0.9 and even with these close values we can observe some sub-family between soft tissues, bones and blood. This filter does not seem to have been sufficient to really erase the inter-tissue differences.

These chosen sites have been used to run DamMet with additional parameters and comparisons between Ancient and Modern samples have been plotted (Figure 16, Supplementary Figure 3). Different coverages have been tested as well as different window sizes around the chosen site. For a window size of 100 centred around the site, RusxNx53 seems to be clustered with one bone and it is close to the second one and the blood if we look at the hierarchical clustering. The correlation values are inferior to 0.70.

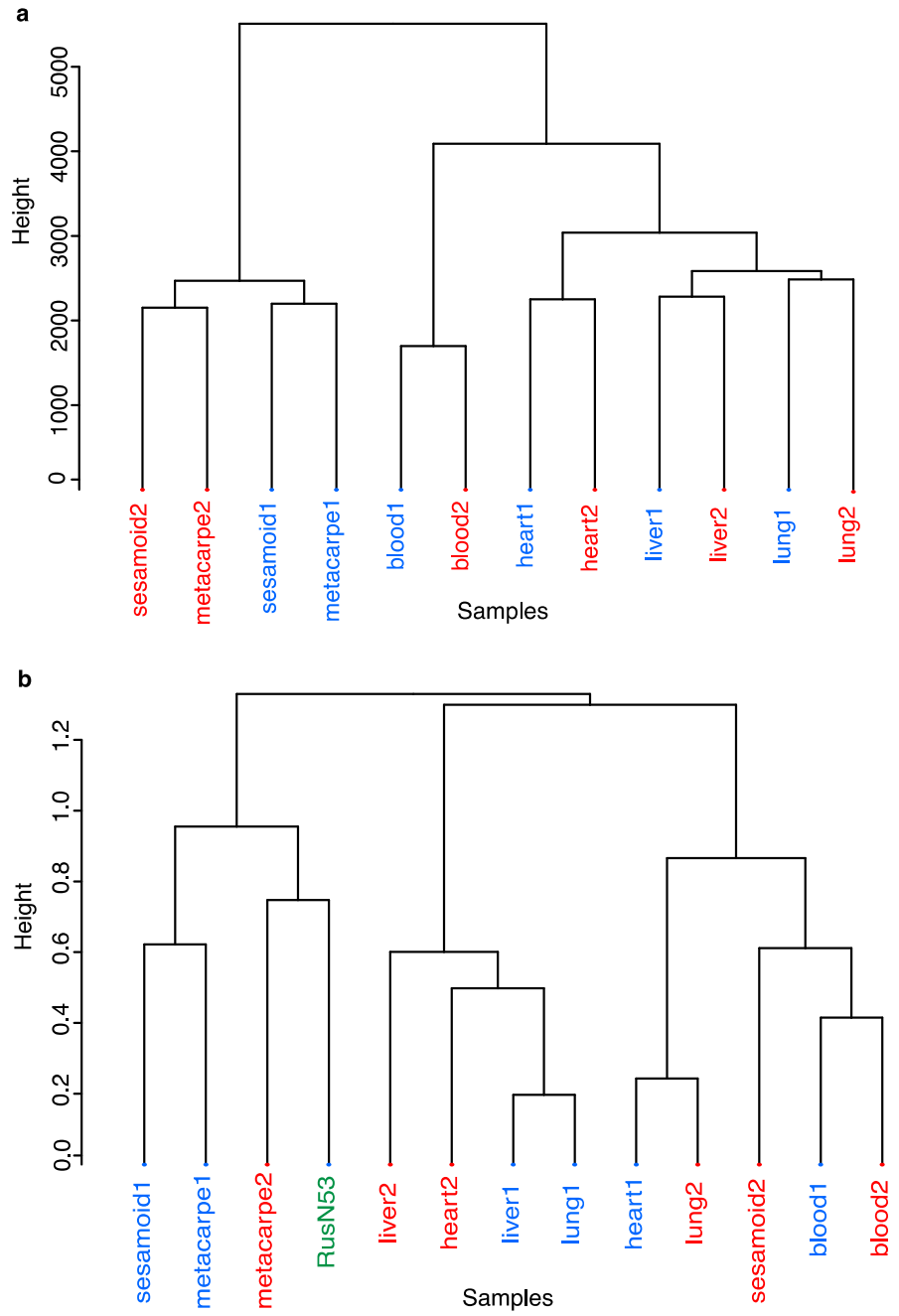


Figure 14: Hierarchical clustering for RRBS and ancient samples using Canberra distance and Ward's methods

(a) Hierarchical clustering across all pairs of RRBS samples. Except for the osseous samples clustered by replica, all other samples are merged with the expected tissue. The colours represent the type of replica. **(b) Hierarchical clustering across all pairs of RRBS samples and RusN53, the ancient sample**. As expected the ancient bone sample is clustered with the majority of the modern osseous tissues. However, the location of all the other modern samples is completely modified.

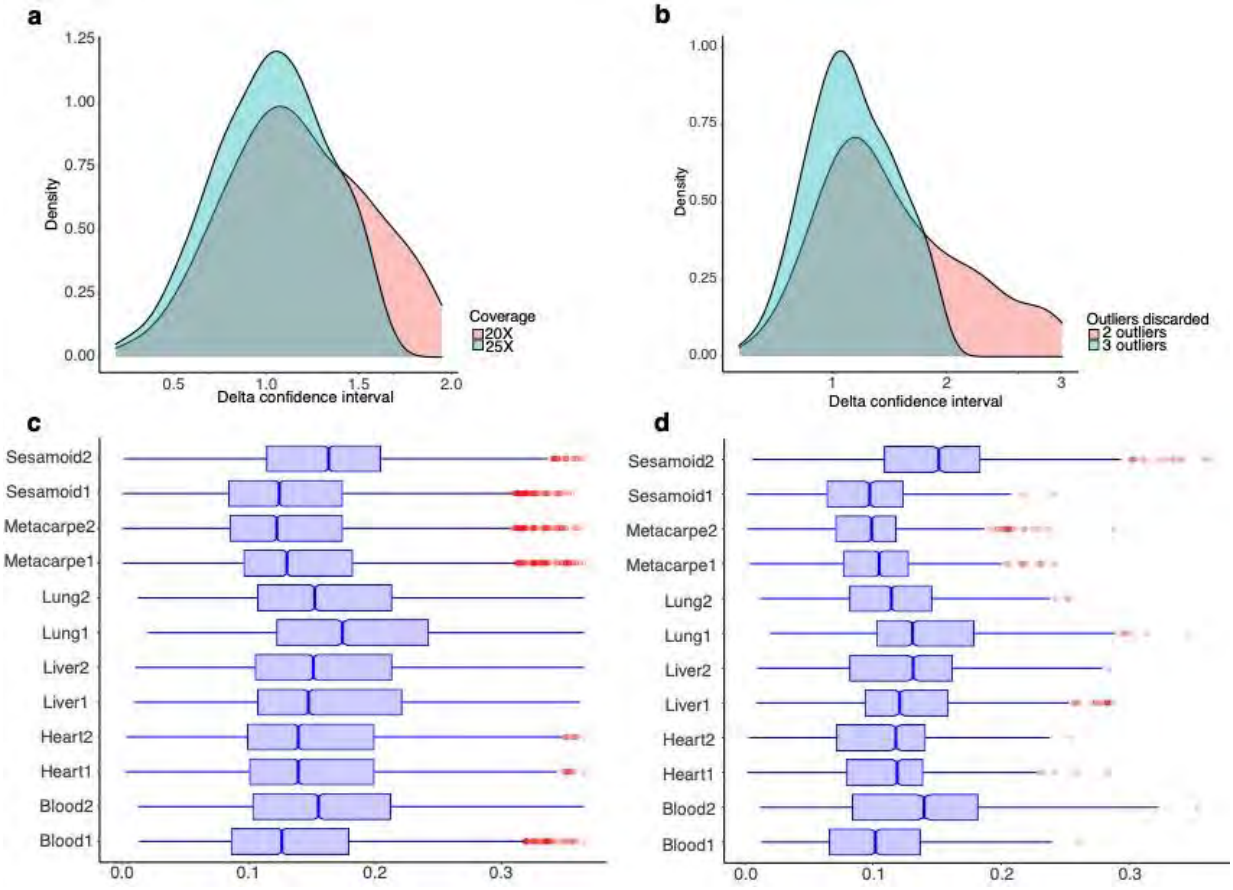


Figure 15: Filters applied to RRBS modern data throughout the extraction of a set of unique and non-overlapping positions.

(a) Distribution of the coverage. Comparison between 20X and 25X coverage for the delta confidence interval filter. The coverage was estimated after filtering alignments for quality scores inferior to 25. The distribution of the 20X coverage curve follows a more centred and less dispersed curve than the 25x coverage curve. **(b) Distribution of the difference observed between the cumulative sum of the methylation score confidence interval.** Comparison of the density curve when two or three outliers are removed from the analyses. The two outliers correspond to Blood2 and Sesamoid2. The three outliers correspond to Blood2, Sesamoid2 and Lung1. This set of filters have been applied to further analyses (25X coverage and three samples removed). **(c) Dataset distribution for all samples across all sites.** **(d) Dataset distribution for all samples across the selected sites.** 1302 CpG sites have been selected. The three outliers correspond to Blood2, Sesamoid2 and Lung1.

4. Discussion

Within this study, we wanted to investigate genomic sites with potential impact on DNA methylation inference and if possible, we wanted to computationally design ancient horse DNA methylation maps. Indeed, although this has been done for other ancient species (Pedersen et al., 2014, Gokhman et al., 2014, Hanghøj et al., 2016), to this day, there is no methylation map regarding horses. However, the characterisation of ancient DNA methylation patterns and their variation in ancient equine specimen seems to show certain experimental and analytical constraints relative to the methodology used.

The reconstruction of ancient DNA methylation patterns and their variation in ancient specimens require high quality and high coverage sequenced genomes to infer reliable epigenetic signals (Hanghøj et al., 2016). These constraints include also dealing with sequencing errors due to post-mortem deamination of cytosine into uraciles even after USER processing on DNA reads. Additionally, it requires dealing with the introduction of significant bias toward ancient DNA templates when using specific sequencing strategies (Suchan et al., submitted). To overcome some of these constraints, use was made of a set of different modern tissues treated with an RRBS treatment. The purpose being to see whether the estimates retrieved in this set of fresh tissues are correlated to those recovered from an ancient horse bone. To do so, we considered as a high-quality set of DNA methylation values to compare, the values consistently estimated to similar levels across all modern tissues.

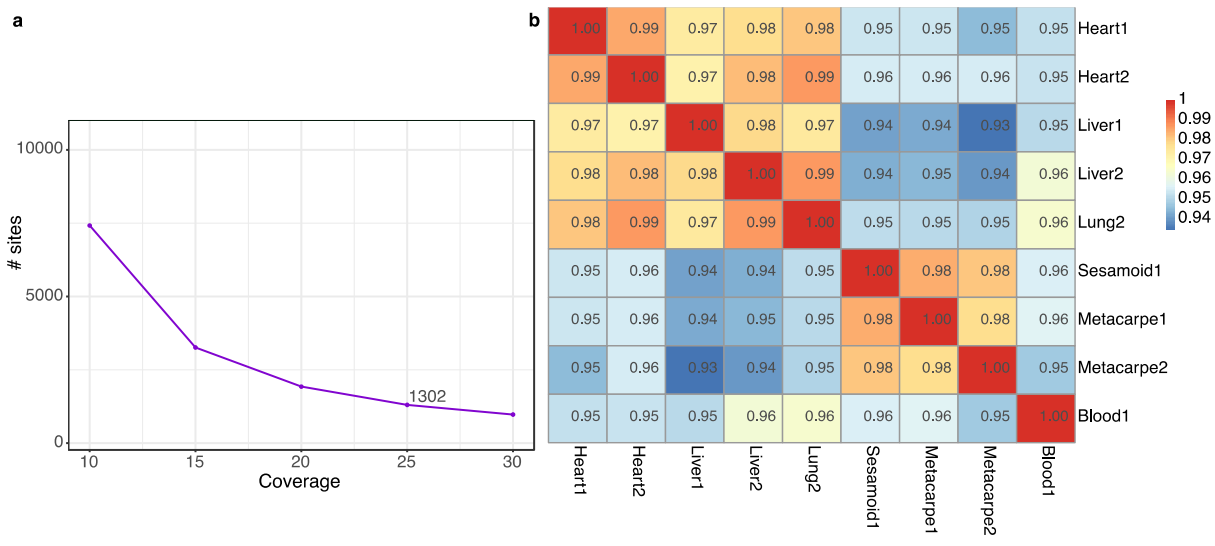


Figure 16: Impact of filtering process on modern DNA methylation inference

(a) Distribution of the number of sites available after the filtering process. The coverage was estimated after filtering alignments for quality scores inferior to 25. The selected coverage for further analysis is 25. This corresponds to 1302 CpG sites. **(b) Correlation matrix between the selected RRBS samples across the selected sites.** The three outliers have been removed. The correlation coefficient was based on Pearson correlation and the use of Euclidean clustering distance.

The validity of the RRBS data have been confirmed with the percent methylation score and read coverage per base analysed across all the twelve modern samples. Indeed, the bimodal distribution of the percent methylation score seems to correspond to what is expected. However, it is still possible to suspect errors in calculating this score that can be due to a sequencing error in high-throughput sequencing experiments, an incomplete bisulfite conversion or even the heterogeneity of samples. This question arises since, during the first hierarchical clustering experiment between all the modern samples and the ancient one, no soft tissues fit a biological signal. Thereby, different filters and parameters were implemented to remove a potential experimental noise coming from the ancient and modern samples and to assess whether better correlation levels could be retrieved between ancient and modern DNA methylation profiles.

First of all, we partially eliminated inter-tissue levels by removing some somatic positions in fresh tissues. However, to completely eliminate these inter-tissue levels, we need to filter many more positions. A first limitation is that the number of remaining CpG sites is too low with this data set to obtain effective results with DamMet (Hanghøj et al., 2019). Furthermore, we are looking for filtered CpGs positions to maximise a known biological signal within this set of tissues. Therefore, after performing the pairwise clustering, the resulted clusters should be first grouped by replica, then by tissues of the same biological origin and then finally by type of tissues (soft or osseous tissues). However, the filters used did not allow enough reliable results for the correlations between modern tissues and RusxNx53. Indeed, for the coverage, for example, the higher it was, the more the correlation values increased. of tissues. Therefore, after performing the pairwise clustering, the resulted clusters should be first grouped by replica, then by tissues of the same biological origin and then finally by type of tissues (soft or osseous tissues). However, the filters used did not allow enough reliable results for the correlations between modern tissues and RusxNx53. Indeed, for the coverage, for example, the higher it was, the more the correlation values increased. To potentially circumvent the above limitations brought by fresh tissues, a future approach could rely on *in silico* DNA

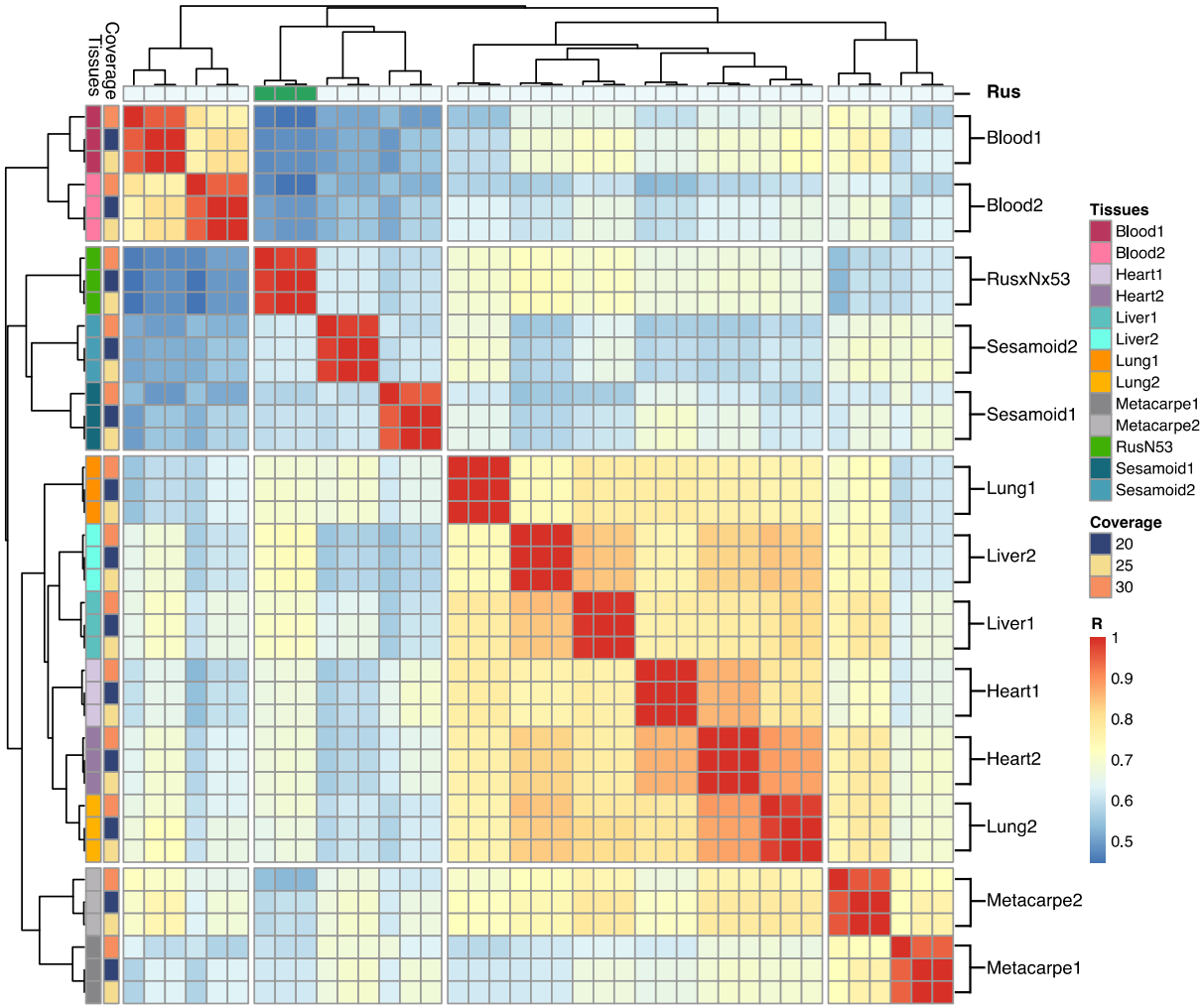


Figure 17: Comparison between all the modern samples and RusxNx53 for a window size of 100 base pairs.

Three different coverage have been tested (20X, 25X, 30X). All the RRBS tissues are tested against RusxNx53 (green). The methylation inference for the ancient sample has been calculated using the selected sites and DamMet. Different window sizes have been investigated for RusxNx53 (Supplementary Figure 3). Each window is centred around the pre-selected RRBS site.

sequence generation. Simulation of ancient DNA data is a powerful technology that could help to evaluate the validity of our methylation inference by reproducing most common ancient DNA features, such as contamination, nucleotide mis-incorporations and read length (Renaud et al., 2016). This could also help us to satisfy the required parameters, such as high quality

and high coverage sequenced genomes. However, in such a study, one limitation could remain in relation to the difficulty to properly simulate USER-treatment (Briggs et al., 2007, Seguin-Orlando et al., 2015b, Article 2). Especially if USER-treatment could enhance the false-positive rate of DNA methylation inference.

In addition, another limitation can be put forward, it concerns the age of the specimens. Indeed, previous studies already considerably linked changes to DNA methylation patterns with aging (Roforth et al., 2015). These specific changes between young and old specimens have been observed with Bisulfite sequencing method. Therefore, analysing individuals of unknown age at the time of death (as for RusxNx53) with individuals of different ages (as for all the modern samples) is an important limitation in this type of study. Finally, although the use of modern tissues was made for the unique purpose of serving as a proxy to obtain high-quality methylation maps in ancient individuals, given the results, it might have been necessary to recover more samples, more specifically of the osseous type. Indeed, the modern RRBS dataset mainly includes soft tissues from young mares. Unfortunately, only a few RRBS data were available on modern horses. Then, another option could be to explore more ancient specimens. In that way, the selected samples would have been more robust to detect epigenetic modifications as subject to similar bias.

V. Perspectives

aDNA analysis has come of age to have a broad scientific scope merging approaches in phylogenomics, population genomics, epigenomics and metagenomics. In addition, emerging technologies and extraction techniques have expanded the range of suitable samples, as they can achieve better mapping, better sensitivity and better performance at lower coverage, opening up exciting new opportunities for the ancient DNA field. Due to these significant improvements, ancient DNA studies are not limited to just fossilised remains and museum samples. Further methodological developments have been such that it is now almost routine to sequence genomes from ancient human individuals, animals, plants and even pathogens. However, improvements are still needed in some areas.

1. Improving ancient DNA read mapping

Depending on the nature of the research plan's objectives and the genetic material to align, the choice of genomic alignment software and parameters can be very different. For aDNA studies, the minute amounts, ultra-short, often contaminated and degraded nature of the genetic material poses a serious challenge (Pääbo et al., 1989). Indeed, sequence alignment algorithms already existing were implemented for longer and undamaged DNA fragments from modern samples. Using such software for analysing ancient samples can only act as a compromise. The shorter the reads, the greater the difficulty of aligning them at lengths typical for aDNA and the lower the accuracy and the reliability. Several studies (Schubert et al. 2012, Cahill et al. 2018, Renaud et al. 2018, Oliva et al., 2021) have already benchmarked the current gold-standard mapping strategies for aDNA studies and have given strategies and recommendations tailored for each type of study. These recommendations include using a USER treatment upstream, the removal of DNA fragments whose quality is too low and the use of parameters specific to each software. However, all of these strategies have the same disadvantage, which is the loss of mapped reads. Then, this loss can impact reference bias and downstream analyses such as D statistic, PCA or methylation calls (Oliva et al., 2021). Finding a trade-off between the quantity and the quality of mapped reads is essential to choose a good filter. In this way, since the growth of the field of palaeogenomics during the past decades has

been close to exponential, at present, it is possible to imagine the implementation of its own sequence alignment algorithms to study the large datasets of ancient genomes that starts to be available. Nevertheless, such implementation is not possible for everyone. Then, we must turn our attention to the new sequencing technologies nowadays available (van Dijk et al., 2018). Third generation sequencing methods could then appear to ancient DNA studies as a new game changer (Orlando et al., 2021). It seems difficult to imagine using long-read sequencers from Pacific Bioscience (PacBio) and Oxford Nanopore as the main characteristic of ancient DNA relies on its high level of fragmentation. However, using a combination of PacBio and Illumina technologies could produce genome assemblies of unprecedented quality and offer new opportunities for ancient DNA analyses. Furthermore, these third-generation sequencing methods can provide epigenetic information previously unapproachable to short-read sequencing technologies (Gouil and Keniry, 2019). New advances in reference-free mapping technology such as de novo assembling of ancient genomes and metagenomes opens up new doors to reliable assemble and reconstruct large ancient microbial genomes and metagenomes (Borry et al., 2021).

2. Detecting epigenetic marks

Molecular and computational advances in ancient DNA research have revealed that genome-scale epigenetic information can be retrieved from subfossil material. Some chromatin features and regulatory epigenetic markings present at the time of death can be preserved within ancient DNA extracts (Pedersen et al., 2014). However, as we have seen in previous chapters, only a small number of these epigenetic processes are studied in ancient DNA research (Pedersen et al., 2014, Hanghøj et al., 2016, 2019, Gokhman et al., 2017, 2020), most commonly including histone modifications (Bell et al., 2011), nucleosome positioning (Struhl and Segal, 2013), DNA methylation (Plongthongkum et al., 2014) and non-coding RNA interactions (Chen et al., 2016). New epigenetic methodologies are constantly being developed, improving our ability to understand the power of epigenetic in biological processes. They encompass techniques related but not limited to whole-genome sequencing, higher order chromatin analyses, ATAC-sequencing and high-throughput chromosome conformation capture (Hi-C, Davis et al., 2020). The latter technology could help us further understanding the three-dimensional organisation of genomes and the frequency of some interesting genomic loci interactions. This could also serve to establish the genomic topology and capture the physical structure of the chromatin, the network of histones and gene

regulatory mechanisms. Further investigations on these different epigenetic marks and methodologies could help characterise temporal paths from past individuals and domestic animals who lived prior to and following critical evolutionary transitions.

3. Accessing ancient metagenomes

In recent years, extensive studies have successfully characterised the genetic material of several ancient pathogens that may have infected individuals during life (Warinner et al., 2017, Spyrou et al., 2019) but also some microbial communities that can reveal the diet of extinct individuals (Adler et al., 2013, Mann et al., 2018, Borry et al., 2020). Indeed, aDNA extracts can play host to an entire metagenomic diversity of microorganisms that mainly colonised the material after death. In general, pathogenomics is primarily concerned with individual disease-causing microorganisms, including some of the deadliest pathogens in human histories, such as those causing plague (Y. pestis, Spyrou et al., 2018, Susat et al., 2021), tuberculosis (M. tuberculosis, Bos et al., 2014), malaria (P. falciparum, Marciniak et al., 2016) or leprosy (M. leprae, Roffey et al., 2017, Schuenemann et al., 2018), whereas microbiome studies focus more on the distribution and diversity of the microbes and microorganisms to a given host and their role in host functions. Recently, the etiological agent of the plague, Yersinia pestis, has been identified from a 5,000-year-old hunter-fisher-gatherer buried in Latvia (Susat et al., 2021) and predated the earliest historical record of plague in human populations by several thousand years. This very early Y. pestis form emerged ~ 7,000 years ago and coincided with the beginning of the Neolithic period in western Eurasia rather than with its decline, suggesting isolated zoonotic events as triggering factors (Rascovan et al., 2019, Susat et al., 2021). Such study shows the strength of this ancient DNA technology to detect ancient pathogens that do not leave indelible marks on bones.

The number of published metagenomes sampled directly from the environment (such as ancient ice core, lake sediments and surface soils, Pedersen et al., 2015) is also on the rise since the recent discovery of excellent microbial DNA preservation in ancient dental calculus (Mann et al., 2018). Due to recent advances in molecular and sequencing technologies, one small sediment sample can yield a comprehensive overview of a past ecosystem, indicating the presence of species from microbes to mammals. Recent work has demonstrated the practicability of sequencing and merging ancient Neanderthal mitochondrial and nuclear

DNA from three different Pleistocene cave sediments to get a fuller picture of human evolution (Vernot et al., 2021). Using a new hybridisation capture design that targeted regions of the hominid nuclear genome with high sequence divergence across mammals, they have been able to discriminate two distinct radiations of Neanderthal populations, the first ~ 100,000 years ago and the second ~ 135,000 years ago. They could be associated with changes in climate and environmental conditions during the last interglacial, but time-series data will be necessary to confirm it. Another study analysing 16S rRNA sequence for characterisation of the gut microbial community in domestic and Przewalski's horse has indeed shown that these horses have different and more diverse faecal microbiomes than domestic horses, but also than Przewalski's horses born in zoos (Metcalf et al., 2017). Future work carrying out a rigorous sampling of horse microbiomes through time might reveal whether the domestication process was followed by important dietary and/or microbial shifts.

Following such work, the study of coprolites or paleofeces reservoirs should be put forward. They are the nonmineralized remains of dung from extant and extinct fauna. They represent a surprisingly large proportion of fossil remains recovered from cave sites across the world and can contain the DNA of the defecator as well as the DNA of ingested plant and animal remains. A recent study has been able to accurately reconstruct known diets and habitats of the extant caribou using faecal samples of extinct megafauna (Polling et al., 2021). They have also extended this approach to Holocene and Pleistocene megafauna including horse, steppe bison and woolly mammoth thanks to the first integrated analysis of plant DNA, macrofossil and pollen from permafrost faeces. This showed, in contrast with previous reconstruction, the presence of a mosaic of habitats in the Pleistocene landscape. Another study has revealed the phylogenetic position of the extinct Shasta ground sloth (*Nothrotheriops shastensis*) from Pleistocene sloth coprolites in a Nevada cave (Poinar et al., 2003). This demonstrates that DNA from coprolites could also survive in specimen preserved in warm arid climates. This source opens up the possibility of detecting the presence of organisms even when no macroscopically identifiable evidence was present and at larger scale, can help to reconstruct entire paleo ecosystems, understand past hominin occupations and how they relate to their environment. Related to this, a recent and large-scale environmental DNA study from across the Arctic has reported the survival of horses and other megafaunal species into the Holocene in the Americas (Wang et al., 2021). This strongly challenges the cause of megafaunal extinctions since it suggests that humans and megafaunal species have coexisted for thousands of years. Such study truly highlights the power of ancient environmental DNA to reconstruct population histories and biotic interactions (Wang et al., 2021). In a similar way, focusing on

mapping past population and extinction dynamics through space and time in the absence of bones could truly advance our understanding of the horse domestication process (Hayle et al., 2009, Wang et al., 2021), which is poorly documented yet in some regions that proved difficult to obtain bone material from (e.g. the Balkans, China and more).

VI. Appendix

Appendix 1: Supplementary information - Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes

---

Supplementary Figures

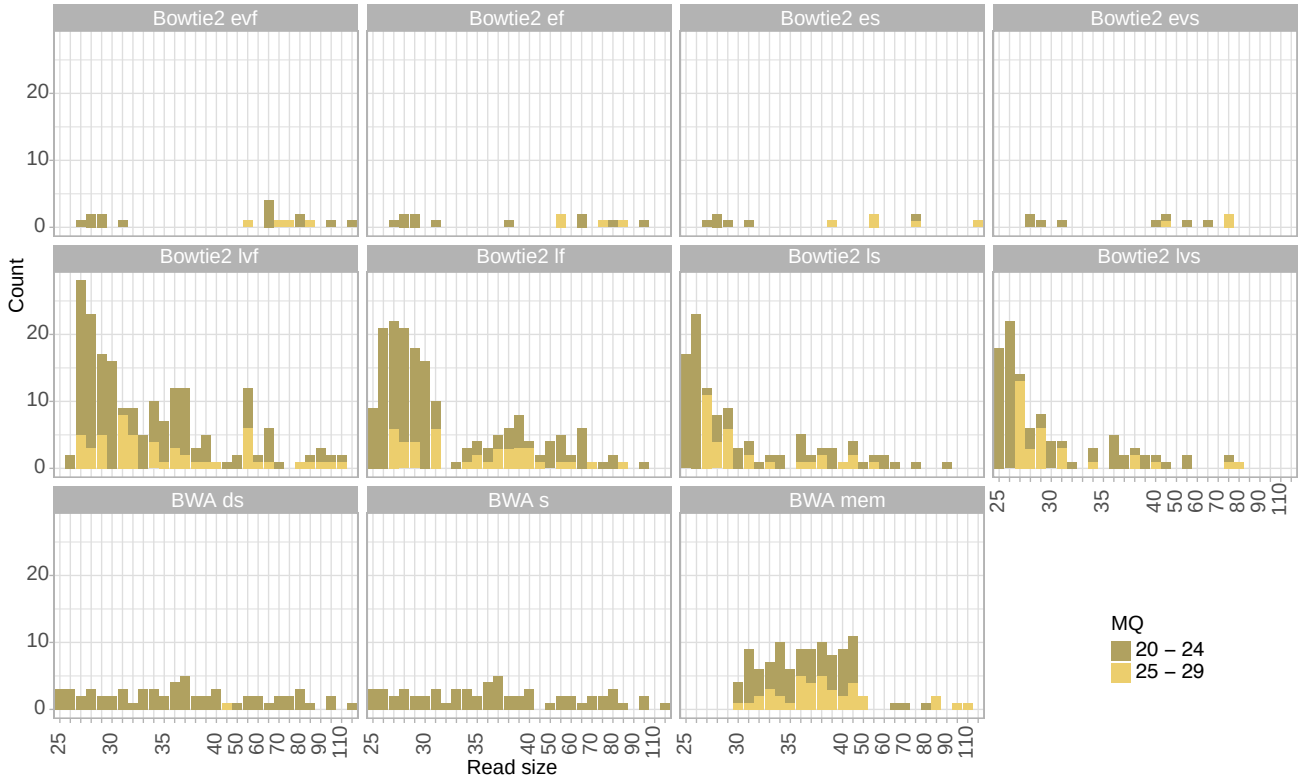
Supplementary Figure 1. Alignment performance of simulated data.

A total of 100,000 reads were simulated for each size category considered, in the presence of typical Illumina sequencing errors as well as nucleotide mis-incorporations remaining following USER-treatment. MQ refers to the mapping quality scores of the read alignments. This figure is the same as Figure 2, except that the respective contributions of each individual mapping quality scores are indicated. (A) Fractions of true positive, false positive and false negative alignments. (B) Mapping quality scores of false positive alignments. (C) Mapping quality scores of true positive alignments.

A
Bowtie2 evf
Bowtie2 ef
Bowtie2 es
Bowtie2 evs
Bowtie2 lvf
Bowtie2 lf
Bowtie2 ls
Bowtie2 lvs
BWA ds
BWA s
BWA mem
Count
Read size
True positives
False positives
False negatives
B
MQ
< 20
20 – 24
25 – 29
MQ >= 30
C
MQ
<= 24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
>= 46

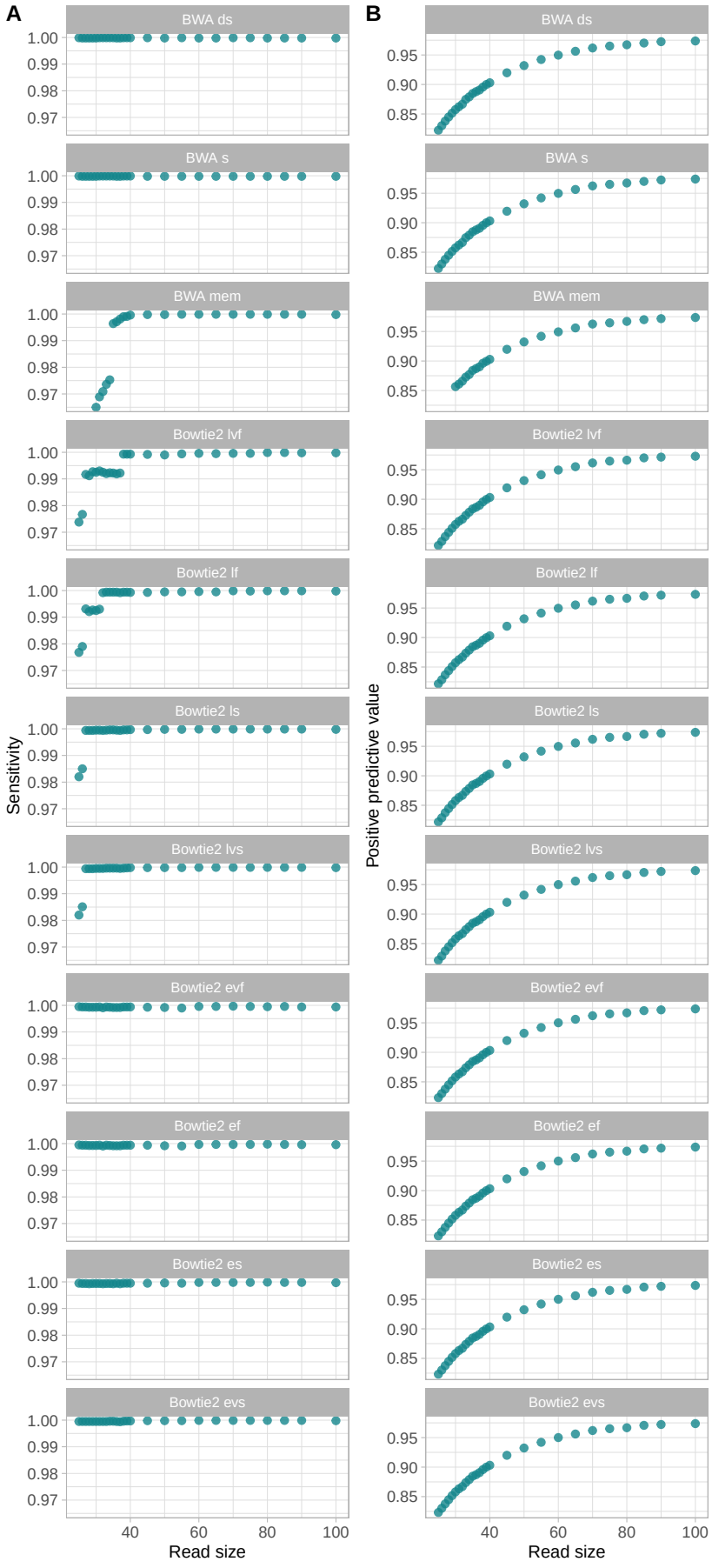
Supplementary Figure 2. Mapping quality scores of false positive alignments.

This figure zooms the fraction of false positive alignments with mapping quality scores between 20 and 30. It corresponds to the same information as shown in Figure S1B, however, the scale of the Y-axis now allows to identify the relevant classes of sequencing reads.



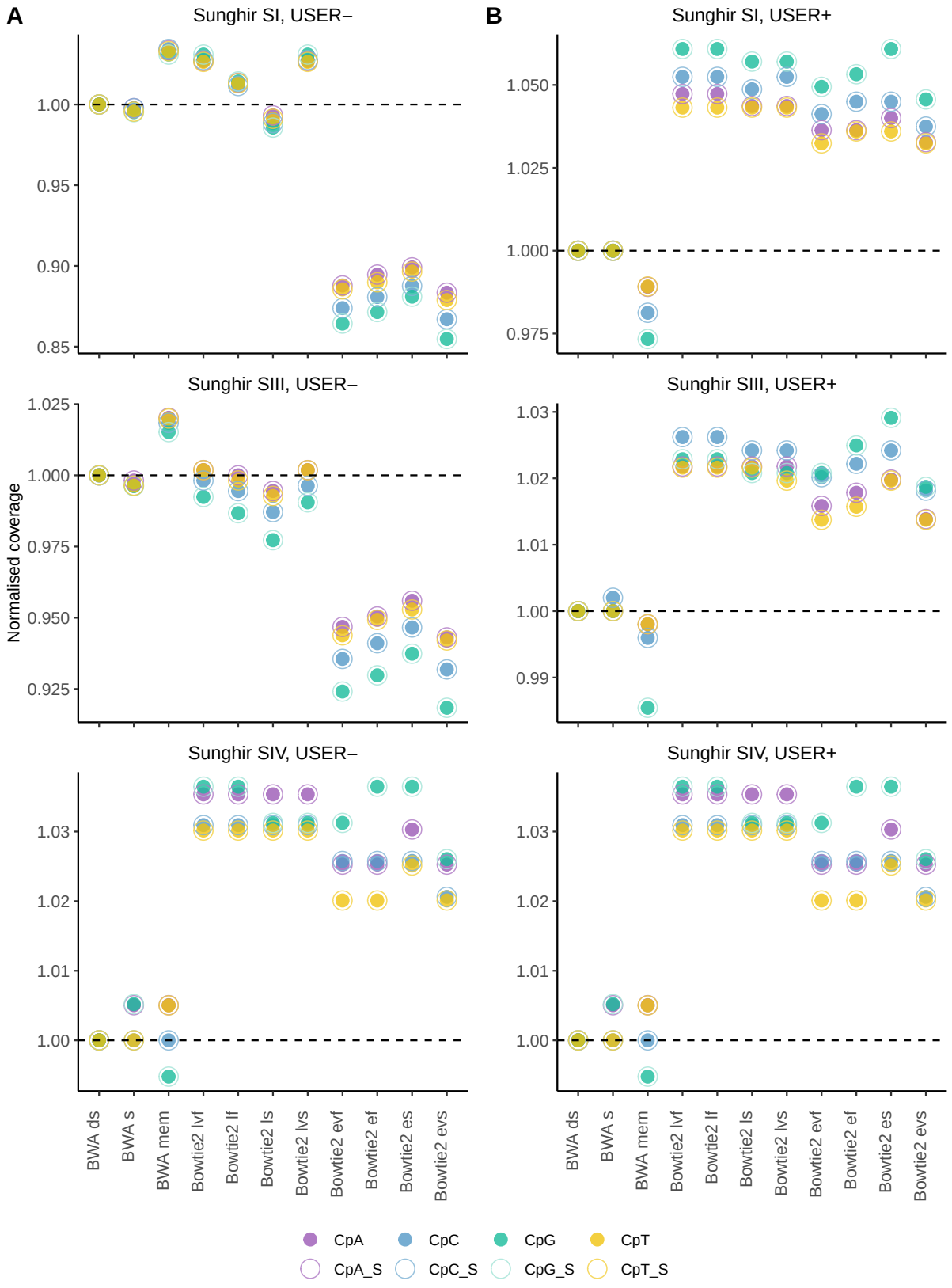
Supplementary Figure 3. Read alignment sensitivity and positive predictive values.

The performance metrics were calculated using simulated sequence data in the presence of typical Illumina sequencing errors as well as nucleotide mis-incorporations remaining following USER treatment. Alignments were filtered for minimal mapping quality scores of 30 and PCR duplicates. (A) Alignment sensitivity (ie true positives / (true positives + false negatives). (B) Alignment positive predictive value (ie true positives / (true positives + false positives)).


A
B
BWA ds
BWA s
BWA mem
Bowtie2 lvf
Bowtie2 lf
Bowtie2 ls
Bowtie2 lvs
Bowtie2 evf
Bowtie2 ef
Bowtie2 es
Bowtie2 evs
Sensitivity
Positive predictive value
Read size
Read size

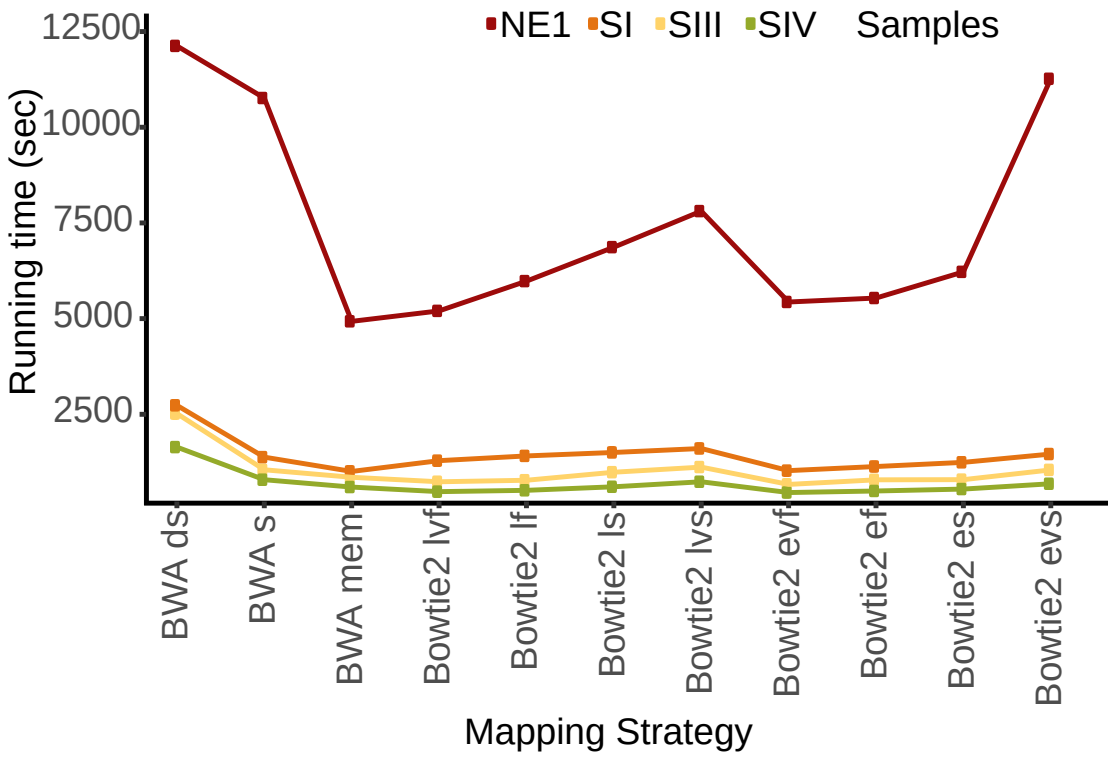
Supplementary Figure 4. Average depth-of-coverage in four dinucleotide contexts with soft-clipped bases (real data).

The average depth-of-coverage was estimated filtering alignments for minimal mapping quality scores of 30 (MQ≥30) and removing PCR duplicates. Coverage values are calculated in the dinucleotide sequence context most affected by DNA methylation (CpG), as well as the three other dinucleotides potentially affected by post-mortem Cytosine deamination at the same position (ie CpA, CpC and CpT). The differences observed are not due to soft-clipped bases as the values returned in the presence or not of soft-clipping masking are identical.


A
Sunghir SI, USER−
B
Sunghir SI, USER+
Sunghir SIII, USER−
Sunghir SIII, USER+
Sunghir SIV, USER−
Sunghir SIV, USER+
Normalised coverage
BWA ds
BWA s
BWA mem
Bowtie2 lvf
Bowtie2 lf
Bowtie2 ls
Bowtie2 lvs
Bowtie2 evf
Bowtie2 ef
Bowtie2 es
Bowtie2 evs
CpA
CpC
CpG
CpT
CpA_S
CpC_S
CpG_S
CpT_S

Supplementary Figure 5. Computational running times.

The times provided represent the running time that was necessary for Paleomix to process the sequence data of each specimen or individual indicated, using the same number of CPUs (E5-2683 v4 at 2.10GHz). Total running times were added for those individuals showing sequencing data generated both in the absence (USER-) and in the presence of USER treatment (USER+) in order to improve the figure readability.



Appendix 2: Supplementary information - Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA

MOLECULAR ECOLOGY

Supplemental Information for:

Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA

Tomasz Suchan[1,2*], Lorelei Chauvey[1*], Marine Poullet[1], Laure Tonasso-Calvière[1], Stéphanie Schiavinato[1], Pierre Clavel[1], Benoit Clavel[3], Sébastien Lepetz[3], Andaine Seguin-Orlando[1], Ludovic Orlando[1$]

*These authors equally contributed to this work

$Correspondence should be sent to Ludovic Orlando, ludovic.orlando@univ-tlse3.fr

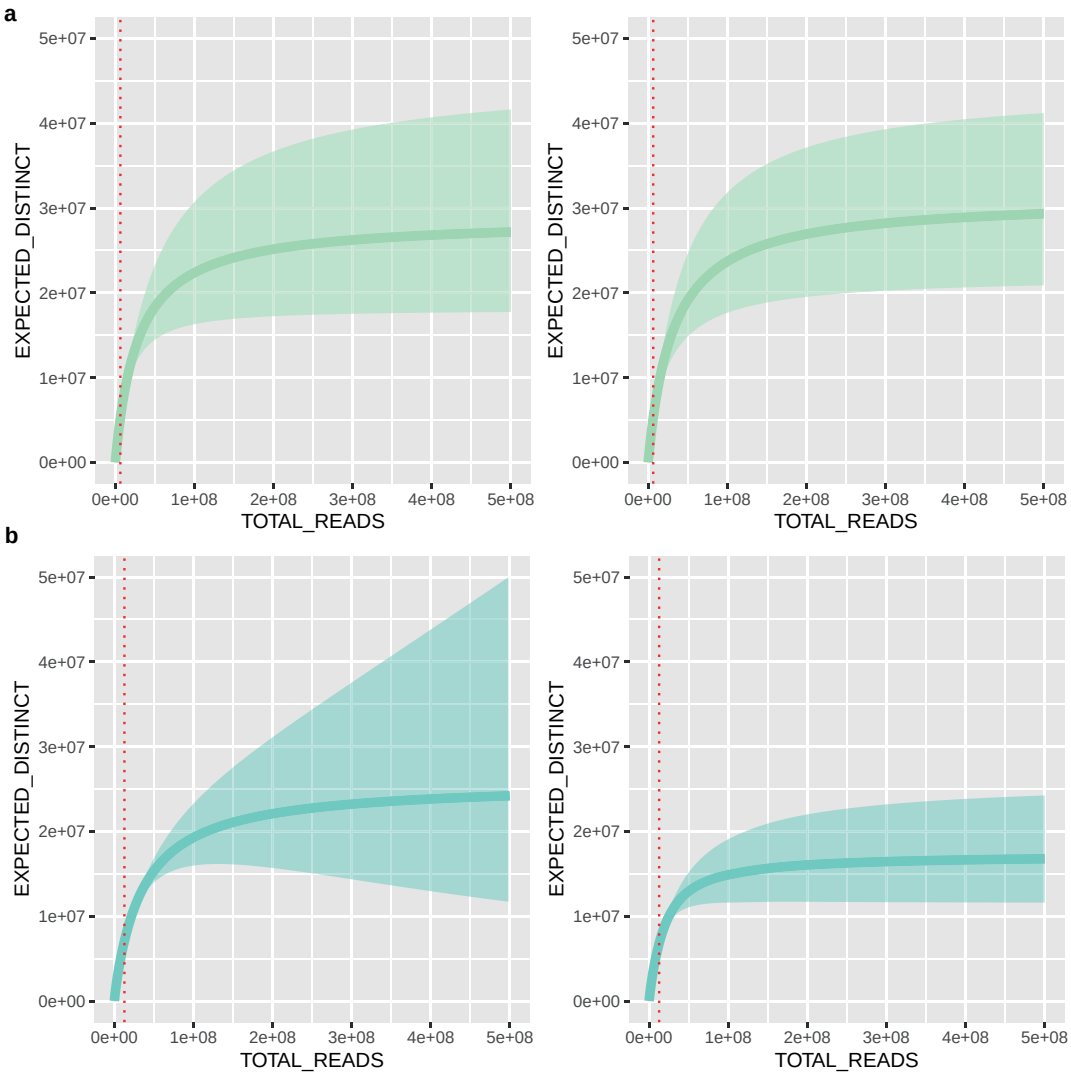
[1]Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT), Université Paul Sabatier, Faculté de Médecine Purpan, Bâtiment A, 37 allées Jules Guesde, 31000 Toulouse, France

[2]W. Szafer Institute of Botany, Polish Academy of Sciences, Lubicz 46, 31-512 Kraków, Poland

[3]Archéozoologie, Archéobotanique: sociétés, pratiques et environnements (AASPE), Muséum national d'histoire naturelle, CNRS, CP 56, 55 rue Buffon 75005 Paris, France

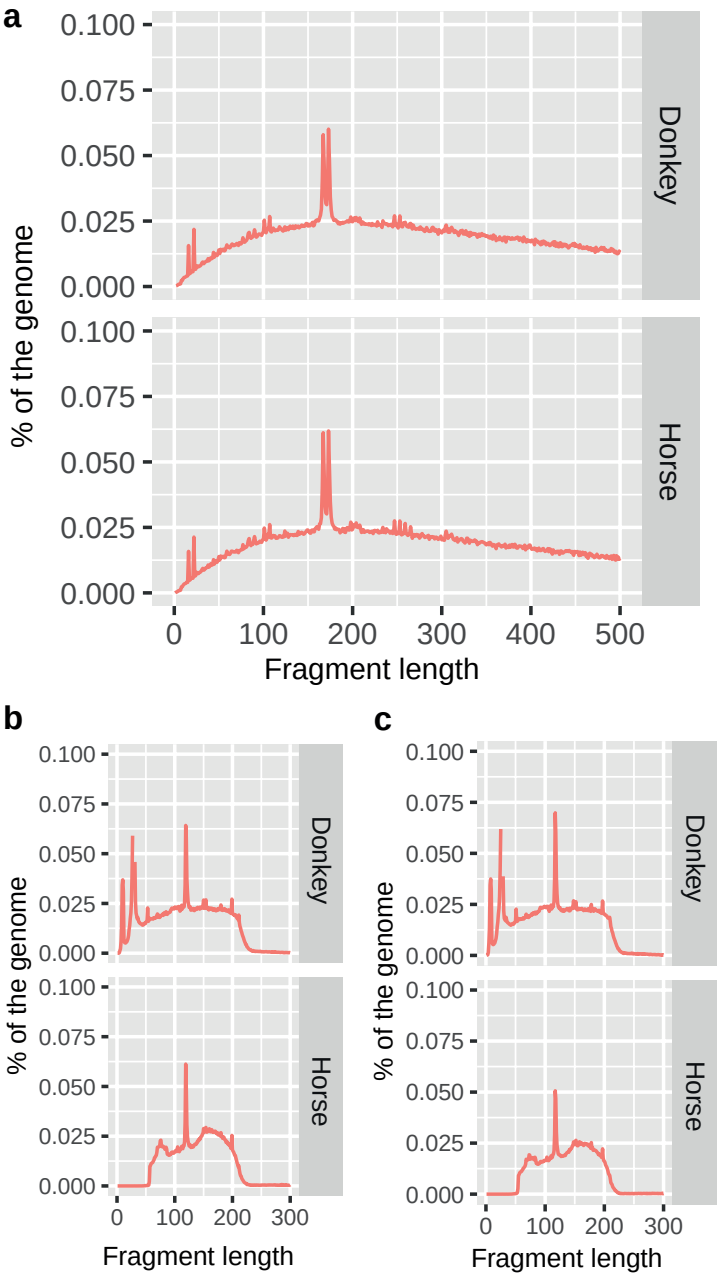
MOLECULAR ECOLOGY

Supplementary Figure 1. PreSeq calculation of the probe library content.
Panel A: The probe DNA library that was prepared from modern horse DNA extracts was subjected to shallow sequencing (red dashed line). Reads were then aligned against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019) (left), or against the donkey reference genome (Wang et al. 2021) (right). Panel B: Same as panel A, except that the probe DNA library was prepared from modern donkey DNA extracts. The PreSeq statistical model indicated that the horse probe DNA library contained an average of ~25-30 million genomic fragments versus 15-25 million for the donkey probe DNA library. Solid lines report the average number of unique genomic fragments that can be expected to be discovered with increasing sequencing efforts. The prediction confidence range is also provided.



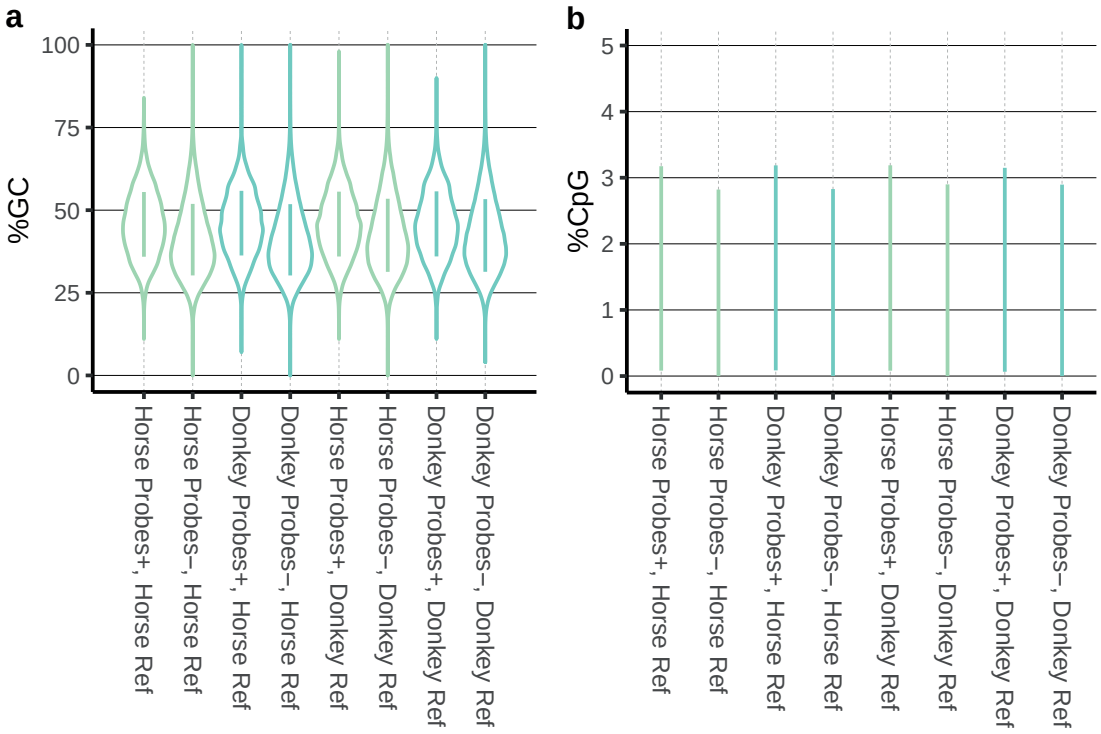
MOLECULAR ECOLOGY

Supplementary Figure 2. Probe size distributions.
Panel A: Size distribution of expected DNA templates following in Silico double restriction of the donkey (top) and horse (bottom) reference genomes. Panel B: Size distribution of the probe genomic fragments. Paired-end sequence data generated from horse and donkey hyRAD probe libraries were mapped against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019). Panel C: Same as panel B, except that mapping was performed against the donkey reference genome (Wang et al. 2021).



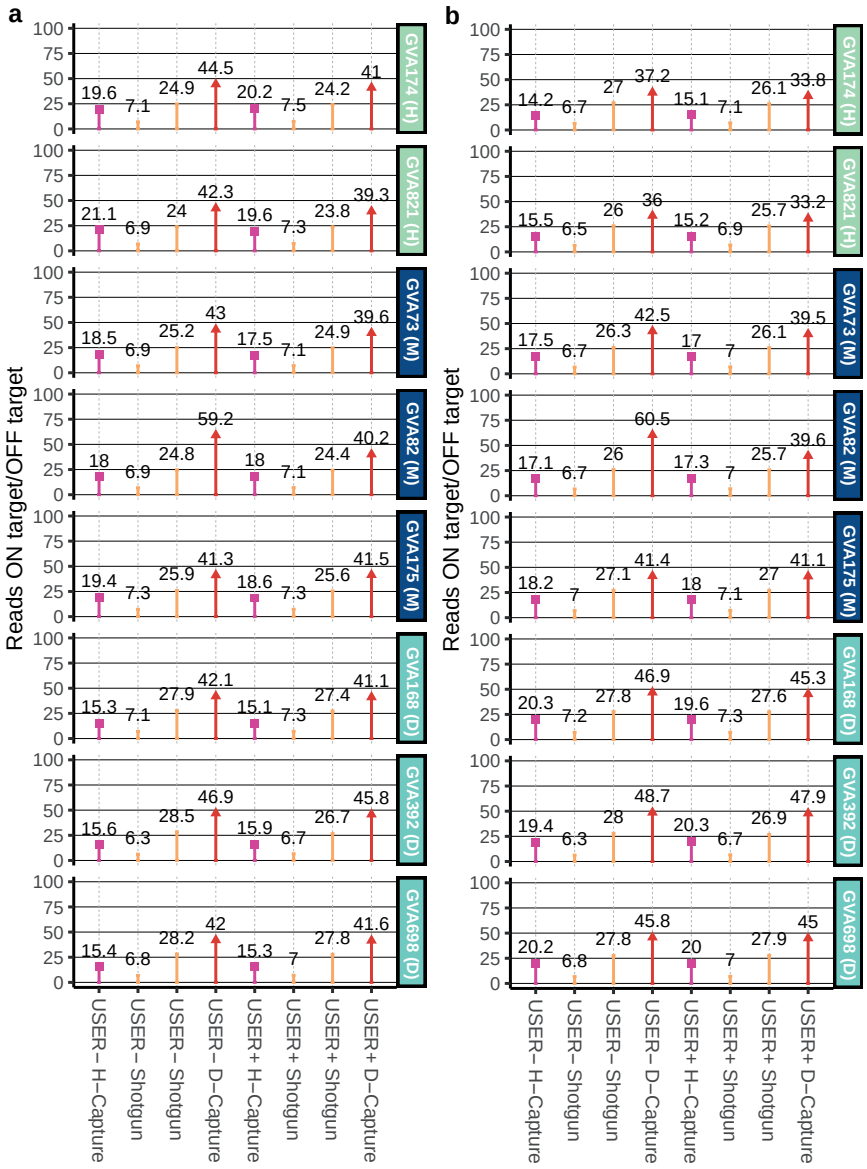
MOLECULAR ECOLOGY

Supplementary Figure 3. Probe base composition.
Panel A: %GC content of the genomic regions encompassing (Probes+) or not (Probes-) the hyRAD probes used for capture. Read alignment was carried out against both the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), and the donkey reference genome (Wang et al. 2021). Panel B: Same as panel A, except that the fractions of CpG dinucleotides within the genomic regions encompassing (Probes+) or not (Probes-) are provided.



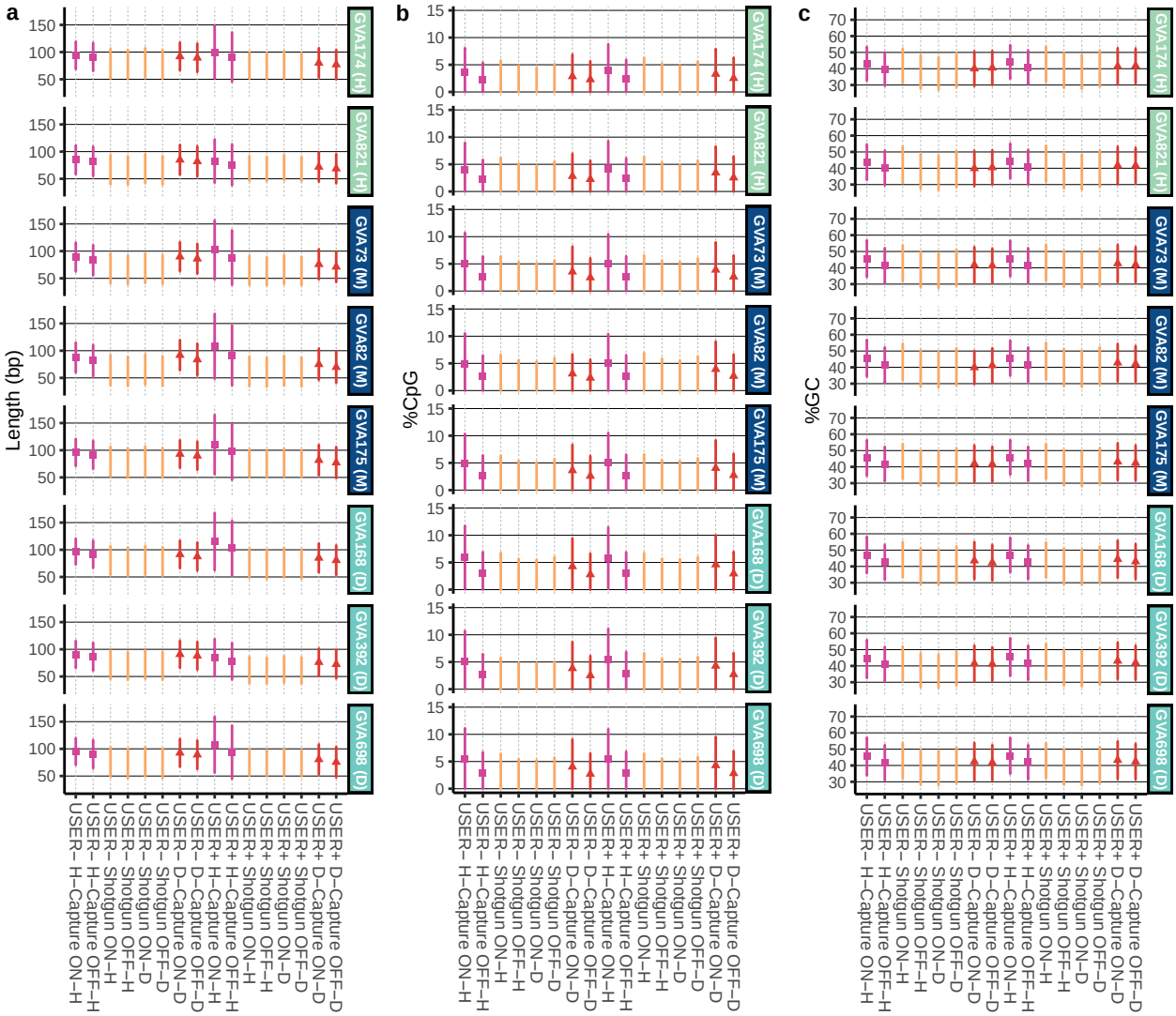
MOLECULAR ECOLOGY

Supplementary Figure 4. Ratios of ON-target and OFF-target read alignments.
Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were either shotgun sequenced (yellow) or subjected to capture with hyRAD horse (H-capture, blue) or donkey (D-capture, red) probes. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Panel A: Reads were aligned against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-chromosome contigs from Felkel and colleagues (Felkel et al. 2019), and filtered for PCR duplicates and minimal mapping quality of 25. Panel B: Same as panel A, except that the sequence alignments against the donkey reference genome (Wang et al. 2021) were used. Resulting ON-target enrichment folds are provided on Figure 1 by dividing the number of read alignments located ON-target following capture relative to shotgun sequencing.



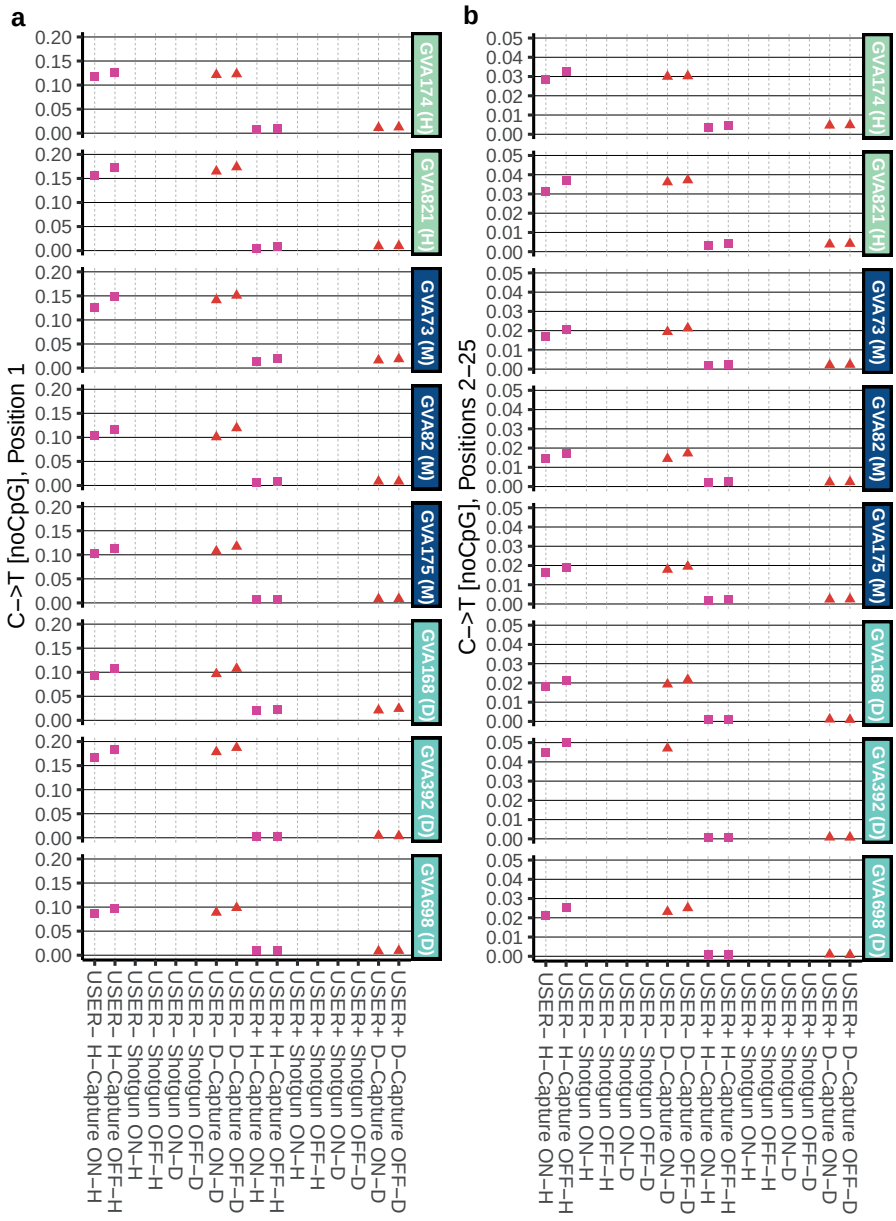
MOLECULAR ECOLOGY

Supplementary Figure 5. Size and base composition of ancient DNA templates following shotgun sequencing and hyRAD capture.
Panel A: Size distribution of endogenous DNA templates. Panel B: %GC content of the genomic regions covered by endogenous DNA templates. Panel C: Proportions of CpG dinucleotides present in the genomic regions covered by endogenous DNA templates. Read pairs that could not be collapsed into full-length templates were not used for mapping and only unique, high-quality alignments were considered. The donkey reference genome (Wang et al. 2021) was used for read alignment. Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green, (H)), 3 donkeys (deep blue, (D)) and 3 mules (light blue, (M)) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.



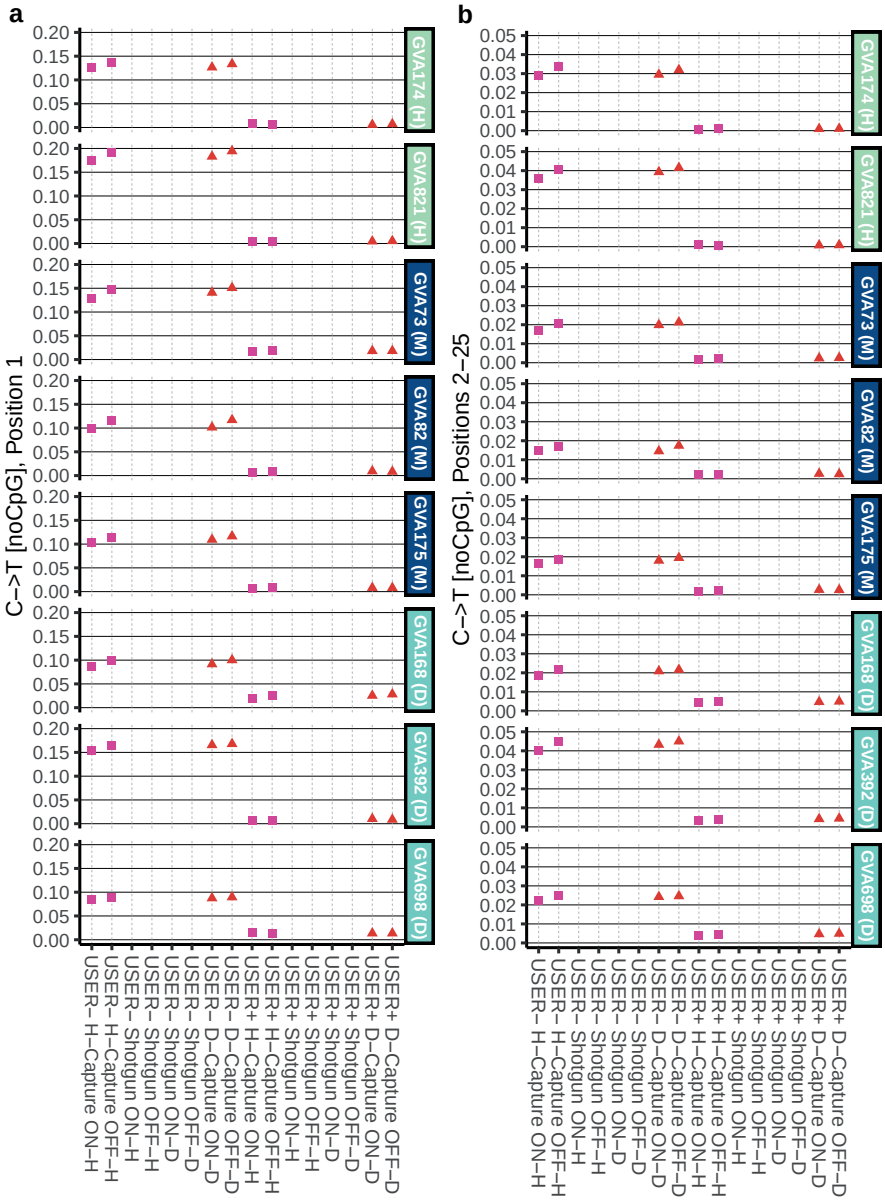
MOLECULAR ECOLOGY

Supplementary Figure 6. CpG→TpG conversion rates.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG→TpG conversions report the relative fraction of read alignments in which a CpG dinucleotide is found in the donkey reference genome considered but a TpG dinucleotide is found in the sequence data. Calculations were conditioned on those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.



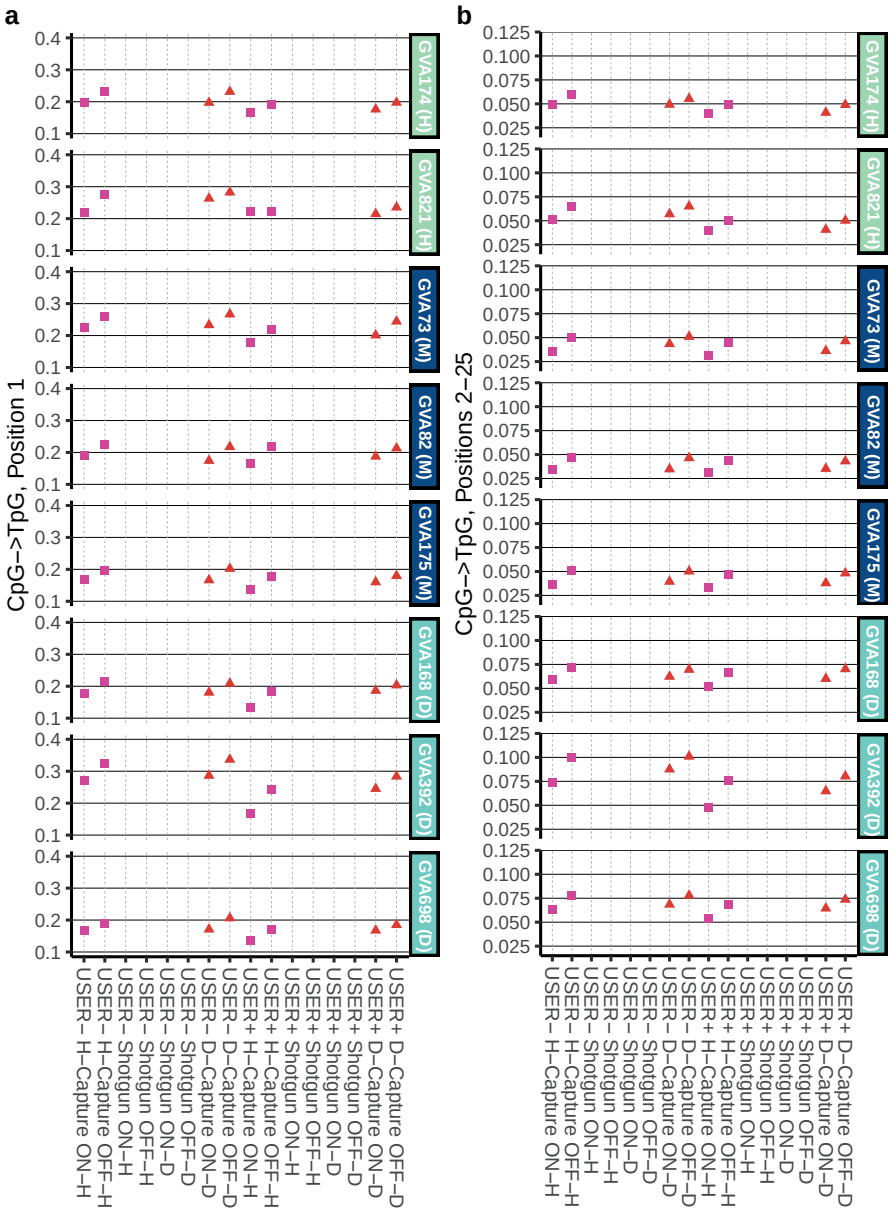
MOLECULAR ECOLOGY

Supplementary Figure 7. C→T conversion rates outside CpG dinucleotides.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG→TpG conversions report the relative fraction of read alignments in which a C is found in the horse reference genome considered but a T is found in the sequence data. Calculations were conditioned on non-CpG dinucleotide sites and those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.



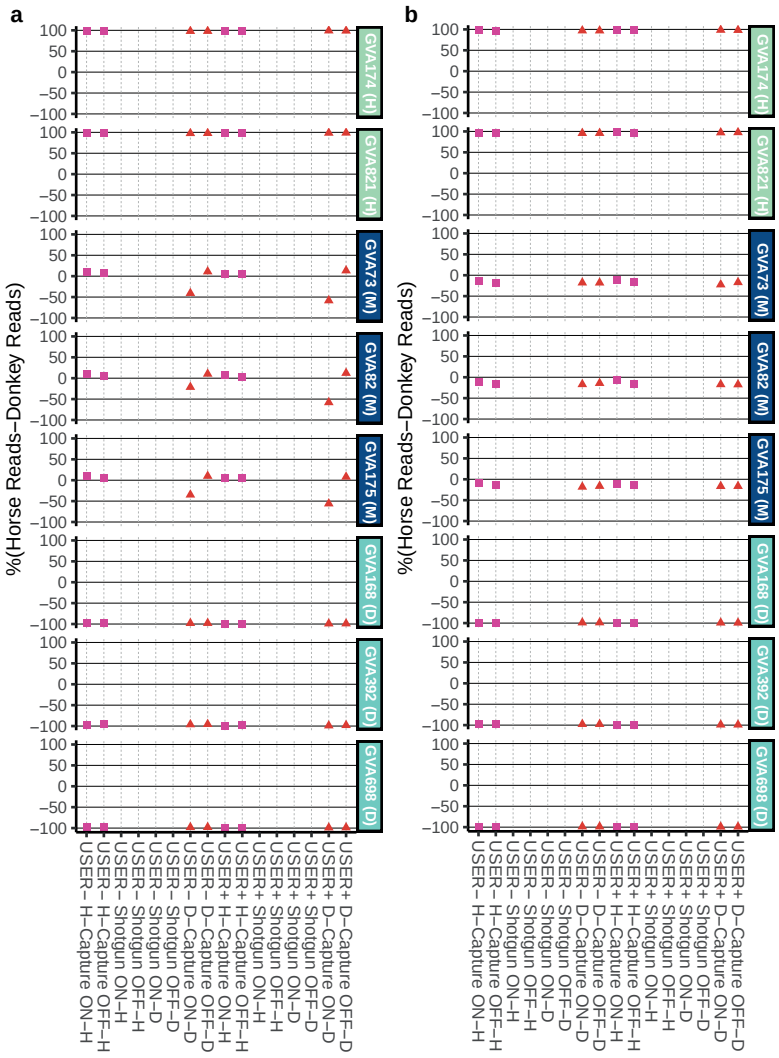
MOLECULAR ECOLOGY

Supplementary Figure 8. C→T conversion rates outside CpG dinucleotides.
Panel A: First read position. Panel B: Read positions 2 to 25. The proportions of CpG→TpG conversions report the relative fraction of read alignments in which a C is found in the donkey reference genome considered but a T is found in the sequence data. Calculations were conditioned on non-CpG dinucleotide sites and those read alignments located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.



MOLECULAR ECOLOGY

Supplementary Figure 9. Proportions of horse-specific and donkey-specific alleles identified following shotgun and hyRAD-capture.
Panel A: Alignments against the horse reference genome (Kalbfleish et al. 2019), supplemented with the Y-contigs from Felkel and colleagues (Felkel et al. 2019). Panel B: Alignments against the donkey reference genome (Wang et al. 2021). Calculations indicate the difference between the number of reads carrying horse-specific and donkey-specific alleles, respectively. Both transversion and transition alleles were considered. Positive (negative) fractions indicate an over-representation of horse-specific (donkey-specific) alleles in the sequence data. Calculations were conditioned on those read alignments against the reference genome that were located ON-target (ON-H when horse hyRAD probes were used for capture, and ON-D when donkey probes were used) or OFF-target (OFF-H and OFF-D, respectively). Ancient DNA libraries were prepared from extracts that were treated with the USER enzymatic mix (USER+) or from raw extracts (USER-). Ancient DNA libraries from 2 horses (light green), 3 donkeys (deep blue) and 3 mules (light blue) specimens were shotgun sequenced (Shotgun, yellow) or subjected to capture with hyRAD horse (H-capture, pink) or donkey (D-capture, red) probes.



Supplementary Material

Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA

Tomasz Suchan[1,2*], Lorelei Chauvey[1*], Marine Poullet[1], Laure Tonasso-Calvière[1], Stéphanie Schiavinato[1], Benoit Clavel[3], Sébastien Lepetz[3], Andaine Seguin-Orlando[1], Ludovic Orlando[1$]

The following commands provide the instructions for automated washing and purification of the DNA enriched post-hyRAD capture on the Opentrons OT2 liquid-handling instrument.

```
 from opentrons import protocol_api

# metadata
metadata = {
    'protocolName': 'Hybridization capture protocol',
    'author': 'Tomasz Suchan <t.suchan@botany.pl>', 'Lorelei Chauvey
<lorelei.chauvey@univ-tlse3.fr>'
    'description': 'Protocol for washing the DNA enriched with MyBaits v4 updated',
    'apiLevel': '2.7'
}


def run(protocol: protocol_api.ProtocolContext):
    # NUMBER OF SAMPLES:
    nsamples = 8

    # modules
    module_pcr = protocol.load_module('thermocycler', 7)
    pcr = module_pcr.load_labware('biorad_96_wellplate_200ul_pcr')

    module_magnet = protocol.load_module('magdeck', 4)
    magnet = module_magnet.load_labware('biorad_96_wellplate_200ul_pcr')

    # labware
    tiprack_multi = [protocol.load_labware('opentrons_96_filtertiprack_200ul',
slot) for slot in ["1","5"]]
    tiprack_single = protocol.load_labware('opentrons_96_tiprack_300ul', 6)
    trash = protocol.load_labware('agilent_1_reservoir_290ml', 9)
    plate_wash = protocol.load_labware('reservoir_21ml', 2)
    tubes_rack_s =
protocol.load_labware('opentrons_24_tuberack_eppendorf_1.5ml_safelock_snapcap', 3)

    # pipettes
    pipette_multi = protocol.load_instrument('p300_multi', 'left',
tip_racks=tiprack_multi)
    pipette_single = protocol.load_instrument('p300_single', 'right',
tip_racks=[tiprack_single])
    pipette_single.flow_rate.aspirate=50
    pipette_single.flow_rate.dispense=50
    pipette_multi.flow_rate.aspirate=30
    pipette_multi.flow_rate.dispense=50

    # commands
    module_pcr.close_lid()
    module_pcr.set_lid_temperature(95)
    module_pcr.set_block_temperature(95)
    protocol.delay(minutes=5)
    module_pcr.set_lid_temperature(85)
    module_pcr.set_block_temperature(55)
    module_pcr.open_lid()
    protocol.pause('Load hybridization mix in column 1 on the PCR and pres
resume.')
    module_pcr.close_lid()
    protocol.delay(minutes=5)
    module_pcr.open_lid()
```

```
    pipette_multi.transfer(18, pcr.columns_by_name()['1'],
                           pcr.columns_by_name()['2'],
                           mix_after=(5, 15))
    module_pcr.close_lid()

    protocol.pause('Incubation started. Place wash buffers in falcon tubes (A1 -
wash 1, A2 - wash 2), and 1.5 ml tubes (A1 - elution buffer, A2 - dynabeads) and
press resume to start.')
    module_pcr.open_lid()
    pipette_multi.transfer(130, plate_wash.columns_by_name()['1'],
                           pcr.columns_by_name()['7'],
                           air_gap=20)# wash buffer1
    pipette_multi.transfer(130, plate_wash.columns_by_name()['2'],
                           pcr.columns_by_name()['8'],
                           air_gap=20)# wash buffer2
    pipette_multi.transfer(130, plate_wash.columns_by_name()['2'],
                           pcr.columns_by_name()['9'],
                           air_gap=20)# wash buffer2
    pipette_multi.transfer(130, plate_wash.columns_by_name()['2'],
                           pcr.columns_by_name()['10'],
                           air_gap=20)# wash buffer2
    pipette_single.distribute(40,  tubes_rack_s.wells_by_name()['A1'],
[pcr.wells_by_name()[well_name] for well_name in
['A11','B11','C11','D11','E11','F11','G11','H11']], disposal_volume=20) # elution
buffer
    pipette_single.flow_rate.aspirate=100
    pipette_single.flow_rate.dispense=150
    pipette_single.pick_up_tip()
    pipette_single.mix(5, 200, tubes_rack_s.wells_by_name()["A2"])
    pipette_single.drop_tip()
    pipette_single.flow_rate.aspirate=50
    pipette_single.flow_rate.dispense=50
    pipette_single.distribute(75,  tubes_rack_s.wells_by_name()['A2'],
[pcr.wells_by_name()[well_name] for well_name in
['A12','B12','C12','D12','E12','F12','G12','H12']], touch_tip=True,
disposal_volume=5)
    protocol.pause("Check if all OK and pres resume.")
    module_pcr.close_lid()
    protocol.delay(minutes=5)
    module_pcr.open_lid()
    pipette_multi.transfer(70, pcr.columns_by_name()['12'],
                           pcr.columns_by_name()['2'],
                           mix_before=(5, 60),
                           mix_after=(5, 60))
    for _ in range(5):
        module_pcr.close_lid()
        protocol.delay(minutes=5)
        module_pcr.open_lid()
        pipette_multi.pick_up_tip()
        pipette_multi.mix(5, 50, pcr.wells_by_name()["A2"])
        pipette_multi.drop_tip()
    module_pcr.close_lid()
    protocol.delay(minutes=5)
    module_pcr.open_lid()

    def wash(step, time, vol_in, vol_out):
        module_pcr.open_lid()
        pipette_multi.pick_up_tip()
        pipette_multi.transfer(vol_in, pcr.columns()[1],
                               magnet.columns()[1],
                               aspirate_speed=50,
                               new_tip='never',
                               mix_before=(5, 90))
        module_magnet.engage()
        protocol.delay(seconds=30)
        pipette_multi.transfer(vol_out, magnet.columns()[1],
                               trash.columns()[0],
                               new_tip='never')
```

```
    pipette_multi.drop_tip()
    module_magnet.disengage()
    pipette_multi.pick_up_tip()
    pipette_multi.transfer(vol_in, pcr.columns()[step+5],
                           magnet.columns()[1],
                           new_tip='never',
                           mix_after=(5, 90))
    pipette_multi.transfer(vol_out, magnet.columns()[1],
                           pcr.columns()[1],
                           aspirate_speed=50,
                           new_tip='never')
    pipette_multi.drop_tip()
    module_pcr.close_lid()
    protocol.delay(minutes=time)

# wash steps 1-3
wash(step=1, time=15, vol_in=100, vol_out=150)
for i in range(2,5):
    wash(step=i, time=15, vol_in=100, vol_out=100)

# last wash and elute in final volume
module_pcr.open_lid()
pipette_multi.pick_up_tip()
pipette_multi.transfer(100, pcr.columns()[1],
                       magnet.columns()[1],
                       aspirate_speed=50,
                       new_tip='never',
                       mix_before=(5, 90))
module_magnet.engage()
protocol.delay(seconds=30)
pipette_multi.transfer(180, magnet.columns()[1],
                       trash.columns()[0],
                       new_tip='never')
pipette_multi.drop_tip()
module_magnet.disengage()
pipette_multi.flow_rate.aspirate=50
pipette_multi.flow_rate.dispense=80
pipette_multi.transfer(30, pcr.columns()[10],
                       magnet.columns()[1],
                       mix_after=(5,20))
module_pcr.deactivate()
```

Type	USER	Sample	Taxon	Site	Context	Collapsed
Capture (Horse Probes	-	GVA174	Male Horse	Evreux - Parking de l'Hôtel de Ville	Ancien Régime, 16th-17th c. CE	780084
Capture (Horse Probes	+	GVA174	Male Horse	-	-	458094
Shotgun	-	GVA174	Male Horse	-	-	1203696
Shotgun	+	GVA174	Male Horse	-	-	712053
Capture (Donkey Probe	-	GVA174	Male Horse	-	-	632512
Capture (Donkey Probe	+	GVA174	Male Horse	-	-	636928
Capture (Horse Probes	-	GVA821	Male Horse	Amiens - Rue Legrand Daussy	Roman Period, 2nd-3rd c. CE	939575
Capture (Horse Probes	+	GVA821	Male Horse	-	-	400170
Shotgun	-	GVA821	Male Horse	-	-	1329053
Shotgun	+	GVA821	Male Horse	-	-	556524
Capture (Donkey Probe	-	GVA821	Male Horse	-	-	1062718
Capture (Donkey Probe	+	GVA821	Male Horse	-	-	1056095
Capture (Horse Probes	-	GVA73	Male Mule	Chartres - Boulevard de la Courtille - C2	Roman Period, 1st-2nd c. CE	734981
Capture (Horse Probes	+	GVA73	Male Mule	-	-	342634
Shotgun	-	GVA73	Male Mule	-	-	1097577
Shotgun	+	GVA73	Male Mule	-	-	1508639
Capture (Donkey Probe	-	GVA73	Male Mule	-	-	847891
Capture (Donkey Probe	+	GVA73	Male Mule	-	-	591328
Capture (Horse Probes	-	GVA82	Male Mule	Chartres - Boulevard de la Courtille - C2	Roman Period, 1st-2nd c. CE	925445
Capture (Horse Probes	+	GVA82	Male Mule	-	-	426606
Shotgun	-	GVA82	Male Mule	-	-	1325652
Shotgun	+	GVA82	Male Mule	-	-	658423
Capture (Donkey Probe	-	GVA82	Male Mule	-	-	672097
Capture (Donkey Probe	+	GVA82	Male Mule	-	-	638988
Capture (Horse Probes	-	GVA175	Male Mule	Evreux - Parking de l'Hôtel de Ville	Ancien Régime, 16th-17th c. CE	865705
Capture (Horse Probes	+	GVA175	Male Mule	-	-	431792
Shotgun	-	GVA175	Male Mule	-	-	1088709

Shotgun	+	GVA175	Male Mule	-	-	520045
Capture (Donkey Probe	-	GVA175	Male Mule	-	-	704142
Capture (Donkey Probe	+	GVA175	Male Mule	-	-	575479
Capture (Horse Probes	-	GVA168	Male Donkey	Evreux - Parking de l'Hôtel de Ville	Ancien Régime, 16th-17th c. CE	715943
Capture (Horse Probes	+	GVA168	Male Donkey	-	-	446013
Shotgun	-	GVA168	Male Donkey	-	-	776819
Shotgun	+	GVA168	Male Donkey	-	-	1188260
Capture (Donkey Probe	-	GVA168	Male Donkey	-	-	780186
Capture (Donkey Probe	+	GVA168	Male Donkey	-	-	563158
Capture (Horse Probes	-	GVA392	Male Donkey	Saint-Sornin - Tour de Broue	Second Middle Ages, 11th-13th (	1002583
Capture (Horse Probes	+	GVA392	Male Donkey	-	-	275864
Shotgun	-	GVA392	Male Donkey	-	-	727937
Shotgun	+	GVA392	Male Donkey	-	-	617987
Capture (Donkey Probe	-	GVA392	Male Donkey	-	-	1143197
Capture (Donkey Probe	+	GVA392	Male Donkey	-	-	779401
Capture (Horse Probes	-	GVA698	Male Donkey	Moussy-le-Neuf - 6 rue Pasteur	First Middle Ages, 6th-8th c. CE	1028195
Capture (Horse Probes	+	GVA698	Male Donkey	-	-	414683
Shotgun	-	GVA698	Male Donkey	-	-	903891
Shotgun	+	GVA698	Male Donkey	-	-	836694
Capture (Donkey Probe	-	GVA698	Male Donkey	-	-	733196
Capture (Donkey Probe	+	GVA698	Male Donkey	-	-	758030

*For shotgun sequencing data: the first number provided corresponds to the total number of unique high-quality (U+MQ25) reads mapped again

CollapsedTruncated	Pairs	Pairs (x2)	Endogenous (HR)	Endogenous (DR)	U+MQ25(HR)	ONT(HR)*	OFFT(HR)*	U+MQ25(mtH)
433	367883	735766	**56,1%**	38,4%	850199	127797	653521	78
237	16741	33482	**54,7%**	38,1%	268794	42311	209036	43
1057	205436	410872	**71,0%**	56,6%	1146863	70141/178814	989715/716751	222
579	191333	382666	**61,8%**	51,9%	676255	43354/103227	580413/426767	147
436	219712	439424	**55,8%**	37,5%	598251	145656	327496	59
507	128551	257102	**62,2%**	43,4%	556627	127235	310055	66
649	245142	490284	**45,0%**	29,2%	643802	103318	489065	77
184	12231	24462	**48,6%**	33,7%	206275	31409	160029	35
1964	111908	223816	**61,3%**	47,6%	953309	56508/142106	820318/591433	266
627	265153	530306	**42,8%**	35,6%	465752	29302/69578	398780/292317	156
872	304807	609614	**49,2%**	32,4%	823837	192434	454899	132
760	92881	185762	**56,9%**	39,3%	706934	155688	395785	98
582	320332	640664	**42,6%**	**42,2%**	586254	83774	453856	42
283	13971	27942	**42,1%**	**41,9%**	156290	21761	124077	18
2151	139011	278022	**58,3%**	**58,3%**	802398	47685/124471	690422/493862	151
2531	164439	328878	**64,8%**	**64,8%**	1191266	72904/182946	1020326/736084	257
849	425334	850668	**44,5%**	**44,3%**	755773	178847	415753	69
701	126527	253054	**46,3%**	**46,0%**	391455	86792	219020	35
814	391209	782418	**46,8%**	**46,5%**	799568	111692	620289	79
268	22325	44650	**44,6%**	**44,4%**	210056	30001	166304	21
2705	147757	295514	**61,1%**	**61,2%**	992193	58766/151220	852174/610874	176
596	74849	149698	**68,9%**	**69,1%**	557276	33871/84029	477069/344538	143
665	409008	818016	**47,9%**	**47,9%**	714515	207096	349762	57
586	154284	308568	**49,5%**	**49,2%**	468944	104720	260762	62
517	407431	814862	**52,4%**	**52,0%**	881108	131133	677058	99
202	25232	50464	**47,8%**	**47,5%**	230626	33796	182046	50
1064	184793	369586	**65,9%**	**66,0%**	961738	60253/154930	828170/598542	227

420	205510	411020	**50,2%**	**50,2%**	467201	29388/74607	401260/291687	136
529	206254	412508	**51,2%**	**51,0%**	572082	132547	321302	74
364	149569	299138	**51,8%**	**51,7%**	453519	104442	251674	84
420	294080	588160	42,0%	**61,2%**	547706	66750	437115	15
221	25898	51796	43,0%	**61,6%**	214216	26439	174820	5
706	129737	259474	57,5%	**71,6%**	595932	36635/101807	514427/365400	20
999	162804	325608	65,6%	**78,1%**	993655	62255/167069	854359/610689	61
658	194229	388458	39,5%	**56,2%**	461513	108391	257616	10
331	147766	295532	43,5%	**60,3%**	373630	85729	208472	11
699	211628	423256	25,3%	**40,6%**	361344	45336	289846	43
139	6798	13596	34,1%	**51,2%**	98811	12594	79336	17
1237	86759	173518	31,3%	**41,1%**	282130	15323/48024	245150/168690	87
441	48479	96958	46,3%	**56,0%**	331337	19025/53573	285140/200400	172
1001	273893	547786	24,0%	**37,9%**	406419	103277	219992	46
649	54313	108626	33,9%	**49,2%**	301385	74552	162717	66
737	484201	968402	42,1%	**61,7%**	841333	103038	671033	9
205	17779	35558	42,6%	**61,4%**	191906	23868	155796	5
1106	136973	273946	59,6%	**74,2%**	702319	41549/119999	607084/425489	38
881	334320	668640	46,4%	**55,1%**	698241	42280/118841	602864/426965	21
637	256868	513736	39,8%	**57,1%**	496484	116341	276849	12
522	165307	330614	42,7%	**59,6%**	465017	107406	258131	11

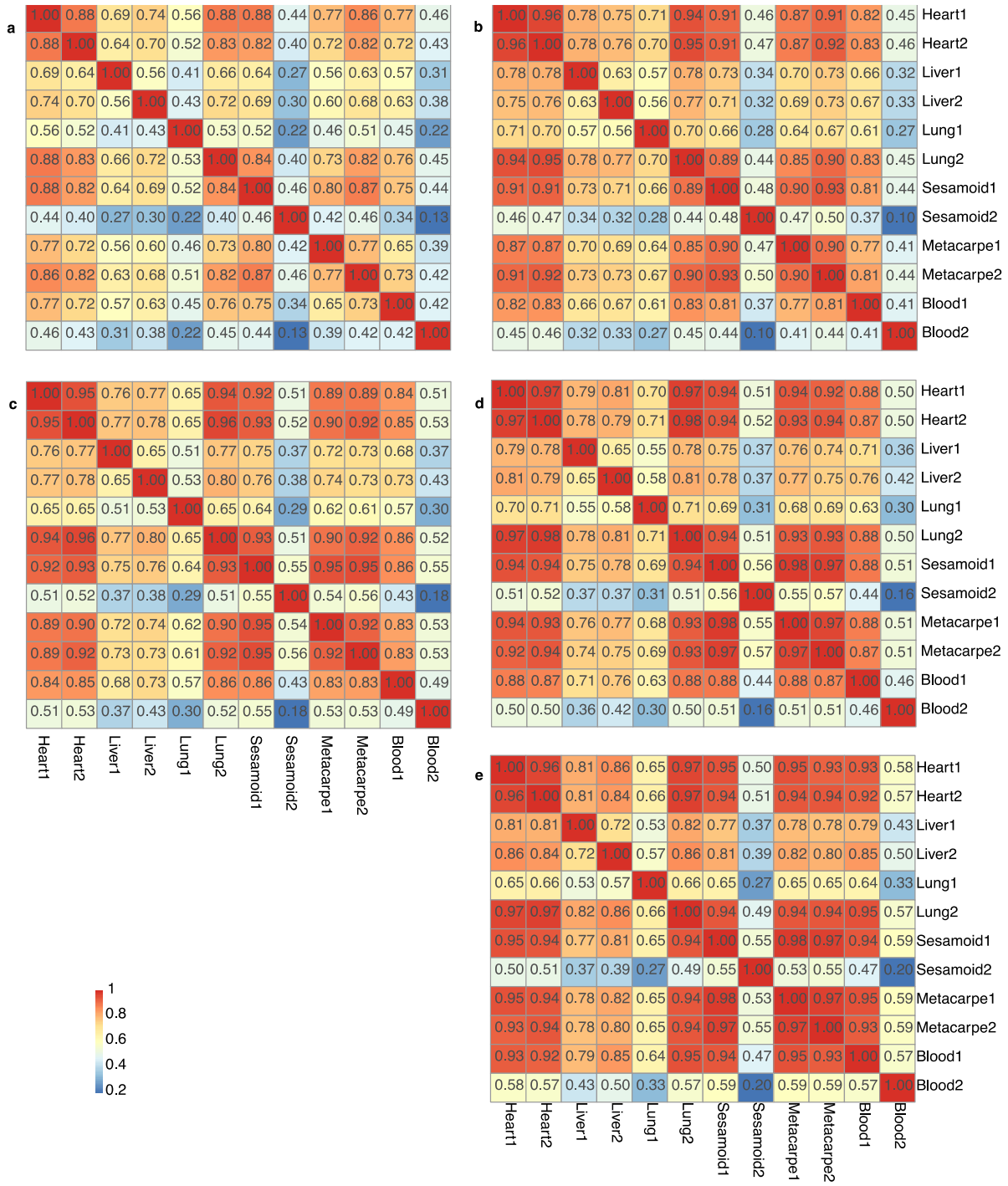
ıst the Horse nuclear reference genome (HR), while the second corresponds to that mapped against the Donkey nuclear reference genome (DR).

ON/OFF	ON target fold-enrichmer		U+MQ25(DR)	ONT(DR)*	OFFT(DR)*	U+MQ25(mtD)	ON/OFF	ON target fold-enrichment	
19,6%	**-**	**2,76**	581952	66962	471259	9	14,2%	-	2,00
20,2%	**-**	**2,71**	187200	23060	152388	9	15,1%	-	2,03
7,1%	**24,9%**	**-**	914035	52886/150012	794771/555332	39	7,1%	24,9%	-
7,5%	**24,2%**	**-**	568318	34999/91225	491179/349205	27	7,5%	24,2%	-
44,5%	**-**	**1,78**	402120	85379	229574	7	37,2%	-	1,49
41,0%	**-**	**1,70**	388555	76411	225891	13	33,8%	-	1,40
21,1%	**-**	**3,07**	417622	51888	334858	18	15,5%	-	2,37
19,6%	**-**	**2,67**	143034	17610	115521	6	15,2%	-	2,19
6,9%	**24,0%**	**-**	739624	41944/115781	640775/446162	38	6,5%	26,0%	-
7,3%	**23,8%**	**-**	387256	23209/60848	334043/236350	25	6,9%	25,7%	-
42,3%	**-**	**1,76**	542639	111910	310703	22	36,0%	-	1,39
39,3%	**-**	**1,65**	488894	93813	282978	15	33,2%	-	1,29
18,5%	**-**	**2,67**	580992	79509	455565	12	**18,5%**	**-**	**2,75**
17,5%	**-**	**2,45**	155239	21155	124268	4	**17,5%**	**-**	**2,51**
6,9%	**25,2%**	**-**	803704	46672/127485	694772/485583	29	**6,7%**	**26,3%**	**-**
7,1%	**24,9%**	**-**	1192330	71611/188415	026294/72298	40	**7,0%**	**26,1%**	**-**
43,0%	**-**	**1,71**	752376	175796	413951	11	**42,5%**		**1,62**
39,6%	**-**	**1,59**	389043	85745	217061	4	**39,5%**		**1,52**
18,0%	**-**	**2,61**	794743	106426	623680	8	**17,1%**		**2,54**
18,0%	**-**	**2,54**	209218	28899	167149	7	**17,3%**		**2,48**
6,9%	**24,8%**	**-**	994137	57738/156050	857875/600202	48	**6,7%**	**26,0%**	**-**
7,1%	**24,4%**	**-**	558464	33421/86977	480336/338154	37	**7,0%**	**25,7%**	**-**
59,2%	**-**	**2,39**	714394	209889	347153	11	**60,5%**		**2,33**
40,2%	**-**	**1,65**	466418	102686	259088	21	**39,6%**		**1,54**
19,4%	**-**	**2,66**	874260	123620	680848	7	**18,2%**		**2,58**
18,6%	**-**	**2,53**	229257	32820	182490	10	**18,0%**		**2,52**
7,3%	**25,9%**	**-**	963516	58771/159434	834008/588479	44	**7,0%**	**27,1%**	**-**

7,3%	**25,6%**	**-**	467691	28834/77313	403683/285942	27	**7,1%**	**27,0%**	**-**
41,3%	**-**	**1,59**	570289	131841	318565	10	**41,4%**		**1,53**
41,5%	**-**	**1,62**	452509	103189	250834	20	**41,1%**		**1,52**
15,3%	-	2,14	797772	124077	609932	50	**20,3%**		**2,84**
15,1%	-	2,08	306737	47305	241555	28	**19,6%**		**2,69**
7,1%	27,9%	-	741990	45973/125795	641468/451877	108	**7,2%**	**27,8%**	**-**
7,3%	27,4%	-	1182362	74136/198988	019309/71997	280	**7,3%**	**27,6%**	**-**
42,1%	-	1,51	656827	166289	354433	68	**46,9%**		**1,69**
41,1%	-	1,50	517489	126826	280065	73	**45,3%**		**1,64**
15,6%	-	2,50	579343	87285	450625	267	**19,4%**		**3,08**
15,9%	-	2,38	148137	23325	115161	91	**20,3%**		**3,02**
6,3%	28,5%	-	370558	20275/61810	322822/220839	438	**6,3%**	**28,0%**	**-**
6,7%	26,7%	-	399622	23125/64511	344838/239480	710	**6,7%**	**26,9%**	**-**
46,9%	-	1,65	641482	166988	342806	516	**48,7%**		**1,74**
45,8%	-	1,71	436980	111498	232602	463	**47,9%**		**1,78**
15,4%	-	2,24	1231297	190359	942397	76	**20,2%**		**2,97**
15,3%	-	2,18	276447	43209	216464	31	**20,0%**		**2,84**
6,8%	28,2%	-	874043	51535/146663	757603/526856	175	**6,8%**	**27,8%**	**-**
7,0%	27,8%	-	830227	50423/140488	718496/503564	144	**7,0%**	**27,9%**	**-**
42,0%	-	1,49	712429	176798	385980	98	**45,8%**		**1,65**
41,6%	-	1,49	649332	158011	351431	78	**45,0%**		**1,61**

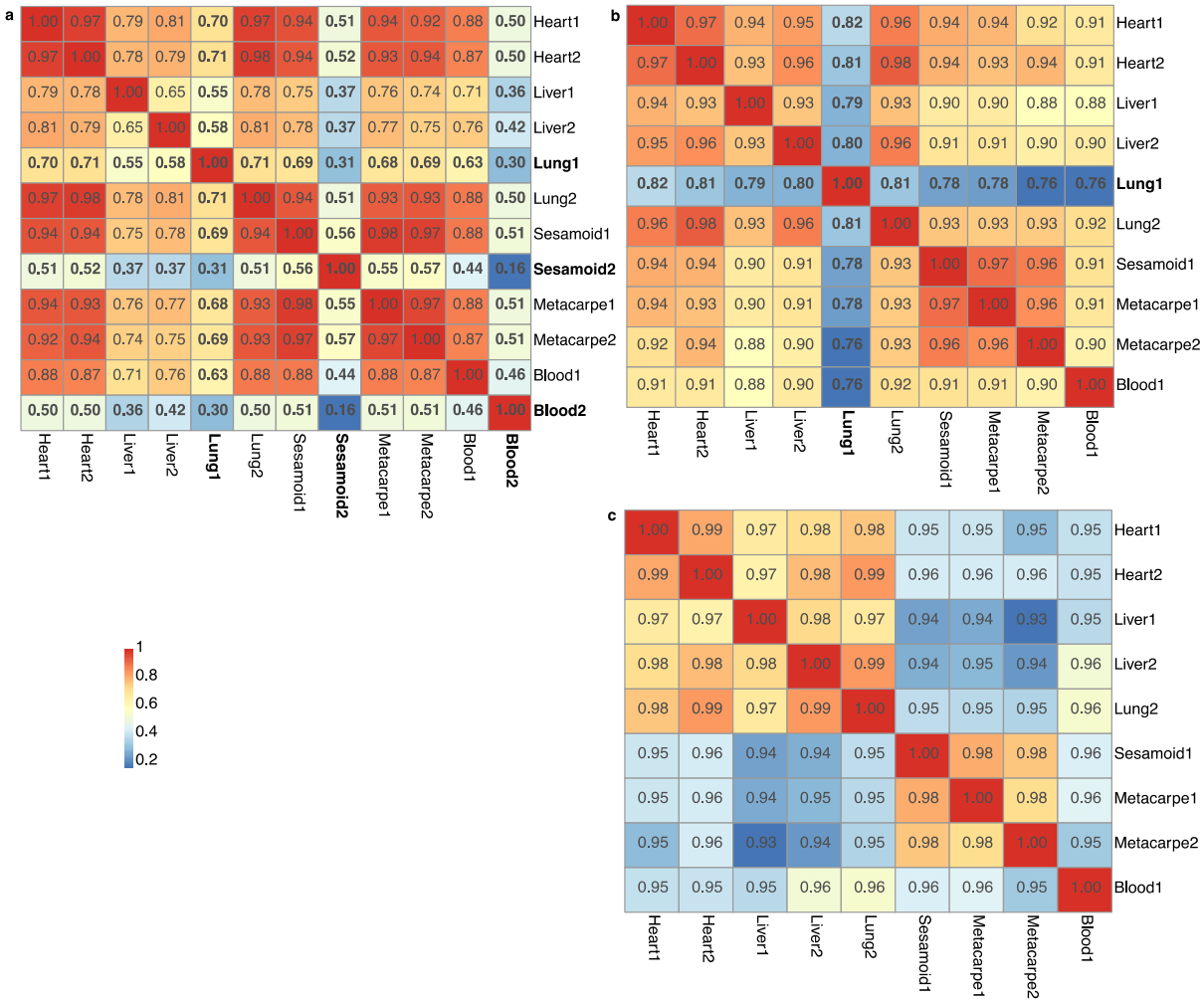
Appendix 3: Supplementary information - Experimental design for the reconstruction of ancient DNA methylation maps in horses

---



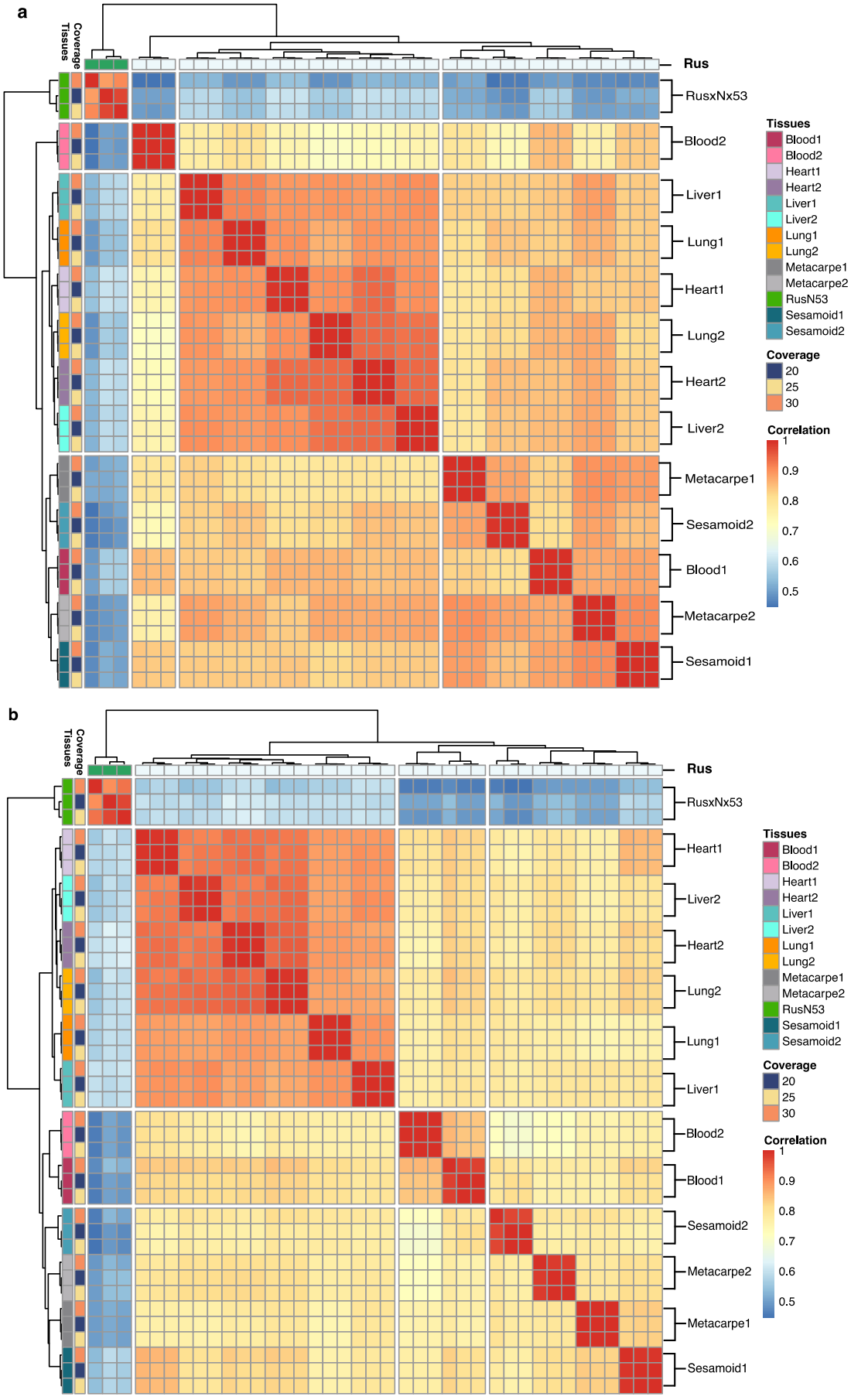
Supplementary Figure 1. Correlation matrix between the selected RRBS samples across different coverage.

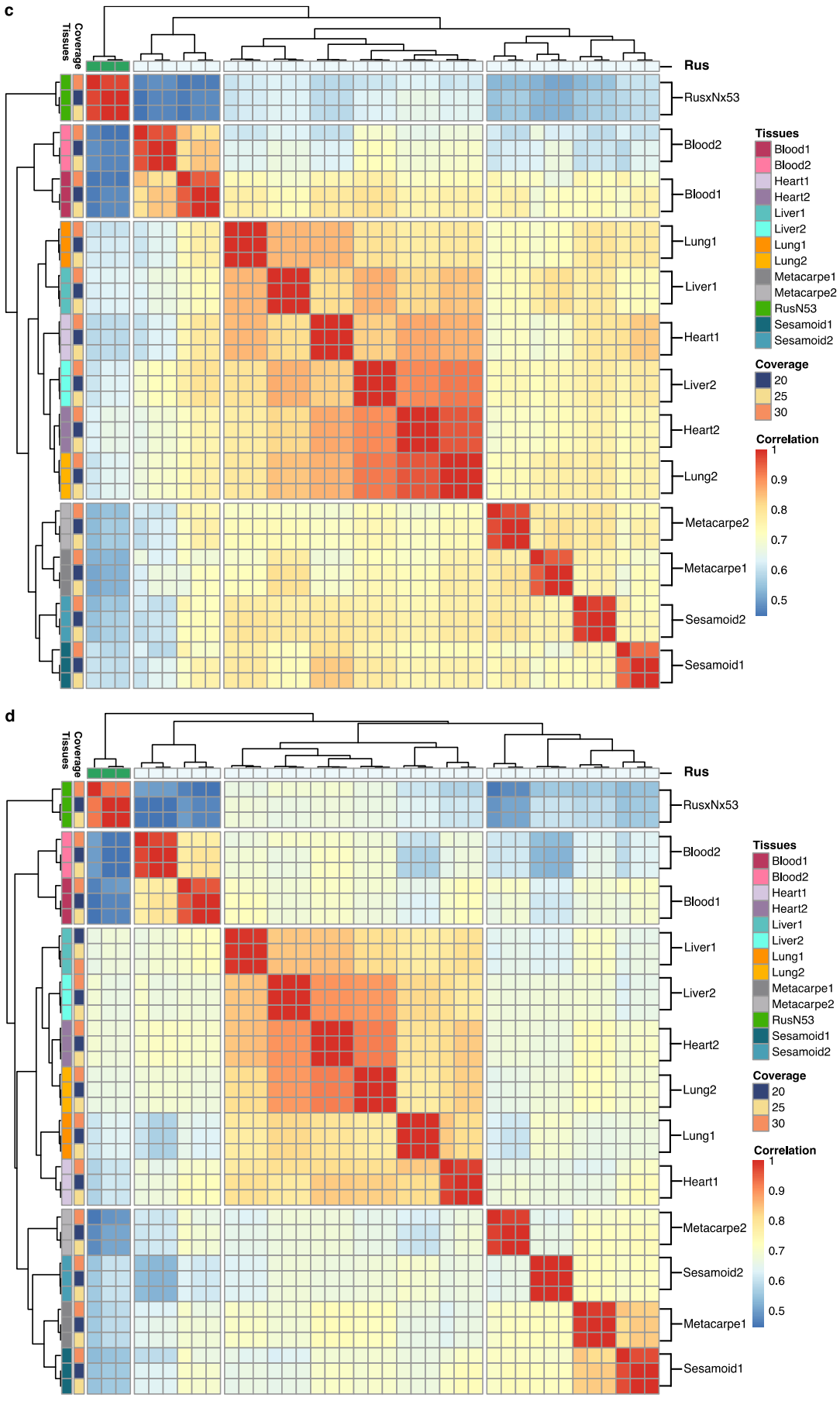
The coverage was estimated after filtering alignments for quality scores inferior to 25. The correlation coefficient was based on Pearson correlation and the use of Euclidean clustering distance. The correlation was done across different coverage: 10X **(a)**, 15X **(b)**, 20X **(c)**, 25X **(d)**, 30X **(e)**. For every test, the higher the coverage values, the better the comparison between each RRBS tissue. The selected coverage for further analysis is 25X.

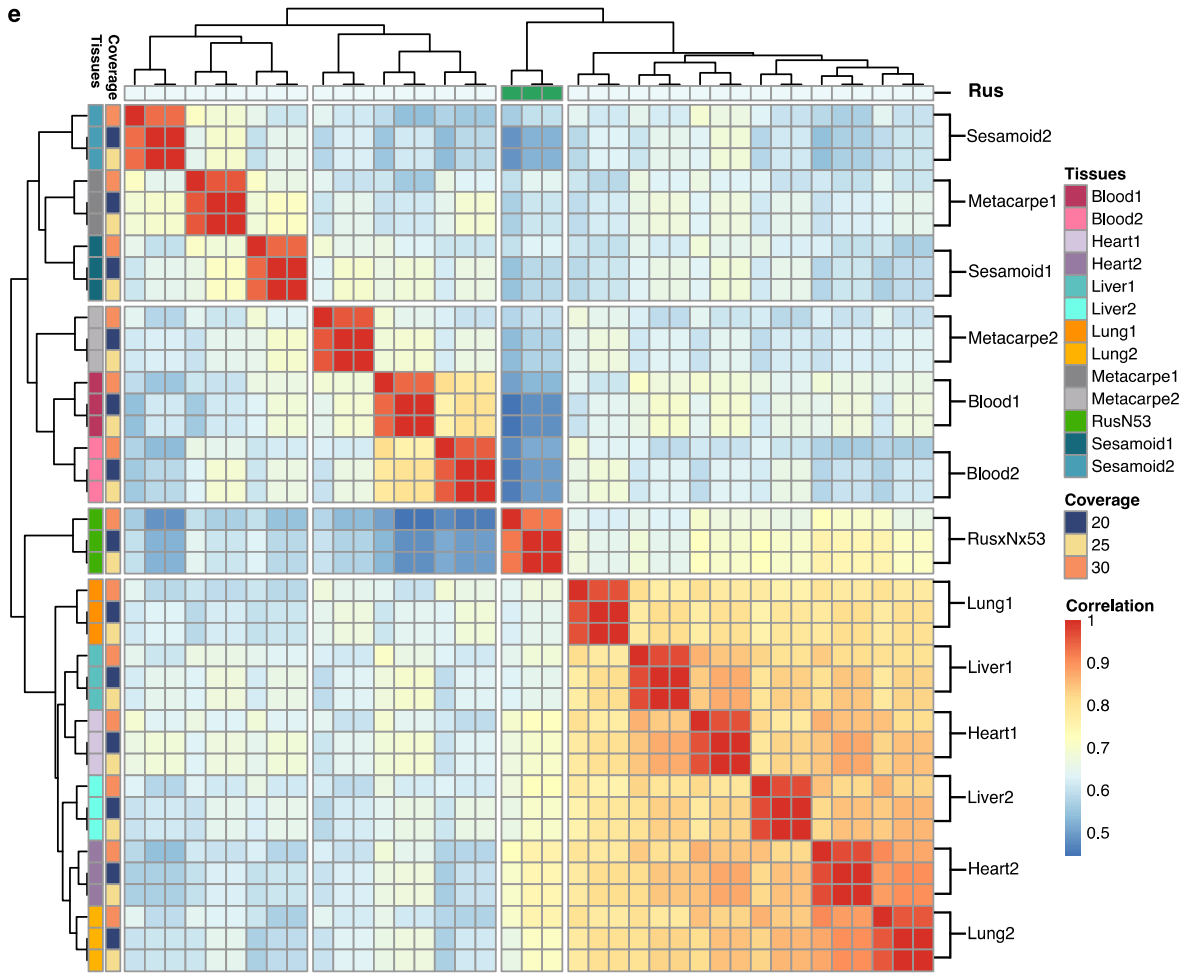


Supplementary Figure 2. Correlation matrix between the selected RRBS samples with different outlier filter.

The coverage was estimated after filtering alignments for quality scores inferior to 25. The selected coverage for further analysis is 25X. The correlation coefficient was based on Pearson correlation and the use of Euclidean clustering distance. For the selected coverage, the correlation was tested after removing progressively outliers' tissues. (**a**) All the tissues are tested. (**b**) Two outliers have been discarded: Blood2 and Sesamoid2. (**c**) Three outliers have been discarded: Blood2, Sesamoid2 and Lung1. The more the outliers are removed, the better the correlation value between each RRBS tissue.


a
Coverage
Tissues
Rus
RusxNx53
Blood2
Liver1
Lung1
Heart1
Lung2
Heart2
Liver2
Metacarpe1
Sesamoid2
Blood1
Metacarpe2
Sesamoid1
Tissues
Blood1
Blood2
Heart1
Heart2
Liver1
Liver2
Lung1
Lung2
Metacarpe1
Metacarpe2
RusN53
Sesamoid1
Sesamoid2
Coverage
20
25
30
Correlation
1
0.9
0.8
0.7
0.6
0.5
b
Coverage
Tissues
Rus
RusxNx53
Heart1
Liver2
Heart2
Lung2
Lung1
Liver1
Blood2
Blood1
Sesamoid2
Metacarpe2
Metacarpe1
Sesamoid1
Tissues
Blood1
Blood2
Heart1
Heart2
Liver1
Liver2
Lung1
Lung2
Metacarpe1
Metacarpe2
RusN53
Sesamoid1
Sesamoid2
Coverage
20
25
30
Correlation
1
0.9
0.8
0.7
0.6
0.5


c
Coverage
Tissues
Rus
RusxNx53
Blood2
Blood1
Lung1
Liver1
Heart1
Liver2
Heart2
Lung2
Metacarpe2
Metacarpe1
Sesamoid2
Sesamoid1
Tissues
Blood1
Blood2
Heart1
Heart2
Liver1
Liver2
Lung1
Lung2
Metacarpe1
Metacarpe2
RusN53
Sesamoid1
Sesamoid2
Coverage
20
25
30
Correlation
1
0.9
0.8
0.7
0.6
0.5
d
Coverage
Tissues
Rus
RusxNx53
Blood2
Blood1
Liver1
Liver2
Heart2
Lung2
Lung1
Heart1
Metacarpe2
Sesamoid2
Metacarpe1
Sesamoid1
Tissues
Blood1
Blood2
Heart1
Heart2
Liver1
Liver2
Lung1
Lung2
Metacarpe1
Metacarpe2
RusN53
Sesamoid1
Sesamoid2
Coverage
20
25
30
Correlation
1
0.9
0.8
0.7
0.6
0.5



Supplementary Figure 3. Comparison between all the modern samples and RusxNx53 for different window sizes.

Three different coverage have been tested (20X, 25X, 30X). All the RRBS tissues are tested against RusxNx53 (green). The methylation inference for the ancient sample has been calculated using the selected sites and DamMet. Different window sizes have been investigated for RusxNx53: 10bp **(a)**, 25bp **(b)**, 50bp **(c)**, 75bp **(d)**, 85bp **(e)** and 101bp (Figure16). Each window is centred around the pre-selected RRBS site. For each test, the correlation values are really low (inferior to 0.7), meaning that this last result is not reliable and should be further investigated.

VII. References

- Adler CJ, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013; 45(4):450-455e1. doi:10.1038/ng.2536
- Akalin A, et al. "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles." Genome Biology. 2012. 13(10), R87.
- Allentoft ME. et al. Population genomics of Bronze Age Eurasia. Nature. 2015. 522, 167–172.
- Allis CD and Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016; 17:487–500. Doi: 10.10 38/nrg.2016.59.
- Ameen C, et al. Specialized sledge dogs accompanied Inuit dispersal across the North American Arctic. Proc. R. Soc. Lond. B Biol. Sci. 2019. 286, 20191929.
- Andersen E, et al. Preadipocytes from obese humans with type 2 diabetes are epigenetically reprogrammed at genes controlling adipose tissue function. Int J Obes 2019. 43, 306–318. Doi: 10.1038/s41366-018-0031-3
- Andersson L, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. Genome Biol. 2015; 16:57. Doi: 10.1186/s13059-015-0622-4.
- Anthony DW. The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world. Princeton NJ: Princeton University Press. 2007.
- Arbuckle BS. Chalcolithic caprines, Dark Age dairy, and Byzantine beef: a first look at animal exploitation at Middle and Late Holocene Çadır Höyük, North Central Turkey. Anatolica. 2009. 35, 179.
- Bai L and Morozov AV. Gene regulation by nucleosome positioning. Trends Genet. 2010; 26:476–83. doi: 1016/j.tig.2010.08. 003.
- Bailey JF, et al. Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor species. Proc R Soc Lond B Biol Sci. 1996; 263(1376):1467–73.
- Bell LS, et al. The speed of post-mortem changes to the human skeleton and its taphonomic significance, Forensic Sci. Int. 1996. 82;129e140.

- Bell LS, et al. Determining isotopic life history trajectories using bone density fractionation and stable isotope measurements: a new approach, Am. J. Phys. Anthropol. 2001; 116; 66e79.
- Bell O, et al. Determinants and dynamics of genome accessibility. Nat Rev Genet. 2011 Jul 12;12(8):554-64. doi: 10.1038/nrg3017. PMID: 21747402.
- Berger SL, et al. An operational definition of epigenetics. Genes Dev. 2009; 23(7):781-783. doi:10.1101/gad.1787609
- Bergström A, et al. Origins and genetic legacy of prehistoric dogs. Science. 2020 Oct 30;370(6516):557-564. doi: 10.1126/science.aba9572. Epub 2020 Oct 29. PMID: 33122379; PMCID: PMC7116352.
- Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nature Biotechnol. **28**, 1045–1048 (2010) https://doi.org/10.1038/nbt1010-1045
- Bird A. Perceptions of epigenetics. Nature 2007 ; 447, 396–398. Doi:10.1038/nature 05913
- Boessenkool S, et al. Combining bleach and mild predigestion improves ancient DNA recovery from bones. Molecular Ecology Resources, 2016; 17(4):742–751. doi: 10.1111 /1755-0998.12623.
- Bon C, et al. Coprolites as a source of information on the genome and diet of the cave hyena. Proc Biol Sci. 2012; 279(1739):2825-2830. doi:10.1098/rspb.2012.0358
- Borry M, et al. CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content. PeerJ. 2020 Apr 17;8:e9001. doi: 10.7717/peerj.9001. PMID: 32337106; PMCID: PMC7169968.
- Borry M, et al. PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly. (2021) PeerJ 9;e11845. Doi: 10. 7717/peerj.11845
- Bos KI, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 2014. 514, 494–497.
- Botigué L, et al. Ancient European dog genomes reveal continuity since the Early Neolithic. Nat Commun 2017. 8, 16082. Doi: 10.1038/ncomms16082
- Boyko A and Kovalchuk I. Genome instability and epigenetic modification--heritable responses to environmental stress? Curr Opin Plant Biol. 2011 Jun;14(3):260-6. doi: 10.1016/j.pbi.2011.03.003. Epub 2011 Mar 26. PMID: 21440490.

- Briggs AW, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci U S A. 2007 Sep 11;104(37):14616-21. doi: 10.1073/pnas.0704665104. Epub 2007 Aug 21. PMID: 17715061; PMCID: PMC1976210.
- Briggs AW, et al. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010 Apr;38(6):e87. doi: 10.1093/nar/gkp1163. Epub 2009 Dec 22. PMID: 20028723; PMCID: PMC2847228.
- Bromberg, R., Grishin, N. V. & Otwinowski, Z. Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. *PLOS Comput. Biol.* **12**, 1–39 (2016).
- Brunson K, and Reich D. "The Promise of Paleogenomics Beyond Our Own Species." Trends in genetics : TIG 35 5 (2019): 319-329.
- Bunce M, et al. Extreme reversed sexual size dimorphism in the extinct New Zealand moa Dinornis. Nature 2003. 425: 172–175.
- Burris H and Baccarelli A. Environmental epigenetics: from novelty to scientific discipline. J. Appl. Toxicol. 2014. 34: 114–116.
- Liu Y, et al. More Arrows in the Ancient DNA Quiver: Use of Paleoepigenomes and Paleomicrobiomes to Investigate Animal Adaptation to Environment, Molecular Biology and Evolution, Volume 37, Issue 2, February 2020, Pages 307–319, doi: 10.1093/mo lbev/msz231
- Cahill JA, et al. Genomic evidence of widespread admixture from polar bears into brown bears during the last ice age. Mol. Biol. Evol. 2018. 35, 1120–1129. doi: 10.1093/molbev/msy018
- Cano RJ, et al. Amplification and sequencing of DNA from a 120–135-million-year-old weevil. Nature 1993. 363:536– 538.
- Cano RJ and Borucki MK. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science, 1995. 268(5213): p.1060-4.
- Capuano F, et al. Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. Analytical Chemistry. 2014; 86:369 7–3702. doi: 10.1021/ac500447w.
- Chandler VL. Paramutation: from maize to mice. Cell. 2007 Feb 23;128(4):641-5. doi: 10.1016/j.cell.2007.02.007. PMID: 17320501.
- Chen J, et al. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol 17, 19 (2016). Doi: 10.11 86/s13059-016-0880-9

- Conolly J, et al. Meta-analysis of zooarchaeological data from SW Asia and SE Europe provides insight into the origins and spread of animal husbandry. J Archaeol Sci. 2011; 38:538–45. Doi:10.1016/j.jas.2010.10 .008.
- Cooper A and Poinar HN. Ancient DNA: do it right or not at all. Science. 2000 Aug 18;289(5482):1139. doi: 10.1126/science.289.5482.1139b. PMID: 10970224.
- Coppinger R and Coppinger L. Dogs: A Startling New Understanding of Canine Origin, Behavior & Evolution. 2001. ISBN 0-684-85530-5
- Cruz-Dávalos DI, et al. Experimental conditions improving in-solution target enrichment for ancient DNA. Mol Ecol Resour, 2017, 17: 508-522. Doi: 10.1111/ 1755-0998.12595
- Dabney J, et al. Ancient DNA damage. Cold Spring Harb Perspect Biol. 2013; 5(7):a012567. Published 2013 Jul 1. doi:10.1101/cshperspect.a012567
- Damgaard P, et al. Improving access to endogenous DNA in ancient bones and teeth. Sci Rep 5, 11184 (2015). Doi: 10.1038/srep11184
- Daniel FI, et al. The role of epigenetic transcription repression and DNA methyltransferases in cancer. Cancer. 2011 Feb 15;117(4):677-87. doi: 10.1002/ cncr.25482. Epub 2010 Oct 13. PMID: 20945317.
- Deans C and Maggert KA. What do you mean, "epigenetic"? Genetics. 2015 Apr;199(4):887-96. doi: 10.1534/genetics.114.173492. PMID: 25855649; PMCID: PMC4391566.
- Dehasque M, et al. Inference of natural selection from ancient DNA. Evol Lett. 2020 Mar 18;4(2):94-108. doi: 10.1002/evl3.165. PMID: 32313686; PMCID: PMC7156104.
- Der Sarkissian C, et al. Ancient genomics. Philos Trans R Soc Lond B Biol Sci. 2015 Jan 19;370(1660):20130387. doi: 10.1098/rstb.2013.0387. PMID: 25487338; PMCID: PMC4275894.
- DeSalle R, et al. DNA Sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications. Science 1992. 257, 1933–1936.
- DeSalle R, et al. PCR jumping in clones of 30-million-year-old DNA fragments from amber preserved termites (Mastotermes electrodominicus). Experientia 1993. 49, 906–909.
- Ding M and Chen ZJ. Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. Curr Opin Plant Biol. 2018 Apr;42:37-48. doi: 10.1016/ j.pbi.2018.02.003. Epub 2018 Mar 7. PMID: 29502038; PMCID: PMC6058195.

- Dolle D, et al. CASCADE: A Custom-Made Archiving System for the Conservation of Ancient DNA Experimental Data. Frontiers in Ecology and Evolution, Frontiers Media S.A, 2020, 8, pp.185. ff10.3389/fevo.2020.00185ff. ffhal-02910989
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. Doi: 10.1038/nature11247.
- Fages A, et al. Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. Cell. 2019 May 30;177(6):1419-1435.e31. doi: 10.1016/j.cell.2019.03.049. Epub 2019 May 2. PMID: 31056281; PMCID: PMC6547883.
- Fagny M, et al. The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat Commun 6, 10047 (2015). Doi: 10.1038/ncomms10047
- Forrest S. The Age of the Horse, An Equine Journey Through Human History, Atlantic Monthly Press, New York 2016, p. 432.
- Fortea J, et al. Excavation protocol of bone remains for Neandertal DNA analysis in El Sidrón Cave (Asturias, Spain). J Hum Evol. 2008 Aug;55(2):353-7. doi: 10.1016/j.jhevol.2008.03.005. Epub 2008 May 16. PMID: 18485447.
- Frantz LAF, et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science. 2016 Jun 3;352(6290):1228-31. doi: 10.1126/science.aaf3161. Epub 2016 Jun 2. PMID: 27257259.
- Frantz LAF, et al. Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. Proc Natl Acad Sci U S A. 2019;11 6(35):17231–8.
- Frantz LAF, et al. Animal domestication in the era of ancient genomics. Nat. Rev. Genet. 21, 449–460 (2020).
- Frith MC, et al. A mostly traditional approach improves alignment of bisulfite-converted DNA. Nucleic Acids Res., 40 (2012), p. e100, 10.1093/nar/gks275
- Fu Q, et al. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A. 2013 Feb 5;110(6):2223-7. doi: 10.1073/pnas.1221359110.
- Fu Q, et al. The genetic history of Ice Age Europe. Nature. 534, 200–205 (2016).
- Gaffney DJ, et al. Controls of nucleosome positioning in the human genome. PLoS Genet. 2012; 8(11):e1003036. doi:10.1371/journal.pgen.1003036
- Gamba C, et al. Genome flux and stasis in a five millennium transect of European prehistory. Nat Commun. 2014 Oct 21;5:5257. doi: 10.1038/ncomms6257. PMID: 25334030; PMCID: PMC4218962.

- Gamba C, et al. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. Mol Ecol Resour. 2016 Mar;16(2):459-69. doi: 10.1111/1755-0998.12470. Epub 2015 Oct 15. PMID: 26401836.
- Gansauge MT, et al. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic Acids Res. 2017 Jun 2;45(10):e79. doi: 10.1093/nar/gkx033. PMID: 28119419; PMCID: PMC5449542.
- Gansauge MT, et al. Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA. Nat Protoc 15, 2279–2300 (2020). Doi: 10.1038/s41596-020-0338-09
- Gaunitz C, et al. Ancient genomes revisit the ancestry of domestic and Przewalski's horses. Science, 2018. 360(6384), 111–114
- Gazagnadou D. La Diffusion des Techniques et les Cultures, Editions Kimé, Paris 2013, p. 200.
- Gehring M. Prodigious plant methylomes. Genome Biol. 2016; 17:197. Doi: 10.1186/s13059-016-1065-2.
- Gerbault P, et al. Storytelling and story testing in domestication. Proc Natl Acad Sci U S A. 2014; 111:6159–64. Doi: 10.1073/pnas.14004 25111.
- Germonpré M, et al. Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. J Archaeol Sci. 2009; 36:473–490. doi: 10.1016/j.jas.2008.09.033.
- Gilbert MTP, et al. Biochemical and physical correlates of DNA contamination in archaeological human bones and teeth excavated at Matera, Italy. J. Archaeol. Sci. 32, 785–793. 2005
- Gilbert MTP, et al. Insights into the processes behind the contamination of degraded human teeth and bone samples with exogenous sources of DNA. Int. J. Osteoarchaeol., 2006. 16: 156-164. doi:10 .1002/oa.832
- Gilbert MTP, et al. Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. Science. 2008 Jun 27;320(5884):1787-9. doi: 10.1126/science.1159750.
- Glocke I and Meyer M. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Res. 2017 Jul;27(7):1230-1237. doi: 10.1101/gr.219675.116.

- Gnecchi-Ruscone GA, et al. Ancient genomic time transect from the Central Asian Steppe unravels the history of the Scythians. Sci Adv 2021 Mar 26; DOI: 10.1126/sciadv.abe4414
- Gokhman D, et al. Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. Science, 344(6183):523–527, 2014. doi: 10.1126/science.1250368.
- Gokhman D, et al. Inferring Past Environments from Ancient Epigenomes, Molecular Biology and Evolution, Volume 34, Issue 10, October 2017, Pages 2429–2438, doi: 10.1093/ molbev/msx211
- Gokhman D, et al. Reconstructing Denisovan Anatomy Using DNA Methylation Maps. Cell, 2019. 179(1), 180–192.e10. doi: 10.1016/j.cell.2019.08.035
- Gokhman D, et al. Differential DNA methylation of vocal and facial anatomy genes in modern humans. Nat Commun 11, 1189 (2020). Doi: 10.1038/ s41467-020-15020-6
- Golenberg EM, et al. Chloroplast DNA from a Miocene Magnolia species. Nature 1990, 344, 656–658.
- Goloubinoff P, et al. Evolution of maize inferred from sequence diversity of an Adh2 gene segment from archaeological specimens. Proc. Natl. Sci. USA 1993. 90, 1997-2001.
- Gonzaga PG. History of the Horse-the Iberian Horse from Ice Age to Antiquity. Robert Hale Limited. 2003.
- Gonzaga PG. A History of the Horse: the Iberian Horse from Ice Age to Antiquity. AbeBooks, 2004.
- Goodwin D. in The Welfare of Horses, N. Waran, Ed. (Springer Netherlands, Dordrecht, 2007), pp. 1–18.
- Gouil Q and Keniry A. Latest techniques to study DNA methylation. *Essays Biochem.* 2019 ;63(6):639-648. doi:10.1042/EBC20190027
- Green RE, et al. A complete neanderthal mitochondrial genome sequence determined by high-throughput sequencing. Cell, 2008. 134, 416–426
- Grehl C, et al. Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants. Frontiers in plant science, 2020, 11, 176. Doi: 10.3389/fpls.2 020.00176
- Günther T and Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet. 2019. 15:e1008302. doi: 10.1371/journal.pgen.1008302

- Guo W, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics, 14 (2013), p. 774, 10.1186/1471-2164-14-774
- Gutierrez G and Marin A. The most ancient DNA recovered from amber-preserved specimen may not be as ancient as it seems. Molecular Biological Evolution, 1998. 15, 926-929. Doi: 10.1093/oxford journals.molbev.a025998
- Haak W, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature, 2015. 522(7555), 207–211. Doi : 10.1038/nature14317
- Hagelberg E, et al. Ancient bone DNA amplified. Nature 342, 485 (1989). Doi:10.1038/342485a0
- Handt O, et al. Ancient DNA: Methodological challenges. Experientia 50, 524–529 (1994). Doi: 10.1007 /BF01921720
- Hanghøj K, et al. Fast, Accurate and Automatic Ancient Nucleosome and Methylation Maps with epiPALEOMIX. Mol Biol Evol. 2016; 33(12):3284-3298. doi: 10.1093/ molbev/msw184
- Hanghøj K and Orlando L. Ancient epigenomics. In: Lindqvist C, Rajora OP, eds. Paleogenomics. Cham: Springer, pp 1-37. 2018.
- Hanghøj K, et al. DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage, GigaScience, Volume 8, Issue 4, April 2019, giz025, doi: 10.1093 /gigascience/giz025
- Hansen A, et al. Statistical evidence for miscoding lesions in ancient DNA templates. Mol. Biol. Evol., 2001. 18, 262—5.
- Harris EY, et al. BRAT: bisulfite-treated reads analysis tool. Bioinformatics 2010; 26(4):572–3.
- Hattori N and Ushijima T. Analysis of gene-specific DNA methylation. In: Tollefsbol T, ed. Handbook of Epigenetics, 2nd ed. San Diego, CA: Academic Press; 2017:113-123.
- Hayatsu H, et al. The addition of sodium bisulfite to uracil and to cytosine. J Am Chem Soc 1970; 92:724–6.
- Hayle J, et al. Ancient DNA reveals late survival of mammoth and horse in interior Alaska. Proceedings of the National Academy of Sciences Dec 2009, 106 (52) 22352-22357; DOI: 10.1073/pnas.0912510106
- Heijmans BT, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci U S A. 2008 Nov 4;105(44):17046-9. doi:

10.1073/pnas.0806560105. Epub 2008 Oct 27. PMID: 18955703; PMCID: PMC2579375.

- Higuchi R, et al. DNA sequences from the quagga, an extinct member of the horse family. Nature 312, 282–284 (1984). doi:10.1038/312282a0
- Higuchi R, et al. "A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions." Nucleic acids research vol. 16,15 (1988): 7351-67.doi:10.1093/nar/ 16.15. 7351
- Hofreiter M, et al. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res. 2001 Dec 1;29(23):4793-9. doi: 10.1093/nar/29.23.4793. PMID: 11726688; PMCID: PMC96698.
- Hong SJ, et al. Age-related methylation patterning of housekeeping genes and tissue-specific genes is distinct between the stomach antrum and body. Epigenomics, 5(3), 283–299. (2013) doi:10.2217/epi.13.17
- Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 14, R115 (2013).
- Höss M, et al. Molecular phylogeny of the extinct ground sloth Mylodon darwinii. Proc Natl Acad Sci U S A. 1996a; 93(1):181-185. doi:10.1073/pnas.93.1.181
- Höss M, et al. DNA damage and DNA sequence retrieval from ancient tissues. Nucleic Acids Res. 1996b;24(7):1304-1307. doi:10.1093/nar/24.7.1304
- Hu CW, et al. Trace analysis of methylated and hydroxymethylated cytosines in DNA by isotope-dilution LC-MS/MS: first evidence of DNA methylation in Caenorhabditis elegans. Biochem J. 2015 Jan 1;465(1):39-47. doi: 10.1042/BJ20140844. PMID: 25299492.
- Hummel S and Hermann B. General aspects of sample preparation, in: B. Hermann, S. Hummel (Eds.), Ancient DNA, Springer Verlag, New York, 1994, pp. 59e68.
- Huynen LC, et al. Nuclear DNA sequences detect species limits in ancient moa. Nature 2003. 425: 175–178. ISSN 0277-3791, doi: 10.1016/j.quascirev.2021. 107084.
- Jablonka E. The evolutionary implications of epigenetic inheritance. Interface focus. 2017; 7(5):20160135. pmid:28839916
- Jaenicke-Despres V, et al. Early allelic selection in maize as revealed by ancient DNA. Science 2003. 302: 1206–1208.

- Jensen P. Adding 'epi-' to behaviour genetics: implications for animal domestication. J Exp Biol. 2015; 218:32–40. doi: 10.1242/jeb.106799.
- Jones P. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13, 484–492 (2012). Doi: 10.1038/nrg3230
- Jónsson H, et al. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29:1682–4.
- Joseph TA and Pe'er I. Inference of population structure from time-series genotype data. Am. J. Hum. Genet. 105, 317–333 (2019).
- Kalbfleisch TS, et al. Erratum: Author Correction: Improved reference genome for the domestic horse increases assembly contiguity and composition. Communications biology, 2019. 2, 342. Doi : 10.1038/s42003-019-0591-3
- Kaplan N, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2009; 458(7236):362–7.
- Kaur G, et al. A systematic review of smoking-related epigenetic alterations. Arch Toxicol. 2019 Oct;93(10):2715-2740. doi: 10.1007/s00204-019-02562-y. Epub 2019 Sep 25. PMID: 31555878.
- Kelekna P, Others. The horse in human history (Cambridge University Press Cambridge, 2009).
- Kemp BM, et al. Evaluation of methods that subdue the effects of polymerase chain reaction inhibitors in the study of ancient and degraded DNA. Journal of Archaeological Science, 2014. 42, 373–380. Doi:10.1016/J.JAS.2013.11.023
- Kingsley NB, et al. "Adopt-a-Tissue" Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare. Front Genet. 2021; 12:649959. Published 2021 Mar 26. doi:10.3389/fgene.2021.649959
- Kistler L, et al. A new model for ancient DNA decay based on paleogenomic meta-analysis, Nucleic Acids Research, Volume 45, Issue 11, 20 June 2017, Pages 6310–6320, doi: 10.1093/nar/gkx361
- Kohn MH and Wayne RK. Facts from feces revisited. Trends Ecol. Evol. 1997. 12: 223–27
- Korneliussen TS, et al. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014; 15:356.

- Krause J, et al. The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol. 2007 Nov 6;17(21):1908-12. doi: 10.1016/j.cub.2007.10.008. Epub 2007 Oct 18. PMID: 17949978.
- Krause J, et al. A complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol. 2010 Feb 9;20(3):231-6. doi: 10.1016/j.cub.2009.11.068. Epub 2009 Dec 31. PMID: 20045327.
- Krueger F and Andrews SR. Bismark: a exible aligner and methylation caller for bisulfite-seq applications. Bioinformatics 2011;27(11):1571–2.
- Krueger F. "Trim galore." A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files 516 (2015): 517.
- Kunde-Ramamoorthy G, et al. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic acids research, 2014. 42(6), e43. Doi: 10.1093/nar/g kt1325
- Kuo H-C, et al. DBCAT: database of CpG islands and analytical tools for identify ing comprehensive methylation profiles in cancer cells. J Comput Biol 2011;18(8):1013–7
- Lahtinen M, et al. Excess protein enabled dog domestication during severe Ice Age winters. Sci Rep 11, 7 (2021). Doi: 10.1038/s41598-020-78214-4
- Lai SR, et al. Epigenetic control of telomerase and modes of telomere maintenance in aging and abnormal systems. Front Biosci 2005;10:1779–96
- Lang J, et al. The Middle Pleistocene tunnel valley at Schöningen as a Paleolithic archive. J. Hum. Evol. 89, 18–26 (2015).
- Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
- Larson G, et al. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science. 2005; 307:1618–21. Doi: 10.1126/science. 1106927.
- Larson G and Bradley DG. How much is that in dog years? The advent of canine population genomics. PLoS Genet. 2014; 10:e1004093. Doi: 10.1371/journal.pgen. 1004093.
- Larson G. et al. Current perspectives and the future of domestication studies. Proc. Natl Acad. Sci. USA 111, 6139–6146 (2014). Doi: 10.1073/pnas. 1323964111.
- Lee JR, et al. Genome-wide analysis of DNA methylation patterns in horse. BMC Genomics. 2014 Jul 15;15(1):598. doi: 10.1186/1471-2164-15-598. PMID: 25027854; PMCID: PMC4117963.

- Leonardi M, et al. Evolutionary Patterns and Processes: Lessons from Ancient DNA. Syst Biol. 2017 Jan 1;66(1):e1-e29. doi: 10.1093/sysbio/syw059. Erratum in: Syst Biol. 2017 Jul 1;66(4):660. PMID: 28173586; PMCID: PMC5410953.
- Leonardi M, et al. Late Quaternary horses in Eurasia in the face of climate and vegetation change. Science Advances, 4(7):eaar5589, July 2018. doi: 10.1126/sciadv.aar5589.
- Li B, et al. The role of chromatin during transcription. Cell. 2007; 128:707–19. Doi: 10.1016/j.cell.2007.01.015.
- Li C, et al. Ancient DNA analysis of desiccated wheat grains excavated from a Bronze Age cemetery in Xinjiang. J. Arc. Sci. 2011. 38, 115–119. doi:10.1016/j.jas.2010.08.016
- Li H, et al. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.
- Li E and Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol.* 2014;6(5):a019133. (2014). doi:10.1101/cshperspect.a019133
- Li H and Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595 (2010).
- Li S, et al. Genetic and environmental causes of variation in epigenetic aging across the lifespan. Clin Epigenet 12, 158 (2020). Doi: 10.1186/s1 3148-020-00950-1
- Li Y and Sasaki H. Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. Cell Res 21, 466–473 (2011). Doi: 10.1038/cr.2011.15
- Librado P, et al. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. PNAS 2015. 112:E6889–97
- Librado P, et al. Ancient genomic changes associated with domestication of the horse. Science. 2017; 356(6336):442–5.
- Librado P, et al. The origins and spread of domestic horses from the Western Eurasian steppes. Nature (2021). Doi: 10.1038/s41586-021-04018-9
- Lindahl T and Andersson A. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. Biochemistry 1972. 11: 3618-3623.
- Lindahl T. Instability and decay of the primary structure of DNA, Nature 362 (1993) 709e715.

- Lippold S, et al, Discovery of lost diversity of paternal horse lineages using ancient DNA. Nat. Commun. 2, 450 (2011).
- Lister A, et al. Ancient and modern DNA in a study of horse domestication. Anc. Biomol. 2, 267–280 (1998).
- Ljungman M and Hanawalt PC. Efficient protection against oxidative DNA damage in chromatin. Mol Carcinog. 1992; 5(4):264-9. doi: 10.1002/mc.2940050406. PMID: 1323299.
- Llamas B, et al. High-resolution analysis of cytosine methylation in ancient DNA. PLoS One. 2012; 7:e30226. Doi: 10.1371/journal. pone.0030226.
- Loog L. Sometimes hidden but always there: the assumptions underlying genetic inference of demographic histories. Phil. Trans. R. Soc. 2021. Doi: 10.1098/ rstb.2019.0719
- MacHugh DE, et al. Taming the past: ancient DNA and the study of animal domestication. Annu Rev Anim Biosci. 2017; 5:329–51. Doi: 10.1146/annurev-ani mal-022516-022747.
- Malaspinas AS, et al. Estimating allele age and selection coefficient from time-serial data. Genetics. 2012; 192(2):599-607. doi:10.1534/g enetics.112.140939
- Mann AE, et al. Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. Sci. Rep. 2018. 8:9822. doi:10.1038/s41598-018-28091-9
- Marciniak S, et al. Plasmodium falciparum malaria in 1 st–2nd century CE southern Italy. Curr. Biol. 26, R1220–R1222 (2016)
- Marciniak S and Perry GH. Harnessing ancient genomes to study the history of human adaptation. Nat Rev Genet. 2017 Nov;18(11):659-674. doi: 10.1038/nrg.2017.65.
- Margaryan A, et al. Ancient pathogen DNA in human teeth and petrous bones. Ecol Evol. 2018; 8(6):3534-3542. Published 2018 Feb 26. doi:10.1002/ece3.3924
- Margulies M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005 Sep 15;437(7057):376-80. doi: 10.1038/nature03959. Epub 2005 Jul 31. Erratum in: Nature. 2006 May 4;441(7089):120. Ho, Chun He [corrected to Ho, Chun Heen]. PMID: 16056220; PMCID: PMC1464427.
- Martiniano R, et al. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol 21, 250 (2020). Doi: 10.1186/ s13059-020-02160-7

- Matsuda S, et al. Accurate estimation of 5-methylcytosine in mammalian mitochondrial DNA. Sci Rep 8, 5801 (2018). doi: 10.1038/s41598-018-24251-z
- McHugo G.P, et al. Unlocking the origins and biology of domestic animals using ancient DNA and paleo-genomics. BMC Biol 17, 98 (2019). doi: 10.1186/s12915-019-0724-7
- McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110.
- McMiken DF. Ancient origins of horsemanship. Equine Vet. J. 22, 73–78 (1990).
- Meinicke, P. Uproc: tools for ultra-fast protein domain classification. *Bioinforma.* **31**, 1382–1388 (2015).
- Metcalf JL, et al. Evaluating the impact of domestication and captivity on the horse gut microbiome. Scientific Reports, 2017. 7(1). Doi: 10.1038/s41598-017-15375-9
- Metzker ML. Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31–46 (2010).
- Meyer M, et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature 531, 504–507 (2016). Doi: 10.1038/ nature17405
- Miller W, et al. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 2008 Nov 20;456(7220):387-90. doi: 10.1038/nature07446. PMID: 19020620.
- Mullis KB and Faloona FA. Specific synthesis of DNA in Vitro via a polymerase-catalyzed chain reaction. Methods in Enzymology. 1987. 155:335-350
- Noonan JP, et al. Genomic sequencing of Pleistocene cave bears. Science. 2005 Jul 22;309(5734):597-9. doi: 10.1126/science.1113485.
- Nunn A, et al. Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. Briefings in bioinformatics, 2021. 22(5), bbab021. Doi: 10.1093/bib/bbab021
- Oliva A, et al. Systematic benchmark of ancient DNA read mapping. Brief Bioinform. 2021 Apr 8:bbab076. doi: 10.1093/bib/bbab076. Epub ahead of print. PMID: 33834210.
- Olova N, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol 19, 33 (2018). Doi: 10.1186/s13059-018-1408-2

- Olsen SL, et al. Horses and Humans: The Evolution of Human-equine Relationships. BAR international series. Archaeopress, 2006.
- Orlando L, et al. Geographic distribution of an extinct equid (Equus hydruntinus: Mammalia, Equidae) revealed by morphological and genetical analyses of fossils. Molecular ecology, 2006. 15(8), 2083–2093. Doi: 10.1111/j.1365-294X.2006.02922.x
- Orlando L, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499, 74–78 (2013). Doi: 10.103 8/nature12323
- Orlando L and Willerslev E. Evolution. An epigenetic window into the past? Science. 2014 Aug 1;345(6196):511-2. doi: 10.1126/science.1256515. PMID: 25082684.
- Orlando L, et al. Reconstructing ancient genomes and epigenomes. Nature reviews. Genetics, 2015. 16(7), 395–408. Doi: 10.1038/nrg3935
- Orlando L. The Evolutionary and Historical Foundation of the Modern Horse: Lessons from Ancient Genomics. Annu Rev Genet. 2020 Nov 23;54:563-581. doi: 10.1146/annurev-genet-021920-011805. Epub 2020 Sep 22. PMID: 32960653
- Orlando L, et al. Ancient DNA analysis. Nat Rev Methods Primers 1, 14 (2021). Doi:10.1038/s43586-020-00011-0
- Outram AK, et al. The Earliest Horse Harnessing and Milking. Science, 2009. 323(5919), 1332–1335. Doi: 10.1126/science.1168594
- Outram AK. in The Oxford Handbook of the Archaeology and Anthropology of Hunter-Gatherers, V. Cummings, P. Jordan, M. Zvelebil, Eds. (Oxford University Press, 2014).
- Pääbo S. Molecular cloning of Ancient Egyptian mummy DNA. Nature 314, 644–645 (1985). Doi: 10.1038/ 314644a0
- Pääbo S and Wilson A. Polymerase chain reaction reveals cloning artefacts. Nature 334, 387–388 (1988). Doi:10. 1038/334387b0
- Pääbo S, et al. Mitochondrial DNA sequences from a 7000-year old brain. Nucleic Acids Res. 1988. 16, 9775–9787. Doi: 10.1093/nar/16.20.9775
- Pääbo S, et al. Ancient DNA and the polymerase chain reaction. J. Biol. Chem. 1989. 264, 9709–9712.
- Pääbo S, et al. DNA damage promotes jumping between templates during enzymatic amplification. J Biol Chem. 1990 Mar 15;265(8):4718-21. PMID: 2307682.

- Pääbo S, et al. Genetic analyses from ancient DNA. Annu Rev Genet. 2004; 38:645-79. doi: 10.1146/annurev.genet.37.110801.1 43214. PMID: 15568989.
- Pachano T, et al. Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. Nat Genet 53, 1036–1049 (2021). Doi: 10.1038/s41588-021-00888-x
- Panchin, A.Y., et al. Preservation of methylated CpG dinucleotides in human CpG islands. *Biol Direct* **11,** 11 (2016). Doi : 10.1186/s13062-016-0113-x
- Park SDE, et al. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. Genome Biol. 2015; 16:234. Doi: 10.1186/s13059-015-0790-2.
- Pedersen B, et al. MethylCoder: software pipeline for bisulfitetreated sequences. Bioinformatics 2011; 27(17):2435–6.
- Pedersen JS, et al. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Res. 2014; 24(3):454-466. doi: 10.1101/gr.163592.113
- Pedersen MW, et al. Ancient and modern environmental DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. 2015. 370:20130383. doi: 10.1098/rstb.2013. 0383
- Peltzer A, et al. EAGER: efficient ancient genome reconstruction. Genome Biol. 17, 60 (2016).
- Perri A. A wolf in dog's clothing: Initial dog domestication and Pleistocene wolf variation. J Archaeol Sci. 2016; 68:1–4. doi: 10.1016/j.jas.2016.02.003.
- Pértille F, et al. Mutation dynamics of CpG dinucleotides during a recent event of vertebrate diversification. Epigenetics. 2019:1–23. pmid:31070073
- Petropoulos S, et al. Single-cell RNA sequencing: revealing human pre-implantation development, pluripotency and germline development. J Intern Med. 2016; 280:252–64. Doi: 10.1111/jo im.12493.
- Peyrégne S and Peter BM. AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination. Genome Biol. 2020 Sep 15;21(1):246. doi: 10.1186/s13059-020-02123-y. PMID: 32933569; PMCID: PMC7490890.
- Pinello L, et al. A motif-independent metric for DNA sequence specificity. *BMC Bioinformatics*. 2011;12:408.

- Pinello L, et al. Applications of alignment-free methods in epigenomics. Brief Bioinform. 2014;15(3):419-430. doi:10.1093/bib/bbt078
- Plongthongkum N, et al. Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet. 2014 Oct;15(10):647-61. doi: 10.1038/nrg3772. Epub 2014 Aug 27. PMID: 25159599.
- Poinar HN, et al. DNA from an extinct plant. Nature 1993. 363, 677.
- Poinar HN, et al. Molecular coproscopy: dung and diet of the extinct ground sloth Nothrotheriops shastensis. Science. 1998 Jul 17;281(5375):402-6. doi: 10.1126/science.281.5375.402. PMID: 9665881.
- Poinar HN, et al., Nuclear gene sequences from a late pleistocene sloth coprolite. Curr. Biol. 2003. 13: 1150–1152.
- Poinar HN, et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science. 2006 Jan 20;311(5759):392-4. doi: 10.1126/science.1123360. Epub 2005 Dec 20. PMID: 16368896.
- Portales-Casamar E, et al. DNA methylation signature of human fetal alcohol spectrum disorder. Epigenetics Chromatin. 2016 Jun 29;9:25. doi: 10.1186/s13072-016-0074-4. PMID: 27358653; PMCID: PMC4926300.
- Potts DT. Cataphractus and kamāndār: Some Thoughts on the Dynamic Evolution of Heavy Cavalry and Mounted Archers in Iran and Central Asia. Bulletin of the Asia Institute. 21, 149–158 (2007).
- Poullet M and Orlando L. Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes. Front. Ecol. Evolut. 8, 105 (2020).
- Prüfer K, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014; 505:43–9. Doi: 10.1038/nature12886.
- Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. doi: 10.1093/bioinformatics/btq033. Epub 2010 Jan 28. PMID: 20110278; PMCID: PMC2832824.
- Qvarnström M, et al. Exceptionally preserved beetles in a Triassic coprolite of putative dinosauriform origin. Curr Biol. 2021 Jun 9:S0960-9822(21)00674-6. doi: 10.1016/j.cub.2021.05.015. Epub ahead of print. PMID: 34197727.
- Ramos-Madrigal J, et al. Palaeogenomic insights into the origins of French grapevine diversity. Nat Plants. 2019 Jun;5(6):595-603. doi: 10.1038/s41477-019-0437-5.

- Rascovan N, et al. Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. Cell. 2019 Jan 10;176(1-2):295-305.e10. doi: 10.1016/j.cell.2018.11.005. Epub 2018 Dec 6. PMID: 30528431.
- Rasmussen M, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010 Feb 11;463(7282):757-62. doi: 10.1038/natu re08835. PMID: 20148029; PMCID: PMC3951495.
- Renaud G, et al. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. 2015; 16:224.
- Renaud G, et al. gargammel: a sequence simulator for ancient DNA. Bioinformatics 33, 577–579 (2017).
- Renaud G, et al. Improved de novo genomic assembly for the domestic donkey. Science advances, 2018. 4(4), eaaq0392. Doi: 10.1126/sciadv.aaq0392
- Richards M, et al. Authenticating DNA extracted from ancient skeletal remains, J. Archeol. Sci. 22 (1995) 291e299
- Roffey S, et al. Investigation of a medieval pilgrim burial excavated from the leprosarium of St Mary Magdalen Winchester. PLoS Negl. Trop. Dis. 11, e0005186 (2017).
- Roforth MM, et al. Global transcriptional profiling using RNA sequencing and DNA methylation patterns in highly enriched mesenchymal cells from young versus elderly women. Bone, 2015. 76, 49–57. Doi: 10.1016/j.bone.2015.03.017
- Rohland N, et al. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos Trans R Soc Lond B Biol Sci. 2015; 370(1660):201 30624. doi:10.1098/rstb.2013.0624
- Rohland N, et al. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat Protoc. 2018 Nov;13(11):2447-2461. doi: 10.1038/s41596-018-0050-5. PMID: 30323185.
- Römpler H, et al. Multiplex amplification of ancient DNA. Nat Protoc. 2006a; 1(2):720-8. doi: 10.1038/nprot.2006.84. PMID: 17406302.
- Römpler H, et al. Nuclear gene indicates coat-color polymorphism in mammoths. Science. 2006b Jul 7;313(5783):62. doi: 10.1126/science.1128994. PMID: 16825562.
- Rossel S, et al. Domestication of the donkey: Timing, processes, and indicators. Proceedings of the National Academy of Sciences, 2008. 105(10), pp.3715-3720.

- Rutten PF and Mill J. Epigenetic mediation of environmental influences in major psychotic disorders. Schizophrenia Research, 2009. 35, 1045-1056.
- Ryan DP and Ehninger D. Bison: bisulfite alignment on nodes of a cluster. BMC Bioinf., 15 (2014), p. 337, 10.1186/1471-2105-15-337
- Schmidt M, et al. Deconvolution of cellular subsets in human tissue based on targeted DNA methylation analysis at individual CpG sites. BMC Biol. 18, 178 (2020).
- Schmidt M, et al. DNA methylation profiling in mummified human remains from the eighteenth-century. Sci Rep 11, 15493 (2021). Doi: 10.1038/s41598-021-95021-7
- Schubert M, et al. Improving ancient DNA read mapping against modern reference genomes. BMC Genomics. 2012; 13:178.
- Schubert M, et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. 9, 1056–1082 (2014).
- Schubert M, et al. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016; 9:88. Published 2016 Feb 12. doi:10.1186/s13104-016-1900-2
- Schuenemann VJ, et al. Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. Nat Commun 8, 15694 (2017). Doi: 10.1038/ncom
- Schuenemann VJ, et al. Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. PLOS Pathogens 2018. 14(5): e1006997. Doi: 10.1371/journal.ppat.100 6997
- Seguin-Orlando A, et al. Pros and cons of methylation-based enrichment methods for ancient DNA. Scientific reports, 5:11826, 2015a. doi: 10.1038/srep11826
- Seguin-Orlando A, et al. Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns, STAR: Science & Technology of Archaeological Research, 2015b. 1:1, 1-9, DOI: 10.1179/2054892315Y.0000000005
- Seguin-Orlando A, et al. Heterogeneous Hunter-Gatherer and Steppe-Related Ancestries in Late Neolithic and Bell Beaker Genomes from Present-Day France. Curr Biol. 2021 Mar 8;31(5):1072-1083.e10. doi: 10.1016/j.cub.2020.12.015. PMID: 33434506.

- Shen W, et al. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PloS one, 2016. 11(10), e0163962. Doi: 10.1371/journal.pone.0163962
- Sherratt A, The secondary exploitation of animals in the Old World. World Archaeol. 15, 90–104 (1983).
- Sidnell P. Warhorse: Cavalry in ancient warfare. London: Continuum International Publishing group. 2006
- Sikora M, et al. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. Science. 2017 Nov 3;358(6363):659-662. doi: 10.1126/science.aao1807. Epub 2017 Oct 5. PMID: 28982795.
- Sirak K, et al. Human auditory ossicles as an alternative optimal source of ancient DNA. Genome Res. 2020 Mar;30(3):427-436. doi: 10.1101/gr.260141.119. Epub 2020 Feb 25. PMID: 32098773; PMCID: PMC7111520.
- Skinner MK. Environmental Epigenetics and a Unified Theory of the Molecular Aspects of Evolution: A Neo-Lamarckian Concept that Facilitates Neo-Darwinian Evolution. Genome biology and evolution, 2015. 7(5), 1296–1302. Doi: 10.1093/gbe/evv073
- Skoglund P, et al. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci U S A. 2014;111:2229–34.
- Skoglund P, et al. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. Curr Biol. 2015 Jun 1;25(11):1515-9. doi: 10.1016/j.cub.2015.04.019. Epub 2015 May 21. PMID: 26004765.
- Slon V, et al. Neandertal and Denisovan DNA from Pleistocene sediments. Science. 2017 May 12;356(6338):605-608. doi: 10.1126/science.aam9695.
- Slotkin RK and Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007 Apr;8(4):272-85. doi: 10.1038/nrg2072. PMID: 17363976.
- Smith RWA, et al. Detection of Cytosine methylation in ancient DNA from five native american populations using bisulfite sequencing. PloS one, 10(5):e0125344, 2015. doi: 10.1371/journal.pone.0125344.
- Snyder MW, et al. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues of origin. Cell 2016. 164:57–68.
- Soltis PS, et al. An rbcL sequence from a Miocene Taxodium (bald cypress). Proc. Natl Acad. Sci. USA 1992. 89, 449–451.

- Spyrou MA, et al. Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. Nat Commun 9, 2234 (2018). Doi: 10.1038/s41467-018-04550-9
- Spyrou MA, et al. Ancient pathogen genomics as an emerging tool for infectious disease research. Nat. Rev. Genet. 2019.20, 323–340. doi:10.1038/s41576-019-0119-1
- Struhl K and Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013 Mar;20(3):267-73. doi: 10.1038/nsmb.2506. PMID: 23463311; PMCID: PMC3740156.
- Suchan T, et al. Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. PloS one, 2016. 11(3), e0151651. Doi: 10.1371/journal.pone.0151651
- Suchan T, et al. Performance and automation of ancient DNA capture with RNA hyRAD probes. Molecular Ecology Resources. 2021. doi: 10.1111/1755-0998.13518
- Sun X, et al. A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data, Bioinformatics, Volume 34, Issue 16, 15 August 2018, Pages 2715–2723, doi: 10.1093/bioinformatics/b ty174
- Susat J, et al. A 5,000-year-old hunter-gatherer already plagued by Yersinia pestis. Cell Rep. 2021 Jun 29;35(13):109278. doi: 10.1016/j.celrep.2021.1 09278. PMID: 34192537.
- Suyama Y, et al. DNA sequence from a fossil pollen of Abies spp. from Pleistocene peat. Genes Genet Syst. 1996 Jun;71(3):145-9. doi: 10.1266/ggs.71.145. PMID: 8828176.
- Taron UH, et al. Testing of Alignment Parameters for Ancient Samples: Evaluating and Optimizing Mapping Parameters for Ancient Samples Using the TAPAS Tool. *Genes (Basel)*. 2018;9(3):157. doi:10.3390/genes9030157
- Taylor WTT and Barrón-Ortiz C.I. Rethinking the evidence for early horse domestication at Botai. Sci Rep 11, 7440 (2021). doi: 10.1038/s41598-021-86832-9
- Thalmann O, et al. Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. Science. 2013 Nov 15;342(6160):871-4. doi: 10.1126/science.1243650. PMID: 24233726.
- Thomas R, et al. DNA phylogeny of the extinct marsupial wolf. Nature 340, 465–467 (1989). Doi: 10.1038/34046 5a0

- Tobi EW, et al. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. Hum Mol Genet. 2009; 18(21):4046-4053. doi:10.1093/hmg/ddp353
- Tobi EW, et al. Biobank-based Integrative Omics Studies Consortium, Slagboom PE, van Zwet EW, Lumey LH, Heijmans BT. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. Sci Adv. 2018 Jan 31;4(1):eaao4364. doi: 10.1126/sciadv.aao4364. PMID: 29399631; PMCID: PMC5792223.
- Tran H, et al. Objective and comprehensive evaluation of bisulfite short read mapping tools. Advances in bioinformatics, 2014, 472045. Doi: 10.1155/2014/472045
- Tran H. Evaluating and Improving Performance of Bisulfite Short Reads Alignment and the Identification of Differentially Methylated Sites. (2018).
- Tsuji J and Weng Z. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. Briefings in bioinformatics, 2016. 17(6), 938–952. Doi: 10.1093/bib/bbv103
- van der Valk T, et al. Million-year-old DNA sheds light on the genomic history of mammoths. Nature. 2021 Mar;591(7849):265-269. doi: 10.1038/s41586-021-03224-9. Epub 2021 Feb 17. PMID: 33597750; PMCID: PMC7116897.
- Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The third revolution in sequencing technology. Trends Genet. 34, 666–681 (2018).
- Verdugo MP, et al. Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. Science. 2019; 365(6449):173–6.
- Vernot B, et al. Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. Science doi: 10.1126/science.abf1 667 (2021).
- Vigne J-D. The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. C. R. Biol. 334, 171–181 (2011).
- Vilà C, et al. Widespread origins of domestic horse lineages. Science. 2001; 291:474–7. Doi: 10.1126/science.2 91.5503.474.
- Voss TC and Hager GL. Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet. 2014 Feb;15(2):69-81. doi: 10.1038/nrg3623. Epub 2013 Dec 17. PMID: 24342920; PMCID: PMC6322398.
- Waddington CH. The epigenotype. Endeavour 1942. 1, 18–20.

- Wagner S, et al. Uncovering signatures of DNA methylation in ancient plant remains from patterns of post-mortem DNA damage. Front. Ecol. Evol. 8, 11 (2020).
- Wales N, et al. New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. Biotechniques. 59, 368–371 (2015).
- Wang, Y. Transcriptome and genome analysis based on alignment-free protocols. Bioinformatics [q-bio.QM]. Université Paris-Saclay, 2021. English. ffNNT : 2021UPASL048ff. fftel-03370851f
- Wang Y, et al. Late Quaternary dynamics of Arctic biota from ancient environmental genomics. Nature 600, 86–92 (2021). https://doi.org/10.1038/s41586-021-04016-x
- Warinner C, et al. Pathogens and host immunity in the ancient human oral cavity. Nat. Genet. 46, 336–344 (2014).
- Warinner C, et al. A Robust Framework for Microbial Archaeology. Annu Rev Genomics Hum Genet. 2017 Aug 31;18:321-356. doi: 10.1146/annurev-genom-091416-035526. Epub 2017 Apr 26. PMID: 28460196; PMCID: PMC5581243.
- Warmuth V, et al. European domestic horses originated in two holocene refugia. PLoS One. 6, e18194 (2011).
- Weber M, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005 Aug;37(8):853-62. doi: 10.1038/ng1598. Epub 2005 Jul 10. PMID: 16007088.
- Weidner CI, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 15, R24 (2014).
- Wickham H. ggplot2: elegant graphics for data analysis. Springer New York. 2009.
- Willerslev E, et al. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science. 2003 May 2;300(5620):791-5. doi: 10.1126/science.1084114. Epub 2003 Apr 17. PMID: 12702808.
- Willerslev E, et al. Long-term persistence of bacterial DNA. Curr. Biol. 2004; 14:R9–R10.
- Woodward SR, et al. DNA sequence from Cretaceous period bone fragments. Science. 1994 Nov 18;266(5188):1229-32. doi: 10.1126/science.7973705. PMID: 7973705.
- Wutke S, et al. Decline of genetic diversity in ancient domestic stallions in Europe. Sci Adv.4, eaap9691 (2018).
- Xi Y and Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 2009; 10:232.

- Xi Y, et al. RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. Bioinformatics 2012; 28(3):430–2.
- Ząbek T, et al. Methylation Marks of Blood Leukocytes of Native Hucul Mares Differentiated in Age. International Journal Of Genomics, 2019, 1-9. doi: 10.1155/2019/2 839614
- Zavala EI, et al. Pleistocene sediment DNA reveals hominin and faunal turnovers at Denisova Cave. Nature (2021). doi:10.1038/s41586-021-03675-0
- Zeder MA. The origins of agriculture in the Near East. Curr Anthropol. 2011; 52(S4):S221–S35.
- Zeder MA. The domestication of animals. J Anthropol Res. 2012 a;68:161–90. Doi: 10.3998/jar.0521004.0068.201.
- Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. Genome Res. **18**, 821–829 (2008).
- Zhang L, et al. The sequence preference of DNA methylation variation in mammalians. PLoS One. 2017;12(10):e0186559. Published 2017 Oct 18. doi: 10.1371/journal.pone.0186559
- Zheng H, Wu H, Li J, et al. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC Med Genomics*. 2013; 6(Suppl. 1):S13.
- Zhenilo S, et al. Epigenetics of Ancient DNA. Acta Naturae, 2016. 8(3), 72-76. doi: 10.32607/20758251-2016-8-3-72-76
- Zhou Z, et al. Gene expression in the addicted brain. Int Rev Neurobiol. 2014; 116:251-73. doi: 10.1016/B978-0-12-801105-8.00010-2. PMID: 25172478; PMCID: PMC4427035.
- Zhu J, et al. On the nature of human housekeeping genes. Trends Genet 2008; 24:481.
- Zielezinski, A., et al. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol **18,** 186 (2017). https://doi.org/10.1186/s13059-017-1319-7
- Zischler H, et al. Detecting dinosaur DNA. Science. 1995 May 26;268(5214):1192-3; author reply 1194. doi: 10.1126/science.7605504. PMID: 7605504.