MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 1780–1798 doi: 10.1111/mec.12690

Shotgun microbial profiling of fossil remains

C. DER SARKISSIAN,*[1] L. ERMINI,*[1] H. JÓNSSON,* A. N. ALEKSEEV,† E. CRUBEZY,‡
B. SHAPIRO§ and L. ORLANDO*

*Centre for Geogenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, Copenhagen K 1350, Denmark, †North-Eastern Federal University, Belinskiy str, 58, suite 312, Yakutsk, Russia, ‡Molecular Anthropology and Image Synthesis Laboratory, Faculté de Médecine, Paul Sabatier University of Toulouse, 37 allées Jules Guesde, Toulouse Cedex 3 31073, France, §Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064, USA

Abstract

Millions to billions of DNA sequences can now be generated from ancient skeletal remains thanks to the massive throughput of next-generation sequencing platforms. Except in cases of exceptional endogenous DNA preservation, most of the sequences isolated from fossil material do not originate from the specimen of interest, but instead reflect environmental organisms that colonized the specimen after death. Here, we characterize the microbial diversity recovered from seven *c.* 200- to 13 000-year-old horse bones collected from northern Siberia. We use a robust, taxonomy-based assignment approach to identify the microorganisms present in ancient DNA extracts and quantify their relative abundance. Our results suggest that molecular preservation niches exist within ancient samples that can potentially be used to characterize the environments from which the remains are recovered. In addition, microbial community profiling of the seven specimens revealed site-specific environmental signatures. These microbial communities appear to comprise mainly organisms that colonized the fossils recently. Our approach significantly extends the amount of useful data that can be recovered from ancient specimens using a shotgun sequencing approach. In future, it may be possible to correlate, for example, the accumulation of *postmortem* DNA damage with the presence and/or abundance of particular microbes.

Keywords: ancient DNA, metagenomics, microbial profiling, next-generation sequencing

Received 22 November 2013; revision received 23 January 2014; accepted 2 February 2014

Introduction

Recent developments in ultra-high-throughput DNA sequencing technologies (next-generation sequencing, NGS) have considerably advanced ancient DNA research, giving rise to the field of palaeogenomics by making possible the characterization of complete nuclear genomes from long-dead organisms (Green *et al.* 2010; Rasmussen *et al.* 2010; Reich *et al.* 2010; Keller *et al.* 2012; Meyer *et al.* 2012; Orlando *et al.* 2013). In the last few years, shotgun sequencing, which allows the recovery of genetic information from the entire population of DNA molecules extracted from a sample, has been used to analyse a variety of archaeological remains, including hair (Miller *et al.* 2008; Rasmussen *et al.* 2010), mummified tissues (Keller *et al.* 2012), calcified teeth and bones (Allentoft *et al.* 2012; Skoglund *et al.* 2012) and herbarium specimens (Martin *et al.* 2013; Yoshida *et al.* 2013). The range of ancient specimens useful for palaeogenomic analysis has also been increasing in recent years. For example, the previously hypothesized limit to DNA survival of *c.* 50 000–100 000 years (Lindahl 1993) has been pushed back by almost one order of magnitude with the publication of a draft genome of an early Middle Pleistocene horse (Orlando *et al.* 2013).

While bones and teeth constitute the densest tissues in vertebrates, the quality and quantity of DNA that can be retrieved from such material is highly variable

Correspondence: Clio Der Sarkissian and Luca Ermini, Fax: 4535322325; and E-mails: clio.dersarkissian@snm.ku.dk; luca.ermini@snm.ku.dk

[1] These authors have contributed equally.

among samples. Cold temperatures are known to favour DNA preservation (Smith *et al.* 2003), but a wide range of other environmental factors has been shown to be equally important (Campos *et al.* 2012). After deposition in aerated soils, colonization of the material by microbes disrupts mineralized tissues almost immediately (Bell *et al.* 1996; Nielsen-Marsh & Hedges 1999; Jans *et al.* 2004). This process facilitates the entry of biotic and abiotic environmental agents that contribute to *postmortem* degradation of DNA and other organic molecules by driving the formation of abasic sites, strand breaks, interstrand cross-links and atypical nucleobases through oxidative and hydrolytic reactions (e.g. Lindahl 1993; Höss *et al.* 1996; Pääbo *et al.* 2004; Dabney *et al.* 2013). Such degradation reactions limit the amount of DNA templates that are accessible for extraction and amplification. Chemical changes along the DNA sequences result in sequence errors, the frequency of which is often greater than biological mutation rates (Ho *et al.* 2005). Because the enzymes used in DNA amplification will preferentially bind to nondamaged molecules, these factors also increase the sensitivity of the genetic assays to contamination by modern DNA.

Colonization by environmental microorganisms is known to increase the porosity of ancient bones and teeth (Jans *et al.* 2004; Campos *et al.* 2012). Increased fossil porosity makes the samples more permeable to water circulating within the depositional environment, which can carry free DNA from other organisms (e.g. DNA leaching, Haile *et al.* 2007). It also increases the risk of contamination during excavation and processing (Gilbert *et al.* 2005). Evidence suggests that teeth (Oota *et al.* 1995; Gilbert *et al.* 2005) are less susceptible to contamination than bones.

In exceptionally well-preserved bone remains, as much as 70% of sequence reads derived from the DNA of the target organism can be retrieved (Reich *et al.* 2010; Meyer *et al.* 2012). However, nonendogenous DNA is generally a serious concern for palaeogenomic research. For most remains, the proportion of endogenous DNA is very small: 1–5% for Neandertal (Green *et al.* 2010) and 0.5–9% for ancient horse DNA extracts (Orlando *et al.* 2011, 2013; Ginolhac *et al.* 2012). Target enrichment approaches have been developed to increase the relative proportion of endogenous versus nonendogenous reads recovered from a DNA extract, with the aim to achieve a cost-effective characterization of mitochondrial genomes (Briggs *et al.* 2009; Orlando *et al.* 2013), exomes (Burbano *et al.* 2010) and even full genomes (Carpenter *et al.* 2013; Schuenemann *et al.* 2013).

It has been proposed that nonendogenous DNA is found mostly near the surfaces of bones and teeth, whereas crystal aggregates in the most interior parts of these samples constitute DNA preservation niches, where endogenous DNA is protected from chemical degradation and microbial attack (Geigl 2002; Malmström *et al.* 2005; Salamon *et al.* 2005). DNA extraction protocols targeting these hypothetical niches aim to increase DNA recovery (Rohland & Hofreiter 2007; Schwarz *et al.* 2009; Orlando *et al.* 2011; Ginolhac *et al.* 2012), with some particularly dedicated to the recovery of very short DNA fragments (Dabney *et al.* 2013). One class of these methods relies on a two-step procedure where the undigested material from a first digestion is pelleted via centrifugation and then redigested. While both digestion fractions can be used for subsequent DNA extraction, pairwise comparisons of those two fractions revealed higher proportions of endogenous ancient DNA following the second digestion and lower levels of DNA damage (Schwarz *et al.* 2009; Orlando *et al.* 2011; Ginolhac *et al.* 2012). Although based on a limited number of samples, and all from permafrost environments, these results imply the existence of protective preservation niches, where endogenous DNA is most efficiently accessed following an initial predigestion step. It also limits the efficiency of nondestructive DNA extraction methods (Rohland *et al.* 2004), which only use DNA recovered from the initial digestion and are therefore likely to be dominated by nonendogenous DNA.

Much work has gone into characterizing endogenous DNA content in the two extraction fractions (Schwarz *et al.* 2009; Orlando *et al.* 2011; Ginolhac *et al.* 2012); however, little is known about the microbial diversity present in these fractions. Assuming that the second fraction is enriched relative to the first fraction for DNA that was protected in preservation niches, but that the entire sample remained at least somewhat permeable to water and microbial attack, no difference would be expected between the two fractions in the overall composition of the microbial community. Alternatively, if different microbial communities are observed, these communities could have very different influences on endogenous DNA, for example by driving different damage kinetics. If the latter is true, then it is possible that microbial community dynamics, and *not* the existence of molecular preservation niches, could explain the lower DNA degradation levels and higher endogenous DNA content observed in the second fraction.

Here, we analyse microbial communities present in the two extraction fractions to test whether differences in microbial communities can explain the better preservation of ancient DNA in second extraction fractions, which presumably reflect DNA preserved more deeply within ancient remains. We use a shotgun sequencing approach and a variety of computational methods (METAPHLAN: Segata *et al.* 2012; LEFSE: Segata *et al.* 2011; PYNAST: Caporaso *et al.* 2010a; QIIME: Caporaso *et al.* 2010b) to profile the microbial genetic composition in seven ancient horse bones and teeth extracted using a two-step digestion protocol (Ginolhac *et al.* 2012). We obtained microbial profiles for both first and second extraction fractions and assess levels of endogenous DNA and contamination by modern humans. In addition, we quantify and compare the type and amount of DNA damage in each fraction and correlate this with differences in the microbial communities.

Material and methods

Samples sites and data sets

We performed metagenomics analyses on one 13 000-year-old and six 200- to 300-year-old horses whose remains were recovered from permafrost-preserved archaeological sites in Siberia (Fig. 1). Sample Taimyr Peninsula (TP) was previously described in Ginolhac *et al.* (2012). All the other samples were collected from fossilized remains of Yakut horses, a native breed from Yakutia, Sakha Republic, Russia (Table 1).

DNA extraction

We performed DNA extraction from ancient horse bones and teeth at the ancient DNA facilities of the Centre for GeoGenetics, University of Copenhagen, Denmark, following strict procedures for limiting contamination by modern DNA. Two extraction fractions per sample were obtained as described in Ginolhac *et al.* (2012) following the silica-based DNA extraction protocol from Orlando *et al.* (2009). In short, 263–4000 mg of tooth/bone powder (Table 2) was first digested overnight in a buffer containing EDTA, *N*-lauroylsarcosine and proteinase K. After centrifugation of the digest, the supernatant and the remaining undigested bone powder were separated and treated separately: (i) the supernatant yielded a first extraction fraction according to the protocol in Ginolhac *et al.* (2012); (ii) the undigested bone powder was digested a second time before being subjected to the same DNA extraction procedure as (i). The second fractions (ii) are hereafter labelled with the suffix ‘RE’.

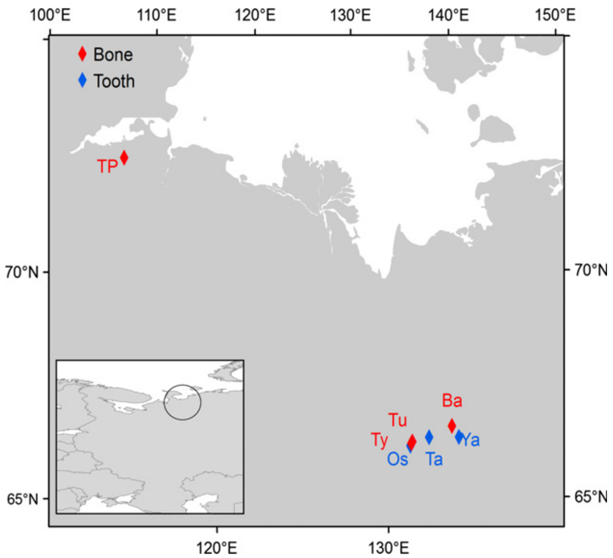

Fig. 1 Map of the sites sampled for ancient horse DNA.

Sequencing

We built shotgun libraries for both the first and second fractions (RE) following the protocols for Illumina blunt-end DNA libraries described in Orlando *et al.* (2013) and Seguin-Orlando *et al.* (2013) and using 500 nM of adapters. DNA contamination from the laboratory and reagents was monitored through mock extractions and DNA libraries that were built at the same time as those for the horse samples. Amplified library concentrations were estimated on a Bioanalyzer instrument using a High-Sensitivity chip (Agilent Technologies) for both controls and ancient samples. In controls, concentrations of the DNA fragments in the expected size range (<125 bp) were too low to be detected by the Bioanalyzer, implying concentration below 5 pg/uL. Concentrations of the libraries constructed from ancient extracts were comprised between 50 and 17 311 pg/uL. As the amount of DNA present in negative controls was orders of magnitude lower than for ancient samples, libraries obtained from blanks were not sequenced. For ancient samples, indexed libraries were sequenced on the Illumina HiSeq2000 platform, except for sample TP for which both Solexa GAIIx and Helicos sequence data were generated (Table 2).

We obtained Helicos true Single Molecule-DNA Sequencing (tSMS) data for samples TP1, TP1RE, TP2, TP2RE from Ginolhac *et al.* (2012; Sequence Read Archive, SRA, Accession nos. SRA045862; SRS264738–41).

Sequence analysis

Solexa and Illumina reads were trimmed and adapter sequences removed using the program ADAPTERREMOVAL version 1.5 (Lindgreen 2012) with the minimal read length set to 25 bp. The DNA sequence data generated in this study have been deposited on the SRA (Accession nos. SRR1030038–SRR1030052). We aligned the trimmed reads against the *Equus caballus* nuclear reference genome (EQUCAB2.0, including chrUn; http://genome-euro.ucsc.edu/cgi-bin/hgGateway?db=equCab2&redirect=auto&source=genome.ucsc.edu, see Wade *et al.*

Table 1 Description of the ancient horse samples

Sample number	ID	Site name and coordinates	Date	Sample type	Context
CGG10023	TP	Taimyr Peninsula (73°2.77′N, 109°43.33′E)	UBA-1679: 14 900–14 044 radiocarbon calibrated (Orlando *et al.* 2013)	Bone	Permafrost washout
CGG101392	OS	OursSire2 (66°50.66′N, 131°45.21′E)	19th century AD	Tooth	Deposit
CGG101393	Ba	Bakhtakh (67°09.37′N, 134°31.01′E)	18th/19th century AD	Bone	Horse grave
CGG101394	Ya	Yakutia (near 66°52.92′N, 131°50.50′E)	19th century AD	Tooth	Domestic fossa
CGG101395	Ty	Tysarastaak2 (66°53.92′N, 131°50.35′E)	18th century AD	Bone	Deposit
CGG101396	Ta	Targana1 (66°59.18′N, 132°59.08′E)	19th century AD	Tooth	Tomb
CGG101397	Tu	Tumeski (66°56.11′N, 131°55.22′E)	19th century AD	Bone	Tomb

BP, before present; AD, Anno Domini.

Table 2 Description of ancient horse DNA extracts and sequencing information

Sample number	ID	Fraction	Quantity (mg)	Sequencing platform	Reference
CGG10023	TP1	1	4000	Helicos tSMS	Ginolhac *et al.* 2012;
	TP1RE	2			
	TP2	1	2160	Helicos tSMS	Ginolhac *et al.* 2012
	TP2RE	2			
	TP3	1	2160 (same extract as TP2)	Illumina GAIIx	This study
	TP3RE	2			
CGG101392	OS	1	605	Illumina HiSeq2000	This study
	OSRE	2			
CGG101393	BaARE	2	673	Illumina HiSeq2000	This study
	BaB	1	263		
	BaBRE	2			
CGG101394	Ya	1	482	Illumina HiSeq2000	This study
	YaRE	2			
CGG101395	Ty	1	326	Illumina HiSeq2000	This study
	TyRE	2			
CGG101396	Ta	1	361	Illumina HiSeq2000	This study
	TaRE	2			
CGG101397	Tu	1	727	Illumina HiSeq2000	This study
	TuRE	2			

2009) and the complete mitochondrial genome sequence (Accession no. NC_001788; Xu & Arnason 1994). For the mitochondrial read alignment, we duplicated the first 30 nucleotides at the end of the reference sequence in order to cope with the circular nature of this genome and to enable read alignment across the first and last positions. We estimated human contamination levels by mapping against the *Homo sapiens* nuclear (hg19 GRCh37; Meyer *et al.* 2013) and mitochondrial genomes (revised Cambridge Reference Sequence; GenBank Accession no. NC_012920.1; Andrews *et al.* 1999). Reads were aligned to the reference sequences using BWA version 0.5.9 (Li & Durbin 2009) with the parameters suggested by Schubert *et al.* 2012: seed option disabled and edit distance relaxed (option $-n$ 0.04). Only reads with a mapping quality superior to 25 were retained. PCR duplicates were removed using the *MarkDuplicates* function of PICARD TOOLS version 1.82 (http://picard.sourceforge.net/) and the SAMtools *view* utility (Li *et al.* 2009), as described in Schubert *et al.* (2014).

Comparison between first and second fractions

We performed a paired t-test to compare percentages of endogenous (horse) and contaminating (human) DNA between first and second fractions. We carried out a two-sample t-test to compare percentages of whole endogenous versus whole contaminating DNA. The same test was also applied after the removal of reads potentially representing genomic regions conserved between horses and humans, that is, reads that mapped to both the horse and the human reference genomes. All tests were performed in the statistical environment R version 3.0.1 (R Core Team 2013).

Analysis of ancient DNA damage

We analysed DNA fragmentation and misincorporation patterns in all fractions using the package MAPDAMAGE version 2.0.1 (Ginolhac *et al.* 2011; Jónsson *et al.* 2013) that first computes misincorporation patterns from NGS data sets and then incorporates a statistical model of *postmortem* DNA damage to estimate degradation parameters using a Bayesian framework. We used default settings with the option –forward (5′ end only) on the BAM alignment files (Li *et al.* 2009) resulting from mapping of each horse data set.

For the most recent samples (OS/OSRE, BaB/BaBRE, Ya/YaRE, Ty/TyRE, Ta/TaRE, Tu/TuRE), we compared the amount of damage in the first and second fractions using a paired t-test on the posterior means of the parameters δ_d (deamination in double strands), δ_s (deamination in single strands) and λ (probability of reads not terminating in overhangs). We excluded sample BaA from this analysis as no data could be generated from the corresponding first fraction (Table 3).

Microbial profiling using METAPHLAN

We profiled the microbial communities present in the extracts using METAPHLAN (METAGENOMIC PHYLOGENETIC ANALYSIS VERSION 1.7.7, February 2013; Segata *et al.* 2012). First, we mapped metagenomic reads from each extract to the markers of the METAPHLAN database using the default parameters of the BOWTIE2 version 2.1.0 aligner (Langmead & Salzberg 2012) and a sensitive global alignment strategy (default –end-to-end mode) for read mapping.

Read clonality, that is, the nonhomogenous and spurious over-representation of sequences arising during presequencing PCR amplification of libraries, could bias the calculation of the microbial relative abundance. Therefore, we identified PCR duplicates as reads whose first base mapped at an identical position in a given METAPHLAN database marker using the *MarkDuplicates* function of PICARD TOOLS version 1.82 (http://picard.sourceforge.net/). Duplicated reads were then excluded using the SAMtools *view* utility (Li *et al.* 2009) before running METAPHLAN. Using the METAPHLAN program, microbial taxonomic groups and their relative abundance were determined at all taxonomic levels (option –tax_lev 'a'; Table S1, Supporting Information). We excluded samples Tu and TP2RE, as none of the nonendogenous sequences could be classified given the existing database.

Low-abundance filtering

In order to reduce the rate of false positives, we arbitrarily excluded low-abundance taxa (<1%) from further analyses. Taxonomic profile analyses for unfiltered data sets are shown in Fig. S1, Supporting Information.

Statistical analyses

We compared taxon abundance, as estimated from METAPHLAN, among extracts at all taxonomic levels using a suite of statistical analyses in R. Relative abundance was visualized through heatmaps (function *heatmap2* in the *gplots* package; http://cran.r-project.org/package=gplots) and stacked barplots (function *ggplot* in the *ggplot2* package; http://cran.r-project.org/package=ggplot2). Shannon diversity indices were computed from relative abundance data using the function *diversity* of the *vegan* package in R (http://cran.r-project.org/package=vegan; Table S2, Supporting Information).

We performed principal component analyses (PCA) of relative abundance using the R function *prcomp* (option scale = TRUE). The first four components accounting for more than 10% of the variance were plotted in R (Fig. S2, Supporting Information).

We calculated Bray–Curtis distances among profiles from taxon relative abundance. The obtained distance matrices were used then for principal coordinate analysis (PCoA, R function *pcoa*; Figs S3 and S4, Supporting Information). We performed hierarchical clustering using the R package *pvclust*, based on the Manhattan metric and average linkage clustering method (Suzuki & Shimodaira 2006). Ten thousand bootstrap iterations were applied to hierarchical clustering in order to estimate P-values (approximately unbiased and bootstrap probabilities) for the support of each cluster.

We tested for consistent differences in taxon abundance among groups of samples using the nonparametric Kruskal–Wallis sum-rank test and the unpaired Wilcoxon test. We also performed a linear discriminant analysis, which estimates the effect size of taxonomical covariates driving the group differences, following the procedure implemented in the LEFSE program (Linear

Table 3 Illumina sequencing of ancient horse DNA extracts and mapping metrics

Sample ID	TP3	OS	BaA	BaB	Ya	Ty	Ta	Tu
First fractions								
Raw reads (raw)	32.0 M	12.5 M	n/a	10.4 M	8.6 M	11.6 M	12.0 M	14.4 M
Post-trimming reads	31.5 M	12.4 M	n/a	9.5 M	8.5 M	11.2 M	11.7 M	12.7 M
Unique horse reads* (nuclear + mitochondrial)	5.4 M	23 828	n/a	52 793	143 873	92 907	48 843	1.2 M
Clonality (horse)	0.23	0.01	n/a	0.07	0.01	0.07	0.02	0.08
Horse nuclear genome coverage	0.16	5.89E-4	n/a	1.32E-3	3.95E-3	2.52E-3	1.40E-3	3.00E-2
Horse mitochondrial genome coverage	4.58	0.16	n/a	0.19	0.13	0.08	0.06	1.34
Unique human reads* (nuclear + mitochondrial)	90 823	1460	n/a	4238	3374	10 586	2657	32 416
Clonality (human)	0.33	0.02	n/a	0.15	0.02	0.08	0.04	0.11
Human mitochondrial genome coverage	0.02	0.01	n/a	0.01	0.00	0.01	0.00	0.02
% Endogenous†	17.07	0.19	n/a	0.56	1.69	0.83	0.42	9.66
% Human contamination‡	0.29	0.01	n/a	0.04	0.04	0.09	0.02	0.25
Sample ID	TP3RE	OSRE	BaARE	BaBRE	YaRE	TyRE	TaRE	TuRE
Second fractions								
Raw reads (raw)	32.0 M	16.4 M	18.3 M	9.8 M	7.2 M	12.2 M	29.4 M	15.1 M
Post-trimming reads	31.5 M	10.8 M	11.9 M	8.9 M	7.0 M	12.0 M	12.4 M	14.9 M
Unique horse reads* (nuclear + mitochondrial)	8.5 M	39 490	292 517	288 544	791 030	319 482	208 182	7.7 M
Clonality (horse)	0.06	0.23	0.05	0.12	0.02	0.04	0.49	0.02
Horse nuclear genome coverage	0.22	1.00E-3	8.84E-3	6.24E-3	2.14E-2	7.74E-3	5.90E-3	1.92E-1
Horse mitochondrial genome coverage	5.14	0.34	1.33	0.57	2.99	0.67	0.84	7.33
Unique human reads* (nuclear + mitochondrial)	140 105	5463	11 763	11 403	11 201	34 436	8691	147 741
Clonality (human)	0.08	0.40	0.19	0.15	0.03	0.08	0.63	0.04
Human mitochondrial genome coverage	0.03	0.02	0.03	0.04	0.04	0.09	0.05	0.14
% Endogenous†	26.88	0.36	2.65	3.22	11.25	2.66	1.68	51.78
% Human contamination‡	0.44	0.05	0.10	0.13	0.16	0.29	0.07	0.99
Sample ID	TP3 (RE)	OS (RE)	BaARE	BaB (RE)	Ya (RE)	Ty (RE)	Ta (RE)	Tu (RE)
Ratio second/first fractions§	1.02	0.44	n/a	2.03	1.66	1.06	1.30	1.38

M, millions.

*After duplicate removal.

†100 × horse reads/post-trimming reads.

‡100 × human reads/post-trimming reads.

§$(\text{Horse/human})_{\text{second fraction}}/(\text{horse/human})_{\text{first fraction}}$.

discriminant analysis Effect Size; Segata *et al.* 2011). Statistical significance alpha values for the factorial Kruskal–Wallis test among classes and for the pairwise Wilcoxon test between subclasses were set to 0.05, and the threshold on logarithmic linear discriminant analysis score for discriminative features was set to 2.0.

Accounting for differences in sequencing depths

Metagenomic data sets vary in number of reads, percentages of endogenous horse reads and possible human contaminants and hence could possibly vary in the representativeness of the microbial diversity of the samples. Potential biases associated with differences in sequencing depths among data sets were accounted for by analysing subsamples of the full read data sets. For each data set, we generated ten subsamples by randomly extracting nonduplicated reads mapping to the METAPHLAN database markers. The number of reads down-sampled was the same for all data sets and corresponded to the minimal number of reads mapping to the METAPHLAN database markers after removing duplicates (i.e. 796 reads for the TP1 data set). METAPHLAN profiling and statistical analyses were performed on all subsamples.

Analysis of 16S ribosomal DNA in the shotgun data sets

We compared results obtained from the microbial profiling of the ancient horse extracts using METAPHLAN to those inferred from the analysis of the microbial 16S ribosomal DNA (16S rDNA) data contained in the shotgun data sets. We extracted microbial 16S rDNA reads by aligning the full shotgun data sets to the Greengenes core set (DeSantis *et al.* 2006) using PYNAST (with minimum percentage of sequence identity to closest BLAST hit, or −p option, set to 90; Caporaso *et al.* 2010a). The taxonomy of the aligned reads was assigned using the RDP classifier (Wang *et al.* 2007) in the QIIME environment (Caporaso *et al.* 2010b). Relative abundance calculated from 16S rDNA data was analysed as described previously.

GC content of mapped sequences in METAPHLAN markers

We calculated the GC content of the aligned regions of the METAPHLAN reference markers using an in-house Perl script available upon request. We assessed differences in the % GC of mapped regions between the first and second fractions using a linear mixed-effect model with the overall mean μ and the fraction type covariate τ_i as fixed effects.

$$H_1 : y_{ijk} = \mu + \tau_i + \nu_j + \varepsilon_{ijk}$$

$$H_0 : y_{ijk} = \mu + \nu_j + \varepsilon_{ijk}$$

The within-sample dependence is modelled as the random factor ν_j; furthermore, the error ε_{ijk} and the ν_j terms are independent and identically distributed, $N(0, \sigma^2)$ and $N(0, \sigma_\nu^2)$, respectively. The mixed-effect model used here allowed correcting for hierarchical or nested structure of the data as we investigate the fraction effect. The *lmer* function in the R package *lme4* (http://cran.r-project.org/package=lme4) was used for the mixed-effects model fitting with the option REML=FALSE. The asymptotic chi-square test was used to assess the statistical significance of the likelihood difference.

Results

Characteristics of the shotgun data sets

We generated NGS metagenomic data from the first and second ('RE') DNA fractions obtained from seven horse bone and tooth fossils preserved in permafrost (Fig. 1; Table 1). For fractions obtained from the Ba horse, only the second fraction (BaARE) yielded sufficient amounts of DNA for sequencing. To ensure that this sample was available for comparison, we performed a second DNA extraction from Ba, from which DNA could be sequenced from both the first (BaB) and second (BaBRE) fractions. In total, we obtained 19 metagenomic shotgun data sets—nine for first fractions and 10 for second fractions—from seven fossilized horse specimens (Table 2), representing a total number of 242.2 million sequences (Table 3).

Endogenous horse DNA content and human contamination

First extraction fractions contained lower percentages of endogenous genomic horse DNA than second fractions (mean first fraction: 4.34%, mean second fraction: 13.97%; *P*-value: 1.3×10^{-3}) as shown in Fig. 2 and Table 3.

All extracts contained more endogenous horse DNA (mean: 9.16%) than contaminating human DNA (mean: 0.20%; *P*-value 1.7×10^{-5}; Fig. 2). Because DNA of horse origin but representing highly conserved genomic regions will align to both the horse and human reference sequences, estimated proportions of human DNA will likely be overestimates (Schubert *et al.* 2012). Similarly to horse DNA, first fractions appear to contain a lower percentage of reads of human origin than second fractions (mean first fraction: 0.11%, mean second fraction: 0.30%; *P*-value: 1.3×10^{-4}). This is most likely an

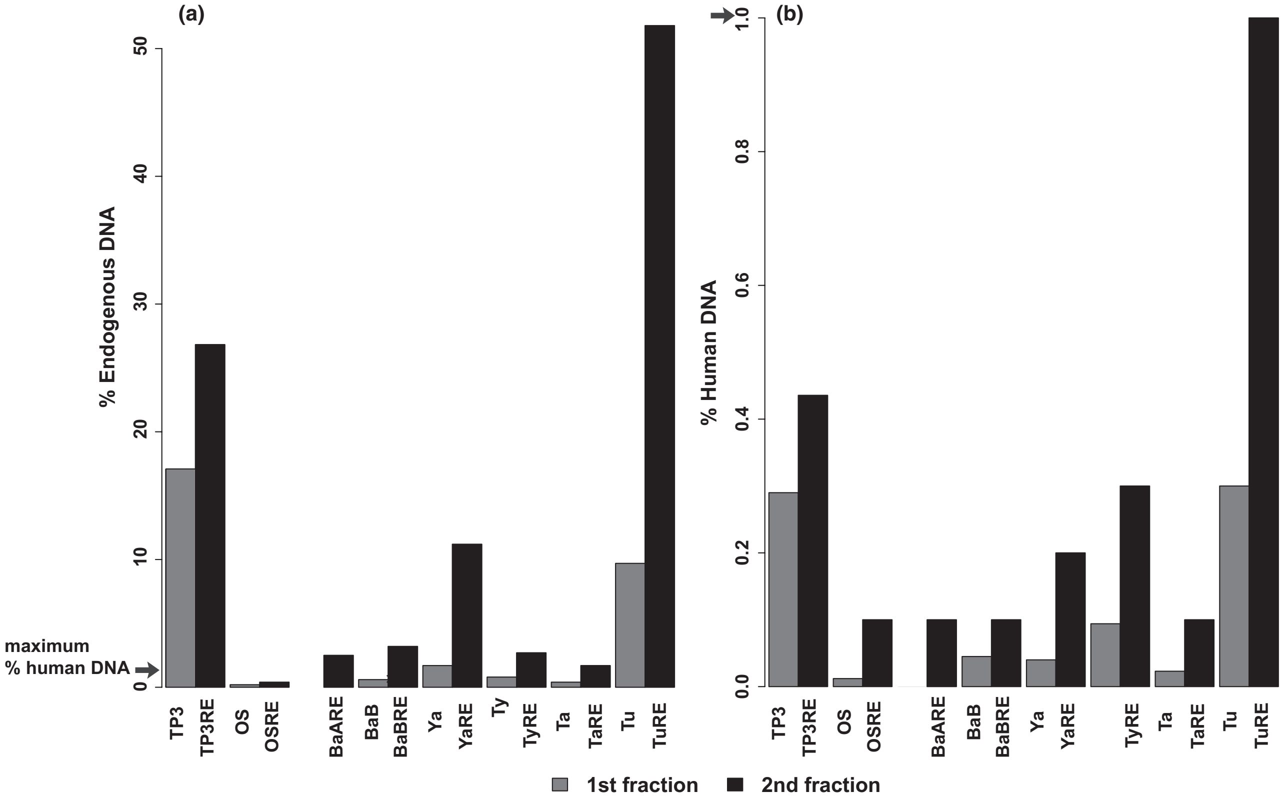

Fig. 2 Percentages of endogenous (a) and human DNA (b) in ancient horse extracts. Note the different scales for the *y*-axis of plots A- and B- (as indicated by the arrow).

artefact of a larger proportion of horse reads mapping to highly conserved regions of the human genome as we observe that the horse/human ratio in the first over the second fraction is approximately close to 1 (Table 3). Similar trends were observed when comparing only regions unique to horses and humans in first and second fractions (Fig. S5 and Table S3, Supporting Information). While no statistical difference was observed between proportions of endogenous DNA preserved in bones versus teeth, we observed a slight trend towards higher proportions of endogenous DNA in bone (Tables 2 and 3).

Analysis of ancient DNA damage

We investigated whether the difference in endogenous DNA content observed between the first and second fractions could be explained by a difference in length distribution of DNA inserts, for example with the second fraction systematically showing an enrichment for shorter endogenous molecules (Fig. S6, Supporting Information). No particular difference in the size distribution of reads unique to horses was observed for samples TP3/TP3RE, OS/OSRE, Ya/YaRE and Ta/TaRE. Interestingly, the size distribution for other samples showed a *c.* 10 bp periodicity, which has been recently suggested as reminiscent of nucleosomal DNA protection (Pedersen *et al.* 2014). However, whereas the second fraction was enriched in slightly shorter reads for samples BaB/BaBRE and Ty/TyRE, the opposite was observed for sample Tu/TuRE (Fig. S6, Supporting Information), which rules out length bias as an explanation for the difference in endogenous DNA content observed between the first and second fractions.

We next used MAPDAMAGE 2.0 to investigate DNA damage patterns from first and second fractions (Jónsson *et al.* 2013). MAPDAMAGE 2.0 takes advantage of high-quality sequence alignments in order to estimate deamination rates within single-stranded and double-stranded parts of DNA molecules as well as one parameter related to the size of overhangs (δ_s, δ_d and λ, respectively). The extracts for samples OS/OSRE, Ty/TyRE and Tu/TuRE showed sporadic or no signs of damage on any parameter investigated (Fig. 3), most likely as a result of their relatively young age (18–19th centuries; Table 1). Although similar in age, all extracts for samples BaARE/BaB/BaBRE, Ta/TaRE and Ya/YaRE display molecular signatures of DNA damage, probably due to different taphonomic conditions.

Higher levels of damage were observed in the significantly older TP3/TP3RE sample (13 000 years old).

For samples BaB/BaBRE, Ta/TaRE and Ya/YaRE, we observed similar parameter estimates for first and second fractions (δ_d *P*-value = 0.89; δ_s *P*-value = 0.82; λ *P*-value = 0.54). In contrast, when considering the most ancient sample (TP3/TP3RE), δ_s was found to be 2.4-fold higher in the first fraction than in the second fraction, in agreement with previous reports (Ginolhac *et al.* 2012). This finding may suggest a temporal dependency of the differences in DNA damage levels observed in the first and second fractions; however, older samples would need to be investigated to test this hypothesis. Finally, the extent of DNA damage in the ancient horse samples does not seem to correlate with tissue type or excavation context.

Microbial profiling of ancient horse extract metagenomic data

We found no differences between first and second fractions at the genus level (Table 4; Table S2, Supporting Information) either in the number of identified taxa, in the Shannon's diversity index or in the cumulative percentage of low-abundance taxa <1%.

In all extracts (Table S1, Supporting Information), *Actinobacteria* (22.7–97.0%) and *Proteobacteria* (1.6–72.8%) were the most abundant phyla. *Bacteroidetes* (0.0–21.0%) and *Chloroflexi* (0.0–9.7%) were also found in most extracts (Fig. 4). *Actinobacteria* was the most taxonomically diverse phylum with *Mycobacterium* (4.8–40.4%), *Arthrobacter* (0.0–66.5%), *Brevibacterium* (0.0–51.9%), *Rhodococcus* (0.0–22.5%), *Frankia* (0.0–14.3%) and *Janibacter* (0.0–10.6%) the most ubiquitous and abundant genera across samples (Fig. 5).

The bacterial phyla *Actinobacteria*, *Proteobacteria*, *Bacteroidetes* and *Chloroflexi* have been identified previously from a wide range of soils using molecular approaches (Janssen 2006; Fierer *et al.* 2012). When comparing the microbial taxonomic profiles from horse extracts to those from soil samples (shotgun sequencing data from Fierer *et al.* 2012), PCoA showed that horse samples fell within the diversity range of nonwarm desert soils (i.e. among temperate forests, boreal forest, prairie, tropical forest and cold desert; Fig. S7, Supporting Information). The differentiation of the microbial profiles obtained from the different TP horse shotgun data sets and a polar soil data set was caused by high relative abundance of the genus *Arthrobacter*, which has been long known for being ubiquitous and particularly abundant in soils (Conn 1928). Altogether, these results support the hypothesis that the microbial diversity observed in the ancient horse extracts represents a complex subset

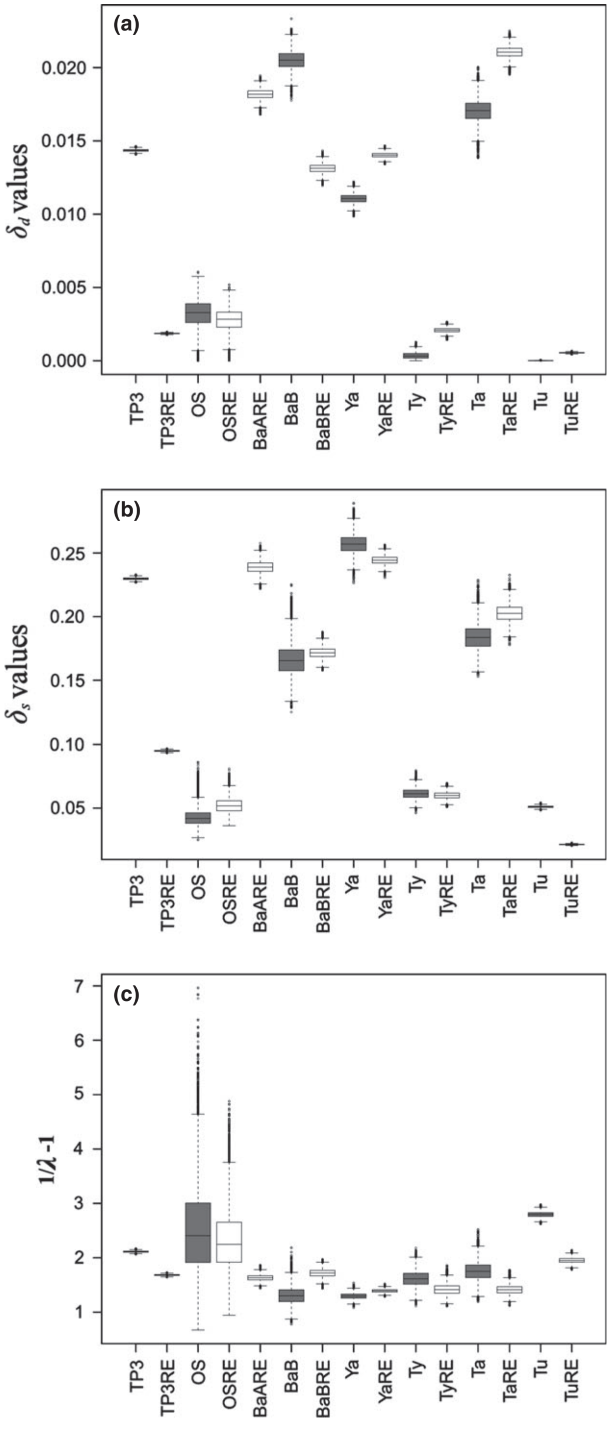

Fig. 3 DNA damage in ancient horse DNA. The boxplot shows the simulated posterior distribution of the three DNA damage model parameters δ_d, δ_s and λ. (a) δ_d, deamination in double strands. (b) δ_s, deamination in single strand. (c) λ, probability of reads not terminating in overhangs, converted in a size estimate of average overhang length (in bp) using the following formula: $1/\lambda - 1$. In grey, first fractions; in white, second fractions.

of the microbial diversity of the depositional soil environment. This is in line with previous results based on low-depth 16S rDNA sequencing that showed the presence of microbes typical of Antarctic and Arctic permafrost environments in 100- to 300-year-old human remains collected in Yakutia (Thèves *et al.* 2011).

Distribution of the microbial diversity in ancient horse extracts

We found no statistical differences between first and second fractions using either Bray–Curtis distance PCoA (Fig. 6) or PCA (Fig. S2, Supporting Information) of the microbial relative abundance (Figs S3 and S4, Supporting Information). Instead, the data cluster by sample into three groups: cluster I (TP3, TP3RE, TP1, TP1RE, TP2), cluster II (Ya, YaRE) and cluster III (OS, OSRE, BaARE, BaB, BaBRE, Ty, TyRE, Ta, TaRE, TuRE). This clustering is further supported by high bootstrap *P*-values in hierarchical clustering analyses of Manhattan distances (Fig. 7). This structure does not seem to be driven by other factors such as sample geographical origin, age, type (bone/tooth) or find context (deposit/tomb).

We further compared the microbial relative abundance in the first versus the second fraction using the linear discriminant analysis procedure implemented in LEFSE (Segata *et al.* 2011), which identifies and quantifies the importance of taxonomical groups driving possible differences among predefined groups. This confirmed the absence of significant differences across fractions. Biomarkers, that is, features quantitatively responsible for the differences observed among the clusters, were identified for each cluster: the *Arthrobacter* and *Rhodopseudomonas* genera for cluster I, the *Brevibacterium*, *Brachybacterium*, *Dietzia*, *Rhodococcus*, and *Streptomyces* genera for cluster II and the *Mycobacterium*, *Pseudomonas*, *Mesorhizobium* genera and an unclassified genera of the *Methylocystaceae* family for cluster III (Fig. 8).

Additionally, we observed no significant differences between the microbial profiles generated from the four extracts available for the TP horse individual, three of which sequenced on the Helicos (TP1, TP1RE, TP2, TP2RE) and one on the Illumina GAIIx (TP3, TP3RE) sequencing platforms. This suggests that the molecular tools used for constructing and amplifying Illumina DNA libraries did not affect the underlying metagenomic composition reflected in the absence of library preparation and amplification with tSMS. Finally, PCoA for microbial profiles of multiple subsamples confirmed results obtained previously (Fig. S8, Supporting Information), thus showing that the observed structure of the microbial diversity into three clusters was not an artefact arising from differences in sequencing depth.

Validation of the METAPHLAN-based microbial profiling

The accuracy of the METAPHLAN microbial profiles depends on the diversity and distribution of microbial sequences in the reference database and the vast fraction of the microbial genomic variation present on Earth remains uncharacterized. To test whether METAPHLAN could be applied reliably to environmental samples, we compared the microbial profiles from METAPHLAN to profiles established by analysing only the sequences of the shotgun data sets that aligned to taxonomy-informative microbial 16S rDNA sequences, for which large comparative databases are available. Comparable class abundance was estimated using the METAPHLAN and 16S rDNA approaches (Fig. S9, Supporting Information). While 16S rDNA-based profiles did not differ between first and second fractions, they segregated samples into the same clusters as those defined by METAPHLAN using PCoA (Fig. S10, Supporting Information), suggesting that the results are not an artefact of the METAPHLAN approach.

Table 4 Characteristics of the microbial DNA profiles

Sample ID	TP1	TP2	TP3	OS	BaA	BaB	Ya	Ty	Ta	Tu
First fractions										
Number of taxa identified	5	4	20	21	n/a	19	13	20	25	n/a
Shannon diversity index	1.07	1.08	1.56	2.51	n/a	2.25	1.54	2.19	2.63	n/a
% Low-abundance taxa	0.00	0.00	9.90	5.19	n/a	2.88	2.53	1.66	7.12	n/a
Sample ID	TP1RE	TP2RE	TP3RE	OSRE	BaARE	BaBRE	YaRE	TyRE	TaRE	TuRE
Second fractions										
Number of taxa identified	17	n/a	16	20	21	31	16	19	18	10
Shannon diversity index	1.58	n/a	1.64	2.26	2.37	2.84	1.54	2.20	2.56	2.07
% Low-abundance taxa	4.57	n/a	5.76	4.20	5.09	3.96	4.78	4.64	1.13	0.00

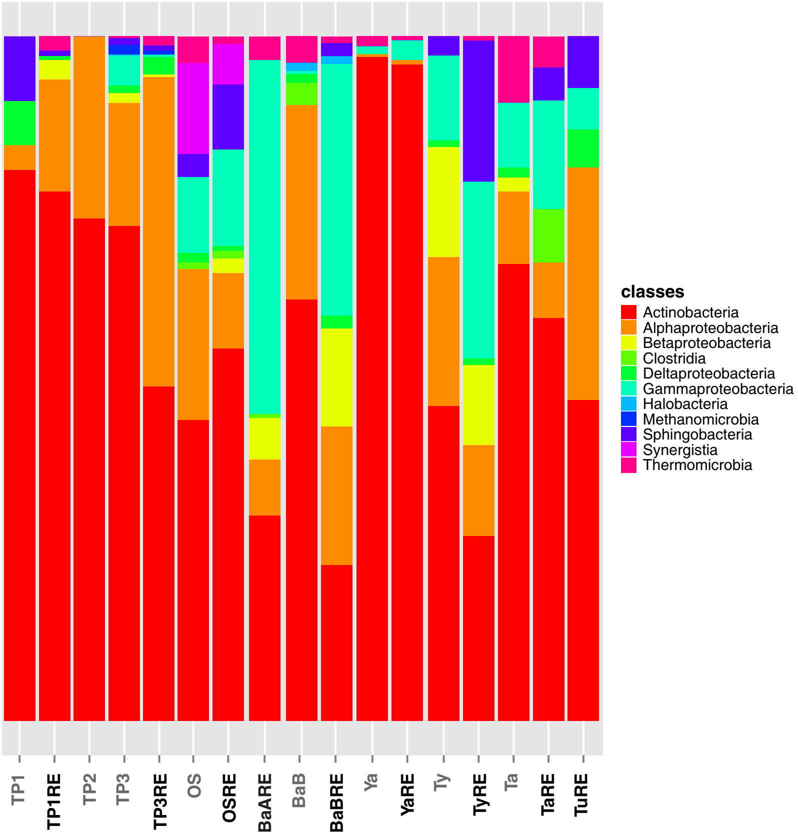

Fig. 4 Relative abundance of microbial classes in ancient horse DNA extracts.

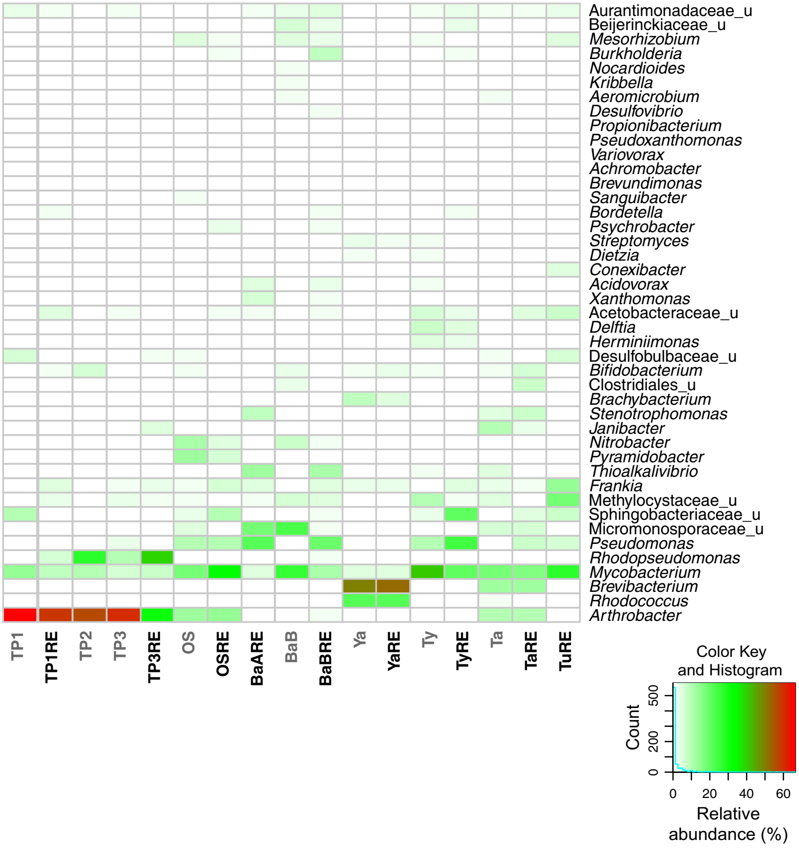

Fig. 5 Heatmap showing relative abundance of microbial genera in ancient horse DNA extracts. '_u', unclassified.

Comparison of % GC content

We verified that populations of reads that aligned to METAPHLAN database markers from first and second fractions were characterized by similar GC contents (non-significant *P*-value of 0.49 for the mixed-effect model-based likelihood ratio test correcting for nested structure).

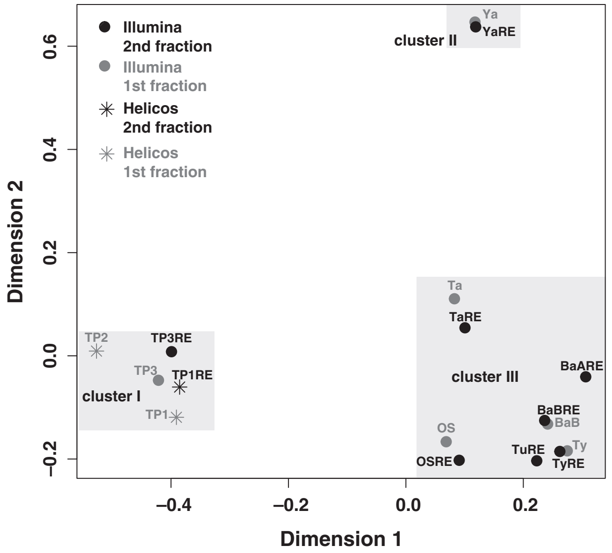

Fig. 6 Principal coordinate analysis of Bray–Curtis distances between microbial DNA profiles at the genus level in ancient horse extracts.

Damage of microbial DNA in ancient horse DNA extracts

In order to test whether it was possible to infer the age of the microbes colonizing the ancient horse remains, and therefore understand when these microbes first entered the bone or tooth, we assessed the amount of *postmortem* damage in the microbial DNA reads (Fig. 9). Shotgun reads were first mapped to a set of microbial genomes representing the most abundant genera (<1%) on the basis of METAPHLAN profiling. We characterized DNA damage patterns of the three bacteria for which the highest genome coverage was obtained. Results show an absence of microbial damage patterns in all samples, even for those where typical DNA damage patterns are observed from endogenous horse reads (BaARE/BaB/BaBRE, Ta/TaRE, TP3/TP3RE and Ya/YaRE; Fig. 3). This suggests that the microbial DNA sequenced from these extracts may postdate the age of the horse specimen, and thus likely may reflect recent colonization from the environment.

Discussion

We present a thorough taxonomic profiling of microbial reads in shotgun data sets generated from two extraction fractions of ancient specimens. We found no substantial differences in microbial diversity between the first and second fractions in the ancient Yakut horse samples, but we observed a higher percentage of

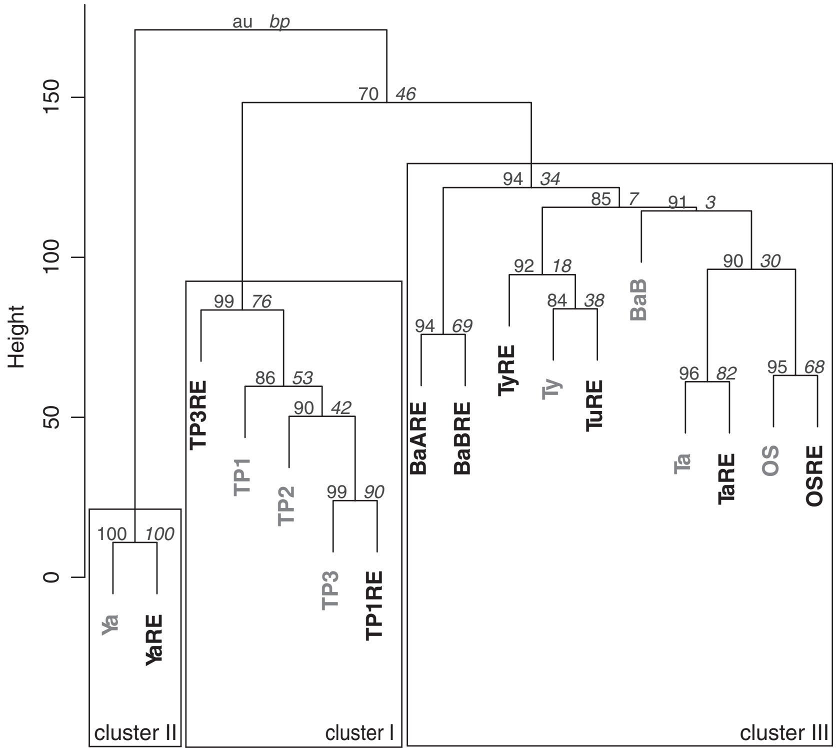

Fig. 7 Hierarchical clustering of Manhattan distances between microbial DNA profiles at the genus level in ancient horse extracts (10 000 bootstraps).

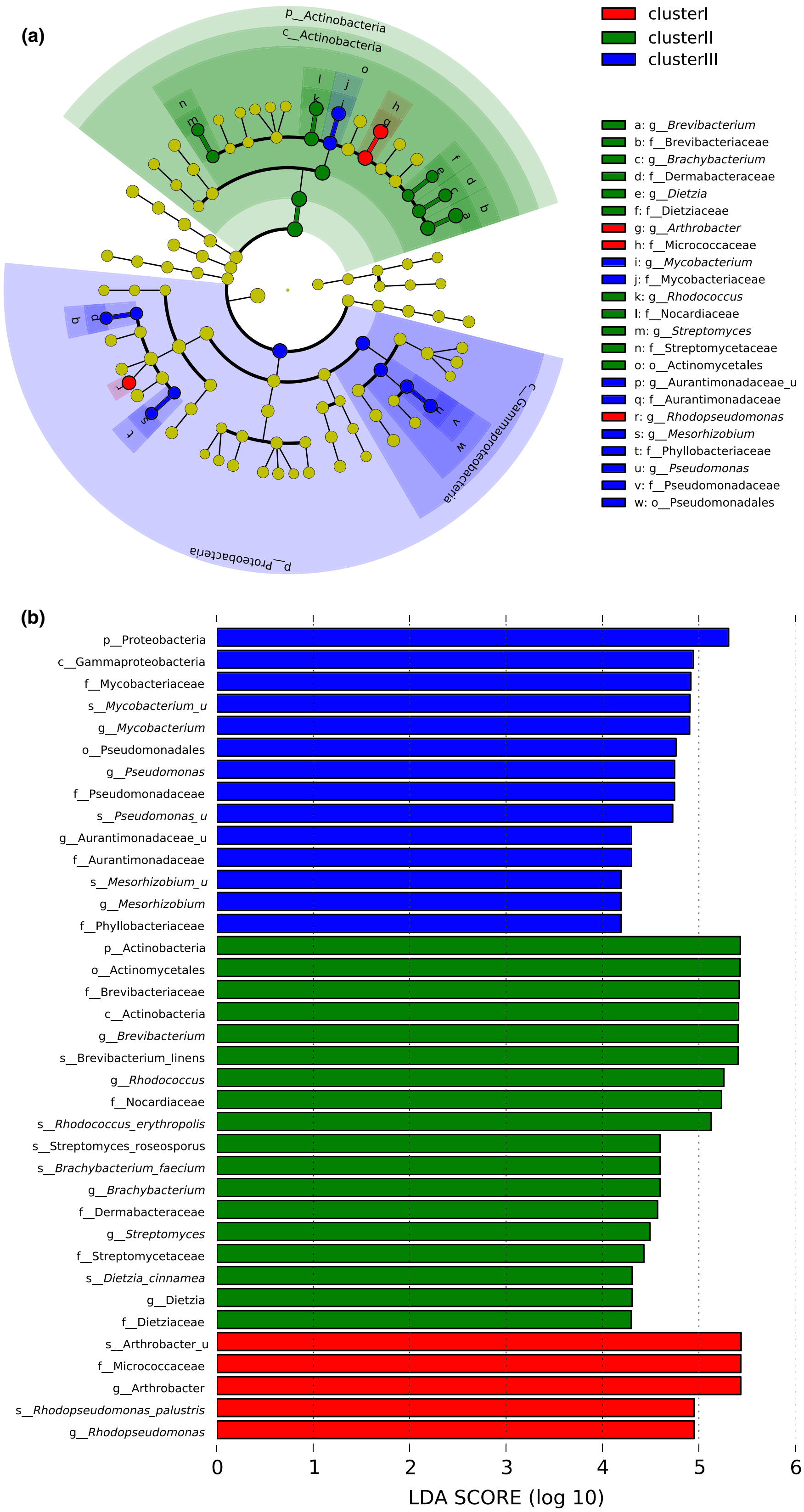

Fig. 8 Differentiating microbial features (biomarkers) among clusters of ancient horse DNA extracts, as identified and quantified by LEFSE. (a) Biomarkers ranked by effect size. (b) Cladogram of biomarkers. '_u', unclassified.

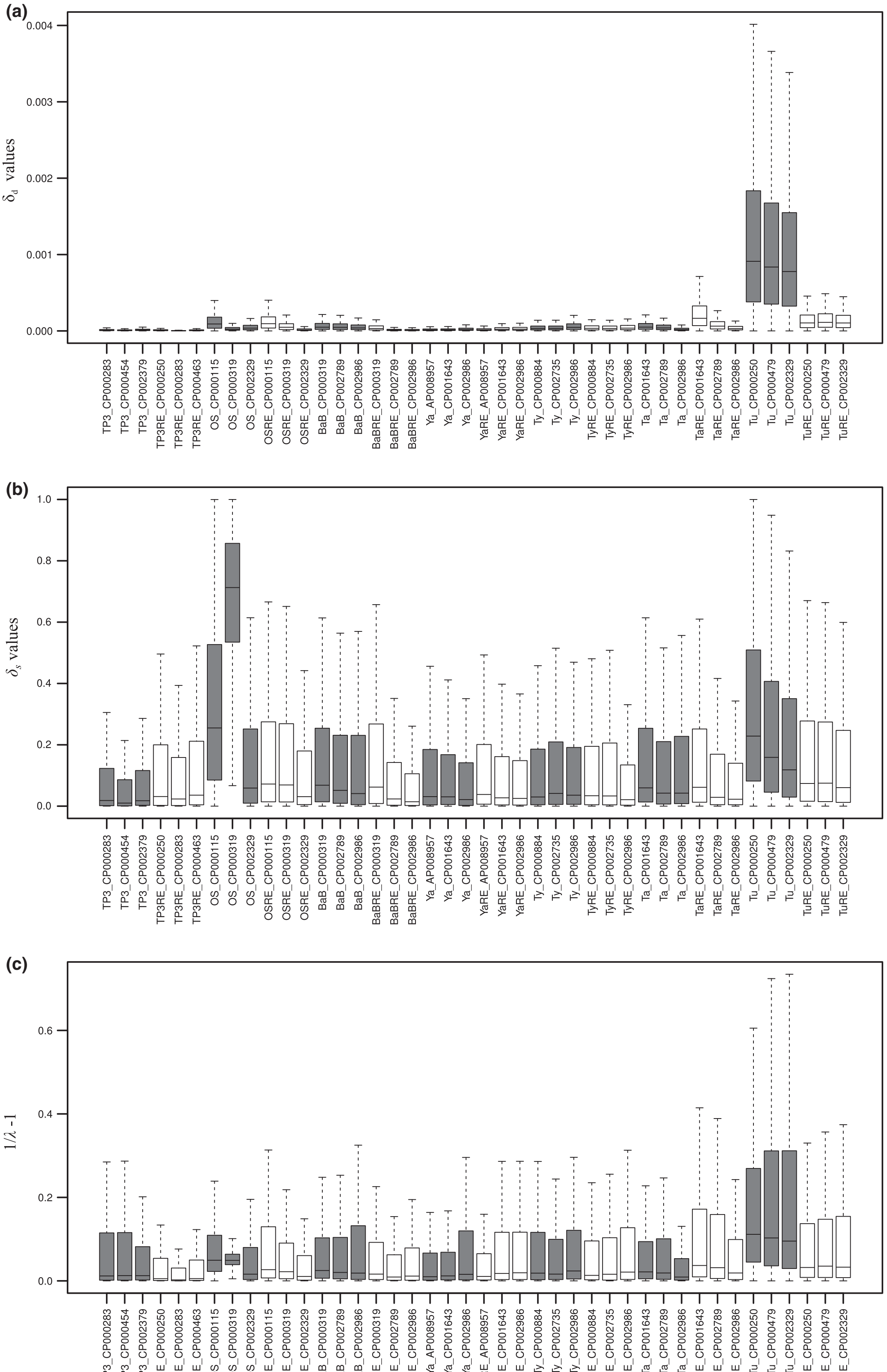
(a)
δ_d values
0.004
0.003
0.002
0.001
0.000
TP3_CP000283
TP3_CP000454
TP3_CP002379
TP3RE_CP000250
TP3RE_CP000283
TP3RE_CP000463
OS_CP000115
OS_CP000319
OS_CP002329
OSRE_CP000115
OSRE_CP000319
OSRE_CP002329
BaB_CP000319
BaB_CP002789
BaB_CP002986
BaBRE_CP000319
BaBRE_CP002789
BaBRE_CP002986
Ya_AP008957
Ya_CP001643
Ya_CP002986
YaRE_AP008957
YaRE_CP001643
YaRE_CP002986
Ty_CP000884
Ty_CP002735
Ty_CP002986
TyRE_CP000884
TyRE_CP002735
TyRE_CP002986
Ta_CP001643
Ta_CP002789
Ta_CP002986
TaRE_CP001643
TaRE_CP002789
TaRE_CP002986
Tu_CP000250
Tu_CP000479
Tu_CP002329
TuRE_CP000250
TuRE_CP000479
TuRE_CP002329
(b)
δ_s values
1.0
0.8
0.6
0.4
0.2
0.0
TP3_CP000283
TP3_CP000454
TP3_CP002379
TP3RE_CP000250
TP3RE_CP000283
TP3RE_CP000463
OS_CP000115
OS_CP000319
OS_CP002329
OSRE_CP000115
OSRE_CP000319
OSRE_CP002329
BaB_CP000319
BaB_CP002789
BaB_CP002986
BaBRE_CP000319
BaBRE_CP002789
BaBRE_CP002986
Ya_AP008957
Ya_CP001643
Ya_CP002986
YaRE_AP008957
YaRE_CP001643
YaRE_CP002986
Ty_CP000884
Ty_CP002735
Ty_CP002986
TyRE_CP000884
TyRE_CP002735
TyRE_CP002986
Ta_CP001643
Ta_CP002789
Ta_CP002986
TaRE_CP001643
TaRE_CP002789
TaRE_CP002986
Tu_CP000250
Tu_CP000479
Tu_CP002329
TuRE_CP000250
TuRE_CP000479
TuRE_CP002329
(c)
$1/\lambda$ -1
0.6
0.4
0.2
0.0
TP3_CP000283
TP3_CP000454
TP3_CP002379
TP3RE_CP000250
TP3RE_CP000283
TP3RE_CP000463
OS_CP000115
OS_CP000319
OS_CP002329
OSRE_CP000115
OSRE_CP000319
OSRE_CP002329
BaB_CP000319
BaB_CP002789
BaB_CP002986
BaBRE_CP000319
BaBRE_CP002789
BaBRE_CP002986
Ya_AP008957
Ya_CP001643
Ya_CP002986
YaRE_AP008957
YaRE_CP001643
YaRE_CP002986
Ty_CP000884
Ty_CP002735
Ty_CP002986
TyRE_CP000884
TyRE_CP002735
TyRE_CP002986
Ta_CP001643
Ta_CP002789
Ta_CP002986
TaRE_CP001643
TaRE_CP002789
TaRE_CP002986
Tu_CP000250
Tu_CP000479
Tu_CP002329
TuRE_CP000250
TuRE_CP000479
TuRE_CP002329

Fig. 9 DNA damage in selected microbial species. The boxplot shows the simulated posterior distribution of the three DNA damage model parameters δ_d, δ_s and λ. (a) δ_d, deamination in double strand. (b) δ_s, deamination in single strand. (c) λ probability of reads not terminating in overhangs, converted in a size estimate of average overhang length (in bp) using the following formula: $1/\lambda - 1$. In grey, first fractions; in white, second fractions. Each sample is labelled with the bacteria NCBI number against which shotgun reads were mapped: AP008957: *Rhodococcus erythropolis PR4*; CP000115: *Nitrobacter winogradskyi Nb-255*; CP000250: *Rhodopseudomonas palustris HaA2*; CP000283: *Rhodopseudomonas palustris BisB5*; CP000319: *Nitrobacter hamburgensis X14*; CP000454: *Arthrobacter* sp. *FB24*; CP000463: *Rhodopseudomonas palustris BisA53*; CP000479: *Mycobacterium avium* 104; CP000884*: Delftia acidovorans SPH-1*; CP001643: *Brachybacterium faecium DSM 4810* strain 6–10; CP002329: *Mycobacterium* sp. *JDM601*; CP002379: *Arthrobacter phenanthrenivorans Sphe3*; CP002735: *Delftia* sp. *Cs1–4*; CP002789: *Xanthomonas campestris pv. raphani 756C*; CP002986: *Stenotrophomonas maltophilia JV3*.

endogenous DNA in second fractions. The oldest sample (TP3/TP3RE) additionally showed higher levels of DNA damage in the first fraction compared to the second fraction, but no difference in microbial community composition. This suggests that the presence of specific microbial taxa is not the main driver of the different amounts of damage in this specimen. Overall, our results support the contention that the second fraction is enriched in molecular preservation niches, which limit the penetrability of the sample to microorganisms (and water).

The existence of molecular preservation niches was hypothesized previously because comparison of DNA extraction fractions suggested that endogenous DNA fragments are smaller (Schwarz *et al.* 2009), more damaged and less concentrated (Orlando *et al.* 2011) in first than in second fractions. Building on these results, a two-step DNA extraction method was applied to a 700 000-year-old horse specimen, the full genome of which was successfully sequenced (Orlando *et al.* 2013). Also in line with our findings, Schuenemann *et al.* (2011) recently observed that EDTA supernatants from ancient bone/tooth extracts are enriched in the pathogen *Yersinia pestis* DNA compared to pellets. In this study, the second extraction provided a 1.6- to 5.5-fold enrichment in endogenous horse DNA content relative to environmental DNA, demonstrating the utility of this approach for sequencing ancient genomes.

Microbial profiles were found to cluster into three main groups that reflect the microenvironmental conditions surrounding the remains during deposition. One cluster consisted of the extracts from sample TP, which was excavated thousands of kilometres away from Yakutia, the source region for all other samples. The second cluster included samples excavated from burial sites and was distinct from sample Ya that was found in a pit where pieces of animals were left decomposing. This result supports the absence of significant postexcavation contamination of the samples, which would most likely have masked any microbe-driven structure among samples. Such a pattern is impossible to reconcile with the presence of significant levels of bacterial contamination.

We identify a near absence of DNA damage patterns in DNA from the most abundant microbial taxa recovered from the ancient horse samples, including from those samples with substantial amounts of damaged endogenous DNA. This suggests that microbial colonization of the remains may have occurred recently. Interestingly, this includes microbial sequences that were recovered from the innermost structures of bones and teeth. Our result is in agreement with previous results showing the absence of DNA damage in *Actinobacteria* recovered from Neandertal bones (Zaremba-Niedźwiedzka & Andersson 2013). However, some caution is required as levels of DNA damage for medieval samples have been found to be lower in *Mycobacterium leprae* sequences than in human sequences generated from the same medieval samples (Schuenemann *et al.* 2013). Mycolic acid that constitutes the cell wall of *Mycobacterium* organisms has been proposed to protect *Mycobacterium* DNA from degradation, thus explaining the particular pattern of DNA preservation observed in *Mycobacterium leprae* DNA (Schuenemann *et al.* 2013). Further studies exploring differences in DNA preservation among microbial taxa are thus needed to fully understand the different levels of DNA damage in microbes.

We did not identify in the microbial profiles obtained from the seven ancient horse extracts any typical gut bacteria involved in *postmortem* decomposition. During the early stages of the decomposition processes, macromolecules are degraded by (facultative) gut anaerobes, mainly belonging to the *Lactobacillaceae* and *Bacteroidaceae* families (Mondor *et al.* 2012; Metcalf *et al.* 2013; Hyde *et al.* 2013). After the rupture of the abdominal cavity, gut bacteria progress throughout the body via the blood vessels, and subsequent exposition to oxygen (facultative) aerobes, such as members of the *Phyllobacteriaceae*, *Hyphomicrobiaceae*, *Brucellaceae* and *Enterobacteriaceae* families, partake to the decomposition of cadavers (Hyde *et al.* 2013; Metcalf *et al.* 2013). Instead, the microbial taxa identified in the ancient horse specimens are typical of soil environments, suggesting an environmental origin of the microbial DNA. The taxonomic differences between the recovered microbial communities likely reflect ecological differences among depositional environments.

These results support a scenario in which the endogenous microbiome associated with postmortem decay either does not penetrate within bones and teeth, or is diluted by environmental microbes in later stages of the taphonomic process. Our results are also consistent with observations that the distribution of microbial taxa tends to become more similar across sites in the latest phases of the decomposition process, due to colonization by microorganisms of the depositional environment (Metcalf *et al.* 2013). In particular, no particular microbial taxa were found to be systematically specific to either bones or teeth, although no general conclusion can be drawn until a broader range of ancient samples and depositional conditions are investigated.

Our methodology is based on METAPHLAN and relies on a similarity-based classification of metagenomic reads by comparison with preidentified and curated clade-specific reference marker sequences (Segata *et al.* 2013), rather than on global similarity searches performed against a full catalogue of reference genomes (i.e. using BLASTN; Altschul *et al.* 1997). The specificity of the reference markers limits bias related to evolutionary uninformative conserved regions and horizontal gene transfer. Additionally, the exclusion of multicopy genes (e.g. 16S rDNA), whose numbers vary across species, also reduces noise in estimating relative abundance. Other factors will, however, limit the sensitivity of our approach. First, despite the large number of shotgun sequence reads generated from each extract (7.3–32.0 millions of reads), we have probably undersampled the microbial diversity present in each sample, which would limit our capacity to identify subtle differences among low-abundance taxa. Second, only sequences matching database markers can be classified, precluding the identification of the fraction of microbes currently uncharacterized at the genome level. We anticipate that the plethora of microbial genome projects currently ongoing such as the Human Microbiome Project (http://www.hmpdacc.org/), my.microbes (http://my.microbes.eu/index.cgi), the Earth Microbiome Project (http://www.earthmicrobiome.org/) and efforts from the International Human Microbiome Consortium (http://www.human-microbiome.org/) will enable expanding the taxonomic coverage represented in the METAPHLAN database and thereby will limit the extent of such problems.

A number of ancient remains, such as ancient dental calculus (Adler *et al.* 2013) and coprolites (Tito *et al.* 2012), have been profiled for their 16S rDNA sequence diversity, revealing striking changes in the human diet in phase with major technological innovations and society transformations. Here, we have presented a framework for achieving comparable precision in describing the structure of microbial communities using shotgun sequence data sets, as shown previously by the analysis of soils (Fierer *et al.* 2012). This result demonstrates that low-depth shotgun sequencing of ancient DNA extracts can not only provide a range of information about the ancient individual of interest (Skoglund *et al.* 2012) and the level of DNA damage (Jónsson *et al.* 2013), but also from its environment and potentially, its diet.

The methodology and results presented here are of particular relevance to the study of ancient pathogens, as a number of pathogenic bacterial species belong to the same genera as environmental species, thus increasing significantly the risk of exogenous contamination (Gilbert *et al.* 2004; Bouwman *et al.* 2012). In particular, in the samples studied here, 4.8–40.4% of the microbial diversity was represented by the single genus *Mycobacterium*, which comprises the human pathogens responsible for tuberculosis and leprosy. A two-step DNA extraction, and/or damage analysis of the microbial derived sequence reads, furthermore, could be adopted as a valuable resource for assessing the authenticity of ancient pathogen DNA. Sample controls from the soil of the excavation site could be also sequenced to control for possible environmental contamination. When studying organisms genetically distinct from environmental microorganisms, a targeted capture approach prior to sequencing increases the relative amount of endogenous DNA of interest compared to environment-derived microbial DNA (e.g. the Yersinia pestis genome by Bos *et al.* 2011).

To conclude, our framework improves the amount of information that can be gathered from the diversity of molecules preserved in fossils. Its application to a broad range of shotgun sequence data sets produced across a variety of archaeological samples and environments promises to advance our understanding of how microbial taxonomic profiles could affect the quality of endogenous DNA, both in the context of depositional environments and museum collections. Such knowledge could be ultimately used to develop new methods tailored to the extraction and analysis of endogenous DNA templates.

Acknowledgements

We thank Tina Brand and Pernille Olsen, the laboratory technicians at the Danish High-throughput DNA Sequencing Centre for technical assistance; Aurélien Ginolhac and Mikkel Schubert for support with read mapping; Ruth Fernandez and Julia Vilstrup for assistance and fruitful discussions. This work was supported by the French Archaeological Mission in Oriental Siberia (Ministry of Foreign and European Affairs, France), the North-Eastern Federal University (Yakutsk, Sakha Republic), the Human Adaptation program of the French Polar Institute Paul Emile Victor, the Danish National Research Foundation

(DNRF94), Marie-Curie Intra-European Fellowship IEF 302617, the Marie-Curie Career Integration Grant CIG-293845, a grant from the Geological Survey of Denmark and Greenland (GEUS), a Danish National Research Foundation FNU grant attributed to Ludovic Orlando.

References

Adler CJ, Dobney K, Weyrich LS *et al.* (2013) Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics*, **45**, 450–455.

Allentoft ME, Collins M, Harker D *et al.* (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings Biological Sciences*, **279**, 4724–4733.

Altschul SF, Madden TL, Schäffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Andrews RM, Kubacka I, Chinnery PF *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, **23**, 147.

Bell LS, Skinner MF, Jones SJ (1996) The speed of post-mortem change to the human skeleton and its taphonomic significance. *Forensic Science International*, **82**, 129–149.

Bos KI, Schuenemann VJ, Golding GB *et al.* (2011) A draft genome of Yersinia pestis from victims of the Black Death. *Nature*, **478**, 506–510.

Bouwman AS, Kennedy SL, Müller R *et al.* (2012) Genotype of a historic strain of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 18511–18516.

Briggs AW, Good JM, Green RE *et al.* (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, **325**, 318–321.

Burbano HA, Hodges E, Green RE *et al.* (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*, **328**, 723–725.

Campos PF, Craig OE, Turner-Walker G *et al.* (2012) DNA in ancient bone—where is it located and how should we extract it? *Annals of Anatomy*, **194**, 7–16.

Caporaso JG, Bittinger K, Bushman FD *et al.* (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, **26**, 266–267.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010b) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Carpenter ML, Buenrostro JD, Valdiosera C *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**, 852–864.

Conn HJ (1928) A type of bacteria abundant of productive soils, but apparently lacking in certain soils of low productivity. *New York State Agricultural Experimental Station Technical Bulletin No*, **138**, 3–26.

Dabney J, Knapp M, Glocke I *et al.* (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 15758–15763.

DeSantis TZ, Hugenholtz P, Larsen N *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environment Microbiology*, **72**, 5069–5072.

Fierer N, Leff JW, Adams BJ *et al.* (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 21390–21395.

Geigl E-M (2002) On the circumstances surrounding the preservation and analysis of very old DNA. *Archaeometry*, **44**, 337–342.

Gilbert MT, Cuccui J, White W *et al.* (2004) Absence of *Yersinia pestis*-specific DNA in human teeth from five European excavations of putative plague victims. *Microbiology*, **150**, 341–354.

Gilbert MTP, Rudbeck L, Willerslev E *et al.* (2005) Biochemical and physical correlates of DNA contamination in archaeological human bones and teeth excavated at Matera, Italy. *Journal of Archaeological Science*, **32**, 785–793.

Ginolhac A, Rasmussen M, Gilbert MT *et al.* (2011) mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, **27**, 2153–2155.

Ginolhac A, Vilstrup J, Stenderup J *et al.* (2012) Improving the performance of true single molecule sequencing for ancient DNA. *BMC Genomics*, **13**, 177.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Haile J, Holdaway R, Oliver K *et al.* (2007) Ancient DNA chronology within sediment deposits: are paleobiological reconstructions possible and is DNA leaching a factor? *Molecular Biology and Evolution*, **24**, 982–989.

Ho SY, Phillips MJ, Cooper A *et al.* (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution*, **22**, 1561–1568.

Höss M, Jaruga P, Zastawny TH *et al.* (1996) DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Research*, **24**, 1304–1307.

Hyde ER, Haarmann DP, Lynne AM *et al.* (2013) The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. *PLoS ONE*, **8**, e77733.

Jans M, Nielsen-Marsh C, Smith C *et al.* (2004) Characterisation of microbial attack on archaeological bone. *Journal of Archaeological Science*, **31**, 87–95.

Janssen PH (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology*, **72**, 1719–1728.

Jónsson H, Ginolhac A, Schubert M *et al.* (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, **29**, 1682–1684.

Keller A, Graefen A, Ball M *et al.* (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications*, **3**, 698.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

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

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.

Lindgreen S (2012) AdapterRemoval: easy cleaning of next generation sequencing reads. *BMC Research Notes*, **5**, 337.

Malmström H, Stora J, Dalen L *et al.* (2005) Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Molecular Biology and Evolution*, **22**, 2040–2047.

Martin MD, Cappellini E, Samaniego JA *et al.* (2013) Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. *Nature Communications*, **4**, 2172.

Metcalf JL, Parfrey LW, Gonzalez A *et al.* (2013) A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *eLife*, **2**, e01104.

Meyer M, Kircher M, Gansauge MT *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science*, **338**, 222–226.

Meyer LR, Zweig AS, Hinrichs AS *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Research*, **41**, D64–D69.

Miller W, Drautz DI, Ratan A *et al.* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456**, 387–390.

Mondor EB, Tremblay MN, Tomberlin JK *et al.* (2012) The ecology of carrion decomposition. *Nature Education Knowledge*, **3**, 21.

Nielsen-Marsh CM, Hedges REM (1999) Bone porosity and the use of mercury intrusion porosimetry in bone diagenesis studies. *Archaeometry*, **41**, 165–174.

Oota H, Saitou N, Matsushita T *et al.* (1995) A genetic study of 2000 year old human remains from Japan using mitochondrial DNA sequences. *American Journal of Physical Anthropology*, **98**, 133–145.

Orlando L, Metcalf JL, Alberdi MT *et al.* (2009) Revising the recent evolutionary history of equids using ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 21754–21759.

Orlando L, Ginolhac A, Raghavan M *et al.* (2011) True single-molecule DNA sequencing of a Pleistocene horse bone. *Genome Research*, **21**, 1705–1719.

Orlando L, Ginolhac A, Zhang G *et al.* (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, **499**, 74–78.

Pääbo S, Poinar H, Serre D *et al.* (2004) Genetic analyses from ancient DNA. *Annual Review of Genetics*, **38**, 645–679.

Pedersen JS, Valen E, Velazquez AMV *et al.* (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Research*, gr-163592. DOI: 10.1101/gr.163592.113.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/.

Rasmussen M, Li Y, Lindgreen S *et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, **463**, 757–762.

Reich D, Green RE, Kircher M *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**, 1053–1060.

Rohland N, Hofreiter M (2007) Ancient DNA extraction from bones and teeth. *Nature Protocols*, **2**, 1756–1762.

Rohland N, Siedel H, Hofreiter M (2004) Nondestructive DNA extraction method for mitochondrial DNA analyses of museum specimens. *BioTechniques*, **36**, 814–816, 818–821.

Salamon M, Tuross N, Arensburg B *et al.* (2005) Relatively well preserved DNA is present in the crystal aggregates of fossil bones. *Proceedings of the National Academy of Science United States of America*, **102**, 13783–13788.

Sawyer S, Krause J, Guschanski K *et al.* (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, **7**, e34131.

Schubert M, Ginolhac A, Lindgreen S *et al.* (2012) Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, **13**, 178.

Schubert M, Ermini L, Der Sarkissian C *et al.* (2014) Characterization of ancient and modern genomes by sequencing, SNP detection, phylogenomic and metagenomic analysis. *Nature Protocols*. In press.

Schuenemann V, Bos K, DeWitte S (2011) Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, E746–E752.

Schuenemann VJ, Singh P, Mendum TA *et al.* (2013) Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science*, **341**, 179–183.

Schwarz C, Debruyne R, Kuch M *et al.* (2009) New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Research*, **37**, 3215–3229.

Segata N, Izard J, Waldron L *et al.* (2011) Metagenomic biomarker discovery and explanation. *Genome Biology*, **12**, R60.

Segata N, Waldron L, Ballarini A *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, **9**, 811–814.

Segata N, Boernigen D, Tickle TL *et al.* (2013) Computational meta'omics for microbial community studies. *Molecular Systems Biology*, **9**, 666.

Seguin-Orlando A, Schubert M, Clary J *et al.* (2013) Ligation bias in Illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PLoS ONE*, **8**, e78575.

Skoglund P, Malmström H, Raghavan M *et al.* (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, **336**, 466–469.

Smith CI, Chamberlain AT, Riley MS *et al.* (2003) The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of Human Evolution*, **45**, 203–217.

Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.

Thèves C, Senescau A, Vanin S *et al.* (2011) Molecular identification of bacteria by total sequence screening: determining the cause of death in ancient human subjects. *PLoS ONE*, **6**, e21733.

Tito RY, Knights D, Metcalf J *et al.* (2012) Insights from characterizing extinct human gut microbiomes. *PLoS ONE*, **7**, e51146.

Wade CM, Giulotto E, Sigurdsson S *et al.* (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, **326**, 865–867.

Wang Q, Garrity GM, Tiedje JM *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Xu X, Arnason U (1994) The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene*, **148**, 357–362.

Yoshida K, Schuenemann VJ, Cano LM *et al.* (2013) The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife*, **2**, e00731.

Zaremba-Niedźwiedzka K, Andersson SGE (2013) No ancient DNA damage in *Actinobacteria* from the Neanderthal bone. *PLoS ONE*, **8**, e62799.

L.O., C.D. and L.E. designed research, L.O. generated ancient DNA data, L.O., C.D., L.E. and H.J. analysed data, L.O., C.D. and L.E. wrote the manuscript, B.S. revised the manuscript, B.S., A.N.A. and E.C. provided samples and contextual information.

Data accessibility

The DNA sequence data generated in this study can be accessed from the SRA (Accession nos.: SRR1030038–SRR1030052). BAM files containing the alignments of shotgun DNA reads to the horse and human reference genome sequences, as well the output files of the METAPHLAN and MAPDAMAGE 2.0 programs, can be accessed from the Dryad digital repository (http://doi.org/10.5061/dryad.r0056).

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Principal coordinate analysis (A) and hierarchical clustering (B) of microbial genus relative abundance unfiltered for low-abundance taxa from ancient horse shotgun data sets.

Fig. S2 Principal component analysis of microbial genus relative abundance for ancient horse shotgun data sets.

Fig. S3 Principal coordinate analysis of relative abundance for ancient horse shotgun data sets at all taxonomic levels: (A) phylum, (B) class, (C) order, (D) family, (E) genus, Fspecies.

Fig. S4 Principal coordinate analysis of dimensions 3 and 4 for microbial genus relative abundance in ancient horse shotgun data sets.

Fig. S5 Percentages of endogenous (A) and human nonconserved DNA (B) in ancient horse extracts.

Fig. S6 Length distribution of DNA unique to horse in first (grey) and second fraction (black).

Fig. S7 Principal coordinate analysis of microbial genus relative abundance for soil and ancient horse shotgun data sets.

Fig. S8 Principal coordinate analysis of microbial genus relative abundance in subsampled ($N = 796$ mapped reads) ancient horse shotgun data sets.

Fig. S9 Relative abundance of microbial classes from ancient horse shotgun data sets, as estimated on the basis of shotgun data (METAPHLAN database) and 16S rDNA ('_16S', Greengenes database).

Fig. S10 Principal coordinate analysis of microbial family relative abundance from ancient horse shotgun data sets.

Table S1 Relative abundances of micro-organisms in ancient horse extracts.

Table S2 Characteristics of microbial diversity at all taxonomic levels (A-number of identified taxa, B- Shannon diversity index, C-cumulative percentage of low-abundance taxa <1%).

Table S3 Illumina sequencing of ancient horse DNA extracts and mapping metrics for nonconserved regions.

Table S4 Mapping coverage obtained for the alignments to microbial genomes used in microbial DNA damage analyses.