Zonkey : a simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages

Mikkel Schubert, Marjan Mashkour, Charleen Gaunitz, Antoine Fages, Andaine Seguin-Orlando, Shiva Sheikhi, Ahmed Alfarhan, Saleh Alquraishi, Khaled A.S. Al-Rasheid, Richard Chuang, et al.

Journal of Archaeological Science xxx (2017) xxx-xxx

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: www.elsevier.com

Zonkey: A simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages

Mikkel Schubert [a], Marjan Mashkour [b], Charleen Gaunitz [a], Antoine Fages [a, c], Andaine Seguin-Orlando [a, d], Shiva Sheikhi [b], Ahmed H. Alfarhan [e], Saleh A. Alquraishi [e], Khaled A.S. Al-Rasheid [e], Richard Chuang [f], Luca Ermini [a], Cristina Gamba [a], Jaco Weinstock [f], Onar Vedat [g], Ludovic Orlando [a, c, *]

[a] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350, Copenhagen K, Denmark
[b] Archéozoologie, Archéobotanique (UMR 7209), CNRS, MNHN, UPMC, Sorbonne Universités, France
[c] Université de Toulouse, University Paul Sabatier (UPS), Laboratoire AMIS, CNRS UMR 5288, Toulouse, France
[d] Danish National High-Throughput DNA Sequencing Center, Natural History Museum of Denmark, University of Copenhagen, Øster Farimagsgade 2D, 1353K, Copenhagen, Denmark
[e] Zoology Department, College of Science, King Saud University, Riyadh, 11451, Saudi Arabia
[f] Faculty of Humanities, University of Southampton, Avenue Campus, Highfield, Southampton, SO17 1BF, United Kingdom
[g] Anatomy Department, Istanbul University Faculty of Veterinary Medicine, Dalı, 34320, Avcılar, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 18 August 2016
Received in revised form 8 November 2016
Accepted 15 December 2016
Available online xxx

Keywords:
Ancient DNA
High-throughput sequencing
Equid
F1-hybrid
Mule
Hinny

ABSTRACT

Horses, asses and zebras, can produce first-generation F1-hybrids, despite their striking karyotypic and phenotypic differences. Such F1-hybrids are mostly infertile, but often present characters of considerable interest to breeders. They were extremely valued in antiquity, and commonly represented in art and on coinage. However, hybrids appear relatively rarely in archaeological faunal assemblages, mostly because identification based on morphometric data alone is extremely difficult. Here, we developed a methodological framework that exploits high-throughput sequencing data retrieved from archaeological material to identify F1-equine hybrids. Our computational methodology is distributed in the open-source Zonkey pipeline, now part of PALEOMIX (https://github.com/MikkelSchubert/paleomix), together with full documentation and examples. Using both synthetic and real sequence datasets, from living and ancient F1-hybrids, we find that Zonkey shows high sensitivity and specificity, even with limited sequencing efforts. Zonkey is thus well suited to the identification of equine F1-hybrids in the archaeological record, even in cases where DNA preservation is limited. Zonkey can also help determine the sex of ancient animals, and allows species identification, which advantageously complements morphological data in cases where material is fragmentary and/or multiple candidate equine species coexisted in sympatry.

Abbreviations

HTS	High-throughput DNA Sequencing
PCR	Polymerase Chain Reaction
PCA	Principal Component Analysis
aDNA	Ancient DNA
mtDNA	mitochondrial DNA

1. Introduction

Historical sources, both written and pictorial, leave no doubts that the cross-breeding of mammal species – equids in particular – has been practised for thousands of years (Clutton-Brock, 1999). One example is provided by the *kunga*, mentioned in Syro-Mesopotamian documents from the mid- and late 3rd millennium [mill.] BCE, and generally interpreted as a hybrid between a hemione (*Equus hemionus*) and a domestic donkey (*Equus asinus*) (Postgate, 1986). Apparently used to pull chariots associated with messengers, soldiers and officials, it is considered to be represented in the Standard of Ur as well as ceilings from Tell Brak and Tell Beydar (Weber, 2008). Additionally, mules – the offspring of a jack (*Equus asinus*) and a mare (*Equus caballus*) – were often present in Mesopotamian art of the first mill. BCE (Clutton-Brock, 1999) and were essential to Roman society (Johnstone, 2008).

Hybrids between closely related species were greatly valued for their morphological and behavioural traits, thanks to hybrid vigour. Mules are more sure-footed than horses, thrive on cheaper food, have stronger working capacities, longer life spans, are more resistant to disease, and can carry more weight than horses (Tegetmeier et al., 1895). Therefore, mules became essential in trade, commerce, transport and military; this is true for ancient times (Armitage and Chapman, 1979; Peddie, 1987; Roth, 1999) and for more recent times as well, up until the early 20th century (Campbell Smith, 2008; Tegetmeier et al., 1895) when hundreds of thousand mules were used by the British Army in World War I (Singleton, 1993).

* Corresponding author. Centre for GeoGenetics, Natural History Museum of Denmark, Øster Voldgade 5-7, 1350K, Copenhagen, Denmark.
Email address: Lorlando@snm.ku.dk (L. Orlando)

http://dx.doi.org/10.1016/j.jas.2016.12.005
0305-4403/

As mules are generally sterile (although exceptions occur (Steiner and Ryder, 2013)), their breeding requires expert knowledge and considerable financial investment, as described for mules in Columella's *De Re Rustica* (VI.3.6), and Varro's *De Rustica* (II.8). Accordingly, hybrids were invaluable commodities and generally commanded much higher prices than purebred equids, both in antiquity and early modern times (Konrad, 1980; Laurence, 1999). Estimating the prevalence of hybrids in archaeological assemblages would thus reveal important facets of ancient societies, specifically regarding transport, trade and economy.

The importance attached to hybrid equids in ancient times strongly contrasts with their relatively limited appearance in archaeological faunal assemblages (Johnstone, 2008). This apparent disparity largely lies in the difficulty of attaining unambiguous taxonomic identification on the sole basis of morphology and/or morphometry (Chuang, 2016). The following non-mutually exclusive factors make equine hybrid identification particularly difficult: *(i)* the great morphological similarity between bones of the parental species (Peters, 1998); *(ii)* their co-occurrence in particular regions, like Southwest Asia (Twiss et al., 2016), where four equine species (the donkey, the hemione, the horse and the recently extinct hydruntine, *Equus hydruntinus*) co-existed until very recently (Eisenmann and Mashkour, 1999; Mashkour, 2002, 2003; Vila, 2006; Orlando et al., 2006); *(iii)* the great morphological variation within domestic horses; and *(iv)* our limited knowledge of the hybrid morphological space, due to the scarcity of modern reference material (Baxter, 1998; Chuang, 2016; Johnstone, 2004). While some 'diagnostic' morphological traits have been postulated in different equine species, including mules (Davis, 1980; Eisenmann, 1986; Peters, 1998; Uerpmann and Uerpmann, 1994), these are not unanimously considered as valid (Baxter, 1998; Chuang, 2016; Groves and Willoughby, 1981; Twiss et al., 2016). Finally, the equid remains commonly recovered from archaeological sites are fragmentary, thus reducing the number of diagnostic traits available for taxonomic identification (Baxter, 1998; Zeder, 1986).

In contrast, first generation hybrids, so-called F_1-hybrids, can easily be identified based on genetic information, since each parental species provides one set of chromosomes. The maternally transmitted mitochondrial DNA (mtDNA) can help identify the maternal species. Genetic information could therefore reveal the proportion of mules and hinnies (the offspring of stallions and jennets) in archaeological assemblages. Recent advances in ancient DNA (aDNA) research and high-throughput DNA sequencing (HTS) have made the retrieval of genome-scale data from minute amounts of archaeological material cost-effective and almost routine (Orlando et al., 2015). In this study, we developed Zonkey, a user-friendly and open-source pipeline that exploits low-depth HTS data from archaeological material to identify F_1-equine hybrids. Using simulations, we demonstrate that Zonkey shows extremely high specificity from as few as 1000 sequences mapped to the horse reference genome. Applying Zonkey to 18 archaeological specimens spanning the last ~6000–8000 years, we identify seven mules from Roman and Byzantine assemblages where morphological analyses were inconclusive. Zonkey works on Linux and is freely available online at https://github.com/MikkelSchubert/paleomix as part of the PALEOMIX pipeline (Schubert et al., 2014).

1.1. Computational analyses

1.1.1. Reference panel

The reference panel consists of nine equine genomes, using alignments published in (Orlando et al., 2013), (Jónsson et al., 2014) and (Der Sarkissian et al., 2015). These include the complete genomes of two caballine individuals (a Przewalski's horse, *Equus przewalski*, and a Franches-Montagnes horse, *Equus caballus*) as well as seven non-caballine individuals: a domestic donkey (*Equus africanus asinus*), an African wild ass (*Equus africanus somaliensis*), a Grant's zebra (*Equus quagga boehmi*), a Grevyi's zebra (*Equus grevyi*), a Hartmann's mountain zebra (*Equus zebra hartmannae*), an onager (*Equus hemionus onager*), and a Tibetan kiang (*Equus kiang*). Samples were genotyped using PALEOMIX (Schubert et al., 2014) as described in (Jónsson et al. (2014)), and bi-allelic sites called in all samples were collected, for a total of ~36.5 million autosomal sites, representing ~15,000 sites per Mb. We further constructed a multiple mtDNA alignment of all species included in the reference panel, using a selection of complete mitochondrial sequences made available for the same species by (Orlando et al., 2013) (Vilstrup et al., 2013), and (Der Sarkissian et al., 2015) (GenBank Accession Numbers: JX312719, JX312722, JX312730, JX312732, KM881680, KM881681, KT368746.1, KX669267, and KX669268). The repetitive region covering positions 16,129 to 16,371 of the horse mitochondrial genome (NC_001640.1) was masked.

1.1.2. Hybridization report

Zonkey automates a complete suite of analyses aimed at evaluating whether the (low-depth) sequence data generated from an ancient equine specimen belongs to a F1-hybrid (Fig. 1). The entire set of analyses requires BAM alignments against both the mitochondrial and the nuclear genomes, but partial sets of analyses can be performed if only one such alignment is available.

For each individual BAM file aligned against the *Equus caballus* nuclear reference genome (EquCab2), reads overlapping sites included in the reference panel are located, and a single nucleotide is sampled at each site in order to generate a pseudo-haploid sequence. Sites falling outside the variation represented in the reference panel are excluded. Two SNP panels, including or excluding transitions, are generated per sample (excluding transitions aims at reducing the impact on downstream analyses of post-mortem DNA damage, which mainly consist of C→T and G→A transitions (Briggs et al., 2007)). The resulting tables are processed using PLINK (Chang et al., 2015), to generate the intermediate files required for downstream analyses, carried out on both panels of SNPs. These include Principal Component Analyses (PCAs), the profiling of main ancestry components and phylogenetic reconstructions.

PCAs are carried out using EIGENSOFT 'SmartPCA' (Price et al., 2006, Patterson et al., 2006). The main estimation of ancestry components is performed using ADMIXTURE (Alexander et al., 2009) and a partially supervised approach in which reference samples are assigned to either one of two (caballine and non-caballine equids) or three (asses, horses, and zebras) groups. The input sample is left unassigned and the ancestry proportions of each supervised group present in the sample are thereby estimated. Phylogenetic reconstructions are performed using TreeMix (Pickrell and Pritchard, 2012), assuming either zero or one migration edges. Coverage statistics of autosomes and the X-chromosome are provided to enable an approximate estimation of the sex (males and females are expected to show X-to autosomal coverage ratios of ~0.5 and ~1.0, respectively). By default, these analyses are restricted to 1,000,000 randomly selected reads.

Mitochondrial phylogenies are constructed from a BAM alignment against one of the mitochondrial sequences included in the reference panel, or against the horse mtDNA sequence (GenBank Accession Number NC_001640.1). A consensus sequence for the sample is constructed by selecting the majority base at each position in the alignment, masking positions with no majority base. The consensus sequence is added into the multiple-sequence alignment by mapping each position of the target sequence to the corresponding base in the reference panel. A maximum likelihood phylogeny is inferred us-

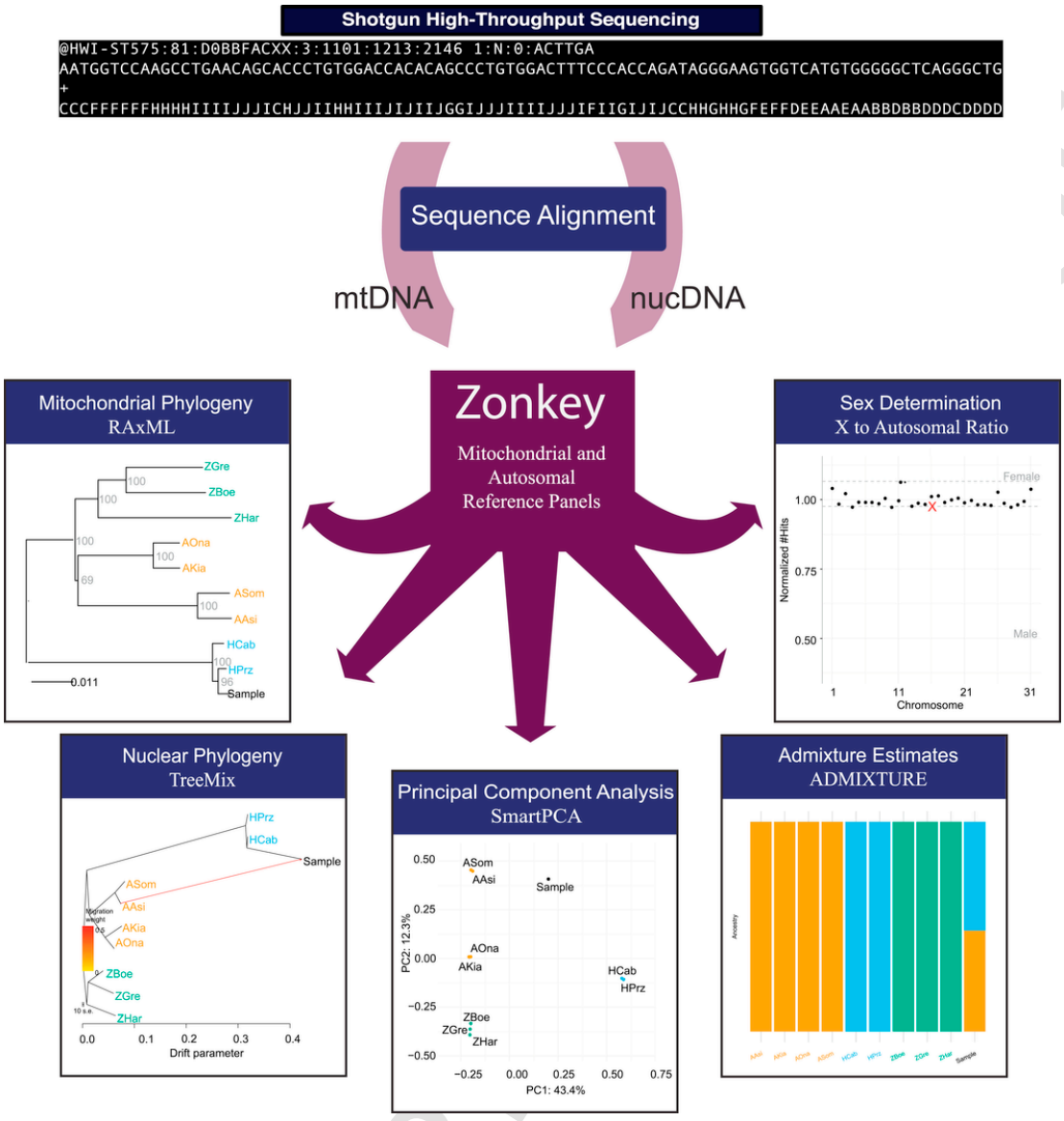

Fig. 1. Overview of the Zonkey pipeline. Zonkey was developed to identify equine F1-hybrids from limited amounts of shotgun sequencing data generated on Illumina HTS platforms. Mitochondrial information provides a taxonomic identification of the maternal lineage, while TreeMix, PCA and ADMIXTURE analyses based on autosomal SNPs help assess the taxonomy of both parents. Normalized X-to autosome coverage ratios allow sex identification.

ing RAxML (Stamatakis, 2006) and visualized using R packages 'Ape' (Paradis et al., 2004) and 'GGPlot2' (Wickham, 2009).

1.1.3. ADMIXTURE significance

The significance of ADMIXTURE results is estimated from an empirical distribution of admixture percentages calculated using synthetic hybrids generated using a panel of 13 equine genomes, representing four asses, six horses, and three zebras (Sequence Read Archive: SRS431817, SRS438157, SRS441443, SRS439179, SRS431663; European Nucleotide Archive: SAMN03010637, SAMEA2802531, SAMEA3542243, SAMEA3499831, SAMEA4364757, SAMEA4364756, SAMEA4364755, SAMEA4364754). Two of these samples – the domestic donkey and the Somali wild ass – were also used in the reference panel, since no other genomes generated using Illumina technology were available. Synthetic hybrids were generated for all possible combinations involving two different groups (Horse x Zebra, Horse x Ass, Ass x Zebra), by sampling a fixed number of reads (N = 1,000, 10,000, 100,000, or 1,000,000) from the two samples used. Each of the N reads were sampled from one of the two sources with equal probability and without replacement. The resulting files were analyzed using Zonkey, and ADMIXTURE results were recorded as the empirical distributions of absolute deviations from the expected admixture percentages (50%) for each of the possible combination of groups (ie. Horse x Zebra, Horse x Ass, Ass x Zebra). A total of 100 replicates were generated for each combination of samples. These empirical distributions are used within Zonkey to assess whether the admixture proportions observed for a given sample significantly deviate from the 50:50 expectation, where the distribution quantile provides an empirical *p*-value. The false positive rate for non-hybrids was estimated by calculating the admixture proportions on pure samples generated by sampling 100 replicates of N reads from each of the 13 genomes listed above. No false positives, defined as having an unex-

pected admixture proportion greater than 0.00002, were observed in non-hybrids.

2. Ancient DNA analyses

2.1. Samples

We selected 18 ancient equine archaeological remains from Yenikapi, Dangstetten and Mehr Ali (Table 1). The Yenikapi excavations are located in the Yenikapi district of Istanbul, covering an area of ~50,000 m^2 situated 1.5 km inland from the Marmara Sea (Onar et al., 2012). A total of 20,881 bone specimens were excavated and morphologically identified at this site, including 32.6% horses. Here, we focused on 12 skull remains, which were measured for a number of morphological characters using Angela von den Driesch's nomenclature (von den Driesch, 1976). A total of 18 measurements were then converted into the Eisenmann system (Eisenmann, 1986) and compared to a reference collection consisting of, depending on the characters considered, 17–21 horses, 75–98 donkeys, 23–24 mules and 13–14 hinnies (Table S2). The comparative method was based on Ratio diagrams (Eisenmann, 1986) using a reference of 46 Persian onagers to reveal differences in size and proportion (Fig. S1).

The Dangstetten samples originate from the short-lived eponymous Roman military site in southern Germany, dated to 15 BCE - 9 BCE. The morphology of the osteological equid material, in particular their possible identification as mules, has been discussed in (Uerpmann and Uerpmann, 1994).

Mehr Ali is located in the province of Fars (southwestern Iran). The site belongs to the Chalcolithic Lapui culture (6th-4th mill. BCE), and shows an over-representation of domestic animals such as cattle, sheep and goats but also a significant number of wild herbivores, such as hemiones and gazelles (Sardari, 2013; Sheikhi et al., 2012). One of the three teeth analyzed by Zonkey was identified as belonging to a pure hemione (Iran2_CGG_1_017448, Table 1). We measured three morphological characters, including the length of the molar (L = 25 mm), the breadth (l = 27 mm) and the length of the protocone (LPt = 12 mm), and compared these to a reference panel of 185 modern individuals, available from http://www.vera-eisenmann.com/ and consisting of 102 horses, 27 Persian onagers and 56 donkeys (Fig. 5).

2.2. Extraction and sequencing

DNA extractions were performed in the aDNA facilities of the Centre for GeoGenetics, University of Copenhagen (Denmark), using two different methods. DNA extraction for the Yenikapi samples was carried out following method 'Y' described in (Gamba et al., 2015), with slight modifications. 115–620 mg of bone powder were digested for 1 h at 37 °C in 1–5 ml of lysis buffer consisting of 0.45M EDTA, N-lauryl 0.5% Sarcosyl and 0.25 mg/ml Proteinase K. DNA extracts were then generated from the fraction of undigested pellets, following an overnight incubation with agitation at 42 °C. Prior to DNA library preparation, 22.75 μl of DNA extracts were incubated for 3 h at 37 °C with 7 μl of USER enzyme (NEB®), consisting of a mixture of Uracil-DNA Glycosylase and Endonuclease VIII.

DNA extractions from the Dangstetten and Mehr Ali samples were performed following (Dabney et al., 2013), using 170–405 mg of bone/tooth powder and a lysis buffer consisting of 0.45M EDTA and 0.25 mg/ml Proteinase K. DNA extracts were obtained using the second digestion fraction recovered from the supernatant of an overnight incubation with agitation at 37 °C and were not USER treated.

DNA libraries were constructed following (Meyer and Kircher, 2010) in a final volume of 25 μl per reaction. We used 14.9 μl of USER-treated DNA extract for the Yenikapi samples and 21.3 μl of DNA extracts otherwise. Following the Bst elongation step, a 20× dilution of unpurified DNA library was subjected to quantitative Polymerase Chain Reaction (PCR) to determine the minimal number of PCR cycles required for subsequent amplification. Real-time amplification was performed in 20 μl (Yenikapi) or 25 μl (Dangstetten and Mehr Ali) reactions on a Roche LightCycler 480 Real-time PCR System, following (Gamba et al., 2015). The Yenikapi DNA libraries were then indexed and amplified in 25 μl using AccuPrime™ Pfx DNA polymerase and 3 μl of the initial unpurified DNA library. We used AmpliTaq Gold® DNA polymerase in 50 μl reactions with 12.5 μl of unpurified DNA library for other samples. Finally, DNA libraries were purified using MinElute kits (QIAgen) and eluted in 25 μl of Elution Buffer (10 mM Tris-Cl, pH 8.5). Indexed libraries were quantified on the TapeStation 2200 instrument (Agilent technologies) using a 10× dilution and were pooled before sequencing at

Table 1
Samples analyzed and genetic results.

Sample name	Lab number	Location	%Endogenous	Nuclear reads	mtDNA reads	Sex	Parents (♂x♀)
Tur140	CGG_1_018706	Yenikapi	75.06	715,934	174	♂	Horse x Horse
Tur141	CGG_1_018707	Yenikapi	75.36	721,316	345	♂	Horse x Horse
Tur142	CGG_1_018708	Yenikapi	73.69	709,604	163	♂	Horse x Horse
Tur144	CGG_1_018710	Yenikapi	73.88	717,463	195	♀	Donkey x Horse
Tur147	CGG_1_018713	Yenikapi	75.02	732,769	175	♀	Donkey x Horse
Tur149	CGG_1_018715	Yenikapi	71.17	698,721	173	♀	Donkey x Horse
Tur171	CGG_1_018737	Yenikapi	66.20	643,504	295	♂	Horse x Horse
Tur189	CGG_1_018755	Yenikapi	73.09	701,674	171	♀	Donkey x Horse
Tur191	CGG_1_018757	Yenikapi	73.86	723,722	138	♂	Donkey x Horse
Tur193	CGG_1_018759	Yenikapi	73.03	709,352	158	♂	Horse x Horse
Tur194	CGG_1_018760	Yenikapi	71.35	682,386	275	♂	Horse x Horse
Tur206	CGG_1_018772	Yenikapi	72.35	695,512	152	♀	Donkey x Horse
R10DA_1104	CGG_1_017516	Dangstetten	0.25	290,313	124	♂	Horse x Horse
R13DA_217	CGG_1_017519	Dangstetten	0.09	96,483	94	♀	Horse x Horse
R14DA_959	CGG_1_017520	Dangstetten	0.36	18,779	39	♀	Donkey x Horse
Iran1	CGG_1_017447	Mehr Ali	0.07	6510	24	♀	Horse x Horse
Iran2	CGG_1_017448	Mehr Ali	0.08	84,294	336	♀	Hemione x Hemione
Iran3_Bijar	CGG_1_017449	Mehr Ali	0.41	637,366	1596	♂	Horse x Horse

The fraction of unique equine sequences identified following alignment against the horse reference genome is reported as %Endogenous. The total numbers of high-quality unique alignments are indicated, along with the molecular identification of the sample sex and the taxonomic status of their parents.

the Danish National High-Throughput DNA Sequencing Centre on the Illumina HiSeq 2500 instrument, using 94–126 single-end cycles.

3. Results and discussion

3.1. Methods validation

We tested the performance of Zonkey using RNA-seq data for three previously published domestic donkeys (SAMN00809366-SAMN00809368), three mules (SAMN00809369-SAMN00809371) and one hinny (SAMN00631159) (Wang et al., 2013). The sequencing data were mapped against the horse nuclear and mitochondrial genomes as described in (Schubert et al., 2014). We then applied Zonkey with default parameters to these alignments, thereby limiting the analysis to at most 1,000,000 randomly selected reads. The resulting reports are provided in Supplemental file 1 and a representative subset are shown in Fig. 2. All donkeys were found to cluster together with the reference donkey, regardless of the analyses performed (PCA, phylogenetic reconstructions, and ADMIXTURE) and the loci considered (mtDNA and nuclear genomes), thus confirming their identification as pure donkeys.

Visual inspection of the admixture analyses clearly identified the three mules and the hinny as hybrids. In the PCA, these individuals did not cluster with any member of the Zonkey reference panel but were found intermediate between horses and donkeys, as expected. Phylogenetic analyses of the mules and the hinny supported that their mtDNA belonged to the horse and the donkey, respectively. In the absence of admixture, TreeMix phylogenetic reconstructions also showed that the samples placed together with horses and shared large residuals with donkeys, reflecting a strong donkey ancestry. This was confirmed when allowing for one migration edge, which dramatically reduced the magnitude of residuals. Corresponding TreeMix results were found for the hinny.

Admixture estimates, however, deviated from the expected 50:50 donkey to horse proportions, as much as 41:59 and 45:55, for two and three ancestral groups, respectively, and only the estimates involving three ancestral groups fell inside the empirical distribution (p-value = 0.07–0.14). This most likely reflects a combination of biases. Firstly, the data for these samples was generated using RNA-seq, which shows differential parental gene expression (Wang et al., 2013), and represents an extremely skewed representation of the genome compared to whole-genome shotgun sequencing data used for the Zonkey reference panel and used to synthesize 50:50 hybrids for the empirical admixture distributions. Secondly, the reference panel used in Zonkey is unbalanced between caballine (two) and non-caballine individuals (seven) and unequal sample sizes are known to affect ADMIXTURE estimates (Shringarpure and Xing, 2014). Finally, mapping against the horse reference genome facilitates the identification of genetic variants segregating in horses versus other species. Aligning reads against an outgroup could reduce this bias, but is impracticable for equids since they diverged from their closest living phylogenetic relatives ∼55 myrs ago (Steiner and Ryder, 2011).

In order to estimate false negative and false positive rates, we used Zonkey on synthetic sequence datasets consisting of a 50:50 mixture of every possible combination of parental species, as well datasets consisting purely of a single sample. To explore the minimal amounts of reads required for a positive identification of hybrids, we considered datasets of 1,000, 10,000, 100,000 and 1,000,000 sequences. The results of all combinations tested are presented in Table S1, and in Fig. 3 for the particular case of mules. In this case, we considered that a hybrid was positively detected as long as the donkey ADMIXTURE ancestry was within the 0.25–0.75 range. For mules, we found a false negative rate equal to 0.4% with 1000 reads, and equal to zero otherwise (Fig. 3). Across all combinations, we observed a small number of false positives when analyzing 50:50 hybrids using only 1000 sequences, excluding transitions, and considering three ancestry groups (k = 3). In those cases, 0–1.2% of tests identified a proportion of the genome belonging to the third ancestry group (Supplementary Table S1). Such cases are therefore marked with a warning by Zonkey. As noted above, the estimated ancestry proportions deviated from the 50:50 expectations, and varied significantly depending on the population background and/or horse breed considered in the synthetic dataset. This suggests that the limited number of reference individuals in Zonkey do not capture the entire variation within equine species.

Overall, our simulations based on synthetic datasets show that Zonkey achieves maximal specificity and sensitivity for a mere ∼10,000 equine reads. Zonkey is thus well suited for the detection of F1-hybrids in equine archaeological material, where DNA preservation and endogenous DNA content (i.e. the fraction of equine reads identified post-sequencing) are often limited. With endogenous content of 1%, as little as one million raw sequencing reads are expected to provide a robust diagnostic for mules and hinnies. This translates to experimental costs as low as ∼50€ and 130€ per sample, including DNA extraction, library preparation, amplification and DNA sequencing, when sequencing up to 125 samples on a HiSeq2500 (80bp SE), or up to 15 samples on a MiSeq sequencing kit (v2; 2 × 150 bp), respectively. Salary costs are not included in the calculations.

3.2. Application to archaeological remains

We extracted aDNA from 18 archaeological remains originating from three excavation sites (Table 1). The Yenikapi material consisted of 12 Byzantine (4-15th cent. AD) equine petrosal bones (Onar et al., 2008, 2012, 2015). Three equine bones were obtained from the Roman (15 BCE-9 BCE) military site of Dangstetten, southern Germany. Finally, three additional teeth belonging to the Chacolithic Lapui culture (6th-4th mill. BCE) were excavated at Mehr Ali, southwestern Iran. We constructed single-indexed Illumina DNA libraries and generated 1,000,000–107,799,314 sequences per sample, which were subsequently processed through Zonkey.

While six Yenikapi specimens were confirmed to be purebred horses, all six remaining consisted of mules (Table 1), for which phylogenetic reconstruction showed mitochondrial clustering with horses, while PCA, TreeMix reconstructions and ADMIXTURE confirmed shared autosomal ancestry between horses and donkeys. This was confirmed when disregarding transitions, suggesting the taxonomic identification to be robust to post-mortem DNA damage (Briggs et al., 2007), in line with the USER-treatment performed on raw extracts. The Zonkey report is provided as Supplemental File S2, and one purebred horse and one mule are shown in Fig. 4. Interestingly, the morphological analysis of 18 skull measurements based on ratio diagrams (Eisenmann, 1986) revealed that Yenikapi horse specimens are smaller than the reference material for most measurements (Fig. S1). Both Yenikapi horses and mules, therefore, largely overlap the morphological range of horses and donkeys represented in our morphological reference panel, precluding direct morphological identification of hybrids with this method. This demonstrates that our genetic approach can successfully differentiate between purebred and F1-hybrids when morphological information is inconclusive.

One of the Roman bones (R14DA_959_CGG_1_017520) was also identified as a mule (Supplementary File S2). This is in agreement with previous morphological analyses (Uerpmann and

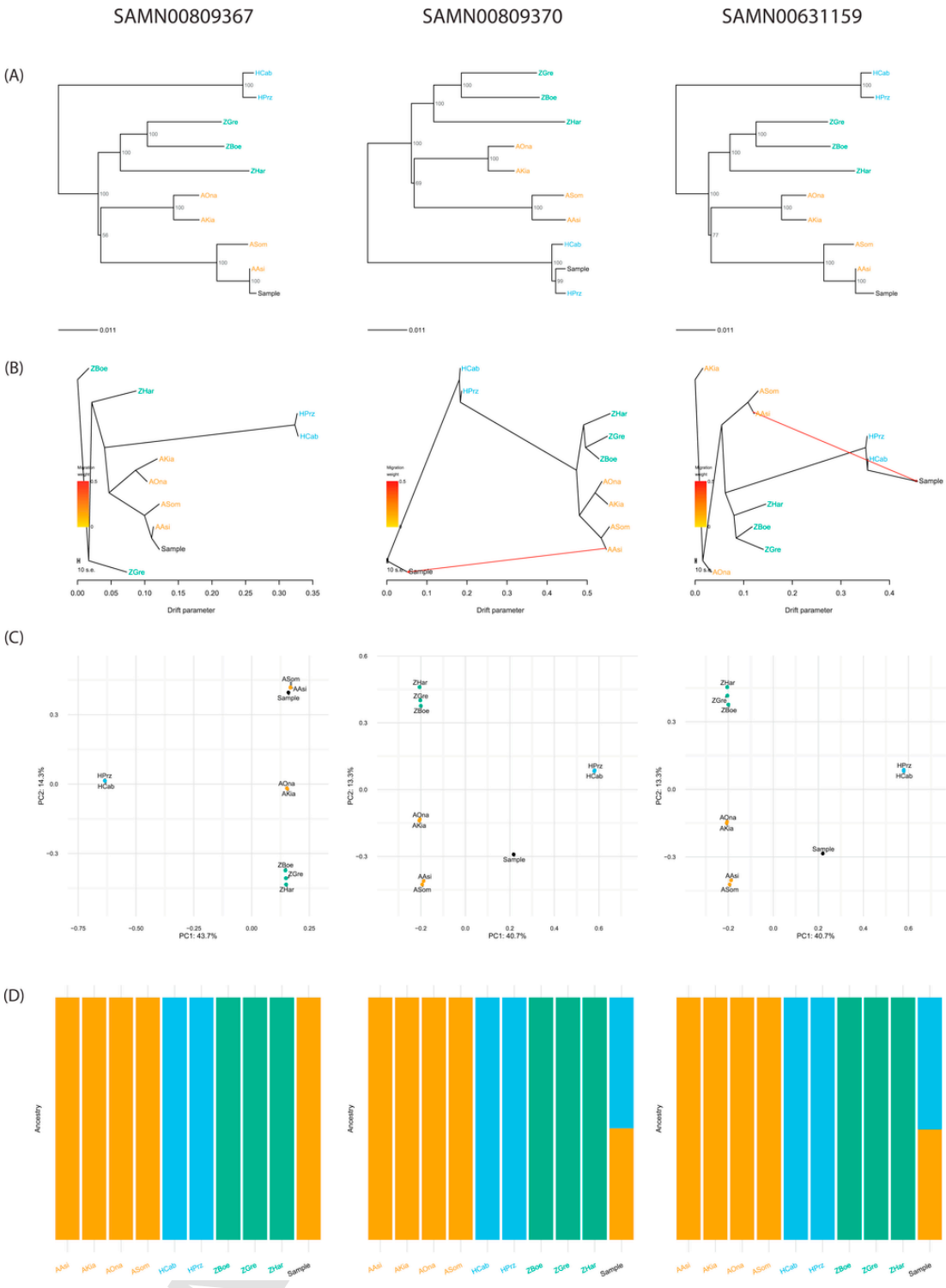

Fig. 2. Zonkey reports for RNA-seq data from donkey, mule and hinny individuals. **Panel A)** Mitochondrial phylogeny. **Panel B)** TreeMix phylogeny, considering one migration edge for samples SAMN00809370 and SAMN00631159. **Panel C)** PCA. **Panel D)** ADMIXTURE plot representing three pre-defined ancestry components (ie. Horses: Hcab, HPrz; Asses: ASom, AAsi, AKia and AOna, and; Zebras: ZBoe, ZGre, ZHar). *Left:* donkey. *Centre*: mule. *Right*: hinny. The sequencing data were originally released by (Wang et al., 2013). Analyses are based on all substitution types.

Uerpmann, 1994), which mainly exploited enamel fold patterns in the maxillar and mandibular dentition. One animal (R10DA_1104_CG-G_1_017516), morphologically identified as a horse based on dental and post cranial traits (personal communication from Prof. Hans-Peter Uerpmann, Tuebingen, Germany), was confirmed to be a horse. However, one putative morphological hybrid (R13DA_217_CG-G_1_017519) was found genetically to represent a purebred horse, which illustrates the limitations of morphology for hybrid identification.

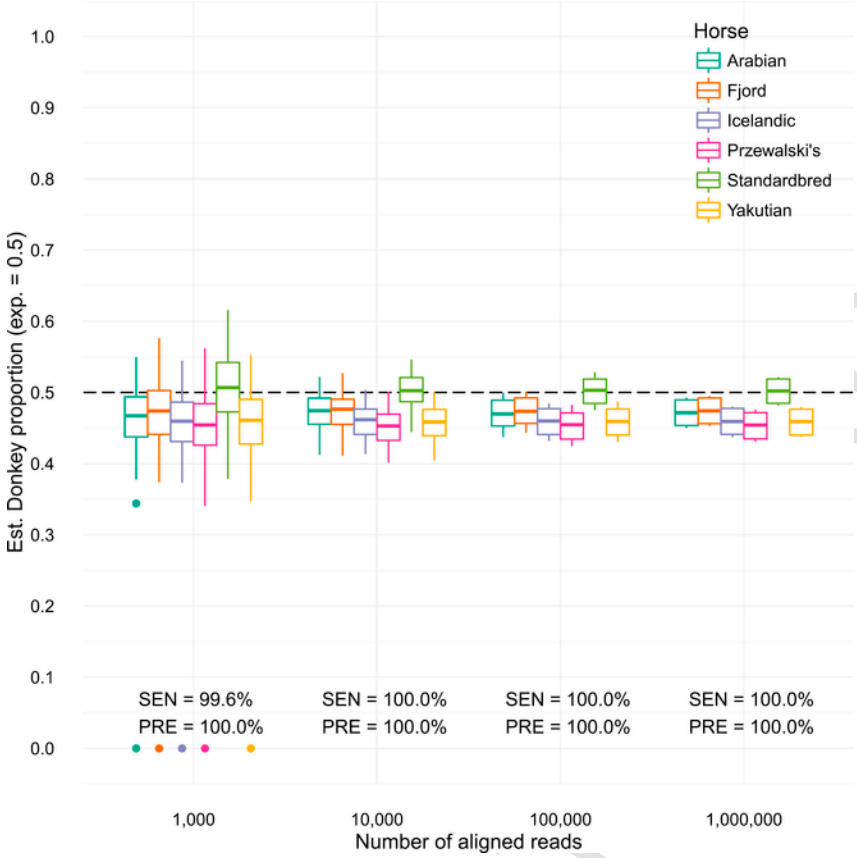

Fig. 3. ADMIXTURE proportion of donkey ancestry in synthetic donkey x horse F1-hybrids. Synthetic hybrids were generated by sampling 50:50 donkey and horse Illumina reads. Donkey sequence data were taken from a single individual, and combined with six different horses (Arabian, Fjord, Icelandic, Przewalski's, Standardbred and Yakutian). 100 random pseudo-replicates were generated for each combination of sample size and horse. The proportion of donkey ancestry was estimated by Zonkey using ADMIXTURE for k = 2 and 3, considering all substitution types. SEN = Sensitivity (ie. true positive rate); PRE = Precision (ie. positive predictive value).

Finally, none of the Mehr Ali samples were identified as F1-hybrids. Interestingly, while two of these were identified as horses (Iran1_CGG_1_017447 and Iran3_Bijar_CGG_1_017449), the third was identified as a hemione (Iran2_CGG_1_017448, Fig. 4). Morphological analyses of three tooth morphological measurements for this sample showed that the tooth lays within the range of variation for hemiones and horses, which largely overlap (Fig. 5). This again, demonstrates the utility of our approach when only fragmentary material is available and different candidate species coexisted in sympatry, as was the case for four equine species in Southwest Asia.

4. Conclusions

In this study, we have presented a pipeline based on low-depth HTS data that can identify F1-equine hybrids in archaeological assemblages, with high sensitivity and specificity. All related computational tools are included in Zonkey, as part of the open-source and freely available PALEOMIX pipeline. Using this methodology, we were able to identify seven mules from 12 Byzantine and three Roman equine archaeological remains. Our approach also allowed us to identify a female hemione from a Chalcolithic site in southwestern Iran, where classical tooth measurements could not tease apart horses and hemiones. Assuming that genome-scale sequence data becomes available for other closely-related mammal species, the framework presented here can easily be applied to the identification of F1-hybrids in other groups. As reduced sensitivity was observed for synthetic ass/zebra crosses (Supplementary Table S1), we propose that Zonkey could be useful for clades that split prior to ~2 million years ago (Jónsson et al., 2014), though this estimate assumes similar levels of genetic differences in the groups of interest. The details of constructing additional reference panels are described in the Zonkey documentation. For now, it provides a cost-effective method to investigate the true importance of equine hybrids in past societies.

Conflicts of interest

The authors declare that they have no conflict of interest.

Funding

This work was supported by the Danish Council for Independent Research Natural Sciences (Grant 4002-00152B); the Danish National Research Foundation (Grant DNRF94); the Villum Fonden (Grant miGENEPI); the International Research Group Program (Grant IRG14-08), Deanship of Scientific Research, King Saud University, and; the “Chaires d'Attractivité 2014” IDEX, University of Toulouse, France (OURASI). LE and CGam were supported by the Marie-Curie Intra-European Fellowship program (FP7-IEF-302617 and FP7-IEF-328024, respectively). JW was supported by a British Academy/Leverhulme award (Grant SG121966).

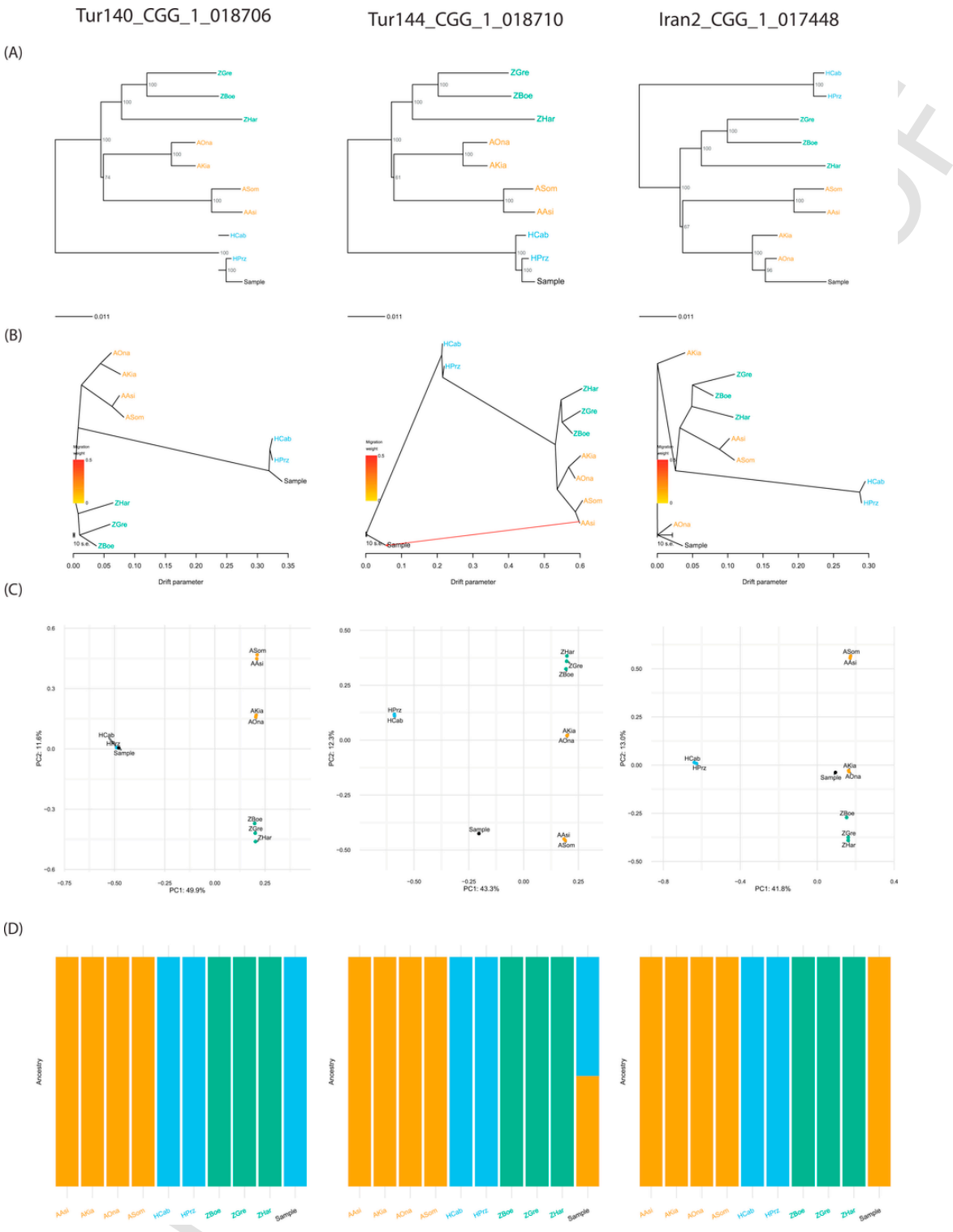

Fig. 4. Zonkey report for a purebred horse (left), a mule (middle) and an hemione (right). For additional information, see Fig. 2 captions.

Acknowledgements

We thank the staff of the Danish National High-Throughput DNA Sequencing Centre for their technical assistance and members of the Paleomix group at the Centre for GeoGenetics for fruitful discussions. MM, SS and all co-authors are particularly grateful to AR. Sardari at the Iranian Center for Archaeological Research (ICAR) and Iranian Cultural Heritage, Handicraft and Tourism Organisation (ICHHTO) for his authorisation to study the faunal material of Mehr Ali Tepe.

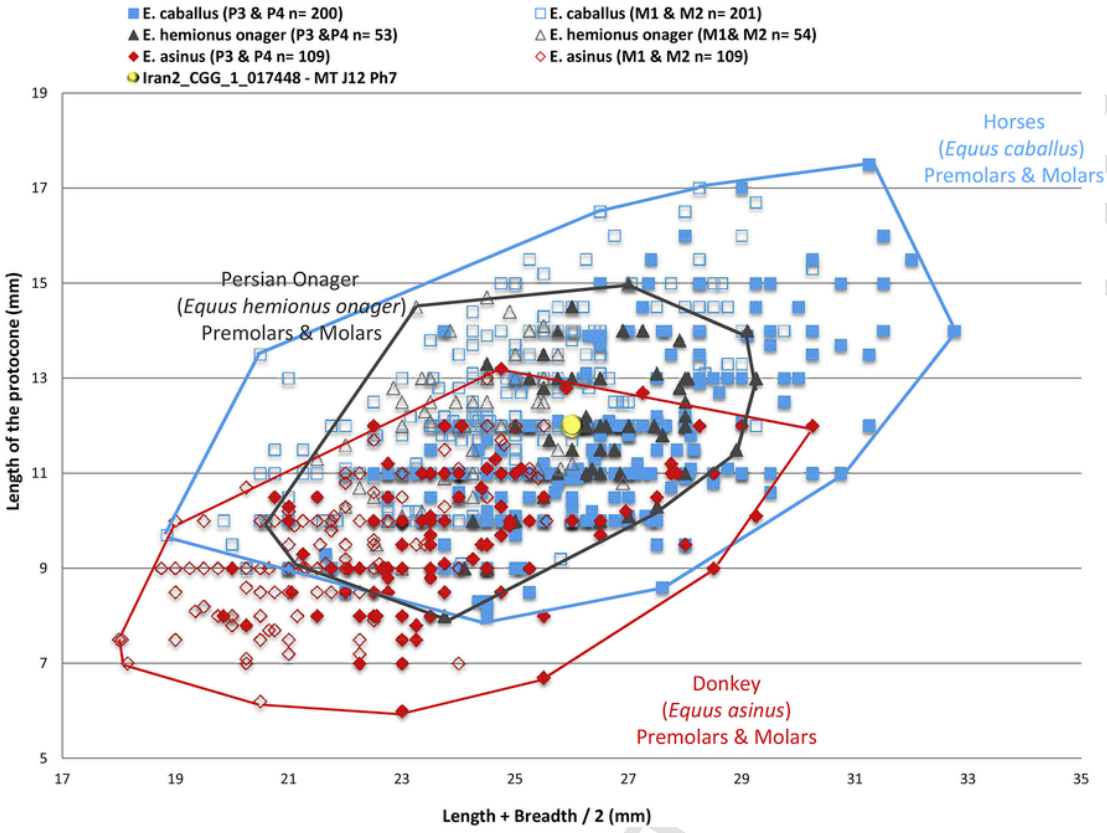

Fig. 5. Tooth morphological range of donkeys, horses and hemiones. Sample **Iran2_CGG_1_017448** (yellow) is genetically a purebred hemione. L = length of the molar. l = Breadth of the Protocone. LPt (y-axis) = Length of the Protocone. P = Premolar. M = Molar. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jas.2016.12.005.

Uncited reference

References

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664.

Armitage, P.L., Chapman, H., 1979. Roman Mules. London Archaeologist Association, 3.

Baxter, I.L., 1998. Species identification of equids from Western European archaeological deposits: methodologies, techniques and problems. In: Current and Recent Research in Osteoarchaeology, Proceedings of the Third Meeting of the Osteoarchaeology Research Group. Oxbow, Oxford, pp. 3–17.

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., Pääbo, S., 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. U. S. A. 104, 14616–14621.

.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.

Chuang, R., 2016. The Acquisition of Domestic Equids in Roman Britain: the Identification of Domestic Equids and Case Study with Isotopic Analysis. PhD. Thesis University of Southampton.

Clutton-Brock, J., 1999. A Natural History of Domesticated Mammals, second ed. Cambridge University Press.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., Garcia, N., Pääbo, S., Arsuaga, J.-L., Meyer, M., 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. U. S. A. 110, 15758–15763.

Davis, S.J., 1980. Late pleistocene and holocene equid remains from Israel. Zool. J. Linn. Soc. 70, 289–312.

Der Sarkissian, C., Ermini, L., Schubert, M., Yang, M.A., Librado, P., Fumagalli, M., Jónsson, H., Bar-Gal, G.K., Albrechtsen, A., Vieira, F.G., Petersen, B., Ginolhac, A., Seguin-Orlando, A., Magnussen, K., Fages, A., Gamba, C., Lorente-Galdos, B., Polani, S., Steiner, C., Neuditschko, M., Jagannathan, V., Feh, C., Greenblatt, C.L., Ludwig, A., Abramson, N.I., Zimmermann, W., Schafberg, R., Tikhonov, A., Sicheritz-Ponten, T., Willerslev, E., Marques-Bonet, T., Ryder, O.A., McCue, M., Rieder, S., Leeb, T., Slatkin, M., Orlando, L., 2015. Evolutionary genomics and conservation of the endangered Przewalski's horse. Curr. Biol. 25, 2577–2583.

Eisenmann, V., 1986. Comparative osteology of modern and fossil horses, halfasses and asses. Equids anc. World 67–116. In: Uerpmann, H.-P., Meadow, R.H. (Eds.), Equids in the Ancient World. Ludwig Reichert Verlag, Wiesbaden, pp. 67–116.

Eisenmann, V., Mashkour, M., 1999. The small equids of Binagady (Azerbaidjan) and Qazvin (Iran): E. Hemionus Binagadensis nov. subsp. and E. Hydruntinus. Geobios Mem. Spec. 32, 105–122.

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Bradley, D.G., Orlando, L., 2015. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. Mol. Ecol. Resour. 16, 459–469.

Groves, C.P., Willoughby, D.P., 1981. Studies on the taxonomy and phylogeny of the genus Equus., 1. Subgeneric classification of the recent species. Mammalia 45, 321–354.

Johnstone, C., 2004. A Biometric Study of Equids in the Roman World. University of York.

Johnstone, C., 2008. Commodities or logistics? The role of equids in Roman supply networks. In: Feeding the Roman Army: the Archaeology of Production and Supply in NW Europe. pp. 128–145.

Jónsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., Albrechtsen, A., Petersen, B., Korneliussen, T.S., Vilstrup, J.T., Lear, T., Myka, J.L., Lundquist, J., Miller, D.C., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Stagegaard, J., Strauss, G., Bertelsen, M.F., Sicheritz-Ponten,

T., Antczak, D.F., Bailey, E., Nielsen, R., Willerslev, E., Orlando, L., 2014. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc. Natl. Acad. Sci. U. S. A. 111, 18655–18660.
Konrad, H.W., 1980. A Jesuit Hacienda in Colonial Mexico: Santa Lucia, 1576–1767. Stanford University Press.
Laurence, R., 1999. The Roads of Roman Italy: Mobility and Cultural Change. Routledge, London.
Mashkour, M., 2002. Chasse et élevage au nord du Plateau central iranien entre le Néolithique et l'Âge du Fer. Paléorient 28, 27–42.
Mashkour, M., 2003. Equids in the northern part of the iranian central plateau from the neolithic to iron age: new zoogeographic evidence. In: Levine, M., Colin, R. Renfrew, Boyle, K. (Eds.), Prehistoric Steppe Adaptation and the Horse. McDonald Institute for Archaeological Research, Cambridge, pp. 129–138.
Meyer, M., Kircher, M., 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb. Protoc.
Onar, V., Pazvant, G., Armutak, A., 2008. Radiocarbon dating results of the animal remains uncovered at Yenikapi Excavations. In: ISTANBUL Archaeol. MUSEUMS Proc. 1st Symp. Marmaray–metro Salvage Excav. 5th–6th, Istanbul. pp. 249–256.
Onar, V., Çakirlar, C., Janeczek, M., Kiziltan, Z., 2012. Skull typology of byzantine dogs from the theodosius harbour at Yenikapi, Istanbul. Anat. Histol. Embryol. 41, 341–352.
Onar, V., Pazvant, G., Pasicka, E., Armutak, A., et al., 2015. Byzantine horse skeletons of Theodosius Harbour: 2. Withers height estimation. Rev. Med. 32–40.
Orlando, L., Mashkour, M., Burke, A., Douady, C.J., Eisenmann, V., Hänni, C., 2006. Geographic distribution of an extinct equid (Equus hydruntinus: mammalia, Equidae) revealed by morphological and genetical analyses of fossils: the phylogenetic origin of Equus hydruntinus. Mol. Ecol. 15, 2083–2093.
Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P.L.F., Fumagalli, M., Vilstrup, J.T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C.D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A.M.V., Cahill, J., Rasmussen, M., Wang, X., Min, J., Zazula, G.D., Seguin-Orlando, A., Mortensen, C., Magnussen, K., Thompson, J.F., Weinstock, J., Gregersen, K., Rø, K.H., Eisenmann, V., Rubin, C.J., Miller, D.C., Antczak, D.F., Bertelsen, M.F., Brunak, S., Al-Rasheid, K.A.S., Ryder, O., Andersson, L., Mundy, J., Krogh, A., Gilbert, M.T.P., Kjær, K., Sicheritz-Ponten, T., Jensen, L.J., Olsen, J.V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J., Willerslev, E., 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499, 74–78.
Orlando, L., Gilbert, M.T.P., Willerslev, E., 2015. Reconstructing ancient genomes and epigenomes. Nat. Rev. Genet. 16, 395–408.
Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and Evolution in R language. Bioinformatics 20, 289–290.
Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet. 2, e190.
Peddie, J., 1987. Invasion: the Roman Invasion of Britain in the Year Ad 43 and the Events Leading to Their Occupation of the West Country. Sutton Publishing Ltd.
Peters, J., 1998. Römische Tierhaltung und Tierzucht: eine Synthese aus archäozoologischer Untersuchung und schriftlich-bildlicher Überlieferung. M. Leidorf, Rahden.
Pickrell, J.K., Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8, e1002967.
Postgate, J.N., 1986. The equids of Sumer. In: Uerpmann, H.-P., Meadow, R.H. (Eds.), Equids in the Ancient World. Ludwig Reichert Verlag, Wiesbaden, pp. 194–206.
Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909.
Roth, J.P., 1999. The Logistics of the Roman Army at War: 264 B.C. - A.D. 235, Columbia Studies in the Classical Tradition/Columbia Studies in the Classical Tradition. Brill, Leiden.
Sardari, A., 2013. Northern Fars in the fourth millennium BC: cultural developments during the Lapui phase. In: Petire, C. (Ed.), Ancient Iran and its Neighbours: Local Developments and Long-range Interactions in the 4th Millennium BC. Oxbow Books, Oxford, pp. 190–201.
Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., Willerslev, E., Orlando, L., 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. 9, 1056–1082.
Sheikhi, S., Mashkour, M.A.S., Sardari Zarchi, A., 2012. Subsistence Economy of the Lapui settlement of Tepe Mehr Ali Fars on the basis of the archaeozoological analysis. [Baresi eghtesad zisti sakenan doreyeh lapui mohavateyeh Mehr Ali Fars bar asas bazmandehaye ostkhanhaye janevari]. J. Iran. Archaeol. 2, 39–60.
Shringarpure, S., Xing, E.P., 2014. Effects of sample selection bias on the accuracy of population structure and ancestry inference. G3 Genes|Genomes|Genetics 4, 901–911.
Singleton, J., 1993. Britain's military use of horses 1914-1918. Past. Present 139, 178–203.
Smith, D.C., 2008. Book of Mules: Selecting, Breeding, and Caring for Equine Hybrids. Lyons Press.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Steiner, C.C., Ryder, O.A., 2011. Molecular phylogeny and evolution of the Perissodactyla. Zool. J. Linn. Soc. 163, 1289–1303.
Steiner, C.C., Ryder, O.A., 2013. Characterization of Prdm9 in equids and sterility in mules. PLoS One 8, e61746.
Tegetmeier, W.B., Sutherland, C.L., et al., 1895. Horses, Asses, Zebras, Mules and Mule Breeding. H. Cox, London.
Twiss, K.C., Wolfhagen, J., Madgwick, R., Foster, H., Demirergi, G.A., Russell, N., Everhart, J.L., Pearson, J., Mulville, J., 2016. Horses, hemiones, Hydruntines? Assessing the reliability of dental criteria for assigning species to southwest asian equid remains. Int. J. Osteoarchaeol http://dx.doi.org/10.1002/oa.2524.
Uerpmann, M., Uerpmann, H.-P., 1994. Animal bone finds from excavation 520 at Qala'at al-Bahrain. In: In: Hojlund, F., Andersen, H. (Eds.), Qala'at Al-Bahrain. 1 the Northern City Wall and the Islamic Fortress. vol. 1. Jutland Archaeolological Society Publications, pp. 417–444.
Vila, E., 2006. Data on equids from late fourth and third millenium sites in Northern Syria. In: Mashkour, M. (Ed.), Equids in Time and Space. Oxbow, Chippenham, pp. 101–123.
Vilstrup, J.T., Seguin-Orlando, A., Stiller, M., Ginolhac, A., Raghavan, M., Nielsen, S.C.A., Weinstock, J., Froese, D., Vasiliev, S.K., Ovodov, N.D., Clary, J., Helgen, K.M., Fleischer, R.C., Cooper, A., Shapiro, B., Orlando, L., 2013. Mitochondrial phylogenomics of modern and ancient equids. PLoS One 8, e55950.
von den Driesch, A., 1976. A Guide to the Measurement of Animal Bones from Archaeological Sites. Harvard University, Peabody Museum of Archaeology and Ethnology, Peabody Museum Bulletin 1.
Wang, X., Miller, D.C., Harman, R., Antczak, D.F., Clark, A.G., 2013. Paternally expressed genes predominate in the placenta. Proc. Natl. Acad. Sci. U. S. A. 110.
Weber, J.A., 2008. Elite equids: redefining equid burials of the mid-to late 3rd millennium BC from Umm el-Marra, Syria. In: In: Vila, E., Gourichon, L., Choyke, A.M., Buitenhuis, H. (Eds.), Trav. la Maison l'Orient la Méditerranée. vol. 49. Archaeozoology of the Near East VIII. TMO 49, Lyon, pp. 499–519.
Wickham, H., 2009. Ggplot2 Elegant Graphics for Data Analysis (Use R!), first ed. Springer, New York.
Zeder, M.A., 1986. The equid remains from Tal-e Malyan. In: Uerpmann, H.-P., Meadow, R.H. (Eds.), Equids in the Ancient World. Ludwig Reichert Verlag, Wiesbaden, pp. 366–412.