Journal of Archaeological Science 78 (2017) 147–157

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: http://www.elsevier.com/locate/jas

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, Université 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 5 January 2017

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.

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.

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

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

Abbreviations

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

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 (Smith, 2008; Tegetmeier et al., 1895) when hundreds of thousand mules were used by the British Army in World War I (Singleton, 1993).

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 now 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 1,000 sequences mapped to the horse reference genome. Applying Zonkey to 18 archaeological specimens spanning the last ~6,000–8,000 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