A '*long-fus*e domestication' of the horse? Tooth shape suggests explosive change in modern breeds compared to extinct populations and living Przewalski horses

Krish Seetah 1†, Andrea Cardini 2,3 & Graeme Barker 4

[1†] Corresponding author:
Department of Anthropology, Stanford University, Stanford, CA 94305, USA.
Email: kseetah@stanford.edu
Tel: 1 650 721 5939

[2]Dipartimento di Scienze Chimiche e Geologiche, Universita` di Modena e Reggio Emilia,
Via Campi 103, 41125 Modena, Italy

[3]Centre for Forensic Science, The University of Western Australia, 35 Stirling Highway,
Crawley, WA 6009, Australia.

[4]McDonald Institute for Archaeological Research, Cambridge, CB2 3ER, United Kingdom.

Abstract

Archaeological and molecular data suggests that horses were domesticated comparatively recently, the genetic evidence indicating that this was from several maternal haplotypes but only a single paternal one. However, although central to our understanding of how humans and environmental conditions shaped animals during domestication, the phenotypic changes associated with this idiosyncratic domestication process remain unclear. Using geometric morphometrics on a sample of horse teeth including Pleistocene wild horses, modern Icelandic and Thoroughbred domestic horses, Przewalski wild horses of recent age, and domestic horses of different ages through the Holocene, we show that, despite variations in size likely related to allometry (changes to bone size in proportion to body size), natural and artificial selective pressures, and geographic and temporal heterogeneity, the *shape* of horse teeth has changed surprisingly little over thousands of years across Eurasia: the shapes of the premolars of prehistoric and historic domestic horses largely resemble those of Pleistocene and recent wild horses. However, this changed dramatically after the end of the Iron Age with an explosive increase in the pace and scale of variation in the past two millennia, ultimately resulting in a two-fold jump in the magnitude of shape divergence in modern breeds. Our findings indicate that the pace of change during domestication may vary even within the same structure with shape, but not size, suggesting a 'long-fuse' model of phenotypic modification, where an initial lengthy period of invariance is followed by an explosive increase in the phenotypic change. These observations support a testable model that is applicable to other traits and species, and add a new layer of complexity to the study of interactions between humans and the organisms they domesticated.

Key words

Artificial selection; disparity; morphology; phenotype; Procrustes; teeth.

Introduction

Horse domestication had a profound effect on humans (Clutton-Brock 1999; Anthony 2007; Olsen 2006). The timing of horse domestication around 3500 BCE (Outram et al. 2009), and the subsequent extirpation of wild progenitors, potentially correlates with important environmental shifts i.e. expansion and contraction of forest cover (Warmuth et al. 2011). However, reconstructing the history of this event has proved challenging (Lippold et al. 2011; Wade et al. 2009; Steiner 2013); a deeper knowledge of the processes that underpinned horse domestication has important implications for our understanding of both the human groups that first undertook this particular domestication event (or events), as well as the ecological setting within which domestication was initiated. Critically, with the expansion of domestic populations our ability to distinguish between discrete groups of wild and domestic horses diminishes to nil (Bendry 2012).

Maternally inherited mtDNA of modern horse breeds demonstrates a high degree of variability, with an exponential population expansion starting 6000-8000 years ago (Lippold et al. 2011). Genomic analyses confirm that horses did not undergo strong genetic bottlenecks and indicate close relationships among breeds and even among domestic horses and their closest living relative, the Przewalski wild horse, *Equus przewalskii* (Wade et al. 2009). The domestication process is strongly sex-biased (Wade et al. 2009): differences in paternal-inherited DNA are small and all modern breeds share the same Y chromosome haplotype (Lingren et al. 2004). Overall, the genetic evidence suggests that mares from multiple populations contributed to the gene pool but that only a few stallions were domesticated.

Archaeological data supporting western central Eurasia as a likely centre of early horse domestication (the presence of mares-milk and carcass residues in ceramic vessels, alongside pathological markers and bit-wear indicative of horse riding *c*.3500 BCE: Outram et al. 2009) are now corroborated by autosomal genotype data (Warmuth et al. 2011; 2012). These same archaeological lines of evidence point to an established and well-developed economy that included the exploitation of a range of secondary products (Outram et al. 2009; Bendry 2011), not just the horse’s ability to transport people and goods. Thus, despite strong consensus that the horse was domesticated for riding, we still need a better assessment of the overall process of domestication and how both the socio-economic and environmental setting influenced this event over the course of time and across space.

Investigating morphological responses

As selection acts on phenotypes, the manner in which behaviour and morphology have been shaped by human-animal interactions, and with the environment, is fundamental to our understanding of the process of domestication. Although genetic approaches have elucidated phenotypic features such as coat colour (Ludwig et al. 2009) and gait variability (Andersson et al. 2012), we have yet to make strong connections between the behavioural and anatomical correlates of genetic changes. Furthermore, at least some aspects of caballine equid morphology potentially reflect adaptation: larger phalanges in glacial horses might be an adaptation to heavy grounds (Bignon & Eismann 2012); shorter muzzles in cold environments follow the prediction of Allen's rule (Eismann & Baylac 2000).

Teeth offer an excellent starting point to address adaptive phenotypic responses. They are one of the most frequently recovered faunal elements within an archaeological setting, they are shaped by natural selection to meet functional demands in relation to diet, and they often mirror important features of a species' environmental setting (Kaiser & Schulz 2006; Evans 2013). Indeed, the term 'dental ecology' has been used to illustrate the power of teeth for informing on responses to the environment (Cuozzo & Sauther 2012). Using geometric morphometrics (henceforth GMM: Adams et al. 2013), we explored premolar variation in a unique dataset, in terms of spatio-temporal variation. The samples span Eurasia, with the majority of the archaeological specimens dating to *c*.. 4500-2000 years ago (Fig. 1a; sample details in Supplementary Information). This period is well after initial domestication, but during a time when populations of wild horses were still present (Warmuth et al. 2011). GMM provides a robust framework within which to study both size and shape data, resulting in crucial phenotypic evidence that can serve as an important complement to molecular techniques such as DNA and isotope analysis. The method offers other advantages: for example, it is non-destructive, and large samples can be assessed with minimal expense, as the technique itself is low-cost. GMM offers a significant, and relatively new, addition to the arsenal of techniques that capitalizes on an abundant primary archaeological resource, faunal remains, to address key issues about human / animal / environmental interactions. Thus, GMM offers a powerful toolkit to assess how much a biometric marker (Houle et al. 2010), such as the premolar, has been shaped by interactions with humans and the environment during domestication.

>> Fig. 1. (a) Geographic distribution and age of samples. (b) Premolar landmark configuration. (c) Number (N) of individuals per sample, box-plots of premolar centroid size and icons (not to scale) emphasizing premolar size clusters: small (light grey), medium (grey), large (black).

>> Tab. 1: Tests for mean size differences: the percentage of variance explained by differences is under the main diagonal; P is above it. Two P values are shown: for the first one, Student's t is tested using 10000 permutations, and, for the second one, the parametric Student's t with no assumption of equal variances is used; significant Ps are in italics and Bonferroni corrected significant Ps (<0.0018) are both in italics and underscored. Comparisons between 'archaic' samples are emphasized using a light grey background.

Materials and Methods

Samples summary

Our comparative framework for the archaeological samples consisted of wild horses and two modern breeds, Icelandic (ICE: n = 50) and Thoroughbred (THB: n = 18). The wild specimens belong to an extinct Late Pleistocene population from, Šandalja, Croatia (CRO: n = 19), dated to 40,000 to 8200 yrs uncal BP and potentially *Equus ferus* – although there is no consensus on the number of wild horse species from this time period – and a small sample of modern Przewalski horses (PRZ: n = 4). The latter is the only living wild horse and likely retains primitive traits (Groves & Ryder 2000). The remaining archaeological samples were predominantly of domestic individuals. These include: Bronze Age specimens from Berel, in the Katonkaragaray Eastern Kazakh Oblast, excavated from Kurgan burial mounds (KAZ: n = 22); Bell-Beaker period materials from Szigetszentmiklós, Hungary (HUN: n = 10); samples from China dated to the East Zhou Dynasty (alongside two Pleistocene specimens from this region) (CHI: n = 16) and Iron Age materials from the site of Slepushka, (and one Bronze Age specimen from Ust'e), Russia (RUS: n = 17). We use the term 'archaic' to refer to the archaeological specimens, the Pleistocene wild horses, and the Przewalski horses. Anatomical landmarks are shown in Figure 1b. Overall, 156 individual horses were studied, with an average sample size of 20 specimens per sample (Fig.1c and SI). Sex differences were negligible (Seetah et al. 2014).

All animals were adult with a similar degree of tooth wear. Because of variation in the enamel folding between the apex of the crown and the cervical margin, an individual tooth will show an apparent change in shape of the folding as the tooth wears. Tooth wear also impacts on the proportions of the teeth themselves (Gidley 1901). For the latter, this situation is compounded by the fact that tooth wear affects the individual teeth differently; thus, tooth wear impacts on the proportions of a P4 differently to an M3. As we only used P4 teeth, modifications to the proportions of individual teeth, consequential to wear, should be uniform across our assemblage. Further, to the best of our ability we minimized variations in enamel folding across the assemblage by selecting teeth with a similar crown height (as per Gidley 1901: 97) and rejecting those that showed excess wear.

Methods

Size and shape variables were computed using a Procrustes superimposition (Rohlf & Slice 1990) on the raw landmark data (Seetah et al. 2014) digitized on high resolution digital images in TPSDig (Rohlf 2015). Groups were tested using 10,000 permutations for mean differences in size (absolute difference – i.e., the sign is not considered and the test probability is two-tailed) and shape (Procrustes distance between mean shapes; two-tailed test). Results were double-checked using t-tests for samples with unequal variance (size) or their multivariate equivalent (shape) using James' statistics (Dryden 2013). Tests were performed in MorphoJ (Klingenberg 2011), PAST (Hammer et al. 2001), NTSYSpc (Rohlf 2013) and R (R Development Core Team 2005). Allometry was tested in MorphoJ using a multivariate regression of shape onto the natural logarithm of centroid size and significance assessed with 10,000 permutations for the percentage of variance explained by size. Variation within and among samples was summarized in PAST with box-plots for size and between group principal component scatterplots for shape (BG-PCA - Seetah et al 2012). A BG-PCA projects the specimens data onto the eigenvectors of the mean shapes variance-covariance matrix.

Focusing exclusively on mean shapes, variation was summarized in R using principal component analysis (PCA) with 95% confidence envelopes computed using 1000 bootstraps in NTSYSpc. For clarity, as two types of principal component analysis were used (BG-PCA and 'standard' PCA), we stress here that, every time we refer to the between group analysis, abbreviations are preceded by the acronym BG- (i.e., BG-PCA or BG-PCs); if this acronym is missing, however, PCA and PCs simply refer to a principal component analysis summarizing total variance in a sample regardless of groups. Size and shape similarity relationships among sample means were also summarized in NTSYSpc with distance-based trees (neighbour joining using Euclidean distances for size and Procrustes distances for shape). To take into account uncertainties in estimates of means, node repeatability was assessed by bootstrapping each sample, computing the corresponding pseudo-means and the resulting tree, and finally computing a 50% majority rule consensus tree from all 1000 bootstrapped trees. Group separation was also assessed by estimating cross-validated classification accuracy using the first eight PCs of premolar shape (91.7% of variance) in a discriminant analysis (DA). The general methodological framework used in this study is described in Cardini (Cardini 2013) and exemplified in Viscosi & Cardini (2011). The bootstrap procedures used to compute the confidence envelopes and the percentages of node repeatability are detailed in Cardini & Elton (2008). The protocol used to perform the DA, including the estimate of random chance thresholds and the sensitivity analysis using a balanced design, is described in Evin et al. (2013).

Disparity (morphological variance in modern *versus* archaic horses) was tested using methods modified from the protocol described by Drake & Klingenberg (2010; for an extensive review on disparity see also Foote 1997). For univariate data, this implied a simple test for the similarity of variances, which is the permutational version of the Levene's test used by Nagorsen & Cardini (2009). The same test can be adapted to test multivariate shape variances estimated by the mean squared Procrustes distance of each individual to its group mean (D1) or equivalently the sum of variances of all shape coordinates (Drake & Klingenberg 2010; Nagorsen & Cardini 2009). Informally, this is akin to measuring the sum of the squared sides of the multivariate 'box' which contains the data. An alternative test statistics (D2) is based on estimates of the squared volume of the 'box' occupied by the data. The product of the eigenvalues of the within-group shape variance covariance matrix, which is the same as its determinant if all eigenvalues are used, provides one way to do this estimate. However, similarly to Drake and Klingenberg (2010), who used an alternative method based on convex hulls, we estimated the volume occupied by either modern or archaic individuals within the sub-space of the first three PCs of the total shape dataset. These PCs account for two thirds of the total variance and, by using them to estimate the volumes of the shape space occupied by each group, we reduced issues with numerical precision using very small numbers, such as those generated by the product of eigenvalues from Procrustes shape data (Drake and Klingenberg 2010). D2 is therefore computed in a subspace of the total shape space, whereas D1 uses all available shape information (i.e., there is no dimensionality reduction). Thus, to summarize D2, its computation meant performing two eigen-decompositions on the shape data: i) the first one used all individuals regardless of group and was performed on shape coordinates (from a common superimposition), as it simply aimed at building a subspace of the total shape space in which to compute disparity as a 'volume'; ii) the second one took the shape variables corresponding to the subspace built in i) (i.e., PC1, PC2 and PC3 scores); it was performed within-group and used the product of the three eigenvalues of one group to estimate its disparity in that subspace.

Results

Moderate premolar size differences were present despite large overlaps across samples (Fig. 1c, Table 1). Using means, the resulting tree was well supported with three main size subdivisions suggested by the longest branches of the dendrogram (Fig. 2): small (ICE, HUN and RUS), medium (THB and KAZ) and large (CHI, CRO and PRZ). Bonferroni-corrected tests were significant (with ca. 37% of variance explained on average) only when they involved groups of small-toothed versus large-toothed animals, with the largest premolars being about 15% bigger than the smallest ones. Size groups were made of a mix of samples from different periods and localities. There was no clear pattern in terms of time or geography: for instance, modern ICE were similar in size to Iron Age RUS and Bronze Age HUN. With the exception of the large-toothed CHI sample, however, the largest premolars belong to the two groups of wild horses (Pleistocene CRO and modern PRZ).

For CHI, the large mean size was not due to a bias in relation to the two Pleistocene individuals in the sample, as their size was actually very close to but slightly (ca. 1 mm) less than the mean size of Iron Age CHI.

>> Fig. 2. Summary of mean size variation: neighbour joining tree of mean sizes with percentages of support for branches. Three main size subdivisions are emphasized using different grey tones, as in Fig. 1. Arrows help visualizing pairwise comparisons significant after a Bonferroni correction.

Shape had little covariation with size (Table 2). Thus, allometry seems to have had minor and generally negligible effect both within and across samples. In contrast, in the scatterplots capturing most of the shape variance (Fig. 3, 82.7% of between group variance, corresponding to 50.7% of total sample variance; Fig. 4a, 91.0% of variance in means), variation was highly structured and the pattern looked unexpectedly simple: modern breeds showed very clear differences between each other and when compared to all archaic horses; archaic horses largely overlapped with the four archaeological samples having similar premolar shape and also resembling extinct (CRO) and living (PRZ) wild horses. Shape components other than those shown in the scatterplots did not suggest any group separation. In terms of the shape features that characterized the three main clusters (Fig. 3), ICE had a relatively short and thick premolar; THB was comparatively long and narrow; archaic samples had a somewhat intermediate shape.

>> Tab. 2: Tests for allometry within and among (using means) samples: multivariate regression of shape onto the natural logarithm of centroid size (10000 permutation test for the percentage of explained variance; significant Ps are in italics and in italics and underscored if significant after a Bonferroni correction - $P < 0.0063$).

Tests of mean shape differences (Table 3, Fig. 4b) provided results in agreement with the pattern suggested by the scatterplots. Virtually all Bonferroni-corrected comparisons involving ICE and THB were significant. In contrast, out of 15 possible pairwise comparisons between pairs of archaic samples, only five were significant. The variance explained by differences with modern breeds (on average 27.7%) was almost twice as much as that found in archaic horses (on average 15.6%). The average shape distance of either ICE or THB to any of the archaic horses was more than double that which was noted among pairs of archaic samples (respectively ≥ 0.0859 *versus* 0.0425 units of Procrustes shape distance).

>> Tab. 3: Pairwise tests (10000 permutations) for mean shape differences: the percentage of variance explained by differences is under the main diagonal; P is above it. Two P values are shown: for the first one, the Procrustes distance between sample means is the test statistics, and for the second one, James' statistics, which does not assume homoscedasticity, is used; significant Ps are in italics and Bonferroni corrected significant Ps (<0.0018) are both in italics and underscored. Comparisons between 'archaic' samples are emphasized using a light grey background.

Because sample size was heterogeneous and the PRZ sample was particularly small, results may have been affected by unbalanced sampling and differences in statistical power. To explore the sensitivity to sampling, we excluded PRZ and repeated all tests on shape using a perfectly balanced design with subsamples of 10 random individuals from each original sample. (N = 10 was chosen as this was the sample size of HUN, the second smallest sample in the study.) Results from the balanced design are briefly summarized here, as they show no appreciable difference compared to the analysis including all specimens: even with a Bonferroni correction, all 11 tests involving modern breeds (ICE and THB) were significant; however, with the same correction, only four of the 10 comparisons between pairs of archaic samples showed significant differences.

Figure 4b shows a neighbour joining tree rooted using PRZ as an 'outgroup'. Phylogenetic inference using Procrustes shape data is problematic (Adams et al. 2013) and was not our aim. The tree is better seen as another type of graphical summary of mean shape similarity relationships, a summary with a very high cophenetic correlation (0.971, an index of high accuracy in reproducing the pattern of full multivariate shape distances) and the potential of rooting the tree to explore directional change. When this was done, the tree topology suggested a progressive trend in change from wild horses to modern breeds with archaeological samples in between and modern breeds at the tips of very long branches. However, only two clusters were well supported when inaccuracies in estimates of mean shape were taken into account. The two clusters suggest the same sharp contrast between modern and archaic animals as in the scatterplots and pairwise tests.

The magnitude and direction of group shape differences were further explored using cross-validated discriminant analyses (Table 4). The baselines to assess classification accuracy were the estimated average and the 95th percentile of correct classification by random chance, which was respectively 26.3% and 30.1%. Modern breeds with the highest accuracies (more than 83% of correctly classified individuals) were thus about three times better than chance. Classification accuracy varied in archaic samples but it was lower and averaged 43%, with a small increase to 56% if the two smallest samples (HUN and PRZ) were not considered. Thus, although on average better than chance, about half of archaic individuals were misclassified into other archaic samples; however, none were misclassified as modern. With a perfectly balanced design (not shown) to control for sample size heterogeneity (the same design as in the balanced permutation tests), modern breeds would have a slightly lower average classification accuracy (75%) and archaic samples a slightly higher one (58%), but the general outcome was similar and no archaic individuals were misclassified as modern breeds. To put these percentages into context, if they were interpreted using an arbitrary but common criterion employed by taxonomists for wild populations, modern breeds would be consistently above the arbitrary 75% threshold in classification accuracy for a valid subspecies (Patten et al. 2002) and archaic samples would be mostly below it.

The overwhelming signal from shape data analysed and summarized with a multiplicity of approaches is that modern breeds are highly distinctive with long and narrow premolars (THB) or short and thick ones (ICE), whereas archaic populations tend to have an intermediate, probably more primitive, shape, that has varied little over time and across localities. These observations lead us to predict an increase in disparity before and after modern breeds were selected.

>> Tab. 4: Cross-validated percentages of individuals classified using a DA on premolar shape. Correctly classified individuals are on the main diagonal; percentages higher than 95% random chance are emphasized in red italics underscored.

>> Tab. 5: Disparity analysis: disparity 1 (D1) is based on variance (size) or sum of variances (shape coordinates) and the test statistics is the F ratio; disparity 2 (D2) is the product of the first 3 eigenvalues of the shape variance covariance matrix and the test statistics is their absolute difference; D1 and D2 are tested using 10000 permutations.

With only two modern breeds in the dataset, our samples likely underestimate their disparity and any test has to be considered preliminary and largely exploratory. Bearing this caveat in mind, we performed a disparity analysis. First, we pooled modern breeds and tested disparity differences between them, together, and the pooled sample of all archaic individuals. Then, we compared disparity between archaic horses (pooled) and each of the two modern breeds (one at a time). In this second series of tests, we expected similar amounts of disparity if variation in archaic horses, regardless of geographic origin and period, was comparable to that found within a single modern breed. For comparative purposes, disparity was also tested, using the same design, on premolar size.

We found that disparity in modern breeds (pooled), estimated by the sum of all eigenvalues (D1, Table 5), is 1.2 times larger than in archaic individuals. This is not significant, or only marginally so ($P = 0.09$). Disparity estimated using the product of the first three eigenvalues (D2, Table 4) is, however, highly significantly different ($P = 0.0016$) and 4.7 times larger in modern horses. As expected, when archaic horses are compared to one or the other modern breed (one sample at a time), disparity is generally larger in the archaic group but never reaches significance ($P > 0.05$). Results from shape data were at odds with those on size, which indicated that archaic individuals vary in size three times or more than modern breeds even when pooled ($P < 0.0001$, D1, Table 4). Thus, disparity analysis strongly suggests a discordant pattern of change in size and shape, with size varying in the archaic horses and shape being constrained within a relatively small region of the morphospace until the appearance of highly derived modern breeds.

Potential sources of error

Taxonomic and especially palaeontological analyses can be strongly affected by sampling error because of the limited availability of individuals in one or more samples. In our study, sample size is very heterogeneous, with ICE (N=50) and PRZ (N=4) falling at opposite extremes in terms of number of specimens. We assessed the impact of sampling error in a previous study using rarefaction analyses and resampling methods in this same horse teeth dataset (Cardini et al. 2015). In that work, we concluded (p. 149) that "likely, centroid size and shape variance require no less than 15–20 specimens to achieve a reasonable degree of accuracy". In the same study, largely in agreement with a previous one on monkeys (Cardini and Elton 2008), we also showed that estimates of mean shapes may require 20 or more specimens, while those of mean size can be fairly accurate with just 10 individuals. Thus, although most of our samples have sizes above or close to this minimum putative requirement, a degree of caution must be exercised in the interpretation of results and this is especially true for the smallest samples. Crucially, however, in terms of what the populations we sampled allow us to say (see also below), the main conclusion of our study seems robust. This is suggested by Figure 7a of Cardini et al. (2015) that indicates how, despite a remarkable inflation of differences in the smallest random samples from the rarefaction analysis of ICE, there is no overlap between ICE means and those of all other groups. This is because differences are so large that small samples do not change the general pattern, and this most likely holds also for THB, whose sample is relatively big and whose shape is as distinctive as that of ICE. As for the analysis of mean shapes, results from the disparity analysis are robust, because archaic samples are pooled, thus strongly mitigating the problems with small samples.

Another source of inaccuracy is related to the inclusion of three allochronic individuals (compared to the others from the same region) in two archaeological samples: two Pleistocene specimens in CHI and one Bronze Age individual in RUS. This happened because, at the time analyses were performed, we had no accurate information to age all individuals and assumed that they were of similar age as the others from the same geographical region. After we found out that CHI and RUS included 1-2 allochronic individuals each, we investigated whether their inclusion in the analyses could have altered results. To this aim, as detailed in the Supplementary Information, we showed that in terms of both size and shape, the allochronic individuals were well within the range of the corresponding samples and they did not appreciably alter the pattern of population differences in means and variances (whose correlations before and after excluding those three specimens are virtually 1). Thus, although the best choice would have been to leave those three specimens out, we acknowledged the problem, demonstrated that it has no practical consequences and kept them in the study. This avoided redoing all analyses, which is a requirement when data are analysed in a common Procrustes shape space, which is specific to the sample (and configuration) being analysed.

A third source of uncertainty is the limited number of samples in the analysis. Our samples are representative of the whole Palearctic range, include the only living population of true wild horse, the Przewalski horse, as well as an extinct one from the Pleistocene, and also include two modern breeds and archaeological material from four different localities, over a time span that corresponds to early and later stages of horse domestication. However, clearly our results will have to be corroborated by future studies on other modern and extinct populations. If, with additional samples from more regions, time periods and also more modern breeds, the vast majority of extinct populations have small differences among them and compared to Przewalski whereas most modern species have much more distinctive shapes, then our hypothesis will be corroborated. If not, it will be rejected.

Discussion

With the provisos discussed above, our data indicate that three complimentary lines of evidence (mean differences, classification accuracy and morphological disparity) suggest that the *size* of horse premolars has changed to a variable degree over time and space in both archaic and modern horses (Fig. 1c-2 and SI), but changes to *shape* have been modest, with variation mostly overlapping amongst the archaic samples (Figs. 3 & 4) until the development of modern breeds in recent centuries, when shape has become hugely distinctive (Fig. 3 & 4). With an estimated time since the divergence of PRZ and the lineage leading to domestic horses of more than 100,000 years (Steiner et al. 2013; Goto et al. 2011; but see Der Sarkissian et al. 2015 for an alternative perspective), the data suggest that conservativeness may have characterized premolar shape for almost 99% of the history of *Equus caballus*, *whereas* in the last 1000-2000 years, under conditions of strong selective breeding, shape differences have more than doubled compared to those observed in archaic horses (Fig. 3). Thus, we propose, for premolar shape, an hypothesis of a 'long-fuse' model of phenotypic change in domestication, whereby a long initial period of small variation was followed by an explosive acceleration in the magnitude of shape change.

>> Fig. 3. Between group principal component analysis of shape (BG-PCA). Scatterplots of the first two components (BG-PC1, 48.7%, and BG-PC2, 34.0%, of between group variance) with groups emphasized using convex hulls. Mean shapes of ICE, THB and

archaic samples are shown using rendering of contours and deformation grids with expansion factors, as in Viscosi & Cardini (2011) and Seetah et al. (2012).

>> Fig. 4. Summary of mean shape variation: (a) 3D scatterplot of the first three principal components (91.0% of total variance) with 95% confidence ellipsoids around means; (b) neighbour joining tree of mean shapes (using Procrustes distances and rooted at PRZ) with percentages of support for branches and arrows to emphasize pairwise comparisons significant after a Bonferroni correction.

We anticipated large phenotypic changes in modern breeds consequential to artificial selection, although not specifically targeting tooth morphology. Differences in premolars are in fact small in size but remarkably large in shape when ICE and THB are compared (Fig. 3). Although of smaller magnitude, a degree of variation in premolar size across all groups was also predictable across all samples based simply on variability in body mass (e.g., the small PRZ and ICE vs. the medium-large THB), geographical distances, time heterogeneity and likely genetic differences. These factors should, however, also affect shape. In contrast, we found a clear disconnection between patterns of size and shape variation: while size varies over time and space, showing no clear trend except a moderate degree of reduction in most domesticated samples compared to wild horses, the amount of shape change of archaic horses in almost 100,000 years is overall modest and especially small if compared to the massive differences shown by modern breeds. This relative phenotypic conservativeness in shape, despite size changes, is unexpected and potentially might provide insight into both ecological and anthropological factors.

Dramatic ecological changes occurred after the end of the last glaciation *c.*11,500 years ago. Over the next 2000-3000 years, during the Early Holocene, temperature rose, forests expanded and open habitats favoured by horses shrank (Warmuth et al. 2011). Humans hunted horses and, with the expansion of agriculture in the mid Holocene, *c*. 8000-5000 years ago, contributed to modifying the environment. Regardless of when and how domestication occurred, horses had to cope with a variety of selective pressures, whose combined effect was profound for the wild counterpart, resulting in the extinction of those populations. Further, mtDNA diversity was greater before *c*. 2800 years ago rather than after (Lippold et al. 2011). Thus, as we anticipated, both environmental and genetic factors suggest that change would be expected, an expectation that is in agreement with our findings regarding size but at odds with those for shape. In fact, as premolar size varies in several archaic samples, allometry on its own, with its typically pervasive effect on bone shape (Klingenberg 1998), would predict concomitant shape differences, which we have not found.

Large shape changes seem to appear suddenly (in geological terms) in populations dated to after the first millennium BCE. This might be consistent with a *refining* phase during domestication that profoundly impacted on breeding practices because of important socio-economic developments in horse husbandry and greater management of individual populations. The almost complete extinction of wild horses in much of Eurasia may have been approximately concomitant with this stage: a reduction in wild populations, resulting from increased habitat encroachment by humans and a warming climate, potentially catalysed a reassessment of the symbolic and economic value of the domesticated form of the horse. Technological developments may also have played a role in promoting change, for example, with the development of the metal bit (Bendry 2011). In this complex scenario of cultural innovation and environmental change, highly distinct modern morphological types emerged with a major contribution, according to genetic evidence, likely due to an independent second wave of domestication in Western Europe (Lindgren et al. 2004; Achilli et al. 2012; Cieslak et al. 2010).

Conclusion

Domestic animals offer some of the best examples of how strong directional selection can change morphology (Drake & Klingenberg 2010). Our results support the notion that domestication is a highly dynamic process (Larson et al. 2014; Marshall et al. 2014), with morphological changes occurring over an extended period of time (Dobney & Larson 2006). However, the amount of change can vary remarkably over time and more importantly may vary sharply depending on the trait or even just the aspect of the character being studied, as shown by our tooth size and shape data. An 'explosion' of morphological disparity is expected as artificial selection became more intense in the last few centuries Interestingly, however, we also show that, if horses were first domesticated around 6000 years ago (Outram et al. 2009), the shape of horse teeth then likely remained unchanged for over half of the subsequent history of the *domestic* horse, until transformations that took place over the past 2000-3000 years. To our knowledge, this has never been reported and suggests a new layer of complexity, in which shape variation is decoupled from size, in the study of human-animal interactions. These outcomes reinforce the singular nature of horse domestication *per se*, and support the notion that the special relationship between horses and humans has existed for some considerable time. Within a well-established and powerful statistical framework (Adams et al. 2013), our work provides a testable hypothesis, which we term the 'long-fuse' model of phenotypic change in domestication, and which can be verified in new samples and on other anatomical structures, and different species.

References

Adams D., Rohlf F. & Slice D. A field comes of age: geometric morphometrics in the 21st century. *Hystrix, the Italian J. of Mammalogy* **24**(1), 7-15 (2013).

Achilli A. *et al*. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl Acad. Sci. USA* **109**(7), 2449–2454 doi:10.1073/pnas.1111637109 (2012).

Andersson L. S. *et al*. Mutations in *DMRT3* affect locomotion in horses and spinal circuit function in mice. *Nature* **488**, 642-646 (2012).

Anthony D. W. *The Horse, the Wheel and Language*. Princeton University Press. NJ (2007).

Bendrey R. Identification of metal residues associated with bit-use on prehistoric horse teeth by scanning electron microscopy with energy dispersive X-ray microanalysis. *J. of Arch. Sci.* **38**, 2989-2994 (2011).

Bendrey R. From wild horses to domestic horses: a European perspective. *World Arch.* **44**(1), 135–157. doi:10.1080/00438243.2012.647571 (2012).

Bignon O. & Eisenmann V. in *Equids in Time and Space* (ed Mashkour, M.) (Oxbow Books, pp.161-171, 2002)

Cardini A. in *Encyclopedia of Life Support Systems* (eds UNESCO-EOLSS Joint Committee) (EOLSS 2013).

Cardini A. & Elton S. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. *Bio. J. of the Linnean Society* **93**(4), 813–834. doi:10.1111/j.1095-8312.2008.01011.x (2008).

Cieslak M. *et al*. Origin and history of mitochondrial DNA lineages in domestic horses.

PLoS One **5**(12), e15311 (2010).

Clutton-Brock J. *A Natural History of Domesticated Mammals*. Cambridge University Press, (1999).

Cuozzo F. P., Sauther M.L. What is dental ecology? *Am. J. Phys. Anthropol*. 148(2): 163–170. doi:10.1002/ajpa.21656 (2012).

Der Sarkissian C. *et al*. Mitochondrial genomes reveal the extinct Hippidion as an outgroup to all living equids. *Biol. Lett*. 11: 20141058. http://dx.doi.org/10.1098/rsbl.2014.1058.

Dobney K. & Larson G. Genetics and animal domestication: new windows on an elusive process. *J. of Zool.* **269**(2), 261–271. doi:10.1111/j.1469-7998.2006.00042.x (2006).

Drake A. G., & Klingenberg C. P. Large scale diversification of skull shape in domestic dogs: disparity and modularity. *The American Naturalist* **175**(3), 289–301 (2010).

Dryden I. L. shapes: Statistical shape analysis (Version 1.1-9). Retrieved from http://cran.r-project.org/web/packages/shapes/index.html (2013).

Eisenmann V. & Baylac M. Extant and fossil Equus (Mammalia, Perissodactyla) skulls: a morphometric definition of the subgenus Equus. *Zoologica Scripta* **29**(2), 89–100. doi:10.1046/j.1463-6409.2000.00034.x (2000).

Evin A. *et al*. The long and winding road: identifying pig domestication through molar size and shape. *J. of Arch. Sci*. doi:10.1016/j.jas.2012.08.005 (2013)

Evans A. R. Shape descriptors as ecometrics in dental ecology. *Hystrix, the Italian J. of Mammalogy* **24**(1), 133-140. (2013).

Foote M. The evolution of morphological diversity. *Ann. Rev. of Ecol. and Systematics* **28**,129-152 (1997).

Gidley J. W. Tooth characters and revision of the North American species of the Genus *Equus*. *Bull. Am. Mus. Nat. Hist*. **14**, 91-141 (1901).

Goto H. *et al*. A massively parallel sequencing approach uncovers ancient origins and high genetic variability of endangered Przewalski's horses. *Genome Biol. & Evol*. **3**, 1096–1106. doi:10.1093/gbe/evr067 (2011).

Groves C. P. & Ryder O. A. in *The Genetics of the Horse* (eds Bowling, A.V. & Ruvinsky, A.) (CABI Publishing, pp. 1-24, 2000).

Hammer O. Harper D. & Ryan P. PAST: Paleontological statistics software package for education and data analysis. *Paleontol. Electron*. **4** (1): 1–9. (2001).

Houle D. Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nature Reviews / Genetics*. **11**, 855-866 (2010).

Kaiser, T. M. & Shulz. E. Tooth wear gradients in zebras as an environmental proxy - a pilot study. *Mitt. hamb. zool. Mus. Inst.* 103, S187-210.

Klingenberg C. P. Heterochrony and allometry: the analysis of evolutionary change in ontogeny. *Biological Reviews* **73**(1), 79–123. doi:10.1111/j.1469-185X.1997.tb00026.x

(1998).

Klingenberg C. P. MorphoJ: an integrated software package for geometric morphometrics. *Mol. Ecol. Res.* **11**(2), 353–357. doi:10.1111/j.1755-0998.2010.02924.x (2011).

Larson G *et al*. Current perspectives and the future of domestication studies. *Proc. Natl Acad. Sci. USA* **111**(17), 6139-6146 doi/10.1073/pnas.1323964111 (2014).

Lindgren G. *et al*. Limited number of patrilines in horse domestication. *Nature Genetics* **36**(4), 335–336. doi:10.1038/ng1326 (2004).

Lippold S. *et al*. Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. *BMC Evol. Biol.* **11**(1), 328. doi:10.1186/1471-2148-11-328 (2011).

Ludwig A. *et al*. Coat color variation at the beginning of horse domestication. *Science* **324**(5926), 485–485. doi:10.1126/science.1172750 (2009).

Marshall F. B. *et al*. Evaluating the roles of directed breeding and gene flow in animal domestication. *Proc. Natl Acad. Sci. USA* **111**(17), 6153-6158 doi/10.1073/pnas.1323964111 (2014).

Nagorsen D. W., & Cardini A. Tempo and mode of evolutionary divergence in modern and Holocene Vancouver Island marmots (Marmota vancouverensis) (Mammalia, Rodentia). *J. of Zoo. Syst. and Evol. Res.* **47**(3), 258–267. doi:10.1111/j.1439-0469.2008.00503.x (2009).

Olsen, S.L. in *Domestication: New Genetic and Archaeological Paradigms* (eds Zeder, M., Bradley, D.G., Emshwiller, E., Smith, B.D.) (University of California Press, pp.245-269**,** 2006).

Outram A. K *et al*. The earliest horse harnessing and milking. *Science* **323**, 1332-1335 (2009).

R Development Core Team. R: a language and environment for statistical computing. (http://www.r-project.org) (2005).

Rohlf F. J, & Slice, D. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Syst. Zoo.* **39**(1), 40–59. doi:10.2307/2992207 (1990).

Rohlf F. The tps series of software. *Hystrix*. **26**(1), 9-12 (2015).

Rohlf F. NTSYSpc: Numerical taxonomy system, ver. 2.1. *Setauket, New York: Exeter Publishing, Ltd* (2013).

Seetah T. K. Cardini A, & Miracle P. T. Can morphospace shed light on cave bear spatial-temporal variation? Population dynamics of *Ursus spelaeus* from Romualdova pećina and Vindija, (Croatia). *J. of Arch. Sci.* **39**(2), 500–510. doi:10.1016/j.jas.2011.10.005 (2012).

Seetah K. Cucchi, T., Dobney, K., & Barker, G. A geometric morphometric re-evaluation of the use of dental form to explore differences in horse (*Equus caballus*) populations and its potential zooarchaeological application. *J. of Arch. Sci.* **41**, 904–910 (2013).

Steiner C. C. Makova, K. D. & Ryder, O. A. Mitochondrial genome: Clues about the evolution of extant equids and genomic diversity of horse breeds. *Equine Genomics* 311-

321 (2013).

Viscosi V. & Cardini A. Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. *PLoS ONE* **6**(10), e25630. doi:10.1371/journal.pone.0025630 (2011).

Wade C. M. *et al*. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**(5954), 865–867. doi:10.1126/science.1178158 (2009).

Warmuth, V. *et al*. European domestic horses originated in two Holocene refugia. *PLoS ONE* **6**(3), e18194. doi:10.1371/journal.pone.0018194 (2011).

Warmuth V. *et al*. Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proc. Natl Acad. Sci. USA* **109**(21), 8202–8206. doi:10.1073/pnas.1111122109 (2012).

Conflicting interests

We have no conflicting interests.

Author contributions

KS & AC designed the study. GB oversaw the study. KS coordinated and collected samples, and performed initial landmarking. AC performed statistical analysis and validation. KS & AC wrote the article with essential input from GB. All authors gave final approval for publication.

Funding statement

Funding was provided to GB from the Leverhulme Trust project grant scheme (F/09 757/B) and to KS & AC from the Lang Fund for Human-Environmental Anthropology, Dept. of Anthropology, Stanford.

Acknowledgments

We are extremely grateful to the following for facilitating access to materials and providing essential input for specific samples: ICE and PRZ: F. Mayer (Museum für Naturkunde, Berlin); THB and PRZ: R. Sabin & L. Tomsett (Natural History Museum, London); CHI: T. Yaqi & H. Songmei (Shaanxi Archaeological Research Institute, Xi'an), KAZ: Z. Samashev & K. Kashkanbajev (Institute of Oriental Studies, Almaty), RUS: P. Kotsinsev

(Institute of Ecology of Plants and Animals, Yekaterinburg) & B. Hanks (Department of Archaeology, Pittsburgh); CRO: D. Brajkovic (Institute for Quaternary Palaeontology and Geology, Zagreb) & P. Miracle (Department of Archaeology, Cambridge) and HUN: A. Choyke (Department of Medieval Studies, Budapest) & A. Endrődi (Budapest History Museum, Budapest). We thank N. Vibla and J. Li for assistance in recovering samples and translation; A. Evin for providing a set of duplicate samples from Berlin and T. Cucchi and K. Dobney for essential analytical input. We are deeply grateful to I. Dryden (University of Nottingham), P. Gunz (Max Planck Institute for Evolutionary Anthropology, Leipzig), L. Monteiro (Universidade Estadual do Norte Fluminense), S. Schlager (University of Freiburg) and other MORPHMET subscribers for help and suggestions with R scripts and packages; to A. Drake for discussing with us methodological and conceptual issues about morphological disparity in dogs; to P. Mitteroecker (University of Vienna) for his important feedback on how to estimate multivariate variance; to P.D. Polly (Indiana University, Bloomington) for discussions on teeth evolution and shape analysis which greatly improved our interpretations. Finally, we are sincerely grateful to P.D. Polly and an anonymous referee for insightful comments on an early draft of this manuscript.