Evaluating demographic models for goat domestication using mtDNA sequences

Pascale GERBAULT
Research Department of Genetics, Evolution and Environment,
University College London, Gower Street, London WC1E 6BT (United Kingdom)
p.gerbault@ucl.ac.uk

Adam POWELL
Research Department of Genetics, Evolution and Environment
University College London, Gower Street, London WC1E 6BT (United Kingdom)

UCL Genetics Institute (UGI)
University College London, Gower Street, London WC1E 6BT (United Kingdom)

Mark G. THOMAS
Research Department of Genetics, Evolution and Environment
University College London, Gower Street, London WC1E 6BT (United Kingdom)

Department of Evolutionary Biology, Evolutionary Biology Centre
Uppsala University, Norbyvagen 18D, SE-752 36 Uppsala (Sweden)

Routes of migration and exchange are important factors in the debate about how the Neolithic transition spread into Europe. Studying the genetic diversity of livestock can help in tracing back some of these past events. Notably, domestic goat (*Capra hircus*) did not have any wild progenitors (*Capra aegagrus*) in Europe before their arrival from the Near East. Studies of mitochondrial DNA have shown that the diversity in European domesticated goats is a subset of that in the wild, underlining the ancestral relationship between both populations. Additionally, an ancient DNA study on Neolithic goat remains has indicated that a high level of genetic diversity was already present early in the Neolithic in northwestern Mediterranean sites. We used coalescent simulations and approximate Bayesian computation, conditioned on patterns of modern and ancient mitochondrial DNA diversity in domesticated and wild goats, to test a series of simplified models of the goat domestication process. Specifically, we ask if domestic goats descend from populations that were distinct prior to domestication. Although the models we present require further analyses, preliminary results indicate that wild and domestic goats are more likely to descend from a single ancestral wild population that was managed 11,500 years before present, and that serial founding events characterise the spread of *Capra hircus* into Europe.

Gerbault P., Powell A. & Thomas M.G.

RÉSUMÉ

Modèles d'évaluation démographique du processus de domestication des chèvres par l'ADN mitochondrial

Les routes migratoires et les voies d'échanges empruntées sont des facteurs importants dans le débat sur la diffusion du Néolithique en Europe. L'étude de la diversité génétique des animaux d'élevage contribue à la compréhension de ces évènements passés. Plus particulièrement, la chèvre domestique (*Capra hircus*) ne possédait pas d'ancêtres sauvages (*Capra aegagrus*) en Europe avant leur arrivée du Proche Orient. Les études génétiques de l'ADN mitochondrial ont montré que la diversité des chèvres domestiques européennes représentait un sous-ensemble de la diversité des chèvres sauvages, soulignant l'apparentement entre ces deux populations. De plus, les résultats d'une étude d'ADN ancien sur des restes fossiles de chèvres Néolithiques ont suggéré que la diversité génétique était élevée dès les débuts du Néolithique du nord-ouest méditerranéen. Afin de tester un série de simples modèles du processus de domestication des chèvres, nous avons utilisé une approche de coalescence ainsi qu'une analyse d'approximate Bayesian computation (ABC) conditionnées sur les patrons de diversité de l'ADN mitochondrial des chèvres domestiques et sauvages. Nous souhaitons plus particulièrement savoir si les chèvres domestiques descendent de populations qui étaient déjà distinctes avant leurs domestications. Bien que les modèles que nous presentont requièrent plus d'analyses, nos résultats préliminaires indiquent que les chèvres sauvages et domestiques descendent probablement d'une seule population ancestrale sauvage dont le contrôle a débuté il y a 11 500 ans, et qu'une série d'effets fondateurs caractérise la diffusion de *Capra hircus* en Europe.

INTRODUCTION

The arrival of the Neolithic in Europe was dependent on earlier developments in the Near East and Anatolia (Whittle & Cummings 2007), an observation well illustrated by the spread of domesticated goat (*Capra hircus*). Indeed *C. hircus* did not have any progenitors in Europe prior to the arrival of the Neolithic there. Instead its wild progenitors (*C. aegagrus*) are found on the Southern slopes of the Zagros and Taurus mountains, a region that also reveals the earliest archaeological evidence for goat domestication (Peters *et al.* 2005, Uerpmann 1996). This evidence dates back to 11,000-10,000 years before present (BP), after which goats appear to have spread further away from their origin location (Legge 1996, Peters *et al.* 2005, Zeder 2008). The number of centres in which *C. aegagrus* was domesticated is one of the main archaeozoological questions in the de-

bate on goat domesticate (Legge 1996, Peters *et al.* 1999, Peters *et al.* 2005, Uerpmann 1996; Zeder & Hesse 2000).

The spread of *C. hircus* also exemplifies how human movement and trade over the Neolithic have affected domesticated species (Bruford *et al.* 2003; Zeder *et al.* 2006). Demographic histories (e.g. bottleneck, founder effects, range expansions, admixture) shape patterns of genetic diversity (Currat & Excoffier 2005; Excoffier & Ray 2008; Excoffier & Schneider 1999; Goldstein & Chikhi 2002; Murray *et al.* 2010). It is thus expected that the expansion of *C. hircus* from the Near East into Europe has left signatures in contemporary genetic diversity. To understand how this demographic expansion associated with the spread of farming may have influenced patterns of genetic diversity in *C. hircus*, we have used published data on the

uniparentally-transmitted mitochondrial DNA (mtDNA). The advantages of using mtDNA in the study of recent genetic history is threefold: (i) its uniparental inheritance makes its genealogy relatively easy to model and estimate; (ii) its mutation rate (especially in the Hyper Variable Regions – HVR –) is high enough to assess the effect of potential population structures; and (iii) as many copies are present in every cell, it is relatively easier to obtain from ancient samples, in which DNA degradation might be considerable. Population genetic tools can be used to compare patterns of diversity based on the DNA sequences gathered. Different measures (hereafter referred to as 'summary statistics') can be obtained to describe the patterns of genetic diversity in the population samples studied. These summary statistics provide indirect information on the demographic parameters (*i.e.* ancestral population size(s), bottleneck size(s), population structures) that have shaped the observed patterns in genetic data.

One of the first studies on goat genetic diversity categorised the mtDNA sequences in *C. hircus* into four highly divergent lineages (Luikart *et al.* 2001). These lineages are arbitrarily defined for reasons of taxonomic convenience and are called haplogroups; sequences belonging to a same haplogroup share a number of key mutations. However, the relationship between the distribution of these haplogroups and the demographic processes that have shaped that distribution is not one-to-one, and not always straightforward to infer (Barbujani *et al.* 1998; Belle *et al.* 2006; Gerbault *et al.* 2011; Goldstein & Chikhi 2002; Nielsen & Beaumont 2009). Nonetheless, based on these four haplogroups, the authors proposed two alternative hypotheses to explain such high divergence: either *C. hircus* populations originated from four distinct domestication events or centres, *i.e.* one independent domestication for each mtDNA lineage observed; or such diversity already existed in the wild population (Luikart *et al.* 2001).

The authors considered the latter unlikely as they estimated that a large effective population size of about 38,000 to 82,000 females would have been necessary to maintain such lineage diversity. However, a few years later, an ancient DNA (aDNA) study showed that two of these haplogroups were already present in an ancient sample from southwestern France dating back to between 7,300 and 7,000 years calibrated (cal.) before present (BP), early in the Neolithic for that region (Fernandez *et al.* 2006). Furthermore, the genetic diversity of domestic goats has been shown to be a subset of the diversity observed in wild goats, even though different haplogroup frequencies are found within and amongst wild and domestic populations (Naderi *et al.* 2008; Naderi *et al.* 2007). Based on mtDNA sequence sharing between wild and domesticated populations, Naderi *et al.* (2008) inferred specific locations where domesticates could have originated. However, it should be noted that any observed gene genealogy is only one amongst a very large number of possible genealogies that can arise under the same demographic history (Nielsen and Beaumont 2009).

In order to assess whether the patterns of mtDNA diversity in modern goats are compatible with a domestication process involving populations that were already genetically (and possibly geographically) distinct 11,500 years BP (*i.e.* potentially distinct domestication centres), as opposed to serial founding events from a single domesticated stock, we have used a coalescent framework coupled with an approximate Bayesian computation approach. This methodology allows us to simulate expected genetic datasets under a range of demographic scenarios and then to compare them to the observed genetic data in order to identify those scenarios (hypotheses) where the simulated data looks most similar to the observed data (Bertorelle *et al.* 2010; Csillery *et al.* 2010; Bollongino *et al.* 2012). We used this methodology and observed genetic diversity patterns to compare a number

of different models, of differing complexity, to investigate goat domestication. We also present some preliminary estimates of ancestral population sizes.

MATERIAL AND METHODS

SAMPLES

We considered 130 base pairs of the mtDNA HVR, from 1904 published sequences (Fernandez *et al.* 2006; Naderi *et al.* 2008; Naderi *et al.* 2007). Note that 130 base pairs is the maximum length available for the 19 ancient samples used (Fernandez *et al.* 2006), although the modern sequences are somewhat longer. We grouped these sequences as follow: The European *C. hircus* population was represented by a sample of 1046 mtDNA sequences; The Middle-Eastern *C. hircus* population was represented by a sample of 368 mtDNA sequences; The *C. aegagrus* population was represented by a sample of 471 sequences (Naderi *et al.* 2008; Naderi *et al.* 2007). The remaining 19 sequences come from an ancient DNA sample from southwestern France, dated from 7,300-7,000 years cal. BP (Fernandez *et al.* 2006). When using the term "Middle East" for both *C. hircus* and *C. aegagrus* samples, we refer to samples coming from five countries, namely Azerbaijan, Dagestan, Iran, Pakistan and Turkey. We restricted our Middle-Eastern *C. hircus* samples to the same geographic region as *C. aegagrus* samples for simplicity of comparisons.

SUMMARY STATISTICS

Summary statistics describing patterns of genetic diversity within and between samples were computed using the software ARLEQUIN v. 3.1 (Excoffier *et al.* 2005). Five summary statistics (number of haplotypes, number of segregating sites, haplotype diversity, Tajima's D and average number of pairwise differences) were calculated for each sample individ-

ually, and four summary statistics (number of private haplotypes to one sample, number of private haplotypes to the other sample, average number of pairwise differences and pairwise Fst) were calculated for pairwise comparisons of population samples. In total, 44 summary statistics were used to describe and compare the genetic diversity within and between the four goat samples (Table 1).

COALESCENT SIMULATIONS AND DEMOGRAPHIC MODELS SIMULATED

The coalescent is a retrospective model of gene genealogies for a sample under a defined population history (Wakeley 2008). At each generation going back in time, samples are randomly attributed a parent in the population, and whenever a pair are attributed to the same parent, they coalesce into a single lineage. All the lineages eventually coalesce into one, called the most recent common ancestor (MRCA) of the gene copies in the sample considered. This marks the end of the coalescent process. Once a gene genealogy has been simulated and the MRCA found, mutations can be added onto the genealogy. One important simplifying assumption in modelling mutations within the coalescent is that all variation is selectively neutral (Wakeley 2008). Because selectively neutral mutations do not affect reproduction, they can be placed randomly on branches of the genealogy afterwards. A Poisson process 'sprinkles' the mutations forward in time along branches of the simulated genealogy to produce simulated genetic data at the tips (i.e. the samples).

Coalescent simulation allows the examination of a range of different hypotheses about demographic history. First, genealogies are simulated for a given demographic model, with its associated parameter values. Once the genealogies have been created, simulated genetic data are generated through the genealogy. Finally, summary statistics describing patterns of genetic diversity in the simulated data under that model are calculated. We performed these simulations with

Summary statistics	C. aegagrus (sample 1)	Middle-Eastern C. hircus (sample 2)	European C. hircus (sample 3)	European ancient DNA (sample 4)
Number of different sequences (haplotypes)	147	205	324	4
Number of segregating sites	61	54	59	25
Number of pairwise differences	14.63	5.59	5.63	11.22
Haplotype diversity	0.98	0.99	0.98	0.58
Tajima's D	1.8	-0.94	-0.75	2.23
Number of private haplotypes (per pairwise comparisons)				
Sample 1		140	141	147
Sample 2	198		169	204
Sample 3	318	288		321
Sample 4	4	3	1	
Pairwise differences (above diagonal)/ BSSC Pairwise F_{st} (below diagonal)				
Sample 1		16.63	16.43	15.29
Sample 2	0.39		5.81	15.63
Sample 3	0.38	0.03		15.32
Sample 4	0.16	0.46	0.55	

Table 1.– Values and names of the 44 summary statistics used to describe the genetic diversity between and within the four samples: *C. aegagrus* (sample 1); Middle Eastern *C. hircus* (sample 2); European *C. hircus* (sample 3) and ancient DNA sample (sample 4).

the software Bayesian Serial Simcoal (Anderson *et al.* 2005; Excoffier *et al.* 2000), as it allows the incorporation of both modern and ancient DNA samples (serial coalescent). We defined five demographic models to be simulated using Bayesian Serial SimCoal (BSSC) software, and performed 10,000,000 simulations for each model.

Figure 1 depicts the five models as well as the corresponding uniform prior parameter ranges that were considered when little information was available. These models can be subdivided into three major groups, named according to the number of populations the samples are drawn from. The four samples can either be drawn from one single expanding population (model 1); from two genetically distinct populations (models 2a and 2b); or from three genetically distinct populations (models 3a and 3b). All five models assume the wild sample (sample 1) descends from a wild population that has

been affected by anthropogenic degradation of its niche prior to domestication and/or by early management in the wild (Legge 1996; Peters *et al.* 2005; Uerpmann 1996), some 11,500 years BP (Peters *et al.* 2005; Zeder 2008).

Model 1 hypothesises that contemporaneous domestic and wild goat populations behave as a single, random mating meta-population that was affected by anthropogenic factors 11,500 years BP (Uerpmann 1996). It is equivalent to a model of population expansion without any population structuring. Models 2a, 2b, 3a and 3b, consider the *C. aegagrus* population as distinct from the *C. hircus* population(s). However, bidirectional migration between the wild and domesticate population is allowed in models 2a and 2b, as it is between the wild and Middle Eastern domestic populations, and between the two domestic populations in models 3a and 3b. Models 2a and 2b assume that there is no genetic
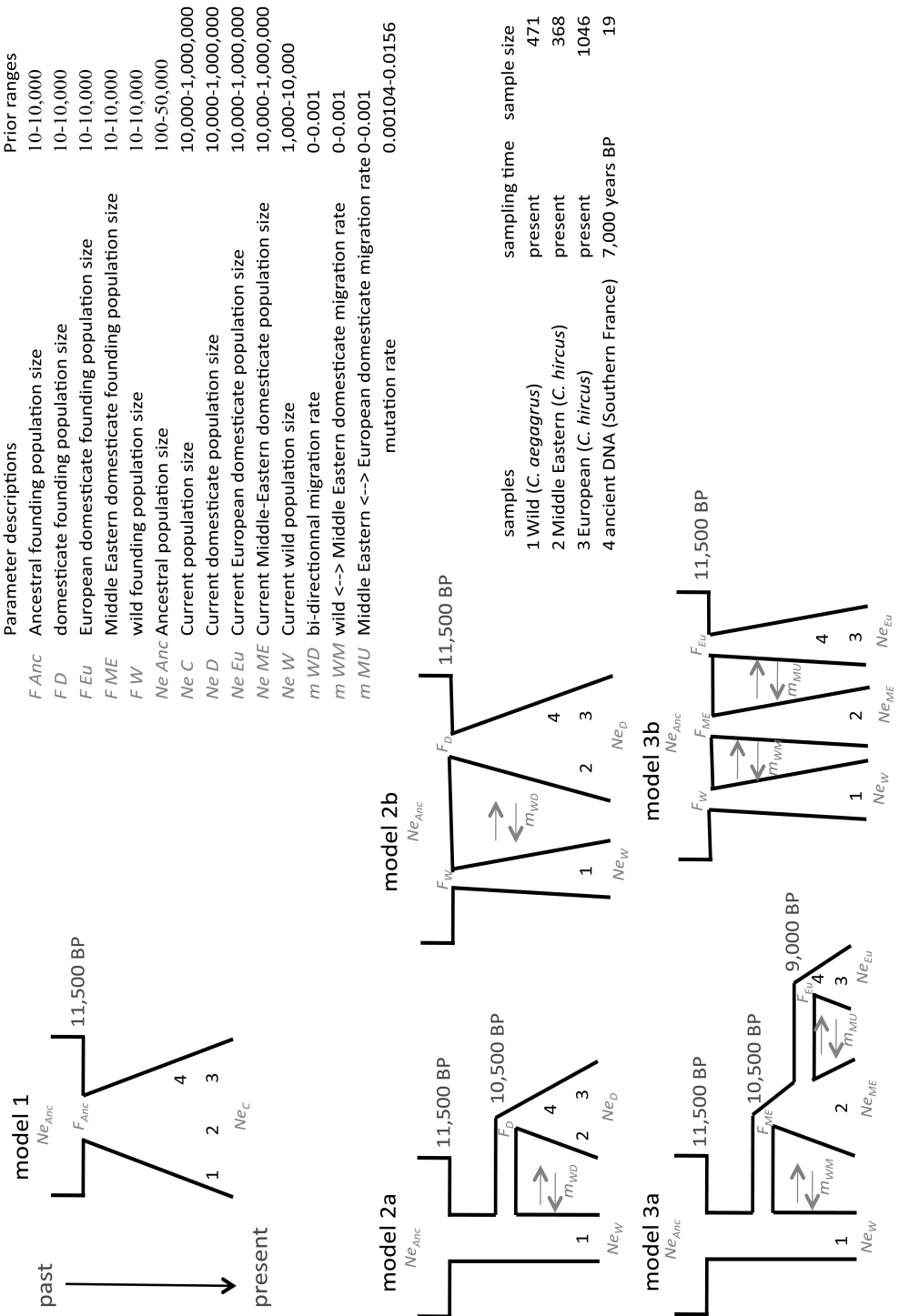
Parameter descriptions		Prior ranges
F_{Anc}	Ancestral founding population size	10-10,000
F_D	domesticate founding population size	10-10,000
F_{Eu}	European domesticate founding population size	10-10,000
F_{ME}	Middle Eastern domesticate founding population size	10-10,000
F_W	wild founding population size	10-10,000
Ne_{Anc}	Ancestral population size	100-50,000
Ne_C	Current population size	10,000-1,000,000
Ne_D	Current domesticate population size	10,000-1,000,000
Ne_{Eu}	Current European domesticate population size	10,000-1,000,000
Ne_{ME}	Current Middle-Eastern domesticate population size	10,000-1,000,000
Ne_W	Current wild population size	1,000-10,000
m_{WD}	bi-directionnal migration rate	0-0.001
m_{WM}	wild <--> Middle Eastern domesticate migration rate	0-0.001
m_{MU}	Middle Eastern <--> European domesticate migration rate	0-0.001
	mutation rate	0.00104-0.0156

samples	sampling time	sample size
1 Wild (*C. aegagrus*)	present	471
2 Middle Eastern (*C. hircus*)	present	368
3 European (*C. hircus*)	present	1046
4 ancient DNA (Southern France)	7,000 years BP	19

Fig. 1.— Schematic representation of the demographic models of the goat domestication process considered. Note that in coalescent simulation, time goes backward (arrows on the left). Parallel lines means the population size remains constant through time, while non-parallel lines are used for expanding populations. Letters within the models highlight the unknown parameters for which ranges of possible values were used. The parameter names and prior ranges used in the simulation are detailed in the upper right table. The sample names, sizes and their respective sampling times are detailed in the lower right table..

Evaluating demographic models for goat domestication using mtDNA sequences

structure within domestic goats, and all three domesticate samples are drawn from a single growing population. In models 3a and 3b population structure between Middle Eastern and European domestic goats is assumed, and the three domesticate samples are drawn from these two distinct populations (two from European domestic goats and one from Middle Eastern domestic goats). The additional population and parameters in models 3 (a and b), when compared to models 2 (a and b), are included to examine the role of continental-scale geographic structuring of goat populations on patterns of genetic diversity.

There are some similarities between 'a' models (2a and 3a) on the one hand, and 'b' models (2b and 3b) on the other. Both 'a' models assume that the *C. aegagrus* sample is drawn from a population that has remained constant in size after the beginnings of the domestication process, some 11,500 years BP (Peters *et al.* 2005; Uerpmann 1996; Zeder 2008). The rationale underlying this assumption is that as human settlement expanded with the intensification of agriculture, *C. aegagrus* underwent a geographical range restriction and consequent population reduction. Both 'a' models also assume the C. *hircus* population(s) descend from an ancestral wild population that had already been reduced for 1,000 years, prior to their domestication, as suggested by archaeozoological data (Legge 1996; Peters *et al.* 2005). On the other hand both 'b' models assume that the modern *C. aegagrus* population arose as an event distinct from those that gave rise to the domesticated population(s) 11,500 years BP; and has been expanding since then. In other words, the 'b' models would presuppose that (i) human developments associated with the Neolithic would have affected *C. aegagrus* genetic diversity less than in the 'a' models; and (ii) that the modern C. *hircus* populations descend from an ancestral wild population via a domestication bottleneck (in the case of model 2b), or two domestication bottlenecks (in the case of model 3b), that are distinct from the management bottleneck that led to modern *C. aegagrus* (all bottlenecks occurring 11,500 BP). In the 'a' models *C. hircus* population(s) descend from serial founder event(s) from a previously managed ancestral wild population. Models 2a and 3a nonetheless differ in the way that *C. hircus* populations are structured. Model 2a considers that both Middle-Eastern and European *C. hircus* samples are drawn from a single expanding population, founded 10,500 years BP, at a time corresponding to the median date for the earliest archaeozoological evidence for domestic goats (Peters *et al.* 2005), while model 3a assumes two distinct expanding populations of *C. hircus*, with European *C. hircus* founded from Middle-Eastern *C. hircus* 9,000 years BP. Model 3a therefore conforms to the hypothesis that sequential founder effects have shaped the genetic diversity of these two *C. hircus* populations differently, and that Middle-Eastern *C. hircus* is older than European *C. hircus*.

Testing the fit to the observed data

To compare models and present estimates of key parameters of those models, we have used an approach belonging to a family of techniques called approximate Bayesian computation (ABC). ABC techniques have been widely used in population genetic studies (Beaumont *et al.* 2002; Fagundes *et al.* 2007; Wegmann *et al.* 2009). These techniques involve the retention of simulated datasets that are 'closest' to some observed data in terms of a single, or set of summary statistic values (i.e. the 'best-fitting'). In our case we used a set of summary statistics (see above) and normalised each statistic by its mean and standard deviation over all 50,000,000 coalescent simulations (Fagundes *et al.* 2007). These normalized summary statistics were then used to calculate a Euclidean distance between the outcome of each coalescent simulation and the same statistics for the observed data. The simulated datasets were then ordered by their normalised Euclidean distance from the observed data. Finally we examined what proportion of the simulations giving
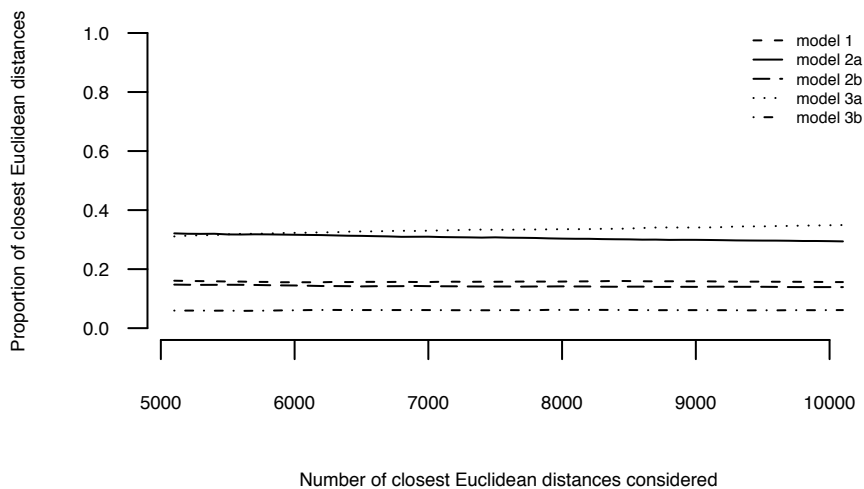
ANTHROPOZOOLOGICA • 2012 • 47.2.

69

Gerbault P., Powell A. & Thomas M.G.

Fig. 2.– Marginal posterior probabilities of each model tested. Lines represent the proportion of n (indicated on the x-axis) best simulations that come from each model that gave the smallest Euclidean distances over all the 50,000,000 simulations performed. These are analogous to Bayes factors, based on 10,000,000 simulations of each model, with a comparison based on the ranking of the closer simulated summary statistics to the observed ones.

the smallest Euclidean distances come from each model, to estimate the Bayes factor (Kass & Raftery 1995; Ray *et al.* 2009).

For the two best-supported models we also estimate key population size parameters. This was done by taking the 0.1% smallest Euclidean distances from each model and plotting the parameter values for those simulations. This procedure is known as rejection-only ABC.

RESULTS

Figure 2 shows the proportion of simulations from each model giving the 10,000 smallest Euclidean distances. This is analogous to the Bayes factor comparison extended to a case with more than two models (Kass and Raftery 1995; Ray *et al.* 2009). This process gives the approximate marginal posterior probability of each model (assuming that these are the only possible 5 models), and can be interpreted as the relative support from the data for each

model (Fagundes *et al.* 2007; Ray *et al.* 2009). Of the models tested the best fitting (across all parameter values considered) is model 3a followed by model 2a, while model 3b appears to be the worst fitting of the models tested. When the best fitting models have been determined, inferences of the unknown parameters (e.g. ancestral effective population sizes, modern effective population sizes, bottleneck or founder event sizes) can be made. We used the simulations that gave the 0.1% smallest Euclidean distances for models 2a and 3a in order to generate posterior estimates of the effective number of *C. aegagrus* females, and the effective number of founding *C. hircus* females (Fig. 3).
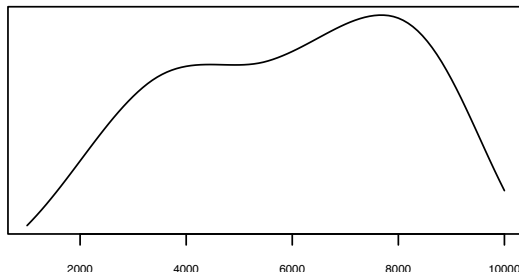
For model 3a the posterior mode for the long-term number of *C. aegagrus* females (Ne_W) is around 8000, while the posterior mode for model 2a is around 2000 (Fig. 3). However, these posterior distributions have wide and similar credibility intervals, and on closer in-
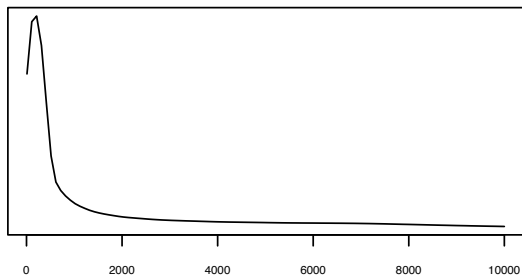
Model 2a
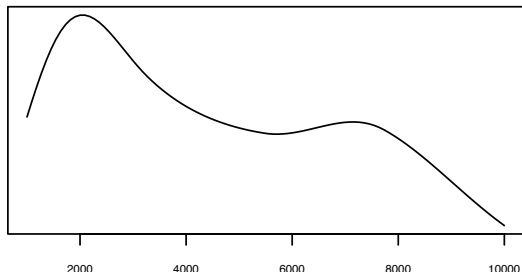
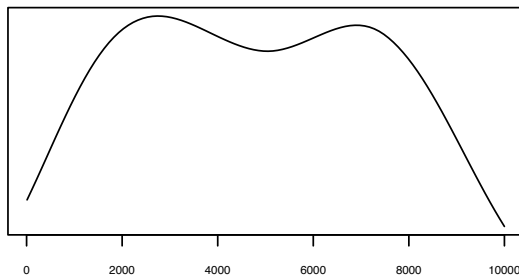

Ne$_W$: Number of effective *C.aegagrus* females

Model 3a

Ne$_W$: Number of effective *C.aegagrus* females

Model 2a

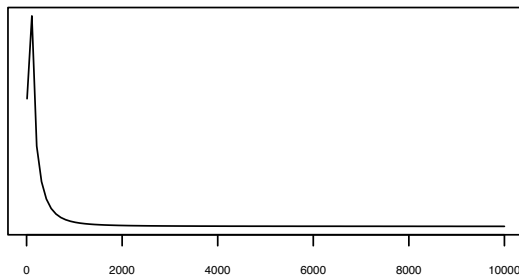*F$_D$*: Number of effective founding *C.hircus* females

Model 3a

F$_{ME}$: Number of effective founding Middle Eastern *C.hircus* females

Fig. 3.– Posterior density distributions of two and three parameters of models 2a and 3a, respectively, based on the 0.1% smallest Euclidean distances of those models. The two parameters of model 2a represented are the number of effective *C. aegagrus* females since domestication (Ne$_W$) and the number of effective *C. hircus* founding females (Ne$_D$) that were compatible with the genetic data observed. The three parameters of model 3a represented here are the number of effective *C. aegagrus* females since domestication (Ne$_W$), note that this is a common parameter of both models 2a and 3a), and the number of both effective founding Middle Eastern *C. hircus* females (Ne$_{ME}$) and effective founding European *C. hircus* females (Ne$_{Eu}$) that most were compatible with the genetic data observed. The x-axes plotted correspond to the prior parameter ranges as used in the simulations. Note that these posterior distributions are by no means definitive estimates but preliminary approximations (see text for further details and discussion on the subject).

Model 3a

F$_{Eu}$: Number of effective founding European *C.hircus* females

spection, do not differ greatly. Additionally, for model 3a the estimated effective number of domesticated females (F$_{Eu}$) required to explain European *C. hircus* genetic diversity is much smaller than the number required to explain Middle Eastern *C. hircus* genetic diversity (F$_{ME}$) (Fig. 3). The European domestic founding size

parameter (F$_{Eu}$) seems to have an important effect on the fit of the model to the data. This is based on the observation that model 2a gives a posterior estimate of the number of effective founding *C. hircus* females (both Middle Eastern and European, F$_D$) that is closer to the size estimated for the number of effective females

for European *C. hircus* (F_{Eu}) under model 3a than it is to the estimated number of effective founding Middle Eastern *C. hircus* females (F_{ME}) (Fig. 3). This observation suggests that the small founding population size estimates (F_D in model 2a, and F_{Eu} in model 3a) are being driven by the European sample, which is around 3 times as large as the Middle Eastern *C. hircus* sample. However, since model 2a provides a poorer fit to the observed data than model 3a (Fig. 2), it seems that Middle Eastern and European *C. hircus* should not be considered as a single population.

DISCUSSION AND CONCLUSION

The Neolithic transition entailed diverse changes in culture (e.g. structured societies), environments (e.g. farming-like landscapes) and genetic diversity of both domesticate species and human populations (Currat & Excoffier 2005; Francois *et al.* 2010; Whittle 2007). As such past evolutionary processes shape patterns of genetic variation in modern populations, modern genetic data, particularly when combined with aDNA data, can be used to detect the influence of the Neolithic transition on domesticates. Analyses of genetic diversity have attempted to address the relative role of founder events, local admixture with wild progenitors, and migrations, on their genetic diversity (Bollongino *et al.* 2008; Larson *et al.* 2007; Luikart *et al.* 2001). In this context, phylogeographic studies have attempted to identify founding female (using mtDNA) or male (using Y chromosomes) lineages. However, the arrival of a population into a geographical region does not necessarily equate to the ages of selected genealogical lineages observed in that region (Barbujani *et al.* 1998; Gerbault *et al.* 2011; Nielsen & Beaumont 2009).

The results of our analyses of *C. aegagrus* and *C. hircus* mtDNA data show that serial founder events from a previously managed popula-tion (model 2a and, in particular, model 3a) are better supported than *C. aegagrus* and *C. hircus* populations being formed by distinct founding events around 11,500 years BP (model 2b and 3b). This emphasizes that a certain level of population structure between wild and domestic populations is needed to explain the observed mtDNA diversity, but that this structure does not need to date back to the beginnings of domestication. In other words, as models 2b and 3b have been shown not to explain summary statistics of the observed data as well as models 2a and 3a, it seems that any structure is more likely to have arisen after 11,500 BP, probably by sequential founding events with subsequent gene flow, rather than domestication processes involving distinct populations (Luikart *et al.* 2001). These preliminary results seem to corroborate archaeozoologically supported scenarios for goat domestication (Legge 1996; Peters *et al.* 2005; Tresset & Vigne 2007). This does not exclude the possibility of distinct domestication founding events, but does indicate that if such distinct events occurred, then they did not cause structuring of goat genetic variation as reflected in the observed data. For example, if goats were domesticated multiple times in distinct locations, but from a panmitic wild population, then we would not expect this to be distinguishable from a single larger domestication event without very fine resolution aDNA data.

From the posterior distributions presented here (Fig. 3), these preliminary results also suggest that the founding population sizes of *C. aegagrus* and *C. hircus* were large, but may not have needed to be as large as previously suggested (Luikart *et al.* 2001). Even though the number of *C. aegagrus* samples in Naderi *et al.* (2008) is good for a wild animal species, it is possible that the regions sampled do not represent the original distribution of *C. aegagrus* populations during the Holocene (Uerpmann 1996). Furthermore, *C. aegagrus* may

also occur further south of the area sampled by Naderi *et al.* (2008). However, one of the strengths of the coalescent approach is that it only accounts for the sequences sampled, building distinct genealogies for those samples only. Therefore, if the genetic diversity carried by the *C. aegagrus* sample was too low to account for the diversity in the domesticate samples, we might expect models 2b and 3b (with populations distinct from 11,500 years BP) to provide a better fit to the data than those that involve serial founder effects (models 2a and 3a). However, to be more confident on this point, further demographic models involving the time of divergence between wild and domesticates should be formally tested and compared against those presented here. It should also be noted that, because of the advantage of the coalescent mentioned above, even though additional aDNA data would provide a higher resolution of the goat mtDNA genealogy (and therefore higher resolution of the demography) and enable exploration of more complex scenarios, it may not substantially change our inferences.

Nonetheless, this analysis of highly dimensional data sets by ABC relies on both the tolerance applied to the accepted Euclidean distance and the choice of the summary statistics used to describe the data. Ideally, the set of summary statistics would be 'sufficient' for the data, meaning that a substantial proportion of the information on the model parameters in the data is captured by the summary statistics. However, outside of (overly) simple demographic models there is no way to guarantee that a chosen set of statistics is sufficient. With 'insufficient' summary statistics the ABC model choice procedure may fail to recover the true population history, as convergence of the approximation to the true Bayes factor is not guaranteed (Robert *et al.* 2011). Several methods have been suggested to overcome these issues (Bertorelle *et al.* 2010; Csillery *et al.* 2010; Joyce & Marjoram 2008; Nunes & Balding 2010; Robert *et al.* 2011; Wegmann

et al. 2009; Wegmann *et al.* 2010) that were not used in the present analysis. Therefore, the posterior estimates of the population sizes provided here should be considered as a first order approximation and further analyses are required for these estimates and model comparisons to be more conclusive.

Finally, the best model among a set tested does not necessarily equate to the one reflecting what actually occurred. All useful models are necessarily reduced and any model that we could test would fail to capture the true complexity of a population's history. This is exemplified when studying structured populations (Goldstein and Chikhi 2002; Mona *et al.* 2010; Wilson *et al.* 2003). Furthermore, it is important to emphasize that several models can be consistent with both archaeological and genetic data, and thus efficient statistical tools should be used to compare alternatives. In that context, approximate Bayesian computation (ABC) methods have been very useful to contrast complex demographic histories of diverse populations (Fagundes *et al.* 2007; Ray *et al.* 2009). These methods make use of the simpler models as a baseline to compare more complex models against. The results presented here provide us with information that can be used to refine models and to sharpen intuitions on the goat domestication process. The next steps are twofold. First, the demographic parameters for which prior ranges were used can now be estimated using one of the statistical methods (Wegmann *et al.* 2009; Wegmann *et al.* 2010) that have been suggested to mitigate the methodological issues discussed above, particularly the number and choice of informative summary statistics. Then, further models can be envisaged in order to provide complementary information on the domestication process of *C. hircus*. This approach can potentially be extended to other domesticates species and highlights how both genetics and archaeology can work in synergy.

REFERENCES

ANDERSON C.N., RAMAKRISHNAN U., CHAN Y.L. & HADLY E.A. 2005. — Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics* 21(8): 1733-1734.

BARBUJANI G., BERTORELLE G. & CHIKHI L. 1998. — Evidence for Paleolithic and Neolithic gene flow in Europe. *Am. J. Hum. Genet.* 62(2): 488-492.

Beaumont M.A., Zhang W. & Balding D.J. 2002. — Approximate Bayesian computation in population genetics. *Genetics* 162(4): 2025-2035.

BELLE E.M., LANDRY P.A. & BARBUJANI G. 2006. — Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proc. Biol. Sci.* 273(1594): 1595-1602.

BERTORELLE G., BENAZZO A. & MONA S. 2010. — ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Eco*l. 19(13): 2609-2625.

BOLLONGINO R, BURGER J, POWELL A, MASHKOUR M, VIGNE J.-D., THOMAS MG 2012 . — Modern taurine cattle descended from small number of near-eastern founders. *Mol. Biol. Evol.* 29(9): 2101-2104.

BOLLONGINO R., ELSNER J., VIGNE J.-D. & BURGER J. 2008. — Y-SNPs do not indicate hybridisation between European aurochs and domestic cattle. *PLoS ONE* 3(10): e3418.

BRUFORD M.W., BRADLEY D.G. & LUIKART G. 2003. — DNA markers reveal the complexity of livestock domestication. *Nat. Rev. Genet.* 4(11): 900-910.

CSILLERY K., BLUM M.G., GAGGIOTTI O.E. & FRANCOIS O. 2010. — Approximate Bayesian Computation (ABC) in practice. *Trends. Ecol. Evol.* 25(7): 410-418.

CURRAT M. & EXCOFFIER L. 2005. — The effect of the Neolithic expansion on European molecular diversity. *Proc. Biol. Sci.* 272(1564): 679-688.

EXCOFFIER L., LAVAL G. & SCHNEIDER S. 2005. — Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1: 47-50.

EXCOFFIER L., NOVEMBRE J. & SCHNEIDER S. 2000. — SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J. Hered.* 91(6): 506-509.

EXCOFFIER L. & RAY N. 2008. — Surfing during population expansions promotes genetic revolutions and structuration. *Trends. Ecol. Evol.* 23(7): 347-351.

EXCOFFIER L. & SCHNEIDER S. 1999. — Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc. Natl. Acad. Sci. USA* 96(19): 10597-10602.

FAGUNDES N.J., RAY N., BEAUMONT M., NEUENSCHWANDER S., SALZANO F.M., BONATTO S.L. & EXCOFFIER L. 2007. — Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA* 104(45): 17614-17619.

FERNANDEZ H., HUGHES S., VIGNE J.-D., HELMER D., HODGINS G., MIQUEL C., HANNI C., LUIKART G. & TABERLET P. 2006. — Divergent mtDNA lineages of goats in an Early Neolithic site, far from the initial domestication areas. *Proc. Natl. Acad. Sci. USA* 103(42): 15375-15379.

FRANCOIS O., CURRAT M., RAY N., HAN E., EXCOFFIER L. & NOVEMBRE J. 2010. — Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture. *Mol. Biol. Evol.* 27(6): 1257-1268.

GERBAULT P., LIEBERT A., ITAN Y., POWELL A., CURRAT M., BURGER J., SWALLOW D.M. & THOMAS M.G. 2011. — Evolution of lactase persistence: an example of human niche construction. Philos. *Trans. R. Soc. Lond. B. Biol. Sci.* 366(1566): 863-877.

GOLDSTEIN D.B. & CHIKHI L. 2002. — Human migrations and population structure: what we know and why it matters. *Annu. Rev. Genomics Hum. Genet.* 3: 129-152.

JOYCE P. & MARJORAM P. 2008. — Approximately sufficient statistics and bayesian computation. *Stat. Appl. Genet. Mol. Biol.* 7(1): Article 26.

Kass R.E. & Raftery A.E. 1995. — Bayes Factor. *J. Am. Stat. Ass.* 90(430): 773-795.

Larson G., Albarella U., Dobney K., Rowley-Conwy P., Schibler J., Tresset A., Vigne J.-D., Edwards C.J., Schlumbaum A., Dinu A., Balacsescu A., Dolman G., Tagliacozzo A., Manaseryan N., Miracle P., Van Wijngaarden-Bakker L., Masseti M., Bradley D.G. & Cooper A. 2007. — Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc. Natl. Acad. Sci. USA* 104(39): 15276-15281.

Legge T. 1996. — The beginning of caprine domestication in Southwest Asia, in Harris D.R. (ed.), *The origins and spread of agriculture and pastoralism in Eurasia*. UCL Press Limited, London: 238-262.

Luikart G., Gielly L., Excoffier L., Vigne J.-D., Bouvet J. & Taberlet P. 2001. — Multiple maternal origins and weak phylogeographic structure in domestic goats. *Proc. Natl. Acad. Sci. USA* 98(10): 5927-5932.

Mona S., Catalano G., Lari M., Larson G., Boscato P., Casoli A., Sineo L., Di Patti C., Pecchioli E., Caramelli D. & Bertorelle G. 2010. — Population dynamic of the extinct European aurochs: genetic evidence of a north-south differentiation pattern and no evidence of post-glacial expansion. *BMC Evol. Biol.* 10: 83.

Murray C., Huerta-Sanchez E., Casey F. & Bradley D.G. 2010. — Cattle demographic history modelled from autosomal sequence variation. Philos. *Trans. R. Soc. Lond. B. Biol. Sci.* 365(1552): 2531-2539.

Naderi S., Rezaei H.R., Pompanon F., Blum M.G., Negrini R., Naghash H.R., Balkiz O., Mashkour M, Gaggiotti OE, Ajmone-Marsan P., Kence A., Vigne J.-D. & Taberlet P. 2008. — The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proc. Natl. Acad. Sci. USA* 105(46): 17659-17664.

Naderi S., Rezaei H., Taberlet P., Zundel S., Rafat S.A., Naghash H.R., el-Barody M.A., Ertugrul O. & Pompanon F. 2007. — Large-scale mitochondrial DNA analysis of the do-mestic goat reveals six haplogroups with high diversity. *PLoS ONE* 2(10): e1012.

Nielsen R. & Beaumont M.A. 2009. — Statistical inferences in phylogeography. *Mol. Ecol.* 18(6): 1034-1047.

Nunes M.A. & Balding D.J. 2010. — On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* 9(1): Article 34.

Peters J., Helmer D., Von Den Driech A. & Sana Segui M. 1999. — Early Animal Husbandry in the Northern Levant. *Paleorient* 25: 27-47.

Peters J., Von Den Driech A. & Helmer D. 2005. — The Upper Euphrates-Tigris Basin: cradle of agro-pastoralism?, in Vigne J.-D., Helmer D. & Peters J. (eds), *First steps of animal domestication. New archaeozoological approach*. Oxford, Oxbow: 96-124.

Ray N., Wegmann D., Fagundes N.J., Wang S., Ruiz-Linares A. & Excoffier L. 2009. — A statistical evaluation of models for the initial settlement of the american continent emphasizes the importance of gene flow with Asia. *Mol. Biol. Evol.* 27(2): 337-345.

Robert C.P., Cornuet J.M., Marin J.M. & Pillai N.S. 2011 . — Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl. Acad. Sci. USA* 108(37):15112-15117.

Tresset A. & Vigne J.-D. 2007. — Substitution of species, techniques and symbols at the Mesolithic-Neolithic transition in Western Europe, in Whittle A. & Cummings V. (eds), *Going Over: The Mesolithic-Neolithic Transition in the North-West Europe*. Oxford University Press, Oxford New York: 189-210.

Uerpmann H.P. 1996. — Animal domestication - accident or intention?, in Harris D.R. (ed.), *The origins and spread of agriculture and pastoralism in Eurasia*. UCL Press Limited, London: 227-237.

Wakeley J. 2008). — *Coalescent Theory: An Introduction*. Roberts & Compagny Publishers, Greenwood Village, Colorado.

Wegmann D., Leuenberger C. & Excoffier L. 2009. — Efficient approximate Bayesian com-

putation coupled with Markov chain Monte Carlo without likelihood. *Genetics* 182(4): 1207-1218.

Wegmann D., Leuenberger C, Neuenschwander S. & Excoffier L. 2010. — ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* 11: 116.

Whittle A. 2007. — Going over: people and their times, in Whittle A. & Cummings V. (eds), *Going Over: The Mesolithic-Neolithic Transition in the North-West Europe*. Oxford University Press, Oxford New York: 617-628.

Whittle A. & Cummings V. 2007. — *Going over: The Mesolithic-Neolithic transition in North-West Europe*. Oxford University Press, New York.

Wilson I.J., Weale M.E. & Balding D.J. 2003. — Inferences from DNA Data: Population Histories, Evolutionary Processes and Forensic Match Probabilities. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 166(2): 155-201.

Zeder M.A. 2008. — Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci. USA* 105(33): 11597-11604.

Zeder M.A. , Emshwiller E., Smith B.D. & Bradley D. 2006. — Documenting domestication: the intersection of genetics and archaeology. *Trends in Genetics* 22(3): 1-17.

Zeder M.A. & Hesse B. 2000. — The initial domestication of goats (*Capra hircus*) in the Zagros mountains 10,000 years ago. *Science* 287(5461): 2254-2257.