UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

From modern mitogenomes to archaeogenomics: exploring the past of human and animal populations

Marco Rosario Capodiferro

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXI – A.A. 2015-2018

Cover image has been created for this thesis by

Cristiano Corbucci

UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie “L. Spallanzani”

From modern mitogenomes to archaeogenomics: exploring the past of human and animal populations

Marco Rosario Capodiferro

Supervised by Proff. Alessandro Achilli and Ornella Semino

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXI – A.A. 2015-2018

Abstract

During my three-year Doctoral Fellowship, I was involved in various aspects in the field of population genetics, in line with the original theme of my PhD project: "Employment of genetic markers to study the variability of human and animal populations." I have explored the genetic variation of six different species from three continents, with a scientific role in all three major steps of a population genetics study: data production (4,953 mtDNA mammalian sequences deposited in GenBank), data elaboration and manuscript writing.

I was a first co-author in the paper (Wang et al., 2017) on the complete mitochondrial variation of the Asian swamp buffalo (*Bubalus bubalis carabanensis*). Even if this animal is one of the most important livestock species in several Asian countries and most people in the world depend on it for their livelihoods more than any other domestic species, its genetic history has been poorly investigated and only two complete mitochondrial DNA (mtDNA) sequences were published prior to this work. I obtained 107 complete novel mitogenomes using both Sanger sequencing and an original Next Generation Sequencing (NGS) protocol. A total of 109 swamp buffalo mitogenomes were eventually analyzed providing the first swamp phylogeny based on complete mitogenomes. In this tree, we were able to (i) define novel haplogroups and sub-haplogroups, (ii) date all the ancestral nodes with different approaches (Maximum Likelihood and Bayesian computations), (iii) estimate the variation of the effective population size through time.

The reconstructed genetic scenario shows that the ancestral swamp buffalo mitogenome could be dated to around 200 thousand years ago (kya) with a demographic history clearly linked to glacial events. Two major macro-lineages diverged during the 2nd Pleistocene Glacial Period (~200-130 kya), but most of the current matrilines (~99%) derive from only two ancestors that lived around the Last Glacial Maximum (LGM; ~26-19 kya). During the late Holocene Optimum (11-6 kya) lineages differentiated further and at least eight matrilines were domesticated around 7-3 kya. Haplotype distributions support an initial domestication process in Southeast Asia, while subsequent catching of wild females probably introduced four additional lineages, which are rare today.

Similarly, the current mitogenomes of domestic goats (*Capra hircus*) descend from a limited number of founding lineages that underwent domestication after surviving the LGM, but in western Eurasian refuges, as reported in another paper I was involved in at the very beginning of my PhD program (Colli et al., 2015). A different domestication pattern has been depicted by studying the horse mitochondrial variation with at least 17 matrilineal lineages that underwent domestication, probably throughout the Neolithic period, in multiple Eurasian locations. Certainly, the first phases of domestication and breeding processes have drastically influenced the genetic variability that is observed in modern breeds even at the micro-geographic level. In particular, the climatic and cultural diversity of the Italian Peninsula triggered, over time, the development of a great variety of horse breeds. I was part of the first comprehensive reassessment of the mitochondrial genetic relationships among the main native Italian hotblood/warmblood horses and ponies (*Equus caballus*) (Cardinali et al., 2016) and among the Podolian cattle breeds (*Bos taurus*) (Di Lorenzo et al., 2018). These papers also highlighted the importance of the mitogenome as a fundamental tool to reconstruct the maternal ancestral origins of local breeds, which represent unique and often endangered sources of genetic variability, particularly when confined to isolated geographical areas. This is the case of the Maltese breed of cattle that, despite its assumed ancient pre-historic origin, nowadays consists of only 12 males and 19 females. Therefore, in another study on livestock variability (Lancioni et al., 2016), we identified only two mtDNA and one Y-chromosome lineages in the current Maltese cattle that disproved the hypothesis of a local domestication, but concurrently confirmed a maternal influence from Northern Europe and a recent paternal introgression of the Chianina breed. Another human-related species, although not involved in the Neolithic farming revolution, is the tiger mosquito (*Aedes albopictus*), which is indigenous to East Asia and has colonized every continent except Antarctica in the last 40 years with major implications for human health. Consequently, I have also contributed to a phylogenetic analysis of 25 novel tiger mosquito mitogenomes that, together with two published ones, revealed a high level of sequence differentiation clarifying some aspects of its rapid spread. For instance, a peculiar lineage now common in Italy, most likely arose in North America from an ancestral Japanese source (Battaglia et al., 2016).

This work bears witness to the recent expansion of population genomics to the analysis of different species, but it also mirrors recent improvements of technological, statistical, and bioinformatics tools that are characterizing the entire field. A paradigmatic example of this "technical evolution" is the sub-field of archaeogenetics (the study of human past using molecular genetics techniques) whose analyses were initially restricted to mtDNA and mostly to samples derived from living populations. Most recently, major technical problems in ancient DNA production have been resolved and the "archaeogenomics era" has begun with a large number of ancient genomes from sites all over the world that are continuously retrieved, analyzed and published. Likewise, during my research work on humans

(*Homo sapiens*), I switched from the analysis of modern mitogenomes, which were the main objective of the Sardinia study (Olivieri et al., 2017) indicating a Mesolithic human presence on the island, to the analysis of ancient entire genomes from the tropical region of Panama (Capodiferro et al., in preparation). In fact, by applying the most updated techniques to extract and sequence aDNA from the petrous bone (a "gold mine" in archaeogenomics), I was able to obtain the first nine low coverage genomes from ancient remains excavated in Panama City and dated to pre-Columbian time, from 1.4 to 0.5 kya. Initially, I analyzed the nine ancient mitogenomes that were retrieved from the entire output and compared them to 162 novel mitogenomes from the five tribes (Ngäbe, Kuna, Emberà, BriBri, Naso) present nowadays in Panama in order to refine the female history of the Isthmus. This dataset of modern and ancient mtDNAs confirmed a mitochondrial legacy of the first Paleo-Americans that colonized the region, but also revealed an increase of the mtDNA variation in the last five thousand years with local specific sub-haplogroups linked to different tribes arising and spreading together with the development of agriculture. Then, we moved to the analysis of the autosomal DNA that allowed us to cluster our samples with modern Chibchan-speaking tribes living in the Isthmo-Colombian area. This cluster is characterized by two major ancestral components, a Pan-American one probably related to the first peopling of the continent and another specific to the Isthmo-Colombian area, that underlines the peculiarities of the region.

Acknowledgements

I would like to thank prof. Alessandro Achilli. It has been a long road together. I would like to thank him for his aid and teaching, but also for have given me responsibilities. In these last years a special thanks for giving me the opportunity to make great experiences.

I would like to thank Prof. Ornella Semino for the support and all the population genetics group of Pavia, Prof.s Antonio Torroni, Anna Olivieri and Luca Ferretti, Dr.s Viola Grugni and Enza Battaglia, and to all the students who have taken turns in these 3 years.

A special, warm and affectionate thanks goes to the group of the “youngs”, Linda Ongaro, Alessandro Raveane, Nicola Rambaldi Migliore, Francesca Bastaroli, Stefania Brandini and Abir Hussein, to whom I am bound by a deep sense of friendship (and not only).

An equally affectionate gratitude goes to Prof. Hovirag Lancioni and Dr. Irene Cardinali who initiated me to this work and from which the friendship has continued for now 5 years.

Special thanks to the three research centers that welcomed me in my abroad experiences:

Thanks to Prof. Walter Parson and Dr. Martin Bodner and their fantastic group at Institute for Legal Medicine at the Medical University of Innsbruck, Austria.

Thanks to Prof. Ripan Malhi and his group in particular Dr. Hongjie Li of the Department of Anthropology & Animal Biology and the Institute for Genomic Biology (Carl R. Woese) at the University of Illinois of Urbana Champaign, Illinois, USA

Thanks to Dr.s Mait Metspalu, Freddi Scheib, Luca Pagani, Francesco Montinaro and all the others guys of the Estonian Biocentre at Tartu, Estonia.

Thanks to all the members of the “An ARTery of EMPIRE” project, in particular to Prof. Bethany Aram that gives me the opportunity to work with ancient samples from Panama.

Thanks to all co-authors of the 7 papers and collaborators of ongoing projects.

I would like to thank Dr. Ugo A. Perego for revising the English for part of this thesis.

Thanks to the Mobility Grant of the University of Pavia that founded me for the first two visiting periods.

These are professional thanks. For the personal ones I will wait to give them in private ways.

As for the master's degree a particular thought goes to my grandfather, who as then continues to follow me and help in everything.

Abbreviations

1KGP: 1000 Genomes project

AD: Anno Domini

aDNA: ancient DNA

AGM: Ancestral Goat Mitogenome

AMH: Anatomically modern human

Anc-A: Ancestry A

Anc-B: Ancestry B

Arg: Arginine

AWM: Ancestral Water-Buffalo Mitogenome

BAM: Binary Alignment Map

BCL: binary base call

BEAST: Bayesian evolutionary analysis of sampling trees

BSP: Bayesian skyline-plot

BWA: Burrows-Wheeler Aligner

EBC: Estonian Biocentre

GATK: Genome Analysis ToolKit

Gb: Giga bases

HG: Haplogroup

HGDP: Human Genome Diversity Project

HGP: Human Genome Project

HPC: high-performance computing

HTS: High Throughput sequencing

HVS: Hyper Variable Sequences

IBD: Identity by Descent

kbp: Kilo base pairs

ky: kilo years

kya: kilo years ago

LD: Linkage Disequilibrium

LGM: Last Glacial Maximum

MAF: Minor Allele Frequency

MCMC. Markov chain Monte-Carlo

ML: Maximum Likelihood

MP: Maximum Parsimony

MRCA: Most Recent Common Ancestor

MSY: Male Specific region of Y chromosome

mtDNA: Mitochondrial DNA

mya: million years ago

N_e: Effective population size

NEB: New England Biolab

NGS: Next Generation Sequencing

NJ: Neighbor joining

NRY: Non-Recombining region of Y chromosome

Numts: Nuclear sequences of mitochondrial origin

OXPHOS: Oxidative phosphorylation

PAML: Phylogenetic Analysis by Maximum Likelihood

PAR: Pseudo Autosomal Regions

PCA: Principal Component Analyses

PCR: Polymerase Chain Reaction

PCs: Principal Components

PGM: Personal Genome Machine

rCRS: revised Cambridge reference sequence

RFLP: restriction fragment length polymorphism,

ROS: Reactive Oxygen Species

SAM: Sequence Alignment Map

SGS: Second Generation Sequencing

SNP: Single Nucleotide Polymorphisms

SSH: Sardinian-Specific Haplogroup

Sub-hg: Sub-Haplogroup

TE: Tris-EDTA

UPGMA: Unweighted Pair Group Method with Arithmetic Mean

VFC: Variants Calling File

Contents

1. Introduction

1.1 Methods and tools for population genetics

Who are we? Where do we come from? These are questions that most people ask themselves throughout their entire life and scholars from different fields try to answer in their studies. Actually, different disciplines are interested in the study of evolution; in particular, those focused on humans, such as anthropology, linguistics, archaeology, demography, history, evolutionary medicine and others. Eventually, the final picture comes from a complementary and multidisciplinary approach that takes into account the information derived from different fields (Cavalli-Sforza and Feldman, 2003). In this scenario, the biological point of view is extremely important. Talking about biological evolution, there is a molecule that links every life form and all generations, from the early beginning to present day: the DeoxyriboNucleic Acid (DNA). Inside this molecule, in its diversity (or similarity) across individuals and species, we could find some answers regarding our past and our origins. In particular, population geneticists try to "ask these questions directly to the DNA" by studying how its sequence evolved through time and how it was inherited through generations. In the history of population genetics three main components marked the progress of the field (Sunnucks, 2000):

i) Efficient molecular genetic methodologies to examine numerous informative DNA sites on a large number of samples.

ii) Statistical models to understand the information hidden into DNA sequences.

iii) Bioinformatics tools and computing power to use the models and to analyze a large amount of data simultaneously.

1.1.1 Molecular genetic methodologies

Genetic variations are caused by many types of mutations. Nowadays, the most studied and useful molecular markers are Single Nucleotide Polymorphisms (SNPs), but also microsatellites are largely used, especially while studying Y-chromosomes. Early approaches to the study of genetic variation among populations were based on blood group antigen proteins ("*classical markers*") recognized with antibodies. The first paper was published in 1919 (Hirszfeld and Hirszfeld, 1919) and focused on the *ABO* gene. The number of informative sites was increased by the introduction of electrophoresis by Pauling and colleagues in 1949 (Pauling et al., 1949), but still limited to serological assays. The first book about allele (a variant form of a given gene) frequencies was published in 1954 (Mourant and Mad, 1954), while in 1967 Ceppellini (Ceppellini et al., 1967) and colleagues proposed the term "*Haplotype*" to define a multi-locus combination of alleles on a chromosome. Another step forward was marked by the identification of mutations in the restriction sites (Botstein et al., 1980) that increased the available number of markers that could highlight the diversification among populations.

An amazing revolution in all molecular biology occurred when Kary Mullis in 1987 developed the polymerase chain reaction (PCR) protocol (Mullis and Faloona, 1987), with the possibility of obtaining numerous copies of a specific DNA fragment, which increased the capacity for targeted DNA variation studies.

Another effect of the PCR approach was an advancement of sequencing methods that in the middle of the '90s became automated and was pushed by the Human Genome Project. The era of the High Throughput Sequencing (HTS) (Figure 1) started at the beginning of the new millennia with the Next Generation Sequencing (NGS) and led to the study of complete genomes.

These advancements produced an incredible yield of data that together with the reduced times and costs now make possible to sequence a whole human genome in a few days and with less than 1000$.

Nowadays the genomic screening takes into consideration the entire dataset of molecular variants and also how specific combinations of alleles within a chromosome (nuclear haplotypes) are maintained throughout generations, due to a high Linkage Disequilibrium (LD) (see below). The most recent applications in population genetics study how these haplotypes are inherited, in the so-called Haplotype Based Methods.

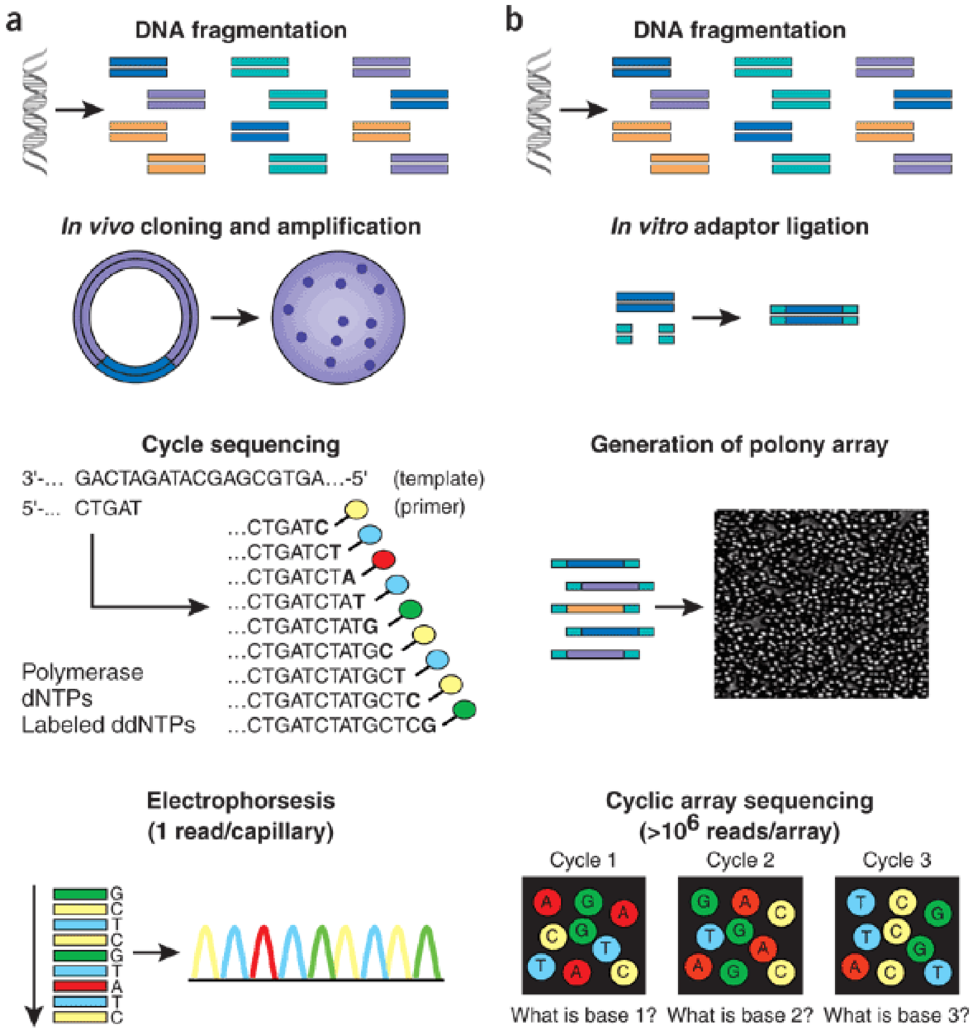


Figure 1. Differences between the traditional sequencing method (a) and the next generation sequencing approach (b) (Shendure and Ji, 2008).

1.1.2 Statistical models and bioinformatics tools

The sophistication of laboratory techniques that allows to obtain a large amount of data is accompanied by the need of new models and statistical approaches able to unveil the hidden information kept in molecular data and to define the forces that shaped them through time.

On the way to understanding our past, it was also clear that there are two main evolutionary forces (Cavalli-Sforza and Feldman, 2003):

i) Natural selection.

ii) Genetic drift.

The first is a non-random process that considers the probability of survival/reproduction of alleles/genotypes in a population, while the second is a random process that leads to a change in the gene pool of small populations due to a finite number of individuals participating in the formation of the next generation. Natural selection could be really important in understanding the evolution of specific genes or haplotypes, for example, the increased heterozygosity of the hemoglobin gene in Africa due to malaria resistance (Przeworski et al., 2000). The genetic drift is crucial to understand the main events that shaped the (neutral) genetic pool of populations, such as migrations, isolations, bottlenecks and founder effects.

The first statistical model that tried to explain and predict evolution due to genetic drift has been the Wright-Fisher (Fisher, 1923, Wright, 1931) model (Figure 2) that is still widely used. The model postulates that in a finite population the allele frequencies in the offspring derive from process of random sampling in the previous generation. They proposed an equation to predict how the allele frequencies change by chance through generations, thus explaining the allele fixation or loss in a population. The model was later updated in order to relate to real populations. Eventually, Kimura in 1968 demonstrated that genetic drift (and not the natural selection) is the main evolutionary force (Kimura, 1968).

Another important theory in population genetics is coalescence, introduced by Kingman in 1982 (Kingman, 1982), that suggests that all present alleles derive from a common ancestral molecule, named “Most Recent Common Ancestor” (MRCA). It is also possible to calculate the coalescence time, in terms of number of generations, by considering that the probability that two alleles coalesce in the previous generation is $1/(2N_e)$, while the probability that our pair of alleles fail to coalesce is $1-1/(2N_e)$, where N_e is the effective population size.

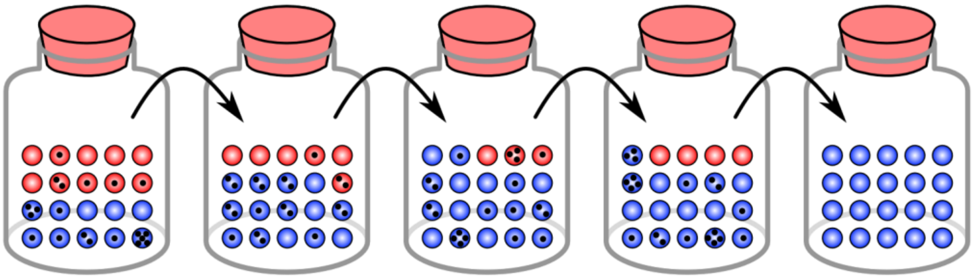


Figure 2. Example of the Wright-Fisher model: fixation of the blue allele in five generations (Wikipedia.com).

The mutation rate is the probability that a mutation occurs in a molecule or gene in each generation. Mutation rates differ among species and even among different regions of the same genome. These different rates are usually measured as “probability of a neutral substitution per site per generation”. The molecular clock is a useful method for estimating evolutionary timescales. It is based on the inference that DNA and protein sequences evolve at a rate that is relatively constant over time. Therefore, the number of differences between cells, individual organisms or groups of organisms is proportional to the time since they last shared a common ancestor (Ho, 2008). The calibration of a molecular clock requires a comparison between a phylogenetic tree (including sequences from different individuals belonging to the same or different species) and an outgroup whose divergence time is exactly known from other sources.

The molecular clock and coalescent theory are fundamental in population genetics to answer the question "when?". These two concepts (and their relative equations) allow us to date events that could be extrapolated from the data, such as migrations and the time to a MRCA.

The development of models and statistics able to identify and depict common histories amongst individuals or populations was one of the most important advancements in this field. Phylogenetic trees and principal component analyses (PCA) are probably the most used representations of genetic population data.

A phylogenetic tree is a diagram representation that allows to reconstruct the history of populations via the analysis of variations that cluster the samples in groups (taxa), which derived from a common ancestor. It is based on branches and internal nodes that show respectively the genetic distance among individuals or populations, in other words the degree of relationship, and the MRCA of each group. This technique is very important to draw the genetic relationships between populations based on uniparental markers, and to show those relationships without considering admixture events, due to the lack of recombination.

One of the first phylogenetic trees was proposed by Cavalli-Sforza and colleagues (Cavalli-Sforza et al., 1964) who reconstructed human evolution using very little information (Figure 3).

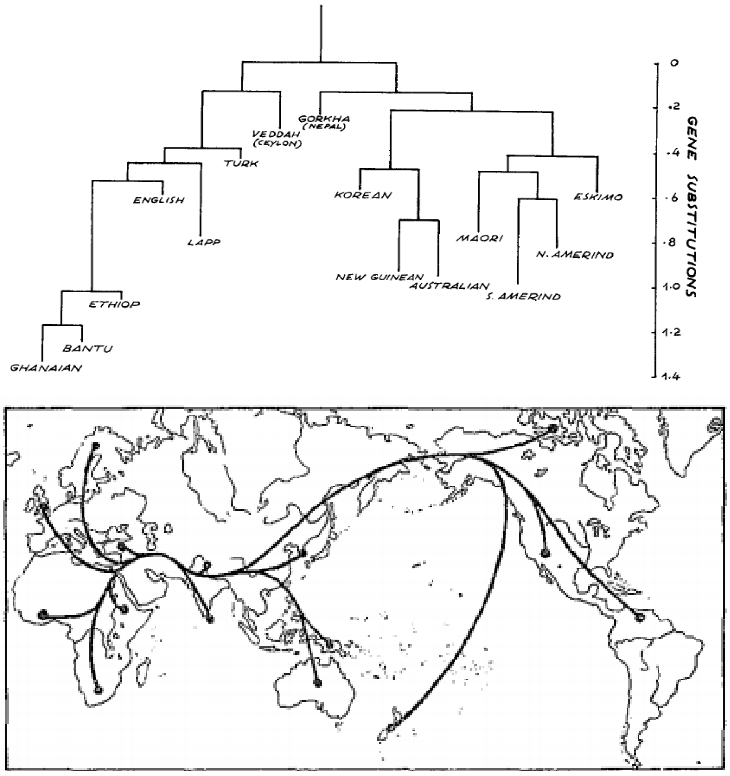


Figure 3. Representation of one of the first phylogenetic tree on a worldwide map (Cavalli-Sforza et al., 1964).

There are four different popular phylogenetic tree building algorithms (Table 1). Two are distance-based methods, using a pairwise distance matrix: i) the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) that generates a unique phylogenetic tree by the progressive combination of the two taxa with the lowest genetic distance between them; ii) the Neighbor Joining (NJ) that starts with the generation of the minimum evolution tree, an unresolved phylogeny which present the shortest sum of the branch lengths, all the posterior distance-based iterations lead to the progressive establishment of the branches reaching the final phylogeny. Two are the character-based methods: iii) the Maximum Parsimony (MP) that generates a tree which uses the smallest number of evolutionary changes; iv) the Maximum Likelihood (ML) that evaluates different tree topologies, but, assuming an evolutionary model, finds the tree explaining the data with the maximum likelihood (Sharma et al., 2018).

Table 1. Summary of the four algorithms for building of a phylogenetic tree. Modified from (Sharma et al., 2018).

Method	**Advantage**	**Disadvantage**
UPGMA	Fast	More assumptions
Neighbor joining	Fast	Unreliable due to loss of information
Maximum parsimony	Fast, robust	Performance is not satisfactory
Maximum likelihood	Phylogeny is clear	Slow

The study of the geographical distribution of the branches (clades or lineages) within a phylogeny is called phylogeography (Avise et al., 2000). This approach requires the combination of three elements: a phylogenetic tree, geographic distribution and coalescence time of lineages, especially those that are restricted to a particular area (Soares et al., 2010).

If the phylogeographic field tries to answer to the question "where?", it is also possible to answer to the question "when?", as described above, considering the molecular clock and calculating the MRCA of the internal nodes. There are three methods commonly used to calculate the age of the nodes: i) rho estimate or ρ statistics, is linearly related to mutation rate and time, assuming constancy of the rate across the tree branches and considering the average number of differing sites between a set of sequences and their MRCA; ii) ML analysis in which the average number is calculated considering a maximum likelihood statistics approach; iii) Bayesian analyses that are based on Markov chain Monte-Carlo (MCMC) simulations with prior estimates to generate the phylogeny with the highest posterior probability. The last two methods basically use two different statistical approaches: the first gives the probability of the data (such as the sequence alignment) given the hypothesis (the tree), the Bayesian approach determines the probability of the hypothesis (the tree) given the data (the sequence alignment). The Bayesian analysis can be also used to infer the timing of the most important demographic events that have shaped the gene pool of groups of individuals.

Phylogenetic trees are powerful tools to understand genetic relationships, but different trees could be obtained if we consider different loci, which might describe various genetic histories. PCA is a multivariate statistical method to summarize the molecular information considering multiple recombining loci (Figure 4).

This technique applies eigenvector decomposition to reduce high-dimensional data to small number of independent variables, the principal components (PCs), that explain a large portion of variation. This portion is usually reported in the plot as percentage of each component. Usually it is used to visualize genetic distances among individuals or populations plotting most informative PCs (Wangkumhang and Hellenthal, 2018) (Figure 5). One of the first comprehensive PCA was performed by Ammerman and Cavalli-Sforza (Ammerman and Cavalli-Sforza, 1984) that studied the Neolithic “revolution” in Europe using 94 alleles at 34 loci in 16 chromosomes.

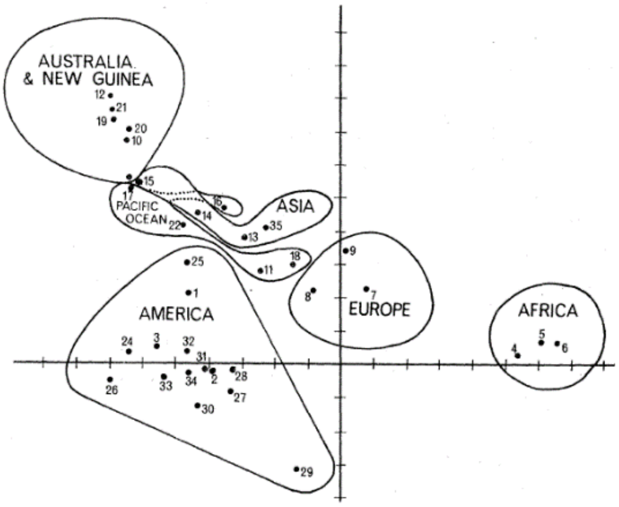
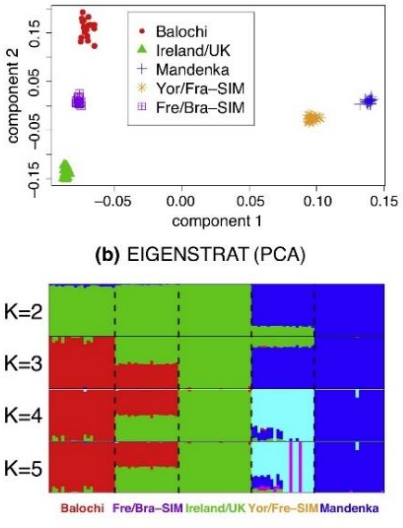


Figure 4. Representation of one of the first PCA based on human genetic data (Cavalli-Sforza et al., 1964).

Other algorithms were recently implemented in the genetic clustering methods, which can be used to identify ancestry-based groups or clusters able to explain the observed genetic variation. A leading method of analysis uses mathematical algorithms to establish the degree of genetic similarity between individuals and groups in order to infer population structures and assign individuals to hypothesized ancestral groups.

In recent years the most commonly used methods to identify clusters have been based on Bayesian clustering algorithms, such as STRUCTURE, initially introduced by Pritchard et al., (Pritchard et al., 2000) and more recently improved in the software ADMIXTURE (Alexander et al., 2009, Alexander and Lange, 2011). These algorithms have the peculiarity of grouping genetic samples into *K* ancestral populations (Figure 5c), sometimes called ancestries or components, on the basis of genetic affinities. Individuals belonging to the same group would be genetically more similar to each other than with samples of other components.



(b) EIGENSTRAT (PCA)

(c) ADMIXTURE

Figure 5. Examples of EIGENSTRAT PCA: (b) each dot represents an individual and (c) ADMIXTURE for cluster numbers K= 2–5 (columns = individuals/populations, colors = clusters). Modified from (Wangkumhang and Hellenthal, 2018).

Additional models have also been developed to analyze the increasing amount of available data. Nowadays, *f*- and D-statistics and Haplotype-Based Methods are used for the analysis of genomic data in order to answer more detailed and complex questions about our history (Pickrell and Reich, 2014, Schraiber and Akey, 2015).

In the *f*-statistics, introduced by Reich and colleagues (Reich et al., 2009), inferences are based on the so-called "shared genetic drift" between sets of populations, under the hypothesis that shared drift implies a shared evolutionary history. Starting from the Wright's fixation index (Wright, 1931), the *f*-statistics try to give an answer to a series of questions about admixture, population structure and genetic closeness (Peter, 2016). Under population phylogeny, the three *f*-statistics, labelled F2, F3, and F4 in Figure 6, have interpretations as branch lengths (or dissimilarities) (Figure 6A) between two, three, and four taxa/populations (P1, P2, P3 and P4), respectively (Px usually denote an admixed population); while in Figure 6B the population phylogeny is extended to account for gene flows (red line) and admixture events (dotted red line).

The *f*-statistics might have different interpretations, some of which discussed hereafter. The main purpose of *f*2 is simply to measure genetic dissimilarity between two populations (P1 and P2) or within the same population at different times (P0 and Pt). A large *f*2 value implies that populations are highly diverged. The main parameter is the genetic drift and its impact on the genetic diversity measured in terms of (i) heterozygosity, (ii) allele frequencies, (iii) covariance, (iv) identity by descent (IBD) probabilities, and (v) coalescence.

The *f*3 and *f*4/D-statistics, which consider respectively three and four populations, are more used to check for admixture events. A common application is the so-called “*outgroup” f*3 statistics, defined as *f*3(PO; Pu, Pi) that correspond respectively to *f3*(Px; P1, P2) and the yellow line in Figure 6A. Considering PO as an outgroup, the objective is to find among a panel of k extant populations (Pi with i = 1; 2,…, k;) which one could be the most closely related to an unknown population (Pu). This measures the shared drift (or “shared branch”) and high *f*3 (long shared branch) values imply close relatedness.

However, *f*3 is also used to identify admixture events. Considering the three population model in Figure 6 (Px; P1, P2) and *p*x, *p*1 and *p*2 as the allele frequency of a variant at a locus in the different populations, if the product of this multiplication (*p*x - *p*1) (*p*x – *p*2) is negative Px descends from an admixture event between P1 and P2 or their descendants, because *p*x falls between *p*1 and *p*2.

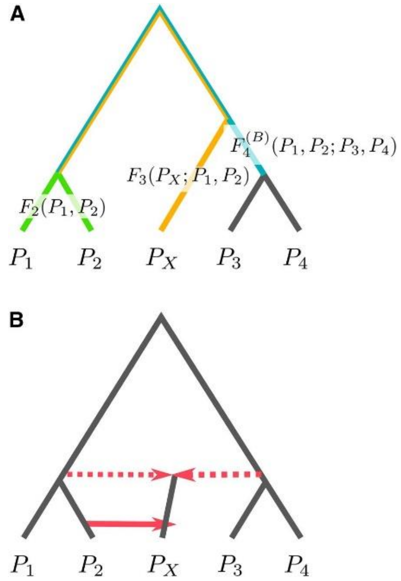


Figure 6. A) A population phylogeny with branches corresponding to F2 (green), F3 (yellow) and F4 (blue). B) A population phylogeny that includes gene flow (red line) and admixture events (dotted red lines). Modified from (Peter, 2016).

Therefore, in the simplest *f*3 admixture model, an ancestral population splits into different populations and then at time *t* the populations mix to form Px (Figure 6B). The power of *f*3 to detect admixture is large if (1) the admixture proportion is close to 50%; or (2) the ratio between the times of the original split and the time of secondary contact is large; or (3) the size of Px is large.

The *f*4 admixture statistics is defined as *f*4 (P1, P2; P3; P4), and *p*1, *p*2, *p*3 and *p*4 are the respective allele frequencies and, by calculating (*p*1 - *p*2)(*p*3 – *p*4), it is possible to calculate if P1/P2 and P3/P4 form clades. If the value is close to zero, the proposed topology is confirmed instead if it is negative P1 is more related to P3 and P2 to P4.

Major applications of *f*4 are the following. The Rank Test estimates the branch length by obtaining a lower bound for the number of admixture events. Comparing two sets of test populations, S1 and S2, with two "unadmixed" reference populations for each set, R1 and R2 (Figure 7A), then *f*4(S1, R1; S2; R2) should be zero (Reich et al., 2012), while each admixture event introduces at most one additional branch.

Another application is to estimate admixture proportion α between closely related populations (Green et al., 2010) as α= F4(P0,PI;Px,P1)/F4(P0,PI;P2,P1) where Px is the admixed population; P1 and P2 are the potential contributors (with proportions α and 1-α, respectively); PO is a reference outgroup; PI is a reference populations with no direct contribution to PX, but more closely related to one of P1 or P2 than the other. Basically, α measures how much closer the common ancestor of PX and PI is to the common ancestor of PI and P1 than to the common ancestor of PI and P2 (Figures 7B and C).

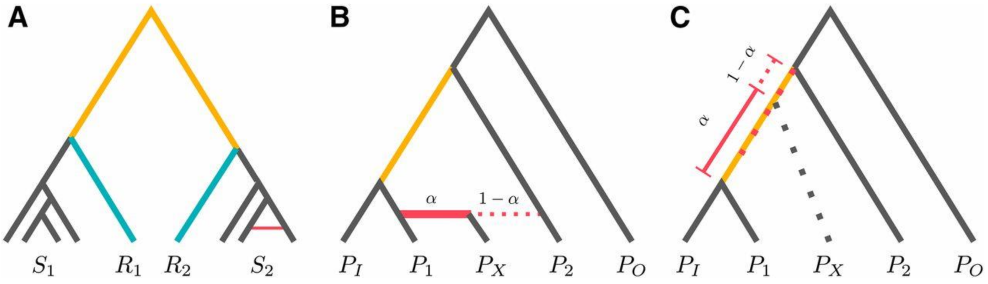


Figure 7. F4 statistics and admixture events. Modified from (Peter, 2016).

The *f* statistics is also used to detect tree-based inference about split and admixture events that occurred amongst a group of populations. An example is qpGraph (Reich et al., 2009, Patterson et al., 2012) that analyzes a user-specified tree with or without admixture events and gives back a likelihood value. Another tree-based method largely used to infer patterns of population splits and mixtures in the history of a set of population is TreeMix that uses a multivariate normal distribution to relate observed allele frequencies among populations.

Another important concept introduced in the middle of the '60s was the LD, the non-random association of alleles present at different loci. Starting with the Lewontin proposal of the D´ statistics for quantifying the LD, nowadays it is possible to use computers to model systems of multiple loci under selection (Lewontin, 1964).

It is clear that each individual has inherited different fragments of chromosome from each parent (haplotypes) that in turn had inherited fragments from their ancestors, therefore considering an entire population there are fragments inherited by multiple individuals that came from the same common ancestor, (identity by descent, IBD).

The recombination reduced these fragments by breaking some of them and maintaining others. The study of these blocks of LD and their common inherited history, by using the so-called haplotypes-based methods, is a new important development in population genetics that tries to improve the resolution of population structure and minimizes the biases of the SNP analysis. The concept behind the use of haplotypes is more or less the same of the SNPs. An important improvement is to obtain more precision in dating admixture events, not considering pairs of SNPs, but haplotypes. Actually it is possible to date admixture events also using SNPs pairs, with software such as *ALDER* (Loh et al., 2013), , considering the genetic distance among them that is larger when multiple admixtures occurred. To increase the precision it is possible to obtain pairs of haplotypes with *CHROMOPAINTER* and then using *GLOBETROTTER* to identify how much those inherited haplotypes are "broken" (Hellenthal et al 2014).

The new field of bioinformatics triggered the development of user-friendly software that facilitates the use of these statistical methods to evaluate affinities between populations and to shed light on the internal population genetic structure by identifying possible ancestral sources and by dating the events that produced the current gene pool. New algorithms and high performing computational machines make it also possible to analyze the increasing number of complete genomes and genome-wide data currently available. In 1964, Cavalli-Sforza limited his study to 15 populations and few loci, not only for the lack of molecular resolution, but also for the lack of bioinformatics tools (Cavalli-Sforza et al., 1964). A possible limitation of the big data management is the risk of losing some information while trying to test a predefined model, therefore the next era in population genetics is the supervised machine learning approach (Schrider and Kern, 2018), in which the models are developed by the machine while analyzing the data.

1.2 Different genetic systems and molecular markers

Molecular genetic markers represent one of the most powerful tools for the analysis of genomes (Duran et al., 2009). Different classes of genetic markers (Figure 8) have been used in turn to reconstruct affinities and diversities among animal (including humans) and plant species: biparental markers (autosomes), X chromosome and uniparental systems, i.e. mitochondrial DNA (mtDNA) and Y chromosome in animals, plus plastid DNA in plants (Underhill and Kivisild, 2007, Ajmone-Marsan et al., 2010, Ferradini et al., 2017).

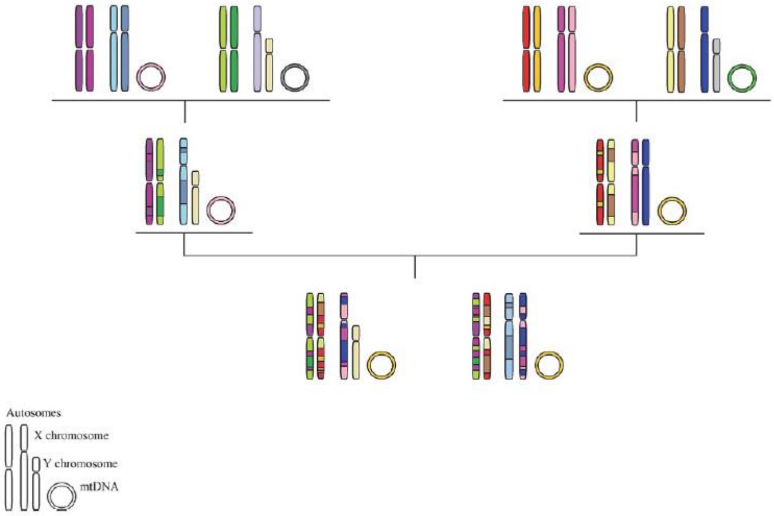


Figure 8. Differences between uniparental and biparental transmission (Pereira and Gusmão, 2017).

These genetic systems contain different pieces of information, therefore only considering them together and integrating evidence from different fields and various species (coevolution), it is possible to reveal the entire puzzle hidden in the genome. A paradigmatic example is the Neanderthal history reconstructed by using both autosomal and mitochondrial markers (Posth et al., 2017).

Until recently, the uniparental systems played a major role in population studies, but over the years, the improvement in DNA sequencing and bioinformatics allowed to increase the region of the genome that could be analyzed, up to the whole genome. Moreover, in the last years, the development of Paleogenomics, based on the analysis of ancient DNA (aDNA) recovered from archaeological remains, is revolutionizing the entire population genetics field, placing itself as a key to better understand the history of humans and other species and to refine all the information previously obtained only from modern DNA (Ottoni et al., 2017, Achilli et al., 2018, Skoglund and Mathieson, 2018).

1.2.1 Uniparental systems

The analysis of uniparentally-inherited portions of the genome finds important applications in a wide range of fields, including evolutionary anthropology, population history, medical genetics, genetic genealogy and forensic science (van Oven and Kayser, 2009).

The maternally inherited mitochondrial DNA and the paternally transmitted Y chromosome have been extensively employed to determine origin and diversity of species. Even though more recently autosomal genetic markers have offered new important tools to reveal population genetic sub-structures and human demographic histories, the mtDNA and the Y chromosome, with their unique patterns of inheritance, continue to be important sources of information for reconstructing traceable molecular genealogies (Wilkins, 2006, Henn et al., 2010, Kivisild, 2015).

The mtDNA is a circular DNA molecule, stored in the mitochondria, inside the cytoplasm. It is around 17 kilo base pare (kbp) long in mammals and inherited through the maternal line.

The structure of the Y chromosome consists in a major region called Male Specific region (MSY), more than 37Mb long, also known as the Non-Recombining region of Y chromosome (NRY) and three Pseudo Autosomal Regions (PAR) (Veerappa et al., 2013, Hughes and Page, 2015). The latter are the only ones that, during meiosis, undergo homologous recombination with their counterparts on the X chromosome; instead, the MSY is transmitted clonally in a holandric way.

The uniparental systems, over the time, underwent processes of random mutation without recombination, that generated monophyletic groups, called haplogroups (Hg), clades or lineages, characterized by the same combination (equal position and order) of molecular markers (Jobling et al., 1997, Jobling and Tyler-Smith, 2003); therefore, all modern mtDNAs and MSY regions coalesce back to one ancestral sequence (the so- calls Eva and Adam molecules, respectively) at some point in the past, while variants belonging to different haplogroups showed a separation at a certain time in the past, from which a series of mutations occurred independently and accumulated differentiating various haplogroups and sub-haplogroups (sub-Hg). Because the process of molecular differentiation is relatively fast and occurred mainly during and after the process of human dispersal into different parts of the World, these haplogroups usually tend to be restricted to particular geographic areas and populations (Torroni et al., 2006). Uniparental transmission and the lack of recombination allowed the uniparental data to be easily combined into the shape of a phylogenetic tree (Kivisild, 2015). They were largely used in the phylogeographic approach, combining phylogenies, geographic distributions and divergence times, to reconstruct the origin and the migrations of various populations.

Moreover the coalescent approach is easily applicable to these systems and represents an easy and rapid way to date migration events (at population level) and to analyze demographic changes that occurred during time, with ML and Bayesian analyses (Soares et al., 2009, Posth et al., 2016).

The genealogies reconstructed from the two uniparental systems might reveal similarities and/or differences among the cultural, social and evolutionary histories of different species and/or population groups. Among the two uniparental systems, only the mitochondrial variability has been investigated in this PhD work.

1.2.1.1 Molecular and genetic peculiarities of the mitochondrial DNA

Mitochondria occupy a central role in the biology of cells and are crucial for energy production and consequently for cell life (Ballard and Whitlock, 2004). Mitochondrial DNA is organized as a circular, double-stranded molecule (Figure 9).

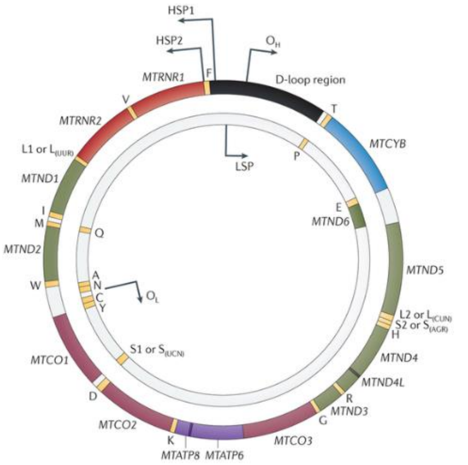


Figure 9. Map of the human mitochondrial genome. Loci are colored according to functional groupings. Gene identifiers on the outside of the map are transcribed on the heavy strand and gene identifiers on the inside of the map are transcribed on the light strand. Transfer RNA loci are designated by the single letter code of their specific amino acid. The non-coding D-loop is shown at the top of the map (in black) (Stewart and Chinnery, 2015).

The two strands are denoted as H (heavy) and L (light), the first is guanine-rich while the L-strand is cytosine-rich. The 16-17 kbs of the mammal mitochondrial genome (the main group considered in this thesis) consist mainly of a coding region. The only non-coding part is limited to the "control region" (or D-Loop). The control region is involved in the regulation of transcription and replication of the molecule (Brandstätter et al., 2004).

In the D-loop there are regions that are more conserved and regions that are highly variable at the population level if compared to the rest of the genome. There are different number of these regions among species, the human contains three Hyper Variable Sequences (HVS), while for instance the bovine mtDNA contains only one HVS. The coding region has a very compact structure without introns, and encodes for 37 genes, (in humans 28 on the H-strand and 9 on the L-strand), all of which are essential for normal mitochondrial function.

Thirteen of these genes encode for enzymes involved in the OXPHOS, while the remaining for two ribosomal RNAs (rRNAs 12S and 16S) and 22 transfer RNAs (tRNAs), which are required and sufficient for the synthesis of mitochondrial proteins (Anderson et al., 1981, Wallace, 1994, DiMauro and Schon, 2003). Transcription of the mtDNA is 'prokaryotic like', based on three polycistronic RNA (Taylor and Turnbull, 2005). The mitochondrial DNA translation machinery uses an alternative genetic code that deviates from the traditional one with four major differences (Barrell et al., 1979, Anderson et al., 1981, Knight et al., 2001) (Table 2).

Table 2. Mitochondrial genetic code variation in mammals, fruit flies and yeasts.

RNA codon	Nuclear genetic code	mtDNA genetic code		
		Mammals	**Drosophila**	**Yeasts**
UGA	STOP	Tryptophan	Tryptophan	Tryptophan
AGA **AGG**	Arginine	STOP	Serine	Arginine
AUA	Isoleucine	Methionine	Methionine	Methionine
AUU	Isoleucine	Methionine	Methionine	Methionine
CUU **CUC** **CUA** **CUG**	Leucine	Leucine	Leucine	Threonine

There are some peculiar features that brought the mtDNA at the focus of evolutionary and population genetics studies.

As already mentioned, one of the main features of the mtDNA is that it is usually inherited along the maternal line, as observed for the first time in 1980 (Giles et al., 1980). Different theories of how the paternal mtDNA is excluded were proposed, however it is clear that in natural conditions the maternal inheritance is strongly controlled by the selective elimination of paternal mtDNA (Sato and Sato, 2013, Pyle et al., 2015, Zhou et al., 2016), even if in some pathological cases a paternal transmission has been observed (Pyle et al., 2015; Luo et al., 2018).

Another important mtDNA feature is the high copy number per cell, with thousands of molecules in each cell (and a wide variability among tissues). This characteristic was very important in the pre-PCR era, when the medical genetics studies were conducted on placenta cells. It is still very important but in other two fields, archaeogenomics and forensic science, in which the starting biological substrates (bones, teeth, hairs etc.) are highly degraded and contain with very few DNA molecule.

Usually the mtDNA sequences in a cell and in an organism are identical and this condition is called "homoplasmy". However, sometimes wild type and mutated molecules can coexist and this situation is named "heteroplasmy". The percentage of heteroplasmy can vary among different individuals and populations (Irwin et al., 2009), but also between organs or cells in the same individual (Calloway et al., 2000).

This mixture of wild type and mutated mtDNAs is often correlated with clinical expressions (Avital et al., 2012, Gasparre et al., 2013, Sobenin et al., 2013); in these cases the wild type must exceed a tissue specific threshold before a cell expresses a biochemical defect in the respiratory chain (Schon et al., 1997, Wallace et al., 1998). An increasing of published heteroplasmic variants in the last decade is due to the application of NGS, in fact the high coverage obtained for the mtDNA might reveal also rare alleles (Wallace and Chalkia, 2013). Studies using this approach revealed that 25-65% of the general population has at least one heteroplasmy across the entire mitochondrial genome (Consortium, 2010, Li et al., 2010, Sosa et al., 2012).

Another key feature of the mtDNA is represented by high mutation rate, 10-20 times higher than that of nuclear genes (Brown et al., 1979). This higher instability is due to several reasons, among which the higher frequency of replication than the nuclear DNA, the lack of an efficient DNA repair mechanism as well as protective proteins such as the histones (Clayton et al., 1974, Tao et al., 2014) and the high exposure to damage of one strand of the mtDNA molecule during the replication and/or transcription processes with the presence of reactive oxygen species (ROS) in the mitochondria. The mutation rate can vary not only along the same species, from individual to individual, but also from cell to cell in the same individual. The main challenge in obtaining a calibrated molecular clock for the mtDNA is due to the presence of hypervariable segments, therefore, the rate in the non-coding control region is about 10 times higher than that of the coding region (Pakendorf and Stoneking, 2005, Howell et al., 2007, van Oven and Kayser, 2009).

The first estimation of the mtDNA mutation rate used the chimpanzee as an outgroup (1.70 x 10 -8 substitution / site / year) (Horai, 1995, Ingman and Gyllensten, 2001). The phylogenetic approach with a distant outgroup generates mutational rate estimates which were at odds with the mutation rates estimated from pedigree data (Kivisild, 2015). This is probably due to the detection of heteroplasmic states that are rarely fixed in the germ lines. During the years, a wide range of molecular clock models and methods, implemented in various statistical phylogenetic settings, have been proposed (Ho and Duchêne, 2014), but the most commonly used for humans is the time-dependent clock established on modern mitogenomes, which corrects for the effect of selection and is routinely applied in phylogeographic studies ($2.33 \pm 0.2 \times 10^{-8}$ base substitution per nucleotide per year) (Soares et al., 2009). Recently another mutation rate, based on the divergence between ancient and modern human mitogenomes, has been introduced and it consists in a linear clock obtained using 66 radiocarbon dated ancient mitogenomes as tip calibration points (2.70 ± 0.2 $x10^{-8}$ base substitution per nucleotide per year) (Posth et al., 2016).

1.2.2 Autosomal markers: the new "evolution"

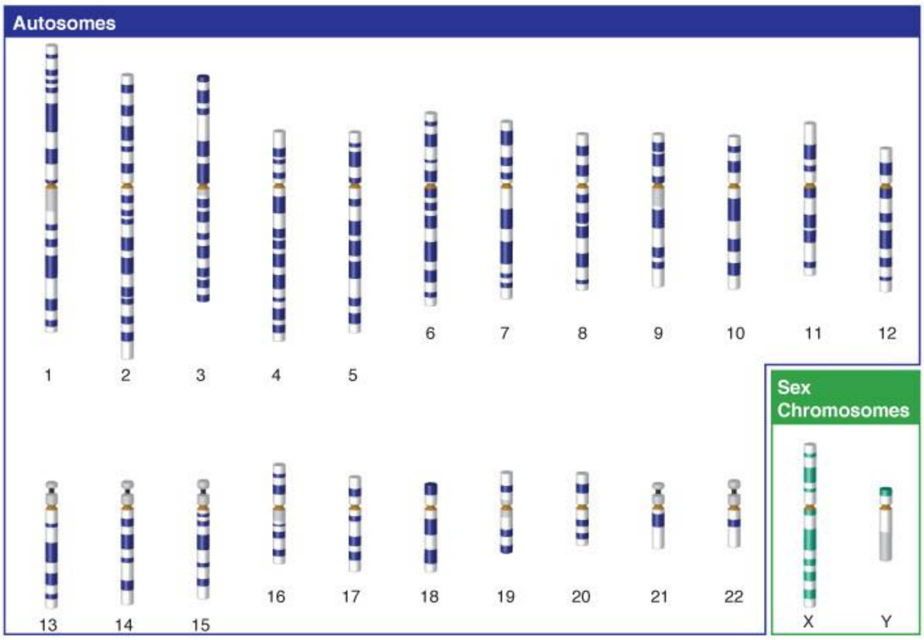


Figure 10. Overview of the Human genome (from the Human Genome Project web site).

The uniparental systems could be considered as two loci, each with its MRCA, able to reconstruct the female and male part of the human history; they can actually describe only two ancestors of the thousands that participated in the genetic heritage of modern populations. The genomic representation of a larger number of ancestors is encrypted in the autosomal markers.

The length of the haploid human nuclear genome is 3.2 billion base pairs organized in 23 separate molecules, the chromosomes (Figure 10). Therefore, in each diploid cell there are 46 chromosomes organized in 23 pairs; each member of the pair is inherited from a parent. Twenty-two pairs are called autosomes, while the remaining two are known as the sex chromosomes (X and Y), due to their differential presence in males and females. Differently from the uniparental systems, autosomes undergo recombination during meiosis (Jeffreys et al., 2001). Variants at loci of different chromosomes are unlinked because inherited from parents in an independent way. On the contrary, variants at loci close together on the same chromosome are linked together and can be co-inherited from the same parent. As described above about the haplotype-based methods, recombination can interrupt these blocks, based on the distance between loci, and after meiosis the offspring might inherit a different combination of variants (Figure 8).

The whole genome represents the highest resolution possible for studying the variability among people, populations and/or species. In the whole genome all the variants are present, but the recombination makes it difficult to rebuild the entire historical puzzle. In fact, each locus has its own MRCA and tells us part of the past. The study of autosomal markers started early, during '50s and '60s (Cavalli-Sforza and Feldman, 2003), and various techniques were developed to discover and compare more markers, i.e. population-specific polymorphisms.

One of the first evidence coming from studies on autosomal markers was that the genetic diversity of humans is relatively low if compared to the other great apes, with the exception of western chimpanzees and eastern gorillas (Prado-Martinez et al., 2013, Xue et al., 2015). Low genetic diversity means that it is difficult to find informative SNPs in a single gene, but chances increase if the whole genome is sequenced. This has been possible since the year 2000 thanks to the automation of Sanger sequencing (Metzker, 2010) and to the complete human genome published by the Human Genome Project (HGP) in 2004 (Consortium, 2004). In the current genomic era, the variability can be represented by whole genomes or by genome-wide data. The first is the analysis of all bases present in a genome, while in the second one only specific polymorphisms (previously identified with the whole genome analysis) are screened. The genome-wide SNPs are informative for various disciplines, such as medicine and population genetics, and are tested through chip-arrays or molecular capture/enrichment techniques (especially for ancient DNA). The targets are specific regions that contain informative SNPs with a cost-effective advantage and two major features: minor allele frequencies (MAF) above 5% and recombination association or LD. Nowadays there are arrays that can detect up to 5 million of SNPs in humans, developed principally by Illumina (Human OMNI5) and

Affymetrix (Axiom™ Genome-Wide Human Origins 1). Additional chips were also commercialized to study the variability of other species, especially domestic animals.

Different projects started to study the human variability by analyzing complete genomes or large number of informative SNPs. The 1000 Genomes project (1KGP), the Human Genome Diversity Project (HGDP) and the Simon Project are the most important ones, and their data are publicly available. Several consortia were also born with the aim to sequence the complete genome of different species, including most of domesticated animals.

1.2.3 The emerging field of Archaeogenomics

The genomic era has revolutionized the study of human evolution, but another important factor that enormously increased the knowledge of our past is the "ancient DNA genomics" (Achilli et al., 2018, Skoglund and Mathieson, 2018).

As for the whole genome analysis, until few years ago, the use of ancient DNA was a daydream for population geneticists. It is easy to understand that with time part of the genetic information can be lost and the surrogate of ancient genomes reconstructed from modern data are only a hypothesis of what was present in a place at a specific time. Whereas, the study of ancient DNA makes it possible to investigate the genome before and after historic events and to observe a real-time evolution (Tuross and Campana, 2018).

After the cell death, the DNA is subject to various physical and chemical modifications, including fragmentation and post-mortem damage. The most recurrent chemical modification is the deamination of the cytosine in uracil at the end of each DNA fragment (Figure 11). The uracil is then copied as thymine, this leads to the introduction of phantom GC→AT mutations (Hofreiter et al., 2001). In other words, DNA molecules in ancient remains are very few, highly fragmented and modified (particularly at their ends).

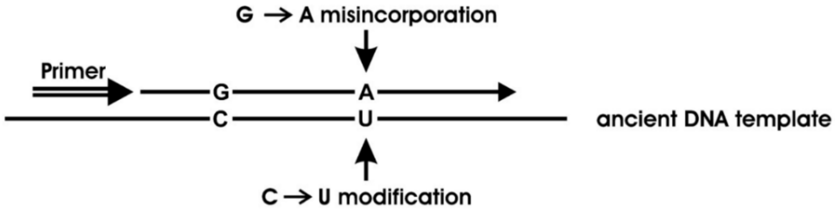


Figure 11. Cytosine deamination. The U in the ancient DNA template represents a deoxyuridine residue, a 5-hydroxydeoxyuridine residue or any other modified deoxycytidine residue read as T by Taq polymerases. Modified from (Hofreiter et al., 2001).

The first attempts to get DNA from archaeological remains (museum specimens) are dated back to the middle of the '80s, when Higuchi (Higuchi et al., 1984) first and then Pääbo (Pääbo, 1985) tried to extract DNA respectively from a 140 year-old extinct zebra (the quagga, *Equus quagga quagga*), which inhabited South Africa until the end of the nineteenth century, and from a 2400 year-old Egyptian mummy. In both studies the aDNA were end-repaired, ligated into plasmids, and replicated within bacteria. Thanks to these two papers on Nature, the aDNA expectations started steeply and increased with the development of PCR-based methods. However, in the first '90s it was clear that there were a lot of problems with the PCR approach. Confounding factors, such as inhibitors, chemical modifications and short fragments, limited the technique and led to the publication of papers including aDNA sequences with a large number of contaminations.

The "endogenous" DNA can be masked by two possible sources of contamination: human modern DNA, which is longer and less degraded, and exogenous DNA from bacteria and other microorganisms (Hofreiter et al., 2001, Pääbo et al., 2004). Golden rules for the authentication of ancient DNA data (Cooper and Poinar, 2000) were published, asking in particular that a subset of results should be verified by independent replications in another laboratory. Yet soon after, disillusion dominated the ancient DNA field (Figure 11) until that was overcome with the high-throughput of the NGS technologies that allowed parallel sequencing of billions of DNA fragments in a single reaction, thereby dramatically increasing the amount of genetic information that could be obtained from each microliter of extract (Orlando et al., 2015).

The main advantage of the NGS is that the PCR is not mandatory, thus bringing two major outcomes:

1) it is possible to sequence those short endogenous fragments that were previously lost;
2) common patterns of post-mortem damage can be verified on the reads *ex-post* to confirm the authenticity of the ancient DNA.

NGS triggered the beginning of the "golden age" of aDNA and soon after the commercialization of the first high-throughput machine a draft of the first woolly mammoth (*Mammuthus primigenius*) genome was generated (Miller et al., 2008).

In 2010 the first ancient genomes from Neanderthals and modern humans appeared (Green et al., 2010, Rasmussen et al., 2010). These studies revealed that ancient populations often had ancestries not fully represented in present-day populations, a preview of the way in which ancient DNA would become critical for reconstructing the genetic history and evolution of modern humans (Skoglund and Mathieson, 2018). So far, whole-genome shotgun and targeted capture data from thousands of ancient human remains were published, dating as far back as 45 kya. These data impacted the knowledge of our past. Paradigmatic examples are: the discovery of the introgression of Neanderthal sequences in the genome of anatomically modern humans; the identification of the ancestral components of Eurasian and Native American populations; and the possibility to recalibrate molecular clocks using radiocarbon-dated samples (Fu et al., 2014, Helgason et al., 2015, Posth et al., 2016).

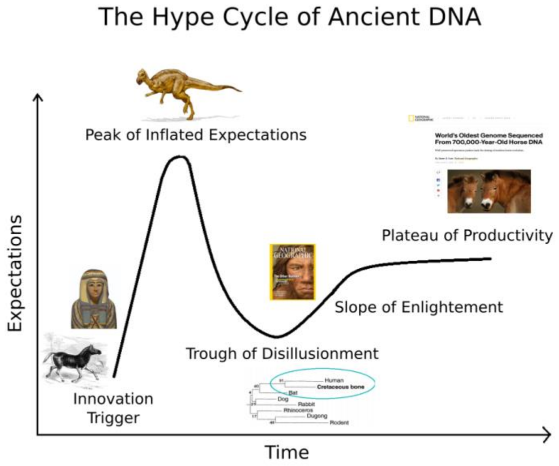

The Hype Cycle of Ancient DNA
Peak of Inflated Expectations
World's Oldest Genome Sequenced From 700,000-Year-Old Horse DNA
Plateau of Productivity
Expectations
Slope of Enlightement
Trough of Disillusionment
Innovation Trigger
Human
Cretaceous bone
Bat
Dog
Rabbit
Rhinoceros
Dugong
Rodent
Time

Figure 12. Expectation in archaeogenomics through time (molecularecologist.com/2017/04/the-hype-cycle-of-ancient-dna/).

Even if NGS has revolutionized archaeogenomics, there are still some challenges in the field due to the poor preservation of DNA in very hot/humid environments (Kistler et al., 2017b). More recently, also this problem seems to be overcome by the identification of the petrous bone as the part of the skeleton with best ancient DNA preservation (Pinhasi et al., 2015).

As a final remark, it is worth mentioning that even if the amount of endogenous DNA sequenced from an ancient human remain is less than 1% (Schuenemann et al., 2011), the other reads might be from pathogens, soil bacteria and other organisms, thus becoming very informative also for other research fields.

1.3 Reconstructing the history of human populations: from the Out-of-Africa exit into the Americas

1.3.1 The mitochondrial perspective

One of the most hotly debated issues in palaeoanthropology (the study of human origins) focuses on the origins of anatomically modern humans (AMH). This issue is the focus of great deliberation between two schools of thought: one that stresses "multiregional continuity" and the other that suggests a single origin for modern humans, the previously mentioned "Out of Africa" theory (Figure 13).

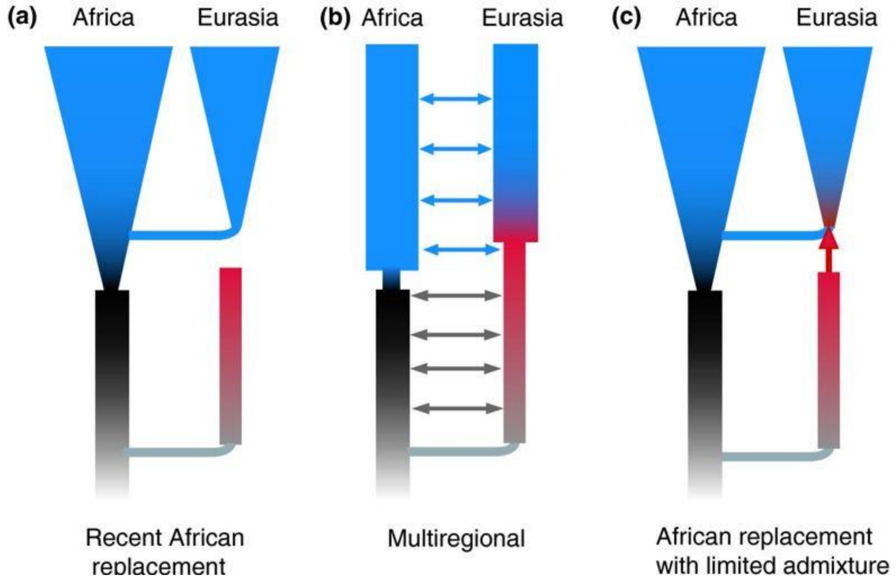

(a)
Africa
Eurasia
(b)
Africa
Eurasia
(c)
Africa
Eurasia
Recent African replacement
Multiregional
African replacement with limited admixture

Figure 13. Models of Human Origin. Modern humans are designated in blue and other extinct Eurasian archaic human species in red: a) "Out of Africa" proposes that anatomically modern humans originated in Africa, expanding into Eurasia relatively recently and replacing other human species; b) Multiregional proposes that the evolution of modern humans occurred in both Africa and Eurasia, maintaining local genetic continuity but with populations united by gene flow; c) Assimilation proposes a recent African origin for the bulk of the human genome with limited admixture with existing populations out of Africa (Hodgson and Disotell, 2008).

The multiregional model proposed by the anthropologist Weidenreich, implies that the transition from *Homo erectus* to AMH took place in several areas of the Old World, with many modern human characteristics arising at different times in different places (Weidenreich, 1940). Usually the "Multiregional" term also defines the Candelabra model, proposed by Coon in the 1962 and analyzed by Templeton

2007 (Coon, 1963, Templeton, 2007). The Candelabra model assumes an independent origin of AMH from separated groups of *H. erectus* that went out of Africa around one million of years ago. An elaboration of this model assumes also interactions and gene flow between groups of *H. erectus*. These “Multiregional” theories were widely rejected since the first mtDNA phylogenetic tree have been published showing a deep root in Africa and a following out of Africa exit (Cann et al., 1987, Vigilant et al., 1991, Horai, 1995, Ingman et al., 2000).

In the first years of the ’90s, the basal branches of the mtDNA phylogeny were defined using restriction fragment length polymorphism (RFLP) analysis and starting from Native Americans (Figure 14), then also Europe, Asia and finally Africa were studied (Torroni et al., 1994, Chen et al., 1995, Torroni et al., 1996). After the first haplogroups were defined, the dissection of the mitochondrial variation increased (Figure 15) from D-loop sequencing to the highest molecular resolution, with thousands of complete mtDNAs published so far, leading to a more detailed mitochondrial DNA tree (Phylotree 17).

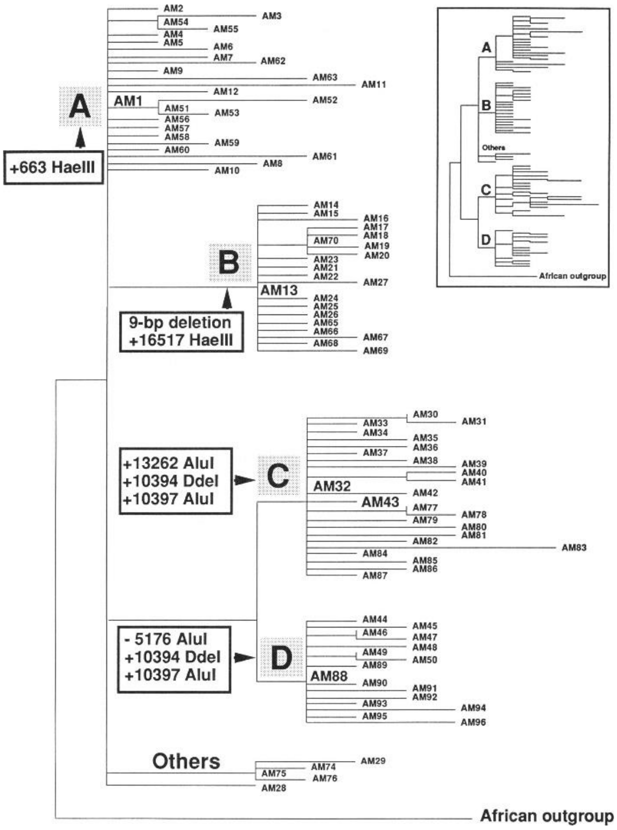


Figure 14. First mtDNA haplogroups identified in Native Americans. (Torroni et al., 1993).

The human mtDNA tree initially splits into branches that carry exclusively African sequences belonging to haplogroup L. In Africa it is also possible to observe the highest mtDNA diversity, whereas the lowest value is found in Native Americans. The haplogroup L has been subdivided into seven main branches (L0-L6) (Kivisild et al., 2004, Behar et al., 2008) and only one of these, Hg L3, has derivatives also outside Africa (Kivisild et al., 2006). The African-specific L3 sub-lineages are also found in Mediterranean Europe, West Asia, and the Americas showing more recent contacts and migrations from Africa (Cerezo et al., 2012).

The L3 is dated to 95-62 kya (Fu et al., 2013b), while its two main derivatives outside of Africa, M and N, branching out from the root of haplogroup L3 (Figure 16) are dated between 70 and 50 kya. In the timeframe modern humans left Africa, probably after the Eruption of Mount Toba in Sumatra 74 kya (Petraglia and Allchin, 2007, Soares et al., 2012).

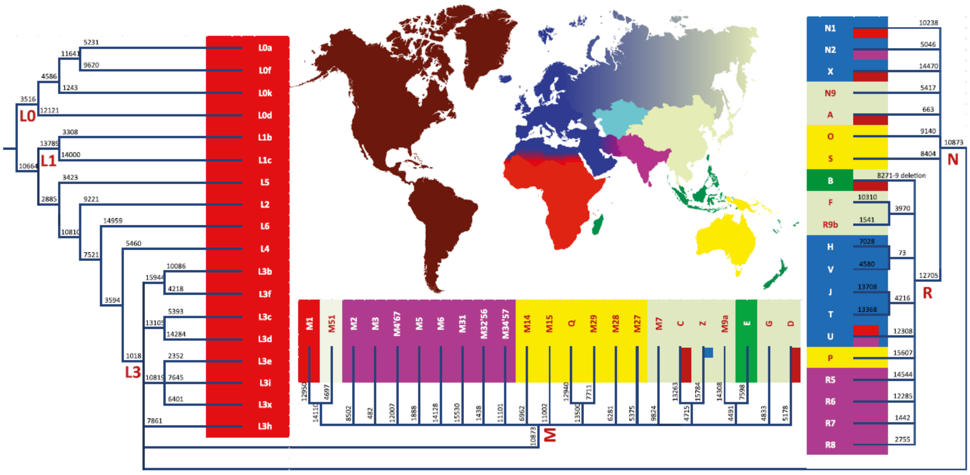


Figure 15. MtDNA haplogroup tree and its distribution map. Haplogroup labels are reported according to the Phylotree 16. Only a single branch-defining marker, preferably from the coding region, is shown. The differential geographic haplogroup distribution are highlighted with different colors (Kivisild, 2015).

The number of extant non-African founder haplogroups has been later extended to include a third member, haplogroup R, which is a daughter-clade of N. An ancient sample dated 45 kya, Ust-Ishim, falls at the root of haplogroup R and confirms the time of Out of Africa (Fu et al., 2014).

The fact that virtually all non-African mtDNA lineages derive from just one of the two sub-clades of the African haplogroup L3 has been interpreted as an evidence of a major bottleneck of mtDNA diversity at the onset of the out of Africa dispersal (Underhill and Kivisild, 2007). The magnitude of this bottleneck has been estimated from whole mtDNA data yielding an effective population size that ranges between several hundreds (Macaulay et al., 2005) and only few dozen females (Lippold et al., 2014).

Two alternative scenarios have been proposed to explain the presence of the two sub-branches of the mtDNA haplogroup L3 in both Europe and Asia. The first postulates a ‘Levantine route’ from northeast Africa to the Levant across the Sinai Peninsula ~45 kya (Stringer and Andrews, 1988, Prugnolle et al., 2005). However, the route along the Levantine corridor did not explain why adjacent Europe was settled thousands of years later than distant Australia (Forster and Matsumura, 2005). The alternative “southern route model” suggests that the dispersal probably started ~70 kya from the Horn of Africa to the Persian Gulf and further along the tropical coast of the Indian Ocean to Southeast Asia and Australasia. This second scenario is strongly supported by paleo-environmental evidence, confirming that a northern migration would have been impossible during the glacial period extending from ~70 to 50 kya (Forster and Matsumura, 2005, Macaulay et al., 2005, Mellars, 2006, Torroni et al., 2006).

The main sub-branches of M, N and R are distributed in a specific regional configuration (Kivisild, 2015). For example the Eurasian lineages, HV, JT, N1, N2, U and X are diffused in Europe, South-West Asia and North Africa (Soares et al., 2010), R5-R8, M2-M6 and M4’67 are virtually confined to South Asia (Chaubey et al., 2007), while haplogroups A-G, Z and M7-M9 are widespread in East Asia (Stoneking and Delfin, 2010).

Actually, the richest basal variation of the founder haplogroups M, N and R is found along the southern stretch of Eurasia, particularly in the Indian subcontinent (Palanichamy et al., 2004, Sun et al., 2005, Chaubey et al., 2008) and a similarly high basal diversification is present in Southeast Asia (Macaulay et al., 2005, Kong et al., 2006, Hill et al., 2007). These data support the “southern route model” indicating a rapid colonization along the southern coast of Asia, reaching Sahul 60 kya. The expansion northwards to fill the heartland of the continent occurred only later, ~45 kya, when a combination of technology and climatic conditions enabled the exploration of the interior of Eurasia. This explains the presence of mtDNA haplogroups M14-M15, M27-M29, Q, P S only in Australia and Melanesia. A more recent Holocene migration of populations speaking Austronesian languages despairs the Hg B4a1a1 in Oceania (Kayser, 2010).

Europeans have a high level of haplogroup diversity within haplogroups N and R (H, HV, N1, J-T, U, I, W, and X) but lack haplogroup M almost entirely (Underhill and Kivisild, 2007, Soares et al., 2010). The first peopling of Europe by modern humans occurred about 45 kya (Gamble et al., 2004, Mellars, 2006). Members of mtDNA haplogroup U5 probably marked the first Upper Palaeolithic entry in Europe from the Near East, while populations bearing U6 (and M1) entered North Africa (Olivieri et al., 2006, Pennarun et al., 2012). Recent ancient DNA data have shown that Palaeolithic Europeans derive from a single ancestral population, but that this long-term genetic continuity was in part interrupted by the appearance of a novel genetic component ~14 kya, during the first significant warming period (Bølling-Allerød interstadial) after the Last Glacial Maximum (LGM) (Fu et al. 2016). Actually, some European mtDNA haplogroups have been dated back to the late Pleistocene, suggesting possible post-glacial expansions (Soares et al., 2010, Kivisild, 2015): haplogroups V (Torroni et al., 1998, Torroni et al., 2001), H1, H3 (Achilli et al., 2004, Pereira et al., 2005), H5 and U5b1b (Tambets et al., 2004, Soares et al., 2010) from the Franco-Cantabrian refugium; U5b3 from the Italian Peninsula (Pala et al., 2009); U4 and U5a from the East European Plain (Malyarchuk et al., 2008, Malyarchuk et al., 2011); and J, T, I and W from Near Eastern refugia (Pala et al., 2012, Olivieri et al., 2013).

However, an ongoing debate concerns the relative amount of genetic input into modern Europeans from Palaeolithic *versus* Neolithic waves of settlement. In fact, the advent of agriculture and pastoralism interested Europe since the arrival of the Early Neolithic material culture in Greece ~8.5 kya (Manning et al. 2014).

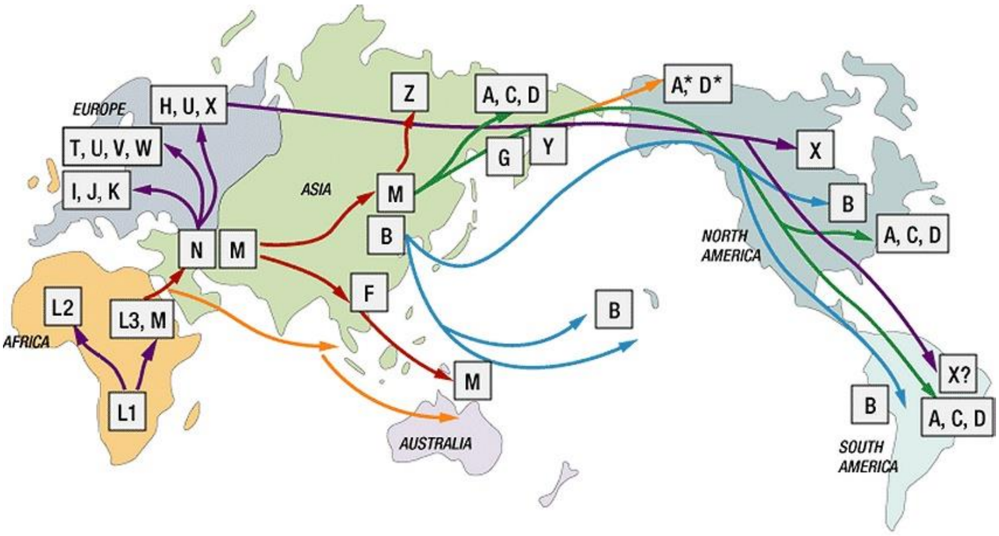


Figure 16. MtDNA haplogroup migration patterns. The map shows the migration patterns of the main mtDNA haplogroups. From FamilyTree webpage.

The prevailing conclusion, supported by ancient DNA studies, is that Palaeolithic and Mesolithic hunter-gatherer European populations differed genetically from early Neolithic farmers, in turn implying that there was a wide-scale replacement across

Europe from the Near East in the early Neolithic, with limited assimilation of native Europeans (Pinhasi et al., 2012, Lazaridis et al., 2014, Omrak et al., 2016, Posth et al., 2016). However, ancient genomes (Hofmanová et al., 2016) have also confirmed the previously mentioned Late Glacial/ Postglacial recolonization of Europe (particularly the Mediterranean area) from the Near East before the migration waves associated with the onset of farming, as hypothesized by modern mtDNA studies (Pala et al., 2012, Olivieri et al., 2013) and by two K1c mitogenomes from Mesolithic Greece (Hofmanová et al. 2016).

Together with the migration towards the western part of Eurasia around 45 kya, the migratory flow probably also took a more southern route, back again into Africa, across the Mediterranean side. This dispersal event has been hypothesized studying the distribution of the mtDNA haplogroups M1 and U6, which are found virtually only in Africa and whose arrival would have temporally overlapped with the event that led to the peopling of Europe by modern humans (Olivieri et al., 2006, González et al., 2007). Similarly, three other potentially Eurasian ancient mtDNA clades have been identified in Eastern Africa: N1a1a, HV1 and R0a, providing additional evidence of a ‘back to Africa’ migration. This scenario has been recently confirmed also by two ancient DNA studies. The first concerning a 35,000-year-old individual from Romania with a mitogenome belonging to haplogroup U6*, with a haplotype not previously found in ancient or present-day humans (Hervella et al., 2015). This finding is in line with a migration event from Western Asia to Africa during which haplogroup U6 diversified until the emergence of the present-day African lineages but indicates that U6 moved at the same time also to Europe, where it later disappeared.

The second paper, by van de Loosdrecht et al. (van de Loosdrecht et al., 2018), reported genomic data from seven individuals from Grotte des Pigeons near Taforalt in eastern Morocco, dated between 15.1 and 13.9 kya. Of the seven mitogenomes, six belong to U6 and one to M1, confirming that these two haplogroups were markers of an Upper Palaeolithic migration from the Levant to North Africa.

The last continent to be colonized was the Americas from a subset of East Asian diversity. Intriguingly, the first mtDNA haplogroups were identified by studying Native Americans (Figure 14) and took the first four letters of the alphabet (A-D) (Torroni et al., 1993). In the last decade, many mitogenome studies of modern and ancient samples increased the number of Native American founding lineages from the initial four to sixteen (Figure 17) (Tamm et al., 2007, Achilli et al., 2008, Perego et al., 2009, O'Rourke and Raff, 2010, Brandini et al., 2018). Among these, eight (A2, B2, C1b, C1c, C1d, C1d1, D1 and D4h3a) are called Pan-American haplogroups, as they are distributed across the double continent, while the others are less frequent and generally show a distribution restricted to specific geographic areas, i.e. North America (A2a, A2b, C4c, D2a, D3, D4e1, X2a and X2g) (Perego et al., 2010).

The origin and diffusion of these sub-lineages is still object of study and a very contentious issue is whether the settlement of the Americas occurred by means of a single or multiple streams of migration(s) and if the native populations underwent an incubation phase in Beringia.

The mitogenome perspective on the first peopling of the America indicates that Native Americans trace their ancestry to Asian groups who colonized northeast Siberia – including part of Beringia – prior to the LGM. The pre-LGM Asian haplotypes evolved in Beringian enclaves, some were lost by genetic drift, while novel ones arose *in situ* often becoming predominant due to founder events.

Further genomic analyses suggest that the initial Northeast Asian source population was highly structured with differential relatedness to present-day East Asians and to ancient central Asians (as testified by the 24,000 year old Mal'ta remains), but also to present-day Australo-Melanesians and Andaman Islanders *("Australasians"*), as identified in present-day populations of Amazonia (Zegura et al., 2004, Schroeder et al., 2007, Tamm et al., 2007, Wang et al., 2007, Kemp and Schurr, 2010, Hoffecker et al., 2014, Raghavan et al., 2014a, Hoffecker et al., 2016).

As for the American peopling, the mtDNA variation has been associated with at least three distinct migratory waves:

i. A primary pacific coastal route undertaken about 18-15 kya by the first Paleoindians that eventually reached South America in few millennia along the Pacific coast (probably only 1.5 kilo years), as confirmed by early archaeological findings and marked by nine Pan-American founders A2*, B2*, C1b, C1c, C1d*, C1d1, D1, D4h3a, and D4e1c;
ii. Approximately at the same time, Native American ancestors of C4c, X2a and X2g lineages entered America through an interior route of dispersal toward the mainland of North America;
iii. The spread of Paleo-Eskimo D2a lineages around 5 kya along the Arctic through Northern Canada and Greenland, which were replaced, in the same region, by the spread of Neo-Eskimos carrying A2a, A2b, and D3 lineages.

A recent study (Brandini et al., 2018) provided genetic evidence that South America was populated as early as about 14 kya, according to the coalescence age of the oldest sub-haplogroups that arose in northern South America. Once arrived into South America, the first settler population(s) might have undergone an early split in the northern part of South America with a later diffusion along both the Pacific and

Atlantic coastal regions, as suggested by mitogenome surveys (Wang et al., 2007, Perego et al., 2009, Bodner et al., 2012, de Saint Pierre et al., 2012a, de Saint Pierre et al., 2012b, Brandini et al., 2018).

Those independent and geographically distant paths determined a genetic isolation that could have caused differences between eastern and western archaic cultures with high levels of complexity, as testified for instance by the Huaca Prieta site in the west (Dillehay et al., 2017) and by the archaeological record of Lapa do Santo in east-central Brazil (Strauss et al., 2016).

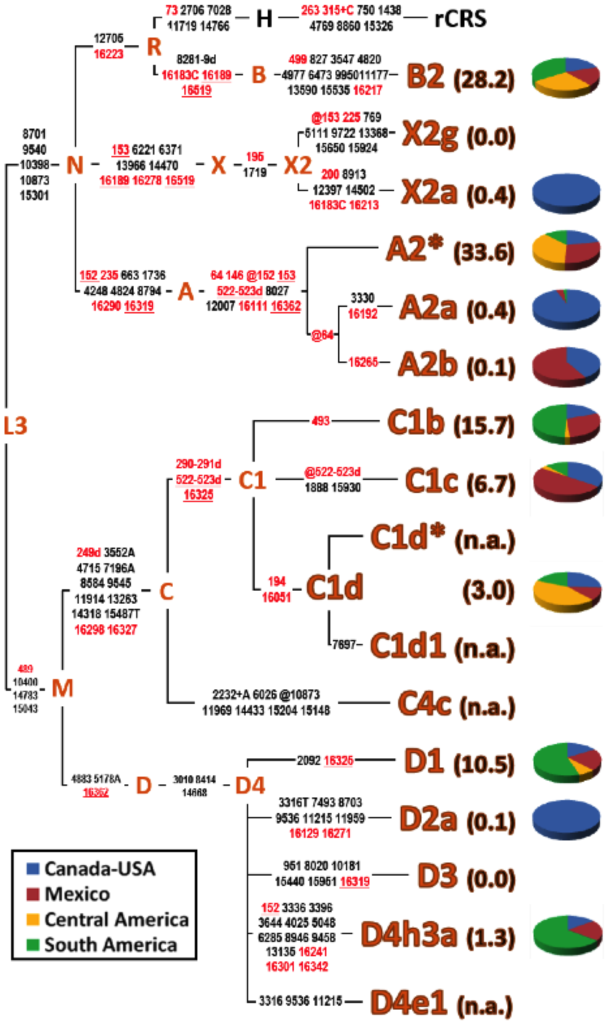


Figure 17. MtDNA phylogenetic tree encompassing the 16 Native American founding lineages (Perego et al., 2010).

1.3.2 The current (Archaeo)genomic scenario

1.3.2.1 “Out of Africa”

The mtDNA “Out of Africa” model described above, which proposes a single and relatively recent transition from archaic hominins to AMH in Africa followed by a later migration into the rest of the world replacing other extant hominin populations, have been confirmed also by studies on genome-wide microsatellite and SNPs (Stringer and Andrews, 1988, Ingman et al., 2000, Stringer, 2002, Cavalli-Sforza and Feldman, 2003, Relethford, 2008, Tattersall, 2009). This is highlighted by the greater amount of genetic diversity found in African than in non-African populations (Figure 18); for example, two chromosomes of a person with recent African ancestry show greater sequence divergence than two chromosomes, ancient or modern, from any two people whose ancestry is from outside Africa (Prugnolle et al., 2005, Ramachandran et al., 2005).

The earliest evidence of anatomically modern humans comes from fossils located in Ethiopia that can be dated to about 190,000–150,000 years ago. Beyond Africa, fossil evidence of anatomically modern humans has been reported as early as about 100 kya in the Middle East and about 80 kya in southern China. However, other hominins, such as Neanderthals, which disappeared from the fossil record about 40 kya, have been found throughout Eurasia as far back as 400 kya.

As already mentioned, the initial out-of-Africa dispersal of a sub-group of the large African metapopulation of *Homo sapiens* (Skoglund et al., 2017) was a notable event that left a strong signature on the genetic variation of all non-African populations, including lower levels of diversity and higher levels of linkage disequilibrium (Nielsen et al., 2017). However, the number, the geographic origin and migratory routes and the timing of major dispersals remain elusive. For instance, there is evidence to support the origins of modern humans in eastern, central and southern Africa, single and multiple dispersals out of Africa, a north or south dispersal route and estimates for the timing of dispersals occurring about 100-50 kya.

The exact origin of anatomically modern humans in Africa remains unknown, mainly because of the scarcity of fossil and archaeological data in the tropical regions of the continent. Two are the hypothesis more debated, one suggests the origin in a region of East Africa, as suggested by mtDNA data (Metspalu et al., 2004, Torroni et al., 2006), the other proposes a birth place in South Africa (Gronau et al., 2011, Veeramah et al., 2011, Pickrell and Pritchard, 2012, Schlebusch et al., 2012) like some recent data from nuclear genomes point out. More recent data also propose to postpone the origin of modern humans to 260 kya (Schlebusch et al., 2017).

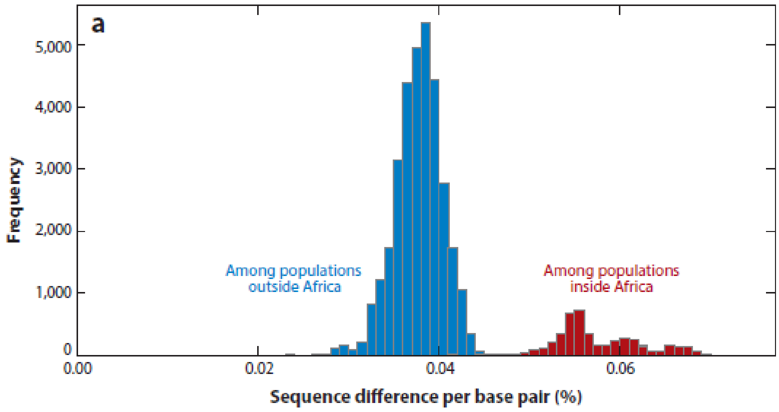


Figure 18. Genomic variability in African (blue) and non-African (red) populations (Skoglund and Mathieson, 2018).

In 2016, the issue n° 538 of Nature published three different works that collected high quality whole genomes from 270 populations across the globe (Malaspinas et al., 2016, Mallick et al., 2016, Pagani et al., 2016) providing a high-resolution portrait of human genetic diversity, allowing new inferences to refine and extend current models of historical human migration out of Africa. These papers support the scenario that all contemporary non-Africans branched off from a single ancestral population, possibly with minor genetic contributions from an earlier modern human migration wave into Oceania (Pagani et al., 2016).

Soon after the “Out of Africa”, the AMH met, perhaps in southwestern Asia, the Neanderthals (Prüfer et al., 2014) and a gene flow between them occurred. The Neanderthals contributed up to around 2% of the genome of all non-Africans, but other gene flows with archaic hominins in different regions around Eurasia took placed, as showed by the presence of up to 5% of Denisova genome (an extinct archaic human who lived about 40 kya found in the Denisova Cave in the Altai Mountains in Siberia) in present-day Oceanian populations (Reich et al., 2010, Meyer et al., 2016, Skoglund and Reich, 2016, Vernot et al., 2016). These contributions imply that not all present-day non-African ancestry is nested within African variation, referred to as a “leaky replacement” model of Eurasian human population history (Gibbons, 2011).

After the out of Africa, the ancestral populations expanded across Eurasia and occupied different lands (Figure 19) with multiple founder effects and bottlenecks, which explains the reduction of genetic variability from Africa to Eurasia (from west to east in Eurasia) and then in Oceania and the Americas (Ramachandran et al., 2005, Liu et al., 2006, DeGiorgio et al., 2009).

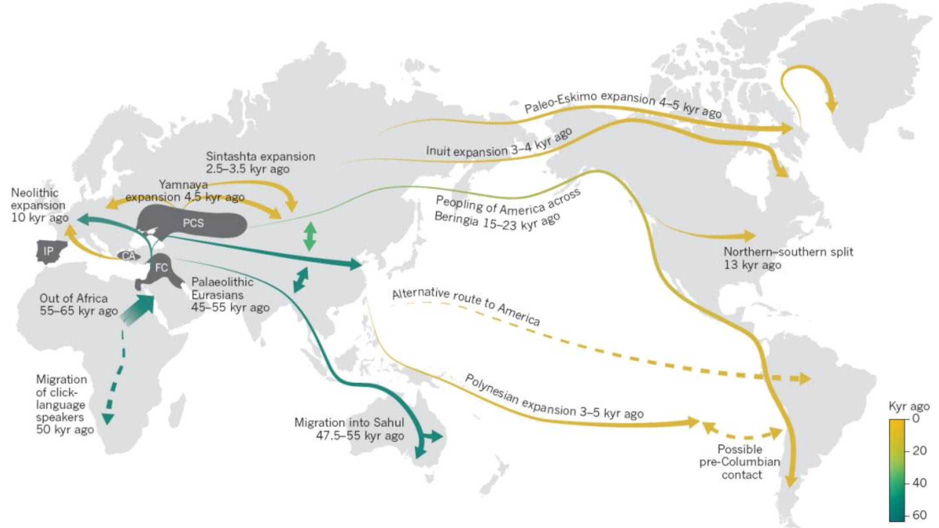


Figure 19. Major human migrations across the world inferred through analyses of genomic data. Some migration routes remain under debate. For example, there is still some uncertainty regarding the migration routes used to populate the Americas. Genomic data are limited in their resolution to determine paths of migration because further population movements, subsequent to the initial migrations, may obscure the geographic patterns that can be discerned from the genomic data. Proposed routes of migration that remain controversial are indicated by dashed lines. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe (Nielsen et al., 2017).

Nowadays, archaeogenomics data allow to clarify the overall scenario directly analyzing ancient remains, providing calibration points for the molecular clock, and allowing for a subtler understanding of the admixture with other hominins (Skoglund and Mathieson, 2018).

After the meeting with Neanderthals, dated back to around 55 kya, the AMH divided in different, at least three, branches: one populated Europe (Upper Palaeolithic Europeans), another ancestral to East Asians and Aboriginal Australians, the third is an extinct linage, found in the Ust'-Ishim remains, a ~45,000-year-old modern human male from Siberia. This split occurred before the admixture of the eastern branch with the Denisova, dated around 50 kya and after the origin of Ust'-Ishim, as also confirmed by different molecular clocks: 50-45 kya for the mtDNA (Fu et al., 2013b, Posth et al., 2017), 47-55 for Y-chromosome (Karmin et al., 2015, Poznik et al., 2016) and 47-50 for genomic data (Pagani et al., 2016, Terhorst et al., 2017).

1.3.2.2 Europe

Europe is the most studied continent and the genetic structure of modern Europeans have been explored by analyzing both uniparental and biparental markers, highlighting multiple admixtures and replacement events that occurred in the last 40 kya (Figure 20). The current genomic picture points to three (or more) genetic components in the gene pool of European populations mirroring different migration events in the continent (Schlebusch et al., 2012, Gamba et al., 2014, Lazaridis et al., 2014, Allentoft et al., 2015, Günther et al., 2015, Haak et al., 2015).

1) An early modern human from the Peştera cu Oase (Romania) (Oase 1, dated 37.6-41.6 kya) confirms that AMHs reached Europe as early as 43 kya. However, this ancestry does not represent the Upper Palaeolithic Eurasian component that we can observe now, because major transformations occurred in the Palaeolithic period (as showed also by the mtDNA). The first major replacement occurred during the Gravettian culture period (31–26 kya), followed by the dispersion of Western European Hunter-gatherers in the Bølling-Allerød interstadial warm period, around 15 kya, which also marked genetic gene flows between European and Near Eastern populations.
2) After the LGM, about 11 kya, a new lifestyle based on animal husbandry, agriculture and sedentarism, known as a Neolithic revolution, started to emerge in several sub-regions of the Fertile Crescent. Analyses of ancient DNA showed that these farmers expanded from Central Anatolia into Europe, while other regions around the Fertile Crescent showed a limited genetic contribution to early European farmers. These events, already detected by Cavalli-Sforza in the '64 (Cavalli-Sforza et al., 1964), did not lead to a complete replacement of the European hunter-gatherers, but to a dynamic cultural and genetic admixture. An important (unknown) postulated population that arrived in Europe from the Near-East during Neolithic is the so-called “Basal Eurasians”. The absence of any Neanderthal ancestry also reduced the archaic contribution in modern Europeans. Due to the lack of aDNA evidence about this population, the debate about its origin is still ongoing (Lazaridis, 2018).
3) The third component (or Steppe component) is associated to a massive migration of the bearers of the Yamnaya culture (Pit-grave culture) that expanded around 4.5 kya from the Eurasian steppe. This component was itself a mixture of Mesolithic hunter-gatherers (Eastern European hunter-gatherers, EHU), Caucasus hunter-gatherers (CHU) and farmers population from Iran and modern Armenia. The Yamnaya “invasion” has supposedly led to the formation of the Corded Ware cultures (a characteristic pottery of the era) in central Europe and thereby to the dispersal of Proto-Indo-European languages in Europe. In fact, this component was distributed across Europe during the population movements of Copper and Bronze Ages.

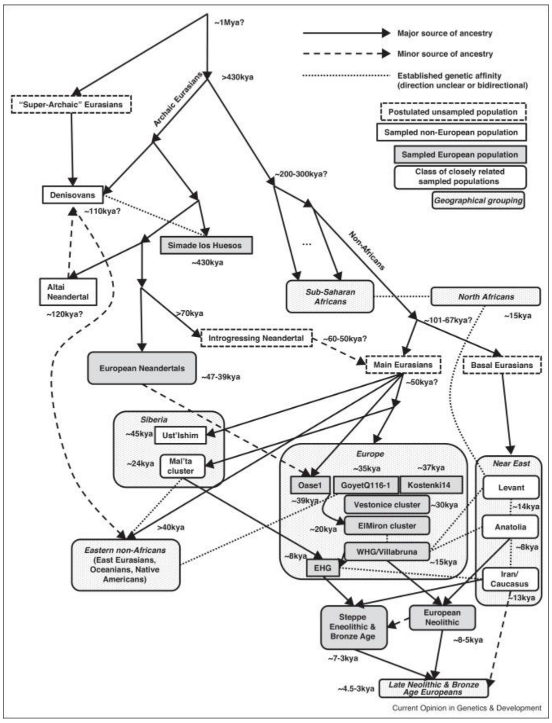


Figure 20. A sketch of European evolutionary history based on ancient DNA that shows the complexity of the European ancestry (Lazaridis, 2018).

1.3.2.3 Asia and Oceania

It is still under debate how many migration waves of AMH colonized Asia and Oceania. Even if assuming a first migration from Africa to Oceania, based on mtDNA data, the main settlement occurred after this "Out of Africa" wave.

The peopling and the genetic structure of Central and East Asia is, so far, a big mystery principally due to the lack of samples, especially aDNAs. The representative ancient sample of the eastern branch of the Out of Africa split is a 40 ky old individual from China (Tianyuan) (Fu et al., 2013a). This genome, together with the aDNAs representative of the Western branch mentioned above, demonstrated that the Eastern Upper Palaeolithic populations originated around 45-36 kya. However, also the Western branch has reached the eastern part Asia, in particular Siberia, as demonstrated by two recently published genomes from Yana RHS site in North-eastern Siberia dated ~31.6 kya (Sikora et al., 2018). This ancestral population later differentiated into the Mal'ta culture, represented by Mal'ta and Afontova Gora ancient genomes (~18 kya) (Fu et al., 2016). Then, populations from Central/East Asia reshaped the genetic structure of Siberia, admixing with the first settlers of the region and giving rise to the Ancient Paleosiberians, represented by a 9.8 kya skeleton from Kolyma River (Sikora et al., 2018). Later, the western Eurasian component also contributed to the final genetic structure of Asia with two recent population expansions from Europe. One is associated with the Steppe component (~5 kya) that expanded towards both west and east; the eastward expansion is linked to the Afansaievo culture that also brought the Proto-European languages into India. Finally, about 3.5-2.5 kya, the Sinthashta culture moved from Urals and Europe and admixed with East Asians.

Oceania was colonized about 55-47.5 kya, seemingly throughout a single migration wave. After the first peopling, Papuans and Aboriginal Australians differentiated and remained isolated until recent times. Many years later, a new migration from East Asia was associated with the Lapita culture, around 3.5 kya. A link between Polynesians and Native Americans was observed, but the origin of this connection is still under debate (see below).

1.3.2.4 America

The first peopling of the Americas (Figure 21), the migrations within the continent before the arrival of European colonists and the genetic structure of current inhabitants after the admixture with other populations arrived in the last five centuries are largely debated and studied. A comparison with other fields, such as archaeology and linguistics, is necessary to completely assess the first peopling of America.

Morphological analyses of early archaeological remains in the Americas have suggested that characteristics of some Pleistocene and early Holocene skeletons are different from present-day Native Americans and fall within the variation of present-day indigenous people in the South Pacific, such as Australians, Melanesians, Polynesians, and Negritos (Neves and Pucciarelli, 1991, Owsley and Jantz, 2014). This led to the hypothesis of a common ancestry between Paleoamericans and Australasians. A significamt replacement by Northeast Asian populations in the early Holocene left some legacies of the original chracteristics only in some locations of South America, such as the extinct Pericúes (aboriginal inhabitants of the Cape Region, the southernmost portion of Baja California Sur, Mexico) and Fuego-Patagonians (in the southern cone). Another archaeological hypothesis based on the affinity between the Clovis lithic technologies (the earliest well-characterized archaeological assemblage) of North America and the European Solutrean culture brought to the conclusion that around 21-17 kya modern humans reached the American continent along the pack ice of the North Atlantic Ocean (Stanford and Bradley, 2012). However, neither the Paleoamerican nor the Solutrean models have been supported by genetic researches.

As previously mentioned, crucial information about the early genetic history of America was based on uniparental markers, but the emergence of modern and ancient genomic data is improving the knowledge about the first settlement and the following migrations (Nielsen et al., 2017, Achilli et al., 2018, Skoglund and Mathieson, 2018).

Many analyses of Native American genetic diversity suggest a/some migratory wave(s) from an ancestral population that lived in Beringia during the LGM, probably coming from East/Central Siberia (Zegura et al., 2004, Schroeder et al., 2007, Tamm et al., 2007, Wang et al., 2007, Kemp and Schurr, 2010, Hoffecker et al., 2014, Hoffecker et al., 2016). The idea that the Siberian component remained in Asia until the LGM, as initially suggested by the Afontova-Gora genomes (Raghavan et al., 2014a, Fu et al., 2016, Skoglund and Mathieson, 2018), was recently overcome by the identification of the Kolyma1 (9.8 kya) as the most closely related ancient genome to all Native American populations. This ancient individual from Northeastern Siberia, represents the ancient Paleosiberians described above (Sikora et al., 2018).

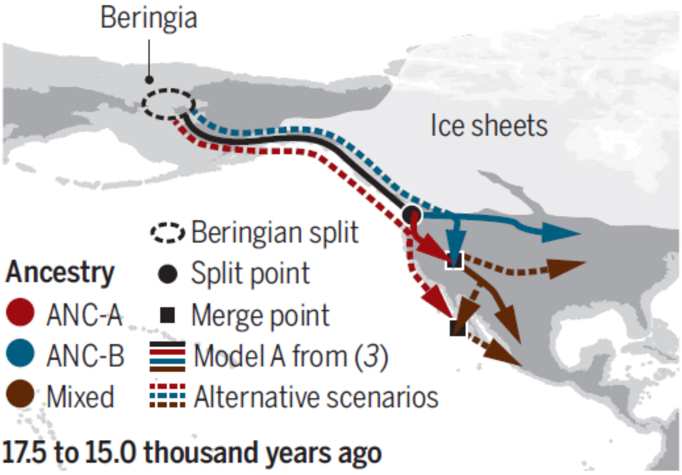


Figure 21. First peopling of America. Ancient Native American genomes suggest different models for the initial settlement of the Americas (Achilli et al., 2018).

Probably, the Paleosiberians diversified in Beringia, experiencing also one or more admixture events with one or more unknown East Asian populations, inside or outside the American continent (Nielsen et al., 2017). This incubation period, named Beringia Standstill (Tamm et al., 2007), continued until the end of the LGM. This differentiation was characterized by various founder events in different isolated Beringian refuge areas (Goebel and Potter, 2016). Therefore, some novel components arose *in situ* and characterized the genetic structure(s) of the first populations that peopled the American continent. Controversial anthropometric and linguistic data, postulates that the Americas were settled through three separate population movements, the so called "tripartite migration model". Originally, it was proposed in 1980' and mtDNA evidence, as showed above, supported this theory. This model recognizes three population groups deriving from different migration waves and expressed in linguistic terms as Amerinds, Na-Dene, and Eskimo–Aleut speakers (Reich et al., 2012, Achilli et al., 2013, Raghavan et al., 2014b, Raghavan et al., 2015).

However, this "tripartite migration model" seems too simplistic when considering the increasing genomic and archaeogenomics data. In particular, four papers, published in Nature, Science and Cell, have recently added more details to the overall scenario of the first peopling of America (Moreno-Mayar et al., 2018a, Moreno-Mayar et al., 2018b, Posth et al., 2018, Scheib et al., 2018). The first Paleoindian wave is testified by the ancient genome of a male infant (Anzick-1) recovered from the Anzick burial site in western Montana, dated to 12.6 kya (Rasmussen et al., 2014), and associated with the Clovis culture, the first well-characterized prehistoric Native American culture (Jenkins et al., 2012). The genetic source represented by Anzick-1 is usually named with different appellatives, such as South Native American (SNA) or Ancestry-A (Anc-A) and could be tentatively associated with the "Amerindian dispersal". The early Paleo-Americans rapidly spread throughout the double continent, diversifying into many hundreds of culturally distinct populations and tribes. The timeframe and exact routes are still matters of debate, and the model faces continuous challenges. However, this initial costal migration wave along the entire double continent is also supported by some informative archaeological sites in South America, such as Pikimachay in Peru (dated ~ 14-12 kya) and Monte Verde in Chile (dated ~ 20-14 kya) (Gilbert et al., 2008, Dillehay et al., 2015). Another ancestry, present only in North America, is represented by "Kennewick man", an ancient genome founded in the Washington state and dated to 9.5 kya (Rasmussen et al., 2015). The ancestral component represented by Kennewick is named North Native American (NNA) or Ancestry-B (Anc-B) and is now represented also by a group of ancient genomes from South Ontario (ASO) in Canada dated around 4 kya (Scheib et al., 2018). The route followed by this second ancestral population(s) as well as any possible linguistic association is still under debate. Similarly, the geographic location of the initial split of these two ancestries is still unclear. Two ancient samples found in the Upon San River site in Alaska and dated around 11 kya (Moreno-Mayar et al., 2018a) have been associated to an Ancestral Beringian (AB) component, which was probably not involved in colonization of the continent, but might support the hypothesis that the initial split occurred to the southeast of Beringia, but north to the North American ice sheets ~17.5-14.6 kya (Lindo et al., 2017, Posth et al., 2018). Other two models suggest that the split happened south of the ice sheets (Scheib et al., 2018) with a later migration of Anc-B (~9kya) towards north (Moreno-Mayar et al., 2018b). Even if the original Anc-A component was involved in the first peopling of the entire double continent (light blue line in Figure 22), it was later replaced only in central and south America by a new (unknown) component derived from Anc-A at least 9 kya (Moreno-Mayar et al., 2018b, Posth et al., 2018) (green line in Figure 22). More recently, around 5 kya, a Mesoamerican population spread towards both North and South America (as testified in the south by the so-called Late Ancestral Andes, orange line in Figure 22) changing again the genetic structure of Native Americans (Moreno-Mayar et al., 2018b, Posth et al., 2018).

Some genetic studies have also revealed a genetic affinity between Amazonian and Australo-Melanesian populations (Raghavan et al., 2015, Skoglund et al., 2015). Since a trans-Pacific migration wave seems unlikely, this finding suggests that the Native American ancestors might have genetic similarities with an eastern Asian population(s) also related to modern Australo-Melanesians (Skoglund et al., 2015, Yang et al., 2017).

Actually, the genome of a 40 ky old individual from China (Tianyuan) shows affinities with the Amazonians (Skoglund and Mathieson, 2018), thus lending support to the idea of an ancient substructure in East Asia that contributed to the connection between Native Americans and Australo-Melanesians, still detectable in some isolated populations of Amazonia (Yang et al., 2017). This component brought in America by a "ghost lineage named *Population Y*" was already present 10 kya in Brazil but is too ephemeral to demonstrate differences between ancient and modern Native Americans (Moreno-Mayar et al., 2018b).

Later, two additional migration waves affected only North America, mostly in the northernmost part. The first ancient genome ever published was from a 4-kilo-year-old individual, associated with the Saqqaq culture in Greenland, that showed more genetic affinity with Siberian populations than with modern Inuits (Rasmussen et al., 2010). A different origin for Paleo- and Neo-Eskimos has been later supported by the analysis of additional genomes (Raghavan et al., 2014a). Therefore, the final model postulates a dual migration in circumpolar regions, where the initial Paleo-Eskimos populations were replaced by the advent of Neo-Eskimos.

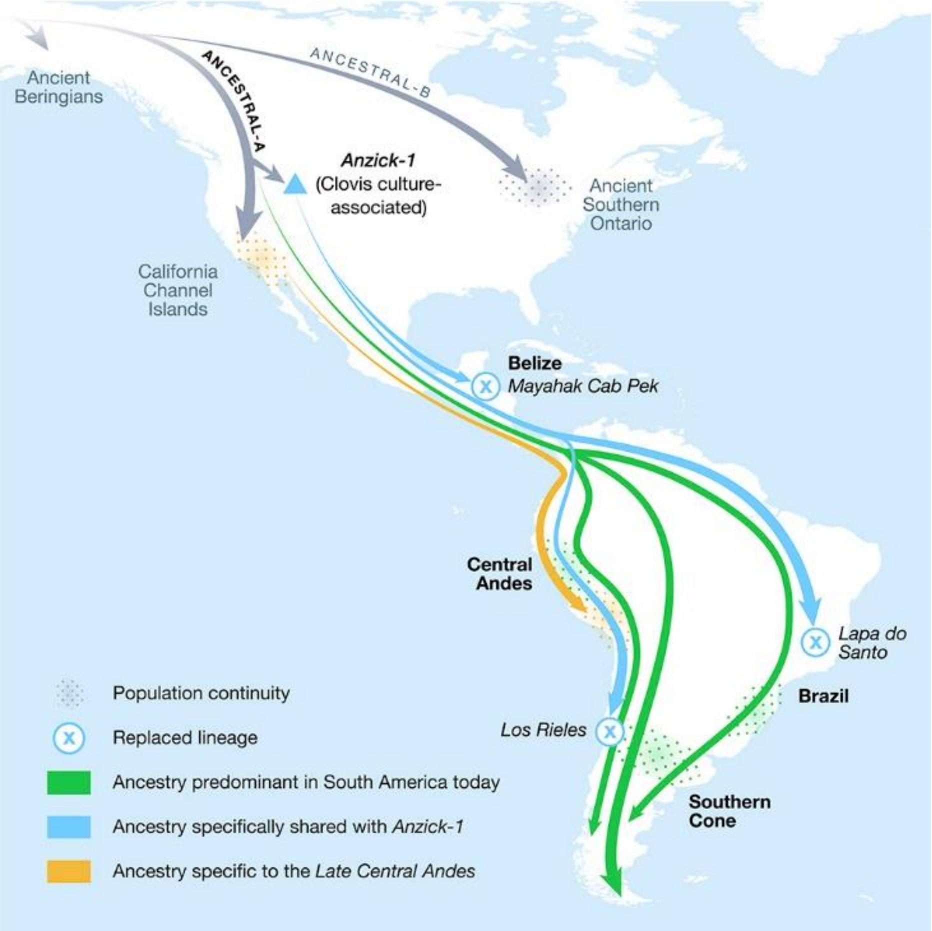

ANCESTRAL-B
ANCESTRAL-A
Ancient Beringians
Anzick-1 (Clovis culture-associated)
Ancient Southern Ontario
California Channel Islands
Belize
Mayahak Cab Pek
Central Andes
Lapa do Santo
Brazil
Los Rieles
Southern Cone
Population continuity
Replaced lineage
Ancestry predominant in South America today
Ancestry specifically shared with Anzick-1
Ancestry specific to the Late Central Andes

Figure 22. Migrations involved in the first peopling of the American continent and following ancient replacements (Posth et al., 2018).

1.4 A strong link between modern humans and other animals

A very important event of the Neolithic period was the advent of agriculture and pastoralism that created a tight link between humans and animals. After domestication, they often undertook parallel evolutionary paths whose signs can still be retrieved from modern and ancient spectra of DNA variation.

Although humans migrated to every habitable continent before the advent of plant and animal domestication, population sizes were small and most people lived nomadically. The development of agriculture and farming allowed for a dramatic expansion of the global human population size (Bocquet-Appel, 2011). Dissecting the domestication processes that accompanied this demographic shift could be a key to understanding the origins of modern human societies (Larson and Burger, 2013).

The domestication of plants and animals began at least 15 kya with wolf (*Canis lupus*) and triggered a rapid and profound shift in the evolution, ecology and demography of both *Homo sapiens* and numerous animal and plant species (Larson et al., 2014). Zooarchaeological evidence for the early domestication centers in Southwestern Asia suggests that goats (*Capra hircus*), sheep (*Ovis aries*), cattle (*Bos taurus*) and pigs (*Sus scrofa*) were among the first domesticated livestock, approximately 10 kya (Conolly et al., 2011, Asouti et al., 2013, Larson and Burger, 2013), while the horse (*Equus caballus*) was probably domesticated in Central Asia approximately 5.5 kya (Outram et al., 2009) and its spread was linked with the human diffusion from steppe.

Domestication was a complex and gradual process, which altered the behaviour and the morphological characteristics of ancestral animals. This event was likely triggered by the ubiquitous tendency of hunter-gatherers to tame and manage wild animals (Diamond, 2002). However, at the end of the Palaeolithic, the process of domestication got underway, since changes in the climate, which became more unpredictable and warmer, led to localized expansion of human populations.

The ancient wild ancestors of the majority of livestock have now been identified. It is also known that many domestic animals originated from more than one wild population and that in some cases genetic admixture or introgression took place after the initial domestication. Once agricultural societies emerged, they often migrated away from the domestication centers, taking their domestic partners with them. However, the domestic populations were small relative to the surrounding wild groups, thus, repeated hybridizations between the two eventually led to the derived domestic population (trough introgression) becoming more genetically divergent from its original source population (Currat et al., 2008).

Certainly, the domesticated animals were bred in captivity and largely modified from their wild ancestors to make them more useful to humans, who controlled their reproduction, care and food supply (Diamond, 2002, Mignon-Grasteau et al., 2005).

In the last decades, zoo-archaeologists and geneticists have focused on understanding the genetic and phenotypic changes that have accompanied the domestication process and on documenting the centers of domestication (Zeder et al., 2006, Driscoll et al., 2009).

Three possible domestication models have been described to define presence and size of bottleneck(s) and to estimate the number and geographic distribution of potential ancestral populations (Zeder, 2012) (Figure 23).

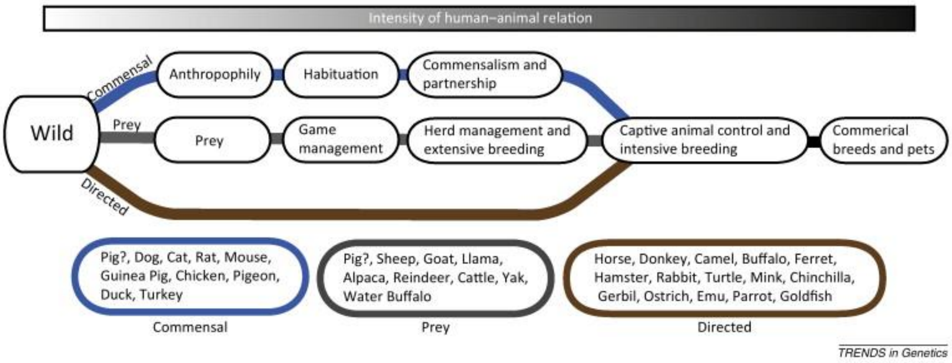

Intensity of human–animal relation
Wild
Commensal
Prey
Directed
Anthropophily
Habituation
Commensalism and partnership
Prey
Game management
Herd management and extensive breeding
Captive animal control and intensive breeding
Commerical breeds and pets
Pig?, Dog, Cat, Rat, Mouse, Guinea Pig, Chicken, Pigeon, Duck, Turkey
Commensal
Pig?, Sheep, Goat, Llama, Alpaca, Reindeer, Cattle, Yak, Water Buffalo
Prey
Horse, Donkey, Camel, Buffalo, Ferret, Hamster, Rabbit, Turtle, Mink, Chinchilla, Gerbil, Ostrich, Emu, Parrot, Goldfish
Directed
TRENDS in Genetics

Figure 23. Representation of the three different domestication models: the commensal (blue), prey (grey) and directed (brown) pathways (Larson and Burger, 2013).

In the commensal pathway, the initial phase involved a gradual adaptation of wild animals to the human presence; it is typical of the first domestication processes.

In the prey pathway, the animals were first hunted away from human settlements and then more directly managed as they were brought into closer proximity to human communities. Because prey animals were typically larger and more difficult to handle than those associated with other pathways, the bottlenecks are expected to be more severe and the process possibly took place over relatively shorter time frames (Bollongino et al., 2012).

In the third scenario, called directed pathway, the domestication happened after humans had been living with livestock for millennia; this group includes several household pets (e.g., hamsters) that were domesticated during the 20th century. This pathway skips the early phases of management and begins with the capture of wild animals with the intention of controlling their breeding. It took place over much shorter time frames and was accompanied by a dramatic bottleneck.

Although the domestication of each animal occurred in different geographical and temporal contexts, these three categories allow for a greater understanding and development of appropriate population genetic models underlying each pathway (Larson and Burger, 2013). It should be noticed that apparently independent and geographically distant domestication events were not necessarily culturally different. Some independent domestication events may have shown the movement of few domesticated animals into a new area, with the genetic signatures of the introduced founders subsequently submerged by the recruitment of local wild animals (Zeder et al., 2006). Alternatively, ancient signatures of local/secondary domestication events may be hidden by more recent arrivals of livestock from other centers of origin.

Osteometric information from archaeological sites and ancient livestock DNA studies represent important tools to address questions about domestication.

Livestock domestication is now thought to have occurred in at least 12 areas of the world. While doubts still surround the existence of certain domestication centers for some species, the following geographic areas (Figure 24) are important primary centers of origin (and diversity) of livestock species (Dobney and Larson, 2006):

- Andean mountains of Southern America (llamas, alpacas, guinea pigs)
- Central America (turkeys, Muscovy ducks)
- Northeastern Africa (donkeys and perhaps cattle,
- Southwestern Asia including the Fertile Crescent (cattle, sheep, goats, pigs)
- Indus valley region (cattle, goats, chickens, riverine buffaloes)
- Southeastern Asia (chickens, Bali cattle)
- Eastern China (pigs, chicken, swamp buffaloes)
- Himalayan plateau (yaks)
- Northern Asia (reindeer).

Additionally, the southern part of the Arabian Peninsula is thought to be the region of origin of the dromedary camel, the Bactrian camel may originate from the area that is now the area of modern-day Iran, and the horse from the Eurasian steppes (MacHugh and Bradley, 2001) (Figure 24).

While domestication occurred in several places, it also happened at different times. However, exact dating of domestication events has proved that the particularly challenging and modern domestic animal populations do not always display a strong phylogeographic structure. For example, this is the case of both the autosomal microsatellites and mtDNAs of horses (Vila et al., 2001, Achilli et al., 2012) and dromedaries (Almathen et al., 2016), probably because the capacity for large-scale dispersal and human movements along transcontinental trade routes have homogenized the original geographic population structure of these species. The discrepancy between divergence and domestication times results from several factors, since contemporary wild populations are not the direct ancestors of domesticated animals and do not necessarily descend from them, as a significant population structure may have existed prior to the onset of the domestication process (MacHugh and Bradley, 2001). Furthermore, the divergence time estimate can reflect other population processes rather than a singular domestication (MacHugh and Bradley, 2001).

However, new archaeological and genetic insights are constantly improving our understanding of the livestock origin. The main factors at the root of the early dispersion of livestock species were the expansion of agriculture, trade and military conquests. The exact mechanisms through which agricultural expansion occurred are still debated. The process probably varied from one region to another (Diamond and Bellwood, 2003). It certainly involved both the movement of human populations and cultural exchanges between populations, as illustrated by the adoption of farming by many hunter-gatherer societies.

Neolithic expansions of early farmers from western Asia represent important examples of processes that brought cattle, sheep and goats into Europe. Domesticated livestock followed two distinct major routes into Europe: the Danubian and the Mediterranean corridors (Bogucki, 1996, Cymbron et al., 2005).

As already mentioned, the first animal to be domesticated was the dog and this probably occurred at least 14 kya. It is unclear where the initial domestication took place, but many maternal lineages have been found in modern dogs, indicating multiple introgressions from their wild ancestor, the grey wolf (*Canis lupus*), but only in Eurasia, because the mitochondrial lineages identified so far in the Americas were of European origin.

Although many genomic studies are still in progress in order to provide a more detailed and comprehensive picture of such extant livestock variability, several levels of phylogeographic structuring have been proposed, deepening phylogenetic branching of the tree topologies of both the mtDNA and the Y chromosome. In particular, thanks to its peculiar features, mtDNA polymorphisms have been extensively used in phylogenetic and genetic diversity analyses, contributing to a significant progress in the initial understanding of livestock domestication (Bruford et al., 2003, Achilli et al., 2012, Achilli et al., 2013).

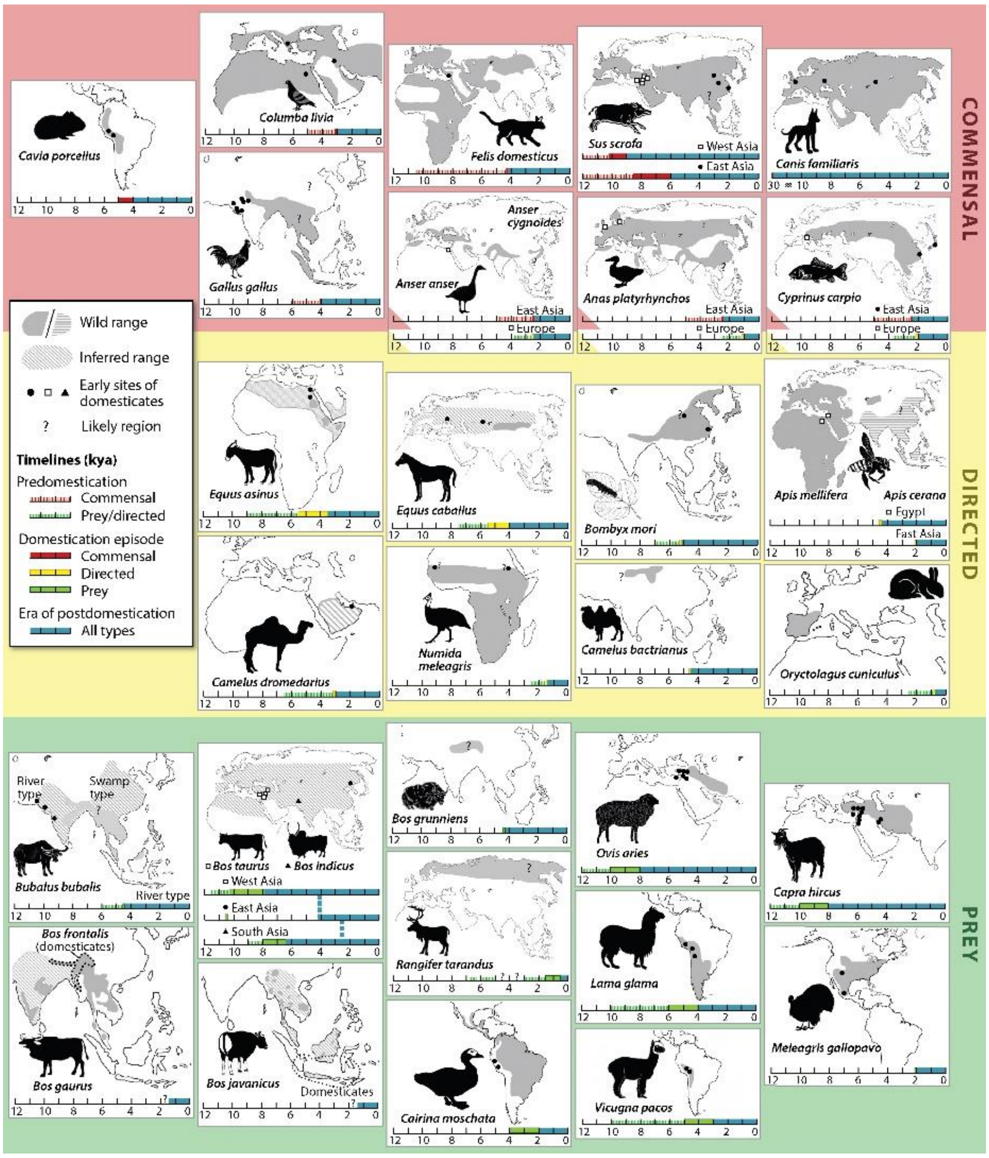


Figure 24. The principal centers of domestication for most animal species. Modified from (Larson and Fuller, 2014).

2. Aim of the thesis

The initial title of my PhD project was “employment of genetic markers to study the variability of human and animal populations.” Consistently with these objectives, during the three years of my research I used the most updated phylogenetic, statistical and molecular tools for “exploring the past of human and animal populations.” Moreover, as the final title of my thesis indicates, I moved forward “from modern mitogenomes to archaeogenomics” analyses. In general, the genetic link between humans and other animals was a constant thread throughout my work.

Mitogenome analyses were initially carried out to improve the maternal phylogenies at the highest level of molecular resolution, the complete mtDNA, in different species and to address various scientific issues. In the first work, presented hereafter, I explored the sequence variation of complete buffalo mitogenomes from China and Southeast Asia to shed light on the domestication and demographic history of the swamp buffalo. A similar approach was employed to build the first mitogenome phylogenies of goats and tiger mosquitos. In other instances, when detailed mtDNA phylogenies were already available, I used the mitogenome tool to investigate the maternal origin of peculiar breeds, such as the Podolian cattle and Italian horses, and to verify the amount of residual variability of an isolated breed, i.e. the Maltese cattle.

As for the humans, I focused on two peculiar microgeographic contexts. The first is the Sardinia Island, in the middle of the Mediterranean Sea, to verify if it had been populated since Mesolithic times, as suggested by some archaeological findings. The second (and main) research objective was the Panamanian Isthmus, a mandatory route between North and South America since Palaeolithic times, a strategic node of the Spanish Empire and a crucial site for early modern globalization. Taking into account the complexity of the Panamanian topic and the recent technological improvements in population genetics and molecular anthropology that nowadays allow for the analysis of whole genomes from ancient samples (archaeogenomics), I spent most of my last year of the PhD program aiming to obtain and analyze the first whole genomes from ancient Panamanian samples.

It is worth mentioning that during my PhD program I have also visited several research institutions in three different countries outside Italy (Austria, Estonia and the state of Illinois in the US) with a double aim: (1) to acquire additional lab expertise moving from classical to next generation sequencing technologies, from mitogenome to nuclear genome and from modern to ancient DNA, and (2) to extend my expertise from basic to computational statistics and bioinformatics.

3. Mitochondrial DNA studies on modern populations

3.1 Methods for analyzing modern mitogenomes

3.1.1 Extraction and amplification

Genomic DNA was extracted from different sources depending on the analyzed species; saliva was the most used biological substrate for humans whereas blood, hair and eventually ear or muscle tissue were used for animals. The DNA was isolated from the other biomolecules following various protocols. Most of the samples were extracted either with commercial kits (e.g. Promega Wizard® Genomic DNA Purification Kit) or with the Promega's Maxwell RSC Instrument for Automated Nucleic Acid Purification. Following the extraction, DNAs were quantified with the Promega Quantus™ Fluorometer and stored at -20°C.

All samples were amplified at least for the D-loop (control region) using species-specific primer pairs; amplifications were verified using 2% agarose gels. The D-loop haplotypes were used to phylogenetically select the most divergent samples for complete sequencing. Eventually, the complete mitogenome was amplified with a Long Range (LR) PCR in two overlapping amplicons (~9 kbp each). The amplification success was evaluated using a 1% agarose gel.

3.1.2 Sequencing methodologies

3.1.2.1 Sanger sequencing

Sanger technique was mostly used for sequencing the mtDNA control region or to cover small gaps in complete mitogenomes generated from NGS sequencing. For this purpose, each amplicon was purified using the ThermoFisher ExoSAP-IT® enzymatic system (Exonuclease I and Shrimp Alkaline Phosphatase) in order to degrade unincorporated nucleotides and residual primer dimers. Species-specific internal primers were designed and used for Sanger sequencing in outsourcing. The row data were provided as *.ab1* (Applied Biosystem) files.

3.1.2.2 Next generation sequencing

Next generation sequencing was used for the analyses of complete mitogenomes. The initial purification of the LR amplicons was performed with the 96 Well PCR Cleanup Kit (Geneaid Presto^{TM}, Taiwan) or Promega ReliaPrep^{TM} (USA) columns. During my PhD program I took part in several projects involving Second Generation Sequencing (SGS) technologies. In my second year I spent three months at the Legal Medicine Institute of University of Innsbruck (Austria) to sequence highly degraded DNA samples from Central Italy using the Precision ID mtDNA Whole Genome

Panel protocol (ThermoFisher, USA) for the Ion Torrent Personal Genome Machine (PGM).

These preliminary results are not presented in this thesis, where most of the mitogenomes were produced with the SGS Illumina technology. The libraries were initially prepared using the Illumina Nextera® XT DNA library preparation kit, a system that uses transposase to cut DNA and to add adapters to the fragments and allows the multiplexing of up to 384 samples with 39 primer pairs. The libraries were verified and quantified using the Agilent Bioanalyzer 2100, a micro-capillary based electrophoretic cell that allows rapid and sensitive investigation of nucleic acid samples and run on the Illumina® MiSeq platform at the Fondazione Mondino (Istituto Neurologico Nazionale a Carattere Scientifico) in Pavia (Prof. Cereda group). This bench machine is specific for small genome and can generate, depending on the reagent kit used (150 or 300 cycles), a pair-end output between ~5 and ~15 Giga bases (Gb). Considering our multiplexing strategy to sequence 288 mtDNAs (17 kbp each) in a single run, we obtained an average coverage depth of ~900-1000X.

3.1.3 Sequence analyses

3.1.3.1 Sanger data

The .ab1 files containing the chromatograms were assembled, aligned and compared to the species-specific reference sequence using the software Sequencher™ 5.0 (Gene Codes). Eventually, the control-region haplotypes were determined and used to obtain the haplogroup classification. For humans they were classified through HaploGrep 2 (Weissensteiner et al., 2016). This program provides an automatic mtDNA haplogroup classification using PhyloTree 17, the latest version of the mtDNA classification tree estimated from worldwide data (van Oven and Kayser, 2009). For other animal species that already have a phylogeny, the classification was carried out building a network and through an accurate analysis of diagnostic mutational motifs identified in the control-region haplotypes. In absence of an already defined phylogeny the network was used to cluster haplotypes in haplogroups that were later confirmed and deepened with complete mitogenomes analysis.

3.1.3.2 Next generation data

The Illumina MiSeq system generates raw data files in binary base call (BCL) format. These BCL files were converted to FASTQ files, one for each sample (demultiplexed), using the Illumina *bcl2fastq* conversion software. The FASTQ files

were then analyzed through a Unix-based pipeline based on free software. A pipeline is a sequence of processes chained together by their standard streams, so that the output of each process (stdout) feeds directly as input (stdin) to the next one.
The first phase of the pipeline is the trimming of Illumina adapters with the software *trim_galore* (Krueger, 2015). The *-q 30* flag allows trimming low quality ends from reads, Phred score lower than 30, in addition to adapter removal:

Command 1:

```
trim_galore --paired -q 30 --nextera
${filename}${seqR1} ${filename}${seqR2}
```

The trimmed FASTQs were checked with the software FastQC (Andrews, 2010). It aims to provide simple quality checks on raw sequence data; in particular, we are interested in the presence of adapters and k-mers (words of k nucleotides). For the analyses of k-mers the specific option was turned-on in the *limits.txt* files (leaving the default values for the k size) present in the software configuration folder.

Command 2:

```
fastqc --threads 64 -f fastq
${filename}_L001_R*_001_val_* -o $fastqc
```

Trimmed FASTQs passing the quality check were mapped against the reference sequence (depending on the analyzed species) using the software Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010), with the algorithm MEM that is specific for long reads (>70bp), generating a Sequence Alignment Map (SAM) file. The *-R* option is used to change the header of the reads instead *-t* is for the parallel analysis in four threads of execution.

Command 3:

```
bwa mem \
-R
"@RG\tID:${filename}\tSM:${filename}\tLB:Illumina
\tPL:Illumina" \
-t 4 \
$mtDNA \
$trm/${filename}_L001_R1_001_val_1.fq.gz \
$trm/${filename}_L001_R2_001_val_2.fq.gz >
$sam_folder/${filename}.sam
```

With SAMtools (Li et al., 2009) the SAM file has been converted to Binary Alignment Map (BAM) file and duplicates mainly produced during the PCR phases have been removed. Considering that all our libraries where built starting from PCRs, a better duplicates removal was then made with another software package Picard (broadinstitute.github.io/picard), using the *MarkDuplicates* implement.

Command 4:

```
java -jar /picard-tools-1.54/MarkDuplicates.jar \
I=${filename}.nSrt.bam \
O=${filename}.picard.bam \
REMOVE_DUPLICATES=true \
METRICS_FILE=$metrics/picardmetrics.${filename}.t
xt
```

When the BAM files became ready two different calls were performed, one with Genome Analysis ToolKit (GATK) (McKenna et al., 2010), that is a platform with various tools to analyze NGS data developed by the Broad Institute (https://www.broadinstitute.org/). The application used for the variant calling is HaplotypeCaller, alongside flag -ploidy 1 for the haploid genome feature of our mtDNA of interest. The output of this step is a Variants Calling File (VCF).

Command 5:

```
java -Xmx20g -jar GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R mtDNA \
-I ${filename}.bam \
-o ${filename}_variants.vcf \
-ploidy 1
```

GATK is also used for defining the covered regions of the genome, with the tool *FindCoveredIntervals*

Command 6:

```
java -Xmx20g -jar $GATK/GenomeAnalysisTK.jar \
-T FindCoveredIntervals \
-R mtDNA \
-I ${filename}. bam \
-o ${filename}.FindCoveredIntervals
```

Another call is made with the option *mpileup* of *SAMtools* and the final files with the haplotype and the coverage range are generated using a homemade software produced by the group of Prof. Luisa Pereira at the University of Porto.

Information about total reads, coverage and average coverage were obtained again using mpileup option in SAMtools, but for a control also the *coverage and depth* options (to calculate both genome coverage and coverage depth) in the *paleomix* panel (Schubert et al., 2014).

Exclusively for human samples, the merged VCF file obtained with GATK and .txt file from mpileup were used to classify the samples with the stand-alone version of HaploGrep 2.0 (github.com/seppinho/haplogrep-cmd), adding the flag *lineage* that produce a *.txt* file for each sample with its phylogeny root from the reference. For the other species, the classification required further analyses, as described below.

3.1.4 Evaluating genetic diversity indexes

Analyses on the genetic variability in the control region and in the complete mtDNA were carried out using the software DnaSP 6 (DNA sequence polymorphism v.6). Among the possible outputs of the software, the haplotype diversity (Hd) with its standard deviation, the nucleotide diversity π (pi) and the average number of nucleotide differences (k) were determined.

Haplotype diversity (also known as gene diversity) represents the probability that two randomly sampled haplotypes are different in a given population (Nei, 1987). Nucleotide diversity (π) is defined as the average number of nucleotide differences per site in pairwise comparisons among DNA sequences and was estimated by assessing windows of 100 bps with step size of 50 bps cantered at the midpoint (Nei, 1987). The average number of nucleotide differences (k) measures the degree of polymorphism within a population and is defined as the mean number of nucleotide differences per site between two DNA sequences randomly chosen from the sample population (Tajima, 1983).

3.1.5 Phylogeographic examination

3.1.5.1 Tree building

Tree building is the first mandatory step in a phylogenetic analysis. I built several trees with different software in order to reveal the phylogenetic relations between samples. A preliminary phase is the sequence alignment, which was mostly obtained via the software Sequencher 5.0 and corrected by hand to resolve misalignment due to indels. For each species, different trees were built with the software Molecular Evolutionary Genetics Analysis (MEGA) (Kumar et al., 2018), with 1000 bootstrapping replicates and using different algorithms (Table 1). In addition, ML trees were obtained with Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang, 1997, Xu and Yang, 2013), while Bayesian inferences were obtained with the software Bayesian Evolutionary Analysis Sampling Trees (BEAST) (Drummond and Rambaut, 2007, Suchard et al., 2018). A specific tool for mtDNA analyses, mtPhyl (eltsov.org/mtphyl.aspx), was used to build MP trees. The classification of human mtDNAs was based on Phylotree 17, while for other species novel phylogenetic trees were manually implemented in this software as a .txt file.

3.1.5.2 Geographic patterns

The phylogeographic approach requires large sample datasets to cover the mtDNA variation of various geographic regions. Until recently, most mtDNA analyses were based only on control-region data, which were re-classified using the most updated level of phylogenetic resolution. Afterwards, haplogroups frequencies were re-calculated and geographically mapped with the program Surfer 9 (Golden Software, goldensoftware.com/products/surfer). Additional information on the genetic proximity of populations were directly obtained from PCA analyses and additional genetic indexes (e.g. Fst values), or by building trees based on complete mtDNAs.

3.1.5.3 Time estimates and demographic reconstructions

The final step is dating the nodes of the tree to understand timing of divergence (or coalescence) of each node/haplogroup. Those inferences are particularly robust when considering numerous samples and the entire mitochondrial sequence. Various statistical approaches could be used to obtain this information. The simplest (but still useful) one is the summary statistics rho (or ρ), i.e. the mean number of base substitutions from a common ancestor among a set of derived sequences (Morral et al., 1994, Forster et al., 1996). The error of this estimate is given by the parameter sigma (or σ), which is an estimator of the variance. In particular, the variance of ρ is calculated as $\sigma^2=(\Delta\rho)^2$, then the "2 sigma rule" gives the 95% confidence interval. Both ρ and σ values are converted into years once calibrating the molecular clock; some mtDNA evolution rates were already available (.eg. for humans), while others were estimated directly from the analyzed datasets, i.e. for buffalos and goats. This approach can be applied also to study the variability of specific mtDNA regions, such as D-loop or protein-coding genes, considering the evolution rate of each specific segment.

Novel and perhaps more accurate methods to assess molecular divergences are Maximum Likelihood (ML) and Bayesian statistics. The ML analyses were performed using the software PAML, while for the Bayesian approach we used BEAST. The best evolution model was selected using the software ModelGenerator v.85 (Keane et al., 2006). Again, the molecular clocks allowed converting molecular divergence into time. In order to increase the accuracies (posterior probabilities) of the Bayesian computations different calibration points were used, such as radiocarbon dates of ancient samples or estimates of ancestral nodes available in the literature (fossil data). Such priors allowed also to calculate the evolution rate if not available yet, as for the swamp buffalo. Obviously, different genomic regions, genes or even specific alleles with different mutation rates will have diverse rates of evolution, e.g. faster in the control region than in coding one or for synonymous than for non-synonymous mutations. For this reason, we compared different estimates considering various partitions of the mitogenome (Achilli et al., 2013).

Another important output of BEAST is the demographic trend, which is represented by the posterior distribution of N_e through time in a Bayesian skyline plot (BSP). The plot can be initially visualized with Tracer v1.5, but then converted into more realistic trends considering the different generation time of each species, for example 25 years for humans and 6 years for bovines (Bollongino et al., 2012).

Additional details about materials and methods are provided in the original publications (attached to this thesis) or available on-line as supplementary information of each paper.

3.2 Whole mitogenomes reveal the history of swamp buffalo: initially shaped by glacial periods and eventually modelled by domestication

3.2.1 Background

During my PhD work one of my major projects was about swamp buffalo (*Bubalus bubalis* subsp. *kerabau or carabanensis)* mitogenome (Wang et al., 2017). This subspecies of domestic water buffalo (*Bubalus bubalis*) is one of the most important livestock species in several Asian countries and is used for the production of milk and meat and for draft power in rice cultivation. Most people in the world depend on it for their livelihoods than any other livestock species (FAO, 2014). The domestic water buffalo in Asia is generally divided in two major subspecies, the dairy river buffalo (*Bubalus bubalis* subsp. *Bubalis*, 2n=50) and the draft swamp buffalo (2n=48), which differ in morphology, behaviour and number of chromosomes (Cockrill, 1981, Kumar et al., 2007a). The river buffalo has been selected as a dairy animal with several recognized breeds, spread from the Indian subcontinent to the eastern Mediterranean countries (the Balkans, Italy, and Egypt) and recently have been imported to eastern Asia, southern America and central Africa. (Cockrill, 1974, Kierstein et al., 2004). The swamp buffalo has primarily been used for draught power in a wide area, ranging from eastern India (Assam region), throughout south-eastern Asia and Indonesia to eastern China in Yangtze River valley (Zhang et al., 2016b), and was recently introduced (in the 20th century) into Australia and southern America (Cockrill, 1974). There are no formally recognized swamp buffalo breeds, but regional populations are subdivided into types based on local adaptation or geographic distribution.

Lau and colleagues hypothesized that the wild Asian buffalo (*Bubalus arnee*) originated in mainland Southeast Asia and spread north toward China and west toward the Indian subcontinent, where the river type was probably domesticated (Lau et al., 1998). Details on the domestication dynamics have been debated for years, with two contrasting hypotheses envisioning either a single (Kierstein 2004) or two independent domestication events for river and swamp buffaloes. Mitochondrial DNA polymorphisms, Y-chromosomal markers and nuclear microsatellite data showed a deep genetic divergence of swamp and river buffalo (Barker et al., 1997a, Barker et al., 1997b, Lau et al., 1998, Navani et al., 2002, Kierstein et al., 2004, Kumar et al., 2007b, Kumar et al., 2007a, Lei et al., 2007a, Lei et al., 2007b, Yue et al., 2013, Mishra et al., 2015, Zhang et al., 2016b), which might indicate two independent domestications (Lei et al., 2007b, Yindee et al., 2010). Domestication of river buffalo most likely took place in the Indian subcontinent (Kumar et al., 2007b), whereas swamp buffalo was proposed to originate from the border region between south China and north Indochina (Yindee et al., 2010, Zhang et al., 2016b), although there is no general agreement on the timing of these events. From their domestication center, swamp buffaloes likely dispersed southwestward to Thailand and Indonesia, and northward to central and eastern China (Zhang et al., 2016b), wherefrom they further spread to the Philippines (Zhang et al., 2016b). According to Epstein (Epstein, 1969), Swamp buffaloes were probably introduced to China from bordering areas of south-eastern Asia. At first it appeared in south-western China, in the Yunnan region during the first century Anno Domini (AD), then it gradually spread to the rest of the country (Yue et al., 2013). Yue and colleagues also hypothesized that the south-western Silk Road, connecting Sichuan (central China) via Yunnan (southern China) and Burma (Myanmar) with southern Asia, may have played a role in the trade of livestock, including water buffaloes.

In spite of the importance of this domestic animal, only three complete and reliable mtDNA sequences were deposited in GenBank (NC006295, AF547270 and JN632607) previous to this work; whereas the fragmentary buffalo mtDNA sequences from rDNA, COII, and cytochrome b loci (Amano et al., 1994, Lau et al., 1998, Lei et al., 2007b, Zhang et al., 2016b) might be meaningless to reconstruct any genetic and population history.

In this scenario, I began to investigate the complete mitogenome of water buffalo by completely sequencing the mtDNA of 107 (KX758296-KX758402) Southeast-Asian swamp from China (46), Laos (23), Myanmar (3), Thailand (14), Vietnam (16) and India (5) and one Chinese river buffalo (KX758295). As reported hereafter, these data were analyzed in order to establish the first mitochondrial phylogeny based on complete mtDNAs and to reconstruct the origin, domestication and demographic history of swamp buffalo; further details are available in the attached original version of the paper (Wang et al., 2017).

3.2.2 Results and discussion

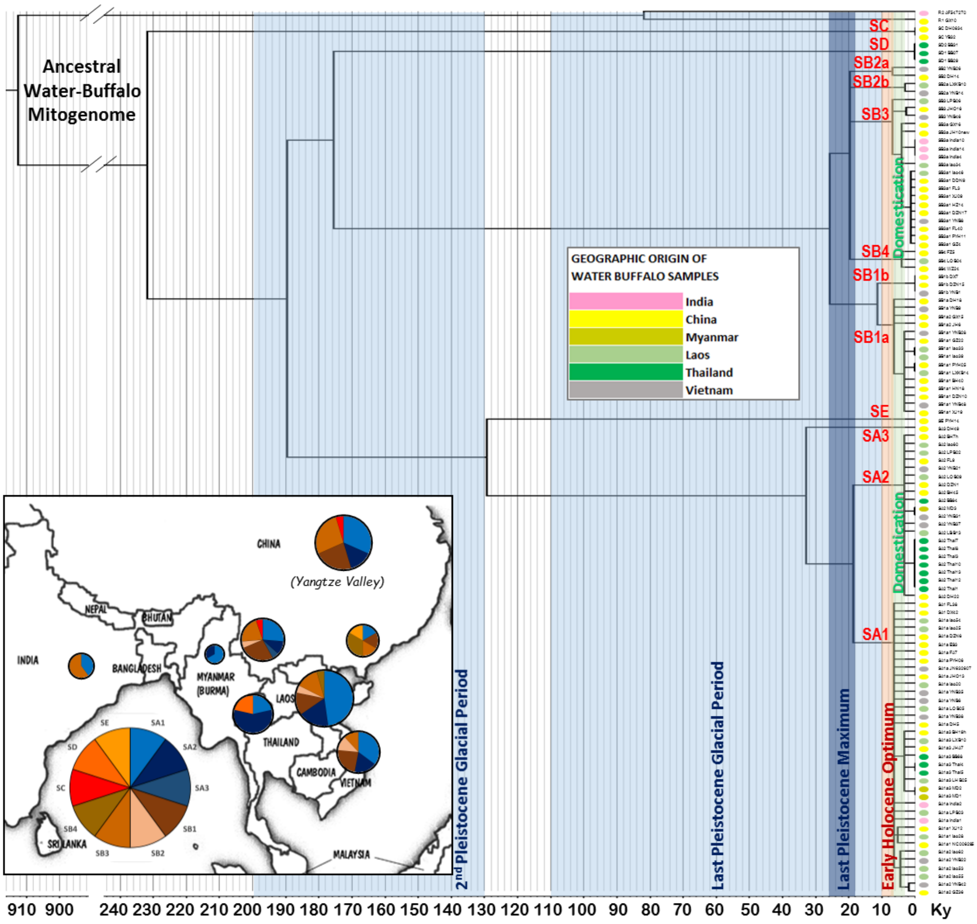


Figure 25. MP phylogeny of complete mtDNAs from 111 buffalo mitogenomes. A maximum likelihood time scale, based on synonymous substitutions, is indicated below the tree. The insert shows the geography distribution of major haplogroups based on these complete mtDNAs.

Our maximum parsimony tree (Figure 25) of water buffalo complete mitogenomes, (111 sequences) shows an initial split of river and swamp buffalo, separated by 300 substitutions. Remarkably, the maximum likelihood and Bayesian trees retrieved similar topologies. The swamp branch largely confirmed previous phylogenies (Yue et al., 2013, Zhang et al., 2016b), but also recognized two novel haplogroups (SA3 and SB4) and 14 new sub-Hgs (SA1a, SA1a1, SA1a2, SA1a3, SB1a, SB1a1, SB1a2, SB1b, SB2a, SB2b, SB3a, SB3a1, SD1 and SD2). Out of a total of 87 swamp haplotypes, 48 belong to haplogroup A (55.2%) and 34 to lineage B (39.1%). Lineages SA and SB are divided into three (SA1 to SA3; SA1’2 as an ancestral node) and four (SB1-SB4; SB2’3’4 as an ancestral node) sub-lineages, respectively. The remaining 5 haplotypes belong to the rare Hgs SC (2), SD (2) or SE (1). The phylogeny confirms that the split-off of lineage SC preceded the SA-SB divergence and that SD is a sister clade of SB, but also indicates that SE and SA are sister clades. The divergence pattern evaluated on all 114 bovine mitogenomes confirms a saturation of the D-loop variation, emphasizing that whole mitogenome data are essential for quantitative analysis (Figure 26).

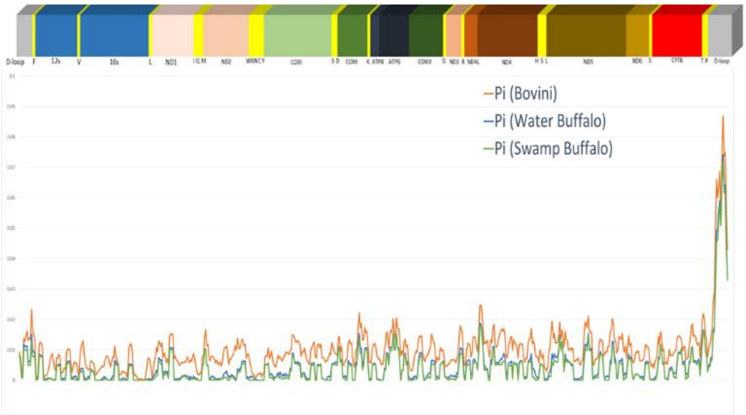


Figure 26. Linearized gene map and nucleotide diversity along the entire mtDNA.

Previously reported estimates of the divergence time between river and swamp range from 1.7 mya (million years ago) to 10 kya (Amano et al., 1994, Tanaka et al., 1995, Tanaka et al., 1996, Barker et al., 1997a, Lau et al., 1998, Kierstein et al., 2004, Kumar et al., 2007a, Lei et al., 2007b, Yue et al., 2013, Mishra et al., 2015), partially because different mtDNA segments were analyzed and different evolution rates were applied.

In the present study, the mtDNA molecular clock was recalibrated by building a buffalo tree, with 111 water buffalo mitogenomes and one African buffalo (*Syncerus caffer*; NC020617), rooted with one *Bos taurus* (V00654.1) and one ancient *Bos primigenius* (GU985279) mitogenome (Figure 27). A fossil age estimate of the *Bovini* tribe of 8.8 mya (Bibi, 2013) and the age (6.7 ky) of the ancient *Bos primigenius* mtDNA (Edwards et al., 2010) were used as external and internal calibration points, providing a final ML estimate of 3.75±0.47 × 10^{-5} synonymous substitutions per nucleotide per ky for 3790 amino acid codons (or 1 synonymous substitution every ~7.030 ky). This value was used to date at ~913±78 kya the root of the water buffalo tree, representing the ancestral water buffalo mitogenome (AWM, Figure 25), as well as all the major nodes of the swamp phylogeny, which were also estimated by considering additional clocks (Table 3).

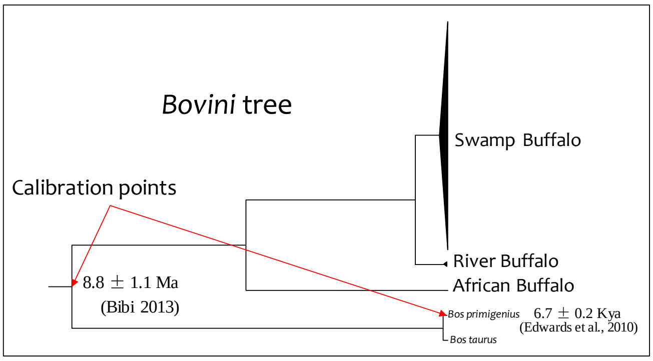


Figure 27. The *Bovini* tree used to calibrate the mtDNA molecular clock.

Table 3. Age estimates of major buffalo branches based on different mitochondrial datasets.

Node	**N**	**ML (syn.s sub.s)**		**ML (only coding region)**		**ML (all sub.s)**[a]		**Beast (all sub.s)**[a]	
		T(ka)	ΔT(ka)	T(ka)	ΔT(ka)	T(ka)	ΔT(ka)	T(ka)	ΔT(ka)
AWM	111	912.6	78.3	1044.3	71.4	672.0	101.7	721.6	59.1
River	2	81.9	18.3	109.3	20.7	63.7	13.2	68.6	11.6
Swamp	109	231.8	35.3	280.1	28.8	194.2	31.4	204.4	20.0
SC	2	0.0	59.1	3.8	3.8	2.9	1.4	1.5	1.5
SA'B'D'E	107	189.6	27.2	222.9	25.0	166.6	26.8	176.6	17.7
SA'E	60	129.3	21.2	152.7	20.9	120.6	20.7	129.1	15.5
SE	1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
SA	59	33.0	9.2	40.5	9.9	35.9	7.9	37.5	7.3
SA1'2	56	18.8	6.5	23.9	7.4	18.5	4.6	18.4	4.5
SA1	37	6.4	4.2	8.1	5.4	8.4	2.2	7.2	1.8
SA1a	35	6.4	1.5	8.1	1.7	6.7	1.5	5.5	0.9
SA1a1	3	5.2	1.6	6.8	1.8	4.1	1.2	4.8	1.0
SA1a2	6	4.4	1.7	5.7	1.9	4.2	1.0	4.5	1.1
SA1a3	9	3.3	2.4	4.4	2.5	4.2	1.0	3.9	1.2
SA2	21	3.3	1.5	3.6	1.6	7.0	1.8	5.1	1.2
SA3	1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
SB'D	47	175.5	27.1	210.0	25.4	146.4	24.3	157.7	16.9
SD	3	0.0	44.6	0.0	44.8	3.2	1.3	0.9	1.0
SB	44	25.8	7.0	35.1	8.3	31.1	6.8	30.7	5.9
SB1	18	11.4	6.2	12.2	4.9	8.9	2.7	8.6	2.4
SB1a	15	6.4	4.4	8.0	2.7	6.0	1.4	5.8	1.4
SB1a1	11	3.3	2.0	4.2	2.1	4.5	1.0	3.7	1.2
SB1a2	2	6.4	4.5	6.5	2.4	3.8	1.0	4.9	1.4
SB1b	3	0.0	13.0	0.0	13.1	3.5	1.3	1.2	1.4
SB2'3'4	26	19.8	5.7	26.1	6.7	23.4	5.3	23.3	4.7
SB2	4	19.8	10.6	26.1	12.5	10.8	4.0	15.7	4.5
SB2a	2	3.1	3.1	6.7	4.8	3.9	1.4	6.0	2.9
SB2b	2	6.8	6.8	7.0	7.0	5.2	2.3	9.9	4.5
SB3	19	6.9	3.4	7.2	3.5	6.6	1.7	4.7	1.8
SB3a	16	4.1	3.1	4.6	3.0	5.0	1.1	2.4	1.0
SB3a1	10	1.2	0.9	1.8	1.1	3.9	0.8	1.4	0.6
SB4	3	4.1	2.9	8.7	4.2	4.7	1.7	5.1	2.2

[a] The entire genome was partitioned into coding and control region.

Notably, at least 12 different mtDNA ancestral Hts of the current swamp buffalos were present in Southeast Asia during the early Neolithic (10-8 kya), which overlaps with the initial phase of domestication (Patel e Meadow, 1998). These haplotypes were the ancestors of the eight sub-Hgs SA1, SA2, SB1a, SB1b, SB2a, SB2b, SB3 and SB4, still common in modern herds, and of the rare SA3, SC, SD and SE.

A Bayesian approach was also employed to analyze demographic trends over time. The final BSP, based on swamp buffalo mitogenomes, shows three major changes in the effective female population size (Figure 28):

i) a slight decrease between around 200 and 130 kya

ii) a more recent decrease starting around 25-20 kya and much steeper during the early Neolithic (10-3 kya)

iii) a rapid increase from 3 kya. This recent increase explains the star-like topology of some Hgs (e.g. SA1, SA2, SB1a, SB3 and SB4; Figure 25).

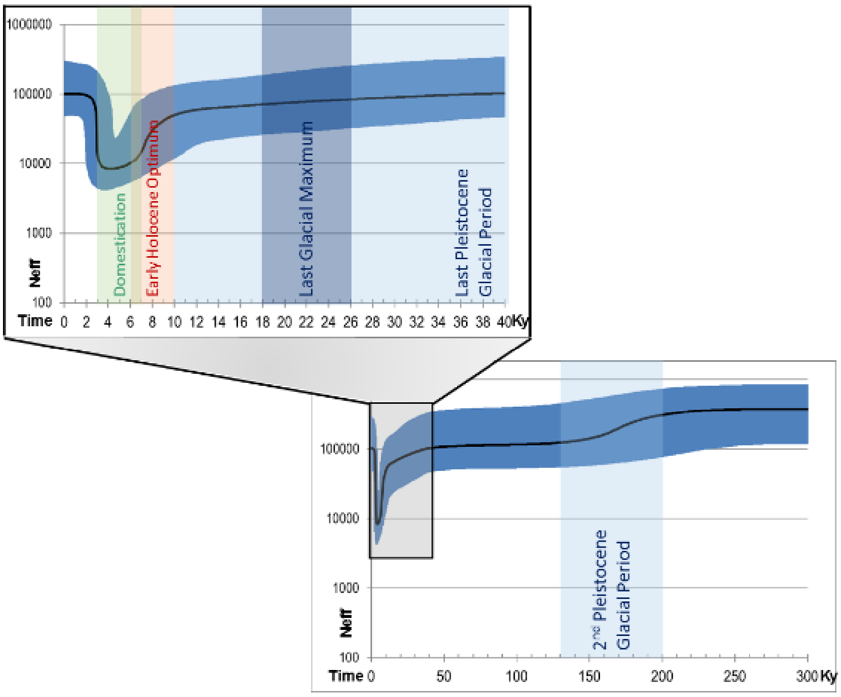


Figure 28. Bayesian Skyline Plot showing the swamp buffalo population size trend. The Y axis indicates the effective number of females, as inferred from our mitogenome dataset considering a generation time of six years (Bollongino et al., 2012). The black solid line is the median estimate and the blue shading shows the 95% highest posterior density limits.

An analysis of the geographic distribution of swamp haplogroups in Southeast Asia, based on the control-region data currently available in literature (Zhang et al., 2016b) or deposited in GenBank, and integrated with our mitogenomes (Figure 29), confirms a predominant and widespread distribution of SA1. However, a geographic differentiation of some swamp lineages has been also revealed, as shown by the contrasting patterns of SA2, SB and SB3.

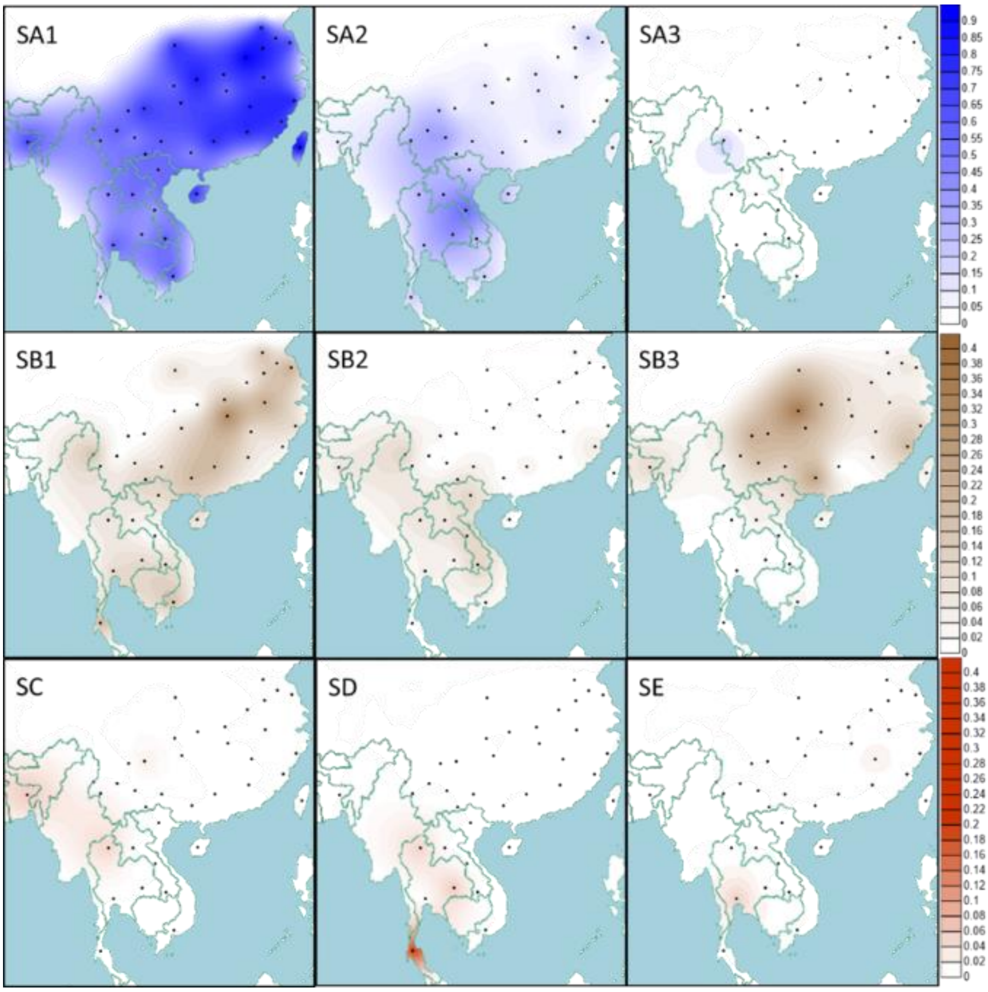


Figure 29. Spatial frequency distributions of swamp buffalo mtDNA haplogroups in different geographic areas based on previously published control-region data e and integrated with our complete mitogenome dataset.

3.2.3 Conclusion

More than 94% of current mitogenomes are derived from only two ancestors (SA1’2 and SB), which are both dated around the LGM (Table 3). This is one of the signatures that climatic events left in the female demographic history of swamp populations. During the glacial periods, the decrease of temperature leads to ice sheet expansion and subsequent fall of sea level, but also to other changes in ecological habitats, e.g. dry seasons largely modified the forest area (Figure 30), which played an important role in the distribution of plants and animals (De Deckker et al., 2003, Woodruff, 2010).

By overlapping our phylogenetic and demographic data with glacial periods, we identified five major phases during the last 200 thousand years (Figure 30) shaping the extant mitogenome variation of swamp buffalo.

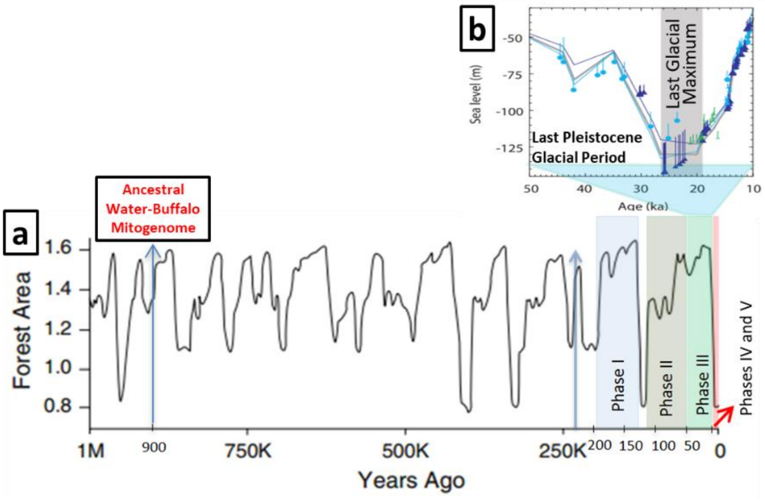


Figure 30. Proposed five major phases during the last million years compared with forest area (a) and sea level (b) fluctuations over time.

(I) (~200 to 130 kya) During the 2nd Pleistocene Glacial Period the first decline of population was observed and the division of two macro-haplogroups (SA'E and SB'D) was completed.

(II) (~110 to 50 kya) The first phase of the last glacial period was still comparatively moderate, the sea level was about -50m (Woodruff, 2010) and the population size remained almost unchanged, while only one divergence event (SA-SE).

(III) (~50 to 10 kya) During the second phase of the last glacial period the population began to decline. After the LGM the major Hgs SA and SB differentiated into 9 sub-Hgs (SA1, SA2, SA3, SB1a, SB1b, SB2a, SB2b, SB3 and SB4).

(IV) (~10 to 3 kya) The improvement of climatic conditions during the Holocene in southern China, known as the early Holocene optimum (between 10 and 6 kya) overlapped with the first phases of the rice cultivation, which is believed to have triggered the domestication of the swamp buffalo at 7-3 kya.

(V) (~3 kya to present) The rapid increase of population size was probably due to expansion of the domestic buffalo to the large current distribution range, harbouring the several present populations with distinct haplogroup distributions.

This link between swamp buffalo phylogenetic/demographic history and glacial events could be explained by a molecular differentiation of the wild populations that probably survived in few and perhaps isolated refuge areas in the extremely harsh environment and reduced suitable habitat during glacial periods. Actually, as a major global biodiversity hotspot (Cannon, 2015), many species of Southeast Asia were able to survive environmental challenges in the Pleistocene refuges (Woodruff, 2010). In fact, the surface of the ocean in this region is the warmest on Earth (Yan et al., 1992) and around this tropic Indo-Pacific Warm Pool the temperature decrease was limited even during the last glacial maximum and many species (including the swamp buffalo) were able to survive in the glacial refuge areas of Southeast Asia. The divergence of the current domestic sub-Hgs SA1, SA2, SB1a, SB1b, SB2a, SB2, SB3 and SB4 most likely occurred between the Holocene Optimum and the beginning of domestication, which is a minimum estimate of the swamp buffalo diversity captured by the first farmers. The high diversity of domestic SA and SB haplotypes in the China-Vietnam border region supports an initial major domestication event of swamp buffalo in Southeast Asia. Capture and domestication of a small part of the wild population may at least partially explain the decrease of the effective population during the period from 10 to 4 kya. The subsequent increase of the population size can be ascribed to the expansion of the domestic swamp buffalo, reaching a large part of the area where rice was cultivated. This was accompanied by a further differentiation of haplotypes within the domestic population. The finding of SC, SD and SE haplotypes almost exclusively in Thailand and Bangladesh suggests introgression of wild cows after the domestic buffaloes had reached the west bank of the Mekong river (Zhang et al., 2016b). Thus, it is plausible that in various periods and at different locations the expansion of domestic buffalo involved capture of wild females. A similar scenario has been proposed for domestic horse (Lindgren et al., 2004, Achilli et al., 2012). The clear difference between haplogroup distributions of nearby regions, which are larger than observed in other domestic species, indicates a low gene flow. This has also been observed for paternal lineages (Zhang et al., 2016b) and is the more remarkable because swamp buffaloes display hardly any variation in coat color or other visible attributes, again in contrast to almost all other domestic species.

More recently a genome wide study (Colli et al., 2018) reached similar conclusions about domestication and lack of gene flow by analyzing more than 20K SNPs. The high heterozygosity at the border between South China and North Indochina supports this area as the primary domestication center, from where two routes expanded (Figure 31), one to the south leading to the colonization of Sumatra and then moving eastwards to the rest of Indonesia, and a northern route spreading first to China and subsequently bending southwards into the Philippines.

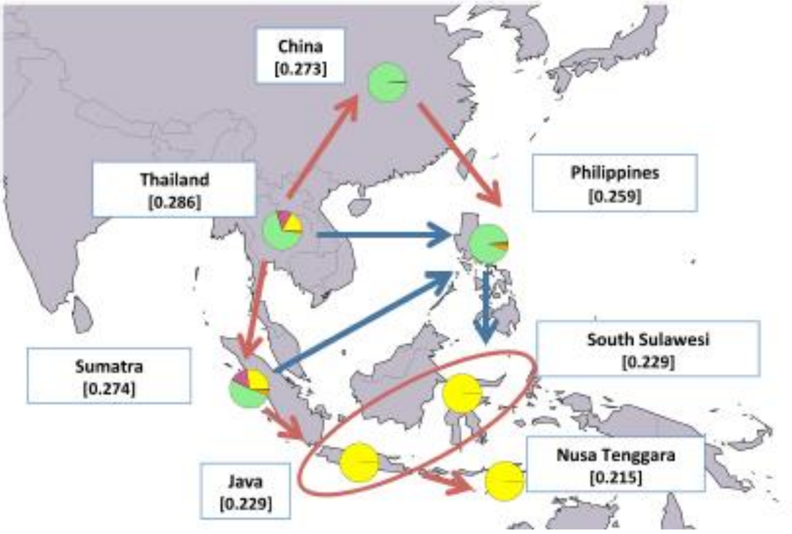


Figure 31. Map showing average expected heterozygosity values calculated after grouping swamp populations according to the geographical area. In particular, the red arrows indicate the most likely direction of significant post-domestication migrations. For each area the average membership coefficients corresponding to the results of *ADMIXTURE* software at K=6 are also shown in the pie plot (Colli et al., 2018).

3.3 Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability

Among the most important and diffused livestock species worldwide, the domestic goat (*Capra hircus*) represents an invaluable source of milk, meat, skin and fiber in developing countries. The current wide distribution is mainly the result of its medium size and high adaptability, while its genetic variability was influenced by early domestication practices. However, a fundamental question that has not yet been completely addressed nor resolved, both from the archaeological and genetic points of view, concerns the timing and mode in which humans began to benefit from the employment of these animals.

The answer to these questions was mainly searched using mitochondrial control region (Luikart et al., 2001, Sultana et al., 2003, Sardina et al., 2006, Naderi et al., 2007). This is because the highest level of molecular resolution is particularly difficult to obtain for goats as attested by major pitfalls affecting the mitochondrial genomes deposited in GenBank, such as chimeric sources, numts (i.e. nuclear sequences of mitochondrial origin) and sequencing errors. Two papers tried to overcome these drawbacks but failed to reach and explore the completeness of the mitochondrial molecule and/or focused only on a very limited number of samples (Parma et al., 2003, Hassanin et al., 2010). In our study we present the first extensive survey based on 86 complete mitogenomes, selected from an initial collection of 758 animals and including also representative of wild matrilineal variation from the Iranian bezoar (*Capra aegagrus*).During the very beginning of my PhD program I was involved in the phylogenetic analysis of this informative and unique dataset of novel complete mtDNAs from 78 domestic and 5 wild goats by using the same successful tools applied to other species presented in this thesis, in order to: i) accurately define (sub-)haplogroups involved in the initial domestication process(es), and ii) provide both coalescence estimates and expansion patterns of these domesticated lineages.

The initial network of D-loop haplotypes evidenced a high number of homoplasies and crosslinks even if the indexes of variability were similar to other animals. Therefore, the phylogenetic tree (Figure 32) was built only considering more stable coding-region sequence, confirming all previously known control-region branches (A-G) with the exception of the very rare haplogroup F, identified so far only in three domestic goats from Sicily (Sardina et al., 2006).

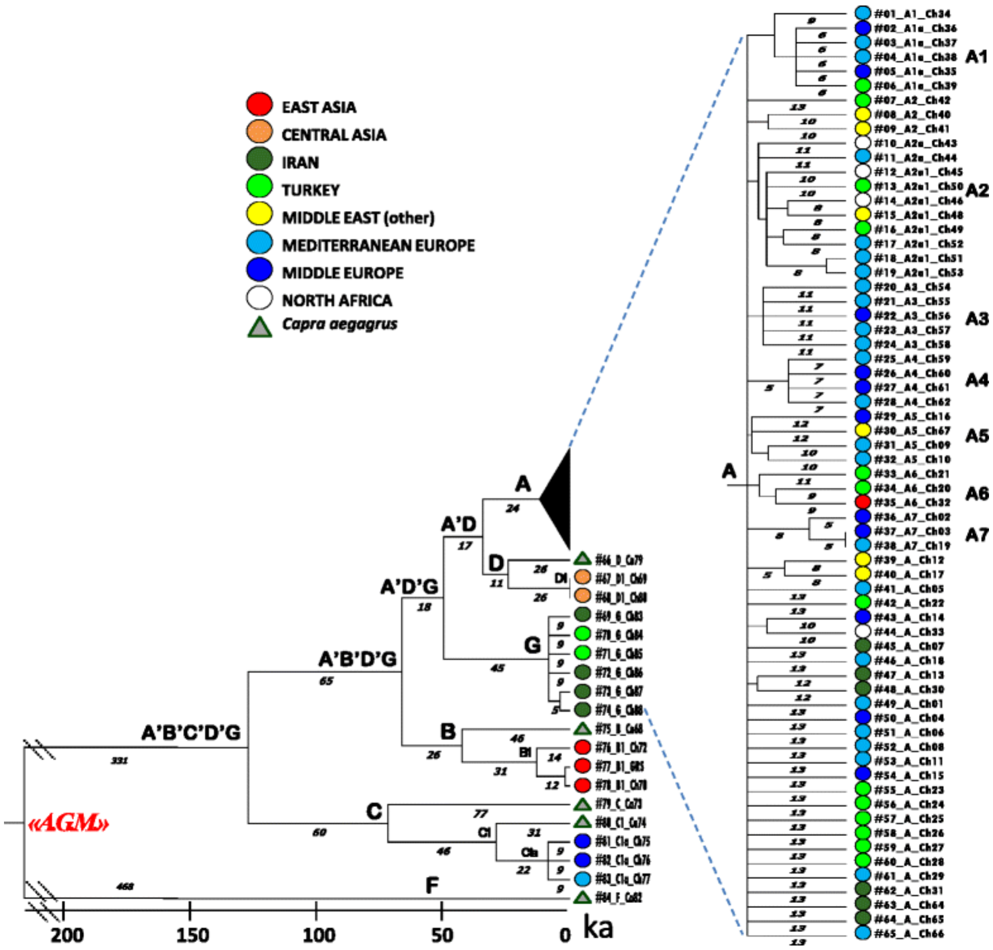


Figure 32. Schematic phylogeny of complete mtDNAs from modern domestic and wild goats. The tree encompasses 84 sequences and was rooted by using a published sheep (*O. aries*) sequence (not displayed). "A.G.M." indicates the reconstructed Ancestral Goat Mitogenome. The topology was inferred by a maximum parsimony approach, while maximum likelihood (ML) time divergences based on synonymous substitutions are shown below the branches. The right inset shows the complete A branch rooted in D.

The increase of resolution allows us to reveal many different sub-branches of the major Hg A, which includes more than 90% of domestic goats. Seven novel sub-branches, named A1 to A7, were identified, all marked by coding-region transitions. The molecular divergence of these and other branches was calculated as ML estimates. In order to calibrate the molecular clock, the goat mtDNA sequences were compared with a published complete mitogenome (KF302445) from a Comisana sheep (*Ovis aries*) (Ho et al., 2011), used as an outgroup. The molecular clock calculated on the synonymous variations (7.77 × 10^{-8} substitutions per year or 1 substitution every 3397 years) was then used to convert mutational distances into time. The ancestral goat mitogenome (AGM) was dated back to ~460 kya (Figure 32). Similar to buffalo, glacial periods played an important role also in goat evolution. In fact, the ancient wild populations might have passed through a severe bottleneck during the Late Saalian glacial maximum (~160–130 kya), with a single sequence (A’B’C’D’G) from that period as the common ancestor of most goat mitogenomes. Many lineages did not survive the following Würm glacial period (~71-12 kya) and the drastic climatic changes during the LGM, but some of them were preserved in the Near East refuge areas and survived the severe drop in temperature known as the Younger Dryas (~12.7–11.5 ka ago) and provided the necessary substrate of variability for later goat domestication and management Our phylogeny indicate that this process involved at least five lineages Hgs A, B1, C1a, D1 and G, mostly nested within wild goat branches and almost concomitantly diverging at the interface between the Epipaleolithic and early Neolithic periods, and took place in an area from the Zagros Mountains in Iran to Southeastern Anatolia, as testified by the abundance of goat remains in Neolithic sites from that region. In particular, the low-frequency goat clusters suggests secondary domestication foci, also including a more eastward Iranian center involving the furthest phylogenetically-related haplotype C1.

Additional details on this scenario were recently provided by ancient DNA analyses of ancient goat specimens ranging from hundreds to thousands of years in age. Daly and colleagues (Daly et al., 2018) provide evidence for a *multilocus* process of domestication in the Near East. Even if the mitochondrial type A spread and became dominant worldwide, at the whole-genome level, modern goat populations are a mix of goats from different sources, which means that multiple wild populations contributed to the origin of modern goats during the Neolithic (Figure 33). Furthermore, the patterns described support the idea of multiple dispersal routes out of the Fertile Crescent region by domesticated animals and their human counterparts.

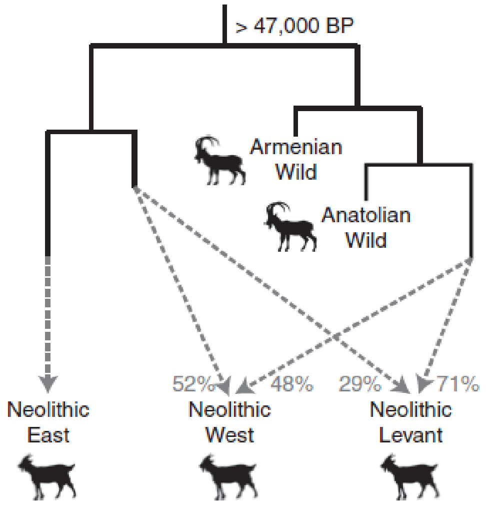


Figure 33. Admixture graph reconstructing the population history of pre-Neolithic and Neolithic goats. Relative inputs from divergent sources into early domestic herds are represented by gray dashed arrows. (Daly et al., 2018).

3.4 Survey of uniparental genetic markers in the Maltese cattle breed reveals a significant founder effect but does not indicate local domestication

The previously described phylogeographic pattern of goats suggests similarities with *Bos primigenius* domestication. Most of present-day cattle and goat mtDNAs belong, in fact, to a single macro-haplogroup (T and A, respectively), whose sequence coalescence time corresponds to a very similar time estimate of about 15-13 ka (Achilli et al., 2008, Colli et al., 2015). As for the major goat haplogroup A, mtDNA data supported a Neolithic origin for the major cattle lineage T3 from the Fertile Crescent, as well as for the other T lineages. However, differently from low-frequency goat clusters suggesting secondary domestication events in the same area, the recently identified cattle minor clades, P, Q and R (Bonfiglio et al., 2010), point to possible European domestication or introgression events.

In this context, Late Pleistocene oxen skeletal remains and Neolithic representations of primitive cattle discovered in the archipelago of Malta, located in the middle of the Mediterranean Sea, have been suggested as proof of a possible local domestication involving the Maltese cattle (Anati and Anati, 1988). Historical sources have remarked that this ancient breed, also known as "*Il-Gendus Malti*", is little known beyond Malta (Borg, 1915). It has been described as phenotypically characterized by large size and a coat of short reddish hair (with underlines slightly lighter) (MacGill, 1839) and employed exclusively for agricultural purposes, but not as a dairy animal. Only few bulls were kept from one generation to the next, just enough to propagate the breed, thereby the genetic pool of Maltese cattle was greatly homogenized and concentrated due to significant inbreeding (Adams, 1866). However, the introduction of other breeds during the British rule could have slowly modified the local gene pool resulting in a vanishing of the original genetic profile. In the early nineties, only three Maltese cows were left and a focused conservation program started through the introduction of the Chianina breed via artificial insemination. The back breeding between offspring was used to increase numbers within the herd. The present adult population consists of 12 males and 19 females divided into two herds. Likewise the buffalo and goat studies, I was also involved in the analysis of the entire mtDNA and of the Y-chromosome (Lancioni et al. 2016) of the Maltese cattle breed aiming to investigate its current genetic variability and to contribute to the preservation of its genetic uniqueness. The 81-bp insertion in intron 26 of USP9Y showed that all 12 bulls carried the haplogroup Y2, that is present in the Mediterranean region, also in the Chianina cattle. The mtDNA control-region analysis performed on the entire Maltese cattle population identified only two different mtDNAs, one encompasses about 90% of the current population and confirms a strong founder effect on the mitochondrial gene pool; the remaining 10% seems to testify for the importation of British cattle, documented in historical records since 1809. The complete mtDNA has allowed defining two novel clades (T3c and T3d, Figure 34), both dated to ~9.5 kya, encompassing only Maltese cattle and few other breeds of Northern European ancestry. These data finally disprove the local domestication hypothesis and confirm the introgression of British cattle.

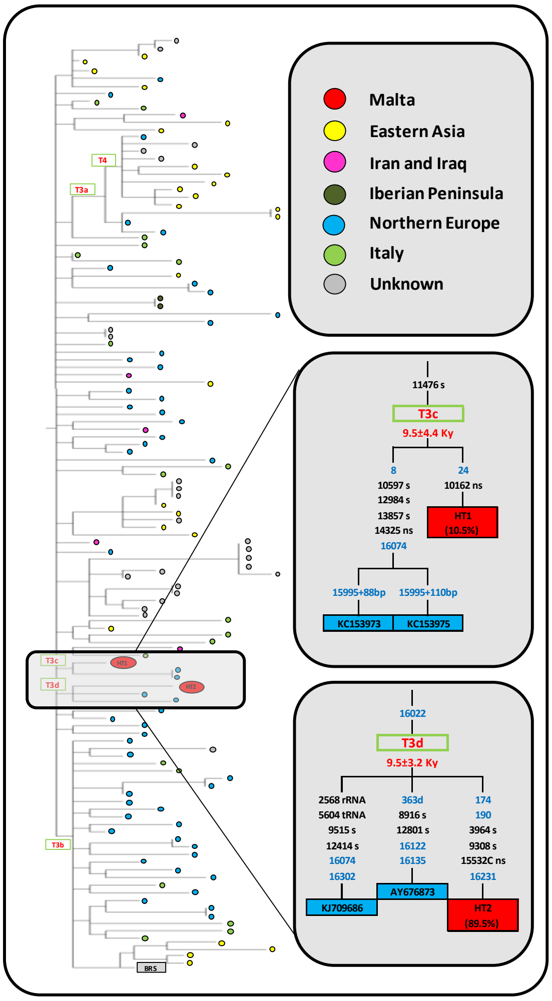


Figure 34. Schematic MP Tree of the mtDNA haplogroup T3. This tree encompasses 132 GenBank sequences, the two novel Maltese mitogenomes (red circles) and the Italian aurochs T3 genome (*). The insets show the novel clades T3c and T3d.

3.5 Mitochondrial DNA variants of Podolian cattle breeds testify for a dual maternal origin

Another approach to better understand the history of a livestock species and its link to human populations is by focusing on a specific breed. To this purpose, we genetically dissected the Podolian cattle breeds, which are among the oldest and still very widespread cattles in Europe. The history of Podolian breeds is still debated; the name of the group indicates a possible origin from Podolia, a region of what is the present-day Ukraine; today this type of cattle spread from Anatolia to the Balkans and the Italian peninsula. Podolian cattle might have spread from the eastern steppe southward into Anatolia and westward into the Balkans and Italy in historical times (3rd-5th century AD) along with East-European Barbarian people (Maretto et al., 2012); other authors suggest a more ancient migration (~3 kya BP) from the Near East to Central Italy through the Mediterranean Sea (Pellecchia et al., 2007), together with a possible contribution from local wild aurochs through a secondary local domestication/introgression events (Beja-Pereira et al., 2006, Bonfiglio et al., 2010). To provide further insights on the debated origin of Podolian breeds, we explored the maternal origin of those breeds through a comprehensive overview of the mtDNA control region variability of 18 Podolian breeds that were compared among them and to nine non-Podolian breeds (Di Lorenzo et al., 2018). A total of 1,957 samples were investigated for variability indexes. A similar nucleotide diversity was identified in Podolian and non-Podolian breeds, while a lower haplotype diversity was found in the Podolians. In particular, the highest variability was recognized in Chianina, Marchigiana and Turkish Grey and the lowest in the Bianca di Val Padana, Calvana, and Slavonian Syrmian Podolian. As expected the results obtained from the classification of the control-region sequences in haplogroups showed a prevalence (82.07%) of T3, the most common haplogroup in western Eurasia (Achilli et al., 2009, Bonfiglio et al., 2012, Olivieri et al., 2015). The high incidence of T1 in central e southern Podolian and non-Podolian breeds clearly shows a Mediterranean connection with Africa and Near East. In general, our survey points to significant differences in the haplogroup distribution between Podolian and non-Podolian breeds (mostly due to the extraordinary high frequency of T2 in Podolian cattle) and to a low genetic differentiation among the Italian Podolian breeds (three times less than in other Podolian breeds or in the non-Podolian group) that in turn might indicate a common (and perhaps peculiar) origin. In order to summarize these data and graphically display the different haplogroup distributions among Podolian breeds, we performed a PCA analysis (Figure 35). The first component clearly separates the breeds in two groups, one with five Italian beef cattles (Chianina, Marchigiana, Maremmana, Podolica Italiana and Romagnola) mostly reared in Central Italy and the Turkish Grey, while the other group very close to other European Podolic breeds.

This pattern might support the hypothesis of a dual ancestral contribution to the present gene pool of Podolian breeds, one deriving from Eastern European cattle through an inland migration, the other arising from the arrival of Middle Eastern cattle into Central Italy through a different route, perhaps by sea, ferried by Etruscan boats (Pellecchia et al. 2007). The historical migration of Podolian cattle from North Eastern Europe to Italy has not cancelled the mtDNA footprints of this ancient migration.

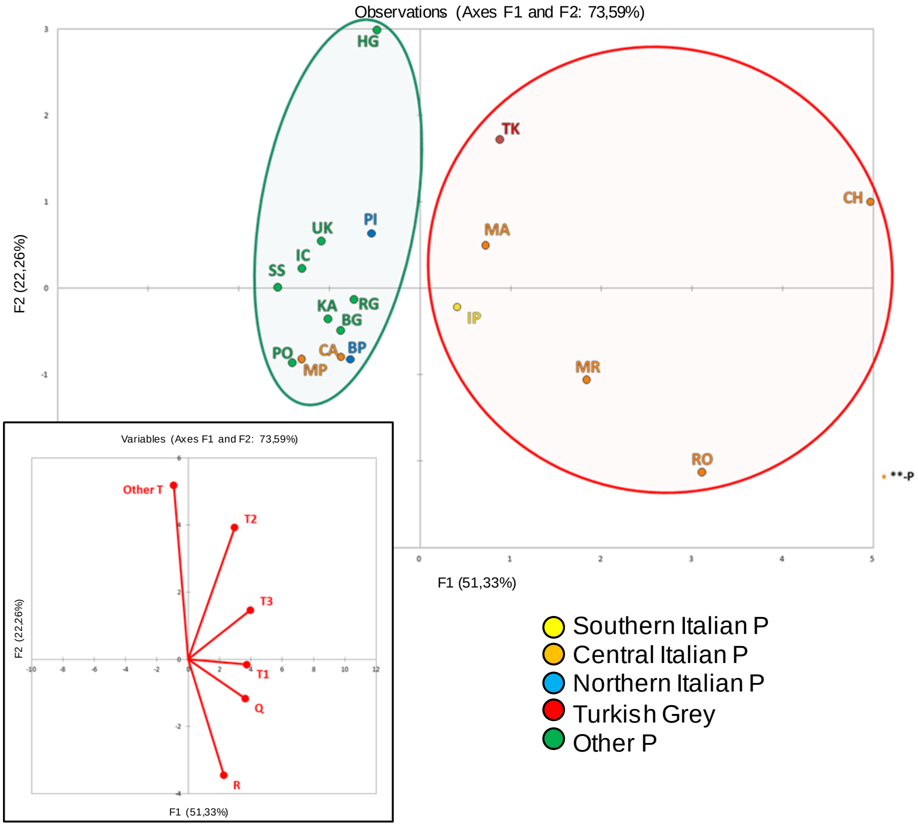


Figure 35. PCA of all Podolian breeds. Below is the plot of the contribution of each haplogroup to the first and second PC (projections of the axes of the original variables).

3.6 An overview of ten Italian horse breeds through mitochondrial DNA

The above-mentioned DNA analyses of domesticated species (i.e. buffalo, goat and cattle) revealed that modern livestock derive from a limited number of animals that were domesticated in just a few places around 10 kya. However, the horse mtDNA tells a different story (Achilli et al., 2012). The "equine Eve", the ancestral mother of all modern horses, probably lived about 140 kya, but modern horse mitochondrial genomes showed a high diversity with 17 different haplogroups identified in domestic breeds and spread in different geographic areas. This seems to imply that the domestication of wild horses was a widespread process that might have persisted for several thousands of years (Gaunitz et al., 2018). This high variability allows also, differently from the Podolian cattle work, to perform microgeographic study focused on the current breeds of a specific region. During my PhD I participated in the first comprehensive reassessment of the mitochondrial genetic relationships among the main native Italian hotblood/warmblood horses and ponies (Cardinali et al., 2016).

The climatic and cultural diversity of the Italian Peninsula triggered, over time, the development of a great variety of horse breeds. Nevertheless, multiple aspects of modern breeds' origin and history remain unclear. In this complex landscape, phenotypic traits and genealogical data are often coupled with molecular screening. In order to provide a comprehensive overview of the genetic variability observed in Italy, 407 control region mtDNA haplotypes from ten of the most important Italian riding horses and ponies were phylogenetically analyzed. An additional collection of 36 Arabian horses, were also included to evaluate the genetic consequences of its common use for the improvement of some local breeds.

The high haplotype diversity shown together with the presence of all domestic lineages (A-R) demonstrates a widespread mitochondrial variability in Italy. The mtDNA genetic landscape of Eurasia depicted by a principal component analysis (Figure 36) shows a clear geographic pattern and highlights a group of closely related intermediate breeds mostly from the Italian peninsula. This finding probably reflects the overall mtDNA legacy of the ancestral local mares that were probably used at the initial stage of breeding selections. They were preserved during the final establishment of pure breeds that was mainly reached through sex-biased breeding practices, which often involved the intensive use of few selected external stallions. Thus, the impact on the original mtDNA gene pool could have been very slight, as also testified by the only four haplotypes shared between the Arabian Horses and the

Italian breeds in spite of the well-recognized use of the Arabian stallions to revitalize some Italian breeds.

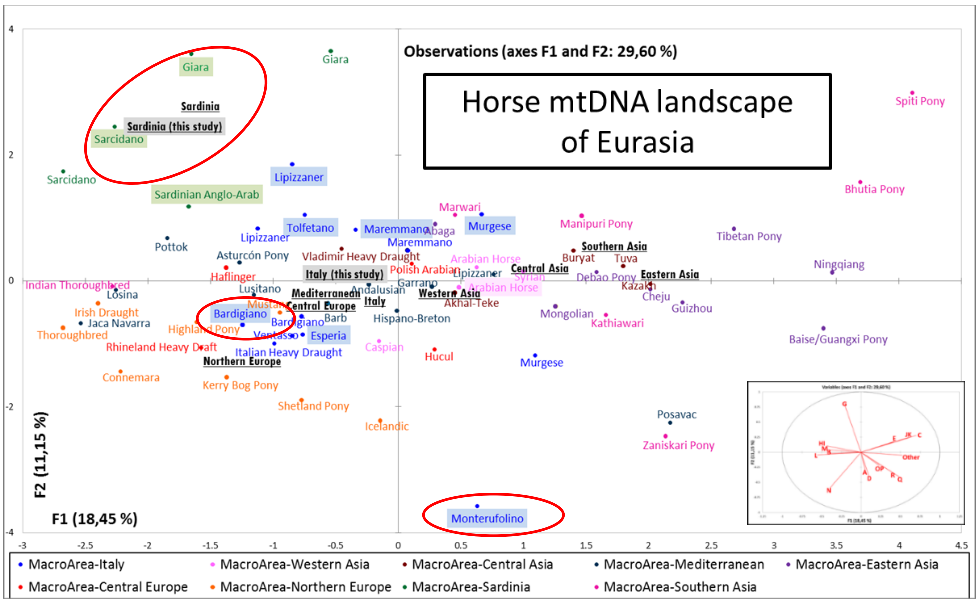


Figure 36. A two-dimensional region-based PCA plot obtained by including the available horse mtDNA data. The eleven breeds analyzed in (Cardinali et al. 2016) (and corresponding macroareas) are highlighted. The breeds discussed in the text are circled in red. Below is the plot of the contribution of each haplogroup to the first and second PC (projections of the axes of the original variables).

As for the recent times, our mtDNA data lend also genetic support to some historical theories about the origin of some Italian breeds, as testified by the peculiar position of some of them in the PCA. A strong founder effect is evident in the Sardinian breeds, due to geographic isolation, and in Monterufolino, whose total population counted less than ten individuals in the Nineties. The peculiar localization of the Bardigiano pony, which falls within a Northern European genetic context in the Eurasian PCA, could be explained by its history. The Bardigiano is considered indigenous of Italy, but its origin could be traced back to the horses ridden by northern invaders during their incursions into the Italian Peninsula in the V century. This original maternal legacy survived the recent dilution process due to the introduction of a diverse range of stallions from various breeds after World War II, especially *Franches Montagnes* breed.

Overall, these findings confirm that the mitogenome could be considered an appropriate resource in studies aiming to reconstruct the maternal ancestral origins of local breeds and to evaluate genetic continuity with the original stocks, thus assessing whether selection differentially affected mitochondrial genome variants during the development of economically important breeds.

3.7 The worldwide spread of the tiger mosquito as revealed by mitogenome haplogroup diversity

The mitochondrial DNA could be a molecular tool to reconstruct the history of other species, not only livestock. Among those organisms whose spread was human-mediated, tiger mosquitos (*Aedes albopictus*), originally indigenous to East Asia, has colonized every continent except Antarctica in the last 40 years (Benedict et al., 2007, Paupy et al., 2009, Bonizzoni et al., 2013, Kraemer et al., 2015). The spread of the tiger mosquitos has major implications for human health being a competent vector for many arboviruses which cause lethal or debilitating human diseases, including the dengue, chikungunya, West Nile viruses (Gasperi et al., 2012, Bonizzoni et al., 2013). The details of this rapid spread have hitherto remained unclarified. Therefore, I was involved in the first study aiming to perform a systematic phylogeographic survey based on mtDNA coding-region sequences from *Ae. albopictus* with the objective of defining its genetic origin and diffusion (Battaglia et al., 2016). Previous to our work, only two nuclear genomes (Bellini et al., 2007, Chen et al., 2015, Dritsou et al., 2015) and two complete mitogenomes (Asian tiger mosquito Reference Sequence, NC006817 and KR068634) (Zhang et al., 2016a) were published. Virtually all previous mitochondrial DNA studies of *Ae. albopictus* were restricted to short segments of the mitochondrial genome, suggesting a limited phylogeographic differentiation among populations. The mtDNA studies, including those described above, have shown that the variation seen in short mtDNA segments may be inadequate to both identify and phylogenetically link haplogroups (Figure 26) (Torroni et al., 2006, Achilli et al., 2008, Achilli et al., 2012). This is especially true for the insect mtDNA control region due to its peculiar features: high A+T content and reduced substitution rate, variable size and high length mutation rate, concerted evolution of tandem repeats and directional mutation pressure (Zhang and Hewitt, 1997).

In this study we have developed a new protocol for the amplification and sequencing of the tiger mosquitos mtDNA and eventually sequenced the entire coding-region (14896 bp) of 25 tiger mosquitoes from representative populations of Asia, America and Europe. The exclusion of the control region is due to the overall length and tandem repeat composition that make the PCR amplification and sequencing of the entire *Ae. albopictus* control region extremely difficult. Phylogenetic analyses of these 25 mitogenomes (together with two previously published) revealed a much higher level of sequence differentiation than previously reported.

The mitogenomes (24 distinct haplotypes) cluster into three major branches named haplogroups A1, A2 and A3 (Figure 37).

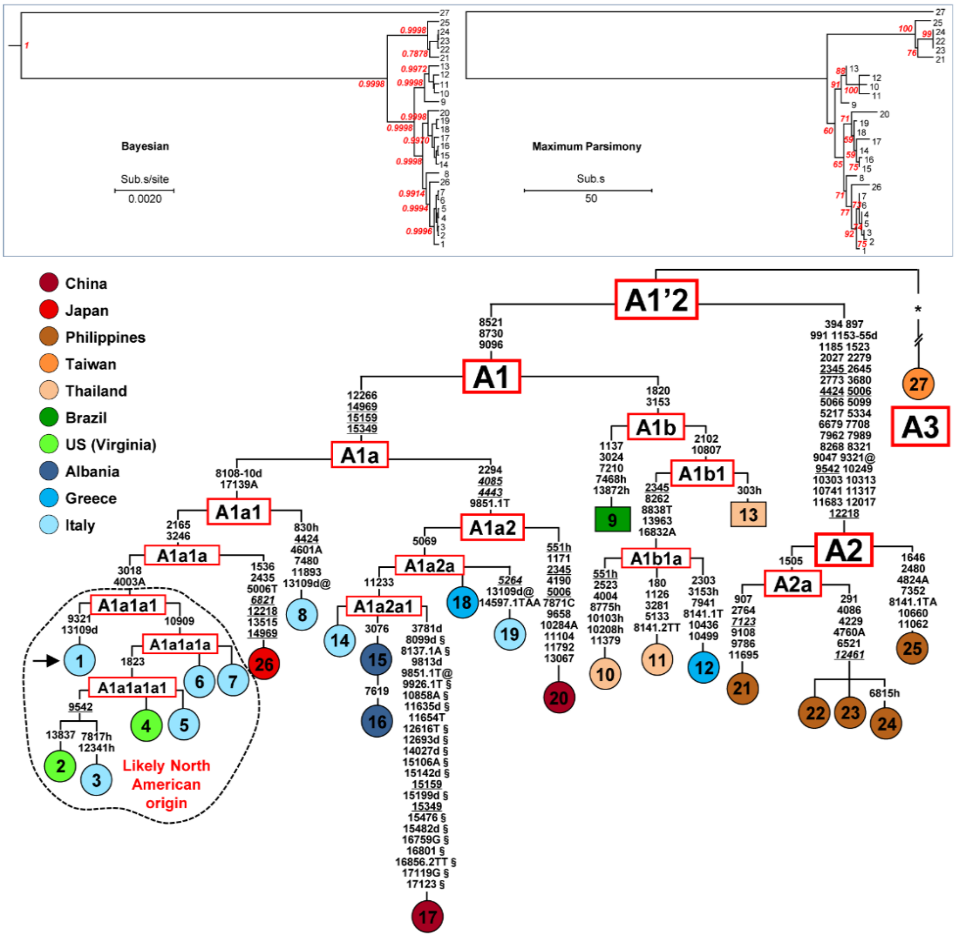


Figure 37. Phylogeny of *Ae. albopictus* mitogenomes. The Bayesian (left) and MP (right) trees are shown in the top inset. These trees encompass 25 novel and two previously published sequences. A magnified MP tree is also shown reporting all mutations that characterize the 27 mitogenomes except for those linking mitogenome #27 to the A1'2 node.

The Hg A1 is the most abundant with 21 mitogenomes from different geographic regions. As already shown for buffalo and goat, the new phylogeny allowed the identification of diagnostic motifs able to classify also samples genotyped for shorter mitogenome fragments. A population survey of the newly identified haplogroup markers allowed the affiliation into haplogroups of 1170 previously published tiger mosquito mtDNAs. No samples were classified as A3 while the A2 distribution is restricted to the insular Southeast Asia. Instead our analysis indicates that only three

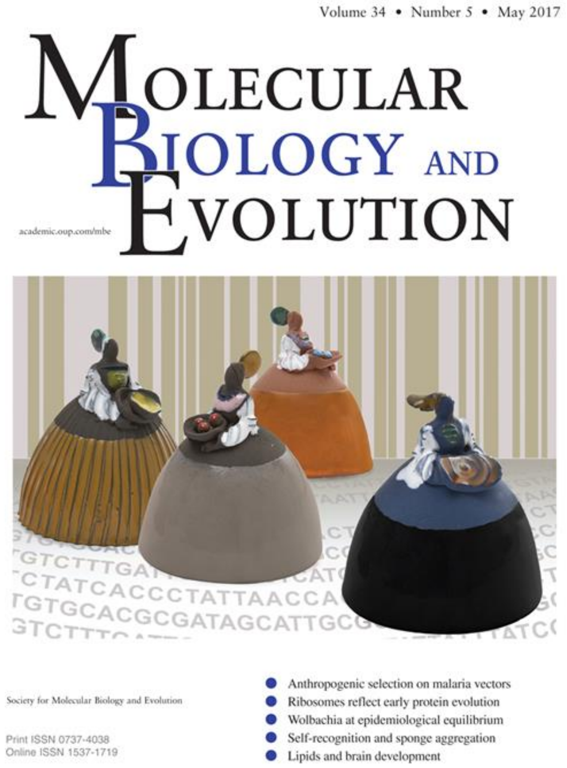
A1 sub-Hgs (A1a1, A1a2 and A1b) were involved in the recent worldwide spread. In particular, the derived lineage (A1a1a1) within A1a1, which is now common in Italy, most likely arose in North America from an ancestral Japanese source. The ancestral homeland of haplogroup A1a2 might have been a temperate area, possibly Japan or northern Asia, rather than the tropical range, in contrast, haplogroup A1b appears to mainly characterize the tropical belt. Nevertheless, the ancestral genetic sources now coexist (and interbreed) in many of the recently colonized areas. This occurs not only in the field but also in the laboratory as attested by our detection of both A1a1a1 and A1a2a1 mitogenomes in the Rimini maintained-strain, thus creating novel genomic combinations that might be one of the causes of the continuous and apparently growing capability of *Ae. albopictus* to expand its geographical range.

With a precise identification of the source populations in Asia it will become possible to evaluate the extent and nature of their nuclear genome diversity and the possible selective advantages (e.g. production of cold or desiccation-resistant eggs, zoophilic versus anthropophilic changes in feeding behavior) relative to other Asian *Ae. albopictus* populations that instead have not spread. Fine scale mitogenome surveys, encompassing worldwide populations, may prove to be an essential pre-requisite for controlling the diffusion of this mosquito and limiting its social, medical and economic effects.

3.8 Mitogenome diversity in Sardinians: a genetic window onto an island’s past

In the complex and debated scenario about the genetic origins of Europeans, described in the introduction, intriguing observations concern Sardinians, always considered as an outlier population in the general European genetic landscape. For evolutionary biologists, islands are geographically isolated pockets with unique populations that can be ripe for exploration. Sardinia sits at a crossroads in the Mediterranean Sea, the second largest island next to Sicily. Surrounded by sparkling turquoise waters, this Mediterranean jewel lies northwest of the toe of the Italian peninsula boot, about 350 kilometers to the west of Rome. This island remained unconnected with the mainland even when the sea level was at its lowest during the LGM (Shackleton et al., 1984) and that was probably the last of the large Mediterranean islands to be colonized by modern humans (Sondaar et al., 1998). Modern Sardinians, a unique reservoir of distinct genetic signatures (Cavalli-Sforza et al., 1994, Pala et al., 2009, Francalacci et al., 2013, Sidore et al., 2015), on one hand apparently harbour the highest levels of nuclear genome similarity with European Neolithic farmers (Lazaridis et al., 2014) and an extensive similarity with the Late Neolithic/Chalcolithic Tyrolean Iceman (Keller et al., 2012, Sikora et al., 2014) but, on the other hand, they differ substantially from Near Eastern Neolithic farmers including those from Anatolia (Lazaridis et al., 2016). Anyway, archaeological and genetic studies on the Y-Chromosome indicated the presence of humans in the island before the Neolithic colonization or an admixture in the mainland of the first incoming farmers (Hofmeijer et al., 1989, Dyson and Rowland Jr, 2007, Broodbank, 2013, Francalacci et al., 2013, Chiang et al., 2016).

To learn more about the genetic ancestry of Sardinians, I participated in the analysis of 3,491 modern mitogenomes (Figure 38) (Olivieri et al., 2017). The subjects were sampled from various areas of the island, encompassing each Sardinian province. In addition, 21 ancient mitogenomes dated to all cultural phases of Sardinia between the Neolithic and the Nuragic Final Bronze Age (5.3-2.9 kya), were analyzed.



Volume 34 • Number 5 • May 2017

MOLECULAR BIOLOGY AND EVOLUTION

academic.oup.com/mbe

Society for Molecular Biology and Evolution

Print ISSN 0737-4038
Online ISSN 1537-1719

- Anthropogenic selection on malaria vectors
- Ribosomes reflect early protein evolution
- Wolbachia at epidemiological equilibrium
- Self-recognition and sponge aggregation
- Lipids and brain development

Figure 38. Cover page of the Molecular Biology and Evolution journal dedicated to our work.

Our phylogenetic analysis revealed that almost 80% of modern Sardinian mitogenomes belong to branches that cannot be found anywhere else outside the island. Thus, they were defined as Sardinian-Specific Haplogroups (SSHs) that most likely arose in the island after its initial occupation. In this study I was involved mainly in the age estimates of all haplogroups and sub-Hgs, using ML and BEAST with two different molecular clocks, the Soares mutation rate (Soares et al., 2009), based on modern samples and the other recently estimated using radiocarbon dated ancient mitogenomes as tip calibration points (Posth et al., 2016). Almost all SSHs coalesce in the post-Nuragic (<2 kya), Nuragic (~4-2 kya) and Neolithic-Copper Age (~7.8-4 kya) periods. However, three SSHs showed (with all approaches) a coalescence age >7.8 ky (Figure 39), the postulated archeologically-based starting time of the Neolithic in the island (Berger and Guilaine, 2009).

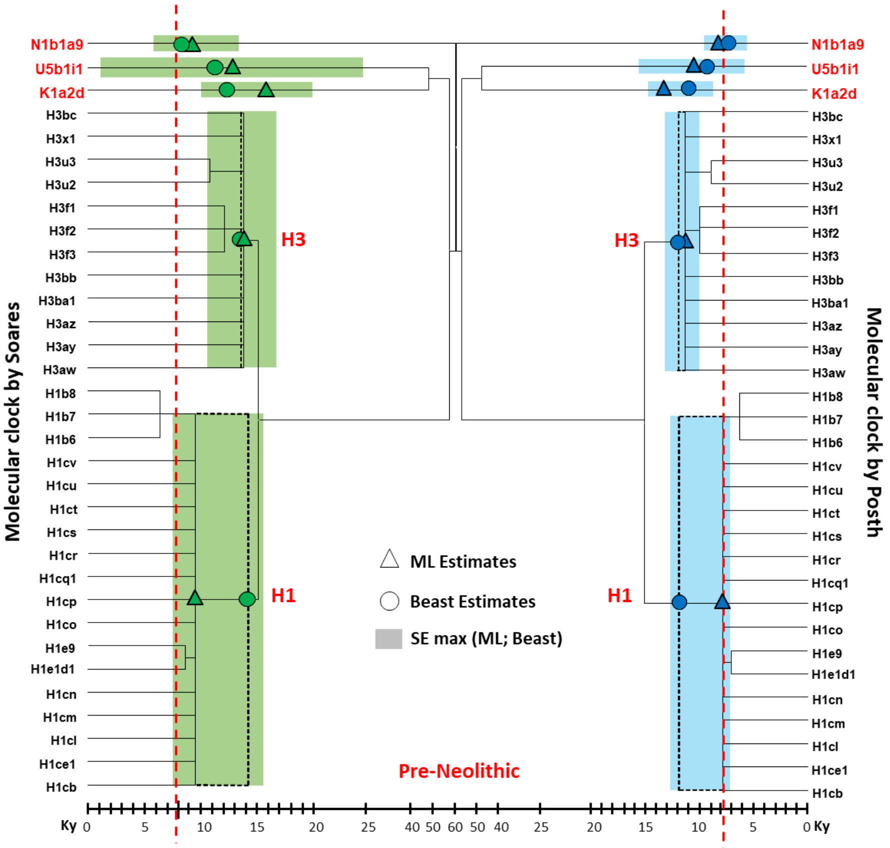


Figure 39. Specular schematic trees encompassing the three Sardinian-specific haplogroups (N1b1a9, U5b1i1, K1a2d) whose age estimates might predate the Neolithic (>7.8 Kya) and the Sardinian haplogroups H1 and H3. Age estimates were calculated by employing two mutation rates, by Soares (Soares et al., 2009) (tree on the left) and by Posth (Posth et al., 2016) (tree on the right). Triangles and continuous lines indicate ML estimates. Circles and dashed lines indicate BEAST estimates. Ages are according to the (non-linear) time scale on the bottom. Coloured shadings show the largest confidential intervals of age estimates.

This finding not only supports archaeological evidence of a Mesolithic human presence on the island but also reveals a dual ancestral origin of the first Sardinians. Indeed, one of the SSHs (U5b1i1) harbours deep ancestral roots in Palaeolithic Western Europe, overlapping the patrilineal source of the very frequent (38.9%) Y-chromosome haplogroup I2a1a-M26 both in terms of geography and timing (Francalacci et al., 2013). The other two (K1a2d and N1b1a9) are most likely of Late Palaeolithic Near Eastern ancestry, confirmed also by ancient samples and among those that are often assumed to have spread from Anatolia only with the Neolithic. The three pre-Neolithic SSHs comprise only 3.1% of modern Sardinians, but the genetic legacy of Mesolithic Sardinians could be much higher. Indeed, a large fraction of the SSHs belong to H1 and H3, the two most common haplogroups in modern Europeans that are dated to Pre-Neolithic time. These SSHs are only one mutation away from the H1 and H3 founding nodes and the confidence interval of their age estimates in the island spans toward the Mesolithic (Figure 39). Moreover, the distribution of the H3 (Figure 4 in Olivieri et al., 2017) suggests a Western Mediterranean source, probably the same of U5b1i1 and Y-chromosome haplogroup I2a1a-M26. Thus, even if H3 (and H1) had arrived in Sardinia only with the Neolithic, it most likely came either from either Spain or elsewhere in the western Mediterranean, but not from the Near East.

Although in the past the stress has often been on the spread of the Neolithic, genetic studies are beginning to emphasize the complexity and mosaic nature of human ancestry in the Mediterranean. Certainly, this work provided evidence that contemporary Sardinians harbour a unique genetic heritage, as a result of their distinct history and relative isolation from the demographic upheavals of continental Europe. It now seems plausible that human mobility, inter-communication and gene flow around the Mediterranean from Late Glacial times onwards might well have left signatures that survive to this day. Some of these signals are still retained in modern Sardinians. Future work on ancient DNA should be able to test directly to what extent this more complex model is supported by genetic evidence, and whether our predictions of Mesolithic ancestry in contemporary Sardinians can be sustained.

4. From modern mitogenomes to Archaeogenomics: the Panama project

4.1 The crucial role of Panama in the ancient and recent history of Native Americans

As illustrated in the introduction, the first Paleo-Indian settlers of the Americas entered a vast uninhabited area over a quite short time interval and then apparently remained relatively isolated from other human contacts for a considerable period of time, when developing a high degree of cultural diversity, linguistic complexity and biological variation. Therefore, the first founders left the greatest genetic mark in the double continent, but the original genetic pool of Native Americans was subsequently reshaped by additional inputs from abroad and local population dynamics, making any attempt to reconstruct the entire population dynamics in the Americas over the last 18-15 ky difficult to test, thus remaining an objective of scientific debates and a common target of multidisciplinary studies aiming to provide a final comprehensive picture.

In this context, the Isthmus of Panama is of great importance. This narrow neck of land connecting North to South America was a mandatory crossroad for the Paleo-Indians moving southward. Panama was also central in colonial times. European incursions onto the narrow isthmian pass that divides and connects the Atlantic and Pacific oceans made it a strategic node of the Spanish Empire during the XVI-XVII centuries and a crucial site for early modern globalization. Thus, the original gene pool underwent a multitude of post-Columbian admixture events, which involved different groups with different ancestries, mostly from Europe (European colonialism) and from Africa (Atlantic slave trade and subsequent African immigrations from Caribbean Islands).

The earliest well documented cultural remains in Panama refer to the Clovis tradition dated to ~11 kya (Cooke et al., 2013). In general, archaeological evidence and lake sediments concurred that some descendants of the initial colonizing population remained in the region adapting to the changing environmental conditions of the Late Glacial-Holocene transition, while others moved southward (Dillehay, 2009). Actually, the land bridge position, vis-`a-vis tropical atmospheric circulation, and its orography, influenced by the proximity of multiple plate junctions, have created a multitude of isthmian and insular landscapes that favoured the development of a high degree of endemism and diversity, also among human groups (Barrantes et al., 1990, Umaña, 1991, Herlihy, 1997, Cropp and Boinski, 2000, Anderson and Handley, 2002).

Agriculture prospered across the Isthmo-Colombian Area after ~8 kya. Although most cultivars, such as maize (*Zea mays*), squash (*Cucurbita spp*), and manioc (or cassava) (*Manihot esculenta*), and the few domesticated animals, like muscovy duck (*Cairina moschata*), were first domesticated north or south of the Panama land-bridge, there is no clear evidence for their initial introduction having been accompanied by major population displacements (Cooke 2005; Piperno 2011). Nevertheless, this area was certainly subjected to a series of migrations from north, south and east (with the Caribbean islands), along both overland and maritime routes, as testified by agriculture and pottery variations and other cultural changes (Figure 40). The maritime route was documented since 6.2-5.6 kya (Martín et al., 2016). Gold and cacao trades with Mesoamerica are documented by first Europeans, but a more ancient connection with Aztec and Maya populations is also possible (Lothrop, 1942, Lothrop, 1952, Cooke et al., 2003). Connections with continental Colombia, occurred at least until 1.4 kya, but perhaps also during the last 500-700 years (Mason and Johnson, 1940).

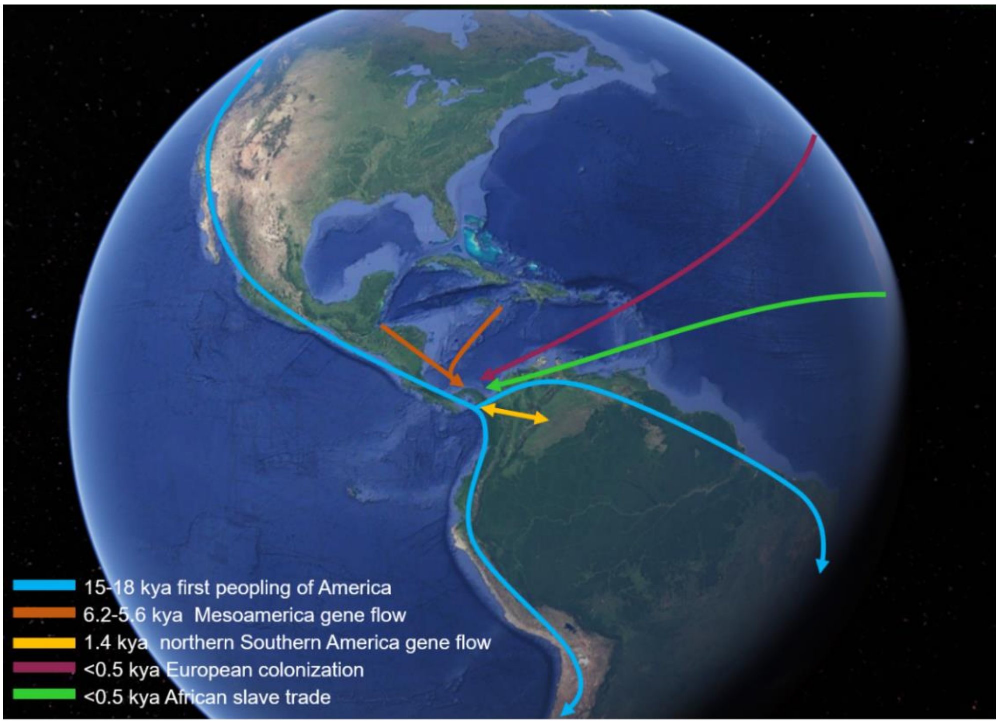

🟦 15-18 kya first peopling of America
🟧 6.2-5.6 kya Mesoamerica gene flow
🟨 1.4 kya northern Southern America gene flow
🟪 <0.5 kya European colonization
🟩 <0.5 kya African slave trade

Figure 40. Map of migrations that involved the Panama Isthmus suggested by historical and archaeological data.

In the last five centuries the arrival of Hispanic colonialists wiped out some autochthonous cultures, but Native resistance was more effective on the extensively forested western Panama, thus allowing for a higher degree of survival of pre-Hispanic cultures (Quilter and Hoopes, 2003, Rojas, 2012). Nowadays, Panama population comprises Native and mixed individuals. The ethnic groups make around 12% of the total population, according to the last census (2010). The main ethnic groups are Ngäbe (62.3%), Kuna (19.3%) and Emberá (including Wounáan) (7.5%) followed by smaller cultural groups: Bribri and Naso Djërdi (or Teribe) (Perego et al., 2012, Grugni et al., 2015).

Nuclear Chibchan (also Chibchan, Chibchano) languages are the most widely spoken linguistic group in Panama today. This language family includes the majority of surviving, or recently extinct languages, spoken in the Isthmo-Colombian area, which include Honduras, Nicaragua, Costa Rica, Panama and northern Colombia. The Chibchan name derives from an extinct language, called Chibcha, once spoken in Colombia. However, genetic and linguistic data now indicate that the root of this family is dated about 10 kya in a "core area" between southern Costa Rica and western/central Panama, where one finds the greatest variety of Chibchan languages. Concerning genetic data, the first study of Central American populations based on autosomal loci observed an isolation of the Chibchan in the pre-contact period (Barrantes et al., 1990).

Population dynamics involving the Macro-Chibchan (or Chibchan-Paezan) cluster, which includes Chibchan and Paezan (a hypothetical language family of Colombia and Ecuador), were also investigated and discussed in a more recent comprehensive study on the genome-wide variation of the current Native Americans (Reich et al., 2012) that analyzed 52 tribes and confirmed a possible correlation between genetics, geography and linguistic clusters. The presented tree (Figure 41a) shows a series of splits in an approximate north–south direction, from the Arctic to South America, thus confirming the mitochondrial scenario of costal route alongside the Pacific Ocean. In this study, the branch of Chibchan-Paezan speakers seems to confirm the hypothesis of a more recent migration across the Isthmus, even if specific samples from Panama were not included in the survey. In fact, as shown in the admixture graph, the Costa Rica and Panama Isthmus is the only region in which populations share ancestry component from both North and South America (Figure 41b).

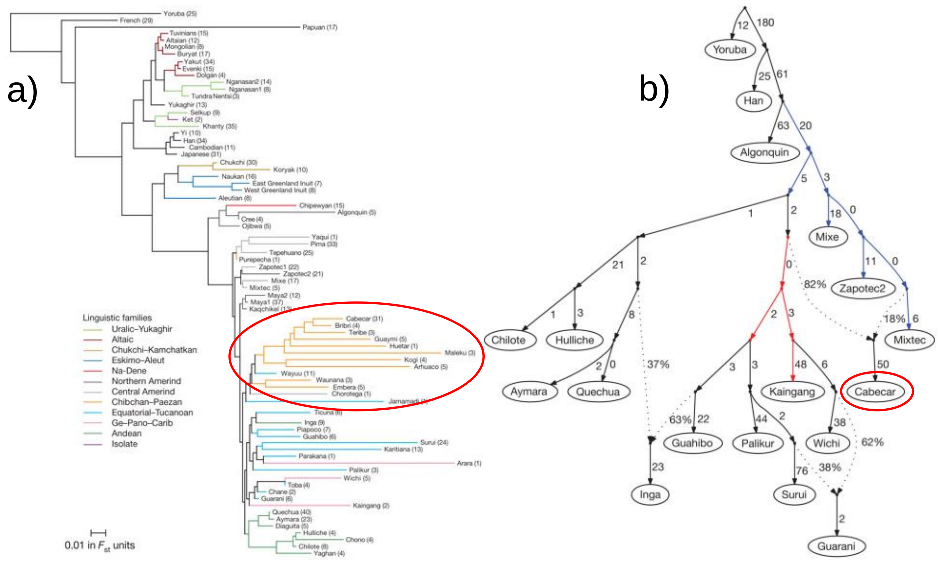


Figure 41. a) Neighbour-joining tree based on Fst distances relating Native American to selected non-American populations (sample sizes in parentheses); b) Admixture graph depicting the relationships of 16 selected Native American populations (Reich et al., 2012). The Chibchan-Paezan branch is circled in red.

Two detailed studies on the genetic composition of general populations in Panama have been performed analyzing uniparental systems, which evidenced different histories for males and females.

The Y-chromosome Native American component, represented by the haplogroup Q (Figure 42), was found with a frequency greater than 50% only in three populations facing the Caribbean Sea: the comarca of Kuna Yala and the province of Bocas del Toro where Chibchan languages are spoken by the majority; and the province of Colón where many Kuna and people of mixed indigenous-African and European descent currently live (Grugni et al., 2015). In the rest of Panama, the main component is represented by western Eurasian haplogroups, reflecting the strong male genetic impact of European colonization, and by African lineages, a consequence of the Atlantic slave trade and of more recent migrations connected to the railroad and canal building. Nevertheless, the high haplotype diversity of Q has shown that Panama was rapidly inhabited by the Paleo-Indians, but the dramatic event of past five centuries have largely influenced the survival of the pre-Hispanic Native lineages. An asymmetric mating between male newcomers and Native American women probably lead to a substantial decrease of Native American Y-chromosome (paternal) component, which was exacerbated by the high male mortality among Native groups due to unequal conflicts and forced transportation of surviving males to Panamanian mines and to other areas subjugated in the early Spanish Empire.

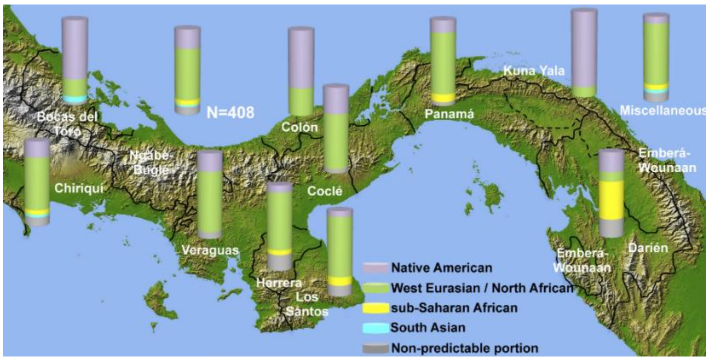


Figure 42. Y-chromosome haplogroups distribution in Panama (Grugni et al., 2015).

The hypothesis of a strong unidirectional sex bias in European-Native American admixture was already suggested by the mitogenome analysis (Perego et al., 2012) showing that the mitochondrial Native American component in Panama is higher than 80% (Figure 43). This percentage is composed by all the common "Pan-American" haplogroups (A2, B2, C1, D1), while none of the rare Native American lineages are present. The main haplogroup is A2 (>50%), the most common native lineage in Central America dated ~19-15 ky (Fagundes et al., 2008, Perego et al., 2009, Kumar et al., 2011). The most common A2 sub-clade in Panama, A2af, has its founder age at more than 10 kya. These data are in concordance with importance of the Pacific coastal route during the first peopling of the double continent and testify for an ancient colonization of the Isthmus by the first Paleo-Indians who left a strong maternal legacy in modern Panamanians (Perego et al., 2012).

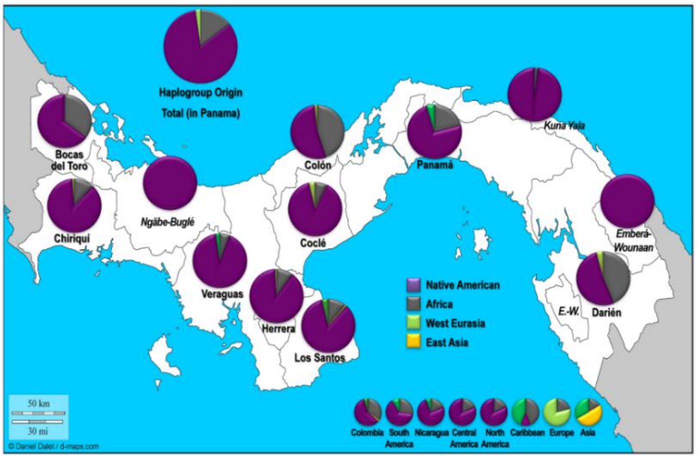


Figure 43. mtDNA haplogroups distribution in Panama (Perego et al., 2012).

These two genetic analyses of uniparental markers revealed that the gene pool of the currently “mixed” population of Panama is characterized by an opposite trend: a limited presence of paternal Native lineages, but an overwhelming legacy of maternal Native haplogroups, with some dating to the Paleo-Indian period. However, mitochondrial DNA investigations were limited to the control region variation and involved only the contemporary general population and the nuclear autosomal variation remained unexplored. In the current project, we are trying to increase the accuracy of our genetic reconstructions by extending the analysis to the entire mitogenome and eventually to the nuclear genome of contemporary ethnic groups and by analyzing ancient remains excavated in Panama City within the ERC Horizon 2020 project (CoG 648535) named “An ARTery of EMPIRE: Conquest, commerce, crisis, culture and the Panama Junction (1513-1671).”

4.2 Refining the maternal genetic history of Panama through modern and ancient mitogenomes

Figure 44. Cover image of the modern section of the Panama project.

4.2.1 Modern Panamanian Samples

In order to increase the population size of our Panamanian collection (Figure 44), especially for the tribal counterpart, a total 476 modern samples were collected in the 2016 sampling campaign (leaded by Dr. Maribel Tribaldos) that was organized thanks to a bilateral agreement between the University of Pavia (represented by Dr. Alessandro Achilli and Dr. Ornella Semino) and the Gorgas Memorial Institute for Health Studies of Panama (represented by Dr. George Motta).

All individuals filled-out a consent form, a genealogical pedigree and a declaration of their ethnic affiliation. The final dataset includes all the Native tribes, Ngäbe, Kuna, Emberá, Bribri, and Naso (Figure 45), and one additional African tribe, the "Moreno".

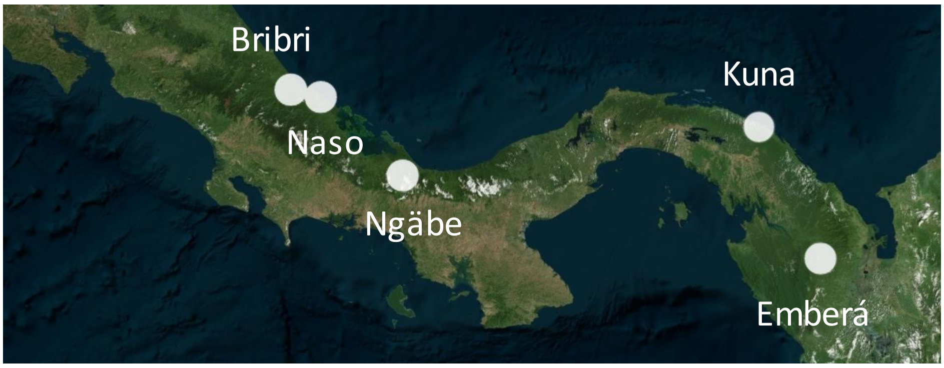


Figure 45. Graphical locations of the five Native American tribes in Panama.

Genealogical analyses allowed us to identify those individuals that were related to each other in order to exclude them from further analyses. Among the 397 unrelated subjects, we tried to sequence the entire mitogenome of those self-declaring to belong to a specific ethnic group (174 in total) (Table 4). After quality checks, we obtained 162 complete mtDNA sequences from unrelated tribal subjects (average coverage depth: 647X).

Table 4. Number of individuals for each Panamanian ethnic group used in this work.

Ethnic Group	Samples selected	Samples sequenced
"Moreno" (African origin)	6	6
BriBri	7	7
Emberá	45	43
Kuna	48	42
Ngäbe	58	56
Naso	10	8
Total	**174**	**162**

4.2.2 Results and discussion

4.2.2.1 Control-region analyses

This project started as a follow-up of the previous work on the mitochondrial control-region variability of the Panamanian general population (Perego et al., 2012). Thus, D-loop sequences were extracted from our 162 mitogenomes and compared to those from that study. Perego and colleagues analyzed 1,565 control regions, 1,422 from the general population (81%) and 143 from ethnic groups (9%), collected in 2011. The samples analyzed for this thesis work are 156 tribal (96%) and six mixed (4%) people.

All the 1,727 samples were divided into various subsets to evaluate differences of various genetic indices, e.g. Hd and Pi (Table 5). The high Hd of the general population is probably due to all non-Native lineages brought by gene flows from Europe and Africa since Columbian time. On the contrary, nucleotide diversity shows higher Pi values in all tribal datasets.

A $\chi 2$ test showed that the differences observed between the ethnic mtDNA gene pools of the two sample collections are not statistically significant (*P-value:* 0.689). In fact, haplogroup distributions are comparable and quite homogeneous (Figure 46). Therefore, the two tribal datasets were merged together to make a comparison with the general population (Table 5 and Figure 46)

Table 5. Estimates of genetic diversity indexes in different subsets of samples.

Group	N. Samples	S*	N. Haplotypes	Hd**	Str Hd	Pi***	k****
Tribal (Perego et al., 2012)	143	69	39	0.933	0.011	0.01	11.372
Tribal Native (Perego et al., 2012)	142	58	38	0.932	0.011	0.0098	11.148
All (Perego et al., 2012)	1,565	225	367	0.9706	0.0019	0.00769	8.4
Tribal Native This Study	154	69	71	0.964	0.007	0.00984	11.17
Tribal This Study	156	73	73	0.965	0.007	0.00987	11.207
All Tribal	299	90	94	0.956	0.006	0.01	11.33
All Tribal Native	296	78	91	0.955	0.006	0.00989	11.207
General Native	1,111	147	186	0.95992	0.0028	0.00728	7.981
General Population	1,428	223	358	0.9709	0.002	0.0077	8.414
TMA Panama	1,578	209	343	0.9716	0.0018	0.00803	8.772
All	1,727	230	406	0.9727	0.0017	0.00769	8.378

**S: Number of polymorphic sites; **Hd: Haplotype diversity; ***Pi: Nucleotide diversity; ****k: Average number of nucleotide differences.*

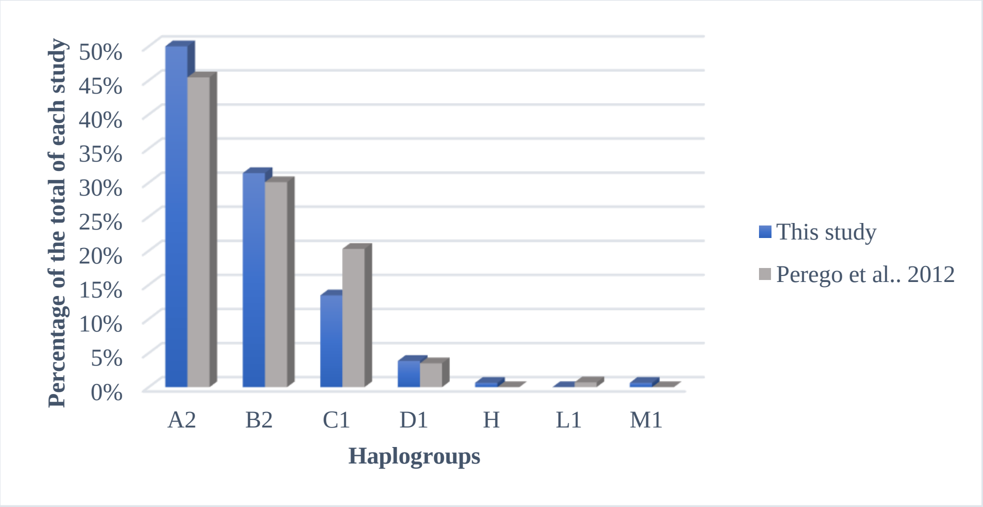


Figure 46. Graphical view of the haplogroup distribution of tribal samples in the two studies.

The resulting significant differences between mixed and ethnic samples (*P-value*: 2.35x10^{-13}) could be mainly due to the presence of African (all L haplogroups) and European (H, V, J, T, U, K) lineages in the general population (Figure 47).

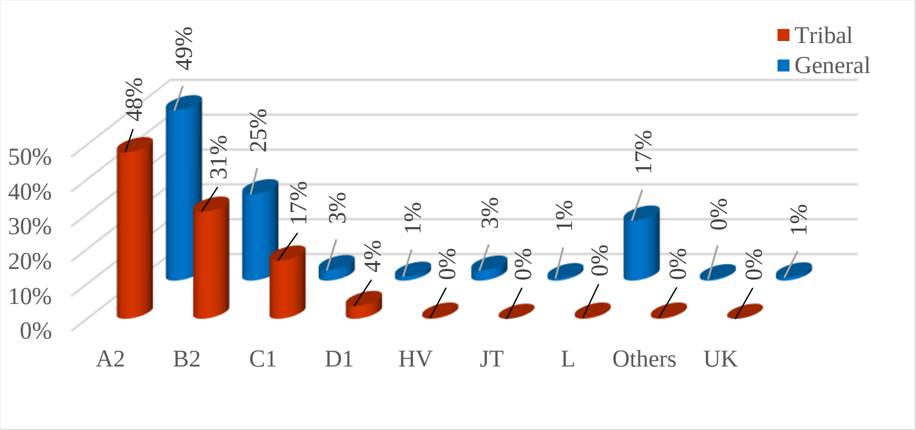


Figure 47. A comparison of haplogroup distributions between general population and tribal groups. The percentage are expressed as fractions of the total in each group. Non-Native American haplogroups have been grouped according to phylogenetic information. The category "Others" comprises less represented lineages: A5, F1, G1, I2'3, M1 and M5.

The main haplogroups are A2 and B2, which might derive from the Pacific migration wave of the first settlers moving southward and are currently represented by various sub-lineages in present-day Panamanian ethnic groups (Table 6).

Table 6. Haplogroup frequency in tribal population.

HG	N. Samples	Frequency in Tribes
A2	28	0.095
A2af	75	0.253
A2al	5	0.017
A2d1a	2	0.007
A2w1	33	0.111
B2	3	0.010
B2c2b	3	0.010
B2d	86	0.291
C1	3	0.010
C1c	4	0.014
C1d	43	0.145
D1	7	0.024
D1f	3	0.010
D1j1a2	1	0.003
Total	296	1.000

The Hg A2af and A2w1, which together encompass more than the 26% of tribal samples, have two different distributions (Figure 48). The A2af is highly represented in Bribri, Naso and Kuna populations, currently preeminent on the Atlantic coast, whereas the A2w1 is almost exclusively located among the Ngäbe. The only B2 subclade, B2d, is equally distributed in almost all tribal populations, with the exception of the Kuna where it reaches the 15% of frequency. The clades D1, that encompasses only the 0.037% of tribal samples, is found only in the Emberá and could mark a gene flows from south America where D1 is highly represented (Perego et al., 2010). These data show a differential distribution of the founder lineages that arose *in situ* and suggest a lack of gene flow among tribes, even when speaking similar languages.

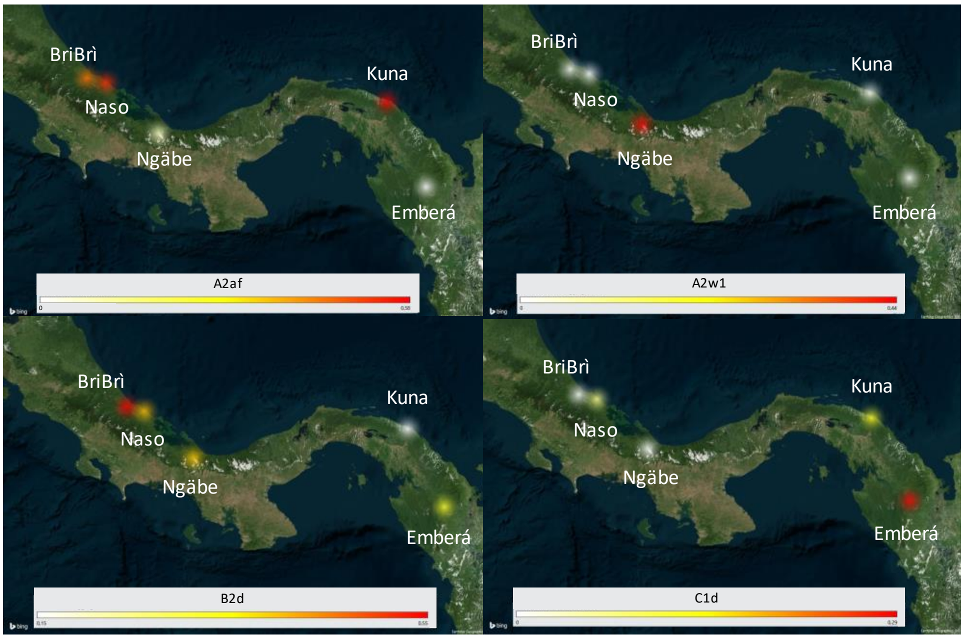


Figure 48. A comparison of the four haplogroups with highest frequencies in tribal population.

4.2.2.2 Complete mitogenome analyses

Extending the analyses to the entire mtDNA, we were able to increase the molecular and phylogenetic resolution. An initial estimate of the genetic diversity was performed on the complete mitogenomes in order to make a comparison with the control-region sequences (Table 7).

Table 7. A comparison between estimates of genetic diversities in control regions and complete mitogenomes.

Group	N. Samples	S*	N. Haplotypes	Hd**	Str Hd	Pi***	k****
Control region	162	84	78	0.968	0.006	0.01012	11.491
Complete	162	310	101	0.988	0.0031	0.00211	34.895

**S: Number of polymorphic sites; **Hd: Haplotype diversity; ***Pi: Nucleotide diversity; ****k: Average number of nucleotide differences.*

The results confirm the higher molecular resolution that can be reached with the complete sequences, as testified by a higher number of polymorphic sites, haplotypes and nucleotide differences. It is also verified that the control region is the highest variable portion of the mitogenome, as attested by the higher nucleotide diversity.

The entire mtDNA haplotypes (data not shown) allowed for a more detailed classification of our 162 samples into different sub-branches and for a phylogenetic comparison with all mitogenome sequences available in GenBank (2,663 sequences) and 1000 Genome Project (152 sequences) databases, belonging to the same haplogroups of our samples (A2, B2, C1, D1) (Figure 49).

Eventually, we were able to define 14 new sub-Hgs and to improve the phylogeny of already existing lineages with the information of the newly sequenced samples, also identifying 11 novel Panama-specific sub-Hgs. All these branches were dated using the ρ statistics and the mtDNA evolution rate of Soares (Soares et al., 2009) (Table 8).

Estimates based on other methods and statistics are generally used in addition to and to complement this method. Thus, the age estimates based on ρ parameter will be confirmed using dating approaches based on different statistics in the future work on the Panama project.

The Panama-specific sub-Hgs A2af1b1c and A2af1b3a, common in the Chibchan speaking groups and dated around 3 kya, support a local differentiation after the agriculture development. The first is present only in Ngäbe individuals, whereas the second one is found only in the Naso population. Regarding the haplogroup B2d two new sub-Hgs were identified.

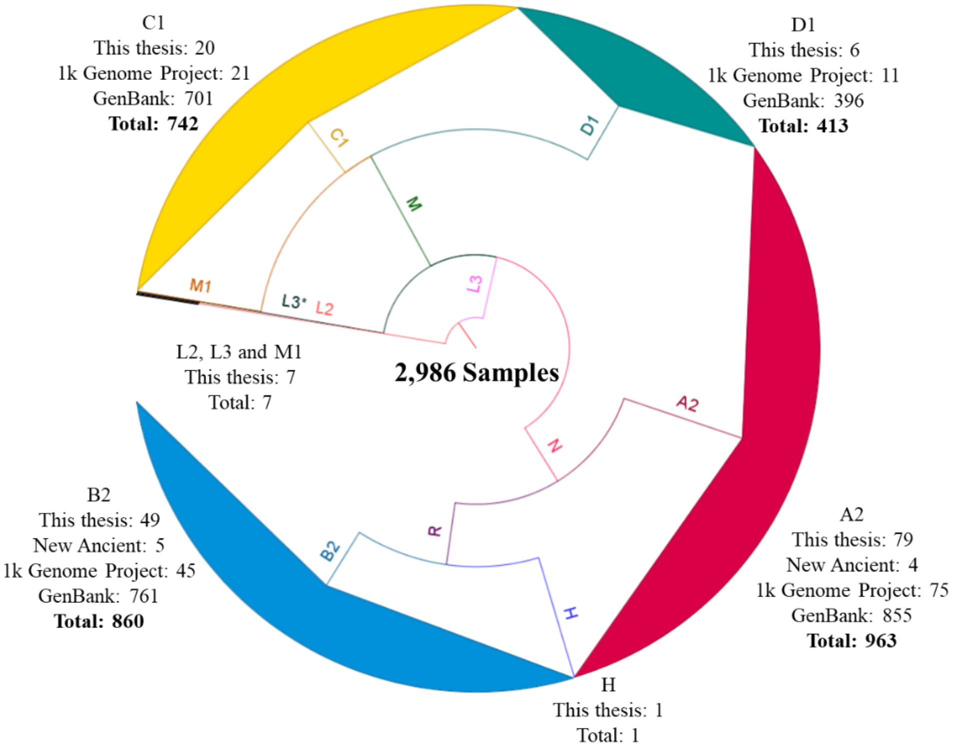


Figure 49. Overall phylogenetic tree, built with the software MEGA, including all mitogenomes analyzed in this work. For each haplogroup, the total number of samples, divided by their sources, is reported.

The Panama-specific B2d1 is dated around 4 kya, while B2af, which comprises also an individual from Ecuador, suggests a gene flow between Central and South America. Likewise, the D1f haplogroup comprises samples from Panama and Ecuador, but is dated to the Pleistocene, thus probably marking the first migration wave along the Pacific coast. On the contrary, D1l, another Panama-specific sub-Hg, probably diverged more recently and differentiated locally in the Panama Isthmus.

Finally, the new C1d sequences contributed further details to the history of this founder lineage. It was initially identified in two tribal populations of Mexico and Argentina, (Perego et al., 2010) and more recently of Amazonian Colombia (Arias et al., 2018). Here we found additional sub-branches that are Panama specific. Two of them closely related to the Emberá samples that neighbor Colombian tribes.

Table 8. Age estimates (in years) based on the entire mitogenome of the new sub-Hgs identified. The Panama-specific haplogroups are underline.

Hg	Age estimate	Error	Diagnostic Motif
A2af1b1c	3,449	4,215	12280, 13191
A2af1b3	5,668	5,005	7515
A2af1b3a	1,546	1,777	6548
A2w1a	6,763	5,406	8572, 12366, 14693
A2w1a1	4,211	1,757	13681
A2w1b	13,565	6,218	8896
A2w1b1	12,815	6,139	6221
A2ar	3,449	4,215	14971, 16274, !16362
B2af	9,926	6,416	4245, 9995, 16301
B2d1	3,030	4,153	3786
B2d2	1,287	1,469	2056, 7269
C1d1g	1,883	3,256	1393, 7302
C1d1h	2,585	2,709	6332, 8555, 15805, 16094, 16526
D1l	3,239	4,620	14968, 15927, 16086, 16311, 16322

4.2.2.3 A diachronic comparison between modern and ancient mitogenomes

The ancient DNA project described below allowed a diachronic comparison of our 162 modern mitogenomes with nine ancient mitogenomes, retrieved from Pre-Columbian remains excavated in Panama City, dated between 500 and 1,400 ybp, and presented here for the first time (Table 9).

Table 9. Summary of the nine ancient samples, whose data were used in this work. In particular, the mtDNA haplogroup, the mtDNA coverage, the bone from which the DNA was extracted, the age (expressed in calendar years before present calibrated with radiocarbon C14 dating) and the historical period are reported.

Samples ID	mtDNA Hg	mtDNA coverage	Bone	Cal C14 BP	Period
Pa10	B2d	53	Petrous Bone	660-520	Pre-Colonial
Pa16A	A2af1a1	30	Petrous Bone	1,070-930	Pre-Colonial
Pa24A	B2d	24	Petrous Bone	1,420-1,300	Pre-Colonial
Pa25	B2b	15	Petrous Bone	650-530	Pre-Colonial
Pa29	A2w	9	Petrous Bone	660-540	Pre-Colonial
Pa30A	A2af1b	11	Petrous Bone	660-530	Pre-Colonial
Pa09	A2af1	11	Petrous Bone	500-300	Pre-Colonial
Pa02	B2	15	Petrous Bone	680 - 550	Pre-Colonial
Pa28A	B2d	8	Petrous Bone	1,080-940	Pre-Colonial

Despite the presence of some missing positions, an accurate phylogenetic analysis based on the mutational motifs allowed to classify the ancient samples in Hgs (Table 9, Figure 50).
All ancient samples belong to Hgs which are quite common in the Isthmo-Colombian area Panama, i.e. A2af, A2w, B2, B2b and B2d (Figure 50). None of them belong to Panama-specific clades but are closely related to modern haplotypes.

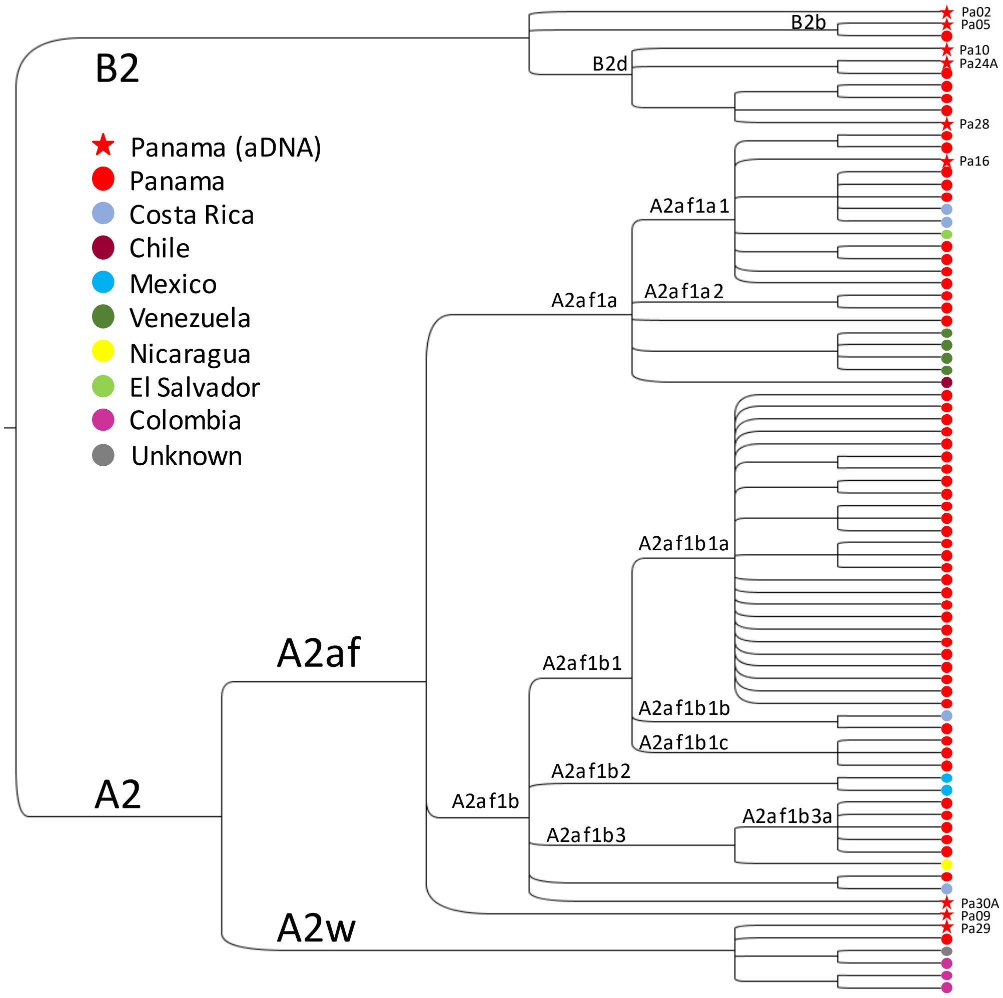


Figure 50. Phylogenetic relationships between the nine ancient mitogenomes (red star) and their most closely related mtDNAs in the analyzed dataset (Figure 49). Note that this tree is not dated and the branch length is uninformative.

4.2.3 Conclusion

In the research work on the complete mitogenome variation of Panamanians, 162 ethnic individuals were sequenced and analyzed, both at the level of the control region (D-loop) and of the complete mtDNA sequence, to refine the fascinating genetic history of the Isthmus of Panama.

Our control-region sequences were initially compared with other 1,565 obtained by a previous study (Perego et al., 2012). In that study, an overwhelming legacy of maternal Native American haplogroups was observed in the general "mixed" population, with some dating to the Paleo-Indian period. We tested these findings on the current ethnic groups of the Panama Isthmus that, at least from a maternal point of view, were less involved in the recent gene flows from Europe and Africa since the Columbian times. Thus, they could be in genetic continuity with descendants of first Paleo-Indians that settled and evolved in this region. Moreover, the complete mitogenome sequencing allowed us to increase the molecular resolution of the mtDNA at its maximum, thus implementing the sequence variation information available for further phylogenetic analyses. Our 162 novel complete sequences were compared to all available modern mitogenomes belonging to the Pan-American haplogroups A2, B2, C1, D1 (2,663 from GenBank and 152 from 1000 Genome Project) and to nine pre-Columbian Panamanian mtDNAs, obtained in our laboratory (see below).

In summary, the overwhelming native maternal legacy in today's Panama, particularly evident in the current ethnic groups, has been confirmed by the analysis of complete mitogenomes, which allowed also to enrich the Native American phylogeny with 14 novel branches. Among them, the 11 Panama-specific clades have detailed the genetic (mitochondrial) history of the Isthmus. The overall scenario indicates an initial settlement by the first Paleo-Indians while moving south along the Pacific coastal path in the late Pleistocene (testified by A2af and A2w), a more recent local differentiation of the Panama-specific clades with demographic expansions and maternal gene flows (largely attested by B2d) limited to the surrounding regions of Central America and to the most northern part of South America. This demographic evolution, dated to less than five thousand years ago, was probably triggered and facilitated by a shift to a predominant sedentary and agricultural subsistence during the late Holocene. Finally, the phylogenetic proximity of all nine ancient mitogenomes to the mtDNAs from modern Panamanians confirms a maternal genetic continuity between the current Panamanian ethnic groups and their ancestors, whose mitochondrial gene pool was only marginally impacted by post-Columbian migrations from Europe and sub-Saharan Africa.

4.3 Exploring the genetic history of Panamanians through ancient genomes

Figure 51. Cover image of the aDNA section of the Panama project

4.3.1 Ancient Panamanian samples

In collaboration with the University Pablo de Olavide of Seville (Spain) and the University of Norte of Barranquilla (Colombia) (under the ERC project "An ARTery of EMPIRE: Conquest, commerce, crisis, culture and the Panama Junction (1513-1671)"; PI: Dr. Bethany Aram) we collected 43 samples excavated in eight different archaeological sites of Panama City (Figures 51 and 52). The number of samples from each site is present in the Figure 52.

Different bones (femur, humerus and petrous bone) and teeth were available for the DNA analyses, for a total of 84 sources. A total of 18 samples have been radiocarbon dated at the University of Mannheim confirming an archaeological context ranging from 1,420 to 300 years ago. Additional archaeological information will be recorded in the ArtEmpire database (in preparation) that will soon be available on the project website (upo.es/investigacion/artempire).

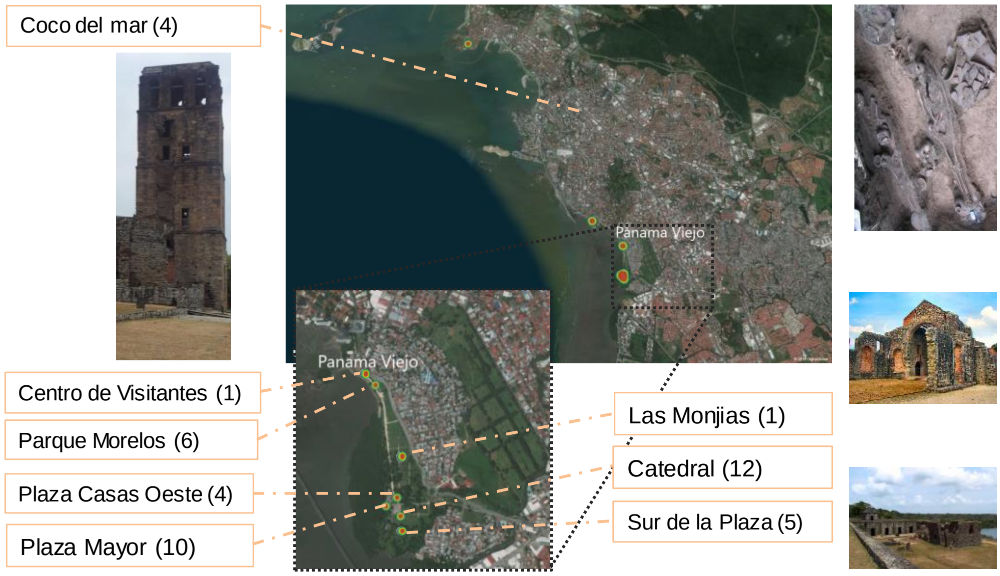


Figure 52. Location of the eight archaeological sites where samples analyzed in this study were excavated. In brackets the number of samples per site.

4.3.2 Methods for analyzing ancient genomes

4.3.2.1 Extraction of ancient DNA

The extraction of DNA from ancient remains requires a specific laboratory, with important precaution to reduce the contamination with modern DNA, hence the DNA from Panama was extracted in the “ancient DNA laboratory facility” of the Carl R. Woese Institute for Genomic Biology at the University of Illinois, headed by Prof. Ripan Malhi.

We have tried to extract DNA from 29 of the 43 available samples using different sources (Table 10).

In the first step, we removed the surface contamination from teeth and bones (from now on all sources will be called bone) by soaking them in 100% bleach for 3 minutes. The bleach was then removed with deionized water and isopropanol. Finally, each sample was dried under UV light for 15 minutes.

Table 10. Samples tried for DNA extraction.

Code	Library ID	Sources	Associated Site	Cal 14C BP	Period
19		clavicle, tooth	Sur de la Plaza		Colonial
26	**Pa01**	cranium, tooth	Sur de la Plaza		Colonial
27	**PaC18**	femur, tooth	Sur de la Plaza		Colonial
39		femur, tooth	Sur de la Plaza		Colonial
40		femur	Sur de la Plaza		Colonial
52	**Pa03**	femur, tooth	Catedral		Colonial
53	**PaC21**	mandible, tooth	Catedral		Colonial
54		femur, tooth	Catedral		Colonial
57	**PaC22**	femur, tooth	Catedral		Colonial
61	**PaC05**	femur, tooth	Catedral		Colonial
62		femur, tooth	Catedral		Colonial
93	**Pa07**	cranium, tooth	Catedral		Colonial
106	**Pa11**	femur, tooth	Centro de Visitantes		Prehispanic
109	**Pa12**	femur, tooth	Plaza Mayor	720-650	Prehispanic
110		cranium, tooth	Plaza Mayor	940-760	Prehispanic
114	**Pa09**	cranium, tooth	Plaza Mayor	500-300	Prehispanic
117	**Pa24A**	cranium, tooth	Plaza Mayor	1420-1300	Prehispanic
118	**Pa25**	cranium, tooth	Plaza Mayor	650-530	Prehispanic
128		femur	Plaza Casas Oeste	1420-1300	Prehispanic
134		femur, tooth	Plaza Casas Oeste	540-490	Prehispanic
137	**Pa26**	femur, tooth	Plaza Casas Oeste		Prehispanic
143		cranium, tooth	Parque Morelos	1050-920	Prehispanic
146	**Pa16A, Pa14**	cranium, tooth	Parque Morelos	1070-930	Prehispanic
156	**Pa27**	humerus, tooth	Parque Morelos		Prehispanic
167	**Pa28A**	cranium, tooth	Parque Morelos	1080-940	Prehispanic
172	**Pa02**	cranium, tooth	Coco del Mar	680 - 550	Prehispanic
173	**Pa10**	cranium, tooth	Coco del Mar	660-520	Prehispanic
174	**Pa29**	cranium, tooth	Coco del Mar	660-540	Prehispanic
175	**Pa30A**	cranium, tooth	Coco del Mar	660-530	Prehispanic

Each cleaned bone was drilled in a specific cabinet to collect 100mg of internal powder. The powder was digested in an Incubator with tube roller at 56°C for 24 hours.

Extraction solution recipe:

- 1ml of 0.5M EDTA
- 100ul of 33.3mg/ml proteinase K (up to 10mg can be used)
- 50ul of 10% N-lauryl sarcosine

After the digestion the remaining powder was pelleted by centrifugation for 2 min at maximum speed (4400 rpm, up to 17k g) and the supernatant was pured in an Amicon centrifugal filter (Sigma-Aldrich Corporation, US) and concentrated by centrifugation at maximum speed to approximately 100µl. The undigested powder was kept separated to be reused for another digestion cycle.

DNA was than extracted from the concentrated solution using MiniElute PCR purification Kit (Qiagen, Germany) for purification of up to 5µg of DNA (70 bp to 4 kb) in low elution volumes, following the standard protocol. We used 70µl of fresh prepared Tris-EDTA (TE) buffer for final elution of DNA from the silica columns.

The extracted DNA was preliminary screened using the Qubit Fluorometric Quantitation (Thermo Fisher Scientific, US) with the dsDNA High-Sensity (HS) Assay kit and performing three PCR amplifications of the mtDNA control region (Table 11).

Table 11. Control-region fragments amplified for preliminary screening.

PCR name	Position Primer Forward	Position Primer Reverse
D-loop1	15986F	16153R
D-loop2	16106F	16215R
D-loop3	16190F	16335R

4.3.2.2 Library preparation for shotgun sequencing

Samples showing at least 5ng of fragmented DNA and with at least one successful PCR amplification were selected for shotgun sequencing. The genomic library was built in the ancient DNA laboratory, using the NEBNext® DNA Library Prep Master Mix Set for Illumina® (New England Biolabs) using 50µl of DNA extract and following the protocol provided by the company.

Since the aDNA is already fragmented the first step was a long end-repair phase generating blunt-ends able to link specific adapters for Illumina sequencing (Figure 53).

End Prep recipe:

End Prep Enzyme Mix	3.0µl
End Repair Reaction Buffer (10X)	6.5µl
Fragmented DNA	55.5µl

Program:

- 30 minutes @ 20°C (room temperature, RT)
- 30 minutes @ 65°C (on heat block)

Since aDNA concentration is presumably low, the adapters were used to a dilution 1:30 to reduce the large amount of adapter dimers created during the library building. The following components were directly added to the End Prep reaction mixture.

Adaptor ligation protocol:

• NEBNext Illumina Adaptor	2.5µl (diluted 1:30)
• Ligation Enhancer	1µl
• Blunt/TA Ligase Master Mix	15µl
• Incubate at 20°C for 15 minutes (RT)	
• Add USER™ enzyme	3µl
• Incubate @ 37°C for up to 3 hours	

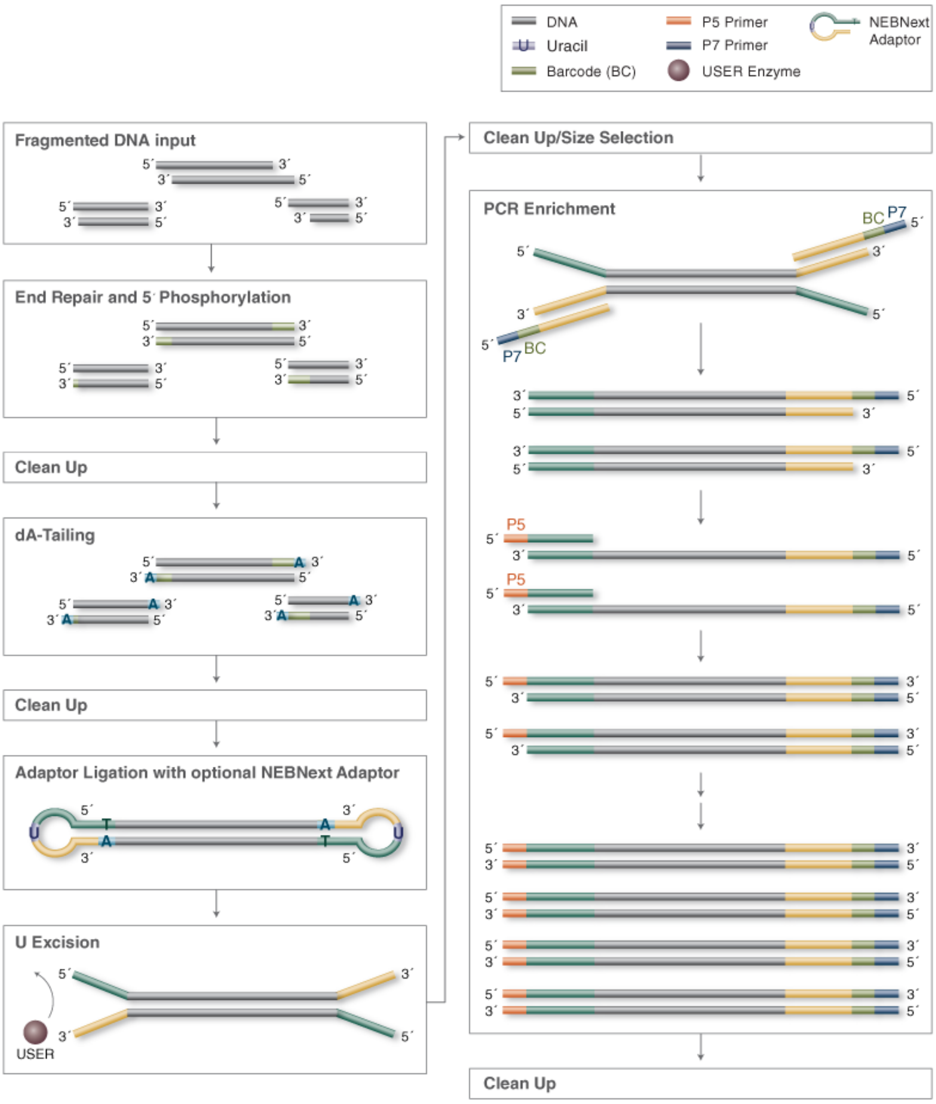


Figure 53. End repair and adapter ligation workflow.

After MinElute clean-up of adaptor-ligated DNA, a limited-cycle PCR amplification of genomic libraries was prepared in the ancient DNA laboratory and then run on thermocyclers in the "modern DNA laboratory".

1st PCR mix recipe:

Adaptor Ligated DNA Fragments	15µl
NEBNext High Fidelity Master Mix	25µl
Index Primer 5xx (white, ~65bps)	5µl
Index Primer 7xx (orange ~65bps)	5µl

1st PCR program:

Initial Denaturation	98°C 30 sec x1
Denaturation	98°C 10 sec x12
Annealing/Extension	65°C 75 sec x12
Final Extension	65°C 5 min x1
Hold	4°C ∞

AMPure XP beads (Beckman Coulter Life Sciences) were used for PCR purification clean-up and size selection to remove the adapter-dimers. The libraries were quantified using the Quibit and finally checked on the E-Gel (E-Gel Precast Agarose Electrophoresis System, Thermo Fisher Scientific, US) (Figure 54).

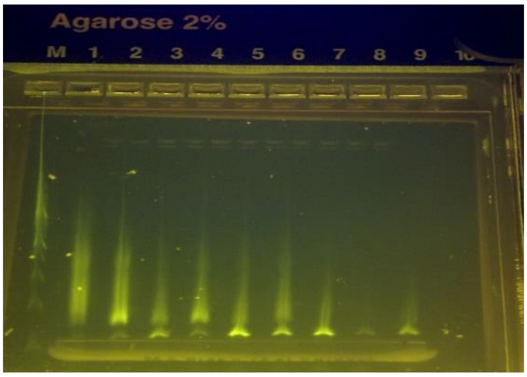


Figure 54. Example of a run on the E-Gel. The bands at the bottom of each samples are adapter dimers.

If the Genomic library band was at a low concentrate, a second PCR run was performed using the NEB Phusion® High-Fidelity PCR Master Mix and primers complementary to the Illumina adapters as in (Meyer and Kircher, 2010, Lindo et al., 2017)

2nd PCR recipe:

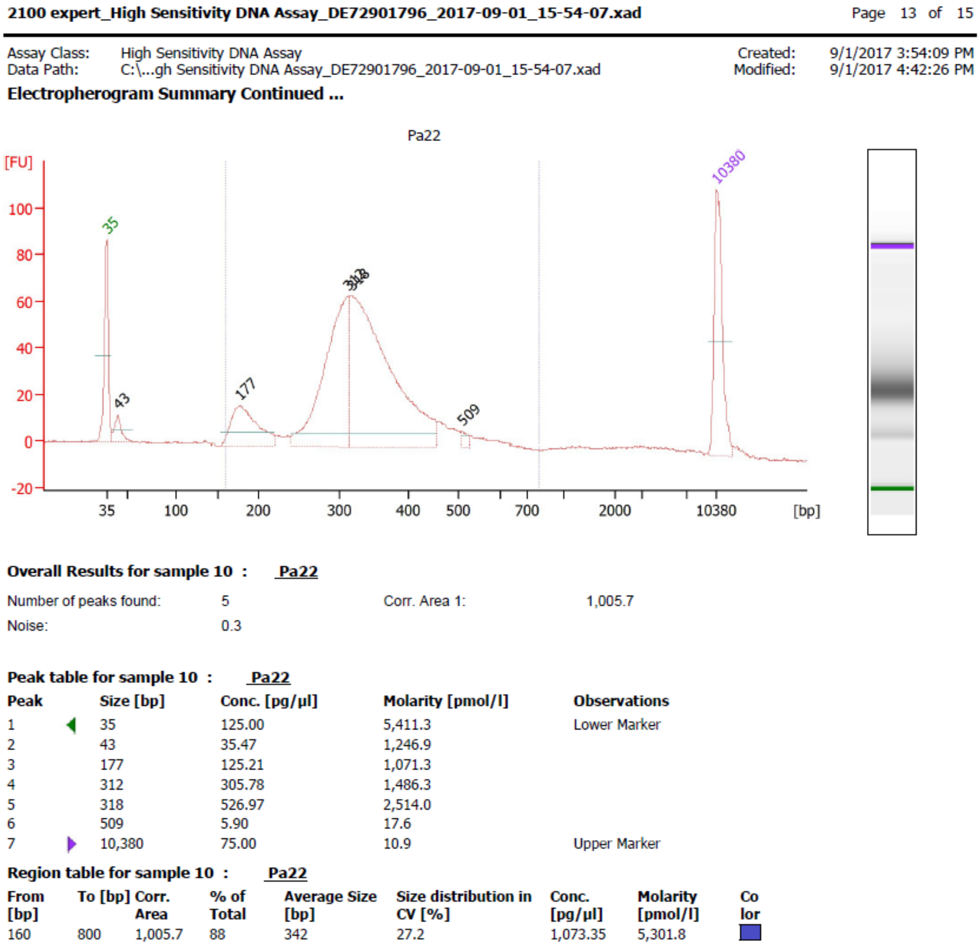
Master Mix	25μl
H2O	14.5μl
IS5 Primer	1.5μl
IS7 Primer	1.5μl
DMSO	1.5μl
BSA	1μl
DNA (1st PCR)	5 ul

2nd PCR program (limited-cycles):

Initial Denaturation	95°C 4 min x1
Denaturation	95°C 15 sec x12
Annealing	65°C 30 sec x12
Extension	68°C 30 sec x12
Hold	10°C ∞

After a second run of clean-up, the library quality was finally assessed with the Agilent 2100 Bioanalayzer (Agilent Technologies, US), using the High Sensitivity DNA Kit (Figure 55).

Eventually 21 samples were selected for whole genome sequencing on the Illumina HiSeq4000 (Single Read 100nt for a total of 150-200M single reads) at the Roy J. Carver Biotechnology Center of the University of Illinois. A pool of eight samples with a final concentration of 10nM were run toghter in one lane of the HiSeq flow cell (average output: 40-60 Gb per lane).



2100 expert_High Sensitivity DNA Assay_DE72901796_2017-09-01_15-54-07.xad Page 13 of 15

Assay Class: High Sensitivity DNA Assay
Data Path: C:\...gh Sensitivity DNA Assay_DE72901796_2017-09-01_15-54-07.xad
Created: 9/1/2017 3:54:09 PM
Modified: 9/1/2017 4:42:26 PM

Electropherogram Summary Continued ...

Overall Results for sample 10 : Pa22

Number of peaks found:	5	Corr. Area 1:	1,005.7
Noise:	0.3		

Peak table for sample 10 : Pa22

Peak		Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1	◀	35	125.00	5,411.3	Lower Marker
2		43	35.47	1,246.9	
3		177	125.21	1,071.3	
4		312	305.78	1,486.3	
5		318	526.97	2,514.0	
6		509	5.90	17.6	
7	▶	10,380	75.00	10.9	Upper Marker

Region table for sample 10 : Pa22

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Co lor
160	800	1,005.7	88	342	27.2	1,073.35	5,301.8	🟦

Figure 55. Example of an electropherogram summary obtained from the Agilent 2100 Bioanalayzer.

4.3.2.3 Ancient sequence analyses

The analyses of the row data were performed in collaboration with the Estonian Biocentre (EBC) at the University of Tartu and were run on the high-performance computing (HPC or cluster) platform owned by EBC.

4.3.2.3.1 Cleaning and mapping

Sequences were provided from the Sequencing Facility in the form of compressed FASTQ.GZ files. The universal Illumina adapters were removed using CutAdapt (Martin, 2011), as in the following command (Command 7). Various flags are added in order to reduce the sequencing artefacts and to remove reads with low length or quality. The command -*m* 29 deletes all the reads that are less than 29 bps in lenght, while *-e* indicates the tolerance of the error rate (0.2) to be kept; errors refer to mismatches and indels found by the software while searching the adapters. The minimum Phred quality score accepted (*-q*) is 30. Finally, a maximum of two N calls are accepted.

Command 7:

```
cutadapt -a
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNNATCTCGT
ATGCCGTCTTCTGCTTG -m 29 -e 0.2 -q 30 --trim-n --
max-n 2
```

The trimmed FASTQ was then checked with *fastqc*. High quantity K-mers were found, mostly poli-A. Probably these poli-A were generated during the blunt-end phase. In order to remove these artefacts, CutAdapt was used again trying to remove poly-A and poly-T. An example is the following command (Command 8).

Command 8:

```
cutadapt -b "A{50}" -m 29 -q 30 --trim-n --max-n
2
```

Trimmed reads were mapped against hg19 build 37.1 as well as *versus* the revised Cambridge Reference Sequence (rCRS) in a separate run using *bwa v0.6.1* (Li and Durbin, 2010). Differently from modern mtDNA data (Command 3), aDNA reads were aligned using the algorithm *aln* (Command 9), which is specific for short reads (<70bps, as expected in aDNA). The flag -l allows a large number of mismatches and is typically used for ancient DNA data that contains numerous errors.

Command 9:

```
bwa aln -l 1000
```

aln gives reads mapped in *sai* format that was converted in *sam* with *bwa sampe* and then in *bam* with SAMTools v1.19 (Li et al., 2009). With the tool MarkDuplicates, which is part of the package Picard Tools (http://broadinstitute.github.io/picard), the numerous duplicates, generated from the library amplification, were removed.

The molecular sex of individuals was than identified using a python script published by Pontus Skoglund and colleagues in 2013 (Skoglund et al., 2013), specific for low-coverage DNA sequencing. This script calculates a ratio of sequences aligned to the X and Y-chromosomes taking also into account the signature of cytosine deamination to reduce the bias due to the present-day contamination. This contamination has also been calculated considering the principle that known polymorphic sites on haploid genomes and their adjacent sites should have the same error rate unless some human contamination is present. The script used for this analysis were provided by Dr. Christiana Lyn Scheib and published in her paper in 2018 (Scheib et al., 2018).

The obtained ancient mtDNA contigs were then re-analyzed with the pipeline used for modern samples and analyzed together with the modern mitochondrial genomes, as described in the previous section.

4.3.2.3.2 Validation of aDNA reads

As already mentioned in the introduction, nowadays, the validation of the ancient DNA is based on the survey of the common ancient DNA damage pattern, which is based on two major changes: i) the "C->T" and G->A substitutions at the end of the reads due to the deamination of the cytosine, recognized as uracil and paired with adenine instead of guanine, therefore in the following replication the cytosine is substituted with a thymine (Figure 11); ii) the length of the reads that in aDNA are typically shorter than in modern samples due to the DNA fragmentation that starts soon after the death of the individual (Figure 56) (Dabney et al., 2013).

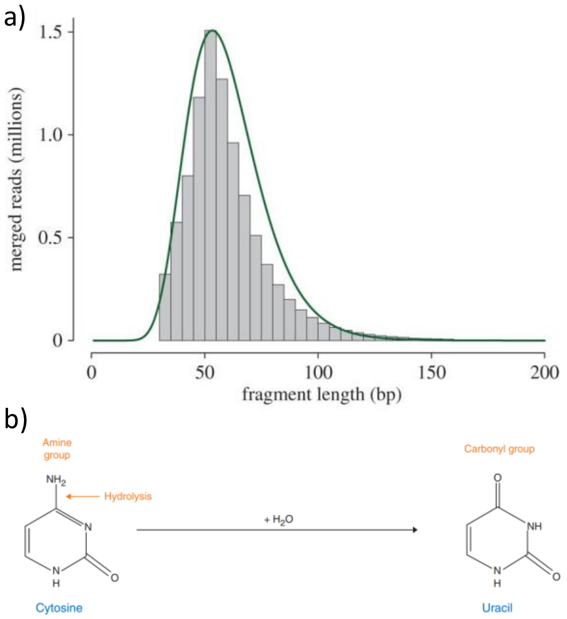


Figure 56. Common damage pattern in ancient DNA, a) Distribution of fragment lengths of merged reads from an ancient sample (NY1365354). The green line shows the fit between the empirical and the lognormal distribution (Weiß et al., 2016). b) the deamination of the cytosine that produces an uracil (Dabney et al., 2013).

Damage patterns of trimmed *bam* files were analyzed with MapDamage2.0 (Jónsson et al., 2013) that returns different *.txt* files with statistical information to check the damage.

In order to observe the molecular decay after death, we started from the length of the reads and used an excel file for automatic calculation, following the protocol of Allentoft and colleagues (Allentoft et al., 2012). Considering the debate about the possibility of dating the fragments with this algorithm (Figure 57), we used these data only to compare the decay of our samples with other published American ancient DNAs.

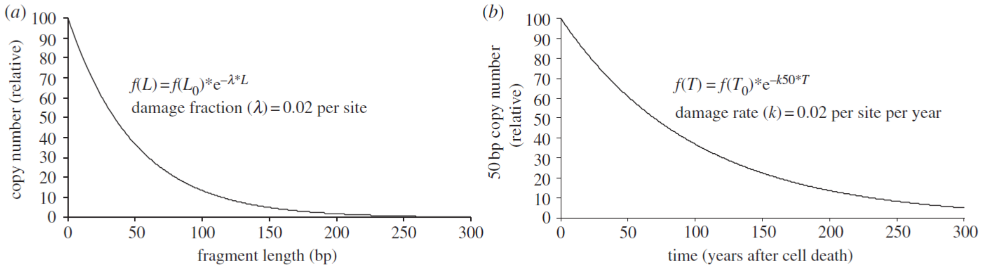


Figure 57. DNA fragmentation theory. *(a)* The exponential relationship caused by random fragmentation of DNA. *Post-mortem*, the template fragment length (*L*) distribution follows an exponential decline determined by the proportion of damaged sites (λ). This relationship has been described from both modern and ancient samples (Deagle et al., 2006, Schwarz et al., 2009, Brotherton et al., 2007). Here, a fragment size distribution representing *λ* = 0.02 (2% of the bonds in the DNA backbone are broken). *(b)* A hypothetical signal of temporal DNA decay, which has, prior to this study, been extremely difficult to demonstrate. The model assumes that the observed damage fraction (*λ*) can be converted to a rate of decay (*k*) when the age (*T*) of a sample is known. It implies that the number of DNA copies of a given length (*L*) will decline exponentially with time—hence the notion that DNA has a half-life. Here, the theoretical decay kinetics of a 50 bp DNA fragment, assuming a *k* of 2% per site per year. *k* is converted to a 50 bp decay rate (k_{50}), according to a Poisson distribution as: $k_{50} = 1 – (e^{-0.02*50})$ (Allentoft et al., 2012).

Finally, we used another parameter to confirm the antiquity of our reads, the error rate. It calculates the excess of derived alleles observed in an ancient genome (compared to a high-quality genome) taking into account that all anatomically modern humans should have the same percentage of derivate alleles. The error rate analysis were performed comparing our trimmed FASTQ with a specific tool of ANGSD program (Korneliussen et al., 2014), using the chimp genome as outgroup and the genome NA12778 (from the 1000 genomes project) as an error free sample (Schroeder et al., 2018).

4.3.2.3.3 Genotyping

SNP sites were called considering the Reich (Reich et al., 2012) Native American chip dataset and using the tool ANGSD (Analysis of Next Generation Sequencing Data), *--haplocall 1* option, which picks a random read starting from a set of input locations. The output was converted to PLINK *tped* format using ANGSD and merged with the comparative dataset using PLINK – 1.9 (Purcell et al., 2007).

The ancient Panamanian DNAs from this work were compared with a worldwide dataset of 846 samples available from literature (Table 12).

This dataset was initially pruned with PLINK following the parameters used by Scheib and colleagues (Scheib et al., 2018) Command 10. The term *-maf* means major allele frequency and we include only SNPs with an allele frequency higher than 0.05. The command *-indep-pairwise* searches for linkage disequilibrium considering window size of 200 variants and shifting by 25 variants. The third parameter *0.4* is a pairwise r^2 threshold and means that pair of variants with a squared correlation greater than 0.4 are deleted from the dataset. The *-geno* option includes SNPs on the basis of missing genotype rate and we impose to conserve SNPs with a 40% of genotyping rate (present in the 40% of the individuals), instead the *-mind* is to exclude samples with a missing genotype higher than 98%. Eventually, all samples were kept and 112453 of the initial 352972 SNPs were conserved.

Command 10:

```
--maf 0.05 and –indep-pairwise 200 25 0.4 –geno
0.6 –mind 0.98
```

Table 12. Comparative dataset used in this study.

Populations	Reference	Number of samples
Alaska	(Lindo et al., 2017)	1
Alaskan_Athabaskan	(Scheib et al., 2018)	1
Aleutians	(Reich et al., 2012)	5
Algonquin	(Reich et al., 2012)	5
Altai	(Rasmussen et al., 2010)	12
Arara	(Reich et al., 2012)	1
Arhuaco	(Reich et al., 2012)	5
Aymara	(Reich et al., 2012)	23
Baja	(Scheib et al., 2018)	2
Bribri	(Reich et al., 2012)	4
BritishColumbia	(Raghavan et al., 2015)	1
Buryats	(Rasmussen et al., 2010)	17
Cabecar	(Reich et al., 2012)	31
California	(Scheib et al., 2018)	2
Catalina	(Scheib et al., 2018)	2
Chane	(Reich et al., 2012)	2
Chilote	(Reich et al., 2012)	8
Chipewyan	(Reich et al., 2012)	15
Chono	(Reich et al., 2012)	4
Chorotega	(Reich et al., 2012)	1
Chukchis	(Rasmussen et al., 2010); (Reich et al., 2012)	11; 19
Chumash	(Scheib et al., 2018)	13
Clovis	(Rasmussen et al., 2014)	1
Colonist	(Scheib et al., 2018)	14
Cree	(Reich et al., 2012)	4
Diaguita	(Reich et al., 2012)	5
Dolgans	(Rasmussen et al., 2010)	4
EarlySanNicolas	(Scheib et al., 2018)	13
East_Greenlandic_Inuit	(Rasmussen et al., 2010)	7
Embera	(Reich et al., 2012)	5
Evenkis	(Rasmussen et al., 2010)	15
French	(HGDP)	28
Guahibo	(Reich et al., 2012)	6
Guarani	(Reich et al., 2012)	6
Guaymi	(Reich et al., 2012)	5
Han	(HGDP)	34
Han.NChina	(HGDP)	10
Huetar	(Reich et al., 2012)	1
Huilliche	(Reich et al., 2012)	4
Huron-Wendat	(Scheib et al., 2018)	2
Inga	(Reich et al., 2012)	9
Ipai	(Scheib et al., 2018)	1
Jamamadi	(Reich et al., 2012)	1
Kaingang	(Reich et al., 2012)	2
Kaqchikel	(Reich et al., 2012)	13
Kennewick	(Rasmussen et al., 2015)	1
Kets	(Rasmussen et al., 2010)	2
Khanty	(Reich et al., 2012)	35

Kitanemuk	(Scheib et al., 2018)	1
Kogi	(Reich et al., 2012)	4
Koryaks	(Rasmussen et al., 2010)	10
LaJollans	(Scheib et al., 2018)	8
LateSanNicolas	(Scheib et al., 2018)	15
Lucier	(Scheib et al., 2018)	5
Lucy Island	(Raghavan et al., 2015)	1
Maleku	(Reich et al., 2012)	3
Maya1	(MGDP); (Reich et al., 2012)	19; 1
Maya2	(MGDP)	12
MbutiPygmy	(HGDP)	13
Mixe	(Reich et al., 2012)	17
Mixtec	(Reich et al., 2012)	5
Mongolians	(Rasmussen et al., 2010)	8
Naukan	(Reich et al., 2012)	16
Nganasan2	(Reich et al., 2012)	14
Nganassans	(Rasmussen et al., 2010)	8
Ojibwa	(Reich et al., 2012)	5
Palikur	(Reich et al., 2012)	3
Parakana	(Reich et al., 2012)	1
Pericu	(Raghavan et al., 2015); (Scheib et al., 2018)	2; 1
Piaui	(Raghavan et al., 2015)	1
Pima	(Reich et al., 2012)	21
Prince Rupert Harbor	(Lindo et al., 2017)	2
Purepecha	(Reich et al., 2012)	1
Quechua	(Reich et al., 2012)	40
San_Francisco_Bay	(Scheib et al., 2018)	1
SanClemente	(Scheib et al., 2018)	5
Saqqaq	(Rasmussen et al., 2010)	1
Selkup	(Rasmussen et al., 2010)	9
Southwest	(Scheib et al., 2018)	5
Surui	(Reich et al., 2012)	16
Tepehuano	(MGDP); (Reich et al., 2012)	20; 5
Teribe	(Reich et al., 2012)	3
Ticuna	(Reich et al., 2012)	6
Toba	(Reich et al., 2012)	4
Tundra_Nentsi	(Reich et al., 2012)	3
Tuvinians	(Rasmussen et al., 2010)	15
Upward Sun River	(Moreno-Mayar et al., 2018a)	2
Waunana	(Reich et al., 2012)	3
Wayuu	(Reich et al., 2012)	11
West_Greenlandic_Inuit	(Rasmussen et al., 2010)	8
Wichi	(Reich et al., 2012)	5
Yaghan	(Reich et al., 2012)	4
Yakut	(Reich et al., 2012)	13
Yaqui	(Reich et al., 2012)	1
Yukaghir	(Reich et al., 2012)	13
Zapotec1	(Reich et al., 2012)	22
Zapotec2	(MGDP); (Reich et al., 2012)	4; 17
	Total	**846**

4.3.2.4 Whole-genome analysis

4.3.2.4.1 PCA

The unpruned dataset was initially used for a PCA analysis. Principle Component plots were generated using EIGENSOFT v 7.2.0 (Patterson et al., 2006, Price et al., 2006) (Command 11) with the options "lsqproject: YES" and "autoshrink: YES". The autoshrink allows to project the ancient genomes onto modern variation, which is needed due to the low coverage of ancient data and the high presence of missing segments; instead with lsqproject PCA projections are carried out by solving least squares equations rather than an orthogonal projection and it is desired to project samples with a lot of missing data onto the top of the PCs. Four components are calculated and we set 20 as the number of standard deviations that an individual must exceed (for each component) to be removed as an outlier.

Command 11:

```
evecoutname: America.shrink.evec.txt
evaloutname: America.shrink.eval
altnormstyle: NO
numoutevec: 4
familynames: NO
outliersigmathresh: 20
poplistname: pops.txt
lsqproject: YES
autoshrink: YES
```

Several PCAs were performed considering a worldwide dataset and different sub-datasets. Here we present only the data of the entire dataset and a "continental PCA" encompassing a subset of Native American populations.

4.3.2.4.2 Admixture

The pruned dataset was used to perform a biogeographical ancestry analysis with the ADMIXTURE v. 1.23 program (Alexander et al., 2009), which is a maximum likelihood estimation of individual ancestries from multi-locus SNP genotype datasets. We performed ten independent runs for each K, from *K*1 to *K*20, adding the *–cv* flag to identify the 5-fold cross-validation error at each *K*. The average cross-validation (cv) value for each *K* were plotted to select the most likelihood model. Ten different runs were analyzed together with the online software *CLUMPAK* (Kopelman et al., 2015). This software is specific for combining different runs of clustering programs.

4.3.2.4.3 *f3*-outgroup

AdmixTools v 4.1 (Patterson et al., 2006, Price et al., 2006) was used to perform a three populations test in order to analyze the closeness of the Panamanian aDNAs to all populations in the worldwide dataset using Mbuti as an outgroup. *f3- outgroup* analysis form: (*Panama, X; Mbuti)*. The X group could be any population in the dataset. Initially, our aDNA samples were clustered considering their age to create various sub-populations, but since no major differences were observed, they were eventually tested as a single population.

4.3.3 Results and discussion

4.3.3.1 Validation of ancient Panamanian genomes

In the ancient DNA laboratory of the University of Illinois, following the protocol described above, we were able to obtain DNA from 21 ancient samples that were eventually sequenced in three different runs (Table 13).

Table 13. Sequencing results of 21 ancient Panamanian samples.

Sample	WG Coverage	% Human Reads	Sex	Contamination Estimate	%GC	Seq-Run
Pa01	0.027	0.063	M	0.032	51	1
Pa02	0.111	0.064	F	0.008	59	2
Pa03	0.001	0.001	F	NA	60	1
Pa07	0.012	0.009	M	NA	60	2
Pa09	0.133	0.120	M	0.007	54	1
Pa10	0.583	0.391	M	0.009	51	1
Pa11	0.001	0.001	NA	NA	56	1
Pa12	0.001	0.001	F	NA	58	1
Pa14	0.000	0.000	M*	NA	54	2
Pa16A	0.429	0.361	M	0.013	48	1
Pa24A	0.265	0.176	M	0.004	56	1
Pa25	0.180	0.331	M	0.016	43	2
Pa26	0.008	0.008	M	0.028	56	2
Pa27	0.000	0.000	F*	NA	55	2
Pa28A	0.085	0.079	M	0.004	56	3
Pa29	0.176	0.158	M	0.003	52	2
Pa30A	0.134	0.106	M	0.010	55	2
PaC05	0.002	0.002	M	NA	55	3
PaC18	0.005	0.007	F	NA	51	3
PaC21	0.002	0.002	F	0.016	58	3
PaC22	0.000	0.001	NA	0.012	56	3

Only nine samples, out of the 21 initially sequenced (43%), showed enough coverage (~0.1-0.6X) and low contamination to allow analyses on the complete genome

(Figure 58). It is worth mentioning that the DNA of all nine "good samples" was extracted from petrous bone covering a time range from 1.4 to 0.5 kya (Figure 59), therefore all sequences used in the following analyses are from pre-Columbian samples.

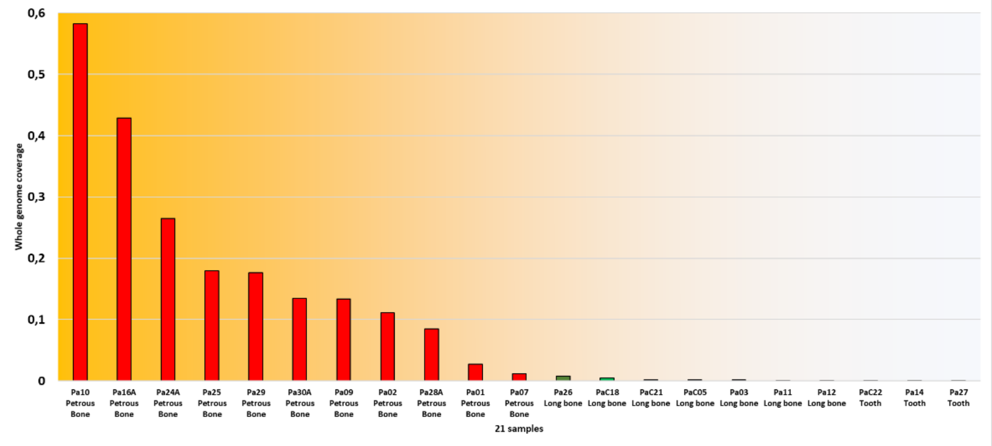


Figure 58. Mapping results. Different colors correspond to different bone sources: red from petrous bone; green from long bone, blue from tooth.

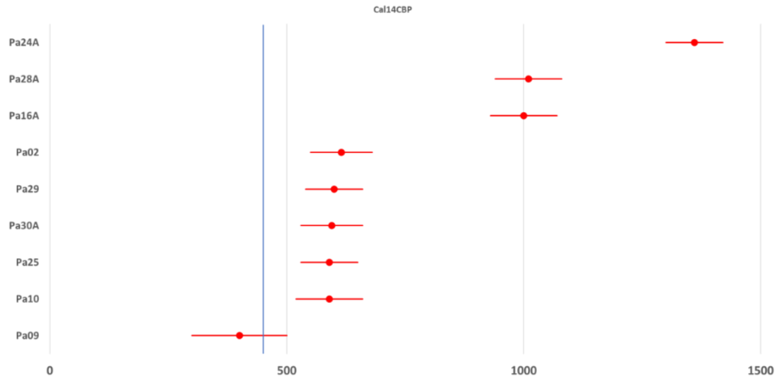


Figure 59. Radio carbon dates of the 9 samples used for whole-genome analyses. The blue line represents the beginning of the European colonialism, thus imaginary dividing pre- and post-Columbian times.

The validation analysis confirmed the common damage pattern of aDNA, such as the excess of misincorporations at the ends of the reads and the short length that is around 50-60 bps (Figure 60). These data confirmed that we were able to obtain real aDNA sequences from a tropical area. As expected only the petrous bone was able to conserve endogenous DNA in this extreme warm environment where the survival of DNA after death is very difficult.

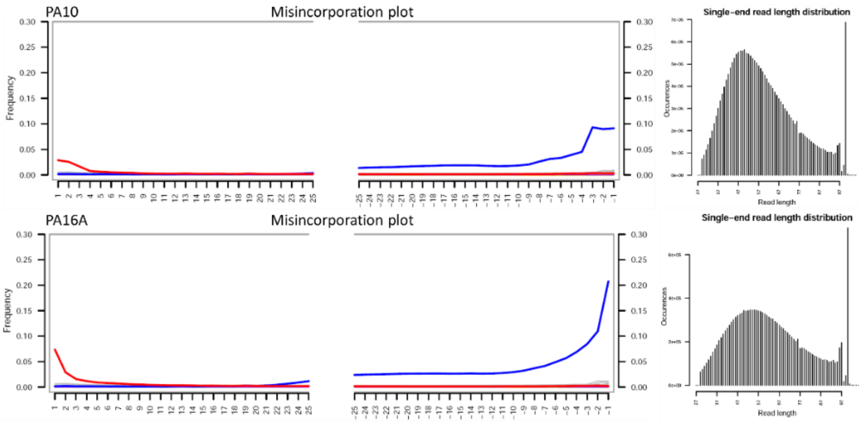


Figure 60. Two examples of the damage pattern found in our samples. The plots on the right are misincorporation frequencies (C to T in red; G to A in blue) at specific positions in terms of nucleotide distance from the 5" (left) or the 3" (right) end. The two plots on the left show the read length distribution.

The low preservation condition in Panama is then assessed by analyzing the molecular decay of our samples in comparison to three ancient Native American genomes published so far (Figure 61).

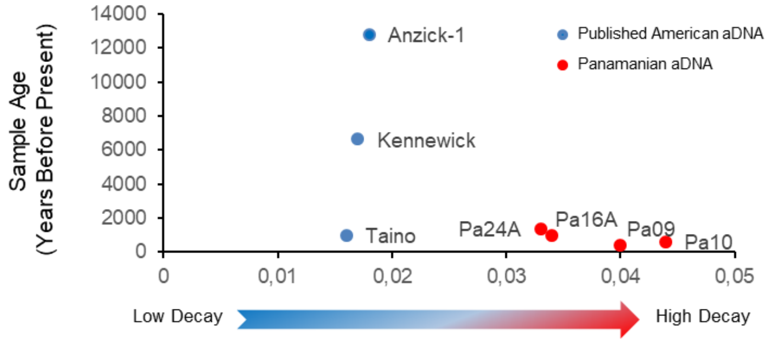


Figure 61. Comparison of molecular decay and age estimate between our samples and other three Native American ancient genomes.

In Panama and especially in the sites where our samples were excavated, close to the Pacific Ocean, the decay is higher than older samples (Anzick-1 and Kennewick) from colder regions, and also higher than the Caribbean ancient genome of Taino with a comparable age and from a similar environment (in the Caribbean Sea). Probably the latitude and the microclimate present in the Preacher's Cave played a fundamental role for well preserving the Taino sample. On the contrary our samples were found close to the ocean and remained submerged for a long time (due to tide) making even more complicated the DNA preservation. Despite the higher decay of our samples, the error rate (i.e. false sequencing calls due to the damage) is analogues to other published Native American aDNAs (Kistler et al., 2017a), thus validating our data (Figure 62).

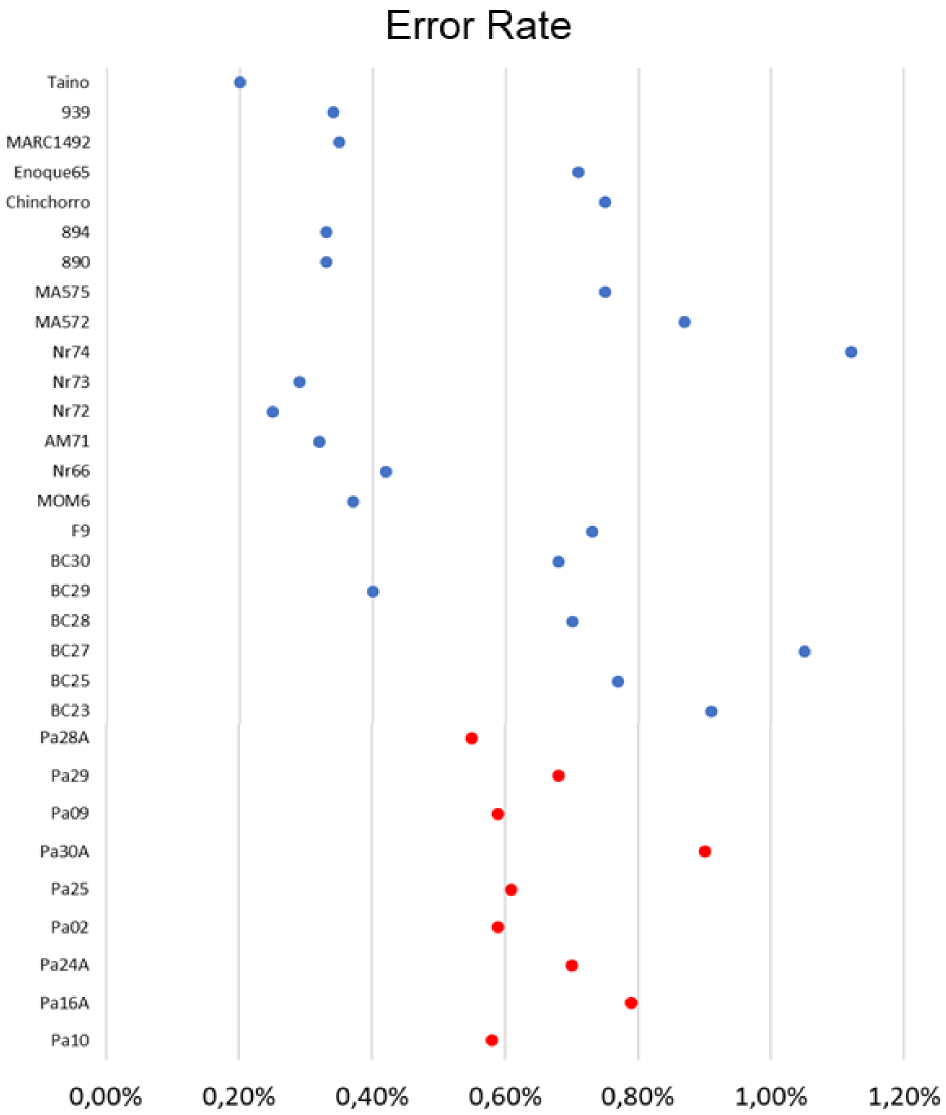


Figure 62. Comparison of overall error rate of ours samples with other Native American aDNAs.

4.3.3.2 A diachronic comparison between modern and ancient genomes

Once the reliability of our data was confirmed, we compared our low coverage Panamanian genomes with the available dataset of ancient and modern data (Table 12), trying to genetically place these new ancient Panamanian samples in a worldwide context. As in previous studies (Scheib et al., 2018, Schroeder et al., 2018), the dataset used for the PCA encompasses samples typed with the 350K SNPs of Reich et al. (Reich et al., 2012). Plotting the first two components in a world-wide context (Figure 63) the Panama aDNAs map among the Native American populations, which in turn are closer to Asian samples than to Europeans and Africans.

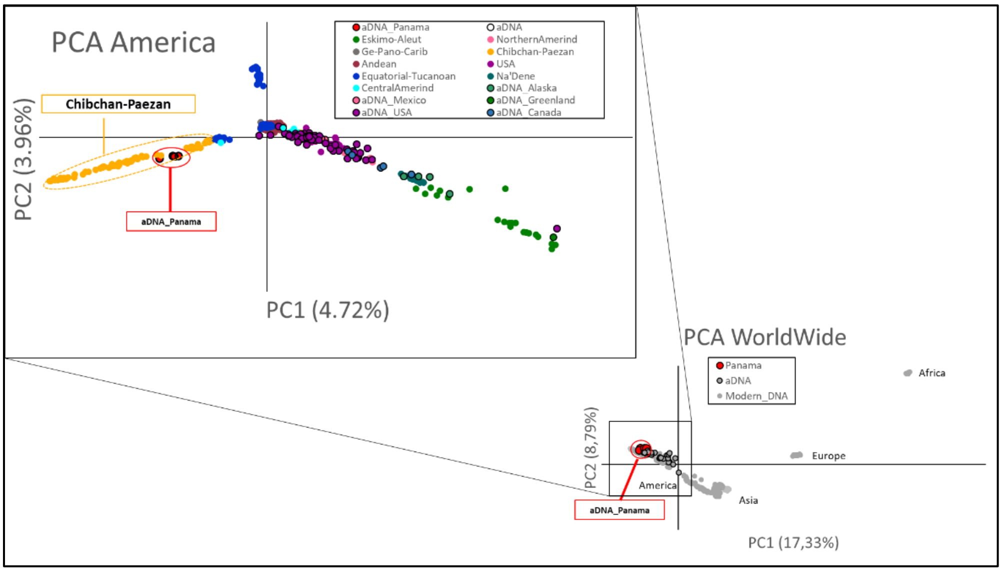


Figure 63. PCA analysis. For the bottom right PCA the entire dataset was used, while the top left PCA encompasses only Native American populations.

Zooming to a "continental PCA" based only on Native American samples (Figure 63) the ancient Panamanians cluster among the modern Chibchan speakers, which are separated from all other Native populations by the first component. Close to the Chibchan cluster there are the Wayuu and Chorotega samples, which speak other languages (respectively Arawakan and Mangue in the Equatorial-Tucanon group) but are geographically related to the Isthmus. This proximity might be explained by two hypotheses: Either the genetic link predated (and survived) the following separation in speaking groups or more recenty there was a gene flow between tribes with different languages.

The affinity between the Panamanian ancient samples and the Chibchan speaking group is also confirmed by the admixture analysis. The plot with the highest likelihood is the one representing K9 (Figure 64).

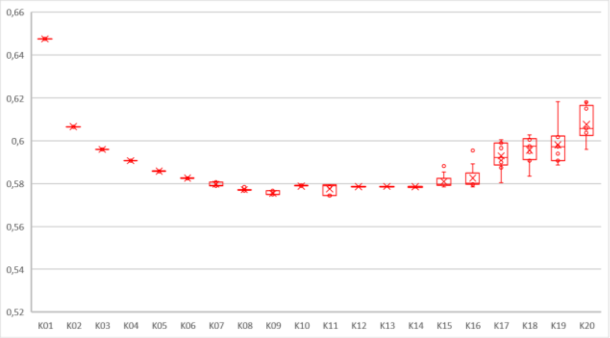


Figure 64. Admixture CV box plot. For each K we performed 10 runs. The horizontal lines are medians, while the X are means. Dots represent CV values, 25-75% of them included in the red box, while 9-91% represented by the vertical lines.

Most of the genetic ancestry of our Panamanian ancient samples is explained by two major ancestral components, K1 and K4. The K1 is shared by almost all the Native populations (Figure 65), with the highest value in the Andean population. The component that characterizes the Panamanian aDNAs and all the Chibchan tribes is K4. Even if it is quite dispersed in the surrounding Native Americans, K4 represents a high proportion of the genetic structure of the Chibchan, with the highest level in the Cabecar. It is interesting to point out that this component reaches ~30% also in the Wayuu and Chorotega, confirming the genetic link between these groups and the Chibchan, already observed in the PCA. The geographic distribution of these two components (Figure 66) shows that K1 seems to mirror the first peopling of America, with a long stretched distribution all over the double continent and some picks on the Pacific coast, the highest in the Andean populations. On the contrary, the K4 component has a restricted geographic distribution in the Isthmo-Colombian area, possibly overlapping with the distribution of the Chibchan culture dispersion.

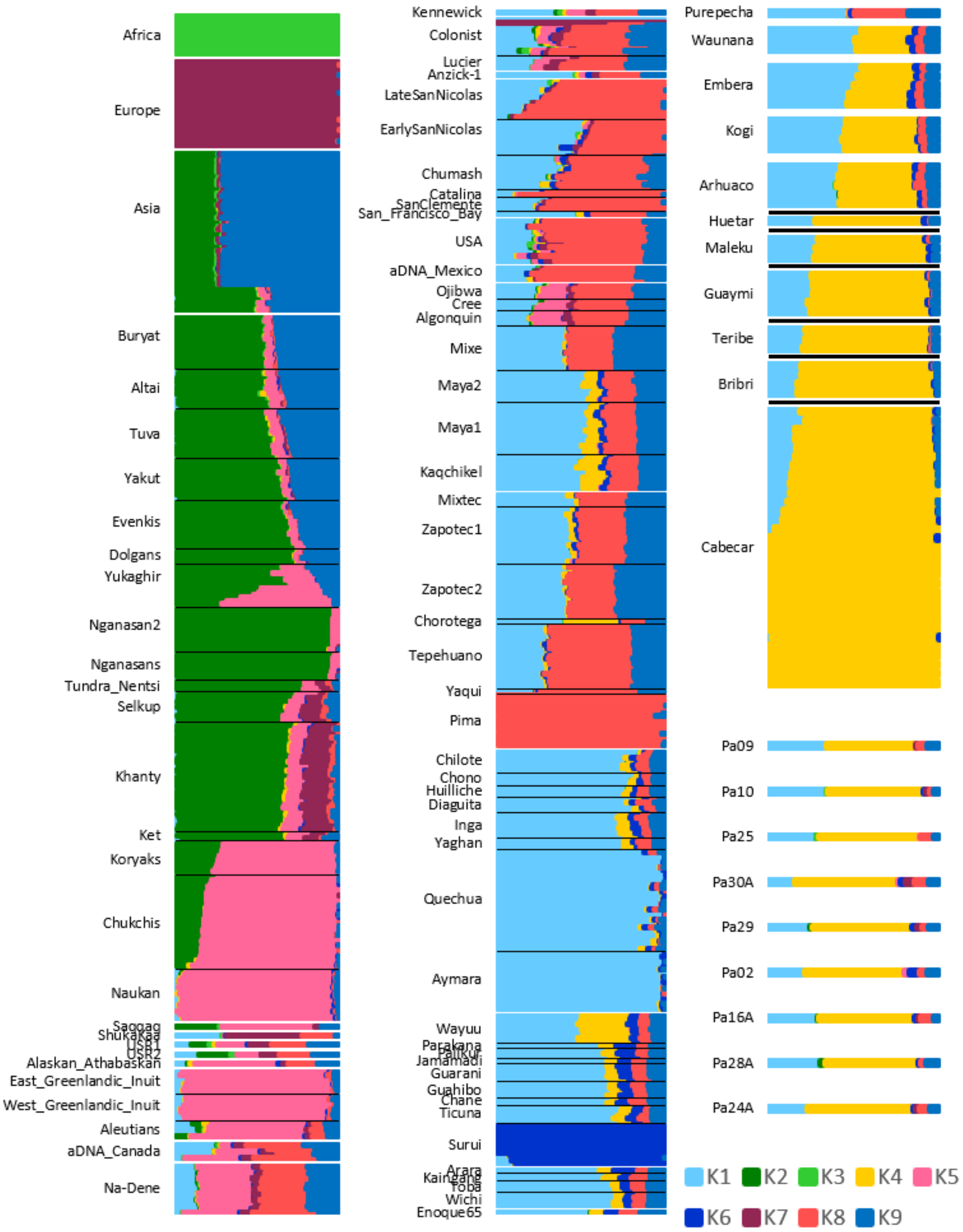


Figure 65. Admixture plot of the K9.

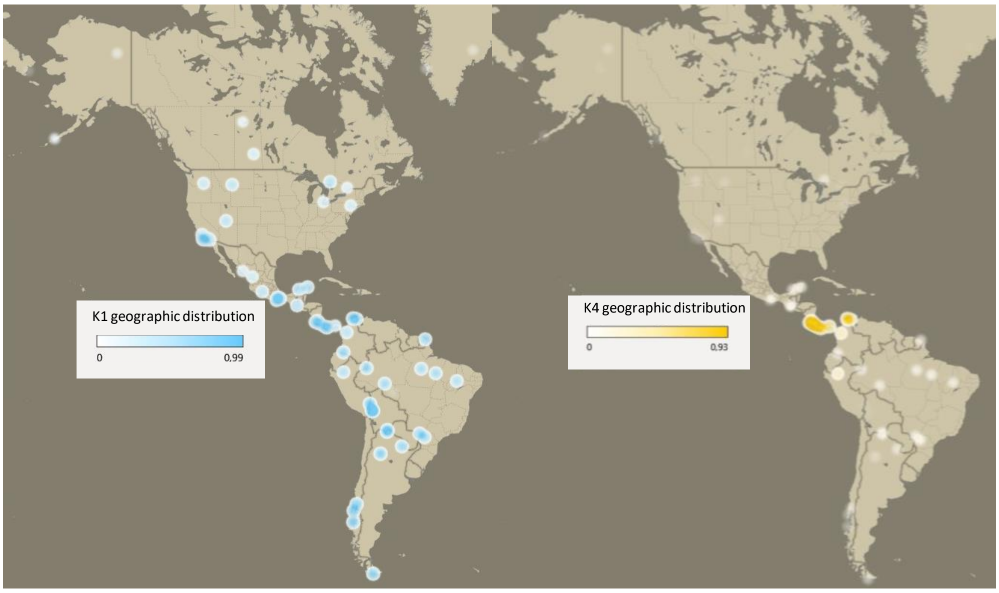


Figure 66. Geographic distribution of the two major components, K1 in blue and K4 in yellow, in our samples.

To highlight the shared genetic drift of our samples with other Native populations we performed an *f3-outgroup* analysis (Figure 67). The results confirm a common genetic history shared between ancient Panamanians and Chibchan speakers. A genetic proximity with the Chorotega is also confirmed, while no relationships with the Wayuu group was observed, probably due to the low number of SNPs retained in the comparison with our ancient data. The geographic representation of the *f3* values (Figure 67) highlights the affinities of the Panamanian samples with the tribes of Isthmo-Colombian area, the highest in Costa-Rica, but still very high in the northern part South America (Colombia and Venezuela). The decrease pattern seems to be higher towards North America than to the south, which might be the result of a backward migration from South to Central America.

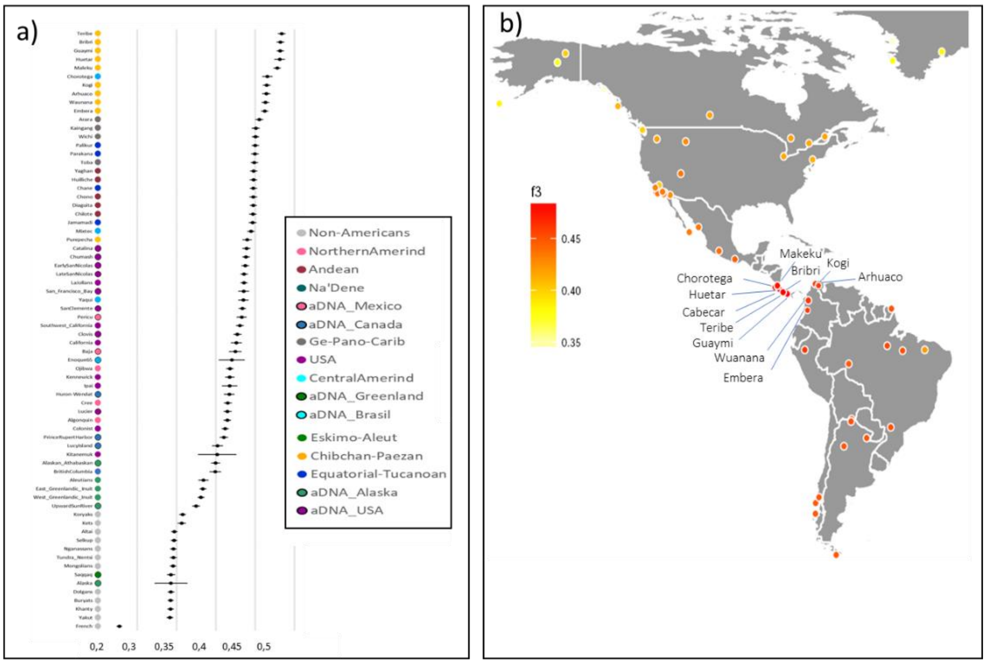


Figure 67. *f3-outgroup* results. A) *f3-outgroup* values obtain by testing (aDNA_Panama,X;Mbuti) where X could be any Native American population; B) Geographic heat map of *f3- statistic outgroup* test; warmer colors indicate higher levels of allele sharing.

4.3.4 Conclusion

In this PhD thesis, we present the first nine ancient genomes from Panama. Despite the tropical environment of the Isthmus, challenging any preservation of aDNA as confirmed by the high molecular decay, the availability of the petrous bone and employment on NGS techniques, allowed us to obtain low-coverage genomes from ancient samples covering a time range from 1.4 to 0.5 kya (pre-Columbian period). After verifying that no major changes occurred in the genetic structure of Panama over the one thousand years represented by our ancient samples (also confirmed by the mitogenome analyses until present days), the nine ancient genomes were compared together with a worldwide dataset of human populations and in particular with a selected group of 77 ancient and modern Native American populations.

In our analyses, the ancient Panamanians show affinities, in terms of genetic proximity (PCA), ancestries (admixture) and shared genetic drift (*f3-statistics*), with the Chibchan speaking tribes currently living in the Isthmo-Colombian area from southern Nicaragua to Colombia. Therefore, our genomes can also be considered the first ancient DNAs from Chibchans and the high similarity between ancient and modern genomes suggests a genetic continuity in this linguistic group. Thus, the proposed genome-wide uniqueness of modern Chibchans in the Native American context (Reich et al., 2012) can now be extended over the last fifteen centuries, as testified by the Chibchan cluster of our PCA plot. Markedly, we were able to identify a specific ancestral component (K4 in the admixture analysis) that left a distinctive sign in the Chibchan genomes of the Isthmo-Colombian area, while the other major component (K1) probably arrived earlier in the region with the first Paleo-Indian settlers and is currently widespread on the double continent. This "*Chibchan*" ancient component is probably responsible also for the affinities with two other Central American tribes (the Wayuu and the Chorotega) belonging to other linguistic groups (Arawakan and Mangue), thus indicating that this component predates the differentiation of native languages and was partially maintained at a later time, even if the hypothesis of recent gene flows cannot be completely ruled out.

Finally, the higher level of shared genetic drift with some South American groups than with most of the North American ones, might suggest two different and both fascinating hypotheses in the pre-Columbian history of Native Americans: i) an ancient back migration from South to Central America (and *vice versa*); ii) additional streams of gene flows from Asia that "partially diluted" the original gene pool of North American populations, without influencing the central and southern groups; the genetic legacy of first migrants is still significant (e.g. the highest K1 component of the Andes) and was only partially reshaped by local differentiations and migrations (e.g. the K4 ancestry in the Isthmo-Columbian area). Further analyses regarding the admixture events and the reconstruction of an autosomal tree will clarify these hypotheses, also providing further details on the genetic history of the peculiar Chibchan group.

References

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V et al. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet. 2004;75(5):910-918

Achilli A, Olivieri A, Pellecchia M, Uboldi C, Colli L, Al-Zahery N, Accetturo M, Pala M, Hooshiar Kashani B, Perego UA et al. Mitochondrial genomes of extinct aurochs survive in domestic cattle. Curr Biol. 2008;18(4):R157-158

Achilli A, Bonfiglio S, Olivieri A, Malusa A, Pala M, Kashani BH, Perego UA, Ajmone-Marsan P, Liotta L, Semino O et al. The multifaceted origin of taurine cattle reflected by the mitochondrial genome. PLoS One**.** 2009;4(6):e5753

Achilli A, Olivieri A, Soares P, Lancioni H, Hooshiar Kashani B, Perego UA, Nergadze SG, Carossa V, Santagostino M, Capomaccio S et al. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc Natl Acad Sci U S A. 2012;109(7):2449-2454

Achilli A, Perego UA, Lancioni H, Olivieri A, Gandini F, Hooshiar Kashani B, Battaglia V, Grugni V, Angerhofer N, Rogers MP et al. Reconciling migration models to the Americas with the variation of North American native mitogenomes. Proc Natl Acad Sci U S A. 2013;110(35):14308-13

Achilli A, Olivieri A, Semino O, Torroni A. Ancient human genomes-keys to understanding our past. Science. 2018;360(6392):964-965

Adams AL. On the discovery of remains of Halitherium in the Miocene deposits of Malta. J Geol Soc. 1866;22(1-2):595-596

Ajmone-Marsan P, Garcia JF, Lenstra JA. On the origin of cattle: how aurochs became cattle and colonized the world. Evol Anthropol. 2010;19(4):148-157

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–1664

Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics. 2011;12(1):246

Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert MTP, Willerslev E. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc Biol Sci. 2012;rspb20121745

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L et al. Population genomics of bronze age Eurasia. Nature. 2015;522(7555):167

Almathen F, Charruau P, Mohandesan E, Mwacharo JM, Orozco-terWengel P, Pitt D, Abdussamad AM, Uerpmann M, Uerpmann H-P, De Cupere B et al. Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. Proc Natl Acad Sci U S A. 2016;113(24):6707-6712

Amano T, Miyakoshi Y, Takada T, Kikkawa Y, Suzuki H. Genetic variants of ribosomal DNA and mitochondrial DNA between swamp and river buffaloes. Anim Genet. 1994;25(S1):29-36

Ammerman AJ, Cavalli-Sforza LL. The Neolithic transition and the genetics of populations in Europe. Princeton University Press. 1984

Anati AF, Anati E. Missione a Malta: ricerche e studi sulla preistoria dell'arcipelago maltese nel contesto mediterraneo. Editoriale Jaca Book. 1988

Anderson RP, Handley CO. Dwarfism in insular sloths: biogeography, selection, and evolutionary rate. Evolution. 2002;56(5):1045-58

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290(5806):457

Andrews S. FastQC: a quality control tool for high throughput sequence data. Cambridge, UK: Babraham Institute. 2010

Arias L, Barbieri C, Barreto G, Stoneking M, Pakendorf B. High-resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia. Am J Phys Anthropol. 2018;165(2):238-255

Asouti E, Fuller DQ, Barker G, Finlayson B, Matthews R, Fazeli Nashli H, McCorriston J, Riehl S, Rosen AM. A contextual approach to the emergence of agriculture in Southwest Asia: reconstructing Early Neolithic plant-food production. Curr Anthropol. 2013;54(3):299-345

Avise JC, Nelson WS, Bowen BW, Walker D. Phylogeography of colonially nesting seabirds, with special reference to global matrilineal patterns in the sooty tern (*Sterna fuscata*). Mol Ecol. 2000;9(11):1783-1792

Avital G, Buchshtav M, Zhidkov I, Tuval J, Dadon S, Rubin E, Glass D, Spector TD, Mishmar D. Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins. Hum Mol Genet. 2012;21(19):4214-4224

Ballard JWO, Whitlock MC. The incomplete natural history of mitochondria. Mol Ecol. 2004;13(4):729-744

Barker JSF, Moore SS, Hetzel DJS, Evans D, Tan SG, Byrne K. Genetic diversity of Asian water buffalo (*Bubalus bubalis*): microsatellite variation and a comparison with protein-coding loci. Anim Genet. 1997a;28(2):103-115

Barker JSF, Tan SG, Selvaraj OS, Mukherjee TK. Genetic variation within and relationships among populations of Asian water buffalo (*Bubalus bubalis*). Anim Genet. 1997b;28(1):1-13

Barrantes R, Smouse PE, Mohrenweiser HW, Gershowitz H, Azofeifa J, Arias TD, Neel JV. Microevolution in lower Central America: genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. Am J Hum Genet. 1990;46(1):63

Barrell B, Bankier A, Drouin J. A different genetic code in human mitochondria. Nature. 1979;282(5735):189

Battaglia V, Gabrieli P, Brandini S, Capodiferro MR, Javier PA, Chen X-G, Achilli A, Semino O, Gomulski LM, Malacrida AR et al. The Worldwide spread of the Tiger mosquito as revealed by mitogenome haplogroup diversity. Front Genet. 2016;7(208)

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D et al. The dawn of human matrilineal diversity. Am J Hum Genet. 2008;82(5):1130-1140

Beja-Pereira A, Caramelli D, Lalueza-Fox C, Vernesi C, Ferrand N, Casoli A, Goyache F, Royo LJ, Conti S, Lari M et al. The origin of European cattle: evidence from modern and ancient DNA. Proc Natl Acad Sci U S A. 2006;103(21):8113-8118

Bellini R, Calvitti M, Medici A, Carrieri M, Celli G, Maini S. Use of the sterile insect technique against *Aedes albopictus* in Italy: first results of a pilot trial. Area-Wide Control of Insect Pests. Dordrecht : Springer. 2007;505-515

Benedict MQ, Levine RS, Hawley WA, Lounibos LP. Spread of the tiger: global risk of invasion by the mosquito *Aedes albopictus*. Vector Borne Zoonotic Dis. 2007;7(1):76-85

Berger J-F, Guilaine J. The 8200 cal BP abrupt environmental change and the Neolithic transition: a Mediterranean perspective. Quat Int. 2009;200(1-2):31-49

Bibi F. A multi-calibrated mitochondrial phylogeny of extant Bovidae (*Artiodactyla, Ruminantia*) and the importance of the fossil record to systematics. BMC Evol Biol. 2013;13(1):1-15

Bocquet-Appel J-P. The agricultural demographic transition during and after the agriculture inventions. Curr Anthropol. 2011;52(S4):S497-S510

Bodner M, Perego UA, Huber G, Fendt L, Rock AW, Zimmermann B, Olivieri A, Gomez-Carballa A, Lancioni H, Angerhofer N et al. Rapid coastal spread of First Americans: Novel insights from South America's Southern Cone mitochondrial genomes. Genome Res. 2012;22(5):811-820

Bogucki P. The spread of early farming in Europe. Am Sci. 1996;84(3):242-253

Bollongino R, Burger J, Powell A, Mashkour M, Vigne JD, Thomas MG. Modern taurine cattle descended from small number of Near-Eastern founders. Mol Biol Evol. 2012;29:1185-1192

Bonfiglio S, Achilli A, Olivieri A, Negrini R, Colli L, Liotta L, Ajmone-Marsan P, Torroni A, Ferretti L. The enigmatic origin of bovine mtDNA haplogroup R: sporadic interbreeding or an independent event of *Bos primigenius* domestication in Italy?. PLoS One**.** 2010;5(12):e1576

Bonfiglio S, Ginja C, De Gaetano A, Achilli A, Olivieri A, Colli L, Tesfaye K, Agha SH, Gama LT, Cattonaro F et al. Origin and spread of *Bos taurus*: new clues from mitochondrial genomes belonging to haplogroup T1. PLoS One**.** 2012;7(6):e38601

Bonizzoni M, Gasperi G, Chen X, James AA. The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. Trends Parasitol. 2013;29(9):460-468

Borg J. Agriculture and horticulture in Malta. Malta and Gibraltar illustrated. London England. 1915

Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32(3):314

Brandini S, Bergamaschi P, Fernando Cerna M, Gandini F, Bastaroli F, Bertolini E, Cereda C, Ferretti L, Gomez-Carballa A, Battaglia V et al. The Paleo-Indian entry into South America according to Mitogenomes. Mol Biol Evol. 2018;35(2):299-311

Brandstätter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, Parsons TJ. Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. Int J Legal Med. 2004;118(5):294-306

Broodbank C. The making of the middle sea: a history of the Mediterranean from the beginning to the emergence of the classical world. London: Oxford University Press. 2013

Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A. Novel high-resolution characterization of ancient DNA reveals C> U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Res. 2007;35(17):5717-5728

Brown WM, George M, Wilson AC. Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci U S A. 1979;76(4):1967-1971

Bruford MW, Bradley DG, Luikart G. DNA markers reveal the complexity of livestock domestication. Nat Rev Genet. 2003;4:900

Calloway CD, Reynolds RL, Herrin Jr GL, Anderson WW. The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. Am J Hum Genet. 2000;66(4):1384-1397

Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987;325(6099):31-36

Cannon C. The ecology of tropical East Asia. The Quarterly Review of Biology. Oxford University Press. 2015;90(4):433-433

Cardinali I, Lancioni H, Giontella A, Capodiferro MR, Capomaccio S, Buttazzoni L, Biggio GP, Cherchi R, Albertini E, Olivieri A et al. An overview of ten Italian horse breeds through mitochondrial DNA. PLoS One**.** 2016;11(4):e0153004

Cavalli-Sforza LL, Barrai I, Edwards A. Analysis of human evolution under random genetic drift. Cold Spring Harb Symp Quant Biol. 1964;29:9-20

Cavalli-Sforza LL, Menozzi P, Cavalli-Sforza L, Piazza A, Cavalli-Sforza L. The history and geography of human genes. Princeton university press. 1994

Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet.. 2003;33:266

Ceppellini R, Curtoni E, Mattiuz P, Miggiano V, Scudeller G, Serra A. Genetics of leukocyte antigens: a family study of segregation and linkage. Histocompatibility testing. Kopenhagen: Munksgaard. 1967;149

Cerezo M, Achilli A, Olivieri A, Perego UA, Gómez-Carballa A, Brisighelli F, Lancioni H, Woodward SR, López-Soto M, Carracedo Á et al. Reconstructing ancient mitochondrial DNA links between Africa and Europe. Genome Res. 2012;22(5):821-826

Chaubey G, Metspalu M, Kivisild T, Villems R. Peopling of South Asia: investigating the caste–tribe continuum in India. Bioessays. 2007;29(1):91-100

Chaubey G, Karmin M, Metspalu E, Metspalu M, Selvi-Rani D, Singh VK, Parik J, Solnik A, Naidu BP, Kumar A et al. Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. BMC Evol Biol. 2008;8(1):227

Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas J et al. Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci U S A. 2015;112(44):E5907-E5915

Chen Y-S, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet. 1995;57(1):133

Chiang CW, Marcus JH, Sidore C, Al-Asadi H, Zoledziewska M, Pitzalis M, Busonero F, Maschio A, Pistis G, Steri M et al. Population history of the Sardinian people inferred from whole-genome sequencing. bioRxiv. 2016;92148

Clayton DA, Doda JN, Friedberg EC. The absence of a pyrimidine dimer repair mechanism in mammalian mitochondria. Proc Natl Acad Sci U S A. 1974;71(7):2777-2781

Cockrill WR. The husbandry and health of the domestic buffalo. Food and agricultural organization of the United nations. Rome. 1974

Cockrill WR. The water buffalo: a review. Br Vet J. 1981;137(1):8-16

Colli L, Lancioni H, Cardinali I, Olivieri A, Capodiferro MR, Pellecchia M, Rzepus M, Zamani W, Naderi S, Gandini F et al. Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. BMC Genomics. 2015;16(1):1-12

Colli L, Milanesi M, Vajana E, Iamartino D, Bomba L, Puglisi F, Del Corvo M, Nicolazzi EL, Ahmed SS, Herrera JR et al. New insights on water buffalo genomic diversity and post-domestication migration routes from medium density SNP chip data. Front Genet. 2018;9:53

Conolly J, Colledge S, Dobney K, Vigne J-D, Peters J, Stopp B, Manning K, Shennan S. Meta-analysis of zooarchaeological data from SW Asia and SE Europe provides insight into the origins and spread of animal husbandry. J Archaeol Sci. 2011;38(3):538-545

Cooke R, Isaza I, Griggs J, Desjardins B, Sánchez LA. Who crafted, exchanged, and displayed gold in Pre-Columbian Panama?. Gold and Power in Ancient Costa Rica, Panama, and Colombia. Washington DC:Dumbarton Oaks. 2003;91-158

Cooke R, Ranere A, Pearson G, Dickau R. Radiocarbon chronology of early human settlement on the Isthmus of Panama (13,000–7000 BP) with comments on cultural affinities, environments, subsistence, and technological change. Quat Int. 2013;301:3-22

Coon CS. The rock art of Africa. Science. 1963;142(3600):1642-1645

Cooper A, Poinar HN. Ancient DNA: do it right or not at all. Science. 2000;289(5482):1139-1139

Cropp S, Boinski S. The Central American squirrel monkey (*Saimiri oerstedii*): introduced hybrid or endemic species?. Mol Phylogenet Evol. 2000;16(3):350-365

Currat M, Ruedi M, Petit RJ, Excoffier L. The hidden side of invasions: massive introgression by local genes. Evolution. 2008;62(8):1908-1920

Cymbron T, Freeman AR, Malheiro MI, Vigne J-D, Bradley DG. Microsatellite diversity suggests different histories for Mediterranean and Northern European cattle populations. Proc Biol Sci. 2005;272(1574):1837-1843

Dabney J, Meyer M, Pääbo S. Ancient DNA damage. Cold Spring Harb Perspect Biol. 2013;a012567

Daly KG, Maisano Delser P, Mullin VE, Scheu A, Mattiangeli V, Teasdale MD, Hare AJ, Burger J, Verdugo MP, Collins MJ et al. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. Science. 2018;361(6397):85-88

De Deckker P, Tapper NJ, van der Kaars S. The status of the Indo-Pacific Warm Pool and adjacent land at the Last Glacial Maximum. Glob Planet Change. 2003;35(1–2):25-35

de Saint Pierre M, Bravi CM, Motti JM, Fuku N, Tanaka M, Llop E, Bonatto SL, Moraga M. An alternative model for the early peopling of southern South America revealed by analyses of three mitochondrial DNA haplogroups. PLoS One**.** 2012a;7(9):e43486

de Saint Pierre M, Gandini F, Perego UA, Bodner M, Gómez-Carballa A, Corach D, Angerhofer N, Woodward SR, Semino O, Salas A et al. Arrival of Paleo-Indians to the Southern Cone of South America: new clues from mitogenomes. PLoS One**.** 2012b;7(12):e51311

Deagle BE, Eveson JP, Jarman SN. Quantification of damage in DNA recovered from highly degraded samples–a case study on DNA in faeces. Front Zool. 2006;3(1):11

DeGiorgio M, Jakobsson M, Rosenberg NA. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proc Natl Acad Sci U S A. 2009;106(38):16057-16062

Di Lorenzo P, Lancioni H, Ceccobelli S, Colli L, Cardinali I, Karsli T, Capodiferro MR, Sahin E, Ferretti L, Ajmone Marsan P et al. Mitochondrial DNA variants of Podolian cattle breeds testify for a dual maternal origin. PLoS One**.** 2018;13(2):e0192567

Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002;418(6898):700

Diamond J, Bellwood P. Farmers and their languages: the first expansions. Science. 2003;300(5619):597-603

Dillehay TD. Probing deeper into first American studies. Proc Natl Acad Sci U S A. 2009;106(4):971-878

Dillehay TD, Ocampo C, Saavedra J, Sawakuchi AO, Vega RM, Pino M, Collins MB, Scott Cummings L, Arregui I, Villagran XS et al. New archaeological evidence for an early Human presence at Monte Verde, Chile. PLoS One**.** 2015;10(11):e0141923

Dillehay TD, Goodbred S, Pino M, Vásquez Sánchez VF, Tham TR, Adovasio J, Collins MB, Netherly PJ, Hastorf CA, Chiou KL et al. Simple technologies and diverse food strategies of the Late Pleistocene and Early Holocene at Huaca Prieta, Coastal Peru. Sci Adv. 2017;3(5):e1602778

DiMauro S, Schon EA. Mitochondrial respiratory-chain diseases. N Engl J Med. 2003;348(26):2656-2668

Dobney K, Larson G. Genetics and animal domestication: new windows on an elusive process. J Zool. 2006;269(2):261-271

Driscoll CA, Macdonald DW, O'Brien SJ. From wild animals to domestic pets, an evolutionary view of domestication. Proc Natl Acad Sci U S A. 2009;106Supplement1:9971-9978

Dritsou V, Topalis P, Windbichler N, Simoni A, Hall A, Lawson D, Hinsley M, Hughes D, Napolioni V, Crucianelli F et al. A draft genome sequence of an invasive mosquito: an Italian *Aedes albopictus*. Pathog Glob Health. 2015;109(5):207-220

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214

Duran C, Appleby N, Edwards D, Batley J. Molecular genetic markers: discovery, applications, data storage and visualisation. Curr Bioinform. 2009;4(1):16-27

Dyson SL, Rowland Jr RJ. Archaeology and history in Sardinia from the Stone Age to the Middle Ages: Shepherds, sailors, and conquerors. University of Pennsylvania Museum of Archaeology and Anthropology. 2007

Edwards CJ, Magee DA, Park SDE, McGettigan PA, Lohan AJ, Murphy A, Finlay EK, Shapiro B, Chamberlain AT, Richards MB et al. A complete mitochondrial genome sequence from a Mesolithic wild aurochs (*Bos primigenius*). PLoS One**.** 2010;5(2):e9255

Epstein H. Domestic animals of China. Winterton: Commonwealth Agricultural Bureaux. 1969

Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA, Zago MA, Ribeiro-dos-Santos AK et al. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet. 2008;82(3):583-592

FAO. http://dad.fao.org. /. 2014

Ferradini N, Lancioni H, Torricelli R, Russi L, Dalla Ragione I, Cardinali I, Marconi G, Gramaccia M, Concezzi L, Achilli A et al. Characterization and phylogenetic analysis of ancient Italian landraces of pear. Front Plant Sci. 2017;8:751

Fisher RA. XXI.—on the dominance ratio. Proc R Soc Edinb. 1923;42:321-341

Forster P, Harding R, Torroni A, Bandelt HJ. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet. 1996;59(4):935-945

Forster P, Matsumura S. Did early humans go north or south?. Science. 2005;308(5724):965-966

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science. 2013;341(6145):565-569

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prüfer K, de Filippo C et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514(7523):445

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A et al. The genetic history of ice age Europe. Nature. 2016;534(7606):200

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Pääbo S. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A. 2013a;110(6):2223-2227

Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol. 2013b;23(7):553-559

Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, Domboróczki L, Kővári I, Pap I, Anders A et al. Genome flux and stasis in a five millennium transect of European prehistory. Nat Commun. 2014;5:5257

Gamble C, Davies W, Pettitt P, Richards M. Climate change and evolving human diversity in Europe during the last glacial. Philos Trans R Soc Lond B Biol Sci. 2004;359(1442):243-254

Gasparre G, Porcelli AM, Lenaz G, Romeo G. Relevance of mitochondrial genetics and metabolism in cancer development. Cold Spring Harb Perspect Biol. 2013;5(2):a011411

Gasperi G, Bellini R, Malacrida AR, Crisanti A, Dottori M, Aksoy S. A new threat looming over the Mediterranean basin: emergence of viral diseases transmitted by *Aedes albopictus* mosquitoes. PLoS Negl Trop Dis. 2012;6(9):e1836

Gaunitz C, Fages A, Hanghøj K, Albrechtsen A, Khan N, Schubert M, Seguin-Orlando A, Owens IJ, Felkel S, Bignon-Lau O et al. Ancient genomes revisit the ancestry of domestic and Przewalski's horses. Science. 2018;360(6384):111-114

Gibbons A. Anthropology. A new view of the birth of *Homo sapiens*. Science. 2011;331(6016):392-394

Gilbert MTP, Jenkins DL, Götherstrom A, Naveran N, Sanchez JJ, Hofreiter M, Thomsen PF, Binladen J, Higham TF, Yohe RM et al. DNA from pre-Clovis human coprolites in Oregon, North America. Science. 2008;320(5877):786-789

Giles RE, Blanc H, Cann HM, Wallace DC. Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A. 1980;77(11):6715-6719

Goebel T, Potter B. First traces> Late pleistocene human settlement if the arctic. The Oxford Handbook of the Prehistoric Arctic. Oxford: Oxford University Press. 2016;223

González AM, Larruga JM, Abu-Amero KK, Shi Y, Pestano J, Cabrera VM. Mitochondrial lineage M1 traces an early human backflow to Africa. BMC Genomics. 2007;8(1):223

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-722

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet.. 2011;43(10):1031

Grugni V, Battaglia V, Perego UA, Raveane A, Lancioni H, Olivieri A, Ferretti L, Woodward SR, Pascale JM, Cooke R et al. Exploring the Y chromosomal ancestry of modern Panamanians. PLoS One**.** 2015;10(12):e0144223

Günther T, Valdiosera C, Malmström H, Ureña I, Rodriguez-Varela R, Sverrisdóttir ÓO, Daskalaki EA, Skoglund P, Naidoo T, Svensson EM et al. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. Proc Natl Acad Sci U S A. 2015;112(38):11917-11922

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522(7555):207

Hassanin A, Bonillo C, Nguyen BX, Cruaud C. Comparisons between mitochondrial genomes of domestic goat (*Capra hircus*) reveal the presence of numts and multiple sequencing errors. Mitochondrial DNA. 2010;21(3-4):68-76

Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, Jagadeesan A, Ebenesersdóttir SS, Kong A, Stefánsson K. The Y-chromosome point mutation rate in humans. Nat Genet.. 2015;47(5):453-457

Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD. Fine-scale population structure and the era of next-generation sequencing. Hum Mol Genet. 2010;19(R2):R221-R226

Herlihy PH. Central American Indian peoples and lands today. Central America: A Natural and Cultural History. New Haven: Yale University Press. 1997;215-240

Hervella M, Rotea M, Izagirre N, Constantinescu M, Alonso S, Ioana M, Lazăr C, Ridiche F, Soficaru AD, Netea MG et al. Ancient DNA from South-East Europe reveals different events during Early and Middle Neolithic influencing the European genetic heritage. PLoS One**.** 2015;10(6):e0128810

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. Nature. 1984;312(5991):282

Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S et al. A mitochondrial stratigraphy for island southeast Asia. Am J Hum Genet. 2007;80(1):29-43

Hirszfeld L, Hirszfeld H. Essai d'application des methods au probléme des racesp. L' anthropologie. 1919;29:505–537

Ho S. The molecular clock and estimating species divergence. Nat Educ. 2008;1(1):1-2

Ho SY, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. Time-dependent rates of molecular evolution. Mol Ecol. 2011;20(15):3087-3101

Ho SY, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol. 2014;23(24):5947-5965

Hodgson JA, Disotell TR. No evidence of a Neanderthal contribution to modern human diversity. Genome Biol. 2008;9(2):206

Hoffecker JF, Elias SA, O'rourke DH. Out of Beringia?. Science. 2014;343(6174):979-980

Hoffecker JF, Elias SA, O'Rourke DH, Scott GR, Bigelow NH. Beringia and the global dispersal of modern humans. Evol Anthropol. 2016;25(2):64-78

Hofmanová Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Díez-Del-Molino D, van Dorp L, López S, Kousathanas A, Link V et al. Early farmers from across Europe directly descended from Neolithic Aegeans. Proc Natl Acad Sci U S A. 2016;113(25):6886-6891

Hofmeijer GK, Alderliesten C, Van Der Borg K, Houston CM, de Jong AFM, Martini F, Sanges M, Sondaar PY, de Visser JA. Dating of the Upper Pleistocene lithic industry of Sardinia. Radiocarbon. 1989;31(3):986-991

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA. Nat Rev Genet. 2001;2(5):353

Horai S. Evolution and the origins of man: clues from complete sequences of hominoid mitochondrial DNA. Southeast Asian J Trop Med Public Health. 1995;26Supplement1:146-154

Howell N, Elson JL, Howell C, Turnbull DM. Relative rates of evolution in the coding and control regions of African mtDNAs. Mol Biol Evol. 2007;24(10):2213-2221

Hughes JF, Page DC. The biology and evolution of mammalian Y chromosomes. Annu Rev Genet. 2015;49:507-527

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000;408(6813):708

Ingman M, Gyllensten U. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J Hered. 2001;92(6):454-461

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-945

Irwin JA, Saunier JL, Niederstätter H, Strouss KM, Sturk KA, Diegoli TM, Brandstätter A, Parson W, Parsons TJ. Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J Mol Evol. 2009;68(5):516-527

Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet.. 2001;29(2):217

Jenkins DL, Davis LG, Stafford TW, Campos PF, Hockett B, Jones GT, Cummings LS, Yost C, Connolly TJ, Yohe RM et al. Clovis age Western Stemmed projectile points and human coprolites at the Paisley Caves. Science. 2012;337(6091):223-228

Jobling MA, Pandya A, Tyler-Smith C. The Y chromosome in forensic analysis and paternity testing. Int J Legal Med. 1997;110(3):118-124

Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003;4(8):598

Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29(13):1682-1684

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, Rootsi S, Ilumäe AM, Mägi R, Mitt M et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015;25(4):459-466

Kayser M. The human genetic history of Oceania: near and remote views of dispersal. Curr Biol. 2010;20(4):R194-R201

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol. 2006;6(1):29

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 2012;3:698

Kemp BM, Schurr TG. Ancient and modern genetic variation in the Americas. Human Variation in the Americas: The Integration of Archaeology and Biological Anthropology. Carbondale: Southern Illinois Univeristy. 2010;18598

Kierstein G, Vallinoto M, Silva A, Schneider MP, Iannuzzi L, Brenig B. Analysis of mitochondrial D-loop region casts new light on domestic water buffalo (*Bubalus bubalis*) phylogeny. Mol Phylogenet Evol. 2004;30(2):308-324

Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-626

Kingman JFC. The coalescent. Stoch Process Their Appl. 1982;13(3):235-248

Kistler L, Ware R, Smith O, Collins M, Allaby RG. A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res. 2017;45(11):6310-6320

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet. 2004;75(5):752-770

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A et al. The role of selection in the evolution of human mitochondrial genomes. Genetics. 2006;172(1):373-387

Kivisild T. Maternal ancestry and population history from whole mitochondrial genomes. Investig Genet. 2015;6(1):3

Knight RD, Freeland SJ, Landweber LF. Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet. 2001;2(1):49

Kong Q-P, Bandelt H-J, Sun C, Yao Y-G, Salas A, Achilli A, Wang C-Y, Zhong L, Zhu C-L, Wu S-F et al. Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet. 2006;15(13):2076-2086

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour. 2015;15(5):1179-1191

Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC bioinformatics. 2014;15(1):356

Kraemer MU, Sinka ME, Duda KA, Mylne A, Shearer FM, Brady OJ, Messina JP, Barker CM, Moore CG, Carvalho RG et al. The global compendium of *Aedes aegypt*i and *Ae. albopictus* occurrence. Sci Data. 2015;2:150035

Krueger F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. /. 2015

Kumar S, Bellis C, Zlojutro M, Melton PE, Blangero J, Curran JE. Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. BMC Evol Biol. 2011;11:293

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547-1549

Kumar S, Nagarajan M, Sandhu J, Kumar N, Behl V, Nishanth G. Mitochondrial DNA analyses of Indian water buffalo support a distinct genetic origin of river and swamp buffalo. Anim Genet. 2007a;38(3):227-232

Kumar S, Nagarajan M, Sandhu JS, Kumar N, Behl V. Phylogeography and domestication of Indian river buffalo. BMC Evol Biol. 2007b;7(1):1-8

Lancioni H, Di Lorenzo P, Cardinali I, Ceccobelli S, Capodiferro MR, Fichera A, Grugni V, Semino O, Ferretti L, Gruppetta A et al. Survey of uniparental genetic markers in the Maltese cattle breed reveals a significant founder effect but does not indicate local domestication. Anim Genet. 2016;47(2):267-269

Larson G, Burger J. A population genetics view of animal domestication. Trends Genet. 2013;29(4):197-205

Larson G, Fuller DQ. The evolution of animal domestication. Annu Rev Ecol Evol Syst. 2014;45:115-136

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, Barton L, Vigueira CC, Denham T, Dobney K et al. Current perspectives and the future of domestication studies. Proc Natl Acad Sci U S A. 2014;111(17):6139-6146

Lau CH, Drinkwater RD, Yusoff K, Tan SG, Hetzel DJS, Barker JSF. Genetic diversity of Asian water buffalo (*Bubalus bubalis*): mitochondrial DNA D-loop and cytochrome b sequence variation. Anim Genet. 1998;29(4):253-264

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513(7518):409

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536(7617):419-424

Lazaridis I. The evolutionary history of human populations in Europe. Curr Opin Genet Dev. 2018;53:21-27

Lei C, Zhang W, Chen H, Lu F, Ge Q, Liu R, Dang R, Yao Y, Yao L, Lu Z et al. Two maternal lineages revealed by mitochondrial DNA D-loop sequences in Chinese native water buffaloes (*Bubalus bubalis*). Asian-Australas J Anim Sci. 2007a;20(4):471

Lei CZ, Zhang W, Chen H, Lu F, Liu RY, Yang XY, Zhang HC, Liu ZG, Yao LB, Lu ZF et al. Independent maternal origin of Chinese swamp buffalo (*Bubalus bubalis*). Anim Genet. 2007b;38:97-102

Lewontin R. The interaction of selection and linkage. I. General considerations;heterotic models. Genetics. 1964;49(1):49

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-595

Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet. 2010;87(2):237-249

Lindgren G, Backstrom N, Swinburne J, Hellborg L, Einarsson A, Sandberg K, Cothran G, Vila C, Binns M, Ellegren H et al. Limited number of patrilines in horse domestication. Nat Genet.. 2004;36(4):335-336

Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schröder R, Stoneking M. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investig Genet. 2014;5(1):13

Liu RY, Yang GS, Lei CZ. The genetic diversity of mtDNA D-loop and the origin of Chinese goats. Acta Genetica Sin. 2006;33(5):420-428

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;193(4):1233-1254

Lothrop SK. Coclé: an archaeological study of central Panama. Peabody Museum of Archaeology and Ethnology. Harvard University. 1942

Lothrop SK. Metals from the cenote of sacrifice, Chichen Itza, Yucatan. The Museum. Cambridge. 1952

Luikart G, Giellly L, Excoffier L, Vigne JD, Bouuvet J, Taberlet P. Multiple maternal origins and weak phylogeographic structure in domestic goats. Proc Natl Acad Sci U S A. 2001;98

Luo S, Valencia CA, Zhang J, Lee NC, Slone J, Gui B, Wang X, Li Z, Dell S, Brown J et al. Biparental inheritance of mitochondrial DNA in Humans. Proc Natl Acad Sci U S A. 2018 201810946

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science. 2005;308(5724):1034-1036

MacGill T. A hand book, or guide, for strangers visiting Malta. Malta. L Tonna. 1839

MacHugh DE, Bradley DG. Livestock genetic origins: goats buck the trend. Proc Natl Acad Sci U S A. 2001;98(10):5382-5384

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE et al. A genomic history of Aboriginal Australia. Nature. 2016;538(7624):207

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201

Malyarchuk B, Grzybowski T, Derenko M, Perkova M, Vanecek T, Lazur J, Gomolcak P, Tsybovsky I. Mitochondrial DNA phylogeny in eastern and western Slavs. Mol Biol Evol. 2008;25(8):1651-1658

Malyarchuk B, Derenko M, Denisova G, Maksimov A, Wozniak M, Grzybowski T, Dambueva I, Zakharov I. Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a. J Hum Genet. 2011;56(8):583-588

Maretto F, Ramljak J, Sbarra F, Penasa M, Mantovani R, Ivanković A, Bittante G. Genetic relationships among Italian and Croatian Podolian cattle breeds assessed by microsatellite markers. Livest Sci. 2012;150(1-3):256-264

Martín JG, Cooke RG, Bustamante F, Holst I, Lara A, Redwood S. Ocupaciones prehispánicas en isla Pedro González, archipiélago de las Perlas, Panamá: aproximación a una cronología con comentarios sobre las conexiones externas. Lat Am Antiq. 2016;27(3):378-396

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):pp10-12

Mason JA, Johnson F. The native languages of Middle America. The Maya and their neighors. Cambridge: Harvard University Press. 1940;52-87

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303

Mellars P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. Proc Natl Acad Sci U S A. 2006;103(25):9381-9386

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MTP et al. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet. 2004;5(1):26

Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11(1):31

Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;(6):pdb.prot5448

Meyer M, Arsuaga J-L, de Filippo C, Nagel S, Aximu-Petri A, Nickel B, Martínez I, Gracia A, de Castro JMB, Carbonell E et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature. 2016;531(7595):504

Mignon-Grasteau S, Boissy A, Bouix J, Faure J-M, Fisher AD, Hinch GN, Jensen P, Le Neindre P, Mormède P, Prunet P et al. Genetics of adaptation and domestication in livestock. Livest Prod Sci. 2005;93(1):3-14

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A et al. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 2008;456(7220):387

Mishra BP, Dubey PK, Prakash B, Kathiravan P, Goyal S, Sadana DK, Das GC, Goswami RN, Bhasin V, Joshi BK et al. Genetic analysis of river, swamp and hybrid buffaloes of north-east India throw new light on phylogeography of water buffalo (*Bubalus bubalis*). J Anim Breed Genet. 2015;132(6):454-466

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas A-S, Sikora M et al. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. Nature. 2018a;553(7687):203

Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T et al. Early human dispersals within the Americas. Science. 2018b;aav2621

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Giménez J, Reis A, Varon-Mateeva R, Macek M, Kalaydjieva L et al. The origin of the major cystic fibrosis mutation (delta F508) in European populations. Nat Genet.. 1994;7(2):169-175

Mourant AE, Mad P. The distribution of the human blood groups. Oxford: Blackwell Scientific Publications. 1954

Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Recombinant DNA Methodology. Methods Enzymol. 1987;155:189-204

Naderi S, Rezaei HR, Taberlet P, Zundel S, Rafat SA, Naghash HR, el-Barody MA, Ertugrul O, Pompanon F, Econogene Consortium. Large-scale mitochondrial DNA analysis of the domestic goat reveals six haplogroups with high diversity. PLoS One**.** 2007;2(10):e1012

Navani N, Jain PK, Gupta S, Sisodia BS, Kumar S. A set of cattle microsatellite DNA markers for genome analysis of riverine buffalo (*Bubalus bubalis*). Anim Genet. 2002;33(2):149-154

Nei M. Molecular evolutionary genetics. New York: Columbia University Press. 1987

Neves WA, Pucciarelli HM. Morphological affinities of the first Americans: an exploratory analysis based on early South American human remains. J Hum Evol. 1991;21(4):261-273

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017;541(7637):302

Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM et al. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science. 2006;314(5806):1767-1770

Olivieri A, Pala M, Gandini F, Kashani BH, Perego UA, Woodward SR, Grugni V, Battaglia V, Semino O, Achilli A et al. Mitogenomes from two uncommon haplogroups mark late glacial/postglacial expansions from the near east and neolithic dispersals within Europe. PLoS One**.** 2013;8(7):e70492

Olivieri A, Gandini F, Achilli A, Fichera A, Rizzi E, Bonfiglio S, Battaglia V, Brandini S, De Gaetano A, El-Beltagi A et al. Mitogenomes from Egyptian cattle breeds: new clues on the origin of haplogroup Q and the early spread of *Bos taurus* from the Near East. PLoS One**.** 2015;10(10):e0141170

Olivieri A, Sidore C, Achilli A, Angius A, Posth C, Furtwängler A, Brandini S, Capodiferro MR, Gandini F, Zoledziewska M et al. Mitogenome diversity in Sardinians: a genetic window onto an island's past. Mol Biol Evol. 2017;34(5):1230-1239

Omrak A, Günther T, Valdiosera C, Svensson EM, Malmström H, Kiesewetter H, Aylward W, Storå J, Jakobsson M, Götherström A et al. Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. Curr Biol. 2016;26(2):270-275

Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 2015;16(7):395

O'Rourke DH, Raff JA. The human genetic history of the Americas: the final frontier. Curr Biol. 2010;20(4):R202-R207

Ottoni C, Van Neer W, De Cupere B, Daligault J, Guimaraes S, Peters J, Spassov N, Prendergast ME, Boivin N, Morales-Muñiz A et al. The palaeogenetics of cat dispersal in the ancient world. Nat Ecol Evol. 2017;1:139

Outram AK, Stear NA, Bendrey R, Olsen S, Kasparov A, Zaibert V, Thorpe N, Evershed RP. The earliest horse harnessing and milking. Science. 2009;323(5919):1332-1335

Owsley DW, Jantz RL. Kennewick Man: the scientific investigation of an ancient American skeleton. Texas A&M University Press. 2014

Pääbo S. Molecular cloning of ancient Egyptian mummy DNA. Nature. 1985;314(6012):644-645

Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M et al. Genetic analyses from ancient DNA. Annu Rev Genet. 2004;38:645-679

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016;538(7624):238

Pakendorf B, Stoneking M. Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet. 2005;6:165-183

Pala M, Achilli A, Olivieri A, Kashani BH, Perego UA, Sanna D, Metspalu E, Tambets K, Tamm E, Accetturo M et al. Mitochondrial haplogroup U5b3: a distant echo of the Epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet. 2009;84(6):814-821

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, Tamm E, Karmin M, Reisberg T, Hooshiar Kashani B et al. Mitochondrial DNA signals of late glacial recolonization of Europe from Near Eastern refugia. Am J Hum Genet. 2012;90(5):915-924

Palanichamy MG, Sun C, Agrawal S, Bandelt H-J, Kong Q-P, Khan F, Wang C-Y, Chaudhuri TK, Palla V, Zhang Y-P et al. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet. 2004;75(6):966-978

Parma P, Feligini M, Greeppi G, Enne G. The complete nucleotide sequence of goat (*Capra hircus*) mitochondrial genome. Goat mitochondrial genome. DNA Seq. 2003;14(3):199-203

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient admixture in human history. Genetics. 2012;192(3):1065-1093

Pauling L, Itano HA, Singer SJ, Wells IC. Sickle cell anemia, a molecular disease. Science. 1949;110(2865):543

Paupy C, Delatte H, Bagny L, Corbel V, Fontenille D. *Aedes albopictus*, an arbovirus vector: from the darkness to the light. Microbes Infect. 2009;11(14-15):1177-1185

Pellecchia M, Negrini R, Colli L, Patrini M, Milanesi E, Achilli A, Bertorelle G, Cavalli-Sforza LL, Piazza A, Torroni A et al. The mystery of Etruscan origins: novel clues from *Bos taurus* mitochondrial DNA. Proc Biol Sci. 2007;274(1614):1175-1179

Pennarun E, Kivisild T, Metspalu E, Metspalu M, Reisberg T, Moisan J-P, Behar DM, Jones SC, Villems R. Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. BMC Evol Biol. 2012;12(1):234

Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Hooshiar Kashani B, Ritchie KH, Scozzari R, Kong QP et al. Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol. 2009;19(1):1-8

Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Kashani BH, Carossa V, Ekins JE, Gómez-Carballa A, Huber G et al. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. Genome Res. 2010;20(9):1174-1179

Perego UA, Lancioni H, Tribaldos M, Angerhofer N, Ekins JE, Olivieri A, Woodward SR, Pascale JM, Cooke R, Motta J et al. Decrypting the mitochondrial gene pool of modern Panamanians. PLoS One**.** 2012;7(6):e38337

Pereira L, Richards M, Goios A, Alonso A, Albarrán C, Garcia O, Behar DM, Gölge M, Hatina J, Al-Gazali L et al. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res. 2005;15(1):19-24

Pereira V, Gusmão L. Types of genomes, sequences and genetic markers (repeats, SNPs, indels, haplotypes). Handbook of Forensic Genetics: Biodiversity And Heredity In Civil and Criminal Investigation. Word Scientific. 2017;163-191

Peter BM. Admixture, population structure and F-statistics. Genetics. 2016;202(4):1485-1501

Petraglia MD, Allchin B. The evolution and history of human populations in South Asia: Inter-disciplinary studies in archaeology, biological anthropology, linguistics and genetics. Springer. 2007

Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8(11):e1002967

Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends Genet. 2014;30(9):377-389

Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J. The genetic history of Europeans. Trends Genet. 2012;28(10):496-505

Pinhasi R, Fernandes D, Sirak K, Novak M, Connell S, Alpaslan-Roodenberg S, Gerritsen F, Moiseyev V, Gromov A, Raczky P et al. Optimal ancient DNA yields from the inner ear part of the human petrous bone. PLoS One**.** 2015;10(6):e0129102

Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, Valentin F, Thevenet C, Furtwängler A, Wißing C et al. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. Curr Biol. 2016;26(6):827-833

Posth C, Wißing C, Kitagawa K, Pagani L, van Holstein L, Racimo F, Wehrberger K, Conard NJ, Kind CJ, Bocherens H et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. Nat Commun. 2017;8:16046

Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E et al. Reconstructing the deep population history of Central and South America. Cell. 2018;175(5):1185-1197.e22

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet.. 2016;48(6):593-599

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'connor TD, Santpere G et al. Great ape genetic diversity and population history. Nature. 2013;499(7459):471

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet.. 2006;38(8):904-909

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945-959

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, De Filippo C et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43

Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Curr Biol. 2005;15(5):R159-R160

Przeworski M, Hudson RR, Di Rienzo A. Adjusting the focus on human variation. Trends Genet. 2000;16(7):296-302

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-575

Pyle A, Hudson G, Wilson IJ, Coxhead J, Smertenko T, Herbert M, Santibanez-Koref M, Chinnery PF. Extreme-depth re-sequencing of mitochondrial DNA finds no evidence of paternal transmission in humans. PLoS Genet. 2015;11(5):e1005040

Quilter J, Hoopes JW. Gold and power in ancient Costa Rica, Panama, and Colombia. Dumbarton Oaks Research Library and Collection. 2003

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 2015;349(6250):aab3884

Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, Grønnow B, Appelt M, Gulløv HC, Friesen TM et al. The genetic prehistory of the New World Arctic. Science. 2014a;345(6200):1255832

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford Jr TW, Orlando L, Metspalu E et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014b;505(7481):87

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005;102(44):15942-15947

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463(7282):757-762

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford Jr TW, Rasmussen S, Moltke I, Albrechtsen A, Doyle SM et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014;506(7487):225

Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, Zollikofer CPE, de León MP, Allentoft ME, Moltke I et al. The ancestry and affiliations of Kennewick Man. Nature. 2015;523(7561):455-458

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468(7327):1053

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N et al. Reconstructing Native American population history. Nature. 2012;488(7411):370-374

Relethford JH. Genetic evidence and the modern human origins debate. Heredity. 2008;100(6):555

Rojas EI. Pueblos que capturan: esclavitud indígena al sur de América Central del siglo XVI al XIX. /. 2012

Sardina MT, Ballester M, Marmi J, Finocchiaro R, van Kaam JB, Portolano B, Folch JM. Phylogenetic analysis of Sicilian goats reveals a new mtDNA lineage. Anim Genet. 2006;37(4):376-378

Sato M, Sato K. Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. Biochim Biophys Acta Mol Cell Res. 2013;1833(8):1979-1984

Scheib C, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Mörseburg A, Johnson JR, Potter A et al. Ancient human parallel lineages within North America contributed to a coastal expansion. Science. 2018;360(6392):1024-1027

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MG et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012;1227721

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, Vicente M, Steyn M, Soodyall H et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 2017;358(6363):652-655

Schon EA, Bonilla E, DiMauro S. Mitochondrial DNA mutations and pathogenesis. J Bioenerg Biomembr. 1997;29(2):131-149

Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history. Nat Rev Genet. 2015;16(12):727

Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. Trends Genet. 2018;34(4):301-312

Schroeder H, Sikora M, Gopalakrishnan S, Cassidy LM, Delser PM, Velasco MS, Schraiber JG, Rasmussen S, Homburger JR, Ávila-Arcos MC et al. Origins and genetic legacies of the Caribbean Taino. Proc Natl Acad Sci U S A. 2018;115(10):2341-2346

Schroeder K, Schurr TG, Long J, Rosenberg N, Crawford M, Tarskaia L, Osipova L, Zhadanov S, Smith DG. A private allele ubiquitous in the Americas. **Biol Lett.** 2007;3(2):218-223

Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat Protoc. 2014;9(5):1056-1082

Schuenemann VJ, Bos K, DeWitte S, Schmedes S, Jamieson J, Mittnik A, Forrest S, Coombes BK, Wood JW, Earn DJ et al. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. Proc Natl Acad Sci U S A. 2011;108(38):E746-E752

Schwarz C, Debruyne R, Kuch M, McNally E, Schwarcz H, Aubrey AD, Bada J, Poinar H. New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. Nucleic Acids Res. 2009;37(10):3215-3229

Shackleton JC, van Andel TH, Runnels CN. Coastal Paleogeography of the Central and Western Mediterranean during the Last 125,000 Years and Its Archaeological Implications. J Field Archaeol. 1984;11(3):307-314

Sharma A, Jaloree S, Thakur RS. Review of clustering methods: toward phylogenetic tree constructions. Proceedings of International Conference on Recent Advancement on Computer and Communication. Singapore: Springer. 2018;475-480

Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnol. 2008;26(10):1135

Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet.. 2015;47(11):1272

Sikora M, Carpenter ML, Moreno-Estrada A, Henn BM, Underhill PA, Sánchez-Quinto F, Zara I, Pitzalis M, Sidore C, Busonero F et al. Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. PLoS Genet. 2014;10(5):e1004353

Sikora M, Pitulko V, Sousa V, Allentoft ME, Vinner L, Rasmussen S, Margaryan A, de Barros Damgaard P, de la Fuente Castro C, Renaud G et al. The population history of northeastern Siberia since the Pleistocene. bioRxiv. 2018;448829

Skoglund P, Storå J, Götherström A, Jakobsson M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. J Archaeol Sci. 2013;40(12):4477-4482

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D. Genetic evidence for two founding populations of the Americas. Nature. 2015;525(7567):104-108

Skoglund P, Reich D. A genomic view of the peopling of the Americas. Curr Opin Genet Dev. 2016;41:27-35

Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A et al. Reconstructing prehistoric African population structure. Cell. 2017;171(1):59-71.e21

Skoglund P, Mathieson I. Ancient Human genomics: the first decade. Annu Rev Genomics Hum Genet. 2018;19:381-404

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009;84(6):740-759

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H-J, Torroni A, Richards MB. The archaeogenetics of Europe. Curr Biol. 2010;20(4):R174-R183

Soares PM, Cardoso RM, Miranda PM, Viterbo P, Belo-Pereira M. Assessment of the ENSEMBLES regional climate models in the representation of precipitation variability and extremes over Portugal. J Geophys Res Atmos. 2012;117(D7)

Sobenin IA, Sazonova MA, Postnov AY, Bobryshev YV, Orekhov AN. Changes of mitochondria in atherosclerosis: possible determinant in the pathogenesis of the disease. Atherosclerosis. 2013;227(2):283-288

Sondaar P, Dermitzakis M, Drinia H, de Vos J. Paleoecological factors that controlled the survival and adaptation of the Pleistocene man on the Mediterranean islands. Ann Geol des Pays Hellenique. 1998;25-35

Sosa MX, Sivakumar IA, Maragh S, Veeramachaneni V, Hariharan R, Parulekar M, Fredrikson KM, Harkins TT, Lin J, Feldman AB et al. Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol. 2012;8(10):e1002737

Stanford DJ, Bradley BA. Across Atlantic ice: the origin of America's Clovis culture. Univeristy of California Press. 2012

Stoneking M, Delfin F. The human genetic history of East Asia: weaving a complex tapestry. Curr Biol. 2010;20(4):R188-R193

Strauss A, Oliveira RE, Villagran XS, Bernardo DV, Salazar-García DC, Bissaro MC, Pugliese F, Hermenegildo T, Santos R, Barioni A et al. Early Holocene ritual complexity in South America: the archaeological record of Lapa do Santo (east-central Brazil). Antiquity. 2016;90(354):1454-1473

Stringer C. Chronological and biogeographic perspectives on later human evolution. Neandertals and Modern Humans in Western Asia. Springer. 2002;29-37

Stringer CB, Andrews P. Genetic and fossil evidence for the origin of modern humans. Neandertals and modern humans in western Asia. Science. 1988;239(4845):1263-1268

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4(1):vey016

Sultana S, Mannen H, Tsuji S. Mitochondrial DNA diversity of Pakistani goats. Anim Genet. 2003;34(6):417-421

Sun C, Kong Q-P, Palanichamy Mg, Agrawal S, Bandelt H-j, Yao Y-G, Khan F, Zhu C-L, Chaudhuri TK, Zhang Y-p et al. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. Mol Biol Evol. 2005;23(3):683-690

Sunnucks P. Efficient genetic markers for population biology. **Trends** Ecol Evol**.** 2000;15(5):199-203

Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437-460

Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogväli E-L, Tolk H-V, Reidla M, Metspalu E, Pliss L et al. The western and eastern roots of the Saami—the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. Am J Hum Genet. 2004;74(4):661-682

Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK et al. Beringian standstill and spread of Native American founders. PLoS One**.** 2007;2(9):e829

Tanaka K, Yamagata T, Masangkay JS, Faruque MO, Vu-Binh D, Salundik, Mansjoer SS, Kawamoto Y, Namikawa T. Nucleotide diversity of mitochondrial DNAs between the swamp and the river types of domestic water buffaloes, *Bubalus bubalis*, based on restriction endonuclease cleavage patterns. Biochem Genet. 1995;33(5):137-148

Tanaka K, Solis CD, Masangkay JS, Maeda K-i, Kawamoto Y, Namikawa T. Phylogenetic relationship among all living species of the genus *Bubalus* based on DNA sequences of the Cytochrome b gene. Biochem Genet. 1996;34(11):443-452

Tao M, You C-P, Zhao R-R, Liu S-J, Zhang Z-H, Zhang C, Liu Y. Animal mitochondria: evolution, function, and disease. Curr Mol Med. 2014;14(1):115-124

Tattersall I. Human origins: out of Africa. Proc Natl Acad Sci U S A. 2009;106(38):16018-16021

Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet. 2005;6(5):389

Templeton AR. Genetics and recent human evolution. Evolution. 2007;61(7):1507-1519

Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet.. 2017;49(2):303-309

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, Wallace DC. Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet. 1993;53(3):563-590

Torroni A, Lott MT, Cabell MF, Chen Y-S, Lavergne L, Wallace DC. mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am J Hum Genet. 1994;55(4):760

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus M-L, Wallace DC. Classification of European mtDNAs from an analysis of three European populations. Genetics. 1996;144(4):1835-1850

Torroni A, Bandelt H-J, D'urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus M-L, Bonné-Tamir B et al. mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet. 1998;62(5):1137-1152

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M et al. Do the four clades of the mtDNA haplogroup L2 evolve at different rates?. Am J Hum Genet. 2001;69(6):1348-1356

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. Harvesting the fruit of the human mtDNA tree. Trends Genet. 2006;22(6):339-345

Tuross N, Campana MG. Ancient DNA. The Science of Roman History: Biology, Climate, and the Future of the Past. Princeton University Press. 2018

Umaña AC. Las lenguas del área intermedia: Introducción a su estudio areal. Editorial Universidad de Costa Rica. 1991

Underhill PA, Kivisild T. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 2007;41:539-564

van de Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, Nickel B, Nagel S, Talbi EH, El Hajraoui MA et al. Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. Science. 2018;360(6388):548-552

van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30(2):E386-94

Veeramah KR, Wegmann D, Woerner A, Mendez FL, Watkins JC, Destro-Bisol G, Soodyall H, Louie L, Hammer MF. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol Biol Evol. 2011;29(2):617-630

Veerappa AM, Padakannaya P, Ramachandra NB. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. Funct Integr Genomics. 2013;13(3):285-293

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science. 2016;aad9416

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. African populations and the evolution of human mitochondrial DNA. Science. 1991;253(5027):1503-1507

Vila C, Leonard JA, Gotherstrom A, Marklund S, Sandberg K, Liden K, Wayne RK, Ellegren H. Widespread origins of domestic horse lineages. Science. 2001;291(5503):474-477

Wallace DC. Mitochondrial DNA sequence variation in human evolution and disease. Proc Natl Acad Sci U S A. 1994;91(19):8739-8746

Wallace DC, Brown MD, Melov S, Graham B, Lott M. Mitochondrial biology, degenerative diseases and aging. Biofactors. 1998;7(3):187-190

Wallace DC, Chalkia D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harb Perspect Biol. 2013;5(11):a021220

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C et al. Genetic variation and population structure in native Americans. PLoS Genet. 2007;3(11):e185

Wang S, Chen N, Capodiferro MR, Zhang T, Lancioni H, Zhang H, Miao Y, Chanthakhoun V, Wanapat M, Yindee M et al. Whole mitogenomes reveal the history of swamp buffalo: initially shaped by glacial periods and eventually modelled by domestication. Sci Rep. 2017;7(1):4708

Wangkumhang P, Hellenthal G. Statistical methods for detecting admixture. Curr Opin Genet Dev. 2018;53:121-127

Weidenreich F. Some problems dealing with ancient man. Am Anthropol. 1940;42(3):375-383

Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R Soc Open Sci. 2016;3(6)

Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schönherr S. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016;44(W1):W58-63

Wilkins JF. Unraveling male and female histories from human genetic data. Curr Opin Genet Dev. 2006;16(6):611-617

Woodruff DS. Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today’s patterns and the future of the remaining refugial-phase biodiversity. Biodivers Conserv. 2010;19(4):919-941

Wright S. Evolution in mendelian populations. Genetics. 1931;16(2):97

Xu B, Yang Z. PAMLX: a graphical user interface for PAML. Mol Biol Evol. 2013;30(12):2723-2724

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science. 2015;348(6231):242-245

Yan X-H, Ho C-R, Zheng Q, Klemas V. Temperature and size variabilities of the Western Pacific Warm Pool. Science. 1992;258(5088):1643

Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, Slatkin M, Meyer M, Pääbo S, Kelso J et al. 40,000-Year-Old Individual from Asia provides insight into early population structure in Eurasia. Curr Biol. 2017;27(20):3202-3208.e9

Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997;13(5):555-556

Yindee M, Vlamings BH, Wajjwalku W, Techakumphu M, Lohachit C, Sirivaidyapong S, Thitaram C, Amarasinghe AA, Alexander PA, Colenbrander B et al. Y-chromosomal variation confirms independent domestications of swamp and river buffalo. Anim Genet. 2010;41(4):433--435

Yue XP, Li R, Xie WM, Xu P, Chang TC, Liu L, Cheng F, Zhang RF, Lan XY, Chen H et al. Phylogeography and domestication of Chinese swamp buffalo. PLoS One**.** 2013;8(2):e56552

Zeder MA, Emshwiller E, Smith BD, Bradley DG. Documenting domestication: the intersection of genetics and archaeology. Trends Genet. 2006;22(3):139-155

Zeder MA. Pathways to animal domestication. Biodiversity in Agriculture: Domestication, Evolution, and Sustainability. New York: Cambridge University Press. 2012;227-259

Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. Mol Biol Evol. 2004;21(1):164-175

Zhang D-X, Hewitt GM. Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. Biochem Syst Ecol. 1997;25(2):99-120

Zhang H, Xing D, Wang G, Li C, Zhao T. Sequencing and analysis of the complete mitochondrial genome of *Aedes albopictus (Diptera Culicidae*) in China. Mitochondrial DNA A DNA Mapp Seq Anal. 2016a;PartA27(4):2787-2788

Zhang Y, Lu Y, Yindee M, Li KY, Kuo HY, Ju YT, Ye S, Faruque MO, Li Q, Wang Y et al. Strong and stable geographic differentiation of swamp buffalo maternal and paternal lineages indicates domestication in the China/Indochina border region. Mol Ecol. 2016b;25(7):1530-1550

Zhou Q, Li H, Li H, Nakagawa A, Lin JL, Lee E-S, Harry BL, Skeen-Gaar RR, Suehiro Y, William D et al. Mitochondrial endonuclease G mediates breakdown of paternal mitochondria upon fertilization. Science. 2016;353(6297):394-399

List of original publications

1. P. Di Lorenzo, H. Lancioni, S. Ceccobelli, L. Colli, I. Cardinali, T. Karsli, **M.R. Capodiferro**, E. Sahin, L. Ferretti, P. Ajmone Marsan, F.M. Sarti, E. Lasagna, F. Panella, A. Achilli. Mitochondrial DNA variants of Podolian cattle breeds testify for a dual maternal origin. PLoS One. 2018 Feb 20;13(2):e0192567.
2. S. Wang*, N. Chen*, **M.R. Capodiferro***, T. Zhang, H. Lancioni, H. Zhang, Y. Miao, V. Chanthakhoun, M. Wanapat, M. Yindee, Y. Zhang, H. Lu, L. Caporali, R. Dang, Y. Huang, X. Lan, M. Plath, H. Chen, J. A. Lenstra, A. Achilli, C. Lei. Whole Mitogenomes Reveal the History of Swamp Buffalo: Initially Shaped by Glacial Periods and Eventually Modelled by Domestication. Sci Rep. 2017 Jul 5;7(1):4708.
3. A. Olivieri, C. Sidore, A. Achilli, A. Angius, C. Posth, A. Furtwängler, S. Brandini, **M.R. Capodiferro**, F. Gandini, M. Zoledziewska, M. Pitzalis, A. Maschio, F. Busonero, L. Lai, R. Skeates, M.G. Gradoli, J. Beckett, M. Marongiu, V. Mazzarello, P. Marongiu, S. Rubino, T. Rito, V. Macaulay, O. Semino, M. Pala, G. Abecasis, D. Schlessinger, P. Soares, M.B. Richards, F. Cucca, A. Torroni. Mitogenome Diversity in Sardinians: a Genetic Window onto an Island's Past. Mol Biol Evol. 2017 May 1;34(5):1230-1239.
4. V. Battaglia, P. Gabrieli, S. Brandini, **M. R. Capodiferro**, P. J. Javier, X. G. Chen, A. Achilli, O. Semino, L. M. Gomulsky, A. R. Malacrida, G. Gasperi, A. Torroni A. Olivieri. The worldwide spread of the tiger mosquito as revealed by mitogenome haplogroup diversity. Front Genet. 2016 Nov 23, 7:208.
5. I. Cardinali, H. Lancioni, A. Giontella, **M.R. Capodiferro**, S. Capomaccio, L. Buttazzoni, G.P. Biggio, R. Cherchi, E. Albertini, A. Olivieri, K. Cappelli, A. Achilli, M. Silvestrelli. An overview of ten Italian horse breeds through mitochondrial DNA. Plos One. 2016 Apr 7;11(4):e0153004.
6. H. Lancioni, P. Di Lorenzo, I. Cardinali, S. Ceccobelli, **M.R. Capodiferro**, A. Fichera, V. Grugni, O. Semino, L. Ferretti, A. Gruppetta, G. Attard, A. Achilli, E. Lasagna. Survey of uniparental genetic markers in the Maltese cattle breed reveals a significant founder effect but does not indicate local domestication. Anim Genet. 2016 Apr;47(2):267-9.
7. L. Colli, H. Lancioni, I Cardinali, A. Olivieri, **M.R. Capodiferro**, M. Pellecchia, M. Rzepus, W. Zamani, S. Naderi, F. Gandini, S.M. Vahidi, S. Agha, E. Randi, V. Battaglia, M.T. Sardina, B. Portolano, H.R. Rezaei, P. Lymberakis, F. Boyer, E. Coic, F. Pompanon, P. Taberlet, P. Ajmone Marsan, A. Achilli. Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. BMC Genomics. 2015 Dec 29;16(1):111

Colli *et al. BMC Genomics* (2015) 16:1115
DOI 10.1186/s12864-015-2342-2

BMC Genomics

RESEARCH ARTICLE Open Access

Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability

Licia Colli[1,2†], Hovirag Lancioni[3†], Irene Cardinali[3], Anna Olivieri[4], Marco Rosario Capodiferro[3,4], Marco Pellecchia[1], Marcin Rzepus[1,5], Wahid Zamani[6,7], Saeid Naderi[8], Francesca Gandini[4,9], Seyed Mohammad Farhad Vahidi[10], Saif Agha[11], Ettore Randi[12,13], Vincenza Battaglia[4], Maria Teresa Sardina[14], Baldassare Portolano[14], Hamid Reza Rezaei[15], Petros Lymberakis[16], Frédéric Boyer[6], Eric Coissac[6], François Pompanon[6], Pierre Taberlet[6], Paolo Ajmone Marsan[1,2] and Alessandro Achilli[3,4*]

Abstract

Background: The current extensive use of the domestic goat (*Capra hircus*) is the result of its medium size and high adaptability as multiple breeds. The extent to which its genetic variability was influenced by early domestication practices is largely unknown. A common standard by which to analyze maternally-inherited variability of livestock species is through complete sequencing of the entire mitogenome (mitochondrial DNA, mtDNA).

Results: We present the first extensive survey of goat mitogenomic variability based on 84 complete sequences selected from an initial collection of 758 samples that represent 60 different breeds of *C. hircus*, as well as its wild sister species, bezoar (*Capra aegagrus*) from Iran. Our phylogenetic analyses dated the most recent common ancestor of *C. hircus* to ~460,000 years (ka) ago and identified five distinctive domestic haplogroups (A, B1, C1a, D1 and G). More than 90 % of goats examined were in haplogroup A. These domestic lineages are predominantly nested within *C. aegagrus* branches, diverged concomitantly at the interface between the Epipaleolithic and early Neolithic periods, and underwent a dramatic expansion starting from ~12–10 ka ago.

Conclusions: Domestic goat mitogenomes descended from a small number of founding haplotypes that underwent domestication after surviving the last glacial maximum in the Near Eastern refuges. All modern haplotypes A probably descended from a single (or at most a few closely related) female *C. aegagrus*. Zooarchaelogical data indicate that domestication first occurred in Southeastern Anatolia. Goats accompanying the first Neolithic migration waves into the Mediterranean were already characterized by two ancestral A and C variants. The ancient separation of the C branch (~130 ka ago) suggests a genetically distinct population that could have been involved in a second event of domestication. The novel diagnostic mutational motifs defined here, which distinguish wild and domestic haplogroups, could be used to understand phylogenetic relationships among modern breeds and ancient remains and to evaluate whether selection differentially affected mitochondrial genome variants during the development of economically important breeds.

Keywords: Goat mitochondrial genome, mtDNA haplogroups, Domestication, Origin of *Capra hircus*, *Capra aegagrus*

* Correspondence: alessandro.achilli@unipv.it
†Equal contributors
[3]Dipartimento di Chimica, Biologia e Biotecnologie, Università di Perugia, Perugia 06123, Italy
[4]Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università di Pavia, Pavia 27100, Italy
Full list of author information is available at the end of the article

Background

The domestic goat (*Capra hircus*), which counts a worldwide population of more than 800,000,000 specimens and about 1200 breeds described (http://dad.fao.org), is among the "big five" livestock species defined by the Food and Agriculture Organization (FAO) [1] and is an invaluable source of milk, meat, skin and fiber for poor small holders and shepherds in developing countries and marginal areas [2].

On the basis of bone morphological changes associated with progressive taming [3], zooarchaeology suggests that goat domestication began about 11,000 years (ka) ago in an area stretching from the high Euphrates valleys in Southeastern Anatolia (Turkey) to the Zagros Mountains in Central Iran [4, 5] and located within the natural distribution of the wild ancestor species, the bezoar *Capra aegagrus* [6, 7]. This event is now considered as the final outcome of a gradual change from hunting to management of wild captive animals [3, 5, 8]. Due to their high rusticity and adaptability to harsh environments, goats represented a key resource during the Neolithic agricultural revolution and for the human migration waves that spread Neolithic culture out of the Fertile Crescent [9].

Previous studies of mitochondrial control-region haplotypes described six highly divergent haplogroups (Hg.s) in domestic goats: A, B, C [10], D [11], F [12] and G [13]. An additional haplogroup, named E, has been described by Joshi et al. [9] on the basis of two highly divergent control-region haplotypes. This was recognized as a sub-clade of haplogroup A when compared to a larger dataset [13]. The reported weak geographical structuring of goat mitochondrial variability was often interpreted as a consequence of the frequent transportation of goats along terrestrial and maritime routes of migration and commerce, probably during the early domestication phases [7, 10, 14, 15]. A subsequent comparison with wild stocks confirmed the presence of all "domestic" haplogroups in the current *C. aegagrus* populations, which could result from early translocations of animals and/or feralization before the worldwide spread of goats [7]. The haplogroup A is largely predominant (>90 %) among domestic goats, but rare (6 %) in the bezoar and never observed in the Iranian Zagros Mountains. The probable origin of haplogroup A occurred in Eastern Anatolia, where it is still present among wild populations, and its presence in Eastern Iran probably is the result of a subsequent feralization of domestic goats. The most frequent haplogroup in wild populations is C (39 %) detected in most of the bezoar distribution area and more common in Southern Zagros/Central Iranian Plateau. The evidence that C control-region haplotypes from Pakistan are the farthest from the domestic-related ones [7] disproved the Indus Valley domestication hypothesis suggested by archaeological remains from Mehrgarh (Baluchistan, Pakistan) [5]. Haplogroup F is still found in wild populations (from Northern Caucasus to lower Indus Valley), but it is very rare in domestic goats (<0.2 %), as it was identified only in three Sicilian samples [12]. The other haplogroups were found only in Iranian (D and G) or in both Northern Iranian and Eastern Anatolian bezoars (B). It has been proposed that these haplotypes might have entered the domestic goat gene pool either during the early spread of domestic goats, or due to small-scale domestication events. These findings indicate that the process of goat domestication occurred not only in Eastern Anatolia, as marked by haplogroup A and supported by zooarchaeological data [5, 8], but possibly also in Central Iran (Zagros Mountains and Iranian Plateau). This additional easternmost domestication event has been marked by haplogroup C, although it led only to a small contribution detectable in the mitochondrial gene pool of current domestic goats (1.5 %) and no archaeological substantiations [7].

MtDNA haplotypes belonging to haplogroups A and C have been found in ancient goat samples retrieved from an early Neolithic site in Southern France [16]. The two haplogroups occurred with almost the same frequency among the analyzed bones (i.e. 8 samples carrying A haplotypes and 11 samples carrying C haplotypes), suggesting that domestic goat populations were already characterized by the mtDNA variants A and C [16] during the first colonization waves that brought Neolithic farmers into the Mediterranean area about 7.5 ka ago [3].

Haplogroup divergence times calculated on different control-region datasets, usually by employing different calibration points, span from 100 to 940 ka [9–11, 17], thus largely predating the domestication events (~11 ka). These data suggest that many sub-haplogroups were already present among the bezoar populations and, therefore, that many lineages were included in the domesticated stocks. Previous studies on livestock species showed that phylogenies based on short mitochondrial sequences can be heavily affected by the confounding effects of homoplasies [18–21] and mitochondrial pseudogenes [22–24], which blur the real extent of lineage divergences/similarities. The available goat sequencing data are usually restricted to a few hundreds of control-region base pairs (bps), spanning from np 15431 to np 16643. Moreover, several complete goat mitochondrial genomes deposited in GenBank are probably chimeric or affected by NuMtS (i.e. nuclear sequences of mitochondrial origin) [22], including the previously adopted mtDNA goat reference sequence NC_005044 [22, 25]. In order to overcome these drawbacks, two recent papers have already extended the analysis to the mtDNA coding genes, but either failed to explore the complete mitochondrial molecule [17] or focused only on a new mtDNA haplotype [26].

We present the first extensive survey of the entire mitogenome variability based on 81 novel complete sequences from domestic goats ($n = 76$) and wild relatives ($n = 5$). This analysis allowed us to accurately define those (sub-) haplogroups involved in the domestication process(es), and to provide haplogroup coalescence estimates falling at the interface between Epipaleolithic and early Neolithic periods and expansion times falling into the Neolithic.

Results

The phylogeny of goat mitochondrial genomes

An initial collection of 758 mtDNA samples from *C. aegagrus* ($n = 19$) and *C. hircus* ($n = 739$; mostly from Western Eurasian breeds) was preliminarily characterized through control-region sequencing (Additional file 1: Table S1). This dataset, including 70 previously published samples [7], was used to build a haplotype network (data not shown). Overall, the network evidenced a high number of homoplasies and crosslinks, but it was useful to select 81 samples (from 76 domestic goats and five bezoars) for complete sequencing (Additional file 1: Table S2) using the criterion of including the widest possible range of mtDNA variation. The selected mitogenomes belonged to all known haplogroups, with few notable exceptions: i) the very rare haplogroup F, identified so far only in three domestic goats from Sicily [12], was not represented within our initial dataset of domestic goats; ii) the single control-region haplotype A in our *C. aegagrus* samples was re-classified as C by preliminary sequencing of some informative coding-region segments, which encompass diagnostic single nucleotide polymorphism (SNP) markers of A (at nps 3194/7839) and C (at nps 2885/3002/3131/3293/7657); iii) lastly, the few G control-region haplotypes were obtained from degraded DNA molecules, which allowed only partial coding-region sequencing that most likely confirmed the G affiliation (Additional file 1: Table S3). The 81 novel mitogenomes were compared with the revised goat reference sequence (GRS; NC_005044.2 – Additional file 1: Table S4). Various measures of molecular diversity were evaluated on the final dataset of 84 mitogenomes, which included 83 different haplotypes. After excluding ambiguous sites and indels (insertions/deletions), we identified a total number of 1003 variant sites (Table 1): 774 in the coding region (15414 nucleotides) and 229 in the D-loop (1213 nucleotides). Overall we observed an average number of 60.470 ± 16.628 nucleotide differences between two randomly chosen sequences. Figure S1 illustrates the distributions of nucleotide diversity (π) and total number of substitutions (continuous and dotted lines, respectively) along the mitogenome (Additional file 1: Figure S1). As expected, the highest diversity was observed around the HVS-I segment (hypervariable segment-I, from np 15,707 to np 16,187), with a peak of $\pi = 0.059$. The latter value is higher than those previously reported in horses [21]. Within the coding region, the highest number of variant sites was found in protein-coding genes ($n = 655$), mostly synonymous mutations. These data are consistent with previous studies on human and horse mitochondrial genomes [21, 27–29].

Table 1 Distribution and recurrence of mutations in the 84 goat mtDNA sequences

	CONTROL	CODING		
	D-loop	rRNA	tRNA	mRNA
Length in base pairs[a]	1213	2529	1513	11370
Invariable sites	984	2456	1467	10715
No. of variable sites	229	73	46	655
Proportion of variable sites	0.233	0.030	0.031	0.061
Sites with a single hit	96	50	36	470
Sites with two hits	25	10	4	51
Sites with three or more hits	108	13	6	134
Transitions	221	70	45	627
Transversions	12	3	1	28
Transition/Transversion ratio	18.4	23.3	45.0	22.4

[a] Lengths of protein-coding genes were readjusted and extended by considering the overlapping segments. Regarding the tRNA loci, the overlapping portions were counted only once

A parsimony approach was applied to infer evolutionary relationships from the final dataset of 84 complete mitogenomes. Eventually, only coding-region substitutions were included in the tree (Additional file 1: Figure S2) because of the extraordinary control-region variability (mainly around HVS-I, see Additional file 1: Figure S1) and high indels' instability. The obtained topology confirmed all previously known control-region branches (A-G), but also revealed many different sub-branches, particularly within the major branch A. Seven novel sub-branches, named A1 to A7, were identified (at least three different haplotypes were required here to nominate a new clade, with the only notable exception of D1), all marked by coding-region transitions. In order to convert mutational distances into time over the entire mitogenome, the goat mtDNA sequences were compared with a published complete mitogenome (KF302445) from a Comisana sheep (*Ovis aries*) [30], used as an outgroup. Diverse maximum likelihood (ML) and Bayesian analyses were employed, as described in Materials and Methods. Initially, an ML tree based on synonymous mutations alone was estimated (Fig. 1). In fact, contrary to the number of potential factors causing time-dependent rates [31], synonymous mutations are virtually neutral and not subject to the effect of purifying selection, even though saturation might still be an issue with long time frames. Using CODEML and the

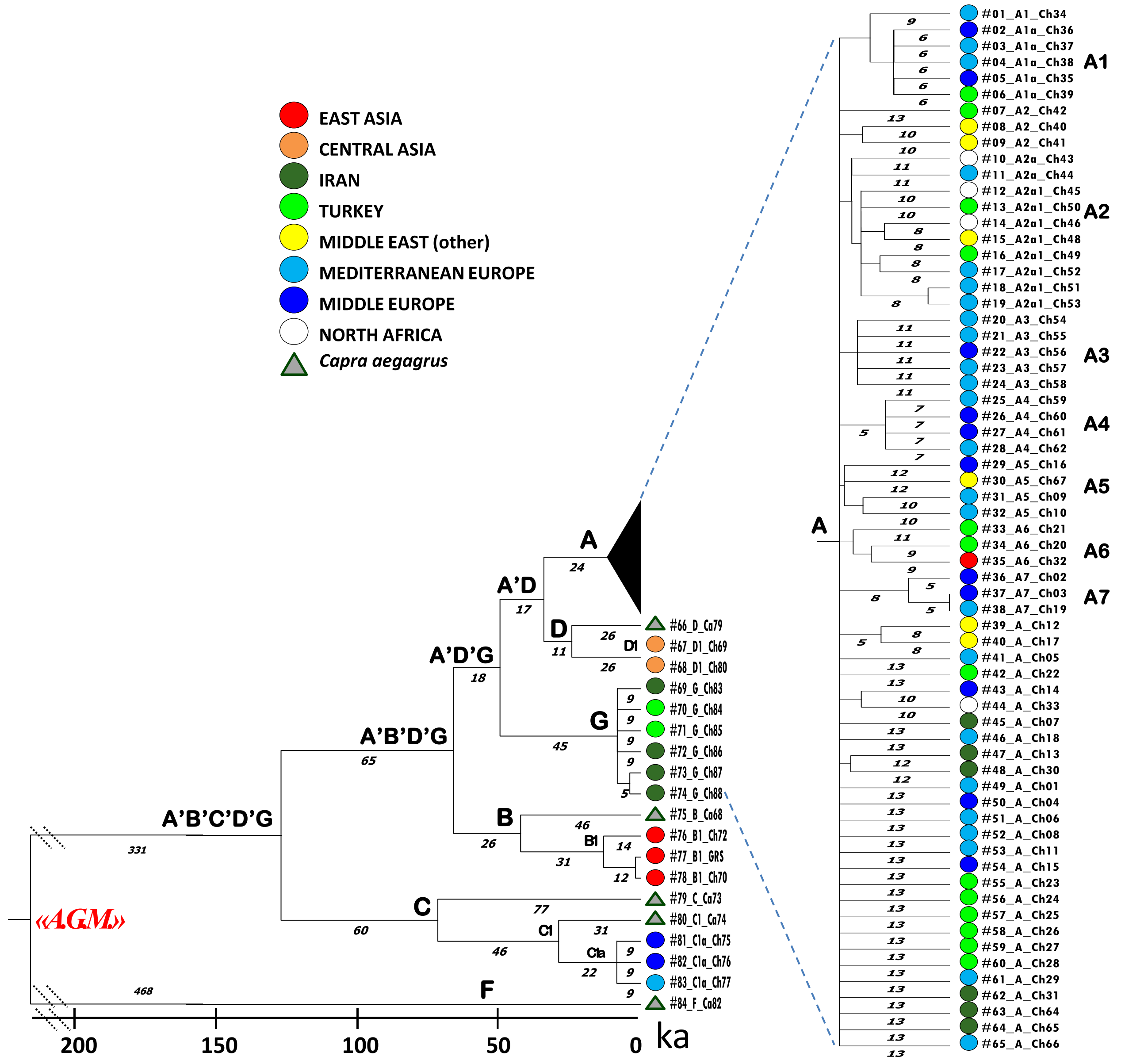


Fig. 1 Schematic phylogeny of complete mtDNAs from modern domestic and wild goats. The tree encompasses 84 sequences and was rooted by using a published sheep (*O. aries*) sequence (not displayed). "A.G.M." indicates the reconstructed Ancestral Goat Mitogenome. The topology was inferred by a maximum parsimony approach, while maximum likelihood (ML) time divergences based on synonymous substitutions are shown below the branches. The right inset shows the complete A branch rooted in D. Further details are given in Additional file 1: Table S4 and Figure S2

mammalian mtDNA genetic code, we calibrated the synonymous molecular clock at 7.77 × 10^{-8} substitutions per year (at 3790 codons) or 1 substitution every 3397 years, after verifying the clock model hypothesis (*p-value* = 0.214). This mutation rate was also used to convert into time the rho estimates based on synonymous mutations (Table 2). The deepest node corresponds to the single Ancestral Goat Mitogenome (AGM), from which all modern goat mtDNA sequences derive, and was dated ~460 ka. The F bezoar mitogenome radiates first in the tree, then a major split (~130 ka ago) separates haplogroup C (dated ~80 ka) from the remaining mtDNA haplotypes. A subsequent branching separates two sister clades (B and A'D'G) both dated about 50 ka. Haplogroup B encompasses four samples, including GRS and one bezoar. The remaining wild samples belong to haplogroup D together with two domestic goats from Kyrgyzstan (with the same haplotype); affiliation to haplogroup D was also confirmed by considering recently published partial coding sequences [17]. The other bezoar mitogenomes, within haplogroups B and C, are ancestral to their most closely related domestic clusters B1 and C1a, dated 14.2 and 9.2 ka, respectively. These estimates are very

Table 2 Age estimates based on different datasets

	ML (synonymous sub.s)		Rho (synonymous sub.s)		ML (all substitutions)[a]		ML (only coding region)		Rho (only coding region)	
Node	T(ka)	ΔT(ka)	T(ka)	ΔT(ka)	T(ka)	ΔT(ka)	T(ka)	ΔT(ka)	T(ka)	ΔT(ka)
A.G.M.[b]	467.7	59.2	457.9	37.3	652.0	34.0	734.9	61.1	715.5	51.2
A'B'C'D'G	136.7	21.3	125.7	17.4	186.6	12.2	214.3	24.0	187.4	23.0
A'B'D'G	71.2	12.2	57.3	10.6	115.9	8.3	124.8	15.6	95.6	14.9
A'D'G	53.5	9.8	42.5	8.8	89.3	6.8	93.1	12.8	68.5	12.0
A'D	36.9	8.1	25.4	5.8	66.0	5.9	63.6	11.3	42.0	8.1
A	12.8	1.9	14.1	1.4	23.4	1.4	23.0	2.2	23.7	1.9
A1	9.2	2.8	10.2	3.5	20.9	1.9	18.6	3.5	19.2	4.9
A2	12.8	2.9	16.2	3.9	19.7	1.4	21.0	2.5	26.3	4.9
A3	10.7	1.9	15.6	3.3	19.7	1.5	19.5	2.7	24.7	4.5
A4	7.4	2.5	6.8	2.4	16.8	1.9	16.1	3.9	14.4	3.9
A5	12.8	3.7	12.7	3.5	21.5	1.8	19.9	3.4	19.6	4.7
A6	11.2	2.1	15.9	4.5	17.1	1.9	20.4	3.0	24.7	6.1
A7	4.8	4.2	3.4	4.4	11.1	1.9	9.3	4.0	6.9	3.6
B	45.7	9.4	47.6	9.1	70.9	6.2	76.7	11.9	78.3	12.7
B1	14.2	5.1	12.5	4.4	23.9	3.4	24.9	7.1	23.4	6.9
C	77.2	14.1	78.1	12.4	120.1	9.1	122.3	16.9	114.5	16.2
C1	31.4	7.6	31.4	7.3	54.3	5.2	51.6	10.1	49.4	9.8
C1a	9.2	3.4	9.1	3.2	15.9	2.2	17.4	5.0	16.5	4.7
D	26.3	7.2	32.8	7.8	45.2	5.6	49.4	10.9	60.4	11.8
D1	0.0	11.0	0.0	5.1	0.0	7.9	0.0	16.5	0.0	6.2[c]
G	9.0	2.7	9.1	2.5	23.1	2.2	22.1	4.6	22.7	4.3

[a] The entire genome was partitioned into six datasets: 1st, 2nd and 3rd positions of the codons, RNAs, HVS-I and the remainder of the control region
[b] Ancestral Goat Mitogenome
[c] A 95 % C.I. for the age of D1 is 0 to ln(20)/n in units of 4120 years

similar to those of the A (12.8 ka) and G (9.0 ka) clusters, which include only *C. hircus* mitogenomes. In summary, all domestic clusters (A, B1, C1a and G) originated between 14 and 9 ka ago, as confirmed by both ML and rho statistics estimates (Table 2).

A Bayesian Skyline Plot (BSP) analysis was carried out to assess population expansions. The overall BSP points to a steep increase of the female effective population size about 12–10 ka ago (Fig. 2). This analysis was performed on the complete molecule after establishing six different partitions (1^{st}, 2^{nd} and 3^{rd} codon positions, RNA genes, HVS-I and other control-region segments). The same partitions were also considered to perform ML analyses on the entire mitogenome. The final outcomes were an overall mutational rate of 3.95×10^{-8} substitutions per nucleotide per year (1 mutation every 1522 years) on the entire molecule and 1.57×10^{-8} substitutions per nucleotide per year (or 1 mutation every 4120 years) on the coding region alone. Both analyses revealed that the molecular clock could not be rejected (*p-values* > 0.05) when employing the complex GTR/REV model. A clear sign of purifying selection is apparent when calculating the non-synonymous/synonymous ratio, i.e. (ω) = 0.190

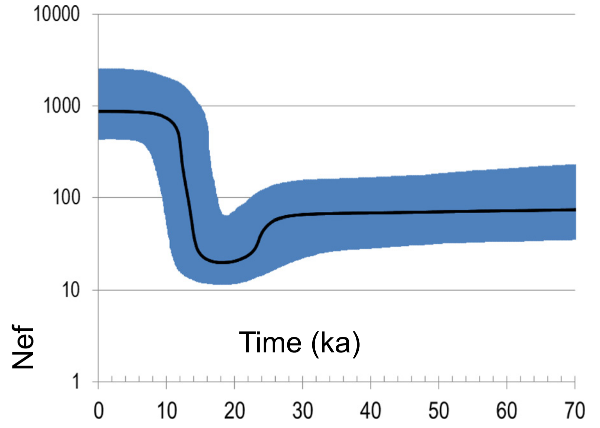


Fig. 2 Bayesian Skyline Plot showing the goat population size trend with a generation time of 4.5 years [60]. The Y axis indicates the effective number of females. The thick solid line is the median estimate and the grey shading shows the 95 % highest posterior density limits. The time axis is limited to 100 ka, beyond that time the curve remains linear

(Table 3). As expected, this ratio is significantly lower (p-$value << 0.001$; Fisher's exact test) in the deep portion of the tree and the comparison among domestic branches reveals slightly significant differences ($\chi 2$ p-$value = 0.026$) (Table 3); in particular the ω ratio of A is much higher than previously reported ($\omega = 0.049$) [17]. Similarly, age values of younger clades are much higher when considering the entire panel of mutations rather than those based on synonymous changes only (Fig. 3), probably because the effect of purifying selection is incomplete and all the negatively selectable characters are still included [29].

Geographic distributions of goat mtDNA haplogroups

An analysis of the geographic distribution of goat haplogroups in Eurasia and Africa (Fig. 4), based on the control-region data currently available in literature or deposited in GenBank, confirms an overwhelming predominance of haplogroup A (~90.5 %) among domestic goats all over the world (including the Americas, Additional file 2: Table S5). The second most frequent haplogroup is B (~6.5 %), more common in Asia and Southern Africa, but also present in Europe. Haplogroup G (0.9 %) is restricted to North-central Africa and Asia, while the presence of haplogroups C and D is limited to Eurasia with an average frequency of 1.4 % and 0.6 %, respectively. Finally, haplogroup F was identified only in three Sicilian domestic samples. The bezoar samples analyzed so far are mainly from the Middle East (Iran, Turkey and Jordan) and South Asia (India, Pakistan and Bangladesh), where they are mostly characterized by the occurrence of haplogroups C (37.4 and 42.4 %) and F (28.9 and 18.2 %). *C. aegagrus* is also represented by haplogroups D (12.9 %), A (6.4 %), B (3.7 %) and G (2.2 %) in the Middle East. Wild goats are also found on some Mediterranean islands (*C. aegagrus cretica*), even though the widely accepted opinion is that they derive from the feralization of very early domestic animals [32]. Finally, the few ancient mtDNA haplotypes (Additional file 2: Table S5) were found in goat remains excavated in Central/East Asia (A, B and D), the Near East (only A) and Europe (B in mainland; A and C in the Mediterranean area). The presence of A and C in the Mediterranean area since ancient times could be due to the Neolithic spread attested by the first appearance of Cardial pottery in the Eastern Adriatic since 8.5 ka ago [33–35].

Table 3 Rates of non-synonymous/synonymous differences (dN/dS) on the goat phylogeny

	dN	dS	dN/dS
Entire Phylogeny	114	600	0.190
Pre-domestic branches	46	361	0.127
Domestic branches	68	239	0.284
	Comparisons within domestic haplogroups		
A	46	199	0.231
B1	3	9	0.333
C1a	3	8	0.375
D1	4	9	0.444
G	12	14	0.857

Discussion and conclusions

Domestication of *C. aegagrus*

Wild goats were widely distributed throughout Southwestern Eurasia during the Middle Paleolithic (300 to 45 ka ago). In an attempt to reconstruct past demographic histories based on phylogenetic inferences, the tree in Fig. 1 suggests that a drastic bottleneck occurred during the Late Eurasian Saalian glacial (~130–160 ka ago) [36–39], which is compatible with the age of the major macro-haplogroup A'B'C'D'G. During the following Riss-Würm interglacial relapse (~115–130 ka ago), the A'B'D'G lineages evolved and expanded independently from the C-derived populations. Many lineages did not survive the following Würm glacial period (~12–71 ka ago) and the drastic climatic changes during the Last Glacial Maximum (LGM, ~19.5–25 ka ago), but some of them, corresponding to the principal goat haplogroups, were preserved in the Near East refuge areas [40], survived the severe drop in temperature known as the Younger Dryas (~11.5–12.7 ka ago) [41, 42], and provided the necessary substrate of variability for later goat domestication and management in an area from the Zagros Mountains to Southeastern Anatolia, as testified by the abundance of goat remains in Neolithic sites from that regions [8, 43].

The main goat haplogroups were identified through control-region analyses following an approach similar to that adopted for other domesticated species, such as sheep [3, 30, 44–46]. However, in the case of both horses and bovines, the mtDNA haplogroup classification received a major makeover from the exploration of the entire mitogenome variability [21, 47–49]. A common feature of all livestock phylogenies is that the control-region molecular clocks turned out to be very inadequate for an accurate dating of mitochondrial lineage divergences. When mtDNA protein-coding genes were considered [17], goat haplogroups were dated to the Middle Paleolithic, thus suggesting that multiple related lineages were domesticated ~11 ka ago in an area spanning from Southeastern Anatolia in Turkey to Zagros Mountains in Iran, by incorporating pre-existing variation [7–9, 17]. This domestication center, further confirmed by zooarchaeological data [43], left a clear signature in the mtDNA gene pool of current domestic breeds, as attested by the large predominance of A haplotypes. Our analyses allowed us to assess the coding-

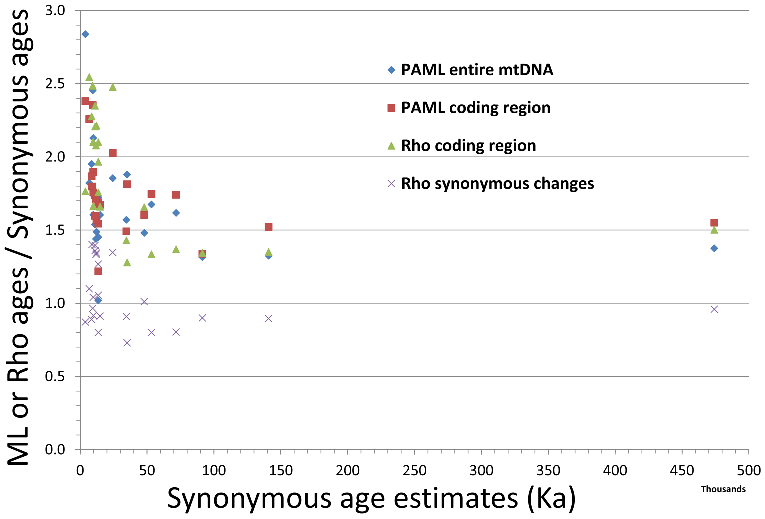


Fig. 3 Comparison of ML and rho-statistics ages relative to the ML synonymous estimates. See Table 2 for further details

region variability within this clade (Additional file 2: Table S6) and to date for the first time this major goat haplogroup to the Epipaleolithic period; a few A sub-haplogroups were also phylogenetically defined and dated to the early Neolithic. The implication is that more than 90 % of current domestic goats might descend from a single foundress, represented by the internal node A in our phylogeny. Obviously, the absence of A bezoars in our dataset is an issue, as well as the long-term calibration point, which might represent an underestimate of the true sheep/goat divergence. Thus, by also considering the alternative hypothesis of many founder lineages within haplogroup A independently proposed by Naderi et al. [7] and Nomura et al. [17], we might suggest that at least seven of them have been identified in our dataset as represented by the ancestral mitogenomes of sub-haplogroups A1–A7.

Moreover, in some instances, our phylogenetic reconstruction clearly shows that other domestic clusters are nested within wild branches and each of them descends from a unique wild haplotype. Therefore, excluding the rare but possible occurrence of stochastic backcrossing between wild and domestic animals, the first domestication process probably involved only four additional female goat populations corresponding to one founder haplotype for each of the other domestic lineages, i.e. B1, D1, C1a and G,. We cannot exclude that, when analyzing further *C. aegagrus* samples at the maximum level of resolution, several additional sub-groups could be identified within each haplogroup, as previously discussed for the A lineages. Likewise, the possibility that some of the mitochondrial haplotypes belonging to lineages B1, D1 and C1a and presently found in domestic animals might derive from introgression of wild lineages cannot be completely ruled out. Yet, events of introgression from wild animals into domestic herds are incidental and usually male-mediated [50], and they were not identified so far even when analyzing a larger dataset of wild and domestic control-region sequences [7, 13]. We have also verified that all available control-region haplotypes (including those from the present study) belonging to haplogroups B, C and D could be specifically assigned either to wild or domestic branches as expected from individual phenotypes (i.e. wild or domestic), without finding any notable exceptions.

The Bayesian analysis of our goat sequences also shows that, after domestication, the early domestic populations soon experienced a drastic demographic expansion about 12–10 ka ago (Fig. 2). However, since our tree includes samples from a very wide geographical area, we cannot exclude that additional and perhaps more recent signals of local demographic expansions could be detected by analyzing larger and more geographically structured datasets, particularly along the domestication routes.

A more fascinating scenario could also be envisioned for haplogroup C since this haplogroup might represent a marker of a concomitant secondary domestication, as already suggested by control-region analyses [7]. To support this scenario, our age estimate for C1a is compatible with the first domestication wave and, more importantly, the C clade represents an independent branch in our tree, which is separated from the A'B'D'G cluster by over one hundred thousand years, thus suggesting a genetically distinct population origin.

The prevalence of haplogroup A: a close similarity between cattle and goat domestication

Phylogenetic analyses based on entire mitogenomes revealed that, similarly to horses and taurine cattle, the ancient goat populations might have passed through a severe bottleneck during the Late Saalian glacial maximum with a single sequence (A'B'C'D'G) from that period as the common ancestor of most goat mitogenomes. However, the recent goat evolutionary history suggests more similarities with *Bos primigenius* domestication (Additional file 2: Table S7). Present-day cattle and goat mitogenomes belong to very few haplogroups, with most of them (>85 % in Eurasia) falling within a single branch (T3 and A, respectively), whose sequence coalescence time corresponds to a very similar time estimate of about 11–13 ka. As in the case of the cattle T3 branch, this study and other published data support a Neolithic origin for all goat A mitogenomes (possibly differentiated into some sub-haplogroups, e.g. A1–A7) from a Middle Eastern bezoar population. At least four other minor lineages (B1, C1a, D1 and G) directly related to bezoar ancestors (i.e. B, C1 and D) were domesticated. However, unlike the bovine minor clades P and Q that suggest a possible European domestication or introgression, the geographical distribution of the low-frequency goat clusters indicates that secondary domestication *foci* were located in the same area between Anatolian Turkey and Iran. This leaves open the possibility of a more eastward Iranian center involving the furthest phylogenetically-related haplotype C1. Moreover, the new coding-region diagnostic mutational motifs defined in the present study (Additional file 2: Table S6) could be employed to revise phylogenetic relationships between modern breeds and ancient remains with the overall objective of testing the proposed scenario of secondary early domestication processes that took place before the occurrence of morphological modifications, and evaluating whether selection differentially affected mitogenome

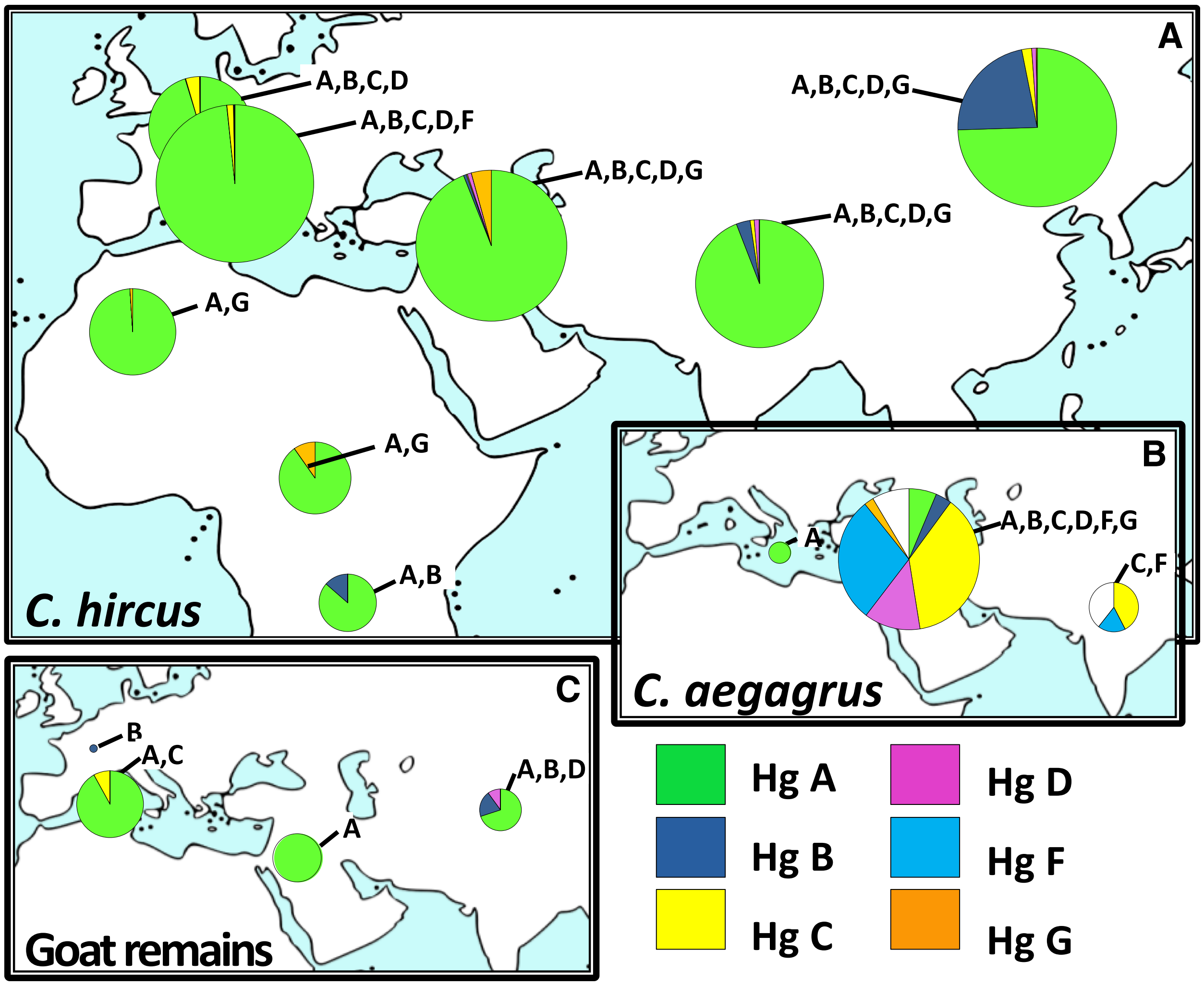


Fig. 4 Spatial frequency distributions of goat mtDNA haplogroups in different geographic areas based on different datasets: modern breeds (*C. hircus*) **a**; wild goats (*C. aegagrus*) **b**; and ancient goat remains **c**. See Additional file 2 (Table S5) for more information

variants during the development of economically important breeds.

Methods

Sequencing of goat mitochondrial genomes

All experimental procedures were reviewed and approved by the Animal Research Ethics Committee of the University of Pavia, Prot. 2/2010 (October 15th, 2010), in accordance with the European Union Directive 86/609. Prior to the sequencing of entire mtDNA molecules, a preliminary sequence analysis of the control regions was performed. This allowed for the selection of 81 samples with good quality DNA and encompassing the widest range of mutational motifs. During the first phases of the experimental work we used the established Sanger sequencing approach rather than new sequencing technologies since, at that time, the latter did not yet guarantee complete coverage and could be still prone to artificial ambiguities [51]. The sequencing protocol was similar to those previously and successfully used for human and livestock mitogenomes [21, 30, 47, 52]. The oligonucleotides used to amplify and sequence the goat mitochondrial genome are reported in Additional file 2: Table S8. They were checked through GenBank BLAST to avoid amplification of nuclear insertions of mitochondrial sequences (NuMtS) [53]. In the last phases of the experimental work, thanks to the acquired affability and accuracy of the next generation techniques, additional 28 samples were amplified with the same oligonucleotide pairs, pooled together (5 μg of DNA in total) and sequenced on the Illumina Genome Analyzer IIx, platform at the IGA Technology Services, Udine, Italy.

Several parameters of the mtDNA sequence variation were estimated by using DnaSP 5.1. The variation of nucleotide diversity (π) along the entire mtDNA was estimated by assessing windows of 100 bps with step size of 50 bps centered at the midpoint. For an estimation of the synonymous/non-synonymous sites we created an alignment containing only the protein-coding genes, with the ND6 gene adjusted to present the same reading direction as the other genes. The “stop codons” were excluded from the analysis. Overlapping loci were counted twice leading to a final alignment of protein-coding genes equal to 11370 bps.

Phylogeny construction and molecular divergence

The phylogeny construction was performed as described elsewhere [21, 30, 47, 52] and confirmed using an adapted version of mtPhyl 3.0 for a maximum parsimony (MP) analysis [54]. The modified .txt files are available upon request.

We constructed an MP tree including 84 goat mitogenomes (without D-loop) rooted on sheep, *O. aries*, mitogenome (KF302445). A first maximum likelihood (ML) analysis was performed using PAML X [55] by considering only the protein coding genes (and synonymous mutations). As already mentioned, the ND6 gene was reverse-complemented to present the same reading direction as the other genes and, under the vertebrate mitochondrial genetic code, the non-synonymous substitutions were excluded from the alignment and replaced with the ancestral base pairs. The “stop codons” were excluded from the analysis. Lengths of protein-coding genes were readjusted and extended by considering the overlapping segments. The final alignment (11370 bps long) was analyzed with CODEML, which calculates a synonymous mutation rate taking into account the mitochondrial genetic code. The second survey was carried out by considering six partitions in the molecule: one corresponds to the RNA genes (tDNA and 12S/16S rDNA), one to each codon position of the protein-coding genes (CDS), one to HVS-I, and one to the remaining D-loop sequences. In the final alignment, the rDNA and tDNA, CDS, HVS-I and D-loop segments were 4042 (overlapping sites between rDNA and tDNA were counted once), 11370 (see above), 481 and 732 bps long respectively. The non-coding region between np 5160 and np 5191 was not considered, but we checked that it was invariable. The best model able to describe the phylogenetic relationships among taxa was selected by using jModelTest [56]. Eventually, separate analyses were performed on the coding region and on the entire mitogenome by assuming a GTR/REV mutation model, a molecular clock and gamma-distributed rates, approximated by a discrete distribution with 32 categories. In order to check the clock hypothesis, likelihood ratio tests were applied with and without molecular clocks. The ML estimates were also compared with those directly obtained on the MP trees by using mtPhyl as averaged distances (ρ) of the haplotypes of a clade to the respective root haplotype, also known as rho-statistics [57], accompanied by a heuristic estimate of the standard error (σ).

We also obtained a Bayesian skyline plot (BSP) [58] from the goat phylogeny using BEAST 2.2.1 software [59]. We run 10,000,000 iterations with samples drawn every 5000 steps and used a generation time of 4.5 years [60]. BSPs provided a good visualization of the increase in diversity in the tree by estimating effective population sizes through time.

Calibrating the goat mtDNA molecular clock

For the calibration point in the maximum likelihood analyses, we assumed an estimated bifurcation time between sheep and goat of 6,000,000 years (assuming a 95 % interval of 5–7,000,000 years in the BEAST analysis) based on fossil evidence as already used by Sultana et al. [11]. Internal calibration points were not available. Considering the time-dependency of molecular rate estimates [31], the use of a paleontological calibration point means that we are prone to possible biases mainly generated by non-synonymous substitutions and mutations affecting tRNA/rRNA genes (purifying selection) and the control region (tendency to saturation due to the high evolution rate). This long-term calibration point probably represents an underestimate of the true sheep/goat divergence and might have had a considerable impact on the date estimates.

Availability of supporting data

Sequences of the novel goat mitogenomes have been deposited in GenBank under accession numbers KR059146 - KR059226 (81 complete mtDNAs) and KR059227 - KR059851 (625 mtDNA control regions). Phylogenetic data have been deposited in TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:S18595).

Additional files

Additional file 1: Table S1. Sources for the 758 goat control-region sequences. **Table S2.** Control-region haplotypes and haplogroup classification of the 758 mtDNA sequences from *Capra aegagrus* (*n* = 19) and *Capra hircus* (*n* = 739). **Table S3.** Partial coding-region haplotypes and haplogroup classification of two bezoar mtDNAs. **Table S4.** Source and haplogroup affiliation of the goat complete mtDNA sequences. **Figure S1.** Nucleotide diversity and total number of substitutions along the entire mtDNA. **Figure S2.** A putative most parsimonious tree of 84 complete mtDNA sequences from goats. (XLSX 1268 kb)

Additional file 2: Table S5. Goat haplogroup frequencies based on modern and ancient control-region mtDNA data from this study and downloaded from GenBank[a]. **Table S6.** Diagnostic mutational motifs of goat mtDNA haplogroups and sub-haplogroups. **Table S7.** A comparison of the phylogeographic features of goat, taurine and horse mtDNA haplogroups identified by analyzing domestic breeds from Eurasia. **Table S8.** Oligonucleotides used to amplify and to sequence (Sanger method) the goat mitochondrial genome. (PDF 652 kb)

Abbreviations

AGM: Ancestral Goat Mitogenome; BEAST: Bayesian Evolutionary Analysis Sampling Trees; BLAST: Basic Local Alignment Search Tool; Bps: Base pairs; BSP: Bayesian Skyline Plot; CDS: Coding DNA Sequence; D-loop: Displacement Loop; dN/dS: Non-synonymous/synonymous ratio; DnaSP: DNA Sequence Polymorphism; FAO: Food and Agriculture Organization; GRS: Goat Reference Sequence; GTR/REV: General Time Reversible; Hg: Haplogroup; HVS-I: Hypervariable segment-I; Indels: Insertions/deletions; Ka: Kilo annos; LGM: Last Glacial Maximum; ML: Maximum Likelihood; mtDNA: Mitochondrial DNA; ND6: NADH, ubiquinone oxidoreductase core subunit 6; NuMt: Nuclear Mitochondrial DNA; PAML: Phylogenetic Analysis Using Maximum Likelihood; rDNA: Ribosomal DNA; rRNA: Ribosomal RNA; SNP: Single Nucleotide Polymorphism; tDNA: Transfer DNA; tRNA: Transfer RNA.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

This study was conceived by LC, HL, FP, PT, PAM and AA. LC, HL, IC and AA designed and performed molecular experiments. LC, HL, IC, AO, MRC, MP, MR, FG, VB, PAM and AA conducted the genetic analysis. ER, MTS, BP, FB, EC, MP, MR, FG, PAM and AA contributed reagents/materials/analysis tools. WZ, SN, SMFV, SA, HRR and PL performed the collection of biological samples. LC, HL, PAM and AA wrote the paper. All authors read and approved the final manuscript.

Acknowledgments

We thank Dr. Ugo A. Perego and Dr. Ivan Fowler for proofreading the manuscript. This research received support from the Italian Ministry of Education, University and Research: Progetti FIRB-Futuro in Ricerca 2008 and 2012 (to AA and AO), and PRIN-Progetti Ricerca di Interesse Nazionale 2007 (to PAM), 2009 and 2012 (to AA); and from the EU FP7 project NextGen (Grant KBBE-2009-1-1-03).

Author details

[1]Institute of Zootechnics, Università Cattolica del S. Cuore, Piacenza 29122, Italy. [2]Research Center on Biodiversity and Ancient DNA – BioDNA, Università Cattolica del S. Cuore, Piacenza 29122, Italy. [3]Dipartimento di Chimica, Biologia e Biotecnologie, Università di Perugia, Perugia 06123, Italy. [4]Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università di Pavia, Pavia 27100, Italy. [5]Institute of Food Science and Nutrition - ISAN, Università Cattolica del S. Cuore, Piacenza 29122, Italy. [6]Université Grenoble Alpes, Laboratoire d'Ecologie Alpine, Grenoble 38041, France. [7]Department of Environmental Sciences, Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University, Noor, Mazandaran 46414-356, Iran. [8]Natural Resources Faculty, University of Guilan, Guilan 41335-1914, Iran. [9]School of Applied Sciences, University of Huddersfield, Huddersfield HD1 3DH, UK. [10]Agricultural Biotechnology Research Institute of Iran (ABRII), North Branch, Rasht 41635-4115, Iran. [11]Department of Animal Production, Faculty of Agriculture, Ain Shams University, Cairo 11241, Egypt. [12]Laboratorio di Genetica, Istituto per la Protezione e la Ricerca Ambientale (ISPRA), Bologna 40064, Italy. [13]Department 18/Section of Environmental Engineering, Aalborg University, Aalborg DK-9000, Denmark. [14]Dipartimento Scienze Agrarie e Forestali, Università degli Studi di Palermo, Palermo 90128, Italy. [15]Environmental Sciences Department, Gorgan University of Agriculture and Natural Resources, Gorgan 49138-15739, Iran. [16]Natural History Museum of Crete, University of Crete, Iraklio, Crete 71409, Greece.

Received: 30 June 2015 Accepted: 22 December 2015
Published online: 29 December 2015

References

1. FAO. The State of the World's Animal Genetic Resources for Food and Agriculture. In: Pilling D, Rischkowsky B, editors. Rome; 2007.
2. MacHugh DE, Bradley DG. Livestock genetic origins: goats buck the trend. Proc Natl Acad Sci U S A. 2001;98(10):5382–4.
3. Zeder MA. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci U S A. 2008;105(33):11597–604.
4. Zeder MA, Emshwiller E, Smith BD, Bradley DG. Documenting domestication: the intersection of genetics and archaeology. Trends Genet. 2006;22(3):139–55.
5. Vigne JD, Peters J, Helmer D. The first steps of animal domestication: new archaeozoological approaches. Oxford: Oxbow; 2005.
6. Taberlet P, Coissac E, Pansu J, Pompanon F. Conservation genetics of cattle, sheep, and goats. C R Biol. 2011;334(3):247–54.
7. Naderi S, Rezaei HR, Pompanon F, Blum MG, Negrini R, Naghash HR, et al. The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. Proc Natl Acad Sci U S A. 2008;105(46):17659–64.
8. Zeder MA, Hesse B. The initial domestication of goats (*Capra hircus*) in the Zagros mountains 10,000 years ago. Science. 2000;287(5461):2254–7.
9. Joshi MB, Rout PK, Mandal AK, Tyler-Smith C, Singh L, Thangaraj K. Phylogeography and origin of Indian domestic goats. Mol Biol Evol. 2004;21(3):454–62.
10. Luikart G, Gielly L, Excoffier L, Vigne JD, Bouvet J, Taberlet P. Multiple maternal origins and weak phylogeographic structure in domestic goats. Proc Natl Acad Sci U S A. 2001;98(10):5927–32.
11. Sultana S, Mannen H, Tsuji S. Mitochondrial DNA diversity of Pakistani goats. Anim Genet. 2003;34(6):417–21.
12. Sardina MT, Ballester M, Marmi J, Finocchiaro R, van Kaam JB, Portolano B, et al. Phylogenetic analysis of Sicilian goats reveals a new mtDNA lineage. Anim Genet. 2006;37(4):376–8.

13. Naderi S, Rezaei HR, Taberlet P, Zundel S, Rafat SA, Naghash HR, et al. Large-scale mitochondrial DNA analysis of the domestic goat reveals six haplogroups with high diversity. PLoS One. 2007;2(10):e1012.
14. Pereira F, Queiros S, Gusmao L, Nijman IJ, Cuppen E, Lenstra JA, et al. Tracing the history of goat pastoralism: new clues from mitochondrial and Y chromosome DNA in North Africa. Mol Biol Evol. 2009;26(12): 2765–73.
15. Piras D, Doro MG, Casu G, Melis PM, Vaccargiu S, Piras I, et al. Haplotype affinities resolve a major component of goat (*Capra hircus*) MtDNA D-loop diversity and reveal specific features of the Sardinian stock. PLoS One. 2012; 7(2):e30785.
16. Fernández H, Hughes S, Vigne JD, Helmer D, Hodgins G, Miquel C, et al. Divergent mtDNA lineages of goats in an Early Neolithic site, far from the initial domestication areas. Proc Natl Acad Sci U S A. 2006;103(42):15375–9.
17. Nomura K, Yonezawa T, Mano S, Kawakami S, Shedlock AM, Hasegawa M, et al. Domestication process of the goat revealed by an analysis of the nearly complete mitochondrial protein-encoding genes. PLoS One. 2013;8(8): e67775.
18. McCracken K, Sorenson M. Is homoplasy or lineage sorting the source of incongruent mtDNA and nuclear gene trees in the stiff-tailed ducks (Nomonyx-Oxyura)? Syst Biol. 2005;54(1):35–55.
19. Achilli A, Bonfiglio S, Olivieri A, Malusa A, Pala M, Kashani BH, et al. The multifaceted origin of taurine cattle reflected by the mitochondrial genome. PLoS One. 2009;4(6):e5753.
20. Bonfiglio S, Ginja C, De Gaetano A, Achilli A, Olivieri A, Colli L, et al. Origin and spread of *Bos taurus*: new clues from mitochondrial genomes belonging to haplogroup T1. PLoS One. 2012;7(6):e38601.
21. Achilli A, Olivieri A, Soares P, Lancioni H, Kashani BH, Perego UA, et al. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc Natl Acad Sci U S A. 2012;109(7):2449–54.
22. Hassanin A, Bonillo C, Nguyen BX, Cruaud C. Comparisons between mitochondrial genomes of domestic goat (*Capra hircus*) reveal the presence of numts and multiple sequencing errors. Mitochondrial DNA. 2010;21(3–4): 68–76.
23. Kim H, Lee T, Park W, Lee JW, Kim J, Lee BY, et al. Peeling back the evolutionary layers of molecular mechanisms responsive to exercise-stress in the skeletal muscle of the racing horse. DNA Res. 2013;20(3): 287–98.
24. Moyle RG, Jones RM, Andersen MJ. A reconsideration of Gallicolumba (Aves: *Columbidae*) relationships using fresh source material reveals pseudogenes, chimeras, and a novel phylogenetic hypothesis. Mol Phylogenet Evol. 2013; 66(3):1060–6.
25. Parma P, Feligini M, Greeppi G, Enne G. The complete nucleotide sequence of goat (*Capra hircus*) mitochondrial genome. Goat mitochondrial genome. DNA Seq. 2003;14(3):199–203.
26. Doro MG, Piras D, Leoni GG, Casu G, Vaccargiu S, Parracciani D, et al. Phylogeny and patterns of diversity of goat mtDNA haplogroup A revealed by resequencing complete mitogenomes. PLoS One. 2014;9(4):e95969.
27. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, et al. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A. 2003;100(1):171–6.
28. Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, et al. The role of selection in the evolution of human mitochondrial genomes. Genetics. 2006;172(1): 373–87.
29. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009;84(6):740–59.
30. Lancioni H, Di Lorenzo P, Ceccobelli S, Perego UA, Miglio A, Landi V, et al. Phylogenetic relationships of three Italian merino-derived sheep breeds evaluated through a complete mitogenome analysis. PLoS One. 2013;8(9): e73712.
31. Ho SY, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, et al. Time-dependent rates of molecular evolution. Mol Ecol. 2011;20(15): 3087–101.
32. Kahila Bar-Gal G, Smith P, Tchernov E, Greenblatt C, Ducos P, Gardeisen A, et al. Genetic evidence for the origin of the agrimi goat (*Capra aegagrus cretica*). J Zool. 2002;256(3):369–77.
33. Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, et al. Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. Eur J Hum Genet. 2009;17(6):820–30.
34. Paschou P, Drineas P, Yannaki E, Razou A, Kanaki K, Tsetsos F, et al. Maritime route of colonization of Europe. Proc Natl Acad Sci U S A. 2014;111(25): 9211–6.
35. Fernández E, Pérez-Pérez A, Gamba C, Prats E, Cuesta P, Anfruns J, et al. Ancient DNA analysis of 8000 B.C. near eastern farmers supports an early neolithic pioneer maritime colonization of Mainland Europe through Cyprus and the Aegean Islands. PLoS Genet. 2014;10(6): e1004401.
36. Richter J. When Did the Middle Paleolithic Begin? In: Conard NJ, Richter J, editors. Neanderthal lifeways, subsistence and technology: one hundred fifty years of Neanderthal study Vertebrate Paleobiology and Paleoanthropology. Berlin: Springer; 2011.
37. Ehlers J, Gibbard PL, Hughes PD. Quaternary glaciations - extent and chronology: a closer look. Amsterdam: Elsevier; 2011.
38. Colleoni F. On the Late Saalian glaciation (160–140 ka) – a climate modeling study. Stockholm: Stockholm University; 2009.
39. Ferrigno JG. Glaciers of the Middle East and Africa - Glaciers of Iran. In: Williams RS, Ferrigno JG, editors. Satellite Image Atlas of Glaciers of the World. Washington, DC: Dept. of the Interior, U.S. Geological Survey; 1991.
40. Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, et al. Mitochondrial DNA signals of late glacial recolonization of Europe from Near Eastern refugia. Am J Hum Genet. 2012;90(5):915–24.
41. Clutton-Brock J. A natural history of domesticated mammals, 2nd Edition edn. Cambridge: Cambridge University Press; 1999.
42. MacFadden BJ. Fossil horses: systematics, paleobiology, and evolution of the family Equidae. Cambridge: Cambridge University Press; 1992.
43. Groves C, Leslie D, Huffman B, Valdez R, Habibi K, Weinberg P, et al. Bovidae (Hollow-Horned Ruminants). In: Wilson DE, Mittermeier RA, editors. Handbook of the Mammals of the World Volume 2 Hoofed Mammals. Barcelona: Lynx Edicions; 2011.
44. Driscoll CA, Macdonald DW, O'Brien SJ. From wild animals to domestic pets, an evolutionary view of domestication. Proc Natl Acad Sci U S A. 2009; 106(1):9971–8.
45. Groeneveld LF, Lenstra JA, Eding H, Toro MA, Scherf B, Pilling D, et al. Genetic diversity in farm animals–a review. Anim Genet. 2010;41(1):6–31.
46. Taberlet P, Valentini A, Rezaei HR, Naderi S, Pompanon F, Negrini R, et al. Are cattle, sheep, and goats endangered species? Mol Ecol. 2008;17(1):275–84.
47. Achilli A, Olivieri A, Pellecchia M, Uboldi C, Colli L, Al-Zahery N, et al. Mitochondrial genomes of extinct aurochs survive in domestic cattle. Curr Biol. 2008;18(4):R157–8.
48. Lippold S, Matzke NJ, Reissmann M, Hofreiter M. Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. BMC Evol Biol. 2011;11:328.
49. Bonfiglio S, Achilli A, Olivieri A, Negrini R, Colli L, Liotta L, et al. The enigmatic origin of bovine mtDNA haplogroup R: sporadic interbreeding or an independent event of *Bos primigenius* domestication in Italy? PLoS One. 2010;5(12):e15760.
50. Kikkawa Y, Takada T, Sutopo, Nomura K, Namikawa T, Yonekawa H, et al. Phylogenies using mtDNA and SRY provide evidence for male-mediated introgression in Asian domestic cattle. Anim Genet. 2003; 34(2):96–101.
51. Bandelt HJ, Salas A. Current next generation sequencing technology may not meet forensic standards. Forensic Sci Int Genet. 2012;6(1):143–5.
52. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, et al. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science. 2006;314(5806):1767–70.
53. Nergadze SG, Lupotto M, Pellanda P, Santagostino M, Vitelli V, Giulotto E. Mitochondrial DNA insertions in the nuclear horse genome. Anim Genet. 2010;41(2):176–85.
54. Eltsov NP, Volodko NV. mtPhyl program. 2011. http://eltsov.org.
55. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997;13(5):555–6.
56. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol. 2004;53(5): 793–808.
57. Forster P, Harding R, Torroni A, Bandelt HJ. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet. 1996;59(4):935–45.

58. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005;22(5):1185–92.
59. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214.
60. Thirstrup JP, Bach LA, Loeschcke V, Pertoldi C. Population viability analysis on domestic horse breeds (*Equus caballus*). J Anim Sci. 2009;87(11):3525–35.

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

BRIEF NOTES

doi: 10.1111/age.12408

Survey of uniparental genetic markers in the Maltese cattle breed reveals a significant founder effect but does not indicate local domestication

Hovirag Lancioni[*], Piera Di Lorenzo[†], Irene Cardinali[*], Simone Ceccobelli[†], Marco Rosario Capodiferro[*,‡], Alessandro Fichera[‡], Viola Grugni[‡], Ornella Semino[‡], Luca Ferretti[‡], Anthony Gruppetta[§], George Attard[§,1], Alessandro Achilli[*,‡,1] and Emiliano Lasagna[†,1]

[*]Dipartimento di Chimica, Biologia e Biotecnologie, Università degli Studi di Perugia, Perugia 06123, Italy; [†]Dipartimento di Scienze Agrarie, Alimentari e Ambientali, Università degli Studi di Perugia, Perugia 06121, Italy; [‡]Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università degli Studi di Pavia, Pavia 27100, Italy; [§]Department of Rural Sciences and Food Systems, Institute of Earth Systems, University of Malta, Msida, MSD 2080, Malta

Accepted for publication 24 November 2015

Description: Local breeds often represent unique and endangered sources of genetic variability, particularly when confined to isolated geographic areas. The Maltese breed of cattle is considered to be of ancient origin. Late Pleistocene oxen skeletal remains and Neolithic representations of primitive cattle are argued as proof of local domestication. However, the origin of the Maltese cattle has always been shrouded in mystery. In addition, subsequent stochastic or intentional mating with other stocks may have eroded the original genetic profile (Appendix S1). The results of this study reflect the probable recent history of the Maltese cattle, given that the last authentic Maltese bull was culled in 1990 and semen from the Chianina breed, which shows phenotypic traits similar to the Maltese breed, was used to propagate the breed. The present adult population consists of 12 males and 19 females divided into two herds.

In order to contribute to the preservation of Maltese cattle genetic uniqueness, the Y chromosomal bi-allelic markers of paternal lineages[1] in 10 males were analyzed and their entire mitochondrial genome sequences were determined as markers of maternal lineages together with 13 female samples (Fig. S1).[2]

Y chromosomes: A 81-bp insertion in intron 26 of *USP9Y* showed that all 10 bulls carried the haplogroup Y2 (Table 1). This haplogroup is reported as prevalent in most central Mediterranean and Iberian breeds and represents the only Y haplogroup found in various Italian breeds, including Chianina.[1,3]

Mitogenomes: Control-region sequencing revealed only two mitochondrial haplotypes. Both belong to the T3 haplogroup (Table 1), with the prevalent unique haplotype (HT2) present in 91% of the samples, suggesting a strong maternal founder effect. The second control-region haplotype (HT1) was identified in only two samples. When compared to the other control-region sequences deposited in GenBank, the HT1 turned out to be identical to a sequence derived from an Ayrshire individual.

One sample for each haplotype was sequenced entirely (GenBank Accession Nos. KT343748 and KT343749) and compared with all available complete mtDNAs (Table S1). The Maltese mitogenomes create two novel subclades (Fig. 1, T3c and T3d), both dated back to 9500 years and encompassing, besides the Maltese cattle, only four animals of British ancestry (White Park, Angus) and one German Red Mountain cattle.

Conclusions: The parental lineages of the extant Maltese cattle herd are represented by only one Y chromosome haplotype and two mitogenomes, which are different from the currently available mtDNA sequences. Our findings reveal a significant founder effect, which is quite common worldwide,[4] but the hypothesis of a local domestication in Malta is not supported. In fact, the two Maltese cattle mitogenomes belong to the T3 haplogroup, whose domestication most likely took place in the Near East ~10–12 thousand years ago,[2,5–7] although alternative scenarios have been proposed.[8] Moreover, the Maltese T3 variants also occur in northern or central Europe, but have so far not been found in Mediterranean cattle.

References

1 Bonfiglio S. *et al.* (2012) *Anim Genet* **43**, 611–3.
2 Achilli A. *et al.* (2008) *Curr Biol* **18**, R157–8.
3 Edwards CJ. *et al.* (2011) *PLoS One* **6**, e15922.
4 Lenstra JA. *et al.* (2014) *Diversity* **6**, 178–87.
5 Olivieri A. *et al.* (2015) *PLoS One* **10**, e0141170.
6 Park SD. *et al.* (2015) *Genome Biol* **16**, 234.
7 Scheu A. *et al.* (2015) *BMC Genet* **16**, 54.
8 Lari M. *et al.* (2011) *BMC Genet* **11**, 32.

Correspondence: H. Lancioni (hovirag.lancioni@unipg.it)

[1] These authors contributed equally to the manuscript.

Table 1 Sources of Maltese cattle samples. Haplotypes (Ht) and haplogroup (Hg) classification of mitochondrial genomes and Y-chromosome (Y-chr) haplogroup affiliation are reported.

			Mitochondrial DNA					
Samples			Control-region (range 15823-215)			Complete mtDNA[1]		Y-chr
ID	Sex	Farm	Haplotype	Ht-ID	Hg	Haplotype	Hg	Hg
MLT01	F	Qormi[2]	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT02	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT03	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT04	F	Qormi	24, 169	HT1	T3	24, 169, 221+C, 587+C, 1600d, 2536A, 9682C, 10162, 11476, 13310C	T3c	
MLT05	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT06	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT07	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT08	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT09	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT10	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	169, 174, 190, 587+C, 1600d, 2536A, 3964, 9308, 9682C, 13310C, 15532C, 16022, 16231	T3d	Y2
MLT11	M	Qormi	24, 169	HT1	T3	–		Y2
MLT12	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT13	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT14	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT15	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT16	F	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT17	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT18	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT19	M	Qormi	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT20	M	Marsascala[3]	16022, 16231, 169, 174, 190	HT2	T3	–		Y2
MLT21	F	Marsascala	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT22	F	Marsascala	16022, 16231, 169, 174, 190	HT2	T3	–		
MLT23	F	Marsascala	16022, 16231, 169, 174, 190	HT2	T3	–		

[1] GenBank Accession numbers: KT343748 and KT343749 for HT1 and HT2, respectively.

[2] Government Experimental Farm at Qormi.

[3] Saliba's holding at Marsascala.

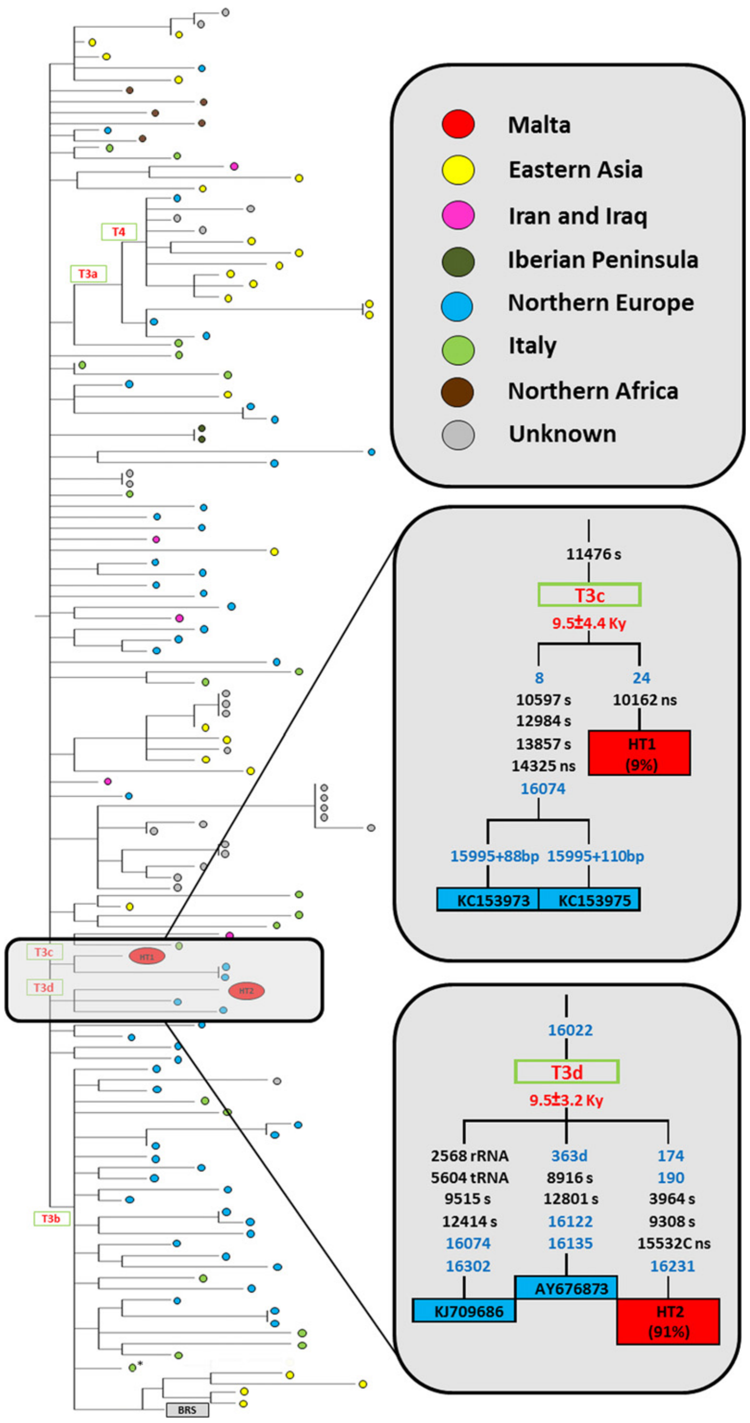


Figure 1 Schematic MP Tree of the mtDNA haplogroup T3. This tree encompasses 132 GenBank sequences (Table S1), the two novel Maltese mitogenomes (red circles) and the Italian aurochs T3 genome (*).[8] The insets show the novel clades T3c and T3d. See Appendix S1 for further details.

Supporting information

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

Figure S1 Geographical localization of Maltese Islands and farms (star denotes Government Experimental Farm at Qormi, whereas triangle denotes Saliba's holding at Marsascala). On the right, one of the three founder females of the extant herd is reported. The picture shows the peculiar phenotypic traits of the Maltese breed.

Table S1 Sources and haplogroup affiliation for the *Bos taurus* complete mtDNA sequences: haplogroup classification within clade T3 was inferred from the MP tree reported in Figure 1.

Appendix S1 The history of Maltese cattle and other supporting information.

PLOS | ONE

RESEARCH ARTICLE

An Overview of Ten Italian Horse Breeds through Mitochondrial DNA

Irene Cardinali[1☯], Hovirag Lancioni[1☯], Andrea Giontella[2☯], Marco Rosario Capodiferro[3], Stefano Capomaccio[2], Luca Buttazzoni[4], Giovanni Paolo Biggio[5], Raffaele Cherchi[5], Emidio Albertini[6], Anna Olivieri[3], Katia Cappelli[2], Alessandro Achilli[1,3‡]*, Maurizio Silvestrelli[2‡]

1 Dipartimento di Chimica, Biologia e Biotecnologie, Università di Perugia, Perugia, Italy, **2** Centro di Studio del Cavallo Sportivo, Dipartimento di Medicina Veterinaria, Università di Perugia, Perugia, Italy, **3** Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università di Pavia, Pavia, Italy, **4** Centro di ricerca per la produzione delle carni e il miglioramento genetico, Sede centrale–Monterotondo, Roma, Italy, **5** Agenzia per la ricerca in agricoltura–AGRIS Sardegna, Sassari, Italy, **6** Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università di Perugia, Perugia, Italy

☯ These authors contributed equally to this work.
‡ These authors jointly supervised this work.
* alessandro.achilli@unipv.it

Abstract

Background

The climatic and cultural diversity of the Italian Peninsula triggered, over time, the development of a great variety of horse breeds, whose origin and history are still unclear. To clarify this issue, analyses on phenotypic traits and genealogical data were recently coupled with molecular screening.

Methodology

To provide a comprehensive overview of the horse genetic variability in Italy, we produced and phylogenetically analyzed 407 mitochondrial DNA (mtDNA) control-region sequences from ten of the most important Italian riding horse and pony breeds: Bardigiano, Esperia, Giara, Lipizzan, Maremmano, Monterufolino, Murgese, Sarcidano, Sardinian Anglo-Arab, and Tolfetano. A collection of 36 Arabian horses was also evaluated to assess the genetic consequences of their common use for the improvement of some local breeds.

Conclusions

In Italian horses, all previously described domestic mtDNA haplogroups were detected as well as a high haplotype diversity. These findings indicate that the ancestral local mares harbored an extensive genetic diversity. Moreover, the limited haplotype sharing (11%) with the Arabian horse reveals that its impact on the autochthonous mitochondrial gene pools during the final establishment of pure breeds was marginal, if any. The only significant signs of genetic structure and differentiation were detected in the geographically most isolated contexts (i.e. Monterufolino and Sardinian breeds). Such a geographic effect was also confirmed in a wider breed setting, where the Italian pool stands in an intermediate position together with most of the other Mediterranean stocks. However, some notable exceptions

OPEN ACCESS

Citation: Cardinali I, Lancioni H, Giontella A, Capodiferro MR, Capomaccio S, Buttazzoni L, et al. (2016) An Overview of Ten Italian Horse Breeds through Mitochondrial DNA. PLoS ONE 11(4): e0153004. doi:10.1371/journal.pone.0153004

Editor: David Caramelli, University of Florence, ITALY

Received: December 2, 2015

Accepted: March 21, 2016

Published: April 7, 2016

Data Availability Statement: All relevant data are within the paper (and/or its Supporting Information files) and available from the NCBI database (Genbank Accession Numbers KU711082-KU711507).

Funding: This study received support from the Italian Ministry of Education, University and Research: FIR2012—RBFR126B8I (to AA and AO) and PRIN2012-2012JA4BTY (to AA); AGRIS Sardegna: "Monitoraggio delle caratteristiche dell'allevamento del cavallo sportivo in Sardegna e creazione della banca dati per gli indici genetici"; CRA-PCM:

"Allevamento Statale Cavallo Lipizzano: conservazione e valorizzazione".

Competing Interests: The authors have declared that no competing interests exist.

and peculiar genetic proximities lend genetic support to historical theories about the origin of specific Italian breeds.

Introduction

A great variety of horse breeds developed, over time, in various Italian cultural contexts and geographic habitats. Light horses (hotblood/warmblood; withers height: 148–170 cm) are typical of the drier central and southern regions, while the northern wet regions are characterized by heavy horses (coldblood; withers height: 148–165 cm). Harsh conditions of marginal and insular areas fostered the smaller size horses (ponies; withers height: 115–147 cm). Until the 1940s horse breeding was mainly linked to the production of animals for military purposes, agricultural labors, forestry and local carriages. Beginning in the fifties, the mechanization of agriculture and transportation caused a rapid decline of horse breeding; such trend has been currently mitigated by a renewed cultural interest in rural life. Most recently, the increased leisure-time physical activities have resulted in a growing consideration and demand for "riding horses"; riding refers to the use of horses for leisure/pleasure purposes including competition events (jumping, driving, flat racing, etc.). In Italy, the demand for riding horses includes: cosmopolitan breeds (Thoroughbreds and Arabs), many autochthonous Italian breeds described in Studbooks, many local Italian populations with "Anagraphic Register of equine populations identifiable as local ethnic groups" and several crossbreedings between all of them.

Phenotypic traits and genealogical data are often insufficient to ascertain the horse history and origin. Molecular analyses provide a needful and reliable tool that can be employed along with the morphometric approach and traditional breeding strategies for an efficient management of genetic resources [1]. Due to its high mutation rate, lack of recombination and maternal inheritance, the control region of the mitochondrial DNA (mtDNA) is a powerful marker system for phylogenetic and phylogeographic studies. MtDNA studies on horses have proved to be capable to identify intra- and interbreed relationships [2–9], particularly when combined with historical information [2, 10, 11]. Unfortunately, most previous studies have been carried out on a very short and hypervariable segment (~350 bp) of the control region (HVSI: nucleotide positions 15,469–15,834) [10, 12–15]. In 2013 Khanshour and Cothran [9] have shown in Arabian horse populations that the degree of informativeness can be extensively improved by increasing the length of the analyzed mtDNA control-region sequence. Most recently, similar to many other livestock species [16–18] also the sequence variation of the entire equine mitogenome was investigated [19–21], contributing extensively to our current understanding of the domestication process. Seventeen different mtDNA haplogroups were identified in domestic breeds leading to the conclusion that the domestication of the wild horse, *Equus ferus*, has been a widespread process that persisted for several thousands of years (throughout the Neolithic) and occurred at different places, mostly centered in the Western Eurasian steppes [22], as also suggested by archeological evidences [23]; but possibly also in Western Europe [19]. The spread of domestic herds across Eurasia involved an extensive introgression from the wild; in particular, it has been proposed that the horse was introduced in Italy with the arrival of Indo-European populations in the Bronze Age and used for military, riding and agricultural purposes [24].

Despite the pivotal role that horses have played in human society's development, multiple aspects of modern breeds' origin and history remain unclear. In Italy, several local breeds have reached a national recognition due to their phenotypic characteristics and to particular socio-cultural and productive peculiarities (a complete list is available at http://www.fao.org/dad-is/).

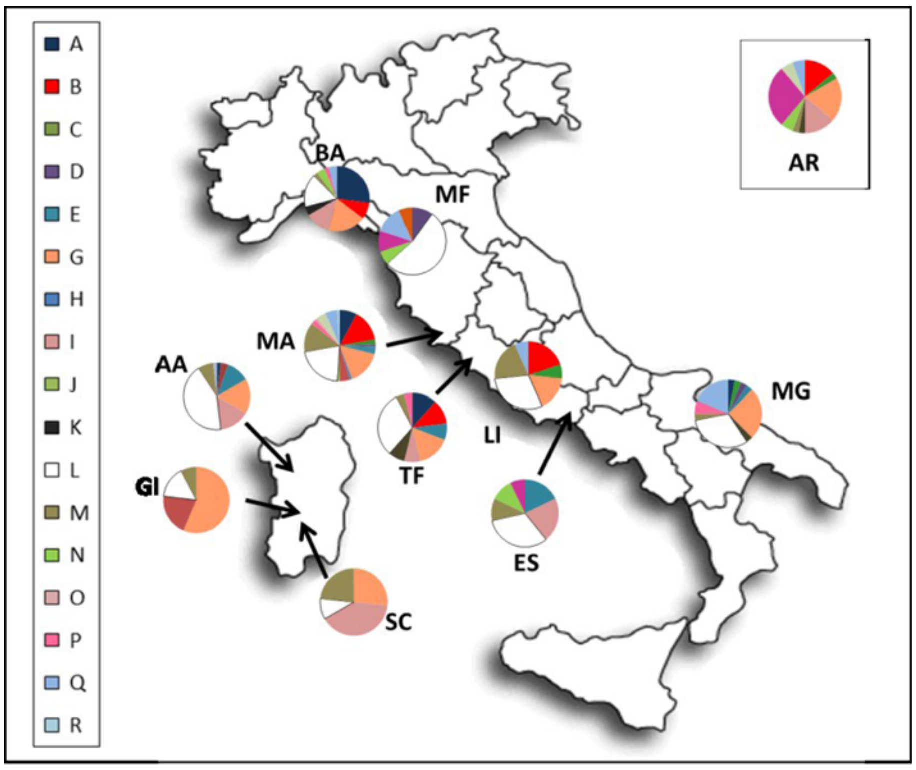


Fig 1. Sampling locations and frequency distributions of mitochondrial haplogroups. Breed code as in Table 2.

doi:10.1371/journal.pone.0153004.g001

However, genetic studies of Italian horse breeds are still limited [25–28] and there are only a few examples of maternal inheritance investigations, but they generally focused on a specific geographic area [14, 29, 30] or included a limited number of samples per breed [31, 32].

To obtain a more comprehensive overview of the Italian horse mitochondrial gene pool we have here determined and phylogenetically analyzed the mtDNA control-region variation of 407 horses from ten of the most important Italian riding horses (including hotblood/warmblood horses and ponies): Bardigiano, Esperia, Giara, Lipizzan, Maremmano, Monterufolino, Murgese, Sarcidano, Sardinian Anglo-Arab and Tolfetano (Fig 1 and Table 1).

Results and Discussion

An overview of the mtDNA sequence variation

More than half of the mtDNA control region, precisely 610 bps (from np 15491 to np 16100), was sequenced in all 407 Italian samples. An additional collection of 36 Arabian horses, which were heavily used in the improvement of some Italian breeds, was analyzed and used as an external reference group. Overall, we identified from seven to 52 haplotypes in the different Italian breeds and 14 in the Arabian horses, summing up to a total of 126 distinct haplotypes. Seventy-eight were unique (found only in a single Italian breed) while 34 were shared among different Italian breeds. Only four haplotypes were in common between Italian and Arabian horses (S1 Table) and these might represent the legacy of recent maternal gene flow from Arabian horses into Italian breeds. Taking into account that the four haplotypes encompass only eleven horses [Maremmano (5), Lipizzan (3) and Sardinian Anglo-Arab (1) horses, Bardigiano (1) and Esperia ponies (1)], this observation indicates that the Arabian horse contributed at

Table 1. List of Italian breeds analyzed in this paper.

Breed	Current Distribution in Italy	Breeding Conditions	Breed Consistency[a]	Genealogical Records	Withers	Breed Classification (AIA)[b]	Breed Classification (FAO)[c]
Bardigiano	North (Emilian Apennine)	Controlled	2606	Studbook	Pony	Autochthonous breed	National breed
Monterufolino	Centre (Tuscany)	Semi-feral[d]	252	Anagr. Reg.	Pony	Autochthonous breed	Local breed (Critical)
Maremmano	Centre (Tuscany and Latium)	Controlled	6300	Studbook	Horse	Autochthonous breed	National breed
Tolfetano	Centre (Latium)	Semi-feral	1518	Anagr. Reg.	Pony	Autochthonous breed	Local breed (Endangered)
Lipizzan	Centre (Latium)	Controlled	433	Studbook	Horse		International breed
Esperia	Centre (South Latium)	Semi-feral	1373	Anagr. Reg.	Pony	Autochthonous breed	Local breed (Endangered)
Murgese	South (Apulia)	Controlled	5564	Studbook	Horse	Autochthonous breed	National breed
Sardinian Anglo-Arab	Sardinia	Controlled	9606	Studbook	Horse	Autochthonous breed	National breed
Giara	Sardinia	Semi-feral	518	Anagr. Reg.	Pony	Autochthonous breed	Local breed (Endangered)
Sarcidano	Sardinia	Semi-feral	110	Anagr. Reg.	Pony	Autochthonous breed	Local breed

[a] Breed consistency = number of individuals recorded in Studbooks or Anagraphic Registers as in 2015 [33, 34].

[b] Italian breeders' association. The Ministerial Decree n. 1598 of January 23, 2015 reported fifteen indigenous horse breeds recorded in the Italian Registry of Autochthonous Equine Breeds.

[c] Based on the establishment of Studbooks or Anagraphic Registers.

[d] Free-roaming horse breeds of domesticated ancestry requiring minimal human management (e.g. supplemental feeding, vaccinations).

doi:10.1371/journal.pone.0153004.t001

most marginally in the formation of the modern mtDNA gene pools of these breeds; this is in agreement with the scenario that the introgression from the Arabian horse was stallion-mediated.

The overall sequence alignment of Italian samples revealed 91 polymorphic sites (S), represented by 90 transitions and three indels (two deletions at nps 15532 and 15868, and one insertion at np 16063; we found also a transition at nps 15868 and 16063) (Table 2).

Nucleotide diversity (π) across all Italian horses was estimated at 0.020. Haplotype diversity was also very high (Hd = 0.979), confirming what already seen in previous horse mtDNA studies [8, 29, 31, 32, 35]. We detected the highest haplotype diversity in the Maremmano horse (Hd = 0.980), followed by the Sardinian Anglo-Arab (Hd = 0.970). The lowest value (Hd = 0.796) was registered in the Monterufolino breed.

The analysis of molecular variance (AMOVA) established that the majority of the observed variance is attributable to differences among samples within breeds (93.57%). However, the remaining among-breeds' component of genetic variation (6.43%) could be associated with a significant value of the fixation index ($\Phi_{ST} = 0.064$, *p-value* < 0.001). We examined different possible structures by establishing and comparing different population groups, which were artificially created by considering various features in turn, such as: breeding conditions (semi-feral *vs* controlled); height at the withers (ponies *vs* others); geographic prevalence (e.g. indigenous of Sardinia *vs* others). Actually, the only significant sign of genetic differentiation was found between the two local Sardinian breeds (Giara and Sarcidano) and the other breeds (Table 3), particularly when considering Monterufolino as a third independent group ($\Phi_{CT} = 0.063$, *p-value* < 0.001).

Table 2. Estimates of genetic diversity[a].

Breed	CODE	N	π	Nh	Hd	S
Bardigiano	BA	48	0.0197	24	0.957	59
Monterufolino	MF	30	0.0220	7	0.796	49
Maremmano	MA	90	0.0206	53	0.980	75
Tolfetano	TF	26	0.0202	18	0.966	46
Lipizzan	LI	30	0.0214	11	0.916	46
Esperia	ES	28	0.0174	8	0.869	37
Murgese	MG	32	0.0219	21	0.968	55
Sardinian Anglo-Arab	AA	54	0.0192	31	0.970	49
Giara	GI	39	0.0166	9	0.888	33
Sarcidano	SC	30	0.0174	8	0.839	33
Italian Breeds		**407**	**0.0200**	**116**	**0.979**	**91**
Arabian horse	AR	36	0.0175	14	0.881	44
Total		**443**	**0.0200**	**126**	**0.981**	**93**

[a] N = number of analyzed samples; π = nucleotide diversity; Nh = number of haplotypes; Hd = haplotype diversity; S = number of polymorphic sites.

doi:10.1371/journal.pone.0153004.t002

This is consistent with the genetic distances between populations: Monterufolino is genetically the most distant breed, while Giara and Sarcidano are confirmed as the most closely related (S1 Fig; pairwise distances above diagonal and Nei's distances below diagonal).

Phylogenetic analyses and haplogroup classification

The reconstructed network of the control-region sequences (Fig 2) clearly defines some major branches corresponding to the horse haplogroups identified so far [19].

The haplogroup classification was confirmed and refined through an accurate analysis of diagnostic mutational motifs identified in the control-region haplotypes (S1 Table). As expected, the Przewalski's specific haplogroup F was absent in our batch of domestic horses. The stochastic distribution of our haplotypes among the remaining 17 haplogroups confirms that it is not possible to identify breed-specific mitochondrial clades, at least at this level of resolution. About one fourth (N = 109) of the 407 Italian samples carries the haplogroup L mutational motif (nps 15494, 15495 and 15496), which was often reported as the most common in a wide range of Italian (Bardigiano, Giara, Haflinger, Italian Heavy Draught, Italian Trotter, Lipizzan, Maremmano, Murgese, Sanfratellano, Sarcidano, Sicilian Indigenous and Ventasso horse) and Western Eurasian breeds [6, 8, 19, 29–32, 36–38]. Haplogroup L is also the most common in seven Italian breeds analyzed in this work, while it is absent among the Arabian samples (Table 4).

Table 3. Hierarchical AMOVA table.

Source of variation	df	Variance component	Variance (%)	Fixation index[a]	P-value[b]
Between areas (Sardinia vs others)	1	0.354	5.18	$\Phi_{CT} = 0.052$	0.029*
Among populations within areas	8	0.311	4.54	$\Phi_{SC} = 0.048$	0.000***
Within populations	397	6.172	90.28	$\Phi_{ST} = 0.097$	0.000***

[a] Φ_{CT} = variation among groups divided by total variation, Φ_{SC} = variation among sub-groups divided by the sum of variation among sub-groups within groups and variation within sub-groups, Φ_{ST} = the sum of variation groups divided by total variation.

[b] ns = $P > 0.05$

* = $P \leq 0.05$

*** = $P \leq 0.001$.

doi:10.1371/journal.pone.0153004.t003

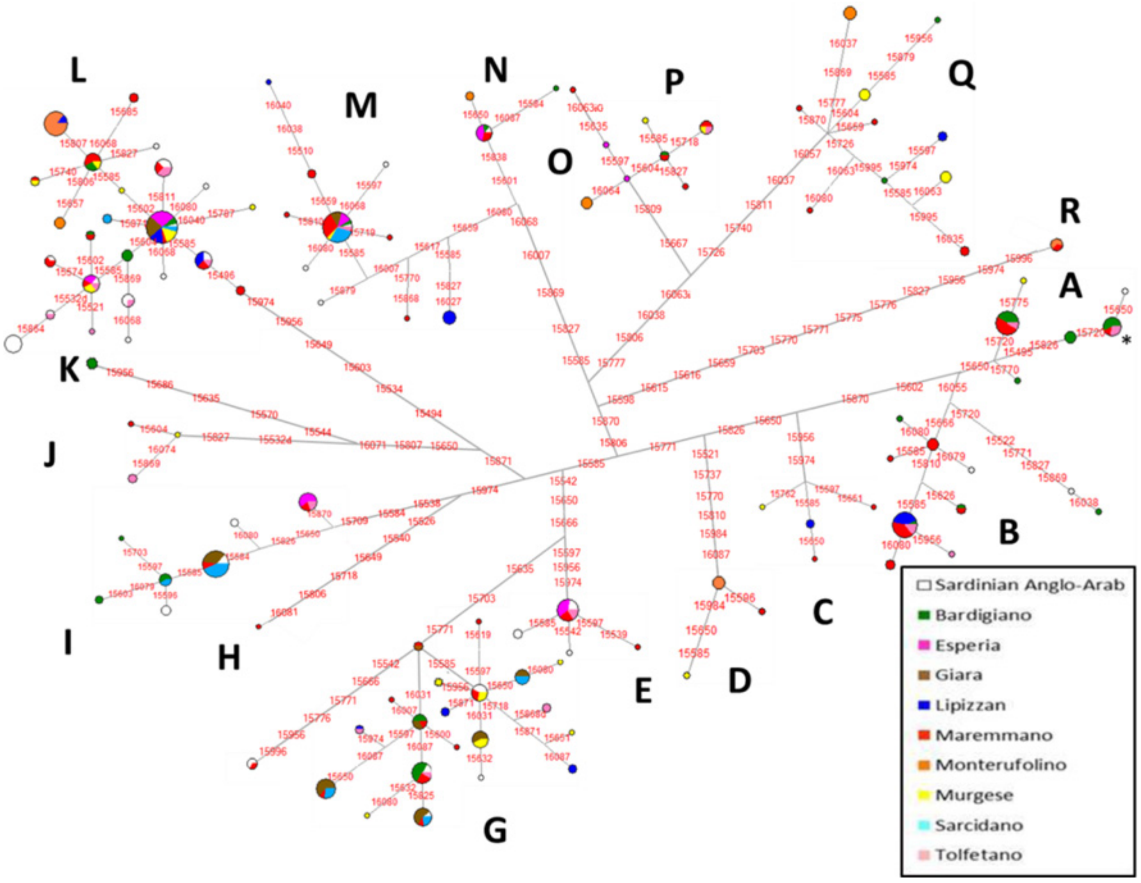


Fig 2. Median-Joining Network based on control-region sequences of ten Italian horse breeds. The asterisk indicates the haplotype identical to ERS.

doi:10.1371/journal.pone.0153004.g002

The second most common haplogroup was G (19.4%) with the highest values in Giara (56.4%) and Sarcidano (26.7%), followed by I (11.3%), which peaks in Sarcidano (40.0%),

Table 4. Haplogroup frequencies (%) in ten Italian breeds and Arabian horses[a].

Breed	A	B	C	D	E	G	H	I	J	K	L	M	N	O	P	Q	R	N Hg[a]
Bardigiano	27.1	8.3				18.8		12.5		4.2	16.7	2.1	4.2		2.1	4.2		**10**
Monterufolino				1.0							53.3		6.7	1.0		13.3	6.7	**6**
Maremmano	7.7	14.3	2.2	1.1	3.3	15.6	1.1	4.4	1.1		20.9	13.2	2.2	1.1	4.4	5.5	1.1	**16**
Tolfetano	11.5	11.5			7.7	15.4		7.7	7.7		30.8	3.8			3.8			**9**
Lipizzan		2.0	6.7			16.7					3.0	2.0				6.7		**6**
Esperia					17.9			21.4			32.1	10.7	10.7	7.1				**6**
Murgese	3.1		3.1	3.1	3.1	25.0			3.1		31.3	3.1			6.3	18.8		**10**
Sardinian Anglo-Arab	1.9	3.7			11.1	16.7		14.8			42.6	7.4	1.9					**8**
Giara						56.4		20.5			15.4	7.7						**4**
Sarcidano						26.7		40.0			10.0	23.3						**4**
Italian Breeds	6.1	6.9	1.2	1.2	4.2	19.4	0.2	11.3	1.0	0.5	27.2	9.3	2.5	1.5	2.0	4.7	0.7	**17**
Arabian horse		13.9	2.8			19.4		13.9	2.8			2.8	5.6	27.8	5.6	5.6		**10**
Total	5.6	7.4	1.4	1.1	3.8	19.4	0.2	11.5	1.1	0.5	25.0	8.8	2.7	3.6	2.3	4.7	0.7	**17**

[a] Total number of haplogroups for each breed. Haplogroup affiliation is according to Achilli et al. [19].

doi:10.1371/journal.pone.0153004.t004

followed by Giara (20.5%) and Esperia ponies (21.4%). According to the literature, haplogroups G and I should be more common in Asia and the Middle East, respectively [19]. The highest number of haplogroups was identified in the Maremmano breed (N = 16), followed by Bardigiano (N = 10) and Murgese (N = 10). As for the "insular" stocks, Giara and Sarcidano present only the major haplogroups (G, I, L, and M), while Sardinian Anglo-Arab displays a wider range of haplogroups, including A (1.9%), B (3.7%), E (11.1%) and N (1.9%). These data confirm the close genetic relationships among the Sardinian horse populations, especially between the Sarcidano and Giara breeds that share the same haplogroups and often the same haplotypes, as displayed in the presented network (Fig 2). Such a reconstructed network, based only on local Italian breeds and control-region data, allowed to date the mtDNA haplogroups to very ancient times (Table 5).

In order to graphically display (and summarize) the mitochondrial relationships among the analyzed breeds, we performed a principal component analysis (PCA)–a method that considers each haplogroup as a discrete variable and allows a summary of the initial dataset into principal components (PCs). After variables reduction to PCs (haplogroup frequencies based on different haplotypes, S2 Table), the coordinates of the observations for the eleven populations were reported in a two-dimensional plot representing the horse genetic landscape of Italy (Fig 3).

The outlier position of Monterufolino is confirmed particularly along the first PC, while the second PC splits the Arabian horses from the other breeds. Moreover, Sardinian breeds clearly separate from Italian ones as also shown by the centroids (the centroid is the geometric center of a two-dimensional shape, as depicted here by breeds typical of a certain macro-geographic area, and it is calculated as the arithmetic average position of all points/breeds). It is well known that the mtDNA inheritance might be influenced by major stochastic processes, which in turn can be amplified by local bottlenecks and founder effects. Actually, the gene pools of geographically isolated populations are dramatically shaped by initial founding events (particularly in a uniparental system such as the mtDNA) that usually lead to low level of within-population genetic distances, as those reported for Giara and Sarcidano by both the PCA and the AMOVA (Table 3), in agreement with some previous studies [31]. The ostensible partial

Table 5. MtDNA haplogroup ages based only on Italian control-region data.

Hg	N	N haplotypes	Rho estimate	Sigma	T (ka)	ΔT
A	25	6	1.84	1.02	10.2	5.6
B	33	10	2.96	1.35	16.4	7.5
C	6	5	1.40	0.72	7.8	4.0
D	5	3	0.80	0.40	4.4	2.2
E	17	4	0.29	0.16	1.6	0.9
G	86	21	2.51	0.22	13.9	1.2
H	1	1	n.a.	n.a.	n.a.	n.a.
I	51	10	3.00	1.38	16.6	7.6
J	5	4	1.25	0.75	6.9	4.2
K	2	1	n.a.	n.a.	n.a.	n.a.
L	111	25	3.49	0.18	19.3	1.0
M	39	11	3.29	1.48	18.2	8.2
N	12	3	0.40	0.24	2.2	1.4
O	16	4	1.00	0.62	5.5	3.5
P	10	6	0.75	0.53	4.2	2.9
Q[a]	21	13	2.95	0.52	16.3	2.9

[a] Calculated by hand.

doi:10.1371/journal.pone.0153004.t005

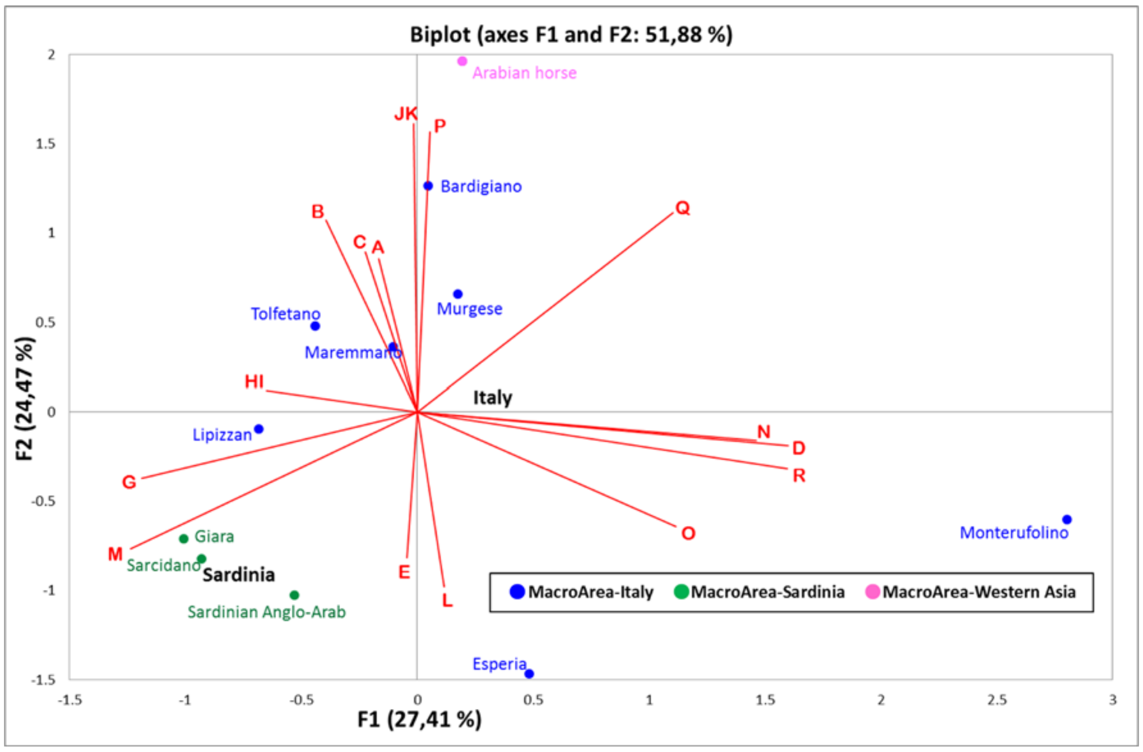


Fig 3. A two-dimensional breed-based bi-plot of mtDNA haplogroup profiles (S2 Table) from the eleven breeds analyzed in this study. The rarest haplogroups (with overall frequencies ≤ 0.5%) H and K were phylogenetically grouped with the corresponding sister clades I and J, respectively. The geographic labels, indicated in bold, represent the centroids of breeds typical of Italy (in blue) and Sardinia (in green).

doi:10.1371/journal.pone.0153004.g003

disagreement with the results reported by Morelli et al. [29], which considered Giara and Sarcidano as two distinct gene pools, could reside in the absence of two of the four haplogroups (I and M) shared by our Giara and Sarcidano samples. Moreover, we identified six different haplotypes shared by Giara and Sarcidano horses (one restricted only to these two breeds), which sum up to 84% of total samples (58 out of 69; S1 Table and Fig 2).

In order to determine whether the overall haplogroup frequencies in the Italian horse populations were indeed different from those of other populations worldwide, we repeated the PCA by including other GenBank data (S3 and S4 Tables). The overall plot, depicted by PCs 1 and 2 (Fig 4) confirms the outlier position of Monterufolino and the Sardinian horses, but at the same time highlights an overall geographic pattern from Northern Europe to Eastern Asia, as shown by the centroids position of each macrogeographic area.

The Italian breeds stand in an intermediate position together with most of the other Mediterranean stocks. The only notable exceptions are represented by the Bardigiano, which shows possible influences from Northern Europe, and particularly by the Murgese that seems to be closely related to the Asian breeds.

The mtDNA peculiarities of some Italian breeds

A strong founder effect is evident in Monterufolino, the only Italian breed with a haplotype diversity lower than 0.8 and placed in an outlier position in both the Italian and the Eurasian population contexts (Figs 3 and 4). Such a peculiar gene pool could be easily connected to the breed’s history. In the nineties, its total population counted less than ten individuals [34] and we were able to identify the considerable number of seven distinctive founding mares.

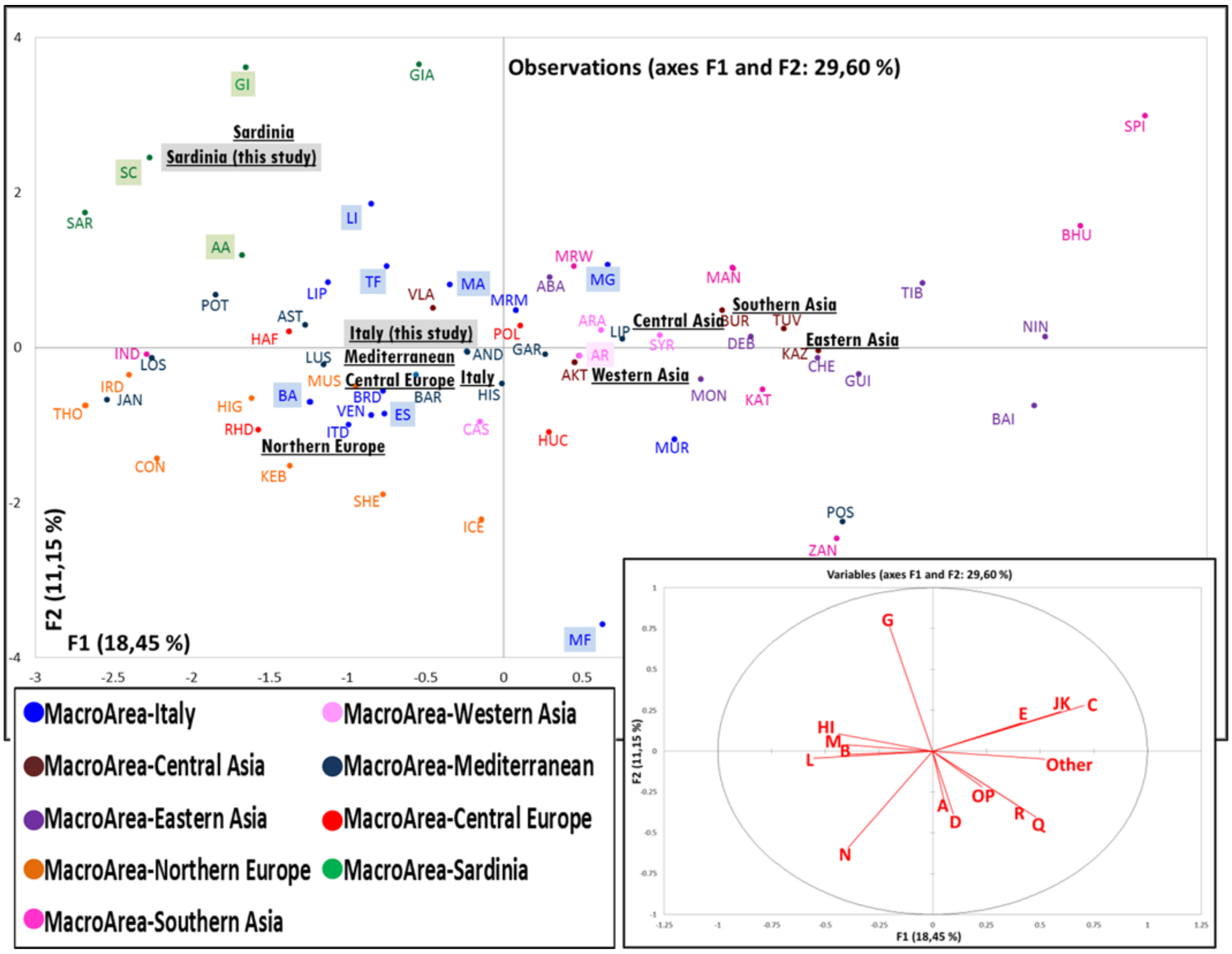


Fig 4. A two-dimensional region-based PCA plot obtained by including the available horse mtDNA data (S3 and S4 Tables). The eleven breeds analyzed in this study (and corresponding macroareas) are highlighted. The macrogeographic labels, indicated in bold and underlined, represent the centroids of breeds from the area. Only those breeds with at least 15 different haplotypes were considered statistically significant and included in the final PC analysis. Below is the plot of the contribution of each haplogroup to the first and second PC (projections of the axes of the original variables).

doi:10.1371/journal.pone.0153004.g004

The PCA analysis also revealed a peculiar localization of the Bardigiano pony within a Northern European genetic context, which was never reported in previous analyses (Sabbioni et al. 2005) [39]. This uniqueness among the Italian breeds could be explained by both its phenotype and its history. The Bardigiano is considered indigenous of Italy [34], but its origin could be traced back to the horses ridden by northern invaders during their incursions into the Italian Peninsula in the V century [40]. This original maternal legacy survived the recent dilution process due to the introduction of a diverse range of stallions from various breeds after World War II, especially *Franches Montagnes*.

Another peculiar position among Western Asian breeds is occupied by the Murgese horse, an ancient breed originated in Apulia during the Spanish domination (XVI-XVIII centuries). It is thought that the breed was developed by crossing a Spanish stock (partially Arab) with native horses, which share the same origin with the Neapolitan horse. Afterward a strict selection began in the early nineties and probably some matrilines from abroad were introduced. We identified 21 different haplotypes from the 46 presumed founding mares and based on our data they were mostly brought from Asia.

A further interesting finding is the clear separation between the Lipizzan horses from Italy and those from abroad (Fig 4). The Lipizzan breed dates back to the XVI century, when it was bred at Lipica (now in Slovenia). In the following centuries several maternal lines have been developed from eight traditional Lipizzan studs [4, 41]. Strict breeding rules were followed to keep separate different genetic reserves as demonstrated from the above mentioned peculiar PCA position of the Lipizzan horses from the Italian breeding farm of Monterotondo, whose eleven founding maternal lines are completely represented by the eleven different haplotypes reported in S1 Table.

Conclusion

Besides confirming a widespread mitochondrial variability in Italy, as already reported [29, 31, 32], this study provides a more comprehensive reassessment of the mitochondrial genetic relationships among ten typical Italian hotblood/warmblood horse and pony breeds. The different mtDNA haplotypes are not preferentially distributed among breeds. The only significant haplotype-based population structure was recognized when considering as a possible differentiation factor the (geographic) isolation of the Monterufolino and Sardinian breeds. The same four haplogroups were identified in the Giara and Sarcidano breeds (often along with the same haplotypes), whose mitochondrial similarities were confirmed in a wider Eurasian context through the PC analysis. The outcoming mtDNA genetic landscape of Eurasia shows a clear geographic pattern and highlights a group of closely related intermediate breeds mostly from the Italian Peninsula. This genetic feature likely reflects the geographic position of Italy, in the center of the Mediterranean Sea, and its cultural/economic past as a crossroad of migratory waves from the Western Asian coasts to Continental Europe. It is worth nothing that Italian breeds show a frequency of haplogroup L (23.9%) which is intermediate between those recorded in Western Asia (18.1%) and in Continental Europe (31.1%) (S5 Table). Moreover, an additional clue of a putative east-west direction of the gene flow is given by the overall haplogroup frequencies of Italian horses, which are somehow more similar to the breeds from South-West Asia (χ^2: 27.5; *p-value*: 0.006) than to those from Continental Europe (χ^2: 74.8; *p-value*: <0.001), as already indicated [32]. These findings probably reflect the overall mtDNA legacy of the ancestral mares (of eastern origins) that long time ago (see age estimates in Table 5) were probably used at the initial stages of breeding selections. Those mitochondrial lineages were also preserved during the final establishment of pure breeds that was mainly reached through sex-biased breeding practices [42], which often involved the intensive use of few selected external stallions [43, 44]. Thus, the impact on the original mtDNA gene pool could have been marginal, as also testified by the only four haplotypes shared between the Arabian horses and the ten Italian breeds here analyzed in spite of the well-recognized use of the Arabian stallions to revitalize some Italian breeds. As for the recent times, our mtDNA data lend also genetic support to some historical theories about the origin of some Italian breeds.

In conclusion, we confirm that the mitogenome is an appropriate resource in studies aiming to reconstruct the maternal ancestral origins of local breeds and to evaluate genetic continuity with the original stocks.

Materials and Methods

Ethics statement

All experimental procedures were reviewed and approved by the Animal Research Ethics Committee of the Universities of Perugia and Pavia in accordance with the European Union Directive 86/609.

Sample collection

DNA was extracted from 1,2 ml of peripheral blood samples of 407 specimens belonging to ten Italian native breeds: Sardinian Anglo-Arab (Anglo-Arabo Sardo, AA; n = 54), Bardigiano (BA; n = 48), Esperia (ES; n = 28), Giara (GI; n = 39), Lipizzan (Lipizzano, LI; n = 30), Maremmano (MA; n = 90), Monterufolino (MF; n = 30), Murgese (MG; n = 32), Sarcidano (SC; n = 30), Tolfetano (TF; n = 26). Also 36 samples of Arabian horses were included (Arabo, AR; n = 36). Horses were sampled from different Italian regions: Emilian Appennines, Latium, Apulia, Tuscany, Sardinia (Fig 1). Overall, 266 were females, 112 were males and six were geldings; no gender information was available for 59 specimens.

For the ten Italian breeds analyzed in this study, genealogical data are recorded in Studbooks (Bardigiano, Lipizzan, Maremmano, Murgese and Sardinian Anglo-Arab) or Anagraphic Registers (Esperia, Giara, Monterufolino, Sarcidano and Tolfetano). Genealogical information was considered, when available (i.e. for Lipizzan, Sardinian Anglo-Arab, Maremmano and Murgese), in order to select unrelated animals, while all other breeds (Bardigiano, Esperia, Giara, Monterufolino, Sarcidano and Tolfetano) were randomly sampled.

Total DNA was extracted from blood samples by automated extraction using the Mag-Core® Automated Nucleic Acid Extractor, following the provided protocol.

PCR amplification and sequencing of the mtDNA control region

For all animals, the mtDNA region comprised between nps 15364 and 563 was amplified by using the following oligonuclotides: forward 5'-AAACCAGAAAAGGGGAAAA-3'; reverse 5'-TGGCGAATAGCTTTGTTGTG-3'. Oligonucleotides were designed employing the GenBank published Equine Reference Sequence (ERS) NC001640 (derived from X79547) [45]. The PCR fragment of 1192 bp encompassing the entire mtDNA control region (15469–16660) was purified using exonuclease I and alkaline phosphatase (ExoSAP-IT® enzymatic system-USB Corporation, Cleveland, OH, USA) and then sent to BMR-Genomics srl (www.bmrgenomics.com) for Sanger sequencing with the primer forward 5'-CACCCAAAGCTGAAATTCTA-3'.

Mitochondrial DNA sequence analyses

Sequences (610 bps from np 15491 to np 16100) were assembled and aligned to ERS using Sequencher™ 5.10 (Gene Codes Corporation). Whenever electropherograms showed ambiguities, new PCR amplifications and sequencing reactions were performed. All mtDNA D-loop sequences determined in this study were deposited in GenBank with accession numbers KU711082-KU711507.

Several mtDNA sequence variation parameters were estimated by using DnaSP 5.1 software [46]. Analysis of MOlecular VAriance (AMOVA) and pairwise Fst calculations were performed using the Arlequin v. 3.5 software package [47]. The statistical significance of the values was estimated by permutation analysis using 100 replications. Intra- as well as inter-population comparisons were performed based on the number of pairwise differences between sequences and figured using an Arlequin integrated R script (http://www.rproject.org/).

The evolutionary relationships among haplotypes were visualized through the construction of different median-joining networks using Network 4.6 (www.fluxusengineering.com), one for each haplogroup (C, D, E, G, L, Q, and R) and macro-haplogroup (A'B, H'I, J'K, M'N, and O'P), then parsimoniously connected by hand according to mutational diagnostic motifs identified by Achilli et al. [19]. The evolutionary distances were computed as averaged distance (ρ) of the haplotypes within a clade from the respective root haplotype, accompanied by a heuristic estimate of SE (σ). All positions containing gaps and ambiguous data were eliminated from the dataset. Estimate of the time to the most recent common ancestor for each cluster was calculated using a corrected age estimate of about 2.96 x 10^{-7} per nucleotide per year in the whole control region [19], which corresponds to 5,540 years per substitution over the sequenced region of 610 bps.

Principal component analyses (PCA) were performed using Excel software implemented by XLSTAT, as described elsewhere [48]. Two PCA were carried out one by considering only our sample; the other by including the available horse mtDNA records obtained from GenBank. The PCA is a widely used dimension-reduction method which seeks to explain the variance of multivariate data by a smaller number of variables (the principal components, PCs), which are linear functions of the original variables, which in this case are the haplogroup frequencies. Considering the high degree inbreeding, which mostly characterizes common selection strategies, the haplogroup frequencies used as source data for the PCA were calculated by considering only different haplotypes within the same breed. The rarest haplogroups were phylogenetically grouped and among the large plethora of available data, only those represented by at least 15 different haplotypes were included in the analysis in order to increase the statistical significance. After having reduced the variables (haplogroups) to PCs, we reported the coordinates of the observations (breeds here and elsewhere analyzed) in two-dimensional graphics representing the genetic landscape of Italy and West Eurasia.

Supporting Information

S1 Fig. Plot of pairwise population genetic distances obtained by the concomitant analysis of all Italian breeds. Breed code as in Table 2.
(TIF)

S1 Table. Control-region haplotypes and haplogroup classification of the 443 horse mtDNAs from Italian breeds (n = 407) and Arabian horses (n = 36).
(XLSX)

S2 Table. Source data for the PCA of ten Italian breeds and Arabian horses here analyzed. Haplogroup frequencies are calculated on different haplotypes.
(XLSX)

S3 Table. Source data for the PCA of Eurasian breeds. Haplogroup frequencies are calculated on different haplotypes.
(XLSX)

S4 Table. A summary of the available horse mtDNA data.
(DOCX)

S5 Table. A geographic comparison of haplogroup frequencies (%).
(DOCX)

Acknowledgments

We thank Edoardo Battista, Riccardo Bozzi, Annamaria Cito, Fabio Frattini, Francesca Martuzzi, and Elisa Pieragostini for their help in providing biological samples.

This work has been possible thanks to a research agreement between “Dipartimento di Chimica, Biologia e Biotecnologie (University of Perugia)” and “Centro di Studio del Cavallo Sportivo (University of Perugia)”.

Author Contributions

Conceived and designed the experiments: AA MS. Performed the experiments: IC HL AG MRC SC. Analyzed the data: IC HL AG MRC AO AA. Contributed reagents/materials/analysis tools: LB GPB RC AO KC AA MS. Wrote the paper: IC HL AG SC EA KC AA MS.

References

1. Dovc P, Kavar T, Solkner H, Achmann R. Development of the Lipizzan horse breed. Reprod Domest Anim. 2006; 41(4):280–5. doi: 10.1111/j.1439-0531.2006.00726.x PMID: 16869882.
2. Hill EW, Bradley DG, Al-Barody M, Ertugrul O, Splan RK, Zakharov I, et al. History and integrity of thoroughbred dam lines revealed in equine mtDNA variation. Anim Genet. 2002; 33(4):287–94. PMID: 12139508
3. Jansen T, Forster P, Levine MA, Oelke H, Hurles M, Renfrew C, et al. Mitochondrial DNA and the origins of the domestic horse. Proc Natl Acad Sci U S A. 2002; 99(16):10905–10. doi: 10.1073/pnas.152330099 PMID: 12130666; PubMed Central PMCID: PMC125071.
4. Kavar T, Brem G, Habe F, Solkner J, Dovc P. History of Lipizzan horse maternal lines as revealed by mtDNA analysis. Genet Sel Evol. 2002; 34(5):635–48. doi: 10.1051/gse:2002028 PMID: 12427390; PubMed Central PMCID: PMC2705438.
5. Lopes MS, Mendonca D, Cymbron T, Valera M, da Costa-Ferreira J, Machado Ada C. The Lusitano horse maternal lineage based on mitochondrial D-loop sequence variation. Anim Genet. 2005; 36 (3):196–202. doi: 10.1111/j.1365-2052.2005.01279.x PMID: 15932397.
6. Royo LJ, Alvarez I, Beja-Pereira A, Molina A, Fernandez I, Jordana J, et al. The origins of Iberian horses assessed via mitochondrial DNA. J Hered. 2005; 96(6):663–9. doi: 10.1093/jhered/esi116 PMID: 16251517.
7. McGahern A, Bower MA, Edwards CJ, Brophy PO, Sulimova G, Zakharov I, et al. Evidence for biogeographic patterning of mitochondrial DNA sequences in Eastern horse populations. Anim Genet. 2006; 37(5):494–7. doi: 10.1111/j.1365-2052.2006.01495.x PMID: 16978180.
8. Moridi M, Masoudi AA, Vaez Torshizi R, Hill EW. Mitochondrial DNA D-loop sequence variation in maternal lineages of Iranian native horses. Anim Genet. 2013; 44(2):209–13. doi: 10.1111/j.1365-2052.2012.02389.x PMID: 22732008.
9. Khanshour AM, Cothran EG. Maternal phylogenetic relationships and genetic variation among Arabian horse populations using whole mitochondrial DNA D-loop sequencing. BMC Genet. 2013; 14:83. doi: 10.1186/1471-2156-14-83 PMID: 24034565; PubMed Central PMCID: PMC3847362.
10. Bowling AT, Del Valle A, Bowling M. A pedigree-based study of mitochondrial D-loop DNA sequence variation among Arabian horses. Anim Genet. 2000; 31(1):1–7. PMID: 10690354
11. Głażewska I. Speculations on the origin of the Arabian horse breed. Livest Sci. 2010; 129:49–55.
12. Cothran EG, Juras R, Macijauskiene V. Mitochondrial DNA D-loop sequence variation among 5 maternal lines of the Zemaitukai horse breed. Genet Mol Biol. 2005; 28(4):677–81.
13. Glazewska I, Wysocka A, Gralak B, Sell J. A new view on dam lines in Polish Arabian horses based on mtDNA analysis. Genet Sel Evol. 2007; 39(5):609–19. doi: 10.1051/gse:2007025 PMID: 17897600.
14. Guastella AM, Zuccaro A, Criscione A, Marletta D, Bordonaro S. Genetic analysis of Sicilian autochthonous horse breeds using nuclear and mitochondrial DNA markers. J Hered. 2011; 102(6):753–8. doi: 10.1093/jhered/esr091 PMID: 21914666.
15. Ivanković A, Ramljak J, Konjačić M, Kelava N, Dovč P, Mijić P. Mitochondrial D-loop sequence variation among autochthonous horse breeds in Croatia. Czech J Anim Sci. 2009; 54(3):101–11.
16. Achilli A, Olivieri A, Pellecchia M, Uboldi C, Colli L, Al-Zahery N, et al. Mitochondrial genomes of extinct aurochs survive in domestic cattle. Curr Biol. 2008; 18(4):R157–8. doi: 10.1016/j.cub.2008.01.019 PMID: 18302915.
17. Colli L, Lancioni H, Cardinali I, Olivieri A, Capodiferro MR, Pellecchia M, et al. Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. BMC genomics. 2015; 16 (1):1115. doi: 10.1186/s12864-015-2342-2 PMID: MEDLINE:26714643.
18. Lancioni H, Di Lorenzo P, Ceccobelli S, Perego UA, Miglio A, Landi V, et al. Phylogenetic relationships of three Italian merino-derived sheep breeds evaluated through a complete mitogenome analysis. PLoS One. 2013; 8(9):e73712. doi: 10.1371/journal.pone.0073712 PMID: 24040036; PubMed Central PMCID: PMCPMC3767607.
19. Achilli A, Olivieri A, Soares P, Lancioni H, Hooshiar Kashani B, Perego UA, et al. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc Natl Acad Sci U S A. 2012; 109(7):2449–54. doi: 10.1073/pnas.1111637109 PMID: 22308342; PubMed Central PMCID: PMCPMC3289334.
20. Vilstrup JT, Seguin-Orlando A, Stiller M, Ginolhac A, Raghavan M, Nielsen SC, et al. Mitochondrial phylogenomics of modern and ancient equids. PLoS One. 2013; 8(2):e55950. doi: 10.1371/journal.pone.0055950 PMID: 23437078; PubMed Central PMCID: PMCPMC3577844.

21. Lippold S, Knapp M, Kuznetsova T, Leonard JA, Benecke N, Ludwig A, et al. Discovery of lost diversity of paternal horse lineages using ancient DNA. Nat Commun. 2011; 2:450. doi: 10.1038/ncomms1447 PMID: 21863017.

22. Warmuth V, Eriksson A, Bower MA, Barker G, Barrett E, Hanks BK, et al. Reconstructing the origin and spread of horse domestication in the Eurasian steppe. Proc Natl Acad Sci U S A. 2012; 109(21):8202–6. doi: 10.1073/pnas.1111122109 PMID: 22566639; PubMed Central PMCID: PMCPMC3361400.

23. Outram AK, Stear NA, Bendrey R, Olsen S, Kasparov A, Zaibert V, et al. The earliest horse harnessing and milking. Science. 2009; 323(5919):1332–5. doi: 10.1126/science.1168594 PMID: 19265018.

24. Bigi D, Zanon A. Atlante delle razze autoctone italiane: Bovini, Equini, Ovicaprini, Suini allevati in Italia. Milano2008.

25. Pieragostini E, Rizzi R, Bramante G, Perrotta G, Caroli A. Genetic study of Murgese horse from genealogical data and microsatellites. Ital J Anim Sci. 2005; 4:197–202.

26. Felicetti M, Lopes MS, Verini-Supplizi A, Machado Ada C, Silvestrelli M, Mendonca D, et al. Genetic diversity in the Maremmano horse and its relationship with other European horse breeds. Anim Genet. 2010; 41 Suppl 2:53–5. doi: 10.1111/j.1365-2052.2010.02102.x PMID: 21070276.

27. Maretto F, Mantovani R. Genetic variability of Italian Heavy Draught Horse. Ital J Anim Sci. 2009; 8 (3):95–7

28. Bigi D, Perrotta G. Genetic structure and differentiation of the Italian catria horse. J Hered. 2012; 103 (1):134–9. doi: 10.1093/jhered/esr121 PMID: 22156056.

29. Morelli L, Useli A, Sanna D, Barbato M, Contu D, Pala M, et al. Mitochondrial DNA lineages of Italian Giara and Sarcidano horses. Genet Mol Res. 2014; 13(4):8241–57. doi: 10.4238/2014.October.20.1 PMID: 25366719.

30. Zuccaro A, Bordonaro S, Guastella AM, Longeri M, Cozzi MC, Guastella AM, et al. Mitochondrial DNA control region variation in Sanfratellano horse and two other Sicilian autochthonous breeds. Ital J Anim Sci. 2009; 8(2):180–2

31. Cozzi MC, Strillacci MG, Valiati P, Bighignoli B, Cancedda M, Zanotti M. Mitochondrial D-loop sequence variation among Italian horse breeds. Genet Sel Evol. 2004; 36(6):663–72. doi: 10.1051/gse:2004023 PMID: 15496286; PubMed Central PMCID: PMC2697199.

32. Bigi D, Perrotta G, Zambonelli P. Genetic analysis of seven Italian horse breeds based on mitochondrial DNA D-loop variation. Anim Genet. 2014; 45(4):593–5. doi: 10.1111/age.12156 PMID: 24702170.

33. http://www.aia.it/ [Web Site]. 2015.

34. http://dad.fao.org/ [Web Site]. 2014.

35. Yue XP, Qin F, Campana MG, Liu DH, Mao CC, Wang XB, et al. Characterization of cytochrome b diversity in Chinese domestic horses. Anim Genet. 2012; 43(5):624–6. doi: 10.1111/j.1365-2052.2011.02298.x PMID: 22497593

36. Kakoi H, Tozaki T, Gawahara H. Molecular analysis using mitochondrial DNA and microsatellites to infer the formation process of Japanese native horse populations. Biochem Genet. 2007; 45(3–4):375–95. doi: 10.1007/s10528-007-9083-0 PMID: 17265183.

37. Cieslak M, Pruvost M, Benecke N, Hofreiter M, Morales A, Reissmann M, et al. Origin and history of mitochondrial DNA lineages in domestic horses. PLoS One. 2010; 5(12):e15311. doi: 10.1371/journal.pone.0015311 PMID: 21187961; PubMed Central PMCID: PMC3004868.

38. Alvarez I, Fernandez I, Cuervo M, Martin D, Lorenzo L, Goyache F. Short communication. Mitochondrial DNA diversity of the founder populations of the Asturcón pony. Spanish Journal of Agricultural Research. 2013; 11(3):702. doi: 10.5424/sjar/2013113-4127

39. Di Stasio L, Perrotta G, Blasi M, Lisa C. Genetic characterization of the Bardigiano horse using microsatellite markers. Ital J Anim Sci. 2008; 7:243–50.

40. Bongianni M. Simon & Schuster's Guide to Horses & Ponies of the World. Simon & Schuster Building, New York: Simon & Schuster Inc., New York; 1988.

41. Zechner P, Sölkner J, Bodo I, Druml T, Baumung R, Achmann R, et al. Analysis of diversity and population structure in the Lipizzan horse breed based on pedigree information. Livest Prod Sci. 2002; 77(2–3):137–46. http://dx.doi.org/10.1016/S0301-6226(02)00079-9.

42. Vila C, Leonard JA, Gotherstrom A, Marklund S, Sandberg K, Liden K, et al. Widespread origins of domestic horse lineages. Science. 2001; 291(5503):474–7. doi: 10.1126/science.291.5503.474 PMID: 11161199.

43. Lindgren G, Backstrom N, Swinburne J, Hellborg L, Einarsson A, Sandberg K, et al. Limited number of patrilines in horse domestication. Nat Genet. 2004; 36(4):335–6. doi: 10.1038/ng1326 PMID: 15034578.

44. Wallner B, Vogl C, Shukla P, Burgstaller JP, Druml T, Brem G. Identification of genetic variation on the horse y chromosome and the tracing of male founder lineages in modern breeds. PLoS One. 2013; 8 (4):e60015. doi: 10.1371/journal.pone.0060015 PMID: 23573227; PubMed Central PMCID: PMC3616054.

45. Xu X, Arnason U. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. Gene. 1994; 148(2):357–62. PMID: 7958969

46. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11):1451–2. doi: 10.1093/bioinformatics/btp187 PMID: 19346325

47. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2005; 1:47–50

48. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, Battaglia V, et al. Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am J Hum Genet. 2007; 80(4):759–68. doi: 10.1086/512822 PMID: 17357081; PubMed Central PMCID: PMCPMC1852723.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 November 2016
doi: 10.3389/fgene.2016.00208

The Worldwide Spread of the Tiger Mosquito as Revealed by Mitogenome Haplogroup Diversity

Vincenza Battaglia[1], Paolo Gabrieli[1], Stefania Brandini[1], Marco R. Capodiferro[1], Pio A. Javier[2], Xiao-Guang Chen[3], Alessandro Achilli[1], Ornella Semino[1], Ludvik M. Gomulski[1], Anna R. Malacrida[1], Giuliano Gasperi[1], Antonio Torroni[1]* and Anna Olivieri[1]**

[1] *Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università di Pavia, Pavia, Italy,* [2] *Crop Protection Cluster, College of Agriculture, University of the Philippines Los Baños, Los Baños, Philippines,* [3] *Department of Pathogen Biology, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, China*

OPEN ACCESS

Edited by:
Badri Padhukasahasram,
Illumina, USA

Reviewed by:
Gonzalo Gajardo,
University of Los Lagos, Chile
Octavio Salgueiro Paulo,
University of Lisbon, Portugal
Luisa Pereira,
Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Portugal

***Correspondence:**
Anna Olivieri
anna.olivieri@unipv.it
Antonio Torroni
antonio.torroni@unipv.it
Giuliano Gasperi
giuliano.gasperi@unipv.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *03 September 2016*
Accepted: *09 November 2016*
Published: *23 November 2016*

Citation:
Battaglia V, Gabrieli P, Brandini S, Capodiferro MR, Javier PA, Chen X-G, Achilli A, Semino O, Gomulski LM, Malacrida AR, Gasperi G, Torroni A and Olivieri A (2016) The Worldwide Spread of the Tiger Mosquito as Revealed by Mitogenome Haplogroup Diversity. Front. Genet. 7:208. doi: 10.3389/fgene.2016.00208

In the last 40 years, the Asian tiger mosquito *Aedes albopictus*, indigenous to East Asia, has colonized every continent except Antarctica. Its spread is a major public health concern, given that this species is a competent vector for numerous arboviruses, including those causing dengue, chikungunya, West Nile, and the recently emerged Zika fever. To acquire more information on the ancestral source(s) of adventive populations and the overall diffusion process from its native range, we analyzed the mitogenome variation of 27 individuals from representative populations of Asia, the Americas, and Europe. Phylogenetic analyses revealed five haplogroups in Asia, but population surveys appear to indicate that only three of these (A1a1, A1a2, and A1b) were involved in the recent worldwide spread. We also found out that a derived lineage (A1a1a1) within A1a1, which is now common in Italy, most likely arose in North America from an ancestral Japanese source. These different genetic sources now coexist in many of the recently colonized areas, thus probably creating novel genomic combinations which might be one of the causes of the apparently growing ability of *A. albopictus* to expand its geographical range.

Keywords: *Aedes albopictus*, tiger mosquito, mitochondrial DNA, mitogenomes, haplogroups

INTRODUCTION

The genus *Aedes* includes five highly invasive species, *A. albopictus*, *A. aegypti*, *A. j. japonicus*, *A. koreicus*, and *A. atropalpus*. Of these, *A. albopictus* and *A. j. japonicus* are the most widespread across the globe, and *A. aegypti* and *A. albopictus*, being competent vectors for several human tropical diseases, have a major impact on human health. The North American species, *A. atropalpus*, arrived in Europe (Italy, France, and the Netherlands) through international trade (Romi et al., 1997), but it was subsequently exterminated in Italy and France and is unlikely to have established in the Netherlands due to climatic conditions (Scholte et al., 2009). *Aedes j. japonicus* and *A. koreicus*, native to East Asia (Japan, Korea, China, Russia), have both colonized central Europe and *A. j. japonicus* is also widely distributed in the US (Medlock et al., 2015). *Aedes aegypti* originated in sub-Saharan Africa and is now considered the main vector of dengue. Climate appears to be the decisive factor limiting the distribution of *A. aegypti* to tropical and sub-tropical regions, with few incursions into Europe and North America (Powell and Tabachnick, 2013; Khormi and Kumar, 2014).

In contrast, *A. albopictus*, indigenous to East Asia, is not so restrained by climatic factors, and in the last 40 years has successfully colonized the tropical and temperate regions of all continents (Benedict et al., 2007; Paupy et al., 2009; Bonizzoni et al., 2013; Kraemer et al., 2015). This mosquito has become a growing public health concern, being a competent vector for many arboviruses which cause lethal or debilitating human diseases, including the dengue (DEN), chikungunya (CHIK), West Nile (WN) viruses (Gasperi et al., 2012; Bonizzoni et al., 2013), and the recently emerged ZIKA virus (Wong et al., 2013; Chouin-Carneiro et al., 2016). Although it has been considered a less efficient vector than *A. aegypti*, this species is the sole vector of recent DENV outbreaks in Southern China, Hawaii, the Indian Ocean and Gabon and the first autochthonous DENV transmission in France and Croatia (Paupy et al., 2009; Wu et al., 2010; Peng et al., 2012; Rezza, 2012, 2014; Schaffner et al., 2013). *A. albopictus* was most likely the main DENV vector in Asia prior to the introduction of *A. aegypti* in the mid nineteenth century (Gubler, 2006).

Introduced into Europe (Albania) in 1979, *A. albopictus* has now colonized all Mediterranean countries from Spain to Syria and has been reported in Central Europe (Medlock et al., 2012). It was introduced into Hawaii at the end of the 19th century (Rai, 1991) and into continental USA (Texas) in 1985, and is now well established in 32 states. Its presence has been reported in Mexico (first recorded in 1988), Central and South America (first recorded in Brazil, 1986), and Africa (first recorded in South Africa, 1989; Bonizzoni et al., 2013). Its ability to spread from the native range and adapt to local environments is probably due to its ecological characteristics, drought-resistant eggs with the ability to diapause, daylight biting habit, aggressive and opportunistic feeding behavior, and capacity to achieve high population densities (Paupy et al., 2009).

The nuclear genome of the tiger mosquito from two laboratory strains, the Italian Fellini (an isofemale line derived from the Rimini strain; Bellini et al., 2007; Dritsou et al., 2015) and the Chinese Foshan strain (Chen et al., 2015), was recently published, but only two complete mitogenomes (∼16.7 kb) are available in GenBank, from Taiwan (Asian tiger mosquito Reference Sequence, NC006817) and from Nanjing, Jiangsu Province, China (KR068634; Zhang et al., 2015). Despite the availability of mitogenomes, virtually all *A. albopictus* mitochondrial DNA (mtDNA) surveys were restricted to short segments of the cytochrome c oxidase subunit 1 (COI) and/or NADH dehydrogenase subunit 5 (ND5) genes, suggesting a limited phylogeographic differentiation among populations, possibly also caused by the inclusion in these studies of laboratory stocks or sibling eggs (Delatte et al., 2011; Kamgang et al., 2011, 2013; Porretta et al., 2012; Zhong et al., 2013; Zawani et al., 2014; Futami et al., 2015) and the postulated cytoplasmatic sweep caused by *Wolbachia* infection (Armbruster et al., 2003). However, more extensive sequencing of the COI gene has revealed more variation than previously thought (Goubert et al., 2016), a scenario also partially supported by microsatellite studies that highlighted slight genetic diversity between native and adventive populations with high variability within populations (Manni et al., 2015).

Previous studies have shown that the variation seen in short mtDNA segments may be inadequate to both identify and phylogenetically link haplogroups (Torroni et al., 2006; Achilli et al., 2008, 2012). This is especially true for the insect mtDNA control region due to its peculiar features: high A+T content and reduced substitution rate, variable size and high length mutation rate, concerted evolution of tandem repeats and directional mutation pressure (Zhang and Hewitt, 1997).

To identify the ancestral source(s) of *A. albopictus* adventive populations, overcoming previous limitations, we here determined and analyzed the sequence variation at the level of entire coding regions of 27 mitogenomes (25 novel and two previously published) from Eastern and Southeastern Asian, American, and European populations. Our analyses reveal that only three of the five identified Asian haplogroups, which are differentially distributed in Asian populations living in temperate and tropical regions, were involved in the recent worldwide spread. These different ancestral sources from Asia now coexist in many adventive populations with possible implications for the adaptive capability of the species.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

A total of 25 novel mitogenomes were included in this study. Twenty-two were from wild populations collected in Europe, Asia, and the Americas (**Table 1**; **Figure 1**). Three were from the Americas (two from Virginia and one from Brazil). Nine were from Asia: three from Thailand (one from Hang Chat district, Lampang province in the North; one from Ban Rai district, Uthai Thani province in the West; one from Phato district, Chumphon province in the South), five from Los Baños, Laguna, Philippines and one from Wakayama prefecture, Japan. Ten were from Europe: two from Tirana (Albania), two from Athens (Greece), two from Cesena and two from Pavia (Northern Italy), one from Cassino (Central Italy), and one from Reggio Calabria (Southern Italy). This study also included three adult laboratory-maintained strain mosquitoes: two from the Italian Rimini strain (Bellini et al., 2007; Manni et al., 2015), established at CAA (Centro Agricoltura Ambiente "G. Nicoli," Crevalcore, Italy) from mosquitoes collected in Rimini, Italy, and one from the Chinese Foshan strain (Center for Disease Control and Prevention of Guangdong Province; **Table 1**). The study did not involve protected species and specimens were not collected at sites protected by law.

Morphological keys (Rueda, 2004) and/or PCRs with species-specific primers for internal transcribed spacer regions (ITS1 and ITS2) of ribosomal DNA (rRNA; Higa et al., 2010) were used to identify the specimens. For the Philippine samples, eggs were collected using ovitraps, and the emerging adults were reared in an insectary under standard conditions of temperature (27°C), humidity (60–80%), and photoperiod (12:12 h). Samples were preserved in 80% ethanol and stored at −20°C until

TABLE 1 | Origin and haplogroup affiliation of *A. albopictus* mitogenomes considered in this study.

Sequence ID#[a]	Original name	Continent	Country (place of collection)	Haplogroup	GenBank ID	Number of type I repeats[b]	Number of type II repeats[b]	Reference
1	Rim1[c]	Europe	Italy, Rimini	A1a1a1	KX383916	7	5	this study
2	Vir1	America	US, Virginia	A1a1a1a1	KX383917	N.D.	6	this study
3	Rc1	Europe	Italy, Reggio Calabria	A1a1a1a1	KX383918	N.D.	6	this study
4	Vir2	America	US, Virginia	A1a1a1a1	KX383919	N.D.	6	this study
5	Ces1	Europe	Italy, Cesena	A1a1a1a1	KX383920	N.D.	6	this study
6	Cas1	Europe	Italy, Cassino	A1a1a1a	KX383921	N.D.	6	this study
7	Pav3	Europe	Italy, Pavia	A1a1a1a	KX383922	N.D.	6	this study
8	Ces2	Europe	Italy, Cesena	A1a1	KX383923	N.D.	4	this study
9	Bra	America	Brazil	A1b	KX383924	N.D.	N.D.	this study
10	Lam2	Asia	Thailand, Lampang, Hang Chat	A1b1a	KX383925	N.D.	3	this study
11	Ban7	Asia	Thailand, Uthai Thani, Ban Rai	A1b1a	KX383926	N.D.	3	this study
12	Ath1	Europe	Greece, Athens	A1b1a	KX383927	N.D.	3	this study
13	Chu3	Asia	Thailand, Chumphon, Phato	A1b1	KX383928	N.D.	3	this study
14	Rim4[c]	Europe	Italy, Rimini	A1a2a1	KX383929	N.D.	4	this study
15	Tir1	Europe	Albania, Tirana	A1a2a1	KX383930	N.D.	4	this study
16	Tir2	Europe	Albania, Tirana	A1a2a1	KX383931	N.D.	4	this study
17	–	Asia	China, Jiangsu, Nanjing	A1a2a1	KR068634	5	4	Zhang et al., 2015
18	Ath2	Europe	Greece, Athens	A1a2a	KX383932	N.D.	4	this study
19	Pav4	Europe	Italy, Pavia	A1a2a	KX383933	N.D.	4	this study
20	Fo2[c]	Asia	China, Foshan	A1a2	KX383934	N.D.	4	this study
21	Los1	Asia	Philippines, Laguna, Los Baños	A2a	KX383935	N.D.	3	this study
22	Los2	Asia	Philippines, Laguna, Los Baños	A2a	KX809761	N.D.	3	this study
23	Los3	Asia	Philippines, Laguna, Los Baños	A2a	KX809762	N.D.	3	this study
24	Los5	Asia	Philippines, Laguna, Los Baños	A2a	KX809764	N.D.	3	this study
25	Los4	Asia	Philippines, Laguna, Los Baños	A2	KX809763	N.D.	3	this study
26	J-Wa1	Asia	Japan, Wakayama	A1a1a	KX809765	N.D.	4	this study
27	–	Asia	Taiwan, Taipei	A3	NC006817	5	4	–

[a] *ID numbers correspond to those in* ***Figure 1***.
[b] *N.D. = not determined.*
[c] *Laboratory-maintained strain.*

DNA extraction. DNA extraction was carried out using the Wizard Genomic DNA Purification Kit (Promega) following the manufacturer's protocol.

Sequencing of the *A. albopictus* Mitochondrial Coding and Control Regions

A primer set was designed to amplify the entire *A. albopictus* mitogenome in four overlapping PCR fragments (protocol I). The *A. albopictus* Reference Sequence (NC006817) was used to design primer sets. The coding region (nps 1-14893) was amplified by two long PCR fragments whereas the control region (nps 14894-16665) was amplified by two rather short PCR fragments (Supplementary Table 1) because of its high A+T content and the presence of repeated elements requiring distinctive PCR procedures.

Coding region long PCRs (Supplementary Table 1) were carried out in 50 μl reaction mixture containing 1X GoTaq Long PCR Master Mix (Promega), 0.2 μM of each primer and 10–20 ng of DNA template using the following PCR conditions: 94°C (2 min); 30 cycles of 94°C (30 s), 59°C (30 s), 65°C (9 min); and a final extension of 72°C (10 min). Alternatively, a set of nine overlapping PCR fragments (Supplementary Table 2) covering the *A. albopictus* coding region was also designed (protocol II). PCRs were carried out in 25 μl reaction with a standard reaction mix containing 1X White Buffer (1.5 mM $MgCl_2$), 0.2 mM of each dNTP mix, 0.6 U of GoTaq G2 Polymerase (Promega), 0.2 μM of

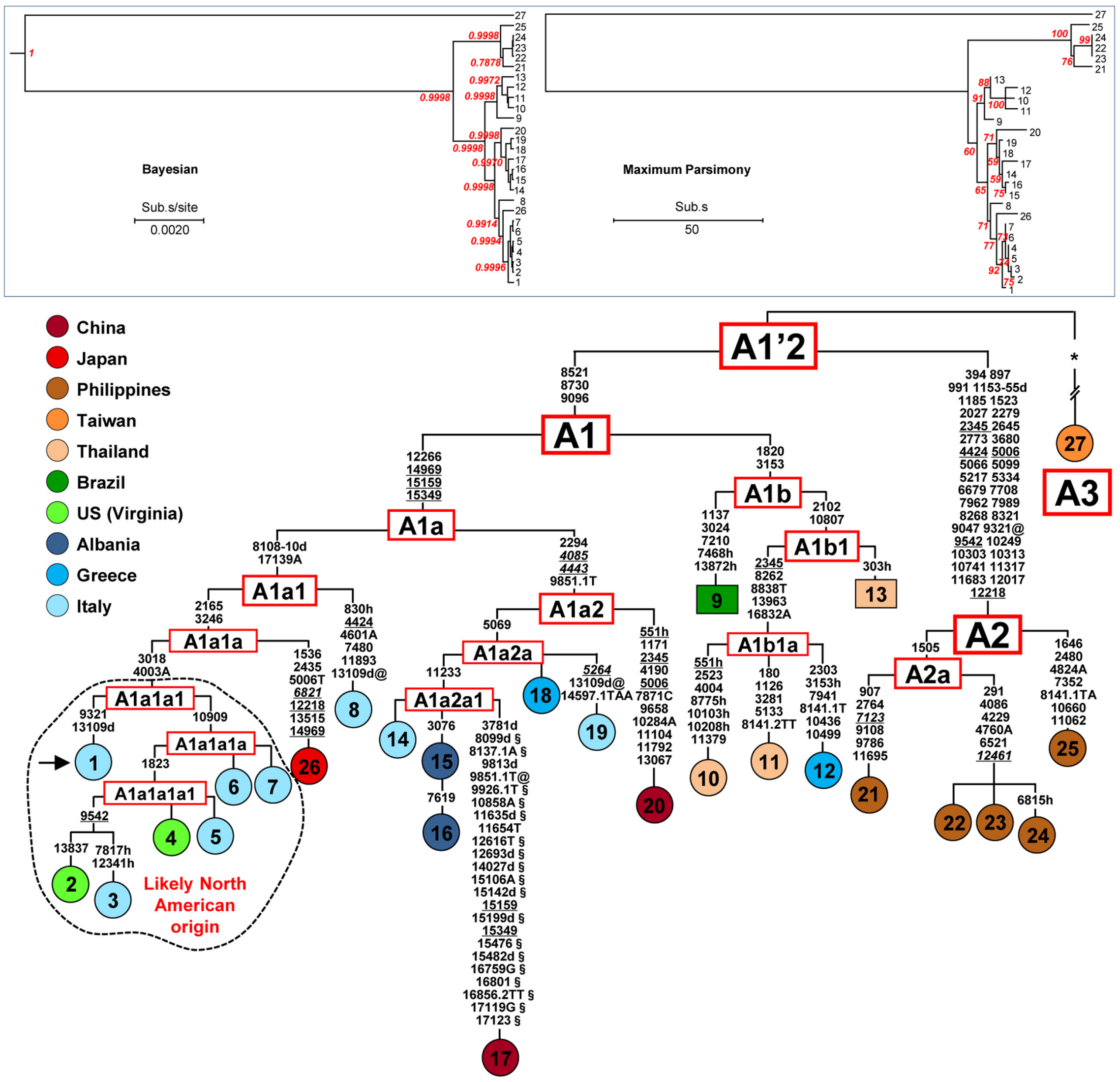

Bayesian
Sub.s/site
0.0020
1
0.9998
0.7878
0.9972
0.9998
0.9998
0.9998
0.9998
0.9970
0.9998
0.9914
0.9994
0.9996
Maximum Parsimony
Sub.s
50
100
99
76
88
91
100
71
60
59
59
65
75
71
77
73
74
92
75
China
Japan
Philippines
Taiwan
Thailand
Brazil
US (Virginia)
Albania
Greece
Italy
A1'2
8521 8730 9096
A1
12266 14969 15159 15349
A1a
8108-10d 17139A
A1a1
2165 3246
A1a1a
3018 4003A
A1a1a1
9321 13109d
1
10909
A1a1a1a
1823
A1a1a1a1
9542
13837
2
7817h 12341h
3
4
5
6
7
Likely North American origin
1536 2435 5006T 6821 12218 13515 14969
26
830h 4424 4601A 7480 11893 13109d@
8
2294 4085 4443 9851.1T
A1a2
5069
A1a2a
11233
A1a2a1
14
3076
15
7619
16
3781d 8099d § 8137.1A § 9813d 9851.1T@ 9926.1T § 10858A § 11635d § 11654T 12616T § 12693d § 14027d § 15106A § 15142d § 15159 15199d § 15349 15476 § 15482d § 16759G § 16801 § 16856.2TT § 17119G § 17123 §
17
18
5264 13109d@ 14597.1TAA
19
551h 1171 2345 4190 5006 7871C 9658 10284A 11104 11792 13067
20
1820 3153
A1b
1137 3024 7210 7468h 13872h
9
2102 10807
A1b1
2345 8262 8838T 13963 16832A
A1b1a
551h 2523 4004 8775h 10103h 10208h 11379
10
180 1126 3281 5133 8141.2TT
11
2303 3153h 7941 8141.1T 10436 10499
12
303h
13
394 897 991 1153-55d 1185 1523 2027 2279 2345 2645 2773 3680 4424 5006 5066 5099 5217 5334 6679 7708 7962 7989 8268 8321 9047 9321@ 9542 10249 10303 10313 10741 11317 11683 12017 12218
A2
1505
A2a
907 2764 7123 9108 9786 11695
21
291 4086 4229 4760A 6521 12461
22
23
6815h
24
1646 2480 4824A 7352 8141.1TA 10660 11062
25
*
27
A3

FIGURE 1 | Phylogeny of *A. albopictus* mitogenomes. The Bayesian (left) and MP (right) trees are shown in the top inset. The posterior probability for the major nodes in the Bayesian tree is reported in red, whereas bootstrap values (1000 replications) are reported in red in the MP tree. These trees encompass 25 novel and two previously published sequences (**Table 1**). A magnified MP tree is also shown reporting all mutations that characterize the 27 mitogenomes except those linking mitogenome #27 to the A1′2 node (see below). The asterisk (*) indicates the location of these mutations (347 in the coding region and 16 in control region), which are listed in Supplementary information for Figure 1. For the phylogeny construction, the entire coding region variation of all mitogenomes was included as well as some control region mutations (see "Material and Methods"). The published sequences (#17 and #27, **Table 1**) are from the Jiangsu Province, China (Zhang et al., 2015) and from Taipei, Taiwan, respectively. The mitogenome from a mosquito of the Italian Rimini strain (#1, marked by the arrow) was employed to number the mutations shown on the branches. Mutations are transitions unless a base is explicitly indicated for transversions (to A, G, C, or T) or a suffix for indels (0.1, d). Heteroplasmic positions are marked by an "h." Recurrent mutations within the phylogeny are underlined (and in italics if present in mitogenome #27) and back mutations are marked with the suffix @. The numerous mutations shared only by the published mitogenomes #17 and #27 are marked with the suffix §. Taking also into account that, despite their extensive coding region differentiation, mitogenomes #17 and #27 intriguingly harbor virtually identical control region sequences, it is likely that at least some of the mutations marked with the suffix § are mistakes. Colors illustrate geographic origins. **Table 1** provides additional information concerning the geographic origin and haplogroup affiliation of each sample. Length variation (insertions/deletions) in a poly-A stretch beginning at np 3808 was not considered. Note that for mitogenomes #9 and #13 (in squared boxes) the sequence variation was assessed only partially for the coding region and not at all for the control region (Supplementary information for Figure 1). The sub-haplogroup A1a1a1 and its derivatives of probable North American origin are encircled.

each primer and 20–30 ng of DNA template, using the following PCR conditions: 94°C (2 min); 35 cycles of 94°C (30 s), 55°C (30 s), 72°C (2 min); and a final extension of 72°C (5 min).

PCR primers used to amplify the control region in two overlapping fragments (Supplementary Table 1) were the same in both protocols. The control region PCRs were carried out using the following PCR conditions: 94°C (2 min); 35 cycles of 94°C (30 s), 54°C (30 s), 60°C (2 min); and a final extension of 60°C (5 min) for PCR #3, and 94°C (2 min); 40 cycles of 94°C (30 s), 55°C (30 s), 60°C (1 min); and a final extension of 60°C (10 min) for PCR #4. PCR products were visualized on a 1–2% agarose gel and successful amplicons were sequenced with standard dideoxy sequencing using Big Dye v3.1 Chemistry (Applied Biosystems) on 3730xl and 3130xl Genetic Analyzer (Applied Biosystems) following the manufacturer's protocol. Sets of 28 and 29 oligonucleotides were designed to sequence the *A. albopictus* mtDNA coding region starting from protocol I (Supplementary Table 3) or protocol II (Supplementary Table 4), respectively. Sequences were assembled and aligned using Sequencher 5.0 (Gene Codes) comparing them with the Reference Sequence (NC006817) from a Taipei sample, Taiwan.

Cloning and Sequencing of Mitogenome #1 (Rimini Strain) Control Region

Given the different copy numbers of repeated elements contained in the *A. albopictus* control region, PCR fragments #3 and #4 (Supplementary Table 1) yielded products of different length in different mosquitoes. The two types of tandem repeats, I and II (**Figure 2**), were amplified with PCR #3 and PCR #4, respectively. Overall, the size of the amplified fragments ranged from ∼1,800 to ∼2,500 bp for PCR #3 and from ∼800 to ∼900 bp for PCR #4. As for one of the two mitogenomes from the Rimini strain (mitogenome #1, **Figure 1**; **Table 1**), PCR #3 and #4 yielded products of ∼2,000 bp and ∼900 bp, respectively. The product of PCR #4 was directly cloned in the pCR2.1 TOPO vector (Invitrogen) following the manufacturer's protocol. White colonies were PCR-screened for the insert length and desired clones were sequenced bi-directionally using the M13 universal primers. The procedure described above is inefficient for longer fragments such as that of ∼2,000 bp, therefore a different strategy was developed to sequence the region amplified by PCR #3. It was re-amplified in two overlapping PCR fragments (PCRs I and II), each with one primer within the tandemly repeated elements of type I (Supplementary Table 5). Multiple amplicons were obtained for each PCR and were cloned as described above. Only the two clones containing the longest fragments deriving from each PCR were sequenced bi-directionally. Sequences for each clone were assembled and aligned using Sequencher 5.0 (Gene Codes). Two additional internal primers were then newly designed (data not shown) to confirm the sequence obtained by cloning.

Phylogeny Construction

The obtained complete sequence of mitogenome #1 (accession number KX383916) was used to assemble and number, with Sequencher 5.0 (Gene Codes), the other 24 novel mitogenomes (accession numbers KX383917-35, KX809761-65). We aligned the novel sequences and the two previously published mitogenomes – #17 from China (Zhang et al., 2015) and #27 from Taiwan (NC006817) – by performing a Multiple-Sequence Alignment with the Clustal algorithm (Chenna et al., 2003) implemented by Sequencher 5.0.

Most Parsimonious (MP) trees (1000 bootstrap replications) encompassing the 27 mitogenomes were built by using MEGA7 (Kumar et al., 2016), employing the Tree-Bisection-Regrafting (TBR) algorithm (Nei and Kumar, 2000). Modelgenerator v.85 indicated for our dataset HKY+I as the best-supported model according to the AIC1, AIC2, and BIC criteria. The obtained settings were selected to infer maximum likelihood (ML) and Bayesian trees for the *A. albopictus* dataset. The ML tree was built using PAMLX (Yang, 2007) and assuming the HKY85 mutation model (two parameters in the model of DNA evolution) with gamma-distributed rates (approximated by a discrete distribution with eight categories). The Bayesian tree was obtained using BEAST 1.8.3 (Drummond and Rambaut, 2007) and running 50,000,000 iterations, with samples drawn every 10,000 Markov chain Monte Carlo (MCMC) steps. It was visualized using FigTree v.1.4.2. Phylogeny reconstruction was performed considering all the nucleotide substitutions (excluding indels and heteroplasmies) in the coding region (from np 1 to np 14896, relative to mitogenome #1) and the five informative control region mutations 14969, 15159, 15349, 16832A, and 17139A. Thirteen additional control region mutations (see Supplementary information for Figure 1), ten shared exclusively by the previously published mitogenomes #17 and #27 and whose reliability is doubtful, and three (private mutations) seen only in mitogenome #27, were simply superimposed on the magnified MP tree (**Figure 1**).

Survey of Published Partial COI and ND5 mtDNA Sequences

The identification of haplogroup diagnostic mutations allowed us to survey 1170 *A. albopictus* partial COI and ND5 available in the literature for the presence/absence of these mutations. Of these, 284 mtDNAs are from Asia, 349 from the Americas, 32 from Europe, 153 from Africa, and 352 from Oceania (**Table 2**), thus encompassing populations from both native and non-native areas, and living at different climatic conditions.

RESULTS

The Variation of *A. albopictus* mtDNA Control Region

The first step in our study was to sequence the entire mtDNA (mitogenome #1) from one mosquito of the Italian (Rimini) laboratory-maintained strain (**Table 1**; **Figure 1**), whose nuclear genome was recently sequenced (Dritsou et al., 2015). This mitogenome sequence confirmed that the *A. albopictus* control region belongs to group 2 of insect control regions (Zhang and Hewitt, 1997), like that of *A. aegypti* (Dueñas et al., 2006). Three conserved blocks are positioned along the region, which contains two different types (I and II) of tandem repeats (**Figure 2**). Type I consists of a ∼190 bp repeat unit, whereas type II is made up of a short unit of ∼42 bp. The number of type I and type II tandem repeats that we observed in mitogenome #1 was different from those in sequence NC006817 from Taiwan (mitogenome #27 in **Figures 1** and **2**) and variable among mtDNAs (**Table 1** and data not shown). Moreover, between these two types of repeats, delimited by two conserved blocks, lies an A+T rich region of variable length. The overall length and tandem repeat composition make the PCR amplification and sequencing of

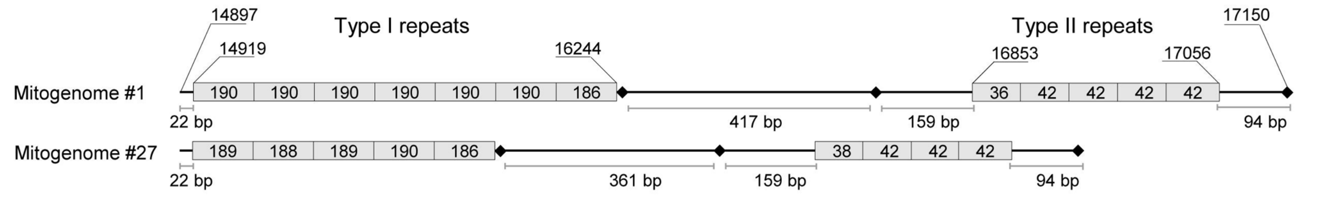


FIGURE 2 | Organization of tandem repeats in the control region of *A. albopictus* mitogenomes. This schematic representation is based on the complete sequence information of mitogenomes #1 (Rimini strain) and #27 (**Figure 1**; **Table 1**). Two types of repeats (I and II) were observed with copy number differing for both repeats in the two mitogenomes. Type I consists of a ~190 bp repeat unit whereas type II is made up of a shorter unit of ~42 bp. The exact length (in bp) of each repeat is reported inside the box. The size of boxes representing repeats is not proportional to repeat lengths. Shown nucleotide positions are relative to sequence #1. Variation in size due to variable number of tandem repeats in the two regions, when available, is provided in **Table 1**. Black diamonds represent in the order from left to right: a poly-T stretch motif, a poly-T stretch motif followed by a GC-rich block, and a short poly-A stretch motif. All of these are conserved in different mosquito species (Dueñas et al., 2006).

the entire *A. albopictus* control region extremely difficult (see "Material and Methods"). For the reasons outlined above, and difficulties originating from the impossibility of distinguishing heteroplasmy from PCR artifacts (due to replication slippage), we restricted our sequencing of *A. albopictus* mitogenomes to the coding region (from np 1 to np 14893, NC006817) and nearby control region segments. This approach was employed to obtain the coding region sequences of the additional 24 mitogenomes.

MtDNA Haplogroups in *A. albopictus*

Figure 1 illustrates the Bayesian and MP trees derived from the coding regions of the 27 *A. albopictus* mitogenomes (25 novel and two previously published). The overall tree structure is virtually identical with the two approaches and is supported by the ML tree (Supplementary Figure 1), indicating a high degree of internal consistency for all major branches. A magnified MP tree is shown in the lower part of **Figure 1** in order to illustrate the branch location of the identified mutations.

The mitogenomes (24 distinct haplotypes) cluster into three major branches that we named haplogroups A1, A2, and A3. Haplogroup A1 includes 21 mitogenomes, A2 consists of the five Philippine samples, while A3 encompasses only one mitogenome (#27) from Taiwan. Haplogroups A1 and A2 are rather close to each other and to the A1′2 node. In contrast, the A3 mitogenome differs by 363 mutations from the same node.

Overall, this phylogeny reveals an extensive and previously unreported mitogenome differentiation within *A. albopictus*. Indeed, when calculated on the standard COI sequence (Hebert et al., 2003) employed for DNA barcoding, the maximum intraspecific divergence was 0.012 (eight mutations in 658 bp), a value in line with those recently reported for *A. scutellaris* (0.008) and *A. aegypti* (0.022), but much greater than the value previously reported for *A. albopictus* (0.002; Sumruayphol et al., 2016).

Haplogroup A1, which encompasses most of the mitogenomes in the phylogeny, is subdivided into two branches that we termed A1a and A1b, with the former further split into A1a1 and A1a2. In our phylogeny these clades and subclades appear to be correlated with different geographic distributions. The branch A1a1 includes the single mitogenome from Japan, the two mitogenomes from the US (Virginia) and many of the mitogenomes from Italy, including one (#1) of the two detected in the Rimini laboratory strain. The mitogenome from Japan (#26) departs from the node A1a1a and its sister clade A1a1a1 contains a sub-branch, A1a1a1a1, of particular interest. It consists of four mitogenomes, two from Italy and the two from the US mentioned above. One of the US mitogenomes (#4) is identical to one from Northern Italy (#5) while the second (#2) is closely related to the mitogenome #3 from Southern Italy.

The sister branch A1a2 is formed by mitogenomes from different regions of Southern Europe (Italy, Albania, and Greece), including the second mitogenome (#14) from the Rimini strain, the previously published Chinese sequence [#17, KR068634 (Zhang et al., 2015)] and the mitogenome (#20) from the Chinese Foshan strain, a laboratory-maintained colony founded in 1981 from mosquitoes from Southeast China. It is worth mentioning that the presence of two distinct haplotypes in different subjects of the Rimini laboratory-maintained strain, one belonging to A1a1a1 and the second to A1a2a1, reveals that at least two females contributed to the genetic formation of the strain. Haplogroup A1b contains all the mitogenomes from Thailand, which cluster in its A1b1 sub-branch, as well as one from Greece and one from Brazil. Finally, haplogroup A2 consists of multiple haplotypes, all from the Philippines.

Haplogroup Affiliation of Worldwide mtDNA Sequences from *A. albopictus*

The phylogenetic analysis not only allowed the identification of *A. albopictus* haplogroups and sub-haplogroups but also the definition of their distinguishing mutations (**Figure 1**). These include some diagnostic markers that are located in COI and ND5 partial sequences whose variation has been extensively assessed by published studies and can be retrieved from GenBank (**Table 2**).

By surveying these sequences for the presence or absence of these mutations, we were able to determine the most likely haplogroup affiliation for most of the 1170 tiger mosquito mtDNAs from populations worldwide. **Table 2** reports the

TABLE 2 | Frequencies of *A. albopictus* mtDNA haplogroups in worldwide populations.

Geographic origin	*N*	Haplogroup frequencies[a]						Reference
		A1a1a1	**A1a2**	**A1b**	**A2**	**A3**	**Others[b]**	
America	349	94 (*0.27*)	59 (*0.17*)	141 (*0.40*)	0	0	55 (*0.16*)	
USA (New Jersey)	30	28 *(0.93)*	0	0	0	0	2 (*0.07*)	Zhong et al., 2013
USA (California)	49	0	29 (*0.59*)	5 (*0.10*)	0	0	15 (*0.31*)	Zhong et al., 2013
USA (Texas)	31	4 (*0.77*)	0	2 (*0.07*)	0	0	5 (*0.16*)	Zhong et al., 2013
USA (Hawaii)	32	0	27 (*0.84*)	0	0	0	5 (*0.16*)	Zhong et al., 2013
Costa Rica	57	29 (*0.51*)	0	0	0	0	28 (*0.49*)	Futami et al., 2015
Panama	16	13 (*0.81*)	*3 (0.19)*	0	0	0	0	Futami et al., 2015
Brazil[c]	134	0	0	134 (*1.0*)	0	0	0	Birungi and Munstermann, 2002
Europe	32	10 (*0.31*)	20 (*0.63*)	0	0	0	2 (*0.06*)	
Italy (Trento)	32	10 (*0.31*)	20 (*0.63*)	0	0	0	2 (*0.06*)	Zhong et al., 2013
Africa	153	0	0	153 (*1.0*)	0	0	0	
Cameroon	153	0	0	153 (*1.0*)	0	0	0	Kamgang et al., 2011
Asia	284	0	100 (*0.35*)	102 (*0.36*)	25 (*0.09*)	0	57 (*0.20*)	
China[d]	61	0	39 (*0.64*)	0	0	0	22 (*0.36*)	Zhong et al., 2013
China[e] (strain)	30	0	*23 (0.77)*	0	0	0	7 (*0.23*)	Zhong et al., 2013
Japan	15	0	15 (*1.00*)	0	0	0	0	Zhong et al., 2013
Taiwan	30	0	4 (*0.13*)	0	0	0	26 (*0.87*)	Zhong et al., 2013
Thailand	10	0	0	10 (*1.00*)	0	0	0	Sumruayphol et al., 2016
Malaysia	77	0	0	77 (*1.00*)	0	0	0	Zawani et al., 2014
Singapore	36	0	19 (*0.53*)	15 (*0.42*)	0	0	2 (*0.05*)	Zhong et al., 2013
Indonesia (Java)	8	0	0	0	8 (*1.00*)	0	0	Beebe et al., 2013
Indonesia (Timor-Leste)	17	0	0	0	17 (*1.00*)	0	0	Beebe et al., 2013
Oceania	352	0	0	236 (*0.67*)	115 (*0.33*)	0	1 (*<0.01*)	
Australia (Torres Strait)	115	0	0	42 (*0.36*)	72 (*0.63*)	0	1 (*0.01*)	Beebe et al., 2013
Papua New Guinea	170	0	0	*162 (0.95)*	8 (*0.05*)	0	0	Beebe et al., 2013
Papua New Guinea (Southern Fly)	67	0	0	32 (*0.48*)	35 (*0.52*)	0	0	Beebe et al., 2013

[a] *Haplogroup affiliation is based on the sequence variation of the cytochrome c oxidase subunit I (COI) gene, except for the Brazilian samples (see footnote c). The diagnostic mutations are the lack of the transitions at nps 2165 and 1536 for haplogroup A1a1a1, the presence of the transition at np 2294 for haplogroup A1a2, the presence of the transition at np 1820 for haplogroup A1b, and the presence of the transition at np 2027 for haplogroup A2.*

[b] *These samples were classified as "others" because they harbor the transition at np 2165 (except those from Australia, in which that nucleotide position was not sequenced), thus they do not belong to haplogroup A1a1a. Moreover, they lack the transitions at nps 2294, 1820, and 2027, thus they are not members of haplogroups A1a2, A1b, nor A2. Some additional mutations found in these samples are shared between specimens from the same or different geographic areas. In most cases they are likely due to multiple mutational events (recurrent mutations), but some might instead mark additional and not yet identified haplogroups.*

[c] *Samples from Brazil were classified on the basis of their NADH dehydrogenase subunit 5 (ND5) gene sequences. They were considered members of haplogroup A1b because they harbored the transition at np 7210.*

[d] *Samples are from Guangdong and Fujian provinces.*

[e] *Laboratory-maintained strain, Jiangsu province.*

haplogroup frequencies for A1a1a1, A1a2, A1b, A2, and A3 obtained from this survey, as well as a category termed "others" that includes mtDNAs that we could not classify and might encompass haplogroups not represented in our phylogeny. **Figure 3** provides an overview of the worldwide spatial distribution of these haplogroups.

DISCUSSION

Geographical Distribution of *A. albopictus* mtDNA Haplogroups

Phylogenetic analyses revealed that our *A. albopictus* mitogenomes cluster into the three main haplogroups A1, A2, and A3. Intriguingly, the population screening of the COI mutations (1503, 1578C, 1676C, 1704, 1964) found in mitogenome #27 from Taiwan, the only one belonging to haplogroup A3 in our phylogeny, did not reveal any match in the 1170 tiger mosquito mtDNAs from worldwide populations, not even in the population sample ($N = 30$) from Taiwan (**Table 2**). This suggests that *A. albopictus* mosquitoes with A3 mtDNAs were probably not involved in the recent worldwide spread of the species, raising the possibility that this haplogroup might be rare and/or with a restricted geographical distribution.

In contrast, the survey for COI mutations characterizing the Philippine mitogenomes #21-25 allowed the identification of many other mtDNAs belonging to haplogroup A2 (**Table 2**), but all in Insular Southeast Asia, suggesting that this haplogroup

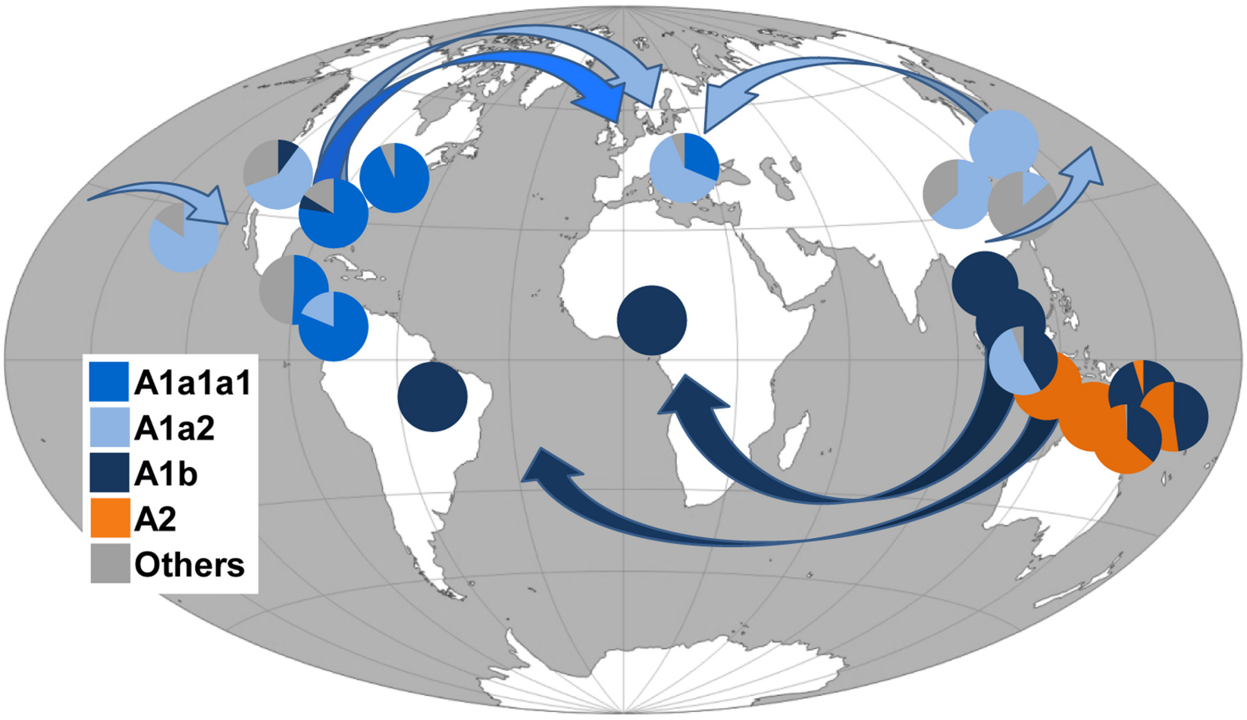


FIGURE 3 | Frequency (%) of *A. albopictus* mtDNA haplogroups in worldwide populations and possible diffusion routes. Arrows indicate possible diffusion routes from the native home-range (South-East Asia) and subsequent dispersals. Populations and corresponding frequency values are listed in **Table 2**. Note that the Chinese laboratory strain was not included. The original world map is from the website (http://www.freeworldmaps.net).

might be typical and possibly limited to the Philippines, Indonesia, Papua New Guinea, and Northern Australia (**Table 2**). Therefore, haplogroup A2 appears to have played a role in the spread of *A. albopictus* from South-East Asia (**Figure 3**) restricted to the context of Oceania. Indeed *A. albopictus* is thought to have spread from Indonesia by human-mediated transportation (Beebe et al., 2013). The low frequency of A2 observed on the Papua New Guinean mainland further supports this scenario, whereas the presence of both A1b and A2 mtDNAs along the North Australian border (**Table 2**) suggests multiple arrivals from distinct geographical sources (Beebe et al., 2013).

Figure 3 shows that, in contrast to the situation described above, members of the other three Asian haplogroups A1a1, A1a2, and A1b are detected in many adventive populations worldwide. This finding identifies these as the Asian mtDNA lineages mainly associated with the recent global spread.

As for haplogroup A1a1, seen in the Japanese mitogenome #26, it is widely distributed in Italy (**Figure 1**) and shows high frequencies in Central America and Eastern USA. Haplogroup A1a2 is present with frequencies higher than 50% in Japan, Southern China, Singapore, Hawaii, California and Italy, whereas A1b is fixed or almost fixed in Thailand, Malaysia, the Papua New Guinea mainland, Cameroon, and Brazil, but present at much lower frequencies also in California and Texas (**Table 2**). Even though these geographical distributions are based on the limited population sampling reported in **Table 2**, some preliminary conclusions can be drawn.

It appears that the ancestral homeland of haplogroup A1a2 might have been a temperate area, possibly Japan or Northern Asia, rather than the tropical range, in agreement with early allozyme studies (Kambhampati et al., 1991). In contrast haplogroup A1b appears to mainly characterize the tropical belt (**Figure 3**). This may imply that genetic and physiological traits make populations with A1b most suited to the colonization of tropical areas (Birungi and Munstermann, 2002; Kamgang et al., 2011). In our phylogeny the Brazilian A1b mitogenome (#9) harbors a mutational motif that includes the transition at np 7210 in the ND5 gene, also found in all Brazilian samples retrieved from the literature and already identified as a marker for Brazilian *A. albopictus* (Birungi and Munstermann, 2002). The same transition was found in two ND5 haplotypes retrieved from the literature, one from Phuntsholing in Southern Bhutan (JQ436953) and one from Chiang Mai in Thailand (JQ436956; Porretta et al., 2012), suggesting a probable route of invasion from Indochina (Goubert et al., 2016). Instead the absence of this transition in the samples from Cameroon (**Table 2**) suggests that A1b mtDNAs arrived in Cameroon from a different tropical source. Interestingly, the absence of photoperiodic diapause in Brazilian mosquitoes supports their origin in tropical Asia, while the diapause in US populations is in agreement with the scenario of an ancestry in the Asian temperate regions (Mousson et al., 2005; Urbanski et al., 2010).

As for haplogroup A1a1, its distribution could not be fully assessed because of the lack of informative marker mutations. However, the absence of the COI transitions at nps 1536 and 2165 distinguishes the members of its main sub-branch, A1a1a1 (**Figure 1**), from all other mitogenomes in the phylogeny. The survey of these transitions in published data sets revealed that haplogroup A1a1a1 is the most common in Costa Rica, Panama, Texas, and New Jersey (**Figure 3**) and widespread in Italy (**Table 2**). This distribution suggests that A1a1a1 and/or its derivatives A1a1a1a and A1a1a1a1 arose recently in an adventive non-Asian population, probably from an ancestral Japanese source, and reached a high frequency because of genetic drift or founder events. From the earliest adventive non-Asian population(s) they then further spread to other distant regions. Such a possibility is in agreement with some previous observations, in particular with the suggestion that Italian tiger mosquitoes, whose first presence was documented in Northern Italy in 1990 (Sabatini et al., 1990), have a dual origin: a possibly direct Northern American source and a probably indirect (through Albania and Greece) Eastern Asian source, with the former related to the international trade of used tires from the eastern coast of the US (Dalla Pozza et al., 1994; Dritsou et al., 2015).

An origin of haplogroup A1a1a1 and/or its derived sub-branches A1a1a1a and A1a1a1a1 in North America and a subsequent arrival to Italy from the US of multiple haplotypes is supported by (i) the detection of four partial sequences from Texas with the transition at np 1823 (Zhong et al., 2013), which defines sub-branch A1a1a1a1, and (ii) our findings that mitogenomes #4 and #5, the first from Virginia and the second from Italy, are identical, and that mitogenomes #2 and #3, again one from Virginia and one from Italy, are closely related (**Figure 1**). Finally, the scenario that the ancestral source of A1a1a1 might be a northern temperate area such as Japan is further supported by the presence of the COI transitions 1536 and 2435, which characterize mitogenome #26, in five published mtDNAs from Kyoto (JQ004524; Xu and Fonseca, 2011), and is in agreement with allozyme data that have highlighted genetic links between North American, Italian and Japanese populations (Urbanelli et al., 2000).

CONCLUSION

Through our analyses, based on complete coding regions, the phylogeny of the *A. albopictus* mitogenomes was charted, and the most likely Asian sources of some adventive populations were identified. The worldwide spread of *A. albopictus* appears to be associated with three mtDNA haplogroups (A1a1, A1a2, and A1b), differently distributed in Asian populations living in temperate and tropical regions, whereas a fourth Asian haplogroup (A2) appears to be restricted to Insular South-East Asia. These ancestral genetic sources now coexist (and interbreed) in many of the recently colonized areas. This occurs not only in the field but also in the laboratory as attested by our detection of both A1a1a1 and A1a2a1 mitogenomes in the Rimini maintained strain, thus creating novel genomic combinations that might be one of the causes of the continuous and apparently growing capability of *A. albopictus* to expand its geographical range.

Note that fine scale mitogenome surveys, encompassing multiple specimens from a wide range of East Asian populations might prove to be an essential pre-requisite to controlling the spread of this mosquito and limiting its social, medical, and economic implications. With a precise identification of the source populations in Asia it will become possible to evaluate the extent and nature of their nuclear genome diversity and the possible selective advantages (e.g., production of cold or desiccation - resistant eggs, zoophilic versus anthropophilic changes in feeding behavior) relative to other Asian *A. albopictus* populations that, by contrast, have not spread.

AUTHOR CONTRIBUTIONS

VB, PG, AT, and AO conceived and designed the experiments. VB, PG, and SB performed the experiments. VB, PG, MRC, AA, LMG, AT, and AO analyzed the data. PG, PAJ, X-GC, AA, OS, LMG, ARM, GG, AT, and AO contributed reagents/materials/analysis tools. VB, PG, GG, AT, and AO wrote the manuscript. All authors contributed to interpretation of data, reviewed and approved the manuscript.

FUNDING

This work was supported by the University of Pavia strategic theme “Toward a governance model for international migration: an interdisciplinary and diachronic perspective” (MIGRAT-IN-G), the EU-FP7 Research Infrastructures project (INFRAVEC) Grant 228421 (to GG), the National Natural Science Foundation of China (81420108024) (to X-GC), the Italian Ministry of Health (RF-2010-2318965) (to ARM), the Italian Ministry of Education, University and Research: Progetti Futuro in Ricerca 2012 (RBFR126B8I) (to AA and AO), and Progetti Ricerca Interesse Nazionale 2012 (2012JA4BTY) (to AA, AT, and OS). PG was supported by a fellowship from the Department of Biology and Biotechnology, University of Pavia.

ACKNOWLEDGMENT

The authors are grateful to Dr. Anca Macovei (Dipartimento di Biologia e Biotecnologie “L. Spallanzani,” Pavia, Italy) and Dr. Kentaro Itokawa (Department of Medical Entomology, National Institute of Infectious Diseases, Tokyo, Japan), for their contributions to the collection of the *A. albopictus* samples from the Philippines and Japan, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00208/full#supplementary-material

REFERENCES

Achilli, A., Olivieri, A., Pellecchia, M., Uboldi, C., Colli, L., Al-Zahery, N., et al. (2012). Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2449–2454. doi: 10.1073/pnas.1111637109

Achilli, A., Olivieri, A., Soares, P., Lancioni, H., Hooshiar Kashani, B., Perego, U. A., et al. (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.* 18, R157–R158. doi: 10.1016/j.cub.2008.01.019

Armbruster, P., Damsky, W. E., Giordano, R., Birungi, J., Munstermann, L. E., and Conn, J. E. (2003). Infection of new- and old-world *Aedes albopictus* (Diptera: Culicidae) by the intracellular parasite *Wolbachia*: implications for host mitochondrial DNA evolution. *J. Med. Entomol.* 40, 356–360. doi: 10.1603/0022-2585-40.3.356

Beebe, N. W., Ambrose, L., Hill, L. A., Davis, J. B., Hapgood, G., Cooper, R. D., et al. (2013). Tracing the tiger: population genetics provides valuable insights into the *Aedes (Stegomyia) albopictus* invasion of the Australasian region. *PLoS Negl. Trop. Dis.* 7:e2361. doi: 10.1371/journal.pntd.0002361

Bellini, R., Calvitti, M., Medici, A., Carrieri, M., Celli, G., and Maini, S. (2007). "Use of the sterile insect technique against *Aedes albopictus* in Italy: first results of a pilot trial," in *Area-Wide Control of Insect Pests, From Research to Field Implementation*, eds M. J. B. Vreysen, A. S. Robinson, and J. Hendrichs (Dordrecht: Springer), 505–515.

Benedict, M. Q., Levine, R. S., Hawley, W. A., and Lounibos, L. P. (2007). Spread of the tiger: global risk of invasion by the mosquito *Aedes albopictus*. *Vector Borne Zoonotic Dis.* 7, 76–85. doi: 10.1089/vbz.2006.0562

Birungi, J., and Munstermann, L. E. (2002). Genetic structure of *Aedes albopictus* (Diptera: Culicidae) populations based on mitochondrial ND5 sequences: evidence for an independent invasion into Brazil and United States. *Ann. Entomol. Soc. Am.* 95, 125–132. doi: 10.1603/0013-8746(2002)095[0125:GSOAAD]2.0.CO;2

Bonizzoni, M., Gasperi, G., Chen, X., and James, A. A. (2013). The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. *Trends Parasitol.* 29, 460–468. doi: 10.1016/j.pt.2013.07.003

Chen, X. G., Jiang, X., Gu, J., Xu, M., Wu, Y., Deng, Y., et al. (2015). Genome sequence of the Asian tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5907–E5915. doi: 10.1073/pnas.1516410112

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., et al. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31, 3497–3500. doi: 10.1093/nar/gkg500

Chouin-Carneiro, T., Vega-Rua, A., Vazeille, M., Yebakima, A., Girod, R., Goindin, D., et al. (2016). Differential susceptibilities of *Aedes aegypti* and *Aedes albopictus* from the Americas to Zika virus. *PLoS Negl. Trop. Dis.* 10:e0004543. doi: 10.1371/journal.pntd.0004543

Dalla Pozza, G. L., Romi, R., and Severini, C. (1994). Source and spread of *Aedes albopictus* in the Veneto region of Italy. *J. Am. Mosq. Control Assoc.* 10, 589–592.

Delatte, H., Bagny, L., Brengue, C., Bouetard, A., Paupy, C., and Fontenille, D. (2011). The invaders: phylogeography of dengue and chikungunya viruses *Aedes* vectors, on the South West islands of the Indian Ocean. *Infect. Genet. Evol.* 11, 1769–1781. doi: 10.1016/j.meegid.2011.07.016

Dritsou, V., Topalis, P., Windbichler, N., Simoni, A., Hall, A., Lawson, D., et al. (2015). A draft genome sequence of an invasive mosquito: an Italian *Aedes albopictus*. *Pathog. Glob. Health* 109, 207–220. doi: 10.1179/2047773215Y.0000000031

Drummond, A. J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Dueñas, J. C., Gardenal, C. N., Llinás, G. A., and Panzetta-Dutari, G. M. (2006). Structural organization of the mitochondrial DNA control region in *Aedes aegypti*. *Genome* 49, 931–937. doi: 10.1139/G06-053

Futami, K., Valderrama, A., Baldi, M., Minakawa, N., Marín Rodríguez, R., and Chaves, L. F. (2015). New and common haplotypes shape genetic diversity in Asian tiger mosquito populations from Costa Rica and Panamá. *J. Econ. Entomol.* 108, 761–768. doi: 10.1093/jee/tou028

Gasperi, G., Bellini, R., Malacrida, A. R., Crisanti, A., Dottori, M., and Aksoy, S. (2012). A new threat looming over the Mediterranean basin: emergence of viral diseases transmitted by *Aedes albopictus* mosquitoes. *PLoS Negl. Trop. Dis.* 6:e1836. doi: 10.1371/journal.pntd.0001836

Goubert, C., Minard, G., Vieira, C., and Boulesteix, M. (2016). Population genetics of the Asian tiger mosquito *Aedes albopictus*, an invasive vector of human diseases. *Heredity (Edinb).* 117, 125–134. doi: 10.1038/hdy.2016.35

Gubler, D. J. (2006). Dengue/dengue haemorrhagic fever: history and current status. *Novartis Found. Symp.* 277, 3–16,discussion16–22,71–13,251–253. doi: 10.1002/0470058005.ch2

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Higa, Y., Toma, T., Tsuda, Y., and Miyagi, I. (2010). A multiplex PCR-based molecular identification of five morphologically related, medically important subgenus *Stegomyia mosquitoes* from the genus *Aedes* (Diptera: Culicidae) found in the Ryukyu Archipelago. *Jpn. J. Infect. Dis.* 63, 312–316.

Kambhampati, S., Black, W. C., and Rai, K. S. (1991). Geographic origin of the US and Brazilian *Aedes albopictus* inferred from allozyme analysis. *Heredity (Edinb.)* 67, 85–93. doi: 10.1038/hdy.1991.67

Kamgang, B., Brengues, C., Fontenille, D., Njiokou, F., Simard, F., and Paupy, C. (2011). Genetic structure of the tiger mosquito, *Aedes albopictus*, in Cameroon (Central Africa). *PLoS ONE* 6:e20257. doi: 10.1371/journal.pone.0020257

Kamgang, B., Ngoagouni, C., Manirakiza, A., Nakouné, E., Paupy, C., and Kazanji, M. (2013). Temporal patterns of abundance of *Aedes aegypti* and *Aedes albopictus* (Diptera: Culicidae) and mitochondrial DNA analysis of *Ae. albopictus* in the Central African Republic. *PLoS Negl. Trop. Dis.* 7:e2590. doi: 10.1371/journal.pntd.0002590

Khormi, H. M., and Kumar, L. (2014). Climate change and the potential global distribution of *Aedes aegypti*: spatial modelling using GIS and CLIMEX. *Geospat. Health* 8, 405–415. doi: 10.4081/gh.2014.29

Kraemer, M. U., Sinka, M. E., Duda, K. A., Mylne, A., Shearer, F. M., Brady, O. J., et al. (2015). The global compendium of *Aedes aegypti* and *Ae. albopictus* occurrence. *Sci. Data* 2:150035. doi: 10.1038/sdata.2015.35

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Manni, M., Gomulski, L. M., Aketarawong, N., Tait, G., Scolari, F., Somboon, P., et al. (2015). Molecular markers for analyses of intraspecific genetic diversity in the Asian tiger mosquito, *Aedes albopictus*. *Parasit. Vectors* 8:188. doi: 10.1186/s13071-015-0794-5

Medlock, J. M., Hansford, K. M., Schaffner, F., Versteirt, V., Hendrickx, G., Zeller, H., et al. (2012). A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options. *Vector Borne Zoonotic Dis.* 12, 435–447. doi: 10.1089/vbz.2011.0814

Medlock, J. M., Hansford, K. M., Versteirt, V., Cull, B., Kampen, H., Fontenille, D., et al. (2015). An entomological review of invasive mosquitoes in Europe. *Bull. Entomol. Res.* 105, 637–663. doi: 10.1017/S0007485315000103

Mousson, L., Dauga, C., Garrigues, T., Schaffner, F., Vazeille, M., and Failloux, A. B. (2005). Phylogeography of *Aedes (Stegomyia) aegypti* (L.) and *Aedes (Stegomyia) albopictus* (Skuse) (Diptera: Culicidae) based on mitochondrial DNA variations. *Genet. Res.* 86, 1–11. doi: 10.1017/S0016672305007627

Nei, M., and Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. New York, NY: Oxford University Press.

Paupy, C., Delatte, H., Bagny, L., Corbel, V., and Fontenille, D. (2009). *Aedes albopictus*, an arbovirus vector: from the darkness to the light. *Microbes Infect.* 11, 1177–1185. doi: 10.1016/j.micinf.2009.05.005

Peng, H. J., Lai, H. B., Zhang, Q. L., Xu, B. Y., Zhang, H., Liu, W. H., et al. (2012). A local outbreak of dengue caused by an imported case in Dongguan China. *BMC Public Health* 12:83. doi: 10.1186/1471-2458-12-83

Porretta, D., Mastrantonio, V., Bellini, R., Somboon, P., and Urbanelli, S. (2012). Glacial history of a modern invader: phylogeography and species distribution modelling of the Asian tiger mosquito *Aedes albopictus*. *PLoS ONE* 7:e44515. doi: 10.1371/journal.pone.0044515

Powell, J. R., and Tabachnick, W. J. (2013). History of domestication and spread of *Aedes aegypti*-a review. *Mem. Inst. Oswaldo Cruz* 108, 11–17. doi: 10.1590/0074-0276130395

Rai, K. S. (1991). *Aedes albopictus* in the Americas. *Annu. Rev. Entomol.* 36, 459–484. doi: 10.1146/annurev.en.36.010191.002331

Rezza, G. (2012). *Aedes albopictus* and the reemergence of dengue. *BMC Public Health* 12:72. doi: 10.1186/1471-2458-12-72

Rezza, G. (2014). Dengue and chikungunya: long-distance spread and outbreaks in naïve areas. *Pathog. Glob. Health* 108, 349–355. doi: 10.1179/2047773214Y.0000000163

Romi, R., Sabatinelli, G., Savelli, L. G., Raris, M., Zago, M., and Malatesta, R. (1997). Identification of a north American mosquito species, *Aedes atropalpus* (Diptera: Culicidae), in Italy. *J. Am. Mosq. Control. Assoc.* 13, 245–246.

Rueda, L. M. (2004). Pictorial keys for the identification of mosquitoes (Diptera: Culicidae) associated with dengue virus transmission. *Zootaxa* 589, 1–60. doi: 10.11646/zootaxa.589.1.1

Sabatini, A., Raineri, V., Trovato, G., and Coluzzi, M. (1990). *Aedes albopictus* in Italy and possible diffusion of the species into the Mediterranean area. *Parassitologia* 32, 301–304.

Schaffner, F., Medlock, J. M., and Van Bortel, W. (2013). Public health significance of invasive mosquitoes in Europe. *Clin. Microbiol. Infect.* 19, 685–692. doi: 10.1111/1469-0691.12189

Scholte, E. J., Den Hartogm, W., Braks, M., Reusken, C., Dik, M., and Hessels, A. (2009). First report of a north American invasive mosquito species *Ochlerotatus atropalpus* (Coquillett) in the Netherlands, 2009. *Euro Surveill.* 14:19400.

Sumruayphol, S., Apiwathnasorn, C., Ruangsittichai, J., Sriwichai, P., Attrapadung, S., Samung, Y., et al. (2016). DNA barcoding and wing morphometrics to distinguish three *Aedes* vectors in Thailand. *Acta Trop.* 159, 1–10. doi: 10.1016/j.actatropica.2016.03.010

Torroni, A., Achilli, A., Macaulay, V., Richards, M., and Bandelt, H. J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends Genet.* 22, 339–345. doi: 10.1016/j.tig.2006.04.001

Urbanelli, S., Bellini, R., Carrieri, M., Sallicandro, P., and Celli, G. (2000). Population structure of *Aedes albopictus* (Skuse): the mosquito which is colonizing Mediterranean countries. *Heredity (Edinb)* 84, 331–337. doi: 10.1046/j.1365-2540.2000.00676.x

Urbanski, J. M., Benoit, J. B., Michaud, M. R., Denlinger, D. L., and Armbruster, P. (2010). The molecular physiology of increased egg desiccation resistance during diapause in the invasive mosquito, *Aedes albopictus*. *Proc. Biol. Sci.* 277, 2683–2692. doi: 10.1098/rspb.2010.0362

Wong, P. S., Li, M. Z., Chong, C. S., Ng, L. C., and Tan, C. H. (2013). *Aedes (Stegomyia) albopictus* (Skuse): a potential vector of Zika virus in Singapore. *PLoS Negl. Trop. Dis.* 7:e2348. doi: 10.1371/journal.pntd.0002348

Wu, J. Y., Lun, Z. R., James, A. A., and Chen, X. G. (2010). Dengue fever in mainland China. *Am. J. Trop. Med. Hyg.* 83, 664–671. doi: 10.4269/ajtmh.2010.09-0755

Xu, J., and Fonseca, D. M. (2011). One-way sequencing of multiple amplicons from tandem repetitive mitochondrial DNA control region. *Mitochondrial DNA* 22, 155–158. doi: 10.3109/19401736.2011.636434

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Zawani, M. K., Abu, H. A., Sazaly, A. B., Zary, S. Y., and Darlina, M. N. (2014). Population genetic structure of *Aedes albopictus* in Penang, Malaysia. *Genet. Mol. Res.* 13, 8184–8196. doi: 10.4238/2014.October.7.13

Zhang, D.-X., and Hewitt, G. M. (1997). Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochem. Syst. Ecol.* 25, 99–120. doi: 10.1016/S0305-1978(96)00042-7

Zhang, H., Xing, D., Wang, G., Li, C., and Zhao, T. (2015). Sequencing and analysis of the complete mitochondrial genome of *Aedes albopictus* (Diptera: Culicidae) in China. *Mitochondrial DNA A DNA Mapp Seq Anal.* 27, 2787–2788. doi: 10.3109/19401736.2015.1053067

Zhong, D., Lo, E., Hu, R., Metzger, M. E., Cummings, R., Bonizzoni, M., et al. (2013). Genetic analysis of invasive *Aedes albopictus* populations in Los Angeles County, California and its potential public health impact. *PLoS ONE* 8:e68586. doi: 10.1371/journal.pone.0068586

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Mitogenome Diversity in Sardinians: A Genetic Window onto an Island's Past

Anna Olivieri,[†,1] Carlo Sidore,[†,2,3,4] Alessandro Achilli,[†,1] Andrea Angius,[2,4,5] Cosimo Posth,[6,7] Anja Furtwängler,[7] Stefania Brandini,[1] Marco Rosario Capodiferro,[1] Francesca Gandini,[1,8] Magdalena Zoledziewska,[2] Maristella Pitzalis,[2] Andrea Maschio,[2,3] Fabio Busonero,[2,3] Luca Lai,[9] Robin Skeates,[10] Maria Giuseppina Gradoli,[11] Jessica Beckett,[12] Michele Marongiu,[2] Vittorio Mazzarello,[4] Patrizia Marongiu,[4] Salvatore Rubino,[4] Teresa Rito,[13] Vincent Macaulay,[14] Ornella Semino,[1] Maria Pala,[8] Gonçalo R. Abecasis,[3] David Schlessinger,[15] Eduardo Conde-Sousa,[16] Pedro Soares,[16] Martin B. Richards,[8] Francesco Cucca,[*,2,4] and Antonio Torroni[*,1]

[1]Dipartimento di Biologia e Biotecnologie, Università di Pavia, Pavia, Italy

[2]Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy

[3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI

[4]Dipartimento di Scienze Biomediche, Università di Sassari, Sassari, Italy

[5]Center for Advanced Studies, Research and Development in Sardinia (CRS4), AGCT Program, Parco Scientifico e Tecnologico della Sardegna, Pula, Italy

[6]Max Planck Institute for the Science of Human History, Jena, Germany

[7]Institute for Archaeological Sciences, Archaeo- and Palaeogenetics, University of Tübingen, Tübingen, Germany

[8]Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Huddersfield, Queensgate, United Kingdom

[9]Department of Anthropology, University of South Florida, Tampa, FL

[10]Department of Archaeology, Durham University, Durham, United Kingdom

[11]School of Archaeology and Ancient History, University of Leicester, Leicester, United Kingdom

[12]Independent Contractor, Cagliari, Italy

[13]Life and Health Sciences Research Institute (ICVS), School of Health Sciences & ICVS/3B's-PT Government Associate Laboratory, University of Minho, Braga, Portugal

[14]School of Mathematics and Statistics, University of Glasgow, Glasgow, United Kingdom

[15]Laboratory of Genetics, National Institute on Aging US National Institutes of Health, Baltimore, Maryland, MD

[16]CBMA (Centre of Molecular and Environmental Biology), Department of Biology, University of Minho, Campus de Gualtar, Braga, Portugal

[†]These authors contributed equally to this work.

*__Corresponding authors:__ E-mails: fcucca@uniss.it; antonio.torroni@unipv.it.

Associate editor: Connie Mulligan

Abstract

Sardinians are "outliers" in the European genetic landscape and, according to paleogenomic nuclear data, the closest to early European Neolithic farmers. To learn more about their genetic ancestry, we analyzed 3,491 modern and 21 ancient mitogenomes from Sardinia. We observed that 78.4% of modern mitogenomes cluster into 89 haplogroups that most likely arose *in situ*. For each Sardinian-specific haplogroup (SSH), we also identified the upstream node in the phylogeny, from which non-Sardinian mitogenomes radiate. This provided minimum and maximum time estimates for the presence of each SSH on the island. In agreement with demographic evidence, almost all SSHs coalesce in the post-Nuragic, Nuragic and Neolithic-Copper Age periods. For some rare SSHs, however, we could not dismiss the possibility that they might have been on the island prior to the Neolithic, a scenario that would be in agreement with archeological evidence of a Mesolithic occupation of Sardinia.

Key words: **mitochondrial genomes, mitochondrial DNA phylogeny, haplogroups, prehistory of Sardinia, origins of Europeans.**

Main Text

Sardinia is an island that remained unconnected with the mainland even when the sea level was at its lowest during the LGM (Shackleton et al. 1984) and probably was the last of the large Mediterranean islands to be colonized by modern humans (Sondaar 1998). Modern Sardinians, a unique

doi:10.1093/molbev/msx082 Advance Access publication February 8, 2017

reservoir of distinct genetic signatures (Cavalli-Sforza et al. 1994; Pala et al. 2009; Francalacci et al. 2013; Sidore et al. 2015), on one hand apparently harbor the highest levels of nuclear genome similarity with European Neolithic farmers (Lazaridis et al. 2014) and an extensive similarity with the Late Neolithic/Chalcolithic Tyrolean Iceman (Keller et al. 2012; Sikora et al. 2014) but, on the other hand, they differ substantially from Near Eastern Neolithic farmers including those from Anatolia (Lazaridis et al. 2016). These findings have led to the view that, in modern Europe, Sardinians may have best preserved the gene pool of Neolithic farmers, possibly because their ancestors were less affected by subsequent Bronze Age dispersals across Europe (Haak et al. 2015).

Note that the above view does not necessarily imply that the first Sardinians were Neolithic farmers. On the contrary, there is archeological evidence indicating that humans were present on the island by at least 13 Kya (Hofmeijer et al. 1989; Dyson and Rowland 2007; Broodbank 2013). Moreover, a European pre-Neolithic origin for Y-chromosome haplogroup I2a1a1-M26, by far the most common in modern Sardinian males (38.9%) (Francalacci et al. 2013), has been postulated (Rootsi et al. 2004). Finally, a massive survey of whole-genome sequences from modern Sardinians has recently shown that the population of the mountainous Gennargentu region harbors higher levels of both hunter-gatherer and Neolithic farmer components relative to other Sardinian groups from less isolated areas. This has been interpreted as indicating that the hunter-gatherer component did not reach the island with more recent migrations from the continent, but it was either already present on the island when farmers arrived or due to previous admixture of the first incoming farmers with hunter-gatherers on the mainland (Chiang et al. 2016).

In this complex and partially contradictory scenario, the genetic perspective of the maternally transmitted mitochondrial genome is still almost completely unexplored. Therefore, to learn more about the ancestry of Sardinians and their genetic links with modern and ancient European (and other) populations, we analyzed a large dataset of 3,491 novel complete mitogenomes from modern islanders as well as 21 mitogenomes from ancient specimens. Among the modern samples, we removed 1,355 maternally related samples on the bases of pedigree data or kinship evaluation of nuclear genomes, and 44 samples with non-Sardinian maternal origins. We then assessed the phylogenetic relationships of the remaining mitogenomes (2,092 out of the initial 3,491), plus 124 previously published Sardinian mitogenomes, with all publicly available worldwide mitogenomes (more than 26,000 data not shown).

Sardinian-Specific Haplogroups and Their Coalescence Ages

Our phylogenetic analyses revealed that 1,737 Sardinian mitogenomes (78.4%) clustered into 89 Sardinian-Specific Haplogroups (SSHs; see Methods for defining criteria) (supplementary table S1, supplementary fig. S1, Supplementary Material online). For each SSH, using non-Sardinian mitogenomes, we also identified the upstream node in the phylogeny, from which non-Sardinian Closest External Mitogenomes (CEMs) radiate (supplementary table S2, supplementary fig. S2, Supplementary Material online).

The finding that 78.4% of Sardinians harbor ethnic-specific haplogroups might appear surprising in the European context. However, a similar—though not so extreme—scenario is seen in the Basque-speaking regions of Spain, where a survey of haplogroup H mitogenomes (54.1% of the populations) identified six autochthonous sub-haplogroups encompassing 29.0% of all mitochondrial DNAs (Behar et al. 2012).

The 89 SSHs, 80 of which are defined here for the first time, include descendants from all major macro-haplogroups of the human mitochondrial DNA (mtDNA) tree (L, M, N and R) and are defined by a total of 104 mutations (supplementary table S3, Supplementary Material online). About 51% of modern Sardinian-specific mtDNAs fall into HV, 27% into JT, 17% into U, and 5% into other lineages. These frequencies are close to those reported in typical western European populations. However, when assessed at a higher level of haplogroup resolution, they differ substantially from those in continental Europe. This is most marked for H1 and H3, with peak values on the island of 18.5% and 18.4%, respectively (supplementary table S4, Supplementary Material online).

It is most likely that the SSHs and their distinguishing mutational motifs arose *in situ* (fig. 1*A*), even though the possibility that some of these motifs arose outside the island and, after their arrival in Sardinia, were lost in the ancestral sources, should not be overlooked. However, even in this case, the coalescence age of the SSH corresponds to the minimum time estimate for the presence of its founder mutational motif on the island (fig. 1*B*). An overestimation of the SSH arrival/presence time in Sardinia would occur only if the founder haplotype as well as a derived haplotype both moved to Sardinia, but were also both lost in the ancestral source (fig. 1*D*). This is not a very likely scenario, not only because it requires multiple events but also given the great diversity of some of the deep-rooting founding lineages—for example, within the predominant haplogroups H1 and H3, where lineages are very sharply partitioned between the island and the mainland.

For all 89 SSHs, we assessed these minimum coalescence ages with both Maximum Likelihood (ML) and BEAST (Bayesian Evolutionary Analysis Sampling Trees) computations, employing two different mutation rates: one established on modern mitogenomes, which corrects for the effect of selection and is routinely applied in phylogeographic studies (Soares et al. 2009); the other using radiocarbon dated ancient mitogenomes as tip calibration points (Posth et al. 2016) (supplementary table S3; supplementary fig. S2, Supplementary Material online). In agreement with historical demographic evidence (Francalacci et al. 2013), all estimates indicate that more than 50% of SSHs coalesce in the post-Nuragic (<2 Kya) and Nuragic (~2–4 Kya) archaeological periods. However, not all of the remainder fall in the Neolithic-Copper Age period (~4–7.8 Kya) (supplementary fig. S2, Supplementary Material online). In particular, three rather rare SSHs, K1a2d, N1b1a9 and U5b1i1, corresponding to 3.1% of modern Sardinian-specific mitogenomes, showed

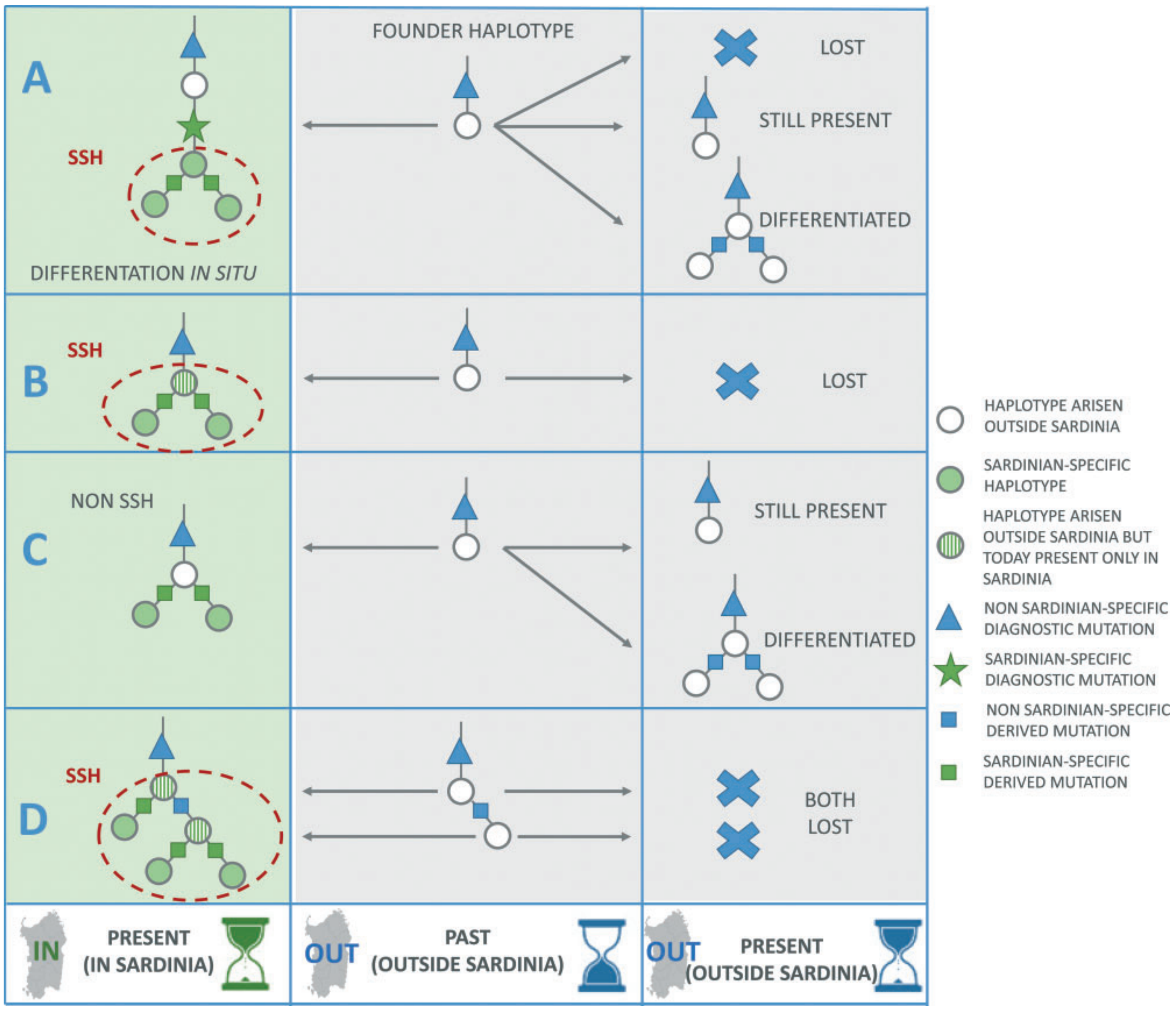

A
SSH
DIFFERENTATION IN SITU
FOUNDER HAPLOTYPE
LOST
STILL PRESENT
DIFFERENTIATED
B
SSH
LOST
C
NON SSH
STILL PRESENT
DIFFERENTIATED
D
SSH
BOTH
LOST
IN
PRESENT
(IN SARDINIA)
OUT
PAST
(OUTSIDE SARDINIA)
OUT
PRESENT
(OUTSIDE SARDINIA)
HAPLOTYPE ARISEN OUTSIDE SARDINIA
SARDINIAN-SPECIFIC HAPLOTYPE
HAPLOTYPE ARISEN OUTSIDE SARDINIA BUT TODAY PRESENT ONLY IN SARDINIA
NON SARDINIAN-SPECIFIC DIAGNOSTIC MUTATION
SARDINIAN-SPECIFIC DIAGNOSTIC MUTATION
NON SARDINIAN-SPECIFIC DERIVED MUTATION
SARDINIAN-SPECIFIC DERIVED MUTATION

Fig. 1. Schematic representation of possible scenarios (*A*–*D*) for the differentiation of a founder mtDNA haplotype into a Sardinian-Specific Haplogroup (SSH) and implications for the SSH age estimate. (*A*) The founder haplotype (from outside Sardinia) acquired one novel mutation (or more) in Sardinia (*in situ*) giving rise to the SSH. At the present time, the founder haplotype could have been lost, still be present, and/or have differentiated into a new haplotype(s) outside Sardinia. (*B*) The founder haplotype arrived to Sardinia and gave rise to Sardinian-specific haplotypes, but it was lost outside Sardinia. (*C*) The founder haplotype arrived to Sardinia and gave rise to Sardinian-specific haplotypes and is still present and/or differentiated outside Sardinia. (*D*) The founder haplotype diverged outside Sardinia and both the founder and the derived haplotype arrived in Sardinia where they both differentiated into Sardinian-specific haplotypes, whilst both were lost outside Sardinia. Scenarios *A*, *B* and *D* would give rise to what we defined as "Sardinian-specific haplogroups", but only scenario *D* would lead to an overestimation of the SSH presence/arrival time on the island.

with all or some of the approaches a mean coalescence age >7.8 Ky (table 1).

A Pre-Neolithic Presence in the Island for Some Sardinian-Specific Haplogroups?

The postulated archeologically-based starting time of the Neolithic in Sardinia is 7.8 Kya (Berger and Guilaine 2009). Taking into account that coalescence ages correspond to the lower bound for their presence on the island, the observation that some SSHs might coalesce prior to that boundary raises the possibility that their founding haplotypes were already on the island during the Mesolithic (fig. 2; supplementary fig. S2, Supplementary Material online).

Haplogroup K1a2d includes nine mitogenomes and, with both ML and BEAST and both mutation rates, shows coalescence ages (11–16 Ky) that, even when standard errors are included, predate the Neolithic (table 1). K1a2d is only one mutation away from K1a2, the previously defined node (supplementary table S3, supplementary fig. S1, Supplementary Material online). From this node several other sub-branches depart (fig. 3), and the members of these branches are CEMs to the Sardinian-specific branch. Most of the K1a2 sub-branches (K1a2a-d) encompass only European mitogenomes (Costa et al. 2013), but several descending directly from the root of K1a2 have been identified also in the Near East and include ancient samples. The oldest are two K1a2

Table 1. Maximum Likelihood (ML) and Bayesian Age Estimates for the Three Sardinian-specific Haplogroups (SSHs) Whose Age Estimates are >7.8 Ky.

SSH	N[a]	ML age estimates (Ky)				BEAST age estimates (Ky)				Ancestral geographic source
		Soares et al. (2009)		Posth et al. (2016)		Soares et al. (2009)		Posth et al. (2016)		
		T	SE	T	SE	T	SE	T	SE	
K1a2d	9	16.02	1.98	13.10	1.55	12.79	2.71	10.99	2.35	Near East
U5b1i1	41	12.98	6.07	10.70	4.82	11.32	2.69	9.69	2.37	Western Europe
N1b1a9	4	9.39	1.89	7.83	1.53	8.46	2.39	7.25	2.01	Near East

[a]Number of mitogenomes included in the corresponding SSH.

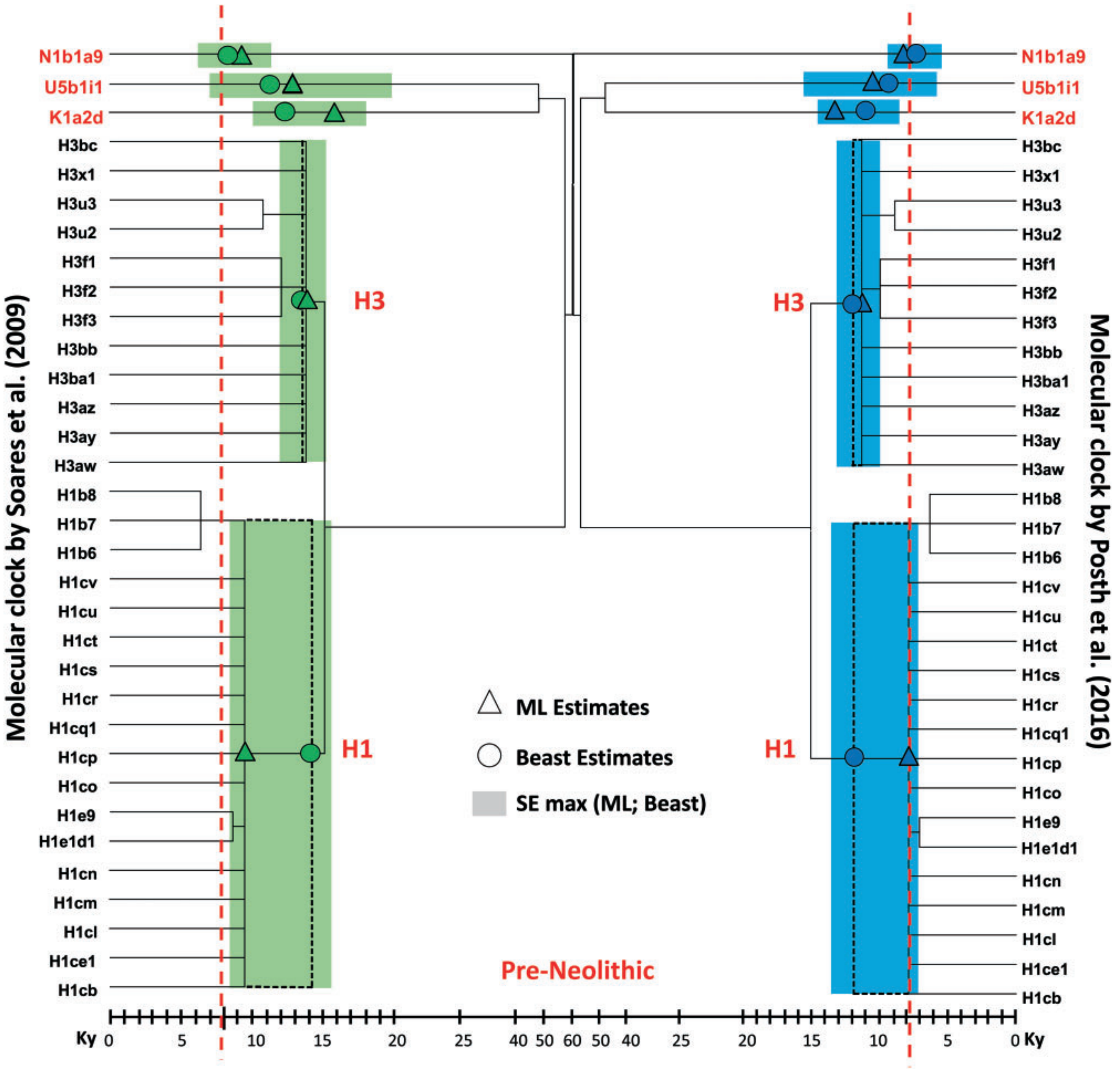


Fig. 2. Specular schematic trees encompassing the three Sardinian-specific haplogroups (N1b1a9, U5b1i1, K1a2d) whose age estimates might predate the Neolithic (>7.8 Kya) and the Sardinian haplogroups H1 and H3. Age estimates were calculated by employing two mutation rates, by Soares et al. (2009) (tree on the left) and by Posth et al. (2016) (tree on the right). Triangles and continuous lines indicate ML estimates. Circles and dashed lines indicate BEAST estimates. Ages are according to the (non-linear) time scale on the bottom. Colored shadings show the largest confidential intervals of age estimates. For details concerning the age estimates of all SSHs see supplementary figure S2, Supplementary Material online.

mitogenomes from early Anatolian farmers, radiocarbon-dated to 8.3 and 8.0 Kya (supplementary table S5, Supplementary Material online). These observations indicate the Near East as the most likely ancestral source of K1a2. If so, our observations could be interpreted as indicating that K1a2 (female) carriers of Near Eastern ancestry arrived in Sardinia in the time frame between 18.7–14.5 Kya (ages of K1a2) and 16.0–11.0 Kya (ages of K1a2d)—that is, in Late Glacial times (supplementary table S3, Supplementary Material online).

Haplogroup U5b1i1 includes 41 mitogenomes. Its mean coalescence ages are in the range of 9.7-13.0 Ky, although in most cases standard errors overlap with the arrival of the Neolithic in Sardinia (table 1). In modern populations, the CEMs to the Sardinian-specific haplogroup U5b1i1 are those departing from the newly identified U5b1i node (with an estimated age in the range of 10.7-13.5 Ky) and are all from western Europe (1 from Germany, 2 from the UK) (fig. 3), the same geographic origin as the U5b1 mitogenomes (upstream of U5b1i) from ancient samples (supplementary table S5, Supplementary Material online). Thus, in contrast with K1a2d, this matrilineal genetic component harbors deep ancestral roots in Western Europe.

Haplogroup N1b1a9 includes only four mitogenomes. Its mean coalescence ages are in the range of 7.3-9.4 Ky with

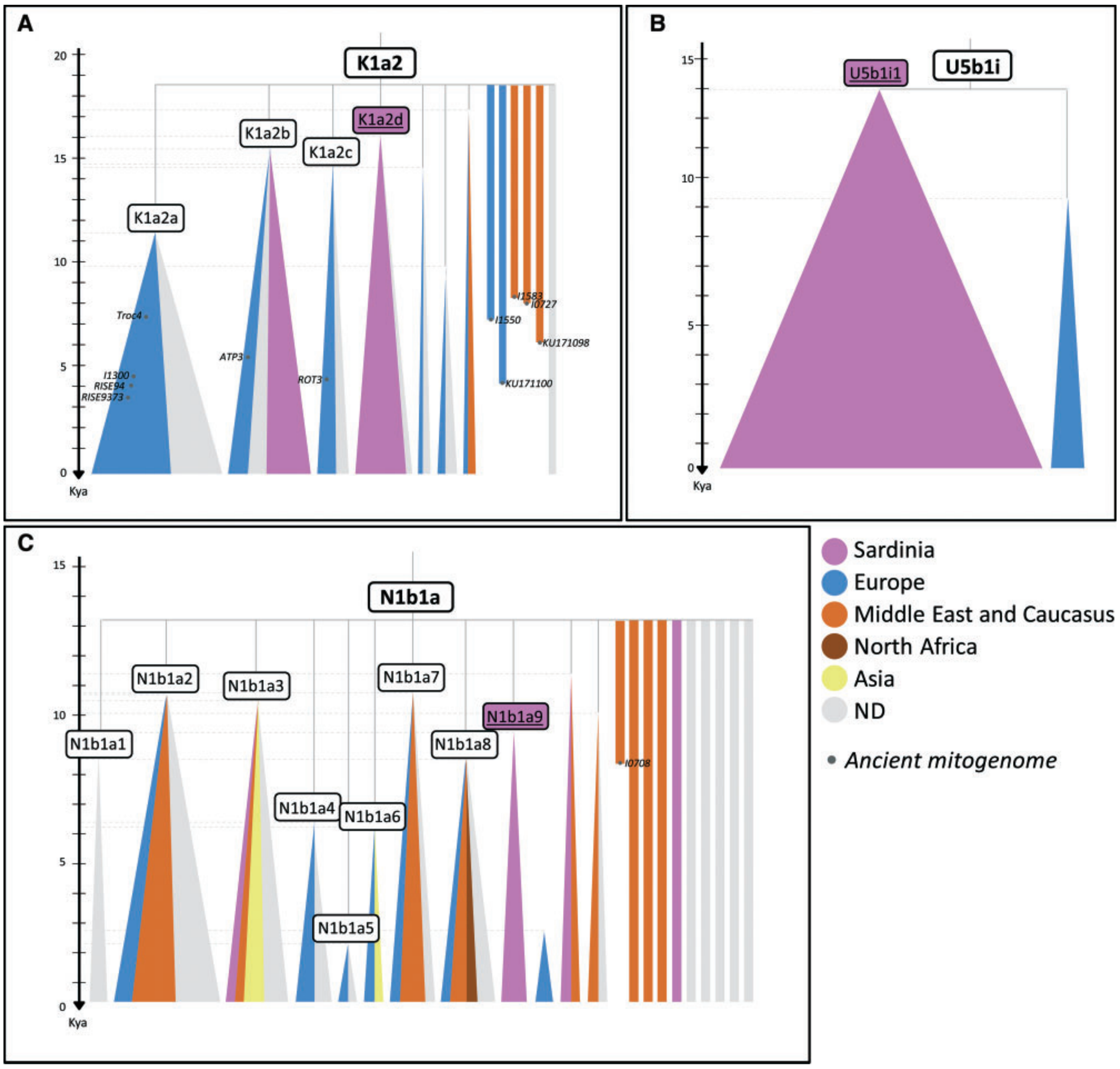


FIG. 3. Schematic representation of K1a2 (panel A), U5b1i (panel B), and N1b1a (panel C) phylogenies. Subclades are represented by triangles, while singletons by lines. The width of triangles is proportional to the number of both modern and ancient mitogenomes, while the height to the age of the clades (Kya) estimated with ML and the molecular clock proposed by Soares et al. (2009). These and the ages obtained with other methods/rates are listed in supplementary table S3, Supplementary Material online. Colours indicate the geographical origin of samples according to the legend. Ancient samples (whose codes are those reported in supplementary table S7 and supplementary fig. S5, Supplementary Material online) are placed in correspondence of their radiocarbon calibrated ages. The name of Sardinian-specific haplogroups is underlined and in a purple field .

standard errors always overlapping with the Neolithic (table 1). N1b1a9 is only one mutation away from the previously defined node N1b1a (supplementary table S3, supplementary fig. S1, Supplementary Material online), from which numerous other branches depart (fig. 3). Most of these branches are shared between Europeans and Near Easterners, indicating the Near East as the likely homeland of N1b1a. This source is further supported by the geographical origin (Anatolia) of the only ancient N1b1a mitogenome (8.3 Ky) recovered so far (supplementary table S5, Supplementary Material online). Thus, the (female) N1b1a carriers of Near Eastern ancestry might have arrived in Sardinia in the time frame between 17-11 Kya (age of N1b1a) and 9-7 Kya (age of N1b1a9) (supplementary table S3, Supplementary Material online).

Founder Analysis and Coalescent Simulations

To further evaluate the arrival/presence times of U5b1i, K1a2 and N1b1a in Sardinia, we performed a founder analysis (Richards et al. 2000; Macaulay and Richards 2013; Soares et al. 2016). This method assumes a strict division between potential source populations and sink population, and subtracts the diversity within the sink dataset that arose in the source region. In our case, the potential sources for the Sardinian mitogenomes were their modern CEMs (fig. 3). The migration scan, which plots the fraction of arriving lineages against time, showed single primary peaks at 17.0, 13.2, and 8.0 Kya for haplogroups K1a2d, U5b1i1 and N1b1a9, respectively (supplementary fig. S3, Supplementary Material online). Taking into account that founder analysis is conservative in that it provides only minimum estimates for the arrival time of each founder lineage, since the arrival necessarily predates the origin and expansion of the corresponding founder cluster, these peaks tend to support a pre-Neolithic presence of U5b1i1 and K1a2d in Sardinia. For N1b1a9, the result does not completely rule out such a possibility, but makes it less plausible.

We then performed coalescent simulations under different demographic models (Hudson 2002) to test the two alternative scenarios: i) a single Neolithic occupation of the island at 8 Kya; ii) a first entry in the Late Paleolithic followed by another in the Neolithic. These simulations supported a dual migration scenario, with a first migration event in the Late Paleolithic at 12–15 Kya (effective population size between 500 and 1,500), followed by an Early Neolithic migration at about 8 Kya (effective population size of 35,000) (supplementary table S6, Supplementary Material online).

Ancient Sardinian Mitogenomes

To further investigate the genetic ancestry of Sardinians, 21 prehistoric mitogenomes, from skeletal remains collected in a number of different rock-cut tombs, megalithic tombs, caves and rock shelters (supplementary materials, Supplementary Material online), were also reconstructed and analyzed. Unfortunately, they were from the cultural phases of Sardinia between the Neolithic and the Nuragic Final Bronze Age with radiocarbon datings in the range of 6.1 to 3.0 Kya, thus they could not shed further light on the issue of the potential pre-Neolithic presence of some SSHs on the island.

They harbored 21 distinct haplotypes in 19 sub-haplogroups belonging to macro-haplogroups R0, JT and U (supplementary table S7; supplementary fig. S4, Supplementary Material online). These haplotypes were compared with those from modern Sardinians and with 417 ancient mitogenomes available in the literature (supplementary table S5; supplementary fig. S5, Supplementary Material online). The sub-haplogroups observed in ancient Sardinians are also present in modern Sardinians at the same (N = 15) or at a very close (N = 4) level of haplogroup resolution (supplementary table S8, Supplementary Material online). None of the ancient Sardinian mitogenomes clustered within a Sardinian-specific haplogroup, but four were closely related (supplementary fig. S6, Supplementary Material online). A Sardinian Bell Beaker mitogenome (MA108) of ~4.3 Kya turned out to be a member of a novel branch (HV0j1: 6.6 ± 1.3 Ky), which was found in both a modern Sardinian and a continental Italian, and derives from a node (HV0j: 10.0 ± 2.1 Ky) from which two other Sardinian mitogenomes diverge (supplementary fig. S6, Supplementary Material online). The phylogenetic age of HV0j1 is thus fully compatible with the radiocarbon dating of MA108. A similar conclusion emerges in all other cases in which an informative phylogenetic link between modern and ancient Sardinian samples was established: (i) MA78 (Early Bronze Age, ~4.0 Kya) is a direct molecular ancestor of the Sardinian-specific haplogroup H3u2 (~3.2 Kya); (ii) MA104 (Early Bronze Age, ~4.5 Kya) harbors one of the diagnostic mutations of the SSH K1a32 (~6.8 Kya); and (iii) MA88 (Early Bronze Age, ~4.2 Kya) shows the mutational motif of haplogroup U5b2b5 node (~12.6 Kya), from which the SSH U5b2b5a (~3.2 Kya) as well as other mtDNAs from Sardinia, Italy and the UK descend (supplementary fig. S6, Supplementary Material online). Ancient DNA links extend beyond Sardinia. A Copper Age (~5 Kya) mitogenome from Northern Spain (ATP16 in Günther et al. 2015) identifies Iberia as the likely homeland of the molecular ancestor of the SSH X2c2a (supplementary fig. S6, Supplementary Material online), and indicates that the founder mtDNA arrived in Sardinia between 8.9 ± 1.7 Kya (age of X2c2) and 5.7 ± 1.8 Kya (age of X2c2a) (supplementary fig. S6, Supplementary Material online).

We also compared modern and ancient Sardinian mitogenomes with the mitogenome (haplogroup K1f) of the Late Neolithic/Chalcolithic Tyrolean Iceman (radiocarbon-dated to ~5.3 Kya) (Ermini et al. 2008). One Sardinian-specific haplogroup (K1g1), present in ~2.1% of Sardinians, is indeed related to Otzi's mitogenome as well as to other mitogenomes found in modern and ancient Europeans. However, the link is extremely distant in time, at the level of a very early node (K1 + 16362), which is only one mutation away from the root of K1 and is dated ~23.5 Ky (supplementary fig. S7, Supplementary Material online).

Origins of the Most Ancient Sardinian-Specific Haplogroups

As described above, our analyses raise the possibility that several SSHs may have already been present on the island prior to the Neolithic. The most plausible candidates would include K1a2d and U5b1i1, which together comprise almost 3% of modern Sardinians, and possibly others that might have arrived at an early date but expanded with the Neolithic. This scenario remains uncertain for two reasons: (i) K1a2d is the only SSH for which the standard errors of the coalescence ages never overlap with the arrival of the Neolithic in Sardinia; (ii) the possibility illustrated in panel D of figure 1 (the founder haplotype as well as a derived haplotype both moved to Sardinia, but were also both lost in the ancestral source), which, even if rather unlikely, can not be ruled out. However, at the same time our analyses show that the scenario of a pre-Neolithic presence of one or more SSHs in Sardinia cannot be easily dismissed either.

Such a scenario would not only support archeological evidence of a Mesolithic occupation of Sardinia, and recent genome-wide studies, but could also suggest a dual ancestral origin of its first inhabitants. K1a2d is of Late Paleolithic Near Eastern ancestry, whereas U5b1i1 harbors deep ancestral roots in Paleolithic Western Europe, possibly paralleling the patrilineal source of the very frequent (38.9%) Y-chromosome haplogroup I2a1a1-M26 both in terms of geography and timing (Francalacci et al. 2013).

Recent genome-wide data from ancient specimens have shown that Palaeolithic Europeans from ∼37 to ∼14 Kya derive from a single ancestral population, but that this long-term genetic continuity was in part interrupted by the appearance of a novel genetic component related to modern Near Easterners ∼14 Kya, during the first significant warming period (Bølling-Allerød interstadial) after the Last Glacial Maximum (LGM) (Fu et al. 2016). The notion of a possible genetic input from the Near East into and across Europe in the late Pleistocene prior to the arrival of the Early Neolithic material culture in Greece ∼8.5 Kya (Manning et al. 2014) is a novelty in human paleogenomics. Indeed, the prevailing conclusion of ancient DNA studies has been so far that Palaeolithic and Mesolithic hunter-gatherer European populations (characterized essentially only by haplogroups U8, U5 and U2 in terms of mitogenomes) differed genetically from early farmers, in turn implying that there was a wide-scale replacement across Europe from the Near East in the Early Neolithic, with limited assimilation of native Europeans (Pinhasi et al. 2012; Lazaridis et al. 2014; Omrak et al. 2016).

A pre-Neolithic genetic input from the Near East is, however, not a novelty in the context of phylogeographic studies of modern mtDNA variation. These have proposed that many modern European mtDNA lineages within haplogroups J, T, I, W and R0a, and possibly others, entered Europe in the Late Glacial and postglacial periods from the Near East before the migration waves associated with the onset of farming (Pala et al. 2012; Olivieri et al. 2013; Gandini et al. 2016; Richards et al. 2016); and that these haplogroups typical of modern Europeans, often assumed to have dispersed from Anatolia only with the advent of the Neolithic, were instead already present in Mesolithic Mediterranean Europe, particularly in Italy (Pereira et al. 2017). This scenario is in line with our findings concerning K1a2d in Sardinia as well as with the recent detection of two K1c mitogenomes in Mesolithic Greece (Hofmanová et al. 2016).

The potential genetic legacy of Mesolithic Sardinians could be even higher than the ∼3% represented by K1a2d and U5b1i1. We should stress that the assumptions of our analysis are highly conservative in this regard, because every migration of a lineage away from Sardinia to the continent would here be recorded instead as the signal of a Sardinian founder event, thus reducing the age of the SSH. For example, a large fraction of the SSHs within H1 and H3 (supplementary fig. S2, Supplementary Material online), the two most common haplogroups in modern Sardinians (18.5% and 18.4%, respectively; supplementary table S4, Supplementary Material online), are only one (sometimes fast-evolving) mutation away from the H1 and H3 founding nodes and/or have CEMs departing from the H1 and H3 nodes. Therefore, their estimated coalescence ages represent upper bounds for the presence of H1 and H3 mitogenomes in the island.

As shown in figure 2, the ages of H1 and H3 leave open the possibility that both were also present in Sardinia prior to the Neolithic. Notably, the frequency of H3 in Sardinia (18.4%) is the highest reported till now, and haplogroup H3 harbors a very peculiar geographical distribution. The highest frequencies are in western Mediterranean (Sardinians, Basques and other Iberians), with a sharp decrease towards Central and Eastern Europe and only very few occurrences in the Near East (fig. 4; supplementary table S9, Supplementary Material online), which founder analyses explain as recent incursions. Given that the population size trends for the Sardinian H3 mtDNAs indicate an expansion beginning between 9.0 and 10.5 Kya (fig. 4), it is tempting to link such an expansion to a pre-Neolithic arrival and diffusion of H3 on the island, most likely from a Western Mediterranean source, as previously suggested (Achilli et al. 2004; Torroni et al. 2006; Soares et al. 2010), possibly the same ancestral source of the ancestors of U5b1i1 and Y-chromosome haplogroup I2a1a1-M26 (Francalacci et al. 2015).

Our detection of potential pre-Neolithic signals in the mitogenome pool of contemporary Sardinians remains to be tested with studies of ancient DNA. If confirmed, a pre-Neolithic presence of H3 (and possibly also of H1 and other lineages, for example within JT) on the island alongside K1a2d and U5b1i1 (and most likely other lineages that we have not detected) would indicate a more substantial genetic legacy of Mesolithic Sardinians to the modern people of Sardinia. However, it is also important to realise that even if H3 (and H1) arrived in Sardinia only with the Neolithic, they most likely came from either Spain or elsewhere in the western Mediterranean, and not from the Near East. This would imply that they are likely the result of autochthonous west Mediterranean Mesolithic acculturation, in the wider European context.

In conclusion, contemporary Sardinians harbor a unique genetic heritage as a result of their distinct history and relative

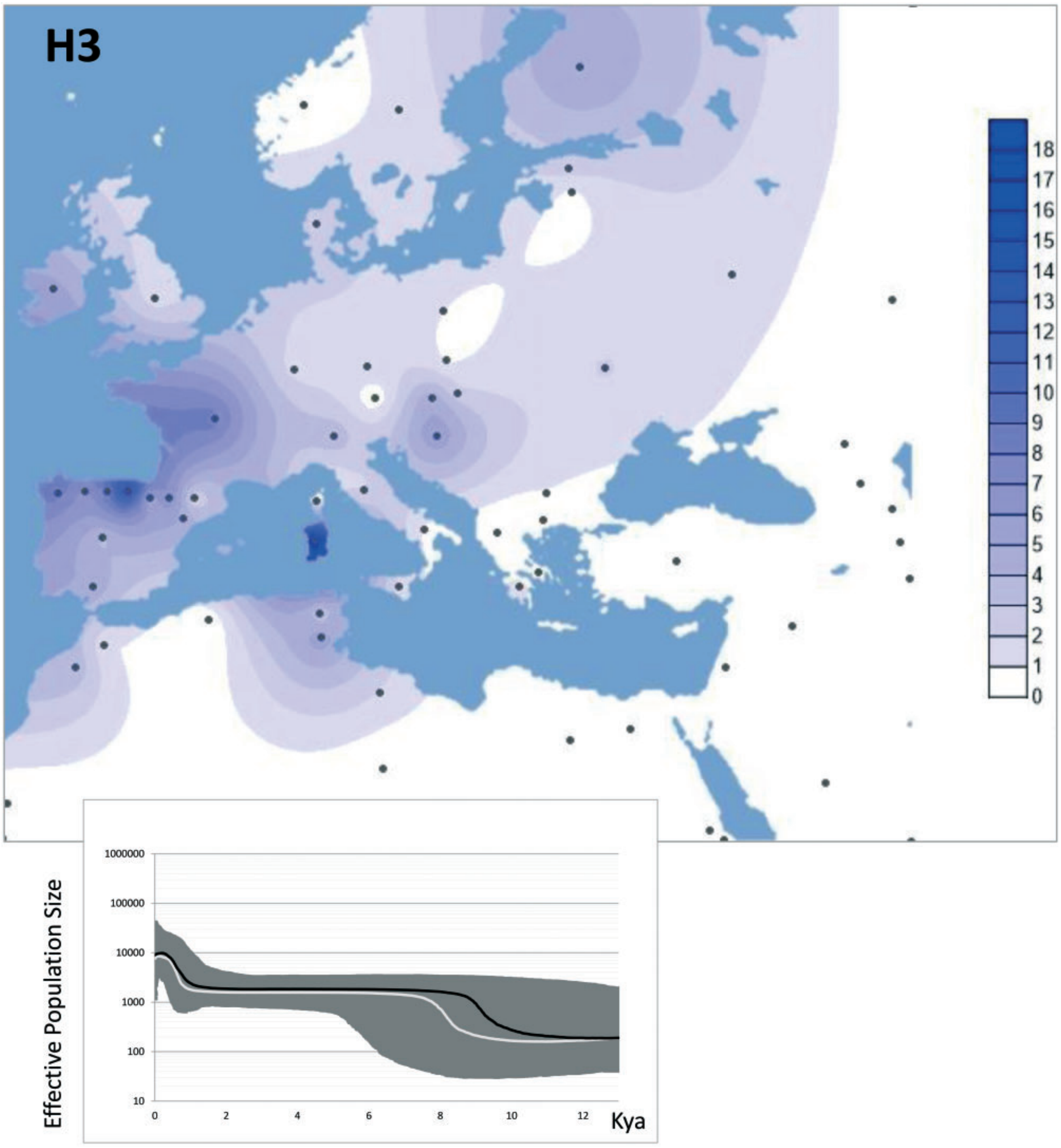


Fig. 4. Spatial frequency distribution map of haplogroup H3. Dots indicate the geographic locations of the surveyed populations. Population frequencies (%) are provided in supplementary table S9, Supplementary Material online. We constructed spatial frequency distribution plots with the program Surfer 9 (Golden Software, http://www.goldensoftware.com/products/surfer). The inset shows the Bayesian skyline plot (BSP) showing effective population size trends of Sardinian H3 mtDNAs. The black and white lines are the median estimates obtained by employing the mutation rates proposed by Soares et al. (2009) and Posth et al. (2016), respectively; the grey shading shows the highest posterior density limits.

isolation from the demographic upheavals of continental Europe. Whilst the major signal appears to be the legacy of the first farmers on the island, our results hint at the possibility that the situation might have been much more complex, both for Sardinia but also, by implication, for Europe as a whole. It now seems plausible that human mobility, intercommunication and gene flow around the Mediterranean from Late Glacial times onwards may well have left signatures that survive to this day.

Archeological evidence indicates that Mesolithic refugia persisted for many centuries in Italy and Iberia (Broodbank 2013), which, like the Near East and Caucasus, may have acted as a long-term refugial zone, as Gamble et al. (2004) has suggested. These populations may have contributed to varying extents to the ancestry of the populations of subsequent millennia, not only around the Mediterranean but also into the heart of the continent. Although in the past the stress has often been on the spread of the Neolithic, genetic studies too are beginning to emphasize the complexity and mosaic nature of human ancestry in the Mediterranean, and indeed in Europe more widely. Future work on ancient DNA should be able to test directly to what extent this more complex model is supported by genetic evidence, and whether our predictions of Mesolithic ancestry in contemporary Sardinians can be sustained.

Materials and Methods

All experimental and analytical procedures are described in the supplementary material file, Supplementary Material online.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to all the volunteers who generously participated in this study and made this research possible. This study received support from the University of Pavia strategic theme "Towards a governance model for international migration: an interdisciplinary and diachronic perspective" (MIGRAT-IN-G) (to A.O., A.A., O.S., A.T.), the British Academy Research Development Award n. 53097 (to R.S.), the Italian Ministry of Education, University and Research: Progetti Futuro in Ricerca 2012 (RBFR126B8I) (to A.O. and A.A.) and Progetti Ricerca Interesse Nazionale 2012 (to A.A., O.S. and A.T.). The sequencing data reported in the study were supported by the National Human Genome Research Institute grants HG005581, HG005552, HG006513, HG007022, and HG007089; the National Heart Lung and Blood Institute grant HL117626; the Intramural Research Program of the NIH, National Institute on Aging, with contracts N01-AG-1-2109 and HHSN271201100005C; the Sardinian Autonomous Region (L.R. no. 7/2009) grant cRP3-154. P.S. was supported by FCT, ESF and POPH through the FCT Investigator Programme (IF/01641/2013). P.S. and E.C.-S. acknowledge FCT IP and ERDF (COMPETE2020 – POCI) for the CBMA programme UID/BIA/04050/2013 (POCI-01-0145-FEDER-007569). T.R. was supported by an FCT grant (SFRH/BPD/108126/2015). M.P., P.S. and M.B.R. acknowledge FCT support through project PTDC/EPH-ARQ/4164/2014 partially funded by FEDER funds (COMPETE 2020 project 016899). M.B.R. received support from a Leverhulme Doctoral Scholarship programme. C.P. was supported by the Baden Württemberg Foundation and the Max Planck Society. The sequence data for the modern (N = 2,136) and ancient (N = 21) Sardinian mitogenomes are available in GenBank with accession numbers KY408145−KY410236, KY399164−KY399207, and KY399143−KY399163.

References

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, et al. 2004. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am J Hum Genet.* 75:910–918.

Behar DM, Harmant C, Manry J, van Oven M, Haak W, Martinez-Cruz B, Salaberria J, Oyharçabal B, Bauduer F, Comas D, et al. 2012. The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times. *Am J Hum Genet.* 90:486–493.

Berger J-F, Guilaine J. 2009. The 8200 cal BP abrupt environmental change and the Neolithic transition: a Mediterranean perspective. *Quat Int.* 200:31–49.

Broodbank C. 2013. The making of the middle sea: a history of the Mediterranean from the beginning to the emergence of the classical world. London: Thames & Hudson.

Costa MD, Pereira JB, Pala M, Fernandes V, Olivieri A, Achilli A, Perego UA, Rychkov S, Naumova O, Hatina J, et al. 2013. A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. *Nat Commun.* 4:2543.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton, New Jersey, USA: Princeton University Press.

Chiang CWK, Marcus JH, Sidore C, Al-Asadi H, Zoledziewska M, Pitzalis M, Busonero F, Maschio A, Pistis G, Steri M, et al. 2016. Population history of the Sardinian people inferred from whole-genome sequencing. *bioRxiv* doi: https://doi.org/10.1101/092148.

Dyson SL, Rowland RJ. 2007. Archaeology and history in Sardinia from the tone Age to the Middle Ages: shepherds, sailors, and conquerors. Philadelphia: University of Pennsylvania Press.

Ermini L, Olivieri C, Rizzi E, Corti G, Bonnal R, Soares P, Luciani S, Marota I, De Bellis G, Richards MB, et al. 2008. Complete mitochondrial genome sequence of the Tyrolean Iceman. *Curr Biol.* 18:1687–1693.

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, et al. 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* 341:565–569.

Francalacci P, Sanna D, Useli A, Berutti R, Barbato M, Whalen MB, Angius A, Sidore C, Alonso S, Tofanelli S, et al. 2015. Detection of phylogenetically informative polymorphisms in the entire euchromatic portion of human Y chromosome from a Sardinian sample. *BMC Res Notes* 8:174.

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A, et al. 2016. The genetic history of Ice Age Europe. *Nature* 534:200–205.

Gandini F, Achilli A, Pala M, Bodner M, Brandini S, Huber G, Egyed B, Ferretti L, Gómez-Carballa A, Salas A, et al. 2016. Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes. *Sci Rep.* 6:25472.

Gamble C, Davies W, Pettitt P, Richards M. 2004. Climate change and evolving human diversity in Europe during the last glacial. *Philos Trans R Soc Lond B Biol Sci.* 359:243–253.

Günther T, Valdiosera C, Malmström H, Ureña I, Rodriguez-Varela R, Sverrisdóttir ÓO, Daskalaki EA, Skoglund P, Naidoo T, Svensson EM, et al. 2015. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc Natl Acad Sci U S A.* 112:11917–11922.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207–211.

Hofmanová Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Díez-Del-Molino D, van Dorp L, López S, Kousathanas A, Link V, et al. 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc Natl Acad Sci U S A.* 113:6886–6891.

Hofmeijer KG, Alderliesten C, van der Borg K, Houston CM, de Jong AFM, Martini F, Sanges M, Sondaar PY, de Visser JA. 1989. Dating of the upper Pleistocene lithic industry of Sardinia. *Radiocarbon* 31:986–991.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. *Nature* 536:419–424.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409–413.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun.* 3:698.

Macaulay V, Richards MB. 2013. Mitochondrial genome sequences and their phylogeographic interpretation. In Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. doi: 10.1002/9780470015902.20843.pub2.

Manning K, Timpson A, Colledge S, Crema E, Edinborough K, Kerig T, Shennan S. 2014. The chronology of culture: a comparative assessment of European Neolithic dating approaches. *Antiquity* 88:1065–1080.

Olivieri A, Pala M, Gandini F, Hooshiar Kashani B, Perego UA, Woodward SR, Grugni V, Battaglia V, Semino O, Achilli A, et al. 2013. Mitogenomes from two uncommon haplogroups mark late glacial/postglacial expansions from the near east and neolithic dispersals within Europe. *PLoS One* 8:e70492.

Omrak A, Günther T, Valdiosera C, Svensson EM, Malmström H, Kiesewetter H, Aylward W, Storå J, Jakobsson M, Götherström A. 2016. Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. *Curr Biol.* 26:270–275.

Pala M, Achilli A, Olivieri A, Hooshiar Kashani B, Perego UA, Sanna D, Metspalu E, Tambets K, Tamm E, Accetturo M, et al. 2009. Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. *Am J Hum Genet.* 84:814–821.

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, Tamm E, Karmin M, Reisberg T, Hooshiar Kashani B, et al. 2012. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. *Am J Hum Genet.* 90:915–924.

Pereira JB, Costa MD, Vieira D, Pala M, Bamford L, Harrich N, Cherni L, Alshamali F, Hatina J, Rychkov S, Stefanescu G, et al. 2017. Reconciling evidence from ancient and contemporary genomes: a major source for the European Neolithic within Mediterranean Europe. *Proc Biol Sci*. Forthcoming.

Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J. 2012. The genetic history of Europeans. *Trends Genet.* 28:496–505.

Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, Valentin F, Thevenet C, Furtwängler A, Wißing C, et al. 2016. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. *Curr Biol.* 26:827–833.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, et al. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet.* 67:1251–1276.

Richards MB, Soares P, Torroni A. 2016. Palaeogenomics: mitogenomes and migrations in Europe's past. *Curr* . 26:R243–R246.

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barać L, Peričić M, Balanovsky O, et al. 2004. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am J Hum Genet.* 75:128–137.

Shackleton JC, van Andel TH, Runnels CN. 1984. Coastal paleogeography of the Central and Western Mediterranean during the last 125,000 years and its archaeological implications. *J Field Archaeol.* 11:307–314.

Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F, et al. 2015. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet.* 47:1272–1281.

Sikora M, Carpenter ML, Moreno-Estrada A, Henn BM, Underhill PA, Sánchez-Quinto F, Zara I, Pitzalis M, Sidore C, Busonero F, et al. 2014. Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the Genetic structure of Europe. *PLoS Genet.* 10:e1004353.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet.* 84:740–759.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H-J, Torroni A, Richards MB. 2010. The archaeogenetics of Europe. *Curr Biol.* 20:R174–R183.

Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, Mormina M, Brandão A, Fraser RM, Wang TY, et al. 2016. Resolving the ancestry of Austronesian-speaking populations. *Hum Genet.* 135:309–326.

Sondaar PY. 1998. Palaeolithic Sardinians: paleontological evidence and methods. In Balmuth MS, Tykot RH, editors. Sardinian and Aegean chronology. Oxford, UK: Oxbow Books. p. 45–51.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. 2006. Harvesting the fruit of the human mtDNA tree. *Trends Genet.* 22:339–345.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 1 December 2016
Accepted: 22 May 2017
Published online: 05 July 2017

Whole Mitogenomes Reveal the History of Swamp Buffalo: Initially Shaped by Glacial Periods and Eventually Modelled by Domestication

S. Wang[1], N. Chen[1], M. R. Capodiferro[2], T. Zhang[3], H. Lancioni[4], H. Zhang[5], Y. Miao[6], V. Chanthakhoun[7], M. Wanapat[8], M. Yindee[9], Y. Zhang[10], H. Lu[3], L. Caporali[11], R. Dang[1], Y. Huang[1], X. Lan[1], M. Plath[1], H. Chen[1], J. A. Lenstra[12], A. Achilli[2] & C. Lei[1]

The newly sequenced mitochondrial genomes of 107 Asian swamp buffalo (*Bubalus bubalis carabensis*) allowed the reconstruction of the matrilineal divergence since ~900 Kya. Phylogenetic trees and Bayesian skyline plots suggest a role of the glacial periods in the demographic history of swamp buffalo. The ancestral swamp-buffalo mitogenome is dated ~232 ± 35 Kya. Two major macro-lineages diverged during the 2nd Pleistocene Glacial Period (~200–130 Kya), but most (~99%) of the current matrilines derive from only two ancestors (SA1′2 and SB) that lived around the Last Glacial Maximum (~26–19 Kya). During the late Holocene optimum (11–6 Kya) lineages differentiated further, and at least eight matrilines (SA1, SA2, SB1a, SB1b, SB2a, SB2b, SB3 and SB4) were domesticated around 7–3 Kya. Haplotype distributions support an initial domestication process in Southeast Asia, while subsequent captures of wild females probably introduced some additional rare lineages (SA3, SC, SD and SE). Dispersal of domestic buffaloes created local population bottlenecks and founder events that further differentiated haplogroup distributions. A lack of maternal gene flow between neighboring populations apparently maintained the strong phylogeography of the swamp buffalo matrilines, which is the more remarkable because of an almost complete absence of phenotypic differentiation.

Water buffalo (*Bubalus bubalis*) is one of the most important livestock species in several Asian countries and is used for the production of milk and meat and for draft power in rice cultivation. The domestic water buffalo in Asia is generally divided in two major subspecies, the dairy river buffalo and the draft swamp buffalo, which differ in morphology, behavior and number of chromosomes[1,2]. The river buffalo is found in the Indian subcontinent,

[1]College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi, 712100, China. [2]Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università di Pavia, Pavia, 27100, Italy. [3]School of Bioscience and Engineering, Shaanxi University of Technology, Hanzhong, Shaanxi, 723000, China. [4]Dipartimento di Chimica, Biologia e Biotecnologie, Università di Perugia, Perugia, 06123, Italy. [5]Key Laboratory of Plateau Lake Ecology and Global Change, College of Tourism and Geography, Yunnan Normal University, Kunming, Yunnan, 650500, China. [6]Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming, Yunnan, 650201, China. [7]Department of Animal Science, Faculty of Agriculture and Forest Resource, Souphanouvong University, Luang Prabang, Laos. [8]Tropical Feed Resources Research and Development Center, Department of Animal Science, Faculty of Agriculture, Khon Kaen University, Khon Kaen, 40002, Thailand. [9]Department of Clinical Science and Public Health, Faculty of Veterinary Science, Mahidol University, Kanchanaburi campus, Kanchanaburi, 71150, Thailand. [10]National Engineering Laboratory for Animal Breeding, Key Laboratory of Aniaml Genetics and Breeding and Reproduction of MOA, College of Animal Science and Technology, China Agricultural University, Beijing, 100193, China. [11]IRCCS Institute of Neurological Sciences of Bologna, Bologna, 40139, Italy. [12]Faculty of Veterinary Medicine, Utrecht University, Yalelaan 104, 3584 CM, Utrecht, The Netherlands. S. Wang, N. Chen, M. R. Capodiferro, A. Achilli and C. Lei contributed equally to this work. Correspondence and requests for materials should be addressed to A.A. (email: alessandro.achilli@unipv.it) or C.L. (email: leichuzhao1118@126.com)

South Asia and the Mediterranean area (Italy, Egypt and the Balkans), and sporadically in Australia and South America, whereas the swamp buffalo is kept in Northeast India, China (southern regions and Yangtze valley) and Southeast Asia[1,3]. Both types of water buffalo descend from the wild Asian buffalo (*Bubalus arnee*)[4], which had a widely distribution range in eastern Indian, Sri Lanka and Southeast Asia until the beginning of XIX century[5–8]. Lau *et al.*[7] hypothesized that the wild Asian buffalo originated in mainland of Southeast Asia and spread north toward China and west toward the Indian subcontinent, where the river type was probably domesticated.

Mitochondrial DNA (mtDNA), Y-chromosomal and nuclear microsatellite data showed a deep genetic divergence of swamp and river buffalo[2,5–7,9–16], which indicate two independent domestications[5,17]. Domestication of river buffalo most likely took place in the Indian subcontinent[15], whereas swamp buffalo was proposed to originate from the border region between south China and north Indochina[6,17]. However, the fragmentary buffalo mtDNA sequences from rDNA, COII, and *cytochrome b* loci reported to date[5–7,18], confound quantitative inferences of population history. So far two swamp buffalo mitogenomes (NC006295/AY702618 and JN632607) and one river buffalo mitogenome (AF547270) were deposited in GenBank. In this study, we report the mitogenomes from an additional 107 Southeast-Asian swamp buffaloes, covering most of its current geographic distribution with as outgroup one mitogenome from a Chinese river buffalo in order to establish the haplogroup phylogeny and reconstruct the demographic history of the swamp buffalo.

Results

Sequence variation of swamp buffalo mitogenomes.

The 109 swamp buffalo mitogenomes with a length of 16340 to 16363 bps belong to 87 different haplotypes (Ht.s) and are divided into 21 haplogroups or subhaplogroups (Supplementary Dataset S1). The swamp haplotypes (Hd: 0.992) contain 362 polymorphic sites (π: 0.422) with a pairwise nucleotide difference of 69.0 ± 16.5 and a synonymous/non-synonymous ration of 4.63. Similar values have been reported for other livestock and human mitochondrial genomes[19–21]. Seventy-five swamp haplotypes are observed once, while the most frequent haplotypes HT22 and HT55 occurred five and seven times, respectively. Forty-eight haplotypes with 59 sequences belong to lineage SA and 34 haplotypes with 44 sequences belong to lineage SB. The remaining 5 haplotypes belong to the rare haplogroups SC (2), SD (2) or SE (1). The two river buffalo mitogenomes belong to clades R1 and R2[1,6].

Swamp Buffalo Mitochondrial Phylogeny.

A maximum parsimony (MP) tree based on the 111 water buffalo mitogenomes (89 haplotypes) confirms two distinct branches, river and swamp buffalo, which were separated by 300 substitutions (Supplementary Dataset S2). The river buffalo branch includes 2 haplotypes belonging to lineages R1 and R2. The remaining 87 swamp buffalo haplotypes cluster into five divergent haplogroups (Hg.s), namely SA, SB, SC, SD and SE, with an overwhelming representation of SA (54.1%) and SB (40.4%). We largely confirm previous phylogenies[6,10], but also reveal the novel haplogroups SA3 and SB4 and many different subhaplogroups. Lineages SA and SB are divided into three (SA1 to SA3; SA1′2 as an ancestral node) and four (SB1-SB4; SB2′3′4 as an ancestral node) sublineages, respectively. A single control-region transition at position 16066 defines a major (32.1%) star-like subclade SA1a. For the three rare lineages SC (1.8%), SD (2.8%) and SE (0.9%), the complete mtDNA sequences confirm that the split-off of lineage SC preceded the SA-SB divergence and that SD is a sister clade of SB, but indicate that also SE and SA are sister clades. Maximum likelihood (ML) and Bayesian evolutionary analysis of sampling trees (Beast) retrieved remarkably similar tree topologies (Supplementary Figs S1 and S2).

Molecular Clocks and Age Estimates.

Previously reported estimates of the divergence time between river and swamp range from 10 Kya to 1.7 Mya[2,5,7,9–11,13–16,18,22,23], partially because different mtDNA segments were analyzed and different evolution rates were applied. In the present study, we phylogenetically compared 111 water buffalo mitogenomes with one African buffalo (*Syncerus caffer*; NC020617), while rooting our MP tree with one *Bos taurus* (V00654.1) and one ancient *Bos primigenius* (GU985279) mitogenome. The divergence pattern evaluated on all 114 bovine mitogenomes (Supplementary Fig. S3) confirms saturation of the D-loop divergence, emphasizing that sequencing of whole mitogenomes is essential for quantification.

We first considered the synonymous mutations for a ML estimation of the molecular age of phylogenetic nodes (Fig. 1). Using a fossil age estimate of the *Bovini* tribe of 8.8 Mya[24] and the age (6.7 Ky) of the ancient *Bos primigenius* mtDNA[25] we calculated a rate of $3.75 \pm 0.47 \times 10^{-5}$ synonymous substitutions per nucleotide per Ky for 3790 amino acid codons equalling 1 synonymous substitution every ~7.030 Ky. This yields divergence times of ~913 ± 78 Kya for the water buffalo mitogenome, ~82 ± 18 Kya for the only two river buffalo R1 and R2 haplotypes and ~232 ± 35 Kya for the swamp buffalo haplogroups (Table 1 and Fig. 1). At least 12 different mtDNA ancestral haplotypes of the current swamp buffaloes were present in Southeast Asia during the early Neolithic (11–6 Kya), which overlaps with the initial phase of domestication[26]. These haplotypes were the ancestors of the eight (sub)haplogroups SA1, SA2, SB1a, SB1b, SB2a, SB2b, SB3 and SB4, still common in modern herds, and of the rare SA3, SC, SD and SE.

Time estimates were confirmed by (i) using all open reading frame mutations, (ii) considering all mutations partitioned in coding and control regions and evaluated with both ML and Beast (Table 1). However, slightly deleterious mutations within the open reading frames may lead to overestimations of younger clades (Supplementary Fig. S4)[19,21]. The overall mutation rate was estimated at $2.11 \pm 0.34 \times 10^{-8}$ substitutions per nucleotide per year (1 mutation every ~2900 years) over the entire mitogenome. As for the river buffalo internal variation, a better estimate will be obtained by analyzing more mitogenomes.

Estimating Past and Present Demographic Trends.

Bayesian skyline plot (BSP) of swamp buffalo mitogenomes shows three major changes in the effective female population size: i) a slight decrease between about 200 and 130 Kya; ii) a more recent decrease starting around 25–20 Kya and much steeper during the early

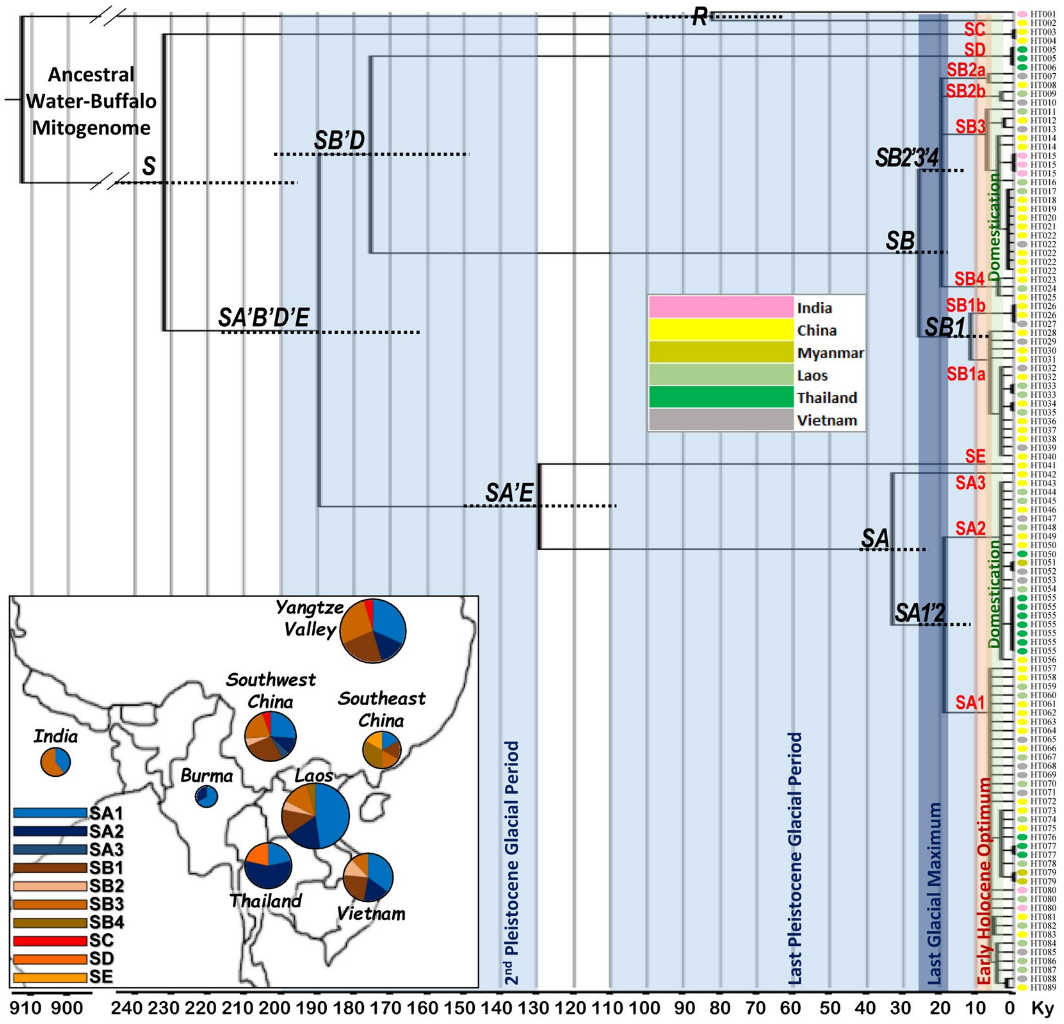


Figure 1. Phylogeny of complete mtDNAs from 111 buffalo mitogenomes. The topology was inferred by maximum parsimony (Supplementary Dataset S2). A maximum likelihood time scale, based on synonymous substitutions, is indicated below the tree. The standard error for the major nodes are represented by dot lines, further details are available in Table 1. Samples are indicated by 89 different haplotype IDs (Supplementary Dataset S1). The insert shows the geographic distribution of major haplogroups based on these complete mtDNAs. Samples from China have been divided into three regions (Yangtze Valley, Southwest China and Southeast China). The map has been drawn by hand in Adobe Photoshop (v. 8.0; http://www.adobe.com).

Neolithic (11–6 Kya); and iii) a rapid increase from 3 Kya (Fig. 2). This recent increase explains the star-like topology of some haplogroups (e.g. SA1, SA2, SB1a, SB3 and SB4; Fig. 1) and is very similar to previously analysis of water buffalo and other bovine domestic species[27].

An analysis of the geographic distribution of swamp haplogroups in Southeast Asia, based on the control-region data currently available in previous literature[6] or deposited in GenBank (Supplementary Table S1) confirms the prevalence of SA or SB mtDNAs (>99%) and a geographic differentiation of subhaplogroups with contrasting geographic distributions of the subhaplogroups SA2, SB1, SB2 and SB3 (Supplementary Fig. S5). The rare haplogroup SC was previously reported to occur in Thailand, Bangladesh and sporadically in Southwest China, while SD and SE were found only in Thailand[6]. We confirmed the presence of SC in the Southwestern Chinese Dehong population, but the haplotypes SC and SE were also found in the Yibin and Poyanghu breeds from the Yangtze Valley (inset in Fig. 1).

Discussion

We sequenced the complete mitogenome of the swamp buffalo in order to reconstruct the phylogenetic relationships of the mitochondrial haplotypes and to obtain a time scale for the phylogeny. Most of the major haplogroups were already identified in previous studies[2,5,7,9–11,13–16,18,22,23], but we recognized the novel haplogroups SA3 and SB4 and defined the subhaplogroups SA1a, SA1a1, SA1a2, SA1a3, SB1a, SB1a1, SB1a2, SB1b, SB2a, SB2b, SB3a, SB3a1, SD1 and SD2 (Supplementary Dataset S1 and S2).

During the glacial periods, drastic changes in ecological and climatic seasons had consecutive major effects on the distribution of plants and animals[28,29]. As part of the Indo-Pacific Warm Pool, Southeast Asia was relatively

Node	N	ML (synonymous subst)		ML (only coding region)		ML (all substitutions)[a]		Beast (all substitutions)[a]	
		T(ky)	SE[b](ky)	T(ky)	SE[b](ky)	T(ky)	SE[b](ky)	T(ky)	SE[b](ky)
A.W.M.[c]	111	912.6	78.3	1044.3	71.4	721.6	59.1	672.0	101.7
River	2	81.9	18.3	109.3	20.7	68.6	11.6	63.7	13.2
Swamp	109	231.8	35.3	280.1	28.8	204.4	20.0	194.2	31.4
SC	2	0.0	59.1	3.8	3.8	1.5	1.5	2.9	1.4
SA′B′D′E	107	189.6	27.2	222.9	25.0	176.6	17.7	166.6	26.8
SA′E	60	129.3	21.2	152.7	20.9	129.1	15.5	120.6	20.7
SE	1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
SA	59	33.0	9.2	40.5	9.9	37.5	7.3	35.9	7.9
SA1′2	56	18.8	6.5	23.9	7.4	18.4	4.5	18.5	4.6
SA1	37	6.4	4.2	8.1	5.4	7.2	1.8	8.4	2.2
SA1a	35	6.4	1.5	8.1	1.7	5.5	0.9	6.7	1.5
SA1a1	3	5.2	1.6	6.8	1.8	4.8	1.0	4.1	1.2
SA1a2	6	4.4	1.7	5.7	1.9	4.5	1.1	4.2	1.0
SA1a3	9	3.3	2.4	4.4	2.5	3.9	1.2	4.2	1.0
SA2	21	3.3	1.5	3.6	1.6	5.1	1.2	7.0	1.8
SA3	1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
SB′D	47	175.5	27.1	210.0	25.4	157.7	16.9	146.4	24.3
SD	3	0.0	44.6	0.0	44.8	0.9	1.0	3.2	1.3
SB	44	25.8	7.0	35.1	8.3	30.7	5.9	31.1	6.8
SB1	18	11.4	6.2	12.2	4.9	8.6	2.4	8.9	2.7
SB1a	15	6.4	4.4	8.0	2.7	5.8	1.4	6.0	1.4
SB1a1	11	3.3	2.0	4.2	2.1	3.7	1.2	4.5	1.0
SB1a2	2	6.4	4.5	6.5	2.4	4.9	1.4	3.8	1.0
SB1b	3	0.0	13.0	0.0	13.1	1.2	1.4	3.5	1.3
SB2′3′4	26	19.8	5.7	26.1	6.7	23.3	4.7	23.4	5.3
SB2	4	19.8	10.6	26.1	12.5	15.7	4.5	10.8	4.0
SB2a	2	3.1	3.1	6.7	4.8	6.0	2.9	3.9	1.4
SB2b	2	6.8	6.8	7.0	7.0	9.9	4.5	5.2	2.3
SB3	19	6.9	3.4	7.2	3.5	4.7	1.8	6.6	1.7
SB3a	16	4.1	3.1	4.6	3.0	2.4	1.0	5.0	1.1
SB3a1	10	1.2	0.9	1.8	1.1	1.4	0.6	3.9	0.8
SB4	3	4.1	2.9	8.7	4.2	5.1	2.2	4.7	1.7

Table 1. Age estimates of major buffalo branches based on different mitochondrial datasets. [a]The entire genome was partitioned into coding and control region. [b]The 95% Confidence Interval (CI) corresponds to 1.96 times the value of the Standard Error (SE) reported here. [c]Ancestral Water-Buffalo Mitogenome.

warm during the Last Glacial Maximum (~26–19 Kya)[30] with a temperature decrease of only 2.5 °C *vs* 5–10 °C globally[31, 32]. This made Southeast Asia a major global biodiversity hotspot[33] in which many species survived (in glacial refuges) the fluctuations of temperature and forest coverage during the Pleistocene (for the latter, see Supplementary Fig. S6)[28].

Estimations of divergence times are inherently imprecise[34], but the accuracy of our data has been optimized by using data from the aurochs as internal calibration point, this in addition to a fossil age of the *Bovini* tribe of 7–11 Mya[24]. According to our estimate, the divergence of the swamp and river types of the water buffalo took place almost at the beginning of a glacial period (~900 to 860 Kya). Remarkably, from the swamp matrilineal diversity that must have formed until 200 Kya only the minor haplogroups SC and the ancestor of all other haplogroups (SA′B′D′E) have survived. We propose to divide the last 200 ka into five phases, correlating glacial periods and estimates of demographic/phylogenetic history (Figs 1 and 2 and Supplementary S6).

1. During 2nd Pleistocene Glacial Period (~200 to 130 Kya) the first decline of population was observed in the BSP while two macro-haplogroups (SA′E and SB′D) diverged (Figs 1 and 2).
2. The first phase of the last Pleistocene glacial period (~110 to 50 Kya) was still comparatively moderate, the population size remained almost unchanged and only one divergence event (SA-SE) has been identified.
3. During the second phase of the last Pleistocene glacial period (~50 to 11 Kya) the population began to decline. The current demographic composition of swamp populations shows that ~99% of current mitogenomes are derived from only two ancestral haplotypes (SA1′2 and SB), both dated around the LGM (~26–19 Kya). Afterwards, the major haplogroups SA1′2 and SB differentiated into 8 haplotypes (SA1, SA2, SB1a, SB1b, SB2a, SB2b, SB3 and SB4).
4. After 11 Kya the increasing temperature raised the sea level, which had a profound impact in the regions

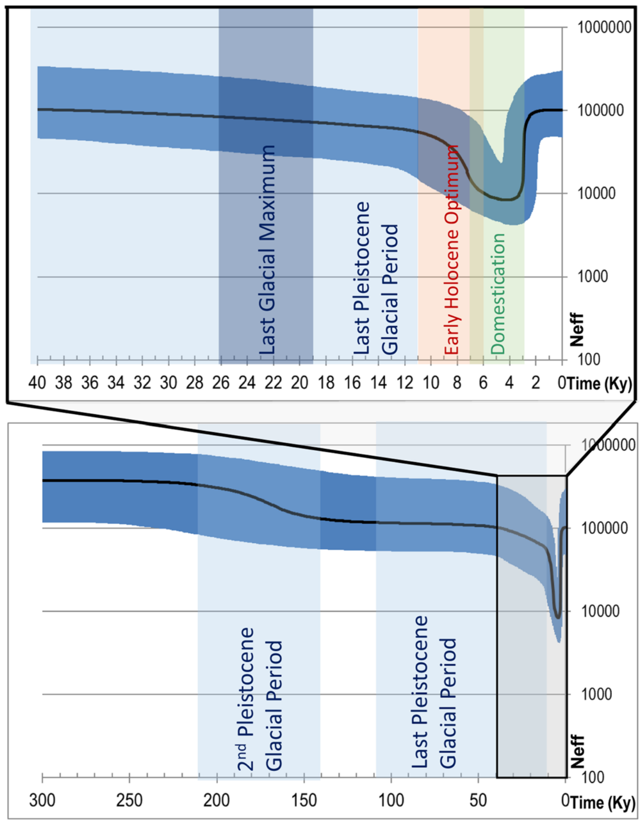


Figure 2. Bayesian skyline plot showing the swamp buffalo population size trend. The Y axis indicates the effective number of females, as inferred from our mitogenome dataset considering a generation time of six years[49]. The black solid line is the median estimate and the blue shading shows the 95% highest posterior density limits.

from Sundaland to Southeast China[35]. Paleoenvironmental data on the Holocene indicate a warm period between 11 and 6 Kya in southern China, known as the early Holocene optimum[36]. This overlapped with the first phases of the rice cultivation, which is believed to have triggered the domestication of the swamp buffalo at 7–3 Kya and an expansion of the water buffalo population[1, 6, 10, 14, 26, 37].

5. Finally, we observe a rapid increase since 3 Kya to the present population size. As previously proposed[27], this was due to expansion of the domestic buffalo to the large current distribution range, harboring the several present populations with distinct haplogroup distributions.

Thus, the demographic history of swamp buffalo seems to be linked to the historic glacial events, establishing isolated refugia of swamp buffalo in which only a limited number of subhaplogroups survived. This may explain why only one divergence event (SA-SE) has been dated to the period of 180–40 Kya. The current demographic composition of swamp populations suggests that ~99% of current mitogenomes are derived from only two ancestors (SA1′2 and SB), which are both dated around the Last Glacial Maximum (~26–19 Kya). The time estimates in Fig. 1 and Table 1 further indicate that also the divergence of the current domestic subhaplogroups SA1, SA2, SB1a, SB1b, SB2a, SB2b, SB3 and SB4 preceded domestication, which gives 8 haplotypes as a minimum estimate of the swamp buffalo diversity captured by the first farmers.

Our account of the pre-domestic history of the swamp buffalo provides a context to previous studies of the diversity of mtDNA control region and cytochrome *b* gene[6, 10], which has been summarized in Supplementary Fig. S5. The high diversity of domestic SA and SB haplotypes in the China-Vietnam border region was proposed as evidence supports an initial major domestication event of swamp buffalo in Southeast Asia, probably between southern China and Vietnam. The finding of SC, SD and SE haplotypes almost exclusively in Thailand and Bangladesh suggests incorporation of these haplotypes after the domestic buffaloes had reached the west bank of the Mekong river[6] in a scenario of recurrent restocking the domestic population with wild females as proposed previously for the horse[20, 38, 39]. Extending the analysis to other loci or even to the entire genome and also a wider sampling covering the entire geographic range of the swamp buffalo is desirable to unravel further the domestication and subsequent demographic history of swamp buffalo and to enable an interesting comparison with the related river buffalo[6].

Methods

Sample Collection.

Most of the samples used for this work were already collected from previous collaborative works[5, 10]. All samples were already classified into mtDNA haplogroups based on control-region data. We selected 107 mtDNA for complete sequencing in order to represent all swamp lineages and to include the highest possible molecular variability avoiding potential redundancies. An additional mitogenome from river buffalo was sequenced to be used as an outgroup.

Mitogenome sequencing. DNA was extracted[10] from 107 swamp buffaloes (blood, ear tissue and hair follicle) from China (46), Laos (23), Myanmar (3), Thailand (14), Vietnam (16) and India (5) and one Chinese river buffalo. Complete mitogenome sequences were obtained by using two different approaches: 1) PCR amplification (with 27 primer pairs, Supplementary Table S2A)[40] and Sanger sequencing; 2) Long-Range PCR amplification (Supplementary Table S2B) and Illumina sequencing[41]. GenBank accession number, sequencing method, coverage and depth of each sample are reported in Supplementary Dataset S1. MtDNA genome sequences were analyzed using DNASTAR 7.0, Sequencher v5, DNAsp v5, Clustal X and GeneSyn packages.

All experimental procedures were performed in accordance with the Regulations for the Administration of Affairs Concerning Experimental Animals approved by the State Council of People's Republic of China. The study was approved by Institutional Animal Care and Use Committee of Northwest A&F University (Permit Number: NWAFAC1019).

Phylogeny Construction and Demographic Inferences. The phylogeny construction was performed following a maximum parsimony (MP) criterion by hand and confirmed using an adapted version of mtPhyl4.015[42], as previously described[20, 21, 43, 44]. The modified.txt files to be loaded in the program are available upon request. The tree was rooted on the *Bos taurus* reference sequence (V00654.1) and on the ancient *Bos primigenius* mtDNA (GU985279). A maximum likelihood (ML) tree was computed using MEGA7.0[45] with 1000 bootstrapping replicates.

A first ML analysis was performed using PAML X[46] by considering only synonymous mutations in the protein coding genes. The ND6 gene was reverse-complemented to present the same reading direction as the other genes and non-synonymous substitutions were replaced with the ancestral base pairs. Stop codons were excluded from the analysis. Total lengths of coding genes were joined together and the final alignment (11370 bps/3790 codons long) was analyzed with CODEML to calculate a synonymous mutation rate. A second tree was calculated in the same way, but considering all coding mutations. The third and fourth trees with molecular ages were calculated by PAML X and BEAST v. 1.8.3 software[47], respectively, while considering two partitions in the molecule corresponding to the coding (including all genes coding for mRNA, rRNA and tRNA) and control regions. Modelgenerator v.85 indicated for our dataset HKY + G + I as the best-supported model according to the AIC2 and BIC criterions. These substitution and site heterogeneity models with 8 gamma categories – the lowest number significantly increasing (>1.0) the likelihood – were selected for the subsequent ML and BEAST estimates. The generalized likelihood ratio statistic was always used to verify the clock hypothesis. In order to calibrate the molecular clock, we built a *Bovini* tree by including one African Buffalo (NC020617) and two *Bos* mitogenomes (one *Bos taurus*, V00654; one ancient *Bos primigenius*, GU985279) used as an outgroup. For the calibration point we used the estimated archaeological age of the *Bovini* tribe (8.8 ± 1.1 My; 95% CI: 7–11 My)[24]. Since multiple calibration points are preferable[24], the age of the ancient *Bos primigenius* (6.7 ± 0.2 Ky; 95% CI: 6.3–7.1 ky) was also used as an internal (recent) calibration point. The major haplogroups were considered as monophyletic in order of being able to calculate their age estimates. The analyses were also repeated excluding the aurochs sequence, but the estimates changed by only ~4% on average. We then obtained a Bayesian skyline plot (BSP)[48] from the swamp buffalo phylogeny by running 50,000,000 iterations with samples drawn every 10,000 steps. We constructed spatial frequency distribution plots with the program Surfer 9 (Golden Software, http://www.goldensoftware.com/products/surfer) by using the control-region data currently available in previous literature[6] or deposited in GenBank (Supplementary Table S1).

Data accessibility. Sequences of the novel water buffalo mitogenomes have been deposited in GenBank under accession numbers KX758295 - KX758402 (108 complete mtDNAs).

References

1. Cockrill, W. R. The water buffalo: a review. *Br. Vet. J.* **137**, 8–16 (1981).
2. Kumar, S. *et al.* Mitochondrial DNA analyses of Indian water buffalo support a distinct genetic origin of river and swamp buffalo. *Anim. Genet.* **38**, 227–232, doi:10.1111/j.1365-2052.2007.01602.x (2007).
3. FAO. http://dad.fao.org/ (2014).
4. Cockrill, W. R. The husbandry and health of the domestic buffalo. (Food and agricultural organization of the United nations, Rome 1974).
5. Lei, C. Z. *et al.* Independent maternal origin of Chinese swamp buffalo (*Bubalus bubalis*). *Anim. Genet.* **38**, 97–102, doi:10.1111/j.1365-2052.2007.01567.x (2007).
6. Zhang, Y. *et al.* Strong and stable geographic differentiation of swamp buffalo maternal and paternal lineages indicates domestication in the China/Indochina border region. *Mol. Ecol.* **25**, 1530–1550, doi:10.1111/mec.13518 (2016).
7. Lau, C. H. *et al.* Genetic diversity of Asian water buffalo (*Bubalus bubalis*): mitochondrial DNA D-loop and cytochrome b sequence variation. *Anim. Genet.* **29**, 253–264, doi:10.1046/j.1365-2052.1998.00309.x (1998).
8. Pandya, P. *et al.* Bacterial diversity in the rumen of Indian Surti buffalo (*Bubalus bubalis*), assessed by 16S rDNA analysis. *J. Appl. Genet.* **51**, 395–402 (2010).
9. Mishra, B. P. *et al.* Genetic analysis of river, swamp and hybrid buffaloes of north-east India throw new light on phylogeography of water buffalo (*Bubalus bubalis*). *J. Anim. Breed. Genet.* **132**, 454–466, doi:10.1111/jbg.12141 (2015).
10. Yue, X.-P. *et al.* Phylogeography and Domestication of Chinese Swamp Buffalo. *PLoS One* **8**, e56552, doi:10.1371/journal.pone.0056552 (2013).
11. Barker, J. S. F. *et al.* Genetic diversity of Asian water buffalo (*Bubalus bubalis*): microsatellite variation and a comparison with protein-coding loci. *Anim. Genet.* **28**, 103–115, doi:10.1111/j.1365-2052.1997.00085.x (1997).
12. Navani, N., Jain, P. K., Gupta, S., Sisodia, B. S. & Kumar, S. A set of cattle microsatellite DNA markers for genome analysis of riverine buffalo (*Bubalus bubalis*). *Anim. Genet.* **33**, 149–154, doi:10.1046/j.1365-2052.2002.00823.x (2002).

13. Barker, J. S. F., Tan, S. G., Selvaraj, O. S. & Mukherjee, T. K. Genetic variation within and relationships among populations of Asian water buffalo (*Bubalus bubalis*). *Anim. Genet.* **28**, 1–13, doi:10.1111/j.1365-2052.1997.00036.x (1997).
14. Kierstein, G. *et al.* Analysis of mitochondrial D-loop region casts new light on domestic water buffalo (*Bubalus bubalis*) phylogeny. *Mol. Phylogenet. Evol.* **30**, 308–324, doi:10.1016/S1055-7903(03)00221-5 (2004).
15. Kumar, S., Nagarajan, M., Sandhu, J. S., Kumar, N. & Behl, V. Phylogeography and domestication of Indian river buffalo. *BMC Evol. Biol.* **7**, 186, doi:10.1186/1471-2148-7-186 (2007).
16. Lei, C. *et al.* Two maternal lineages revealed by mitochondrial DNA D-loop sequences in Chinese native water buffaloes (*Bubalus bubalis*). *Asian-Australas. J. Anim. Sci.* **20**, 471, doi:10.5713/ajas.2007.471 (2007).
17. Yindee, M. *et al.* Y-chromosomal variation confirms independent domestications of swamp and river buffalo. *Anim. Genet.* **41**, 433–435, doi:10.1111/j.1365-2052.2010.02020.x (2010).
18. Amano, T., Miyakoshi, Y., Takada, T., Kikkawa, Y. & Suzuki, H. Genetic variants of ribosomal DNA and mitochondrial DNA between swamp and river buffaloes. *Anim. Genet.* **25**, 29–36, doi:10.1111/j.1365-2052.1994.tb00400.x (1994).
19. Soares, P. *et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**, 740–759, doi:10.1016/j.ajhg.2009.05.001 (2009).
20. Achilli, A. *et al.* Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl. Acad. Sci. USA* **109**, 2449–2454, doi:10.1073/pnas.1111637109 (2012).
21. Colli, L. *et al.* Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. *BMC Genomics* **16**, 1115, doi:10.1186/s12864-015-2342-2 (2015).
22. Tanaka, K. *et al.* Nucleotide diversity of mitochondrial DNAs between the swamp and the river types of domestic water buffaloes, *Bubalus bubalis*, based on restriction endonuclease cleavage patterns. *Biochem. Genet.* **33**, 137–148, doi:10.1007/bf00554726 (1995).
23. Tanaka, K. *et al.* Phylogenetic relationship among all living species of the genus *Bubalus* based on DNA sequences of the cytochrome b gene. *Biochem. Genet.* **34**, 443–452, doi:10.1007/bf00570125 (1996).
24. Bibi, F. A multi-calibrated mitochondrial phylogeny of extant *Bovidae* (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. *BMC Evol. Biol.* **13**, 166, doi:10.1186/1471-2148-13-166 (2013).
25. Edwards, C. J. *et al.* A complete mitochondrial genome sequence from a mesolithic Wild Aurochs (*Bos primigenius*). *PLoS One* **5**, e9255, doi:10.1371/journal.pone.0009255 (2010).
26. Patel, A. K. & Meadow, R. H. In *Archaeology of the near east*: *Proceedings of the third international symposium on the archeozoology of the southwestern Asia and adjacent areas* (eds H. L. Buitenhuis, L. Bartosiewicz, & A. M. Choyke) (ARC- publications, 1998).
27. Finlay, E. K. *et al.* Bayesian inference of population expansions in domestic bovines. *Biol. Lett.* **3**, 449–452, doi:10.1098/rsbl.2007.0146 (2007).
28. Woodruff, D. S. Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity. *Biodivers. Conserv.* **19**, 919–941, doi:10.1007/s10531-010-9783-3 (2010).
29. De Deckker, P., Tapper, N. J. & van der Kaars, S. The status of the Indo-Pacific Warm Pool and adjacent land at the Last Glacial Maximum. *Glob. Planet. Change* **35**, 25–35, doi:10.1016/S0921-8181(02)00089-9 (2003).
30. Clark, P. U. *et al.* Global climate evolution during the last deglaciation. *Proc. Natl. Acad. Sci. USA* **109**, E1134–E1142, doi:10.1073/pnas.1116619109 (2012).
31. Yan, X.-H., Ho, C.-R., Zheng, Q. & Klemas, V. Temperature and Size Variabilities of the Western Pacific Warm Pool. *Science* **258**, 1643 (1992).
32. Crowley, J. T. CLIMAP SSTs re-revisited. *Clim. Dyn.* **16**, 241–255, doi:10.1007/s003820050325 (2000).
33. Cannon, C. The Ecology of Tropical East Asia by Richard T. Corlett. *Q. Rev. Biol.* **90**, 433, doi:10.1086/683725 (2015).
34. Graur, D. & Martin, W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* **20**, 80–86, doi:10.1016/j.tig.2003.12.003 (2004).
35. Pelejero, C., Kienast, M., Wang, L. & Grimalt, J. O. The flooding of Sundaland during the last deglaciation: imprints in hemipelagic sediments from the southern South China Sea. *Earth Planet. Sci. Lett.* **171**, 661–671, doi:10.1016/S0012-821X(99)00178-8 (1999).
36. Zhou, W. *et al.* High-resolution evidence from southern China of an early Holocene optimum and a mid-Holocene dry event during the past 18,000 years. *Quat. Res.* **62**, 39–48, doi:10.1016/j.yqres.2004.05.004 (2004).
37. Nagarajan, M., Nimisha, K. & Kumar, S. Mitochondrial DNA variability of domestic river buffalo (*Bubalus bubalis*) populations: Genetic evidence for domestication of river buffalo in Indian Subcontinent. *Genome Biol. Evol.* **7**, 1252–1259, doi:10.1093/gbe/evv067 (2015).
38. Librado, P. *et al.* The evolutionary origin and genetic makeup of domestic horses. *Genetics* **204**, 423–434, doi:10.1534/genetics.116.194860 (2016).
39. Lindgren, G. *et al.* Limited number of patrilines in horse domestication. *Nat. Genet.* **36**, 335–336, doi:10.1038/ng1326 (2004).
40. Parma, P., Erra-Pujada, M., Feligini, M., Greppi, G. & Enne, G. Water buffalo (*Bubalus bubalis*): Complete nucleotide mitochondrial genome sequence. *DNA Seq.* **15**, 369–373, doi:10.1080/10425170400019318 (2004).
41. Olivieri, A. *et al.* Mitogenomes from Egyptian cattle breeds: new clues on the origin of haplogroup Q and the early spread of *Bos taurus* from the Near East. *PLoS One* **10**, e0141170, doi:10.1371/journal.pone.0141170 (2015).
42. Eltsov, N. P. & Volodko, N. V. In http://eltsov.org (2011).
43. Achilli, A. *et al.* Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.* **18**, R157–158, doi:10.1016/j.cub.2008.01.019 (2008).
44. Lancioni, H. *et al.* Phylogenetic relationships of three Italian merino-derived sheep breeds evaluated through a complete mitogenome analysis. *PLoS One* **8**, e73712, doi:10.1371/journal.pone.0073712 (2013).
45. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729, doi:10.1093/molbev/mst197 (2013).
46. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591, doi:10.1093/molbev/msm088 (2007).
47. Drummond, A. J. & Rambaut, A. BEAST: Bayesian Evolutionary Analysis by Sampling Trees. *BMC Evol. Biol.* **7**, 214, doi:10.1186/1471-2148-7-214 (2007).
48. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192, doi:10.1093/molbev/msi103 (2005).
49. Bollongino, R. *et al.* Modern Taurine Cattle descended from small number of Near-Eastern founders. *Mol. Biol. Evol.* **29**, 2101–2104, doi:10.1093/molbev/mss092 (2012).

Acknowledgements

This research received support from: Natural Science Foundation of China (31272399) and the Program of National Beef Cattle and Yak Industrial Technology System [CARS-38] (to CL); the Italian Ministry of Education, University and Research: Progetti Futuro in Ricerca 2012 (RBFR126B8I) (to AA) and Progetti Ricerca Interesse Nazionale 2012 (2012JA4BTY) (to AA); the University of Pavia strategic theme MIGRATions: towards an "INterdisciplinary Governance model" (MIGRAT.IN.G; to AA), and the Asia Link Project Reproductive biotechnology funded by the European Union.

Author Contributions

Conceived and designed the experiments: S.W., N.C., M.R.C., A.A., C.L.; Performed the experiments: S.W., N.C., M.R.C., L.C. Analyzed the data: S.W., N.C., M.R.C., H.La., A.A., C.L. Contributed reagents/materials/analysis tools: T.Z., A.A., C.L. Performed the collection of biological samples: S.W., N.C., T.Z., H.Z., Y.M., V.C., M.W., M.Y., Y.Z., H.Lu., R.D., Y.H., X.L., M.P., H.C., J.A.L., C.L. Wrote the paper: S.W., N.C., M.R.C., H.La., H.Lu., M.P., J.A.L., A.A., C.L. All authors reviewed and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-04830-2

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

RESEARCH ARTICLE

Mitochondrial DNA variants of Podolian cattle breeds testify for a dual maternal origin

Piera Di Lorenzo[1☯], Hovirag Lancioni[2☯*], Simone Ceccobelli[1], Licia Colli[3,4], Irene Cardinali[2], Taki Karsli[5], Marco Rosario Capodiferro[6], Emine Sahin[7], Luca Ferretti[6], Paolo Ajmone Marsan[3,4], Francesca Maria Sarti[1], Emiliano Lasagna[1], Francesco Panella[1], Alessandro Achilli[6*]

1 Dipartimento di Scienze Agrarie, Alimentari e Ambientali, Università degli Studi di Perugia, Perugia, Italy, **2** Dipartimento di Chimica, Biologia e Biotecnologie, Università degli Studi di Perugia, Perugia, Italy, **3** Institute of Zootechnics, Università Cattolica del S. Cuore, Piacenza, Italy, **4** Biodiversity and Ancient DNA Research Center–BioDNA, Università Cattolica del S. Cuore, Piacenza, Italy, **5** Department of Animal Science, Faculty of Agriculture, University of Akdeniz, Antalya, Turkey, **6** Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università di Pavia, Pavia, Italy, **7** Korkuteli Vocational School, University of Akdeniz, Antalya, Turkey

☯ These authors contributed equally to this work.
* hovirag.lancioni@unipg.it (HL); alessandro.achilli@unipv.it (AA).

Citation: Di Lorenzo P, Lancioni H, Ceccobelli S, Colli L, Cardinali I, Karsli T, et al. (2018) Mitochondrial DNA variants of Podolian cattle breeds testify for a dual maternal origin. PLoS ONE 13(2): e0192567. https://doi.org/10.1371/journal.pone.0192567

Editor: David Caramelli, University of Florence, ITALY

Received: October 17, 2017

Accepted: January 25, 2018

Published: February 20, 2018

Data Availability Statement: Sequences were recorded in GenBank (https://www.ncbi.nlm.nih.gov/genbank/) with accession numbers MF474376-MF475904. All other relevant data are within the paper and its Supporting Information files.

Funding: This research received support from: the Italian Ministry of Education, University and Research: Progetti Futuro in Ricerca 2012 (RBFR126B8I) (to AA) and Progetti Ricerca Interesse Nazionale 2012 (2012JA4BTY) (to AA);

Abstract

Background

Over the past 15 years, 300 out of 6000 breeds of all farm animal species identified by the Food and Agriculture Organization of the United Nations (FAO) have gone extinct. Among cattle, many Podolian breeds are seriously endangered in various European areas. Podolian cattle include a group of very ancient European breeds, phenotypically close to the aurochs ancestors (*Bos primigenius*). The aim of the present study was to assess the genetic diversity of Podolian breeds and to reconstruct their origin.

Methodology

The mitochondrial DNA (mtDNA) control-regions of 18 Podolian breeds have been phylogenetically assessed. Nine non-Podolian breeds have been also included for comparison.

Conclusion

The overall analysis clearly highlights some peculiarities in the mtDNA gene pool of some Podolian breeds. In particular, a principal component analysis point to a genetic proximity between five breeds (*Chianina*, *Marchigiana*, *Maremmana*, *Podolica Italiana* and *Romagnola*) reared in Central Italy and the Turkish Grey. We here propose the suggestive hypothesis of a dual ancestral contribution to the present gene pool of Podolian breeds, one deriving from Eastern European cattle; the other arising from the arrival of Middle Eastern cattle into Central Italy through a different route, perhaps by sea, ferried by Etruscan boats. The historical migration of Podolian cattle from North Eastern Europe towards Italy has not cancelled the mtDNA footprints of this previous ancient migration.

the University of Pavia strategic theme MIGRATions: towards an INterdisciplinary Governance model" (MIGRAT.IN.G; to AA). We thank Dr. Giovanni Capuzzello and Dr. Gilberto Leonelli for their precious help in providing biological samples for Modicana and Bianca di Val Padana breeds respectively.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Over the past 15 years, 300 out of 6000 livestock breeds identified by Food and Agriculture Organization of the United Nations (FAO) have gone extinct [1]. Risk factors for farm animal breeds are mainly: i) a reduction of genetic variability due to strict selection processes; ii) a strong economic pressure focused on specific traits, such as milk production, which leads to the replacement of local less productive breeds with highly productive industrial breeds; iii) an unrestricted and indiscriminate cross-breeding, especially in developing countries [2]. *Bos taurus* is one of the most economically important livestock species [3]. Both in historic and current societies it has fulfilled agricultural, economic, cultural, and even religious key roles, often paralleling human evolution [4]. Among cattle, many Podolian breeds are seriously endangered in various European countries [5]; [6]; [7]. Podolian cattle include a group of very ancient European breeds, with a grey coat colour and long horns, phenotypically close to the aurochs (*Bos primigenius*). According to many traditional notes the name Podolian refers to a common ancestral origin in Podolia (the modern western Ukraine). However place of origin and timing of spread out of the source area are both debated. Alternative hypotheses have been proposed: Podolian cattle might have spread from the eastern steppe southward into Anatolia and westward into the Balkans and Italy in historical times (3^{rd}-5^{th} century AD) along with East-European Barbarian people [8]; other authors suggest a more ancient migration (~3 kya BP) from the Near East to Central Italy through the Mediterranean Sea [9], together with a possible contribution from local wild aurochs through a secondary local domestication/introgression events[10]; [11].

Nowadays, some phenotypic distinctions stand out among Podolian cattle [12]. The noble aurochs-shaped ancient breeds with long horns (such as Hungarian Grey, Katerini, Podolsko, Slavonian Syrmian and Maremmana) are considered as the only true Podolian breeds by some scholars. However, some local breeds (i.e. Podolica Italiana, Ukrainian Grey, Turkish Grey and other Balkan breeds) do not necessarily show the long horns, but maintained some distinctive Podolian traits such as a red coat in calves and light grey in adults [13]. In general, a commercial trait shared by all Podolian cattle is that they are more suitable for beef production rather than for dairy. Because of that some improved beef breeds (Chianina, Marchigiana, Romagnola and Piemontese) are also considered within the Podolian group, although the inclusion of Chianina and Piemontese is still debated [14].

During the last decades, mitochondrial DNA (mtDNA) has been widely used as a molecular tool to investigate genetic origin, history and diversity of livestock species [15] [16] [17] [18] [19] [20]. Following this trend, the aim of the present study is to re-assess the mtDNA diversity of the major Podolian cattle breeds (ten of which classified as endangered or critical by FAO) to obtain additional information on their ancestral origin and ancient dispersal routes.

Results

We have analysed the mitochondrial DNA of 18 Podolian cattle breeds (Fig 1, Table 1, S1–S3 Tables).

The molecular analysis of 221 base pairs of the control region (from np 16042 to np 16262) on the entire dataset of 1,957 samples revealed a total of 247 distinct haplotypes (from four to 70 haplotypes per breed) and 91 polymorphic sites (S), all represented by single nucleotide polymorphisms (SNPs). The average nucleotide diversity (π) was comparable between Podolian and non-Podolian breeds (~0.010–0.011; Table 1), while haplotype diversity was significantly lower (*P-value* < 0.01) in Podolian (Hd = 0.837 ± 0.010) than in non-Podolian breeds (Hd = 0.879 ± 0.013). Both indices varied largely across breeds as already seen in previous mtDNA studies [11]; [21]; [22]; [23]. Among all Podolian, the highest Hd values (≥ 0.90) were

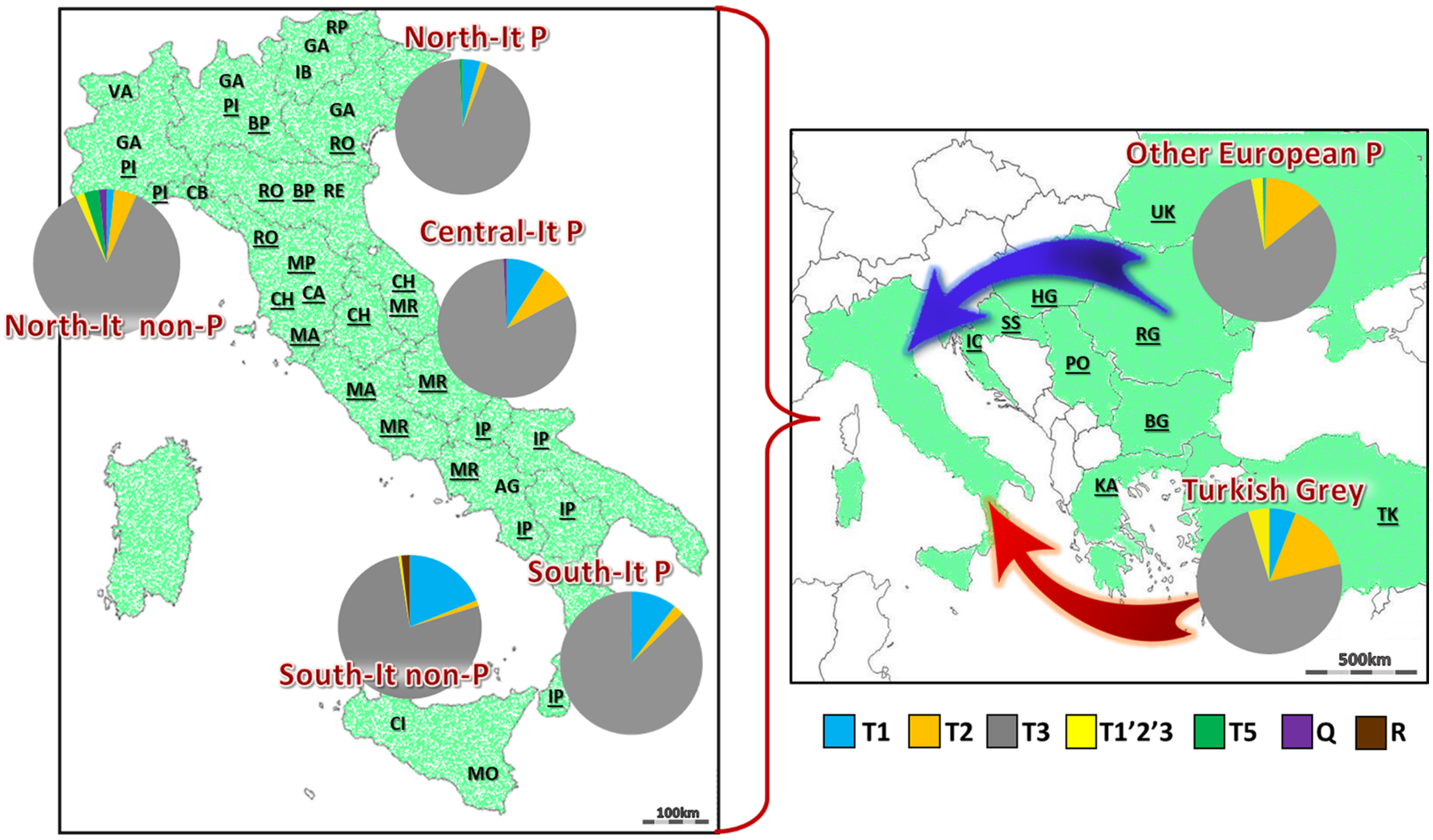


Fig 1. Prevalent locations and frequency distributions of mitochondrial haplogroups in 18 Podolian (P, underlined) and 9 non-Podolian (non-P) breeds analyzed in this study. Breed codes as in Table 1 (see also Table 2 and S3 Table for details). Note that RP and IB are also widespread in Italy. Maps (www.histgeo.ac-aix-marseille.fr/ancien_site/carto/) were used for illustrative purposes only and largely modified by the authors.

https://doi.org/10.1371/journal.pone.0192567.g001

identified in Chianina, Marchigiana and Turkish Grey, while the lowest values of Hd (<0.70), as well as of nucleotide diversity (≤0.008), were scored in Bianca di Val Padana, Calvana, and Slavonian Syrmian Podolian. As for 1,617 sequences, we were able to extend the analysis to a longer control-region fragment encompassing 731 bps (Table 1). The results largely confirmed the same trend, with the only notable exception of Piemontese, Romagnola and Maremmana showing a higher Hd (>0.960) on this extended fragment. It is also interesting that the highest Hd values were identified in Chianina, and Maremmana, which showed values (>0.970) comparable to the Turkish Grey (0.971).

All control-region haplotypes have been classified in haplogroups and sub-haplogroups through an accurate analysis of mutational motifs (Table 2 and S3 Table), according to previously published classification criteria [11]; [24]; [25]; [26]. Haplogroup T3 was the most common (83%) in all breeds, with the highest value in MP and PO (both 100%), followed by PI (96%) (acronyms are listed in Table 1). The second and third most common haplogroups (both 7%) were T1 and T2, which were missing in MP, PO and SS. The frequency of T2 is lower in non-Podolian (3.85%) than in Podolian breeds (8.35%) with extraordinary high peaks in three breeds, KA (42%), RG (24%) and BG (22%), followed by UK (16%) and MA and TK (both 15%). T1 haplogroup was predominantly found among breeds from Central and Southern Italy, both in Podolian (10%) and non-Podolian (19%) groups (Fig 1). Haplogroup T5 was found exclusively in non-Podolian breeds, and was restricted to IB, RP and VA except for one sequence found in SS and one in PI. Finally, haplogroups Q and R showed very low incidences

Table 1. Estimates of genetic diversity[a] on the 27 breeds analyzed in this work.

Breed	Country	Podolian/non-Podolian	ID name	N	π	Hd	N	π	Hd
				Range:	**16042-**	**16262**	**Range:**	**15823-**	**215**
Piemontese	Northern Italy	Podolian	PI	72	0.008	0.833	72	0.005	0.970
Bianca di Val Padana	Northern Italy	Podolian	BP	45	0.006	0.630	45	0.003	0.834
Romagnola	Central Italy	Podolian	RO	225	0.014	0.890	225	0.007	0.965
Mucco Pisano	Central Italy	Podolian	MP	33	0.006	0.788	33	0.004	0.814
Calvana	Central Italy	Podolian	CA	35	0.008	0.662	26	0.004	0.889
Chianina	Central Italy	Podolian	CH	369	0.013	0.895	338	0.006	0.973
Maremmana	Central Italy	Podolian	MA	75	0.013	0.857	62	0.006	0.973
Marchigiana	Central Italy	Podolian	MR	146	0.013	0.903	146	0.006	0.967
Italian Podolian	Southern Italy	Podolian	IP	125	0.008	0.780	91	0.005	0.872
Ukrainian Grey	Ukraine	Podolian	UK	32	0.010	0.856	1		
Romanian Grey	Romania	Podolian	RG	17	0.014	0.890	-		
Hungarian Grey	Hungary	Podolian	HG	93	0.010	0.856	1		
Slavonian Syrmian Pod.	Croatia	Podolian	SS	9	0.004	0.583	-		
Istrian Cattle	Croatia	Podolian	IC	17	0.010	0.794	-		
Podolsko	Serbia	Podolian	PO	11	0.005	0.709	-		
Bulgarian Grey	Bulgaria	Podolian	BG	36	0.014	0.838	30	0.005	0.779
Katerini	Greece	Podolian	KA	12	0.019	0.879	-		
Turkish Grey	Turkey	Podolian	TK	85	0.012	0.922	70	0.006	0.971
Valdostana	Northern Italy	non-Podolian	VA	54	0.011	0.934	54	0.006	0.941
Grey Alpine	Northern Italy	non-Podolian	GA	45	0.012	0.853	45	0.006	0.979
Italian Brown	Northern Italy	non-Podolian	IB	34	0.010	0.852	34	0.005	0.929
Italian Red Pied	Northern Italy	non-Podolian	RP	136	0.010	0.896	125	0.006	0.982
Cabannina	Northern Italy	non-Podolian	CB	55	0.011	0.882	43	0.005	0.928
Reggiana	Northern Italy	non-Podolian	RE	38	0.006	0.713	38	0.003	0.845
Agerolese	Southern Italy	non-Podolian	AG	36	0.014	0.913	36	0.006	0.956
Cinisara	Southern Italy	non-Podolian	CI	81	0.014	0.881	69	0.007	0.966
Modicana	Southern Italy	non-Podolian	MO	41	0.010	0.763	33	0.005	0.864
		All Podolian		**1437**	**0.010**	**0.837**	**1140**	**0.006**	**0.980**
		Non-Podolian		**520**	**0.011**	**0.879**	**477**	**0.006**	**0.963**
		All samples		**1957**	**0.010**	**0.845**	**1617**	**0.006**	**0.977**b

[a] N = number of sequences

π = nucleotide diversity, Hd = haplotype diversity.

https://doi.org/10.1371/journal.pone.0192567.t001

restricted to Italian non-Podolian (Q = 1.15%, and R = 0.58%) and Podolian (Q = 0.77%, and R = 0.49%) breeds. Overall, the haplogroup distribution differed significantly between the Podolian and non-Podolian groups of breeds included in the current analysis (Table 2; *chi-square P-value* < 0.001) with the highest contribution given by the T2 haplogroup. This result was also verified by considering haplogroup frequencies based on different haplotypes (*chi-square P-value* < 0.001) in order to mitigate the effect of inbreeding and recent founder effects.

Thus, we performed a principal component analysis (PCA) to graphically display the different haplogroup distributions among the Podolian breeds. In order to consider as many populations as possible, the dataset based on the short fragment was included. After variables reduction to PCs, the coordinates of the observations for the 18 breeds were displayed in a two-dimensional plot representing the European Podolian genetic landscape (Fig 2). PC1

Table 2. Sources and haplogroup affiliation for the Podolian and non-Podolian mtDNA sequences. Haplogroup frequencies (%) are in parentheses.

Code	Group/Breed	T1	T2	T3	T1'2'3	T5	Q	Q1	Q2	R	R1	R2	TOTAL
	Podolian	**103(7.17)**	**120(8.35)**	**1184(82.39)**	**10(0.70)**	**2(0.14)**	**1(0.07)**	**5(0.35)**	**5(0.35)**	**0(0.00)**	**5(0.35)**	**2(0.14)**	**1437**
BP	Bianca di Val Padana	5(11.11)	-	40(88.89)	-	-	-	-	-	-	-	-	45
PI	Piemontese	-	2(2.78)	69(95.83)	-	1(1.39)	-	-	-	-	-	-	72
CA	Calvana	5(14.29)	-	30(85.71)	-	-	-	-	-	-	-	-	35
CH	Chianina	35(9.49)	36(9.76)	292(79.13)	-	-	1(0.27)	2(0.54)	3(0.81)	-	-	-	369
MR	Marchigiana	18(12.33)	5(3.42)	122(83.56)	-	-	-	-	-	-	-	1(0.68)	146
MA	Maremmana	9(9.33)	11(14.67)	55(73.33)	-	-	-	-	-	-	-	-	75
MP	Mucco Pisano	-	-	33(100.00)	-	-	-	-	-	-	-	-	33
RO	Romagnola	12(5.33)	19(8.44)	183(81.33)	-	-	-	3(1.33)	2(0.89)	-	5(2.22)	1(0.44)	225
IP	Podolica Italiana	13(10.40)	3(2.40)	109(87.20)	-	-	-	-	-	-	-	-	125
HG	Hungarian Grey		8(8.60)	82(88.17)	3(3.23)	-	-	-	-	-	-	-	93
UK	Ukrainian Grey	-	5(15.63)	25(78.13)	2(6.25)	-	-	-	-	-	-	-	32
BG	Bulgarian Grey	1(2.78)	8(22.22)	27(75.00)	-	-	-	-	-	-	-	-	36
IC	Istrian Cattle	-	1(5.88)	15(88.24)	1(5.88)	-	-	-	-	-	-	-	17
PO	Podolsko	-	-	11(100.0)	-	-	-	-	-	-	-	-	11
RG	Romanian Grey	-	4(23.53)	13(76.47)	-	-	-	-	-	-	-	-	17
SS	Slavonian Syrmian Pod.	-	-	8(88.89)	-	1(11.11)	-	-	-	-	-	-	9
KA	Katerini	-	5(41.67)	7(58.33)	-	-	-	-	-	-	-	-	12
TK	Turkish Grey	5(5.88)	13(15.29)	63(74.12)	4(4.71)	-	-	-	-	-	-	-	85
	Non-Podolian	**36(6.92)**	**20(3.85)**	**435(83.65)**	**8(1.54)**	**12(2.31)**	**3(0.57)**	**3(0.58)**	**-**	**2(0.38)**	**1(0.19)**	**-**	**520**
AG	Agerolese	7(19.44)	-	27(75.00)	1(2.78)	-	-	-	-	-	1(2.78)	-	36
CB	Cabannina	-	5(9.09)	40(72.73)	7(12.73)	-	3(5.45)	-	-	-	-	-	55
CI	Cinisara	17(20.99)	2(2.47)	60(74.07)	-	-	-	-	-	2(2.47)	-	-	81
GA	Grigio Alpina	-	2(4.44)	41(91.11)	-	-	-	2(4.44)	-	-	-	-	45
IB	Bruna Italiana	2(5.88)	-	29(85.29)	-	3(8.82)	-	-	-	-	-	-	34
RP	Pezzata Rossa Italiana	2(1.47)	10(7.35)	122(89.71)	-	1(0.74)	-	1(0.74)	-	-	-	-	136
MO	Modicana	6(14.63)	-	35(85.37)	-	-	-	-	-	-	-	-	41
RE	Reggiana	2(5.26)	-	36(94.74)	-	-	-	-	-	-	-	-	38
VA	Valdostana	-	1(1.85)	45(83.33)	-	8(14.81)	-	-	-	-	-	-	54
	Total	**139(7.10)**	**140(7.15)**	**1619(82.73)**	**18(0.92)**	**14(0.712**	**4(0.20)**	**8(0.41)**	**5(0.26)**	**2(0.10)**	**6(0.31)**	**2(0.10)**	**1957**

https://doi.org/10.1371/journal.pone.0192567.t002

clearly separated the Turkish Grey and the five most important Central and Southern Italian beef cattle breeds (CH, RO, MR, MA and IP; S1 Fig) from all the remaining populations, while PC2 contributed to separate the Hungarian and the Turkish Grey.

Because of the peculiar position of some Italian breeds, we used an analysis of molecular variance (AMOVA) to investigate fixation indices in three (artificially created) population groups, one including the Italian Podolian breeds, the other encompassing the European Podolian breeds and the Turkish Grey, and the third group covering the Italian non-Podolian breeds. Most of the variance (about 98%) observed in the Italian Podolian populations explained differences among samples within breeds, while less than 2% represented differences between breeds (Table 3), a value three times lower than in the other Podolian breeds.

Discussion

To date, only a limited number of studies have addressed the genetic composition of Podolian breeds. These investigations were generally limited to few breeds and focused on the nuclear genome [8]; [27]; [28]; [29]; [30]. Mitochondrial DNA data have been previously reported by

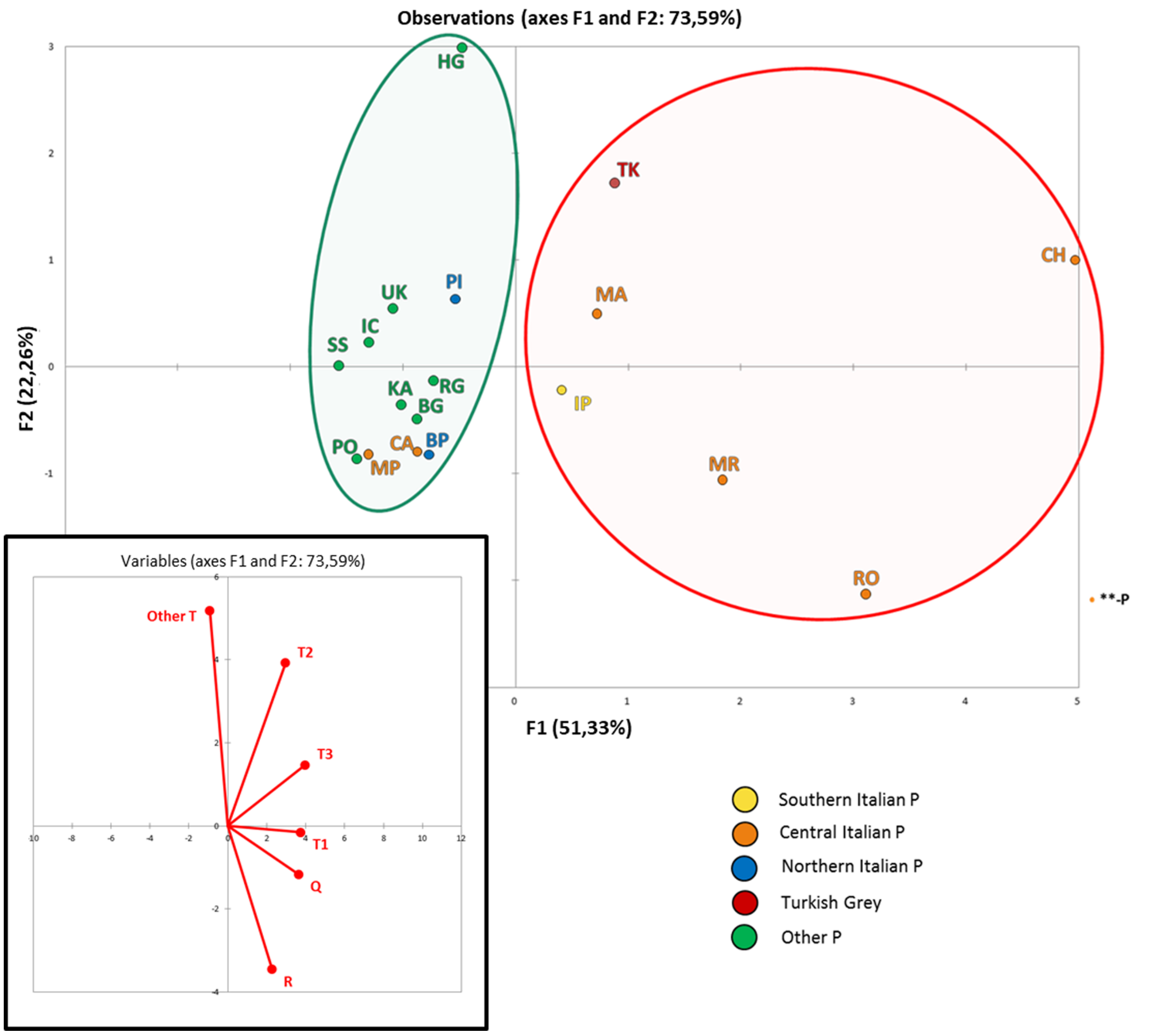


Fig 2. Principal Component Analysis (PCA) of all Podolian breeds. Below is the plot of the contribution of each haplogroup to the first and second PC (projections of the axes of the original variables).

https://doi.org/10.1371/journal.pone.0192567.g002

Ivankovic et al. [6] and Ilie et al. [7] who analysed the control-region sequences of some Croatian and Romanian cattle breeds, respectively. The present study extended the analysis of the mitochondrial genetic variation to 18 Podolian breeds by evaluating their haplogroup distributions, which were eventually compared among them and to nine non-Podolian breeds. Genetic diversity considered in terms of number of haplotypes and nucleotide diversity revealed some peculiarities of several Podolian breeds. The low mtDNA diversities of Bianca di Val Padana,

Table 3. Analysis of molecular variance (AMOVA).

Group	Variation within breeds (%)	Variation among breeds (%)
Italian Podolian breeds	98.08	1.92
Non-Italian Podolian breed	93.85	6.15
Non-Podolian breeds	97.21	2.79

https://doi.org/10.1371/journal.pone.0192567.t003

Calvana and Slavonian Syrmian Podolian could be due to a combination of factors, such as i) a sampling bias depending on the low consistence of current herds (S1 Table), ii) a bottleneck effect caused by the strong reduction in population size experienced by these breeds during the last decades, iii) genetic drift acting on small populations. On the contrary, the large diffusion of Piemontese, Marchigiana and Chianina cattle probably favoured the accumulation and maintenance of a high level of mtDNA variation, which is evident also in the Maremmana breed in spite of its lower consistency (S1 Table). In general, the most common European haplogroup T3 is predominant (83%) in our dataset. High frequencies of T1 in central and southern Italy might be due to the intensive migrations across the Mediterranean Sea, eased by the proximity to northern Africa, where T1 is prevalent [20]; [25]; [26]; [31] and to the Near East where T1 is also present [32, 33]. It is worth noting that the central-Italian Podolian breeds show higher frequencies of T2, while the presence of Q and R within the Podolian group is limited to three Italian breeds: Romagnola, Chianina and Marchigiana, the latter derived from crossbreeding between the first two in the early 20th century.

The significant differences between the haplogroup distributions of Podolian and non-Podolian and the low genetic differentiation among the Italian Podolian breeds (three times less than in other Podolian breeds or in the non-Podolian group) points to a common (and perhaps peculiar) origin. As a matter of fact, according to the first component of the PCA, five Italian beef cattle breeds formed a clearly separated group and were also closer to the Turkish Grey than to any other Podolian breeds. At first this peculiarity could be explained as the effect of a stronger beef-oriented selection carried out on these breeds compared to the other Podolian populations. However, another important Italian beef cattle, the Piemontese, is placed within the other Podolian cluster, which shares the common feature of a strong grey coat. Thus, an alternative explanation might assume a different ancestral origin for the two groups of Podolian breeds, as summarized in Fig 1: a first group, mostly consisting of breeds from East Europe and northern Italy that share a similar mitochondrial gene pool, may have originated from ancestors migrated through an inland route from Podolia across eastern Europe all the way into northern Italy, in accordance with the great wave of cattle migration occurred during the Barbarian invasions. Whereas, a second group, including the white Podolian cattle closely related to the Turkish Grey, may descend from ancestral bovines brought to Italy through a different and likely maritime route crossing the Mediterranean Sea. A previous study [34] suggested also a genetic link between the Turkish Grey and Bulgarian and Hungarian breeds, but our results do not support such hypothesis, highlight instead a stronger maternal relationship between the Turkish Grey and five central-southern Italian Podolian breeds. Those cattle are bred since the medieval time [29] in an area that largely overlaps with the ancient territory of Etruria. This finding further supports and extends another hypothesis, according to which at least part of the maternal ancestry of those breeds could be related to the Etruscan migration from Lydia, a region on the south-western coast of ancient Anatolia [9]; [35]; [36], [37]; [38]. It is worth noting that the five Podolian breeds are also the main Italian beef cattle together with the Piemontese, and that previous studies suggested a possible common genetic origin [39]. Our findings suggest that, in spite of a stronger beef-oriented selection, their mitochondrial gene pool still preserves genetic traces of a different maternal origin, confirming that the selection practices were mostly male-mediated and enforcing the importance of the mtDNA screening to reconstruct the ancestry and history of current breeds.

Material and methods

Ethics statement

All experimental procedures were reviewed and approved by the Animal Research Ethics Committee of the Universities of Perugia and Pavia in accordance with the European Union Directive 86/609.

Samples

The entire dataset analyzed in this study encompasses 1,957 mtDNA control-region sequences including 1,321 from our previous studies [11]; [24]; [40], 428 retrieved from GenBank, and 208 additional samples specifically collected for this study (Table 1, S1–S3 Tables). Piemontese (also called Piedmontese) (PI, n = 72), Romagnola (RO, n = 225), Marchigiana (MR, n = 146), Chianina (CH, n = 369), Maremmana (MA, n = 75), Podolica Italiana (also known as Italian Podolian) (IP, n = 125), Mucco Pisano (MP, n = 33), Calvana (CA, n = 35), Bianca di Val Padana (BP, n = 45), Hungarian Grey (HG, n = 93), Bulgarian Grey (BG, n = 36), Istrian cattle (IC, n = 17), Katerini (KA, n = 12), Romanian Grey (RG, n = 17), Slavonian Syrmian Podolian (SS, n = 9), Turkish Grey (TK, n = 85), Ukrainian Grey (UK, n = 32), Podolsko (PO, n = 11). Moreover, nine unrelated non-Podolian breeds from Italy, are included as a control group: Valdostana (VA, n = 54), Bruna Italiana (also known as Italian Brown) (IB, n = 34), Grigio Alpina (also called Grey Alpine) (GA, n = 45), Pezzata Rossa Italiana (also known as Italian Red Pied) (RP, n = 136), Modicana (MO, n = 41), Reggiana (RE, n = 38), Agerolese (AG, n = 36), Cinisara (CI, n = 81), Cabannina (CB, n = 55).

DNA extraction, amplification and sequencing

As for the 208 novel mtDNAs, blood samples were collected from the jugular vein of each animal in vacutainer tubes, containing EDTA as anticoagulant. These animals were chosen in different farms in order to avoid closely related individuals and gather a representative sample of the breeds. Whole blood was stored at -20°C until DNA extraction. DNA was isolated using the GenElute Blood Genomic DNA kit (Sigma Aldrich, St. Louis, MO, USA) and stored at -20°C until genotyping. PCR amplification of the control region was performed using forward and reverse primers (5′-CCTAAGACTCAAGGAAGAAACTGC-3′ and 3′-AACCTAGAGGGC ATTCTCACTG-5′ respectively) specifically designed on the Bovine Reference Sequence (BRS; GenBank V00654). The 1138 bp PCR fragment encompassed the mtDNA control region from np 15718 to 517. Amplicons were first purified using exonuclease I and alkaline phosphatase (ExoSAP-IT® enzymatic system-USB Corporation, Cleveland, OH, USA), then sequenced with the primer 15757F (5′-CCCCAAAGCTGAAGTTCTAT-3′), as previously described [40]. A dataset of 1,321 sequences was already available in our laboratories. All data were recorded in GenBank with accession numbers MF474376-MF475904 (S3 Table) and compared to those retrieved from the database (S2 Table).

Data analyses

Sequences were aligned to the Bovine Reference Sequence (BRS; V00654) using the software Sequencher^{TM} 5.0. For a total of 1,617 samples, we were able to analyse a 731-bps fragment trimmed from np 15823 to np 215, while only a short fragment (221 bps, from np 16042 to np 16262) was considered in order to include the widest possible number of samples (N = 1,957). Haplotypes were classified in haplogroups and sub-haplogroups according to previously identified mutational motifs [24]. Indices of molecular variation were calculated using the DNAsp 5.1 software [41], while an analysis of molecular variance was computed using AMOVA program implemented in the ARLEQUIN 3.01 package [42]. Finally, principal component analyses (PCA) were performed using Excel software implemented by XLSTAT, as described elsewhere [43, 44]. The PCA is a widely used dimension-reduction method that summarizes the variance of multivariate data in a smaller number of variables (the principal components, PCs), which are linear functions of the original variables, here expressed as haplogroup and sub-haplogroup frequencies, The rarest haplogroups were phylogenetically grouped and frequencies were calculated by considering only different haplotypes within the same breed.

Supporting information

S1 Table. List of Podolian breeds analyzed in this study.
(XLSX)

S2 Table. List of samples retrieved from GenBank.
(XLSX)

S3 Table. List of samples analysed in this study.
(XLSX)

S1 Fig. The five most important Italian beef cattle breeds from central and southern Italy discussed in this paper.
(PDF)

Author Contributions

Conceptualization: Hovirag Lancioni, Emiliano Lasagna, Francesco Panella, Alessandro Achilli.

Data curation: Hovirag Lancioni, Taki Karsli, Marco Rosario Capodiferro, Emine Sahin, Luca Ferretti, Paolo Ajmone Marsan, Emiliano Lasagna.

Formal analysis: Piera Di Lorenzo, Hovirag Lancioni, Simone Ceccobelli, Licia Colli, Irene Cardinali, Taki Karsli, Marco Rosario Capodiferro, Francesca Maria Sarti.

Funding acquisition: Alessandro Achilli.

Methodology: Hovirag Lancioni, Licia Colli.

Resources: Paolo Ajmone Marsan, Alessandro Achilli.

Supervision: Licia Colli, Luca Ferretti, Francesca Maria Sarti, Francesco Panella, Alessandro Achilli.

Validation: Simone Ceccobelli, Alessandro Achilli.

Writing – original draft: Piera Di Lorenzo, Hovirag Lancioni, Alessandro Achilli.

Writing – review & editing: Hovirag Lancioni, Luca Ferretti, Paolo Ajmone Marsan, Emiliano Lasagna, Francesco Panella, Alessandro Achilli.

References

1. Scherf B. World Watch List for Domestic Animal Diversity. 3rd ed. Rome. 2000.
2. Soysal M, Tuna Y, Gurcan E, Ozkan E. Farms in Turkey: sustainable development in the preservation of animal genetic resources in Turkey and in the world. Trakia J Sci. 2004; 2(3):47–53.
3. Cunningham E. Selected Issues in Livestock Industry Development. Washington, DC.1992.

4. Bradley DG, MacHugh DE, Cunningham P, Loftus RT. Mitochondrial diversity and the origins of African and European cattle. Proc Natl Acad Sci U S A. 1996; 93(10):5131–5. PMID: 8643540; PubMed Central PMCID: PMCPMC39419.

5. Pariset L, Joost S, Marsan PA, Valentini A, Econogene Consortium E. Landscape genomics and biased FST approaches reveal single nucleotide polymorphisms under selection in goat breeds of North-East Mediterranean. BMC Genet. 2009; 10:7. Epub 2009/02/19. https://doi.org/10.1186/1471-2156-10-7 PMID: 19228375; PubMed Central PMCID: PMCPMC2663570.

6. Ivankovic A, Paprika S, Ramljak J, Dovc P, M. K. Mitochondrial DNA-based genetic evaluation of autochthonous cattle breeds in Croatia. Czech J Anim Sci 2014; 59(11):519–28.

7. Ilie DE, Cean A, Cziszter LT, Gavojdian D, Ivan A, Kusza S. Microsatellite and Mitochondrial DNA Study of Native Eastern European Cattle Populations: The Case of the Romanian Grey. PLoS One. 2015; 10(9):e0138736. Epub 2015/09/23. https://doi.org/10.1371/journal.pone.0138736 PMID: 26398563; PubMed Central PMCID: PMCPMC4580412.

8. Maretto F, Ramljak J, Sbarra F, Penasa M, Mantovani R, Ivankovic A, et al. Genetic relationship among Italian and Croatian Podolian cattle breeds assessed by microsatellite markers. Livest Sci. 2012;Sept 13; 150 (1–3):256–64.

9. Pellecchia M, Negrini R, Colli L, Patrini M, Milanesi E, Achilli A, et al. The mystery of Etruscan origins: novel clues from *Bos taurus* mitochondrial DNA. Proc Biol Sci. 2007; 274(1614):1175–9. https://doi.org/10.1098/rspb.2006.0258 PMID: 17301019; PubMed Central PMCID: PMCPMC2189563.

10. Beja-Pereira A, Caramelli D, Lalueza-Fox C, Vernesi C, Ferrand N, Casoli A, et al. The origin of European cattle: evidence from modern and ancient DNA. Proc Natl Acad Sci U S A. 2006; 103(21):8113–8. https://doi.org/10.1073/pnas.0509210103 PMID: 16690747; PubMed Central PMCID: PMCPMC1472438.

11. Bonfiglio S, Achilli A, Olivieri A, Negrini R, Colli L, Liotta L, et al. The enigmatic origin of bovine mtDNA haplogroup R: sporadic interbreeding or an independent event of *Bos primigenius* domestication in Italy? PLoS One. 2010; 5(12):e15760. Epub 2011/01/07. https://doi.org/10.1371/journal.pone.0015760 PMID: 21209945.

12. Bodò I, Gera I, Koppàny G. The Hungarian Grey cattle breed. Budapest: Association of the hungarian grey cattle breeders; 2004. 128 p p.

13. Bonadonna T. Etnologia zootecnica. Torino: UTET.

14. Bartosiewicz L. Hungarian Grey cattle in search of origins. Hungarian Agricultural research 1996; (3):13–20.

15. Groeneveld LF, Lenstra JA, Eding H, Toro MA, Scherf B, Pilling D, et al. Genetic diversity in farm animals—a review. Anim Genet. 2010; 41 Suppl 1:6–31. https://doi.org/10.1111/j.1365-2052.2010.02038.x PMID: 20500753.

16. Achilli A, Olivieri A, Soares P, Lancioni H, Hooshiar Kashani B, Perego UA, et al. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc Natl Acad Sci U S A. 2012; 109(7):2449–54. https://doi.org/10.1073/pnas.1111637109 PMID: 22308342; PubMed Central PMCID: PMCPMC3289334.

17. Lenstra JA, Groeneveld LF, Eding H, Kantanen J, Williams JL, Taberlet P, et al. Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. Anim Genet. 2012; 43 (5):483–502. https://doi.org/10.1111/j.1365-2052.2011.02309.x PMID: 22497351.

18. Colli L, Lancioni H, Cardinali I, Olivieri A, Capodiferro MR, Pellecchia M, et al. Whole mitochondrial genomes unveil the impact of domestication on goat matrilineal variability. BMC Genomics. 2015; 16 (1):1115. https://doi.org/10.1186/s12864-015-2342-2 PMID: 26714643; PubMed Central PMCID: PMCPMC4696231.

19. Scheu A, Powell A, Bollongino R, Vigne JD, Tresset A, Çakırlar C, et al. The genetic prehistory of domesticated cattle from their origin to the spread across Europe. BMC Genet. 2015; 16:54. https://doi.org/10.1186/s12863-015-0203-2 PMID: 26018295; PubMed Central PMCID: PMCPMC4445560.

20. Di Lorenzo P, Lancioni H, Ceccobelli S, Curcio L, Panella F, et al. Uniparental genetic systems: a male and a female perspective in the domestic cattle origin and evolution. Electronic Journal of Biotechnology. 2016;Sept; 23:69–78.

21. Lai SJ, Liu YP, Liu YX, Li XW, Yao YG. Genetic diversity and origin of Chinese cattle revealed by mtDNA D-loop sequence variation. Mol Phylogenet Evol. 2006; 38(1):146–54. Epub 2005/07/28. https://doi.org/10.1016/j.ympev.2005.06.013 PMID: 16054846.

22. Dadi H, Tibbo M, Takahashi Y, Nomura K, Hanada H, Amano T. Variation in mitochondrial DNA and maternal genetic ancestry of Ethiopian cattle populations. Anim Genet. 2009; 40(4):556–9. Epub 2009/04/03. https://doi.org/10.1111/j.1365-2052.2009.01866.x PMID: 19397526.

23. Hristov P, Spassov N, Iliev N, Radoslavov G. An independent event of Neolithic cattle domestication on the South-eastern Balkans: evidence from prehistoric aurochs and cattle populations. Mitochondrial DNA A DNA Mapp Seq Anal. 2017; 28(3):383–91. Epub 2015/12/29. https://doi.org/10.3109/19401736.2015.1127361 PMID: 26711535.

24. Achilli A, Bonfiglio S, Olivieri A, Malusà A, Pala M, Hooshiar Kashani B, et al. The multifaceted origin of taurine cattle reflected by the mitochondrial genome. PLoS One. 2009; 4(6):e5753. https://doi.org/10.1371/journal.pone.0005753 PMID: 19484124; PubMed Central PMCID: PMCPMC2684589.

25. Bonfiglio S, Ginja C, De Gaetano A, Achilli A, Olivieri A, Colli L, et al. Origin and spread of *Bos taurus*: new clues from mitochondrial genomes belonging to haplogroup T1. PLoS One. 2012; 7(6):e38601. https://doi.org/10.1371/journal.pone.0038601 PMID: 22685589; PubMed Central PMCID: PMCPMC3369859.

26. Olivieri A, Gandini F, Achilli A, Fichera A, Rizzi E, Bonfiglio S, et al. Mitogenomes from Egyptian Cattle Breeds: New Clues on the Origin of Haplogroup Q and the Early Spread of *Bos taurus* from the Near East. PLoS One. 2015; 10(10):e0141170. https://doi.org/10.1371/journal.pone.0141170 PMID: 26513361.

27. Moioli B, Napolitano F, Catillo G. Genetic diversity between Piedmontese, Maremmana, and Podolica cattle breeds. J Hered. 2004; 95(3):250–6. PMID: 15220392.

28. Pariset L, Mariotti M, Nardone A, Soysal MI, Ozkan E, Williams JL, et al. Relationships between Podolic cattle breeds assessed by single nucleotide polymorphisms (SNPs) genotyping. J Anim Breed Genet. 2010; 127(6):481–8. Epub 2010/10/28. https://doi.org/10.1111/j.1439-0388.2010.00868.x PMID: 21077972.

29. Gargani M, Pariset L, Lenstra JA, De Minicis E, Valentini A, Consortium ECGD. Microsatellite genotyping of medieval cattle from central Italy suggests an old origin of Chianina and Romagnola cattle. Front Genet. 2015; 6:68. https://doi.org/10.3389/fgene.2015.00068 PMID: 25788902; PubMed Central PMCID: PMCPMC4349168.

30. Upadhyay MR, Chen W, Lenstra JA, Goderie CR, MacHugh DE, Park SD, et al. Genetic origin, admixture and population history of aurochs (*Bos primigenius*) and primitive European cattle. Heredity (Edinb). 2017; 118(2):169–76. Epub 2016/09/28. https://doi.org/10.1038/hdy.2016.79 PMID: 27677498; PubMed Central PMCID: PMCPMC5234481.

31. Lenstra J, Ajmone, Marsan P, Beja Pereira A, Bollongino R, Bradley DG, et al. Meta-Analysis of Mitochondrial DNA Reveals Several Population Bottlenecks during Worldwide Migrations of Cattle. Diversity. 2014; 6 178–87.

32. Cerezo M, Achilli A, Olivieri A, Perego UA, Gómez-Carballa A, Brisighelli F, et al. Reconstructing ancient mitochondrial DNA links between Africa and Europe. Genome Res. 2012; 22(5):821–6. https://doi.org/10.1101/gr.134452.111 PMID: 22454235; PubMed Central PMCID: PMCPMC3337428.

33. Ascunce S, Kitchen A, Schmidt P, Miyamoto M, Mulligan C. An Unusual Pattern of Ancient Mitochondrial DNA Haplogroups in Northern African Cattle. Zool Stud. 2007; 46(1):123–5.

34. Mason I. A world dictionary of livestock breeds, types and varieties. 4th ed. 1996.

35. Gómez-Carballa A, Pardo-Seco J, Amigo J, Martinón-Torres F, Salas A. Mitogenomes from The 1000 Genome Project reveal new Near Eastern features in present-day Tuscans. PLoS One. 2015; 10(3): e0119242. Epub 2015/03/18. https://doi.org/10.1371/journal.pone.0119242 PMID: 25786119; PubMed Central PMCID: PMCPMC4365045.

36. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, Battaglia V, et al. Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am J Hum Genet. 2007; 80(4):759–68. Epub 2007/02/06. https://doi.org/10.1086/512822 PMID: 17357081; PubMed Central PMCID: PMCPMC1852723.

37. Brisighelli F, Capelli C, Alvarez-Iglesias V, Onofri V, Paoli G, Tofanelli S, et al. The Etruscan timeline: a recent Anatolian connection. Eur J Hum Genet. 2009; 17(5):693–6. Epub 2008/12/03. https://doi.org/10.1038/ejhg.2008.224 PMID: 19050723; PubMed Central PMCID: PMCPMC2986270.

38. Pardo-Seco J, Gómez-Carballa A, Amigo J, Martinón-Torres F, Salas A. A genome-wide study of modern-day Tuscans: revisiting Herodotus's theory on the origin of the Etruscans. PLoS One. 2014; 9(9): e105920. Epub 2014/09/17. https://doi.org/10.1371/journal.pone.0105920 PMID: 25230205; PubMed Central PMCID: PMCPMC4167696.

39. Rognoni G, Pagnacco G. Atlante Etnografico delle popolazioni bovine allevate in Italia. Rome: Consiglio Nazionale delle Ricerche; 1983.

40. Achilli A, Olivieri A, Pellecchia M, Uboldi C, Colli L, Al-Zahery N, et al. Mitochondrial genomes of extinct aurochs survive in domestic cattle. Curr Biol. 2008; 18(4):R157–8. https://doi.org/10.1016/j.cub.2008.01.019 PMID: 18302915.

41. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11):1451–2. Epub 2009/04/03. https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325.

42. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2007; 1:47–50. Epub 2007/02/23. PMID: 19325852; PubMed Central PMCID: PMCPMC2658868.

43. Lindgren G, Backström N, Swinburne J, Hellborg L, Einarsson A, Sandberg K, et al. Limited number of patrilines in horse domestication. Nat Genet. 2004; 36(4):335–6. https://doi.org/10.1038/ng1326 PMID: 15034578.

44. Cardinali I, Lancioni H, Giontella A, Capodiferro MR, Capomaccio S, Buttazzoni L, et al. An Overview of Ten Italian Horse Breeds through Mitochondrial DNA. PLoS One. 2016; 11(4):e0153004. Epub 2016/04/07. https://doi.org/10.1371/journal.pone.0153004 PMID: 27054850; PubMed Central PMCID: PMCPMC4824442.