Unione europea
Fondo sociale europeo

REPUBBLICA ITALIANA

REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA

POR SARDEGNA
obiettivo competitività regionale e occupazione

UNIVERSITÀ DEGLI STUDI DI SASSARI

Dissertation for the Degree of Doctor of Philosophy in Environmental Biology
presented at Sassari University in 2013

XXVI cycle

ANALISI DI SISTEMI GENETICI A TRASMISSIONE UNIPARENTALE (ANALYSIS OF UNIPARENTALLY TRASMITTED GENETIC SYSTEMS)

PH.D. CANDIDATE: ***Dr. Antonella Useli***

DIRECTOR OF THE SCHOOL: ***Prof. Marco Curini Galletti***

SUPERVISOR: ***Prof. Paolo Francalacci***

La presente tesi è stata prodotta nell'ambito della scuola di dottorato in Scienze della Natura e delle Sue Risorse dell'Università degli Studi di Sassari, a.a. 2010/2011 – XXVI ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività l.3.1.

This PhD Thesis is dedicatet to Laura Morelli

SUMMARY

[ENGLISH] The analysis of the variability of the uniparentally transmitted genetic systems, as the mitochondrial DNA (mtDNA) and the non-recombining region of Y chromosome (NRY), is considered a useful tool to study evolutionary processes. In the present PhD Thesis both haploid systems were analyzed in different organisms by a phylogenetic and phylogeographic approaches. In the first part, the variability of mtDNA has been analyzed in domestic Horse. In particular, two Sardinian native breeds -the Giara and Sarcidano horse- were studied in order to determine their genetic structure and phylogenetic history by a comparison in a global framework. The project regarding the creation of a new informatics tool- *HapSign Software*- to assign mtDNA haplotypes is also reported with the first results of application in horse breeds. The second part encompasses a study focused on the mtDNA variation of a Sardinian endemic genus of Carabidae (Coleoptera) of the subterranean fauna, with the aim to establish the phylogenetic relations among species already known and potentially new ones, coming from some Sardinian caves. Finally, some studies about mtDNA and Y chromosome variation in human populations were conducted; the first one concern the high resolution phylogeny of the NRY of Sardinian population by the sequencing of a large portion of Y chromosome. The other studies has been focused on Italian and African populations in order reconstruct their history from a genetic perspective.

KEY WORDS: mitochondrial DNA, Y chromosome, phylogeny, human populations, horse breeds.

.

RIASSUNTO

[ITALIAN] L'analisi della variabilità del DNA mitocondriale e del cromosoma Y è considerata uno strumento utile nello studio dei processi evolutivi. Nella presente tesi di dottorato, entrambi i sistemi aploidi sono stati analizzati, con un approccio filogenetico e filogeografico. Nella prima parte la variabilità del mtDNA è stata analizzata negli Equidi domestici. In particolare sono state studiate due razze equine native della Sardegna –il cavallo della Giara e del Sarcidano- al fine di determinare la struttura genetica e la storia delle due razze da un punto di vista filogenetico, confrontandole in un contesto globale. E' stato inoltre riportato un progetto riguardante la creazione di un nuovo Software HapSign, per assegnare gli aplotipi del mtDNA e i primi risultati di applicazione nelle razze equine. La seconda parte riguarda uno studio sulla variabilità del mtDNA in un genere endemico della Sardegna appartenete ai Carabidae (Coleoptera) appartenente alla fauna ipogea, al fine di stabilire le relazioni filogenetiche tra specie note e potenzialmente nuove di alcune grotte sarde. Infine, sono stati condotti alcuni studi sulla variabilità del mtDNA e del cromosoma Y nelle popolazioni umane; il primo riguarda la filogenesi ad alta risoluzione della NRY nella popolazione sarda, attraverso il sequenziamento di una larga porzione del chromosoma. Gli altri studi hanno riguardato popolazioni italiane e africane al fine di ricostruire la loro storia dal punto di vista genetico.

PAROLE CHIAVE: DNA mitocondriale, cromosoma Y, filogenesi, razze equine, popolazioni umane.

INTRODUCTION.

The analysis of the variability of the uniparentally transmitted genetic systems, as the mitochondrial DNA (mtDNA) and the non recombining region of Y chromosome (NRY), is considered a useful tool to study evolutionary processes.

Due to their peculiar inheritance way, the mitochondrial genome and the NRY, matrilineal and patrilineal respectively, are not subject to recombination event during meiosis. As a primary consequence, they pass unaltered through generations, except for the occurrence of mutational events, which represent the only source of new variation. As a secondary consequence, the occurred mutations may accumulate sequentially along maternal and paternal lineages over time. In addition, the analysis of variability of haploid genetic systems by an adequate number of mutations, which should be more precisely defined as stable polymorphisms, could allow the reconstruction of the phylogenetic relationships among lineages. Once identified groups of lineages (haplotypes) closely related (haplogroups), sharing mutational patterns derived from a common ancestor, it is possible to study the diversification/ramification process along time on the constructed phylogenetic trees. On the bases of the reconstructed phylogenies it is also possible, going backward along the lineages, to reach the Most Recent Common Ancestor (MRCA), under neutrality conditions. Applying a valid calibration method to the molecular-clock and defining the phylogenetic rate of mutational events, the MRCA and the clades could be located in a temporal framework. In some cases, as in the well known case of human populations, it is evident that the haplogroups are not geographically random distributed. By the application of the phylogeographic approach it is possible to study the genetic variation in space or, in other words, to locate the phylogenies in a geographical framework. These characteristic properties could finally help to make inferences on the evolutionary history of populations or of any other taxonomic level.

In the last decades, several studies based on mtDNA and NRY has been conducted to address a wide range of issues in the evolutionary research field. Beside such kind of specific studies, it is becoming more and more frequent the use of the unilinear transmitted genetic systems in other biological related fields as forensic genetics (Gurney et al. 2010; Caniglia et al. 2013), health sciences (Achilli et al. 2011; Dowling 2013), molecular ecology (Warmuth et al. 2013) and conservation genetics (Alvarez et al. 2012), just to cite the most relevant ones. In the general context of the increasing interest for the haploid molecules, mitochondrial DNA is by far the more studied. In fact, from a certain point of view, the mtDNA can offer a major number of advantages

that, regarding animals, could be summarized as follow: it is a small circular molecule, in general about 15-20 kilo base pairs (bp) long; it is constituted by a coding region, poor of introns, carrying 37 genes in the majority of taxa (22 tRNA, 2 rRNA and 13 sub-units of the oxidative phosphorylation chain), and a control region called displacement loop (d-loop) of about 1200 bp. This last one is composed by two hypervariable segments (HVS) called HVS-I and HVS-II; it is present at a very high number for cell and shows an evolutionary rate higher than nuclear DNA. As expected for any genome portion, the mutation rate is not constant along the molecule, being actually faster in HVS and particularly in HVS-I, which offers a higher number of polymorphisms. Mainly for this last reason, several studies are based on HVS-I variation. Anyway, a higher mutations rate imply a higher rate of reversion and mutational hot spots; as consequence, it is expected a higher number of recurrent mutations occurring in lineages not phylogenetically related. Moreover, depending on the aim of the study and the taxonomic level considered, should be take into account that the amount of variation observed under a too fast mutation rate could represent a short time frame in order to highlight the deep clades of the inferred phylogenetic trees. This effect could also makes the signal of phylogeographic structure too weak to be revealed. For all these reasons the general tendency in this field has been to couple the d-loop sequence by more stable genetic markers of the coding region and, subsequently, to keep information from the whole mitochondrial genome sequence. In the last years, thanks to technical advances of sequencing methods as Next-Generation Sequencing (NGS), accompanied by the decreasing of costs, the number of organisms whose complete mtDNA sequence has become available is dramatically increased. The re-sequencing approach of both mtDNA (Torroni et al. 2006; Behar et al. 2006; Pala et al. 2012) and NRY (Wei et al. 2012; Poznik et al. 2013; Francalacci et al. 2013) is giving important results in human evolutionary studies and more recently even in animal genetics (Bonfiglio et al. 2011; Achilli et al. 2012; Gazave et al. 2013). Anyway, in the field of animal genetics beside this approach we can still observe a variety of studies laying at different stages, regarding the exploration of the potential applications of unilinear systems variation analysis. In this framework, the d-loop variation is still the most analyzed in intraspecific analysis but could be interesting to report how also other fragments of the coding region of the mtDNA are largely used. The proposal of the DNA Barcode system to identify animal taxa done by Hebert and collaborators (2003) has led to the accumulation of a large amount of sequences of a short fragment of the first sub-unit of the Cytochrome c Oxydase gene (COI). The spread of COI fragment analysis, is not

only limited to the barcoding framework, but it represents a signal of an important change in systematic and taxonomic methodologies. In fact, the molecular data are even more frequently used to integrate biological, morphological, ecological and other kind of data by the use of a multidisciplinary approach. The fragments of COI gene, sometimes combined with other mitochondrial or nuclear genes, is largely being studied in various phyla, at different taxonomic levels, ranging from the resolution of phylogenies and biogeography patterns or the recognition of morphological cryptic species, to revealing the connectivity of populations by a phylogeographic perspective (Dailianis et al. 2011; Ribera et al. 2010; Sanna et al. 2013).

In the described composite picture, it is relevant to note that the investigation based on unilinear systems variation of domestic animals is rapidly following the wake of human population evolutionary approach. In fact, beside the huge accumulation of mtDNA HVS-I sequences, the new assemblies and re-sequencing of complete mitochondrial genomes of an increasing number of samples became prevalent in the last few years. As is occurred in human populations molecular studies, data on Y chromosome markers is being discovered with a certain delay but this is likely related to the intrinsic characteristics of the two molecules, as reported above (Lindgreen et al. 2004; Brandariz-Fontes et al. 2013). Anyway, the main reasons of the progress made are probably related to the growing interest in domestic species due to their strength relationships with humans. Starting from the Neolithic revolution, the worldwide development of human societies was strongly dependent on domestication dynamics of livestock species. The study of these events could improve the knowledge on human history, giving a parallel view of human population dynamics too. Presently, the importance of domestic animals it was also referred claiming they represent a relevant source of biodiversity to preserve for the future. Therefore, it was emphasized the importance of the knowledge of the genetic structure of native breeds, particularly if reared in traditional way, in order to identify the correct conservation strategies (Medugorac et al. 2009; Groeneveld et al. 2010; Lenstra et al. 2012).

In the present PhD Thesis the following main research topics has been addressed: the first part is related to the analysis of mitochondrial DNA variation in the domestic species of the genus Equus, particularly in *Equus caballus* (*E. caballus*), while the second part concern the analysis of uniparentally transmitted genetic systems in other species of phylogenetic interest.

Regarding the first part, the variability of mtDNA has been analyzed in the Italian Giara and Sarcidano horse breeds. They are two native breeds from Sardinia Island which are recognize as

breeds with limited diffusion and thus considered important to preserve. The study was conducted order to determine the genetic structure of the two breeds and their phylogenetic history in a global framework, represented by other native horse breeds of the "Old World" (Chapter 1).

Within the same part a second study, reported as work in progress, has been projected and carried out in order to obtain a new automatic tool, which could allow a rapid and accurate assignation of a query mitochondrial haplotype to its haplogroup. This Software would be apply especially to mtDNA analysis of domestic animals. At this stage, the validating tests of the method are carrying in *Equus caballus,* considered as model due to the well definite phylogenesis of mtDNA haplogroups in horse breeds (Chapter 2).

The second part encompass one study focused on the mitochondrial variation of a Sardinian endemic genera of Carabidae (Coleoptera) of the subterranean fauna (Chapter 3) and some studies about mtDNA and Y chromosome variation in human populations (Chapter 4).

The ongoing study about the Carabidae of the Ovobathysciola genus of the Leptodirini tribe (family Leiodidae), has been analyzed by the means of the 3'-end of COI sequence variation. The mtDNA analysis has been done to establish the phylogenetic relations among species already known and potentially new ones, of various Sardinian caves. The integration with the morphological data will allow to establish the phylogenetic position of each species with more precision and to clarify their complex biogeographic pattern.

The studies about human populations can be subdivided in two sections. The first one concern the high resolution phylogeny of the NRY, also called male-specific portion of Y chromosome (MSY) of Sardinian population. This study has been designed to attempt to overcome the limits of the phylogenetic resolution of the Y chromosome tree of human populations, related to the analysis of a limited numbers of markers and samples. A large portion of Y chromosome has been analyzed in about 1200 Sardinian males Y chromosome by a NGS technology. In addition, the use of a phylogenetic rate was applied to dating estimates to calculate the putative age of the nodes of the inferred parsimony tree and of the MRCA of human populations (Paragraph 4.1). The second one concern some studies that has been focused on a regional scale. A wide set of mtDNA and MSY markers has been analyzed in order to update and to improve the knowledge of Italian populations history by a uniparental markers point of view. The presence of a sex-biased structure has been also investigated (Paragraph 4.2). The mtDNA variation has been used to better reveal the genetic structure of populations from Eastern Africa in relation to the complex linguistic pattern of the

region (Paragraph 4.3). The combined analysis of patrilineal markers, genetic and cultural as the NRY and the surnames, has been applied to infer the recent history of an Italian population of a microgeographic area in the Upper Savio Valley (Central Appennines) (Paragraph 4.4).

The variation of both mtDNA and NRY has been investigated in phylogenetic distant organism at an intra- and interspecific level. The different contexts of each study, even in terms of kind and quantity of markers used and methods, could allow making a comparison of the informative power of the analyzed genetic systems.

All the specific aims of the carried studies can be unified in the overall purpose to analyze the information contained in these small portions of the genome, as preferential windows opened over the past, to improve the knowledge of the evolutionary history of species.

BIBLIOGRAPHY.

Achilli A, Olivieri A, Pala M, Hooshiar Kashani B, Carossa V, Perego UA, Gandini F, Santoro A, Battaglia V, Grugni V, Lancioni H, Sirolla C, Bonfigli AR, Cormio A, Boemi M, Testa I, Semino O, Ceriello A, Spazzafumo L, Gadaleta MN, Marra M, Testa R, Franceschi C, Torroni A. (2011). Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole. PLoS One.;6(6):e21029.

Achilli A, Olivieri A, Soares P, Lancioni H, et al. (2012). Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc. Natl. Acad. Sci. USA. 109:2449-54.

Alvarez I, Fernandez I., Lorenzo L., Payeras L., Cuervo M. & Goyache F. (2012) Founder and present maternal diversity in two endangered Spanish horse breeds assessed via pedigree and mitochondrial DNA information J. Anim. Breed. Genet. 129 (2012) 271-279

Antonio Torroni, Alessandro Achilli, Vincent Macaulay, Martin Richards and Hans-Jurgen Bandelt (2006). Harvesting the fruit of the human mtDNA tree. TRENDS in Genetics Vol.22 No.6

Bonfiglio S, Achilli A, Olivieri A, Negrini R, Colli L, Luigi Liotta, Paolo Ajmone-Marsan, Antonio Torroni, Luca Ferretti (2010) The Enigmatic Origin of Bovine mtDNA Haplogroup R: Sporadic Interbreeding or an Independent Event of Bos primigenius Domestication in Italy? PLoS ONE 5(12): e15760.

Brandariz-Fontes C, Leonard JA, Vega-Pla JL, Backstrom N, Lindgren G, Lippold S., Rico C. (2013) Y-Chromosome Analysis in Retuertas Horses. PLoS ONE 8(5): e64985.

Caniglia R, Fabbri E, Mastrogiuseppe L, Randi E. Who is who? Identification of livestock predators using forensic genetic approaches. Forensic Sci Int Genet. (2013) 7(3):397-404.

Dailianis T, Tsigenopoulos CS, Dounas C & Voultsiadou E (2011) Genetic diversity of the imperiled bath sponge Spongia officinalis Linnaeus, 1759 across the Mediterranean Sea: patterns of population differentiation and implications for taxonomy and conservation. Molecular Ecology (2011) 20, 3757-3772.

Doron M. Behar, Ene Metspalu, Toomas Kivisild, Alessandro Achilli, Yarin Hadid, Shay Tzur, Luisa Pereira, Antonio Amorim, Llu?s Quintana-Murci, Kari Majamaa, Corinna Herrnstadt, Neil Howell,8 Oleg Balanovsky, Ildus Kutuev,Andrey Pshenichnov, David Gurwitz, Batsheva Bonne-Tamir, Antonio Torroni, Richard Villems, and Karl Skorecki (2006). The Matrilineal Ancestry of Ashkenazi Jewry: Portrait of a Recent Founder Event. Am. J. Hum. Genet. 2006;78:487-497.

Dowling DK. Evolutionary perspectives on the links between mitochondrial genotype and disease phenotype. (2013) Biochim Biophys Acta. 2013 Nov 15. pii: S0304-4165(13)00496-0.

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, Sanna D, Useli A, Urru MF, Marcelli M, Cusano R,Oppo M, Zoledziewska M, Pitzalis M, Deidda F, Porcu E, Poddie F, Kang HM, Lyons R, Tarrier B, Gresham JB, Li B, Tofanelli S, Alonso S, Dei M, Lai S, Mulas A, Whalen MB, Uzzau S, Jones C, Schlessinger D, Abecasis GR, Sanna S, Sidore C, Cucca F. (2013). Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science. 341(6145):565-9.

Gazave E, Lape´bie P, Renard E, Vacelet J, Rocher C, Alexander V.Ereskovsky, Dennis V. Lavrov, Carole Borchiellin. (2010) Molecular Phylogeny Restores the Supra-Generic Subdivision of Homoscleromorph Sponges (Porifera, Homoscleromorpha). PLoS ONE 5(12): e14290.

Groeneveld L. F., Lenstra J. A., Eding H.,. Toro M. A, Scherf B., Pilling D., Negrini R., Finlay E. K., Jianlin H., Groeneveld E., Weigend S. and The GLOBALDIV Consortium (2010). Genetic diversity in farm animals - a review. Animal Genetics, 41 (Suppl. 1), 6-31.

Gurney SM, Schneider S, Pflugradt R, Barrett E, Forster AC, Brinkmann B, Jansen T, Forster P. Developing equine mtDNA profiling for forensic application (2010). Int J Legal Med. 124(6):617-22.

Lenstra J. A., Groeneveld L. F., Eding H., Kantanen J., Williams J. L., Taberlet P., Nicolazzi E. L., Solkner J., Simianer H., Ciani E., Garcia J. F., Bruford M. W., Ajmone-Marsan P., and Weigend S.. (2012). Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. Review, Animal Genetics, 43, 483-502.

Lindgren G, Backstrom N, Swinburne J, Hellborg L, et al. (2004). Limited number of patrilines in horse domestication. Nat. Genet. 36:335-336

Medugorac I., Medugorac A., Russ I., Veit-Kensch C.E., Taberlet P., Luntz B., Mix H.M., Forster M. (2009) Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. Molecular Ecology 18, 3394-3410.

Morin PA, Archer FI, Foote AD, Vilstrup J, Allen EE, Wade P, Durban J, Parsons K, Pitman R, Li L, et al: Complete mitochondrial genome phylogeographic analysis of killer whales (Orcinus orca) indicates multiple species. Genome Res 2010, 20:908-916.

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, Tamm E,Karmin M, Reisberg T, Hooshiar Kashani B, Perego UA, Carossa V, Gandini F, Pereira JB, Soares P, Angerhofer N, Rychkov S, Al-Zahery N, Carelli V, Sanati MH, Houshmand M, Hatina J, Macaulay V, Pereira L, Woodward SR, Davies W, Gamble C, Baird D, Semino O, Villems R, Torroni A, Richards MB. (2012).Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. Am J Hum Genet. 4;90(5):915-24.

Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA, Bustamante CD. (2013). Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science. 341(6145):562-5.

Ribera, I., Fresneda, J., Bucur, R., Izquierdo, A., Vogler, A.P., Salgado, J.M. & Cieslak, A. (2010) Ancient origin of a Western Mediterranean radiation of subterranean beetles. BMC Evolutionary Biology, 10, 29.

Sanna D, Cossu P, Dedola GL, Scarpa F, Maltagliati F, Castelli A, Franzoi P, Lai T, Cristo B, Curini-Galletti M, Francalacci P, Casu M (2013) Mitochondrial DNA Reveals Genetic Structuring of Pinna nobilis across the Mediterranean Sea. PLoS ONE 8(6): e67372.

Warmuth VM, Campana MG, Eriksson A, Bower M, Barker G, Manica A. Ancient trade routes shaped the genetic structure of horses in eastern Eurasia (2013). Mol Ecol. 22(21):5340-51.

Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C. (2013). A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res.;23(2):388-95.

CHAPTER 1
Mitochondrial DNA lineages of Italian Giara and Sarcidano horses

Running title: mtDNA lineages in Sardinian horses

L. Morelli[1§], **A. Useli**[1✉], **D. Sanna**[1], **M. Barbato**[1,2], **D. Contu**[3], **M. Pala**[4,5], **M. Cancedda**[6] and **P. Francalacci**[1]

[1] Dipartimento di Scienze della Natura e del Territorio - Unità di Zoologia Archeozoologia e Genetica -, Università di Sassari, Sassari, Italy
[2] Cardiff School of Biosciences, Cardiff University, Cardiff, UK
[3] Laboratorio di Immunogenetica, Ospedale Microcitemico, Via Jenner, Cagliari, Italy
[4] Dipartimento di Genetica e Microbiologia, Università di Pavia, Pavia, Italy
[5] School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK
[6] Dipartimento di Biologia Animale, Facoltà di Medicina Veterinaria, Università di Sassari, Sassari, Italy
[§] This paper is dedicated to the memory of Laura Morelli who prematurely passed away before the publication

✉ **Corresponding author**: Antonella Useli
Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, I-07100 Sassari, Italy
Tel. + 39 79 228630, Fax. + 39 79 228665
Email: auseli@uniss.it;

ABSTRACT

Giara and Sarcidano are 2 of the 15 extant native Italian horse breeds with limited dispersal capability that originated from a larger number of individuals. The 2 breeds live in two distinct isolated locations on the island of Sardinia. To determine the genetic structure and evolutionary history of these 2 Sardinian breeds, the first hypervariable segment of the mitochondrial DNA (mtDNA) was sequenced and analyzed in 40 Giara and Sarcidano horses and compared with publicly available mtDNA data from 43 Old World breeds. Four different analyses, including genetic distance, analysis of molecular variance (AMOVA), haplotype sharing, and clustering methods were used to study the genetic relationships between the Sardinian and other horse breeds. The analyses yielded similar results, and the average F_{ST} values indicated that approx. 10% of the total genetic variation was explained by between-breed differences. Consistent with their distinct phenotypes and geographic isolation, the two Sardinian breeds were shown to consist of 2 distinct gene pools that had no gene flow between them. Giara horses were clearly separated from the other breeds examined and showed traces of ancient separation from horses of other breeds that share the same mitochondrial lineage. On the other hand, the data from the Sarcidano horses fit well with

variation among breeds from the Iberian Peninsula and North-West Europe: genetic relationships among Sarcidano and the other breeds are consistent with the documented history of this breed.

Key words: Mitochondrial DNA; Giara horse; Sarcidano horse; Haplogroup attribution; Domestication

INTRODUCTION

Island environments enable the persistence of relic varieties of species or breeds. Among the 15 horse breeds officially recognized by the Italian Registry of Autochthonous Equine Breeds, 2 breeds, which are inbred and phenotypically distinctive, are from the island of Sardinia (Figure 1):

- The Giara horses take their name by the basalt plateau in the central/south region of Sardinia where they live in the wild. The Giara plateau extends over an area of 45 km^2 at an altitude of about 500–600 m above sea level. The steep mountain slopes limit connections with the surrounding valleys and prevent migration of the horses. The average height of Giara horse approaches that of a pony, but it is considered a miniature horse. The bioecological features of this population make it a rich livestock heritage to safeguard and a guarantee of protection for the natural environment where it lives (Gratani, 1980).
- The Sarcidano horses are concentrated on a single farm in Laconi (in central-western Sardinia). The history of how the breed was introduced onto the island is controversial and mostly unknown; nonofficial records suggest a descent from the ancient Spanish horse, an ancestor of the Andalusian breed.

Figure 1. Giara and Sarcidano breeds areal map.

At the end of the 18th century, Cetti (1774) described the presence on Sardinia of 3 different types of horses, the "*selvaticus*" (wild), the "*vulgar*" (common or ordinary), and the "*di razza*" (thoroughbred) horse. Cetti's description of the height and temperament of "*selvaticus*" is consistent with the descriptions of the present-day Giara horse. On the other hand, "*the vulgar*" Cetti's horse description fits well with that of the present population of the Sarcidano breed in terms of character, phenotype, and work attitude. At present, there are 481 registered Giara and 108 Sarcidano horses (http://www.anagrafeequidi.it/index.php?id=217).

The aim of this work was to shed light on the genetic structure of the Giara and Sarcidano horses in a global context to depict their genetic relationships with other present-day breeds. To achieve this goal, we compared a sample of the 2 Sardinian breeds with a large number of horses genotyped to

date. For this comparison, we selected a 247-bp long internal portion of the hyper variable region segment I (HVS-I) of the control displacement loop (d-loop) of mitochondrial DNA (mtDNA). This choice was made for three reasons: (1) sequence variation in these mtDNA regions is solely generated by the sequential accumulation of new mutations along radiating maternal lineages; (2) maternal lineages are presumably more stable because of the practice to move stallions for reproduction and breed improvement, so mtDNA sequences should present breed-specific motifs that relate our samples to a geographical origin; and (3) a large number of d-loop sequences from many horse breeds is available in GenBank.

To date, only limited data (Cozzi et al., 2004, Achilli et al., 2012) on mtDNA variation in Sardinian horse breeds are available. A comparison of a limited number of Sardinian horses with other Italian breeds suggests a reduced relationship with the other Italian populations (Cozzi et al., 2004).

MATERIALS AND METHODS

Materials

Using standard procedures, total DNA was extracted from peripheral blood samples of 24 horses of the Giara (JAR) breed and 16 horses of the Sarcidano (SAR) breed. The Giara and Sarcidano horses had been bred in semiferal conditions, and animals were randomly selected by capture.

The sequences produced in this study were pooled with 5 sequences from Giara (GRH1-5; GenBank accessions AY462426–AY462430) and 5 from Sarcidano (SRH1-5; AY462451–AY462455) breeds previously reported by Cozzi et al (2004), and with 2 additional HVS-I sequences identified in the complete mtDNA genome sequences (Gia01 JN398411 and Gia02, JN398407) reported by Achilli et al (2012). The final sample set used in the analyses was obtained from 31 Giara and 21 Sarcidano horses.

In addition, all the 150 complete mtDNA genome sequences available from literature (NC_001640 [Xu and Arnason, 1994]; EF597512–14 [Xu et al., 2007]; AP012267–70 [Goto et al., 2011]; EU939445 [Jiang et al., 2011]; HQ439441–500 [Lippold et al., 2011]; and JN398377–457 [Achilli et al., 2012]) were used to obtain more reliable and informative HVS-I pattern of sites defining haplogroups.

Finally, we produced a dataset of typical regional breeds by selecting 1,192 HVS-I horse sequences from the literature for which frequency population data were available and which were reported for at least 15 individuals. Taking into account the haplotype frequencies, we obtained 1,232 HVS-I samples belonging to 45 breeds (including Giara and Sarcidano) representing 6 geographic Old

World macro areas: the Iberian Peninsula (N = 220), Central Europe (N = 400), North-west Africa (N = 40), Arabian Peninsula (N = 70), Central Eurasia (N = 113), and the Far East (N = 389).

Methods

The HVS-I of the d-loop region was amplified by the polymerase chain reaction (PCR) using two primers from a published horse sequence (GenBank access. X79547): forward 5′-AACGTTTCCTCCCAAGGACT-3′ and reverse 5′-GTAGTTGGGAGGGTTGCTGA-3′ (Ishida, 1994; Xu and Arnason, 1994). The amplicon obtained was a 397-bp fragment included between the $tRNA^{Pro}$ gene and the large central conserved sequence block from nucleotide position (np) 15382–15778. The PCR products were purified by using ExoSAP-IT (USB Corporation) and sequenced using the BigDye Terminator Kit (Applied Biosystems) on an ABI PRISM 377 DNA Sequencer equipped with the Sequencing Analysis and Sequence Navigator programs (Applied Biosystems). Sequence alignments were performed with the software BioEdit 7.0.5.2 (Hall, 1999).

Intra- and interpopulation level methods were conducted with the Arlequin 3.5. software (Excoffier et al., 2007) (http://cmpg.unibe.ch/software/arlequin3): intrapopulation level variation was estimated with both standard (Gene diversity of Nei, 1987) and molecular indices as pairwise differences (Tajima, 1993) and nucleotide diversity (Tajima, 1983; Nei, 1987). Population genetic structure was obtained by hierarchical analysis of the total variance subdivided in percentage of variance within the breeds, among breeds within groups and among groups by using molecular analysis of variance (AMOVA) (Excoffier et al., 1992) taking into account the number of mutations between molecular haplotypes. In both cases, the F-statistic was set at a significance level of 0.05, obtained by 10,000 permutation tests. The matrix of interpopulation pairwise distances (Tajima, 1993, Arlequin software) was summarized in two dimensions by using multidimensional scaling (MDS) analysis as implemented by the STATISTICA ’99 software (StatSoft, Tulsa, Oklahoma, USA) and plotted on a MS Excel graphic.

Haplogroups attribution was performed following the nomenclature rules of Achilli et al. (2012). To increase the power of imputation of the HVS-I sequences, we pooled together all the available horse complete mtDNA genome sequences. Polymorphic sites occurring among the total of 150 sequences were exported as an Excel spreadsheet and the haplotypes were organized in haplogroups following a hierarchical and parsimonious order, and the haplogroup name was assigned to the unclassified data. Three of them were eliminated because their polymorphisms were not consistent

with phylogeny as previously observed by Lippold et al. (2011) and Achilli et al. (2012). The mutational pattern was dissected in order to define the following variables:

- variation associated univocally with the haplogroup; this variation is due to the most informative mutations since they are present in all of the haplotypes of the same haplogroup and absent in other haplogroups;
- variation associated univocally with the super-haplogroup; this variation is due to mutations that allow allocation to a unique clade represented by individuals that share the same mutation because it is ancestral for all the haplogroups that compose it;
- variation due to mutations occurring in a unique haplogroup but not in all of the haplotypes of this haplogroup;
- variation due to informative mutations in the allelic association;
- variation due to poorly informative mutations that are the result of recurrence or reversion.

The haplotypic pattern of the HVS-I region was extracted and a probabilistic algorithm defining haplogroup attribution was generated when only the HVS-I sequence was available. Haplotype variation and haplotype sharing into the haplogroups were evaluated by clustering in the Network program 5.0.0, and default parameters were used for obtaining the median-joining network trees (Bandelt et al., 1999). A weight of 0 were assigned to mutations classified as recurrent. In addition, mismatch distribution of the number of pairwise differences between haplotypes among haplogroups and associated demographic parameters including Harpending's raggedness index (r), Tajima's *D* (D), and Fu's (Fs) statistics were calculated using Arlequin 3.5.

RESULTS

Diversity indices

The 52 sequences (247-bp long) obtained from the Sardinian horses consisted of 29 different haplotypes on the segment ranging from np 15494 to 15740 (Table 1).

Table 1. Polymorphic sites of mtDNA HVS-I in Sardinian breeds.

We calculated the diversity indices from 43 native breeds scattered in the Old World (see Tables 2 for details).

Table 2. The Old World native breeds.

We compared diversity indices from Old World breeds to those obtained from the Sardinian horses: both the haplotype diversity values of the Giara (0.847 ± 0.053) and Sarcidano (0.905 ± 0.047) breeds were higher than the average estimates for the other 43 breeds, but only the Sarcidano breed was above the median of the distribution (Table 3). Other molecular diversity indices that were also considered (see Table 3) showed that the heterogeneity of the Sarcidano sample (mean number of pairwise differences, 6.738 ± 3.309 and nucleotide diversity, 0.027 ± 0.015) was comparable with the highest values reported for the other breeds, whereas Giara horses showed a lower molecular diversity (mean number of pairwise differences, 3.933 ± 2.025 and nucleotide diversity, 0.016 ± 0.009).

Table 3. Intra-population level variation of Old World native breeds.

Genetic structure

Next, we used AMOVA on the basis of the pairwise difference distance method (Excoffier et al., 1992; Weir, 1996) to determine the genetic structure of the group composed by the Giara and the Sarcidano populations. A relevant and significant percentage of interpopulation variation (45.1%) was detected when compared with the intrapopulation variation (54.9%; $P < 10^{-5}$). Therefore, taking into account that in all of the 45 worldwide diffused breeds analyzed, 25% ($P < 10^{-5}$) of the total variation is allocated to the among-breed source, we can confidently predict the presence of a genetic barrier between the two Sardinian breeds. The same value was obtained after grouping into macrogeographical areas as reported in column 1 of Table 2, and no variance was attributable to the differences between groups. Just a small, but significant, amount of variation was related to the east (Arabian Peninsula, Central Eurasia, Far East) and west (north-west Africa, Iberian Peninsula, Central Europe) groupings (2.89%; $P = 0.01$).

The relationship between the different breeds was inferred by estimating the pairwise differences between breeds and interpolating the data into a MDS graphic (Figure 2).

Figure 2. Multidimensional scaling (MDS) plot computed on the basis of the matrix of the pairwise differences of the mitochondrial HVS-I sequences. Each symbol represents one of the 45 populations tested. Breed codes are as in Table 2. A) in the graph only those populations outside the main cluster (enclosed within the square with the broken line and including breeds with high amount of intra population variation) are named. B) in the graph the square including the main cluster is enlarged and breeds are named. D-star: Raw stress = 30.54; Alienation = 0.12; D-hat: Raw stress = 24.05; Stress = 0.11.

This graphic showed a large presence of outliers representing low variability (Figure 2A and see also Table 3) and clustering of the majority of breeds having higher variation (Figure 2A and 2B enlarged). The Giara sample is in an outlier position because of its low variability. In contrast, the Sarcidano sample is located in an area of generally high variability, included in the Western European group of breeds.

Haplogroup assignment

The phylogenetic analysis involving 147 complete mtDNA genome sequences reported from the literature allowed us to infer the HVS-I haplotype patterns and assign them to the 18 (A–R) haplogroups (Achilli et al., 2012) with better confidence. HVS-I was affected by variation that differed in the quality and degree of information (see methods). The haplogroups D, F, H, I, L, M, N, Q, and R were defined by highly informative mutations linked univocally to their haplogroup (see Table 4).

Table 4. Frequency of HVS-I polymorphisms linked to the haplogroups.

The EFG clade was defined by the 15542T and 15666A mutations. Haplogroup F was distinguished in its clade by the haplogroup F-specific mutation 15595G, whereas haplogroup G was characterized by its association with the 15635T mutation. The clade OPQR was defined by the mutation 15703T, and the inside group OP was identified by the mutation 15667G.

About 80% of the A haplotypes had the mutation 15720G, 20% of which showed an association between 15720G and 15495G. Overall, the 20% of the A haplotypes had no diagnostic sites and the accuracy of their haplogroup attribution could only be confirmed when HVS-I–specific sites were available.

The B haplotypes were attributed -on the basis of an allelic association/exclusion criterion- in 90% of cases on the basis of the 15666A mutation if this mutation was not associated with the 15542T mutation typical of the EFG clade. However, 10% of the B sequences did not contain any diagnostic nucleotide site.

The haplogroup C and the JK clade, well defined by specific variants in the coding region, were associated with hypervariable mutations in HVS-I. For this reason, when these mutations were available, inference of haplogroup attribution was conducted by either a comparison or an exclusion criterion.

Following the algorithm shown in Table 4, we classified the HVS-I mtDNA of the 45 typical breeds. In total, 237 haplotypes were identified from the breed dataset and 229 were attributed to haplogroups. Eight haplotypes (3.4%) were unambiguously attributable (Table S1). The polymorphism 15602C appeared to be phylogenetically recursive in the L lineage and was also observed in all A and B lineages; analogously, the 15585A, 15597G, 15604A, and 15650G mutations were not specifically associated to the haplogroups, and for this reason we attributed a null phylogenetic weight to these markers.

All of the HVS-I sequences of the Sardinian horses were assigned to corresponding haplogroups on the basis of univocal mutations and other sufficiently informative polymorphisms, so further sequencing or genotyping of specific diagnostic sites in the coding region was not required (see the haplogroup attribution in Table 1).

Twenty-six out of 31 HVS-I d-loop sequences from the Giara breed belonged to the G haplogroup. In fact, this haplogroup was associated with the pattern 15542T, 15666A, 15650G, and 15635G in all of the 16 complete mtDNA sequences belonging to this haplogroup, and was more variable in the 15597G and 15703C variants. Overall, the G haplogroup represented 84% of the Giara maternal lineages.

The 21 Sarcidano sequences belonged mainly to haplogroups I (43%) and L (38%). Haplogroup L was defined by the HVS-I pattern 15494C, 15496G, 15534T, 15602C, 15603C, and 15649G present in 30 haplogroup L whole mtDNA genome sequences. On the other hand, the I haplogroup in the HVS-I of the 10 complete mtDNA genome sequences was defined by 15709T and 15538G.

Intrahaplogroup variation of the 3 main lineages found in the Giara and Sarcidano horses (haplogroups I, G, and L) was examined in all the available breed data by using the clustering method of the median-joining network. All of these maternal lineages produced a nascent star-like structure of the networks suggesting recent growth.

Haplogroup G (Figure 3) was poorly differentiated in eastern breeds, while a relevant number of subtypes were present in the breeds from Central Europe and the Iberian Peninsula, indicating differences in demographic growth between the eastern and western macroareas. In particular, the longest branches were often shared between samples from different breeds. Haplogroup G was the main haplogroup in the Giara horses. In this breed, the ancestral form of the haplogroup G evolved into several derivate haplotypes that are scarcely shared with other breeds. By contrast, haplogroup G was rare in the Sarcidano horses and its derivative lineages were shared with samples from the Iberian Peninsula.

The network of haplogroup I (Figure 3) showed a strong signal of population size expansion in breeds from north-west Europe and less differentiation in other macroareas. It was common in the Sarcidano maternal lineages where it was represented by derived and private lineages.

Figure 3. G, I and L intra-haplogroup variation analyzed by neighbor-joining networks.

The network of the haplotypes belonging to the haplogroup L (Figure 3) also showed a recent worldwide growth, but, unlike for haplogroups G and I, the expansion had been similar in eastern and western breeds, even if there were relatively fewer eastern L lineages than western lineages. The location of major evolution was identified as the Iberian Peninsula, but the haplotype sharing of new lineages among macroareas and breeds was very low. The Sarcidano breed displayed emergence of 3 new haplotypes in the evolution of the haplogroup, whereas this was rarely observed in the Giara breed.

Mismatch distribution of the G, I, and L haplogroups in breeds from western and eastern Eurasia and from Sardinia indicated a recent expansion that seems to have occurred earlier in haplogroup L than in the other 2 haplogroups, as indicated by a greater number of pairwise differences (Figure 4).

Figure 4. Intra-haplogroup variation analyzed by mismatch distribution in Western and Eastern Eurasian breeds and in Sardinian breeds. Thin line represents the expected mismatch distribution of a stationary population. The dotted line represents the observed mismatch distribution from segregating sites of the aligned sequences of HVS-I sequences in horse mtDNA. Mismatch distribution established for the haplogroups G, I, and L.

For the same reason, it is postulated that an earlier western Eurasian expansion took place in the haplogroups I and G. However, correlated r, *D*, and Fs statistics gave negative values not significantly different from 0, indicating that rare alleles were not more frequent than expected from a null-neutral hypothesis in an equilibrium population. Moreover, pairwise differences did not fit well into an unimodal mismatch distribution model. This could be explained by a stationary population size or by very slow growth of the population.

DISCUSSION

The mtDNA genetic structure of the lineages observed in the domestic breeds can be used to infer their demographic and domestication history (Kavar et al., 2008; Georgescu et al., 2011; Cieslak et al., 2011). To infer information about the demographic history and origin of the 2 Sardinian Giara and Sarcidano breeds from the genetic data, we compared these data with those from 43 typical breeds of the Old World.

The examined breeds displayed large molecular variation. These differences were not solely due to the total number of different haplotypes because a large amount of this molecular variation was due to the presence of different haplogroups in the same breed. The A–R haplogroups represent extant maternal lineages transmitted from a wild ancestral mare to the present-day mares: these haplogroups were defined on the complete mtDNA phylogeny as the cutoff of lineages that lived 10,000 years ago and whose haplogroups were transmitted during the process of domestication (Achilli et al., 2012). Therefore, the distinction between molecular variation that is produced before and after the formation of the present-day breeds is particularly relevant. Breeds that have more than 1 haplogroup show an ancestral variability that arose long before breed formation. In addition, as observed in the main lineage networks, the phylogeny of the single haplogroups showed a low degree of evolution: the molecular pattern of the HVS-I belonging to the same haplogroup was barely differentiated because only a limited number of new mutations had arisen in the ancestral haplotype. For this reason, breeds with a single haplogroup had lower molecular variation and were placed as outliers in the MDS graphics.

The emerging "star-like" phylogeny of the haplogroups indicated a population bottleneck followed by a small expansion in population size. In fact, the population size of breeds had remained relatively unchanged for a long time as indicated by the non significant values of the *D* and Fs statistics (Figure 4). The presence of a single (and usually frequent) ancestral haplotype shared by the majority of populations suggests founding of recent breeds from the same genetic pool.

As a consequence, all of the breeds that have maintained high haplogroup variability shared the same ancient variation and tended to cluster in the MDS graphic. Furthermore, the partition of the variability determined in the AMOVA confirmed that only a small amount of variation was due to differences between breeds. For this reason, the different geographical areas of the breeds' origins tended to overlap (Figure 2A).

Nevertheless, we observed that the variability of breeds from the western European steppes, where, according to archeological records, the domestication originated (Outram et al., 2009), was central and entirely enclosed in the overall variability. Therefore, this area also represents the point of lineage radiation, and diversification in other geographical macroareas appears to be incipient and has not yet been completed.

In addition to suggesting a recent origin of the current breeds, our data also suggest that the homogeneity of the genetic pool may have been stable until recently when warrior peoples repeatedly migrated in several waves from the Central Eurasian steppes into Europe during the

Middle Age. An east-west distribution of the variation was also apparent, probably generated from isolation by distance and weakly detected as significant by AMOVA tests when the breeds were merged into the two east and west Eurasia groups.

A genetic structuring occurred between the 2 Sardinian breeds (Figure 2) analyzed in this study, which were clearly separated and located in 2 different groups as determined by the mitochondrial lineage variation: the Sarcidano horse breed predominantly consisted of the haplogroups I and L, and other less frequent groups. This haplogroup variability, as mentioned above, is the result of molecular and haplotype diversity, which account for the presence Sarcidano among the most genetically variable breeds. The MDS analysis, which is based on genetic distances between breeds, effectively positioned the Sarcidano in the group with high variation, showing a greater affinity with the Iberian breeds as well as with those from north-west Europe, where the haplogroups L and I are highly represented. Haplogroup L is the most representative of the Iberian Peninsula, and Spanish influence in the Sardinia Island was strong until 1700. Moreover, none of the derivative Sarcidano horse L or I haplotypes is shared with the other continental breeds, suggesting that there was no recent gene flow from outside into the island.

The Giara breed consisted almost exclusively of the haplogroup G, a very common worldwide haplogroup in horses and typical of many other outlier breeds. Unlike other breeds with little haplogroup variation, the Giara sample showed a significant molecular variation. Mismatch distribution in the Giara G haplogroup were comparable to those identified in the western eurasian population group. We therefore conclude that ancestors of the Giara horse in the past may have occupied an area that was larger than the one where they are found now, albeit always within the boundaries of the island. This interpretation is supported by historical records reporting the widespread presence in Sardinia of a horse described as "wild" that was phenotypically similar to the Giara horse (Cetti, 1774).

In conclusion, in this study we have first described the distribution of the current genetic diversity in typical breeds of horses and compared genetic differences among the various breeds and among groups of breeds from different geographical areas. We also inferred that the genetic diversity in the Sardinian Giara and Sarcidano breeds is the result of recent evolution. We further demonstrated that the genetic system used is powerful of discerning past and present evolutionary patterns.

As to the question of horse domestication, we agree with other authors that horses are a notable exception to the theory that holds that domestication is the result of a very small number of independent, often geographically separated taming events, as observed for all domesticated species

(Brudford et al., 2003). The abundance of very differentiated mitochondrial DNA lineages indicates that horse domestication probably involved a large number of wild captures. Our analysis support the hypothesis that the area of horse domestication was originally restricted and then gave rise to the high number of present-day horse mitochondrial lineages (Jansen et al., 2002; Forster et al., 2012); however, the subsequent recruitment of local mares from wild horse populations into domesticated herds is less apparent. According to our results, multiple radiation events of lineages from the original place of wild capture better explain the genetic structure of today's domestic horses.

ACKNOWLEDGMENTS

We thank Barbara Wilkens, Marco Apollonio, and Ercole Contu for useful discussions and helpful comments and breed owners for providing sample collections. This work was supported by MURST ex60% and the "Fondazione del Banco di Sardegna" to LM and PF. AU was supported by the Autonomous Region of Sardinia and by the Programma Operativo Fondo Sociale Europeo 2007-2013.

REFERENCES

Achilli A, Olivieri A, Soares P, Lancioni H, et al. (2012). Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl. Acad. Sci. USA*. 109:2449-54.

Bandelt HJ, Forster P, and Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol.Biol. Evol*. 16: 37–48

Bruford MW, Bradley DG, and Gordon Luikart G (2003). DNA markers reveal the complexity of livestock domestication. *Nature* 4: 900-910

Cetti F (1774). I quadrupedi di Sardegna. In: Storia naturale di Sardegna (2000). Ilisso Biblioteca Sarda, Nuoro.

Cieslak M, Pruvost M, Benecke N, Hofreiter M, et al. (2010). Origin and history of mitochondrial DNA lineages in domestic horses. *PLoS One* 5:e15311

Cozzi MC, Strillacci MG, Valiati P, Bighignoli B, et al. (2004). Mitochondrial d-loop sequence variation among Italian horse breeds. *Genet. Sel. Evol.* 36:663-672

Excoffier L, Laval G, and Schneider S (2007). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1:47-50

Excoffier L, Smouse PE, and Quattro JM (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479-491

Forster P, Hurles ME, Jansen T, Levine M, et al. (2012). Origins of the domestic horse. *Proc. Natl. Acad. Sci. USA*. 109:E3148.

Georgescu SE, Manea MA, Dudu A and Costache M (2011). Phylogenetic relationships of the Hucul horse from Romania inferred from mitochondrial D-loop variation. *Genet. Mol. Res.* 10: 4104 - 4113

Goto H, Ryder OA, Fisher AR, Schultz B, et al. (2011) A massively parallel sequencing approach uncovers ancient origins and high genetic variability of endangered Przewalski's horses. *Genome Biol. Evol.* 3:1096-106.

Gratani L (1980). Cavallo della Giara. Istituto di Incremento Ippico della Sardegna. RAS, Cagliari.
Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98
Ishida N, Hasegawa T, Takeda K, Sakagami M, et al. (1994) Polymorphic sequence in the d-loop region of equine mitochondrial DNA. *Anim. Genet.* 25:215-221
Jansen T, Forster P, Levine MA, Oelke H, et al. (2002) Mitochondrial DNA and the origins of the domestic horse. *Proc. Natl. Acad. Sci. USA* 99:10905-10910
Jiang Q, Wei Y, Huang Y, Jiang H, et al. (2011). The complete mitochondrial genome and phylogenetic analysis of the Debao pony (*Equus caballus*). *Mol. Biol. Rep.* 38:593-599
Kavar T and Dovc P (2008). Domestication of the horse: Genetic relationships between domestic and wild horses. *Liv. Sci.* 116:1-14
Lindgren G, Backstrom N, Swinburne J, Hellborg L, et al. (2004). Limited number of patrilines in horse domestication. *Nat. Genet.* 36:335-336
Ling Y, Ma Y, Guan W, Cheng Y, et al. (2010) Identification of Y chromosome genetic variations in Chinese indigenous horse breeds. *J. Hered.* 101:639-43
Lippold S, Matzke NJ, Reissmann M, and Hofreiter M. (2011)Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. *BMC Evol. Biol.* 11:328.
Nei M (1987). Molecular Evolutionary Genetics. Columbia University Press, New York, NY, USA
Outram AK, Stear NA, Bendrey R, Olsen S, et al. (2009). The earliest horse harnessing and milking. *Science* 323:1332-1335
Tajima F (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-460
Tajima F (1993). Measurement of DNA polymorphism. In: Takahata N, and Clark AG. Mechanisms of Molecular Evolution. Introduction to Molecular Paleopopulation Biology. Sunderland, MA: Japan Scientific Societies Press, Sinauer Associates, Inc., 37-59, Tokyo.
Weir BS (1996) The second National Research Council report on forensic DNA evidence. *Am. J. Hum. Genet.* 59:497-500
Xu S, Luosang J, Hua S, He J, et al. (2007). High altitude adaptation and phylogenetic analysis of Tibetan horse based on the mitochondrial genome. *J. Genet. Genomics*. 34:720-9
Xu X and Arnason U (1994). The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene* 148:357-362.

Tables legends

Table 1. Mitochondrial control region sequence variation in Sardinian horses; 247-bp segment of the control region from 31 Giara and 21 Sarcidano horses aligned to the mtDNA reference X79547 (Xu and Arnason, 1994). Sequence positions based on the same reference sequence are given above each column. Variable nucleotides are indicated. All samples have been assigned to one of the main haplogroups and grouped by the occurrence of haplogroup informative polymorphism. Description of the samples names: JAR, Giara horses from the present work; GRH, Giara horses by Cozzi et al., 2004; GIA, Giara horses by Achilli et al., 2012; SAR, Sarcidano horses from the present work; SRH, Sarcidano horses by Cozzi et al., 2004.

Table 2. The 45 Old World native breeds used for comparison.

Table 3. Intra-population level variation of 45 Old World native breeds. MNPD, mean number of pairwise differences.

Table 4. Relative frequency of HVS-I polymorphisms linked to the haplogroups defined on the whole mitochondrial genome. Mutated positions are ordered on the basis of their specificity in defining haplogroups.

Supplemental Table 1. Database of the HVS-I sequences and haplogroup attribution.

Figure legends

Figure 1. Giara and Sarcidano breeds areal map.

Figure 2. Multidimensional scaling (MDS) plot computed from the matrix of the pairwise differences of the mitochondrial HVS-I sequences. Each symbol represents one of the 45 populations tested. Breed codes are as in Table 2. A) in the graph only those populations outside the main cluster (enclosed within the square with the broken line and including breeds with high amount of intra population variation) are named. B) in the graph the square including the main cluster is enlarged and breeds are named. D-star: Raw stress = 30.54; Alienation = 0.12; D-hat: Raw stress = 24.05; Stress = 0.11.

Figure 3. G, I and L intra-haplogroup variation analyzed by neighbor-joining networks.

Figure 4. Intra-haplogroup variation analyzed by mismatch distribution in Western and Eastern Eurasian breeds and in Sardinian breeds. Thin line represents the expected mismatch distribution of a stationary population. The dotted line represents the observed mismatch distribution from segregating sites of the aligned sequences of HVS-I sequences in horse mtDNA. Mismatch distribution established for the haplogroups G, I, and L.

Table 1. Polymorphic sites of mtDNA HVS-I in Sardinian breeds.

Samples	Giara	Sarcidano	Haplogroup	15494	15495	15496	15505	15507	15526	15534	15538	15540	15542	15569	15575	15576	15585	15589	15597	15599	15601	15602	15603	15604	15617	15618	15619	15623	15627	15633	15634	15635	15639	15644	15649	15650	15659	15665	15666	15667	15672	15678	15680	15692	15694	15695A	15703	15704	15708	15709	15712	15716	15720	15732	15740
reference X79547	#	#		T	T	A	T	C	T	C	A	A	C	T	G	C	G	T	A	T	T	C	T	G	T	A	A	T	C	A	T	C	T	T	A	A	T	T	G	A	G	C	C	A	C	-	T	C	A	C	A	C	G	C	A
JAR38	1	0	B	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.
JAR69	1	0	B	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	A	.	.	G	.	.	A	.	.	.	.	T	.	-	.	.	.	.	.	.	A	.	.
JAR90	1	0	E	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.
JAR40, JAR64, JAR70, JAR72, JAR87, JAR93, JAR95, GRH1, GRH3, GRH5	10	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	-	C	.	.	.	.	.	A	.	.
JAR55	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	.	-	C	.	.	A	.	.	A	.	.
JAR58	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	.	-	C	A	G	.	.	.	A	.	.
JAR59	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	.	-	C	.	.	G	.	.	A	.	.
JAR67	1	0	G	.	C	.	G	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	.	-	C	.	.	.	G	.	A	G	.
JAR71	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	.	-	C	.	.	.	.	.	A	.	C
JAR77	1	0	G	.	C	.	.	T	.	.	.	.	T	G	.	.	A	.	G	.	.	T	.	.	.	.	.	G	.	.	.	T	A	.	G	.	.	A	T	.	.	A	.	.	-	C	.	.	.	.	.	A	.	T
JAR92	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	C	.	.	.	.	.	.	-	C	.	.	.	.	.	A	.	.
JAR81	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	A	G	.	.	A	.	.	.	.	.	.	.	-	C	.	.	.	.	.	A	.	.
GIA02	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	.	-	C	.	.	C	.	.	A	.	.
JAR34	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	A	.	.	.	.	.	.	.	-	C	.	.	.	.	.	A	.	.
JAR39, JAR41, JAR45,GRH2, GRH4	5	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	A	C	.	.	.	.	.	A	.	.
GIA01	1	0	G	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	A	.	.	.	.	.	.	.	-	C	.	.	C	.	.	A	.	.
SAR22	0	1	G	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	G	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	A	.	.	.	.	.	.	.	-	C	.	.	.	.	.	A	.	.
SAR10	0	1	G	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	A	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.
SAR9	0	1	I	.	C	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	-	.	.	.	T	.	.	A	.	.
SAR1, SAR6, SAR26, SAR27, SRH3	0	5	I	.	C	.	.	.	.	.	G	.	.	.	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	-	.	.	.	T	.	.	A	.	.
SRH1	0	1	I	.	C	.	.	.	.	.	G	.	.	.	.	.	A	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	-	.	.	.	T	.	.	A	.	.
SAR24	0	1	I	.	C	.	.	.	.	.	G	.	.	.	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	-	.	.	.	G	.	A	A	.	.
SAR28	0	1	I	.	C	.	G	.	.	.	G	.	.	.	.	.	A	C	.	G	.	T	G	.	.	.	.	.	.	.	.	.	.	A	.	G	.	.	.	.	.	.	.	.	T	-	.	.	.	.	.	.	A	.	.
JAR35, JAR57, SAR13, SAR15, SAR20, SRH4, SRH5	2	5	L	C	C	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.
SAR18	0	1	L	C	C	G	.	.	.	T	.	.	.	.	.	.	.	-	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.
SAR12	0	1	L	C	C	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	T	.	-	.	.	.	.	.	.	A	.	.
SRH2	0	1	L	C	C	G	.	.	.	T	.	G	.	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.
SAR17	0	1	M	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	T	.	.	C	C	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.
SAR4	0	1	M	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.

Table 2. The Old World native breeds used for comparison.

Macroarea	Breed	Code	Sources: references and GenBank accessions
North- west Africa	Barb	BAR	Jansen et al., 2002; EF686021-45
Iberian Peninsula	Andalusian	AND	Mirol et al., 2002; Jansen et al., 2002; Royo et al., 2005; Luis et al., 2006b
	Asturcòn	AST	Mirol et al., 2002; Royo et al., 2005; HQ827083-90
	Caballo de Corro	CCO	Royo et al., 2005; HQ827099-103
	Garrano	GAR	Royo et al., 2005; Luis et al., 2006b; AY246231-4
	Jaca Navarra	JAN	HQ827104-HQ827118
	Losino	LOS	Mirol et al., 2002; Royo et al., 2005; HQ827119-29
	Lusitano	LUS	Jansen et al., 2002; Luis et al., 2006b; AY246242-7
	Marismeno	MAR	Royo et al., 2005; HQ827136-45
	Pottoka	POT	Mirol et al., 2002; Royo et al., 2005; HQ827156-61
	Sorraia	SOR	Jansen et al., 2002; Luis et al., 2006a; Luis et al., 2006b; HQ827162-3; AY246259-65
Central Europe	Exmoor	EXM	Jansen et al., 2002; AY246219-24
	Fell	FEL	Bower et al., 2010
	Giara	GIA	Present work; Cozzi et al., 2004; Achilli et al., 2012
	Irish Draught	IRD	McGahern et al., 2006a
	Kerry Bog Pony	KEB	McGahern et al., 2006a
	Percheron	PER	Kakoi et al., 2007
	Rhineland Heavy Draft	RHD	Jansen et al., 2002
	Sanfratellano	SAN	Zuccaro et al., 2009; Guastella et al., 2011
	Sarcidano	SAR	Present work; Cozzi et al., 2004
	Scottish Highland	SCH	Jansen et al., 2002; Bower et al., 2010
	Senner	SEN	Jansen et al., 2002
	Shetland	SHE	Hill et al., 2002; Jansen et al., 2002; Bower et al., 2010; AY246253-8
	Sicilian Indigenous	SII	Zuccaro et al.,2009; Guastella et al., 2011
	Sicilian Oriental Purebred	SOP	Zuccaro et al.,2009; Guastella et al., 2011
Arabian Peninsula	Arabian	ARA	Bowling et al., 2000; Mirol et al., 2002; Jansen et al., 2002; AY246180-5
Central Eurasia	Akhal-Teke	AKT	McGahern et al., 2006b; AY246174-9
	Anatolian	ANA	Hill et al., 2002
	Kazahk	KAZ	Lei et al., 2009
	Mesenskay	MES	McGahern et al., 2006b
	Vyayskaya	VYA	McGahern et al., 2006b
	Yakut	YAK	McGahern et al., 2006b
Far East	Baise	BAI	GQ203128-GQ203143; GQ222059-60
	Cheju	CHE	Yang et al., 2002; AY246201-8
	Debao	DEB	EU826536; FJ392562-80; GQ203125-7
	Guanzhong	GUA	Lei et al., 2009
	Guizhou	GUI	Lei et al., 2009
	Mongolian	MON	Jansen et al., 2002; McGahern et al., 2006b; Kakoi et al., 2007
	NingQiang	NIN	Lei et al., 2009
	Tibetan	TIB	DQ986464-79
	Hokkaido	HOK	Kakoi et al., 2007
	Misaki	MIS	Kakoi et al., 2007
	Taishu	TAI	Kakoi et al., 2007
	Tokara	TOK	Kakoi et al., 2007
	Yonaguni	YON	Kakoi et al., 2007

Table 3. Intra-population level variation of 45 Old World native breeds.

Breed	Individuals	Haplotypes	Polymorphic sites	Sum of square freqs	Haplotype diversity		Pairwise differences		Nucleotide diversity	
					h	SD	MNPD	SD	π	SD
BAR	40	14	23	0.165	0.856	0.040	4.659	2.332	0.019	0.010
AND	30	15	22	0.104	0.926	0.026	5.400	2.677	0.022	0.012
AST	21	9	21	0.134	0.910	0.035	7.381	3.596	0.030	0.016
CCO	19	4	10	0.357	0.678	0.088	3.719	1.965	0.015	0.009
GAR	18	14	20	0.080	0.974	0.025	6.327	3.147	0.026	0.014
JAN	15	14	22	0.076	0.991	0.028	6.381	3.203	0.026	0.015
LOS	23	15	24	0.081	0.961	0.022	6.719	3.288	0.027	0.015
LUS	21	10	15	0.147	0.895	0.039	5.943	2.953	0.024	0.013
MAR	22	9	15	0.169	0.870	0.044	5.931	2.942	0.024	0.013
POT	21	18	26	0.066	0.981	0.023	6.600	3.247	0.027	0.015
SOR	30	5	12	0.422	0.598	0.059	2.232	1.265	0.009	0.006
EXM	18	6	20	0.247	0.797	0.066	5.209	2.644	0.021	0.012
FEL	17	8	19	0.177	0.875	0.053	5.500	2.783	0.022	0.013
GIA	31	15	31	0.180	0.847	0.053	3.933	2.025	0.016	0.009
IRD	59	28	31	0.070	0.946	0.017	6.373	3.063	0.026	0.014
KEB	39	17	26	0.090	0.934	0.020	5.614	2.753	0.023	0.012
PER	15	3	10	0.662	0.362	0.145	2.800	1.566	0.011	0.007
RHD	24	15	24	0.118	0.920	0.040	6.754	3.298	0.027	0.015
SAN	20	11	20	0.130	0.916	0.041	6.295	3.117	0.025	0.014
SAR	21	13	26	0.138	0.905	0.047	6.738	3.309	0.027	0.015
SCH	31	16	26	0.086	0.944	0.021	6.185	3.021	0.025	0.014
SEN	19	2	1	0.900	0.105	0.092	0.105	0.183	0.000	0.001
SHE	66	15	26	0.134	0.880	0.023	6.441	3.088	0.026	0.014
SII	20	13	30	0.120	0.926	0.043	7.132	3.492	0.029	0.016
SOP	20	1	0	1.000	0.000	0.000	0.000	0.000	0.000	0.000
ARA	70	37	37	0.042	0.972	0.007	6.231	2.995	0.025	0.013
AKT	24	14	27	0.108	0.931	0.033	6.859	3.345	0.028	0.015
ANA	15	11	17	0.102	0.962	0.034	5.333	2.727	0.022	0.012
KAZ	18	16	31	0.068	0.987	0.023	6.131	3.059	0.025	0.014
MES	18	11	26	0.136	0.915	0.050	5.549	2.797	0.022	0.013
VYA	18	10	17	0.124	0.928	0.037	5.216	2.646	0.021	0.012
YAK	20	12	20	0.105	0.942	0.030	6.021	2.995	0.024	0.014
BAI	18	16	37	0.068	0.987	0.023	7.974	3.888	0.032	0.018
CHE	73	15	24	0.114	0.899	0.017	4.974	2.447	0.020	0.011
DEB	23	15	31	0.093	0.949	0.028	6.727	3.292	0.027	0.015
GUA	27	10	20	0.180	0.852	0.039	5.556	2.754	0.022	0.012
GUI	62	27	33	0.082	0.933	0.019	5.799	2.812	0.023	0.013
MON	35	17	27	0.084	0.943	0.019	6.424	3.116	0.026	0.014
NIN	27	16	26	0.092	0.943	0.027	5.880	2.898	0.024	0.013
TIB	16	14	22	0.078	0.983	0.028	6.742	3.355	0.027	0.015
HOK	28	3	13	0.865	0.140	0.087	1.325	0.850	0.005	0.004
MIS	26	3	10	0.731	0.280	0.107	1.563	0.963	0.006	0.004
TAI	16	3	14	0.594	0.433	0.138	3.600	1.927	0.015	0.009
TOK	19	1	0	1.000	0.000	0.000	0.000	0.000	0.000	0.000
YON	19	2	1	0.501	0.526	0.040	0.526	0.460	0.002	0.002

Table 4. Frequency of HVS-I polymorphisms linked to the haplogroups.

Haplogroup	A	B	C	D	E	F	G	H	I	JK	J	K	L	M	N	O	P	Q	R	# associated haplogroups
# sequences	13	12	9	5	1	5	16	2	10	1	2	1	30	10	6	2	8	12	2	
Variant																				
15720G	0.769	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15737C	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15521A	0.000	0.000	0.000	0.800	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15595G	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15540G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15526C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15709T	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15538G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15495T	0.231	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15534T	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15496G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15494C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1
15685G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.000	0.000	0.000	0.000	0.000	0.000	1
15601C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	1
15617C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	1
15740G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1
15726A	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.417	0.000	1
15616G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1
15598C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1
15574A	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.000	0.000	0.000	0.000	0.000	0.000	1
15533G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.000	0.000	0.000	0.000	0.000	0.000	1
15718T	0.000	0.000	0.000	0.200	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2
15596G	0.000	0.000	0.000	0.400	0.000	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2
15603C	0.077	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	2
15659C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.900	0.000	0.000	0.000	0.000	1.000	2
15667G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000	0.000	2
15615G	0.000	0.000	0.000	0.000	0.000	0.000	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	2
15544C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.167	0.000	0.000	0.000	0.000	2
15651A	0.000	0.000	0.111	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2
15602C	0.923	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.433	0.000	0.000	0.000	0.000	0.000	0.000	3
15542T	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3
15635T	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	3
15649G	0.077	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	3
15666A	0.000	0.917	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	4
15703C	0.000	0.000	0.000	0.000	0.000	0.000	0.938	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	5
15604A	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.300	0.000	0.000	0.500	0.625	0.583	0.000	5
15597G	0.000	0.000	0.333	0.000	0.000	0.200	0.875	0.000	0.200	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.625	0.083	0.000	7
15650G	0.000	1.000	1.000	0.000	1.000	1.000	1.000	0.000	0.800	0.000	0.500	1.000	0.000	0.000	0.000	0.000	0.000	0.083	0.000	9
15585A	0.077	0.583	0.222	0.200	1.000	0.000	0.625	1.000	0.900	1.000	1.000	0.000	0.667	0.200	0.500	0.000	0.000	0.250	0.000	14

Supplemental Table 1.

GenBank Accession	Breed *	HG#	Freq.	Diagnostic sites: 15602C	15666A	15650G	15720G	15495T	15737C	15521A	15596G	15542T	15595G	15635T	15597G	15540G	15526C	15718T	15709T	15538G	15534T	15603C	15649G	15496G	15494C	15568C	15601C	15617C	15659C	15703C	15604A	15667G	15740G	15726A	15615G	15616G	15598C	15544C	15574A	15651A	15533G	15585A
EU203753.1	GUI	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AF132589.1	ARA	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AF465997.1	ARA	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203846.1	NIN	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203855.1	NIN	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ327973.1	MES	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328023.1	VYA	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328033.1	VYA	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328030.1	VYA	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328025.1	VYA	A	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EF686042.1	BAR	A	1	C	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EF686035.1	BAR	A	1	C	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
FJ392578.1	DEB	A	1	C	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AF132587.1	ARA	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AF466003.1	ARA	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
GQ203138.1	BAI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
GQ203132.1	BAI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
GQ203140.1	BAI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AF354429.1	CHE	A	4	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
GQ203127.1	DEB	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
FJ392563.1	DEB	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
GQ203125.1	DEB	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
FJ392575.1	DEB	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203762.1	GUI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203763.1	GUI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203765.1	GUI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203766.1	GUI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203772.1	GUI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU203775.1	GUI	A	1	C	G	A	A	G	C	T	G	A	C	A	C	A	A	T	C	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G

HQ827124.1	LOS	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AJ413741.2	LUS	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AJ413743.2	LUS	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AJ413744.2	LUS	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AJ413745.2	LUS	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AY519938.1	MAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AY519934.1	MAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AY519941.1	MAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
HQ827136.1	MAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
HQ827139.1	MAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
HQ827141.1	MAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ327993.1	MON	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ327992.1	MON	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU750720.1	SAN	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU831237.1	SII	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU831238.1	SAN	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AJ413847.2	RHD	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328024.1	VYA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328032.1	VYA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328035.1	VYA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
DQ328029.1	VYA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AF481237.1	ANA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AF132591.1	ARA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AJ413648.2	ARA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AY246185.1	ARA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AY246183.1	ARA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AY246182.1	ARA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AY246181.1	ARA	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
HQ827086.1	AST	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AY519878.1	AST	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AJ413649.2	BAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
GU563631.1	FEL	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
GU563638.1	FEL	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
GU563642.1	FEL	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
GU563643.1	FEL	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
GU563644.1	FEL	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A
AY519920.1	GAR	B	1	C	A	G	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	A

EU203831.1	KAZ	NC	1	T	G	**G**	A	C	T	**A**	A	C	A	C	**G**	A	T	C	C	A	C	T	A	A	T	A	T	T	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AY246222.1	EXM	NC	1	T	G	**G**	**G**	C	T	G	A	C	A	C	**G**	A	T	C	C	A	C	T	A	A	T	A	T	T	T	**C**	G	A	A	G	A	A	T	T	G	G	A	G
DQ327981.1	MES	NC	1	T	G	A	A	C	T	G	A	C	A	C	**G**	A	T	C	C	A	C	T	A	A	T	A	T	T	T	**C**	G	A	A	G	A	A	T	T	G	G	A	G
AF132575.1	ARA	NC	1	T	G	A	A	C	T	G	A	C	A	C	**G**	A	T	C	C	A	C	T	A	A	T	A	T	T	T	**C**	G	A	A	G	A	A	T	T	G	G	A	G
AY136785.1	GUA	NC	1	T	G	A	A	C	T	G	A	C	A	C	**G**	A	T	C	C	A	C	T	A	A	T	A	T	T	T	**C**	G	A	A	G	A	A	T	T	G	G	A	G
EU203739.1	GUA	NC	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	**C**	G	A	A	G	A	A	T	T	G	G	A	G
EU203740.1	GUA	NC	1	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	**C**	G	A	A	G	A	A	T	T	G	G	A	G
AY246208.1	CHE	NC	1	T	G	A	A	C	**C**	**A**	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	**C**	T	T	G	A	A	G	A	A	T	T	G	G	A	G
AF354425.1	CHE	NC	4	T	G	A	A	C	**C**	**A**	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	**C**	T	T	G	A	A	G	A	A	T	T	G	G	A	G
EU750723.1	SAN	NC	2	T	G	A	A	C	T	G	A	C	A	C	A	A	T	C	C	A	C	T	A	A	T	A	T	T	T	**C**	G	A	A	G	A	A	T	T	G	G	A	G

*Breed's codes as in Table2 of the main text
#HG: haplogroup; NC: Not Classified

Figure 1.


Sardinia
Sarcidano breed
Giara breed
Italy
Mediterranean Sea

Figure 2.

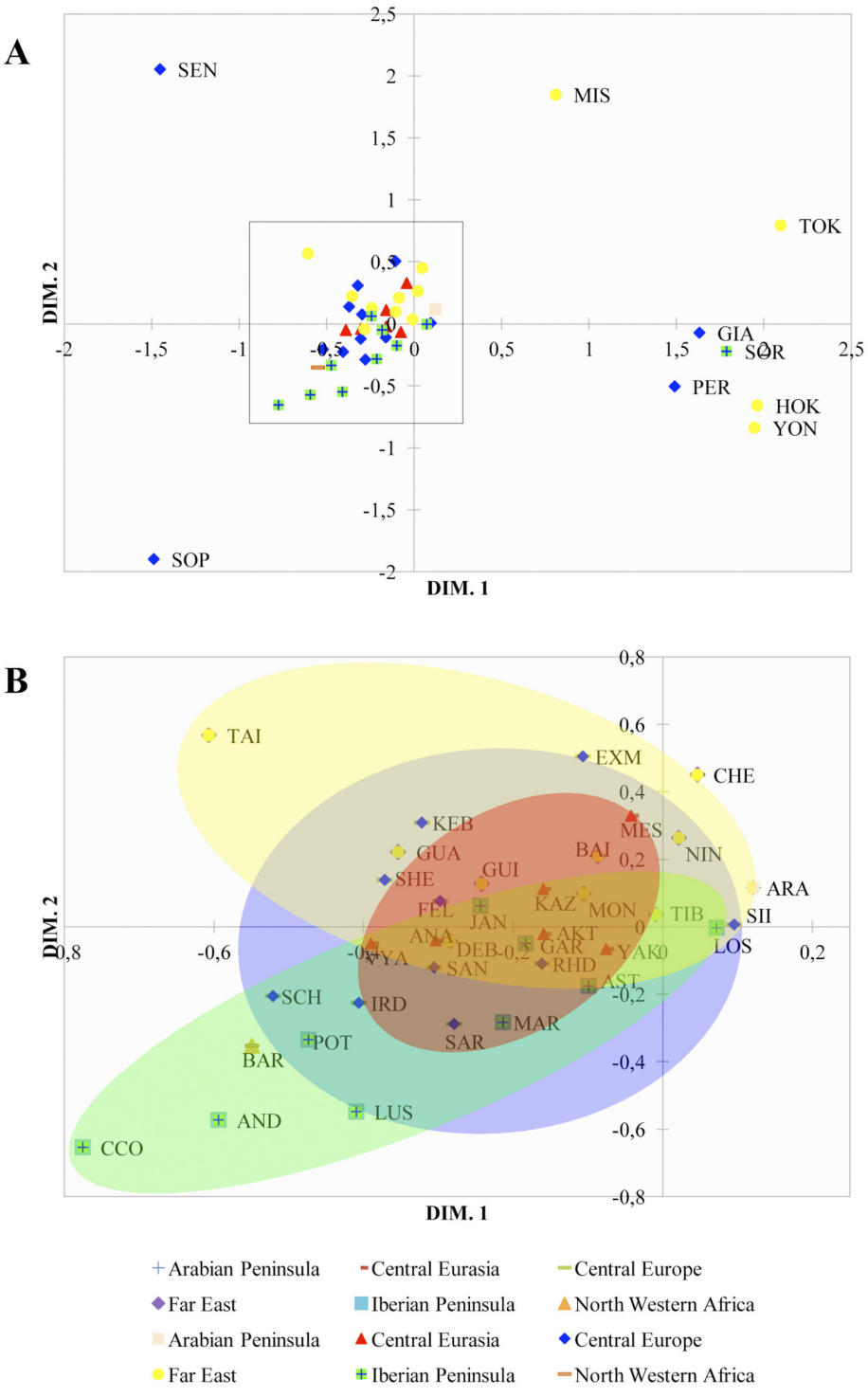

A
SEN
MIS
TOK
GIA
SOR
PER
HOK
YON
SOP
DIM. 2
DIM. 1
B
TAI
EXM
CHE
KEB
MES
GUA
BAI
NIN
GUI
SHE
ARA
KAZ
MON
TIB
SII
FEL
JAN
AKT
ANA
GAR
YAK
LOS
DEB
VYA
SAN
RHD
AST
SCH
IRD
MAR
POT
SAR
BAR
AND
LUS
CCO
DIM. 2
DIM. 1
Arabian Peninsula
Central Eurasia
Central Europe
Far East
Iberian Peninsula
North Western Africa
Arabian Peninsula
Central Eurasia
Central Europe
Far East
Iberian Peninsula
North Western Africa

Figure 3.

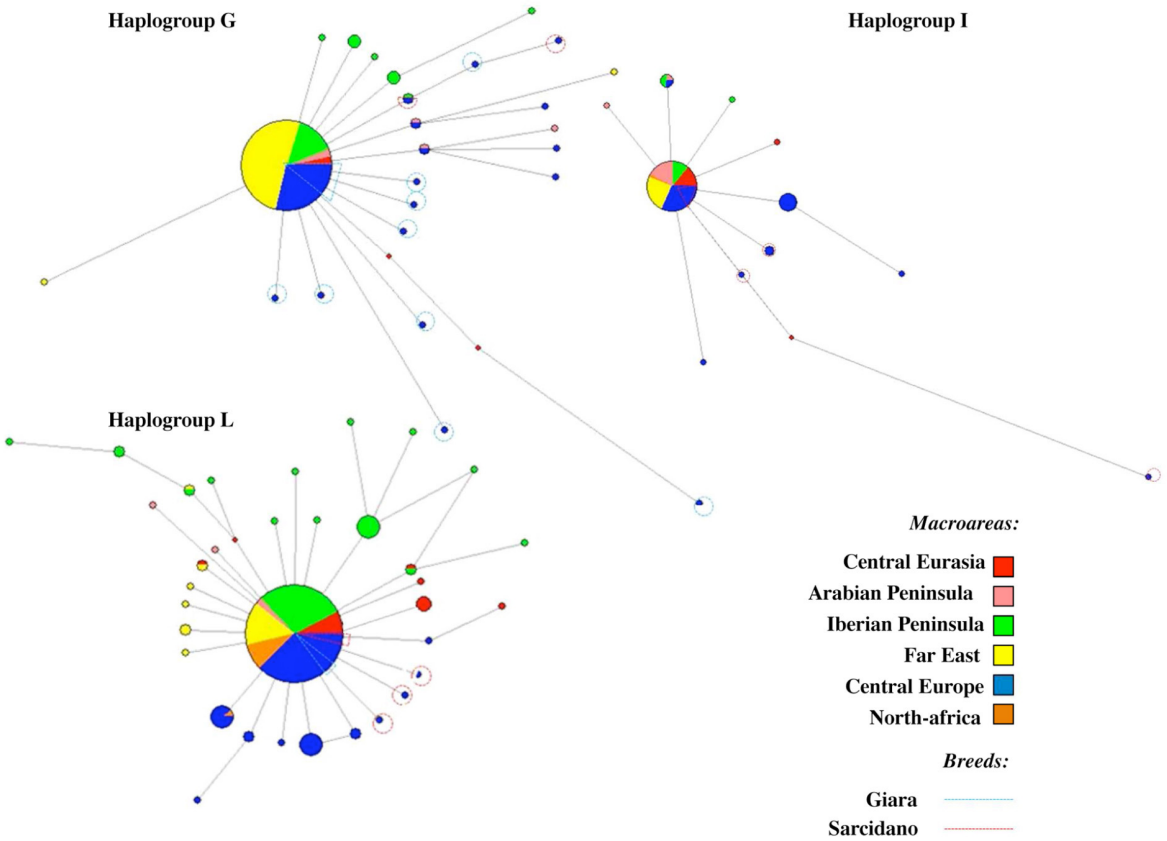
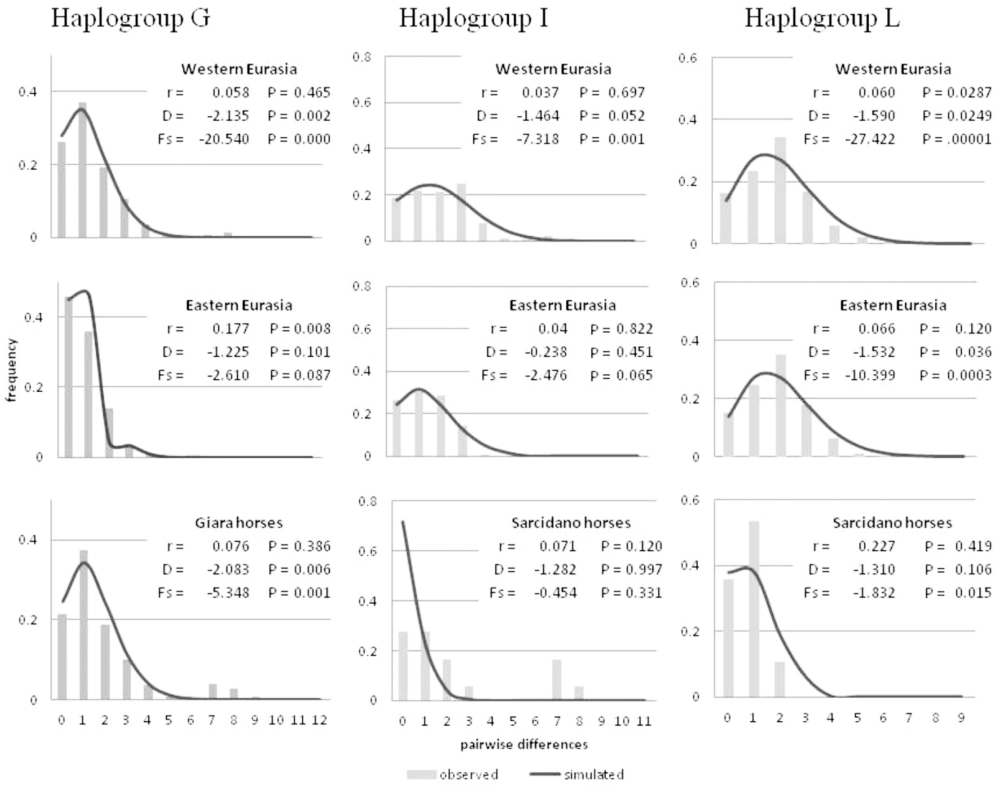

Haplogroup G
Haplogroup I
Haplogroup L
Macroareas:
Central Eurasia
Arabian Peninsula
Iberian Peninsula
Far East
Central Europe
North-africa
Breeds:
Giara
Sarcidano

Figure 4.


Haplogroup G
Haplogroup I
Haplogroup L
Western Eurasia
r = 0.058 P = 0.465
D = -2.135 P = 0.002
Fs = -20.540 P = 0.000
Western Eurasia
r = 0.037 P = 0.697
D = -1.464 P = 0.052
Fs = -7.318 P = 0.001
Western Eurasia
r = 0.060 P = 0.0287
D = -1.590 P = 0.0249
Fs = -27.422 P = .00001
Eastern Eurasia
r = 0.177 P = 0.008
D = -1.225 P = 0.101
Fs = -2.610 P = 0.087
Eastern Eurasia
r = 0.04 P = 0.822
D = -0.238 P = 0.451
Fs = -2.476 P = 0.065
Eastern Eurasia
r = 0.066 P = 0.120
D = -1.532 P = 0.036
Fs = -10.399 P = 0.0003
Giara horses
r = 0.076 P = 0.386
D = -2.083 P = 0.006
Fs = -5.348 P = 0.001
Sarcidano horses
r = 0.071 P = 0.120
D = -1.282 P = 0.997
Fs = -0.454 P = 0.331
Sarcidano horses
r = 0.227 P = 0.419
D = -1.310 P = 0.106
Fs = -1.832 P = 0.015
frequency
pairwise differences
observed
simulated

CHAPTER 2
HapSign: an informatic tool for mitochondrial haplotype assignment.

-Manuscript in preparation-

Authors:
Mario Barbato[1], Daria Sanna[2], Paolo Francalacci[2], Antonella Useli[2]
Affiliations:
[1] Cardiff School of Biosciences, Cardiff University, Cardiff, UK
[2] Dipartimento di Scienze della Natura e del Territorio - Unità di Zoologia, Archeozoologia e Genetica -, Università di Sassari, Sassari, Italy

Keywords: Haplotype assignment, haplogroup, mitochondrial DNA phylogeny, domestic animals, horse breeds.

Abstract.
The investigation based on mitochondrial DNA (mtDNA) and non-recombining region of Y chromosome (NRY) variation of domestic mammals has been subjected to a rapid development. Between the two systems, mtDNA is by far more studied. Currently, thanks to the technical advances and the decreasing of costs, the number of organisms whose complete mtDNA sequence has become available is rapidly increasing. Nearby those extensive studies there is a significant accumulation of mtDNA fragment sequences, especially of the first hypervariable segment (HVS-I). Both approaches are contributing to the increasing of the amount of available data. In addition, the inclination to analyze numerous samples and the diffusions of this kind of analysis for wide range of applications needs a greater expertise than previous request. A new Software tool for automatic assignment of mitochondrial haplotypes was developed. The *HapSign* Software is especially dedicated to domestic animals and has been designed to allow an accurate, fast and easy mtDNA haplotypes assignation.

Introduction.

In the last years, the molecular phylogenetic methods are being applied in several branches of biological sciences. The important findings derived from these kinds of studies are showing that by a multidisciplinary perspective, which include the genetic information, it is possible to significantly improve the knowledge about the evolutionary history of species. In this context, despite the increasing amount of data derived from large sets of whole-genome markers (Elhaik et al. 2013; Prado-Martinez et al. 2013; Petersen et al. 2013), the analysis of mitochondrial DNA (mtDNA) and the non-recombining region of Y chromosome (NRY) markers is currently growing in many fields. Forensic genetics (Gurney et al. 2010; Caniglia et al. 2013), health sciences (Achilli et al. 2011;

Dowling 2013), molecular ecology (Warmuth et al. 2013) and conservation genetics (Alvarez et al. 2012; Mondol et al. 2013; Bagatharia et al. 2013) are giving some examples. In fact, lacking recombination because of their unilinear inheritance, they are considered particularly useful in phylogeny reconstruction and phylogeographic analysis of variation (Soares et al. 2010).

Anyway, between the two systems, mtDNA is by far more studied. Abundant and easy to genotyping, the mtDNA has a small size and a higher mutation rate, in respect to the average of nuclear DNA, particularly in the first hypervariable segment (HVS-I) of the displacement loop (d-loop). Probably related to the high number of polymorphisms that could be found in this short fragment, several studies are based on HVS-I variation. Nevertheless, hyper-variability has a double face that resides in a higher rate of reversion and recurrence, which may appear in lineages of haplotypes not phylogenetically related. For this reason the phylogenetic trees inferred only on the basis of HVS-I polymorphisms have to be intended as provisional, while the information contained in the coding region, more stable, would clarify the relationships between the haplotypes. In order to avoid this problem, the overall tendency has been to combine the d-loop polymorphisms with stable genetic markers of the coding region. In fact, the definition of groups of haplotypes with a common ancestor (haplogroups), by an adequate number of shared stable mutations, and the application of the right phylogenetic mutation rate, could allow making inference about phylogenetic events over time. The analysis of geographic distribution of variation (phylogeography) could be used as a signature of origins and evolutionary dynamics that finally explain the extant genetic structure of populations.

Currently, thanks to the technical advances of sequencing methods as Next-Generation Sequencing (NGS), accompanied by the decreasing of costs, the number of organisms whose complete mtDNA sequence has become available is rapidly increasing. The re-sequencing methodology of mtDNA is giving important results in human evolutionary studies (Torroni et al. 2006; Behar et al. 2006; Pala et al. 2012) and more recently even in animal genetics (Bonfiglio et al. 2011; Achilli et al. 2012; Gazave et al. 2013). It is worthy to note that the investigation based on unilinear systems variation of domestic animals has been subjected to a rapid development.

The sequence of complete mitochondrial genomes of an increasing number of species and samples are thus available in public databases as the National Center for Biotechnology Information (NCBI). The main reasons of those research efforts are probably related to the growing interest in domestic species. Those studies could actually give an important contribute to clarify the history and past demographic dynamics of human populations. In fact, with the changes of the economic

system occurred in the Neolithic (about 10.000 years ago), the worldwide development of human societies was strongly related to domestication dynamics of livestock species (see also Bruford et al. 2003).

As reported by Lenstra and colleagues (2012) in a recent review, the genetic studies of livestock populations focus on questions of domestication, within- and among-breed diversity, breed history and adaptive variation. The study of Achilli and colleagues (2009) for example, confirmed the origins of the main haplogroups of Cattle in the Near East and discovered novel haplogroups. The subsequent study of the same research group (Bonfiglio et al. 2011), support the hypothesis of the existence of an independent center of domestication in the Italian peninsula. In a recent study about goat domestication, Nomura and coll. (2013) indicate that the process of domestication is more complex than may be presently appreciated.

Presently, the importance of domestic animals it was also highlight claiming they represent a relevant source of biodiversity to preserve for the future. Therefore, it was emphasized the importance of the knowledge of the genetic structure of native breeds, particularly if reared in traditional way, in order to identify the correct conservation strategies, in the attempt to solve the pressing biodiversity loss related to global change (Medugorac et al. 2009; Groeneveld et al. 2010; Lenstra et al. 2012).

The pointed out interest about domestic species genetic variation yielded to different kind of product data, depending on specific questions and interest. Nearby the extensive studies mentioned above, whose strategies still remain less applied, there is a significant accumulation of mtDNA fragment sequences, especially of HVS-I. Both approaches are contributing to the increasing of the amount of available data. In addition, the inclination to analyze numerous samples and the diffusions of this kind of analysis for wide range of applications needs a greater expertise than previous request. In human population studies, which can be considered the most advanced in this field, the rapid accumulation of new phylogenetically informative polymorphisms and the growing derived complexity of the phylogenetically constructed trees of mtDNA and NRY, yielded to the continuous revision of the trees topology of both systems. The existence of a well established, hierarchical nomenclature system which could be self-update and the availability of the changing of the trees by consulting the International Society of Genetic Genealogy (http://www.isogg.org/) for NRY and the *PhyloTree* (http://www.phylotree.org/) for mtDNA can help researchers to orienting among haplogroup diagnostic mutations. Nevertheless, there was the necessity to develop new

informatics tools as *HaploGrep* (Kloss-Brandst tter et al. 2010), *MitoTool* (Fan and Yao 2013) or *Haplogroup Predictor* (Athey 2005) to assign haplotypes to their corresponding haplogroup.

To follow what already experienced in human population genetics, the aim of this study was to develop a new informatics tool for mitochondrial haplotypes assignment in a general evolutionary context. The Software would be especially dedicated to domestic animals and has been designed for a wide range of users which depending on the specific purposes and research field, are interested into assign a set of analyzed haplotypes to their haplogroups, but may differ regarding the level of expertise in haplogroups attribution methods. The manual assignment could be difficult, more prone to errors and if the number of samples is high can take long time. The use of *HapSign* could allow an accurate, fast and easy assignation. The rapid and automatic attribution to the belonging haplogroup would present a considerable advantage in terms of time and accuracy but also of cost and efforts of research. In fact, one of the possible interesting applications concern the use of the Software also to make a rapid screening based on the information contained in partial but informative mtDNA sequences. For example, in a study provided for genotyping a large set of samples of a species whose phylogeny based on complete mtDNA sequence is known, also a partial segment as HVS-I sequence in many cases could contain enough diagnostic sites to allow the assignation of the majority of haplotypes. A rapid and automatic assignment of haplotypes will allow limiting further analysis of the coding region only for haplotypes that are not assignable only on the base on the short HVS-I sequence.

Materials and Methods.

The HapSign.

HapSign calculates the haplogroup assignment score (*TotScore*) for one or more nucleotide sequences according to equation (1):

$$(1)\ TotScore = \frac{\sum_{nuc=1}^{totNuc} (pW^{mpW}\ pB^{mpB})}{totNuc}$$

The score assignment depends on the pairwise comparison between each nucleotide position (*nuc*) in the unknown sequence and a consensus sequence built for each of the known haplogroups existing in the database. The database used to produce the consensus sequences is the collection of those concatenated polymorphisms that are diagnostic for the haplogroup.

The nucleotide at position *nuc* with the highest frequency within a haplogroup is chosen to build the consensus and as weight for that nucleotide within that haplogroup (*pW*). Whether is not possible to choose a nucleotide (frequency 0.5) an 'N' is assigned.

A further variable to define the assignment score is the frequency between the haplogroups of the nucleotide chosen for the consensus (*pB*) described by equation (2):

$$(2)\ pB = \frac{freq_{nuc,HG}}{\sum_{HG=1}^{totHG} freq_{nuc,HG}}$$

where *HG* represent an haplogroup of the total number of haplogroups (*totHG*) in the reference dataset, and $freq_{nuc,HG}$ is the frequency of a nucleotide at position *nuc* within haplotype *HG*.

In order to allow the user to adjust the weight of the two parameters *pW* and *pB* in the final assignment, the two of them are provided with an exponent that can be modulated within 0 and 1 (excluded): *mpW* and *mpB*. This is meant to allow different setting to be applied due to the ongoing development of the method.

To define the quality of the assignment for each sequence the difference between the highest and the second highest scores is calculated to produce a delta index (Δ), the larger Δ values are, the better is the assignment.

To confidence in the assignment strongly depends on the quality of the database used as reference.

To appreciate the reliability of the reference database a bootstrap re-sampling of the sequences will be implemented to calculate the variance in the assignment and therefore a p-value, using a randomly picked sequence from each haplogroup in the same database as query sequence. Low values of confidence will lead the user to consider the assignment unreliable for that particular haplogroup, and along with that, will suggest increasing the number of diagnostic sites in the database if possible.

The script has been developed as an EXCEL application in order to make it as user friendly as possible as well as to improve the portability, moreover the computational effort required is minimal and the EXCEL environment is still a widely used workbench in the genetic field.

The input file format is the .rdf, commonly used for the software Network (Bandelt et al. 1999). A converter is included in the application to translate data from the spreadsheet to the .rdf format, this makes *HapSign* easy to get into the daily workflow (Figure 1.a).

The reference database can be modified manually or through the software to include/exclude sequences or nucleotide positions.

Two input files in “rdf” format are needed to run the analysis: the database of the reference haplotypes, and the database of one or more query sequences (Figure 1.b).
Each haplotype within the haplogroups in the reference database can be named directly by the user. To allow the pairwise comparison between matching nucleotide positions, it is required to have the nucleotide position named with the same format in both the database and the query sequence. However, the latter does not need to be filtered for the diagnostic sites, as the software will manage to identify the positions of interest.

First tests of assignment of HVS-I haplotypes in horse breeds.

At this stage, the first tests were made on mtDNA data of horse breeds. Among domestic animals, horse has been long studied. In 1994, Xu and co-workers published the first complete sequence of mtDNA of *E.caballus* but the number of complete sequences deposited in GenBank was limited until few years ago, while there was a high number of sequence data of the HVS-I. Within the several studies interested in the phylogeny of the maternal lineages in *E. caballus* a large part investigated the origins of specific breeds (Kavar et al., 1999; Luis et al., 2005; Royo et al., 2006; Bower et al. 2010), while others aimed to reconstruct the dynamics of domestication. Different models were proposed to explain the maternal variation of current breeds (Vila et al. 2001; Kavar and Dovc 2008). In a recent study (Morelli et al. 2013, in press) the variability of mtDNA has been analyzed in order to reveal the genetic structure and the relationships between two Sardinian native breeds with limited diffusion. The Giara and the Sarcidano horse were compared with other breeds from the *"Old-World"*. The analyses were conducted using a short fragment of 247 bp (np15494-15740) of HVS-I to reach the collection of a high number of published haplotypes.
The current availability of a greater number of complete mitochondrial sequences (Xu et al. 2007; Goto et al. 2011; Jiang et al. 2011; Lippold et al. 2011; Achilli et al. 2012) lead to the revision of the phylogeny of the mitochondrial lineages in *E. caballus*, in particular on the basis of the new phylogeny proposed by Achilli et al. (2012). The combined analysis of the d-loop and the coding region, allowed the definition of a parsimony tree, which explains the relation of 18 mitochondrial main haplogroups and new nomenclature rules were proposed.
In Morelli et al. 2013 (in press), a dataset consisting of 147 complete sequences were used to define a set consisting of 39 diagnostic sites in the HVS-I, which allow the assignment of the haplotypes to haplogroups. The assignment was done by the support of a probabilistic algorithm designed to

speed up the process. Applying this method all the HVS-I haplotypes of the Giara and Sarcidano horses

were assigned to the haplogroups. By the same approach were assigned almost all the haplotypes of a total of 1,232 individuals of other 43 horse breeds, representing six geographic regions of the "Old–World".

The dataset consisting of the 147 haplotypes, used to define the diagnostic nucleotide positions of the HVS-I, represented the database of the reference haplotypes used by *HapSign* to do the tests of assignment of the unknown sequences of 1232 individuals of 45 horse breeds.

A total of 237 HVS-I haplotypes were assigned manually and by the means of *HapSign.* The resulting assignations to the 18 haplogroups of mtDNA phylogeny were compared in order to do the first test of *HapSign* reliability.

Once prepared the two *.rdf files requested, they are imported directly by the utilities HapSign. The results appeared in a separate spreadsheet called HapSign Results in few minutes. Each of the query haplotypes were assigned and its relative assignment score and the Δ index value.

By *HapSign* all the query haplotypes were assigned to the corresponding haplogroup, while by manual methods the 3,4 % were not assigned.

A simple comparison were made between the results obtained by the two methods. The results are coherent in the majority of cases (97%), only a low percentage of differences were found. Moreover, contrasting results encompass all the haplotypes not manually assigned and some others, all presenting some problem of interpretation.

One of these haplotypes, lacking diagnostic sites, was manually assigned to A haplogroup while by *HapSign* to C haplogroup. About 80% of the A haplotypes had the mutation 15720G, 20% of which showed an association between 15720G and 15495G. Overall, the 20% of the A haplotypes had no diagnostic sites and the accuracy of their haplogroup attribution could only be confirmed when HVS-I–specific sites were available. The haplogroup C likely the JK clade, is well defined by specific variants in the coding region, and is associated with hypervariable mutations in HVS-I. For this reason, when these mutations were available, inference of haplogroup attribution was conducted by either a comparison or an exclusion criterion (Morelli et al. 2013, in press). In this case, both methods are not able to assign the haplotype with confidence. Other two haplotypes assigned to J were classified Q (for the unexpected presence of 15703C one of the diagnostic site of OPQR) and C by HapSign. As reported above the JK clade and C is not easily assign only based on

HVS-I and an accurate evaluation is needed. Another example is a haplotype belonging to the OP clade, which was identified by the mutation 15667G. Despite the presence of the 15635T mutation, which should distinguish O from P clade, due to the lacking of the 15597G (present in all haplotypes of the O haplogroup in the reference database) was assigned to P by the *HapSign.*
Anyway, is relevant to note that only two O haplotypes were present in the reference database, thus is not possible to exclude that not all the possible pattern are represented.
Finally, all the haplotypes not assigned manually were assigned to A, B, D, P and Q due to a controversial mutations pattern.
In this analysis the delta index vary from a maximum value of ~0,12 to 0,0015 as the minor. It is interesting to report how the delta index vary in the different haplogroups. As shown in figure 2, for these haplogroups, along the well defined and frequent in the reference database A and the G haplogroups, the delta indexes calculated for the D, F, H, I, L, M, N, Q, and R haplogroups are the highest. These last are defined by highly informative mutations which are linked univocally to all those haplogroups, further confirming that the quality of assignation strongly depends on the nature of the variation of HVS-I which differs in quality and degree of information.

Preliminary conclusion and future perspective.
The preliminary tests conducted by the *HapSign* as a new software tool for mitochondrial haplotype assignment, are giving promising results. Almost all (97%) of the unknown haplotypes were assigned by *HapSign* and manually to the same haplogroup. The discrepancy observed pointed out to the importance of the quality of the database used as reference as a fundamental requisite to confidence in the assignment. An update of the reference database by new complete sequences and the implementation of the Software by a bootstrap re-sampling of the reference sequences with a p-value to calculate the variance in the assignment will be the next step.

ACKNOWLEDGMENT
AU was supported by P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività l.3.1. Sassari University, 2010/2011 – XXVI cycle of Environmental Biology PhD program.

Bibliography.
Bagatharia SB, Joshi MN, Pandya RV, Pandit AS, Patel RP, Desai SM, Sharma A, Panchal O, Jasmani FP, Saxena AK. (2013). Complete mitogenome of Asiatic lion resolves phylogenetic status within Panthera. BMC Genomics. 14:572.

Achilli A, Olivieri A, Soares P, Lancioni H, et al. (2012). Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc. Natl. Acad. Sci. USA. 109:2449-54.

Bandelt HJ, Forster P, and Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. Mol.Biol. Evol. 16: 37-48

Bruford MW, Bradley DG, and Gordon Luikart G (2003). DNA markers reveal the complexity of livestock domestication. Nature 4: 900-910

Elhaik E, Greenspan E, Staats S, Krahn T, Tyler-Smith C, Xue Y, Tofanelli S,Francalacci P, Cucca F, Pagani L, Jin L, Li H, Schurr TG, Greenspan B, Spencer Wells R; Genographic Consortium. (2013) The GenoChip: a new tool for genetic anthropology. Genome Biol Evol. 2013;5(5):1021-31.

Goto H, Ryder OA, Fisher AR, Schultz B, et al. (2011) A massively parallel sequencing approach uncovers ancient origins and high genetic variability of endangered Przewalski's horses. Genome Biol. Evol. 3:1096-106.

Jessica L. Petersen, James R. Mickelson, Aaron K. Rendahl, Stephanie J. Valberg, Lisa S. Andersson, Jeanette Axelsson, Ernie Bailey, Danika Bannasch, Matthew M. Binns, Alexandre S. Borges, Pieter Brama, Artur da Caˆmara Machado, Stefano Capomaccio, Katia Cappelli, E. Gus Cothran, Ottmar Distl, Laura Fox-Clipsham, Kathryn T. Graves, Gerard Guerin, Bianca Haase, Telhisa Hasegawa, Karin Hemmann, Emmeline W. Hill, Tosso Leeb, Gabriella Lindgren, Hannes Lohi, Maria Susana Lopes, Beatrice A. McGivney, Sofia Mikko, Nicholas Orr, M. Cecilia T. Penedo, Richard J. Piercy, Marja Raekallio, Stefan Rieder, Knut H. Røed, June Swinburne, Teruaki Tozaki, Mark Vaudin, Claire M. Wade, Molly E. McCue (2013) Genome-Wide Analysis Reveals Selection for Important Traits in Domestic Horse Breeds. PLoS Genet 9(1): e1003211. doi:10.1371/journal.pgen.1003211

Jiang Q, Wei Y, Huang Y, Jiang H, et al. (2011). The complete mitochondrial genome and phylogenetic analysis of the Debao pony (Equus caballus). Mol. Biol. Rep. 38:593-599

Kavar T and Dovc P (2008). Domestication of the horse: Genetic relationships between domestic and wild horses. Liv. Sci. 116:1-14

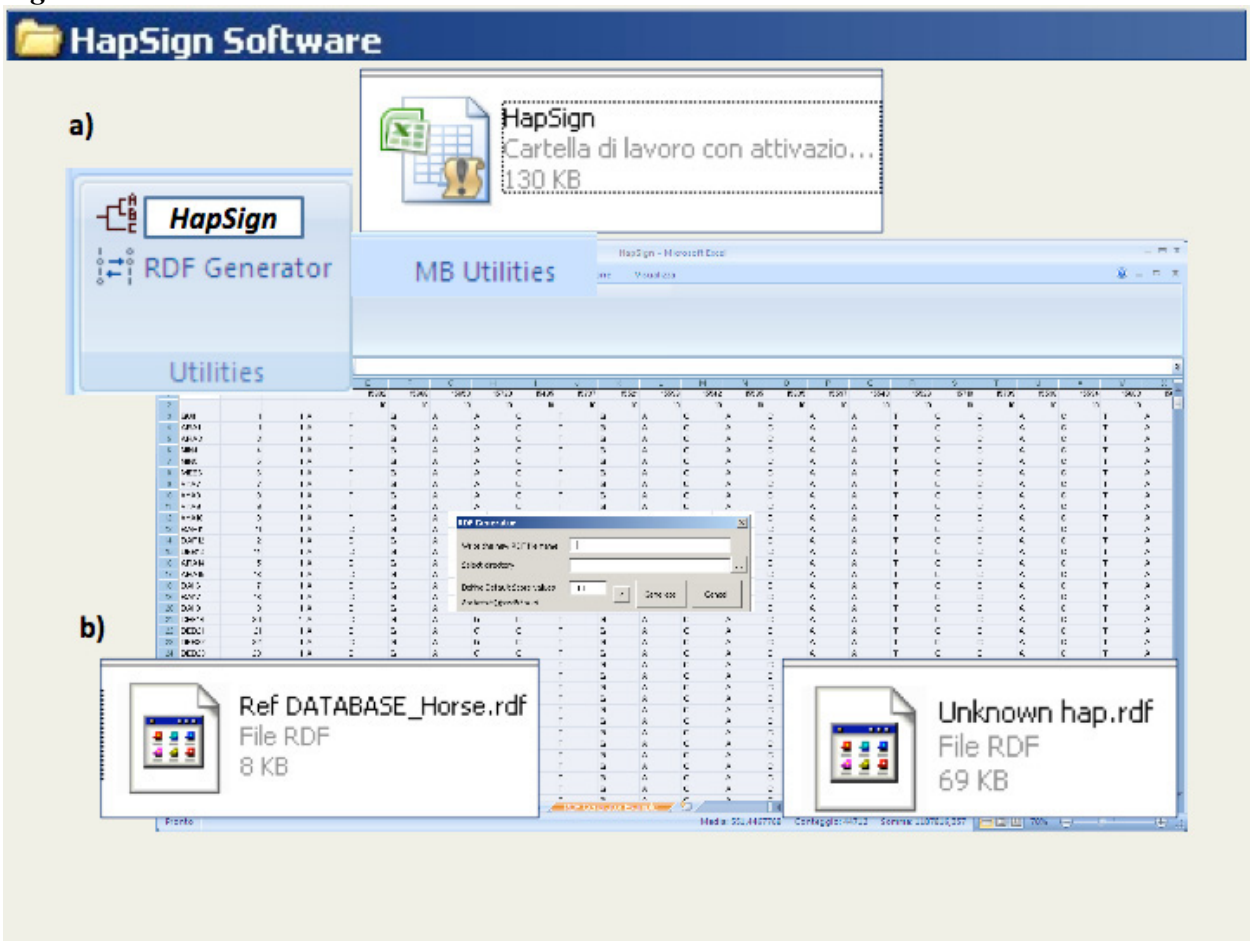
Lippold S, Matzke NJ, Reissmann M, and Hofreiter M. (2011)Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. BMC Evol. Biol. 11:328.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M,Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. . (2013) Great ape genetic diversity and population history. Nature499(7459):471-5.

Soares P, Achilli A, Semino O, Davies W, Macaulays V, et al. (2010) The Archaeogenetics of Europe Curr Biol 20: 174-183.

Xu X and Arnason U (1994). The complete mitochondrial DNA sequence of the horse, Equus caballus: extensive heteroplasmy of the control region. Gene 148:357-362.

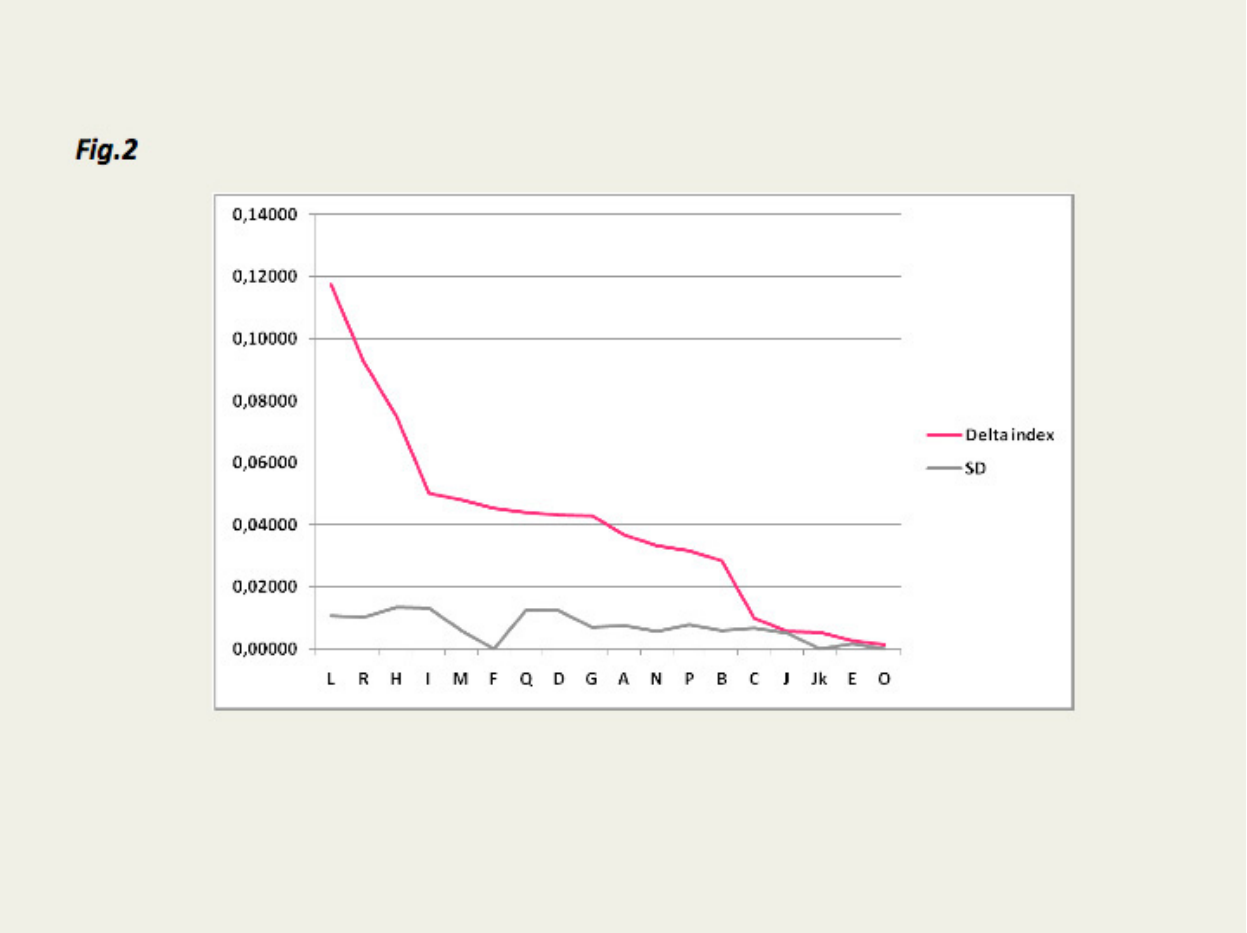
Figure 1.



HapSign feature. The script has been developed as an EXCEL application. A converter is included in the application to translate data from the spreadsheet to the .rdf format. (Figure 1.a).

As example, two input files in rdf format needed to run the analysis are shown: the database of the reference haplotypes, and the database of one or more query sequences (Figure 1.b).

Figure 2.



In the graphic the variation of average values of delta index calculated by HapSign for each of the 18 mtDNA haplogroups of horses are shown. (delta index, red curve). Standard deviation (SD, grey curve).

CHAPTER 3
Analysis of the variability of mitochondrial DNA in the genus *Ovobathysciola*, a Sardinian endemic subterranean *Coleoptera* of the *Leptodirini* tribe
-Manuscript in preparation-

Authors: Antonella Useli, Paolo Marcia, Paolo Francalacci, Achille Casale
Affiliations: Dipartimento di Scienze della Natura e del Territorio - Unità di Zoologia Archeozoologia e Genetica -, Università di Sassari, Sassari, Italy

Keywords: mitochondrial DNA, COI gene, subterranean *Coleoptera.*
Abstract.
In this study was analyzed the variability of the first sub-unit of the Cytochrome c Oxydase gene (COI or cox1) of the mtDNA, to establish the phylogenetic relations among species of *Ovobathysciola* already known and potentially new ones coming from various Sardinian caves.
The tree constructed by the Maximum Likelihood method (Figure 2) showed the relation among the *Ovobathysciola* species analyzed. The tree topology seems to be coherent with the taxonomic units and with the geographic locations. Preliminary results suggest to further investigate the phylogeography pattern. The results produced will be interpreted on the basis of the integration of knowledge gained through the study of the same species with a morphological approach.

Introduction.
Cave organisms have long been considered a model system for testing evolutionary and biogeographic hypotheses because of their isolation, simplicity of community structure and specialization (Juan and Emerson, 2010).
A recent study (Ribera et al. 2010) focused on the origin, the distribution and evolution of diverse lineage of subterranean beetles of the tribe Leptodirini (family Leiodidae) in the western Mediterranean area. This study was conducted by the analysis of the 3'-end of COI sequence, already largely used in phylogenetics studies of insects, in combination with other mitochondrial and nuclear genes.
The subterranean *genus Ovobathysciola* genera (Staphylinoidea, Cholevidae) belonging to the same tribe. Only three species are so far described: *O. majori*, *O. gestroi* and *O. grafittii*, all of them endemic to Sardinia.
In this study was analyzed the variability of the first sub-unit of the Cytochrome c Oxydase gene (COI or cox1) of the mtDNA, to establish the phylogenetic relations among species of *Ovobathysciola* already known and potentially new ones coming from various Sardinian caves. In

addition, the integration with the morphological data will allow to establish the phylogenetic position of each species with more precision and to clarify their complex biogeographic pattern.
In fact, the evolution of this gene is considered to be fast enough to discriminate closely related species, but also phylogeographic groups within single species. In addition, thanks the study of Ribera and co-workers (2010), there are many sequences of the same fragment revealing the relationships of the other groups of the tribe *Leptodirinae.* Moreover, the mtDNA cox1 fragment was already analyzed by Caccone and Sbordoni (2001) in the known species of *Ovobathysciola.* Although only four haplotypes belonging to *O. majori, O. gestroi, O.grafitti* and to one *Ovobathysciola sp.* (Casale, pers. comm. in Caccone and Sbordoni, 2001) are available in GenBank, it will be possible to make comparisons necessary for the analysis of the variability in species in the study.

Materials and Methods.

Samples collection.

The selection of samples was performed by including two of the species already know (*O. majori* and *O. gestroi)*, while the other specimens were potentially *Ovobathysciola* new species (sp.n.) from the locations indicated in figure 1.
The sampling was designed in order to have a picture of the variability of the mtDNA fragment analysed, both between species of the same genus and within species, in order to reveal any possible phylogeographic group. Two specimens of the genus *Bathysciola* , were analysed as out-group for phylogenetic analysis. In addition, other two samples of a related group not already studied were included.
A total of 37 of the collected specimens were analysed for cox 1 variation. The extraction of genomic DNA was made by a modified salting-out method (Miller et al. 1988) which yielded good results in the majority of cases. For the amplification of the *cox 1* fragment a pair of conserved primers (Simon et al. 1994) were used. The amplification was carried out in standard conditions, only varying the annealing temperature (45-48 Cº) to obtain amplification of all the samples.
Whereas in GenBank there is a very limited number of cox 1 sequences of the Ovobathysciola genus and that the analysis concerns potentially new species, the sequence reaction was performed for both strands, in order to obtain a greater certainty of the base attribution.
Sequencing was performed by an external sequencing core service (Macrogen Europe).

Once compared the two obtained sequences for each samples by the alignments Software BioEdit 7.0.5.2 (Hall, 1999), the specificity was also verified through the use of BLAST (Basic Local Alignment Search Tool, http://blast.ncbi.nlm.nih.gov).

Subsequently, a first alignment were made with the complete sequence of the mitochondrial genome of *Drosophila yakuba* (> gi | 5834829 | ref | NC_001322.1 |), which represents the reference for the coordinates of the used primers; this comparison it was necessary to define the nucleotide positions of the fragment of interest. Finally, we proceeded with the multi-alignment of the sequences produced with those of *Ovobathysciola* available in GenBank (GB), for the identification of polymorphic sites in the fragment analyzed. The first evolutionary analysis were carried out by the use of MEGA 5 Software (Tamura et al. 2011).

Maximum Likelihood fits of 24 different nucleotide substitution models were considered. Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (lnL), and the number of parameters (including branch lengths) are also presented (Nei and Kumar, 2000). For estimating ML values, a tree topology was automatically computed. The analysis involved 41 nucleotide sequences. Codon positions included were $1^{st} + 2^{nd}$. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were 541 positions in the final dataset. A tree were constructed by the Maximum Likelihood method based on the Tamura 3-parameter model (T92).

The bootstrap consensus tree inferred from 1000 replicates (Felsenstein J., 1985) was calculated to represent the evolutionary history of the taxa analyzed (Felsenstein J., 1985). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein J., 1985). Initial tree(s) for the euristic search were obtained automatically by applying Neighbour-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.7799)).

Preliminary results.

The tree constructed by the Maximum Likelihood method (Figure 2) showed the relation among the *Ovobathysciola* species analyzed. The tree topology seems to be coherent with the taxonomic units indicated (the species) and with the geographic locations. There are two main clades, in the first there are almost all the species of the central-east area *O. majori* (area 1), *Ovobathysciola* sp.N.3 (area 3) *Ovobathysciola* sp. N.4) (site 4) and one of the west-central Sardinia *Ovobathysciola* sp.n Ac (area 2); in the second there are the species of the south-east of Sardinia *O. gestroi* (area 7). It interesting to note that in latter there is also a sub-clade where the *O. grafittii,* which is known only from one cave in northwestern Sardinia, is closer to the *Ovobathysciola* sp.N.5. (area 5) despite this last one is from the northernmost area.

If we focused on the sub-clades, it is possible to observe that seven main sub-clades correspond of each one of the main sampling areas. The main clades and sub-clades were also well supported by high bootstrap values. Moreover, it is interesting to note that within the main sub-clades, also a certain amount of variation could be revealed, and this structure is related to the cave of origin of each group of samples analyzed. In particular the *Ovobathysciola* sp.N.4 sub-clade show two separate groups supported by a high bootstrap value. Finally, in the clade of *O.gestroi*, one of the sample seems to be differentiated from the other that are in strict relation, on the same branch.

First conclusion and future perspective

The mtDNA cox 1fragment seemed to be informative for the study of the *Ovobathysciola* genus. The preliminary results depicted an interesting framework to study and suggest to further investigating the phylogeography pattern. The results produced will be interpreted on the basis of the integration of knowledge gained through the study of the same species with a morphological approach.

Bibliography.

Caccone A. and Sbordoni V. Molecular Biogeography of cave life: a study using mitochondrial DNA from Bathysciine Beetles.(2001). Evolution, 55(1), 2001, pp. 122-130.

Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95-98

Juan C. and Emerson BC (2010). Evolution underground: shedding light on the diversification of subterranean insects-Mini-review- Journal of Biology , 9:17.

Miller SA, Dykes DD, Polesky HF, A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res, (1988) 16(3):1215.

Ribera, I., Fresneda, J., Bucur, R., Izquierdo, A., Vogler, A.P., Salgado, J.M. & Cieslak, A. (2010) Ancient origin of a Western Mediterranean radiation of subterranean beetles. BMC Evolutionary Biology, 10, 29.

Simons C, Frati F, Bechenbach A, Crespi B, Liu H and Floors P. (1994). Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase Chain Reaction Primers. Ann. Entomol. Soc. Am. 87(6): 651-701.

Tamura K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Molecular Biology and Evolution 9:678-687.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731-2739.

Figure 1. Map of Sardinia Island with the sampling locations.

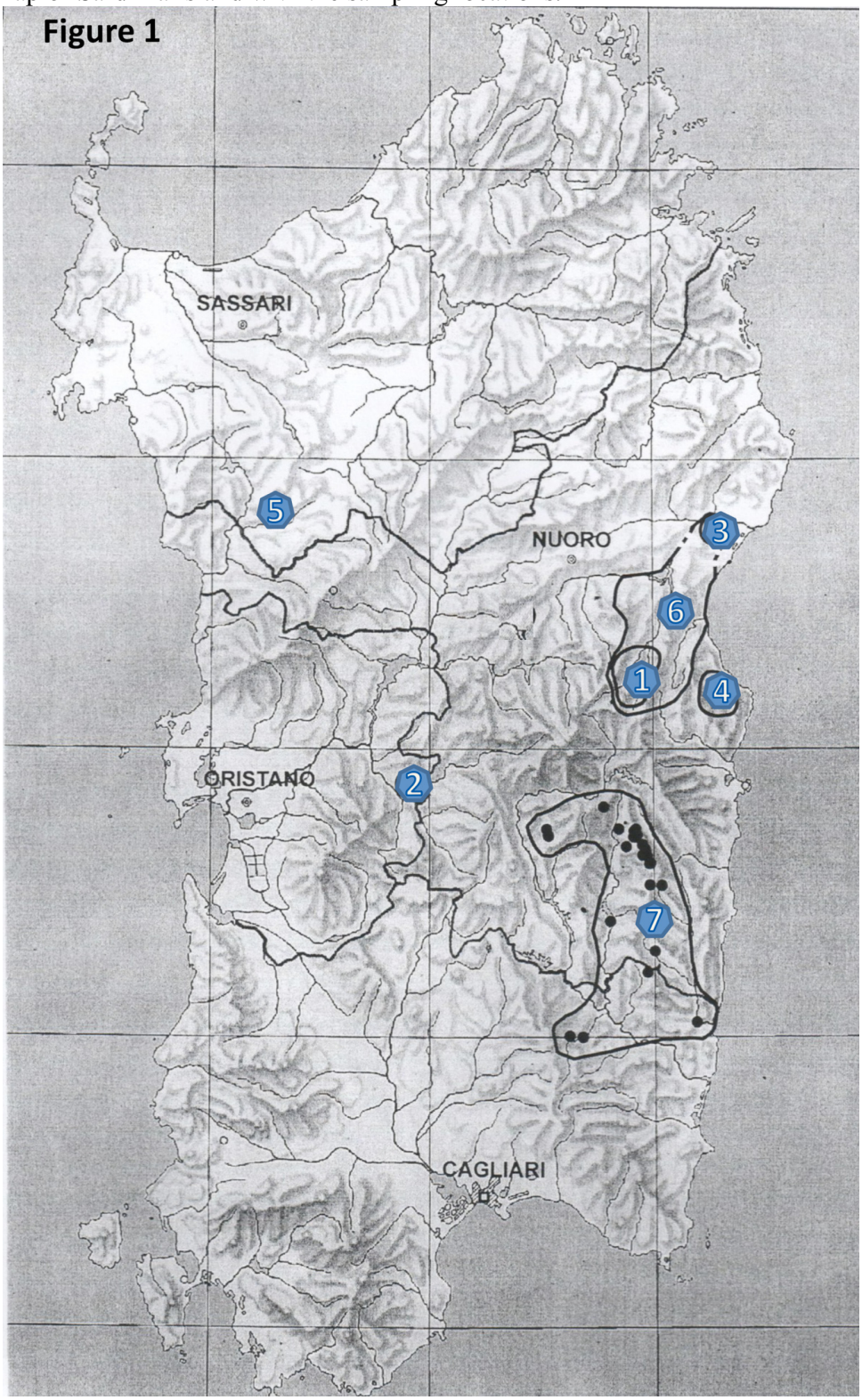


Figure 2. In the figure the Maximum Likelihood tree calculated is shown. The last number of the sample names (.N) represent the sample locations as indicated on the map in figure 1. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches.

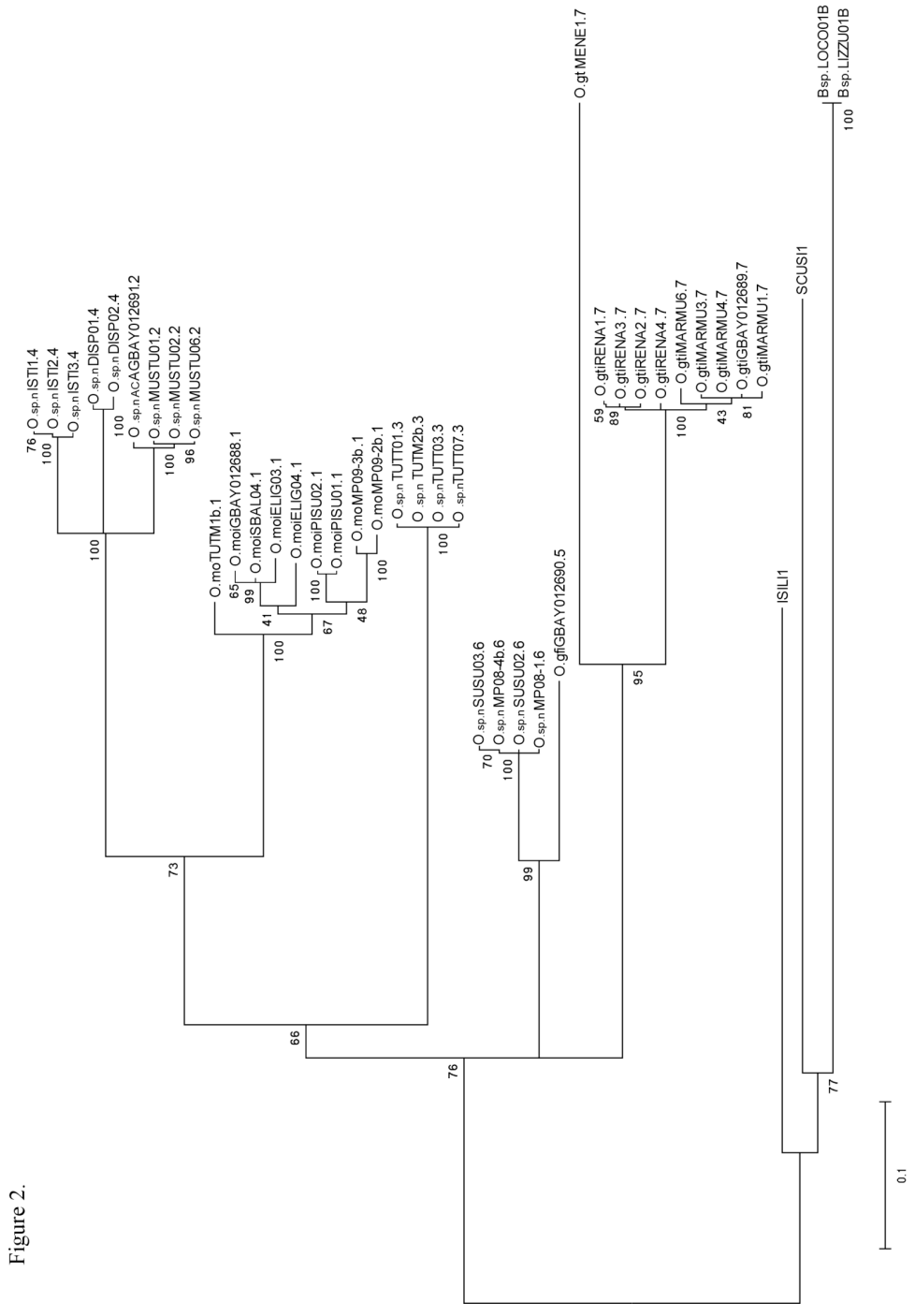


Figure 2.

4.1 Low pass DNA sequencing of 1,200 Sardinians reconstructs European Y chromosome phylogeny

Low pass DNA sequencing of 1,200 Sardinians reconstructs European Y chromosome phylogeny

Paolo Francalacci[1✉], Laura Morelli[1†], Andrea Angius[2,3], Riccardo Berutti[3,4], Frederic Reinier[3], Rossano Atzeni[3], Rosella Pilu[2], Fabio Busonero[2,5], Andrea Maschio[2,5], Ilenia Zara[3], Daria Sanna[1], Antonella Useli[1], Maria Francesca Urru[3], Marco Marcelli[3], Roberto Cusano[3], Manuela Oppo[3], Magdalena Zoledziewska[2,4], Maristella Pitzalis[2,4], Francesca Deidda[2,4], Eleonora Porcu[2,4,5], Fausto Poddie[4], Hyun Min Kang[5], Robert Lyons[6], Brendan Tarrier[6], Jennifer Bragg Gresham[6], Bingshan Li[7], Sergio Tofanelli[8], Santos Alonso[9], Mariano Dei[2], Sandra Lai[2], Antonella Mulas[2], Michael B. Whalen[2], Sergio Uzzau[4,10], Chris Jones[3], David Schlessinger[11], Gonçalo R. Abecasis[5], Serena Sanna[2], Carlo Sidore[2,4,5], and Francesco Cucca[2,4✉]

[1] Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, 07100 Sassari, Italy

[2] Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy

[3] Center for Advanced Studies, Research and Development in Sardinia (CRS4), Pula, Italy

[4] Dipartimento di Scienze Biomediche, Università di Sassari, 07100 Sassari, Italy

[5] Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[6] DNA Sequencing Core, University of Michigan, Ann Arbor, MI, USA

[7] Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA

[8] Dipartimento di Biologia, Università di Pisa, 56126 Pisa, Italy

[9] Departamento de Genética, Antropología Física y Fisiología Animal. Universidad del País Vasco UPV/EHU. 48080 Bilbao, Spain

[10] Porto Conte Ricerche, Località Tramariglio, Alghero, 07041 Sassari, Italy

[11] Laboratory of Genetics, National Institute on Aging, Baltimore, Maryland, USA

✉ Corresponding Authors

† Laura Morelli prematurely passed away on Feb. 20th, 2013. This work is dedicated to her memory

ABSTRACT (~125 words)

Genetic variation within the male specific portion of the Y chromosome (MSY) can clarify the origins of contemporary populations, but previous studies were hampered by partial genetic information. Population sequencing of 1,204 Sardinian males identified 11,763 MSY single nucleotide polymorphisms (SNPs), of which 6,751 have not previously been observed. We constructed a MSY phylogenetic tree containing all main haplogroups found in Europe along with many Sardinian-specific lineage clusters within each haplogroup. The tree was calibrated with archaeological data from the initial expansion of the Sardinian population ~7,700 years ago. The ages of nodes highlight different genetic strata in Sardinia and reveal presumptive timing of coalescence with other human populations. We calculate a putative age for coalescence of ~180-200,000 years ago, consistent with previous mitochondrial DNA (mtDNA) based estimates.

One Sentence Summary: Human demographic history can be inferred from analyses of Sardinian Y chromosome DNA sequences.

New sequencing technologies have provided genomic data sets that can reconstruct past events in human evolution more accurately (1). Sequencing data from the male specific portion of the Y chromosome (MSY) (2), because of its lack of recombination and low mutation, reversion and recurrence rates, can be particularly informative for these evolutionary analyses (3, 4). Recently, high coverage Y chromosome sequencing data in 36 males from different worldwide populations

(5) assessed 6,662 phylogenetically informative variants and estimated the timing of past events, including a putative coalescence time for modern humans of ~101-115 thousand years ago.

MSY sequencing data reported to date still represent a relatively small number of individuals from a few populations. Furthermore, dating estimates are also affected by the calibration of the phylogenetic tree used to establish the rate of molecular change over time. This calibration can either correlate the number of nucleotide substitutions with dates from paleontological/archaeological records (phylogenetic rate) or can use directly observed *de novo* mutations in present-day families (mutation rate). However, both approaches are complicated by several variables (6, 7).

Some of these problems can be resolved by the analysis of MSY sequencing data of many individuals from a genetically informative population where archaeological data are available to provide suitable calibration points. This prompted us to use large-scale MSY sequencing data from the island population of Sardinia for phylogenetic analysis. We generated a high-resolution analysis of the MSY from population sequencing of 1,204 Sardinian males, (8). We used a hierarchical approach and, to be consistent with previous work (5), focused on approximately 8.97 Mbp from the Y chromosome in the X-degenerated region. We inferred 11,763 MSY phylogenetically informative SNPs, detected in at least two individuals and unequivocally associated with specific haplogroups and sub-haplogroups; 6,751 of these SNPs were not thus far reported in public databases.

The informative SNPs were used to construct a parsimony-based phylogenetic tree. To root the tree, we used the chimpanzee genome reference as outgroup and inferred the ancestral status at all SNP sites except for 26 that were discarded in further analysis. The first bifurcation point, and thus the Most Recent Common Ancestor, separates samples 1-7 from the rest of the sample (samples 8-1204) (Table 1). The average number of derived alleles in the 1,204 males is 1,002.6 (±21.2 s.d.)

which, consistent with a neutral evolution of these Y polymorphisms, shows a remarkable uniformity of the branch length.

The Sardinian sequences show a very high degree of inter-individual variation. As shown in a schematic tree (Fig. 1), all of the most common Y chromosome haplogroups previously detected in Europe are present in our sample (Table 1), with the sole exception of the northernmost Uralic haplogroup N. The first bifurcation separates the mostly sub-Saharan haplogroup A (7 individuals, 0.6% in our sample) from the others. Haplogroup E (132, 11.0%) is present with its European clade, characterized by the presence of the M35 marker, together with a small number of individuals belonging to the mainly African clade E1a. The rare haplogroup F (7, 0.6%) is related to haplogroup G (131, 10.9%), which shows a private Sardinian-Corsican clade whose ancient roots have been found in an Eneolithic sample from the Italian Alps (9). Haplogroup I (490, 40.7%) is of special interest because it is mostly represented by the I2a1a clade, identified by the M26 marker, which is at high frequencies in Sardinia (10) but is rare or absent elsewhere (11). Haplogroup J (161, 13.4%) is observed with its main subgroups; and the super-haplogroup K is present with the related L and T branches (36, 3.0%), with a single individual of haplogroup Q (1, 0.08%), and with the more common haplogroup R (239, 19.9%), occurring mostly as the western European M173-M269 branch.

Almost half the discovered SNPs (4,872) make up the skeleton of the phylogenetic tree and constitute the root of the main clades. The skeleton comprises lineages that are unbranched for most of their length, with ramifications only in the terminal portion. This indicates an early separation of the clades followed by new variability generated during subsequent expansion events.

To estimate points of divergence between Sardinian and continental clades, we sequenced two samples from the Basque Country and Northern Italy, belonging to haplogroup I, and two, from Tuscany and Corsica, belonging to haplogroup G. We also analysed the sequence of the so-called

Iceman Ötzi (9), together with 133 publicly available European sequences from the 1,000 Genome database, and those SNPs from the ISOGG database detected outside Sardinia.

Notably the Basque individual separates from the basal position of the I2a1a branch that encompasses 11 Sardinian individuals. The Northern Italian sample, instead, most likely reflecting the last step of I2a1 lineages before their arrival in Sardinia, is at the basal point of most of the remaining I2a1a samples (Fig. 2). Considering two other basal lineages encompassing only Sardinian samples, we can infer that when the I2a1a sub-haplogroup entered Sardinia it had already differentiated into four founder lineages that then accumulated private Sardinian variability. Two other founder clades show similar divergence after entry into the island - one belonging to haplogroup R1b1c (xV35) (whose differentiation is identified contrasting the Sardinian data with the ISOGG and 1000 Genome data); and the other, to haplogroup G2a2b-L166 (identified by divergence from a sequenced Corsican sample).

The branch length uniformity observed in our phylogeny is consistent (Fig. 1) with a relatively constant accumulation of SNPs in different lineages over time. Hence, this accumulation can be effectively used as a molecular clock for the dating of branch points. We calibrated accumulation of Sardinian-specific genetic variation against established Sardinian archaeological records, reviewed in Supplementary Figures 7 and 8 and accompanying text, indicating a putative age of initial demographic expansion ~7,700 years ago (8, 12), that is also supported by mtDNA analyses (13). Importantly, comparison of Sardinian genetic variation with that found elsewhere helped us to establish the amount of variability produced during and after this expansion, resulting in sub-lineages that appear to be unique to the island.

We focused our calibration analyses on the individuals belonging to the I2a1a-δ clade, which is shared by 435 individuals and is best suited to assess the Sardinian specific variability. Taking into account the average variation of all Sardinian individuals in the common I2a1a-δ clade of 37.3

(±7.8) SNPs, a calibration point of 7,700 years ago results in a phylogenetic rate of one new mutation every 205 (±50) years. Considering that our analysis focused on approximately 8.97 Mbp of sequence from the Y chromosome, X-degenerated region, this rate is equivalent to 0.53 x 10^{-9} bp^{-1} $year^{-1}$. Notably, this phylogenetic rate is consistent with the value of 0.617 (0.439-0.707) x 10^{-9} bp^{-1} $year^{-1}$ - from the genome-wide mutation rate observed from *de novo* mutations adjusted for Y chromosome specific variables (14). Our mutation rate is instead lower than the value of 1.0 x 10^{-9} bp^{-1} $year^{-1}$ obtained from *de novo* MSY mutations in a single deep-rooted family (5) which also coincides with that traditionally deduced from the *Homo-Pongo* divergence (15).

Using our phylogenetic rate of 0.53 x 10^{-9} bp^{-1} $year^{-1}$, we estimated the Time to the Most Recent Common Ancestor (TMRCA) of all samples, whose average variability is 1,002.6 (±21.2) SNPs, at ~200,000 years ago. This is older than previously proposed (16) for the Y chromosome, but is in agreement with estimates from a de novo mutation rate in an African Y chromosome lineage (14) and with the revised molecular clock for humans (7) and the TMRCA estimated from analyses of maternally inherited mtDNA (13, 17).

The main non-African super-haplogroup F-R shows an average variation of 534.8 (±28.7) SNPs, corresponding to a TMRCA of ~110,000 years ago, in agreement with fossil remains of archaic *Homo sapiens* out of Africa (7,18) though not with mtDNA, whose M and N super-haplogroups coalesce at a younger age (13). The main European subclades show a differentiation predating the peopling of Sardinia, with an average variation ranging from 70 to 120 SNPs (Table 1), corresponding to a coalescent age between 14,000 and 24,000 years ago, compatible with the post glacial peopling of Europe.

However, the inferred phylogenetic rate and dating estimates presented here remain tentative because the calibration date was deduced from archaeological data, which may be incomplete and typically cover a relatively large temporal interval making it difficult to indicate a specific moment

in time. A more precise calibration point could be obtained by sequencing DNAs from prehistoric Sardinian remains with absolute ^{14}C dating. Further limitations derive from the scarcity of related samples for rare lineages coupled with the low pass sequencing approach we used (8). Low pass sequencing is expected to detect nearly all common variants (frequency >1%) but to miss rare variants. Missed variants have competing effects on estimates of ancestral coalescent times: when they lead to missed differences among haplotypes diverged after the founding of Sardinia they lower our calibrated estimates of mutation rate and increase coalescent time estimates; when they lead to missed differences among ancestral clades, they lower these time estimates. In fact, despite the overall homogeneity of the length of the branches from the MRCA (Fig. 1), those represented by fewer individuals are generally shorter (8). To estimate the effect of missed variants on the age estimates, we sequenced with deep coverage 6 selected individuals, 4 of them belonging to the I2a1a-δ clade, used for calibration and two with the I2a1a-β and J2b2f clades, used as outgroup. Notably, the deep sequencing of the I2a1a-δ samples yielded an average of 45.7 (±2.2) Sardinian specific SNPs among these haplotypes (versus 37.3 (±7.8) in low coverage data), corresponding to a phylogenetic rate of 0.65×10^{-9} bp^{-1} $year^{-1}$. Overall, this re-analysis suggested a slightly more recent TMRCA (lower than 13%) still in substantial agreement with the antiquity of the main Y chromosome haplogroups (8).

Hence, despite current limitations, the calibration used from common haplogroups in over a thousand people from this isolated population, including many island-specific SNPs, permit an estimate of main demographic events during the peopling of Sardinia concordant with the archaeological/historical record and ancient DNA analysis (8). The initial expansion of the Sardinian population, used for calibration, is marked by six clades belonging to three different haplogroups, with an average variation of around 35-40 SNPs, representing the ancient founder core of modern Sardinians.

Our data further reveal a more intricate scenario of Sardinian demographic history, in which additional variability was introduced after the Early Neolithic expansion, to add to the observed great inter-individual variation. Further expansion likely occurred in the Late Neolithic (~5,500-6,000 years ago), associated with some clades of E, R and G that show Sardinian-specific variability of 25-30 SNPs and related to later genetic flow (Table 1). Additional variation putatively arrived with groups of individuals carrying other haplogroups (namely the I clades different from I2a1, J and T). Taken together, the genetic data and demographic expansions are consistent with classical archaeological data indicating that Sardinia reached a considerable population size in prehistoric times; the estimated population during the Nuragic Period (~2,500-3,700 years ago) was > 300,000 inhabitants (19). Finally, the rare, mostly African A1b-M13 and E1a-M44 clades could have come to Sardinia in more recent times up to the historic period corresponding to the Roman and Vandalic dominations, suggested by a private Sardinian variability of 7-10 SNPs.

REFERENCES AND NOTES

1. 1000 Genomes Project Consortium, *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
2. H. Skaletsky, *et al.*, The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** :825-37 (2003).
3. P.A. Underhill *et al.*, Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**, 358–361 (2000).
4. O. Semino, *et al.*, The genetic legacy of Palaeolithic Homo sapiens sapiens in extant Europeans, a Y-chromosome perspective. *Science* **290**, 1155-1159 (2000).
5. W. Wei, *et al.*, A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res* **23**, 388–395(2013).
6. A. Scally, R. Durbin, Revising the human mutation rate:implications for understanding human evolution. *Nat Rev Genet* **13**, 745-753 (2012).
7. A. Gibbons, Turning back the clock: slowing the pace of prehistory. *Science* **338,** 189-191 (2012).
8. Materials and methods are available as supplementary materials on *Science* Online.
9. A. Keller *et al.*, New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Commun* ***2*3**, 698 (2012).
10. P. Francalacci *et al.*, Peopling of three Mediterranean islands (Corsica, Sardinia and Sicily) inferred by Y-chromosome biallelic variability. *Am J Phys Anthrop* **121**, 270-279 (2003).
11. S. Rootsi *et al.*, Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am J Hum Genet* **75**, 128-137 (2004).

12. R.H. Tykot, Radiocarbon Dating and Absolute Chronology in Sardinia and Corsica. In *Radiocarbon dating and Italian Prehistory,* R. Skeates, R. Withehouse, Eds.(Accordia Specialist Studies on Italy, London, 1994), pp. 115-145.

13. A. Olivieri *et al.,* The mtDNA legacy of the Levantine early upper Palaeolithic in Africa. *Science* **314**, 1767-1770 (2006).

14. F. L. Mendez *et al.*, An African American Paternal Lineage Adds an Extremely Ancient Root to the Human Y Chromosome Phylogenetic Tree. *Am J Hum Genet* **92**, 1–6 (2013).

15. A. Hobolth, J.Y. Dutheil, J. Hawks, M.H. Schierup, T. Mailund, Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res* **21**, 349–356 (2011).

16. F. Cruciani *et al.*, A revised root for the human Y chromosomal phylogenetic tree, the origin of patrilineal diversity in Africa. *Am J Hum Genet* **88**, 814-818. (2011).

17. M. Ingman, H. Kaessmann, S. Pääbo, U. Gyllensten, Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).

18. S.J. Armitage *et al*., The Southern Route "Out of Africa":Evidence for an Early Expansion of Modern Humans into Arabia. *Science* **331,** 453–456 (2011).

19. G. Lilliu, *La civiltà nuragica* (Delfino, Sassari, 1982).

20. Funding information. This research was supported in part by NIH contract NO1-AG-1-2109 from the National Institute of Aging (NIA) to the IRGB institute, by the Sardinian Autonomous Region (L.R. n°7/2009) grants cRP3-154 to FC and cRP2-597 to PF, by Fondazione Banco di Sardegna to PF and LM, by Basque Gov. G.I.C. IT-542-10 to SA, and National Human Genome Research Institute grants HG005581, HG005552, HG006513, HG007022 to G.R.A. We are grateful to all the Sardinian donors for providing blood samples. We also thank the 1,000 Genomes consortium for making available their

sequencing data that, in compliance with the Fort Lauderdale principles. We thank H. Skaletsky for detailed information about sequence blocks on the Y chromosome; M. Uda and R. Nagaraja for useful comments; C. Calò, D. Luiselli and C. de la Rua for providing some non-Sardinian samples, E. Garau, M. Rendeli and B. Wilkens for the archaeological background, and the CRS4 HPC group for their IT support, and in particular, L. Leoni and C. Podda.

REFERENCES OF THE SOM

21. G. Pilia, *et al.*, Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**(8),e132 (2006).

22. D. Contu, *et al.*, Sex related bias and exclusion mapping of the non-recombinant portion of chromosome Y in human type 1 diabetes in the isolated founder population of Sardinia. *Diabetes* **51**, 3573-3576 (2002).

23. L. Pireddu, S. Leo, G. Zanetti, MapReducing a genomic sequencing workflow. In *Proceedings of the 20th ACM International Symposium on High Performance Distributed Computing*, pp. 67–74 (2011).

24. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* **26**(5), 589-95 (2010).

25. A. McKenna, *et al.*, The Genome Analysis Toolkit, a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

26. H. Li, *et al.*, The Sequence alignment/map SAM format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

27. Y Chromosome Consortium, A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* **12**, 339-348 (2002).

28. A.J. Vilella, *et al.*, EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**:327-335 (2009).

29. J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, (2005).

30. A. Rambaut,. Fig.Tree. Tree Figure Drawing Tool, Version 1.4.0 (2006-2012). Available from: http://tree.bio.ed.ac.uk/software/figtree/.

31. Sondaar P. et al.,The human colonization of Sardinia, a Late-Pleistocene human fossil from Corbeddu cave. Comp. Rend. Acad. Sci. Paris **320**, 145-150 (1995).

32. Tanda G. Il neolitico antico in Sardegna. In: Blasco Ferrer et al. (Eds) *Iberia e Sardegna, Legami linguistici, archeologici e genetici dal Mesolitico all'Età del Bronzo*. Le Monnier, Firenze, pp.234-249 (2013 in press).

33. Aimar A., Giacobini G. and Tozzi C. Trinità d'Agultu (Sassari). Località Porto Leccio. *Bd. Archaeol*. **45**, 83-87 (1997).

34. Lugliè C. Il Neolitico antico. In: Lugliè C. And Cicilloni R. (Eds) *Atti della XLIV riunione scientifica – La preistoria e la prototoria della Sardegna, Vol. 1.* Istituto Italiano di Preistoria e Prototoria, Firenze, pp. 37-47 (2009)

35. Tykot R.H. Characterization of the Monte Arci (Sardinia) Obsidian Sources. *J. Archaeol. Sci.* **24**, 467-479 (1997).

36. Usai L. Il Neolitico medio. In: Lugliè C. And Cicilloni R. (Eds*) Atti della XLIV riunione scientifica – La preistoria e la prototoria della Sardegna, Vol. 1*. Istituto Italiano di Preistoria e Prototoria, Firenze, pp. 49-58 (2009)

37. Dyson S.L. and Rowland R.J. Jr. *Shepherds, Sailors & Conquerors. Archeology and History in Sardinia from the Stone Age to the Middle Ages.* University of Pennsylvania Museum of Archaeology and Anthropology, Philadelphia, PA (2007).

38. Lilliu G. Prima dei nuraghi. In: AAVV, *La società in Sardegna nei secoli*. ERI, Torino, pp.7-21 (1967).

SUPPLEMENTARY MATERIALS

Materials and Methods

Samples

DNA was extracted from peripheral blood samples of 1,204 adult males from various areas of Sardinia. 873 were unrelated individuals, whereas 331 were paternally related (154 father-son pairs, 5 families of 3 individuals and 2 of 4 individuals). Because of the non-random nature of the sample, the data were used here for phylogenetic purposes (considering males within a nuclear family as a single Y chromosome sample), and no population analysis at a sub-regional level was done. The individuals sequenced were part of two large clinical studies: the SardiNIA study, which assesses quantitative traits of biomedical relevance a general population cohort (21), and a case-control study to detect the genetic factors of risk for autoimmune diseases with no correlation with the Y chromosome (22). An additional 4 individuals from different European regions (1 Basque, 1 Northern Italian, 1 Corsican, 1 Tuscan) of known haplogroup were included to provide the tree with points of divergence from continental populations. We also considered for comparison 133 publicly available European sequences (40 Britons, 8 Iberians, 50 Tuscans and 35 Finns) from the 1,000 Genomes database (http://www.1000genomes.org) together with the sequence of the so-called Iceman Ötzi (9). The present study was approved by the Ethical committees of the ASL 6 in Lanusei and ASL 1 in Sassari, and by the IRB of the National Institute on Aging, NIH. Each participant signed an informed consent form for the samples used.

Sequencing method and analysis

Genomic DNA-Seq Pair-End libraries were generated from 3-5μg of genomic DNA using the Paired-End Genomic Sample Prep Kit (Illumina) according to the manufacturer's instructions. Samples were sequenced on the Illumina Genome Analyzer IIx and Illumina Hiseq2000

instruments at the CRS4 Sequencing and Genotyping Platform Center in Pula, and at the University of Michigan DNA Sequencing Core. Image analysis, base calling and quality scoring were carried out using the Illumina analysis pipeline. Genome libraries were sequenced at an average whole-genome coverage of 4.32-fold. Raw sequencing data were aligned using Seal-0.3.1 (23), a distributed alignment tool based on BWA 0.5.9 (Burrows Wheeler Alignment tool) (24) and the human NCBI GRCh37-Decoy reference assembly of the Genome Reference consortium (1).

Quality scores were recalibrated using the GATK toolkit version 1.2-24 (25). Base alignment quality correction was applied using the SAMtools software (26) assessing alignment quality to avoid false SNP calls due to misalignments. A last pre-processing step was performed using the bamUtils' ClipOverlap tool, which solves the problem created by overlapping reads (i.e., reads whose insert size is shorter than twice the read length, which can result in double counting of PCR errors) by cutting the overlapping part of the read with lower quality.

The sequence coverage and mappability of the MSY regions analysed are shown in fig. S1.

Variant calling

Sequence alignments were used to call genotypes, analysed using a customized version of the glfMultiples tool, specifically modified to call MSY genotypes, which calibrates genotype posterior probabilities using a prior probability of mismatch of 10^{-3} based on the hypothesis of a mutation every 1000 bases. To increase the statistical power to detect variants in low pass sequencing, we used a population-based approach to call all the samples in a single batch. In particular the modified version of glfMultiples contains the following filters: 1) bases were selected only if mapping quality (expressed on the Phred scale) was >=60 for paired-end reads and >=37 for single-end reads; 2) Genotypes were retained if they passed filters and prior correction did not change the resulting genotype call; 3) singleton polymorphisms were preliminarily accepted (but not included in the analyses) only if the non-reference allele was observed at least four times; and 4) heterozygous calls

were imposed to be "no-calls" and positions accounting for more than 20 samples with heterozygous calls were discarded in all samples. This analytic approach to the sequencing data focuses on base pair substitutions (SNPs) and does not allow the detection of length polymorphisms such as STR and In/Dels. After applying this method to all Sardinians and the additional 5 non-Sardinian males, we observed 14,042 polymorphic sites that occurred in at least two individuals.

Hierarchical inference of the variation

Polymorphic nucleotide positions and respective genotypes were listed and the variants appearing in at least two individuals were assigned according to their association to known haplogroups (groups of haplotypes sharing one or more common ancestral SNPs). The variants univocally associated with a known haplogroup, sub-haplogroup or phylogenetically related haplogroups (with a tolerance margin of 1% of recurrence outside the group), were considered informative (fig. S2). The lack of base calls due to the absence of reads at a position in a particular sample was resolved either as an ancestral or derived allele by a hierarchical inferential approach based on sequential accumulation of mutations over time in the MSY, due to the absence of recombination and the low recurrence and reversion rates on this portion of the Y chromosome (3) (fig. S3). Hence, the allelic status of the discovered SNPs is not always experimentally determined for each sequenced individual, but we report an average of 64.8% directly detected and 35.2% inferred SNPs for each individual. It should be noted that such a phylogenetically based inference strategy is inherently less accurate for the rarest lineages, where the chance is higher of missing a polymorphic site or of proposing a singleton SNP when in fact the variant is shared among different individuals (and is hence lost from the analysis). This is particularly critical in the terminal branches of the various lineages. For instance, in our phylogenetic tree (Fig. 1) all the clades encompassing less than 10 individuals show a branch length below the average, which may affect their coalescence times. This is particularly evident for the branch including a single individual (sample 965, haplogroup Q), with

only 765 SNPs (Fig.1). The polymorphic sites that were discovered in multiple individuals but could not be unequivocally assigned to any of the known haplogroups were discarded (fig. S4). A further selection was applied to sites lacking reciprocal correspondence at the lower hierarchical level of the sub-haplogroup.

The overall process yielded 2,282 (16.2%) uninformative SNPs that were discarded (table S1, sheet 4), resulting in 11,763 validated SNPs available for the analysis, of which 23.1% present in the 1000 Genomes Phase 1 release, 8.9% in the dbSNP135 database, and 580 markers of known haplotype attribution reported in the ISOGG list (table S1, sheet 1). Both newly detected and known SNPs (dbSNP135) showed a similar transition/transversion ratio (1.57 and 1.69 respectively).

In addition to these 11,763 informative SNPs detected in at least two individuals, there were 8,314 detected in only one individual (singletons) (table S1, sheet 2 and 3). Since in this case it is not possible to apply a phylogenetic criterion, the more stringent quality control using a threshold of 4 reads yielded 1,290 singletons (with a transition/transversion ratio of 1.3) that passed this strict filter (table S1, sheet 2), 2.1% of them present in the 1000 Genomes Phase 1 release and 11.0% in the dbSNP database.

Phylogenetic analysis

The reference genome is a chimera of at least two individuals (2). It contains a major portion belonging to haplogroup R defined according to the Y-Chromosome Consortium nomenclature (27), with about 1Mb (from 14.3Mb to 15.3Mb) belonging to haplogroup G. To overcome this confounding factor, we referred to ancestral allelic status rather than to a reference genome allele to describe each SNP. The 9 non continuous portions of the Y chromosome here analyzed span from position 2.6 to 28.8 Mbp for a total of ~8.97 Mbp sequenced [see web resources]. The ancestral status of each position was determined by comparison with a chimpanzee sequence using the LASTZ software as in the Ensembl-Compara pipeline (28) according to the method of Wei et al (5).

The 11,763 phylogenetically informative SNPs were used to build a phylogenetic tree using Phylip v3.69 package (29), using the Pars application (Discrete character parsimony algorithm). FigTree v1.4.0 software was used to display the generated tree (30).

Comparison between sequences analysed at low and high coverage

As discussed above, the low pass sequencing strategy used here can lead to incomplete extraction of the genetic variation. To quantify the amount of variation that has been lost in our sample, and to evaluate the accuracy of the hierarchical phylogenetic-based inference, we sequenced at high coverage (average 16.1x in the MSY regions considered, Minimum 13.7 Maximum 18.4) 6 selected individuals (samples 287, 336, 338, 688, 699 and 915), 4 belonging to the I2a1a-δ clade used for calibration, 1 belonging to the I2a1a-β clade, and 1 to the haplogroups J, used as outgroup (table S2).

Having accurate deep sequencing data we are able to estimate the concordance rate of phylogenetic inference. The low pass sequencing followed by hierarchical inference yielded, as a whole for the six individuals, 4,409 genotypes alternative to the reference (NCBI GRCh37), 1,710 of them directly genotyped and 2,699 inferred. The inferred genotypes were concordant with the deep sequence genotyping in 2,695 cases, while in 4 cases the inferred genotype was different from that observed at high coverage (error rate of ~0.001), confirming the accuracy of the inference procedure. It is worthy to notice that the incorrect inference does not result in a wrong haplogroup assignment but a shift of the inferred SNP one node upstream on the phylogenetic tree.

We then compared the tree constructed for these six individuals at both low and high coverage, in order to reveal how the new informative variability is distributed. The deep sequencing discovered 103 additional SNPs not previously observed at low pass. As expected, the variants detected only by deep sequencing were not evenly distributed along the tree (fig. S5). In fact, most of the

common variants in the upper part of the tree were detected by our low pass, population based approach, while the false negative rate is highest in the terminal branches and thus for rare variants.

The average number of SNPs for the four I2a1a-δ samples (downstream the calibration point dated at 7,700 years) is 45.7 (±2.2), which is about 20% higher than the average detected in the general low pass samples. This result corresponds to one new mutation every 168.5 years, equivalent to a phylogenetic rate of 0.65×10^{-9} bp^{-1} $year^{-1}$.

When considering the two other individuals outside the clade used for calibration, the coalescent number of SNPs increases to 1031.8 (±10.9). Using this data with the corrected rate of 1 SNP x 168.5 years, the TMRCA of the six samples results to be 174,000 years old, ~13% younger with respect to the one calculated for the general low pass sequenced samples.

However, this result must be regarded as preliminary, as some of the variability at the root above the IJ node cannot be detected with this analysis, being identical to the reference. Hence, the TMRCA calculated in the deep sequenced samples could be even closer to that we reported for the general low pass sequenced samples.

Ancient DNA analysis

The reliability of the phylogenetic rate was assessed using a sequence coming from an ancient mummified sample (Ötzi) of known age, radiocarbon dated to 5,300 years old (9). This sequence belongs to a clade of haplogroup G encompassing a Tuscan, a Corsican and 8 Sardinian samples (table S3). It shares with the other samples 10 derived SNPs, while it is not possible to assign the allelic status to other 17 SNPs lacking of a readable signal in the ancient sample. Another 8 SNPs, showing the ancestral allele, separate the Ötzi sequences from the Tuscan, Corsican and Sardinian cluster, which presents the derived allele. Subsequently, the Tuscan sequence separates from the Sardo-Corsican ones. In addition, the Ötzi sequence shows 15 singletons, although this number is likely underestimated, considering the poor coverage of ancient DNA. Applying our phylogenetic

rate (one mutation every about 200 years), the date for the MRCA between Ötzi and the Sardo-Corsican samples ranges from about 9,000-12,500 years ago (depending on the assignment of the 17 SNPs of unknown status) (fig. S6). The number of Ötzi's singletons indicates a separate evolution of at least 3,000 years, resulting in an estimated age of 6,000 years at its lower limit, approaching the actual age of the specimen. However, this estimate is older and thus does not conflict with our phylogenetic rate, whereas a younger age would have contradicted it.

Archaeological context

The long history of human settlement in Sardinia is illustrated by the known Mesolithic to Late Neolithic period archaeological sites (Figures S7). The first direct evidence of modern Homo sapiens in Sardinia dates back to upper Paleolithic, with a human phalanx discovered at Corbeddu cave (Central Eastern Sardinia) and radiocarbon dated to ~20,000 years ago (BP) (31). However, based on the available evidence, this early peopling remained isolated. Sardinia began to be substantially inhabited during the Mesolithic (10,500-8,000 years BP), with 6 known settlements (fig. S7a) (32). The relatively limited food resources dominated by a "pika-like" rabbit named Prolagus sardus and mollusks hampered a significant demographic growth by the local hunter-gatherers whose locations remained largely restricted to the coast (33).

The introduction of a productive economy with farming and domesticated animals started in the Early Neolithic, ~7,700 years BP (fig S8) (12, 34), and resulted in an initial population expansion at 7,700 years BP, highlighted by 73 known archaeological sites, 45 of them at open air and 28 in caves or rock shelters (fig. S7b) (33, 34). These sites are found on the coast as well as in the interior. Cardial impressed pottery was developed, and the obsidian trade witnessed contacts with the northwestern shores of the Mediterranean (35). This population evolved into the Middle Neolithic culture of "Bonu Ighinu", with 76 sites known (fig. S7c), characterized by more developed open-air settlements, although caves continued to be used (36). By ~5,200 years BP, the

Late Neolithic “Ozieri” culture, has afforded about 127 known sites (fig. S7d) reflecting a remarkable increase of the population size and internal movements and settling of new territories by the local population (37), although with some contacts with external populations, as suggested by cultural exchanges with the Aegean area (38). The settlements are mainly in newly founded villages, and is characterized by megalithic tombs (“*domus de janas*").

The Eneolithic and Bronze Age periods (4,800-2,900 years BP) culminated in the rich Nuragic culture, named for the imposing stone fortress towers (“*Nuraghe”*).

WEB RESOURCES

The URL for data presented herein is as follows:

1000 Genomes Project,

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/uniq

ISOGG 2012 Y-DNA Haplogroup Tree, http://www.isogg.org, version 8.26, 10 March 2013.

bamUtils' ClipOverlap tool, http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap.

The human NCBI GRCh37-Decoy reference assembly of the Genome Reference consortium, http://www.ncbi.nlm.nih.gov/projects/ genome/assembly/grc/human/index.shtml.

glfMultiples tool, http://genome.sph.umich.edu/wiki/GlfMultiples

DISCLOSURE DECLARATION The Authors declare to have any conflicts of interest

AUTHOR CONTRIBUTIONS

PF, FC designed the study; PF designed and performed the hierarchical analysis for the variant filtering; LM, DSa, AU prepared data for hierarchical analysis; LM, PF performed the TMRCA

analysis; FC, DSc, GRA, PF, SU, CJ provided funding; MZ, MP, FD, AM, MD, SL, FP, FB, AM, MFU, AA, MM, RC, MO, RP, RL, BT, JBG performed sample selection, DNA preparation and sequencing experiments; RB developed analysis programs; SA, ST provided and genotyped non-Sardinian DNA samples; RB, FR, CS, IZ, HMK, BL, EP, RA, SS analyzed sequencing data; PF, FC wrote the manuscript with critical revisions provided by DS, MBW, LM, AA, SS, RB, GRA, SU, CJ and CS.

FIGURE LEGENDS

Figure 1

Phylogenetic tree of the 1,209 (1204 Sardinians and 5 non Sardinians) Y chromosome sequences.

The bifurcations AT, BT, CT, and DE have been inferred due to the absence of individuals belonging to haplogroups B, C and D in our sample. Colored branches represent different Y chromosome haplotypes.

Number of polymorphisms for the main branches is shown in black, while the average number of SNPs of sub-haplogroups is given in blue. The sub-haplogroups are named according to ISOGG nomenclature.

The left axis indicates the number of SNPs from the root. The asterisk indicates the calibration points. The colored dots indicates private Sardinian clusters with an average number of SNPs in the range of: 35-40 in red; 25-30 in green; and 7-12 in blue. The black dots indicate clusters with an average number of SNPs in the range 70-120.

The arrow indicates the position on the tree of the Ötzi, Tuscan, and Corsican samples.

The grey box is enlarged in Figure 2.

Figure 2

Phylogenetic tree of the 492 (490 Sardinians and 2 non Sardinians) Y chromosome sequences belonging to haplogroup I.

Number of polymorphisms for the main branches is written in black. Average number of SNPs of sub-haplogroups is shown in blue. The sub-haplogroups are named according to ISOGG nomenclature.

The red dots indicate Sardinian private clades, labelled in Greek letters as in Table 1. The black dots indicate clusters with an average numbers of SNPs in the range of 70-120.

The arrows indicate the position of the Northern Italian and Basque samples on the tree. The asterisk indicates the calibration point.

SUPPORTING FIGURES AND TABLES

Figure S1

Physical map of Y chromosome, sequence coverage and mappability of the regions included in our analysis

Upper panel: depth distribution (red, left axis) and fraction of reads with map quality = 0 (blue, right axis)

Mid panel: Number of variants (red, left axis) and fraction of reference bases available (blue, right axis)

Lower panel: Schematic representation of chromosome Y regions

Data are plotted along the portion of Y chromosome spanning from 2.5Mb to 28.8 Mb. Grey shaded rectangles indicate the portion of the Y chromosome provided by the 1000G consortium and used in our analysis. Background colours indicate the three classes of MSY euchromatic sequences: X-transposed (yellow), X-degenerate (grey) and ampliconic (pink), as well as heterochromatic (blue) and pseudoautosomal (dark grey) sequences and other (purple).

Figure S2

Example of Informative SNPs (unequivocally associated with established haplogroups or sub-haplogroups)

- SNP 1: associated with two related haplogroups
- SNP 2-3 and 4-5: associated with root haplogroups
- SNPs 6-8 and 9-10: associated with sub-haplogroups
- SNP 8: associated with a sub-haplogroup, at a lower hierarchical level
- SNP 9: associated with an haplogroup, with derived allele (in red) occurring outside the haplogroup at a frequency < 1%

Figure S3

Examples illustrating the criteria for inferring the allelic status in case of lack of data (blanks)

Samples 3-8 belong to haplogroup X defined by SNPs 1-10, with the main trunk defined by SNPs 1-2.

Haplogroup X is subdivided into sub-haplogroup X1 (samples 3-4), defined by SNPs 6-8, and X2 (samples 5-8), defined by SNPs 9-10.

The blanks were resolved either as ancestral allele (e.g. sample 1, SNP 3 - sample 2, SNPs 6, 7, 10 - and analogous – colored in white) or derived allele (e.g. sample 3, SNP 2 - sample 4, SNP 1 - and analogous – colored in light yellow) according to the phylogenetic context.

The calls in the box related to SNPs 3-5 could be assigned either to the derived (a) or to the ancestral (b) allele. Because of the blanks for both individuals belonging to sub-haplogroup X1, the choice between the two alternative states is arbitrary.

In our dataset, we applied scenario (a) to all cases of ambiguity, placing particular SNPs upstream of the sub-haplogroup bifurcation point.

Future inclusion in the analysis of a new sample with a better coverage could resolve ambiguity (c), assigning the derived allele for SNP 4 (haplo X), and the ancestral allele for SNPs 3 and 5 (sub-haplo X2)

Figure S4

Examples of discarded SNPs not showing unequivocal correspondence with established haplogroups or sub-haplogroups

- SNPs 1-2: high frequency, no association with established haplogroup
- SNP 3: high frequency, associated with one haplogroup but also present in other haplogroups with a frequency > 1%

- SNP 4: high frequency, associated with two unrelated haplogroups
- SNPs 5-6: low frequency, not associated with any haplogroup
- SNP 7: low frequency, associated with one haplogroup, but not associated with any established sub-haplogroup

Figure S5

Reduced phylogenetic tree of the six deep sequenced samples

Numbers in black are variants detected at low pass sequencing. Numbers in blue are new variants detected by deep sequencing. The arrow indicates the calibration point.

Figure S6

Phylogenetic tree of the ancient DNA sample "Ötzi"

The dotted lines indicate the branching and the branch length not supported by direct observation on the ancient sample. Notes as in table S3: 1) SNPs ancestral to all samples in the subclade; 1-2) SNPs whose ancestral status in Ötzi is unknown; 2) SNPs not shared by Ötzi; 3) SNPs shared by the Sardo-Corsican samples; 4) beginning of the private Sardinian SNPs; 5) Private SNPs of Ötzi.

Figure S7

Spatial distribution of known archaeological sites from Mesolithic to Late Neolithic

a) Mesolithic (13,000-7,700 years BP)
b) Early Neolithic (7,700-6,000 years BP)
c) Middle Neolithic (6,000-5,400 years BP)
d) Late Neolithic (5,400-4,800 years BP)

Figure S8

Chronology of Early Neolithic sites

Averages and standard deviations of calibrated radiocarbon dates of Ealry Neolithc levels of relevant Sardinian sites (37).

Supplemental Table 1

List of informative, singleton and discarded SNPs

Sheet 1 – Informative (bi-univocal) SNP list

- Column A: SNP-ID
- Column B: Physical position in build 37
- Column C: Reference allele
- Column D: Alternative allele
- Column E: Ancestral allele
- Column F: x denotes presence in the 1000 Genomes Phase 1 release
- Column G: dbSNP rs-code
- Column H: ISOGG marker code
- Column I: Alternative ISOGG marker code
- Column J: Haplotype assignment
- Column K: Number of observed derived alleles
- Column L: Number of inferred derived alleles
- Column M: Total number of derived alleles (observed + inferred)
- Column N: Percentage observed / total derived alleles
- Column O: First sample with the derived allele

- Column P: Last sample with the derived allele
- Column Q: Non Sardinian samples with the derived allele (O= Ötzi, T= Tuscan, B= Basque, C=Corsican, I=Northern Italian)

Sheet 2 – List of singleton (private) SNPs covered by 4 reads or more

- Column A: SNP-ID
- Column B: Physical position in build 37
- Column C: Reference allele
- Column D: Alternative allele
- Column E: Ancestrale allele
- Column F: x denotes presence in the 1000 Genomes Phase 1 release
- Column G: dbSNP rs-code
- Column H: Individual haplogroup
- Column I: Individual # (O= Ötzi, T= Tuscan, B= Basque, C=Corsican, I=Northern Italian)

Sheet 3 – List of singleton (private) SNPs covered by 2 or 3 reads

- Column A: SNP-ID
- Column B: Physical position in build 37
- Column C: Reference allele
- Column D: Alternative allele
- Column E: Ancestral allele
- Column F: Individual haplogroup
- Column G: Individual # (O= Ötzi, T= Tuscan, B= Basque, C=Corsican, I=Northern Italian)

Sheet 4 – Non-Informative (discarded) SNP list

- Column A: SNP-ID
- Column B: Physical position in build 37

- Column C: Reference allele
- Column D: Alternative allele
- Column E: Number of observed alternative alleles
- Column F: Other alternative allele (in case of observed triallelic polymorphisms)
- Column G: Number of observed other alternative alleles
- Column H: Number of total observed alternative alleles

Supplemental Table 2

Comparison between sequences analysed at low and high coverage

Sheets 1-6 – SNPs detected, with respect to the reference (NCBI GRCh37), at low and high coverage in samples 287, 386, 388, 688, 699 and 915

- Column A: SNP-ID
- Column B: Physical position in build 37
- Column C: Reference allele
- Column D: Alternative allele
- Column E: Ancestral allele
- Column F: ISOGG marker code
- Column G: Alternative ISOGG marker code
- Column H: Haplotype assignment
- Column I: SNPs at low coverage (upper case = observed; lower case = inferred; boxed = lack of concordance with high coverage)
- Column J: SNPs at high coverage (framed = lack of concordance with inferred)

Sheets 7 – Statistics for samples 287, 386, 388, 688, 699 and 915

Sheets 8 –SNPs discovered at high coverage

- Column A: SNP-ID

- Column B: Physical position in build 37
- Column C: Reference allele
- Column D: Alternative allele
- Column E: Ancestral allele
- Column F: Individual #

Supplemental Table 3

List of SNPs of the Sardo-Corsican clade including the ancient DNA sample "Ötzi"

- Column A: SNP-ID
- Column B: Physical position in build 37
- Column C: Reference allele
- Column D: Alternative allele
- Column E: Ancestral allele
- Column F: ISOGG marker code
- Column G: Individual # (O= Ötzi, T= Tuscan, C=Corsican, n.a.= not available)
- Column H: Notes as in fig S6: 1) SNPs ancestral to all samples in the subclade; 1-2) SNPs whose ancestral status in Ötzi is unknown; 2) SNPs not shared by Ötzi; 3) SNPs shared by the Sardo-Corsican samples; 4) beginning of the private Sardinian SNPs; 5) Private SNPs of Ötzi

TABLE LEGENDS

Table 1

Super-haplogroups, haplogroups, sub-haplogroups and private Sardinian-Corsican clades.

Here the average number of SNPs defining each class are shown in our 1212 samples. * The average number of SNPs for haplogroups A and E cannot be determined with precision because of the lack in our sample of individuals belonging to haplogroups B, C, and D. Consequently, the number here reported is an overestimate.

The Sardinian samples are progressively numbered from 1 to 1,204, and the non Sardinian samples are labelled as follows: O= Ötzi, T= Tuscan, B= Basque, C=Corsican, I=Northern Italian.

The clades containing only private Sardinian SNPs are indicated in Greek letters (progressively from α to δ within each haplogroup).

Super-haplogroup (individual no.)	Mean SNPs	Haplogroup (individual no.)	Mean SNPs	Sub-haplogroup (individual no.)	Mean SNPs	Private Sardinian clade (individual no.)	Mean SNPs
A-R (1–1204; OTCBI)	1002.6	A (1–7)	879.9*	A1b1b2b (1–7)	11.9	α (1–7)	11.9
				E1a1 (8–13)	7.0	α (8–13)	7.0
				E1b1b1a1 (14–45)	87.4		
E-R (8–1204; OTCBI)		E (8–139)	541.8*	E1b1b1b1 (46–115)	96.1	β (49–115)	15.6
				E1b1b1b2 (116–139)	114.9	γ (116–131)	25.8
		F (140–146)	299.0	F3 (140–146)	79.3		
				G2a2b (147–186; OTC)	109.5	α (C; 155–162)	42.8
F-R (140–1204; OTCBI)	534.8	G (147–277; OTC)	373.8			β (163–186)	29.4
				G2a3 (187–277)	120.3	γ (247–277)	25.0
				I1a3a2 (278–279)	0.0		
				I2a1a (280–744; BI)	106.2	α (280–285)	36.0
						β (286–296)	39.1
						γ (297–314)	34.1
		I (278–767; BI)	353.5			δ (315–744)	37.3
I-J (278–928; BI)	387.0			I2a1b (745–746)	0.0		
				I2a2a (747–756)	38.1		
				I2c (757–767)	72.2		
		J (768–928)	334.3	J1c (768–830)	112.7	α (816–830)	11.0
				J2a (831–905)	125.1		
				J2b (906–928)	91.9		
K-R (929–1204)	375.3	K (929–964)	324.9	L (929–936)	123.7		
				T (937–964)	101.3		
		P (965–1204)	359.1	Q1a3c (965)	0.0		
				R1a1a1 (966–980)	13.8		
				R1b1a2 (981–1165)	37.2	α (981–989)	23.0
		R (966–1204)	241.2			β (991–1165)	29.4
				R1b1c (1166–1194)	75.7	γ (1177–1194)	36.2
				R2a1 (1195–1204)	8.5	δ (1195–1204)	8.5

Figure 1

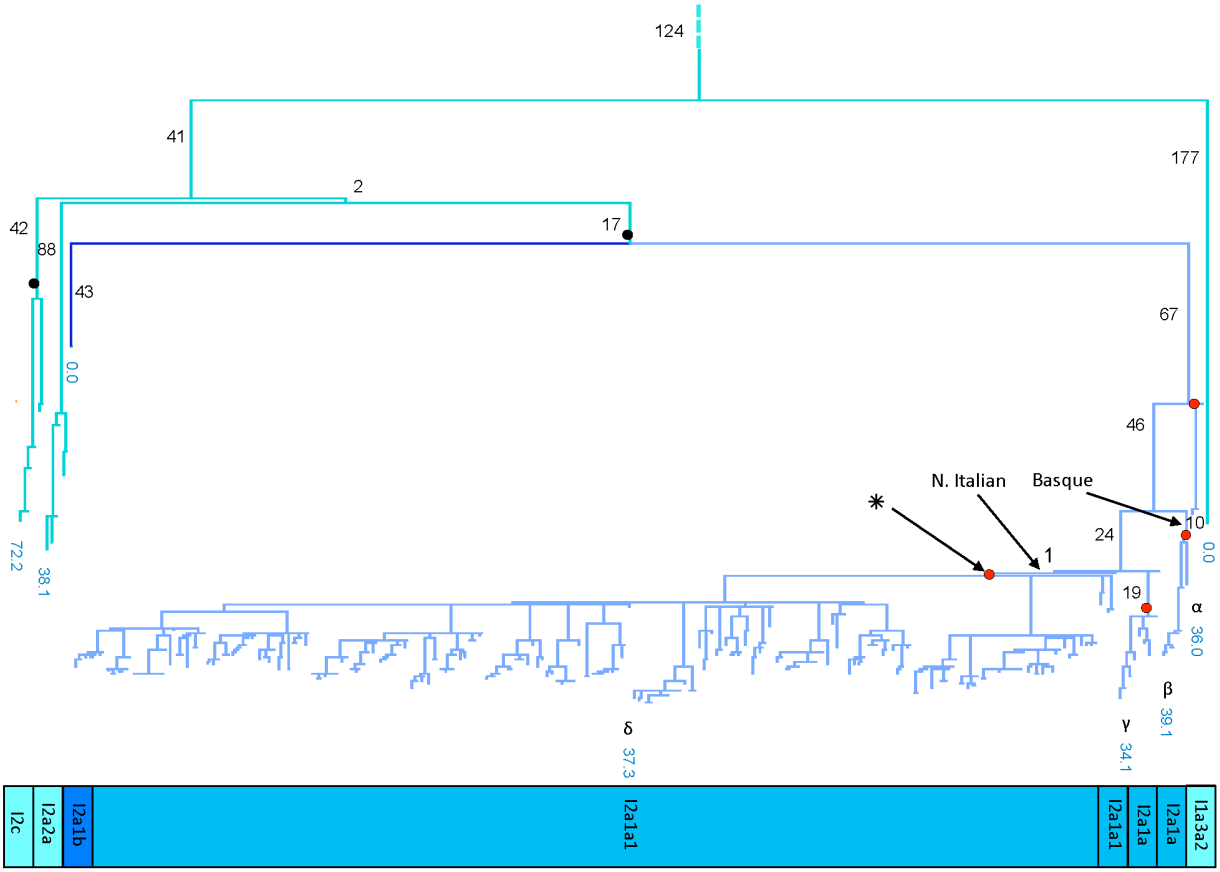

124
41
177
2
17
42
88
43
67
0.0
46
N. Italian
Basque
10
24
1
19
0.0
72.2
38.1
α
36.0
β
39.1
δ
37.3
γ
34.1
I2c
I2a2a
I2a1b
I2a1a1
I2a1a1
I2a1a
I2a1a
I1a3a2

Figure 2

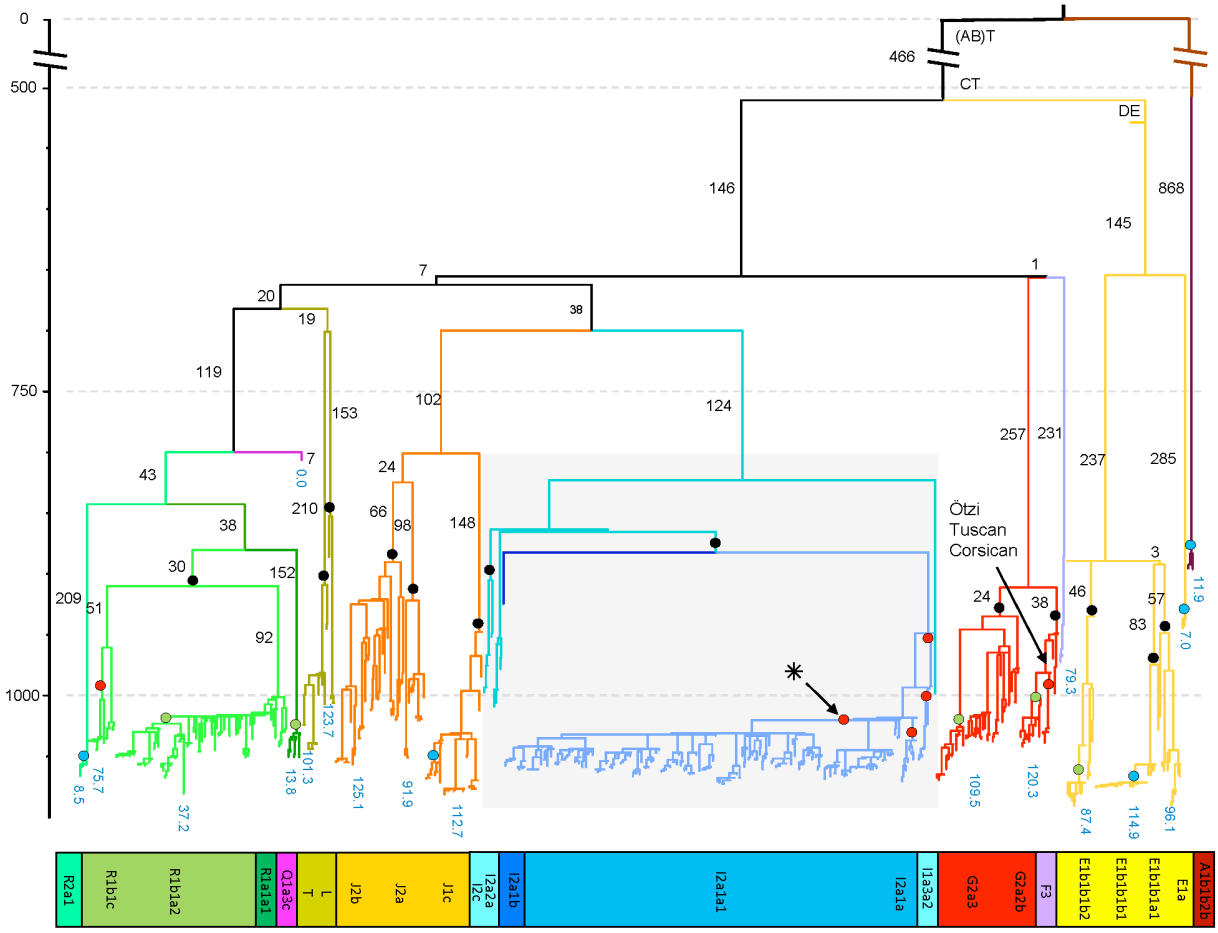

0
500
750
1000
(AB)T
CT
DE
466
146
868
145
1
7
20
19
38
119
153
102
124
257
231
237
285
43
7
0.0
24
210
66
98
148
Ötzi
Tuscan
Corsican
3
30
38
152
24
38
46
57
209
51
92
83
79.3
7.0
11.9
75.7
8.5
37.2
101.3
13.8
123.7
125.1
91.9
112.7
109.5
120.3
87.4
114.9
96.1
R2a1
R1b1c
R1b1a2
R1a1a1
Q1a3c
T
L
J2b
J2a
J1c
I2a2a
I2c
I2a1b
I2a1a1
I2a1a
I1a3a2
G2a3
G2a2b
F3
E1b1b1b2
E1b1b1b1
E1b1b1a1
E1a
A1b1b2b

Figure S1

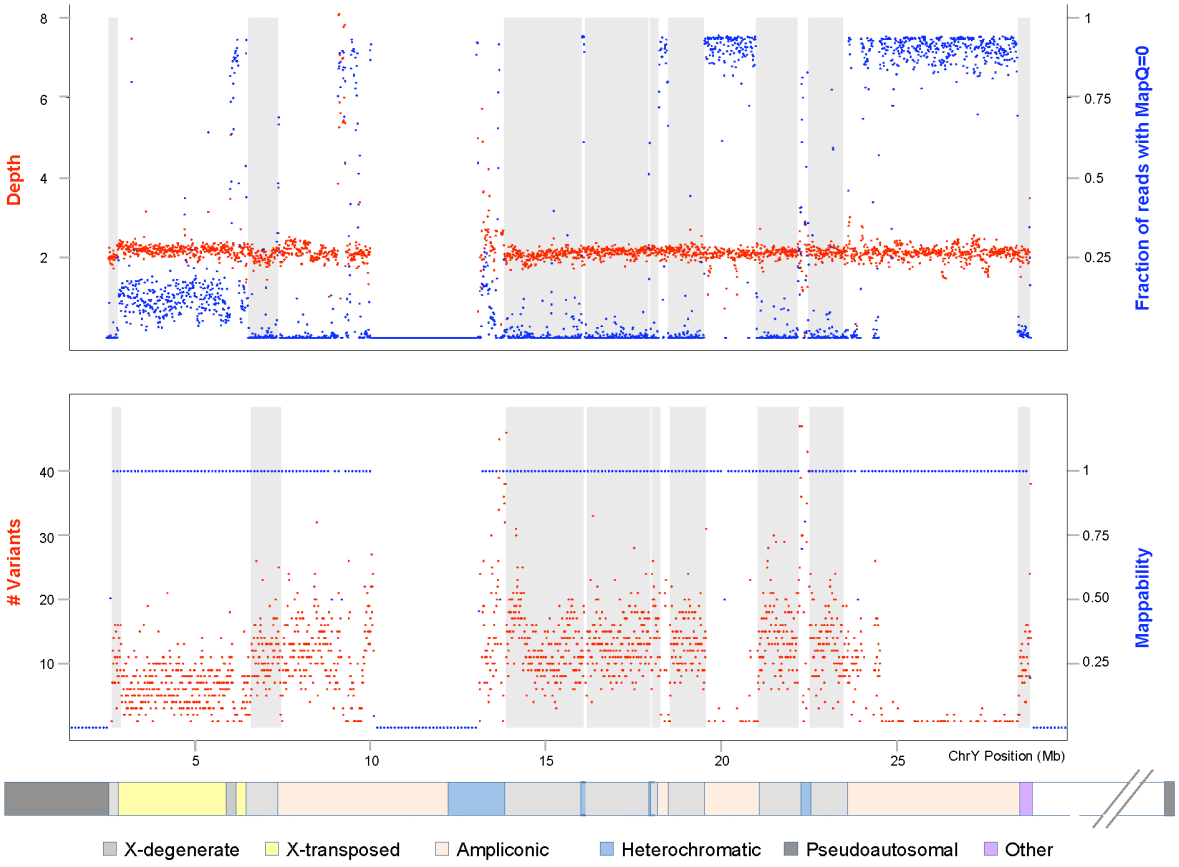

Depth
8
6
4
2
1
0.75
0.5
0.25
Fraction of reads with MapQ=0
Variants
40
30
20
10
1
0.75
0.50
0.25
Mappability
5
10
15
20
25
ChrY Position (Mb)
X-degenerate
X-transposed
Ampliconic
Heterochromatic
Pseudoautosomal
Other

Figure S2

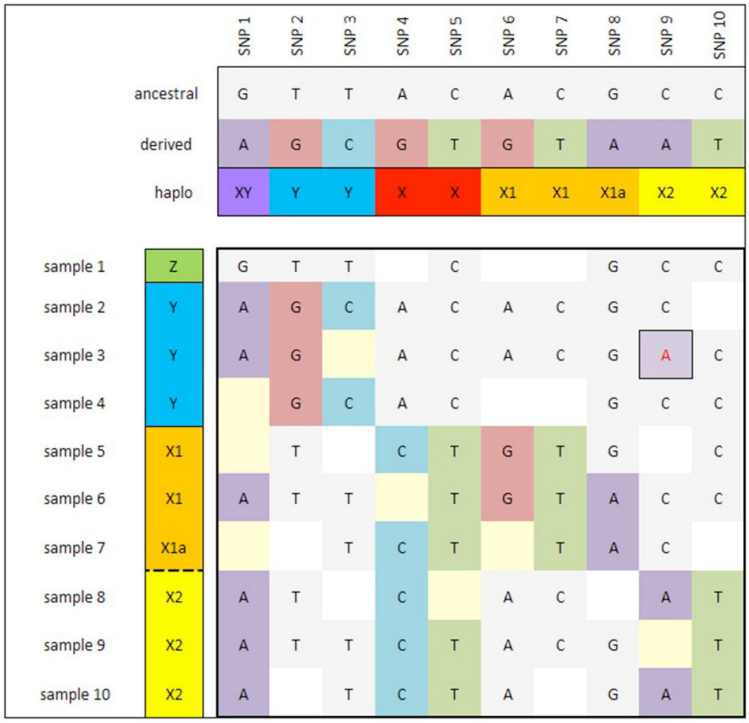

SNP 1
SNP 2
SNP 3
SNP 4
SNP 5
SNP 6
SNP 7
SNP 8
SNP 9
SNP 10
ancestral G T T A C A C G C C
derived A G C G T G T A A T
haplo XY Y Y X X X1 X1 X1a X2 X2
sample 1 Z G T T C G C C
sample 2 Y A G C A C A C G C
sample 3 Y A G A C A C G A C
sample 4 Y G C A C G C C
sample 5 X1 T C T G T G C
sample 6 X1 A T T T G T A C C
sample 7 X1a T C T T A C
sample 8 X2 A T C A C A T
sample 9 X2 A T T C T A C G T
sample 10 X2 A T C T A G A T

Figure S3

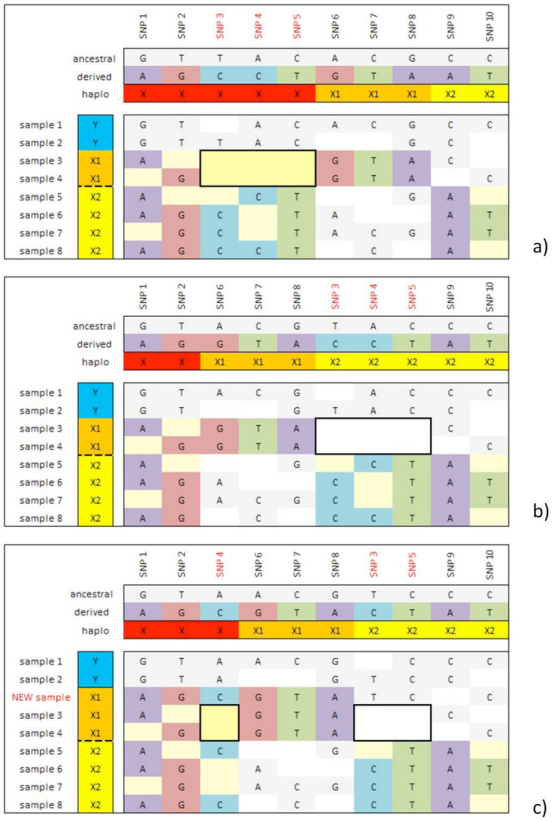
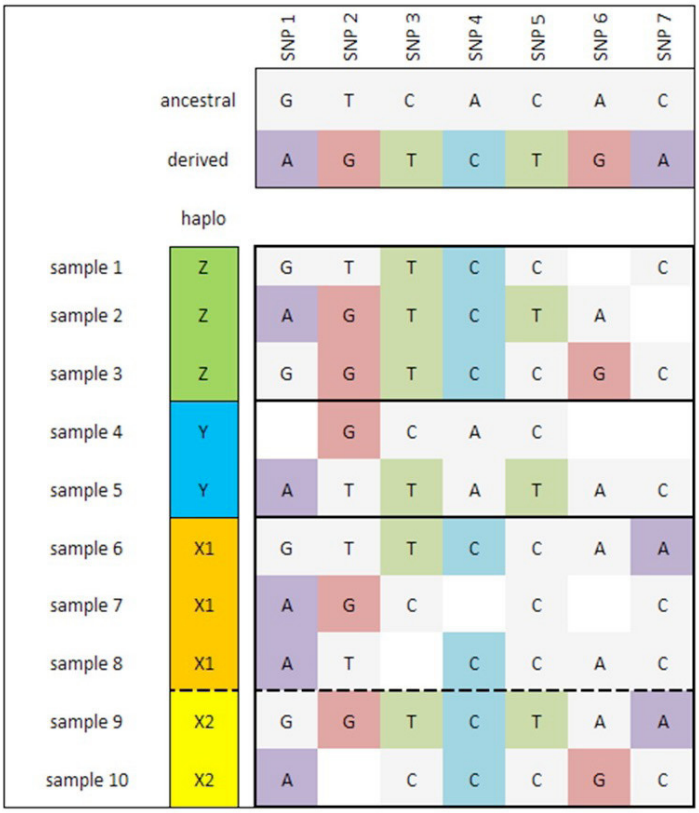

SNP 1 SNP 2 SNP 3 SNP 4 SNP5 SNP6 SNP 7 SNP8 SNP 9 SNP 10
ancestral G T T A C A C G C C
derived A G C C T G T A A T
haplo X X X X X X1 X1 X1 X2 X2
sample 1 Y G T A C A C G C C
sample 2 Y G T T A C G C
sample 3 X1 A G T A C
sample 4 X1 G G T A C
sample 5 X2 A C T G A
sample 6 X2 A G C T A A T
sample 7 X2 G C T A C G A T
sample 8 X2 A G C C T C A
a)
SNP 1 SNP 2 SNP6 SNP 7 SNP 8 SNP 3 SNP 4 SNP5 SNP 9 SNP 10
ancestral G T A C G T A C C C
derived A G G T A C C T A T
haplo X X X1 X1 X1 X2 X2 X2 X2 X2
sample 1 Y G T A C G A C C C
sample 2 Y G T G T A C C
sample 3 X1 A G T A C
sample 4 X1 G G T A C
sample 5 X2 A G C T A
sample 6 X2 A G A C T A T
sample 7 X2 G A C G C T A T
sample 8 X2 A G C C C T A
b)
SNP 1 SNP 2 SNP 4 SNP 6 SNP 7 SNP 8 SNP 3 SNP5 SNP 9 SNP 10
ancestral G T A A C G T C C C
derived A G C G T A C T A T
haplo X X X X1 X1 X1 X2 X2 X2 X2
sample 1 Y G T A A C G C C C
sample 2 Y G T A G T C C
NEW sample X1 A G C G T A T C C
sample 3 X1 A G T A C
sample 4 X1 G G T A C
sample 5 X2 A C G T A
sample 6 X2 A G A C T A T
sample 7 X2 G A C G C T A T
sample 8 X2 A G C C C T A
c)

Figure S4



		SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
	ancestral	G	T	C	A	C	A	C
	derived	A	G	T	C	T	G	A
	haplo							
sample 1	Z	G	T	T	C	C		C
sample 2	Z	A	G	T	C	T	A	
sample 3	Z	G	G	T	C	C	G	C
sample 4	Y		G	C	A	C		
sample 5	Y	A	T	T	A	T	A	C
sample 6	X1	G	T	T	C	C	A	A
sample 7	X1	A	G	C		C		C
sample 8	X1	A	T		C	C	A	C
sample 9	X2	G	G	T	C	T	A	A
sample 10	X2	A		C	C	C	G	C

Figure S5

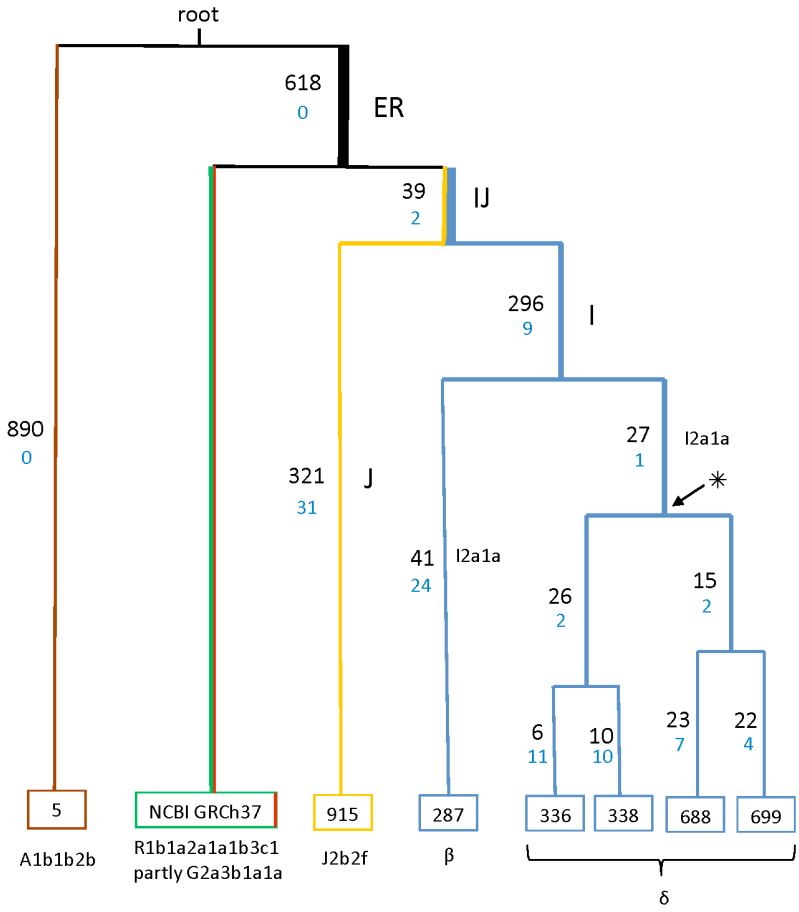

root
618
0
ER
39
2
IJ
296
9
I
890
0
27
1
I2a1a
321
31
J
41
24
I2a1a
26
2
15
2
6
11
10
10
23
7
22
4
5
NCBI GRCh37
915
287
336
338
688
699
A1b1b2b
R1b1a2a1a1b3c1
partly G2a3b1a1a
J2b2f
β
δ

Figure S6

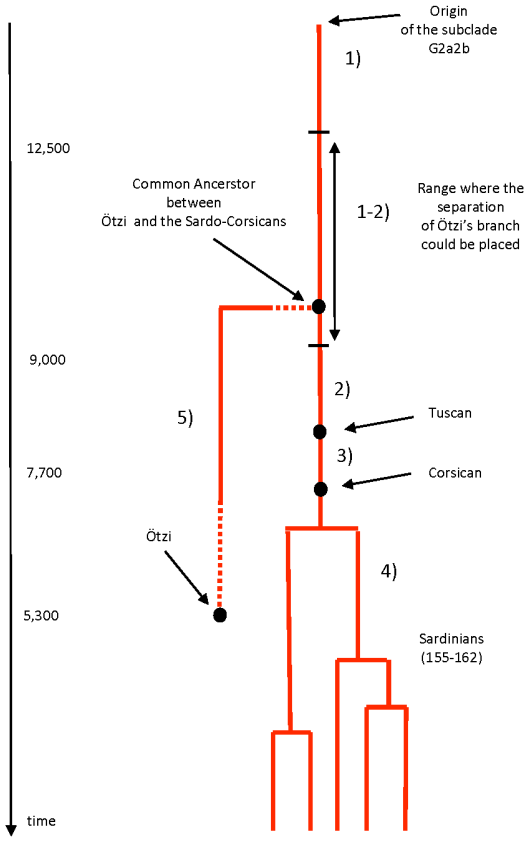

Origin
of the subclade
G2a2b
1)
12,500
Common Ancerstor
between
Ötzi and the Sardo-Corsicans
1-2)
Range where the
separation
of Ötzi's branch
could be placed
9,000
2)
5)
Tuscan
3)
7,700
Corsican
Ötzi
4)
5,300
Sardinians
(155-162)
time

Figure S7

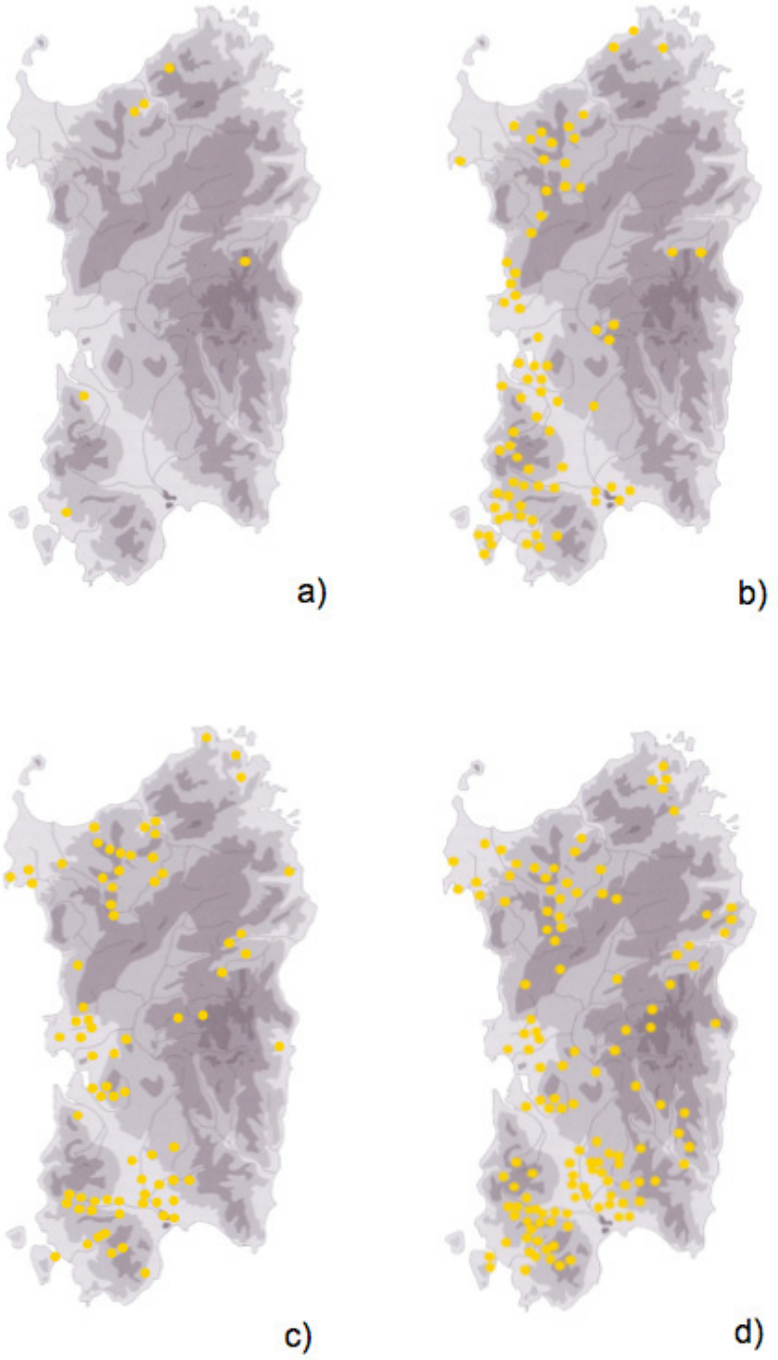
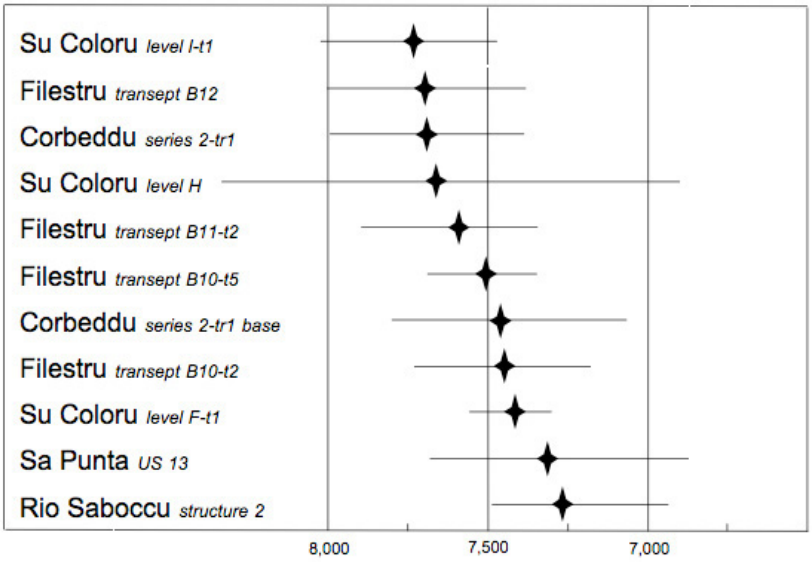


Figure S8



From
European Y-Chromosome Phylogeny Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs.
Paolo Francalacci, Laura Morelli, Andrea Angius, Riccardo Berutti, Frederic Reinier, Rossano Atzeni, Rosella Pilu, Fabio Busonero, Andrea Maschio, Ilenia Zara, Daria Sanna, Antonella Useli, Maria Francesca Urru, Marco Marcelli, Roberto Cusano, Manuela Oppo, Magdalena Zoledziewska, Maristella Pitzalis, Francesca Deidda, Eleonora Porcu, Fausto Poddie, Hyun Min Kang, Robert Lyons, Brendan Tarrier, Jennifer Bragg Gresham, Bingshan Li, Sergio Tofanelli, Santos Alonso, Mariano Dei, Sandra Lai, Antonella Mulas, Michael B. Whalen, Sergio Uzzau, Chris Jones, David Schlessinger, Gonçalo R. Abecasis, Serena Sanna, Carlo Sidore, Francesco Cucca
Science **341**, 565 (2013); Reprinted with permission from AAAS.
Final version of the article can be found on
http://www.sciencemag.org/content/341/6145/565.full
Supplementary Materials
www.sciencemag.org/cgi/content/full/341/6145/565/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S3
References (20–37)

4.2. Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata.

OPEN ACCESS Freely available online

PLOS ONE

Uniparental Markers in Italy Reveal a Sex-Biased Genetic Structure and Different Historical Strata

Alessio Boattini[1☯], Begoña Martinez-Cruz[2☯], Stefania Sarno[1], Christine Harmant[3,4], Antonella Useli[5], Paula Sanz[2], Daniele Yang-Yao[1], Jeremy Manry[3,4], Graziella Ciani[1], Donata Luiselli[1], Lluis Quintana-Murci[3,4], David Comas[2]* , Davide Pettener[1]* , the Genographic Consortium[¶]

1 Laboratorio di Antropologia Molecolare, Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Università di Bologna, Bologna, Italy, **2** Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Spain, **3** Institut Pasteur, Human Evolutionary Genetics Unit, Department of Genomes and Genetics, Paris, France, **4** Centre National de la Recherche Scientifique, Paris, France, **5** Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, Sassari, Italy

Abstract

Located in the center of the Mediterranean landscape and with an extensive coastal line, the territory of what is today Italy has played an important role in the history of human settlements and movements of Southern Europe and the Mediterranean Basin. Populated since Paleolithic times, the complexity of human movements during the Neolithic, the Metal Ages and the most recent history of the two last millennia (involving the overlapping of different cultural and demic strata) has shaped the pattern of the modern Italian genetic structure. With the aim of disentangling this pattern and understanding which processes more importantly shaped the distribution of diversity, we have analyzed the uniparentally-inherited markers in ~900 individuals from an extensive sampling across the Italian peninsula, Sardinia and Sicily. Spatial PCAs and DAPCs revealed a sex-biased pattern indicating different demographic histories for males and females. Besides the genetic outlier position of Sardinians, a North West–South East Y-chromosome structure is found in continental Italy. Such structure is in agreement with recent archeological syntheses indicating two independent and parallel processes of Neolithisation. In addition, date estimates pinpoint the importance of the cultural and demographic events during the late Neolithic and Metal Ages. On the other hand, mitochondrial diversity is distributed more homogeneously in agreement with older population events that might be related to the presence of an Italian Refugium during the last glacial period in Europe.

Citation: Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, et al. (2013) Uniparental Markers in Italy Reveal a Sex-Biased Genetic Structure and Different Historical Strata. PLoS ONE 8(5): e65441. doi:10.1371/journal.pone.0065441

Editor: David Caramelli, University of Florence, Italy

Received January 8, 2013; **Accepted** April 24, 2013; **Published** May 29, 2013

Funding: This study was supported by Strategic Project 2006-09 from the University of Bologna to DP and from MIUR PRIN 2007 and 2009 Grants to DP. The project was also supported by the Spanish Government grant CGL2010-14944/BOS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: david.comas@upf.edu (DC); davide.pettener@unibo.it (DP)

☯ These authors contributed equally to this work.

¶ Membership of the Genographic Consortium is provided in the Acknowledgments.

Introduction

Due to its central position and to the extension of its coastal line (~7,460 Km), the modern Republic of Italy e.g. the Italian Peninsula and the two major islands of Sicily and Sardinia has been one of the focal points in the settlement history of Southern Europe and the Mediterranean Basin.

Populated by early modern humans since approximately 30,000 40,000 years before present (YBP) [1] during the LGM (~25,000 YBP) it was involved in the southward contraction of human groups from Central Europe that rapidly retreated to the Mediterranean coastlines, occupying refuge areas, such as in the well-known cases of Iberia and the Balkans [2 5]. After contributing to the substantial re-shaping of the early Paleolithic genetic composition of glacial Refugia, northward re-peopling processes started approximately 16,000 13,000 YBP [3], [6 9].

Subsequently Italy has received the passage of multiple human groups in prehistoric and historic times, acting both as a step point and an area of expansion during the different major migratory events following the early Paleolithic colonization.

The most recent archaeological syntheses [10] describe the early Neolithisation of Italy as the result of two independent and parallel processes, involving respectively the Adriatic and the Tyrrhenian coasts and dating as early as 8,100 YBP (Apulia, South-Eastern Italy) and 7,900 YBP (Liguria, North-Western Italy).

Italian Late Neolithic and the Metal Ages revealed to be a complicated tapestry of different cultural strata, potentially associated with population movements. During the first millennium BC, Italy hosted a vast set of different peoples whose origins in some cases remain unknown (e.g. Etruscans, Ligurians, Veneti), while in other cases are the result of specific migration processes (Celts in North-Western Italy; Greeks in Southern Italy and Sicily) [11].

In addition, independent and/or intersecting subsequent historic events (related with the trade and expansion of different populations in our era: Phoenician, Greek, Carthaginian, Roman,

Arabic and Barbaric) also contributed to the present genetic composition of Italy. Unlikely to have completely deleted precedent genetic structures, such migrations may have resulted in partially overlapping patterns of diffusion within Italy.

At present, only few studies addressed the reconstruction of the genetic structure and history of Italian populations. Barbujani and colleagues (1995), in a study based on mtDNA variability [12], identified a North-South gradient within the peninsula, confirming what was previously revealed by classical genetic markers [13], while underlying the genetic differentiation between Sardinia and the mainland [12]. More recent studies focused only on specific regions of Italy and revealed a homogeneous pattern of distribution for mtDNA haplogroups. These findings point towards a substantial homogeneity of the mtDNA gene pool within the different areas of the Peninsula [14], [15].

On the paternal perspective, Di Giacomo et al. (2003) carried out an investigation of Y-chromosome diversity in continental Italy [16]. They identified a single decreasing North-South major cline within the Peninsula, while local drift and founder effects were invoked to explain the observed distribution of genetic variation. The study was replicated by Capelli et al. (2007) with a much larger set of genetic markers and a more specific sampling strategy [17]. They observed that more than 70% of the detected diversity was distributed along latitude-related gradients. A certain level of discontinuity was suggested between Northern and Southern portions of the Italian peninsula that, according to the authors, may be related to differential Neolithic/Mesolithic contributes in the two regions [17]. These results North-South clinal patterns related to differential Neolithic contributes were largely confirmed in a recent update of the same study adding more populations and including mtDNA information [18]. Some discontinuity between Northern and Southern Italy was apparent also in genome-wide studies at the European geographical scale [19], [20] and in a specific analysis on Italian samples [21].

Although a common north-south cline has been described for maternal and paternal lineages in Italy, recent data on the Neolithisation of southern Europe [22], [23] suggest a sex-biased Neolithic migration that might account for an asymmetrical pattern of structure in Italy. Eventually more recent migrations could have magnified these sex-biased patterns. For example, this seems to be the case for the first Greek groups in Southern Italy and Sicily, reportedly biased towards a low number of females [11]. Such differential sex-specific demographic events could therefore have affected the genetic structure of Italy in a way that might have been ignored in recent whole-genome analyses.

The present research aims to update our knowledge about Italian population genetic history, by increasing the specificity of sampling strategy and the resolution power of uniparental molecular markers. For the first time, we present an extensive study of both mitochondrial DNA and Y-chromosomal variation in the Italian Peninsula, Sicily and Sardinia. Almost 900 individuals from eight sampling macro-areas have been deeply typed for 136 SNPs and 19 STRs of Y-chromosome, as well as for the whole control region and 39 coding SNPs of mtDNA. We use this detailed and complete dataset to address the following issues. First, we seek to describe the genetic structure of Italy and compare it with the patterns obtained before, in order to distinguish between a clinal and a discontinuous pattern of genetic variation. Second, we want to investigate whether the structure observed is sex-biased and which factors could account for any differential contributes from paternal and maternal lineages. Third, we seek to identify which population movements mostly could be in the origin of the current genetic diversity of the Italian populations.

Materials and Methods

Ethics Statement

For all subjects, a written informed consent was obtained, and Ethics Committees at the Universitat Pompeu Fabra of Barcelona (Spain), and at the Azienda Ospedaliero-Universitaria Policlinico S.Orsola-Malpighi of Bologna (Italy), approved all procedures.

Sample collection

A total of 884 unrelated individuals from continental Italy, Sicily and Sardinia were collected according to the following sampling strategy. Firstly, based on the results of a precedent reconstruction of the surname structure of Italy [24], we defined lists of monophyletic surnames for each of the 96 Italian provinces. Secondly, monophyletic surnames frequencies were used to define eight clusters of homogeneous Italian provinces (sampling macro-areas, Figure S1). Within each sampling macro-area, we selected a set of provinces (sampling points) from a minimum of one to a maximum of three, depending on the geographical extension of the macro-area as well as their historical background. This was done in order to depict a sampling grid able to capture as much genetic variability as possible (given the number of planned samples/sampling points). Within each sampling point, individuals were finally sampled according to the standard 'grandparents' criterion, thus considering as eligible for our study only those individuals whose four grandparents were born in the same sampling macro-area. It is important to underline that individuals within sampling points were not selected by surnames. That way 1) our data are consistent with those from other similar studies; 2) we avoid to introduce a bias between Y-chromosome and mtDNA results.

DNA was extracted from fresh blood by a Salting Out modified protocol [25].

Y-chromosome genotyping

A total of 884 samples were successfully typed for Y-chromosome markers. 121 SNPs in the non-recombining region of the Y chromosome were genotyped using the OpenArray® Real-Time PCR System (Applied Biosystems) as described previously [26]. Six additional SNPs (M91, M139, M60, M186, M175, and M17) were genotyped in a single multiplex, Multiplex2 [27]. Nine additional single SNPs (M227, L22, M458, L48, L2, L20, M320, P77) were typed with individual TaqMan assays. Nomenclature of the haplogroups is in accordance with the Y-Chromosome Consortium [28]. Detailed phylogeny may be found at Y-DNA SNP Index - 2009 (http://isogg.org/tree/ISOGG_YDNA_SNP_Index09.html). For simplicity reasons, we will use asterisks (*) to indicate those chromosomes that are derived at a certain SNP, but ancestral at all the tested downstream SNPs.

All individuals were additionally typed for a set of 19 STRs: 17 using the Yfiler kit (Applied Biosystems) and two (DYS388, DYS426) included in the Multiplex2. As the Yfiler kit amplifies DYS385a/b simultaneously avoiding the determination of each of the two alleles (a or b), DYS385a/b were excluded from all the analyses performed. DYS389b was obtained by subtracting DYS389I from DYS389II [29].

Mitochondrial DNA genotyping

865 samples were successfully sequenced for the whole control region as in Behar et al. (2007) [30], and typed using a 22 coding region SNPs multiplex as described previously [27], [31]. Variable positions throughout the control region were determined between positions 16,001 and 573. Sequences were deposited in the GenBank nucleotide database under accession numbers

KC806300-KC807164. In addition, for haplogroup H, the most frequent in Western Europe [2], [6], we used a specifically designed multiplex (named HPLEX17) in order to resolve 17 distinct sub-lineages [27]. Based on combined HVS sequence and coding region SNP data, individuals were assigned to the major haplogroups of the mtDNA phylogeny with the software Haplogrep [32] that uses Phylotree version 13 [33]. Due to their phylogenetic uncertainty, indels at nucleotide positions 309, 315, and 16193 were not taken into account.

Statistical Analyses

Population structure and genetic variability. Haplogroup frequencies were estimated by direct counting. Standard diversity parameters (haplogroup diversity, number of observed STR haplotypes, sequence diversity values, and mean number of pairwise differences) were calculated with Arlequin 3.5 [34]. FST and RST results were corrected with Bonferroni test for multiple comparisons ($p<0.05$).

The relationships between geographical distances and genetic diversity were investigated by using several spatial analyses. The correlation between geographical distances and genetic distances (Reynolds distance), based on haplogroup frequencies, was evaluated by means of a Mantel test (10,000 replications). In order to distinguish any clinal pattern (Isolation-by-Distance pattern) from any discontinuous genetic structure (both of them can result in significant correlations with geography), geographical distances were plotted against genetic ones. A 2-dimensional kernel density estimation layer [35] was added to the plot in order to highlight the presence of discontinuities in the cloud of points. The analysis was performed with all the samples and then removing the Sardinian ones, given their outlier status previously described in literature [7], [13], [21], [36 38].

To further explore spatial patterns of variation a spatial principal component analysis (sPCA) based on haplogroup frequencies was performed using the R software package *adegenet* [39 41]. Additional information about the sPCA method is provided in Methods S1.

To further test the significance of the structure found with the sPCA analysis, we carried out a series of hierarchical analyses of molecular variance (AMOVA) pooling populations according to the sPCA results. We used haplogroup frequencies (both Y-chromosome and mtDNA), RST distances (Y-STRs) and number of pairwise differences (HVRI-HVRII mtDNA sequences). In order to explore genetic variability within the most frequent haplogroups, and in particular within those identified by sPCA loadings, we applied a Discriminant Analysis of Principal Components (DAPC) to Y-STR haplotypes and mtDNA sequences (see Methods S1 for more details). Analyses were performed using the R software *adegenet* package [39 41]. In addition, for comparison purposes we calculated a Network representation of haplogroup G2a using a Median Joining (MJ) algorithm as implemented in the Network 4.6.1.1 software (http://www.fluxus-engineering.com, [42]), weighting STR loci according to the variance method.

DAPC was first performed using Italian haplotypes only. As a second step, in order to investigate the origin of the genetic diversity for the most common haplogroups in Italy, additional individuals from selected European populations were incorporated into the DAPC of major haplogroups. Unpublished 194 Y-chromosome data from Iberia, Germany and the Balkans were provided by the Genographic Project, while data for Causasus and Western Anatolia were extracted from literature [43], [44]. Comparison data for mtDNA was generated using additional information from Basque [45], Austrian [46] and Balkan samples [46], [47].

Y-chromosome and mtDNA dating. In order to minimize the biasing effect of STRs saturation through time (especially important for rapidly evolving STRs as some of those included in the Yfiler kit, [48]), all Y-chromosome age estimations were calculated selecting the eight markers (DYS448, DYS388, DYS392, DYS426, DYS438, DYS390, DYS393, DYS439) with the highest values of duration of linearity D approximated as in Busby et al. (2011) [49].

Splitting time between the sPCA-identified regions (NWI and SEI, see Results) was estimated with BATWING [50] under a model of exponential growth and splitting from a constant size ancestral population. Two samples (Treviso, Foligno/PG) were excluded from the analysis according to a 5% quantile threshold of the sPC1 scores. Two chains with different starting points were run with a total of $3.5\times10[6]$ samples with an initial burn in of $1.5\times10[6]$ samples and a thinning interval of 10×20. The outfiles were treated with the R package [41] to get the posterior distributions of the parameters of interest. We checked that results were equivalent for both runs and reported the mean values of both analyses for every parameter. We used a prior distribution for mutation rates as proposed by Xue et al. (2006) [51] based on Zhivotovsky et al. (2004) [52]. Such distribution is wide enough to encompass all mutation rates for each of the eight considered Y-STRs. A generation time of 25 years was used [52]. Priors and further information about the BATWING procedure are shown in the Methods S1.

The age of Y-chromosome DAPC clusters exhibiting peaks of frequency higher than 70% in any of the sPCA-identified populations (NWI, SEI, and SAR) with the exception of haplogroup G2a due to its particular relevance in our populations (see Results) and composed by at least ten individuals, as well as the age of the entire haplogroups, were estimated with the standard deviation (SD) estimator [53]. Differently from BATWING, this method does not estimate the population split time, but the amount of time needed to evolve the observed STRs variation within haplotype clusters (or whole haplogroups) at each population. As for mutation rates, we adopted locus-specific rates for each of the eight considered loci as estimated by Ballantyne et al. (2010) [48]. These rates were preferred to the 'evolutionary' one [52] for the following reasons: 1) 'germline' rates are locus-specific and based on the direct observation of transmission between father-son pairs; 2) 'germline' rates share the same magnitude with genealogy based estimates [54] while the 'evolutionary' rate is a magnitude lower; 3) a recent study [43] suggested that family based rates (germline, genealogies) provide a better fit with history and linguistics. The 95% confidence intervals of time estimates were calculated based on the standard error (SE). Only individuals with a membership >99% in their corresponding DAPC clusters were considered. Given that moments like mean and variance hence time estimates based on variance are very sensitive to the presence of outliers (e.g. non-robust), we designed a "jackknife-like" procedure in order to detect possible outlier individuals that could be significantly biasing our estimates (see Methods S1 for details).

TMRCA for the most common mtDNA haplogroups was estimated by means of the ρ (rho) statistic with the calculator proposed by Soares et al. (2009) [55] for the entire control region (that considers a mutation rate corrected for purifying selection of one mutation every 9,058 years).

However, results have to be taken with caution, given that molecular date estimates with ρ can be affected by past demography. Simulations show that error rates tend to increase

with effective size, bottleneck and growth effects [56]. In order to avoid sampling errors, the estimates were calculated only for those haplogroups with absolute frequencies higher than 30 individuals.

Results

Y-chromosome lineages in Italy

Haplogroup frequencies. A total of 884 unrelated individuals from 23 Italian locations (Figure S1) were successfully genotyped for 19 STRs and 136 SNPs, and classified in 46 different haplogroups (including sub-lineages) whose phylogeny ([28]; ISOGG Y-DNA SNP Index – 2009) and frequencies for the whole dataset are detailed in Table S1; Y-STR haplotypes of each individual are provided in Table S2.

The haplotype and haplogroup diversity (h), STR diversity (π_n) and mean number of pairwise differences (π) of the population samples are listed in Table S3. The lowest values for haplogroup diversity (h) are observed in Sardinia, while the Italian peninsula is characterized by a negative correlation between haplogroup diversity and latitude, resulting in a south-north decreasing pattern of variation (Spearman's rho $=-0.463$, p-value $=0.036$). The most frequent haplogroups in Italy are R-U152* (12.1%), G-P15 (11.1%), E-V13 (7.8%) and J-M410* (7.6%). They are followed by three R1b-lineages (R-M269*, R-P312* and R-L2*), whose frequencies ranged from 6.9% to 5.7%; and finally from I-M26, which embraced more than the 4% of total variability. On the whole these haplogroups encompass ~62% of Y-chromosomes lineages, while the remaining 38 haplogroups show frequencies lower or equal to 3.3%. Haplogroups distribution in the considered eight sampling areas is detailed in Table S1.

Paternal population structure. In order to explore the relationship between geographical and paternal genetic distances among the 23 investigated Italian populations a Mantel test was performed. A significant correlation was found (observed value $=0.26$, p-value $=0.006$), even after removing Sardinian samples (observed value $=0.19$, p-value $=0.03$). However, a non-homogeneous distribution of points is apparent when plotting geographical distances against genetic ones (Figure S2), indicating that the genetic structure of Italy is better characterised by discontinuities than by clinal patterns.

These general spatial patterns were further explored by means of sPCA based on haplogroup frequencies. The analysis showed that the Italian genetic structure is characterised by two significant global components (positive eigenvalues) with similar variance values, being sPC1 characterized by a higher spatial autocorrelation (Moran's I) (Figure S3). These observations are further assessed by means of a significant Global test (observed value $=0.08$, p-value $=0.015$) and a non-significant Local test (observed value $=0.06$, p-value $=0.677$).

Geographical patterns of sPC1 and sPC2 are plotted in Figure 1. sPC1 identifies two main groups of populations separated by an almost longitudinal line (Figure 1a). The first group (black squares) is represented by populations from North-Western Italy, including most of the Padana plain and Tuscany. The second group (white squares) includes locations from South-Eastern Italy and the whole Adriatic coast, being represented also in North-Eastern Italy. Nonetheless, these two groups are not separated by a sharp discontinuity, but by some sort of gradient, as it is represented by a few samples from North-Eastern and Central Italy that show very low absolute values of sPC1 scores. However, sPC2 scores differentiate Sardinia from the rest of Italy (Figure 1b). Indeed, scores from these populations show the highest absolute values, while those from the other Italian locations (especially in the South) are much lower. In summary, sPC1 and sPC2 depict a

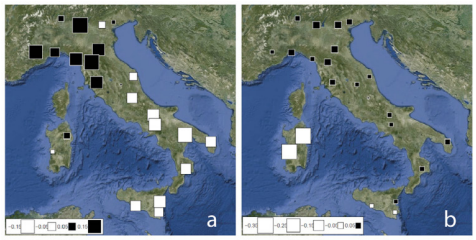


Figure 1. Spatial Principal Component Analysis (sPCA) based on frequencies of Y-chromosome haplogroups. The first two global components, sPC1 (a) and sPC2 (b), are depicted. Positive values are represented by black square; negative values are represented by white squares; the size of the square is proportional to the absolute value of sPC scores.
doi:10.1371/journal.pone.0065441.g001

three-partitioned structure of Italian population: 1) North-Western Italy (from now on NWI), 2) South-Eastern Italy (from now on SEI), and 3) Sardinia (from now on SAR).

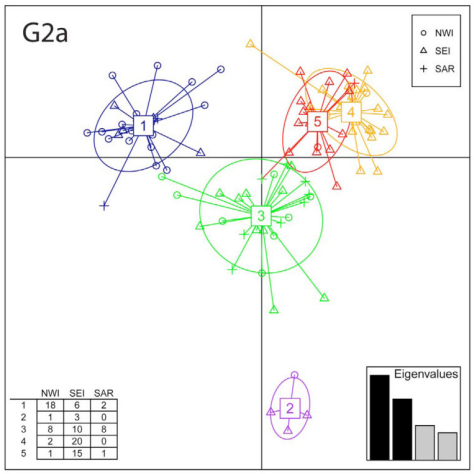
When we tested the reliability of these three groups (NWI, SEI, SAR), by means of AMOVA based both on haplogroup frequencies and STR variability, the proportion of variation between groups (haplogroup frequencies: 3.71%; haplotypes: 4.48%; both p-values <0.001; Table S4) was 1.5 times higher than the variation explained when grouping according to the eight sampling macro-areas (2.62%, p-value <0.001, and 3.11%, p-value <0.001, respectively, Table S4). Interestingly, there is a partial congruence between sPCA-based groups and sampling macro-areas (Figure S1). In particular, SAR coincides with macro-area 8, while macro-areas 1, 3 and 4 are grouped in NWI and macro-areas 6 and 7 are grouped in SWI; macro-areas 2 and 5 are crossed by the sharp gradient that separate NWI from SEI.

To further test the reliability of the mentioned structure, for each of the considered populations we calculated DAPC-based posterior membership probabilities to the considered three groups. Results (Table S5) show that all the populations are characterised by high congruence (membership probability $=\sim 9\%$ or higher) to the given SPCA-group, the only exception being a single population from Central Italy (Foligno/PG), whose intermediate position between NWI and SEI has been already revealed by sPCA.

Interestingly, NWI revealed a high and significant degree of internal differentiation, while SEI is a fairly homogeneous group ($Fst=0.014$, p-value <0.001 and $Fst=0.002$, p-value >0.05, respectively; both estimates are based on haplogroup frequencies).

In order to quantify the contribution of each haplogroup to the genetic structure detected, the loadings values of the sPC1 and sPC2 were calculated and plotted in Figure S4. Lineages contributing more to the differentiation along the first sPC were R-U152*, G-P15 and, with lower loadings values, R-L2* and R-P312* (Figure S4a). On the contrary, sPC2 is influenced primarily and almost exclusively by the haplogroup I-M26 (Figure S4b).

Haplogroup DAPC analysis. DAPC was performed within the most frequent haplogroups (E-V13, G-P15, I-M26, J-M410*, R-P312*, R-U152*, R-L2*). Results (Table 1, Figure 2, Figure S5) show how the seven considered haplogroups disaggregate in 25 clusters, ranging from a minimum of two (I2a-M26) to a maximum of five (E-V13, G2a-P15). Considering a 70% threshold, 13 out of 25 are mostly frequent in one of the sPCA-identified areas (NWI: 7, SEI: 4, SAR: 2) (Table 1).



	NWI	SEI	SAR
1	18	6	2
2	1	3	0
3	8	10	8
4	2	20	0
5	1	15	1

Figure 2. Discriminant Analysis of Principal Components (DAPC) for G2a-P15 haplotypes. Samples are grouped according to their affiliation at the sPCA-identified groups (NWI; SEI; SAR; symbols in the top right table). The table in the bottom left shows the number of haplotypes in each of the five G2a clusters and their geographical distribution in the three Italian areas. DAPC eigenvalues are depicted in the enclosed barplot.
doi:10.1371/journal.pone.0065441.g002

It is noteworthy the structure shown by haplogroup G2a-P15 (Figure 2), which includes clusters with very different spatial distribution: cluster 1 is mostly frequent in NWI, while clusters 4 and 5 – partially overlapping in the DAPC plot – are found in SEI. For comparison purposes, we calculated a Median Joining Network (Figure S6) based on the same haplotypes. While results from both methods are largely overlapping, DAPC offers some advantages compared to the network, namely 1) it outputs clear-cut clusters (while in Network the definition of clusters is in some way arbitrary), 2) it gives probability memberships for each individual. Networks for other haplogroups are not shown.

DAPC comparisons with additional samples (Table S6, Figure S7) suggest differential affinities for some of the considered haplogroups and clusters of haplotypes. Most notably, G2a-P15 haplotypes from NWI cluster mainly with German ones, while haplotypes from SEI seem to indicate wider relationships, going from Iberia to the Balkans and the Caucasus. On the contrary, I2-M26 samples from Sardinia (SAR) cluster in a separate group than Iberians, suggesting a geographical neat separation between continental and Sardinian I2-M26 lineages.

Date estimates for paternal variation. BATWING was used to estimate the age of split between the Italian regions identified by the first sPCA (NWI and SEI, excluding SAR). BATWING modelled population growth starting at 12,890 YBP (95% CI: 3,700–83,070), with a rate of 0.00429 (95% CI: 0.00254–0.01219) per year. Our results suggest that the split happened around 5,490 YBP (95% CI: 1,620–26,830). Since BATWING does not consider migration, admixture between NWI and SEI would likely bias the split time estimate towards more recent dates.

Concerning Y-chromosome lineages, STR variation within the 13 clusters mentioned above suggests that most of them date back to relatively recent times (Table 2). In fact, the ages of the considered clusters (with a peak in one of the considered sPCA groups) fall roughly within the interval from the time of split estimated with BATWING between NWI and SEI and the present. This is consistent with the fact that group-specific clusters of haplotypes (NWI, SEI) are very likely to have emerged after the split within the Italian 'ancestral' population or later. No different patterns of timing are detected between both regions. The time estimates were similar for whole haplogroups with the notable exception of G2-P15, which showed older ages. These results suggest that most of the Y-chromosomal diversity present in modern day Italians was originated from few common ancestors living during late Neolithic times and the Early Metal Ages. However, if we would take into account evolutionary rates, we would observe results three times higher than those above mentioned, meaning that most dates would shift to late Paleolithic.

Mitochondrial DNA lineages in Italy

Haplogroup frequencies. The maternal genetic ancestry of Italian populations was explored by characterizing coding region SNPs and control region sequences from 865 individuals, which yielded to 79 distinct mtDNA haplogroups (including sublineages). Haplogroup frequencies and within-population diversity parameters are shown in Table S7 and Table S3 respectively.

The haplogroup distribution in Italy reflects the typical pattern of mtDNA variability of Western Europe. As described for other European and Italian populations [2], [6], [14], [15], [57] most of the sequences belong to the super-haplogroup H, which includes 44.4% of the Italian mtDNA lineages. In particular, H1 turned out to represent a large proportion of H samples, encompassing the 13.8% of the total variability (10.4% excluding sub-lineages). Compared to H1, sub-haplogroups H3 and H5 represent much smaller fractions of H composition, reaching however noteworthy frequencies (3.9% and 4.3% respectively). Most of the remaining samples belong to haplogroups frequently found in western Eurasia, including U5, K1, J1, J2, T1, T2, and HV. Among the U5 lineages, U5a is the most frequent (3.70%). Haplogroups K1a, HV and J1c take into account respectively the 4.39%, 4.05% and the 3.93% of the total mtDNA variability. The remaining lineages reach frequencies that do not exceed a 3.5% threshold.

Maternal population structure. In contrast to paternal lineages, correlation between geographical and genetic distances was non-significant (Mantel Test: observed value = 0.011, p-value = 0.45). These results point to a strong homogeneity within the Italian Peninsula for the mtDNA gene pool composition. In order to extract further insights into the distribution of mtDNA lineages, a sPCA was performed using haplogroup frequencies. The highest absolute eigenvalues (Figure S8) correspond to the first two positive components (global structure). According to the Global test of significance, the geographical distribution of the genetic variability observed with sPCA was found to be marginally significant (observed value = 0.061, p-value = 0.046).

Scores of the sPC1 and sPC2 are plotted in Figure 3. Both sPC1 and sPC2 highlight the extreme position of Sardinia (large white squares). In addition, sPC1 identifies a North-East centred group that spreads southwards along the Apennines (including most of populations from central Italy), while sPC2 highlights the same East-West pattern observed for Y-chromosome. Loadings of sPC1 and sPC2 (Figure S9) identify lineages H1 and H3 respectively as the haplogroups affecting more the spatial genetic differentiation of Italian populations.

Haplogroup DAPC analysis. DAPC was performed within the eight most frequent haplogroups (H*, H1, H3, H5, HV, J1c, K1a, U5a). They disaggregate in 24 haplotype clusters (Table S8,

Table 1. Frequencies of Y-Chromosome DAPC cluster for each Italian sPCA-identified group.

HG	DAPC CLUSTER	N. HAPLOTYPES				N. INDIVIDUALS				MAX% (GROUP)
		NWI	SEI	SAR	TOT	NWI	SEI	SAR	TOT	
E-V13	1	8	10	1	19	8	10	1	19	53% (SEI)
	2	6	6	0	12	6	6	0	12	50% (NWI, SEI)
	3	3	11	1	15	3	**11**	1	15	**73% (SEI)**
	4	5	6	0	11	5	6	0	11	55% (SEI)
	5	6	6	0	12	6	6	0	12	50% (NWI, SEI)
G2a-P15	1	18	6	2	26	**20**	6	2	28	**71% (NWI)**
	2	1	3	0	4	1	3	0	4	75% (SEI)*
	3	8	10	8	26	8	10	8	26	38% (SEI)
	4	2	20	0	22	2	**20**	0	22	**91% (SEI)**
	5	1	15	1	17	1	**16**	1	18	**89% (SEI)**
I2a-M26	**1**	0	1	18	19	0	1	**19**	20	**95% (SAR)**
	2	2	1	12	15	2	1	**13**	16	**81% (SAR)**
J2a-M410	1	7	9	3	19	7	9	3	19	47% (SEI)
	2	8	18	2	28	8	19	2	29	66% (SEI)
	3	7	11	0	18	7	12	0	19	63% (SEI)
R-P312	**1**	11	**4**	1	16	**12**	**4**	1	17	**71% (NWI)**
	2	13	8	0	21	13	9	0	22	59% (NWI)
	3	6	5	0	11	6	5	0	11	55% (NWI)
R-U152	1	16	7	2	25	16	7	2	25	64% (NWI)
	2	21	1	0	22	**21**	1	0	22	**95% (NWI)**
	3	23	8	2	33	24	10	2	36	67% (NWI)
	4	16	**4**	2	22	**17**	5	2	24	**71% (NWI)**
R-L2	**1**	18	1	1	20	**18**	1	1	20	**90% (NWI)**
	2	18	6	1	25	**18**	6	1	25	**72% (NWI)**
	3	10	**4**	0	14	**10**	**4**	0	14	**71% (NWI)**

*Number of individuals <10
The absolute number of haplotypes and individuals are shown for each DAPC-cluster, and the maximum frequency for each cluster is expressed in percentage (max%).
NWI: North-Western Italy; SEI: Southern and Eastern Italy; SAR: Sardinia.
doi:10.1371/journal.pone.0065441.t001

Figure S10), ranging from a minimum of two (K1a) to a maximum of four (U5a). Most of them are widespread in the whole of Italy, in fact, if we consider a 70% threshold, only nine clusters show traces of geography-related distributions (but six of them are composed by less than 10 individuals). Haplogroup HV is the most important exception, including two clusters located in NWI and SEI, respectively. It is noteworthy a cluster from haplogroup H3 that is almost exclusive of SAR.

Comparisons with other European samples (Table S9, Figure S11) confirm that great part of Italian mtDNA haplotypes share a wide range of affinities spanning from Iberia to Eastern Europe, but haplotypes from H1 and H3 appear to be related mostly with Western and Central Europe.

Date estimates for maternal variation. TMRCA estimates for the most frequent haplogroups (Table 2) could be classified in two groups: "old" haplogroups, predating the Last Glacial Maximum, LGM (~31,600 YBP for HV, ~28,300 YBP for U5a and ~19,500 YBP for J1c), and haplogroups dating after the LGM (~16,200 YBP for H*, ~15,600 YBP for H1, ~15,500 YBP for H3, ~14,700 YBP for H5, ~16,700 YBP for K1a). Estimates for H1 and H3 haplogroups are slightly older than estimates in Western Eurasia for the same haplogroups [2], [4], [5], [55]. These results are in agreement with what has been shown for the Basque region in Iberia [27] and may be related to the length of the mitochondrial region used.

Additionally, we calculated TMRCA for the two DAPC clusters within HV haplogroup (2 and 3), given that they show a clear spatial polarity within continental Italy and Sicily. Their ages fall between the time estimate for the whole haplogroup (~31,600 YBP) and the LGM, suggesting that their differentiation happened during this time frame (Table 2).

Discussion

Previous reconstructions of the genetic structure of Italy agreed on two points: the peculiarity of the population of Sardinia due to a distinct background and a high degree of isolation [58], [59] and the clinal pattern of variation in the Italian Peninsula, which has been explained by differential migration patterns [17], [18] although some genetic discontinuity due to local drift and founder effects have been described [16], [19], [20]. This study represents a significant upgrade on the knowledge of the genetic structure of Italy for the following reasons: the wide sampling coverage (coupled to a detailed sampling strategy), the high number of typed

Table 2. Age estimates (in YBP) of STR and HVS variation for the most common haplogroups in the Italian data set.

Y Chromosome Haplogroups	SD	SE	Age estimate	SE
E-V13	146.46	51.78	3662	1295
Cluster3 (SEI 70.3%)	139.52	49.33	3488	1233
G-P15	600.79	212.41	15020	5310
Cluster1 (NWI 71.4%)	144.31	51.02	3608	1276
Cluster3	505.72	178.80	12643	4470
Cluster4 (SEI 90.9%)	111.40	39.39	2785	985
Cluster5 (SEI 88.9%)	240.62	85.07	6016	2127
I-M26	206.11	72.87	5153	1822
Cluster 1 (SAR 95.0%)	48.26	17.06	1207	427
Cluster 2 (SAR 81.3%)	227.81	80.54	5695	2014
R-U152	137.29	48.54	3432	1214
Cluster2 (NWI 95.5%)	199.16	70.41	4979	1760
Cluster4 (NWI 70.8%)	184.29	65.16	4607	1629
R-L2	129.67	45.85	3242	1146
Cluster1 (NWI 90.0%)	250.32	88.50	6258	2213
Cluster2 (NWI 72.0%)	185.52	65.59	4638	1640
Cluster3 (NWI 71.4%)	148.55	52.52	3714	1313
R-P312	302.55	106.97	7564	2674
Cluster1 (NWI 70.6%)	130.05	45.98	3251	1149
mtDNA Haplogroups	**Rho**	**SE**	**Age estimate**	**SE**
H*	1.79	0.16	16229	2889
H1_whole (including all H1 derivates)	1.72	0.15	15604	2588
H1*	1.43	0.14	12983	2549
H3	1.71	0.28	15452	4954
H5	1.62	0.23	14689	4015
HV	3.49	0.33	31574	5872
Cluster 2 (NWI 75%)	2.00	0.42	18116	7476
Cluster 3 (SEI 85%)	2.33	0.39	21135	7002
U5a	3.13	0.35	28306	6128
K1a	1.84	0.25	16686	4383
Cluster 2 (NWI 71%)	1.33	0.28	12077	4929
J1c	2.15	0.27	19448	4757

Standard deviation (SD) estimator (Sengupta et al. 2006) and ñ statistic calculator (Soares et al. 2009) were used for Y-chromosome and mtDNA haplogroups respectively. Ages were estimated for the entire haplogroups as well as for each DAPC cluster with at least 10 individuals and frequencies >70% in NWI, SEI, or SAR (excepted for G-P15, cluster 2, see Methods).

doi:10.1371/journal.pone.0065441.t002

markers and the innovative methodological approach. Our results show that the Y-chromosomal genetic diversity of Italy is not clinal but structured in three geographical areas: North-Western Italy (NWI), South-Eastern Italy (SEI) and Sardinia (SAR). The outlier position of SAR described in previous studies [21], [58 61] is mainly due to the high frequency of I-M26 haplogroup, that in turn is almost completely absent in continental Italy. In addition, it is noteworthy the scanty haplotype affinities with other European I-M26 lineages as DAPC results seem to indicate (Figure S7, Table S6). However, the structure observed for paternal lineages in continental Italy and Sicily was not characterised by North-South gradients as previously described: our results show a NWI-SEI clustering (Figure 1a), suggesting a shared genetic background between Southern Italy and the Adriatic coast from one side, and between Northern Italy and Tuscany from the other side. Actually, the most accurate description of the discontinuity between NWI and SEI is that of a "belt", that is a restricted portion of territory in which haplogroup frequencies tend to change more rapidly than in the rest of the Italian peninsula. This model was suggested by the presence of a few populations from North-Eastern and Central Italy (Treviso, Foligno/PG) that reveal an intermediate position between the two main groups.

The discontinuous Y-chromosomal structure of continental Italy is also confirmed by the distribution of DAPC haplotype clusters identified for the most frequent haplogroups (Table 1). Haplogroup G2a provides the most compelling case, being widespread in the whole region, but revealing different clusters in NWI and SEI (Figure 2). This is in agreement with a recent G haplogroup

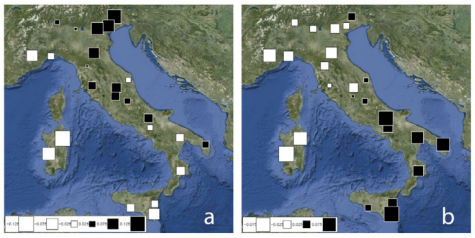


Figure 3. Spatial Principal Component Analysis (sPCA) based on frequencies of mtDNA haplogroups. The first two global components sPC1 (a) and sPC2 (b) are depicted. Positive values are represented by black squares; negative values are represented by white squares; the size of the square is proportional to the absolute value of sPC scores.
doi:10.1371/journal.pone.0065441.g003

survey that revealed the presence of different G2a sub-clades in Italy [62]. Nevertheless, we are not identifying the whole Italian population history with a single haplogroup. In fact, comparisons with other populations taking into account the whole haplogroup spectrum suggest differential patterns of haplotype similarity, implying different genetic histories for the identified sPCA-groups. In particular, NWI is mostly related with Western and Central Europe, while SEI seems to indicate more affinities with the Balkans. In addition, NWI and SEI are characterised by different distributions of genetic variance, the latter showing higher intra-population and lower (not significant) inter-population variability, while the opposite is true for NWI, where significant variation between populations was detected. On the whole, these patterns may be explained by a higher degree of population mobility in SEI, while in NWI local drift effects may have had a greater impact.

In contrast to the results obtained for Y-chromosome, the mtDNA diversity in Italy is characterised by a high degree of homogeneity: the only exception (a marginally significant sPCA global test based on haplogroup frequencies) is due to significant differentiation found in the Sardinian samples compared to continental Italy and Sicily (AMOVA difference between groups = 1.02%, $p<0.05$, Table S4). These results (in agreement with Y chromosome) suggest at least partially different demographic histories for SEI-NWI populations on one hand and SAR on the other hand, the latter being less affected to the gene flow of different migrations occurred in the Italian Peninsula and Sicily. Traces of such processes are visible in sPCA results (Figure 3) and in particular in sPC2, reflecting the same NWI-SEI pattern shown by Y-chromosomal sPC1. Anyway, such differentiation was not significant in the case of mtDNA (AMOVA difference between groups = 0.10%, $p = 0.08$). Analogously, DAPC clusters of mtDNA haplotypes do not show any geographic structure even when compared with other European samples, with clusters of similar haplotypes spanning from Iberia to the Balkans. However, not only uniparental differences in the genetic structure but also in time estimates are shown in the present dataset: our age estimates for the Y-chromosome and the mtDNA haplogroups (as well as the corresponding clusters of haplotypes) highlight significantly different time periods (Table 2), which could reflect multi-layered histories in Italy. Age estimates for mtDNA haplogroups - even if past demographic events affecting error rates cannot be excluded - point almost unanimously to pre-Neolithic times, ranging approximately from ~13,000 (H1*) to ~31,600 (HV) YBP. Although such estimates might reflect the haplogroups pre-existent diversity previous to their establishment in Italy (which could be the case of HV, that includes two DAPC clusters with different geographical distributions and whose ages largely post-date that of the whole haplogroup; Table 2), this does not seem to hold for most of the mtDNA haplogroups analysed. Indeed, most of our mtDNA time estimates are consistent with the hypothesis of the existence of a Glacial Refugium in the Italian Peninsula and its probable role in subsequent post-glacial expansions.

Actually, the role of Italy as a Southern European Glacial Refugium – together with the Iberian and Balkan peninsulas – is demonstrated for a high number of animal and plant species [63–69]. The presence of numerous Epigravettian sites suggests strongly that Italy could have acted as such also for humans [70]. Nevertheless, molecular evidences going in the same direction are still scarce, the only exception being mitochondrial haplogroup U5b3 [8], [9] whose frequency in Italy is relatively low (U5b lineages account for 1.73% in our data). Our results suggest that most of Italian mitochondrial diversity originated during and immediately after LGM. In particular, estimates for H1 and H3 are even older in Italy than in the Franco-Cantabrian area [27] where these clades have been postulated to originate [4]. Furthermore, DAPC comparisons with a wide set of European haplotypes (Table S9) show that Italy, in most cases, is characterised by the highest number of different haplotypes. On the whole, these observations not only are in agreement with the existence of a human Glacial Refugium in Italy, but also suggest that its relevance has been until now largely underrated.

The use of STR variation for dating Y-chromosome lineages or population splits, is a controversial issue, due to the effect that both mutation rates and STR choice has on the temporal scale of age estimates. Following the most recent studies our estimates are based on those STRs that show the highest duration of linearity [49] and by using locus-specific mutation rates (Ballantyne et al. 2010). This is one of the reasons that led us to exclude 'evolutionary' mutation rates (see Methods for details). In addition, we removed 'outlier' haplotypes (see Methods S1), since their presence could inflate significantly the ages of haplogroups and DAPC clusters. However, these results have to be taken with great caution, keeping in mind that 'evolutionary' rates (applied to the same data) would yield time estimates around three times greater. Nonetheless, we observe that two independent methods applied to our data – BATWING and SD-based estimates – yield consistent results. In fact, in contrast to mtDNA age estimates, almost all Y-chromosome estimates fall between late Neolithic and the Bronze Age. This finding supports the hypothesis that group-specific clusters of haplotypes did originate after the split between NWI and SEI (dated with BATWING), even if the confidence interval for BATWING estimate is not tight enough to exclude alternative hypotheses. Interestingly, the NWI and SEI structure detected (Figure 1, Table S4) might be traced back around 5,500 YBP indicating relevant demographic events within continental Italy in this period. Anyway, this value has to be considered as a lower bound, given that the model used does not account for migration that would bias the split time towards recent dates. In fact, given a specific level of populations differentiation, the separation time estimated between these populations has necessarily to be higher (i.e. more ancient) as migration is considered.

According to the most recent syntheses, the Neolithic revolution diffused in Italy following two independent routes along the Adriatic (Eastern) and the Tyrrhenian (Western) coasts. Furthermore, archaeological sites from NWI are characterized by a deeper continuity with earlier Mesolithic cultures and a higher degree of local variability than SEI, while this last area, besides

being culturally more homogeneous, shows clear links with the Southern Balkans [10]. Our Y-chromosome results showing discontinuity between NWI and SEI, higher inter-population variability in NWI, higher homogeneity in SEI coupled with relevant contributes from the Balkans are quite consistent with this model. Thus, we can hypothesize that the NWI-SEI structure detected with paternal lineages could have its origins after these different Neolithic processes. Indeed, comparisons with other European and Near-Eastern populations (Table S6) suggest a stronger affinity between NWI with Iberia and Central Europe, while SEI is more related to the Balkans and Anatolia. The emergence of population structures during the Neolithic has been recently shown in two different studies using Y-chromosome markers, in Near East [71] and in Western Europe [27]. Our results confirm these findings and emphasize the role of demographic expansions and cultural advances related to the Neolithic revolution in shaping human genetic diversity, at least for male lineages. Nonetheless, such pattern might have been further influenced and/or re-shaped also by more recent events.

For instance, the dates of several DAPC clusters fall within the range of the Metal Ages (Table 2). During this long period (third and second millennia BC) Italy underwent important technological and social transformations finally leading to the ethnogenesis of the most important proto-historic Italic peoples. On the whole, our results indicate that these transformations, far from being exclusively cultural phenomena, actually involved relevant population events.

It is worth noting the older age estimate obtained for Y-haplogroup G2-P15 (15,020 YBP) that, coupled with its high frequency (11.09%), makes it the most probable candidate for a continuity with Italian Mesolithic populations (although a Neolithic origin for G2-P15 is discussed, [22], [23]). The most frequent G2-P15 cluster (12,643 YBP, Table 2), besides being evenly diffused in NWI and SEI, it encompasses almost all Sardinian G2-P15 individuals (Figure 2, Table 1). These facts, together with the higher degree of isolation of Sardinia to Neolithic and Post-Neolithic migration processes, support the antiquity of this haplogroup in Italy. Despite obtaining similar time estimates for G2a in Italy (12,899 YBP), Rootsi et al. (2012) [62] explain the diffusion of its main sub-lineages in this country solely as a consequence of Neolithic and Post-Neolithic events.

Conclusions

This study depicts the most complete picture of Italian genetic variability from the point of view of uniparental markers to date. Our analyses revealed that the Y-chromosomal genetic structure of Italy is characterised by discontinuities. Such a structure is defined by three different and well-defined groups of populations: the Sardinia island (SAR), North-Western Italy (NWI) and South-Eastern Italy (SEI). Furthermore, we observed that NWI and SEI are not separated according to latitude but following a longitudinal line. Such discontinuity may date at the Neolithic revolution in Italy, which was characterised by (at least) two independent diffusion processes involving the Western and Eastern coasts, respectively. Mitochondrial DNA, despite showing some correspondence with Y-chromosome results, depicts a substantially homogeneous genetic landscape for the Italian peninsula. Significantly different ages were estimated for mtDNA and Y-chromosome systems. mtDNA variability dates back to Paleolithic and supports the existence of an Italian human Refugium during the last glacial maximum whereas Y-chromosome points to the importance that the demographic events happened during the Neolithic and the Metal Ages had in the male Italian patterns of diversity and distribution.

Supporting Information

Figure S1 Map showing the geographical location of populations sampled in the present study. Colors indicate the eight clusters of homogeneous Italian provinces (sampling macro-areas) identified after a preliminary surname-based analysis [24]. The set of provinces (sampling points) and the number of samples successfully typed for Y-chromosome and mtDNA markers are detailed for each sampling macro-area (table on the left).
(TIF)

Figure S2 Plot of geographical distances against genetic distances (based on frequencies of Y-chromosome haplogroups). A 2-dimensional kernel density estimation layer (Venables and Ripley 2002) was added to the plot. The analysis was performed including (a) and excluding (b) the Sardinian samples.
(TIF)

Figure S3 Eigenvalues of Y-chromosome-based sPCA analysis (A) with their decomposition in spatial and variance components (B). Eigenvalues are obtained maximizing the product of variance and spatial autocorrelation (Moran's I index). They are both positive and negative depending from Moran's I positive or negative values. Large positive components correspond to global structures (cline-like structures); large negative components correspond to local structures (marked genetic differentiation among neighbours).
(TIF)

Figure S4 Loadings of the most informative components (a: sPC1, b: sPC2). These values identify Y-chromosome haplogroups that mostly affect the genetic structure of Italian populations.
(TIF)

Figure S5 DAPC analysis of STRs variation for the most frequent Italian Y-chromosome haplogroups (E-V13, I-M26, J-M410, R-P312*, R-U152*, R-L2). Samples are grouped according to their affiliation to sPCA-identified areas (NWI, SEI, SAR; symbols in the top right legend of each plot). For each plot, the number of different haplotypes per cluster and their geographic distribution in the above areas are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot. Haplogroup I-M26, including two clusters only, is represented by a single discriminant function (no eigenvalues barplot).
(TIF)

Figure S6 Median joining network for Italian G2a-P15 haplotypes. Individuals have been assigned and colored according to the correspondent DAPC-based clusters (Figure 2).
(TIF)

Figure S7 DAPC analysis of STRs variation for the most frequent Y-chromosome haplogroups. Results are based on Italian data and additional comparison samples (NWI; SEI; SAR; IBE: Iberian Peninsula; BAL: Balkan Peninsula; GER: Central-Europe (Germany); CAU: Caucasus; WAN: Western Anatolia; symbols in the legend of each plot). For each plot, the number of different haplotypes per cluster and their geographical distribution are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot.
(TIF)

Figure S8 Eigenvalues of mtDNA-based sPCA analysis (A) with their decomposition in spatial and variance components (B). Eigenvalues are obtained maximizing the product of variance and spatial autocorrelation (Moran's I index), and are both positive and negative, depending from Moran's I positive or negative values. Large positive components correspond to global structures; large negative components correspond to local structures (marked genetic differentiation among neighbours).
(TIF)

Figure S9 Loadings of the most informative components (a: sPC1, b: sPC2). These values identify mtDNA haplogroups that mostly influence the genetic structure of Italian populations.
(TIF)

Figure S10 DAPC analysis of HVS variation for the most frequent mtDNA haplogroups (H*, H1, H3, H5, HV, J1c, K1a, U5a) in the Italian data set. Results have been grouped geographically using the same categories as for Y-Chromosome (NWI; SEI; SAR); "0" codes were attributed to those populations for which Y-chromosome information was not available and whose geographical position lies along the boundary between NWI and SEI (Aviano, Terni). For each plot, the number of different haplotypes per cluster and their geographical distribution are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot. Haplogroup K1a, including two clusters only, is represented by a single discriminant function (no eigenvalues barplot).
(TIF)

Figure S11 DAPC analysis of HVS variation for the most frequent mtDNAhaplogroups. Results are based on Italian data and comparison European populations (ITA: Continental Italy; SAR: Sardinia; BASQ: Iberian Peninsula (Basques); AUST: Central Europe (Austria); MAC: Macedonians; ROM: Romanians; BALK: Balkan Peninsula; symbols in the legend of each plot). For each plot, the number of different haplotypes per cluster and their geographical distribution are shown in the enclosed table. The DAPC eigenvalues are depicted in the enclosed barplot.
(TIF)

Table S1 Frequencies of Y-chromosome haplogroups. Absolute values are reported for the whole Italian data set, while the frequencies within the eight sampling areas (from I to VIII) are expressed in percentage (%).
(XLS)

Table S2 Y-Chromosome STRs haplotypes in the 884 Italian samples of the present study.
(XLS)

Table S3 Diversity indices computed for the different Italian sampling points. Standard diversity parameters were calculated for both Y-chromosome and mtDNA based on haplotype/sequence data and haplogroup frequencies.
(XLS)

Table S4 Analyses of the molecular variance (AMOVA). Apportionment of the variance in %. Samples were grouped according to the geographic clusters (eight macro-areas) and to the sPCA results.
(XLS)

Table S5 DAPC membership probabilities to the SPCA-identified groups.
(XLS)

Table S6 Frequencies of Y-Chromosome DAPC clusters based on Italian data and comparison to other populations. The absolute number of haplotypes and individuals are shown for each population (NWI: sPCA North-Western Italy; SEI: sPCA Southern and Eastern Italy; SAR: Sardinia; IBE: Iberian Peninsula; BAL: Balkan Peninsula; GER: Central-Europe (Germany); CAU: Caucasus; WAN: Western Anatolia).
(XLS)

Table S7 Frequencies of mtDNA haplogroups. Absolute values are reported for the whole Italian data set, while the frequencies within the eight sampling areas (from I to VIII) are expressed in percentage (%).
(XLS)

Table S8 Frequencies of mtDNA DAPC clusters in Italy. Values were calculated both grouping according to the geographical clusters identified with Y-Chromosome sPCA (NWI: Y-sPCA North-Western Italy; SEI: Y-sPCA Southern and Eastern Italy; SAR: Sardinia) as well as considering the continental Italy (including Sicily) altogether (ITA). The absolute number of haplotypes and individuals are shown for each DAPC-cluster, and the maximum frequency for each cluster is expressed in percentage (max%).
(XLS)

Table S9 Frequencies of mtDNA DAPC clusters based on Italian data and comparison to other populations. The absolute number of haplotypes and individuals are shown for each population (ITA: Continental Italy and Sicily; SAR: Sardinia; BASQ: Iberia Peninsula (Basques); AUST: Central Europe (Austria); MAC: Macedonians; ROM: Romanians; BALK: Balkan Peninsula).
(XLS)

Methods S1 Spatial Principal Component Analysis (sPCA). Discriminant Analysis of Principal Components. Batwing analysis. "Jackknife-like" procedure for outliers identification.
(DOC)

Acknowledgments

We are indebted to the Personnel of the Italian Blood Centers and Hospital Centers from: Agrigento, Alessandria, Ancona, Ascoli Piceno, Aviano, Barletta, Bassano del Grappa, Belluno, Benevento, Brescia, Campobasso, Castrovillari, Catania, Catanzaro, Chianciano Terme, Chioggia, Chiusi, Como, Crotone, Cuneo, Enna, Foligno, Genova, Imperia, L'Aquila, La Spezia, Lecce, Macerata, Matera, Montalcino, Novara, Olbia, Oristano, Padova, Pesaro, Pescara, Pistoia, Policoro, Potenza, Ragusa, Sarteano, Sassari, Savona, Teramo, Terni, Trapani, Treviso, Trieste, Varese, Vercelli, Vicenza, Villa d'Agri. Their participation to the Project has been of invaluable help to perform the sampling campaign. We thank Dr. Tosca Corti, Dr. Luisa Stella Dolci and Mr. Carlo Pascucci for helping us in sample collection. We thank all the volunteers who kindly accepted to participate to the study. We are very grateful to CESGA (Centro de Supercomputación de Galicia), where BATWING computational analyses were performed.

The Genographic Consortium includes: Syama Adhikarla[1], Christina J. Adler[2], Elena Balanovska[3], Oleg Balanovsky[3], Jaume Bertranpetit[4], Andrew C. Clarke[5], Alan Cooper[2], Clio S. I. Der Sarkissian[2], Matthew C. Dulik[6], Jill B. Gaieski[6], ArunKumar GaneshPrasad[1], Wolfgang Haak[2], Marc Haber[4,7], Li Jin[8], Matthew E. Kaplan[9], Hui Li[8], Shilin Li[8], Elizabeth A. Matisoo-Smith[5], Nirav C. Merchant[9], R. John Mitchell[10], Amanda C. Owings[6], Laxmi Parida[11], Ramasamy Pitchappan[1], Daniel E. Platt[11], Colin Renfrew[12], Daniela R. Lacerda[13], Ajay K. Royyuru[11], Fabricio R. Santos[13], Theodore G. Schurr[6], Himla Soodyall[14], David F. Soria Hernanz[15], Pandikumar Swamikrishnan[16], Chris Tyler-Smith[17], Arun Varatharajan Santhakumari[1], Pedro Paulo Vieira[18], Miguel G. Vilar[6], R. Spencer Wells[15], Pierre A. Zalloua[7], Janet S. Ziegle[19].

Affiliations for participants: [1]Madurai Kamaraj University, Madurai, Tamil Nadu, India; [2]University of Adelaide, South Australia, Australia; [3]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia; [4]Universitat Pompeu Fabra, Barcelona, Spain; [5]University of Otago, Dunedin, New Zealand; [6]University of Pennsylvania, Philadelphia, PA, USA; [7]Lebanese American University, Chouran, Beirut, Lebanon; [8]Fudan University, Shanghai, China; [9]University of Arizona, Tucson, AZ, USA; [10]La Trobe University, Melbourne, Victoria, Australia; [11]IBM, Yorktown Heights, NY, USA; [12]University of Cambridge, Cambridge, UK; [13]Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [14]National Health Laboratory Service, Johannesburg, South Africa; [15]National Geographic Society, Washington, DC, USA; [16]IBM, Somers, NY, USA; [17]The Wellcome Trust Sanger Institute, Hinxton, UK; [18]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil; [19] Vitapath Genetics, Foster City, CA, USA.

Author Contributions

Conceived and designed the experiments: DP DL DC. Performed the experiments: BMC AU DYY SS GC CH JM LQM PS. Analyzed the data: AB SS BMC AU. Contributed reagents/materials/analysis tools: DL DP DC LQM. Wrote the paper: AB BMC SS DC AU. Performed field work, sampling design and collection: AB DYY AU DL DP.

References

1. Cunliffe B (2001) The Oxford Illustrated History of PreHistoric Europe. Oxford: Oxford University Press. 544.
2. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, 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.
3. Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, 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.
4. Pereira L, Richards M, Goios A, Alonso A, Albarrán C, et al. (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res 15: 19 24.
5. Soares P, Achilli A, Semino O, Davies W, Macaulays V, et al. (2010) The Archaeogenetics of Europe Curr Biol 20: 174 183.
6. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251 1276.
7. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, et al. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science. 290: 1155 1559.
8. Pala M, Achilli A, Olivieri A, Hooshiar Kashani B, Perego UA, 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.
9. Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, et al. (2012) Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. Am J Hum Genet 90: 915 924.
10. Pessina A, Tinè V (2008) Archeologia del Neolitico. L'Italia tra il Vi e il IV millennio a.C. Roma: Carrocci editore. 375.
11. Pesando F (2005) L'Italia antica. Culture e forme del popolamento nel I millennio a. C. Roma: Carocci editore. 326.
12. Barbujani G, Bertorelle G, Capitani G, Scozzari R (1995) Geographical structuring in the mtDNA of Italians. Proc Natl Acad Sci U S A 92: 9171 9175.
13. Cavalli-Sforza L, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton: Princeton University Press. 1088.
14. Turchi C, Buscemi L, Previderé C, Grignani P, Brandstätter A, et al. (2008) Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing. Int J Legal Med 122: 199 204.
15. Ottoni C, Martinez-Labarga C, Vitelli L, Scano G, Fabrini E, et al. (2009) Human mitochondrial DNA variation in Southern Italy. Ann Hum Biol 36: 785 811.
16. Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, et al. (2003) Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Mol Phylogenet Evol 28: 387 395.
17. Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia' A, et al. (2007) Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Mol Phylogenet Evol 44: 228 239.
18. Brisighelli F, Alvarez-Iglesias V, Fondevila M, Blanco-Verea A, Carracedo A, et al. (2012) Uniparental Markers of Contemporary Italian Population Reveals Details on Its Pre-Roman Heritage. PLoS ONE 7: e50794.
19. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18: 1241 1248.
20. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, et al. (2009) Genetic structure of Europeans: a view from the North-East. PLoS One 4: e5472.
21. Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, et al. (2012) An Overview of the Genetic Structure within the Italian Population from Genome-Wide Data. PLoS One 7: e43759.
22. Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, et al. (2011a) Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci U S A 108: 9788 9791.
23. Lacan M, Keyser C, Ricaut FX, Brucato N, Tarrús J, et al. (2011b) Ancient DNA suggests the leading role played by men in the Neolithic dissemination. Proc Natl Acad Sci U S A 108: 18255 18259.
24. Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F (2012) General method to unravel ancient population structures through surnames. Final validation on Italian data. Hum Biol 84: 235 270.
25. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16: 1215.
26. Martínez-Cruz B, Ziegle J, Sanz P, Sotelo G, Anglada R, et al. (2011) Multiplex single-nucleotide polymorphism typing of the human Y chromosome using TaqMan probes. Investig Genet 2: 13.
27. Martínez-Cruz B, Harmant C, Platt DE, Haak W, Manry J, et al. (2012) Evidence of Pre-Roman Tribal Genetic Structure in Basques from Uniparentally Inherited Markers. Mol Biol Evol 29: 2211 2222.
28. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18: 830 838.
29. Gusmão L, Butler JM, Carracedo A, Gill P, Kayser M, et al. (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Forensic Sci Int 157:187 197.
30. Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, et al. (2007) The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3: e104.
31. Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, et al. (2010) Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PloS Biol 8: e1000536.
32. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2010) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32: 25 32.
33. Van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30: 386 394.
34. Excoffier L, Laval G, Schneider S (2007) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1: 47 50.
35. Venables WN, Ripley BD (2002) Modern Applied Statistics with S. New York: Springer495.
36. Caramelli D, Vernesi C, Sanna S, Sampietro L, Lari M, et al. (2007) Genetic variation in prehistoric Sardinia. Hum Genet 122: 327 336.
37. Calò CM, Melis A, Vona G, Piras I (2008) Sardinian population (Italy): a genetic review. International Journal of Modern Anthropology 1: 39 64.
38. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100 1104.
39. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403 1405.
40. Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101: 92 103.
41. R Development Core Team (2008) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.
42. Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37 48.
43. Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, et al. (2011). Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol 28: 2905 2920.
44. King RJ, Di Cristofaro J, Kouvatsi A, Triantaphyllidis C, Scheidel W, et al. (2011) The coming of the Greeks to Provence and Corsica: Y-chromosome models of archaic Greek colonization of the western Mediterranean. BMC Evol Biol 11: 69.
45. Behar DM, Harmant C, Manry J, van Oven M, Haak W, 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.
46. Brandstätter A, Zimmermann B, Wagner J, Göbel T, Röck AW, et al. (2008) Timing and deciphering mitochondrial DNA macro-haplogroup R0 variability in Central Europe and Middle East. BMC Evol Biol 8: 191.
47. Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Drobnic K, Miścicka-Sliwka D (2003) Mitochondrial DNA variability in Bosnians and Slovenians. Ann Hum Genet 67: 412 425.

48. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, et al. (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet 87: 341 353.
49. Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, et al. (2011). The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci 279: 884 892.
50. Wilson I, Weale M, Balding D (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J Roy Stat Soc A 166:155 188.
51. Xue YL, Zejal T, Bao WD, Zhu S, Shu Q, et al. (2006) Male demography in East Asia: A north-south contrast in human population expansion times. Genetics 172: 2431 2439.
52. Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, et al. (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74: 50 61.
53. Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, et al. (2006) Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian Pastoralists. Am J Hum Genet 78: 202 221.
54. King TE, Jobling MA (2009) Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol 26:1093 1102.
55. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. Am J Hum Genet 84: 740 759.
56. Cox MP (2008) Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. Hum Biol 80:335 357.
57. Babalini C, Martínez-Labarga C, Tolk HV, Kivisild T, Giampaolo R, et al. (2005) The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. Eur J Hum Genet 13: 902 912.
58. Contu D, Morelli L, Santoni F, Foster JW, Francalacci P, Cucca F (2008) Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. PLoS One 3: e1430.
59. Pardo LM, Piras G, Asproni R, van der Gaag KJ, Gabbas A, et al. (2012) Dissecting the genetic make-up of North-East Sardinia using a large set of haploid and autosomal markers. Eur J Hum Genet 20: 956 964.
60. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38:556 560.
61. Chiò A, Borghero G, Pugliatti M, Ticca A, Calvo A, et al. (2011) Large proportion of amyotrophic lateral sclerosis cases in Sardinia due to a single founder mutation of the TARDBP gene. Arch Neurol 68:594 598.
62. Rootsi S, Myres NM, Lin AA, Järve M, King RJ, et al. (2012) Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. Eur J Hum Genet 20: 1275 1282.
63. Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol 7: 453 464.
64. Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, et al. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300: 1563 1565.
65. Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. Philos Trans Ser B 359: 183 195.
66. Randi E (2007) Phylogeography of South European Mammals. In: Weiss S, Ferrand N, editors. Phylogeography of Southern European Refugia. Amsterdam: Kluwer Academic Publishers. 101 126.
67. Grassi F, De Mattia F, Zecca G, Sala F, Labra M (2008) Historical isolation and Quaternary range expansion of divergent lineages in wild grapevine. Biological Journal of the Linnean Society 95: 611 619.
68. Grassi F, Minuto L, Casazza G, Labra M, Sala F (2009) Haplotype richness in refugial areas: phylogeographical structure of Saxifraga callosa. Journal of Plant Research 122: 377 387.
69. Zecca G, Casazza G, Labra M, Minuto L, Grassi F (2011) Allopatric divergence and secondary contacts in Euphorbia spinosa L: Influence of climate change on the split of the species. Organisms Diversity and Evolution 11: 357 372.
70. Banks WE, d'Errico F, Peterson AT, Vanhaeren M, Kageyama M, et al (2008) Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. Journal of Archaeological Science 35: 481 491.
71. Haber M, Platt DE, Ashrafian Bonab M, Youhanna SC, Soria-Hernanz DF, et al. (2012) Afghanistan's ethnic groups share a Y-chromosomal heritage structured by historical events. PLoS One 7: e34288.

SUPPLEMENTARY METHODS

Spatial Principal Component Analysis (sPCA)

In order to investigate the spatial distribution of genetic variability within the Italian Peninsula, a spatial principal component analysis (sPCA) was performed on haplogroup frequencies for both Y-Chromosome and Mitochondrial DNA data. Differently from classic PCA, where eigenvalues are calculated by maximizing variance of the data, in sPCA eigenvalues are obtained maximizing the product of variance and spatial autocorrelation (Moran's I index). In order to include spatial information in the analysis, we used a weighting procedure based on a Delaunay connection network [1]. Eigenvalues found by sPCA are both positive and negative, depending from Moran's I positive or negative values. The most informative components are those identified by eigenvalues with the highest absolute values. Large positive components correspond to global structures (cline-like structures); large negative components correspond to local structures (marked genetic differentiation among neighbours). The presence of global or local structures is further assessed by using the Global and Local random test as implemented in the *adegenet* package [2]-[4]. Loadings of the most informative components were used to identify haplogroups that mostly influence the genetic structure of Italian populations.

Discriminant Analysis of Principal Components (DAPC)

The genetic variability of mtDNA and Y-chromosome haplotypes within main haplogroups was explored by means of a DAPC analysis. The DAPC method [5] is aimed to describe the diversity between pre-defined groups of observations. Being designed to investigate individual genetic data, the method can be easily adapted to the study of haplotypes within haplogroups. Preliminarily, data are grouped using k-means, a clustering algorithm which finds a given number of clusters maximizing the variation between groups. The algorithm runs on a transformation of the raw data using Principal Component Analysis (PCA). We retained all the principal components in order to

conserve all the variation in the original data. The optimal number of clusters is identified by running k-means with increasing values of k (up to a maximum, in our case, of 20). Clustering solutions for different k values are compared calculating Bayesian Information Criterion (BIC). The 'best' solution corresponds to the lowest BIC. The actual DAPC procedure consists of two further steps. First, original data (STR haplotypes) are transformed (centred, in our case) and submitted to a PCA. Second, the retained PCs are passed to a Linear Discriminant Analysis based on the groups identified during the preliminary k-means clustering step. As a result, discriminant functions are constructed as linear combinations of the original variables which have the largest between-group variance and the smallest within-group variance. Membership probabilities are based on the retained discriminant functions. Concerning the first step, it is important to observe that retaining too many PCs with respect to the number of populations can lead to over-fitting the discriminant functions, meaning that membership probabilities may become drastically inflated for the best-fitting cluster, resulting in apparent perfect discrimination. As a consequence, we decided to retain as much PCs are needed to represent ~80% of the variation in the original data. The same problem would hold also for the second step, e.g. the number of retained discriminant functions. In our case, given that the number of investigated clusters is relatively low, all the discriminant functions were retained.

Batwing analysis

We established prior distributions covering an expected range congruent with human population history. For mutation rate priors, muprior, these were set to $k = 1.47$ and $\_ = 2173$ for 25 year generations, where the form of the gamma distribution was $f(x;k,\_)dx = \frac{1}{\_(k)}\left(\frac{x}{\_}\right)^{k-1} e^{-x/\_} d\left(\frac{x}{\_}\right)$.

The prior for the ancestral population size was designed to be very flat over the range of likely ancestral values, with $k = 1$, and $\_ = 0.0001$. The population growth rate priors, alpha prior and

betaprior, were set to $k=2$, $\_=400$, and $k=2$, $\_=1$. The number of times parameters were updated between samples was Nbetsamp=10, and the number of times trees were changed before updating parameters was treebetN=20. The number of samples between writing the outfile was picgap=1500000. The total number of samples accumulated in the out file was 3.5 million, and 1 million were excluded as burn-in.

SNP information was integrated for the phylogenetic reconstruction, but it was not considered for posterior estimates. Chain convergence was evaluated by running three independent runs (starting from different seeds) and estimating the Gelman and Geweke diagnostic statistics [6], [7] for the parameters of interest with the R package CODA [4], [8].

"Jackknife-like" procedure for outliers identification

Being the SD-based time estimation of DAPC clusters sensitively affected by the presence of outliers, a jackknife-like procedure for their identification has been designed as follows. For each DAPC cluster of N individuals, the variance-based estimate (SD) was recomputed N times on a set of N-1 haplotypes, leaving out one individual at a time from the original data set. If one of the N estimates, recalculated with the exclusion procedure, is significantly different from the others, we can suspect the presence of an outlier in the original dataset. In this case the best estimation of time will be the one for which the "outlier" haplotype has been excluded. Otherwise, if none of the recomputed estimates differs significantly compared to the others, we can exclude the presence of outliers. In that case, we retain the time estimate calculated on the whole original dataset. The identification of outlier estimates was performed with Grubbs' test [9] using the R software *Outliers* package [10].

References

1. Brassel KE, Reif D. (1979) A procedure to generate Thiessen polygons. Geogr Anal 325:31-36.
2. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403-1405.
3. Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101: 92-103.
4. R Development Core Team (2008) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.
5. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.
6. Gelman A, Rubin DB (1992) Inference from Iterative Simulation Using Multiple Sequences. Stat Sci 7:457-472.
7. Geweke J (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bayesian Statistics 4. Oxford (UK): Clarendon Press.
8. Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R News 6:7-11.
9. Grubbs FE (1950) Sample Criteria for testing outlying observations. Ann Math Stat 21:27-58.
10. Komsta L (2006) Processing data for outliers. R News: 6:10-13.

Figure S1

Area I		Y Chr	mtDNA
1	CUNEO (CN)	30	40
2	SAVONA/GENOVA (SV.GE)	51	43
3	COMO (CO)	41	39
4	BRESCIA (BS)	39	40
Area II			
5	VICENZA (VI)	40	40
6	TREVISO (TV)	33	39
7	AVIANO (PN)	-	29
Area III			
8	BOLOGNA (BO)	29	35
Area IV			
9	LA SPEZIA/MASSA (SP.MS)	24	-
10	GROSSETO/SIENA (GR.SN)	86	36
11	PISTOIA (PT)	13	16
Area V			
12	MACERATA (MC)	40	39
13	FOLIGNO (PG)	37	40
14	TERNI (TR)	-	31
Area VI			
15	L'AQUILA (AQ)	30	25
16	CAMPOBASSO (CB)	30	37
17	BENEVENTO (BN)	36	36
18	MATERA (MT)	25	36
19	LECCE (LE)	39	39
20	COSENZA/CATANZARO/CROTONE (CS.CZ.KR)	38	37
Area VII			
21	AGRIGENTO (AG)	45	42
22	CATANIA (CT)	52	37
23	RAGUSA/SIRACUSA (RG.SR)	44	39
Area VIII			
24	OLBIA/NUORO (OT.NU)	40	31
25	ORISTANO (OR)	42	39
		884	**865**

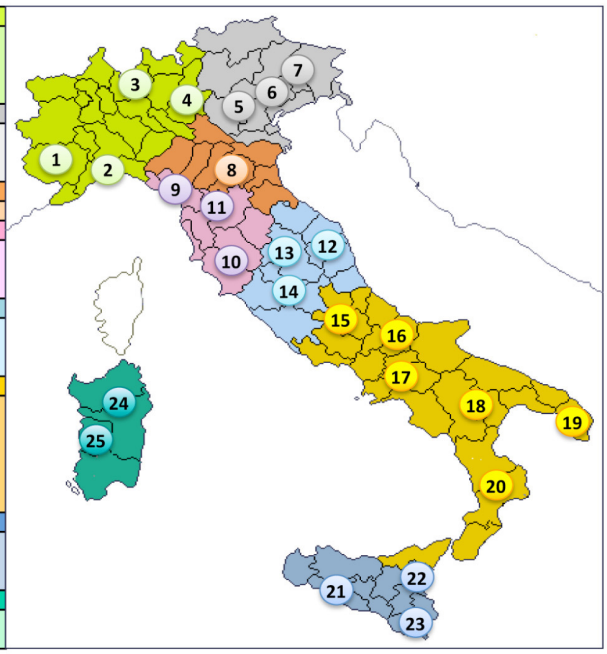


Figure S2

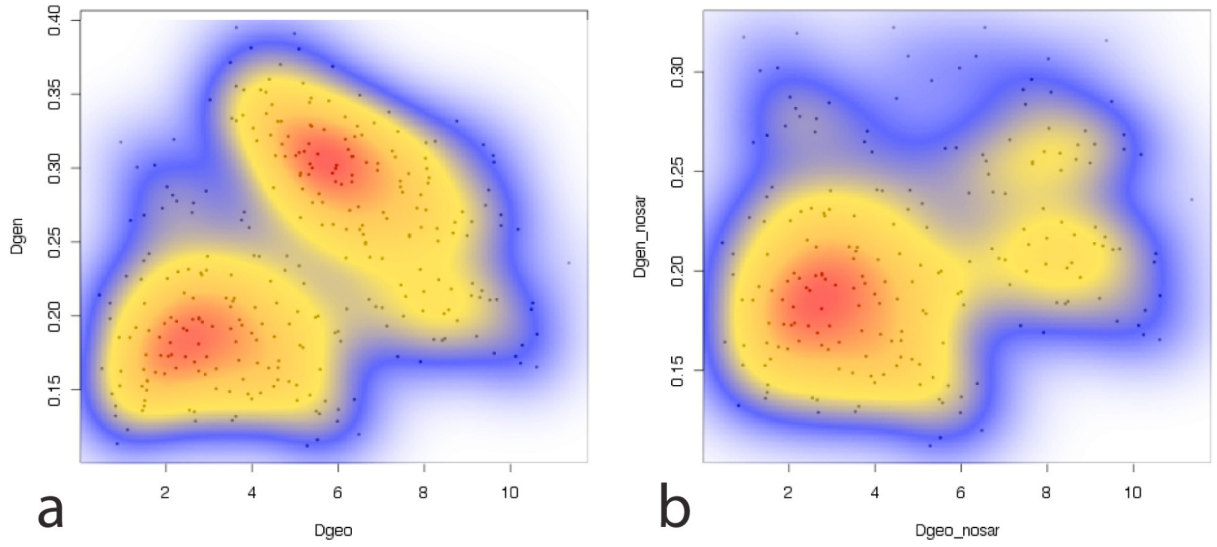


Figure S3

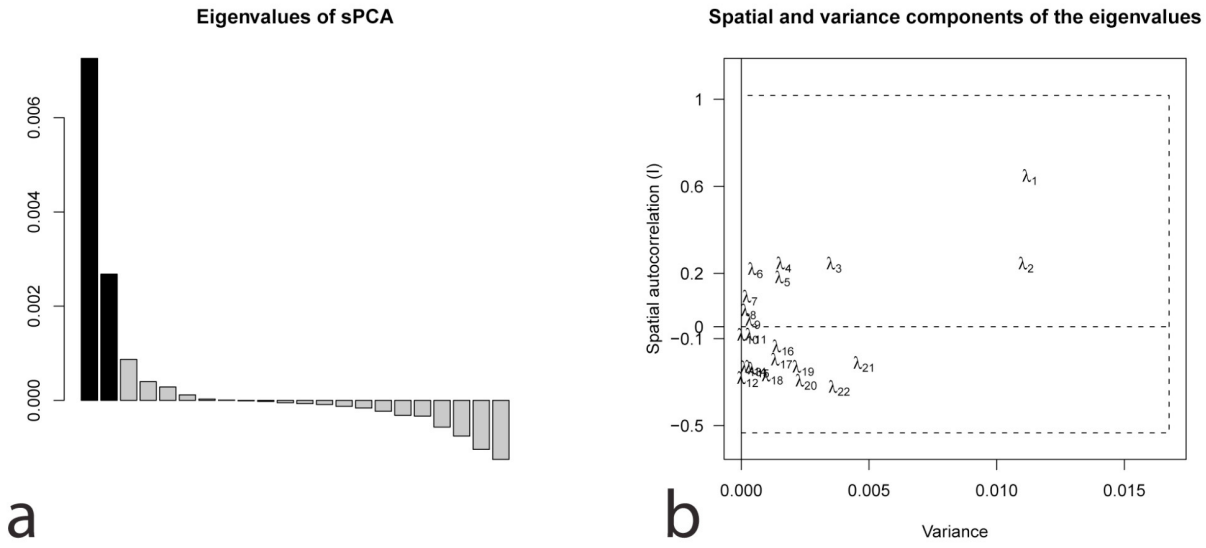


Figure S4

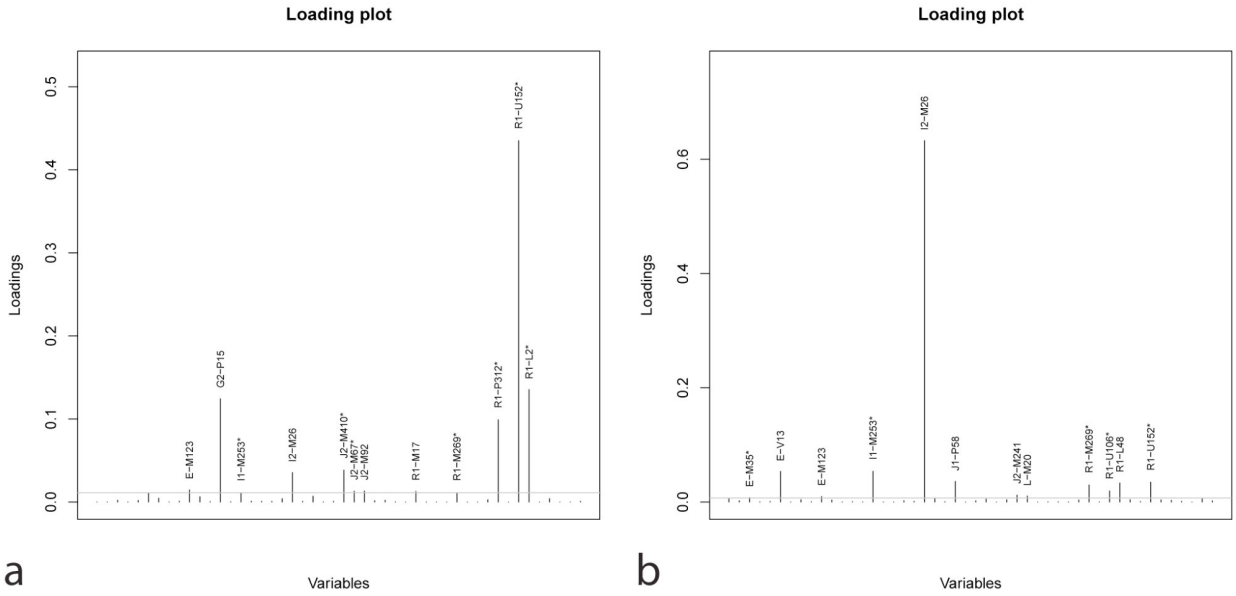


Figure S5

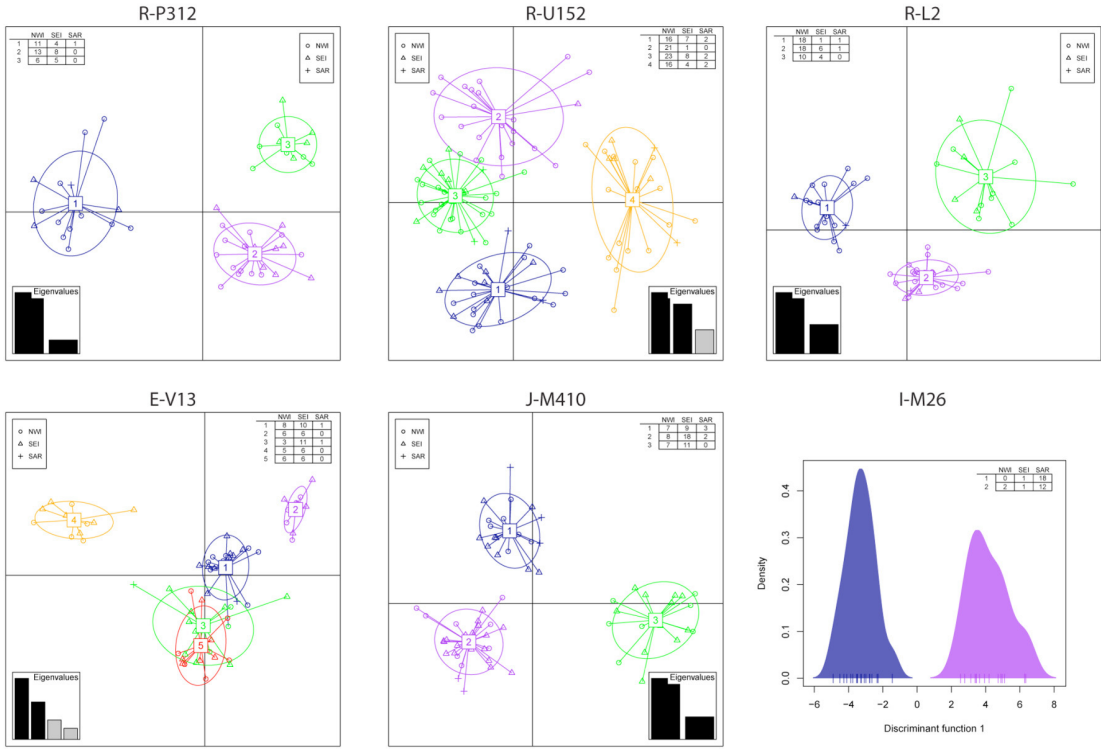


Figure S6

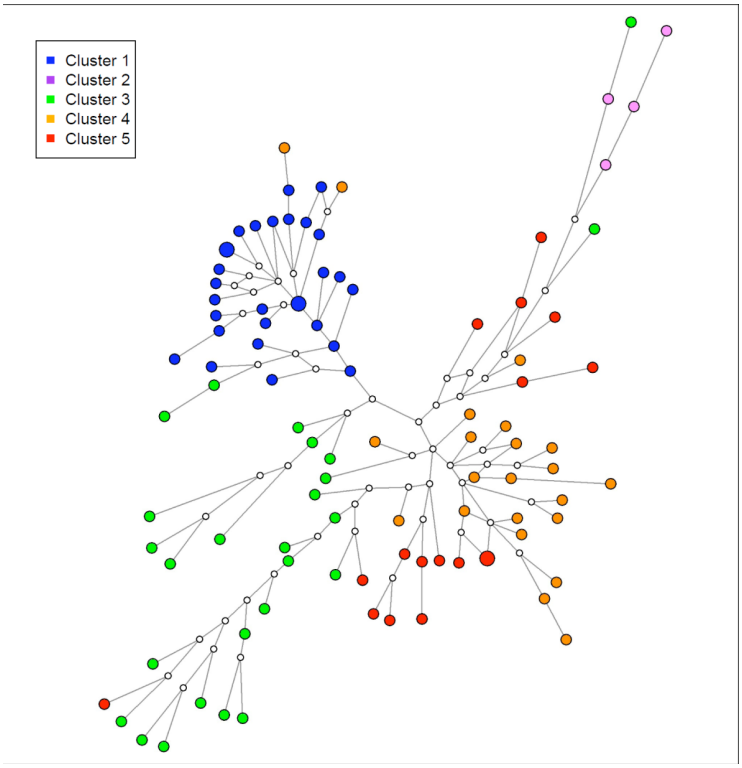


Figure S7

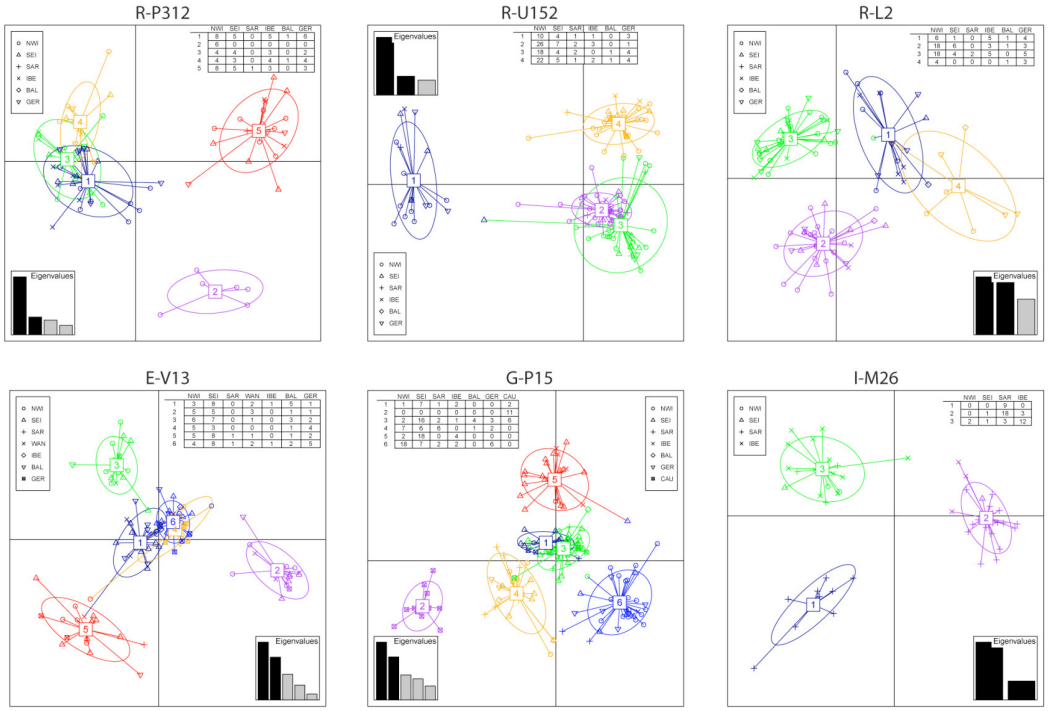


Figure S8

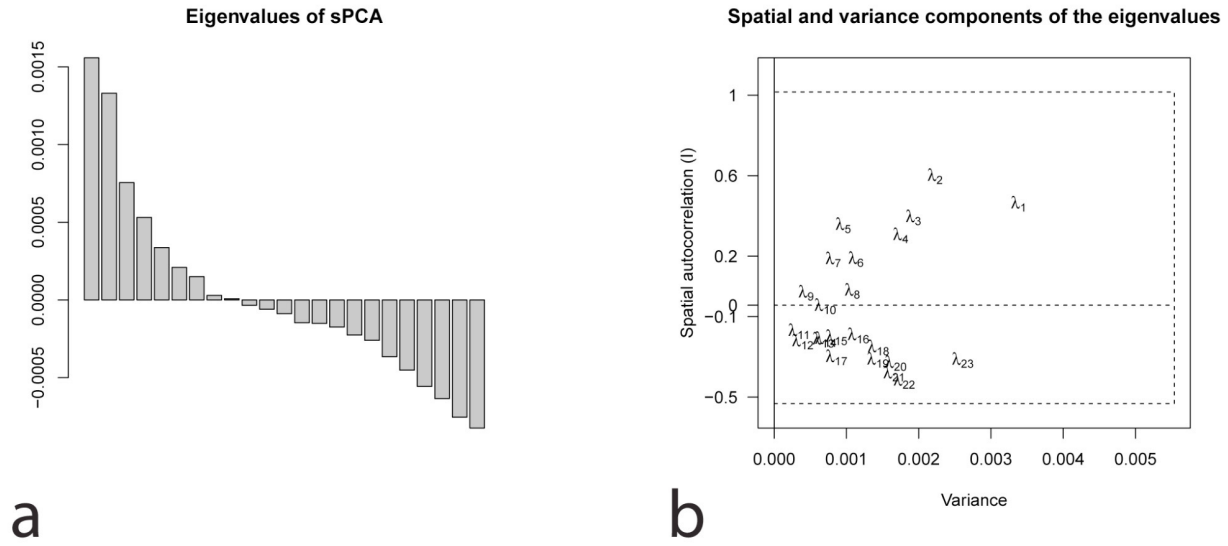


Figure S9

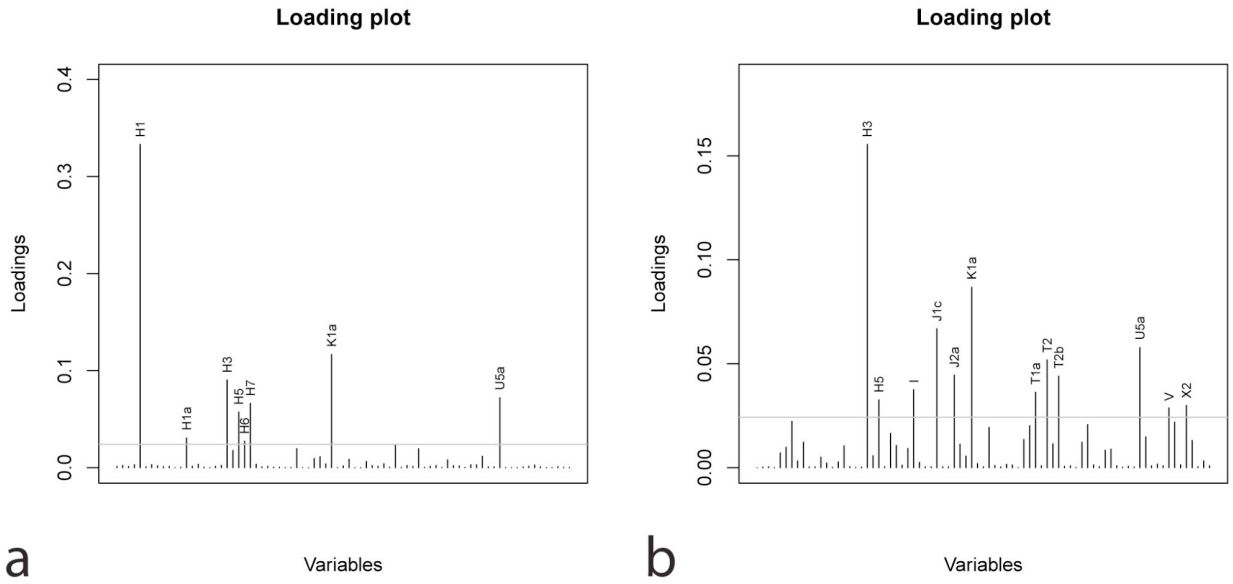


Figure S10

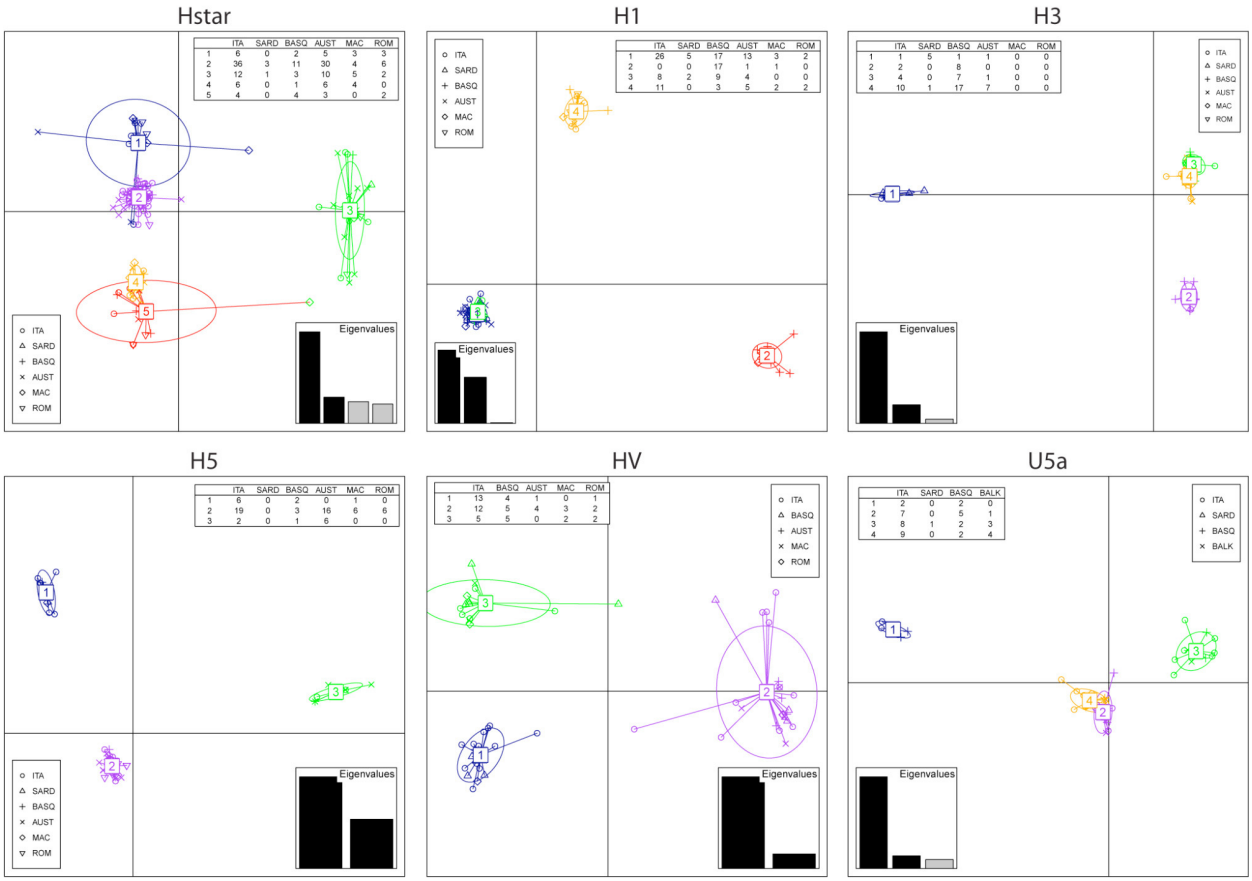
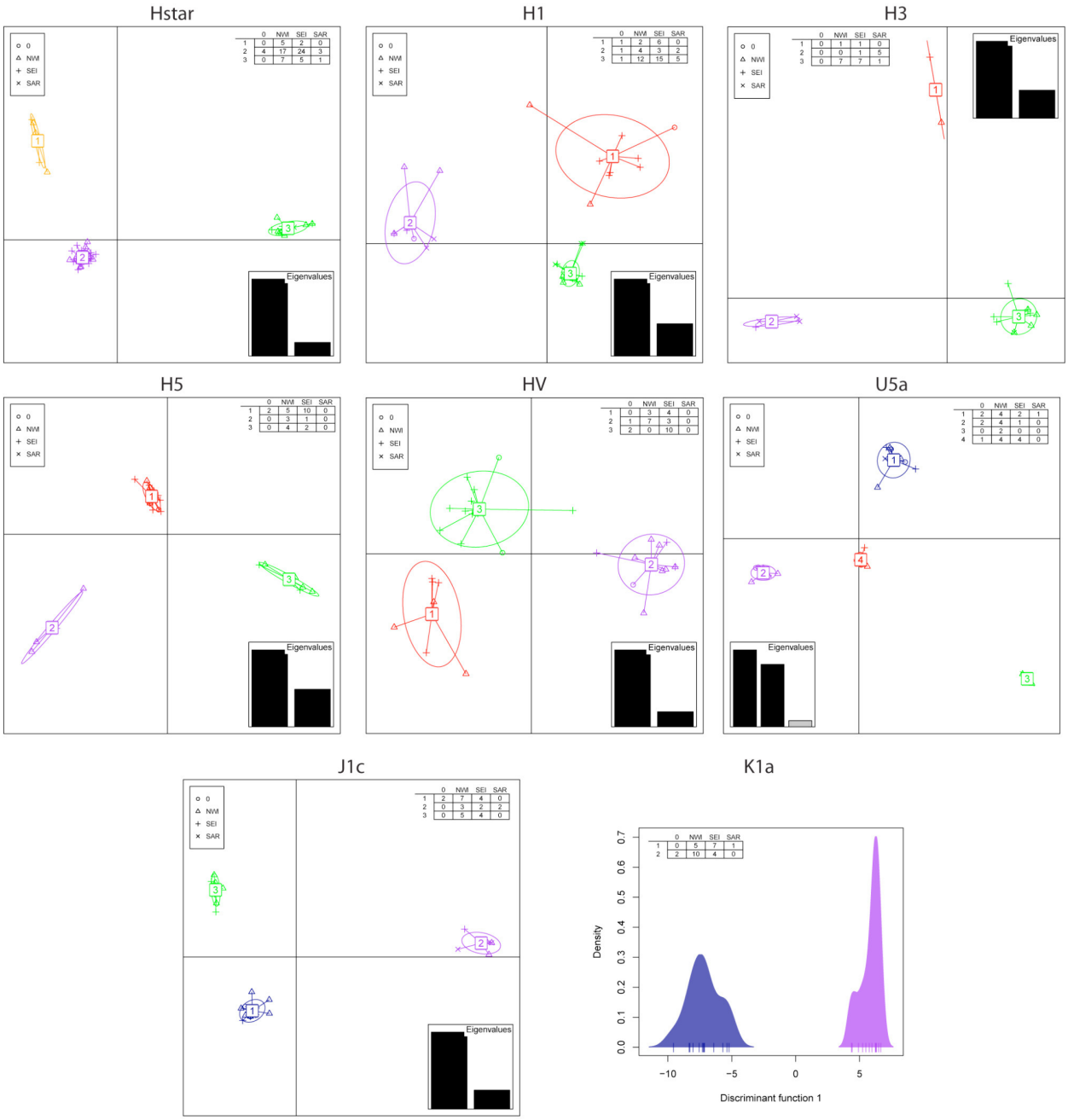

Hstar
ITA SARD BASQ AUST MAC ROM
1 6 0 2 5 3 3
2 36 3 11 30 4 6
3 12 1 3 10 5 2
4 6 0 1 6 4 0
5 4 0 4 3 0 2
ITA
SARD
BASQ
AUST
MAC
ROM
Eigenvalues
H1
ITA SARD BASQ AUST MAC ROM
1 26 5 17 13 3 2
2 0 0 17 1 1 0
3 8 2 9 4 0 0
4 11 0 3 5 2 2
ITA
SARD
BASQ
AUST
MAC
ROM
Eigenvalues
H3
ITA SARD BASQ AUST MAC ROM
1 1 5 1 1 0 0
2 2 0 8 0 0 0
3 4 0 7 1 0 0
4 10 1 17 7 0 0
ITA
SARD
BASQ
AUST
MAC
ROM
Eigenvalues
H5
ITA SARD BASQ AUST MAC ROM
1 6 0 2 0 1 0
2 19 0 3 16 6 6
3 2 0 1 6 0 0
ITA
SARD
BASQ
AUST
MAC
ROM
Eigenvalues
HV
ITA BASQ AUST MAC ROM
1 13 4 1 0 1
2 12 5 4 3 2
3 5 5 0 2 2
ITA
BASQ
AUST
MAC
ROM
Eigenvalues
U5a
ITA SARD BASQ BALK
1 2 0 2 0
2 7 0 5 1
3 8 1 2 3
4 9 0 2 4
ITA
SARD
BASQ
BALK
Eigenvalues

Figure S11



All Supporting Information, included Tables S1-S9 here not shown, can be found on: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0065441#s6

4.3. mtDNA variation in East Africa unravels the history of Afro-Asiatic groups

AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 150:375–385 (2013)

mtDNA Variation in East Africa Unravels the History of Afro-Asiatic Groups

Alessio Boattini,[1]* Loredana Castrì,[1]† Stefania Sarno,[1] Antonella Useli,[1,2] Manuela Cioffi,[1] Marco Sazzini,[1] Paolo Garagnani,[3] Sara De Fanti,[1] Davide Pettener,[1] and Donata Luiselli[1]

[1]*Department of Biological, Geological and Environmental Sciences, Laboratory of Molecular Anthropology, University of Bologna, 40126, Bologna, Italy*
[2]*Department of Science for Nature and Environmental Resources, University of Sassari, 07100 Sassari, Italy*
[3]*Department of Experimental Pathology, University of Bologna, 40126 Bologna, Italy*

KEY WORDS Cushitic; Semitic; Omotic; Horn of Africa; genetic structure

ABSTRACT East Africa (EA) has witnessed pivotal steps in the history of human evolution. Due to its high environmental and cultural variability, and to the long-term human presence there, the genetic structure of modern EA populations is one of the most complicated puzzles in human diversity worldwide. Similarly, the widespread Afro-Asiatic (AA) linguistic phylum reaches its highest levels of internal differentiation in EA. To disentangle this complex ethno-linguistic pattern, we studied mtDNA variability in 1,671 individuals (452 of which were newly typed) from 30 EA populations and compared our data with those from 40 populations (2970 individuals) from Central and Northern Africa and the Levant, affiliated to the AA phylum. The genetic structure of the studied populations—explored using spatial Principal Component Analysis and Model-based clustering—turned out to be composed of four clusters, each with different geographic distribution and/or linguistic affiliation, and signaling different population events in the history of the region. One cluster is widespread in Ethiopia, where it is associated with different AA-speaking populations, and shows shared ancestry with Semitic-speaking groups from Yemen and Egypt and AA-Chadic-speaking groups from Central Africa. Two clusters included populations from Southern Ethiopia, Kenya and Tanzania. Despite high and recent gene-flow (Bantu, Nilo-Saharan pastoralists), one of them is associated with a more ancient AA-Cushitic stratum. Most North-African and Levantine populations (AA-Berber, AA-Semitic) were grouped in a fourth and more differentiated cluster. We therefore conclude that EA genetic variability, although heavily influenced by migration processes, conserves traces of more ancient strata. Am J Phys Anthropol 150:375–385, 2013.

Populations from East Africa (EA) are one of the most compelling puzzles in human diversity worldwide, both from a genetic and a linguistic perspective. In light of the long-term hominid occupation attested by the local fossil record, the Horn of Africa is very likely to have had a major role in the emergence of anatomically modern humans. Furthermore it is one of the most probable gateways for Eurasian colonization by *Homo sapiens*. However, EA complexity could also be the result of more recent events. Ethiopia, for instance, has been involved in a broad net of people movements that extends from the Levant and Arabian Peninsula—via the Bab-el-Mandeb strait—to Central Africa and the Chad Basin—via the Sahel and Southern Sahara (Kivisild et al., 2004; Cerný et al., 2007; 2009; Tishkoff et al., 2009; Cruciani et al., 2010; Musilova et al., 2011; Pagani et al., 2012). Southwards, the area approximately comprised between Kenya, Tanzania, Uganda and Sudan was affected by Bantu expansions and gene flow from Nilo-Saharan-speaking pastoralists starting from ~3,000 years before present (Castrì et al., 2008, 2009; Tishkoff et al., 2009; de Filippo et al., 2011; Gomes et al., 2010). Northwards, the Nile basin has been a privileged way of access to North-Eastern Africa for Neolithic technological innovations (i.e., pastoralism and agriculture) (Newman, 1995).

Language diversity in EA fits well with its complicated genetic history. In Fleming words, "Ethiopia by itself has more languages than all of Europe, even counting all the so-called dialects of the Romance family" (Fleming, 2006). All African linguistic phyla are found in EA: Afro-Asiatic (AA), Nilo-Saharan, Niger-Congo and Khoisan (however, the genealogical unit of Khoisan is no longer generally accepted). Among them, AA is the most differentiated, being represented by three (Omotic, Cushitic, Semitic) of its six major clades (the others being Chadic, Berber and Egyptian). Omotic and Cushitic are considered the deepest clades of AA, and both are found almost exclusively in the Horn of Africa, along with the linguistic relict Ongota that is traditionally assigned to the Cushitic family but whose classification is still widely debated (Fleming, 2006). These observations are in agreement

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: University of Bologna RFO grants 2010 and 2011 to DL.

*Correspondence to: Alessio Boattini, Department of Biological, Geological and Environmental Sciences, Via Selmi 3, 40126, Bologna, Italy.
E-mail: alessio.boattini2@unibo.it

†Diseased

Received 11 July 2012; accepted 19 November 2012

DOI 10.1002/ajpa.22212
Published online 3 January 2013 in Wiley Online Library (wileyonlinelibrary.com).

TABLE 1. List of the sampled ethnic groups from Ethiopia and Kenya with their geographic location, linguistic affiliation, approximate census size and subsistence patterns

Population	Code	Sample Size	Approximate Census size	Country	Language affiliation	Long.	Lat.	Subsistence Patterns
Dawro-Konta	37	137	259,633	Ethiopia	AA-Omotic	37.16676	7.08454	Agro-pastoralists
Ongota	38	19	89	Ethiopia	AA-Cushitic	36.98167	4.83278	Hunter-gatherers, small-scale agriculture
Hamer	39	11	42,466	Ethiopia	AA-Omotic	36.48333	4.96667	Agro-pastoralists
Rendille	40	17	34,700	Kenya	AA-Cushitic	37.46613	2.80588	Seminomadic pastoralists
Elmolo	41	52	700	Kenya	AA-Cushitic	36.71906	2.74827	Fishing
Luo	42	49	4,270,000	Kenya	Nilo-Saharan	34.4751	-0.53832	Agro-pastoralists, fishing
Maasai	43	81	590,000	Kenya	Nilo-Saharan	36.85089	0.70036	Seminomadic pastoralists
Samburu	44	35	174,000	Kenya	Nilo-Saharan	37.06512	1.25783	Seminomadic pastoralists
Turkana	45	51	451,000	Kenya	Nilo-Saharan	35.11231	4.46691	Seminomadic pastoralists

Census size information was extracted from the ethnologue web site (www.ethnologue.com).

with a North-Eastern African origin of the AA languages, most probably in pre-Neolithic times (Ehret, 1979, 1995; Kitchen et al., 2009). The main contender to the African hypothesis is the farming-language dispersal theory (Diamond and Bellwood, 2003), according to which AA languages originated in the Levant and spread in Africa after the Neolithic revolution, along with agriculture and cattle rising. However, this view might contrast with evidence for a local development of farming packages in East Africa such as the Ethiopian ensete, tef and coffee (Phillipson, 1998). Despite the fact that relationships between linguistics and genetics are often elusive, the fact that both languages and genes are influenced by the same kind of evolutionary factors, together with the correlations between genetic and linguistic variation that were observed in several studies, make the use of linguistic affiliations a reasonable way of grouping populations for genetic studies (Scheinfeldt et al., 2010).

This research therefore aims to provide a better knowledge of the mitochondrial genetic structure of populations from EA by taking in consideration both geography and linguistics. More precisely, we are trying to address the following questions: first, is the mtDNA genetic structure of EA more related to geography or linguistics? Second, does the East African high genetic diversity have recent (<3,000 years before present) or ancient origins? Third, from a molecular perspective, which of the two hypotheses about the AA origin is more plausible? To answer all these questions, we have investigated mtDNA variability in individuals belonging to three AA linguistic families (Omotic, Cushitic, Semitic) and two linguistic phyla (Nilo-Saharan, Niger-Congo). We decided not to take into consideration EA Khoisan-speakers, which would behave both as linguistic and genetic outliers in our sample, potentially flattening the diversity observed within AA populations. In addition, these populations are not relevant to the specific purposes of this study. Moreover, in order to better explore the genetic relationships among AA speakers, samples from the Chad basin (Chadic family), North Africa (Berber and Semitic families) and Levant (Semitic family) are included in our analyses, reaching a total of 4,641 EA and AA samples. Materials include 452 HVS-I unpublished sequences from nine populations settled in EA, among which individuals speaking the little known – from the genetic perspective – AA-Omotic family (Dawro Konta, Hamer) and of the controversial AA Ongota language.

MATERIALS AND METHODS

Population samples and locations

Buccal swabs were collected from 167 Ethiopian and 285 Kenyan unrelated apparently healthy individuals belonging to nine ethnic groups and two linguistic phyla: Dawro-Konta (137, AA-Omotic), Hamer (11, AA-Omotic), Ongota (19, AA-Cushitic), Rendille (17, AA-Cushitic), Elmolo (52, AA-Cushitic), Luo (49, Nilo-Saharan), Maasai (81, Nilo-Saharan), Samburu (35, Nilo-Saharan) and Turkana (51, Nilo-Saharan). Their geographic location, linguistic affiliation, census size and subsistence patterns are detailed in Table 1. Census sizes greatly vary between groups, spanning from Kenyan Luo, who exceed 4,000,000, to Elmolo and Ongota, whose sizes are around 700 and only 89, respectively. Not surprisingly, languages spoken by these groups are nowadays almost extinct, Elmolo having shifted to Samburu (Nilo-Saharan) and Ongota to Tsamai (AA-Cushitic). All the considered groups are basically patrilineal and clan systems and/or age-grade institutions of governance are widespread, coupled with exogamy (marriage outside clan) and polygyny, with varying degrees of intensity. For instance, Dawro-Konta people have a very strong clan system, while this is less important among Turkana. As for subsistence patterns, seminomadic pastoralism is frequent in the Lake Turkana area (Turkana, Samburu, Rendille) as well as among Maasai from Kenya. An important exception is represented by Elmolo, who are mainly fishermen (but formerly they were pastoralists, too). Dawro-Konta and Hamer from Ethiopia are agro-pastoralists (with an emphasis on the first term for Dawro-Konta, vice versa for Hamer). Ongota are the only case of hunter-gatherers. Ethnic origin, birthplace and up-to-grandfathers maternal and paternal pedigrees of all individuals were ascertained by oral interview performed in collaboration with local consultants. The collection of biological samples was performed during several expeditions conducted from 1999 to 2010. Written ethical approval for the use of samples from Kenya in this study was provided by the ethics committee of the University of Bologna (record of 2 March 2011; hard copies are available upon request). Samples from Ethiopia were procured in 2007 (Dawro-Konta) and 2010 (Hamer, Ongota) with individual informed consent and following the ethical guidelines stipulated by the research institutions involved in this project. The confidentiality of personal information for each participant to the study was assured.

Reference data from EA include 1,219 HVS-I sequences from individuals belonging to 21 different ethnic groups. A further set of 2,970 HVS-I sequences from North Africa (NA) (1,600), Central Africa (CA) (275) and the Levant (1,095) was established in order to explore the genetic variability of AA-speakers outside EA. The complete reference dataset includes 4,641 HVS-I sequences from 79 different populations. Geographic locations and linguistic affiliations of each of the considered groups are detailed in Supplementary Table 1.

HVS-I sequencing

DNA was extracted by means of a salting out modified protocol (Miller et al., 1988). mtDNA variability was investigated with a focus on the first hypervariable segment (HVS-I), by sequencing a total of 360 base pairs (bp), encompassing nucleotide positions from 16,024 to 16,383.

Polymerase chain reaction (PCR) of the HVS-I region was performed in a T-Gradient Thermocycler (Whatman Biometra, Gottingen, Germany) using L15996 and H16401 primers and following the standard protocol (Vigilant et al., 1991). PCR products were purified by ExoSap-IT® (USB Corporation, Cleveland, OH) and sequenced on an ABI Prism 3730 Genetic Analyzer (Applied Biosystem), using a Big-Dye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), according to the manufacturer's instructions and with the aforementioned primers. To reduce ambiguities in sequence determination the forward and reverse primers were used to sequence both strands of HVS-I. Sequences were then aligned to the reference sequence (Anderson et al., 1981; Andrews et al., 1999) using the DNA Alignment Software 1.3.0.1 (http://www.fluxusengineering.com/align.htm).

To ensure data quality, all sequences were aligned and edited by two researchers independently. The final consensus sequence was then generated by comparing the two independent results. No ambiguities were found.

In order to fit our data with the most updated mtDNA phylogeny (PhyloTree build 15; van Oven, 2009), haplotype motifs were obtained comparing sequences with both the Cambridge Reference Sequence (CRS) and the new Reconstructed Sapiens Reference Sequence (RSRS, Behar et al. 2012). Both CRS- and RSRS-based haplotypes, as well as the corresponding haplogroups, are detailed in Supporting Information Table 2. For comparison purposes, we used the level of phylogenetic resolution adopted in Poloni et al. (2009).

Statistical methods

All following analyses are based on haplotypes and nucleotide differences among haplotypes are taken into account.

Nucleotide diversity and Analysis of Molecular Variance (AMOVA) were calculated using the software Arlequin 3.5 (Excoffier and Lischer, 2010). A Non-Metric Multi-Dimensional Scaling bi-dimensional plot of the examined populations was obtained calculating Nei's distance (Nei, 1972) and using the function isoMDS implemented in the R software package MASS (Cox and Cox, 2001; Venables and Ripley, 2002; R Development Core Team, 2008).

Relationships between geographic coordinates of the populations and genetic variation (HVS-I allelic frequencies) were explored by means of a spatial Principal Component Analysis (sPCA) performed using the R software package adegenet (Jombart, 2008; Jombart et al., 2008; R Development Core Team, 2008). Differently from classic PCA, where eigenvalues are calculated by maximizing variance of the data, in sPCA eigenvalues are obtained maximizing the product of variance and spatial autocorrelation (Moran's I index). In order to include spatial information in the analysis, we used a weighting procedure based on a Delaunay connection network. Eigenvalues obtained by sPCA are both positive and negative, depending from Moran's I positive or negative values. The most informative components are those identified by eigenvalues with the highest absolute values. Large positive components correspond to global structures (i.e., cline-like structures), whereas large negative components correspond to local structures (i.e., marked genetic differentiation among neighbors). Only scores from the most informative components (up to ~80% of the sum of the eigenvalues absolute values) were retained. Therefore, we excluded from downstream analyses those components that convey scarce spatial information and low variance. A Model Based Clustering algorithm, as implemented in the Mclust function included in the R software Mclust package (Fraley and Raftery, 2002; 2006), was then applied to sPCA scores. Mclust explores a set of ten different models for Expectation-Maximization (EM) – each characterized by a different parameterization of the covariance matrix – and for different number of clusters and chooses the best one according to the highest Bayesian Information Criterion (BIC). The output includes the parameters of the maximum-BIC model and the corresponding classification (i.e., affiliation of each population to one of the inferred clusters).

An independent evaluation of membership probabilities for each population to the Mclust-inferred clusters was obtained by means of Discriminant Analysis of Principal Components (DAPC). It is important to note that a certain population – attributed by the above sPCA-Mclust procedure to a certain cluster – not necessarily has its highest DAPC-based membership probability for the same cluster. Actually, DAPC membership probabilities are here used as indicators of how clear-cut genetic clusters are. Accordingly, low values can be interpreted as evidence of admixture with populations from other clusters. The DAPC method (Jombart et al., 2010) aims to describe the diversity between pre-defined groups of observations. Analyses were performed using the R software adegenet package (Jombart 2008; R Development Core Team 2008). Despite being designed to investigate individual genetic data, the method can be easily adapted to population data (and in general to all kinds of tabular data). Briefly, the DAPC procedure consists of two steps. First, original data (e.g., allele frequencies) are centered and submitted to a PCA. Second, the retained PCs are passed to a Linear Discriminant Analysis. As a result, discriminant functions are constructed as linear combinations of the original variables which have the largest between-group variance and the smallest within-group variance. Membership probabilities are based on the retained discriminant functions. Concerning the first step, it is important to observe that retaining too many PCs with respect to the number of populations can lead to over-fitting the discriminant functions, meaning that membership probabilities may become drastically inflated for the best-fitting cluster, resulting in apparent perfect discrimination. The optimal number of retained PCs is evaluated calculating the

α-score, which is the difference between the proportion of successful reassignment of the analysis (observed discrimination) and values obtained using random groups (random discrimination). The procedure relies on repeating DAPC with different randomized groups (the default setting is ten) for different numbers of retained PCs. The 'best' number of retained PCs is the one that optimizes the mean α-score (i.e., the closest to one). The same problem would hold also for the second step, e.g., the number of retained discriminant functions. In our case, given that the number of inspected clusters is low (three), all the discriminant functions were retained.

Surface plots of nucleotide diversity in Africa and the Levant and macro-haplogroup frequencies in EA were obtained with the software Surfer 8 (Golden Software, Golden, CO).

The BayeSSC software (Excoffier et al., 2000; Anderson et al., 2005) was used to perform coalescent simulations of multiple sets of HVS-I mtDNA sequences assuming different demographic scenarios. Simulation sets were used to test the following hypotheses: (1) whether the high genetic EA nucleotide diversity is mainly the result of recent (from ~3,000 YBP) or more ancient events; (2) whether mtDNA results are consistent with an EA origin of AA or with a Levantine one. In both cases, we considered 25 years generations, HVS-I substitution rate of 1.64723×10^{-7} mutations per nucleotide per year (Soares et al., 2009), a Kimura 2-Parameter model with Gamma correction of 0.4 and a transition/transversion bias of 0.91 (Poloni et al., 2009). For hypothesis (1), we proceeded as follows. We simulated four populations corresponding to the four clusters (A, B1, B2, C; see Results) identified using the above described method (sPCA, Mclust); sample sizes are equal to the number of individuals affiliated to each cluster. As a basic demographic model, we assumed three African Sub-Saharan populations (A, B1, B2) splitting from each other ~5,000 generations ago (~125,000 YBP; Garrigan et al., 2007). Experiments with higher values did not yield significant changes in results (not shown). For these clusters we assumed a constant population size. A fourth population (C)—simulating a Levantine/North African cluster—was assumed to split from A 2,400 generations ago (~60,000 YBP) according to a bottleneck scenario (followed by re-expansion) compatible with the parameter space estimated by Gravel et al. (2011). Effective population sizes were introduced in the model as prior uniform distributions varying between 1,500 and 6,500. Within this model, we tested four scenarios with different degrees of population mobility: (a) no migrations (only population splits); (b) instant gene flow (33%) at 120 generations ago (~3,000 YBP) following the direction C → A → B1 → B2; (c) continuous gene flow from 120 generations ago to the present (migration matrix based on mean DAPC membership probabilities per cluster); (d) sum of (b) and (c). For each of the four tested scenarios, we performed 2,000,000 preliminary simulations in BayeSSC. Simulated nucleotide diversity values were further processed for calculating most likely estimates (MLE) of model parameters (population effective sizes) using Approximate Bayesian Computation and retaining the best 5% simulations (Beaumont, 2008). A second set of 100,000 simulations per scenario was run based on MLE. Finally, we calculated AIC (Akaike Information Criterion) for each scenario by comparing simulated and observed nucleotide diversity values.

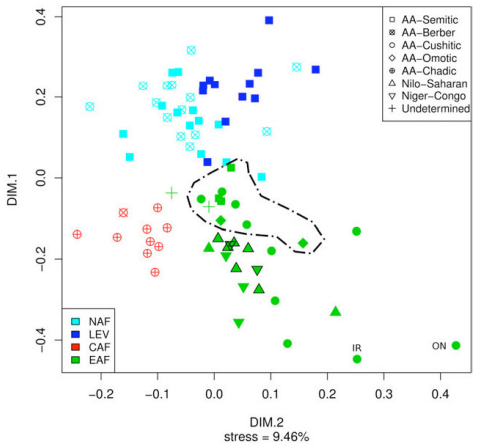


Fig. 1. Non-Metric MDS representation of the 79 examined populations. The plot is rotated right by 90° to better fit the representation with geographic coordinates. Ethiopian samples are enclosed by the dashed-dotted line. Nilo-Saharan and Niger-Congo groups from Kenya are represented by black bordered symbols. Labels in the plot indicate the position of outlier populations (ON: Ongota, IR: Iraqw). Stress value is lower than the cut-off threshold according to Sturrock and Rocha, 2000 (38.8% for two dimensions and 79 objects). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Simulations for hypothesis 2 assume substantial identity between genes and language evolution processes. In other words, these experiments may be read as a mean to understand if mtDNA variability is able to distinguish between two different scenarios (EA or Levantine origin of AA) in a fully idealized case. We simulated 200 HVS-I sequences equally shared between two populations—representing EA and the Levant—evolving independently from a common proto-AA-speaking ancestor. Based on results from precedent simulations, effective population sizes were set to 2,000 and 1,000, respectively, while we assumed a constant population size. We tested two scenarios: (a) EA origin of AA with split between the two populations at 480 generations ago (~12,000 YBP), (b) Levantine origin of AA with split between the two populations at 200 generations ago (~5,000 YBP). 100,000 simulations per scenario were run and empiric distributions for standard genetic parameters (haplotype diversity, nucleotide diversity, Tajima's D) and Nei's distance were compared.

RESULTS

Nucleotide diversity in the whole dataset (Supporting Information Table 1) varies between 0.0115 ± 0.0064 (Libyan Tuareg, 22) and 0.0305 ± 0.0156 (Datoga, 55), the mean being 0.0206. Values higher than the third quartile of the empiric distribution (0.0244) are found almost exclusively in EA, the only exception being a Chadic-speaking population from CA (Hide, 73). The highest values are observed in Kenya and Tanzania, with a decreasing gradient moving towards NA and the Levant (Supporting Information Fig. 1).

As a first overview of mtDNA genetic landscape in the considered populations, we performed a MDS analysis (Fig. 1). Results show that Levantine populations are

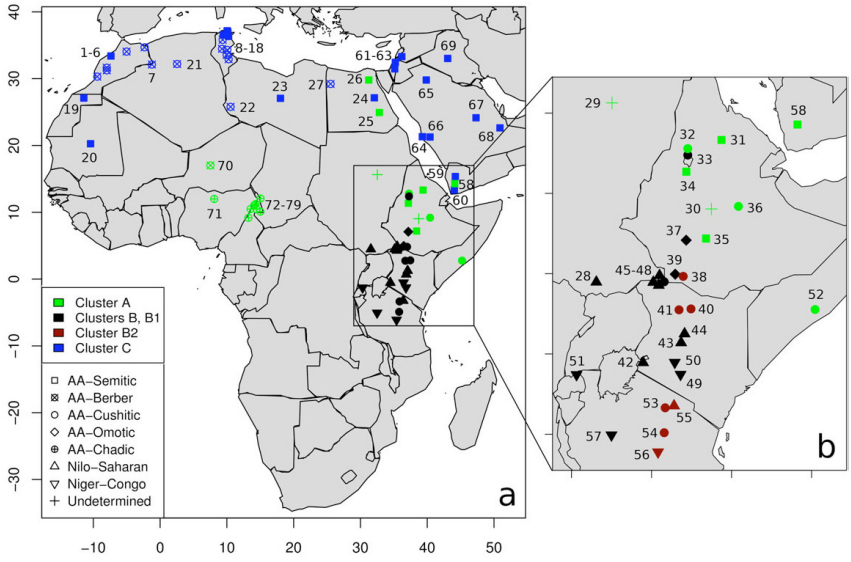


Fig. 2. Geographic distribution of the sPCA-Mclust-based clusters calculated on the whole dataset (a) and on East Africa only (b). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

separated from EA ones along the first dimension, while the second dimension highlights a West-East gradient. Berber and Semitic groups from NA are undistinguishable from the mtDNA point of view, being interspersed between each other. Chadic-speaking populations from CA form a tight and fairly homogeneous group. On the contrary, EA populations show a remarkable degree of variability. Among them, Cushitic speakers are particularly heterogeneous, sometimes behaving as outgroups (Ongota, 38; Iraqw, 53). Groups from Ethiopia fall in the center of the plot (dashed-dotted line), behaving like a bridge between the Levant, NA and EA. Nilo-Saharan and Niger-Congo samples from Kenya form a homogeneous cluster.

sPCA and Mclust analyses were performed at two different geographic scales: (a) on the whole African and Levantine dataset (including all AA-speaking populations); (b) on populations from EA.

In the first case, after having performed sPCA on allelic frequencies data, we retained the first three global sPCs, explaining respectively 73.30%, 10.17%, and 4.37% of the sum of the absolute values of eigenvalues, for a total of 87.83%. sPC scores were fed to Mclust, that settled (as a best model) to an ellipsoidal model with equal shape (VEV) and three clusters (Supporting Information Fig. 2a). Figure 2a shows the geographic distribution of the three inferred clusters (A, B, C).

Cluster A (green) is frequent in EA, mostly in Ethiopia, and in CA. Interestingly, one Yemenite and two Egyptian populations are included in this cluster. Cluster B (black) occupies a relatively tighter area, spanning from Tanzania to Southern Ethiopia. From a linguistic point of view, cluster B includes non-AA speakers (namely Nilo-Saharan and Niger-Congo), as well as AA-Cushitic and Omotic speakers. Cluster C (blue) collects populations from NA and the Levant. From a linguistic point of view, these populations are affiliated to AA-Semitic and AA-Berber families.

The corresponding membership probabilities were calculated using DAPC and retaining eight principal components. Results are detailed in Supplementary Table 1. Cluster C shows the highest mean membership probability (0.897 ± 0.003); individual lowest values are found in a group from Tunisia (14), one from Yemen (59) and two from Saudi Arabia (64, 65). All of them show relevant contributions from cluster A, which can be interpreted as traces of probably recent events of admixture. Mean membership probability in cluster A is 0.810 ± 0.007. It includes two Egyptian populations (25 and 26) that show evidence of admixture with cluster C. Cluster B is characterized by the lowest mean membership probability (0.773 ± 0.015); accordingly, great part of its populations show evidence of admixture with other clusters, most notably with A (in particular 28, 33, 39, 40, 44, 57).

Focusing on EA, and including one population from Yemen (58, cluster A), we retained four sPCs encompassing for 59.32%, 8.23%, 5.40% and 4.82%, respectively, of the sum of the absolute values of eigenvalues, for a total of 77.77%. In that case, Mclust found that the best model was a diagonal, equal shape (VEI) with three clusters (Supporting Information Fig. 2b). As for the geographic location of the clusters, (Fig. 2b) we observe that the 'green' cluster coincides exactly with the EA distribution of cluster A, hence we maintain the same label. More interestingly, cluster B from previous analysis here diverges in two groups, which we call clusters B1 and B2. Cluster B2 (dark red) shows a discontinuous geographic pattern, being divided into a northern portion – located on the border between Ethiopia and Kenya –

TABLE 2. AMOVA results according to geographic-, linguistic-, and mclust-based groupings; (a) whole dataset; (b) East Africa only

Grouping	N° of Groups	N° of Pops	Proportion of variation (%)		
			Among Groups	Among Populations Within Group	Within Populations
(a) Whole Data Base					
All populations	1	79	-	8.80***	91.20***
Geography	4[a]	79	7.66***	3.19***	89.15***
Geography (AA populations only)	4[a]	64	6.22***	3.12***	90.66***
Language	3[b]	76	8.87***	6.17***	84.96***
Language (AA clades)	7[c]	76	7.17***	3.55***	89.28***
Language (AA clades only)	5[d]	64	5.51***	3.88***	90.61***
Mclust	3	79	9.04***	2.96***	88.01***
(b) East Africa Only					
All populations	1	31	–	4.91***	95.09***
Geography	7[e]	31	2.49***	2.90***	94.61***
Geography (AA populations only)	5[f]	16	3.41*	3.38***	93.21***
Language	3[b]	28	1.06*	4.56***	94.38***
Language (AA clades)	5[g]	28	0.92	4.44***	94.63***
Language (AA clades only)	3[h]	16	-0.25	6.18***	94.07***
Mclust	3	31	3.84***	2.43***	93.73***

*** *P*-value < 0.001; ** *P*-value < 0.01; * *P*-value < 0.05.
[a] North Africa, East Africa, Levant, Central Africa.
[b] AA, Nilo-Saharan, Niger-Congo.
[c] AA-Semitic, AA-Berber, AA-Cushitic, AA-Omotic, AA-Chadic, Nilo-Saharan, Niger-Congo.
[d] AA-Semitic, AA-Berber, AA-Cushitic, AA-Omotic, AA-Chadic.
[e] Ethiopia, Kenya, Sudan, Somalia, Rwanda, Tanzania, Yemen.
[f] Ethiopia, Kenya, Somalia, Tanzania, Yemen.
[g] AA-Semitic, AA-Cushitic, AA-Omotic, Nilo-Saharan, Niger-Congo.
[h] AA-Semitic, AA-Cushitic, AA-Omotic.

and a southern one, located in Tanzania. Cluster B2 collects most of the AA-Cushitic speaking populations from these areas (including Ongota). Cluster B1, on the contrary, includes AA-Omotic and non-AA (Nilo-Saharan and Niger-Congo) speaking populations.

Concerning DAPC-based membership probabilities (eight PCs retained), cluster B1 shows the highest mean value (0.949 ± 0.005) and conversely the lowest traces of admixture with other clusters. The opposite is true for cluster B2, revealing low mean membership probability (0.700 ± 0.052) and strong evidences of introgressions from cluster B1 (40, 53, 56). Cluster A is characterized by a mean membership probability equal to 0.863 ± 0.024, showing some relevant contributions from cluster B1 (35).

To further test the reliability of the above clusters, we performed AMOVA at both the full dataset level and the EA level (Table 2). Results were compared with geography- and language-based groupings. In both cases, the highest proportion of variance among groups (9.04%, $p < 0.001$ and 3.84%, $p < 0.001$, respectively) is reached with the Mclust inferred groups. At the full dataset level, significant results were obtained also with geography- and language-based groupings, albeit with lower values. At the EA level, language-based groupings did not yield significant (or only marginally significant, but not after Bonferroni correction) results, while a better score is obtained with geography-based groupings.

Coalescent simulations under different scenarios were performed to test whether the observed patterns of nucleotide diversity within the identified clusters were consistent with recent (~ 3,000 YBP) gene flow. Among the considered scenarios (Table 3, Supporting Information Fig. 3), the least-fitting one (AIC = 1.17) is characterized by absence of migration between populations. Notably, the lowest AIC figure (0.214) was obtained for the scenario involving the highest degree of population mobility, i.e., instant gene flow at 120 generations ago followed by continuous migrations (with rates based on DAPC membership probabilities).

TABLE 3. Akaike Information Criterion (AIC) values calculated comparing observed and simulated nucleotide diversity values assuming four different demographic scenarios

	Scenario			
	Splits	Gen. Flow.	Mig. Mat.	AIC
a	Y	N	N	1.178
b	Y	Y	N	1.062
c	Y	N	Y	0.617
d	Y	Y	Y	0.214

Each scenario is characterized by the presence/absence (Y/N) of the following demographic events: population splits (Splits), instant gene flow (33%) at 120 generations ago (Gen. Flow), continuous gene flow from 120 generations ago to the present (Mig. Mat.). For details see the Methods.

Assuming identity between processes leading to language and genetic variability, a second set of simulations was performed to test whether mtDNA variability may help to discern between an EA and a Levantine origin of AA. Empiric distributions for standard genetic parameters (haplotype diversity, nucleotide diversity, Tajima's D) and Nei's distance (Supporting Information Fig. 4) for the two scenarios show that their confidence intervals are largely overlapping.

Table 4 details distributions of the considered haplogroups in the tree EA clusters, while frequencies for each single EA population are included in Supplementary Table 3. Cluster A is characterized by high frequencies of L2a, M and R0a haplogroups, with lower frequencies of the L0 and L3 lineages, except for L3f.

TABLE 4. Frequencies of the considered mtDNA haplogroups in East African clusters A, B1, and B2

	A	B1	B2
	N (%)	*N* (%)	*N* (%)
H	15 (2.1)	0 (0)	1 (0.4)
HV	20 (2.8)	7 (0.8)	6 (2.7)
I	5 (0.7)	5 (0.6)	16 (7.2)
J	15 (2.1)	2 (0.2)	2 (0.9)
K	20 (2.8)	9 (1.1)	0 (0)
L0	0 (0)	5 (0.6)	0 (0)
L0a	40 (5.6)	108 (12.8)	66 (29.6)
L0b	0 (0)	8 (0.9)	1 (0.4)
L0d	1 (0.1)	2 (0.2)	1 (0.4)
L0f	5 (0.7)	39 (4.6)	32 (14.3)
L0g	0 (0)	5 (0.6)	0 (0)
L0k	2 (0.3)	0 (0)	0 (0)
L1	0 (0)	2 (0.2)	0 (0)
L1b	6 (0.8)	7 (0.8)	0 (0)
L1c	2 (0.3)	11 (1.3)	1 (0.4)
L2	0 (0)	4 (0.5)	0 (0)
L2a	103 (14.4)	78 (9.2)	8 (3.6)
L2b	12 (1.7)	11 (1.3)	1 (0.4)
L2c	2 (0.3)	1 (0.1)	0 (0)
L2d	3 (0.4)	6 (0.7)	0 (0)
L2e	1 (0.1)	0 (0)	0 (0)
L3	2 (0.3)	7 (0.8)	1 (0.4)
L3a	6 (0.8)	17 (2)	23 (10.3)
L3b	6 (0.8)	38 (4.5)	2 (0.9)
L3c	0 (0)	2 (0.2)	2 (0.9)
L3d	20 (2.8)	18 (2.1)	1 (0.4)
L3e	15 (2.1)	36 (4.3)	1 (0.4)
L3f	39 (5.5)	35 (4.1)	2 (0.9)
L3h	19 (2.7)	53 (6.3)	18 (8.1)
L3i	26 (3.6)	36 (4.3)	1 (0.4)
L3x	22 (3.1)	41 (4.8)	2 (0.9)
L4	2 (0.3)	3 (0.4)	0 (0)
L4a	20 (2.8)	11 (1.3)	0 (0)
L4b2	14 (2)	87 (10.3)	22 (9.9)
L5	1 (0.1)	17 (2)	0 (0)
L5a	7 (1)	23 (2.7)	0 (0)
L5b	3 (0.4)	2 (0.2)	0 (0)
L5c	4 (0.6)	29 (3.4)	0 (0)
L6	25 (3.5)	12 (1.4)	4 (1.8)
M	85 (11.9)	41 (4.8)	2 (0.9)
N	34 (4.8)	8 (0.9)	5 (2.2)
R	3 (0.4)	0 (0)	0 (0)
R0a	52 (7.3)	9 (1.1)	1 (0.4)
T	16 (2.2)	2 (0.2)	1 (0.4)
U	28 (3.9)	8 (0.9)	0 (0)
Others	12 (1.7)	1 (0.1)	0 (0)

Cluster B2 shows the highest frequencies of L0a (reaching 29.6%), L0f, L3a, L3h and I, while it has the lowest frequencies of L2 and L5 lineages, as well as M and R0a. Cluster B1 occupies an intermediate position between A and B2, while showing the highest frequencies of L4 lineages. Contour maps of the most frequent macro-haplogroups (L0, L2, L3, L4, M) are reported in Supporting Information Figure 5.

DISCUSSION

In their mtDNA-based survey of East African variability, Poloni and colleagues found "no strong association between linguistically-defined and genetically differentiated groups". Furthermore, they observed that EA "combines a high level of within population-diversity with strong genetic structure among populations". They argue that such results "may be explained [as a consequence of] periodical episodes of admixture in these populations, separated by periods of isolation and genetic drift" (Poloni et al., 2009). Our results largely agree with these observations and, in addition, we were able to uncover traces of an underlying and as yet uncovered genetic structure.

Anyway, a possible drawback of our procedure relies on the fact that sPCA may minimize the role of drift and isolation on single populations, while DAPC maximizes between-group variability, hence underestimating the component of variance generated by gene flow. Another possible source of distortion, however independent from the statistical methods used, could be due to discrepancies in the sampling criteria used in reference studies. Although we cannot exclude some minor effects, we do not observe any detectable relationship between different data sources (i.e., reference studies) and our results.

Our analyses indicate that the structure of EA mtDNA diversity is characterized by three population clusters: A, B1 and B2 (Fig. 2, Supporting Information Table 1). Such structure appears to be related both with geography and linguistic affiliation. On the contrary, to the best of our knowledge there is no evidence of relationships with other socio-cultural variables such as mating behavior (patrilocality is widespread in EA, as well as clan exogamy), social structure (clan-based structures are present in almost all of our samples) and subsistence patterns. The same can be said for demographic dimensions, given that each cluster does include populations with widely differing census sizes (Table 1). Cluster A is centered in Ethiopia and highlights long-range connections of Ethiopian Semitic- and Cushitic-speaking groups with Chadic ones from Central Africa, and Semitic ones from Egypt and the Arabic peninsula. This finding is highly consistent with the role of Ethiopia as a primary hub for recent human migrations already detected in other studies. In fact, movements between Ethiopia and the Arabian peninsula via the Bab-el-Mandeb strait were revealed by mtDNA analyses (Kivisild et al., 2004; Musilova et al., 2011), confirming ancient links between the two coasts of the Red Sea (at least since 8,000 – 9,000 YBP). A reconstruction of the phylogeny of the Semitic linguistic family suggested a single, presumably Levantine origin for Semitic languages in the Horn of Africa (Ethiosemitic), dating their diversification at approximately 2,850 YBP (Kitchen et al., 2009). Evidences of introgressions from the Levant to Ethiopia in the same time frame were indeed revealed by a very recent whole genome study (Pagani et al., 2012). Further population movements along the Nile Valley are suggested by the affiliation to cluster A of two Egyptian populations (25, Gurna and 26, Upper Egypt). They could be related to the spread of Neolithic technologies – according to Newman (1995), the first evidences of pastoralism and agriculture in highland Ethiopia date to ~5,000 YBP – or as the remnants of an ancient AA unity (Egyptian is an extinct branch of AA) extending from EA to Egypt (Stevanovitch et al., 2004). Furthermore, mtDNA (Černý et al., 2009) revealed traces of ancient movements between EA and Central Africa. (Based on Y-chromosome, Cruciani et al., 2010, instead proposed a different route linking Central Africa with North-Eastern Africa). These last migrations were suggested to be responsible for the introduction of Chadic languages (along with their speakers) in the Chad basin area. The high frequencies of haplogroups M and R0a and, to a lesser extent, of T and U (Table 4) – all of them related with the Levant and Asia (Rosa and Brehem, 2011) – fit well with the high mobility patterns detected for this area.

Contrarily to cluster A, clusters B1 and B2 are restricted to EA only, at least in our panel of populations. This means that groups belonging to these clusters do not have 'relatives' among AA-speaking populations outside EA, but maybe they could have them in non-AA groups that were not included in our study. Cluster B2 shows an interesting association with southern Cushitic groups, including the Ongota (38), who are problematic from a linguistic point of view. In fact, the Ongota language, despite being traditionally assigned to the Cushitic family, is suspected to be the remnant of an independent clade of AA (Fleming, 2006), while other scholars (Savà and Tosco, 2003) propose that it may be considered a Cushitic language retaining a Nilo-Saharan substratum. Notably, this last hypothesis implies mixed ancestry for the Ongota, helping to explain their outlying position in genetic space (Fig. 1). In addition, B2 encompasses populations as Rendille (40) and Elmolo (41), that, despite sharing Cushitic languages and the same geographic area (Marsabit district, North-Eastern Kenya), are at present characterized by different subsistence strategies (pastoralism and fishing, respectively, but Elmolo were formerly herders, too). Elmolo's affiliation to cluster B2 is of particular interest, given their current reduced census size (~700) and the fact that their language is almost extinct, being largely substituted by Samburu (44, cluster B1). A second group of B2 populations is located far more south, in Tanzania (53, Iraqw; 54, Burunge; 55, Datog; 56, Turu). Indeed, the association between B2 and AA-Cushitic seems particularly relevant given the discontinuous geographic distribution of the cluster and the fact that Cushitic is considered one of the deepest and most ancient clades of AA (Ehret, 1995); cluster B2, also, shows the highest frequencies of L0 lineages (in particular L0a and L0f), the deepest clade of the human mtDNA phylogeny (Rosa and Brehem, 2011).

Cluster B1 is widespread from the Ethiopian-Kenyan border to Tanzania (Fig. 2), almost encapsulating cluster B2, that, in turn, shows strong signals of admixture with B1. To add other elements to the picture, various studies demonstrated that EA was involved in Bantu expansions (~3,000 YBP, Scheinfeldt et al., 2012), acting both as a point of arrival of migrations from the West and as a starting point for further movements directed to the south, while contributing largely to the diffusion of Niger-Congo languages (Castrì et al., 2008, 2009; de Filippo et al., 2011). Furthermore, a recent Y-chromosome study (Gomes et al., 2010) showed that Kenya and Tanzania (along with Uganda and Sudan) were also affected by a dispersal of pastoralist people speaking Nilo-Saharan languages around 3,000 YBP. Looking at the big picture, it seems reasonable to hypothesize that cluster B1 may be, at least in part, the result of migrations from West of non-AA groups. These migratory processes may have caused both disruption of the geographic distribution of Cushitic cluster B2, as well as admixture, and, at a certain extent, language shifts. For instance, Ehret (1998) argues for a number of now extinct South Cushitic languages in Tanzania on the basis of loan word evidence in local Bantu languages. Further investigations are needed to clarify this important issue.

Omotic speakers, despite the fact that Omotic is generally considered the deepest branch in AA (Blench, 2006), do not show any particular pattern of differentiation, at least from the mtDNA perspective. In fact, they are included in cluster B1, suggesting that wide and ancient phenomena of admixture and/or language shift may have occurred in the area between southern Ethiopia and Kenya. Indeed, high mobility rates are one of the most likely keys to explain the elevated nucleotide diversity observed in EA (Supporting Information Tab. 1, Supporting Information Fig. 1), as well as its (up to now) almost undecipherable genetic structure. Our coalescent simulation experiments reveal indeed that the most likely scenario is the one that implies the most elevated degree of population mobility (Table 3). Such scenario assumes major migration events around 3,000 YBP (corresponding to Bantu and Nilo-Saharan migrations in EA) followed by continuous migratory flows (whose rates are based on DAPC membership probabilities).

Nearly all populations from Northern Africa and the Levant, corresponding to the Semitic and Berber families of AA, are affiliated to cluster C. This finding agrees with contacts and bi-directional migratory exchanges involving the wide corridor between the Maghreb and the Near East already detected for Y-chromosome, autosomal STRs and SNPs (Semino et al., 2004; Tishkoff et al., 2009, Scheinfeldt et al., 2012; Henn et al., 2012). Cluster C is also the most divergent one, showing the highest mean value of membership probability and, consequently, only very limited signals of admixture with other African clusters. These last observations are consistent with the postulated Western Eurasian origin of large part of North African mtDNA lineages (Maca-Meyer et al., 2003; Cherni et al, 2008; Ennafaa et al., 2009, Coudray et al., 2009).

Our results are largely consistent with those of Tishkoff et al. (2009), who adopted a different clustering method using multilocus autosomal data. As in our case, they observe that within EA, clustering is primarily (but not exclusively) associated with language phyla (AA, Nilo-Saharan, Niger-Congo, Khoisan). For instance, their results show that AA-Cushitic-speaking populations from Tanzania as Iraqw (53) and Gorowa cluster with Nilo-Saharan Datoga (55), who are close geographically, mirroring almost perfectly the southern portion of our B2 cluster. Interestingly, according to their results Elmolo (41, cluster B2) are related with formerly Cushitic-speaking groups as Yaaku from Kenya and Akie from Tanzania (not available for our study), strengthening our interpretation of B2 as a Cushitic-specific cluster. As for AA-Chadic speaking groups, Tishkoff et al. (2009) find some shared ancestry between these populations and AA groups from EA, but they conclude that their spread in Central Africa "was not accompanied by large amounts of AA gene flow".

Coming back to the three questions that introduced our study, we are now able to point out the following answers. Concerning the relationships between EA genetic structure and geographic/linguistic affiliations, we observe that both of them contribute to explain our results. Clusters A and B1 are both related to geography, the first being mainly located in Ethiopia and Central Africa, the second in Kenya and Tanzania. Nevertheless, cluster B2 shows an interesting association with some Eastern and all Southern Cushitic populations; cluster A itself is exclusively associated with AA language families, namely Chadic, Cushitic and Semitic. If the arrival of Semitic in EA is relatively recent (~2,850 YBP; Kitchen et al., 2009), Cushitic seems much older (at least ~7,000 YBP, according to Ehret, 1979) and the same holds for proto-Chadic expansions in Central Africa

(~7,000 YBP, according to Ehret, 2002). The fact that Cushitic groups are separated into two different clusters (A and B2) may be an indirect proof of their antiquity. On the whole, these observations indicate that languages had an important role in shaping the matrilineal genetic structure of EA.

As a second point, we asked whether the high mtDNA genetic diversity observed in EA (Supporting Information Table 1, Supporting Information Fig. 1) may be interpreted as the outcome of recent events. According to our results, the answer is yes. It has to be mentioned here that our method for detecting clusters enhances the geographic structure of mtDNA variability, by retaining only those sPCA components with the highest absolute value of eigenvalues. For instance, Ethiopia-centered cluster A can be read as the outcome of recent and repeated migration events, which, from a longitudinal point of view, extend from the Arabian Peninsula to Central Africa (being likely related to the spread of Chadic languages), and from a latitudinal point of view, reach Egypt through the Nile Valley. This is consistent with the local linguistic structure, overlapping in the very same area an ancient and autochthonous AA clade (i.e., Cushitic) with a more recent and exogenous one (i.e., Semitic). Similarly, cluster B, spreading from Southern Ethiopia to Tanzania, was affected by different migration events. In particular, cluster B1 may be interpreted as the result of recent contributions (starting from ~3,000 YBP) from Bantu (Niger-Congo) and Nilo-Saharan pastoralists. Cushitic-specific cluster B2 itself shows clear traces of the same migration events, as suggested by its low mean membership probabilities, as well as by evidences of introgression from B1 (Supporting Information Table 1). In addition, coalescent simulation experiments (Table 3, Supporting Information Fig. 3) suggest that the observed nucleotide diversity patterns can be best explained assuming high population mobility. Nevertheless, our results showed that all these migrations did not manage to completely delete more ancient genetic structures. In particular, cluster B2 seems to be the remnant of an ancient Cushitic continuity between Kenya and Tanzania. On the whole, EA has functioned both as a contact point between already differentiated populations and languages, and as an ancient center of expansion.

Concerning the third point, i.e., the place of origin of AA (EA or the Levant), our results do not allow us to make conclusive statements. Indeed, coalescent simulations of different genetic parameters (Supporting Information Fig. 4) according to the two mentioned hypotheses show that—even assuming complete correlation between languages and mtDNA variability—their confidence intervals largely overlap. Thus, we limit ourselves to the following observations. First, EA shows the highest levels of nucleotide diversity among the studied populations with a decreasing cline towards NA and the Levant (Supporting Information Fig. 1 and Supporting Information Table 1). This is true not only for the Ethiopian cluster A, but also, and especially, for groups belonging to clusters B1 and B2. Second, EA hosts the two deepest clades of AA, Omotic and Cushitic. These families are found exclusively in EA, while the presence of Semitic in this area is much more recent. Third, cluster C – collecting Berber- and Semitic-speaking populations from NA and the Levant – shows only modest signals of admixture with clusters A and B (Fig. 2, Supporting Information Table 1). None of these points, taken by itself, is conclusive, but undoubtedly the hypothesis of origin of AA in EA is the most parsimonious one, if compared to the Levant.

CONCLUSIONS

This study confirms the central role of EA and the Horn of Africa in the genetic and linguistic history of a wide area spanning from Central and Northern Africa to the Levant. Our results confirm high mtDNA diversity and strong genetic structuring in EA. We were indeed able to identify three population clusters (A, B1, B2) that are related both to geography and linguistics, and signaling different population events in the history of the region. The Horn of Africa (cluster A), in accordance with its role as a major gateway between sub-Saharan Africa and the Levant, shows widespread contacts with populations from CA (AA-Chadic speakers), the Arabian peninsula and the Nile Valley. Southwards, Kenya, and Tanzania (clusters B1 and B2), despite being both heavily involved in Bantu and Nilo-Saharan pastoralist expansions, reveal traces of a more ancient genetic stratum associated with Cushitic-speaking groups (cluster B2). Conversely, Berber- and Semitic-speaking populations of NA and the Levant show only marginal traces of admixture with sub-Saharan groups, as well as a different mtDNA genetic background, making the hypothesis of a Levantine origin of AA unlikely. In conclusion, EA genetic structure configures itself as a complicated palimpsest in which more ancient strata (AA-Cushitic-speaking groups) are largely overridden by recent different migration events. Further explorations of AA-Cushitic-speaking populations – both in terms of sampled groups and typed genetic markers – will be of great importance for the reconstruction of the genetic history of EA and AA-speakers.

ACKNOWLEDGMENTS

The authors wish to dedicate this article to the memory of our dear friend and colleague Loredana Castrì. We would like to acknowledge all the participants to the study as well as Francesca Lipeti (Fatima Health Center, Lengesim, Kenya), Samantha Semproli, Serena Tucci and Gianluca Frinchillucci (Perigeo onlus, www.perigeo.org) for their invaluable help in designing and performing the sampling campaigns for this research. AB would like to thank Luca Pagani for his comments that helped to improve the manuscript. This study was in part funded by University of Bologna RFO grants 2010 and 2011 to DL. The authors declare no conflict of interest with the publication of the present study.

LITERATURE CITED

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457–465.

Anderson CNK, Ramakrishnan U, Chan YL, Hadly EA. 2005. Serial SimCoal: a population genetic model for data from multiple populations and points in time. Bioinformatics 21:1733–1734.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

Beaumont MA. 2008. Joint determination of topology, divergence time and immigration in population tree. In: Matsu-

mura S, Forster P, Renfrew C, editors. Simulation, genetics and human prehistory. Cambridge: McDonald Institute for Archaeological Research, University of Cambridge. p 135–154.
Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R. 2012. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90:675–684.
Blench R. 2006. Archaeology, language, and the African past. Lanham, MD: Altamira Press.
Castrì L, Garagnani P, Useli A, Pettener D, Luiselli D. 2008. Kenyan crossroads: migration and gene flow in six ethnic groups from Eastern Africa. J Anthropol Sci 86:189–192.
Castrì L, Tofanelli S, Garagnani P, Bini C, Fosella X, Pelotti S, Paoli G, Pettener D, Luiselli D. 2009. mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. Am J Phys Anthropol 140:302–311.
Cerný V, Fernandes V, Costa MD, Hájek M, Mulligan CJ, Pereira L. 2009. Migration of Chadic speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. BMC Evol Biol 9:63.
Cerný V, Salas A, Hájek M, Zaloudková M, Brdicka R. 2007. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71:433–452.
Cherni L, Fernandes V, Pereira JB, Costa MD, Goios A, Frigi S, Yacoubi-Loueslati B, Amor MB, Slama A, Amorim A, El Gaaied AB, Pereira L. 2009. Post-last glacial maximum expansion from Iberia to North Africa revealed by fine characterization of mtDNA H haplogroup in Tunisia. Am J Phys Anthropol 139:253–260.
Coudray C, Olivieri A, Achilli A, Pala M, Melhaoui M, Cherkaoui M, El-Chennawi F, Kossmann M, Torroni A, Dugoujon JM. 2009. The complex and diversified mitochondrial gene pool of Berber populations. Ann Hum Genet 73:196–214.
Cox TF, Cox MAA. 2001. Multidimensional scaling. Boca Raton: Chapman and Hall.
Cruciani F, Trombetta B, Sellitto D, Massaia A, Destro-Bisol G, Watson E, Colomb EB, Dugoujon JM, Moral P, Scozzari R. 2010. Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. Eur J Hum Genet 18:800–807.
de Filippo C, Barbieri C, Whitten M, Mpoloka SW, Gunnarsdóttir ED, Bostoen K, Nyambe T, Beyer K, Schreiber H, de Knijff P, Luiselli D, Stoneking M, Pakendorf B. 2011. Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol 28:1255–1269.
Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. Science 300:597–603.
Ehret C. 1979. On the antiquity of agriculture in Ethiopia. J Afr Hist 20:161–177.
Ehret C. 1995. Reconstructing proto-Afro-Asiatic. Berkeley, CA: University of California Press.
Ehret C. 1998. An African classical age: eastern and southern Africa in world history, 1000 B.C. to A.D. 400. Charlottesville: University of Virginia Press.
Ehret C. 2002. The civilizations of Africa: a history to 1800. Oxford: James Currey.
Ennafaa H, Cabrera VM, Abu-Amero KK, González AM, Amor MB, Bouhaha R, Dzimiri N, Elgaaïed AB, Larruga JM. 2009. Mitochondrial DNA haplogroup H structure in North Africa. BMC Genet 10:8.
Excoffier L, Novembre J, Schneider S. 2000. SIMCOAL: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. J Hered 91:506–509.
Excoffier L, Lischer HE. 2010. Arlequin suite ver. 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–7.
Fleming HC. 2006. Ongota: a decisive language in African prehistory. Wiesbaden: Harrassowitz.
Fraley C, Raftery AE. 2006. MCLUST Version 3 for R: normal mixture modeling and model-based clustering. Technical Report No. 504, Department of Statistics, University of Washington (revised 2009).
Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis and density estimation. J Am Statist Assoc 97:611–631.
Garrigan D, Kingan SB, Pilkington MM, Wilder JA, Cox MP, Soodyall H, Strassmann B, Destro-Bisol G, de Knijff P, Novelletto A, Friedlaender J, Hammer MF. 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics 177:2195–2207.
Gomes V, Sánchez-Diz P, Amorim A, Carracedo A, Gusmão L. 2010. Digging deeper into East African human Y chromosome lineages. Hum Genet 127:603–13.
Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA,The 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. PNAS 108:11983–11988.
Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, Bustamante CD, Comas D. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet 8:e1002397.
Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci USA 105:10693–10698.
Jombart T, Devillard S and Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94.
Jombart T, Devillard S, Dufour AB, Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity (Edinb) 101:92–103.
Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.
Kitchen A, Ehret C, Assefa S, Mulligan CJ. 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc Biol Sci 276:2703–2710.
Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752–770.
Maca-Meyer N, González AM, Pestano J, Flores C, Larruga JM, Cabrera VM. 2003. Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. BMC Genet 4:15.
Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16:1215.
Musilová E, Fernandes V, Silva NM, Soares P, Alshamali F, Harich N, Cherni L, Gaaied AB, Al-Meeri A, Pereira L, Cerný V. 2011. Population history of the Red Sea: genetic exchanges between the Arabian Peninsula and East Africa signaled in the mitochondrial DNA HV1 haplogroup. Am J Phys Anthropol 145:592–598.
Nei M. 1972. Genetic distances between populations. Am Nat 106:283–292.
Newman JL. 1995. The peopling of Africa: a geographic interpretation. New Haven: Yale University Press.
Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, Bekele E, Bradman N, Balding DJ, Tyler-Smith C. 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am J Hum Genet 91:83–96.
Phillipson DW. 1998. Ancient Ethiopia. Aksum: its antecedents and successors. London: British Museum Press.
Poloni ES, Naciri Y, Bucho R, Niba R, Kervaire B, Excoffier L, Langaney A, Sanchez-Mazas A. 2009. Genetic evidence for complexity in ethnic differentiation and history in East Africa. Ann Hum Genet 73:582–600.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3–900051-07–0, Available at: URLhttp://www.R-project.org.

Rosa A, Brehem A. 2011. African human mtDNA phylogeography at-a-glance. J Anthropol Sci 89:25–58.

Savà G, Tosco M. 2003. The classification of Ongota. In: Bender ML, Takács G, Appleyard DL, editors. Selected comparative-historical Afrasian linguistic studies. Munich, Germany: LINCOM Europa. p 307–316.

Scheinfeldt LB, Soi S, Tishkoff SA. 2010. Colloquium paper: working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. Proc Natl Acad Sci USA 107Suppl 2:8931–8938.

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS. 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74:1023–1034.

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.

Stevanovitch A, Gilles A, Bouzaid E, Kefi R, Paris F, Gayraud RP, Spadoni JL, El-Chenawi F, Béraud-Colomb E. 2004. Mitochondrial DNA sequence diversity in a sedentary population from Egypt. Ann Hum Genet 68:23–39.

Sturrock K, Rocha J. 2000. A multidimensional scaling stress evaluation table. Field Methods 12:49–60.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM. 2009. The genetic structure and history of Africans and African Americans. Science 324:1035–1044.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394. Available at:http://www.phylotree.org.

Venables WN, Ripley BD. 2002. Modern applied statistics with S, 4th ed. New York: Springer.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. 1991. African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507.

Figure S1

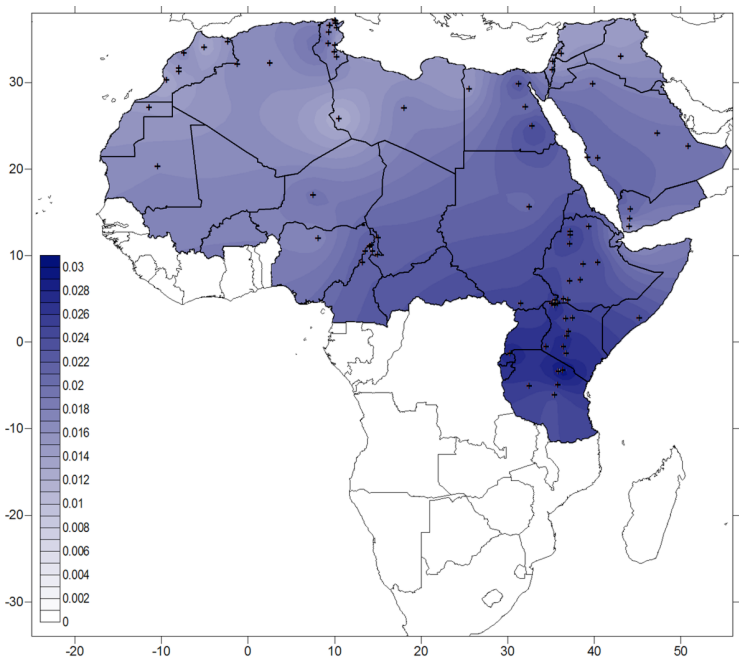


Figure S2

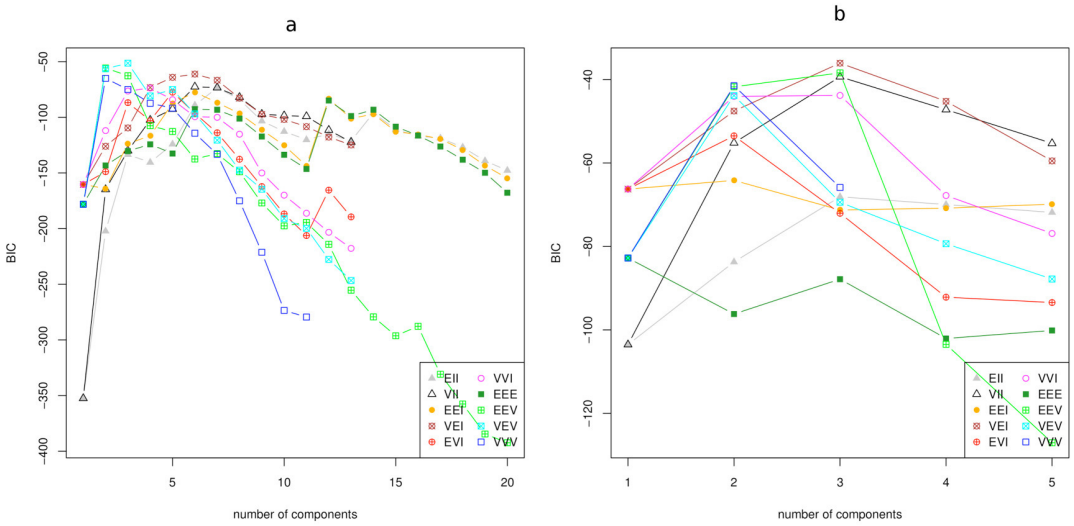


Figure S3

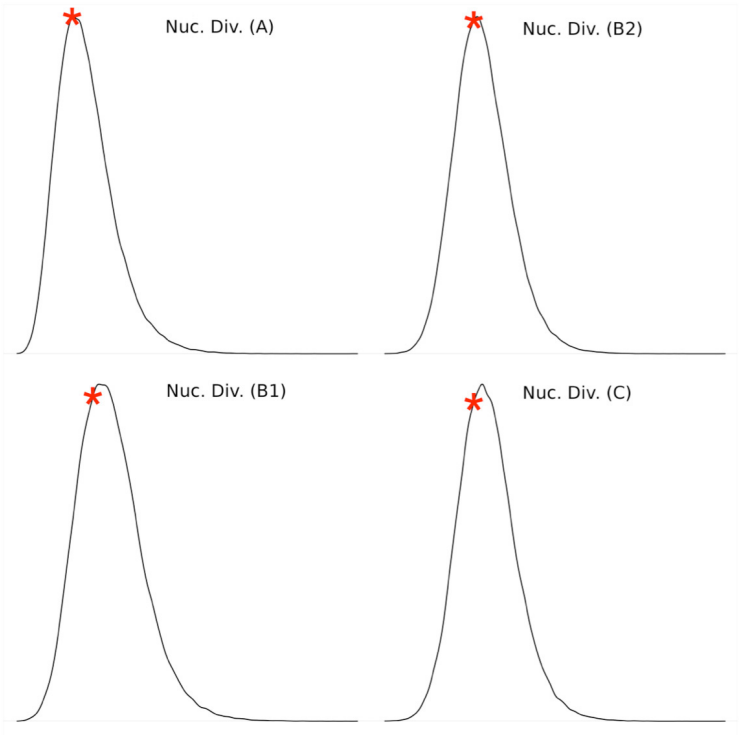


Figure S4

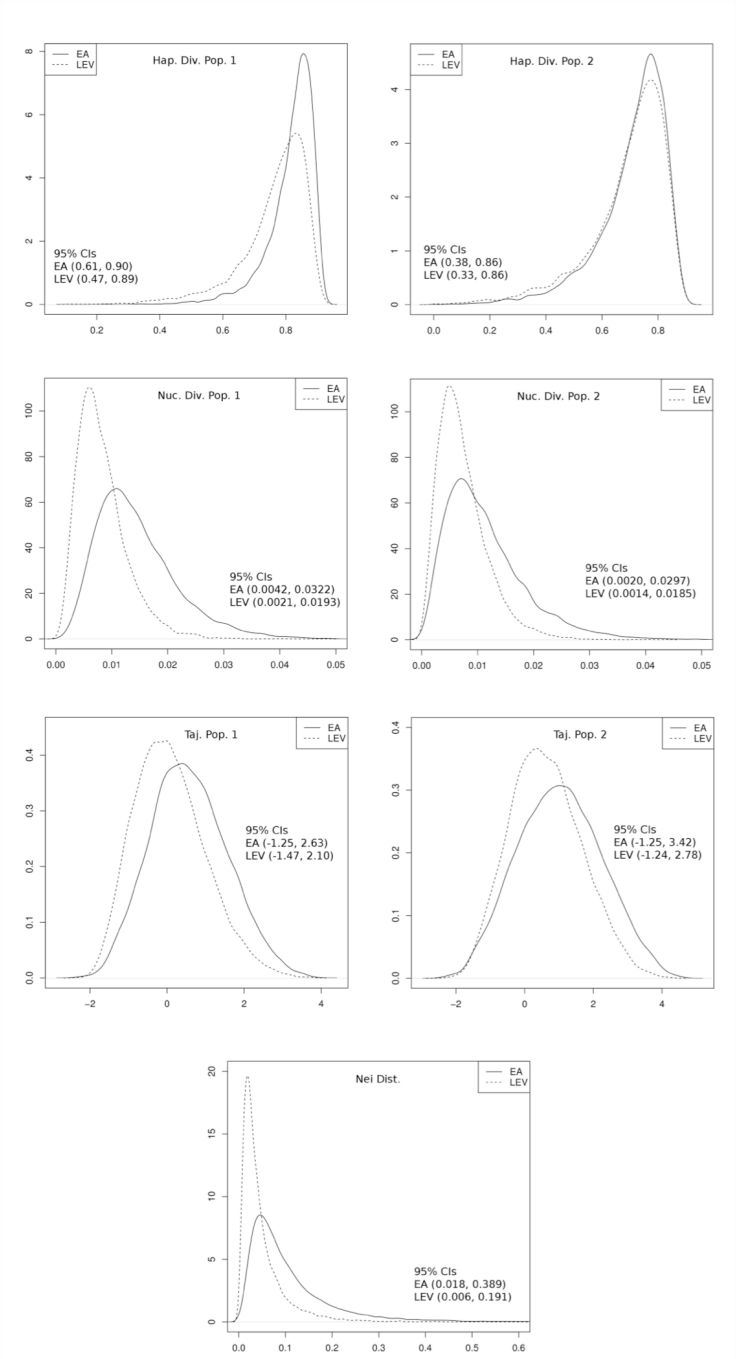
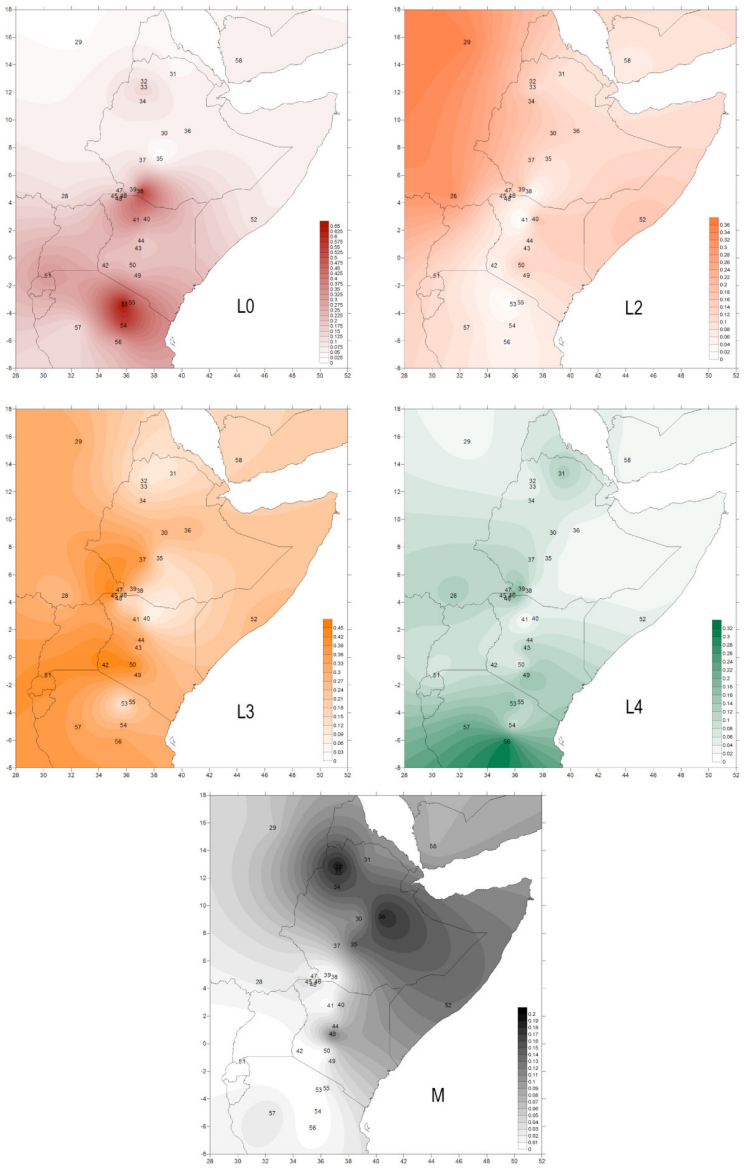

EA
LEV
Hap. Div. Pop. 1
95% CIs
EA (0.61, 0.90)
LEV (0.47, 0.89)
Hap. Div. Pop. 2
95% CIs
EA (0.38, 0.86)
LEV (0.33, 0.86)
Nuc. Div. Pop. 1
95% CIs
EA (0.0042, 0.0322)
LEV (0.0021, 0.0193)
Nuc. Div. Pop. 2
95% CIs
EA (0.0020, 0.0297)
LEV (0.0014, 0.0185)
Taj. Pop. 1
95% CIs
EA (-1.25, 2.63)
LEV (-1.47, 2.10)
Taj. Pop. 2
95% CIs
EA (-1.25, 3.42)
LEV (-1.24, 2.78)
Nei Dist.
95% CIs
EA (0.018, 0.389)
LEV (0.006, 0.191)

Figure S5



Abu-Amero KK, Larruga JM, Cabrera VM, González AM. 2008. Mitochondrial DNA structure in the Arabian Peninsula. BMC Evol Biol 12,8:45.

Behar DM, Metspalu E, Kivisild T, Rosset S, Tzur S, Hadid Y, Yudkovsky G, Rosengarten D, Pereira L, Amorim A, Kutuev I, Gurwitz D, Bonne-Tamir B, Villems R, Skorecki K. 2008. Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. PLoS One 34:e2062.

Brakez Z, Bosch E, Izaabel H, Akhayat O, Comas D, Bertranpetit J, Calafell F. 2001. Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. Ann Hum Biol 28:295-307.

Brandstätter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, Parsons TJ. 2004. Mitochondrial DNA control region sequences from Nairobi Kenya.: inferring phylogenetic parameters for the establishment of a forensic database. Int J Legal Med 1185:294-306.

Castrì L, Tofanelli S, Garagnani P, Bini C, Fosella X, Pelotti S, Paoli G, Pettener D, Luiselli D. 2009. mtDNA variability in two Bantu-speaking populations Shona and Hutu. from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. Am J Phys Anthropol 1402:302-11.

Cerný V, Hájek M, Cmejla R, Brůzek J, Brdicka R. 2004. mtDNA sequences of Chadic-speaking populations from northern Cameroon suggest their affinities with eastern Africa. Ann Hum Biol 315:554-69.

Coia V, Destro-Bisol G, Verginelli F, Battaggia C, Boschi I, Cruciani F, Spedini G, Comas D, Calafell F. 2005. Brief communication: mtDNA variation in North Cameroon: lack of Asian lineages and implications for back migration from Asia to sub-Saharan Africa. Am J Phys Anthropol 1283:678-81.

Cherni L, Fernandes V, Pereira JB, Costa MD, Goios A, Frigi S, Yacoubi-Loueslati B, Amor MB, Slama A, Amorim A, El Gaaied AB, Pereira L. 2009. Post-last glacial maximum expansion from Iberia to North Africa revealed by fine characterization of mtDNA H haplogroup in Tunisia. Am J Phys Anthropol 1392:253-60.

Côrte-Real HB, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, Bertranpetit J, Sykes BC. 1996. Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60Pt4:331-50.

Coudray C, Olivieri A, Achilli A, Pala M, Melhaoui M, Cherkaoui M, El-Chennawi F, Kossmann M, Torroni A, Dugoujon JM. 2009. The complex and diversified mitochondrial gene pool of Berber populations. Ann Hum Genet 732:196-214.

Di Rienzo A, Wilson AC. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci U S A 885:1597-601.

Fadhlaoui-Zid K, Plaza S, Calafell F, Ben Amor M, Comas D, Bennamar El gaaied A. 2004. Mitochondrial DNA heterogeneity in Tunisian Berbers. Ann Hum Genet 68Pt3:222-33.

Fadhlaoui-Zid K, Rodríguez-Botigué L, Naoui N, Benammar-Elgaaied A, Calafell F, Comas D. 2011. Mitochondrial DNA structure in North Africa reveals a genetic discontinuity in the Nile Valley. Am J Phys Anthropol 1451:107-17.

Falchi A, Giovannoni L, Calo CM, Piras IS, Moral P, Paoli G, Vona G, Varesi L. 2006. Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. Am J Hum Genet 511:9-14.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 755:752-70.

Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL. 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr Biol 136:464-73.

Krings M, Salem AE, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, Welsby D, Di Rienzo A, Utermann G, Sajantila A, Pääbo S, Stoneking M. 1999. mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? Am J Hum Genet 644:1166-76.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonné-Tamir B, Sykes B, Torroni A. 1999. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 641:232-49.

Non AL, Al-Meeri A, Raaum RL, Sanchez LF, Mulligan CJ. 2011. Mitochondrial DNA reveals distinct evolutionary histories for Jewish populations in Yemen and Ethiopia. Am J Phys Anthropol 1441:1-10.

Ottoni C, Martínez-Labarga C, Loogväli EL, Pennarun E, Achilli A, De Angelis F, Trucchi E, Contini I, Biondi G, Rickards O. 2009. First genetic insight into Libyan Tuaregs: a maternal perspective. Ann Hum Genet 73Pt4:438-48.

Poloni ES, Naciri Y, Bucho R, Niba R, Kervaire B, Excoffier L, Langaney A, Sanchez-Mazas A. 2009. Genetic evidence for complexity in ethnic differentiation and history in East Africa. Ann Hum Genet 73Pt6:582-600.

Rando JC, Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM, Bandelt HJ 1998. Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. Ann Hum Genet 62Pt6:531-50.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 675:1251-76.

Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, Costa MD, Musilová E, Macaulay V, Richards MB, Cerny V, Pereira L. 2011. The Expansion of mtDNA Haplogroup L3 within and out of Africa. Mol Biol Evol 293:915-27.

Stevanovitch A, Gilles A, Bouzaid E, Kefi R, Paris F, Gayraud RP, Spadoni JL, El-Chenawi F, Béraud-Colomb E. 2004. Mitochondrial DNA sequence diversity in a sedentary population from Egypt. Ann Hum Genet 68Pt1:23-39.

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL. 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 2410:2180-95.

Watson E, Forster P, Richards M, Bandelt HJ. 1997. Mitochondrial footprints of human expansions in Africa. Am J Hum Genet 613:691-704.

All additional Supporting Information may be found in the online version of this article on:
http://onlinelibrary.wiley.com/doi/10.1002/ajpa.22212/suppinfo

4.4. -The family name as sociocultural feature and genetic metaphor: from concepts to methods.

4-1-2012

The family name as socio-cultural feature and genetic metaphor: from concepts to methods

Pierre Darlu
UMR7206, CNRS, Muséum National d'Histoire Naturelle, Université Paris 7 Paris

Gerrit Bloothooft
Utrecht University, Utrecht institute of Linguistics

Alessio Boattini
Dipartimento di Biologia E.S., Area di Antropologia, Universita di Bologna

Leendert Brouwer
Meertens Institute KNAW, Amsterdam

Matthijs Brouwer
Meertens Institute KNAW, Amsterdam

See next page for additional authors

Authors
Pierre Darlu, Gerrit Bloothooft, Alessio Boattini, Leendert Brouwer, Matthijs Brouwer, Guy Brunet, Pascal Chareille, James Cheshire, Richard Coates, Paul Longley, Kathrin Dräger, Bertrand Desjardins, Patrick Hanks, Kees Mandemakers, Pablo Mateos, Davide Pettener, Antonella Useli, and Franz Manni

The family name as socio-cultural feature and genetic metaphor:

from concepts to methods

Pierre Darlu (1), Gerrit Bloothooft (2,3,4), Alessio Boattini (5), Leendert Brouwer (3), Matthijs Brouwer (3), Guy Brunet (6), Pascal Chareille (7), James Cheshire (12), Richard Coates (8), Paul Longley (12), Kathrin Dräger (9), Bertrand Desjardins (10), Patrick Hanks (8), Kees Mandemakers (4), Pablo Mateos (12) Davide Pettener (5), Antonella Useli (5, 11), and Franz Manni (1)

(1) UMR7206, CNRS, Muséum National d'Histoire Naturelle, Université Paris 7 Paris
(2) Utrecht University, Utrecht institute of Linguistics
(3) Meertens Institute KNAW, Amsterdam
(4) International Institute for Social History KNAW, Amsterdam
(5) Dipartimento di Biologia E.S., Area di Antropologia, Università di Bologna
(6) UMR CNRS 5190 – Université Lyon 2
(7) University of Tours, France, Centre d'Études Supérieures de la Renaissance (CESR)
(8) University of the West of England, Bristol
(9) Deutsches Seminar, Albert-Ludwigs-Universität, Freiburg im Breisgau
(10) Département de Démographie, Université de Montréal
(11) Dipartimento di Zoologia e Genetica Evoluzionistica, Università di Sassari
(12) Department of Geography / Center for Advanced Spatial Analysis, University College London (UCL)

Running title: Family names, from concepts to methods.

ABSTRACT

A recent workshop on "Family name between socio-cultural feature and genetic metaphor – From concepts to methods" was held in Paris on the 9th and 10th December 2010, partly sponsored by the Social Science and Humanity Institute (CNRS), and by Human Biology. This workshop was intended to facilitate exchanges on recent questions related to the names of persons and to confront different multidisciplinary approaches in a field of investigation where geneticists and historians, geographers, sociologists and ethnologists have all an active part. Here are the abstracts of some contributions.

In 1983, Human Biology published a special issue devoted to surnames as tools to evaluate average consanguinity, to assess population isolation and structure, and to estimate the intensity and directionality of migrations. At that time, many population geneticists made major contributions to this field, including Crow, Cavalli-Sforza, Morton, Relethford, Lasker, and Barraï (see review in Lasker, 1985, Colantonio et al., 2011).

Since then, most studies have focused on extending knowledge on population structure, isonymy, and migration. A synthesis was recently published in this journal (Colantonio et al., 2003) showing that surname methodologies have now been applied to about 30 societies all around the world. The geographic scope ranges widely, from the household or village to a whole continent. The authors also underlined the recent methods to analyze Y chromosome DNA polymorphisms which allow the examination of the degree of co-segregation of surnames and Y haplotypes, at least in the occidental naming practice.

The present workshop hoped to go beyond this, even if some presentations were closely allied to classical concerns, and to pinpoint some particularly relevant aspects in current research. There are two main strands. The first rests on the exploitation of databases that are increasing in size and exhaustiveness due to the spread of computerization. In this respect, Pablo Mateos and Paul Longley's UCL Worldnames database (http://worldnames.publicprofiler.org/), which includes about 6 million surnames registered in 26 different countries, constitutes an impressive quantity of information and a wonderful tool for future research (Mateos et al., 2011). However, the data are drawn from diverse sources depending on country, such as national electoral registers or telephone directories, raising problems of homogenization and representativeness that need discussion. Moreover, long distance comparisons between stocks of names with totally different historical and linguistic origins are also a challenge. The corpus of names described by Kathrin Dräger (*Deutscher Familiennamenatlas*) based on the telephone directory of the federal Republic of Germany in

2005 contains a set of one million different types of name for about thirty million telephone lines. These can be organized according to phonology (vowels, consonants, morphology) and to surname type (derived from place names, professions, nicknames, first names). These data allow the exploration of regional variations of names in consideration of lexis, phonology, graphemics, and morphology. Regarding the current distribution of surnames it is possible to trace ancient migratory movements in some cases. In the same vein, Gerrit Bloothooft presented the modern set of 16 million family names of the entire Dutch population collected from the Civil Registration. This includes 314,000 different surnames of which the spatial distribution can be studied online, while etymological and onomastic enrichment is available for 100,000 names. Patrick Hanks and Richard Coates's approach is quite different since they have collected names from various sources, such as ancient or recent dictionaries, primary sources of many kinds, and lists of surnames already published in England, Wales, and Scotland. This approach constitutes the *Family Names of the United Kingdom Project*. It aims to reconstruct the etymology of names and to explain their morphological variations through space and time.

Besides these attempts to draw from modern registers the largest number of surnames in wide geographic areas, the second major research strand involved a focus on historical data. The advantage of surnames over genetic data is that they can be available backward in time for consecutive generations, allowing a more accurate description of population dynamics. Thus Gerrit Bloothooft and Kees Mandemakers included information on collected life cycles of 76,000 persons born between 1811 and 1922; Guy Brunet used the almost exhaustive list of about 400,000 baptisms recorded in Québec from 1600 to 1800; and Pascal Chareille studied the surnames in the Normandy currency tax rolls between 1383 to 1515, and also exploited the household census in Burgundy between 1376 and 1610. Davide Pettener and Alessio

Boattini used the conscription list of individuals born between 1808 and 1987 in Italy's Upper Savio Valley.

The large expansion of the available data, both in time and space, has led to the development of new methods and analytical tools. Among them, and now widely used, are automatic geographic representations of surname diversity, which plot either the variations of frequency of a given name or a set of names sharing some phonetic or grammatical features (see Bloothooft's, Dräger's, and Lisa's figures). Some recent statistical methods, although not entirely new, were also presented, for example a Bayesian approach to infer the origins of migrants (Brunet et al.), Self-Organizing Maps to identify names sharing the same geographic origin (Boattini et al.), or naming network clustering into ethno-cultural groups (Mateos et al, 2011) .

Surnames are efficient markers for tracing the movements of people, and therefore most presentations focus on migration. Gerrit Bloothooft compares the distribution of birth places of current inhabitants of a given town and the corresponding distribution for their great-grandfathers. Guy Brunet discusses the origins of migrants who settled in parts of Québec between the beginning and the end of the 18th century. Pascal Chareille extracts from the household census (14th century, Burgundy) annotations indicating movements of people around Dijon. Patrick Hanks, Richard Coates, and Kathrin Dräger, thanks to their databases providing etymological information on names, can localize the most likely geographic origin of a given name.

One can foresee that the future of surname studies lies probably more in the rich information provided by the set of data preserved through the generations (one of the oldest, which include 8500 names, comes from the 9th century (Chareille, 2011) and in well-defined communities, than in the accumulation of surnames on a wider geographical scale. Moreover, the large amounts of time- and geo-referenced data that will be gathered in the future will

require new statistical methods that take into account the inescapable problems of lemmatization (the grouping together of related surnames) and sampling.

However, names are not just a way to identify individuals that is cheaper and more efficient than by analyzing Y chromosome polymorphisms. They also carry social and economical meanings that merit inclusion in any interdisciplinary approach. Historians, linguists, and geographers, as exemplified during this workshop, can play as active a role as biologists, in surname studies and population analysis. And for the future, the trend should be to expand our traditional western-centered field of investigation, in order to investigate other modes of naming in other countries that have both different cultural traditions and large amounts of available data.

6. Reconstructing past genetic structures in recently transformed populations: Surnames and Y-chromosomes in the Upper Savio Valley (Central Apennines, Italy).

[Alessio Boattini, Antonela Useli, Davide Pettener]

Many of the preceding contributors (Bloothooft et al., Brunet et al., Chareille, Coates & Hanks, Dräger) focused on the efficacy of surnames in tracing movements of people as well as in reconstructing historical changes in migration patterns and/or similarity/dissimilarity coefficients between populations. These features make surnames an interesting tool for human population genetics inferences *per se*.

Recently, in the context of molecular anthropology studies focused on the variability of the Y-chromosome – with which surnames share a patrilineal ancestry (King and Jobling, 2009) – the study of surnames found a new field of application. Most frequently, surnames have been advocated to design more careful sampling strategies (Manni et al., 2005, Boattini et al., 2010a). Surnames have been used to increase the 'archaeogenetic' power of genetic studies through the analysis of historical records and pedigrees (Bowden et al., 2008; Boattini et al., 2011). In this way, researchers were able to infer 'past' genetic structures of populations by selecting those individuals who carry surnames that were proved to be present in a certain area at the time of surname introduction. In particular, Manni et al. (2005) introduced a 'general' surname method, based on Self-Organizing Maps (SOMs), that provides an efficient identification of groups of surnames that share a geographic origin and migration history. The method was first tested in the case study of the Netherlands (Manni et al., 2005, Manni et al., 2008), then successfully replicated in microgeographic contexts (Boattini et al., 2010a, 2010b; Rodriguez Diaz & Blanco-Villegas, 2010).

Here we apply the SOMs methodology in order to unravel the genetic structure of a population that was subjected to radical transformations during the last century. The Upper

Savio Valley – a mountain population located in Italian Central Apennines – experienced a series of demographic phenomena that were common to great part of Italian mountain communities: major depopulation and migrations towards the most important urban centers. In this study, we will compare surname clusters identified by SOMs with Y-chromosome variability in the Upper Savio Valley. Our main purposes are: 1) to test the power of the SOMs method to discover 'real' (biologically significant) clusters, and, if this condition is met, 2) to search for historical changes in surname structure of the population and 3) to identify remnants of historic genetic structures within the investigated area.

The data and methods

Surname analysis is based on 10,202 records from conscription lists for the years 1828-2005, corresponding to individuals born between 1808 and 1987. Following historic/geographic criteria, the Upper Savio Valley was subdivided into five areas (A, B, C, D, E), of which A and B correspond to the main urban centers of the valley – where the great part of the population is currently settled – while C, D and E are very rural areas, that nowadays are largely deserted (Figure 16).

Surname distributions were analyzed with SOMs. The SOMs method is a clustering technique through neural networks based on "competitive learning", an adaptive process in which the cells ("neurons") simulating a neural network ("map") gradually become sensitive to different input categories (Kohonen, 1984). The main idea is that different neurons specialize to represent different types of input vectors; in doing so they interact with the neighboring neurons by means of a "neighborhood function". This procedure will result in the differentiation of the whole map-space: a) identical vectors will be mapped at the same neuron, b) slightly different ones at close neurons, while c) very different vectors will be mapped at far neurons. The shape (rectangular or square) and size (number of cells) of the

SOMs are defined by the user. The size of the map determines the maximum number of different clusters; therefore, larger maps will classify items (surnames, in this study) more accurately than smaller ones. Nevertheless, it may happen that some cells remain empty, while others collect many items. Manni et al. (2005) demonstrated that the SOMs method can be considered a "blind" automated approach to identify the geographic origin of surnames. For the study of Y-chromosome variability, we collected peripheral blood samples from 59 individuals who were selected on the basis of a) pertinence of their surname to one of the main SOMs clusters (see below), b) ascertained patrilineal residence in the Upper Savio Valley for the last three generations. For each sample, 31 binary polymorphisms (M213, M9, 92R7, M173, SRY1532, P25, TAT, M22, M70, 12f2, M170, M62, M172, M26, M201, M34, M81, M78, M35, M96, M123, M167, M17, M153, M18, M37, M126, M73, M65, M160) and 12 short tandem repeats [STRs] (DYS391, DYS389I, DYS439, DYS393, DYS390, DYS385a/b, DYS438, DYS437, DYS19, DYS392, DYS389II) were typed.

Results and Discussion

The geographic distribution of surnames was analyzed using SOMs. This revealed four main surname clusters: clusters I (33 items) and II (99 items) are mainly represented in areas C, D and E, thus these groups of surnames may be considered as indigenous to rural areas, while clusters III (72 items) and IV (125 items) are mostly found in areas A and B, thus the corresponding surnames very likely had their origin in the urban centers of the Upper Savio Valley (Figure 17). For some of these, we were able to confirm their inferred place of origin based on 16th-century surname information for two Upper Savio Valley parishes from previous research (Boattini & Pettener, 2005). As a second step, we explored diachronic changes in SOMs cluster frequencies by subdividing our data according to six 30-year

intervals (referring to the year of birth: 1808-1837, 1838-1867, 1868-1897, 1898-1927, 1928-1957, 1958-1987).

All the considered areas show a temporal increase in the degree of within-area surname diversity (Figure 16), particularly for the two more recent periods. These results were confirmed by continuous descending Fst patterns for the Upper Savio Valley for the whole historic interval considered (results not shown) and suggest that our population was characterized by considerable internal mobility (in particular towards the urban areas). These results suggest strongly that social-cultural factors gave rise to a reproductive barrier between inhabitants of the chief towns and those of the surrounding areas, despite their sharing the very same environment. Nevertheless, historical changes in SOMs cluster frequencies and Fst show a shift towards a higher degree of surname homogeneity between areas, meaning that the reproductive barrier has been disappearing, especially during the last two periods (i.e. the second half of the 20th century). Unfortunately, our study was not able to discriminate between monophyletic and polyphyletic surnames, as was the case for Manni et al. (2005), but this was expected given the microgeographic setting of this research; regarding this last point, analogous results were obtained for the Alpine isolate Val di Scalve (Boattini et al., 2010a).

The next step of our research was to verify if SOMs results were confirmed by Y-chromosome analyses. The 59 total samples were divided into two groups corresponding to: 29 individuals whose surnames are included in clusters I and II (rural), and 30 individuals whose surnames are included in clusters III and IV (urban). While haplogroup frequencies between the two sub-populations were not significantly different (with the exception of haplogroup G, that was found almost exclusively in the urban sub-population) (Figure 17), Fst calculations based on STR haplotypes revealed a slight but significant differentiation (Fst = 0.022, p = 0.02). This means that these differences lay mainly within haplogroups, as is

clearly demonstrated by a network representation of haplogroup R1b1-P25 (Figure 2), the most widespread in the Upper Savio Valley, to which corresponds Fst = 0.074, p = 0.02. "Urban" haplotypes mostly cluster in the same branch of the network, while "rural" ones form different branches (stemming from the same "urban" haplotype). Summing up, it seems very likely that the two sub-populations evolved from the same ancestral population, a process that – for historical reasons – probably had its origins during the late middle ages.

In conclusion, we can affirm that surname results, as obtained with the SOMs, are confirmed and enhanced by Y-chromosome data. Furthermore, the combined use of cultural markers (surnames) and molecular markers (Y-chromosomes), enabled us to bring to light a 'fossil' reproductive barrier between two different groups of individuals – urban and rural ones – within the same population and environment. The demographic changes that intervened during the studied period and in particular in the second half of the 20th century (increased population mobility, depopulation of the rural areas), caused that barrier to disappear. At a more general level, this study underlines the contribution that surname analysis can bring to molecular anthropology studies and in particular to those aimed at the reconstruction of genetic histories of populations.

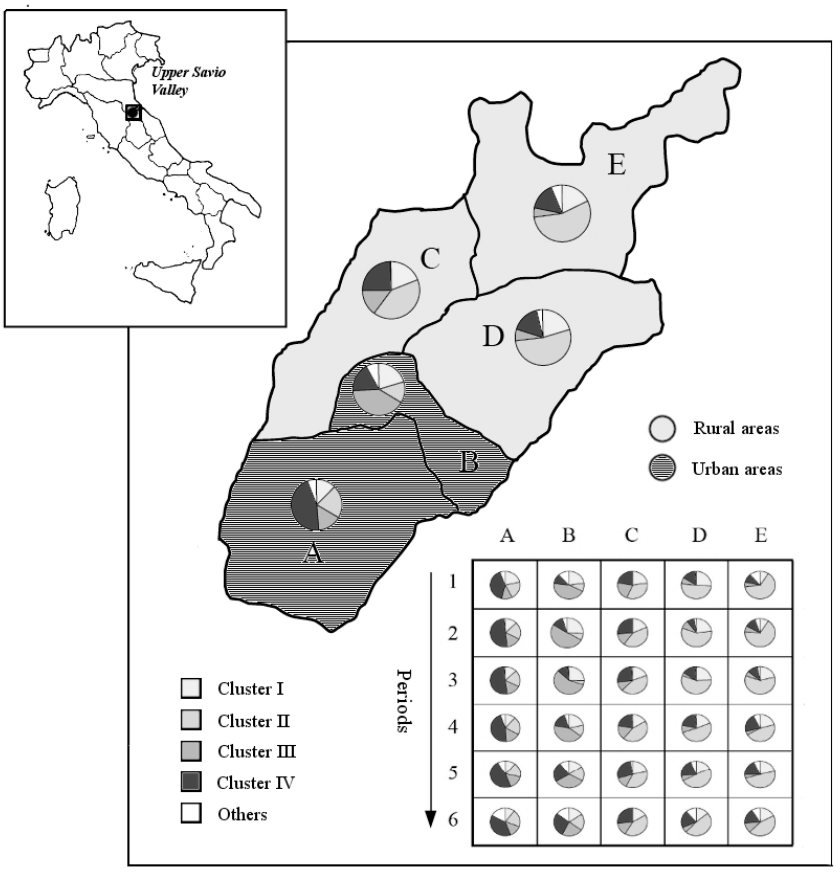


Figure 16. Geographic location and frequencies of the main surname clusters from SOMs with their temporal changes (right, below) in the Upper Savio Valley.

	Frequencies	
Hg	*Rural areas*	*Urban areas*
E1b1b1a-M78	0.07	0.00
E1b1b1b-M81	0.03	0.03
G-M201	0.03	0.23
I*(xI2a2)-M170	0.00	0.07
J*(xJ2)-12f2	0.03	0.00
J2-M172	0.21	0.20
R1*(xR1a,R1b1)-M173	0.03	0.00
R1a1-M17	0.03	0.00
R1b1-P25	0.55	0.43
T-M70	0.00	0.03
Sample size	29	30

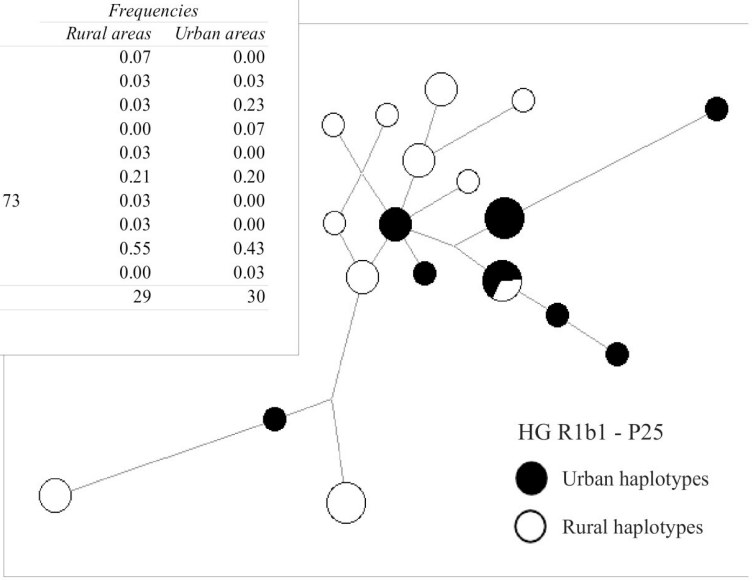


Figure 17. Haplogroup frequencies and network of the R1b1-P25 haplogroup in the rural and urban sub-populations.

Final remarks

Here, we have provided an overview of some ongoing research about surnames to understand population dynamics in Western Europe and Canada. The rather narrow geographical area addressed here is explained only by the venue of the workshop (Paris). While similar studies are conducted in other regions and continents (see Colantonio et al. 2003; and Mateos *forthcoming*), we think that the examples presented here are representative of the kind of research questions that surnames allow; questions that often go beyond the simplistic use of surnames as a proxy to Y-chromosome diversity. In any case, let us start our discussion from this traditional use of surnames, since this contribution will mainly address a readership of anthropologists and population geneticists that are directly involved in the description of human genetic variation on a world scale. From this perspective, our discussion will then expand to other disciplines and applications highlighting the clear need for increased cross-disciplinary study of population dynamics across space and time in order to better understand human diversity.

Continuous technological improvements have made possible the analysis of very large portions of our DNA. Full genome sequencing will soon become an easy and widespread technique allowing very deep inference about regional and microregional genetic differences that can be explained by demographic factors that, in turn, can rely on historical and cultural processes. Family names of patrilineal descent have proved to mirror a single locus on the Y-chromosome (King and Jobling, 2009). However, they have a temporal depth that is very limited (between 4 and ±30 generations) when compared to the scale of demographic processes inferred with molecular markers, and in any case variations in the Y-chromosome represents an extremely small amount of genetic information. *In this context, why should*

anthropologists take into consideration surname information that, albeit easier to collect than DNA data, is sometimes tricky to interpret, as it is suggested in this summary paper?

The easiest answer is that surnames allow a retrospective look at human variation. They permit comparisons between recent and ancient surname corpora, as historical documents often report surname information over several successive generations, and with a degree of polymorphism that (for the moment) is larger than that available with DNA. *Is this not similar to the scientific interest in ancient DNA technology, which is now being applied to past populations?* Once extant human diversity has been satisfactorily described (and at large geographical scales this is not too far away), one of the major questions will be to explain when and how it arose. If nowadays there are already several clues based on statistical analysis of genetic markers, direct evidence is seldom available, and ancient DNA extraction and typing will remain difficult as the molecule inevitably degrades and appropriate bones cannot be found. This is why Boattini (this paper) most appropriately uses the expression of *archaeogenetic power* to define the interest of surnames in anthropology and biodemography.

Today, in an age of global migration (Castles and Miller, 2009), surnames have indeed the potential to allow researchers an intermediate level of access to the recent past and to small geographical scales that are difficult to obtain otherwise. Interest in surname research is ultimately related to their hereditary character in most societies, but also to their group identity function (Alford, 1998), making them very useful to classify populations according to ancestral proximity. Studies in this area are all based on one simple assumption: the distribution of people's names over space and time is far from random, even in today's highly mobile societies. Therefore, surnames have already proved very useful to provide evidence of migration phenomena in different periods making it possible to identify past genetic isolates and population structures that have been modified or disappeared altogether.

This is where the potential of surnames in population studies goes well beyond the traditional paternal lineage demonstrated in Y-chromosome research.

In order to seize the opportunities lying ahead in such surname studies, more cross-disciplinary research is required that addresses the following key research challenges: a) determine the most probable geographical, temporal and cultural origin of surnames; b) distinguish polyphyletic from monophyletic surnames; c) identify common surname linages in variations of spellings; d) establish finely detailed surname frequency distribution across space and time, e) delineate areas of surname origin and barriers to cultural and population interaction, and f) combine the above advances to tease out the different population episodes that have been overlaid across space and over time. It is obvious that such scientific endeavor will only be possible through the close collaboration with disciplines outside population genetics, as most of this paper's contributions clearly show. We encourage researchers is such cognate fields to participate in the exciting challenge of improve understandings of our shared past through future contributions in Human Biology in this direction.

Acknowledgements:

Pascal Chareille describes a work in progress undertaken jointly with Denise Angers on Normandy and Patrice Beck on the Dijonnais region. He is especially indebted to Pierre Darlu for help with the analysis of the data. Alessio Boattini and Davide Pettener wish to thank the Municipality of Bagno di Romagna and the local section of AVIS (Associazione Italiana Volontari del Sangue) for their kind collaboration. The study was partly funded by the Comunità Montana dell'Appennino Cesenate.

Literature cited

Alford R (1988) *Naming and Identity: A Cross-Cultural Study of Personal Naming Practices*. New Haven, CT: Hraf Press.

Angers, D. and P. Chareille. 2010. Patronymes et migrations en Normandie de la fin du XIVe à la fin du XVe siècle : premiers résultats. In *Anthroponymie et migrations dans la Chrétienté médiévale*, M. Bourin and P. Martinez Sopena, eds. Madrid, Spain: Casa de Velázquez (Colección de la Casa de Velázquez 116), 275-316.

Ambekar A, Ward C, Mohammed J, Male S, Skiena S (2009) Name-ethnicity classification from open sources. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; June 28– July 1; Paris, France. pp 49–58. Available: http://delivery.acm.org/10.1145/ 1560000/1557032/p49-ambekar.pdf?key1 = 1557032&key2 = 1502083521&- coll = GUIDE&dl = GUIDE&CFID = 53350992&CFTOKEN=96858509 Accessed 2010 Dec 18.

Archer, Steven (2003) *The British 19th-century surname atlas*. CD-ROM. Dartford: Archer Software. [New edition (2011).]

Beck, P., and P. Chareille. 1997. Espaces migratoires et aire d'influence de la ville de Dijon à la fin du XIVe siècle. *Cahiers de Recherches Médiévales (XIIIe-XVe s.)* 3:17-32.

Beck, P., and P. Chareille. 1998. Sédentarité et mobilité à Dijon à la fin du XIVe siècle. In *La ville au moyen Âge*, vol. 2: *Sociétés et pouvoirs dans la ville*, N. Coulet and O. Guyotjeannin, eds. Paris, France: Éditions du CTHS, 95-104.

Black, George F. (1946) *The surnames of Scotland: their origin, meaning and history*. New York: New York Public Library. [Reprinted Edinburgh: Birlinn (1993).]

Bloothooft et al. (this article)

Bloothooft, G. (2011), 'Linguistics and geography, the surname case', in: W. Zonneveld, H. Quené, and W. Heeren (Eds.), *Sound and Sounds, studies presented to M.E.H. (Bert) Schouten*, Utrecht, UiL-OTS, 9-20.

Bloothooft, G. and D. Onland (2011), 'Socioeconomic determinants of first names', Names 59:1, 25-41.

Bloothooft, G. and L. Groot (2008), 'Name clustering on the basis of parental preferences', Names 56:3, 111-163

Boattini et al. (this article)

Boattini A, Griso C, Pettener D. 2010b. Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). J Anthropol Sci, 89:161-173.

Boattini A, Luiselli D, Sazzini M, Useli A, Tagarelli G, Pettener D. 2011. Linking Italy and the Balkans. A Y-chromosome perspective form the Arbereshe of Calabria. Ann Hum Biol, 38:59-68.

Boattini A, Pedrosi ME, Luiselli D, Pettener D. 2010. Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). Ann Hum Biol, 37:604-609.

Boattini A, Pedrosi ME, Luiselli D, Pettener D. 2010a. Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). Ann Hum Biol, 37:604-609.

Boattini A, Pettener D. 2005. Tra crinali e confini: mobilità matrimoniale e barriere riproduttive in Romagna Toscana (Bagno di Romagna, 1572-1930), in Breschi M, Fornasin A (editors), Il matrimonio in situazioni estreme: isole e isolati demografici, Forum, Udine, pp. 127-142.

Bourin, M., and P. Martínez Sopena, eds. 2010. *Anthroponymie et déplacements dans la Chrétienté médiévale*. Madrid, Spain: Casa de Velázquez (Collection de la Casa de Velázquez, 116).

Bowden GR, Balaresque P, King TE, Hansen Z, Lee AC, Pergl-Wilson G, Hurley E, Roberts SJ, Waite P, Jesch J, Jones AL, Thomas MG, Harding SE, Jobling MA. 2008. Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. Mol Biol Evol. 25:301-9.

Bugelski BR (1961) Assimilation through intermarriage. *Social Forces* 40: 148

Castles, S and Miller, M (2009) *The Age of Migration*, Palgrave Macmillan: London

Cavalli Sforza L.L., Menozzi P., Piazza A. 1994 The History and geography of human genes. Princeton University Press, Princeton, New Jersey, USA.

Cavalli-Sforza L.L., Moroni A. and G. Zei 2004 Consanguinity, Inbreeding and Genetic Dfift in Italy. Princeton University Press, Princeton, New Jersey, USA, pages 90-148.

Chang J, Rosenn I, Backstrom L, Marlow C (2010) ePluribus: Ethnicity on Social Networks. Procceedings of the 4th International AAAI Conference on Weblogs and Social Media23–26 May; Washington. Association for the Advancement of Artificial Intelligence (AAAI). pp 18–25. Available: http:// www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1534/ 1828 Accessed 2011 Feb 03.

Cieślikowa, Aleksandra, ed. (2007-) *Antroponimia polski od XVI do XVIII wieku* [Polish anthroponymy from the XVI to the XVIIIth century], vol. 1 (A-G); vol. 2 (H-Mą). Kraków: Lexis (2007 and 2009 respectively). [Vol. 3 is at press and vol. 4 in preparation.]

Cottle, Basil (1967) *The Penguin dictionary of surnames*. Harmondsworth: Penguin. [3rd, fully revised, edn by John Titford (2009).]

Mis en forme : Anglais (Royaume-Uni)

Chareille (this volume)

Chareille P., and P. Darlu, 2010. Anthroponymie et migration: quelques outils d'analyse et leur application à l'étude des déplacements dans les domaines de Saint-Germain-des-Près au IXe siècle. In *Anthroponymie et migrations dans la chrétienté médiévale*. M. Bourin and P. Martinez Sopena, eds. Madrid, Spain: Casa de Velázquez (Collection de la Casa de Velázquez 116), 41-73.

Chen, K. and L.L. Cavalli-Sforza. 1983. Surnames in Taiwan: interpretations based on geography and history. *Human Biology* 55: 367-374.

Cheshire, J.A., P. Mateos, P.A. Longley (2011). 'Delineating Europe's Cultural Regions: Population Structure and Surname Clustering', Human Biology 83(5):573-598

Cheshire, JA and Longley, PA (2012) Identifying spatial concentrations of surnames, *International Journal of Geographical Information Science,* 26 (2) 309-325

Dancygier, R.M. (2010) *Immigration and Conflict in Europe*, Cambridge University Press: New York.

Dammel, Antje, and Mirjam Schmuck. 2008. Der Deutsche Familiennamenatlas (DFA): Relevanz computergestützter Familiennamengeographie für die Dialektgeographie. In: Elspaß, Stephan, and Werner König (eds.), *Sprachgeographie digital: die neue Generation der Sprachatlanten*, 73–104; 254–260. Hildesheim et al.: Olms.

Darlu P. and Ruffié J., 1992. L'immigration dans les départements français étudiée par la méthode des patronymes. *Population*, 3 : 719-734.

Darlu P., Brunet G., Barbero D., 2011. Spatial and temporal analyses of surname distributions to estimate mobility and changes in historical demography: the example of Savoy (France) from the XVIIIth to XXth century. In: *Navigating Time and space in Population studies*.Gutmann, M.P.; Deane, G.D.; Merchant, E.R.; Sylvester, K.M.

(Eds.)Series. International Studies in Population, vol 9, Springer, 1st Edition, 2011, XII, 245 p.

Darlu, P. and A. Degioanni. 2007. Localisation de l'origine géographique de migrants par la méthode patronymique: exemple de quelques villes de France au début du XXème siècle. *Espace géographique* 3: 251-265.

Darlu, P., Degioanni, A., Ruffié, J. 1997. Quelques statistiques sur la distribution des patronymes en France. *Population*, 3:607-634.

de Bhulbh, Seán (*c.* 1997) *Sloinnte na h-Éireann = Irish surnames*. Faing, Co. Luimnigh: Comhar-Chumann Íde Naofa. [2nd edn titled *Sloinnte uile Éireann = All Ireland surnames* (2002).]

De Felice E. 1978 Dizionario dei cognomi italiani, Mondadori, Milano, Italy.

de Woulfe, Patrick (1906) *Sloinnte Gaedheal is Gall = Irish names and surnames*. Dublin: M. H. Gill [2nd edn (1922/3).]

Degioanni, A. and P. Darlu. 2001. A Bayesian approach to infer geographical origins of migrants through surnames. *Annals of Human Biology* 28: 537-545.

Dräger, Kathrin, and Mirjam Schmuck. 2009. The German Surname Atlas Project - Computer-based surname geography. In: Ahrens, Wolfgang, Embleton, Sheila and Lapierre, André (eds.): *Names in multi-lingual, multi-cultural and multi-ethnic contact. Proceedings of the 23rd International Congress of Onomastic Sciences, August 17-22, 2008, York University, Toronto, Canada*. [CD-Rom].

Emery, R. 1952. The use of the surname in the study of medieval economic history. *Medievalia et Humanistica* 7:43-50.

Emery, R. 1955. A further note on medieval surnames. *Medievalia et Humanistica* 9:104-106.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783-791.

Grünert, Horst. 1958, *Die Altenburgischen Personennamen. Ein Beitrag zur mitteldeutschen Namenforschung*. Tübingen: Niemeyer.

Hanks, Patrick, and Flavia Hodges, eds (1988) *A dictionary of surnames*. Oxford: Oxford University Press.

Hanks, Patrick, ed. (1990) *Dictionary of American family names*, 3 vols. Oxford: Oxford University Press.

Hanks, Patrick, Peter McClure, and Richard Coates (forthcoming 2012) Family Names of the United Kingdom: a new research project in British anthroponomastics. *Proceedings of the 24th International Congress of Onomastic Sciences, Barcelona, Spain, 5-9 September 2011.*

Hellfritzsch, Volkmar. 2007, *Personennamen Südwestsachsens. Die Personennamen der Städte Zwickau und Chemnitz bis zum Jahre 1500 und ihre sprachgeschichtliche Bedeutung*. Leipzig: Leipziger Universitätsverlag.

Hey, David G. (2000) *Family names and family history.* London: Hambledon and London.

ICOS = *Proceedings* of the International Congresses of Onomastic Sciences.

INSEE (1985) *Registre français des noms patronymiques*

Jobling MA (2001) In the name of the father: surnames and genetics. *Trends in Genetics* 17: 353–357.

Karlin, S. and J. McGregor (1967). "The number of mutant forms maintained in a population." Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics,and Probability 4: 415-438.

Kedar, B. 1973. Toponymic surnames as evidence of origin : Some Medieval Views. *Viator* 4:123-129.

King TE, Jobling MA. 2009. What's in a name? Y chromosomes and the genetic genealogy revolution. Trends Genet., 25:351-360.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. Biol Cybern 43:59-69.

Kohonen T. 1984. Self-organization and associative memory. Berlin: Springer.

Kunze, Konrad, and Damaris Nübling. 2007. Der Deutsche Familiennamenatlas (DFA): Konzept, Konturen, Kartenbeispiele. *Beiträge zur Namenforschung (N.F.)* 42/2, 125–172.

Kunze, Konrad. 2004. *dtv-Atlas Namenkunde: Vor- und Familiennamen im deutschen Sprachgebiet*. 4th edition. München: dtv.

Kunze, Konrad; Damaris Nübling (eds.). Deutscher Familiennamenatlas. Berlin, New York: de Gruyter. Bd. 1: Graphematik/Phonologie der Familiennamen I: Vokalismus. Von Christian Bochenek, Kathrin Dräger (2009). Bd. 2: Graphematik/Phonologie der Familiennamen II: Konsonantismus. Von Antje Dammel, Kathrin Dräger, Rita Heuser, Mirjam Schmuck (2011).

Lakha, F., Gorman D, Mateos, P. (2011) Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland, *Public Health* 125 (10) 688-696

Lauderdale, D.S. & Kestenbaum, B., 2000. Asian American ethnic identification by surname. *Population Research and Policy Review,* 19(3) 283-300

Longley, PA and Cheshire, JA and Mateos, P (2011) Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum* , 42 (4) 506 – 516

MacLysaght, Edward (1985) *The surnames of Ireland*, 6th edn. Blackrock, Co. Dublin: Irish Academic Press.

Mandemakers, K. (2000), 'The Netherlands. Historical Sample of the Netherlands', in: P. Kelly Hall, R. McCaa & G. Thorvaldsen (ed.), *Handbook of International Historical*

Microdata for Population Research (Minnesota Population Center Minneapolis 2000), 149-177.

Manni F, Heeringa W, Toupance B, Nerbonne J. 2008. Do surname differences mirror dialect variation? Hum Biol 81:41-64.

Manni F, Toupance B, Sabbagh A, Heyer EDA-F. (2005) New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. *American Journal of Physical Anthropology.*126(2):214-228

Manni F., Toupance B., Sabbagh A., Heyer E. 2005 A new method for surname studies of ancient patrilinear population structures, and possible application to improvement of Y-chromosome sampling. Am. J. Phys. Anthropol., 126:214-228

Mateos, P. (forthcoming) *Ethnicity, geography and populations: Tracing diversity and migration through people's names*, Springer: Heidelberg

Mateos, P. (2007) A review of name-based ethnicity classification methods and their potential in population studies, *Population Space and Place*, 13 (4): 243-263.

Mateos, P., Longley, P.A. and O'Sullivan, D. (2011) Ethnicity and Population Structure in Personal Naming Networks. *PloS ONE* 6 (9) e22943

Mateos, P., R. Webber, P. Longley (2007). 'The Cultural, Ethnic, and Linguistic Classification of Populations and Neighborhoods Using Personal Names', CASA working paper 116, UCL, London http://discovery.ucl.ac.uk/3472/

McKinley, Richard A. (1975) *The surnames of Norfolk & Suffolk.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1977) *The surnames of Oxfordshire.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1981) *The surnames of Lancashire.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1988) *The surnames of Sussex.* Oxford: Leopard's Head Press.

McKinley, Richard A. (1990) *A history of British surnames.* Harlow: Longman.

McKinley, Richard A., George Redmonds, and David Postles (1973-98) A series of county-based surname studies. Oxford: Leopard's Head.

Morgan, T.J., and Prys Morgan (1985) *Welsh surnames*. Cardiff: University of Wales Press.

Nathan, M. (2011) The economics of super-diversity: findings from British cities, 2001-2006. SERC Discussion Paper Series, SERCDP0068, London School of Economics: London.

Nei, M. 1973. Genetic distance between populations. *American Naturalist* 106: 283-292.

Neumann, Isolde. 1970. *Die bäuerlichen Familiennamen des Landkreises Oschatz*. Berlin: Akademie-Verlag.

Neumann, Isolde. 1981. *Die Familiennamen der Stadtbewohner in den Kreisen Oschatz, Riesa und Großenhain bis 1600*. Berlin: Akademie-Verlag.

Nicholls, Kenneth, ed. (1994) *The Irish Fiants of the Tudor Sovereign during the reigns of Henry VIII, Edward VI, Philip and Mary and Elizabeth I*, 4 vols. Dublin: Burke.

Nübling, Damaris, and Konrad Kunze. 2005. Familiennamenforschung morgen: der deutsche Familiennamenatlas (DFA). In: Brendler, Andrea; Brendler, Silvio (eds.), *Namenforschung morgen: Ideen, Perspektiven, Visionen*, 141–151. Hamburg: Baar.

Nübling, Damaris, and Konrad Kunze. 2006. New perspectives on Müller, Meyer, Schmidt: computer-based surname geography and the German Surname Atlas project. *Studia anthroponymica scandinavica* 24, 53–85.

Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL (1987) Migration rates of human populations from surname distribution. *Nature* 329: 714–716.

Postles, David (1995) *The surnames of Devon*. Oxford: Leopard's Head Press.

Postles, David (1998) *The surnames of Leicestershire & Rutland*. Oxford: Leopard's Head Press.

Reaney, Percy H. (1958, 1976) *A dictionary of British surnames*. London: Routledge and Kegan Paul. [Third edn by Reaney and Richard M. Wilson, *A dictionary of English surnames* (1991). R&W.]

Redmonds, George (1973) *The surnames of Yorkshire, West Riding*. Oxford: Leopard's Head Press.

Redmonds, George (2002) *Surnames and genealogy: a new approach*. Bury: Federation of Family History Societies.

Rodriguez Diaz R, Blanco Villegas MJ. 2010. Genetic structure of a rural region in Spain: distribution of surnames and gene flow. Hum Biol. 82:301-314.

Rohlfs G. 1997 Studio e ricerche su lingue e dialetti di Italia, Sansoni Editore, Fireze, Italia.

Saitou, N. and Nei, M. 1987. The Neighbor-Joining Method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406-425.

Schmuck, Mirjam. 2009. Personennamen als Quelle der Grammatikalisierung. Der *ing*-Diminutiv in Mecklenburg-Vorpommern, *Beiträge zur Namenforschung (N.F.)* 44, 35-65.

Tooth, Edgar (2000) *The distinctive surnames of North Staffordshire*, 2 vols. Leek: Churnet Valley Books.

Tyler-Smith, C. & Xue, Y., 2012. A British approach to sampling. European journal of human genetics *European Journal of Human Genetics*, 20(2), p.129-130.

Walther, Hans. 1993. *Zur Namenkunde und Siedlungsgeschichte Sachsens und Thüringens. Ausgewählte Beiträge 1953-1991*. Leipzig: Reprint-Verlag.

Supprimé :

Wijsman, E., G. Zei, Moroni A., Cavalli-Sforza L.L.. (1984). "Surnames in Sardinia. II. Computation of migration matrices from surname distribution in different periods." Annals of Human Genetics 48: 65-78.

Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I, Hutnik K, Leslie S, Nicodemus K, Royrvik EC, Tonks S, Yang X, Cheshire J, Longley P, Mateos P, Groom A, Relton C, Bishop DT, Black K, Northwood E, Parkinson L, Frayling TM, Steele A, Sampson JR, King T, Dixon R, Middleton D, Jennings B, Bowden R, Donnelly P, Bodmer W.(2012). People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *European Journal of Human Genetics*, 20(2), p.203-10.

Wood, J.; Badawood, D.; Dykes, J.; Slingsby, A.(2011) BallotMaps: Detecting Name Bias in Alphabetically Ordered Ballot Papers, *IEEE Transactions on Visualization and Computer Graphics*, 17 (12) 2384 - 2391 doi: 10.1109/TVCG.2011.174

Yasuda, N., L. L. Cavalli-Sforza, Skolnick M., Moroni A. (1974). "The evolution of surnames: an analysis of their distribution and extinction." Theoretical Population Biology 5: 123-142.

Zei G., Barbujani G., Lisa A., Fiorani O., Menozzi P., Siri E. and L. Cavalli-Sforza 1993 Barrier to gene flow estimated by surname distribution in Italy. Ann Hum Genet, 57:123-140.

Zei G., Lisa A., Fiorani O., Magri C., Quintana-Murci L., Semino O. and S. Santachiara-Benerecetti 2003 From surnames to History of Y-chromosome: the Sardinian population as a paradigm. Europ J Hum Genet, 11(10):802-7.

Zei, G., R. Guglielmino, et al. (1983). "Surname in Sardinia. I. Fit od frequency istributions for neutral alleles and genetic population structure." *Annals of Human Genetics* **47**: 329-352

The article,

Darlu, Pierre; Bloothooft, Gerrit; Boattini, Alessio; Brouwer, Leendert; Brouwer, Matthijs; Brunet, Guy; Chareille, Pascal; Cheshire, James; Coates, Richard; Longley, Paul; Dräger, Kathrin; Desjardins, Bertrand; Hanks, Patrick; Mandemakers, Kees; Mateos, Pablo; Pettener, Davide; Useli, Antonella; and Manni, Franz (2012) "The Family Name as Socio-Cultural Feature and Genetic Metaphor: From Concepts to Methods," *Human Biology*: Vol. 84: Iss. 2, Article 5. Available at: http://digitalcommons.wayne.edu/humbiol/vol84/iss2/5

is reported as a part of the
Open Access Pre-Print version
http://digitalcommons.wayne.edu/humbiol_preprints/8/

the Final Published Version can be found on:
http://digitalcommons.wayne.edu/humbiol/vol84/iss2/5/

CONCLUSIONS

In the present PhD Thesis the variation of the uniparentally transmitted genetic systems as the mitochondrial DNA (mtDNA) and the non recombining region of Y chromosome (NRY) has been investigated in *Equus caballus* (*E. caballus*), in the Ovobatysciola genus and in human populations.

For each of studies presented the analyzed genetic systems resulted powerful in order to reach the specific aims: the phylogenetics and phylogeographic studies by the means of haploid systems could actually contribute to knowledge of the evolutionary history of species.