Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
Biologia Evoluzionistica e Ambientale

CICLO XXV

COORDINATORE Prof. Guido Barbujani

Spatial and temporal distribution of mitochondrial lineages in the European wild boar

Settore Scientifico Disciplinare BIO/18

Dottorando	**Tutore**
Dott. Torres Vilaca Sibelle	Prof. Bertorelle Giorgio
(firma)	(firma)

Anni 2010/2012

“– What's my destiny, Momma?

– You're gonna have to figure that out for yourself. Life is a box of chocolates, Forrest.
You never know what you're gonna get.”

Index

Thesis presentation

The use of molecular markers to distinguish genotypes have completely changed our view of nature and its processes. The main breakthrough was the use of DNA-based markers, with the invention of the Polymerase Chain Reaction (PCR). For the first time, genomic regions could be amplified from small quantities of DNA and many individuals, and the genetic variation pattern could be described to reconstruct evolutionary processes.

Further on, the access to genetic information allowed the investigation of the distribution of genetic diversity in a geographical context. In the study by Avise *et al.* (1987), the name *phylogeography* was introduced to designate the geographical placement of genealogical lineages. Since then, the field has been led by the use of mitochondrial DNA (mtDNA) to determine phylogenetic relationships among animal populations, subspecies and species, which may then be plotted on their geographical distribution (Hewitt 2001).

Phylogeography can be considered as a sub-field of a major research area, the biogeography. Phylogeography is a powerful approach for investigating a wide range of issues related to biogeography, including the relative roles of gene flow, bottlenecks, population expansion, and vicariant events in shaping geographical patterns of genetic variation. Comparing the geographical patterns and genetic variation among multiple co-distributed taxa, the biogeographical history of a certain area can be inferred. Several studies have observed a strong correlation in phylogeographical patterns among species inhabiting the same area, which can lead to the conclusion that a certain area have one single history that influenced the co-inhabiting species (Arbogast & Kenagy 2008).

Integrative approaches, such as phylogeography coupled with historical biogeography and geospatial data, can help identify how demographic events coincide with changes in landscape and environmental histories, such as climate variables and the distribution of suitable habitat over time, revealing ecological and evolutionary mechanisms that may underlie population differentiation and explain current patterns of population diversity.

In this dissertation, the demographic history of the wild boar in Europe was investigated. Combining new developed methods in biogeography and species distribution, population genetics and demographic inferences, the history of the European wild boar is suggested. An special emphasis was given to how the Last Glacial Maximum (LGM) affected the distribution of this species and the genetic diversity pattern currently observed. The dissertation is divided in six chapters: Introduction, Objectives, Methods, Results, Discussion and Conclusions.

In the Introduction, firstly a general description of the wild boar, *Sus scrofa*, is presented, including its biology and some general results obtained by previous genetic studies. Then a brief historical review on niche reconstruction methods is presented, followed by a description of recent methods that use only presence records, including the most recent develop method, Maxent, used in this thesis.

Following the introduction, the Objectives approached in this study are presented.

In the Methodology section, the data are presented. Several statistical methods in population genetics, pylogenetics and biogeography will be applied, and each of them is described in a separate subsection.

In the Results section results obtained are presented. For each method used, a separate subsection within Results was done.

In the Discussion, based on the results obtained for the wild boar, a comparison with other European mammals studies is given, with emphasis in game species. Inferences on the dynamics of this species and the consequences of the LGM are presented. Systematic inferences from the results obtained were also done, and how the genetic data can be compared with the current subspecies definition. Possible refuge areas during the LGM and the recolonization of Europe post-LGM are also discussed.

In the Conclusions section, the outcomes of the thesis are given together with some possibilities for follow up studies.

Chapter 1

Introduction

Biology

The wild boar, *Sus scrofa* (Linnaeus, 1758), has one of the most widespread terrestrial distributions of all mammals, occurring in all Europe and Asia (Figure 1). In some areas, boars can cause damages to agricultural cultivations and natural ecosystems; over-hunting and changes in land use have resulted in the range fragmentation and its extermination though the British Isles, Scandinavia, parts of North Africa, Russia, and northern Japan (Groves & Grubb 1993). In recent years, its range has been greatly expanded by humans especially after the Second World War, when the occurrence of the wild boar increased almost everywhere in Europe. Factors like global warming, changes in agricultural practices, restocking and reduced number of predators, directly influenced the recent population growth of the wild boar (Scandura *et al.* 2008).

The wild boar has by far the largest range of all pigs. It occurs throughout the steppe and broadleaved forest regions of the Palearctic, from western Europe to the Far East, extending southward as far as North Africa, the Mediterranean Basin and the Middle East, through India, Indo-china, Japan (including the Ryukyu Chain), Taiwan and the Greater Sunda Islands of southeast Asia (Groves & Grubb 1993). Besides the great geographical distribution, the Eurasian wild boar also occurs in a variety of temperate and tropical habitats, from semi-desert to tropical rain forests, temperate woodlands, grasslands and reed jungles, and often ventures onto agricultural land to forage. It has been extinct in the British Isles since sometime in the 17th century, despite attempted introductions of new stock from Europe, although recently animals have escaped from captivity and have established in the wild (there are at least three small wild populations in England, on the Kent/East Sussex border, in Dorset, and in Hereford (Oliver & Leus 2008)). It is also extinct in southern Scandinavia, over extensive portions of its recent range in west-central and eastern parts of the former Soviet Union, and in northern Japan (Groves & Grubb 1993). The species was last reported in Libya in the 1880s, and it became extinct in Egypt in 1902 (Oliver & Leus 2008).

Depending on the author classification, at least 16 subspecies of wild boar can be identified, reaching up to 26 (Genov 1999; Mayer & Brisbin Jr 2008). The most accepted subspecies list (Groves & Grubb 1993) distinguishes four “subspecies groups”, based on both geographic and morphological criteria:

1. “Western races” in Europe (*scrofa* and *meridionalis*), North Africa (*algira*) and the Middle East (*lybicus*), extending at least as far east as Central Russia (*attila* and *nigripes*);

2. “Indian races” of the sub-Himalayan region from Iran in the west (*davidi*) to north India and adjacent countries as far east as Myanmar and west Thailand (*cristatus*), and south India and Sri Lanka (*affinis* and subsp. nov.);

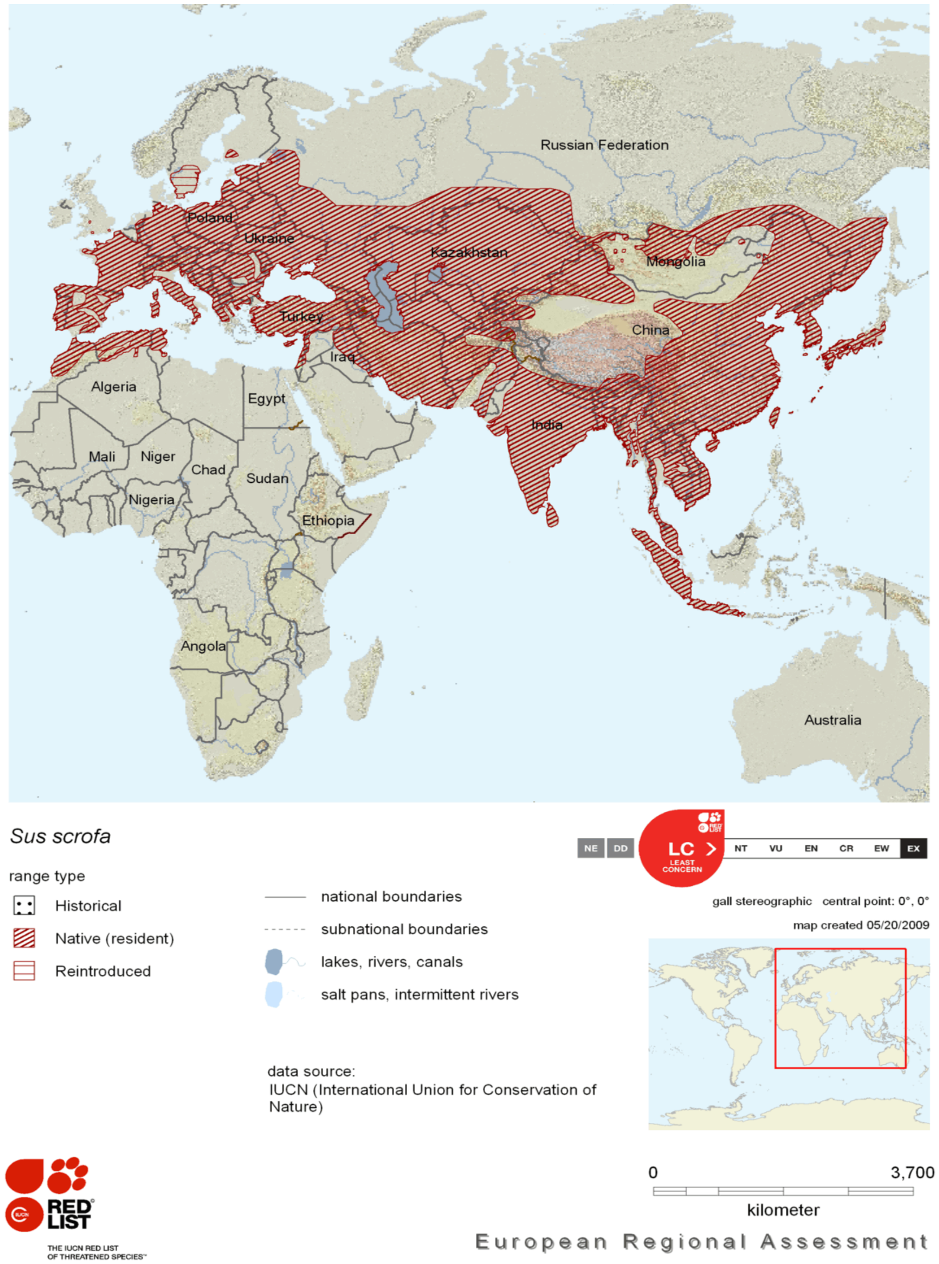


Figure 1 – Distribution of the wild boar in Europe and Asia, with indication of the range type. From Oliver & Leus (2008)

3. “Eastern races” of Mongolia and the Russian Far East (*sibiricus* and *ussuricus*), Japan (*leucomystax* and *riukiuanus*), Taiwan (*taivanus*), to south-east China and Viet Nam (*moupinensis*);

4. “Indonesian race” (or banded pig) from the Malay Peninsular, Sumatra, Java, Bali and certain offshore islands (*vittatus*).

The wild boar is omnivorous, though stomach and fecal contents analyses indicate that vegetable matter, principally fruits, seeds, roots and tubers, constitutes about 90% of the diet (Groves & Grubb 1993). It is believed that the food availability (including the presence of acorns, *Quercus* sp.) is one of the essential factors responsible for shaping year-to-year variation in wild boar population density (Melis *et al.* 2006). Analysis of long-term records of wild boar densities in the Bialowieza forest in Poland and of hunting bags in Germany showed that the presence or absence of the mast of deciduous trees, such as beech *Fagus sylvatica* L. and oak species *Quercus* L. were the dominating factor determining yearly population growth rates (Bieber & Ruf 2005).

The litter size is usually between 4 and 7 piglets in western Europe, and reports from Iraq and Armenia cite 5 to 7-10 piglets as being usual (Sommer *et al.* 2009). Litter sizes of wild boar in Mediterranean countries such as Spain or Italy are generally smaller than those observed in more eastern countries of Europe, probably due to a drier climate and lower resource availability (Servanty *et al.* 2007).

According to Groves & Grubb (1993), boars are normally most active in the early morning and late afternoon, though they become nocturnal in disturbed areas, where activity usually begins shortly before sunset and continues throughout the night. A total of 4 to 8 hours are spent foraging or traveling to feeding areas. Radio telemetry studies indicate that boars generally travel up to 10 km per night. Social groups tend to be faithful to a resting place (Goulding 1998).

In most of its range, the wild boar is considered as an environmental pest, since local populations are reported to be increasing in number, especially in countries where the Islamism is the predominant religion (Oliver & Leus 2008). Wild boars are known to cause damages to agriculture, and in some areas (e.g., in Japan) farmers are paid to hunt them (Oliver & Leus 2008). On the other hand, in some countries it is considered as a valuable game resource and populations are sustained by artificial restocking. Nevertheless, some populations of wild boar are under serious extinction risk. The Japanese Ryukyu pig (*S. s. riukiuanus*) is included in the IUCN Red List of Threatened Animals, where it has been granted the status of “vulnerable” since 1982 (Oliver & Leus 2008).

In Italy, before the 1500's, the wild boar was present throughout the country, but during the end to the 16th century due to hunting pressure it has been progressive declining. Local extinctions were reported in Trentino (17th century), Friuli and Romagna (18th century) and Liguria (1814), with the lowest densities been reported before the Second World War (Vatore *et al.* 2007). After the 1950's a general increasing of the population was observed, in part due to the recovery of forests used before for agriculture, placement of artificial feeding sites, decrease in predator numbers, global warming, reintroductions, and decreased hunting pressure (Vernesi *et al.* 2003). Currently in Italy the wild boar is present in 90 provinces out of 103, with a population estimated in 300,000 to 500,000 individuals (Vatore *et al.* 2007). The few areas that the boar is currently absent are in the Adriatic coast and in the Alpine arc in altitudes beyond the limit of vegetation trees.

Genetic studies

Cytogenetics

European and Asian boars differ in their chromosome number. Asian wild boars have 2n = 38, while Europeans have 2n = 36. The difference between the two karyotypes is a single centric fusion between chromosomes 15 and 17, resulting in the 2n = 36 karyotype. European populations exhibit numerical polymorphism with 36, 37 and 38 chromosomes. Karyotypes 2n = 37 arise from crossings between individuals with karyotype 2n = 36 and 2n = 38. The probable ancestral condition was 2n = 38, which is the most common karyotype in Asia wild (and domestic) pigs. Thus, the chromosomal number 2n = 36 is a derived state that is now very common in Europe (Scandura *et al.* 2011a).

Although no detailed study was done in a continental scale, a recent review on wild boar genetic diversity (Scandura *et al.* 2011a) showed that the distribution of the chromosomal number in Europe does not show a clear geographical border, with the Netherlands, Spain, Italy, Poland, Russia, Lithuania, Byelorussia being polymorphic for chromosome number (2n = 36, 37, 38); while Austria, Germany and France exhibited only 2n = 36; and Corsica and Yugoslavia only 2n = 38.

Genetic structure and variability in molecular markers

Phylogeographic studies have been made in populations worldwide, although only one study (Larson *et al.* 2005) investigated population in a broad geographic scale. Most phylogeographic studies used only one marker of the mitochondrial DNA (mtDNA), the control region (CR), with only a few using the mitochondrial gene Cytochrome B, nuclear microsatellites (Scandura *et al.* 2008; Vernesi *et al.* 2003) or the Y chromosome (Ramirez *et al.* 2009). Recently, also the complete genome of ten wild boars (six Europeans and four Asian) and several domestic pigs were published (Groenen *et al.* 2012). Totaling 60,000 single nucleotide polymorphisms (SNPs), a SNP chip was also developed by Amaral et al. (2011) from genome sequences of European pigs and wild boars.

The most used marker for phylogeographic studies in wild boars is the control region of the mtDNA. Studies that used this this DNA fragment were able to identify numerous geographic groups, several in Asia, one in the Middle East, and two in Europe (Figure 2). Within the Asian group at least eight different subgroups can be observed, while within the European group one large subgroup is widespread in the entire continent (E1), and another subgroup is restricted to Italy (E2) (Larson *et al.* 2005). Besides the European continent, the E1 lineage is present in the Near East (although this region has its specific clade) and North Africa, but was not observed in Asia (Hajji & Zachos 2011; Ramirez *et al.* 2009). Based on molecular clock, the time most recent common ancestor (TMRCA) between the Asian clade and E1 was dated back to 900,000 years before present (BP), while the clades E1 and E2 separated at least 50,000 years BP (Scandura *et al.* 2008).

The strikingly difference between Asian and European wild boars is also observed in the Y-chromosome (Ramirez *et al.* 2009) and nuclear microsatellites. As a general pattern, Asian boars have more diversity than their European con-specifics. According to Scandura et al. (2011a) this pattern of genetic variation may reflect past expansion dynamics from the origin area, the south-eastern Asia.

Within Europe, two main studies were done involving samples of the entire continent (Larson *et al.* 2005; Scandura *et al.* 2008). Both studies, using a small portion of the CR (663 bp and 411 bp, respectively), found one clade distributed from Portugal to Poland and one clade exclusively found in continental Italy and Sardinia (Figure 2). The European clade shows two core lineages, denominated A and C (Larson *et al.* 2005, Figure 3). The haplotypes A and C are found in high frequency in the continent and are separated by one transversion (Figure 3). In a network analysis, Larson et al. (2005, Figure 3) found that all the other E1 haplotypes are distributed in a star-like pattern around these two core haplotypes, which is consistent with a population expansion analogous to the one seen in cattle (Troy *et al.* 2001). In a meta-analysis involving recent published sequences, Scandura et al. (2011a) found that C-side haplotypes (sequences directly derived from the C haplotype)

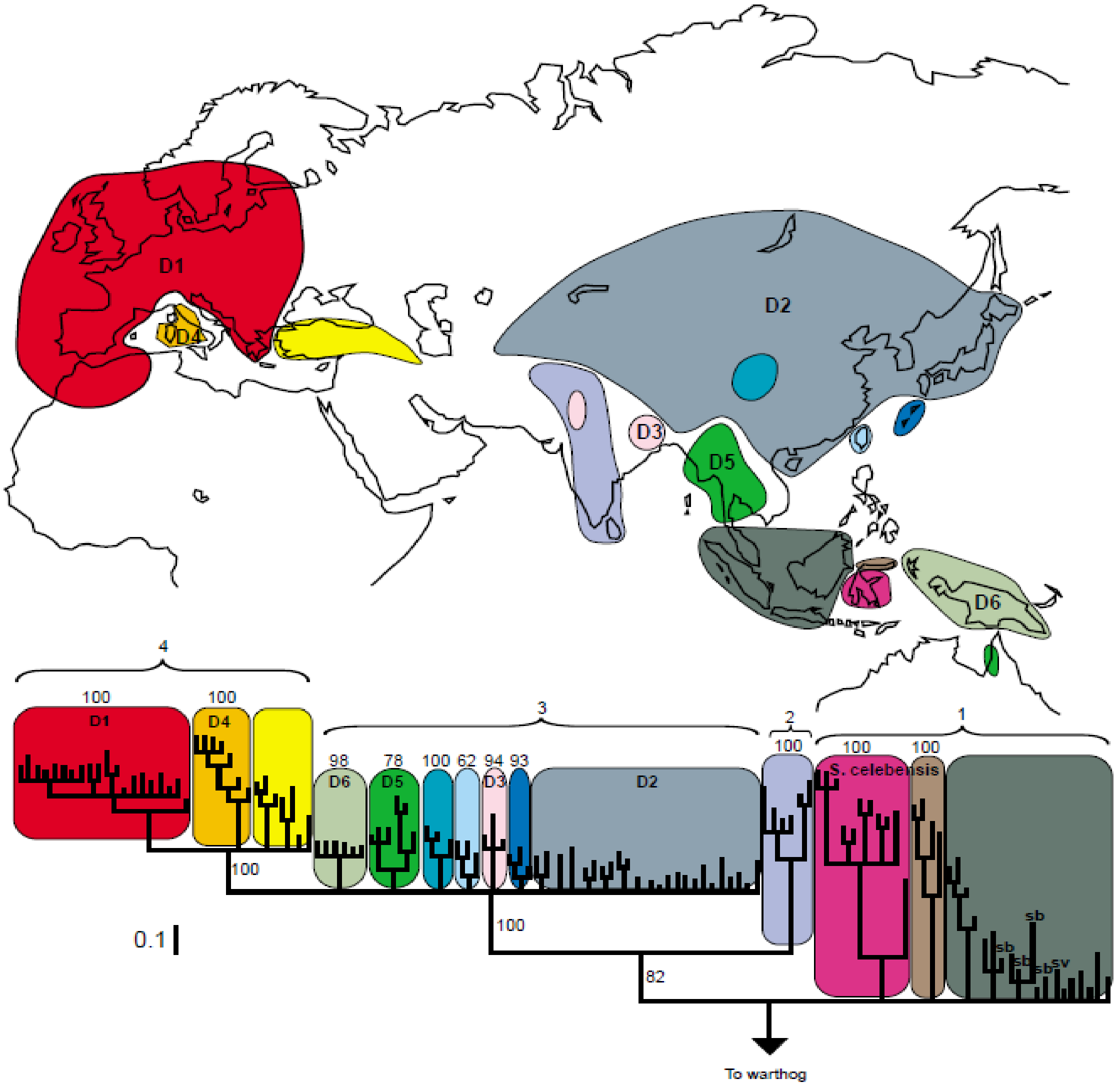


Figure 2 – Worldwide distribution of mitochondrial clades of wild boars. From Larson et al. (2005).

are more commonly found in Iberia and eastern Europe, while A-side haplotypes are found in Central Europe and Italy.

Two studies genotyped diverse microsatellite loci in continental populations. Vernesi et al. (2003) showed that Hungarian and Italian samples from natural parks (Maremma and Castelporziano) were composed by homogenous clusters, while boar from Florence, which was greatly affected by reintroductions, were genetically intermediate with several hybrids. Ramirez et al. (2009) investigated boar populations from Europe, Near East and North Africa and could not find any strong differentiation between these three regions, with several shared alleles.

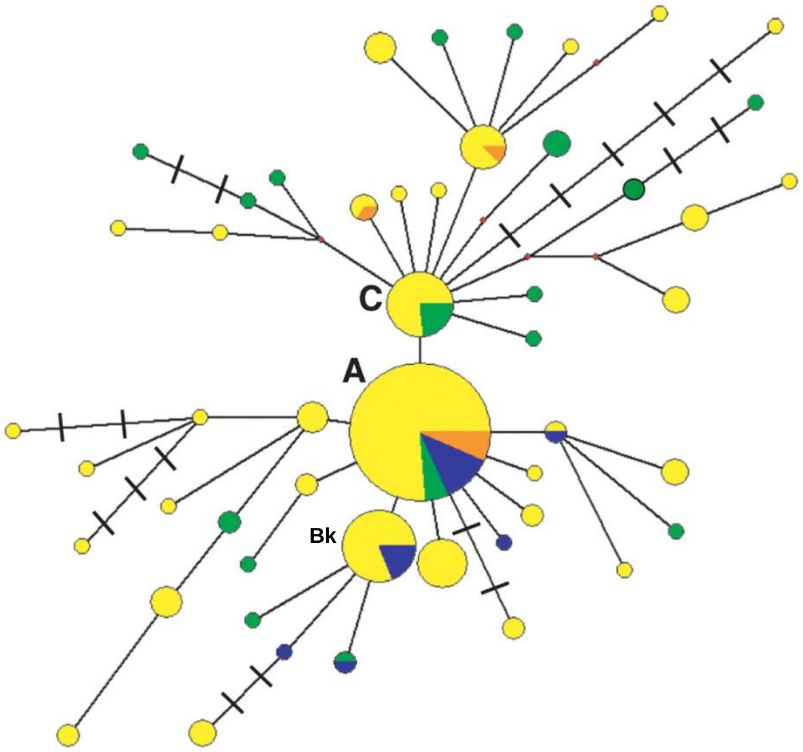


Figure 3 – Network depicting the relationship between the haplotypes found in Europe. Colors within the nodes: yellow, domestic; green, wild; orange, feral; blue, unknown status; and red, inferred intermediate haplotypes not represented by any sampled pigs. From Larson et al. (2005).

In more geographical restricted studies, Alexandri et al. (2012), Ferreira et al. (2009), Velickovic et al. (2012), Nikolov et al. (2009), Scandura et al. (2011b) and Alves et al. (2010) studied diverse European populations. All studies found local geographic groups: Ferreira et al. (2009) using six microsatellite loci found that Portuguese populations could be subdivided in North, Centre and South (result also supported by van Asch et al. (2012) using mtDNA). Alves et al. (2010) using 660 bp of the mitochondrial CR showed that the Iberian population constitute a different gene pool compared to other European populations. Alexandri et al. (2012) using a CR fragment of 637 bp identified a total of 62 unique haplotypes in Balkan wild boars that formed two structured groups: one in central Greece and one in north Greece/Bulgary. Interestingly the Greek island of Samos was composed by only Near Eastern haplotypes, probably due to its geographical proximity to the Anatolian Peninsula. In the three studies, only E1 sequences were found in continental Europe (with an exception of one sample from northern Greece which possessed an Asian haplotype). Within Bulgaria, Nikolov et al. (2009) found two distinct clades typing 10 microsatellites, one north and one south-west of the Thracian Valley. Using only four microsatellite loci, Velickovic et al. (2012) reported that Croatian/Serbian populations were differentiated from Bosnian wild boars. Scandura et al. (2011b) typed more than two hundred wild boars across Sardinia for 10 microsatellites and found three regionally structured groups inside the island.

The Italian clade, E2, have at least five single nucleotide changes in comparison to the rest of the European haplotypes (Scandura *et al.* 2008). Taking into consideration only modern sequences, the E2 clade is only found in the peninsular Italy and Sardinia, with an especially high frequency in two areas: Castelporziano Reserve and Maremma Regional Park. The main reason for this strong genetic discontinuity in Italy is probably the presence of the Alps, which is a physical barrier to the dispersal of individuals (Scandura *et al.* 2011a; Scandura *et al.* 2008). In contrast, Sardinian wild boars are strongly differentiated from any other continental population, including the Italian populations (Scandura *et al.* 2011b; Scandura *et al.* 2008). Using microsatellites markers (Scandura *et al.* 2011b) showed that this island population have almost "private" genetic components, constituting a different group, even if part of this strong differentiation may have arisen due to introgression with exotic wild populations or local domestic stocks. Regarding the mtDNA, Sardinian wild boars have a high frequency of private alleles (Scandura *et al.* 2008).

Several studies at different geographical scales and using different markers tried to estimate the demographic dynamics of the European populations using genetic data. For example, Ferreira et al. (2009) concluded that all three Portuguese populations distinguished by microsatellite markers showed a size decrease, and Alves et al. (2010) analyzing both Portuguese and Spanish populations with mtDNA failed to find any demographic expansion. In eastern European wild boars, Alexandri et al. (2012) failed to find any demographic expansion in Central Greece, although an excess of low-frequency polymorphisms was detected, but for the Greece/Bulgarian population it was possible to see a sign of growth/expansion. In studies using more wide range samples, Larson et al. (2005) when analyzing only E1 haplotypes found a star-like network compatible with a recent population expansion, and neutrality tests using both wild boar and domestic pigs mtDNA sequences were consistent with a population expansion. Scandura et al. (2008) found a sign of expansion when analyzing European sequences (excluding Italy), while the Italian population did not show any past demographic expansion for both E1 and E2 clades. The different results obtained by these studies in different geographical areas exemplify the complexity of the European demographic history and the need of a global analysis.

The publication of the genome of several European wild boars also shed some light into the demography of the wild boar (Groenen *et al.* 2012). Asian wild boars showed much more segregating SNPs in the genome, when compared to European. The differentiation between these two lineages is also clear at the genomic level, with 1,272,737 fixed differences between them. Phylogenomic analysis of complete genome sequences from wild boars and six domestic pigs revealed distinct Asian and European lineages, with an estimated split date during the mid-Pleistocene 1.6 – 0.8 million years ago, an estimation in accordance with previous mtDNA studies.

The contribution of ancient DNA

Only two studies involving museum/ancient samples in European wild boar were published so far (Larson *et al.* 2007a; Larson *et al.* 2005), although other studies that involves the Asian continent (Larson *et al.* 2007b; Larson *et al.* 2010) and the Middle East (Ottoni *et al.* 2012) have been published. In a worldwide phylogeographic study, (Larson *et al.* 2005) included sequences from museum samples, although most of them were from Anatolian or Asian origin. Among the European ones, 21 sequences were included, and only eight were not from Italy, Sardinia or Corsica. Within these three sites, they all showed typical European alleles (A-sided haplotypes). Within continental Italy, six sequences from the Maremma Regional Park were included: they were typical Italian although the haplotypes are not currently found in wild boars from this region. In Corsica and Sardinia, three typical A-sided European haplotypes were found.

Using ancient DNA, Larson et al. (Larson *et al.* 2007a) investigated the Neolithic transition in Europe. Diverse sites throughout Europe from 11,000 BC to 1,500 AD (spanning a total of 13,000 years) were included in this study. Differently from the modern pattern, Italian E2 haplotypes were found in Croatia 9,000 years ago; Near Eastern haplotypes were found in wild boars from 5,900 BC – 3,900 BC in Germany, France and Romania. This means that Near Eastern pigs arrived in Europe during the Neolithic, but no genetic signature is detected in either modern European domestic breeds or wild boar populations. The same is valid for the Italian haplotypes, around 9,000 years ago they could be found outside the Italian peninsula, but no genetic signature is currently detected north of the Alps. Besides these two exceptions, all other haplotypes in Europe during this time span were typical E1.

In a more detailed study, (Ottoni *et al.* 2012) investigated the history of wild boar in Near East and Anatolia. In a previous study, (Larson *et al.* 2007a) showed that by 3,900 BC, all domestic pigs in

Europe possessed haplotypes only found in European wild boar. This pattern was probably due to introgression with local female wild boar into imported domestic stocks. After the genetic turnover in Europe, ancient samples from Armenia indicated that European pigs were present in the Near East by 700 BC until the end of the Iron Age, where they replaced indigenous Near Eastern domestic pigs. Ottoni et al. (Ottoni *et al.* 2012) showed that European wild boars arrived in Near East almost 1,000 years before the estimated data from Larson et al. (Larson *et al.* 2007a). Already at the early Middle Bronze Age European haplotypes were present in Armenia, introduced by humans. They also demonstrated how the frequency of pigs with European ancestry increased rapidly from the 12th century BC, and by the 5th century AD domestic pigs exhibiting a Near Eastern genetic signature had disappeared across Anatolia and southern Caucasus.

Translocations, restocking and hybridization with domestic pigs

In the last few centuries, habitat loss and hunting pressure have led to the decline of several European populations, with extinctions in a few regions such as the British isles, Scandinavia, and several Italian and western Russian areas. In order to restock areas where the population declined, events of translocation and/or release of wild boars from other areas were done. After two centuries of demographic decline, reintroductions of the wild boar in central Italy began in the 1950's (Apollonio *et al.* 1988). Even though registers indicated that individuals from central Europe were used to restock local populations (from Hungary, Czech Republic and Poland), which almost led to the extinction of subspecies *S. s. majori* (Vatore *et al.* 2007), their effective contribution to the Italian genetic pool is not seem when analyzing mitochondrial fragments and microsatellites (Scandura *et al.* 2008; Vernesi *et al.* 2003). In the UK, the wild boar became extinct in 1705, and since then, escapes and deliberate release of captive animals (farmed in the UK since the 1980's) has led to the re-establishment of four known populations totaling 750 – 1000 animals (Hartley 2010). In Spain, the wild boar is still an important game species, with 5,000 boars being shot annually during the hunting season. Due to hunting demand, an unknown number of wild-caught or captive-bred wild boars are imported yearly for restocking, mainly from France (Fernandez-de-Mera *et al.* 2003).

The hybridization with domestic pigs are also a factor to be considered, wild boars and domestic pigs being fully capable of interbreeding. In central Italy, pigs were occasionally crossbred with wild boar in captivity and hybrids were released for hunting purposes (Randi *et al.* 1989). Genov et al. (1991)

reported that the traditional practice of rearing “domestic” pigs in semi-wild conditions in Bulgaria has resulted in their hybridizing with the wild boar populations in the eastern and north-eastern parts of that country, and that genetically pure wild boars now occur only in the south of the country, where domestic pigs are not reared in the wild.

Translocations and restocking are expected to affect the genetic variation in local populations, especially when distinct genetic groups are being admixed. For the wild boar, the effects of translocation and hybridization with domestic pigs seem to have affected the local populations in different degrees. Vernesi et al. (2003) analyzing Italian and Hungarian wild boar populations with microsatellite markers, suggested a limited impact of reintroductions, with patterns of genetic variation at a macro-geographic scale appear to have been only slightly affected by recent human manipulation, since the effects of ancient demographic events are still detectable. In Sardinia, the wild boar population is supposed to have originated when Neolithic pigs escaped from man’s control and became feral (Scandura *et al.* 2011b). In their study, Scandura et al. (2011b) found that 11% of the sampled Sardinian wild boars were immigrants or hybrids with continental ancestors.

Besides the influence on the wild boar genetic diversity, translocations and restocking might also have an implication in disease control and parasites distribution. The translocation of French wild boars into Spain for hunting purposes was associated with the transmission of diverse helminth species, once only found in translocated animals and now also found in wild ones (Fernandez-de-Mera *et al.* 2003). One of the main causes on diseases outbreaks, including swine fever, is associated with the import or translocation of boars and subsequent release into wild for sports hunting (Hartley 2010).

Climate and LGM: the impact on species distribution and genetic variation

Climatic changes in Europe during glaciations

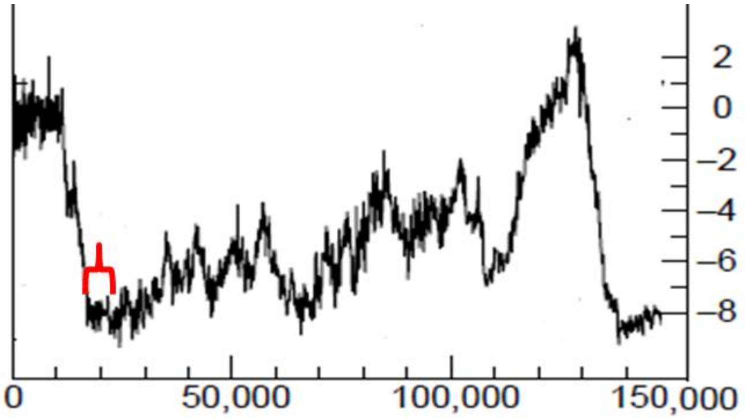
Global climate has fluctuated greatly during the past three million years, leading to recent major ice ages. The Earth’s climate became cooler through the Tertiary (64 million years (Myr)) with frequent oscillations that increased in amplitude and lead to a series of major ice ages during the Quaternary (2.4 Myr to the present). Along with other measures, the extraction of ice cores of 2 km in length are particularly important in the study of the temperature changes. The analysis of annually layered snow for entrapped gases, isotopes, acidity, dust and pollen can cover over 400 kyr (thousands of years ago)

(Petit *et al.* 1999), but most go back 125 kyr from the present to the previous (Eemian) interglacial. The Greenland (Artic) and Vostok (Antartic) ice cores are particularly informative, offering fine temporal resolution and continuity (Figure 4).

While the Antarctic ice cap grew from the Oligocene (35 Myr ago), the Arctic ice cap became established about 2.4 Myr ago. From then until 0.9 Myr ago, the ice sheets advanced and receded with a roughly 41,000 year cycle; thereafter they have followed a 100,000 year cycle and became increasingly dramatic. Such periodicity suggests a controlling mechanism, and the Milankovitch theory proposes that the regular variation in the Earth's orbit around the sun are the pacemakers of the ice-age cycles, with the 41,000 year cycle being caused by a oscillation of earth's axial tilt, and the 100,000 year cycle happening due to a eccentricity of Earth's orbit (Bennett 1990). The Milankovitch cycles control the pace of Quaternary ice ages, with the 100,000 year cycle eccentricity cycle dominate in the Late Quaternary and the 41,000 year cycle dominant in the early Quaternary.

These several climatic oscillations produced great changes in species distributions, with species going to extinction over large part of their range, or dispersing into new locations, or surviving in refugia and then expanding again. The effects of the glaciations varied according to latitude and topography (higher latitudes and altitudes suffered more from the glaciations effects), and temperate regions suffered more than tropical regions (Hewitt 2011). Regarding the European continent, the last full glacial period (from the Eemian interglacial around 135 kyr to present) is the best understood, and in particular the last warming from full glacial conditions around 18,000 years ago (18 kyr) through the current warm interglacial climate (Hewitt 1999).

Fig 4 – Temperature (°C) estimation for the last 150,000 years, from the last interglacial to the present. Temperatures are calculated as changes from the present temperature at the atmospheric level (ΔT). Values corresponding to the last glacial maximum are indicated by red curved brackets. From Petit et al. (1999).



During the Last Glacial Maxima (LGM), which occurred between 23 kyr to 18 kyr, a great part of Europe was covered by an ice sheet that extended to 52°N, and permafrost south to 47°N. All the major mountains chains had extensive ice caps. Between the main ice sheet and the southern mountain blocks (Alps, Pyrenees, Cantabrians, Transylvanians and Caucasus) was a plain of permafrost, tundra and cold steppe, which extended eastwards across Russia and the Urals (Hewitt 1999). From about 16 kyr, the climate warmed and the ice retreated, until 11 kyr when the Younger Dryas period shifted the temperatures and ice readvanced in some areas. Around 13 kyr, there is some evidence of spread of pine, oak and elm in the western Atlantic fringe to Britain, Ireland and Scotland due to water currents or animals (Hewitt 1999). Once this period came to an end around 10 kyr, the climate warmed again and around 6 kyr the vegetation pattern broadly resembled what is seen today. In the far north the Scandinavian ice sheet remained only on the highlands by 8 kyr; similarly, the glacial blocks on southern mountains like the Pyrenees and Alps had shrunk and represented less extreme barriers to cross.

The impact of climate change in species distributions

Recent analysis of deep ice cores in Greenland revealed a dramatic switch in the temperature through the ice age and the Eemian interglacial. Average temperatures would seem to have changed 10-12% over 5-10 years and lasted for periods of 70-5000 years (Hewitt 1999). Such massive temperature change combined with geography greatly influenced the species movements and distributions in Europe during the LGM. Species went extinct over large parts of their ranges, some dispersed to new locations, some survived in refugia areas and then expanded again (Hewitt 2000).

Species are expected to respond to climate variation, reducing or increasing their ranges/abundances accordingly to the temperature variation and tolerance to warm/cold weather. During the LGM, species all over the world responded to the climate change, with less or more intensity depending on the latitude and altitude (Hewitt 2000) and on their cold tolerance. In Europe, the study of the Quaternary phylogeography is based on fossil data, relying on pollen cores and beetle fossil records, to describe the changes in species distributions through the last Ice Ages (Hewitt 1999; Hewitt 2004). This is most detailed for the last glacial cycle (120,000 years ago), which is complemented by recent ice-core data on paleoclimates.

Currently, the reconstruction of past climate data in Europe is heavily based in pollen data, which directly reflects the distribution of trees during the glacial period. One of the methods of reconstruction involves the quantitative reconstruction of climatic parameters from fossil pollen data based on a modern data set (based on plants functional type and biome concepts) for calibration or on reference to modern analogs (Peyron *et al.* 1998).

Fossils from animals can also help to reconstruct how the glacial period influenced species distributions (Figure 5). Fossils are a direct evidence of a glacial refuge: only fossil remains of temperate species[1] dated to the LGM can indicate glacial refuge; the closer the fossils are dated to the culmination point of the LGM, the more significant is the evidence of a refuge (Sommer & Nadachowski 2006).

Combining pollen data, fossil distribution and DNA evidences from ancient sequences, several studies tried to recover possible refuge areas throughout Europe. Three areas are always recovered as glacial refuge for Mediterranean[1] species: Iberia, Italy and the Balkans (Hewitt 2000; Hewitt 1999; Schmitt 2007).

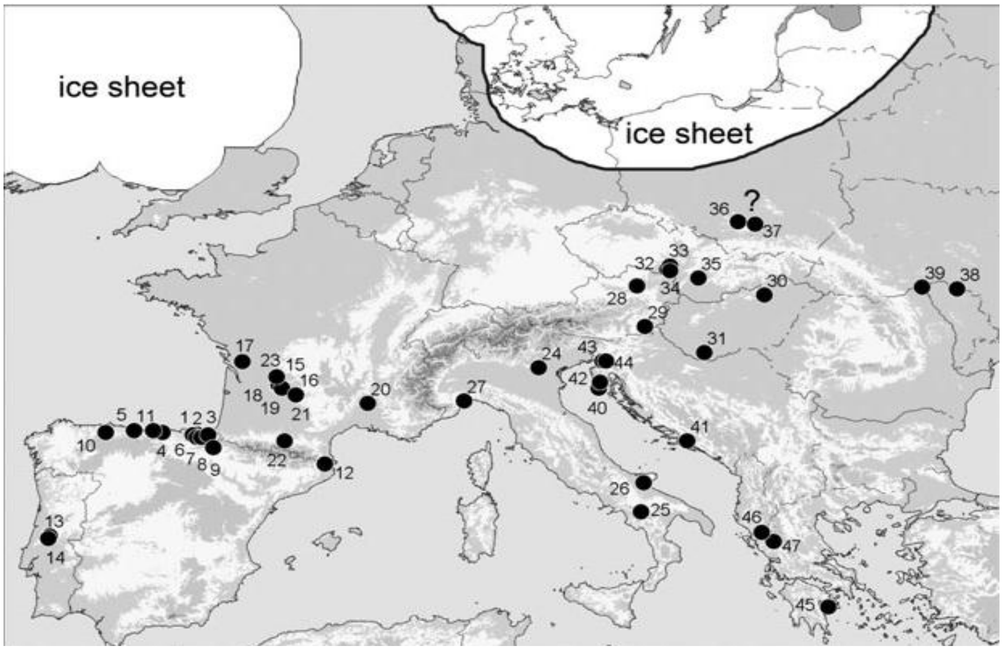


Fig 5 – Forty seven archeological sites around Europe with mammal sites associated with the LGM period (radiocarbon dating). Numbers are associated with locations defined in Sommer & Nadachowski 2006.

[1] Mediterranean species refers to the biogeographic group of last dispersal center, not the recent distribution pattern of a species - the distribution can reach Scandinavia but it is considered Mediterranean if the last refuge was in the Mediterranean basin expanding northwards in the postglacial. Schmitt T (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Front Zool,* 4: 11..

In the case of some cold-tolerant species, northern refuges can also be found, as in the case of several tree and mammal species that were found during the LGM like Slovakia, Hungary, and UK (Stewart & Lister 2001). These cryptic refugia were characterized by small tree patches in buffered local microclimates.

The population shrinking into more southern areas during the LGM is also reflected in the species genetic diversity and the current patterns of diversity distribution across the Europe. Using mainly mitochondrial DNA, several studies tried to investigate the consequences of the temperature decrease during the LGM in European species. These include invertebrates (Habel *et al.* 2010; Weigand *et al.* 2012), mammals (Rebelo *et al.* 2012; Vega *et al.* 2010) and trees (Svenning *et al.* 2011).

Migration patterns post-LGM

When the temperature started to increase again, around 16 kyrago, the ice retreated and species started to expand their ranges northwards, out of the refugia (Hewitt 1999). During this period, the temperature warmed rapidly, and populations at northern limits of the refuge range expanded into more relatively large areas of suitable territory. Each species reacted differently to this warming, according to its own responsiveness to environmental conditions such as barriers, habitat conditions and prior colonizers. The differential response was determinant on how the species expanded, which route they took based on the presence of barriers, suitable habitat, and on how long they took to expand northwards of the refugia.

Four general postglacial patterns from southern refugia are usually defined for the European populations: the grasshopper, hedgehog, bear and the butterfly (Schmitt 2007) (Figure 6). These four paradigms reflect the main expansion routes used by European species to expand to northern areas from the three refugia areas after the LGM, with direct consequences on the distribution of the genetic diversity. The hedgehog paradigm shows a postglacial expansion from all three southern European refugia, the bear paradigm reflects an expansion of the western (Iberia) and eastern (Balkans) lineages but trapping of the Italian lineage; the butterfly exemplifies an expansion of the Italian and eastern lineages but trapping of the western lineages; and the grasshopper dispersion example has one major expansion from the Balkans and trapping of Italian and Iberian lineages.

The four paradigms are frequently repeated in many animal and plant species, with various examples in the literature. One interesting pattern that arises when comparing the fours paradigms is the differential influence of barriers in the species dispersion, especially the presence of mountains and how they limited the migration from refuge areas. Carpathians, Alps and Pyrenees are the three greatest mountain chains in Europe, and directly influenced the post-colonization routes from the Balkans, Italy, and Iberia, respectively. In a review of post-migration routes, Hewitt (1999) estimated that the Alps were a greater barrier when compared to the other two mountain chains, since it blocked the expansion of 8 out of 11 species. The Pyrenees seem to be less of an impediment, blocking 4/11 species, while only two species did not expand from the Balkans but probably due to earlier colonizations from the East (Hewitt 1999).

The fact that the Alps are a greater barrier to post-glacial colonization affected current genetic distribution throughout Europe. For the majority of species studied, Italy is the refuge area with more unique haplotypes, because once the individuals were isolated in the refuge areas and started to differentiate (accumulate unique haplotypes), they remained isolated by the presence of the Alps, while individuals from the Iberian and Balkans refugia expanded north.

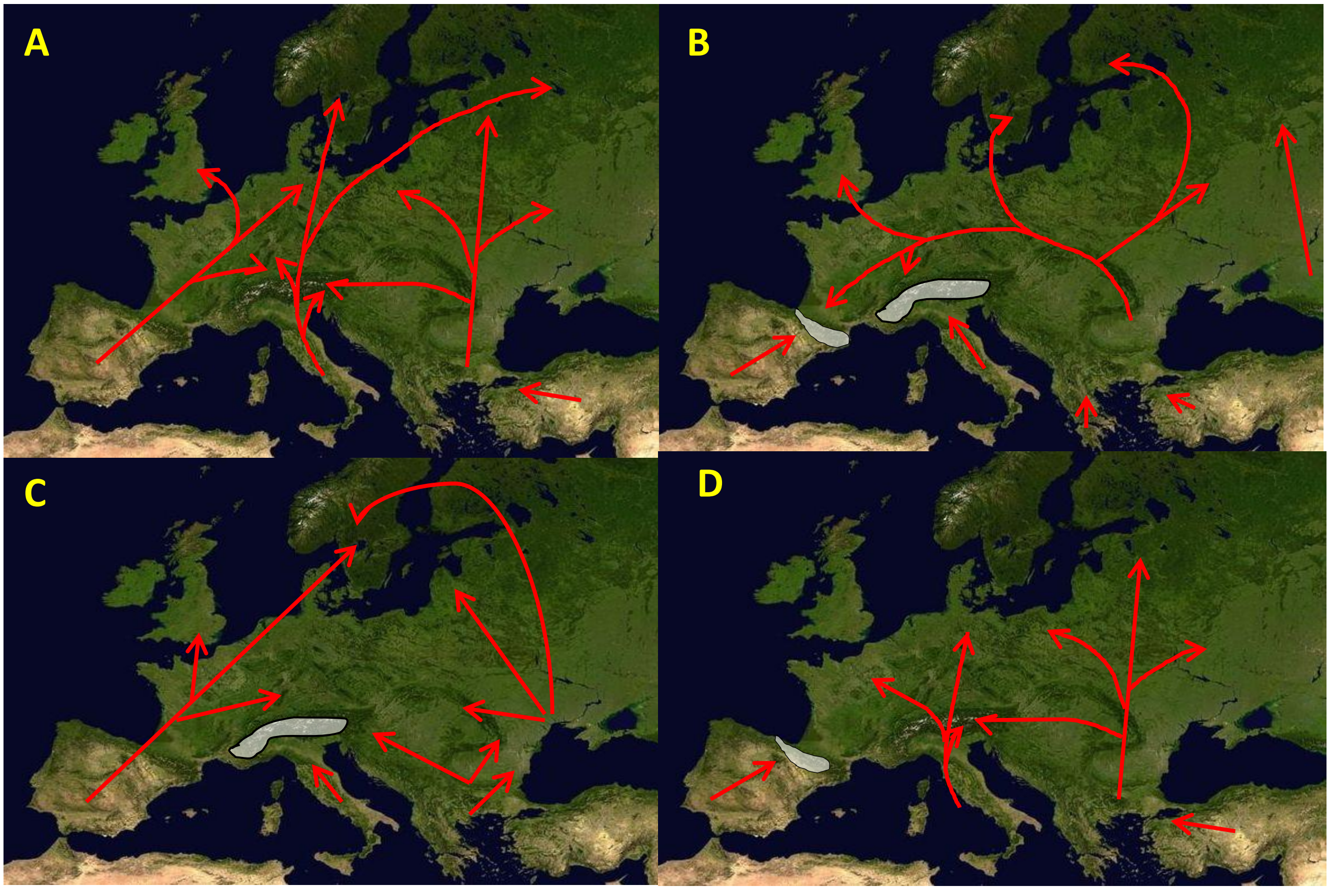


Figure 6 – Patterns of postglacial expansion from southern refugia. Red arrows represent the expansion from each of the refugia, while white forms depict putative barriers. A) Hedgehog, B) Grasshopper, C) Bear, D) Butterfly.

The rapid northward expansion across the European plains occurred when the climatic conditions became more suitable, and different genetic consequences are expected due to the presence of mountains in southern parts. The dispersal at the leading edge of a refuge would likely be by long distance dispersants that set up colonies far ahead from the main population. These pioneers could expand rapidly to fill the area before significant numbers of dispersants arrived, and so their genes would dominate the new population genetic diversity. A series of founder effects would be repeated many times during an expansion distance, leading to a loss of alleles and to higher homozygosity in the new populations. This predicts that a rapid continued expansion would produce large areas of reduced genetic diversity in northern Europe. A rapid expansion from one refugia would also hamper the expansion from other refugia, since the pioneer area would already be colonized and a migrant from the behind the front of expansion would not contribute to the new population nor influence its genetic diversity.

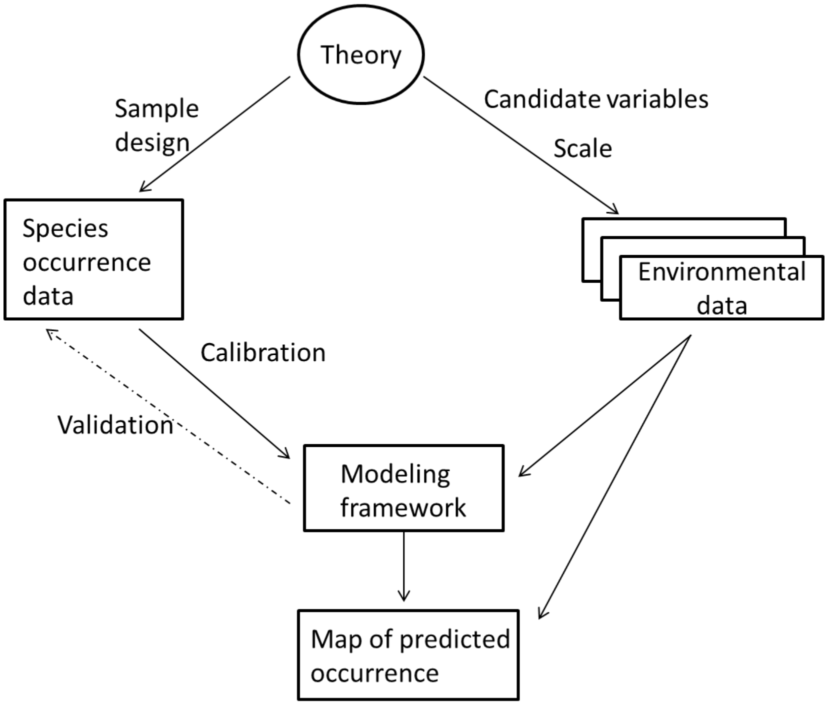
Species Distribution Models (SDMs)

Species distribution models estimate the relationship between species records at sites and the environmental and/or spatial characteristics of those sites (Franklin 2009). They are widely used for many purposes in biogeography, conservation biology and ecology, including species extinctions, speciation mechanisms, plant diversification, ecological niche conservatism, past distribution of different taxa (including extinct ones), location of Pleistocene refugia, biodiversity hotspots and historical migration pathways (Nogués-Bravo 2009).

SDMs are also referred to as habitat suitability models, describing the suitability of habitat to support a species. The concept of habitat suitability is closely related to the idea of a resource selection function from wildlife biology. A resource selection function (RSF) is any function (for example, from a statistical model) that is proportional to the probability of habitat use by an organism. If a RSF is proportional to the probability of use, then a SDM could be said to predict the likelihood of an event (species) occurs at a location, that is, the probability of species presence (Franklin 2009).

SDMs describe empirical correlations between species distributions and environmental variables. In order to successfully model a species distributions the following elements need to be followed:

- A theoretical model controls the species distribution over time and space, and at different scales, and the expected form of the response functions;
- data on species occurrence (location) in geographical space; digital maps of environmental variables representing those factor determining habitat quality (generally derived from remote sensing and stored in GIS);
- a model linking habitat requirements to the environmental variables; tools for applying the model (rules, thresholds, etc);
- data and criteria to validate the predictions and a way to interpret error or uncertainty in the analysis (see diagram below).



But after a map of predicted occurrence is drawn from the data, the modeled distribution reflects which kind of niche concept? In other words, what exactly is being modeled? This topic has been approached in various studies (Franklin 2009; Phillips *et al.* 2006; Symonds & Moussalli 2011), if the fundamental species niche (response of species to environment in absence of biotic interactions such as competition or predation), the realized niche (the environmental dimensions in which species can survive and reproduce, including the effects of biotic interactions) or the probability of habitat use (potential distribution) are the results of the SDMs. Most studies identify a description of the realized niche as the outcome of SDM (Austin 2002; Guisan & Thuiller 2005), because data on actual species occurrence are usually used in modeling, and so the model extrapolates in geographical space those conditions associated with species abundance or occurrence in the environmental "hypervolume" (Figure 7). When this realized niche described by the statistical model is mapped in geographical space it represents the potential distribution or habitat suitability (Araujo & Guisan 2006). But depending on the method of SDM, different outcomes can be proposed in relation to which kind of niche is being modeled. It has been suggested that environmental envelope-type models using presence-only data tend to depict potential distribution (suitable habitat) and are more suitable for extrapolation while more complex models that discriminate presence from absence (logistic regression, GAM, decision trees – such as random forests) tend to predict realized distributions (occupied habitats), and are more suitable for interpolation (Franklin 2009). A different view, discussed in (Phillips *et al.* 2006) suggests that models based on occurrence localities, called niche-based models, represents the realized niche since it is an approximation of the species' realized niche. Niche-based models are drawn for a source habitat, rather than a sink habitat, which may contain a given species without having the conditions necessary to maintain the population without immigration. Then, by definition, environmental conditions at the occurrence localities constitute samples from the realized niche.

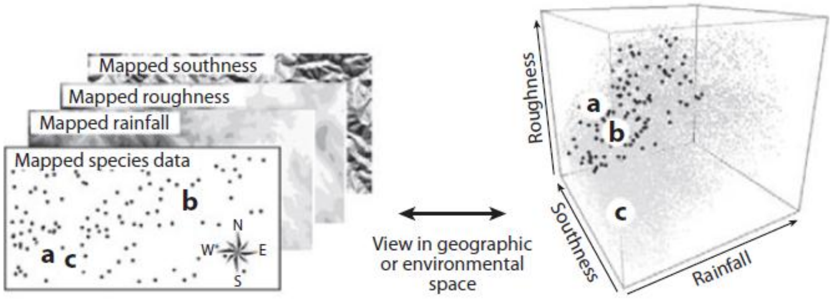


Figure 7 – Example of an environmental "hypervolume". The environmental space (right) is a consequence of mapped species (letters) in environmental data (left). Note that inter-site distance in geographic space might be quite different from those in environmental space – *a* and *c* are geographically close, but not environmentally.

The climatic niches of species are potentially the result of inheriting the climatic niches of their ancestors, and the result of adaptation of species to past and current climatic conditions that allows them to persist. One of the main theoretical assumptions for transferring the projections of SDMs through time (or extrapolation) is the temporal stability of climatic niches (Nogués-Bravo 2009). These models assume a non-significant evolutionary and/or ecological change in a species niche as a response to changing environmental conditions through time. Nevertheless, recent studies suggests that niche shifts have occurred for some species (Pearman *et al.* 2008a), due to changes in the fundamental niche or because of competition with different species over different periods of time; questioning the premise of niche stability. Whether the niche shifts are a general pattern may depend on the scale: in a geographical scale (regional vs. continental) or temporal (100 years; 10,000 years or 1,000,000 years). Although the influence of the temporal scale remains an unanswered question (Nogués-Bravo 2009), several studies that modeled the potential distribution over different time scales proposed different ways to test the niche stability over time, including the presence of fossils (Nogués-Bravo *et al.* 2008), lineage membership (Peterson & Nyari 2008), multivariate techniques (Pearman *et al.* 2008b) and climatic response curves for past and present conditions (Rodriguez-Sanchez & Arroyo 2008).

Presence-only methods

In the last two decades, there have been many developments in the field of species distribution modeling, and multiple methods are now available. A major distinction among methods is the kind of species data they use. Where species data have been collected systematically – for example in formal biological surveys in which a set of sites are surveyed and the presence/absence or abundance of species at each site are recorded – regression methods can be used. However, in most cases where only presence-data is available (no absence records), specific approaches are required (Franklin 2009).

Models that fit presence-only data predict the relative likelihood of a species presence in a given site, or the relative habitat suitability (Franklin 2009). One of the first systems used for species presence modeling, BIOCLIM (Adamic *et al.* 2010; Eckert *et al.* 2008) used a simple "hyper-box" classifier to define species potential range as the multi-dimensional environmental space bounded by the minimum and maximum values for all presences (Figure 8). In other words, the presence of a species is closed inside a box, and its limits are given by the data observed, and the results are a binary classification in suitable or unsuitable habitat. This method is current implemented in the DIVA-GIS software and is still widely used for presence estimation (Bartos *et al.* 2010; Carranza 2010; Maillard *et al.* 2010; Wotschikowsky 2010), although with some recent improvements (Franklin 2009). One refinement of this method is the HABITAT (Kusak & Krapinec 2010), which encloses the presence points in a convex hull (Figure 8).

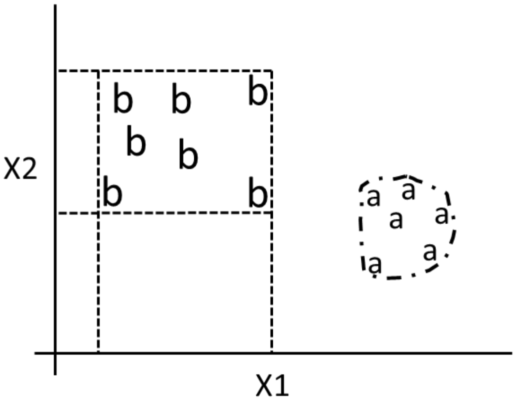


Figure 8 - Graphical representation of environmental envelopes shown for two environmental variables (X1 and X2). The *b*'s represent the model used by BIOCLIM, while the *a*'s represent the model used by HABITAT.

Genetic algorithms are also used for the estimation of a species presence. It is named because a population of classification rules is generated and then the rules "evolve" by a process analogous to natural selection until an optimal solution is reached (Franklin 2009). This approach is useful when there is a large search space for a solution and where there are complex relationships between variables. This genetic algorithm framework is implemented in the stand-alone software Desktop GARP (Genetic-Algorithm for Rule-set Production, Stockwell & Peters 1999). The classification is based in a series of population of rules generated for determining the species presence/absence. Rules are developed by searching for appropriate conditional probabilities and the resulting model is expressed in terms of conditional decision (Franklin 2009). For example, a species X is present if annual rainfall is >20 cm and average July temperature <14° C. The GARP classification that first generates a set of rules is based in four different methods, including a "BIOCLIM-type" rule. Because the output of GARP is stochastic, it is typical to run the model multiple times, and then average a subset of the best models. The proportion of models predicting presence for an observation (a single pixel) can be interpreted as the probability of occurrence.

General purpose statistical methods such as generalized linear models (GLM) and generalize additive models (GAM) are commonly used for modeling with presence-absence datasets (Franklin 2009). Recently, they have been applied to presence-only situations by taking a random sample of pixels from a study area, known as "background pixels" or "pseudo-absences", and using them in place of absences during modeling (REF).

Maxent

One other approach to the presence-only data is the Maximum Entropy modeling (MaxEnt, Phillips *et al.* 2006). As a general purpose, it is a method for making predictions from incomplete information (Baldwin 2009; Phillips *et al.* 2006). The idea of Maxent is to estimate a target probability distribution by finding the probability distribution of maximum entropy (i.e., that is most spread out, or close to uniform), subject to a set of constraints that represent the incomplete information about the target distribution, which is estimated from a set of species presence observations (Phillips *et al.* 2006).

Maxent has been described as a generative modeling approach that models the species distribution directly by estimating the density of environmental covariates conditional on species presence (Elith *et al.* 2011). Environmental factors relevant to habitat suitability (climate, topography, prey presence, etc) can be considered as independent variables, and in a model for estimating a given species potential distribution they can be called covariates, predictors or inputs. In the case of Maxent, basis function and other transformation of available data are termed features (the expanded set of transformations of the original covariates) (Elith *et al.* 2011). The features are formed "behind the scenes", in a similar way as in regression, where the model matrix is augmented by terms specified in the model. Since the fitted function is usually defined over many features, in most models there will be more features than covariates (Elith *et al.* 2011).

Previous descriptions of the Maxent model were based in a strict machine-learning terminology (Phillips *et al.* 2006). However, a recent paper published by (Elith *et al.* 2011) described the model using a statistical terminology and notation, and this was the terminology chosen to describe the Maxent model.

Maxent uses the conditional density of covariates at the presence sites, $f_1(\mathbf{z})$, and the background sample (a finite sample of points from the map with associate covariates), to estimate the ratio $f_1(\mathbf{z})/f(\mathbf{z})$. It does this by making the estimate of $f_1(\mathbf{z})$ that is consistent with the occurrence data; despite many possible distributions, it only chooses the one that is closest to $f(\mathbf{z})$ (the marginal density of covariates across the study area). Minimizing the distance from $f(\mathbf{z})$ is sensible, because $f(\mathbf{z})$ is a null model for $f_1(\mathbf{z})$: without any occurrence data, there would be no reason to expect that a species prefer any particular environmental conditions over another, and as a consequence the prediction would be no better than predict that the species occupies an environmental condition proportional to their availability in the landscape. In Maxent, the distance from $f(\mathbf{z})$ is taken to be the relative entropy of $f_1(\mathbf{z})$ in respect to $f(\mathbf{z})$.

The use of background data informs the model about $f(\mathbf{z})$, the density of covariates in the region, and provides the basis for comparison with density of covariates occupied by the species ($f_1(\mathbf{z})$). Constraints are imposed so that the solution is one that reflect information from the presence records. For example, if one covariate is summer rainfall, then constraints ensure that the mean summer rainfall for the estimate of $f_1(\mathbf{z})$ is close to its mean across the locations with observed presences. The species' distribution is thus estimated by minimizing the distance between $f_1(\mathbf{z})$ and $f(\mathbf{z})$ subject to constraining the mean summer rainfall estimated by f_1 (and the mean of other covariates) to be close to the mean across presence locations (a schematic representation is shown in Figure 9).

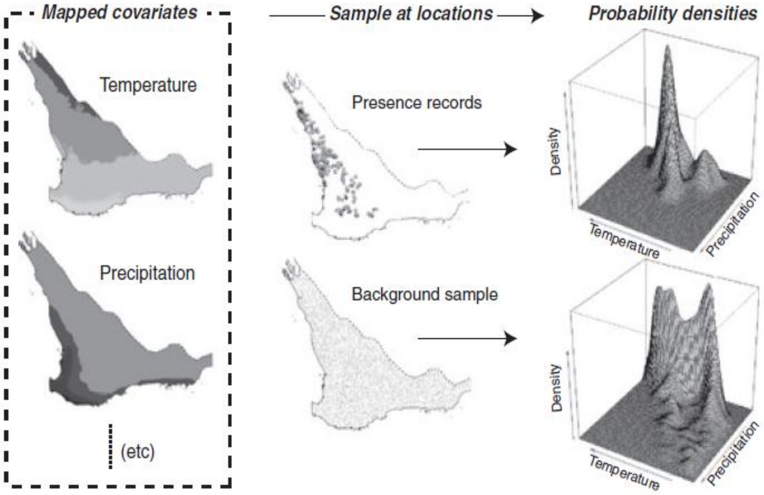


Figure 9 – Schematic representation of the probability densities relevant for the model estimation in Maxent. The maps on the left are examples of covariates. In the center are the locations of the presence and background samples. The density estimates on the right show the distribution of values in covariate space for the presence (top) and background (bottom). In a simple model that could represent, respectively, the densities $f_1(\mathbf{z})$ and $f(\mathbf{z})$. From Elith et al. (2011).

Phillips et al. (2006) outlined some advantages and disadvantages of Maxent for SDM compared to other methods. Maxent only requires presence data plus environmental information for the whole study area. It can use both continuous (e.g. climatic data) and categorical (e. g. predator presence/absence) data. The results are amenable to interpretation on the form of the environmental response functions. Its efficient deterministic algorithms have been developed that are guaranteed to converge to the optimal (maximum entropy) probability distribution, which also makes it very robust to limited amounts of training data (small samples). The output is continuous, allowing fine distinctions to be made between the modeled suitability of different areas. Some drawbacks of the method are: it is not a mature statistical method as GLM or GAM; its probabilistic models can lead to very large predicted values for environmental conditions outside the range present in the study area; and is not available in standard statistical packages (although a stand-alone and user friendly software was developed and recently the R package *dismo* also incorporated Maxent estimations).

Maxent has increasingly been used in ecological studies, for estimating species richness (Fløjgaard *et al.* 2009; Graham & Hijmans 2006), invasive species (Elith *et al.* 2010), climate change effects on species distributions (Symonds & Moussalli 2011), extent of occurrence (Pearson *et al.* 2007), climate constraints in species distributions (Wollan *et al.* 2008). One of the increasing uses of Maxent is to estimate past species distributions, and how past climate changes affected the species distributions in confront to their current distribution (Bartos *et al.* 2010; Nogués-Bravo 2009; Rebelo *et al.* 2012; Svenning *et al.* 2011; Vega *et al.* 2010). Most studies estimate the effect of the Last Glacial Maximum, with only few studies estimating in other time periods, as the late Pliocene (Rodriguez-Sanchez & Arroyo 2008), mid-Miocene (Varela *et al.* 2010) and other periods between the last Interglacial (around 120 ka) and the present (Nogués-Bravo *et al.* 2008).

One reason why Maxent is becoming increasingly popular is because it performs extremely well in predicting occurrences in relation to other common approaches (particularly compared to GARP (Ramayo *et al.* 2011)). In studies comparing the ability of diverse modeling methods to predict potential future/past distributions (i.e. extrapolation) of diverse species, (Fang *et al.* 2009) found that Maxent performed considerably poorer than GARP (but Phillips (2008) questioned the experimental design). Elith et al. (2010) compared diverse methods with Maxent (GLM, GAM and boosted regression tree), which aim to minimize extrapolation errors and assess predictions against prior biological data. Their results show that controlling the fit of models and integrating information from mechanistic models can enhance the reliability of correlative predictions of species in novel settings.

Chapter 2

Objectives

The objectives of this thesis are:

- Describe how the mitochondrial DNA diversity if distributed throughout the European continent
- Identify structured populations/groups
- Characterize how much the European wild population is influenced by translocations and hybridization with pigs
- Check if the current subspecies classification is corroborated by mtDNA analysis
- Describe how the Last Glacial Maximum affected the wild boar populations, and how it reflects in today's genetic distribution
- Identify the main refuge areas for the wild boar during the LGM
- Characterize the populations/groups responsible for the post-LGM colonization of Europe
- Identify the climate variables that most influence the geographic distribution of this species

Chapter 3

Methods

Data collection

Several studies involving the European wild boar have been published, although no study approached the genetic diversity in Europe in a wide range. In order to achieve a full description of the genetic diversity throughout the continent, muscle samples provided by local hunters were obtained for 16 countries: Germany, Luxembourg, France, Portugal, Italy, Greece, Croatia, Bosnia, Serbia, Slovakia, Romania, Bulgaria, Poland, Belarus, Ukraine, and Russia. These samples were sequenced and the data were pooled with previous published sequences (description below) to cover the whole European continent.

The sampling areas included the range of five subspecies listed in the Mammal Species of the World database (Wilson & Reeder 2005): *S. s. scrofa* in central-western Europe, *S. s. meridionalis* in Sardinia and Corsica, *S. s. majori* in central Italy, *S. s. attila* in eastern Europe, *S. s. lybicus* in the Balkans (Figure 10).

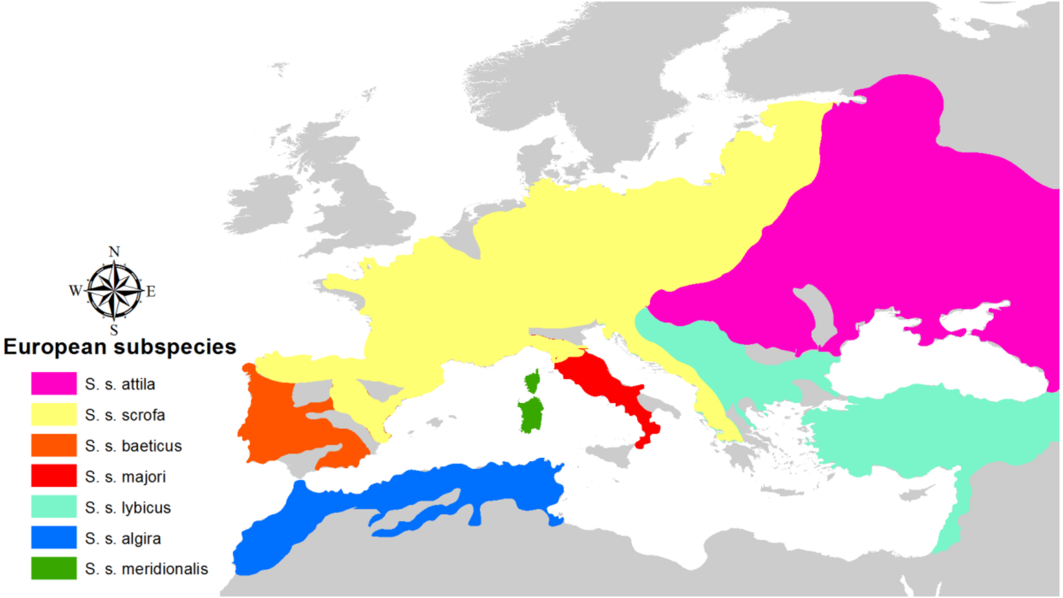


Figure 10 – European wild boar subspecies distribution. Each subspecies is shown in a different color. Subspecies ranges were taken from Groves & Grubb (1993).

Sequencing

A total of 467 wild boar specimens from 36 locations throughout Europe were sequenced. Total genomic DNA was isolated using a commercial DNA isolation kit (Sigma, Qiagen). Laboratory analyses consisted in amplification of almost the entire control region (CR) using two primers developed by (Alves *et al.* 2003): Ss.L-Dloop 5'-CGCCATCAGCACCCAAAGCT3' and Ss.Hext-Dloop 5'-ATTTTGGGAGGTTATTGTGTTGTA3'. Amplifications conditions were set as 35 cycles of 92 °C for 1 min, 62 °C for 1 min, 72 °C for 1 min, followed by a final extension step at 72 °C for 10 min. PCR products were purified by Exo/SAP digestion and a 411-bp fragment was directly sequenced using the forward primer SS.L-Dloop and the BigDye Terminatior kit version 3.1 (Applied Biosystems). Fragments were purified in columns loaded with Sephadex G-50 and run in an ABI Prism 3100 Avant automatic sequencer (Applied Biosystems).

Samples from Tunisia, previously sequenced in a partially overlapping region of the D-loop by Hajji & Zachos (2011), were sequenced for the complementary fragment in order to be fully aligned with European sequences. The sequence protocol was the same as mentioned above.

Electropherograms were visually inspected and base calls edited in FinchTV 1.2. Due to the quality of electropherograms and the shortness of the region, most sequences were obtained with a single (forward) primer. Nonetheless, to assure accuracy of nucleotide identification, a subset of samples was sequenced in the reverse direction and all individuals assigned to singleton haplotypes were sequenced twice, as well as all samples showing doubtful base calls, using the internal reverse primer (Ss.Hint-Dloop 5'-TGGGCGATTTTAGGTGAGATGGT3').

Data description

The 467 new sequences were aligned with sequences available in GenBank (Table 1). Sequences from previous studies were added, forming a database with 1099 individuals.

All 1099 sequences were aligned using the ClustalX algorithm implemented in MEGA v. 4.0 (Tamura *et al.* 2007). The downloaded sequences represented animals classified as wild *Sus scrofa* from Europe, Asia and North Africa.

Table 1 – Summary of all sequences used. *N* refers to the number of samples analysed in each study.

Study	N	GeneBank accession number
Alves et al. (2003)	7	AY232868-AY232874
Giuffra et al. (2000)	5	AF136555-56, AF136563-65
Gongora et al. (2007)	2	AF535163-AF535164
Larson et al. (2005)	65	AY884612-AY884815
Okumura et al. (2001)	164	AB015084-90, AB015094-95, AB041467-73, AB059651-52, D16483, D42171-84
Randi et al. (2007)	3	AJ314542-AJ314544
Fang et al. (2006)	36	DQ379232-DQ379267
Scandura et al. (2008)	145	EU362409-EU362553
Alves et al. (2010)	138	HM747196-HM747215
Hajji & Zachos (2011)	67	-
Kirschning et al. (unpublished data)	156	-
GenBank Total	632	
New - Total	467	
Total	1099	

Network

The alignment, corresponding to 1099 individual wild boars from three continents, was used to study intraspecific phylogenetic relationships among mitochondrial sequences and to identify lineages. Haplotypes were first collapsed for the 411 bp region in Collapse 1.2 (Posada 2011) using gaps as fifth state. Then a median-joining (MJ) network of haplotypes (Bandelt *et al.* 1999) was built in Network 4.6 (Fluxus Technologies Ltd.). In constructing the network, all polymorphic sites were considered equally informative.

Phylogeny

In order to select the most appropriate evolutionary model of nucleotide change for the D-loop sequences, the software jModeltest v. 0.1.1 (Posada 2008) was used. As outgroup, a sequence of *Sus barbatus* was included. The selection of the best model was based both on Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The best model, out of 88 tested, was the Hasegawa-Kishino-Yang (HKY) (Hasegawa *et al.* 1985) model, with gamma-distributed (G) rate variation across sites. The HKY model does not assume equal base frequencies, and accounts for the different rates of transitions and transversions.

Bayesian phylogenetic analyses were performed in MrBayes v. 3.2 (Ronquist & Huelsenbeck 2003) using the HKY+G model of sequence evolution and two independent runs of four Markov chains (1 cold and 3 heated) with 1,000,000 generations and sampling every 100 generations. The first 25% of the sampling trees and estimated parameters were discarded as burn-in. Results of log-likelihood scores were plotted against generation times to identify the point at which log-likelihood values reached stationarity. The final consensus tree was drawn in MEGA4.

European sequences

The dataset of 1099 sequences assembled includes sequences from three continents: Europe, Africa and Asia. In order to investigate the European history of the wild boar, only sequences of European origin were considerate. European sequences without spatial information were also discarded (samples from Finland and Belgium). Taking into account only European wild boars, a total of 763 sequences were obtained (467 newly sequenced, and 296 retrieved from previous published studies). These comprise 74 sampling sites from 19 countries (Figure 11).

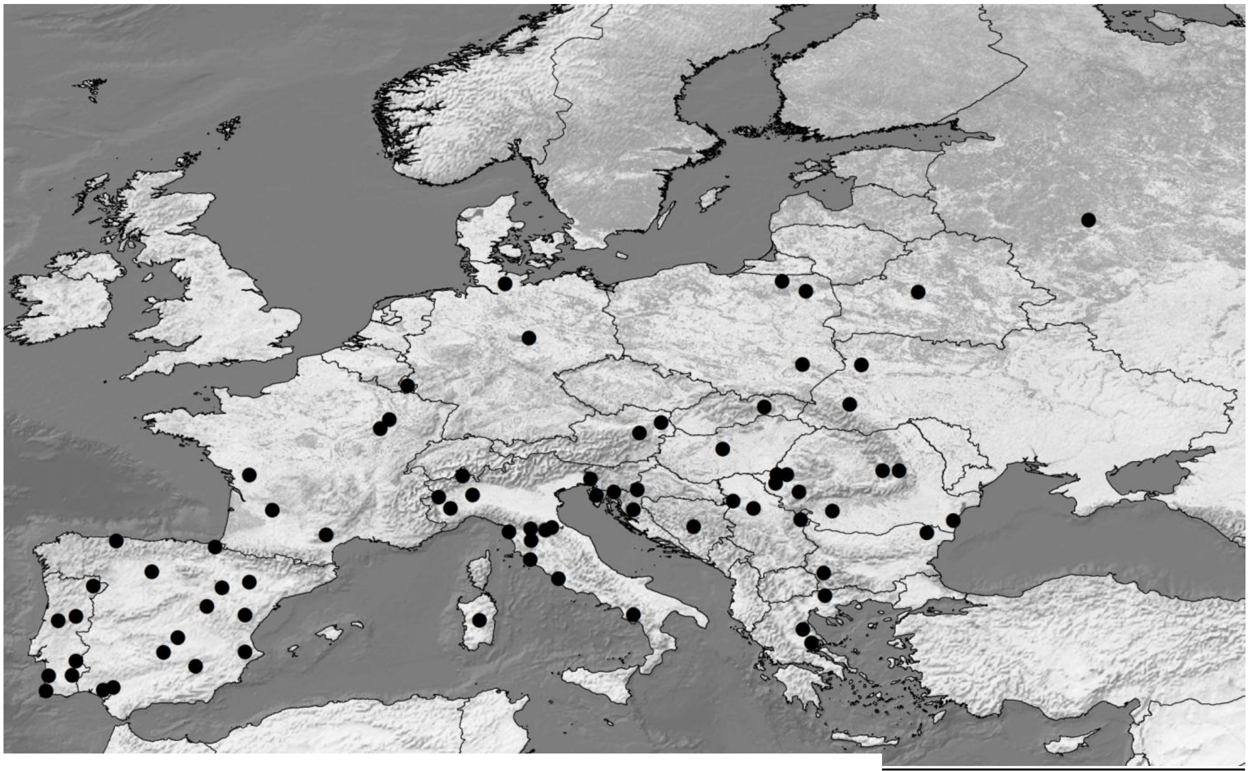


Figure 11 – Sample sites for the 763 sequences.

Since the number of sequences per sample site varied from one to 83, samples with less than 10 individuals were grouped in further analysis. In so doing, haplotype composition of each population was taken into account, to avoid that populations with different colonization histories were joined. One population in Southern Italy (ISal), having n = 7 (after removal of three Asian haplotypes), was kept separated, as its allele composition was highly different from the nearest populations. After the grouping, a total of 39 populations were obtained (Figure 12).

The presence of Asian sequences in European wild boars is considered as a signal of hybridization with domestic pigs. Asian haplotypes belonging to clade A are present at a frequency as high as 29% in European domestic breeds (Fang & Andersson 2006). This is mainly due a historical introgression with Chinese domestic pigs during the 18-19th century (Giuffra *et al.* 2000). Given that four Asian haplotypes were found in the European wild boar (three in Italy and one in Luxembourg), in analysis involving the 39 populations or strictly wild boar samples, this four haplotypes were excluded since they come from a hybridization with domestic pigs, totaling 759 samples.

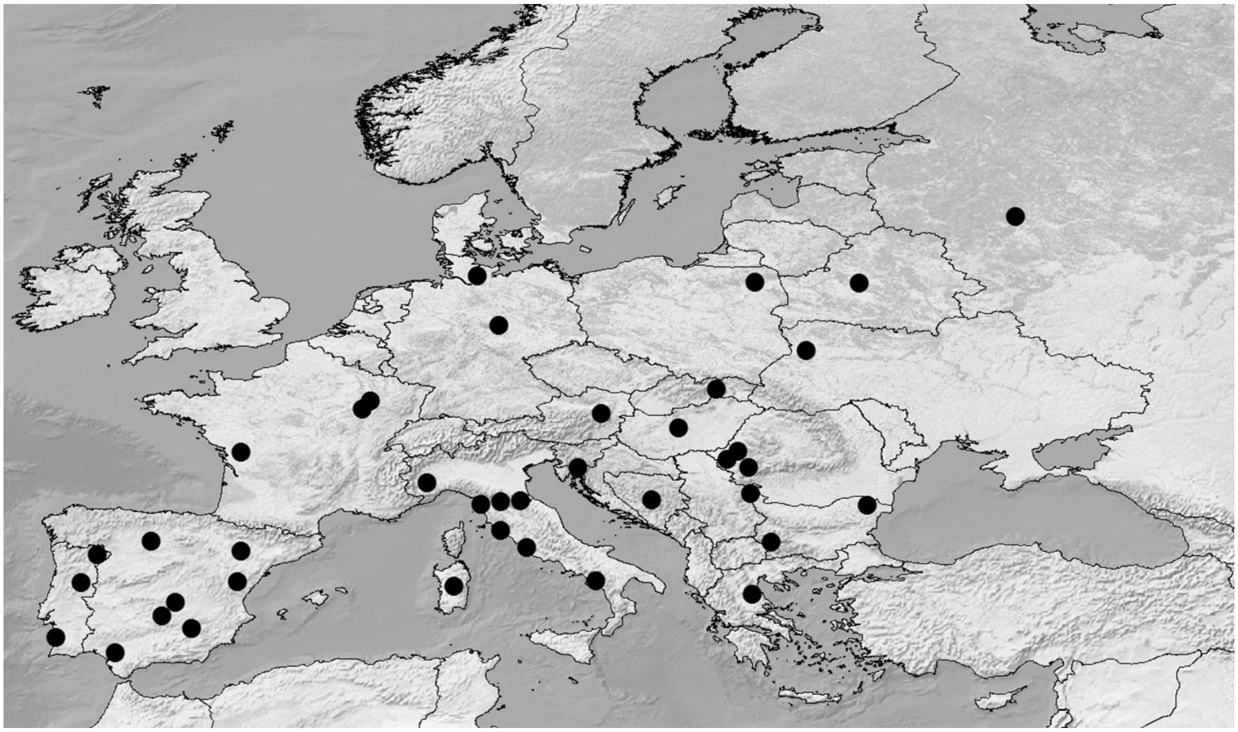


Figure 12 – Sample sites for the 39 populations.

Nested Clade Phylogeographic Analysis

To group haplotypes into geographically coherent haplogroups and see the relationships between the haplotypes, a Nested Clade Phylogeographic Analysis (NCPA, (Templeton *et al.* 1995)) was carried out with the AneCa platform (Panchal 2007), following the procedure described in (Templeton 2005). All 763 were used in this analysis (European and Asian haplotypes), since also the relationships between haplotypes were analysed. The first step in the NCPA is to generate a haplotype network including only connections with >95% probability to represent single substitutions (i.e., <5% chance of multiple substitutions). Once the parsimony network was created, possible loops may arise. To resolve loops, a subset of samples representing different D-loop haplotypes were sequenced for a portion of the Cytochrome B gene (cyt B). Between two equally likely connections, that linking two D-loop haplotypes sharing the same cyt B haplotype was maintained. The remaining loops were resolved using the criteria proposed by (Crandall & Templeton 1993).

The next step of the NCPA is a hierarchical nested design, where single haplotypes representing 0-step clades are grouped into 1-step clades, which in turn are then grouped into 2-step clades and so on, until the last nesting level corresponds to the whole cladogram. Clade distances D_c (geographical range of a given clade), D_n (geographical distance between a given clade and other clades joining the next higher nested clade) are then computed. Geographical coordinates of sampling sites were used to compute distances. Robustness of the association between nesting pattern and geography is tested over permutations of clades against sampling locations. Significance for a nested clade may be interpreted as a non-random association between position of those haplotypes in the cladogram and their geographical distribution. NCPA was not used to infer spatial/demographic processes, as many criticisms have been raised on this part of the analysis (Knowles 2008).

Clustering

Genetic differentiation among populations was anlaysed with a hierarchical clustering method implemented in the package FactoMineR (Le *et al.* 2008) based on the software R (R Core Team 2012). A Hierarchical Clustering on Principal Components (HCPC) was performed using the allele frequencies for each of the 39 populations. The HCPC combines three kinds of methods in order to establish the relationship between individuals/groups: a principal components method is used as pre-processing to evidence the similarities of single populations in the Euclidean space, a partition is performed to agglomerate the populations into groups, and a hierarchical clustering method (represented as a dendrogram) is calculated from the centers of gravity of the partition weighted by the number of populations of each cluster.

A Principal Components Analysis (PCA) was performed prior to the hierarchical clustering, since PCA can be considered as a “denoising” method, with the first dimensions extracting the essential information from the data, allowing the clustering to be more stable than the one obtained with the original distances. Therefore, the hierarchical clustering was performed using the results from a PCA.

When considering the partitioning of clusters, the hierarchical trees are built with Ward's criterion, which decomposes the total variance (multidimensional variance, or inertia) into among-groups and within-groups partitions. This method aggregates clusters minimizing the growth of within-group variance (in other words minimizing the reduction of the between-inertia) at each step of the algorithm. The within inertia characterizes the homogeneous of a cluster. The hierarchy is represented by a dendrogram which is indexed by the gain of within-inertia (Le *et al.* 2008).

From the hierarchical tree built using the Ward's criterion, the number of clusters can be chosen based on the growth ratio of within-group variance by observing the shape of the tree, by choosing the value of inertia growth that minimizes Q clusters or using K-algorithms.

When using the division into Q clusters, a division into Q clusters when the increase of between-inertia between $Q - 1$ and Q clusters is much greater than the one between Q and $Q + 1$ clusters. An empirical criterion can formalize this idea. Let $\Delta(Q)$ the between-inertia increase when moving from $Q - 1$ to Q clusters, the criterion proposed is:

$$\frac{\Delta Q}{\Delta(Q+1)}$$

When considering the K-means, the partition obtained from the cut of the hierarchical tree is introduced as the initial partition of the K-means algorithm and several iterations of this algorithm are done. The partition resulting from this algorithm is finally retained. Usually, the initial partition is never entirely replaced, but rather improved (or "consolidated"). This improvement can be measured by inspecting the [(between inertia)/(total inertia)] ratio.

For determining the best number of clusters, both methods were taking into consideration. A minimum of 2 clusters, and a maximum of 10 were imposed as limits.

Group genetic diversity

After the definition of groups (hereafter called macro-regions), descriptive statistics were calculated for both groups and populations.

The software Arlequin v. 3.5 (Excoffier & Lischer 2010) was used to obtain the gene diversity (*Hk*, (Nei 1987)) and nucleotide diversity (π). The gene diversity is equivalent to the expected heterozygosity when using diploid data. It can be defined as the probability that two randomly chosen haplotypes are different in the sample. It can be estimated using:

$$\widehat{Hk} = \frac{n}{n-1}\left(1 - \sum_{1=1}^{k} p_i^2\right)$$

Where *n* is the number of gene copies in a sample, k is the number of haplotypes, and p_i is the sample frequency of the i-th haplotype.

The nucleotide diversity is equivalent to the gene diversity, but at the nucleotide level. It computes the probability that two randomly chosen homologous sites are different. It can be estimated by the formula:

$$\hat{\pi}_n = \frac{\sum_{i=1}^{k} \sum_{j<i} p_i \ p_j \ d_{ij}}{L}$$

Where d_{ij} is an estimate of the number of mutations having occurred since the divergence of haplotypes *i* and j; k is the number of haplotypes, p_i is the frequency of haplotype *i*, and L is the number of loci.

As sample size differed considerably among areas, an unbiased estimate of allelic richness (*AR*) was also calculated with Contrib 1.0 (Petit *et al.* 1998). Allelic richness refers to the expected numbers of alleles that a sample (a population) would have if the sample size had been g genes instead on N (Morueta-Holme *et al.* 2010). Since the allelic richness of a sample is affected by the size of that sample (large samples are expected to have more alleles), it is used a rarefaction value for estimating how many alleles are expected in a sample of a specified size. Rarefaction can be defined as a statistical method for estimating how many alleles are expected in a sample of specified size (Morueta-Holme *et al.* 2010). When analyzing the 39 populations, the rarefaction was set to 7 (smallest sample size, corresponding to the Salerno population) and for macro-regions it was set as 26 (sample size of the Greek group).

Mismatch distribution

To investigate the demographic history of the European continent, the distribution of pairwise nucleotide differences between haplotypes for each macro-region was calculated. Mismatch distributions tend to be multimodal in stationary populations, while they are unimodal in populations that have passed through a recent demographic expansion (Rogers & Harpending 1992) or through a range expansion with high levels of migration between neighbouring demes (Excoffier 2004). Two hypotheses were tested: sudden demographic expansion and range expansion.

Neutrality tests

To estimate departures from the neutrality, Fu's F_S (Fu 1997) and Tajima's D (Tajima 1989) statistics were calculated in Arlequin, as they are sensitive to departures from population equilibrium due to demographic fluctuations (but also to selection and mutation rate heterogeneity). Significance was assessed by randomly generating samples under the hypothesis of neutrality and calculating the proportion of simulated values lower than or equal to the observed one.

Interpolation of genetic data

To check the spatial trend of the genetic diversity distribution, the values obtained of gene diversity (*Hk*) and nucleotide diversity (π) for the 39 populations were interpolated. The Kriging method was chosen as the interpolation procedure since it is a geostatistical method, and creates a surface that incorporates the statistical properties of the measured data. Kriging methods can produce besides the predicted surface, probability and error surfaces, giving an indication of how good the predictions are. One of the advantages of the Kriging is the production of smoother maps from irregularly spaced data; it will try to express trends suggested by the data, so that, for example, high points might me connected along a ridge rather than isolated by bull's-eye type of contours. This is one of the reasons which makes Kriging a good method for irregularly spaced data, unlike other exact (deterministic) interpolators such as the Spline or the Inverse Distance Weighting. So far no study was published on which is the best Kriging method for estimating the surface when interpolating genetic data, so the Ordinary Kriging with Circular variogram model was used, since it produces a smoother surface and it is the most used method in GIS packages (de Smith *et al.* 2009). For the production of the rasters, the Spatial Analyst extension of ArcGIS 10 (ESRI, Redlands, CA, USA) was used.

Demography through time

To establish the population dynamics in the past, an analysis of Bayesian Skyline Plot (BSP) was done, using the software package Beast v1.6.1 (Drummond & Rambaut 2007). The BSP estimates changes in the effective population size through time, by jointly estimating the genealogy, population history and substitution-model parameters (Ho & Shapiro 2011). As the first step, the genealogy (tree topology) is estimated, along with the divergence times (node times). The population size history is then estimated from the genealogy, and possible size changes depend on the timing of the coalescent events and not on the exact genealogical relationships between the sequences. For example, coalescent events occurring in rapid succession are normally indicative of small population size. Because of the relationships between coalescent events (length of intervals) and population size, the BSP can give an estimate of the population size for each coalescent interval in the estimated genealogy producing a piecewise reconstruction of the demographic history (Ho & Shapiro 2011).

The BSP assumes that populations are isolated. The impact of violating this assumption has not been investigated in great depth, but it is likely that some bias in estimating the model parameters (such as population size and mutation rate) are introduced (Ho & Shapiro 2011). Since panmixia cannot be assumed for the wild boar, and close populations are not isolated, we run this analysis only on macro-regions. Analysis were run for 20,000,000 generations, parameters were sampled every 1000 generations, and was employed a HKY+G model of evolution with four rate categories. The convergence was checked in Tracer (Rambaut & Drummond 2007), to assure that all parameters had an ESS (Effective Sample Size) of at least 200. For the estimates we used a mutation rate estimated for other ungulates of 20% per site per million years (Birungi & Arctander 2000), and a generation time of 1.5 years (Gaillard *et al.* 1992).

Model choice with Approximate Bayesian Computation

Three macro-regions showing some differences in the demographic patterns as reconstructed by the Bayesian Skyline Plots were chosen for an additional analysis of model comparison using Approximate Bayesian Computation (ABC). The groups selected for this analysis were Italy (some evidence of recent decrease in population size), Sardinia (some evidence of recent increase in population size) and Central Europe (stable population size). The rationale behind this analysis was to check is ABC was capable to recover the same pattern observed in BSPs, and if there was enough information in the sequences to distinguish between three simple models.

ABC is a highly flexible method that allows the estimation of parameters under demographic models that are too complex to be estimated under a full-likelihood method (Beaumont *et al.* 2002). Using summary statistics the ABC method compares the observed data with datasets obtained simulating different models using prior distributions of the parameters of interest (Hamilton *et al.* 2005). This methodology permits the choice of a demographic model over another by the calculation of the probability of the most probable scenario (Bertorelle *et al.* 2010; Cornuet *et al.* 2008). This method is characterized by two main features: the use of summary statistics to summarize the data, which significantly reduces the amount of data to handle; and the use of Monte Carlo simulations that avoid the need to use explicit likelihood functions (Cornuet *et al.* 2008). Inference is usually performed in nine steps (Bertorelle *et al.* 2010), which can be broadly divided in three general ones (Cornuet *et al.* 2008). The first one is a simulation step which a *reference table* is produced, which also comprises the correct choice of priors and summary statistics for the simulations. A reference table consists of rows, which corresponds to a simulated dataset and contains the parameters values used to simulate the dataset and summary statistics computed. Parameter values are drawn from prior distributions, set by prior knowledge. In absence of information, a broad, flat distribution should be used as prior. The second step is the rejection step. Euclidean distances between each simulated and the observed dataset in the space of summary statistics are computed and only the simulated datasets closest to the observed dataset are retained. The parameter values used to simulate these selected datasets provide a sample of parameter values approximately distributed according to their own posterior distribution (Cornuet *et al.* 2008).

For each of the groups, three simple scenarios were tested (constant, sudden expansion and bottleneck), for which three parameters were estimated (Figure 13): current population size, ancestral population size and time of population size change (for expansion and bottleneck scenarios). For each scenario, one million simulations were performed for each model, producing a reference table with 3 000 000 simulated datasets. Summary statistics included number of haplotypes, number of segregating sites, mean of pairwise differences, variance of pairwise differences, Tajima's D, mean of numbers of the rarest nucleotide at segregating sites, and variance of numbers of the rarest nucleotide at segregating sites. A uniform prior was used for the current population size, *Ne*, U[100, 1 000 000], for the ancestral population size after the population size change, *Na*, U[100, 1 000 000], and for the time of the size change, *t*, U[100, 10 000]. The mutation rate assumed a gamma distribution varying between 1 x 10^{-7} to 1 x 10^{-8} (Scandura *et al.* 2008), following a HKY mutational model. For the expansion and bottleneck scenarios, a condition was set, for the expansion only parameters which Na< Ne were accepted, while for the bottleneck only Na>Ne was accepted. To estimate the posterior distributions 1% simulated datasets closest to the observed dataset were taken into consideration for the local linear regression, after applying a *logit* transformation to parameter values for the linear regression (Fagundes *et al.* 2007). This regression method uses an adjustment based on the weighted multinomial logistic regression (Bertorelle *et al.* 2010). The closest simulated datasets were also plotted in a PCA to identify how distant the simulations were from the observed dataset. All simulations and parameter estimations were performed using DIY-ABC (Cornuet *et al.* 2008).

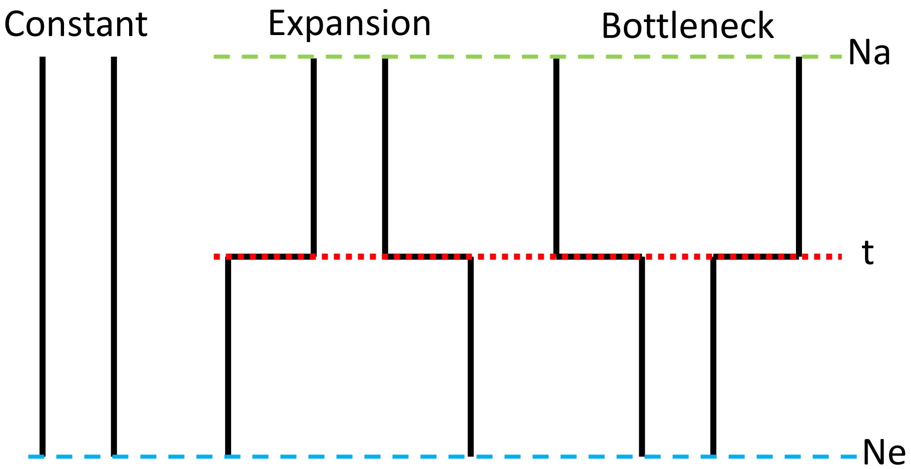


Figure 13 – Models tested for each population: constant, expansion and bottleneck. The different colored lines indicated the estimated parameters for each model (see text).

Modeling present and past (LGM) range

The ecological suitability for the species at the time of the LGM was assessed in order to identify probable refugia. The machine learning method based on maximum entropy, implemented in the program Maxent v.3.3.3 (Phillips *et al.* 2006) was used to predict the wild boar distribution during the LGM and in the present time. Maxent is a generative modeling approach that models the species distribution directly by estimating the density of environmental covariates (temperature, precipitation, etc) conditional on species presence (Franklin 2009). Since the wild boar is distributed throughout Europe, a presence-only modeling technique was more fit because no absence data is available. Therefore, the geographical coordinates of the 74 sampling sites of this study were summed to the ones used in Melis *et al.* (2006), and those available from the GBIF (Global Biodiversity Information Facility) database. Since there was an evident sample bias among the available GBIF locations, a selection was made in order to reach an even density of points across countries, and sampling the different environmental contexts where the wild boar occur today. A total of 215 unique locations were obtained (Figure 14).

As climate represents the driving factor influencing in turn other environmental variables affecting wild boar occurrence (habitat, water and food availability, etc.), climatic variables were used to construct the climate prediction models. Specifically, the following variables were used: annual mean temperature (Bio01), temperature seasonality (Bio04), temperature annual range (Bio07), mean temperature of the warmest quarter (Bio10), mean temperature of the coldest quarter (Bio11), annual precipitation (Bio12), precipitation of the wettest quarter (Bio16), and precipitation of the driest quarter (Bio17). Current and LGM data were downloaded from WorldClim v. 1.4 (Hijmans *et al.* 2005). Two different general circulation models were adopted for the LGM estimations, the Community Climate System Model (CCSM) and the Model for Interdisciplinary Research on Climate (MIROC). These two models were produced using different climate simulation methods, and therefore have slightly different values for the variables. As snow cover is a crucial limiting factor for wild boar (Melis *et al.* 2006), two variables from the Stage Three Project (van Andel 2002) were included, the snow depth (in centimeters) and number of days per year with snow cover (1 to 365 days). The values of both snow presence and depth were interpolated using the inverse distance weight method to derive the final rasters, using the Spatial Analyst in ArcGIS 10 (ESRI, Redlands, CA, USA). Since snow data were not available for the entire European continent, the layers were cropped to span from latitude 68 N 33.8 N and longitude 12 W to 51.7 E. All layers were used in its original spatial resolution (2.5’, ca. 5 km^2).

Models were run in Maxent with 75% of the presence data, and the remaining 25% were used as training. A total of 10 replicates were run, and the average among all runs was considered. The default parameters of the software were used (500 maximum iterations, convergence threshold of 0.00001; 10,000 background points; regularization multiplier of 1), and was also performed a Multivariate Similarity Surface (MESS) analysis when projecting. The MESS calculation represents how similar a point is to a reference set of points, with respect to a set of predictor variables (model training). The result allows to distinguish novel climate regions, or value ranges outside the predictor’s range.

Another measure of error, the most dissimilar variable (MoD), was also estimated. The information on which the variable is driving the MESS value at any given point can be extracted and mapped by finding the MoD. As a result, the variable in a given region that shows a novel climate condition (value outside the training model) can be identified and the certainty of the suitability estimation can be evaluated.

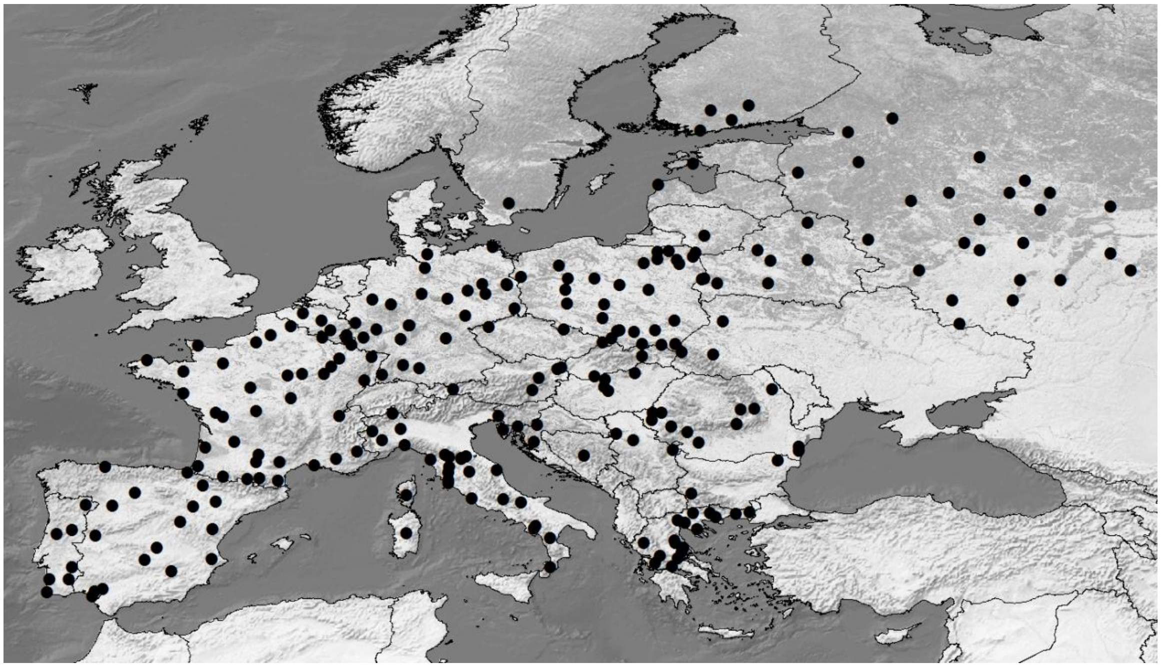


Figure 14 – Wild boar occurrences used for the estimation of ecological suitability using Maxent.

To evaluate model performance, was used the Area Under the Curve (AUC) of the Receiver Operated Characteristics (ROC), which measures the ability of a prediction to discriminate presence from absence (Elith *et al.* 2010) and ranges from 0.5 to 1. An AUC value of 0.5 indicates that the model has no predictive ability, whereas a perfect discrimination between suitable and unsuitable cells will achieve the best possible AUC of 1.0 (Morueta-Holme *et al.* 2010). The standard deviation of the estimations was also analysed, to assure it was not too high. Accuracy was verified using modern data by comparing estimated values with the current distribution of wild boar in Europe. Only when the layers of predicted values fitted the actual species range, the projection into paleoclimatic surfaces was performed. To verify which variables most contributed to the model, a jackknife analysis of gain was performed with the training data.

The paleoclimatic map was validated using fossil records as compiled by William Daves (available on the Stage Three Project website) and Sommer & Nadachowski (2006). Only archaeological sites were considered where the presence of *Sus scrofa* was reported, plotting their geographical coordinates in a map. Two time intervals were considered: LGM (23,000-16,000 BP) and older than 23,000 BP. The occurrence of wild boar remains in a site during the glacial period indicates the presence of a glacial refuge.

Finally, to derive presence/absence maps from the continuous output of Maxent, the 10th percentile training presence threshold was calculated. It predicts the absence of the 10% most extreme presence observations, meaning that the ten percent of records with the lowest predicted model values will fall into the absence regions, and the presence regions will encompass the other 90% of the distribution records. The extreme presence observations may represent recording errors, ephemeral populations, recent migrants, presence of unusual microclimatic conditions) (Morueta-Holme *et al.* 2010). After the reclassification, it was possible to represent the areas of presence/absence of the wild boar as binary maps. The binary maps represent the Minimal Presence Area, which is based on the assumption that a good suitability map should predict an area that is small as possible while still including the maximum number of the species occurrences (Engler *et al.* 2004). Such maps were created for present day data and the LGM (representing areas of possible refugia), according to the two different climatic models adopted.

General linear model

With the aim to understand how present and past suitability in Europe can explain the detected patterns of genetic diversity across the continent, the effect of five predictive variables (latitude, longitude, present suitability, LGM suitability according to MIROC model, and LGM suitability according to CCSM model) on *Hk*, π and *AR* was comparatively tested by a multiple linear regression using the software R (R Core Team 2012). Both single effects and joint effects of more variables were tested and the most parsimonious model was selected on the basis of the Akaike Information Criterion (AIC), using the corrected formula for small sample sizes (AIC_c) (Symonds & Moussalli 2011). In order to rank the models developed, a priori candidate models were treated as plausible hypothesis and investigated the weight of each hypothesis via differences in unbiased AIC_c. The support for each model was assessed by comparing its relative performance (Δ_i) to the best model following the form $\Delta_i = AIC_i – AIC_{min}$. Thus, the Δ_i for the best model, the one with the lower AIC is 0, and other model are ranked in descending order of the Δ_i relative to the best one (Hobbs & Hilborn 2006). Following (Westley *et al.* 2010) models with Δ_i values between 0 and 2 were considerate to have substantial empirical support, models with $\Delta_i > 4$ to have considerably less support, and models with $\Delta_i > 10$ to have essentially no empirical support compared to better models.

In addition, the Akaike weight (w_i) was calculated in order to assess each plausible model (Hobbs & Hilborn 2006). The Akaike weight can be interpreted as the "probability" of a given model being the best model among a set of candidates if the model selection analyses were repeated many times (Westley *et al.* 2010). Thus, the best models will have Akaike weights closer to 1 than the poor models have. This is not a probability in the Bayesian sense of a posterior distribution, but rather is inferred from empirical, bootstrapped simulations where the proportion of times that a model is chosen as the best model is well approximated by the w_i (Hobbs & Hilborn 2006).

Isolation-by-distance (IBD)

To investigate whether an isolation-by-distance (IBD) model could explain the observed structuring in European wild boar D-loop data, Mantel tests (Mantel 1967) were run in Arlequin v. 3.5 (Excoffier & Lischer 2010). The occurrence of IBD patterns was tested for the 39 European populations, by looking at the significance of correlations between spatial distances and Φ_{ST} values. As geographic distance, the linear Euclidean distance between populations and the cost-weighted distance between population pairs along an optimal route (least-cost path) were considered. The distance values were calculated using the Spatial Analyst extension in ArcGIS. A "cost raster" was created assigning coast values to sea and to mountains higher than 2,000 m (classified as absolute barriers for the wild boar dispersal). The high values assigned intend to reflect a great cost to move through each of these landscapes (cells), transforming both features in barriers for the wild boar.

Historical models vs. genetic diversity

One of the fundamental problems in population genetics is the study of the nature of genetic differentiation that is found in real populations and, when possible, to identify the factors that are responsible for the observed spatial structuring of genetic diversity (Foll & Gaggiotti 2006). The use of Mantel tests allows the description of how much of the genetic diversity can be explained by geographical distance, but other factors are not considered. Historical factors, such as migrations, translocations, play a major role in the observe genetic diversity, and cannot be considered when calculating a simple IBD pattern. In order to disentangle history from geography, the historical correlation with the genetic diversity were tested. A Mantel test was performed in Arlequin using the

Φ_{ST} values between the continental populations (Sardinia excluded). As environmental factors, a matrix of historical correlation was constructed (described below), and confronted against Φ_{ST} values between the populations.

The distance matrix was built in three steps: first evaluating which areas could be considered as refuge (R) according to literature and the results obtained, secondly identify the possible post-glacial recolonized areas (C), and finally a dissimilarity weight between all considered populations was computed taking into account the demographic process of dispersal from refuge areas. For the first step, Italy, Iberia and Balkans were considered as source populations, since they are described in literature as the classical refugia during the LGM for European animals and plants. Previous results indicated Croatia and Greece as separated refugia since they possess genetic signatures not present in any other population, therefore were also classified as refuge areas(Alexandri *et al.* 2012; Scandura *et al.* 2008, this study). Populations from Portugal and Spain were assumed to belong to one refugia (Iberia); all populations in Italy (except Sardinia, due to its unique colonization history) were grouped in the Italian refugia; and all other populations in the Balkans not in one of the four mentioned refuges were assumed to belong to the Balkans/Eastern Europe refugia. To define the colonized areas and understand which refugia colonized northern Europe, populations from France and Germany were considered "colonized populations". In the case of Austria, in accordance with other studies, it showed a mixed pattern of haplotype composition, so it was considered also as a "colonized population", independently if it was colonized naturally of by human translocations. Because Sardinia has a very peculiar story, it was not taken into consideration (Scandura *et al.* 2011b).

After separating all the 38 population in refuges/colonized, to construct the weight matrix under each dispersal scenarios, two different approaches were taken:

1) Two discrete weights: for each population comparison, weights were assigned according to the relationships between refuges/colonized populations (blue and red arrows in Figure 15a).. A population that was colonized by a refugia was assumed to have a weight of 0 in relation to its colonizer, because of the close relationship between them (their genetic pools derive from the same source). For the same reason, two populations colonized by the same refugia were considered to have weight 0. On the other hand, two populations colonized by different refugia were considered to have an arbitrary weight of 2 (their genetic pools derive from two distinct

refuge). Two refuges always have a value of 2 between them, since they represent two distinct gene pools. . (Figure 15a)

2) Four discrete weights: in this case, each type of relationship among refugia and colonized populations had a different weight (Figure 15b). A refugia and its colonized population had a value of 1 (red arrow), two refugia had a value of 2 (green arrow), one refugia and a colonized population that was not colonized by this refugia had a value of 3 (blue arrow), and finally, two colonized populations that were colonized by two different refugia had a value of 4. A schematic representation of the weight assignment is given below. In contrast to the two weight approach, the use of four weights allows a more precise description of the similarity between refuges and colonized populations.

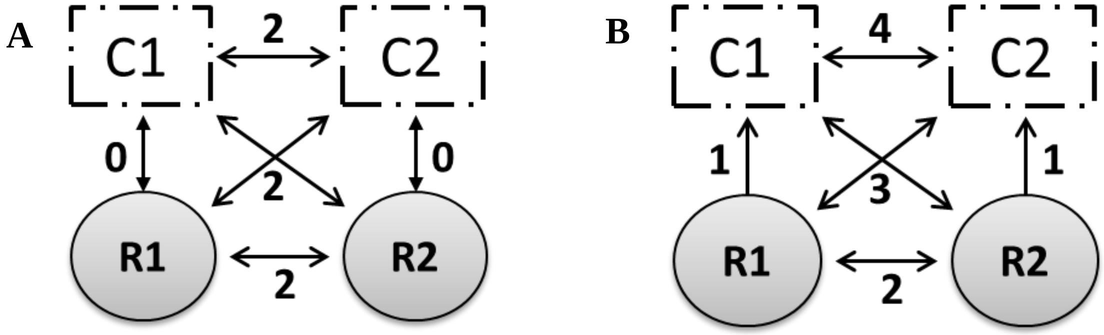


Figure 15 – Schematic representation of the IBD models. A) Schematic model using two weights. B) Schematic model using four weights. Arrows represent the relationship between the refuges/colonized populations. Not all relationships considered to build the matrix are shown.

A total of 10 models were constructed, considering the most plausible relationships between the refuges/colonized populations. For each of the 10 models, the two matrices were built, totaling 20 models tested.

Coalescent phylogeographic reconstruction

To reconstruct the past geographical location for each macro-region and infer their ancestral relationship, a Bayesian phylogeographic model was used. In this analysis, the geography is a discrete character taking on as many states as the number of macro-regions, and its evolution is modeled through time. The Bayesian framework implemented by Lemey et al. (2009) provides an inferential method that allows the joint estimation of the phylogenetic relationships between populations and the geographical position of each common ancestor (tree nodes). The advantage of this method over classical approaches is that it takes into account both the unknown relationships between individuals (phylogenetic uncertainty) and the unknown populations past demography (demographic uncertainty). The probability of a certain geographic state at each tree node can then be calculated, linking the geographic information with time.

Ancestral state reconstruction was conducted in BEAST v. 1.7.4 (Drummond *et al.* 2012) with the same mutational parameters used for the Bayesian Skyline Plots. Parameters for the phylogeographic analysis were set following Lemey et al. (2009).. Discrete states for the ancestral state reconstruction were considered as the macro-regions (Fig 13) and the root of each clade was estimated according to these states. To consider the potential changes of population sizes through time, two demographic models were used: in the “constant” model, the population size does not vary, while in the “Bayesian Skyline” model the population size is free to vary. Under the constant model, a total of two analysis of 100,000,000 generations each were run, and then combined in a single file using LogCombiner.. For the model using the Bayesian Skyline model, five runs of 100,000,000 generations were combined in a single file. The first 10% generations were treated as burn-in. The consensus trees were generated using TreeAnnotator and visualized in Figtree v. 1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Chapter 4

Results

Phylogenies

The median-joining network based on the large alignment including the Asian and North African haplotypes (in total 87 different haplotypes) showed three major groups (Figure 16): the Asian group A (corresponding to clade D2 in Larson et al. (2005), Figure 2), a pan-European group E1 (D4 in Larson et al. (2005)) and the group E2 present only in Italy (D1 in Larson et al. (2005)). Most of the North African sequences grouped within the European E1 clade, sharing one of the most widespread haplotypes (H029). However, another North African sequence (H143) showed an intermediate position between the European and Asian groups, suggesting the possibility of a private North African clade. Within European wild boars, only seven Asian haplotypes were observed, four in Italy and three in Germany/Luxembourg/Belgium.

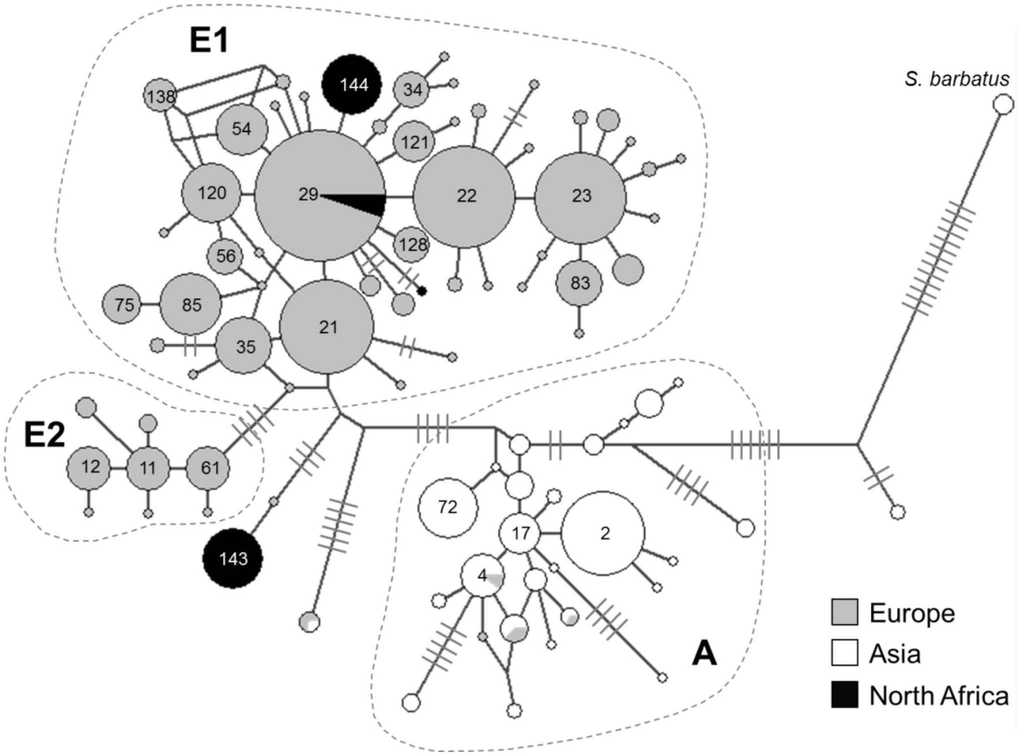


Figure 16 - Median-joining network of European and non-European haplotypes. Circle sizes are proportional to haplotype frequency. Haplotype number designations are indicated inside the circles. Branches with more than one nucleotide change are identified by transversal bars.

When taking into consideration only the European sequences (763 individuals), a total of 50 haplotypes were observed, corresponding to two A, 41 E1 and seven E2. The Bayesian tree restricted to such haplotypes (Figure 17) gave a high (≥95%) posterior probability support to the three main clades, but also indicated an additional structure with the highly supported E1a clade within E1 (Figure 17). This clade correspond to the previously reported A-side group (Alexandri *et al.* 2012; Larson *et al.* 2007a; Scandura *et al.* 2011a). Haplotypes corresponding to the C-side group identified by Larson et al. (2005) with a smaller sample size did not form a monophyletic clade within the E1 group.

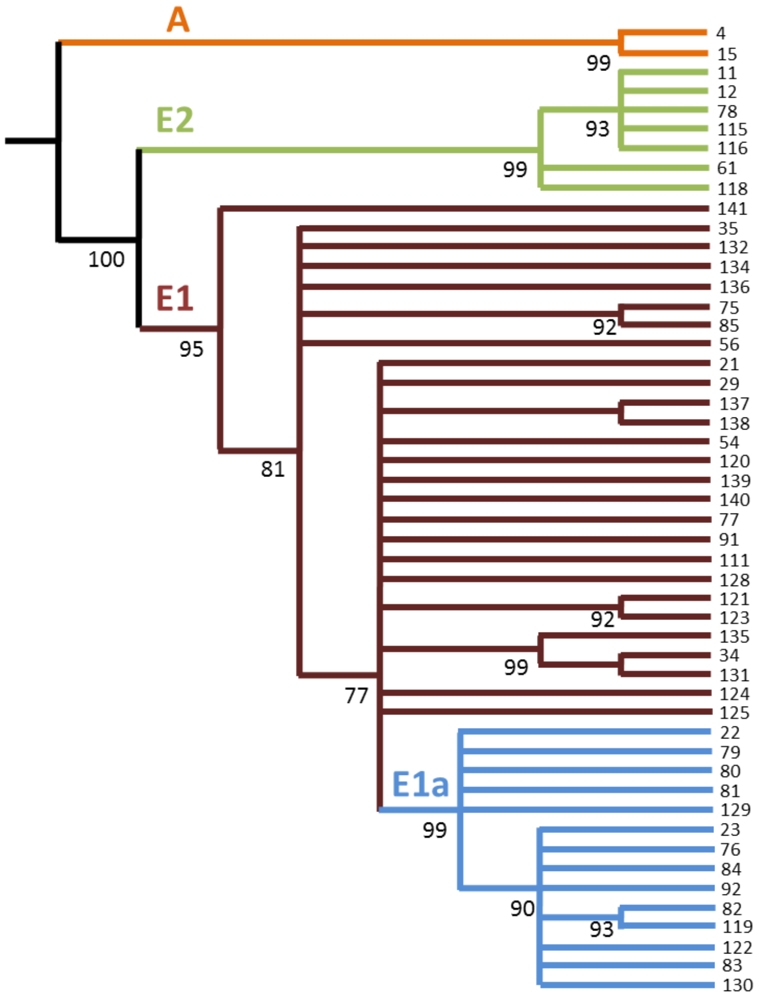


Figure 17 - Bayesian tree of the 50 mitochondrial D-loop haplotypes observed in 763 wild boars sampled throughout Europe. The tree is rooted using a homologous sequence of *Sus barbatus* as outgroup. Posterior probabilities >75 are shown in the internodes. Numbers at the tips of the tree indicate the haplotype number.

Genetic diversity and spatial differentiation

Haplotype distribution

Plotting the haplotype distribution for each of the 74 populations investigated, a pattern clearly emerged (Figure 18). The most common haplotype in East Europe is also shared with Iberia (light pink), while the most common haplotype in Iberia is also found in East Europe (dark blue). Greece and Croatia have private haplotypes (brownish colors).

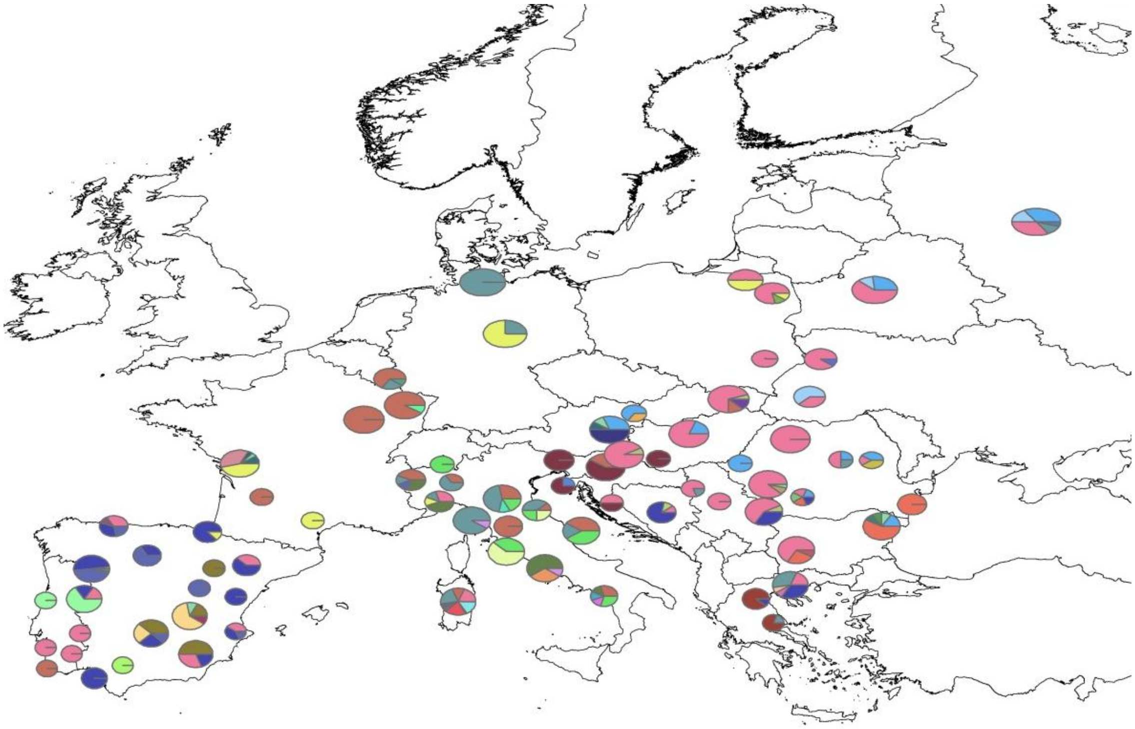


Figure 18 – Haplotype distribution of each of the 74 populations. Circles sizes are proportional do sample size (<5 samples, >5 and <10, or >10). Each color represent a single haplotype.

Regarding the geographic distribution of haplogroups defined by the phylogenetic tree (Figure 19), the typical E2 only occurred in Italy and Sardinia. The E1a haplogroup came out to be the dominant clade in Central Europe, in the Italian peninsula and in north-western Balkans, where H022 and H023 were the major haplotypes (matching A and BK, respectively, in Larson *et al*. (2005)). E1 was widespread, but very common in the East (from Belarus and Russia in the north, to Greece in the south) and in Iberia. The most common haplotype across Europe was H029 (haplotype C in Larson *et al*. (2005), belonging to E1). It had an overall frequency of 28% across Europe and was observed in both Eastern European and Iberian populations, but not in Central Europe. Several private haplotypes occurred in Iberia, but the most common (H021, matching haplotype E in Larson *et al.* (2005)) was shared with eastern populations. Besides the private haplogroup E2, Italian wild boars showed an exclusive E1 haplotype, that was spread across the peninsula (H075). Croatian and Gorizia (Northeast Italy) populations were dominated by a private haplotype (H083, dark brown in Figure 16), which differed by a single mutation from H023. Greece exhibited a large proportion of private haplotypes, one being more common than the others (H128, matching the common haplotype found by Alexandri *et al.* (2012)). France and Germany exhibited a low number of haplotypes when compared to the rest of European populations, but the majority of these haplotypes were shared with Italian populations and belonged to the E1a haplogroup. Sardinia was the population with the largest number of haplotypes (14), most of which were private (64%).

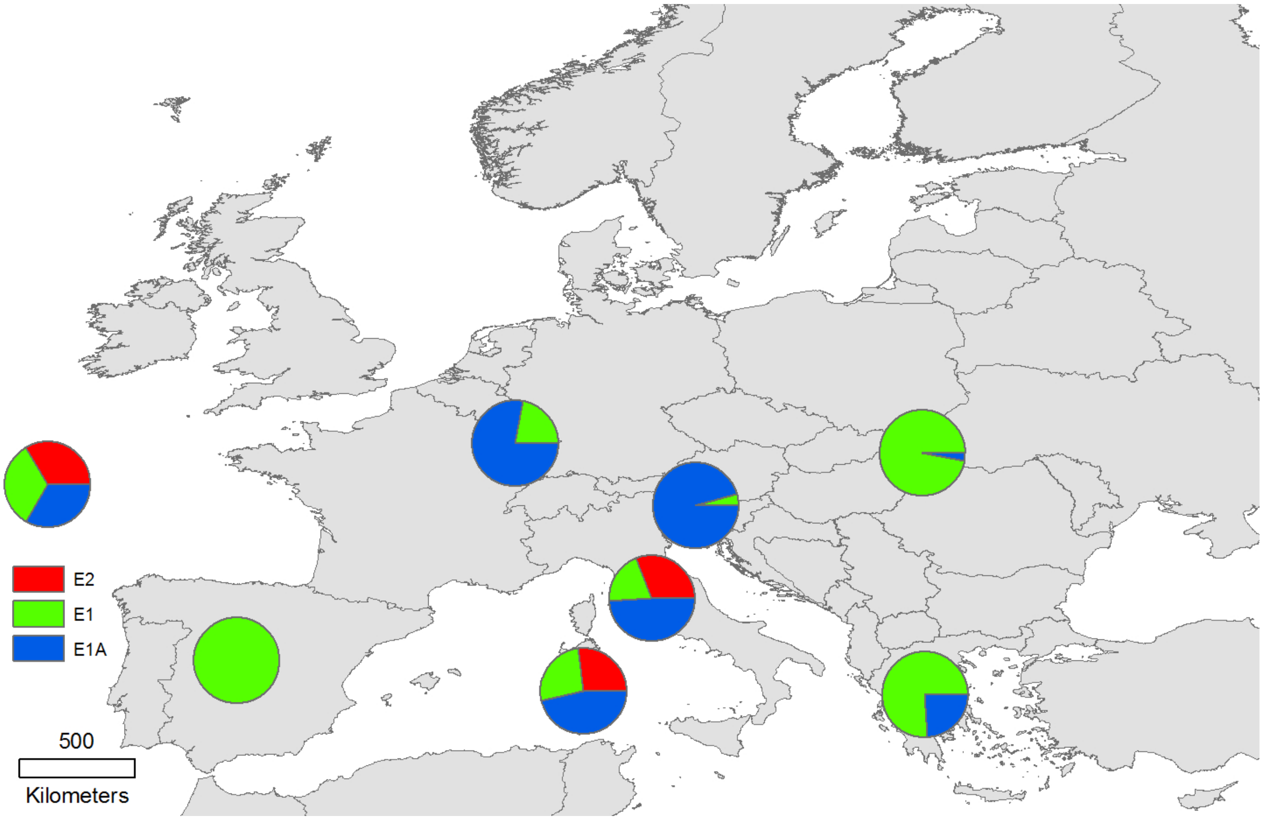


Figure 19 – Frequency of the clades in each of the macro-regions.

When considering the results of the NCPA (Figure 20), considering the third level of grouping, five major sequence groups were obtained: one corresponding to Asian haplotypes (3-1), one corresponding to E2 (Italian haplotypes, level 3-2 – with the exception to the private Austrian haplotype H141), and the E1 group was subdivided in three: one widespread in Italy, France and Germany (3-4); and two shared between Iberia and East Europe in similar proportions (3-3 and 3-5). Although the group subdivision within the E1 clade is not exactly the same as the one obtained in the phylogenetic tree (Figure 15), the same general pattern of genetic sharing between East Europe and Iberia, and Central Italy and Italy is observed.

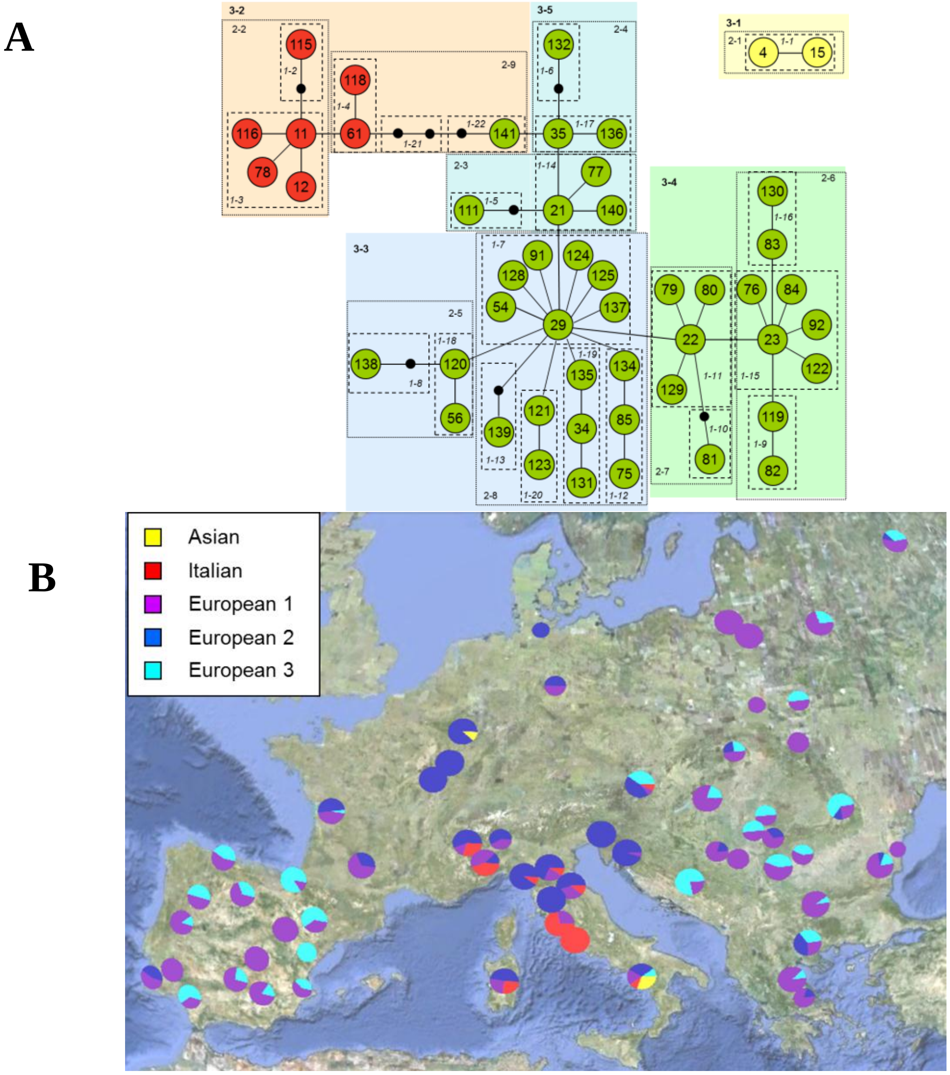


Figure 20 – Results of the NCPA analysis. A) Nested clade diagram for 763 European sequences. Pies are unique haplotypes and numbers correspond to haplotype names. Missing intermediate haplotypes are represented by black dots. Nested clades are labeled with nesting level (prefix) followed by clade number (suffix). B) Distribution of nested clades (third level) in the 74 European populations.

Genetic variation within populations

Differences in the pattern of genetic variation can be observed when homogeneous groups with samples size larger than 10 (with the exception of Salerno) are compared (Table 2). The number of haplotypes ranged from 0 (one population from Germany and one from France) to 14 (Sardinia). Interestingly, the highest levels of Allelic Richness and Haplotype Diversity were found in the populations in Piemonte (comprising the sampling sites of Cuneo, Torino, Val d'Ossola and Alessandria) and Salerno. On the other hand, the highest values of nucleotide diversity, mean number of pairwise differences and total number of polymorphic sites were observed for most Italian populations. This result is due to the presence of E2 haplotypes in Italy only. The only exception to this pattern is the Castelporziano population. Unlike the other Italian populations, Castelporziano has only E2 haplotypes which lead to smaller values of nucleotide indices since there is no mixing between typical European haplotypes E1 and the typical Italian E2.

Table 2 - Genetic diversity in wild boars from 39 sampling areas across Europe. Numbers in bold denote statistically significant values (p<0.05).

# Pop	Populations	Putative subspecies	N	# Polymorphic sites	Ts	Tv	# Haplotypes	Haplotype diversity (H_k)	Allelic Richness (AR) [7]	Tajima's D	Fu's FS	Mean Number Pairwise Differences	Nucleotide Diversity (π)
1	Austria_ALow	*scrofa*	13	6	5	1	5	0.744	2.606	**1.694**	0.795	2.821	0.007
2	Belarus	*attila*	19	4	4	0	3	0.550	1.546	0.185	1.648	1.216	0.003
3	Bulgaria_BLud_RCon	*lybicus*	18	6	5	1	5	0.484	1.807	-1.183	-0.735	1.229	0.003
4	Bulgaria_BRil	*lybicus*	12	2	2	0	3	0.530	1.538	-0.382	-0.362	0.576	0.001
5	Castelporziano	*majori*	10	1	1	0	3	0.600	1.692	**1.303**	0.477	1.000	0.002
6	Croatia_GKo	*lybicus*	26	3	2	1	4	0.397	1.279	-0.965	-1.098	0.554	0.001
7	Cuneo_Torino_ValDossola_Alessandria	*majori?*	19	12	11	1	7	0.860	3.641	1.047	1.080	4.421	0.011
8	Firenze_Siena	*majori*	18	11	10	1	4	0.595	1.922	-1.166	1.865	2.176	0.005
9	Forli_Arezzo	*majori*	18	12	11	1	4	0.765	2.508	0.410	**3.969**	3.876	0.010
10	France_Fang	*scrofa*	28	8	7	1	5	0.669	2.139	0.965	2.246	2.698	0.007
11	France_FHma_Luxembourg	*scrofa*	18	1	1	0	2	0.294	0.798	0.022	0.463	1.544	0.004
12	France_FArc	*scrofa*	10	0	0	0	1	0.000	0.000	**0.000**	0.000	0	0
13	Germany_GSwh	*scrofa*	20	0	0	0	1	0.000	0.000	**0.000**	0.000	0	0
14	Germany_GHar	*scrofa*	20	3	2	1	2	0.395	0.917	1.069	3.433	1.184	0.003

15	Greece	*lybicus*	29	3	2	1	5	0.727	2.498	1.455	0.028	1.300	0.003
16	Hungary	*attila*	10	2	2	0	2	0.356	0.933	0.019	1.523	0.711	0.002
17	Maremma	*majori*	11	9	9	0	2	0.436	0.976	1.169	6.822	3.927	0.010
18	Poland_PLNor	*scrofa*	22	2	2	0	3	0.437	1.204	0.970	0.761	0.784	0.002
19	Portugal_PBra	*scrofa*	36	3	3	0	4	0.567	1.376	1.112	0.786	1.083	0.003
20	Portugal_PSes_PSlo	*scrofa*	14	3	3	0	3	0.484	1.538	0.647	1.145	1.143	0.003
21	Portugal_PCer_PRmo_PVgu_PAlg	*scrofa*	18	2	1	1	2	0.366	0.892	0.632	2.082	0.732	0.002
22	Romania_RCar	*attila*	24	4	3	1	4	0.308	1.091	-1.690	**-1.854**	0.409	0.001
23	Romania_RLip_RTir	*attila*	20	5	2	3	4	0.500	1.689	-0.306	0.556	1.268	0.003
24	Romania_RTim	*attila*	34	2	2	0	2	0.166	0.511	-0.636	0.953	0.332	0.001
25	Russia	*attila*	18	5	4	1	5	0.778	2.777	0.933	0.506	1.993	0.005
26	Salerno	*majori*	7	14	13	1	5	0.905	4.000	-0.459	0.384	5.238	0.013
27	SanRossore	*majori*	10	9	8	1	2	0.200	0.700	**-1.901**	3.672	1.800	0.004
28	Serbia_SEBor_RCra	*lybicus*	17	3	3	0	5	0.713	2.450	-0.110	-1.414	0.971	0.002
29	Serbia_SEVoj_Bosnia_SEBel	*lybicus*	19	3	1	2	4	0.614	1.720	-0.496	-0.859	0.702	0.002
30	Slovakia	*scrofa*	16	5	4	1	4	0.525	1.838	-1.218	-0.374	0.950	0.002
31	Spain_SAst_SBur_SPyr	*scrofa*	22	5	5	0	5	0.684	2.272	-0.355	-0.508	1.208	0.003
32	Spain_SCot_SSev	*scrofa*	14	2	2	0	2	0.495	0.990	**1.554**	**3.641**	1.484	0.004
33	Spain_SCre	*scrofa*	16	4	4	0	4	0.767	2.551	1.543	1.137	1.800	0.004
34	Spain_SJae	*scrofa*	16	2	2	0	3	0.658	1.820	1.085	0.668	0.858	0.002
35	Spain_SMun_sAlb_SGua	*scrofa*	15	2	2	0	3	0.648	1.726	**1.850**	1.029	1.048	0.003
36	Spain_STol	*scrofa*	11	3	3	0	4	0.600	2.164	0.830	-0.228	1.273	0.003
37	Spain_SVeb	*scrofa*	11	2	2	0	3	0.709	1.952	1.339	0.551	0.982	0.002
38	Ukraine_UPla_UMts	*attila*	17	4	4	0	3	0.522	1.371	-0.169	1.556	1.235	0.003
39	Sardinia	*meridionalis*	83	20	19	1	14	0.815	3.281	0.022	-0.326	4.086	0.010

When the values of gene diversity (*Hk*) are interpolated (Figure 21a), a north-south gradient with higher diversity in southern regions is observed. Highest values of diversity are found in Western Spain, Italy and Greece, with a local peak in Russia. The general trend for the nucleotide diversity (Figure 21b) shows a different pattern, clearly controlled by the presence of two phylogenetically distant clades (E1 and E2) in Italy and Sardinia.

The errors for both interpolations were low, with only the areas corresponding to the sea showing the highest error values.

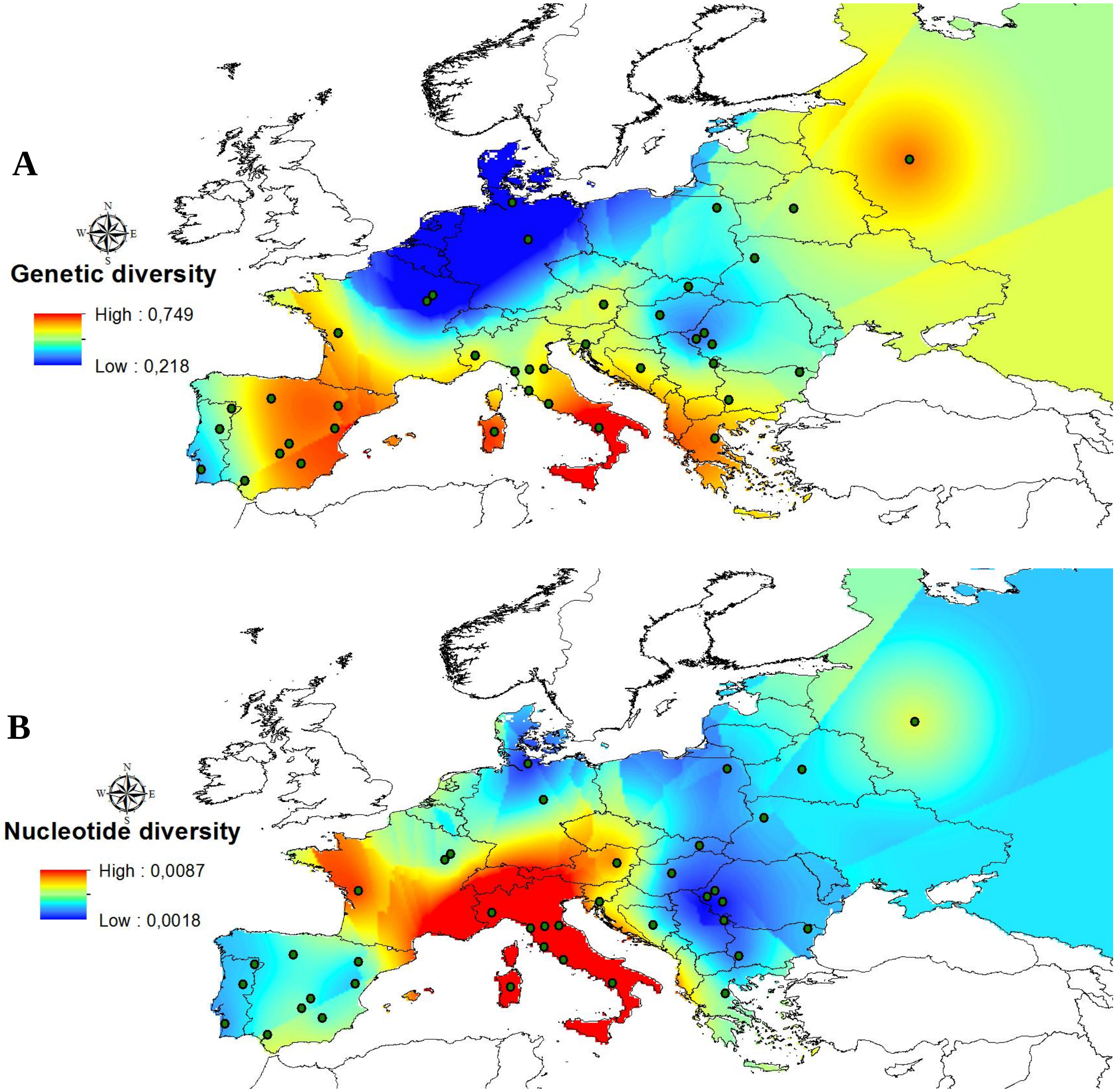


Figure 21 – Interpolation rasters using the Kriging method for the genetic diversity (A) and nucleotide diversity (B). Colors depict higher values (red) or low (blue) for both statistics. Green dots represent the 39 populations used for the interpolation.

Hierarchical clustering on Principal components (HCPC)

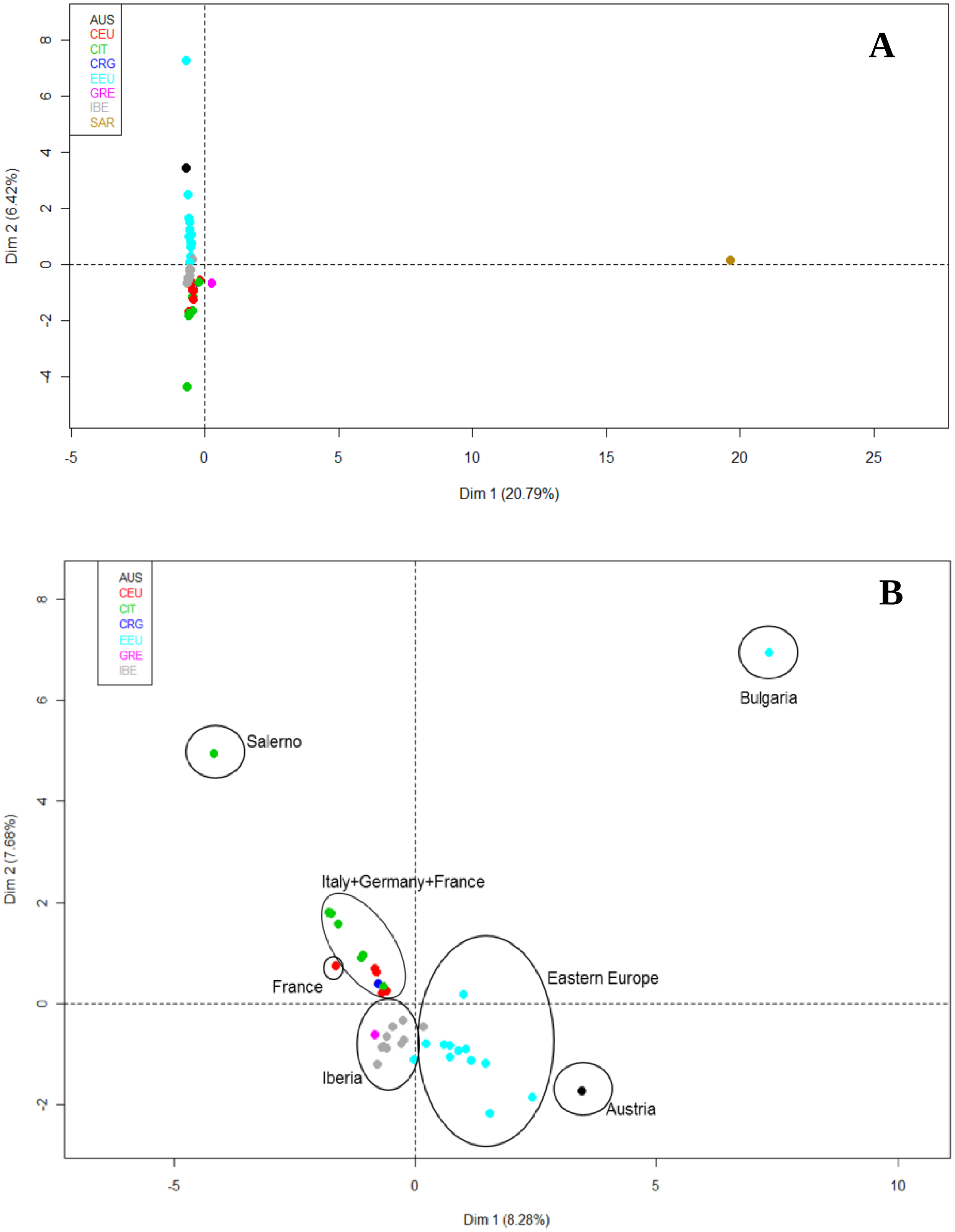
Considering the allele frequencies of the 39 populations and plotting them in a PCA followed by a hierarchical clustering method, only two major groups were identified: Sardinia and the rest of European populations (Figure 22a), with the island showing a remarkable distance from all the other populations. Accordingly, also by virtue of its geographic isolation, Sardinia was excluded in the following HCPC analysis.. Considering the increase of within-group inertia (ΔQ), determined by the number of minimum and maximum clusters defined, the optimal number of clusters (Table 3) was defined by the lowest inertia gain which corresponds to the division in seven clusters (Figure 22b, Table 3).

Table 3 – Number of estimated optimal clusters. The number in bold depict the optimal number.

N clusters	2	3	4	5	6	7	8	9	10
Ratio between two successive within-group inertias	0.7832	0.7845	0.7464	0.6996	0.5997	**0.5981**	0.6824	0.7827	0.7665

As observed in the haplogroup distribution, some level of unexpected genetic similarity between western and eastern areas emerges. One population from Portugal grouped with East European populations, and some from Eastern Europe (Greece, Serbia/Bosnia) fell in the Iberian cluster. Four HCPC-groups are composed by a single population each: Austria, Bulgaria, one population in Italy (Salerno), and one population in western France (FFang). These results are due to the presence of private haplotypes, and possibly reflect sampling errors or recent drift effects in specific areas.

The Bulgarian population, despite exhibiting a similar haplotype composition as the rest of the Eastern European populations, shows a remarkable separation, mainly due to the presence of two exclusive haplotypes. The same holds for Salerno (Italy) which harbors two exclusive haplotypes, although most of its alleles are shared with other Italian populations. When the Bulgarian population is excluded from the HCPC, the populations from Croatia/Gorizia are defined as a separate group.


A
AUS
CEU
CIT
CRG
EEU
GRE
IBE
SAR
Dim 2 (6.42%)
Dim 1 (20.79%)
-5
0
5
10
15
20
25
-4
-2
0
2
4
6
8
B
AUS
CEU
CIT
CRG
EEU
GRE
IBE
Bulgaria
Salerno
Italy+Germany+France
France
Eastern Europe
Iberia
Austria
Dim 2 (7.68%)
Dim 1 (8.28%)
-5
0
5
10
-2
0
2
4
6
8

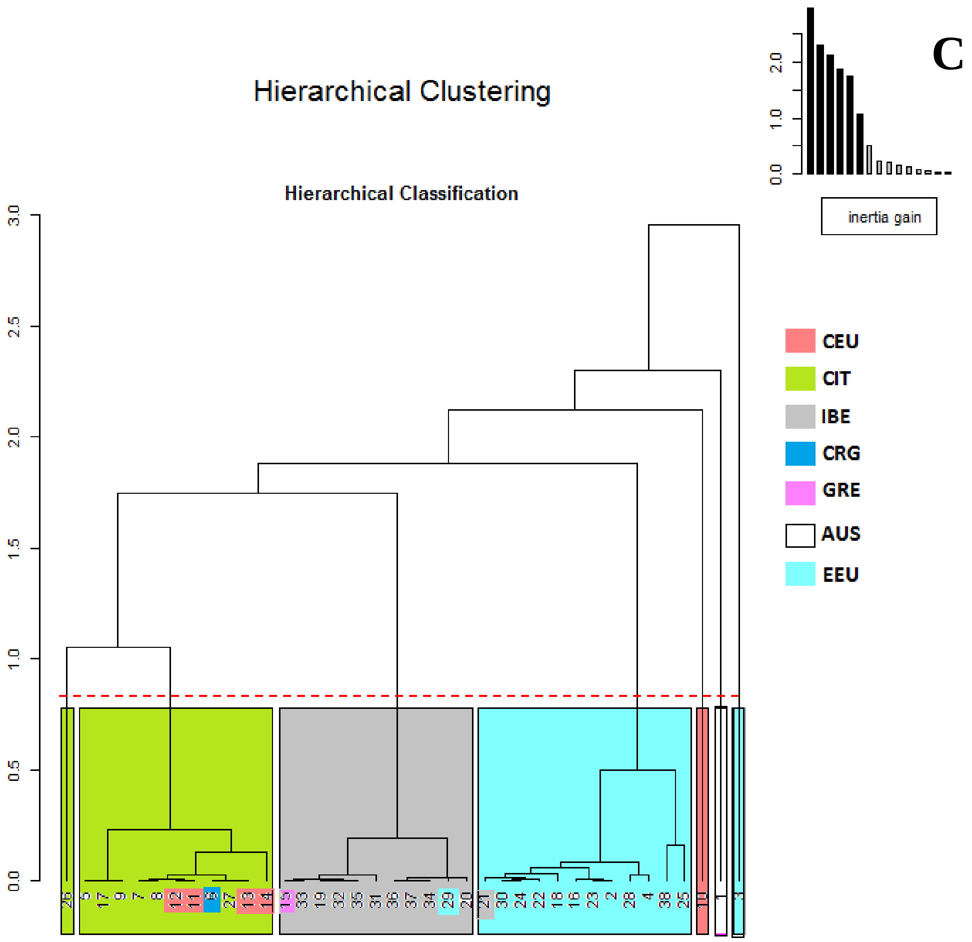


Figure 22 - Bidimensional plotting of sampled populations on the basis of HCPC. Colors are based on biogeographical definition. A) All 39 populations, including Sardinia. B) 38 populations, Sardinia excluded. The populations defined by the HCPC are inside the circles. Names indicate group definitions by the HCPC. C) Hierarquical tree depicting the populations relationships. The red line indicate the optimal group partition. Population numbers refers to identification numbers in Table 2.

Macro-region definition

Previous evidence based on microsatellite markers (Scandura *et al.* 2008) separated populations from France, Gorizia, Spain, Hungary in different clusters based on the clustering algorithm Structure (Pritchard *et al.* 2000). These were the only populations in the study published by (Scandura *et al.* 2008) representing the regions of, respectively, Central Europe, Northern Adriatic, Iberia and East Europe.

Integrating the results of the HCPC analysis, with previous evidence based on microsatellite markers and biogeographical features of the continent (presence/absence of geographical barriers - like mountain chains), seven groups of population were analyzed in detail, corresponding to macro-regions or clearly differentiated areas: 1) Sardinia; 2) Italy: continental Italy excluding the divergent sample from Salerno and the North-Eastern sample from Gorizia found genetically different at microsatellite markers; 3) Iberia: Spain and Portugal, excluding the mixed population from southwest Portugal affiliated with the Eastern Europe cluster in HCPC; 4) Central Europe: France, Luxembourg, and Germany, excluding the HCPC outlier population from France based on published data (Fang *et al.* 2006); 5) North-Adriatic: Croatia and Gorizia; 6) Eastern Europe: Hungary, Slovakia, Poland, Belarus, Russia, Ukraine, Bosnia, and Romania; Serbia and Bulgaria were not included because their clustering with Iberia or outlier position in HCPC; 7) Greece.

Macro-region statistics

The AMOVA results indicated that 27% of the genetic variation is within groups and 73% between groups. The overall Φst value is therefore very high (0.73) and highly significant. Pairwise Φst between the macro-regions (Table 4) ranged from 0.03 (Italy x Sardinia) to 0.70 (North Adriatic x Eastern Europe), with a mean of 0.29. The North Adriatic group showed high values of Φst in all pairwise comparisons (>0.32).

Table 4 – Pairwise Φst values between macroregions. Below diagonal: Φst values. Above diagonal: p values

Phist	Greece	EastEurope	NorthAdriatic	CentralEurope	Iberia	Italy	Sardinia
Greece		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
EastEurope	0.202		<0.001	<0.001	<0.001	<0.001	<0.001
NorthAdriatic	0.675	0.702		<0.001	<0.001	<0.001	<0.001
CentralEurope	0.335	0.409	0.468		<0.001	<0.001	<0.001
Iberia	0.209	0.144	0.588	0.404		<0.001	<0.001
Italy	0.214	0.328	0.330	0.185	0.306		0.02
Sardinia	0.157	0.246	0.326	0.169	0.236	0.031	

The summary statistics within the macro-regions (Table 5), show that although the eastern European and Iberian groups have more samples and more haplotypes when compared to the other macro-regions, the levels of diversity are always higher for the Sardinian and Italian groups. Even if the Sardinian and Italian samples have smaller sample size, and are distributed in a smaller geographical area compared to most of the other macro-regions, the presence of a divergent haplogroup (E2) and several private sequences in Sardinia, raises the diversity levels.

Table 5 – Descriptive statistics within macro-regions. Numbers in bold denote statistically significant values ($p<0.05$).

Macro-regions	N	# Polymorphic sites	Ts	Tv	# Haplotypes	Haplotype diversity (H_k)	Allelic Richness (*AR*) [26]	Tajima's *D*	Fu's *FS*	Mean Number Pairwise Differences	Nucleotide Diversity (π)
Greece	29	3	2	1	5	0.727	3.896	1.455	0.028	1.300	0.002
EastEurope	228	15	8	4	16	0.523	5.476	-1.265	**-9.876**	0,959	0.003
Central Europe	68	4	3	1	3	0.656	2.000	1.743	3.985	1.519	0.004
Iberia	155	9	8	0	9	0.782	5.633	0.297	-0.319	1.673	0.004
Italy	86	14	12	1	9	0.826	5.761	**2.396**	4.151	4.905	0.012
NorthAdriatic	26	3	2	1	4	0.397	0.671	-0.965	-1.098	0.554	0.001
Sardinia	83	22	15	1	14	0.815	6.793	0.022	-0.326	4.086	0.010

Demography

The neutrality tests Fu's *Fs* and Tajima's *D* (Table 5) were generally non-significant for the majority of the populations. The only macro-region with a significant and negative value of *Fs* was Eastern Europe, and negative values of this statistics can be produced by a demographic expansion in the past. For the Tajima's *D*, only the Italian group exhibited a significant (and positive) value. Positive values suggest an excess of common variation, which can be consistent with balancing selection or a population contraction (bottleneck).

A clearly unimodal mismatch distribution was obtained in the Iberian population only, while the shape of the distribution for all other groups was ragged and multimodal (Figure 23). The presence of E1 and E2 haplogroups in Italy clearly affects also this analysis, producing two distinguishable peaks.

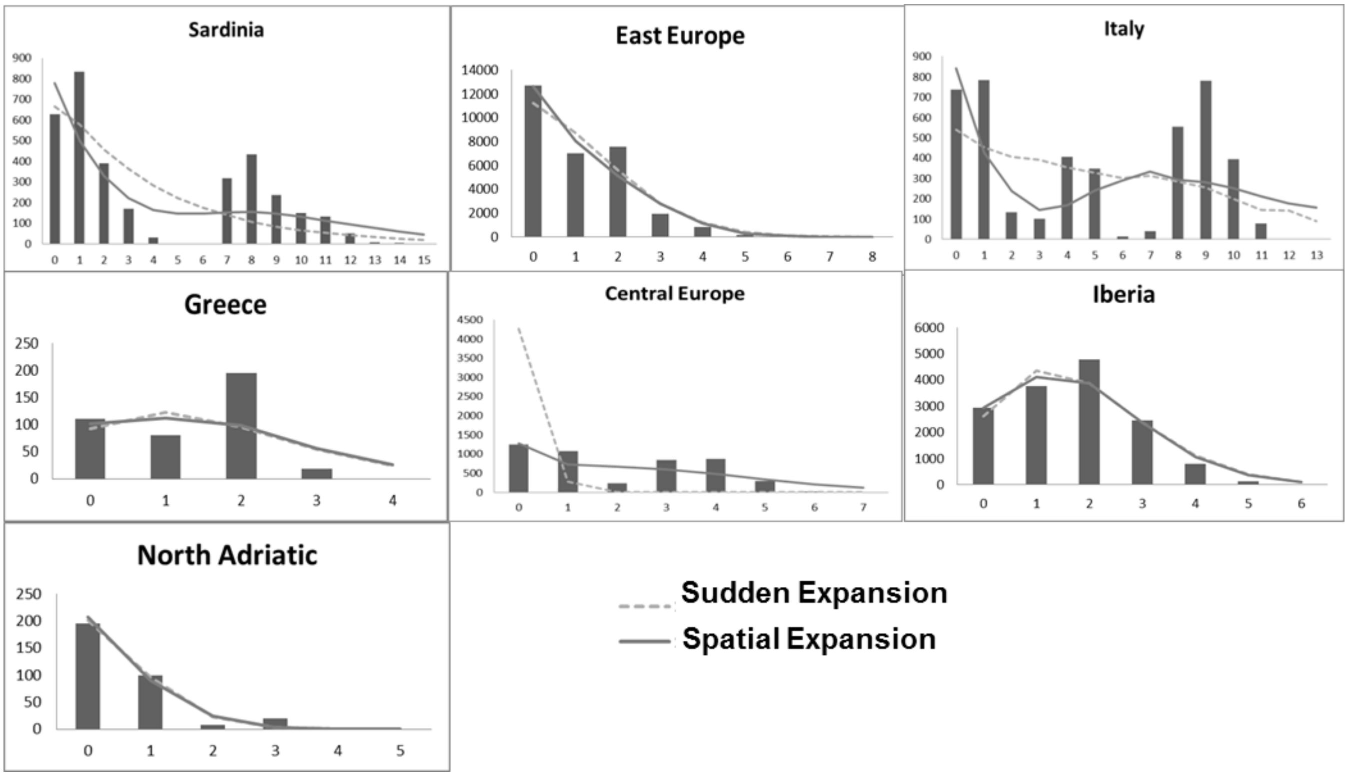


Figure 23 – Mismatch distributions for the seven macro-regions.

The Bayesian Skyline Plots analyses generally yielded constant population sizes through time, with only slight recent changes in a few macro-regions (Figure 24). Thus, the BSP failed to show ancient changes and any bottleneck/expansion. Despite the recent changes were shown in the median values (solid lines) and on the upper 95% interval values (upper dotted lines), the lower 95% values did not show any change in the population size, for any macro-region. The recent and slight changes in the BSPs are often associated with an artifact of the BSP method (A. Drummond, personal communication), so the changes observed in the graphs cannot be considered as true size changes.

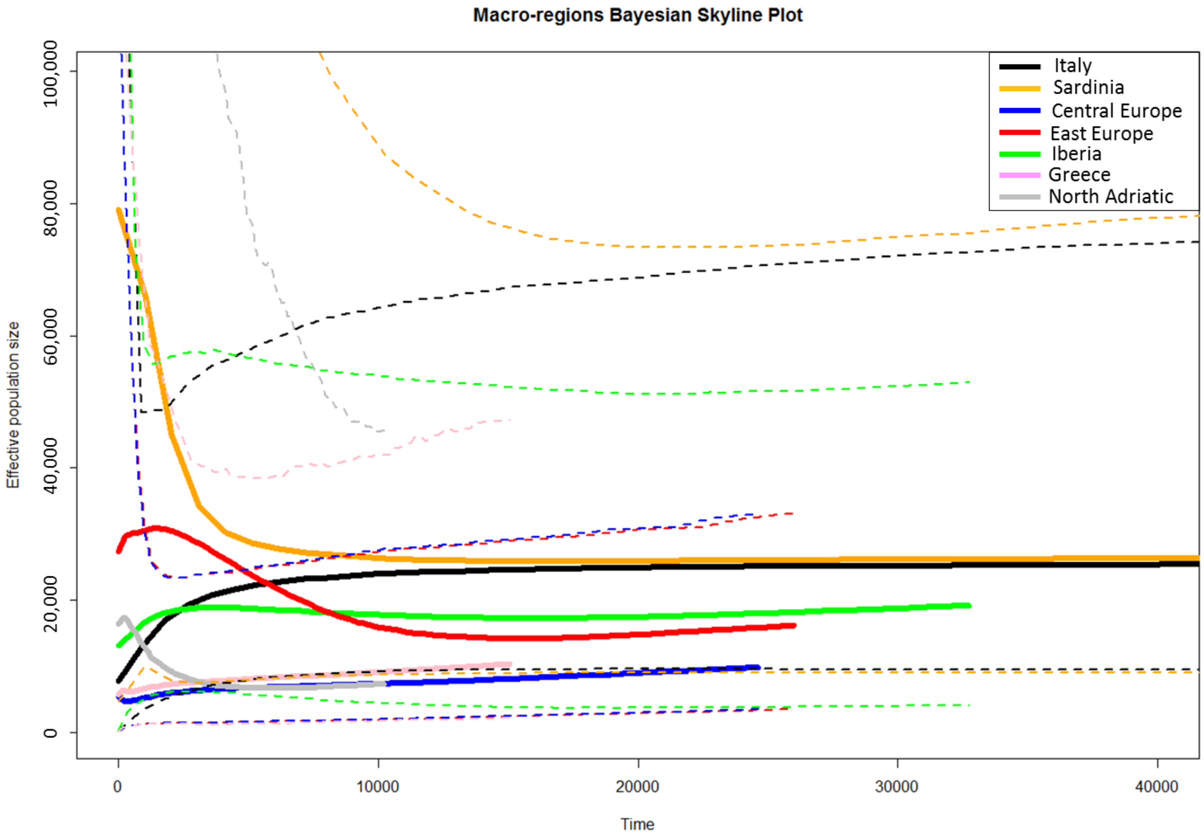


Figure 24 - Bayesian Skyline plots for the macro-regions. Solid lines represent median values, while dashed lines represent the upper and lower limits of the 95% confidence interval

The ABC results for each of the tested macroregions are compatible with those observed with the BSPs. Different demographic models cannot be distinguished with confidence, even if some evidence in favor of the decline and for the expansion are found in Italy and Sardinia, respectively (Figure 25).

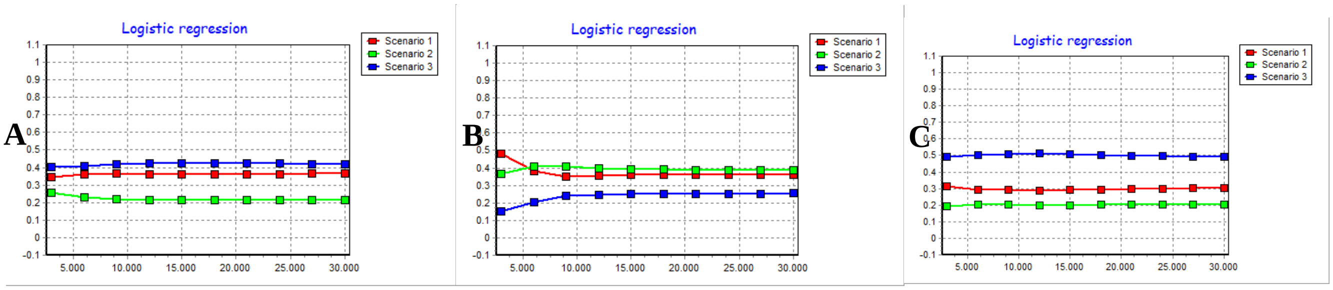


Figure 25 – Model choice for each of the macroregions. Scenario 1 (red) represents the Constant model, Scenario 2 the population expansion (green) and Scenario 3 the bottleneck model (blue). Values

in the X axis represents the number of the closest retained simulations (varying from 30000 to 5000), while the Y axis shows the probabilities of each model. A) Italy, B) Sardinia, C) Central Europe.

To visualize the spatial distribution of the closest simulations to the observed data set, the simulated data was plotted in a PCA. The closest 1000 simulations for each model are shown in Figure 26. As a general pattern, the scenarios overlap and are not well-distinguished, which is the probable the reason why no scenario is strongly preferred.

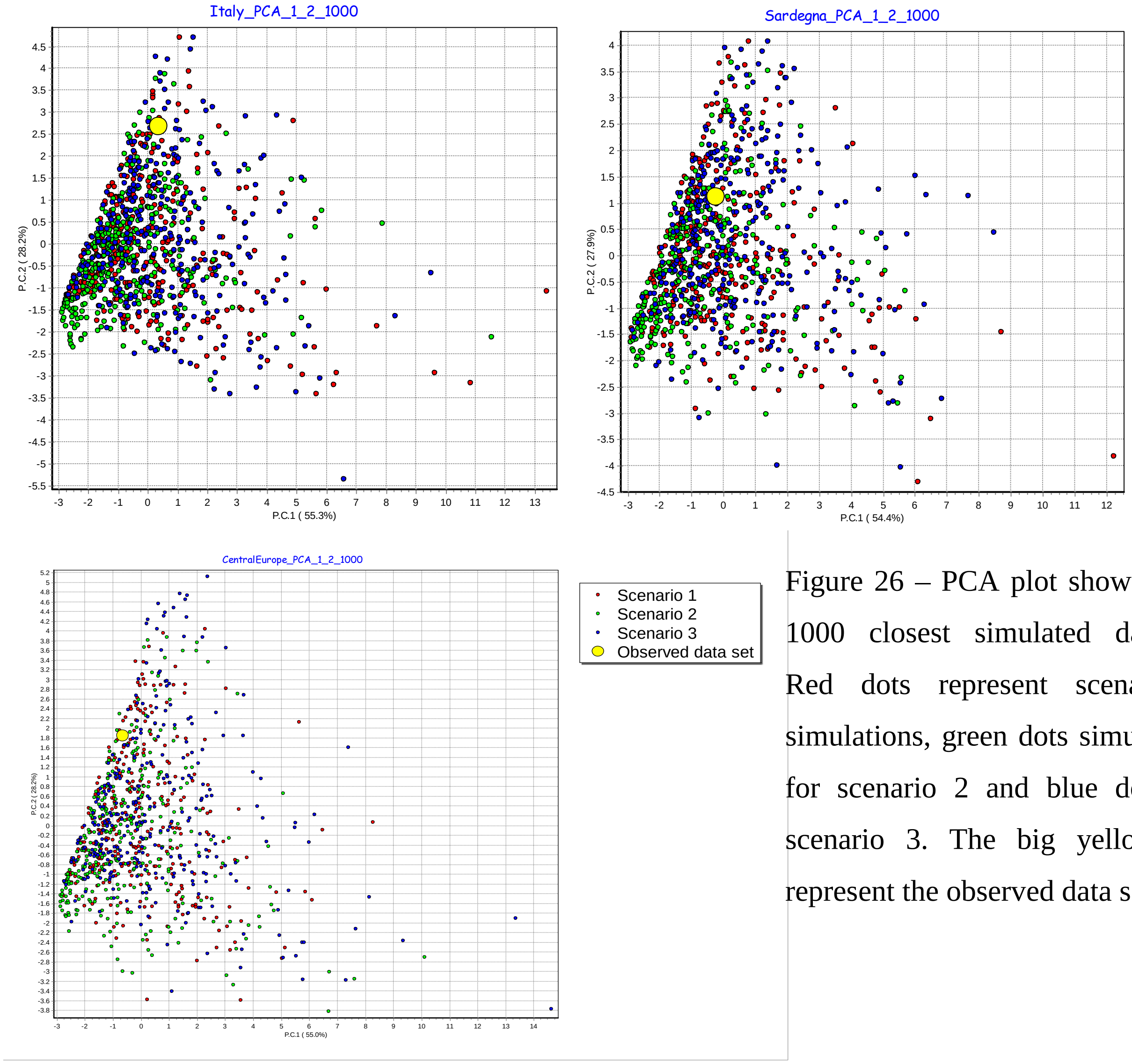


Figure 26 – PCA plot showing the 1000 closest simulated datasets. Red dots represent scenario 1 simulations, green dots simulations for scenario 2 and blue dots for scenario 3. The big yellow dot represent the observed data set.

Present and past species' range

Under the current distribution, the AUC mean values across 10 replicates for the training and test data showed satisfactory values (0.886 and 0.817 respectively), performing better than the null model. The Figure 27 shows the receiver operating curve for both test and training data and also gives the mean AUC values across 10 replicates. The model performs better than the random expectation.

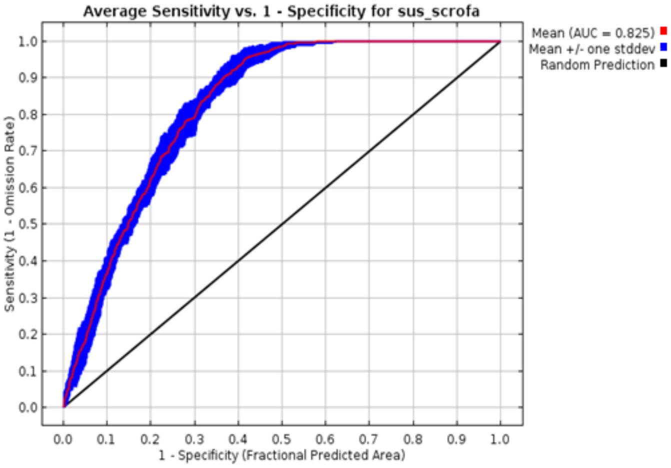


Figure 27 –AUC estimation. The red training line shows the fit of the model to the training data, while the blue interval represents the fit of the model to the testing data (across 10 replicates) and represents the real test of the model predictive power. The black line shows the random prediction.

The Maxent estimation for the present distribution (Figure 28) was consistent with the current distribution of the wild boar, and was able to predict even new recently colonized areas, like Finland and Sweden.

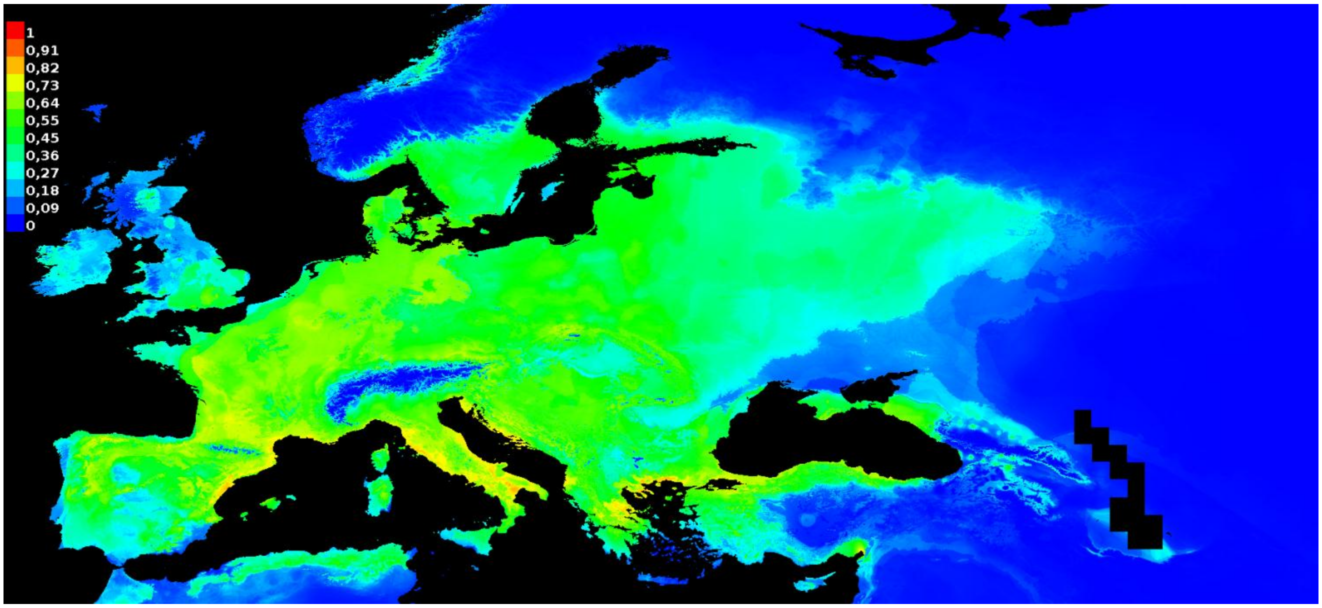


Figure 28 – Maxent prediction for the current climate.

The standard deviation (Figure 29) was relatively small, indicating that there was not much difference in the estimation across the replicates.

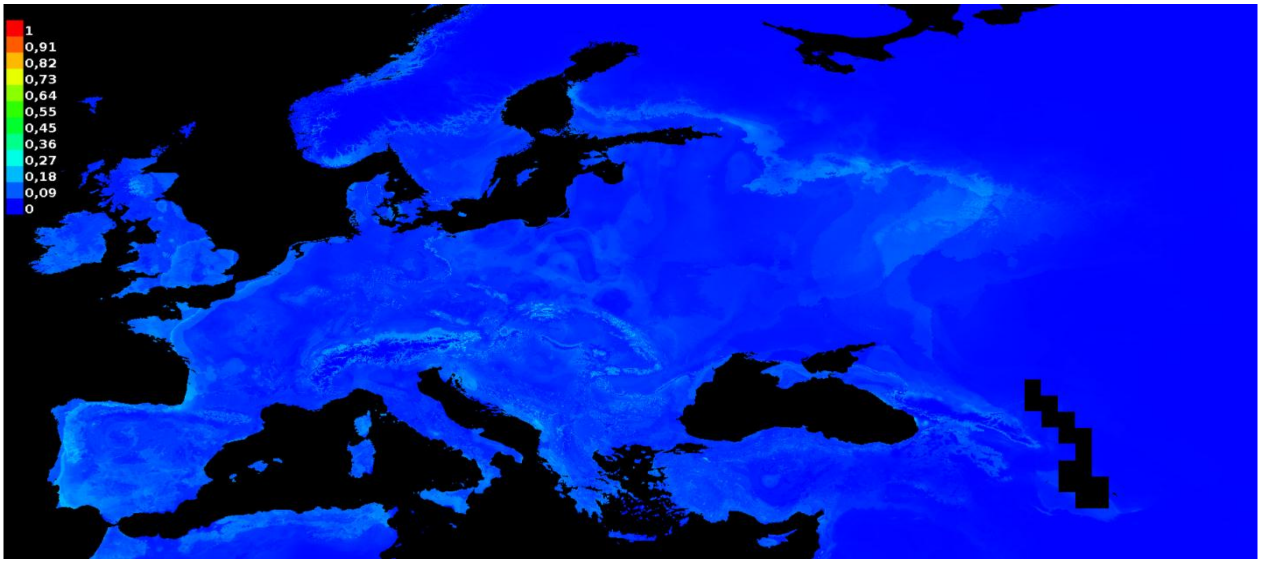


Figure 29 – Standard deviation for the current climate estimation.

Regarding the LGM distribution, both models predicted the occurrence of the wild boar in Iberia, southern France, Italy and Balkans. The CCSM model showed smaller areas of climatic suitability, especially in Italy and France, while the MIROC showed wider refugia for the wild boar, represented by a bigger predicted area (Figure 30).

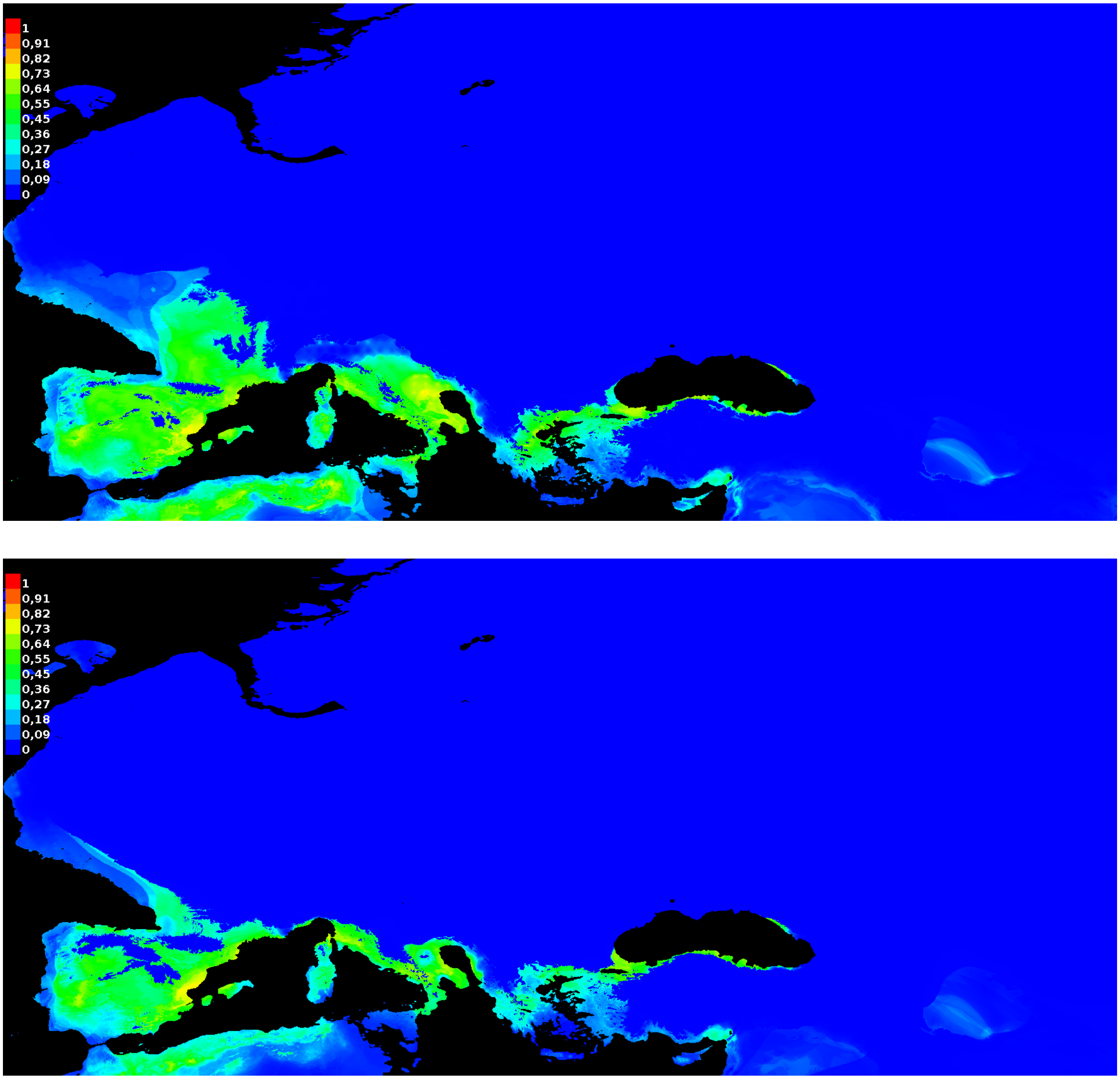


Figure 30 – LGM predictions. Above: MIROC. Below: CCSM. Probability of occurrence are represented by the color scale.

Regarding the standard deviation for the LGM estimations, both MIROC and CCSM models had low values of error (Figure 31).

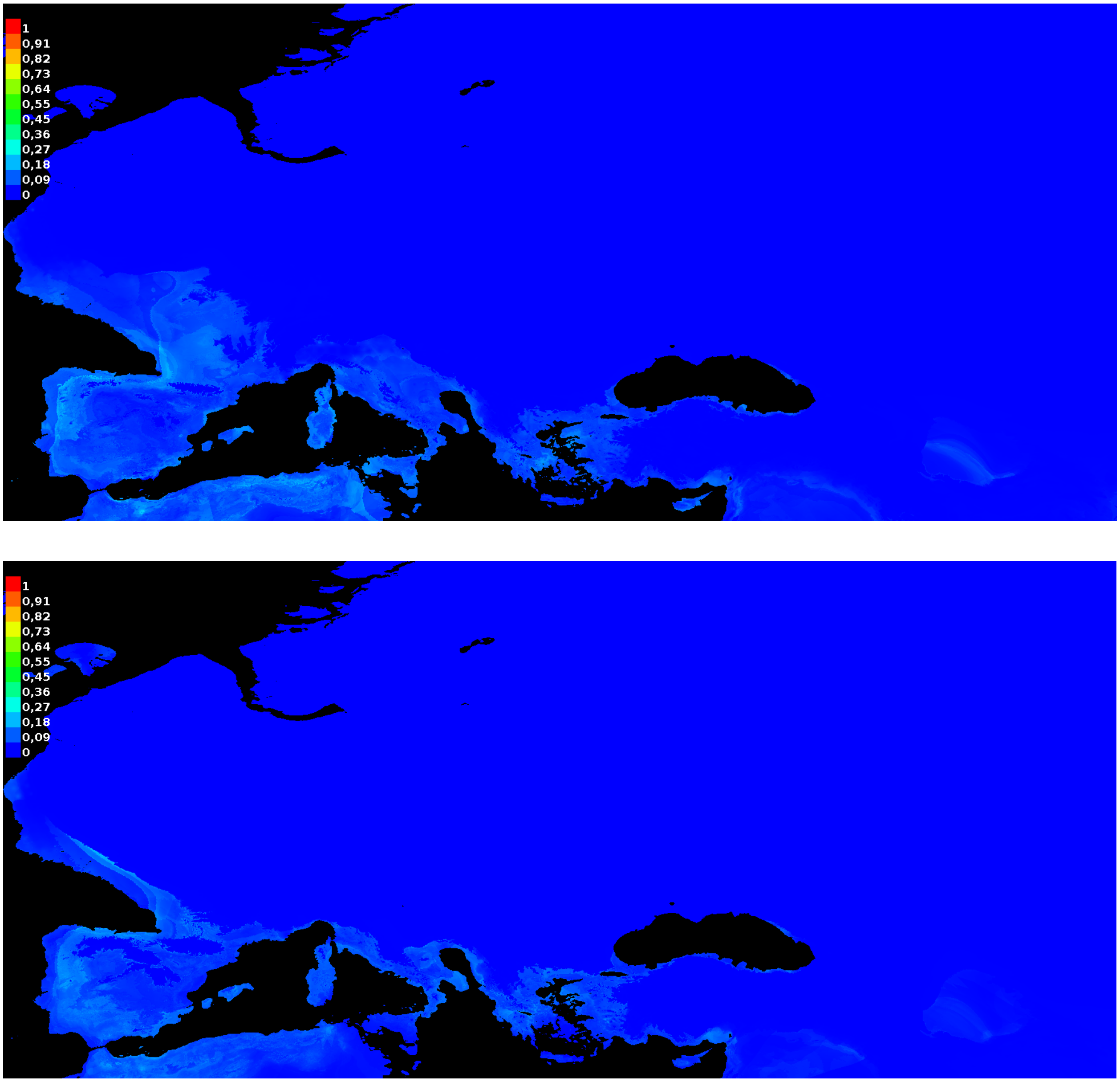


Figure 31 – Standard deviations for the predicted LGM niche. Above: MIROC. Below: CCSM.

When comparing the environmental variables used for projection to those used for training the model, it is possible to estimate the values outside the model training (MESS) and the novel climate conditions (MoD) (Figure 32). In the MESS map, it is possible to observe that most values fall in the training range, with only a few regions falling out the training (red regions). Checking the MoD map, this novel climate conditions are due to different values of snow presence (height and days) that were are observed today but not in the LGM. Most of it is due to regions with snow presence during a great part of the year (more than 300 days a year) and a height reaching up to 30 meters (the values observed today reach up to a few meters).

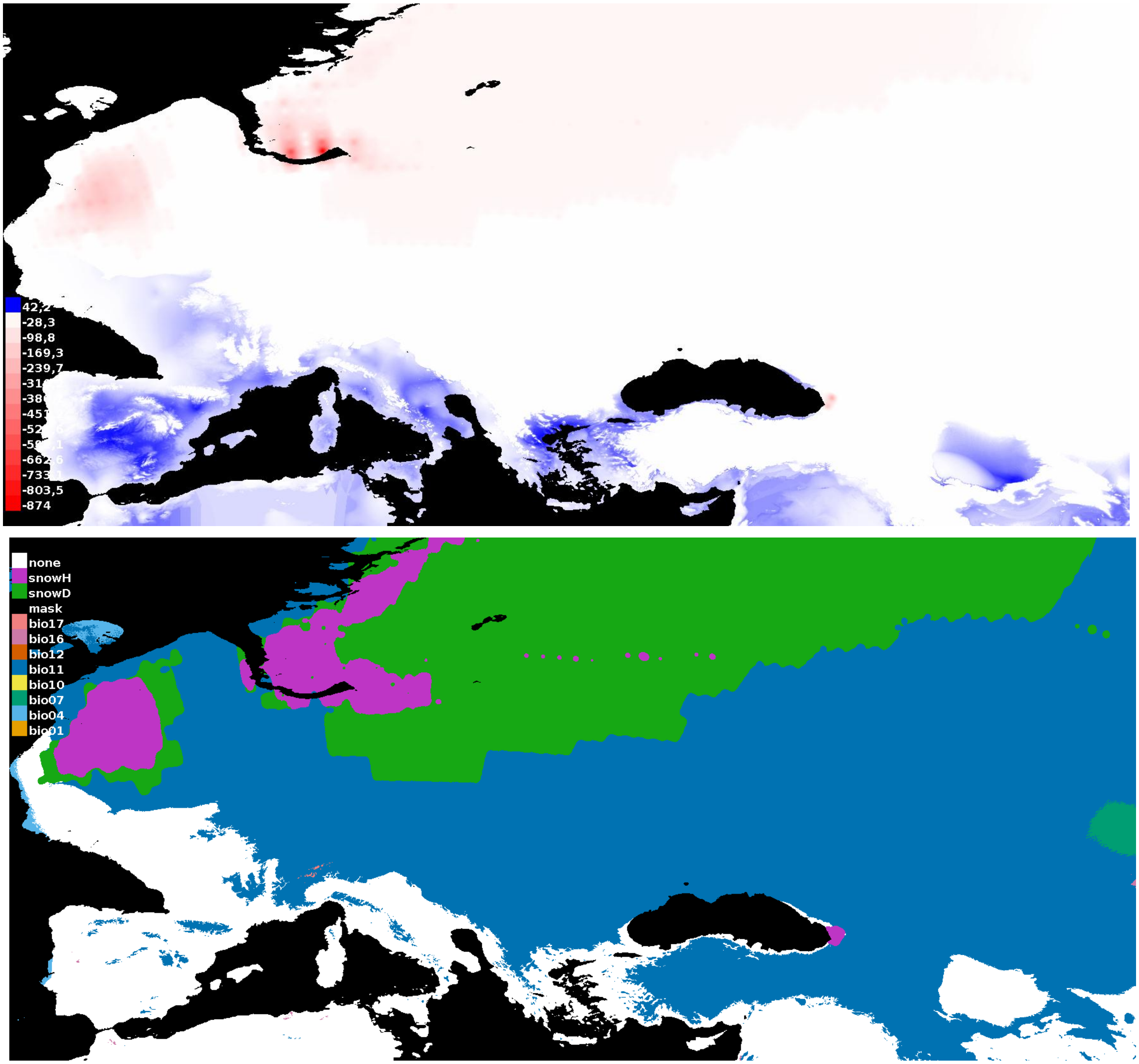


Figure 32 – MESS (above) and MoD (below) maps for the LGM projection.

Regarding the variable contribution for the model construction, a heuristic estimate of the variable contribution gives that the variable that contributed the most was the snow height (34.6%), followed by the mean annual temperature (32.8%) and precipitation of the driest quarter (6.7%). All other variables contribute less than 6% each. An alternative estimate, using a jackknife test to estimate the variable importance, shows that the environmental variable with highest gain when used in isolation is Bio07 (temperature annual range), which therefore appears to have the most useful information by itself (shown by the blue bar, Figure 33a). The environmental variable that most decreases the gain when omitted is Bio17 (precipitation of the driest quarter), which therefore appears to have the most information that is not present in the other variables (shown by the green bar, Figure 33a). When taking into consideration the test gain (instead of the training gain), the contributions of the variables remains the same (Figure 33b).

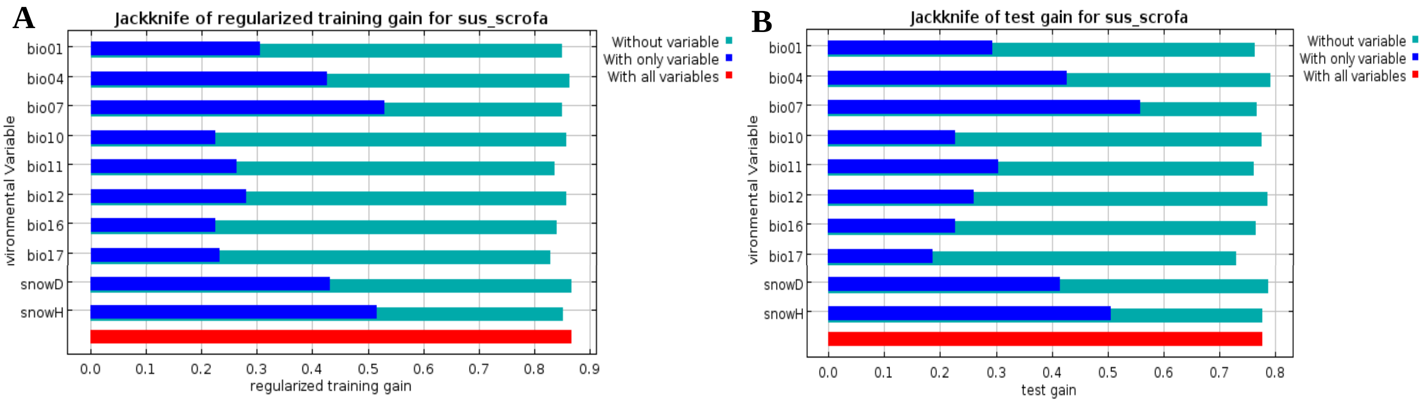


Figure 33 – Jackknife test of variable importance. Each bar represents one environmental variable. A) training gain, B) test gain.

One of the greatest advantages of Maxent is the possibility to estimate the contribution of each variable for the modeled species. This includes the understanding of which variable have the greatest influence in the model and how these variables influence species occurrence. A measure of that is the “response curves”. It is a plot that reflect the dependence of a predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables. Since the environmental variables are all somewhat dependent (mean annual temperature is influenced by the maximum and mean temperatures, for example), a representation of the variable response that is less influenced by the correlation between variables was chosen. In this set of graphs, each curve is made by generating a model using only the corresponding variable, disregarding the other variables.

It is possible to note that the highest probabilities for the wild boar presence are usually in the middle of the values distribution (Figure 34): extreme values of temperature and humidity are not tolerated.

Also the snow presence duration is a strong limiting factor, values bigger than 20 cm of snow height and more than 150 days a year of snow presence are not tolerated.

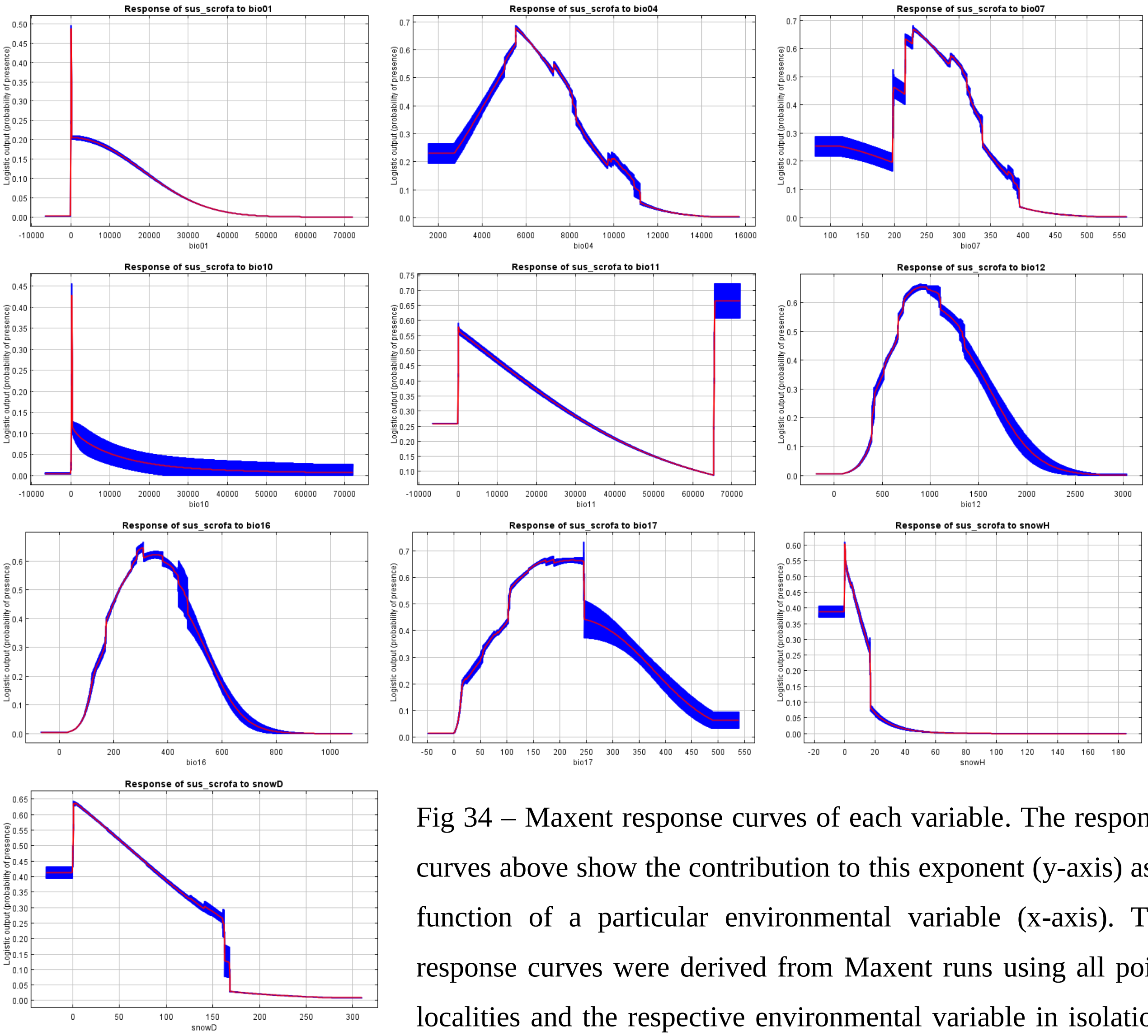


Fig 34 – Maxent response curves of each variable. The response curves above show the contribution to this exponent (y-axis) as a function of a particular environmental variable (x-axis). The response curves were derived from Maxent runs using all point localities and the respective environmental variable in isolation, variables. Environmental variables codes: Bio01: annual mean temperature, Bio04: temperature seasonality, Bio07: temperature annual range, Bio10: mean temperature of the warmest quarter, Bio11: mean temperature of the coldest quarter, Bio12: annual precipitation, Bio16: precipitation of the wettest quarter, Bio17: precipitation of the driest quarter (Bio17).

To transform the continuous predicted maps into binary (presence/absence) maps, the 10% omission of test localities was used as a threshold value. This corresponded to a probability value of 0.29 for the MIROC model and 0.31 for the CCSM model. The binary classification resulted in the estimated current distribution of the wild boar very close to the distribution maps based in the sample records (Figure 35). As observed in the probability maps, the presence during the LGM was estimated in a wider area for the MIROC when compared to the CCSM. This difference between the two models was mainly observed in southern France and continental Italy.

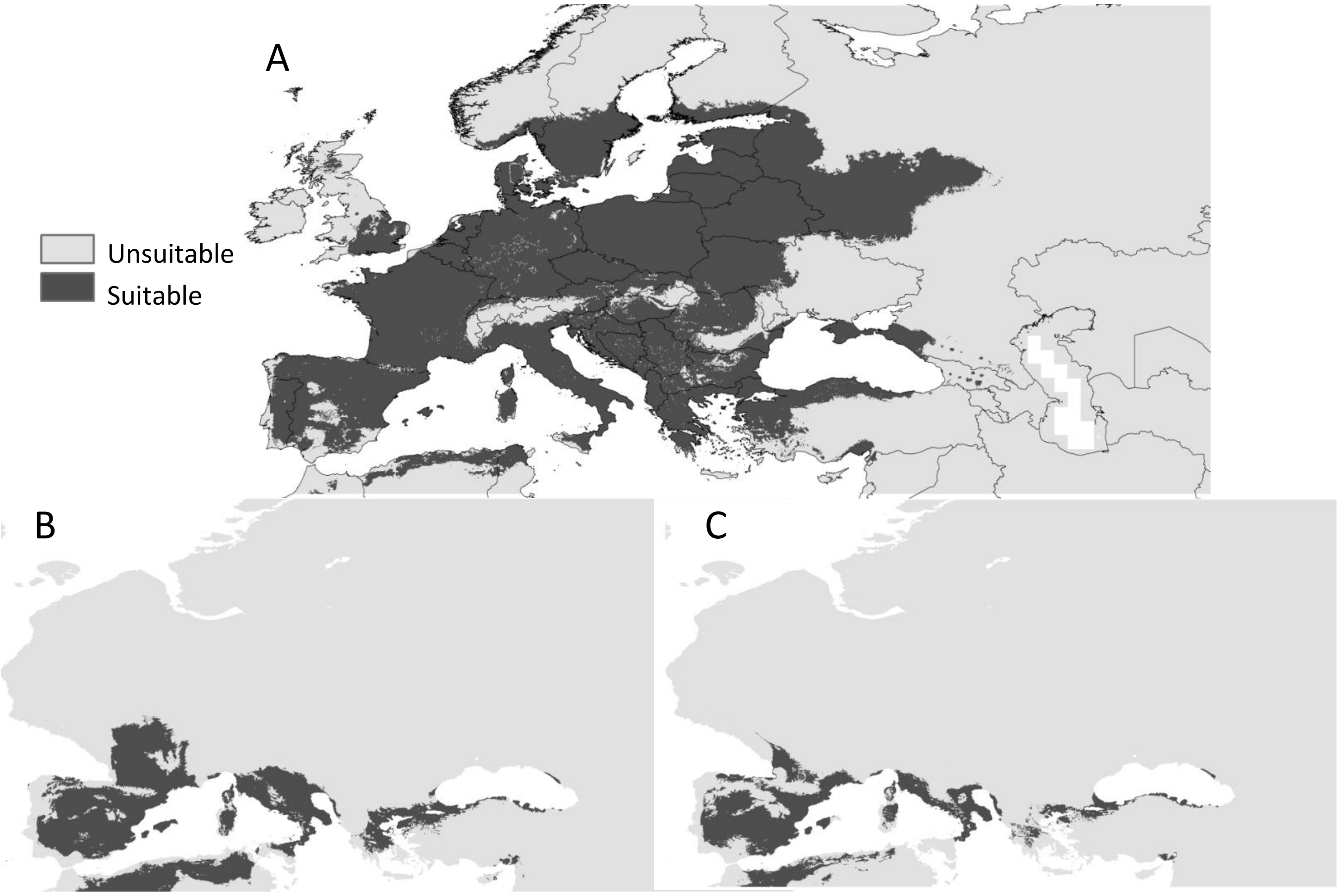


Figure 35 - Wild boar distribution range, as predicted on the basis of climatic suitability data estimated for nowadays and for the last glaciation (LGM). A) Current predicted distribution, B) LGM distribution based on the MIROC model, C) LGM distribution based on the CCSM model.

When using the binary map to plot the fossil records pre-LGM and LGM, it is possible to observe that the predictions for the LGM were consistent with the fossil record of this period in Europe (Figure 36). The MIROC model best reflected the fossil distribution, especially in southern France, where the model estimates a larger suitable area when compared to the CCSM model. Four sites where fossils have been found were not predicted by Maxent, one in Greece, one in north Italy, one in Slovenia and one in Croatia. Even though those areas were not predicted, they are all close to regions which resulted suitable for the wild boar during the LGM. For the fossils in Greece, Slovenia, and Croatia this could be an indication that the suitable areas around today's Adriatic coast were actually larger than predicted by the Maxent model. When considering the fossil in Italy, its location is around the Po River. According to Sala (2005) this entire region around the Po river had all similar climatic condition which may indicate fairly suitable conditions in this region. Excavations around this region shows that wild boar was not the most common species found in this region, indicating that boars occurred in low densities in this region (Sala 2005).

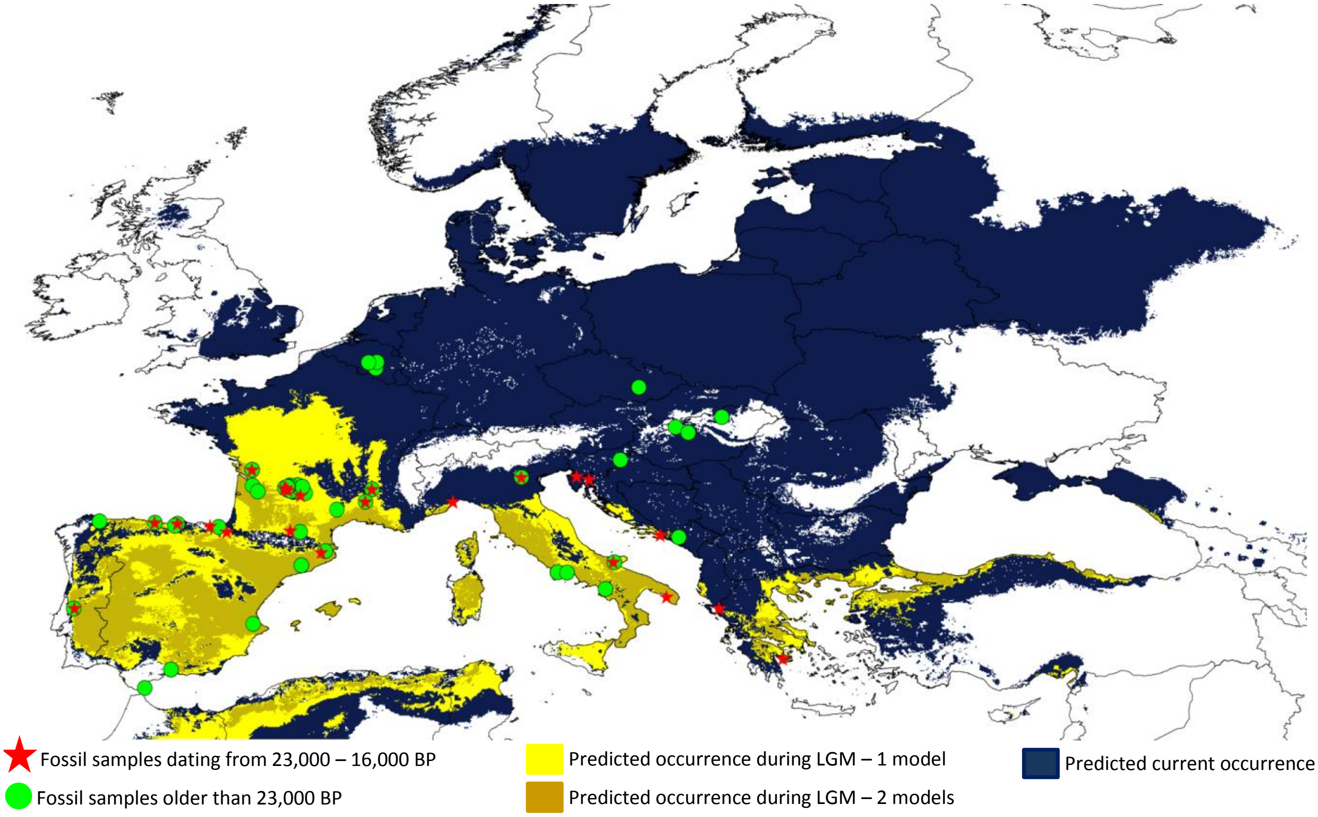

★ Fossil samples dating from 23,000 – 16,000 BP
● Fossil samples older than 23,000 BP
Predicted occurrence during LGM – 1 model
Predicted occurrence during LGM – 2 models
Predicted current occurrence

Figure 36 - Present and LGM wild boar distribution range as predicted by Maxent, compared to fossil records

Linear models

The model including geographical coordinates and the present and past (MIROC) suitability was the best fitting model to explain gene diversity (Table 6), explaining almost 24% of the variation and with strong empirical support ($\Delta_i = 0$) and the most weight of evidence ($w_i = 0.274$). Yet, it is noteworthy that all the best models in Table 3, included MIROC as predictor (corresponding to the Cum w_i of 0.76), although if the MIROC model is considered alone, resulted in a slight loss of predictive power with respect to the best model ($\Delta_i = 0.613$, $w_i = 0.202$). The strongest effect on genetic diversity indexes was indeed produced by the MIROC-predicted suitability for the species at LGM, followed by the current predicted suitability and latitude and finally by the CCSM-predicted suitability. No model showed a $\Delta_i > 10$, which represented no empirical support, but all models that had CCSM as a predictor exhibited $\Delta_i > 4$, which are models with considerably less empirical support.

Given these results, the MIROC suitability model is a better predictor of genetic variability than latitude and longitude alone. This can indicate that, other than the current values of suitability, the past values represented by the refugia locations also influenced in the current values of genetic diversity, maintaining high values of genetic diversity where the populations were more stable through time.

Table 6 – Set of linear regression models with the predicted suitability values for the LGM (MIROC and CCSM models) and for the current distribution, latitude and longitude for the 39 populations as explanatory variables and as dependent. Abbreviations: AIC_c = Akaike's Information Criterion suitable for small samples, Δ_i = ranking of model performance, w_i = Akaike's weight, Cum w_i = cumulative sum of the Akaike's weight values, $adjR^2$ = adjusted R^2.

Model predictors	**AIC_c**	**Δ_i**	**w_i**	**Cum w_i**	**$adjR^2$**
Lat+Long+MIROC+Current	-10.886	0	0.274	0.274	0.239
MIROC	-10.272	0.613	0.202	0.476	0.129
Lat+Long+MIROC	-9.239	1.647	0.120	0.597	0.171
Lat +MIROC+Current	-8.664	2.222	0.090	0.687	0.159
Lat+MIROC	-8.253	2.633	0.074	0.760	0.116
Lat	-8.101	2.786	0.068	0.828	0.079
Lat +Current	-7.056	3.830	0.040	0.869	0.088
Lat+Long+Current	-6.622	4.264	0.033	0.901	0.114
CCSM	-5.996	4.890	0.024	0.925	0.028
Lat+CCSM	-5.690	5.196	0.020	0.946	0.055
Lat+Long+CCSM	-5.289	5.598	0.017	0.962	0.083

IBD-models

When considering only geographic distance, Mantel tests failed to show a significant correlation between geographic vs. genetic distance (R^2 = 0.003, p =0.1). When testing the influence of the historical factors on genetic diversity, 10 colonization hypothesis were tested. For each of the 10 models, the two approaches (two weight values) were tested, totaling 20 models. For each of the sets, the model that considered the Iberia and East Europe as a single pool had the highest correlations (R^2 = 0.15 with two weights and R^2 = 0.17 for four weights), followed by the model that considered that Italy colonized France, Germany and Austria, and that Austria was also colonized by East Europe (R^2 = 0.10) (Figure 37, Table 7).

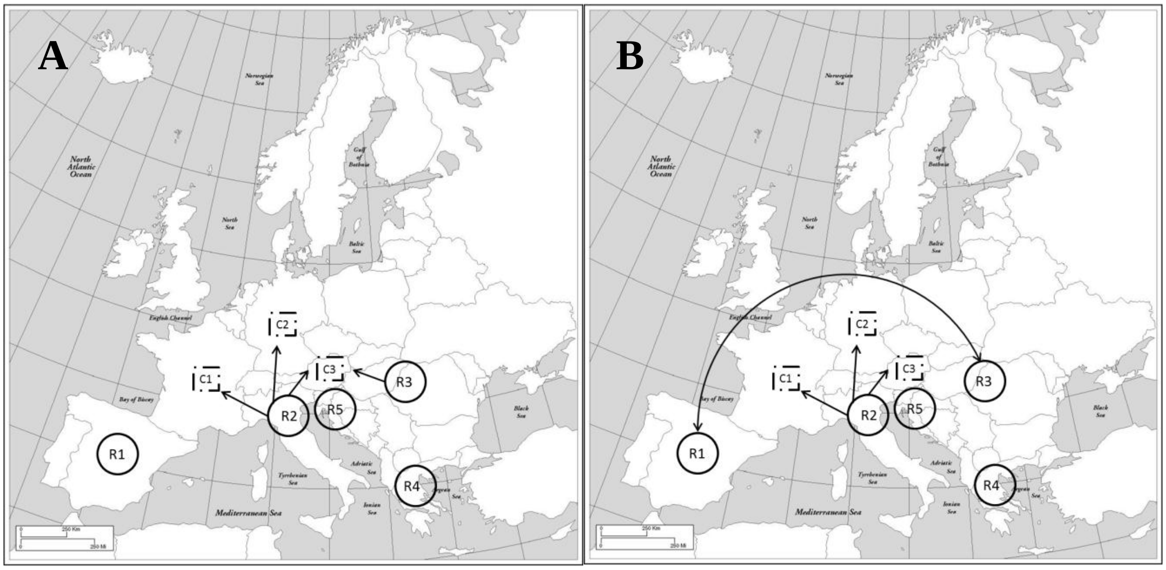


Figure 37 – Schematic representation of the two models with highest correlation values.

Table 7 – Correlation values for each of the colonization hypothesis. Models from 1 to 10 (grey shading) correspond to matrices with two weight values, and models 11 to 20 correspond to the same models but with four weights. Numbers in bold represent the two preferred models for each set of values.

model	Regression coefficient	Correlation coefficient	Determination of Y by X1(%)	P
1	0.049	0.179	0.032	0.0001
2	0.044	0.161	0.026	0.0007
3	0.046	0.168	0.028	0.0015
4	0.022	0.082	0.007	0.1217
5	0.046	0.166	0.027	0.0001
6	0.034	0.127	0.016	0.0093
7	0.037	0.138	0.019	0.0092

8	0.012	0.046	0.002	0.2412
9	0.041	0.153	**0.024**	0.0014
10	0.096	0.387	**0.150**	0.0000
11	0.069	0.290	0.084	0.0000
12	0.068	0.286	0.082	0.0001
13	0.069	0.294	0.086	0.0000
14	0.053	0.216	0.047	0.0003
15	0.068	0.282	0.079	0.0002
16	0.062	0.260	0.068	0.0001
17	0.062	0.256	0.066	0.0001
18	0.065	0.261	0.068	0.0000
19	0.075	0.316	**0.099**	0.0000
20	0.086	0.413	**0.171**	0.0000

Coalescent phylogeographic reconstruction

For the several parameters tested, the runs done with the BSP model did not converge (ESS < 200), even after 500,000,000 generations, suggesting that there is not enough information in the genetic data to infer all the parameters contained in this model. Therefore, only runs with the less parameterized constant model were considered.

The root state was located in Greece, with a posterior probability of 0.12. In fact, the root state posterior probability for all locations was not higher than 0.3, with a few exceptions in more derived branches.

The dates along the tree are somewhat in agreement with the wild boar history. The oldest records of *Sus* in Europe date back to the Early Pleistocene around 1.5-1.0 million years ago (Rook & Martinez-Navarro 2010), the time of the most recent common ancestor (TMRCA) in the Eurasian pig based on mtDNA sequences support a more recent differentiation within the species around 900,000 years ago (Scandura *et al.* 2011a), while Giuffra et al. (2000) estimated their separation in around 500,000 years ago. On the other hand, Scandura et al. (2008) estimated the divergence among the European and Italian clades in at least 50,000. The root of the estimated tree is in Greece, where supposedly the wild boar entered Europe from Near East, and where Alexandri et al. (2012) found the most basal clades of the E1 group. The estimated dates of the separation for the Italian E2 and the European E1 fall within a 95% HPD interval between 33,391-132,840 years ago, with a mean of 80,162 years ago. During the LGM, three groups were already formed, Italy, Iberia and East Europe, the three corresponding to the main refugia during the LGM. The Sardinia group also starts to appear around 6,000 years ago, period in each the wild boars were took by humans to the island, according to fossil presence (Vigne 1992).

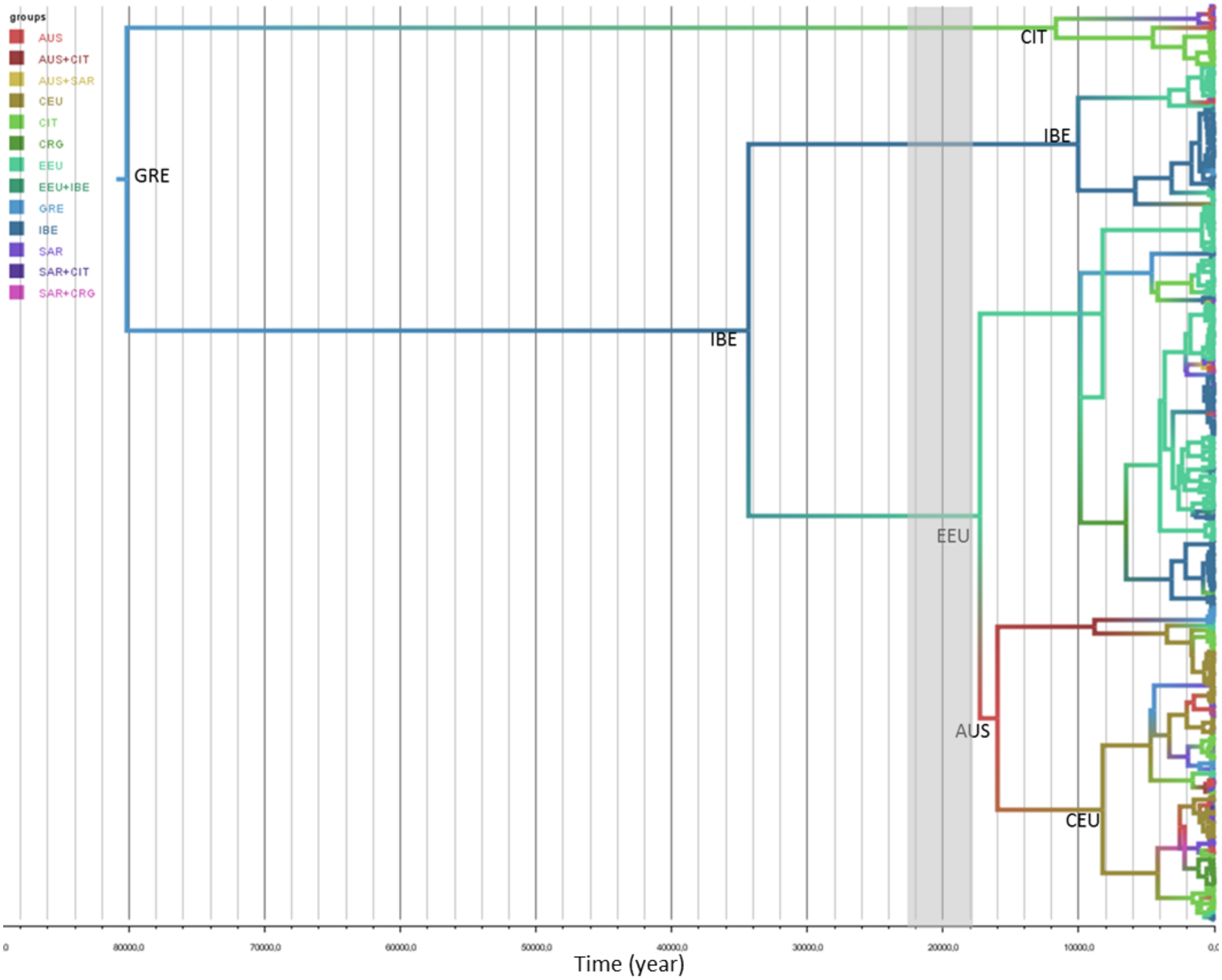


Figure 38 – Phylogeny with branches colored according to the most probable posterior location of their child nodes (see legend for group colors). Gray shading indicates LGM period.

Chapter 5

Discussion

A large database of mtDNA sequences, which included published data and almost 500 newly typed individuals, allowed the reconstruction of the phylogeographic pattern in the European wild boar and the likely evolutionary processes that shaped it.

Distribution of genetic diversity

Previous studies using more restricted sampling or fewer samples in a wide range fail in finding E2 haplotypes outside Italy. When using wider sampling and including new data, besides Italy, almost exclusively E1 haplotypes were detected in Europe. Eastern countries and the southern Balkans were actually expected to show an influence of Asian lineages. Asian wild boars previously sampled in Eastern Russia (Larson *et al.* 2005; Ramayo *et al.* 2011) carried mitochondrial sequences belonging to the main Asian lineage, thus leading to the expectation that a suture zone would exist somewhere between East Russia/North China and the sampled regions of Western Russia.

It was also surprising not finding any Near Eastern haplotype in south Greece, a possible contact zone due to its geographical proximity to Turkey. A different mitochondrial clade, basal to the European E1 and E2, had been found in specimens from Armenia and Iran (Larson *et al.* 2005) and in ancient European *Sus scrofa*, supposedly introduced from the Near East at an early stage of pig domestication (Larson *et al.* 2007a; Ottoni *et al.* 2012). Notably, it was not observed either in contemporary European populations samples nor in a large dataset of continental Greek wild boars (Alexandri *et al.* 2012). On the other hand, typical modern Near Eastern haplotypes were found in a Greek island close to Turkey (Samos island), although it was not found in continental samples. This is particularly interesting as a historical introgression of Near Eastern mtDNA sequences into Balkan populations was previously reported in other terrestrial species (e.g. brown hare, *Lepus europaeus* (Mamuris *et al.* 2010), and bicoloured shrew *Crocidura leucodon* (Dubey *et al.* 2007)) and often attributed to late Pleistocene connections across the Bosphorus.

Similarly surprising was the absence of E2 sequences outside Italy/Sardinia. E2 haplotypes have been detected by Larson and colleagues (2007a) in ancient remains from Croatia, dated 9,000 B.C., and their occurrence in the Balkans was fully plausible on the basis of the wide land bridge connecting the two peninsulas during the LGM. Yet, neither a wider sampling in the Balkan region nor a sample from another study in a close area in Croatia (Cubric-Curik *et al.* 2011) revealed any sequence belonging to E2. Nonetheless mtDNA sequences of wild boars sampled in Croatia and in North-East Italy, close to the boundary with Slovenia, were closer to those of Italian and Central European populations (E1a clade) than to any other region in the Balkans.Geographical distribution of detected subclades and single haplotypes across Europe yielded a general phylogeographic pattern which has not yet been observed in any large-scale study in Europe, although previous studies have indicated a probable similar distribution (Scandura *et al.* 2011a). The two most frequent haplotypes (H21 and H29) are shared between Iberia and Greece/Eastern Europe. Shared haplotypes need not be identical by descent, but can also be a product of homoplasy. However, in the mtDNA sequences used, Cytochrome B sequences associated to identical CR haplotypes observed in Iberia and East Europe came out to match completely (data from Scandura et al. unpublished). Accordingly, shared haplotypes occurring in Eastern and Western Europe have the same origin and sustain a connection between these two distant areas.

Considering the geographical distribution of the genetic diversity indices, like the haplotype diversity (Figure 21a, Table 2), wild boar populations from northern areas showed very low genetic diversity. Southern areas, on the contrary, showed high variation and the Italian peninsula had the highest values of haplotype and nucleotide diversity. This particularly holds true for the Sardinian population, where major mtDNA groups were shared with the Italian mainland, but many haplotypes were exclusive to the island. High levels of diversity were also detected in Iberia (with increasing diversity from Portugal to eastern Spain) and in the Balkans, where Greece showed the highest diversity, in line with Alexandri *et al*. (2012). This south-north distribution patter is usually observed in species which suffered a great impact from the low temperatures during the LGM (Hewitt 2004).

Two continental populations interestingly deviated from this North-South geographic pattern of genetic variation: Russia and Austria. Russia was the northernmost sampling region. This population resulted from aggregating nine different sampling sites, within a radius of 540 km, all having a maximum sample size of 3. Such artificial grouping, coupled with the fact that this area was possibly affected by the colonization from more eastern (unsampled) populations, may have led to high estimates of diversity. Another factor to consider is the low population density in this area (around 0.1 individuals per km^2 (Melis *et al.* 2006)) coupled with low winter temperatures, and winter harshness that is known to impose higher mortality rates in wild boar populations. These population may greatly suffer from yearly fluctuations which may reflect in their genetic differentiation from other populations in East Europe. As for the 13 individuals from Austria, they show an intermediate pattern between Italy and Eastern Europe, with typical haplotypes from both areas. Human translocations may have artificially increased the level of genetic variation in Austria, but it cannot be ruled out the hypothesis that this is actually a contact zone between Italian and Eastern European populations. Additional sampling is needed to clarify the observed pattern.

The finding of contact zones due to LGM expansion is common in European animals. Besides Austria, no other population showed a mixed pattern between two distinct macro-areas, indicating a lack of contact zones for the wild boar. This pattern may be explained by a rapid-expansion towards northern areas, which rapidly occupied the available areas after the ice-sheet retraction (more details in the section *Recolonization post-LGM*). This rapid expansion coupled with a large population number (common in wild boars due to its rapid growth rate) did not allow that neighbor areas came into contact and introgressed. A more refine sampling must be performed in the limits of the macro-regions to confirm this observation.

Genetic diversity in North Africa

The new sequences from Tunisia allowed the data to cover three continents: Africa, Europe and Asia. Among the 67 individuals sequenced, a total of four haplotypes were found: the common H029, two haplotypes that differed by one or two mutations from H029, and one (H0143) that fell outside both E1 and E2 groups (Figure 16). The wild boar is indigenous from North Africa, and in the past its occurrence ranged from Morocco through Egypt, but today they are most extinct in the region. According to the fossil record, the wild boar has been present in the Maghreb fauna from the Pleistocene throughout the Holocene (Hajji & Zachos 2011). No records of translocations from Europe are registered, which lead to the conclusion that the diversity observed in North Africa did not suffer any recent influence from Europe. Therefore, the shared haplotypes between Europe and North Africa may indicate a connection in some time period between these two areas, already reported in several animals between the regions two sides of the Strait of Gibraltar, Maghreb and Iberia (Schmitt 2007). The presence of an exclusive haplotype (H0143) probably also indicates that some population have been evolving in that area, and further studies on the neglected North African populations are necessary to enlighten the evolution of the wild boar on the African continent.

Genetic diversity and subspecies classification

The concept of subspecies is based on recognizable phenotypic differences between geographic groups or local populations of a single species. For the Eurasian wild boar, 16 different subspecies are usually recognized, divided in four groups (Groves & Grubb 1993). Within the "western races", six subspecies are described in Europe. Although subspecies are recognized for their phenotypic differences, some wild boar subspecies can also be identified as geographic groups which are also genetically structured. *S. s. scrofa*, which is distributed in middle-western Europe (from Germany to the Pyrenees), is coincident with the Central European genetic group (CEU) found in this study. The Sardinian population, recognized as *S. s. meridionalis* and phenotypically smaller than other European subspecies (Apollonio *et al.* 1988), shows a remarkable genetic separation from other European populations. *S. s. majori*, in Italy, is coincident with the Central Italian group (CIT). The Iberian subspecies, *S. s. baeticus,* ranging from west Spain to Portugal, has approximately half the distribution of the Iberian clade (IBE), while the eastern part of Spain is classified as *S. s. scrofa*. The Iberian gene pool seems to justify a former classification as a separate subspecies (*S. s. castilianus*, (Herre 1986)), which proposed that the entire Iberia was composed by a single subspecies. Regarding the eastern European races *S. s. attila* and *S. s. lybicus*, the latter has a range from Transcaucasia to Palestine while the former ranges from Hungary, Ukraine and Russia to Kazakhstan. Both subspecies have ranges that comprises the Eastern European clade, the Greek clade, and the typical Near Eastern clade (Larson *et al.* 2005), although the geographical boundaries of the subspecies distribution does not match any of the genetic clade ranges. Especially the subspecies *S. s. lybicus* has a distribution that includes three distinct clades, including also the clade that can only be found in Near East. Therefore, the observed level of homogenization found in the mtDNA does not support the current subspecies classification in East Europe or in the Near East.

A correlation between subspecies definitions and genetic data is suggested through the finding of geographically structured mtDNA haplotypes, although some subspecies distributions are not supported by clade distribution. The influence of LGM in shaping subspecies is recognized for diverse taxa in Europe and North America (Gravlund *et al.* 1998; Hamill *et al.* 2006; Hewitt 2000; Wenink *et al.* 1996). Ancient DNA data on several species indicate that pre-LGM Europe had no clearly structured populations (Hofreiter *et al.* 2004). Thus, the European wild boar subspecies might have also originated during and after the LGM. The only exception was the Central Italian clade/subspecies, which previous a study estimated a separation between E1 and E2 clades in at least 50,000 years (Scandura *et al.* 2008), and the CTMC analysis suggested a separation around 80,000 years, consequently, before the LGM.

Although the results agree with most of the current subspecies definition in Europe, caution is warranted in accepting all subspecies assignments in the European wild boar solely based on mitochondrial DNA phylogeographic divisions; mtDNA is effectively one locus and may not reflect genetic subdivisions at nuclear loci. Typing of nuclear loci in wild boar of all Europe, together with the typing of new central European samples (e.g. Belgium, Netherlands) should help to clarify the relationships of morphologically recognized subspecies and genetic groups.

Genetic diversity and management

The observed phylogeography of the species is not consistent with a large-scale homogenization possibly driven by human translocations. Growing genetic evidence suggests a limited effect of human translocations on wild boar gene pools at a wide scale (Scandura *et al.* 2011a; Vernesi *et al.* 2003), which is consistent with a prevalence of short-range translocations. An alternative explanation is that animals translocated over long distances have lower fitness than resident individuals, and have therefore a limited impact on the local gene pool.

Recent studies on the phylogeography of many European game species showed that as a general pattern, recent translocations did not blur the historical genetic diversity trend, at least when considering the large-scale phylogeographic pattern (Zachos & Hartl 2011). Large distributed mammals, such as the red deer (*Cervus elaphus*) and the roe deer (*Capreolus capreolus*) showed a similar genetic pattern, with diverse clades in Europe. The red deer showed three well distinct clades: Eastern, Western and Sardinian/African lineages (Sommer *et al.* 2008; Zachos & Hartl 2011); while the roe deer had four clades: Eastern, Western, Central and Italy (Sommer & Zachos 2009). The two abovementioned game species have a similar genetic diversity distribution as the wild boar, which an Eastern European, a Western (Iberia), one Central clade (France+Luxembourg+Germany), and two in Italy (continental and Sardinia) were found. For game species it was expected that the observed genetic pattern was influenced by recent human translocations (by showing a homogenization), but the similar patterns and the signal of structure demonstrates that it is still possible to investigate how climatic processes affected these species.

Sardinia can be considered the one of the few populations that originated from boars translocated from the continent, which became feral with the years (Larson *et al.* 2005; Scandura *et al.* 2011b). Despite Asian haplotypes are not found in Sardinia, using nuclear markers, the hybrid status of this population is evident (Scandura *et al.* 2011b; Scandura *et al.* 2008). Regarding other populations, the presence of Asian haplotypes is minimal (present only in Luxembourg and Salerno), even in other European populations (Scandura *et al.* 2008). This indicates that the impact of hybridization is insignificant in Europe and the current genetic pattern observed in the wild boar was not influenced by pig hybridization.

LGM range and refugia

Using paleoclimatic niche modeling, coupled with the observed patterns of geographical distribution of the genetic diversity it was possible to identify possible refuge areas in Europe. The existence of higher genetic diversity at lower latitudes suggests that southern areas had an important role as genetic reservoirs during the LGM. This pattern is also confirmed by the niche modeling and the presence of fossils pre- and during LGM. The results obtained with Maxent showed the following refugia areas: Balkans, Italy, Southern France and Iberia. These refugia were suggested in previous studies based on genetic data (Alexandri *et al.* 2012; Alves *et al.* 2010; Scandura *et al.* 2011a; Scandura *et al.* 2008), but this is the first time that the geographic pattern was rigorously analysed and suitability areas during the LGM were identified.

One of the challenges of species distribution is model validation, i.e. confirming the projection of the climatic niches using independent data (Nogués-Bravo 2009). Using independent data from fossil records, the putative refuge areas recovered with Maxent are in agreement with fossil presence during the LGM. Also the results from phylogeographical analysis can be considered as an independent model validation (Nogués-Bravo *et al.* 2008), with higher genetic diversity in the putative refuge, and the presence of private haplotypes.

The predicted area of past species distribution slightly differed depending on the climatic model. This difference is due to the algorithms used to simulate each of the climatic models. While the MIROC model assumes milder winter temperatures, the CCSM has more extreme temperatures, with also a wider range. Regarding the precipitation values, the MIROC has less rainfall, although the values between the two models do not differ much. One of the values that most differ between the two models is the temperature annual range (Bio07), which in fact one of the most important variables for constructing the wild boar model. The CCSM model has a much wider annual range, mainly due to the extreme winter temperatures. Since one of the limiting factors for the wild boar is low temperatures, together with the snow presence, the result that the predicted distribution with the CCSM model is smaller than the MIROC model is not surprising. Due to the presence of fossils in the area predicted for the MIROC model, combined with the linear regression results that indicated the MIROC better explains the current observed distribution of the genetic diversity, the LGM occurrence of wild boars was probably similar to the one estimated by MIROC.

Several studies on niche modeling involving different widely distributed species in Europe have been performed, including plants (Beatty & Provan 2011; Magri *et al.* 2006; Svenning *et al.* 2008), small mammals (Fløjgaard *et al.* 2009; Rebelo *et al.* 2012; Vega *et al.* 2010), butterflies (Habel *et al.* 2011; Habel *et al.* 2010), gastropods (Cordellier & Pfenninger 2009; Weigand *et al.* 2012), the extinct wooly mammoth (Nogués-Bravo *et al.* 2008) and spotted hyena (Varela *et al.* 2010). Generally, most species had suitable areas in at least one of the classic refugia in the Mediterranean peninsulas (Balkans, Italy, Iberia), and the tolerance to cold determined the occurrence in more northern areas. Regarding game species, only one study was published so far (red deer, Banks *et al.* 2008). Using a modeling method based on genetic algorithms (GARP, Stockwell & Peters 1999) the red deer distribution during the LGM is similar to the one of the wild boar, with predicted refugia in the Balkans, Italy, Southern France and Iberia. Although comparisons between GARP and Maxent results should be seem with caution (Phillips *et al.* 2006), the red deer fossil record during the LGM (Banks *et al.* 2008; Sommer *et al.* 2008) is similar to the one of the wild boar, corroborating the idea that greater genetic diversity exhibit by both animals in southern areas are mainly due to common refugia, and that current differences observed in a broad range phylogeography is due to post-LGM dispersal events.

Recolonization post-LGM

With the ice-sheet retraction after the LGM, animals and plants that were restricted to more southern areas due to severe climatic conditions started to recolonize the northern areas of Europe. Four main paradigms of postglacial re-colonization of Central and Northern Europe by southern refugia areas emerged as a common pattern among several European species (Habel *et al.* 2005; Hewitt 2000). Considering the similarity between individual populations and macro-areas, the recolonization routes followed by the wild boar can be classified as the fourth paradigm ("butterfly" paradigm, Figure 4d). The similarity between the group from France+Luxembourg+Germany and Italy evidences a dispersal to Central Europe from the Italian refugia, and no contribution from East and Western refugia. This similarity is also corroborated by the IBD models tested, the preferred colonization model was the one which Italy colonized France and Germany.

The East European clade was the only one with a significant signal of range expansion (supported by the Fu's *Fs* test): the southern refugia restricted to the Balkans area expanded towards northern areas, colonizing the entire East Europe as far as Russia. Despite the range expansion of the eastern European clade, one haplotype group, typical of south Greece is still restricted to that area, similar to the "grasshopper" paradigm. Even though the neutrality tests showed an expansion of the East European macro-region, no other macro-region showed a clear indication of expansion, with non-significant neutrality tests and, with exception of Sardinia group, no expansion signals are found with the Bayesian Skyline Plot. Also the mismatch distributions did not show a clear unimodal pattern for any population, except for the Iberian macro-region. The attempt to distinguish between scenarios using the ABC recovered a similar trend showed for the BSPs. This could indicate that either the small fragment of the D-loop is not informative enough or that the simple predictions of these models are not sufficiently different and could not be easily discriminated using only the D-loop (Porretta *et al.* 2013), and thus, more loci are necessary. The use of more data, such as the SNPs developed by (Amaral *et al.* 2011) and already typed for more than 3,000 wild and domestic samples (Groenen *et al.* 2012) (data not publicly available yet) may help to distinguish between these and even more complicated scenarios.

The geographic expansion of Italian population is not expected, since the Alps are a major barrier for most species, becoming a major feature in shaping the phylogeographic pattern (Hewitt 2004). Within the four paradigms, both the "hedgehog" and the "butterfly" colonization models have the Italian lineage as contributors of northern regions colonization, but most species do not exhibit any expansion from Italy because of the presence of Alps (Hewitt 1999). The similarity between Italian and Central Europe wild boar populations suggests a colonization by the Italian populations, despite the Alps. This dispersing pattern was suggested to reflect a structure of quickly expanding species (Schmitt 2007), and although the wild boar is a r-strategist (high ecological adaptability, opportunistic feeding and high reproductive potential) (Scandura *et al.* 2008) which can justify the rapid expanding, this recolonization may be also density-dependent. Climatic suitability values are strongly correlated with population density (Oliver *et al.* 2012), and the Maxent results indicate higher suitability values in Italy and

Southern France during the LGM. Higher local densities are expected in this region than other refugia, and since between France and Germany no great physical obstacles are present, once outside the Alps the diffusion process is much easier. Because Italian typical haplotypes (E2) are not found outside the peninsula, a few hypothesis on this dispersion event can be considered: the leading-edge mechanism of recolonization could explain the absence of E2 haplotypes in outgoing stocks (or the following extinction of low frequency haplotypes representing this clade). Another hypothesis would be that the French South-Western refugia was the biggest contributor to the recolonization of central Europe. Current populations located in this refugia area show a majority of Italian (A-like) haplotypes, but also one private haplotype and one haplotype that is found in low-frequency in North-Italian populations, indicating a close relationship with Italy (a probable ancient colonization from Italy or one continuous population), but also a distinct genetic pool. Once the climate became milder, the South-Western French and northern Italian populations expanded towards east, recolonizing a previous unoccupied area.

One pattern which was not observed in any large scale population study in Europe was the genetic similarity between Iberia and East Europe. Two highly frequent haplotypes are shared between Iberia and East Europe (lineage C in Larson *et al.* 2005). Even though both haplotypes may not be identical by descent, and be a product of homoplasy, the associated sequence of Cytochrome B for both haplotypes is the same in Iberia and East Europe. Given this result, it is possible to infer that the same haplotype (with the same origin) occur in both East Europe and Iberia.

What may have caused this unique pattern? Investigating the fossil evidence in Europe pre- and during the LGM (Figure 36), pre-LGM fossil (older than 23,000 years ago) from the wild boar were found in northern areas, while during the LGM fossils are mainly restrict to refugia areas (Balkans, Italy, Southern France and Iberia). In a scenario pre-LGM which the wild boar was distributed throughout Europe, the similarity between East Europe and Iberia may be a reflection of the pre-LGM distribution, where Iberia, Central Europe and East Europe formed one single group, with similar haplotype composition. This scenario is consistent with previous findings which little phylogeographic pattern existed in European mammals pre-LGM (Hofreiter *et al.* 2004). During the LGM, the permafrost reached as far as 45°N (Hewitt 1999), thus isolating the Balkans from Iberia. When the ice cap retreated, the Italian populations in the northern limit expanded their range to more northern areas. Once this population filled the space and occupied the until vacant niche, it was much more difficult for other populations to expand outside their LGM ranges. This isolated the two Europe extremes, although they still hold some level of shared diversity related to pre-LGM gene flow processes.

Chapter 6

Conclusions

Through the study of the control region of the mitochondrial DNA of samples covering the major part of the European continent it was possible to conclude that:

- Three main clades were recovered for European sequences: one Asian, one distributed along the continent, and one exclusive of Italy
- Samples from North Africa show typical European alleles and a possible private clade
- The majority of the subspecies recognized today correspond to a genetic differentiated populations
- Seven genetic macro-regions were identified among the European samples
- Four single populations showed a remarkable separation from the others, probable due to specific processes (drift and/or recent translocations) that affected them
- A south-north decreasing pattern of genetic diversity was observed
- The great majority of macroregions did not show any signal of population expansion
- The snow presence and low temperatures were one of the main limiting factors to the presence of the wild boar in certain regions
- Besides the three main refuge areas already identified in previous studies (Italy, Balkans and Iberia), Southern France and North Adriatic were also identified as possible refugia for the wild boar during the LGM
- After the LGM, the area responsible for colonizing central-northern regions was Italy
- An unexpected pattern of similarity between East Europe and Iberia was observed, probably a reflection of pre-LGM connectivity among these areas.

References

Adamic M, Jerina K, Apollonio M, Andersen R, Putman R (2010) Ungulates and their Management in Slovenia. *In:* APOLLONIO M, ANDERSEN R, PUTMAN R (eds.) *European Ungulates and Their Management in the 21st Century.*

Alexandri P, Triantafyllidis A, Papakostas S, Chatzinikos E, Platis P*, et al.* (2012) The Balkans and the colonization of Europe: the post-glacial range expansion of the wild boar, *Sus scrofa*. *Journal of Biogeography,* 39: 713-723.

Alves E, Ovilo C, Rodriguez MC, Silio L (2003) Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. *Animal Genetics,* 34: 319-324.

Alves PC, Pinheiro I, Godinho R, Vicente J, Gortazar C*, et al.* (2010) Genetic diversity of wild boar populations and domestic pig breeds (*Sus scrofa*) in South-western Europe. *Biological Journal of the Linnean Society,* 101: 797-822.

Amaral AJ, Ferretti L, Megens HJ, Crooijmans RPMA, Nie HS*, et al.* (2011) Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA. *Plos One,* 6.

Apollonio M, Randi E, Toso S (1988) The Systematics of the Wild Boar (*Sus scrofa* L) in Italy. *Bollettino Di Zoologia,* 55: 213-221.

Araujo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography,* 33: 1677-1688.

Arbogast BS, Kenagy G (2008) Comparative phylogeography as an integrative approach to historical biogeography. *Journal of Biogeography,* 28: 819-825.

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling,* 157: 101-118.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T*, et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual review of ecology and systematics*: 489-522.

Baldwin RA (2009) Use of Maximum Entropy Modeling in Wildlife Research. *Entropy,* 11: 854-866.

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution,* 16: 37-48.

Banks WE, D'errico F, Peterson AT, Kageyama M, Colombeau G (2008) Reconstructing ecological niches and geographic distributions of caribou (*Rangifer tarandus*) and red deer (*Cervus elaphus*) during the Last Glacial Maximum. *Quaternary Science Reviews,* 27: 2568-2575.

Bartos L, Kotrba R, Pintir J (2010) Ungulates and their management in the Czech Republic. *In:* APOLLONIO M, ANDERSEN R, PUTMAN R (eds.) *European Ungulates and Their Management in the 21st Century.* Cambridge University Press. Cambridge, UK.

Beatty GE, Provan J (2011) Comparative phylogeography of two related plant species with overlapping ranges in Europe, and the potential effects of climate change on their intraspecific genetic diversity. *Bmc Evolutionary Biology,* 11.

Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics,* 162: 2025-2035.

Bennett KD (1990) Milankovitch Cycles and Their Effects on Species in Ecological and Evolutionary Time. *Paleobiology,* 16: 11-21.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology,* 19: 2609-2625.

Bieber C, Ruf T (2005) Population dynamics in wild boar Sus scrofa: ecology, elasticity of growth rate and implications for the management of pulsed resource consumers. *Journal of Applied Ecology,* 42: 1203-1213.

Birungi J, Arctander P (2000) Large sequence divergence of mitochondrial DNA genotypes of the control region within populations of the African antelope, kob (*Kobus kob*). *Molecular Ecology,* 9: 1997-2008.

Carranza J (2010) Ungulates and their management in Spain. *In:* APOLLONIO M, ANDERSEN R, PUTMAN R (eds.) *European Ungulates and Their Management in the 21st Century.* Cambridge University Press. Cambridge, UK.

Cordellier M, Pfenninger M (2009) Inferring the past to predict the future: climate modelling predictions and phylogeography for the freshwater gastropod *Radix balthica* (Pulmonata, Basommatophora). *Molecular Ecology,* 18: 534-544.

Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM*, et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics,* 24: 2713-2719.

Crandall KA, Templeton AR (1993) Empirical Tests of Some Predictions from Coalescent Theory with Applications to Intraspecific Phylogeny Reconstruction. *Genetics,* 134: 959-969.

Cubric-Curik V, Atlija M, Sprem N, Curik I (2011) Mitochondrial DNA diversity in wild boars from the Istria and Cres Island. *Agriculturae Conspectus Scientificus,* 76: 321-324.

De Smith MJ, Goodchild MF, Longley PA (2009) *Geospatial Analysis: a Comprehensive Guide to Principles, Techniques and Software Tools,* Troubador Publishing Ltd, Leicester, UK.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *Bmc Evolutionary Biology,* 7.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution,* 29: 1969-1973.

Dubey S, Cosson JF, Vohralík V, Kryštufek B, Diker E*, et al.* (2007) Molecular evidence of Pleistocene bidirectional faunal exchange between Europe and the Near East: the case of the bicoloured shrew (Crocidura leucodon, Soricidae). *Journal of Evolutionary Biology,* 20: 1799-1808.

Eckert CG, Samis KE, Lougheed SC (2008) Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Molecular Ecology,* 17: 1170-1188.

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution,* 1: 330-342.

Elith J, Phillips SJ, Hastie T, Dudik M, Chee YE*, et al.* (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions,* 17: 43-57.

Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology,* 41: 263-274.

Excoffier L (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology,* 13: 853-864.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources,* 10: 564-567.

Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM*, et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 104: 17614-17619.

Fang M, Berg F, Ducos A, Andersson L (2006) Mitochondrial haplotypes of European wild boars with 2n = 36 are closely related to those of European domestic pigs with 2n = 38. *Animal Genetics,* 37: 459-464.

Fang MY, Andersson L (2006) Mitochondrial diversity in European and Chinese pigs is consistent with population expansions that occurred prior to domestication. *Proceedings of the Royal Society B-Biological Sciences,* 273: 1803-1810.

Fang MY, Larson G, Ribeiro HS, Li N, Andersson L (2009) Contrasting Mode of Evolution at a Coat Color Locus in Wild and Domestic Pigs. *Plos Genetics,* 5.

Fernandez-De-Mera IG, Gortazar C, Vicente J, Hofle U, Fierro Y (2003) Wild boar helminths: risks in animal translocations. *Veterinary Parasitology,* 115: 335-341.

Ferreira E, Souto L, Soares AMVM, Fonseca C (2009) Genetic structure of the wild boar population in Portugal: Evidence of a recent bottleneck. *Mammalian Biology,* 74: 274-285.

Fløjgaard C, Normand S, Skov F, Svenning JC (2009) Ice age distributions of European small mammals: insights from species distribution modelling. *Journal of Biogeography,* 36: 1152-1163.

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of Populations. *Genetics,* 174: 875-891.

Franklin J (2009) *Mapping Species Distributions: Spatial Inference and Prediction,* Cambridge University Press, Cambridge, UK.

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. *Genetics,* 147.

Gaillard JM, Pontier D, Brandt S, Jullien JM, Allaine D (1992) Sex-Differentiation in Postnatal-Growth Rate - a Test in a Wild Boar Population. *Oecologia,* 90: 167-171.

Genov P, Nikolov H, Massei G, Gerasimov S (1991) Craniometrical Analysis of Bulgarian Wild Boar (*Sus scrofa*) Populations. *Journal of Zoology,* 225: 309-325.

Genov PV (1999) A review of the cranial characteristics of the Wild Boar (*Sus scrofa* Linnaeus 1758), with systematic conclusions. *Mammal Review,* 29: 205-238.

Giuffra E, Kijas JMH, Amarger V, Carlborg O, Jeon JT*, et al.* (2000) The origin of the domestic pig: Independent domestication and subsequent introgression. *Genetics,* 154: 1785-1791.

Goulding MJ 1998. Current Status and Potential Impact of Wild Boar (Sus scrofa) in the English Countryside: a Risk Assessment. Conservation Management Division C, Ministry of Agriculture, Fisheries and Food.

Graham CH, Hijmans RJ (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography,* 15: 578-587.

Gravlund P, Meldgaard M, Paabo S, Arctander P (1998) Polyphyletic origin of the small-bodied, high-arctic subspecies of tundra reindeer (*Rangifer tarandus*). *Molecular Phylogenetics and Evolution,* 10: 151-159.

Groenen MaM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y*, et al.* (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* 491: 393-398.

Groves CP, Grubb P (1993) The Eurasian suids: *Sus* and *Babyrousa*. *In:* OLIVER WLR (ed.) *Pigs, Peccaries and Hippos - Status Survey and Conservation Action Plan.* IUCN/SCC. Gland, Switzerland.

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters,* 8: 993-1009.

Habel JC, Lens L, Rodder D, Schmitt T (2011) From Africa to Europe and back: refugia and range shifts cause high genetic differentiation in the Marbled White butterfly Melanargia galathea. *Bmc Evolutionary Biology,* 11.

Habel JC, Schmitt T, Meyer M, Finger A, Rodder D*, et al.* (2010) Biogeography meets conservation: the genetic structure of the endangered lycaenid butterfly Lycaena helle (Denis & Schiffermuller, 1775). *Biological Journal of the Linnean Society,* 101: 155-168.

Habel JC, Schmitt T, Muller P (2005) The fourth paradigm pattern of post-glacial range expansion of European terrestrial species: the phylogeography of the Marbled White butterfly (Satyrinae, Lepidoptera). *Journal of Biogeography,* 32: 1489-1497.

Hajji GE, Zachos FE (2011) Mitochondrial and nuclear DNA analyses reveal pronounced genetic structuring in Tunisian wild boar *Sus scrofa*. *European Journal of Wildlife Research,* 57: 449-456.

Hamill RM, Doyle D, Duke EJ (2006) Spatial patterns of genetic diversity across European subspecies of the mountain hare, Lepus timidus L. *Heredity,* 97: 355-365.

Hamilton G, Currat M, Ray N, Heckel G, Beaumont M*, et al.* (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics,* 170: 409-417.

Hartley M (2010) Qualitative risk assessment of the role of the feral wild boar (*Sus scrofa*) in the likelihood of incursion and the impacts on effective disease control of selected exotic diseases in England. *European Journal of Wildlife Research,* 56: 401-410.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution,* 22: 160-174.

Herre W (1986) *Sus scrofa* Linnaeus, 1758 - Wildschwein. *In:* NIETHAMMER J, KRAPP F (eds.) *Handbuch der Säugetiere Europas.* AULA – Verlag, Wiesbaden.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature,* 405: 907-913.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society,* 68: 87-112.

Hewitt GM (2001) Speciation, hybrid zones and phylogeography - or seeing genes in space and time. *Molecular Ecology,* 10: 537-549.

Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences,* 359: 183-195.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology,* 25: 1965-1978.

Ho SYW, Shapiro B (2011) Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources,* 11: 423-434.

Hobbs NT, Hilborn R (2006) Alternatives to statistical hypothesis testing in ecology: A guide to self teaching. *Ecological Applications,* 16: 5-19.

Hofreiter M, Serre D, Rohland N, Rabeder G, Nagel D*, et al.* (2004) Lack of phylogeography in European mammals before the last glaciation. *Proceedings of the National Academy of Sciences of the United States of America,* 101: 12963-12968.

Knowles LL (2008) Why Does a Method That Fails Continue to Be Used? *Evolution,* 62: 2713-2717.

Kusak J, Krapinec K (2010) Ungulates and their management in Croatia. *In:* APOLLONIO M, ANDERSEN R, PUTMAN R (eds.) *European ungulates and their management in the 21st century. .* Cambridge University Press. Cambridge.

Larson G, Albarella U, Dobney K, Rowley-Conwy P, Schibler J*, et al.* (2007a) Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proceedings of the National Academy of Sciences of the United States of America,* 104: 15276-15281.

Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J*, et al.* (2007b) Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in island southeast Asia and Oceania. *Proceedings of the National Academy of Sciences of the United States of America,* 104: 4834-4839.

Larson G, Dobney K, Albarella U, Fang MY, Matisoo-Smith E*, et al.* (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science,* 307: 1618-1621.

Larson G, Liu RR, Zhao XB, Yuan J, Fuller D*, et al.* (2010) Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 107: 7686-7691.

Le S, Josse J, Husson F (2008) FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software,* 25: 1-18.

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian Phylogeography Finds Its Roots. *Plos Computational Biology,* 5.

Magri D, Vendramin GG, Comps B, Dupanloup I, Geburek T*, et al.* (2006) A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist,* 171: 199-221.

Maillard D, Gaillard J, Hewison M, Ballon P, Duncan P*, et al.* (2010) Ungulates and their management in France. *European Ungulates and Their Management in the 21st Century*: 441-474.

Mamuris Z, Moutou KA, Stamatis C, Sarafidou T, Suchentrunk F (2010) Y DNA and mitochondrial lineages in European and Asian populations of the brown hare (Lepus europaeus). *Mammalian Biology,* 75: 233-242.

Mantel NA (1967) The detection of disease clustering and a generalized regression approach. *Cancer research,* 27: 209-220.

Mayer JJ, Brisbin Jr IL (2008) *Wild pigs in the United States: their history, comparative morphology, and current status,* University of Georgia Press.

Melis C, Szafranska PA, Jedrzejewska B, Barton K (2006) Biogeographical variation in the population density of wild boar (*Sus scrofa*) in western Eurasia. *Journal of Biogeography,* 33: 803-811.

Morueta-Holme N, Flojgaard C, Svenning JC (2010) Climate Change Risks and Conservation Implications for a Threatened Small-Range Mammal Species. *Plos One,* 5.

Nei M (1987) *Molecular evolutionary genetics,* Columbia University Press, New York.

Nikolov IS, Gum B, Markov G, Kuehn R (2009) Population genetic structure of wild boar *Sus scrofa* in Bulgaria as revealed by microsatellite analysis. *Acta Theriologica,* 54: 193-205.

Nogués-Bravo D (2009) Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography,* 18: 521-531.

Nogués-Bravo D, Rodiguez J, Hortal J, Batra P, Araujo MB (2008) Climate change, humans, and the extinction of the woolly mammoth. *Plos Biology,* 6: 685-692.

Okumura N, Kurosawa Y, Kobayashi E, Watanobe T, Ishiguro N*, et al.* (2001) Genetic relationship amongst the major non-coding regions of mitochondrial DNAs in wild boars and several breeds of domesticated pigs. *Animal Genetics,* 32: 139-147.

Oliver TH, Gillings S, Girardello M, Rapacciuolo G, Brereton TM*, et al.* (2012) Population density but not stability can be predicted from species distribution models. *Journal of Applied Ecology,* 49: 581-590.

Oliver W, Leus K 2008. Sus scrofa. *In:* 2012 I (ed.). IUCN Red List of Threatened Species.

Ottoni C, Flink LG, Evin A, Georg C, De Cupere B*, et al.* (2012) Pig domestication and human-mediated dispersal in western Eurasia revealed through ancient DNA and geometric morphometrics. *Molecular Biology and Evolution*: in press.

Panchal M (2007) The automation of nested clade phylogeographic analysis. *Bioinformatics,* 23: 509-510.

Pearman PB, Guisan A, Broennimann O, Randin CF (2008a) Niche dynamics in space and time. *Trends in Ecology & Evolution,* 23: 149-158.

Pearman PB, Randin CF, Broennimann O, Vittoz P, Van Der Knaap WO*, et al.* (2008b) Prediction of plant species distributions across six millennia. *Ecology Letters,* 11: 357-369.

Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography,* 34: 102-117.

Peterson AT, Nyari AS (2008) Ecological niche conservatism and pleistocene refugia in the thrush-like mourner, *Schiffornis* sp., in the neotropics. *Evolution,* 62: 173-183.

Petit JR, Jouzel J, Raynaud D, Barkov NI, Barnola JM*, et al.* (1999) Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature,* 399: 429-436.

Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology,* 12: 844-855.

Peyron O, Guiot J, Cheddadi R, Tarasov P, Reille M*, et al.* (1998) Climatic reconstruction in Europe for 18,000 yr B.P. from pollen data. *Quaternary Research,* 49: 183-196.

Phillips SJ (2008) Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). *Ecography,* 31: 272-278.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling,* 190: 231-259.

Porretta D, Mastrantonio V, Mona S, Epis S, Montagna M*, et al.* (2013) The integration of multiple independent data reveals an unusual response to Pleistocene climatic changes in the hard tick *Ixodes ricinus*. *Molecular Ecology*: n/a-n/a.

Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution,* 25: 1253-1256.

Posada D 2011. Collapse: Describing haplotypes from sequence alignments. [Online]. Website last modified on May 28, 2011 (accessed on August 11, 2011). Available at http://darwin.uvigo.es/software/collapse.html

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

R Core Team 2012. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ramayo Y, Shemeret'eva IN, Perez-Enciso M (2011) Mitochondrial DNA diversity in wild boar from the Primorsky Krai Region (East Russia). *Animal Genetics,* 42: 96-99.

Rambaut A, Drummond AJ (2007) Tracer v1.5, Available from http://beast.bio.ed.ac.uk/Tracer.

Ramirez O, Ojeda A, Tomas A, Gallardo D, Huang LS*, et al.* (2009) Integrating Y-Chromosome, Mitochondrial, and Autosomal Data to Analyze the Origin of Pig Breeds. *Molecular Biology and Evolution,* 26: 2061-2072.

Randi E, Apollonio M, Toso S (1989) The Systematics of Some Italian Populations of Wild Boar (*Sus scrofa* L) - a Craniometric and Electrophoretic Analysis. *Zeitschrift Fur Saugetierkunde-International Journal of Mammalian Biology,* 54: 40-56.

Rebelo H, Froufe E, Brito JC, Russo D, Cistrone L*, et al.* (2012) Postglacial colonization of Europe by the barbastelle bat: agreement between molecular data and past predictive modelling. *Molecular Ecology,* 21: 2761-2774.

Rodriguez-Sanchez F, Arroyo J (2008) Reconstructing the demise of Tethyan plants: climate-driven range dynamics of *Laurus* since the Pliocene. *Global Ecology and Biogeography,* 17: 685-695.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution,* 9: 552-569.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19: 1572-1574.

Rook L, Martinez-Navarro B (2010) Villafranchian: The long story of a Plio-Pleistocene European large mammal biochronologic unit. *Quaternary International,* 219: 134-144.

Sala B (2005) Mammalian faunas and environment from the Würmian Glacial Maximum of the Italian peninsula (approx. 22±2 ka cal BP). *Annali dell'Università degli Studi di Ferrara, Museologia Scientifica e Naturalistica,* volume speciale 2005: 125-129.

Scandura M, Iacolina L, Apollonio M (2011a) Genetic diversity in the European wild boar *Sus scrofa*: phylogeography, population structure and wild x domestic hybridization. *Mammal Review,* 41: 125-137.

Scandura M, Iacolina L, Cossu A, Apollonio M (2011b) Effects of human perturbation on the genetic make-up of an island population: the case of the Sardinian wild boar. *Heredity,* 106: 1012-1020.

Scandura M, Iacolina L, Crestanello B, Pecchioli E, Di Benedetto MF*, et al.* (2008) Ancient vs. recent processes as factors shaping the genetic variation of the European wild boar: are the effects of the last glaciation still detectable? *Molecular Ecology,* 17: 1745-1762.

Schmitt T (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Front Zool,* 4: 11.

Servanty S, Gaillard JM, Allaine D, Brandt S, Baubet E (2007) Litter size and fetal sex ratio adjustment in a highly polytocous species: the wild boar. *Behavioral Ecology,* 18: 427-432.

Sommer RS, Fahlke JM, Schmolcke U, Benecke N, Zachos FE (2009) Quaternary history of the European roe deer *Capreolus capreolus*. *Mammal Review,* 39: 1-16.

Sommer RS, Nadachowski A (2006) Glacial refugia of mammals in Europe: evidence from fossil records. *Mammal Review,* 36: 251-265.

Sommer RS, Zachos FE (2009) Fossil evidence and phylogeography of temperate species: 'glacial refugia' and post-glacial recolonization. *Journal of Biogeography,* 36: 2013-2020.

Sommer RS, Zachos FE, Street M, Joris O, Skog A*, et al.* (2008) Late Quaternary distribution dynamics and phylogeography of the red deer (*Cervus elaphus*) in Europe. *Quaternary Science Reviews,* 27: 714-733.

Stewart JR, Lister AM (2001) Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology & Evolution,* 16: 608-613.

Stockwell D, Peters D (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science,* 13: 143-158.

Svenning JC, Flojgaard C, Marske KA, Nogues-Bravo D, Normand S (2011) Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews,* 30: 2930-2947.

Svenning JC, Normand S, Kageyama M (2008) Glacial refugia of temperate trees in Europe: insights from species distribution modelling. *Journal of Ecology,* 96: 1117-1127.

Symonds M, Moussalli A (2011) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology,* 65: 13-21.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123: 585-595.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution,* 24: 1596-1599.

Templeton AR (2005) Haplotype trees and modern human origins. *Yearbook of Physical Anthropology, Vol 48,* 48: 33-59.

Templeton AR, Routman E, Phillips CA (1995) Separating Population-Structure from Population History - a Cladistic-Analysis of the Geographical-Distribution of Mitochondrial-DNA Haplotypes in the Tiger Salamander, Ambystoma-Tigrinum. *Genetics,* 140: 767-782.

Troy CS, Machugh DE, Bailey JF, Magee DA, Loftus RT*, et al.* (2001) Genetic evidence for Near-Eastern origins of European cattle. *Nature,* 410: 1088-1091.

Van Andel TH (2002) The climate and landscape of the middle part of the Weichselian glaciation in Europe: The Stage 3 Project. *Quaternary Research,* 57: 2-8.

Van Asch B, Pereira F, Santos LS, Carneiro J, Santos N*, et al.* (2012) Mitochondrial lineages reveal intense gene flow between Iberian wild boars and South Iberian pig breeds. *Animal Genetics,* 43: 35-41.

Varela S, Lobo JM, Rodriguez J, Batra P (2010) Were the Late Pleistocene climatic changes responsible for the disappearance of the European spotted hyena populations? Hindcasting a species geographic distribution across time. *Quaternary Science Reviews,* 29: 2027-2035.

Vatore R, Pignataro C, Vicidomini S (2007) La gestione del cinghiale (*Sus scrofa* L.) in Italia, con cenni su biologia e distribuzione (Mammalia: Suiformes: Suidae). *Il Naturalista Campano,* 32: 1-42.

Vega R, Flojgaard C, Lira-Noriega A, Nakazawa Y, Svenning JC*, et al.* (2010) Northern glacial refugia for the pygmy shrew Sorex minutus in Europe revealed by phylogeographic analyses and species distribution modelling. *Ecography,* 33: 260-271.

Velickovic N, Djan M, Obreht D, Vapa L (2012) Population genetic structure of wild boars in the West Balkan region. *Russian Journal of Genetics,* 48: 859-863.

Vernesi C, Crestanello B, Pecchioli E, Tartari D, Caramelli D*, et al.* (2003) The genetic impact of demographic decline and reintroduction in the wild boar (*Sus scrofa*): A microsatellite analysis. *Molecular Ecology,* 12: 585-595.

Vigne JD (1992) Zooarchaeology and the Biogeographical History of the Mammals of Corsica and Sardinia since the Last Ice-Age. *Mammal Review,* 22: 87-96.

Weigand AM, Pfenninger M, Jochum A, Klussmann-Kolb A (2012) Alpine Crossroads or Origin of Genetic Diversity? Comparative Phylogeography of Two Sympatric Microgastropod Species. *Plos One,* 7.

Wenink PW, Baker AJ, Rosner HU, Tilanus MGJ (1996) Global mitochondrial DNA phylogeography of holarctic breeding dunlins (*Calidris alpina*). *Evolution,* 50: 318-330.

Westley PaH, Schindler DE, Quinn TP, Ruggerone GT, Hilborn R (2010) Natural habitat change, commercial fishing, climate, and dispersal interact to restructure an Alaskan fish metacommunity. *Oecologia,* 163: 471-484.

Wilson DE, Reeder DM (2005) *Mammal Species of the World. A Taxonomic and Geographic Reference* Johns Hopkins University Press.

Wollan AK, Bakkestuen V, Kauserud H, Gulden G, Halvorsen R (2008) Modelling and predicting fungal distribution patterns using herbarium data. *Journal of Biogeography,* 35: 2298-2310.

Wotschikowsky U (2010) Ungulates and their management in Germany. *In:* APOLLONIO M, ANDERSEN R, PUTMAN R (eds.) *European Ungulates and Their Management in the 21st Century.* Cambridge University Press. Cambridge, UK.

Zachos FE, Hartl GB (2011) Phylogeography, population genetics and conservation of the European red deer *Cervus elaphus*. *Mammal Review,* 41: 138-150.