Human population genomics of the Western Mediterranean

Simone Andrea Biagini

TESI DOCTORAL UPF / 2019

DIRECTOR DE LA TESI

Dr. Francesc Calafell & Dr. David Comas

DEPARTAMENT DE CIÈNCES EXPERIMENTALS I DE LA SALUT

” Begin at the beginning, and go on till you come to the end: then stop ”

Lewis Carroll

Alice in Wonderland

Acknowledgments

This work is the result of four years of personal and professional growing. Everything I have achieved comes from the support of all those people who believed in me. I thank all of them, starting with Francesc and David who gave me the chance to achieve this goal, teaching me how to deal with my limits, and letting me learn from my mistakes with patience and trust. I thank my colleagues, now friends, who were there for me sharing the tough parts of our working days always with a smile. I personally thank André, Gerard, Lara, and Alex who were there from the very beginning teaching me the humility of a great working team. I thank Carla, Neus, Barbara, Ana, and Rocio for being such good friends and an essential part of every day I spent with them. I thank my mother and father who always believed in me, supporting all the crazy choices I made in my life, giving me love and care. To my friend Elena who shared with me this long path, no matter the distance between us. To Marcella and Carlotta for being the friends that everybody should have. To my friends from Naples who proved me that, no matter the distance between us, we will always be a family. To Francesca who supported me in many moments of my life. To my friends Maria and Stefano who always cared about me. To my friend Serena who has been by my side in the last twenty years, always in different places but with the same great affection. And thank to you Claudio, only you and I know how hard it has been living the distance between us, I never stopped loving you, and I never will.

Abstract

With its pivotal position, the Western Mediterranean basin represents a unique entity that connects different populations whose individual genetic backgrounds have been explored through different methods, but a global study has not been explored yet. In this thesis, genotype array data provide more information about Spain and France, with the aim to lay the foundation for more comprehensive studies. The results presented concern the Island of Ibiza, whose isolation was detected and related to processes of genetic drift and inbreeding, excluding any genetic continuity with the ancient Phoenician culture. Furthermore, the internal structure of France was explored through haplotype-based methods that, together with the analysis of the admixture events with the surrounding populations, provided the first image of the French genetic landscape.

Riassunto

Con la sua posizione strategica, il Mediterraneo Occidentale rappresenta un'entità unica che connette diverse popolazioni il cui background genetico è stato esplorato avvalendosi dell'utilizzo di diverse metodologie. Tuttavia, uno studio sull'intera area non è stato ancora sviluppato. In questa tesi, dati genome-wide forniscono maggiori informazioni sulla Spagna e la Francia, con lo scopo di gettare le basi per futuri studi sull'intero bacino del Mediterraneo Occidentale. I risultati presentati riguardano l'isola di Ibiza, il cui isolamento è stato identificato ed associato a meccanismi di deriva genetica ed inbreeding, escludendo qualsiasi continuità con l'antica civiltà Fenicia. Inoltre, la struttura interna della Francia è stata esplorata mediante metodi basati su aplotipi che, insieme ad analisi sui meccanismi di admixture con le popolazioni circostanti, ha fornito la prima immagine del pattern genetico di questo Paese.

Preface

The interest that historians, linguists, and anthropologists poured into the study of the dynamics that defined the Western Mediterranean space laid the foundation for the study of the population genetics of this very area.

The historian Fernand Braudel once said "*History may be divided into three movements: what moves rapidly, what moves slowly, and what appears not to move at all*"; this sentence would still make sense if instead of history we talked about population genetics: the processes of migration, admixture, demographic expansion, and the isolation ones, are perfectly described by this definition. The present work wants to explore these processes trying to define the profiles that are missing from the overall image of the Western Mediterranean basin.

Nowadays, many genome-wide and whole-genome studies became possible thanks to the availability of an increasing number of samples. However, a comprehensive image of the Western Mediterranean basin is still missing. Undoubtedly, the reason is the lack of information for some of the geographical regions forming part of this area; several studies have dug into the genetic landscape of Italy, some progresses have been made in disentangling the history of North Africa, only recent advances have been made about the genetics of Spain, while almost nothing has been told about the genetic structure of France.

In this work, we shed light on the genetic landscape of Spain, discovering an unexpected isolate: Ibiza. Furthermore, we present some of the results for the first genome-wide study on the genetic structure of the French population.

Sometimes, observing a part of something is not the same as observing the whole, and it is a purpose of this work to give voice to the missing parts of a wider story before it can be finally told.

CONTENTS

INTRODUCTION

ENTERING THE MEDITERRANEAN

In the Odyssey, Homer describes Ulysses' journey to the discovery of the Mediterranean unexplored space. The thirst for knowledge that the Greek hero exhibits has always been seen as a symbol of desire of broadening our horizons. In the Odissey, Ulysses urges his companions to follow him to the discovery of a new world; very famous is what, in the Divine Comedy, Dante makes Ulysse say:

"Considerate la vostra semenza: fatti non foste a viver come bruti ma per seguir virtute e canoscenza"

"Consider your origin; you were not born to live like brutes, but to follow virtue and knowledge"

(*Canto XXVI vv. 112-120*)

Somehow, Dante condemns these words for the consequences that come with them, but I think Ulysses' words help us to feel less alone in front of the modern desire of shedding light on the unexplored parts of our history.

Just like Ulysses, we will explore the Mediterranean basin focusing our attention on its Western part as much as possible; we will look at the journey our ancestors braved when they first entered this area, analysing the different steps that allowed us to become Mediterranean people.

BECOMING MEDITERRANEANS

The presence of anatomically modern humans in the Mediterranean area dates back to the Paleolithic roughly around 40.000 YBP (Sazzini M., Sarno S., 2014). Our ancestors were hunter-gatherers at that time; to provide themselves with sustenance, they obtained their food by fishing, hunting, scavenging, and gathering wild plants and other edibles. They did not settle in any place, but lived as nomadic groups exploring new territories, a proper demic diffusion that allowed us to reach always new parts of the world (Terrenato, 2007). Multiple migrations and movements have been detected mostly through paleoanthropological and archaeological studies; the archaeological tradition linked to the first movements into the Mediterranean area (and into the wider European space) is represented by the Proto-Aurignacian and the Aurignacian cultures (see Figure 1), both archaeological proxies for the presence of *Homo sapiens* in this very area (Hoffecker, 2009). Generally speaking, elements linked to the Proto-Aurignacian culture are evidence of the earliest modern human existence in Southwestern Europe and, together with elements belonging to the Aurignacian culture, may have represented one of two waves of cultural diffusion, both started in Western Eurasia (Falcucci *et al.*, 2017). Focusing on the Western Mediterranean area, we can find different Proto-Aurignacian sites dating back 37,000–35,000 YBP: Italy with Fumane and Mochi, Spain with Arbreda and Morin, and France with Arcy-sur-Cure, Esquicho-Grapaou, Isturitz, La Laouza, Le Piage, and Mandrin (Teyssandier, 2006). Studies on the technical traditions behind the production of different tools linked to this

period (mostly blades and bladelets) helped to differentiate the Proto-Aurignacian distribution from the Aurignacian culture that seems to be better represented in south-west France (Bon, 1960); Aurignac, in the Haute-Garonne department in south-west France is indeed considered the type site of this culture.

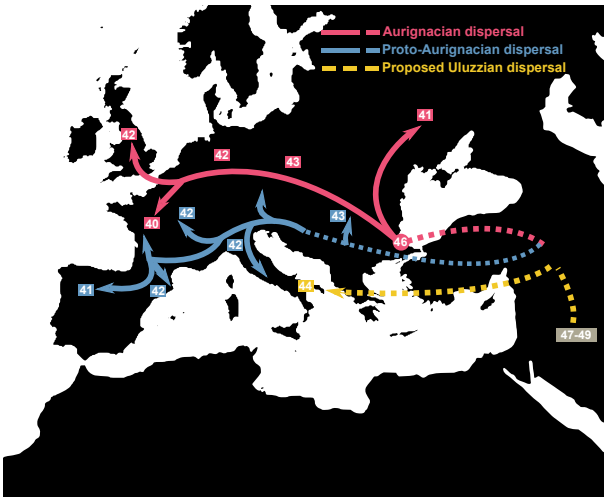


Figure 1: Dispersal routes of anatomically modern humans during the Paleolithic; numbers are years before present (KYBP) (modified from Sazzini M., Sarno S., 2014, original from Mellars, 2011).

The Aurignacian is also linked to the advent of figurative art in Europe as our ancestors began to express some sort of need of communication through several representation that can be found, among the others, in France (Chauvet, l'Aldène, l'Abri Castanet), in the northeast of the Iberian Peninsula (Cueva de Altxerri), and still in Fumane in Italy (Bourrillon *et al.*, 2015). The main subjects of these representations are animals (mostly mammoths, felines, rhinoceros and bears), but also abstract elements like lines or dots, and handprints (captivating examples are represented by the red

ochre hand prints in the Chauvet-Pont-d'Arc Cave in the Ardèche department of southern France). Possibly, this interest in depicting hunted animals had an educational purpose; they might have helped the youngest to learn what animals to hunt and how (Terrenato, 2007). The representation of human figures is very rare in this phase, but it will be more central in the following archaeological industry, the Gravettian, whose type site is again in south-west France in La Gravette (Dordogne department). One of the hallmarks of the Gravettian industry are the Venus figurines which have been interpreted as attempts to celebrate or honour the female image as symbol of fertility, or even as representations of some sort of mother goddesses (Adovasio *et al.*, 2000). The Gravettian industry dates back to 28,000 YBP in the Mediterranean area (Svoboda *et al.*, 2014) and represents the last Upper Paleolithic culture before an impressive worldwide climatic change would have led to the Last Glacial Maximum (LGM). Further evidence about the first signs of anatomically modern humans in the Mediterranean area seems to put the arrival of our ancestors earlier than expected, around 43,000-45,000 YBP (Sazzini M., Sarno S., 2014). Signs of this early events are represented in South-eastern Mediterranean Europe by the Uluzzian technology (see Figure 1), an industry that predates the early Upper Paleolithic finding its expression in different Italian sites (Grotta La Fabbrica in Tuscany, Colle Rotondo in Latium, Grotta La Cala and Grotta di Castelcivita in Campania, and Grotta del Cavallo plus Grotta Bernardini and Grotta di Uluzzo C in Apulia) and in Greece (Peloponnese and Kephalari) up to ~39,500 YBP (Douka, Peresani, Villa). It was around 25,000 KYP that a

great extent of ice sheets covered a substantial part of the northern hemisphere (Sazzini M., Sarno S., 2014); people were scarce and dealing with a deteriorating climate was challenging (Mithen, 2006). Forced by the harsh climatic changes, people from Northern and Central Europe moved to the Southern European regions encountering the already existing occupants of those areas (Sazzini M., Sarno S., 2014). Mostly, human communities survived by retreating to refugia along the Mediterranean coastline, where it was still possible to find sustenance (Mithen, 2006). The Italian, the Balkan, and the Franco-Cantabrian refugia represented a shelter for our species until conditions got improved around 16,000-13,000 KBP (Sazzini M., Sarno S., 2014; see Figure 2).

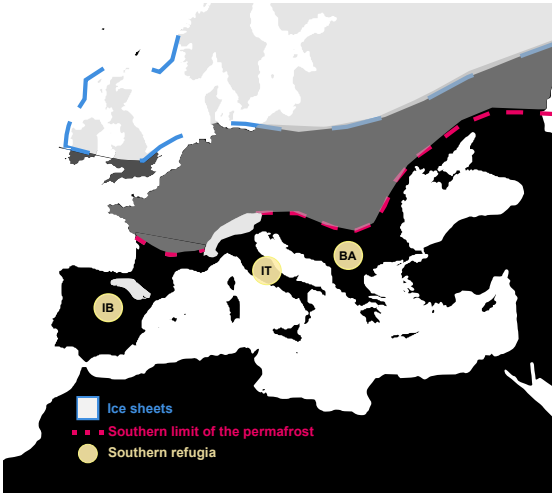


Figure 2: Representation of the three southern refugia during the last glacial maximum (LGM). IT, Italy; BA, Balkans; IB, Iberian Peninsula (modified from Sazzini M., Sarno S., 2014).

The climate improvement allowed people to leave the regions they occupied for so long, and a process of re-expansion began from

these very refugia. Around 10,000 KBP, an even milder climate created the conditions for one of the largest changes that human history have ever experienced (Mithen, 2006). According to the French archaeologist Jean Guilaine, the Mediterranean space did not truly belong to the hunter-gatherer groups; actually, a true Mediterranean identity was reached around 8,000-2,000 YBP, when the process of Neolithization made its way from the Near East through a North-western route, strongly impacting the Mediterranean human populations (Braudel, 1985). Possibly, a primitive Mediterranean civilization arose in the Levant (South Turkey, Syria, and Lebanon) during the 6th millennium BCE (8,000-7,000 YBP) and then spread through the entire Mediterranean basin (Figure 3); the adoption of agricultural innovations such as the domestication of plants and animals allowed the development of a sedentary lifestyle and the establishment of the first villages (Braudel, 1985).

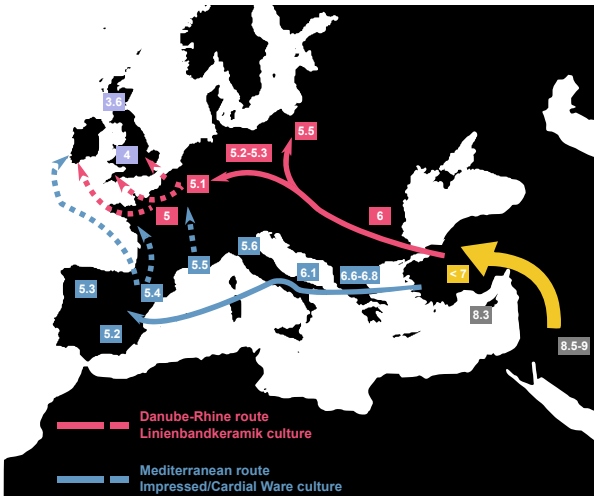


Figure 3: Neolithic spread through the Mediterranean basin (modified from Sazzini M., Sarno S., 2014, original from Tresset & Vigne, 2011).

One of the major controversies about the transition from one lifestyle to another has often regarded whether the spread of agriculture had involved a process of acculturation (farmers introducing indigenous peoples to new techniques), or whether it had resulted from demic diffusion and subsequent gene flow (farmers migrating into areas already occupied by hunter-gatherers) (Figure 4) (Jobling *et al.*, 2014).

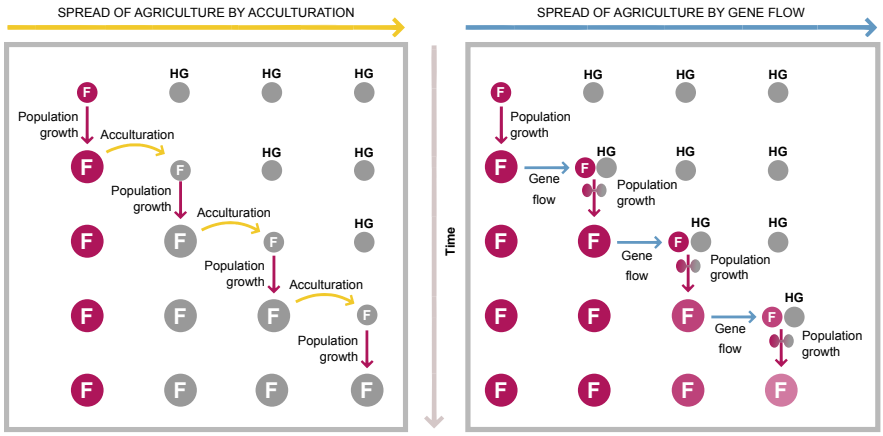


Figure 4: Proposed models for the spread of agriculture (acculturation model on the left panel; gene flow model on the right panel). (modified from Jobling, Hollox, Kivisild, & Tyler-Smith, 2014). HG stands for hunter-gatherers, F stands for farmers.

However, this old controversy has been recently settled by the outcomes of ancient DNA studies in which Europeans has been described as the product of at least three groups admixing in different proportions: western European hunter-gatherers, ancient north Eurasians, and early European farmers (Lazaridis *et al.*, 2014). When genomes from early European Neolithic farmers have been compared to those of Mesolithic hunter-gatherers, the farmers

defined a cluster clearly separated from all the hunter-gatherers that, in contrast, were found differentiated into three distinct groups: Eastern, Western, and Scandinavian (Haak *et al.*, 2015). The comparison to modern populations showed that the ancient north Eurasians contributed to the smallest proportion of modern Europeans' genome, that the western European hunter-gatherer ancestry reached the highest contribution in northern European groups, and that the ancient farmers defined the major ancestral component in southern Europeans (Lazaridis *et al.*, 2014). A further genome-wide analysis of early Neolithic farmers from the Aegean and Anatolia defined that the largest contribution to the genetic ancestry of early European farmers derived mostly from the oldest Anatolian farmers, followed by the Aegean ones, while only a small hunter-gatherer contribution was found, suggesting a low level of admixture between the early European farmers and the local hunter-gatherer groups (Hofmanová *et al.*, 2016). This evidence seems to support the demic diffusion as the main model of farming expansion into Europe (Harris, 2017). Also, studies on the ratios of X chromosome and autosomes agree with a balanced ratio between male and female migrations, suggesting the displacement of entire groups of families (Goldberg *et al.*, 2017). However, after the early Neolithic, an increase in the hunter-gatherer component was found in European samples from the middle and the late Neolithic (Haak *et al.*, 2015; Hofmanová *et al.*, 2016; Omrak *et al.*, 2016; Harris, 2017), suggesting admixture and a decreased immigration of farmers from the Aegean and Anatolia, in favor of matings between farmers and indigenous hunter-gatherers after the early Neolithic

(Harris, 2017). Archaeological evidences find traces of agriculture expansion in two distinctive cultures that co-existed in Early Neolithic: the linear pottery (or *Linienbandkeramik*) spreading from the Danube river, and the *Cardial ware* mostly associated to the spread of agriculture in the Mediterranean basin (Italy, southern France, Spain, Sardinia, and Corsica) (Jobling *et al.*, 2014). These cultures came with different techniques that allowed the separation between the different communities in terms of abilities in mastering distinct materials. Without any doubt, this generated the conditions for a subdivision of the first agricultural Mediterranean in distinct areas, even if we need to wait until the 3rd millennium BCE for the first real compartmentalization of the different communities according to the differential productions arisen during the Bronze Age. It is interesting to point out that, while Italy, Spain, and Southern France were living the advent of metallurgy, in the same period some civilization completely ignored it, developing in contrast a Megalithic culture (Jean Guilaine in Braudel, 1985). Examples of this culture can be found in Portugal and Britain with a production of portal tombs that also spread into Cantabrian Spain and Atlantic France, somehow highlighting the co-existence of different cultures in the same span of time (Arias, 1999). This internal compartmentalisation of the peoples that lived in the lands facing into the Mediterranean basin will be the starting point for the development of a long story of relations and contrasts that made the history of the "sea between the lands".

THE HISTORY THAT SHAPED US

'Tolerance does not rule over the Mediterranean. It is History that has "disfigured" the Mediterranean; and History cannot be erased'

Fernand Braudel

(Une leçon d'histoire, 1975)

As we could understand from the previous paragraph, separating the history of the Western Mediterranean from the whole basin is anything but an easy task. In order to reconstruct the crucial points of this truly complex history, it is undeniable the valuable support I found in the work of one of the greatest historians of the 20th century, the French Fernand Braudel. In one of his most remarkable works, Braudel looks at the Mediterranean space as a doctor observes a body and its functions; there is some sort of poetry in his words when he claims 'What we can be certain of is the architectural unity of which the mountains form the "skeleton": a sprawling, overpowering, ever-present skeleton whose bones show through the skin' (Braudel, 1975). Braudel explores this "skin" to understand its heart, unveiling the processes of interactions between those civilizations that shared the same space for a very long time.

Many different historical events contributed to the intricate patterns of human genetic variation in the Mediterranean area (Sazzini M., Sarno S., 2014); it is undeniable that today Mediterranean populations own an absolutely extraordinary mix of genes formed in populations that in a previous historical era had distinct characteristics (Mirko Drazen Grmek in Braudel, 1985). Several

relevant ancient civilizations flourished in this area, which was dominated for centuries by Phoenicia, Carthage, Greece, Rome, and the Arab empire (Moretti *et al.*, 2016). So far, however, defining the independent contributions of each historical event on the gene pools of the different Mediterranean populations is still a difficult task (Sazzini M., Sarno S., 2014).

Around 3,200 YBP, the Late Bronze Age collapse of Eastern Mediterranean societies laid the foundations for the expansion of new civilizations, redefining the Mediterranean social and economic order (Knapp *et al.*, 2016). From the ancient Levantine region of Canaan (corresponding to modern Lebanon plus parts from Syria, Jordan, and Israel), the Phoenician civilization spread through the entire Mediterranean basin founding cities in North Africa (Carthage in modern Tunisia), South Iberia, the Balearic Islands (Ibiza represented one of the most flourishing colonies), Sardinia, Sicily, Malta, and Cyprus (Zalloua *et al.*, 2018). Always at the same time, from the Aegean Sea, the Greek population expanded through several Mediterranean sites; they founded Magna Graecia (formed by the coastal areas of Southern Italy and Sicily), and also colonized the Iberian and French Mediterranean coastline (Sazzini M., Sarno S., 2014; Moretti *et al.*, 2016). Both Phoenicians and Greeks took part in the economic and political resurgence that came with the Iron Age (Knapp *et al.*, 2016).

Later, the intricated process of colonization carried out by the Romans begun in Central Italy and expanded through the entire

Peninsula during the 4th and 3rd centuries BCE. Internal civil wars during several years harshly impacted the stability of this first phase and, eventually, the Roman Republic was weakened at the end of the first century (Kulikowski, 2017); the Roman Empire was founded and with it the domain of the entire Mediterranean coastline. Romans were the masters of the entire basin that they used to call Mare Nostrum (which was the Latin for "Our Sea"), and their role in the economic and social order became significant, since they represented the power that connected all the populations living in the Mediterranean basin; at least until the Roman Empire fell around 1,500 YBP (476 CE) (Sazzini M., Sarno S., 2014; Moretti *et al.*, 2016).

In the 7th century CE, after the collapse of the Roman Empire, the Mediterranean basin was characterized by the Arab expansion from the Arabian Peninsula. The first wave of colonization involved the entire North African coastline with the submission of the local African Berber groups. A second wave of expansion followed during the 8th century, when the colonization reached the Iberian Peninsula and Sicily (Robert Mantran in Braudel, 1985). Between the 11th and the 15th century, with a territorial disgregation of the power and the rise of local dynasties in North Africa and the Middle East, the Empire redefined its presence in the Mediterranean space (Robert Mantran in Braudel, 1985). The impact this civilization had on the Mediterranean populations was huge; the ethno-linguistic mark they left on the southern and western regions of the Mediterranean basin is unquestionable (Sazzini M., Sarno S., 2014).

What transpires from the panorama just depicted, is that the Mediterranean, as Braudel himself used to claim, is almost like an experimental field; 'there is always', he says, 'a zone of the Mediterranean that employs some sort of dominance on the other ones, somehow feeding on them' (Braudel, 1985). In any case, when it comes to look at what might have affected the patterns of genetic variation in today Mediterranean populations, it does not matter if we are talking about processes of invasion or colonization; they both imply the same event: movements of people and migratory flows.

Migration is central to the history of the Mediterranean, even the one occurred in the last two centuries, together with the most recent processes of worldwide globalization, might be factors currently affecting the Mediterranean populations' genetic variation, possibly leading to the loss of the traces that our ancestors left in our genomes (Sazzini M., Sarno S., 2014), somehow changing the image of the Mediterranean and the peoples that inhabit it.

DID YOU SAY ISLANDS?

In genetics, isolated populations are those groups of individuals that have experienced little gene flow with surrounding populations (Jobling *et al.*, 2014). Normally, they are subpopulations originated from a small number of individuals, possibly isolated as result of bottlenecks and/or founding events (famine, war, social and/or cultural barriers, environmental changes, epidemic diseases,

settlement in a new area) and then closed in their isolation for many generations (Hatzikotoulas *et al.*, 2014). Based on this explanation, the term *island* must not be confused with the proper geographical islands, since genetic isolation can occur in very different environments and for a variety of different reasons. Thus, *island* does not necessarily mean (genetic) isolate, as well (genetic) isolate does not necessarily mean island.

According to the French archaeologist Jean Guilaine, islands are both places where the time moves slowly or faster than any other not isolated reality. For the first case he refers, for example, to the Nuragic civilization developed in Sardinia during the Bronze Age with the megalithic construction represented by the Giants' tomb, or to the Navetas monumental tombs erected on Menorca in the Balearic Islands. On the other hand, he claims that the appearance of a pioneering proto-urban center in the northeastern Aegean Sea, on Poliochini and Thermi, and on the island of Phylakopi in the northern coast of the island of Milos in the Cyclades, anticipated the rise of urban centers before they appeared in the Western Mediterranean basin (Braudel, 1985).

When it comes to genetics, things are not so different. As consequence of isolation, isolates preserve a genetic uniqueness that is not recorded in those not isolated populations in which a greater genetic similarity comes with the exchange of migrants; somehow, it is like the absence of new incoming material from the outside stopped the evolution, or at least slowed it down (Cavalli-Sforza,

1996). On the other hand, the reduced genetic complexity of isolates, the small size of the founder population, and the effects of the drift, can rapidly increase the appearance of rare variants more than in any other not isolated group (Bittles, 2005; Hatzikotoulas *et al.*, 2014; Kurki *et al.*, 2019).

Both bottlenecks and founding events are size-reduction mechanisms defined by the lowering of the amount of genetic variation. In case of bottlenecks, the size reduction of a previously larger population leads to a random loss of the previous variation; in the case of founding events, a subgroup of the source population splits off to establish a colony in which the diversity will be a random subset of the original population (Figure 5a) (Jobling *et al.*, 2014). Over time, apart from the selection-driven changes, the variation of the allele frequencies over generations can increase or decrease by chance; this fluctuation is better known as random genetic drift. This evolutionary mechanism affects populations of all sizes, but its effects are stronger in small populations. In isolates, the limited number of individuals allows a more rapid fluctuation of the allele frequencies, with the following fixation or loss of specific alleles, than in larger populations (Figure 5b). Compared to the neighboring populations, the unusual allele frequencies that isolates show can be interpreted as sign of geographical, linguistic, or cultural isolation (Jobling *et al.*, 2014). Also, it is quite common to observe in these communities the development of endogamous activity that, together with the low levels of incoming external mates from the surrounding populations, allow the increase of a

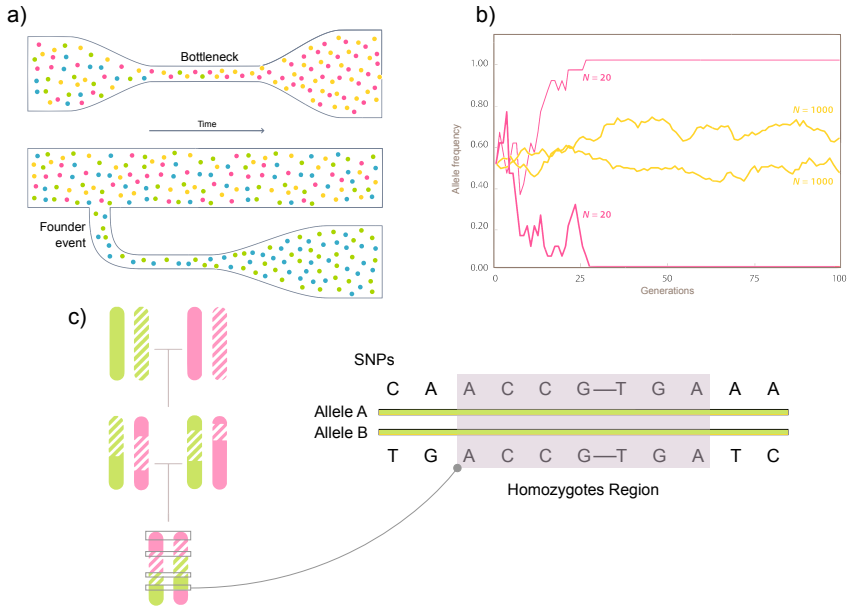


Figure 5: a) Loss of diversity through bottleneck and founder effect (coloured dots are different alleles) (modified from Jobling et al., 2014); **b)** Allele frequency evolution of a binary polymorphism through 100 simulated generations: allele frequency starts at 0.5 in both N=20 and N=1000 sized populations (modified from Jobling et al., 2014); **c)** Extreme case of inbreeding between two siblings highlights the origin of several ROH regions (identical chunks in both chromosomes enclosed by boxes):a zoom into one region evidences the equal sharing of single nucleotide variants in both the alleles.

wide genetic internal homogeneity; this is also reflected by a reduced effective population size (N_e) (Hatzikotoulas *et al.*, 2014). The latter is used to define the amount of genetic drift a population experienced; the smaller N_e, the greater the drift (Jobling *et al.*, 2014). Commonly, the effective population size is based on the loss of genetic diversity through genetic drift or through inbreeding (Husemann *et al.*, 2016). The latter considers the identity by descent (IBD) concept, that is the probability that two alleles within the

same individual descent from the same ancestor (Jobling *et al.*, 2014). The small effective population size that population isolates usually show, increases the levels of homozygosity and those of linkage disequilibrium (LD) (Zeggini, 2014). Higher levels of homozygosity are reflected in the length of runs of homozygosity (ROHs), continuous regions of the genome identically inherited from both parents (Figure 5c). ROHs are widely used in population genetics to dig into the level of relatedness between different individuals; their length is commonly interpreted as function of the nature of this relatedness: longer ROHs are symptomatic of recent inbreeding because of the lack of recombination in the short term, while shorter ROHs are usually linked to a much older story (Kirin *et al.*, 2010). Even the increased levels of LD reflect a similar pattern; because of ancestral recombination events, LD stretches tend to be shorter in length in older isolates compared to those of more recent isolated groups in which LD will span over longer genomic regions (Heutink *et al.*, 2003).

Studies on genetically human isolates are at the base of very different fields. In anthropology, genetic isolation has been often associated to cultural diversity (Capocasa *et al.*, 2014); in medical genetics, isolates have been widely used for mapping genes for rare monogenic disorders as also for complex diseases (Heutink *et al.*, 2003), while in population genetics they are fundamental to understand the genetic structure of human populations, unveiling those factors that account for genome variation (Robledo *et al.*, 2012).

Today, marriages within the same group persist as a rule in some traditional societies. Endogamous marriages are an essential feature of tribal, clan, or caste systems for different cultures (Bittles, 2005). Also, it seems quite common to observe consanguineous marriages especially in religious communities, where the taboo of intra-family marriages does not subsist and where first- and second-cousin marriages are the most usually observed in different cases (Klat *et al.*, 1986). Consanguineous marriages are widespread in different parts of North Africa, Middle East, South and Southeast Asia, where consanguineous and non-consanguineous marriages co-exist and the choice of a partner can occur between neighboring families, or inside the same family (Denic *et al.*, 2011). In modern western not isolated societies, the choice of a partner is mostly personal and no ethnic or community background seems to matter (Bittles, 2005). However, in pre-industrial societies the interchange of genetic material between communities was limited by geographical boundaries, and by a strong positive assortative mating. This means that the choice of a partner was limited not only to the geographical space, but also depended on shared phenotypic characteristics, socioeconomic status, religious beliefs, education, and also political parties. This means that, before moving away from an intra-community mating, we all had roots as isolated communities (Bittles, 2005).

However, finding human isolates in Western countries is not so uncommon. One of the most studied examples is represented in the Western Mediterranean by the case of Sardinia. Since prehistoric

times the island was peopled by many groups and signs of a long history of isolation have been detected in modern inhabitants (Di Gaetano *et al.*, 2014). Its distance from the continental neighbors is a strong sign of differentiation (Grimaldi *et al.*, 2001), and its large internal homogeneity seems to point to a high level of isolation (Di Gaetano *et al.*, 2014). Based on single nucleotide polymorphisms (SNPs), the analysis of ROHs pointed to an ancestral small effective population size; compared to the Italian ones, more extended haplotypes for the shortest ROHs category (0.5-1 Mb and 1-2 Mb) recall the configuration usually detected for isolates and small communities (Di Gaetano *et al.*, 2014). Furthermore, studies on ancient DNA have detected high levels of genetic affinity to ancient Neolithic farming peoples of Europe (EF) (Keller *et al.*, 2012; Skoglund *et al.*, 2012; Lazaridis *et al.*, 2014), who colonized the Island during the Neolithic before remaining isolated from subsequent migrations (Sikora *et al.*, 2014).

A further example of genetically isolated human groups is represented by modern-day Basques whose unique culture and language (*Euskera*) put them at least among culturally and linguistically isolated groups; according to different studies, the region they occupy represented one of the most densely populated European glacial refugia during the Last Glacial Maximum (Flores-Bello et al., 2018). The comparison with early Iberian farmers showed their greatest affinity to modern-day Basques, possibly linking their peculiar origin with the spread of agriculture during the Neolithic; the higher hunter-gatherer ancestry these farmers own

respect to the European early farmers separate them from the Neolithic farmer component detected in Sardinia and the rest of the European groups (Günther *et al.*, 2015).

According to a pairwise F_{st} analysis, both Sardinians and Basques show the highest values of separation from major European groups (Rodríguez-Ezpeleta *et al.*, 2010), enforcing their isolated condition in the wider panorama of the Mediterranean and European countries.

At north of Sardinia, also the French island of Corsica seems to find a place among the isolated groups in the Western Mediterranean basin. Its distance from continental Europe could be at the base of its long-term isolation (Tofanelli *et al.*, 2004). A common genetic founding pool links the Corsicans to the Sardinians; their coasts were invaded several times by the same populations and similar evolutionary forces left marks on their gene pools (isolation, consanguinity, and bottlenecks) (Vona *et al.*, 2003; Latini *et al.*, 2004). A general heterogeneity between Corsica and Sardinia is also supported by the presence of rare ß-globin cluster haplotypes, possibly highlighting a common origin dating back to the Paleolithic that has been interpreted as a founder effect (Latini *et al.*, 2003).

Having a global vision on the vast group of isolated populations in the countries of the Western Mediterranean area is anything but an easy task. However, some valuable mentions are due. A fascinating

example is undoubtedly represented by the Chuetas, descendants of Majorcan Jews, a clear example of cultural isolation characterized by a strong endogamy and whose relation to other Jewish populations (Polish, Askhenazi, and Sephardic Jewish) has been detected, even if signals of admixture with modern Majorcans seems to point to a more recent change in the behavioral patterns, possibly due to their conversion to Christianity in the 15th century, that allowed marriages with the local population (Picornell *et al.*, 1997). A further example is represented by the Arbereshe, linguistic isolated groups of Sicily and Southern Italy (Calabria) migrated from Albania between the 15th and 16th centuries. While the groups from Calabria displayed an evident cultural and genetic continuity with the source populations, Sicilian groups showed a differential pattern, with the Y chromosome results pointing to a continuity with incoming Greeks, while mtDNA linked the Sicilian Arbereshe community to the populations from Southern Italy, thus defining a lower impact of isolation (Capocasa *et al.*, 2014; Sarno *et al.*, 2015).

Without any doubt, the contribution of genetic isolates to the understanding of the genetic structure of human populations is undeniable, as well the relations between cultural factors and genetic variation that comes with them.

Immanuel Kant states that the domain of reason is an island, "the land of truth", circumscribed by "a wide and stormy ocean", where illusion and deception come from (Kant, 1781); to see the truth, maybe it is worth to stay on that island.

THE GENETICS OF THE WESTERN MEDITERRANEAN

Both genetic and genomic markers have been largely used in population genetics in order to shed light on the human evolutionary history. It was about forty years ago that the genetic studies of population structure and history begun with the study of the blood group markers (Menozzi *et al.*, 1978; Henn *et al.*, 2010). Later, uniparental markers such as mtDNA and the non-recombining region of Y chromosome have largely helped to unveil the differential migration patterns for both male and female lineages (Henn *et al.*, 2010). Y chromosomal variation proved to be useful for studies of different populations and mtDNA studies demonstrated that the geographical migration of women has been higher than in men (Seielstad *et al.*, 1998). Furthermore, the increase of genome-wide marker data, especially SNPs, also added valuable information to the knowledge of several populations from both micro- and macro-structural points of view (Jakkula *et al.*, 2008; Jakobsson *et al.*, 2008; Li *et al.*, 2008). However, when it comes to the Western Mediterranean basin, different information has been provided about most of the populations of this area, but a global vision is still missing. Here, we are going to revise the state of the art of the four main groups of the Western Mediterranean area, mostly focusing on Spain and France. Also, even if they are not directly part of this work, we are going to look to some of the information collected so far about Italy and North Africa.

SPAIN

With its position in the Mediterranean basin, Spain has been part of many historical events, coming into contact with different cultures from both the Mediterranean and the Atlantic sides, including European and African populations. Many studies are devoted to this population because of its pivotal role during the last glacial maximum, with the Franco-Cantabrian region representing one of the southern refugia, but also for the mark left by the Umayyad conquest during their long permanence in the 8th century. So far, the outcomes reached through different methods tried to explain the complex history of this country with some striking results.

According to first studies with classical genetic markers (blood groups, proteins, and enzymes), a clear distinction between the Basque region and the rest of Spain was already detected; this result was interpreted as the outcome of a long isolation occurred during the Paleolithic and Mesolithic times that had possibly amplified the separation from the incoming Neolithic farmers (Aguirre *et al.*, 1991; Bertranpetit *et al.*, 1991). Furthermore, a divergence of Catalonia from the rest of the Peninsula was detected and explained as a signal of a first Neolithic wave possibly through the Pyrenees or along the coast. A third observation pointed out a divergence between the Atlantic and the Mediterranean parts, a duality maybe linked to the two different Celtic and Iberian cultures (Bertranpetit *et al.*, 1991). Regarding the Balearic Islands, classical markers have highlighted a higher similarity between Majorca and Menorca,

defining a genetic separation for the island of Ibiza. This outcome was explained as a possible genetic continuity with North African and East Mediterranean populations, and a Carthaginian origin has been proposed due to the long permanence of this culture on Ibiza (Picornell *et al.*, 1996).

The distinctiveness of the Basques from the rest of the Peninsula was also defined through first mtDNA studies, pointing to its isolated condition not only from Spain, but also from the rest of the European groups (Côrte-Real *et al.*, 1996). Moreover, a common origin for all the Spanish groups was proposed and dated back to the Upper Paleolithic (Côrte-Real *et al.*, 1996; Richards *et al.*, 1996), while a further differentiation of Catalonia and Andalusia was also pointed out (Côrte-Real *et al.*, 1996). The majority of the haplotypes detected defined a homogeneous landscape with the rest of the European groups, with the exception of some lineages possibly with a North African root (Côrte-Real *et al.*, 1996).

More advanced studies suggested a stratification for the mtDNA variation in Spain (Barral-Arca *et al.*, 2016). Haplogroup R showed a prevalence in the northern part of Spain; nested in this haplogroup, HV0 displayed a higher frequency in the Basque country, while haplogroup H (also nested in haplogroup R), showed a higher frequency on the Atlantic side of the country, with a decreasing pattern moving to the Mediterranean side, spanning the Cantabrian coast, and reaching higher peaks in Galicia (Barral-Arca *et al.*, 2016). Both HV0 and H fall into the genetic Mesolithic

substratum of Southwestern Europe (Brandt *et al.*, 2015). For haplogroup H a Middle Eastern origin around 30,000-25,000 YBP was proposed, followed by a Paleolithic expansion into Europe, and a central role in the process of re-expansion from the ice-age refugia during the LGM (Achilli *et al.*, 2004). Thus, its presence in the Franco-Cantabrian region, together with HV0, can be seen as a proof for this region to have played a pivotal role during the LGM and the following re-expansion (Barral-Arca *et al.*, 2016), while the higher frequency of haplogroup H in Galicia agree with the hypothesis for this region being an European edge for an ancient central European migration (Salas *et al.*, 1998; Barral-Arca *et al.*, 2016). Among other Mesolithic/Neolithic lineages, haplogroup J showed a higher frequency in the northwestern corner of the Iberian Peninsula and in the Basque country, while haplogroup T was found with a higher prevalence in the Mediterranean area, included the Balearic Islands with a major representation on Ibiza (Picornell *et al.*, 2005; Barral-Arca *et al.*, 2016). In the southern part of the Peninsula, haplogroup L showed a higher distribution, possibly representing an introduction of more recent demographic events such as the Arab conquest of the Iberian Peninsula (Barral-Arca *et al.*, 2016); also, the presence of haplogroup U6 at higher frequencies in southern Andalusia is probably due to the Muslim expansion from North Africa (Hernández *et al.*, 2014). Among the Balearic Islands, Majorca and Menorca showed a high haplotypic diversity, similar to the one detected in other European populations (Picornell *et al.*, 2005; Simón *et al.*, 2017). This finding seems to agree with the hypothesis of a more active gene flow between

Majorca and Menorca with the mainland populations, in comparison to Ibiza whose diversity was found be the lowest (Picornell *et al.*, 2005).

Looking at the male heritage, studies on the Y chromosome pointed out an overall reduced genetic structure, even for the Balearic Islands where a high similarity with the surrounding populations was detected (Flores *et al.*, 2004; Tomàs *et al.*, 2006). Also, the overall Y-chromosome pattern was probably already present during the Paleolithic, leading to the absence of a Neolithic ancestry in the nowadays Spanish male lineages (Flores *et al.*, 2004). Just like for the rest of the Western European countries, the most frequent haplogroup in Spain is R1b, which represents the persistence of Paleolithic Y chromosomes in Europe after the Neolithic expansion from the Near East (Myres *et al.*, 2011). However, the sublineage DF27 of the R1b haplogroup, with a frequency lower than 20% outside of the Iberian Peninsula, seems to be characteristic of the 40-48% of the men in Spain with a 70% of frequency in the Basque population, and a possible origin in the early Bronze Age (Valverde *et al.*, 2016; Solé-Morata *et al.*, 2017; Villaescusa *et al.*, 2017). Also, recent history seems to have affected the modern pattern of variation of the Spanish Y-chromosomal lineages; means of 10.6% and 19.8% have been associated to a North African and a Sephardic Jewish ancestry, respectively. The African lineage E3b2 showed a higher frequency in the western half of the peninsula (Galicia and Northwest Castile), and a lower one in the eastern part of it; this configuration seems to agree with the historical relocations and

expulsion of *moriscos* (Adams *et al.*, 2008). Furthermore, proportions of G, K*, and J lineages, detected at lower frequencies among the different groups of Spain, have been ascribed to eastern Mediterranean populations such as Greeks and Phoenicians (Adams *et al.*, 2008; Zalloua *et al.*, 2008). Additionally, just as for the mtDNA, even a North Western African contribution was detected for the Iberian Y-chromosome pool with the highest percentage of contribution (14%) found in Andalusia (Bosch *et al.*, 2001). This is also supported by a genome-wide study in which Iberia was found to be the European region with the highest IBD sharing with North African populations (Botigue *et al.*, 2013).

When it comes to genome-wide studies, the internal structure of Spain, has been disentangled in a very recent research, in which an internal pattern of variation was described, providing evidence of historical movements for both the Muslim conquest and the following Reconquista; a general east-west pattern of genetic differentiation has been detected, as well as a higher internal similarity along the north-south axis, interpreted as evidence of the historical background of population movements (Figure 6) (Bycroft *et al.*, 2019). This two directional structures have been ascribed to the same period: the north-south gene flow might be the result of the *Reconquista*, when people from the northern Christian kingdoms expanded southwards fighting back the Muslim domain, while the east-west differentiation, highly evident in the northern part of the peninsula, might correspond to the geopolitical and linguistic borders that were defined around the end of the time of

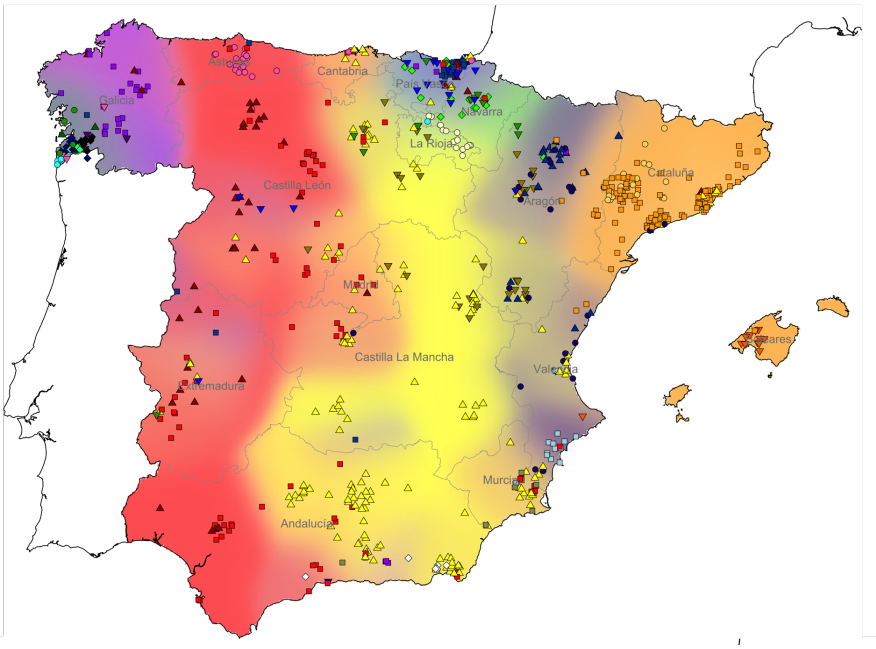


Figure 6: Spatial density distribution for the clusters detected with fineSTRUCTURE. North-south direction is the outcome for the gene flow resulting from the *Reconquista*, while the east-west gradient might point to the northern geopolitical and linguistic structure by the end of the Muslim settlement (Bycroft et al., 2019).

Muslim rule in Iberia and dated from ~930 to 1300 CE (Bycroft *et al.*, 2019). Furthermore, as evidence for the mark left by the Muslim rule, a varying fraction of north-west African ancestry has been detected in modern-day Iberians with a regional variation of 0-11%. This admixture event has been dated to 860–1120 CE. Both Basques and people from Galicia are included in this admixture event, even if they have never been under Muslim rule; a fact that is explained by possible internal migratory flows occurred in more recent times (Bycroft *et al.*, 2019).

FRANCE

The current French population is the result of multiple migrations since Paleolithic times when anatomically modern humans reached this part of Europe. The Franco-Cantabrian region in the southwestern part of the country represented a shelter for our ancestors during the LGM, before the postglacial re-colonization begun. Two different waves of Neolithization interested this area; eastern France was reached by the Linienbandkeramik culture from central Europe, while the Mediterranean side was the target for the Impressed-Cardial Ware culture from Dalmatia and Italy (Prevost *et al.*, 1984; Richard *et al.*, 2007).

In the context of population genetics, France has been poorly analyzed as a whole, especially in very recent times. In my opinion, the mostly recent and unstable internal geographical organization of the country made extremely difficult the comparison between different studies, and a common agreement must be reached on how the samples should be geographically arranged.

First studies with classical markers (AB0, Rh, MHC, serum proteins) found a general heterogeneous pattern when considering different geographical structures (military districts, historical provinces, and regions) (Kherumian *et al.*, 1967; Cambon-Thomsen *et al.*, 1988). This heterogeneity was also suggested by Cavalli-Sforza's synthetic maps; their interpretation defined differences between northern and southern Neolithic influences, as also some

internal substructures, especially for Brittany and the Franco-Cantabrian region (Figure 7) (Cavalli-Sforza, Menozzi, & Piazza, 1994).

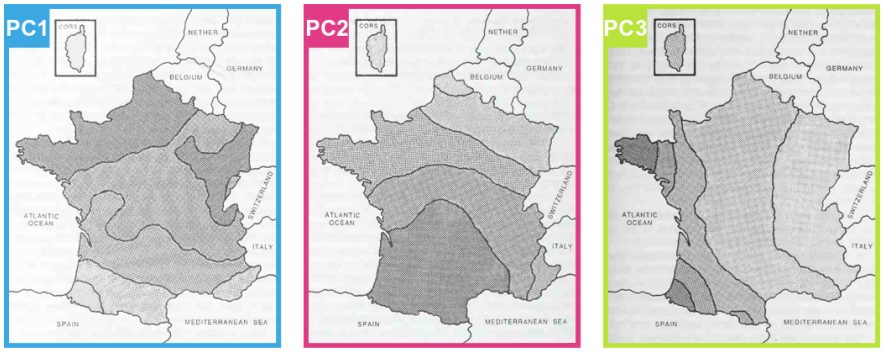


Figure 7: Cavalli-Sforza's synthetic maps interpretation for the first three principal components based on data from Cambon-Thomsen & Ohayon, 1988. In these maps, the authors suggested some pattern of differentiation. In the first principal component a difference between northern and southern European Neolithic people is mainly stressed, while the second principal component mostly highlights the differences between the Upper Paleolithic influence in the southwest and the Neolithic one in both northern and southern parts of France. Finally, in the third principal component, a migratory east-west gradient is detected, with Brittany and the Basque region showing a higher differentiation pattern (modified from L.L. Cavalli-Sforza, Menozzi, & Piazza, 1994).

The analysis of the maternal lineages through mtDNA, redefined the genetic landscape of France, highlighting a general homogeneity on both a regional scale (based on 1982 Metropolitan France) and historic provinces (Dubut *et al.*, 2004; Richard *et al.*, 2007). Apart from some differentiation on a microgeographic scale for both Brittany (that showed affinity with Britain and Scandinavian) and the French Basques (with high frequency of haplogroup H linked

with the Neolithic diffusion in Europe), the mtDNA haplogroup composition of French people did not differentiate neither internally, nor from the surrounding European genetic landscape (Dubut *et al.*, 2004; Richard *et al.*, 2007).

From the paternal point of view, Y-chromosome mostly agreed with the results from the maternal side. On a base of regional distribution (1982 Metropolitan France), the general panorama is a lack of differentiation between the different groups, with the only exception of Brittany, for which a lower Y-chromosome diversity was found, explained by a possible founder effect plus an isolation process (Ramos-Luis *et al.*, 2009).

The story is not so different from the autosomal point of view. A genome-wide study based on regional geographic distribution (1982 Metropolitan France) did not detect any differentiation between the different groups in the study, with the only exception for Western France represented by samples from Brittany. The higher linkage disequilibrium detected for this group suggested a lower effective population size, thus corroborating the hypothesis of isolation inferred by the outcomes of the Y-chromosome analyses (Karakachoff *et al.*, 2015). A comparison with surrounding populations, also supported the admixture between Bretons and individuals from the British isles (Karakachoff *et al.*, 2015).

ITALY

Without any doubt the central position in the Mediterranean basin has been one of the factors that made the Italian genetic landscape truly peculiar. The peopling of Italy dates back to Paleolithic times; the migrations during the Neolithic period, together with the most recent history of relations with the surrounding countries have contributed to its modern genetic structure (Boattini *et al.*, 2013). So far, autosomal genome-wide studies (both allele frequency and haplotype-based ones) have proved the presence of structured four macro-areas across the peninsula (Figure 8), clearly separating Northern, Central, and Southern Italians from the most differentiated Sardinian cluster (Di Gaetano *et al.*, 2012; Fiorito *et al.*, 2016).

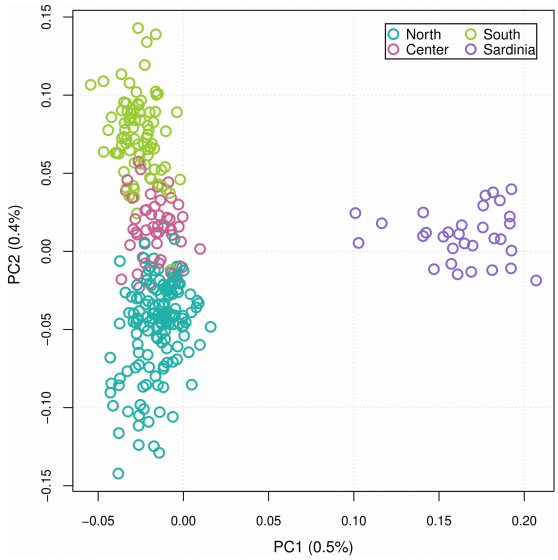


Figure 8: Principal component analysis showing the four Italian macro-areas (plot reproduced with genome-wide data from Di Gaetano et al., 2012; Fiorito et al., 2016)

When the relation with the surrounding populations is considered, differential contributions are detected supporting the presence of a continuous gene flow mostly linking Northern Italians to the European countries, and Southern Italians to the Middle Eastern and North African ones (Fiorito *et al.*, 2016).

The analysis of the paternal lineages showed a discontinuity in the Y chromosome distribution with three groups separating the Sardinian one from the North-western and South-Eastern Italian ones (Boattini *et al.*, 2013; Sarno *et al.*, 2015); the haplogroup distribution defined a south-north decreasing pattern of variation with R-U152* showing a frequency of 12.1%, followed by G-P15 (11.1%), E-V13 (7.8%), and J-M410* (7.6%) (Boattini *et al.*, 2013).

Furthermore, in agreement with the analysis of the autosomal variants, the Southern group showed a higher similarity with Middle Eastern, Southern Balkan, and Anatolian groups, while the Northern Italian group seems to be closer to the North-Western European and Northern Balkan populations (Boattini *et al.*, 2013; Grugni *et al.*, 2018). This findings for the Italian male lineages seem to agree with different patterns of migration during the Neolithic revolution (Boattini *et al.*, 2013; Sarno *et al.*, 2014). On the other hand, the analyses on the matrilineal lineages with mtDNA highlighted a more homogeneous genetic landscape (Boattini *et al.*, 2013; Sarno *et al.*, 2014); even when compared to external groups, any geographic pattern was not detected for the mitochondrial DNA distribution, reflecting the usual variability of Western Europe

(Boattini *et al.*, 2013). With a 13.8% of the global variability, H1 represented the major haplogroup in the peninsula, followed by H3 (3.9%), H5 (4.3%) and, with a lower frequency, U5, K1, J1, J2, T1, T2, and HV, also common in western Eurasia (Boattini *et al.*, 2013). The age estimation for the mtDNA pointed to pre-Neolithic times, supporting the hypothesis of the presence of an Italian Glacial Refugium during the LGM (Boattini *et al.*, 2013), possibly also in the South of the peninsula according to the distribution of the HV4 lineage that, together with the U5b one, has been indicated as a characteristic Italian pre-Neolithic lineage (De Fanti *et al.*, 2015).

NORTH AFRICA

Limited between two natural barriers (the Sahara Desert to the south and the Mediterranean Sea to the North), North Africa has been the focus for different studies with the aim of understanding human history (Van De Loosdrecht *et al.*, 2018). However, behind its complex demographic history lay also a series of different connections with the surrounding populations from sub-Saharan Africa, Europe, and the Middle East (Henn *et al.*, 2012; Arauna *et al.*, 2017; Font-Porterias *et al.*, 2018). A genome-wide study on autosomal markers described three different events involved in the definition of the present-day ancestry of North African populations: an ancient back-to-Africa gene flow represented by an east-to-west increasing gradient possibly related to an autochthonous Maghrebi ancestry, an east-to-west decreasing gradient representing a more recent Near Eastern Arabic ancestry, and a very recent gene flow

from sub-Saharan Africa estimated approximately 1,200 YBP in southern Morocco and about 750 years ago into Egypt (Figure 9) (Henn *et al.*, 2012). A further study based on haplotype methods found that the autochthonous genetic component of North African populations is showing a heterogeneous admixture pattern with other populations (Middle Easterners, sub-Saharan Africans, and also Europeans) (Arauna *et al.*, 2017).

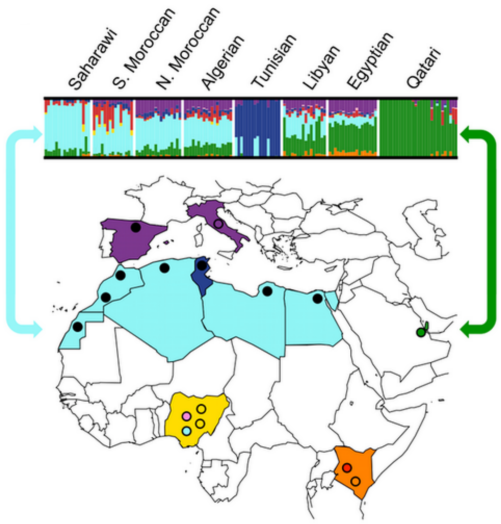


Figure 9: Samples distribution and ADMIXTURE analysis for North African populations and the surrounding countries. In the bar plot the different colours represent ancestral components and their amount is the frequency reached in specific each group. A decreasing pattern of North African ancestry is defined along a west-east gradient (blue colour on map and bar plot). (Henn *et al.*, 2012).

Varying amounts of external gene flow, rather than isolation, seem to be the cause for heterogeneity among Berber groups (Arauna *et al.*, 2017). The analysis of the paternal North African structure identified E-M81 as the most frequent Y-chromosome haplogroup,

possibly with a Middle Eastern root (Solé-Morata, García-Fernández, *et al.*, 2017). A Paleolithic origin for this lineage was first proposed (Bosch *et al.*, 2001), followed by a Neolithic one (Arredi *et al.*, 2004), and finally a new temporal definition placed its root around 2000 YBP (Solé-Morata, García-Fernández, *et al.*, 2017).

Studies on the matrilineal lineages have highlighted that within North Africa the genetic structure is the result of different haplogroup frequency distribution of U6, L and H lineages (Fadhlaoui-Zid *et al.*, 2011). Haplogroup U6 has been inferred as a non-African lineage with western Eurasian origin representing a back-migration to Africa during the Early Upper Paleolithic, and that occurs most frequently in the west of North Africa (Pereira *et al.*, 2010; Fadhlaoui-Zid *et al.*, 2011; Secher *et al.*, 2014). A prehistoric gene flow into some Mediterranean populations has been also hypothesized because of the presence of U6 haplogroups in the Iberian Peninsula (Maca-Meyer *et al.*, 2003). The introduction of L haplogroups in North Africa was proposed to have occurred through gene flow from eastern sub-Saharan populations dating back to 20,000 YBP (Frigi *et al.*, 2010). With its Eurasian origin, haplogroup H represents instead a post LGM expansion (Cherni *et al.*, 2009).

METHODS IN GENOME-WIDE STUDIES

The study of human genetics is undoubtedly driven by the desire to understand the extent of genetic diversity among populations and the will to reconstruct their history over time. Mostly, at the base of population genetics studies, the analysis of the variation in allele frequencies of populations allows to reach these goals. Studies of populations structure are interested in the genetic differences in a group of subjects and the genetic mechanisms that are responsible for the populations to be differentiated.

As the Human Genome Project pointed out, on average, 99.9% of the genome is shared across all humans. Our differences lie in the remaining 0.1% which is represented by variants, commonly at single nucleotides. When more than 1% of a population carry the same variant, this will be called a single nucleotide polymorphism (SNP), genome positions exhibiting two distinct alleles. The proportion of each allele varies within a population and among different populations, and it is on the analysis of this variation that the study of population genetics works (Gardiner, 2001; Laframboise, 2009).

Today, the microarray SNP technology allows genotyping a large number of individuals, making it relatively easy, fast, and cheap to retrieve data for group of subjects within the same population (Laframboise, 2009). Behind the microarray technology lie the simple biochemical principle of complementary binding between

nucleotide bases (A binds to T, and C binds to G). Usually, an array contains hundreds of thousands of unique nucleotide probe sequences designed to recognize and bind to a specific target DNA sequence (Laframboise, 2009). After hybridization, a different signal intensity will be recorded for each probe/target pair (Figure 10); the following SNP calling process allows to infer SNP genotype for every subject at each locus. The resulting output is usually exported to binary file formats (PLINK format) or Variant Call Format (vcf files) and, after applying filters for quality control and removing possible genotyping errors, used for different analyses.

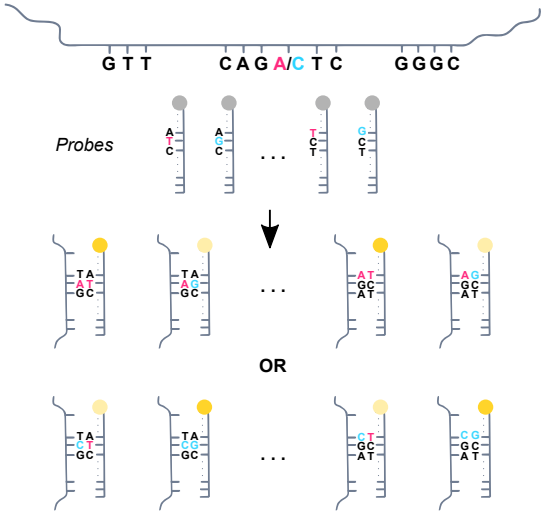


Figure 10: In the example, the SNP A/C is interrogated by the designed probes. The DNA fragments bind to the different probes with a resulting different efficiency level and a distinct intensity of the signal produced by the binding. A perfect bind will result in a more intense signal (darker yellow), while a mismatch of the SNP site will produce a lower signal (lighter yellow) (modified from Laframboise, 2009).

In population genetics, different analytical methods have been widely used for the inference of population structure. The rapid evolution of such methods and their always more precise and specific capability in detecting patterns of population structure allowed to better understand human evolutionary history, migration rates, and times of population splitting (Schraiber *et al.*, 2015).

Standard Principal Component Analysis (PCA) is a method based on independent markers that allows the qualitative description of population structure, drift and admixture of populations (Cavalli-Sforza *et al.*, 1994; Schraiber *et al.*, 2015). To support the information retrieved with PCA, several assignment methods have been developed with the aim to infer clusters of populations defined by the fact that a specific group of subjects form a distinct group based on some genetic criteria, such as shared allele frequencies (Guillot *et al.*, 2009; Schraiber *et al.*, 2015). Tools like STRUCTURE (Pritchard *et al.*, 2000) (Bayesian Estimation based), ADMIXTURE (Alexander *et al.*, 2009), and FRAPPE (Tang *et al.*, 2005) (both Maximum-Likelihood Estimation based) are the most used so far. Nonetheless, to formally test for admixture between populations, the *f-statistics* proved to be convenient tools for evaluating phylogenetic relationship, through the sharing of allele frequencies between populations along tree branches, thus testing the existence of admixture events (Peter, 2016; Schaefer *et al.*, 2016; Wangkumhang *et al.*, 2018).

However, working with unlinked markers can drastically reduce the amount of information that can be retained. Thus, a series of analyses are based on haplotype sharing between samples, and different tools have been developed in order to consider the linked SNPs and linkage information. This means that haplotype-based methods provide a finer structure information than standard PCA, allowing to define more precisely the relation between the different subjects. The haplotype information for different individuals can be summarised through a method implemented in ChromoPainter, a software capable of constructing a representation of the relationships of the subjects in a study. This representation is called "*coancestry matrix*", which is built under a Hidden Markov Model, and represents the expected number of genomic chunks shared between each individual, defining the extent of their relationship (Lawson *et al.*, 2012). Based on ChromoPainter results, the software FineSTRUCTURE can arrange the samples into groups using a Markov chain Monte Carlo algorithm (MCMC) defining populations that are not known a priori (Lawson *et al.*, 2012; Schraiber *et al.*, 2015).

Haplotype information and LD can also be used for inferring population size changes, including recent growth. A non-parametric method is implemented in IBDNe software which estimates the effective population size for the TMRCA (Time to Most Recent Common Ancestor) using inferred long segments of IBD (Gao *et al.*, 2016; Browning *et al.*, 2018). Since this method is IBD-based, and longer segments describe latest events, this approach is

specifically useful for inferring very recent history together with the population size changes at each generation (Gao *et al.*, 2016).

Most of the methods described above can be applied even to ancient samples. When working with ancient DNA (aDNA), damage, contamination, and length of the molecules can highly affect the coverage, and thus the genotype calling (Schaefer *et al.*, 2017). Post-mortem DNA fragmentation is mostly due to depurination processes that affect the length of the molecules available for sequencing. Deamination of cytosine to uracil causes C to T substitutions, leading to an increased level of damage. Lastly, contamination of ancient DNA samples is normally linked to microbial DNA from the environment and/or modern human DNA, usually from the researchers themselves (Orlando *et al.*, 2015). All these processes have an impact on the quality of the DNA that needs to be carefully prepared before the mapping, using tools such as AdapterRemoval (Lindgreen, 2012) for trimming the adapters from the sequences and Samtools (Li *et al.*, 2009) for removing duplicates, and after it, by computing the deamination rate using dedicated tools such as MapDamage (Jónsson *et al.*, 2013). The reliability of the data produced is at the base of a good merge with modern samples; thus, the ancient samples can be projected onto the space of a PCA, used for ADMIXTURE analyses, *f-statistics*, or even for haplotype-based studies. The valuable contribution ancient samples gave to the study of population genetics is undeniable; they certainly increased the information gained today about our past.

RESULTS

PEOPLE FROM IBIZA: AN UNEXPECTED ISOLATE IN THE WESTERN MEDITERRANEAN

Simone Andrea Biagini, Neus Solé-Morata, Elizabeth Matisoo-Smith, Pierre Zalloua, David Comas, Francesc Calafell

European Journal of Human Genetics, 2019.

Biagini SA, Solé-Morata N, Matisoo-Smith E, Zalloua P, Comas D, Calafell F. People from Ibiza: an unexpected isolate in the Western Mediterranean. Eur J Hum Genet. 2019 Jun 1;27(6):941–51. DOI: 10.1038/s41431-019-0361-1

DISENTANGLING THE *HEXAGONE*: THE GENETIC LANDSCAPE OF MODERN FRANCE

Simone Andrea Biagini, David Comas, Francesc Calafell

In preparation

Disentangling the *Hexagone*: the genetic landscape of modern France

Simone Andrea Biagini1, David Comas1, Francesc Calafell1

1. Departament de Ciències Experimentals i de la Salut, Institute of Evolutionary Biology (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Spain.

Abstract

Unlike other European countries, Metropolitan France is surprisingly understudied. In this work, we combined newly genotyped samples from various zones in France with publicly available data and applied both allele frequency and haplotype-based methods in order to describe the internal structure of this country. We found out that French Basques are genetically distinct from all other populations in the *Hexagone* and that the populations from southwest France (namely the Franco-Cantabrian region) are intermediate between Basques ad other populations. Moreover, Bretons slightly separated from the rest of the groups and a link with the historical gene flow from the British Isles has been found. The general background we describe appears to be a mixture of two components, one closer to Southern Italy and the other to Ireland. This combination may be the result of a contact happened in two different moments: in the Early Neolithic, and then Ireland would be a proxy for the continental pathway for the Neolithic wave of

advance and South Italy for the coastal penetration, or the Iron Age, when the Celtic and the Mediterranean worlds met in France.

Introduction

Located in the center of Western Europe, Metropolitan France acts as a bridge connecting Northern Europe to the Mediterranean and the Iberian spaces. Nowadays, France is a cosmopolitan country whose society is shaped by a plurality of life styles and truly different ethno-cultural diversity. Without any doubt, the impact immigrations from colonised countries to mainland France, such as the migration of Arabs and Berbers from Algeria which was the most extensive of all colonial migrations to Western Europe before the 1960s (MacMaster, 1997), enriched the modern genetic landscape of the French territory. However, it is beyond our intention to explore this genetic contribution here, which can be quantified much more precisely with demographic analyses. Instead, we are more keen to dig into a deeper and ancient genetic background. The geographical position of France strongly affected the history of the settlement of the different parts of the territory, whose continuous fragmentation through time is testified by the huge amount of populations and cultures that settled this area. Greeks, Romans and Celtic tribes from central Europe shaped a first internal structure between the 6th and the 1st centuries BCE, while waves of barbarian invasions (Alamanni, Burgundians, Visigoths, Franks, and Celts) harshly impacted the population landscape of

France during the 5th century CE (Haine, 2000). During the 9th and 10th centuries CE, foreign invasions from all sides also influenced the territory: Muslims and Saracens from North Africa coming through Iberia, Hungarian Magyar from the east, and Vikings (*Northmen*) from the north (Haine, 2000). At the light of this complex past, the genetic landscape of France has been poorly analyzed, especially in recent times. First studies with classical markers defined a general heterogeneous pattern considering different geographical arrangements such as military districts, historical provinces, and regions (Kherumian, Moullec and Van Cong, 1967; Cambon-Thomsen and Ohayon, 1988). With his synthetic maps, Cavalli-Sforza proposed that this heterogeneity was a consequence of differential Neolithic influences between northern and southern France, and also pointed out a differentiation for Brittany and the Franco-Cantabrian region (Cavalli-Sforza, Menozzi and Piazza, 1994). More recently, studies on mitochondrial DNA highlighted a general homogeneity when the samples were distributed among the 22 regions established in 1982 and historic provinces (Dubut *et al.*, 2004; Richard *et al.*, 2007). Generally, the mtDNA haplogroup composition of French people did not differentiate neither internally, nor from the surrounding European genetic landscape (Dubut *et al.*, 2004; Richard *et al.*, 2007). On a microgeographical scale, Brittany showed affinity with Scandinavia and Britain, while French Basques stood out for a high frequency of haplogroup H, suggesting a link with the Neolithic diffusion in Europe (Dubut *et al.*, 2004; Richard *et al.*, 2007). In agreement with the homogeneity described by mtDNA studies, Y-

chromosome strongly pointed out a lack of differentiation between the distinct groups when samples were organized on a regional scale. Even in this case, Brittany represented an exception, showing a lower Y-chromosome diversity that was interpreted as consequence of a possible founder effect, plus an isolation process (Ramos-Luis *et al.*, 2014). Based on autosomal variants, a genome-wide study on Western France did not find any differentiation among the distinct groups organized on a regional geographical distribution. Even in this case, the only outlier was Brittany, whose higher linkage disequilibrium suggested a lower effective population size, thus supporting the hypothesis of isolation inferred by the outcomes of the Y-chromosome analyses (Karakachoff *et al.*, 2015). Furthermore, in agreement with mitochondrial studies, Bretons were found to be admixed with individuals from the British Isles (Karakachoff *et al.*, 2015). In this work, we present the first more comprehensive genome-wide study on France with the aim to define patterns of internal differentiation using both allele frequency and haplotype-based methods, presenting a more heterogeneous geogenetical landscape.

Materials and Methods

Dataset arrangement and genotypes

In this study, informed consent was obtained from 331 individuals from different French departments. Internal Review Board approval

for this work was granted by CEIC-PSMAR ref. 2016/6723/I. These samples were first reported in an analysis of Y-chromosome markers in Ramos-Luis et al. 2014. DNA was extracted from blood samples as described in Ramos-Luis et al. (Ramos-Luis *et al.*, 2014). A total of four Axiom ® Genome-Wide Human Origins Arrays (~629 K SNPs) (Patterson *et al.*, 2012) were genotyped at the Centro Nacional de Genotipado - Universidade de Santiago de Compostela facility. Genotype calling was performed running four different batches according to the Affymetrix Best Practices Workflow implemented in the software Axiom™ Analysis Suite 2.0. Out of 331 samples, 52 failed the genotyping process and a total of 279 samples were retained. Three additional samples were removed following an Identity-by-descent analysis (IBD) since they displayed a Proportion IBD value ≥ 0.125 (minimum threshold for removing relatedness equal or higher than a third degree). Eventually, 276 samples were retained. To complete the French dataset, 79 additional samples from a public source (Lazaridis, Nadel, Rollefson, Merrett, *et al.*, 2016) and 60 from unpublished data (from an ongoing study on the Basque Country and the Franco-Cantabrian region) were added to the original 276, leading to a total of 415 samples. In a preliminary part of this work, 20 out of the 276 samples were identified as outliers and removed from the study (see Supplementary_Figure1 and caption). Thus, the complete dataset included 256 newly genotyped samples, plus 139 additional ones, for a final group of 395 samples distributed among 20 different French departments (see Figure 1 for the geographical distribution). As comparison with external populations, 218 samples among

Germany, Norway, Spain, England, Ireland, and Scotland were used from public data (Lazaridis, Nadel, Rollefson, Merrett, *et al.*, 2016), together with 107 samples from the Spanish autonomous communities of Catalonia, Valencian Community, and Balearic islands from public data (Biagini *et al.*, 2019), and 8 additional samples from South Italy (Naples) newly genotyped with Axiom ® Genome-Wide Human Origins Arrays (~629 K SNPs) and presented in this study for the first time.

Data Quality Control

Data were prepared using PLINK1.9 (Purcell *et al.*, 2007). Uniparental markers and X-chromosome variants were excluded. For the French dataset, a preliminary set of filters were applied to each group separately before the merging process. We filtered out all variants with missing call rate greater than 5%, those that failed Hardy-Weinberg test at $p < 10^{-5}$, and samples with more than 10% missing genotype data. After merging, only variants common to the three datasets were retained and SNPs with a minor allele frequency (MAF) below 5% were excluded, resulting in a final 343,884 variants used for haplotype-based methods. For the analyses that needed a set of independent markers, SNPs were pruned setting a pairwise linkage disequilibrium maximum threshold of 0.5, a window of size 200 and a shift step of 25. Eventually, the pruned data retained 142,803 variants. In the analyses that included the external populations, only the pruned dataset, consisting in 154,889, was used.

Statistical analyses

Eigenvectors were computed using the SmartPCA program in Eigenstrat software package (v. 13050) (Patterson, Price and Reich, 2006). Results were plotted in R (v 3.0.1).

The F_{ST} Fixation index was computed using the SmartPCA tool (v. 13050) from the Eigenstrat software package. Results were produced in Rstudio (RStudio Team, 2015) using R version 3.4.4 (R Core Team, 2018). The F_{ST} matrix was used together with a geographic distance matrix produced with The Geographic Distance Matrix Generator (v. 1.2.3, available from http://biodiversityinformatics.amnh.org/open_source/gdmg) in order to perform a Mantel test correlation using the ade4 (Dray and Dufour, 2007) library in R. Results were displayed using ggplot2 (Wickham, 2009) and reshape (Wickham, 2007) libraries.

Based on different hierarchical levels (within Departments, Between Departments within Areas/Regions, Between Areas/Regions; see Supplementary Figure 2 for a visual representation of the used Areas and Regions), AMOVA was performed using the *poppr.amova* function in R package poppr (v. 2.8.1) (Kamvar, Tabima and Grünwald, 2014; Kamvar, Brooks and Grünwald, 2015) and significance was tested with the *randtest* function implemented in R package ade4. For every percentage of variance, a p-value was calculated based on 1000 permutations.

Patterns of population structure were explored using ADMIXTURE (Alexander and Novembre, 2009) testing from K=2 to K=10 ancestral clusters and using 10 independent random seeds. Results were represented using the software pong (Behr *et al.*, 2016). Admixture was formally tested with f3 statistics computed using the qp3Pop function implemented in Admixtools (Patterson *et al.*, 2012).

EEMS (Estimated Effective Migration Surface)

EEMS (Petkova, Novembre and Stephens, 2016) analysis was run on the 395 samples French dataset using 142,803 variants from the pruned file. With a matrix of average pairwise genetic dissimilarities calculated using the internal program bed2diffs, a sample coordinates file, and a habitat coordinates file generated using Google Earth Pro (v. 7.3.2.5495), we performed 10 pilot runs of 6 million MCMC iterations each, with 3 million burn-in, and a thinning interval of 30,000. A second set of 5 runs was then performed restarting the chain with the highest likelihood with 4 million MCMC iterations, 1 million burn-in, and thinning interval of 10,000. The density of the population grid was set to 300 demes, and random seeds were used for each one of the runs. We used the default hyperparameter values but tuned some of the proposal variances to improve convergence in the second set of runs. Results for the chain with the highest likelihood were displayed using eems.plots function in the R package rEEMSplots.

Haplotype-based analysis

Phasing was performed on the 343,884 variants dataset using the software Shapeit (v. v2.r837) (Delaneau *et al.*, 2014; O'Connell *et al.*, 2014). All 395 samples were used as both recipients and donors when running ChromoPainter (Lawson *et al.*, 2012), without any population specification (-a option) and not allowing self-copying. First, the parameters for the switch rate and global mutation probability were estimated using the EM algorithm implemented in ChromoPainter using the parameters -i 15 -in -iM for chromosomes 1, 7, 14, and 20 for all the samples. This step allows to estimate the two parameters that will be then averaged for all chromosomes. The outcome for the average weighted values for the global mutation probability and the switch rate parameters were respectively 0.000745 and 266.67196. In a second step, ChromoPainter was run for all chromosomes using the two fixed parameters. Later, the final coancestry matrices for each chromosome were combined using the tool Chromocombine. The latter also estimates the C parameter which is needed for the normalization of the coancestry matrix data when we run fineSTRUCTURE in order to identify the population structure. The MCMC of fineSTRUCTURE was run using 1000000 burn-in iterations (flag -x), 2000000 iterations sampled (flag -y), and thinning interval of 10000 (flag -z). Eventually, the fineSTRUCTURE tree was estimated running three different seeds and using the flags -X -Y -m T that allow to build the sample relationship tree.

Results

Internal genetic structure in France

In order to define the best geographical partitioning of genetic differentiation, a hierarchical analysis of molecular variance (AMOVA) was performed with areas or regions as major grouping factors. In the first AMOVA we determined the proportion of genetic variation partitioned among geographic areas, among departments within geographic areas, and within departments. In the second one we tested the proportion of genetic variation partitioned among regions (considering the 13 regions established in 2016), among departments within regions, and within departments. A further AMOVA was performed only testing the proportion of genetic variation partitioned among and within departments. As shown in Table 1, in all cases the main contribution to the genetic variance was found at the lowest hierarchical level (variation within departments), while differences among regions resulted in a negative value that could be interpreted as zero, meaning absence of any structure at this level. Conversely, differences among areas displayed positive values, supporting the role of areas as more reliable grouping factors of genetic variations when considering wider samples distributions. Finally, the results for the variation between departments, also supported by significant p-values in all the AMOVA analyses, pointed to the fact that this level of stratification might be a better representation for the minimal unit of genetic differentiation. Based on these results, samples were

distributed on map according to the departmental locations (Figure 1) and all the subsequent analyses considered this grouping factor. A first Principal Component Analysis (PCA) showed two distinct groups separated along the first PC (Figure 2A): the Basque samples on the right part of the plot, against most of the rest of the samples on the left one, within which a structure cannot be defined. These two major groups are connected by a "bridge" of samples represented by non-Basque-speaking individuals from the Franco-Cantabrian region in the southwestern corner of France. When we averaged the eigenvalues for the first two PCs and represented the same PCA, together with standard deviation (SD) values for each group, no evident pattern could still be discerned (Supplementary Figure 3A). When we removed both Basque- and non-Basque-speaking Franco-Cantabrian samples from the analysis (Figure2B), the resulting PCA showed some internal pattern of differentiation, more clearly defined by the average PCA (Supplementary Figure 3B), in which samples from the departments belonging to the northeastern region of Brittany seem to form a cluster on the left part of the plot.

Patterns of gene flow within France

In the genetic variation computed with the F_{ST} analysis, a general homogeneous pattern was found, with fine scale values of differentiation between some departments. The Franco-Cantabrian samples showed the highest values of differentiation with the northwestern departments reaching scores between 0.008 and 0.009

for the Basque-speaking samples, and between 0.004 and 0.006 for the non-Basque-speaking ones (Supplementary Figure 4A, left), followed by lower values of differentiation with the northern and northeastern departments. Without the Franco-Cantabrian samples, the main differentiation was recorded between the northwestern departments and the southeastern corner of the country, with a highest value of differentiation around 0.002 between the southeastern department of Bouches-du-Rhône (BdR) and the northwestern Breton department of Côtes-d'Armor (CdA) (Supplementary Figure 4B, left). Lower levels of differentiation were locally found among the departments in the northwest, among those in the north together with the northeastern ones, and among the Basque-speakers Franco-Cantabrian groups. A Multidimensional Scaling analysis (MDS) based on the F_{ST} matrices clearly showed how the Franco-Cantabrian samples separate from the rest of the groups (Supplementary Figure 4A, right), and how the Breton departments do the same once the Franco-Cantabrian samples are removed (Supplementary Figure 4B, right). A Mantel test of isolation by distance (IBD) between the F_{ST} values and the geographical distances showed a positive and statistically supported correlation (R^2=0.332, P=0.001) (Supplementary Figure 5A), moving to even more positive values when the Franco-Cantabrian samples were removed (R^2=0.432, P=0.001) (Supplementary Figure 5B). Next, we used the EEMS analysis, a method for visualizing genetic diversity patterns, and found that the resulting effective migration surface mirrors the outcomes of genetic differentiation detected by the F_{ST} analyses (Figure 3); a higher effective migration

was locally found in northern, northeastern and northwestern France among departments belonging to the same geographical areas, while a major barrier was discovered along the western side of France.

Haplotype sharing patterns within France

Using haplotype-based methods, we looked for marks of haplotype sharing, illustrating relations between departments. We linked samples belonging to the same branch in the fineSTRUCTURE tree to the same haplotype, then we represented the results on a departmental scale; the outcome is a picture of haplotype distributions within France (Figure 4). The resulting map is clearly consistent with the results seen so far; we could define at least four distinct groups, plus a more widespread component. In the southwestern corner, the Franco-Cantabrian samples clearly split in two groups, represented by the Basque-speaking and non-Basque-speaking subjects, respectively. In the northwestern vertex, the Breton departments exhibit their very own haplotypic signature, in agreement with the lower level of differentiation detected with the F_{ST} analysis and the higher internal effective migration rate detected with EEMS. The same was found for the northern and northeastern departments that display a clearly shared haplotypic configuration. Furthermore, a more generally spread "French haplotype pattern" is found on the north-south axis. Lastly, also the southwestern department of Haute-Garonne (HG) and the southeastern one of Bouches-du-Rhône (BdR) present higher frequencies for some local

haplotypes that in other departments reached only lower frequencies.

Sources of gene flow into France

When we added external sources from the surrounding populations to test for signatures of admixture events, the configuration observed pointed to a general homogeneous picture. The only exception was represented by the samples belonging to the Breton departments whose configuration was more like the one observed for the Irish, Scottish, and English groups. Moving through the different K ancestral components, this behavior clearly characterizes the northwestern departments, separating them from the rest of the French groups since from the very first K ancestral components (Supplementary Figure 6). Thus, we formally tested for admixture events using the f3-statistics with the Test groups being the different departments, and the external surrounding populations as Sources. We only retained the negative f3 values for those departments represented at least by two individuals. Results are shown in Table 2 were only significant Z-scores < -3 are reported, while results for those departments passing all the requested filters but with higher Z-score values are shown in Supplementary Table 1. Notably, in 9 departments, a combination of sources that was highly significant was Ireland-Southern Italy (see discussion below)

Discussion

In order to describe the internal structure of Metropolitan France, in this work we used both allele frequency and haplotype-based methods. While the first described a more homogeneous landscape, the latter unveiled patterns of local differentiation. In previous works about France, samples were differently arranged into the geographical space and no consensus had been reached on what system would work better; apart from less canonical systems of organization like the military districts (Kherumian, Moullec and Van Cong, 1967), historical provinces (Cambon-Thomsen and Ohayon, 1988; Richard *et al.*, 2007) and old regions (Ramos-Luis *et al.*, 2014) are the most used so far. Thus, our first goal was to look for the best geographical level of genetic stratification before arranging our samples on a map. After the French Revolution in 1790, in order to weaken the old loyalties, the ancient provinces of France were subdivided into departments, whose overall configuration has been mostly conserved so far (Forstenzer, 1981). Furthermore, in 1982, a system of 22 regions was established by grouping different departments into wider areas (Sowerwine, 2009). However, in 2016, the number of the regions was reduced to 13, with the consequent rearrangement of the departments (OECD, 2017). Given this background, our AMOVA results support the idea that regions, as a new internal reorganization, are not a suitable model for the genetic compartmentalization and point to the absence of any contribution to the total genetic variation, possibly implying that regions are separating genetically similar departments

into different groups. On the other hand, departments, as result of a more conserved internal geographical structure, represent the best minimal unit of genetic stratification. Using this system of internal organization, we found a truly remarkable connection between the different results we achieved, but it was with fineSTRUCTURE that we could really define the first ever detected internal subdivision of France. A general widespread "French haplotype" moving through the north-south axis was detected; possibly the overall homogeneity found with the principal component analysis can be linked to the fact that on an allele frequency scale, such wide spread pattern may represent a confounding factor. Indeed, only the two Franco-Cantabrian groups were not reached by this common "French haplotype". These two populations clearly differentiated from the rest of the French groups for both allele frequency and haplotype-based methods. It is interesting to notice that the presence of two distinct groups in the Franco-Cantabrian region stressed the outcome of the isolation the Basque-speaking groups experienced, splitting from their non-Basque-speaking neighbors from the very same departments. This finding is in agreement with their recognized distinct cultural entity (Calafell and Bertranpetit, 1994) and their genetic outlier position in the European landscape (Rodríguez-Ezpeleta *et al.*, 2010), as also with the lower internal levels of differentiation we detected with the F_{ST} analysis, and the low effective migration rates evidenced by EEMS, resulting in a barrier to migration in the southwestern corner of France. The latter was found to span along the entire western French coast, defining a second barrier in the northwestern corner, justifying the presence of

another distinct group represented by the Breton departments. This group was firstly detected, on a coarser scale, with the removal of the Franco-Cantabrian samples from the first PCA, and its outstanding position is in agreement with different studies on both uniparental and autosomal markers (Dubut *et al.*, 2004; Richard *et al.*, 2007; Ramos-Luis *et al.*, 2014; Karakachoff *et al.*, 2015). However, based on the fineSTRUCTURE results, in our work we detected a stronger evidence of differentiation based on haplotypic data. Furthermore, the barrier detected with EEMS on the western side of France is supported by the highest level of genetic differentiation between the Breton and the Franco-Cantabrian groups. Such levels of F_{ST} separated geographically more distant groups, in a corner-to-corner scheme inside the *hexagone*, thus justifying the positive results obtained with the Mantel test. Also, the genetic distances the Franco-Cantabrian and Breton samples show, span through the entire geographical space, skewing the more homogeneous samples in the lower part of the distribution (Supplementary Figure 5A), defining two distinct groups, separated by the trend bar. Furthermore, the barrier in the northwestern corner is consistent with the lower levels of allele sharing we found between the Breton departments and the southeastern corner of France. Apart from some possible artifacts due to the presence of unsampled regions (House and Hahn, 2018), we found the EEMS results consistent with the rest of the data collected in our study. Indeed, another group with a higher internal migration rate according to the EEMS results, is represented by the north and the northeastern departments. Samples from these areas showed a

shared haplotypic configuration as detected by fineSTRUCTURE and lower levels of internal differentiation according to the F_{ST} results, pointing to a preferential internal gene flow in this area. In the southern part of France, two departments, Haute-Garonne (HG) in the southwest and Bouches-du-Rhône (BdR) in the southeast also showed higher frequencies for some local haplotypes.

In order to understand whether these internal patterns of differentiation are due to recent events or whether they reflect a more ancient history, we performed an admixture analysis. The homogeneity we detected affected all the French departments, with one important exception: Brittany, whose connection to the Irish samples completely agrees with previous findings (Dubut *et al.*, 2004; Ramos-Luis *et al.*, 2014; Karakachoff *et al.*, 2015). Historical migrations from Ireland to Brittany are well recorded since the 4th century CE (Monnier, 1997), as well as the immigration of Irish people during the War of Ireland (1641-1651) into the present day departments of Finistère (FI) and Côte d'Armor (CdA), within which a higher integration of the Irish immigrants is proved by records of marriage, birth and death certificates (Dubut *et al.*, 2004). Furthermore, a Celtic root for the Breton language links the Breton departments to the ‘‘Insular Celtic" languages from the British Isles (Forster and Toth, 2003). One of the most surprisig outcomes of our work is undoubdetly represented by the f3-statistics; 9 out of 22 distinct targets we tested against different external sources gave significant results with the lowest Z-scores detected for the same couple represented by the Italian South and Irish sources. Z-scores lower than -3 indicate that our test populations are admixed from

sources not necessarily identical but related to the sources we used in the analysis (Lazaridis, Nadel, Rollefson, C.Merrett, *et al.*, 2016). Thus, these two sources may be proxies, respectively, for a Mediterranean and continental components. A further analysis is required to restrict the time frame when this admixture occurred, but we may hypothesize an Iron Age infusion of Mediterranean Greeks and continental Celts, or two different paths followed by the Neolithic expansion, either along the Mediterranean or through the Balkans and Central Europe.. According to this model, the general outcome is a widespread homogenous background that link the entire French territory to pre-historic events; by 600 BCE Greeks established a colony on the Mediterranean coastline of France in the city of Massalia (present-day Marseille) (Fine, 1985); around the same time (ca. 700 BCE) Celtic tribes from central Europe (subsequently known as Gauls) expanded across the entire territory bringing their culture, advanced farming tecniques, and their language (Hubert, 1989). In conclusion, the French genetic landscape seems to be homogeneously dominated by an ancient background, with cultural isolates in the southwest and a more recent gene flow into the northwest, but also with subsequent isolation.

References

1000 Genomes Project Consortium *et al.* (2015) ‘A global reference for human genetic variation.’, *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Alexander, D. H. and Novembre, J. (2009) ‘Fast Model-Based Estimation of Ancestry in Unrelated Individuals’, *Genome Research*, 19(9), pp. 1655–1664. doi: 10.1101/gr.094052.109.vidual.

Behr, A. A. *et al.* (2016) ‘Pong: Fast analysis and visualization of latent clusters in population genetic data’, *Bioinformatics*, 32(18), pp. 2817–2823. doi: 10.1093/bioinformatics/btw327.

Biagini, S. A. *et al.* (2019) ‘People from Ibiza: an unexpected isolate in the Western Mediterranean’, *European Journal of Human Genetics*. Springer US. doi: 10.1038/s41431-019-0361-1.

Calafell, F. and Bertranpetit, J. (1994) ‘Principal component analysis of gene frequencies and the origin of Basques’, *Am J Phys Anthropol*, 93(2), pp. 201–215. doi: 10.1002/ajpa.1330930205.

Cambon-Thomsen, A. and Ohayon, E. (1988) ‘Practical Application of Population Genetics: The Genetic Survey “Provinces Françaises”’, in Mayr, W. (ed.) *Advances in Forensic Haemogenetics. Advances in Forensic Haemogenetics, vol2*. Berlin,

Heidelberg: Springer.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994) *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Delaneau, O. *et al.* (2014) 'Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel', *Nature Communications*. Nature Publishing Group, 5, pp. 1–9. doi: 10.1038/ncomms4934.

Dray, S. and Dufour, A.-B. (2007) 'The **ade4** Package: Implementing the Duality Diagram for Ecologists', *Journal of Statistical Software*, 22(4). doi: 10.18637/jss.v022.i04.

Dubut, V. *et al.* (2004) 'mtDNA polymorphisms in five French groups: Importance of regional sampling', *European Journal of Human Genetics*, 12(4), pp. 293–300. doi: 10.1038/sj.ejhg.5201145.

Fine, J. V. A. (1985) *The Ancient Greeks: A Critical History*. Belknap Press: An Imprint of Harvard University Press.

Forstenzer, T. R. (1981) *French Provincial Police and the Fall of the Second Republic: Social Fear and Counterrevolution*. Princeton University Press.

Forster, P. and Toth, A. (2003) 'Toward a phylogenetic chronology

of ancient Gaulish, Celtic, and Indo-European', *Proceedings of the National Academy of Sciences*, 100(15), pp. 9079–9084. doi: 10.1073/pnas.1331158100.

Haine, W. S. (2000) *The History of France*. Greenwood Press.

House, G. L. and Hahn, M. W. (2018) 'Evaluating methods to visualize patterns of genetic differentiation on a landscape', *Molecular Ecology Resources*. doi: 10.1111/1755-0998.

Hubert, H. (1989) *The Greatness and Decline of the Celts*. Marboro Books.

Kamvar, Z. N., Brooks, J. C. and Grünwald, N. J. (2015) 'Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality', *Frontiers in Genetics*, 6(JUN), pp. 1–10. doi: 10.3389/fgene.2015.00208.

Kamvar, Z. N., Tabima, J. F. and Grünwald, N. J. (2014) 'Poppr : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction ', *PeerJ*, 2, p. e281. doi: 10.7717/peerj.281.

Karakachoff, M. *et al.* (2015) 'Fine-scale human genetic structure in Western France', *European Journal of Human Genetics*, 23(6), pp. 831–836. doi: 10.1038/ejhg.2014.175.

Kherumian, R., Moullec, J. and Van Cong, N. (1967) ‘Groupes sanguins érythrocytaires A 1 , A 2 , BO, MN, Rh (CcDE) et sériques, Hp, Tf, Gm dans quatre régions militaires françaises.’, *Bulletins et Mémoires de la Société d’Anthropologie de Paris*, 1(XII), pp. 377–384. doi: DOI : https://doi.org/10.3406/bmsap.1967.1396.

Lawson, D. J. *et al.* (2012) ‘Inference of Population Structure using Dense Haplotype Data’, *PLoS Genetics*, 8(1), pp. 11–17. doi: 10.1371/journal.pgen.1002453.

Lazaridis, I., Nadel, D., Rollefson, G., C.Merrett, D., *et al.* (2016) ‘Genomic insights into the origin of farming in the ancient Near East’, *Nature*. doi: 10.1038/nature19310.

MacMaster, N. (1997) *Colonial Migrants and Racism: Algerians in France, 1900–62*. Palgrave Macmillan.

Monnier, J. (1997) ‘Chapitre 6 : L’immigration bretonne en Armorique’, in Monnier, J. and Cassard, J. (eds) *Toute l’histoire de Bretagne*. Skol Vreizh, pp. 97–106.

O’Connell, J. *et al.* (2014) ‘A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness’, *PLoS Genetics*, 10(4), p. e1004234. doi: 10.1371/journal.pgen.1004234.

OECD (2017) *OECD Multi-level Governance Studies Multi-level Governance Reforms Overview of OECD Country Experiences*.

Patterson, N. *et al.* (2012) ‘Ancient Admixture in Human History’, *Genetics*, 192(November), pp. 1065–1093. doi: 10.1534/genetics.112.145037.

Patterson, N., Price, A. L. and Reich, D. (2006) ‘Population structure and eigenanalysis’, *PLoS Genetics*, 2(12), pp. 2074–2093. doi: 10.1371/journal.pgen.0020190.

Petkova, D., Novembre, J. and Stephens, M. (2016) ‘Visualizing spatial population structure with estimated effective migration surfaces’, *Nature Genetics*, 48(1), pp. 94–100. doi: 10.1038/ng.3464.Visualizing.

Purcell, S. *et al.* (2007) ‘PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses’, *The American Journal of Human Genetics*, 81(3), pp. 559–575. doi: 10.1086/519795.

R Core Team (2018) ‘R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.’

Ramos-Luis, E. *et al.* (2014) ‘Y-chromosomal DNA analysis in French male lineages’, *Forensic Science International: Genetics*. Elsevier Ireland Ltd, 9(1), pp. 162–168. doi: 10.1016/j.fsigen.2013.12.008.

Richard, C. *et al.* (2007) ‘An mtDNA perspective of French genetic variation’, *Annals of Human Biology*, 34(1), pp. 68–79. doi: 10.1080/03014460601076098.

Rodríguez-Ezpeleta, N. *et al.* (2010) ‘High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations’, *Human Genetics*, 128(1), pp. 113–117. doi: 10.1007/s00439-010-0833-4.

RStudio Team (2015) ‘RStudio Team (2015). RStudio: Integrated Development for R.’ Available at: http://www.rstudio.com/.

Sowerwine, C. (2009) *France since 1870: Culture, Society and the Making of the Republic*. 2nd edn. Palgrave Macmillan.

Wickham, H. (2007) ‘Reshaping Data with the \pkgreshape Package’, *Journal of Statistical Software*, 21(12), pp. 1–20. doi: 10.1016/S0142-1123(99)00007-9.

Wickham, H. (2009) ‘ggplot2: Elegant Graphics for Data Analysis’, *Media*, 35(July), p. 211. doi: 10.1007/978-0-387-98141-3.

Main Figures and Tables

	Groupings	% Total variance	Φ-statistics	p
A	Variations Between Areas	0.07	Φ_{ST}= 0.0007	0.2417
	Variations Between Departments Within Areas	0.17	Φ_{ST}= 0.0017	0.0009
	Variations Within Departments	99.75	Φ_{ST}= 0.0024	0.0009
B	Variations Between Regions	-0.073	Φ_{ST}= -0.0007	0.7462
	Variations Between Departments Within Regions	0.3	Φ_{ST}= 0.003	0.0009
	Variations Within Departments	99.76	Φ_{ST}= 0.0023	0.0009
C	Variations Between Departments	0.23	Φ_{ST}= 0.0023	0.0009
	Variations Within Departments	99.76		

Table 1. Hierarchical analysis of molecular variance (AMOVA). Results for percentage of total variance, Φ-statistics, and p-values are reported for the three distinct analyses. **A)** proportion of genetic variation partitioned among geographic areas, among departments within geographic areas, and within departments; **B)** proportion of genetic variation partitioned among regions, among departments within regions, and within departments; **C)** proportion of genetic variation partitioned among departments and within departments.

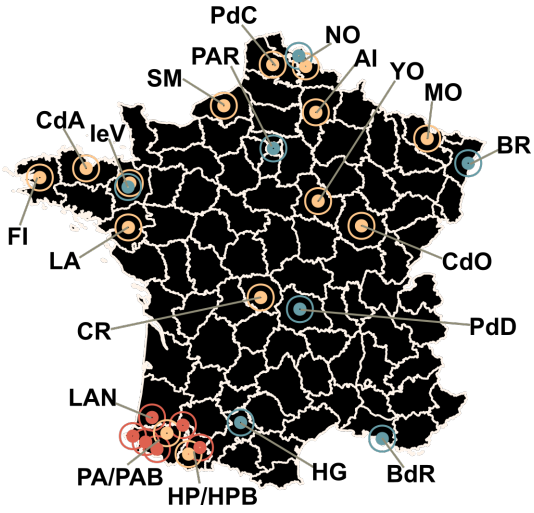


Figure 1. Map showing sample distribution among the different departments. Geographical coordinates are averages among samples. Different colors define the three datasets used in this work (blue dots correspond to the 256 samples genotyped for this work; yellow dots correspond to the 79 samples from Lazaridis et al., 2016; red dots correspond to the 60 samples from unpublished data). Sample size and acronyms for the departments are: PdD, Puy-de-Dôme (33); CR, Creuse (25); CdO, Côte-d'Or (1); AI, Aisne (1); NO, Nord (47); PdC, Pas-de-Calais (4); PAR, Paris (22); YO, Yonne (1); MO, Moselle (8); BR, Bas-Rhin (48); IeV, Ille-et-Vilaine (45); CdA, Côtes-d'Armor (3); FI, Finistère (5); SM, Seine-Maritime (2); LA, Loire-Atlantique (1); BdR, Bouches-du-Rhône (21); LAN, Landes (10); HG, Haute-Garonne (43); PA, Pyrénées-Atlantiques (15); PAB, Pyrénées-Atlantiques Basque (31); HP, Hautes-Pyrénées (9); HPB, Hautes-Pyrénées Basque (20).

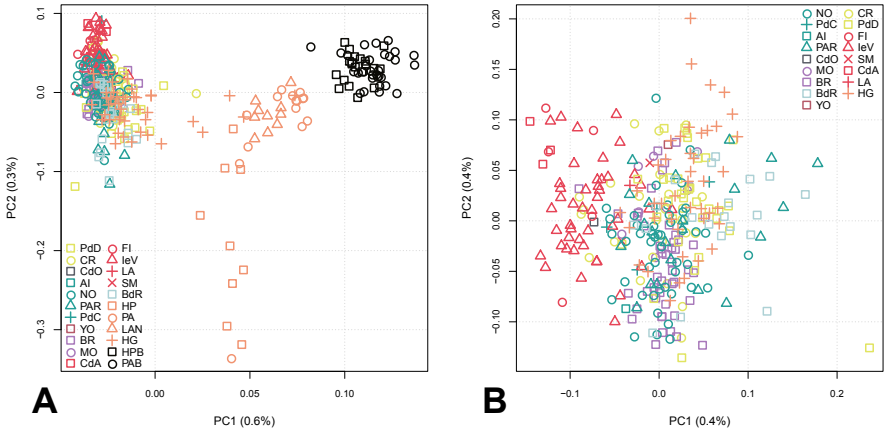


Figure 2. Principal Component Analysis with A) Franco-Cantabrian samples, and B) without them. Colors correspond to distinct geographic areas, while different symbols with the same color represent distinct departments in each area. However, Basques are colored differently than the non-Basque-speaking samples from that same area, but symbols recall the departments they share with the non-Basque-speaking groups.

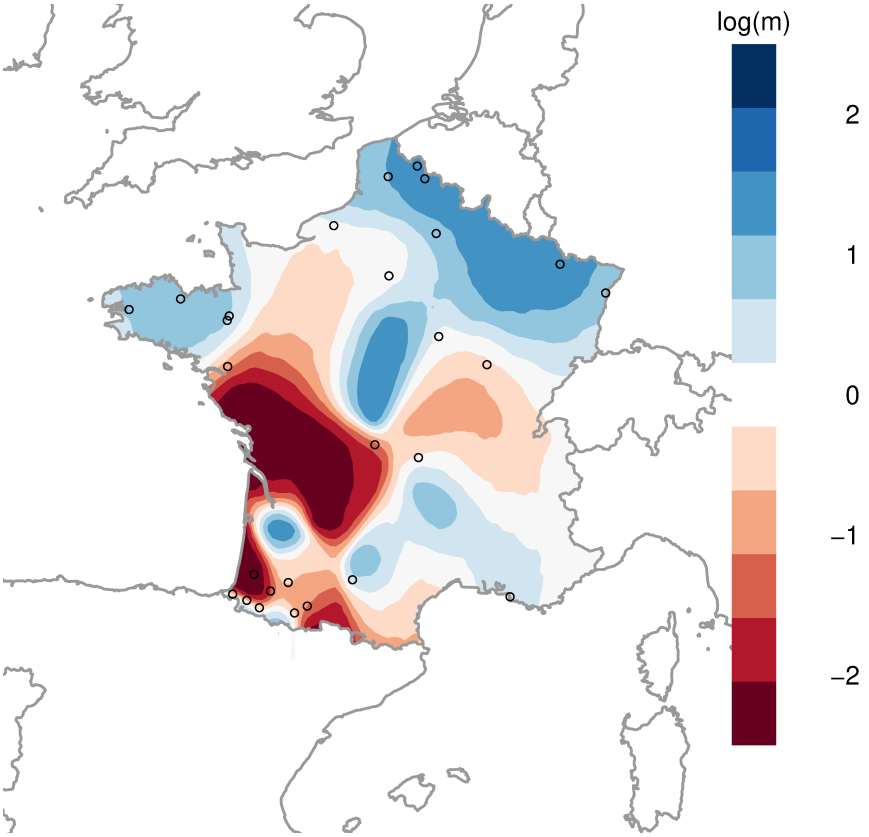


Figure 3. EEMS plot based on 395 French samples. Different shades of the same color represent differential levels of high (blue) or low (red) effective migration rates. The zero value indicates the average effective migration rate. Geographical locations for the different departments are averages of the coordinates among samples.

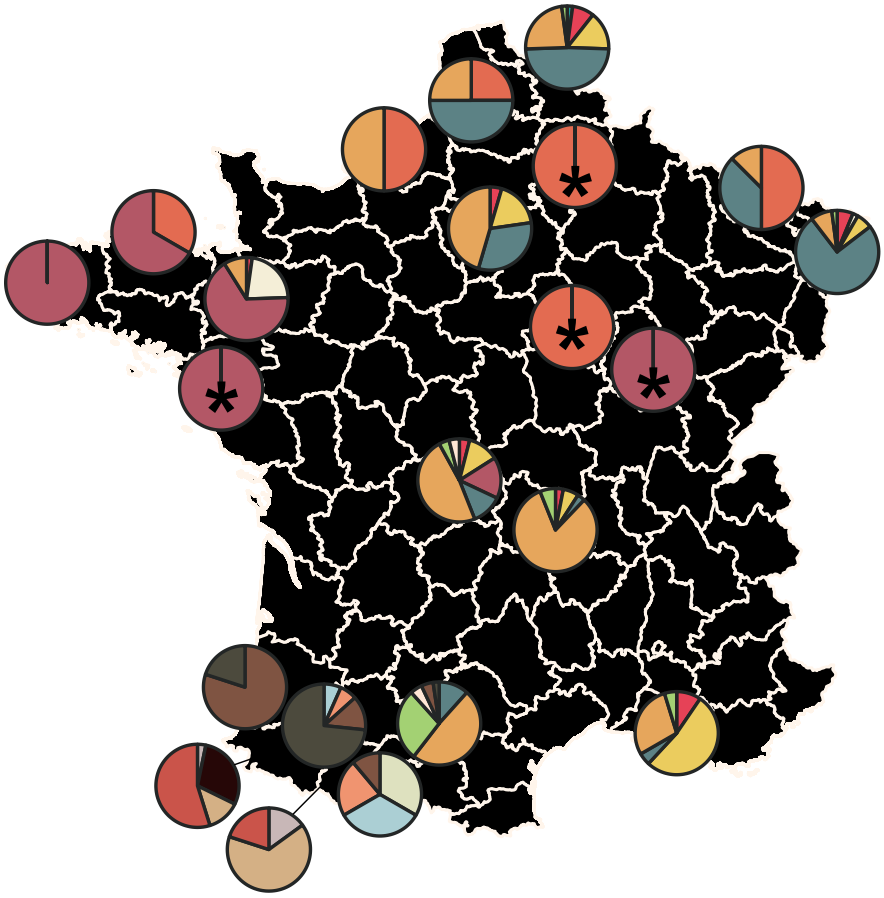


Figure 4. Pie charts showing the spatial distribution of haplotypes inferred by the fineSTRUCTURE tree. See Figure 1 for department names. Asterisks indicate departments with only one sample.

Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
BdR	Italian_South	Irish	-0.001401	-12.843
BdR	Italian_South	Scottish	-0.001232	-12.129
BdR	Italian_South	Basque_Country	-0.000967	-7.427
BdR	Norwegian	Italian_South	-0.000973	-7.001
BdR	German	Italian_South	-0.000942	-6.728
BdR	English	Italian_South	-0.000879	-6.188
BdR	Italian_North	Irish	-0.000626	-5.34
BdR	Italian_North	Scottish	-0.000508	-4.708
BdR	Norwegian	Italian_North	-0.00063	-4.502
BdR	South_Spain	Norwegian	-0.000536	-3.469
BdR	Norwegian	Basque_Country	-0.000475	-3.165
BdR	South_Spain	Scottish	-0.000341	-3.065
BdR	South_Spain	Irish	-0.000371	-3.016
BR	Italian_South	Irish	-0.001114	-14.112
BR	Italian_South	Scottish	-0.000886	-13.011
BR	Italian_North	Irish	-0.000783	-9.907
BR	Italian_North	Scottish	-0.000608	-9.575
BR	Norwegian	Italian_North	-0.000797	-7.74
BR	Norwegian	Italian_South	-0.000694	-6.563
BR	German	Italian_South	-0.000693	-6.457
BR	Norwegian	Mediterranean_Spain	-0.000388	-5.612
BR	English	Italian_South	-0.000499	-4.921
BR	German	Italian_North	-0.000495	-4.832
BR	German	Mediterranean_Spain	-0.0003	-4.252
BR	South_Spain	Norwegian	-0.000506	-4.18
BR	South_Spain	Irish	-0.00033	-3.868
BR	Central_Spain	German	-0.000412	-3.781
BR	Norwegian	Basque_Country	-0.000404	-3.385
BR	South_Spain	Scottish	-0.000243	-3.341
BR	Northwestern_Spain	Scottish	-0.000259	-3.27
BR	German	Northwestern_Spain	-0.000425	-3.266
BR	English	Italian_North	-0.000327	-3.175
BR	Mediterranean_Spain	Irish	-0.000164	-3.043
BR	Mediterranean_Spain	Scottish	-0.000145	-3.005
CR	Italian_South	Irish	-0.001268	-12.482
CR	Italian_South	Scottish	-0.001006	-10.597
CR	Italian_North	Irish	-0.001018	-10.162
CR	Italian_North	Scottish	-0.000808	-8.695
CR	Norwegian	Mediterranean_Spain	-0.000747	-8.007
CR	Norwegian	Italian_North	-0.000993	-7.904
CR	South_Spain	Irish	-0.000814	-7.629
CR	South_Spain	Scottish	-0.000693	-7.451
CR	Norwegian	Central_Spain	-0.00088	-7.091
CR	Mediterranean_Spain	Irish	-0.000562	-7.027
CR	Mediterranean_Spain	Scottish	-0.000509	-6.924
CR	Central_Spain	Irish	-0.000715	-6.91
CR	South_Spain	Norwegian	-0.000951	-6.735
CR	Norwegian	Basque_Country	-0.000987	-6.656
CR	Norwegian	Italian_South	-0.000809	-6.305
CR	Central_Spain	Scottish	-0.000609	-6.245
CR	Central_Spain	English	-0.00073	-5.409
CR	English	Italian_South	-0.000692	-5.33

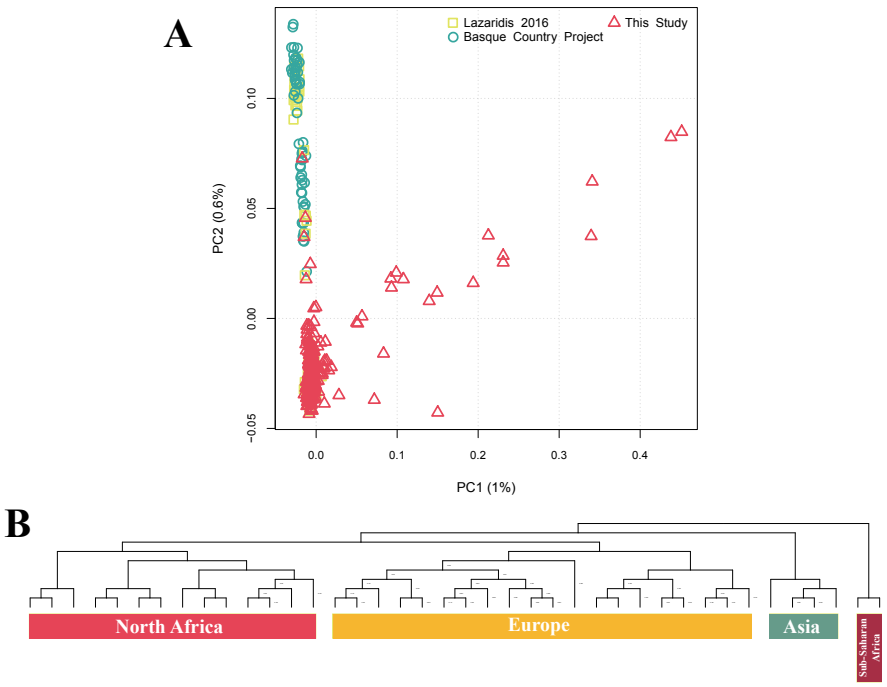
Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
CR	Basque_Country	Irish	-0.000563	-4.905
CR	English	Basque_Country	-0.000718	-4.895
CR	Central_Spain	German	-0.000667	-4.779
CR	South_Spain	English	-0.000652	-4.73
CR	English	Mediterranean_Spain	-0.000471	-4.727
CR	Basque_Country	Scottish	-0.000499	-4.687
CR	English	Italian_North	-0.000601	-4.538
CR	Northwestern_Spain	Scottish	-0.000497	-4.509
CR	Northwestern_Spain	Irish	-0.000533	-4.455
CR	German	Italian_South	-0.000507	-3.794
CR	German	Mediterranean_Spain	-0.000358	-3.559
CR	English	Northwestern_Spain	-0.000522	-3.43
CR	South_Spain	German	-0.000453	-3.238
CR	German	Italian_North	-0.00039	-3.118
CR	German	Basque_Country	-0.000471	-3.104
HG	Italian_South	Irish	-0.000918	-11.28
HG	Italian_South	Scottish	-0.000796	-11.014
HG	Italian_North	Scottish	-0.00057	-8.014
HG	Italian_North	Irish	-0.00064	-7.831
HG	Mediterranean_Spain	Scottish	-0.000364	-7.517
HG	South_Spain	Scottish	-0.00057	-7.472
HG	Basque_Country	Scottish	-0.000586	-7.09
HG	Norwegian	Basque_Country	-0.000847	-6.77
HG	Italian_South	Basque_Country	-0.000667	-6.276
HG	South_Spain	Irish	-0.00055	-6.256
HG	English	Basque_Country	-0.000749	-5.718
HG	Basque_Country	Irish	-0.00051	-5.447
HG	Northwestern_Spain	Scottish	-0.000447	-5.224
HG	Norwegian	Mediterranean_Spain	-0.000375	-5.146
HG	Mediterranean_Spain	Irish	-0.000277	-5.1
HG	South_Spain	Norwegian	-0.0006	-4.935
HG	German	Basque_Country	-0.000601	-4.834
HG	Norwegian	Italian_North	-0.000527	-4.832
HG	Central_Spain	Scottish	-0.00035	-4.65
HG	Italian_North	Basque_Country	-0.000469	-4.505
HG	South_Spain	English	-0.000472	-4.029
HG	English	Italian_South	-0.000424	-3.876
HG	Central_Spain	German	-0.00045	-3.78
HG	Central_Spain	Irish	-0.000315	-3.628
HG	Central_Spain	English	-0.000413	-3.613
HG	Northwestern_Spain	Irish	-0.000342	-3.607
HG	English	Mediterranean_Spain	-0.00027	-3.514
HG	German	Mediterranean_Spain	-0.000256	-3.451
HG	Norwegian	Central_Spain	-0.000393	-3.322
HG	Norwegian	Italian_South	-0.000371	-3.292
HG	English	Northwestern_Spain	-0.000414	-3.097
HG	South_Spain	German	-0.000372	-3.074
IeV	Italian_North	Irish	-0.000493	-6.175
IeV	Italian_South	Irish	-0.00051	-6.05
IeV	South_Spain	Irish	-0.000423	-4.624
IeV	Mediterranean_Spain	Irish	-0.000257	-4.498
IeV	Central_Spain	Irish	-0.000345	-4.365
MO	Italian_South	Irish	-0.001174	-5.138

Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
MO	Italian_South	Scottish	-0.001093	-4.962
MO	South_Spain	Norwegian	-0.001053	-4.098
MO	Norwegian	Italian_South	-0.00098	-4.003
MO	German	Italian_South	-0.000935	-3.941
MO	Norwegian	Italian_North	-0.000855	-3.476
MO	Norwegian	Basque_Country	-0.000885	-3.332
MO	South_Spain	German	-0.000813	-3.316
MO	Norwegian	Mediterranean_Spain	-0.00072	-3.276
MO	South_Spain	Scottish	-0.000713	-3.159
NO	Italian_South	Irish	-0.001196	-14.818
NO	Italian_South	Scottish	-0.001029	-14.549
NO	Italian_North	Scottish	-0.000773	-11.819
NO	Italian_North	Irish	-0.000888	-11.416
NO	Mediterranean_Spain	Scottish	-0.000445	-9.474
NO	Norwegian	Mediterranean_Spain	-0.000657	-9.21
NO	Norwegian	Italian_North	-0.000931	-8.896
NO	Mediterranean_Spain	Irish	-0.000404	-7.64
NO	South_Spain	Scottish	-0.000571	-7.552
NO	Norwegian	Italian_South	-0.000807	-7.383
NO	South_Spain	Irish	-0.000598	-6.817
NO	German	Mediterranean_Spain	-0.000486	-6.797
NO	South_Spain	Norwegian	-0.000803	-6.682
NO	German	Italian_South	-0.000721	-6.65
NO	Central_Spain	Scottish	-0.000449	-6.191
NO	Norwegian	Central_Spain	-0.000693	-6.129
NO	Central_Spain	German	-0.000698	-6.102
NO	English	Italian_South	-0.000581	-5.536
NO	Central_Spain	Irish	-0.00046	-5.515
NO	Norwegian	Basque_Country	-0.000646	-5.249
NO	German	Italian_North	-0.000546	-5.174
NO	Northwestern_Spain	Scottish	-0.000403	-4.954
NO	South_Spain	German	-0.000523	-4.418
NO	Central_Spain	English	-0.000435	-4.092
NO	English	Italian_North	-0.000431	-3.938
NO	English	Mediterranean_Spain	-0.000273	-3.789
NO	Northwestern_Spain	Irish	-0.000344	-3.661
NO	South_Spain	English	-0.000396	-3.408
NO	German	Northwestern_Spain	-0.000453	-3.381
PAR	Italian_South	Irish	-0.001536	-14.131
PAR	Italian_South	Scottish	-0.001384	-13.931
PAR	Norwegian	Italian_South	-0.001225	-9.242
PAR	Italian_North	Irish	-0.000911	-8.574
PAR	Italian_North	Scottish	-0.000811	-8.326
PAR	Norwegian	Italian_North	-0.001032	-8.075
PAR	German	Italian_South	-0.001055	-8.024
PAR	English	Italian_South	-0.001026	-7.991
PAR	Norwegian	Mediterranean_Spain	-0.000686	-6.825
PAR	South_Spain	Norwegian	-0.000877	-6.035
PAR	South_Spain	Scottish	-0.000583	-5.506
PAR	South_Spain	Irish	-0.000595	-5.134
PAR	Mediterranean_Spain	Scottish	-0.000411	-5.098
PAR	Norwegian	Central_Spain	-0.000634	-4.494
PAR	Northwestern_Spain	Scottish	-0.000501	-4.465

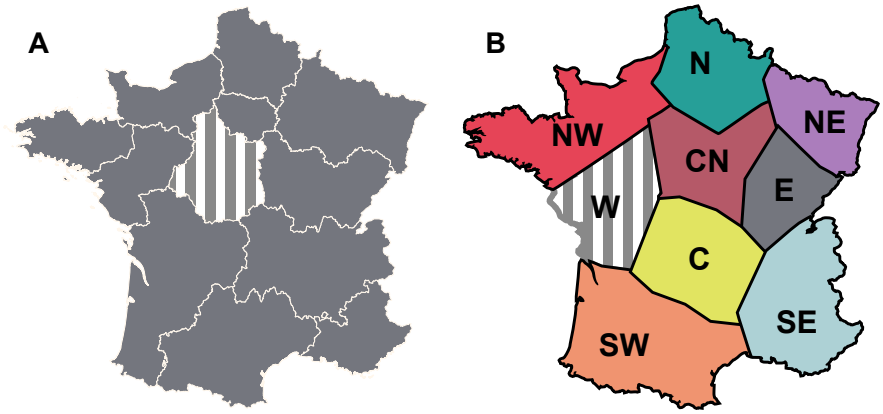
Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
PAR	German	Italian_North	-0.000563	-4.205
PAR	English	Italian_North	-0.000559	-4.15
PAR	Mediterranean_Spain	Irish	-0.000355	-4.036
PAR	German	Mediterranean_Spain	-0.00043	-4.009
PAR	Central_Spain	German	-0.000554	-3.898
PAR	Norwegian	Basque_Country	-0.000601	-3.828
PAR	Italian_South	Basque_Country	-0.000512	-3.755
PAR	South_Spain	German	-0.000513	-3.543
PAR	Northwestern_Spain	Irish	-0.000428	-3.477
PAR	South_Spain	English	-0.000498	-3.439
PAR	German	Northwestern_Spain	-0.000531	-3.341
PAR	English	Mediterranean_Spain	-0.000329	-3.092
PAR	Central_Spain	Scottish	-0.000327	-3.03
PAR	English	Northwestern_Spain	-0.00048	-3.02
PdD	Italian_South	Irish	-0.001146	-12.844
PdD	Italian_South	Scottish	-0.000978	-12.316
PdD	Italian_North	Irish	-0.000688	-7.723
PdD	Italian_North	Scottish	-0.000572	-7.324
PdD	German	Italian_South	-0.000684	-6.048
PdD	Norwegian	Italian_South	-0.000693	-6.016
PdD	Norwegian	Mediterranean_Spain	-0.000466	-5.992
PdD	Mediterranean_Spain	Scottish	-0.000316	-5.786
PdD	Norwegian	Italian_North	-0.000669	-5.677
PdD	South_Spain	Scottish	-0.00045	-5.496
PdD	Italian_South	Basque_Country	-0.000624	-5.473
PdD	English	Italian_South	-0.00062	-5.3
PdD	Norwegian	Basque_Country	-0.000714	-5.239
PdD	South_Spain	Irish	-0.000477	-4.847
PdD	South_Spain	Norwegian	-0.000619	-4.81
PdD	German	Mediterranean_Spain	-0.00037	-4.492
PdD	Central_Spain	German	-0.000553	-4.317
PdD	Mediterranean_Spain	Irish	-0.000275	-4.269
PdD	Norwegian	Central_Spain	-0.000473	-3.945
PdD	Northwestern_Spain	Scottish	-0.000368	-3.932
PdD	German	Basque_Country	-0.000491	-3.584
PdD	English	Basque_Country	-0.000489	-3.455
PdD	Basque_Country	Scottish	-0.000314	-3.455
PdD	South_Spain	German	-0.000415	-3.332
PdD	Central_Spain	Scottish	-0.00029	-3.309
PdD	German	Italian_North	-0.000359	-3.172
PdD	Central_Spain	Irish	-0.000302	-3.124
PdD	German	Northwestern_Spain	-0.000433	-3.044

Table 2. Results for the f_3(Test; Source1, Source2) statistics. Test populations are the departments with at least 2 individuals, Sources are groups of external populations. Only tests with Z scores < -3 are listed.

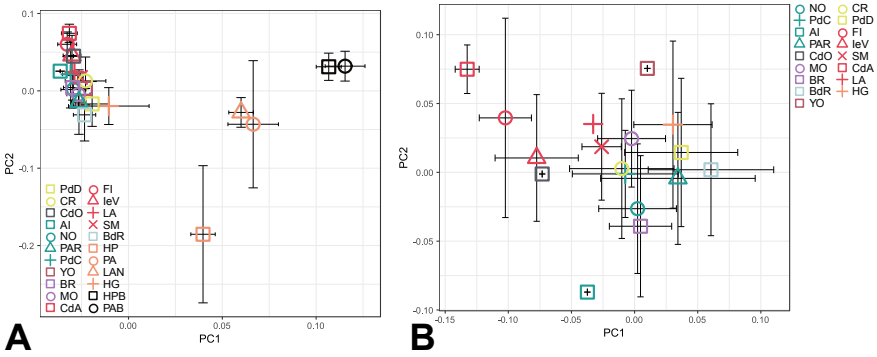
Supplementary Data



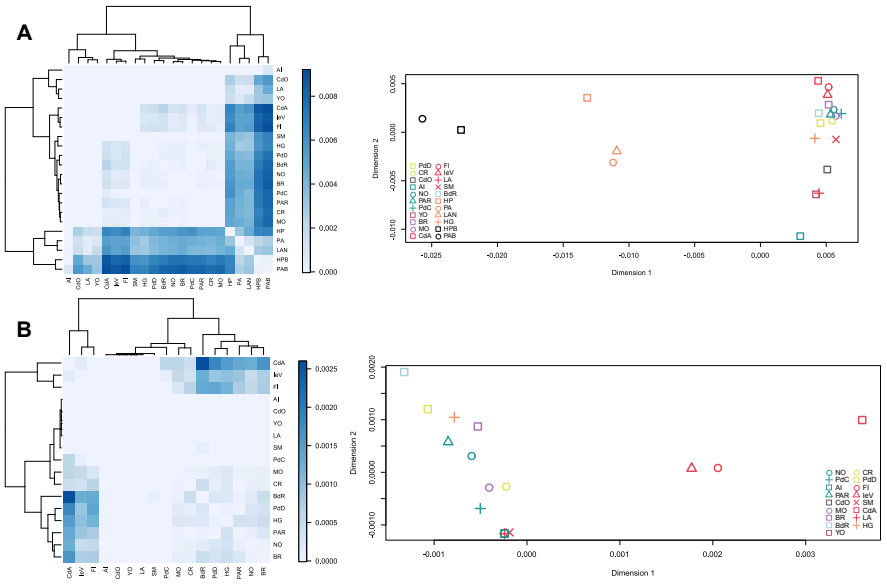
Supplementary Figure 1. PCA with 415 French samples highlighted the presence of outliers clearly skewing the global distribution of the samples (**A**). We assessed the origin of those samples using ChromoPainter and fineSTRUCTURE in the context of external references from three worldwide populations (CEU, YRI, CHB) from the 1000 genomes project (1000 Genomes Project Consortium *et al.*, 2015) and North African samples from published data (Lazaridis, Nadel, Rollefson, C.Merrett, *et al.*, 2016) . Four clusters were defined (**B**), assigning the majority of our samples (395) to the European cluster. The remaining 20 were outliers mainly belonging to the North African cluster (16 samples), 2 samples each were instead assigned to the Asian and the Sub-Saharan African clusters.



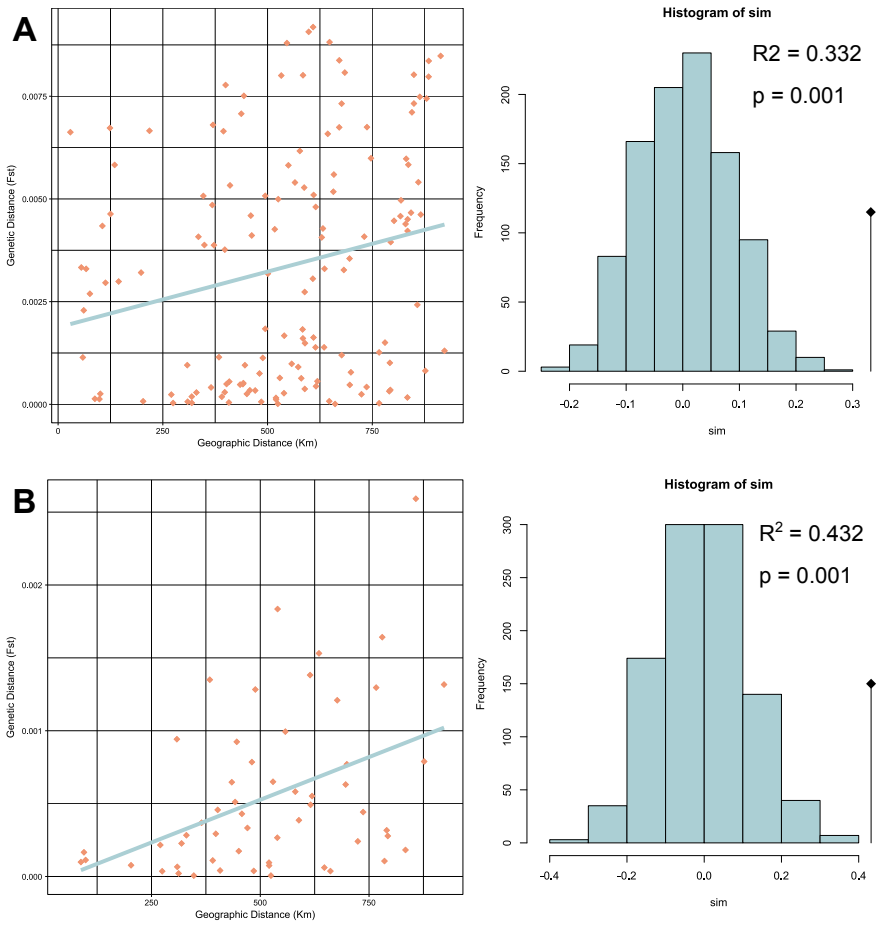
Supplementary Figure 2. Higher hierarchical levels used in the AMOVA analysis for **A)** Regions and **B)** Areas. Grey vertical lines highlight unsampled zones. Acronyms for the Areas are: NW, Northwest; N, North; NE, Northeast; W, West; CN, Central North; E, East; C, Center; SW, Southwest; SE, Southeast.



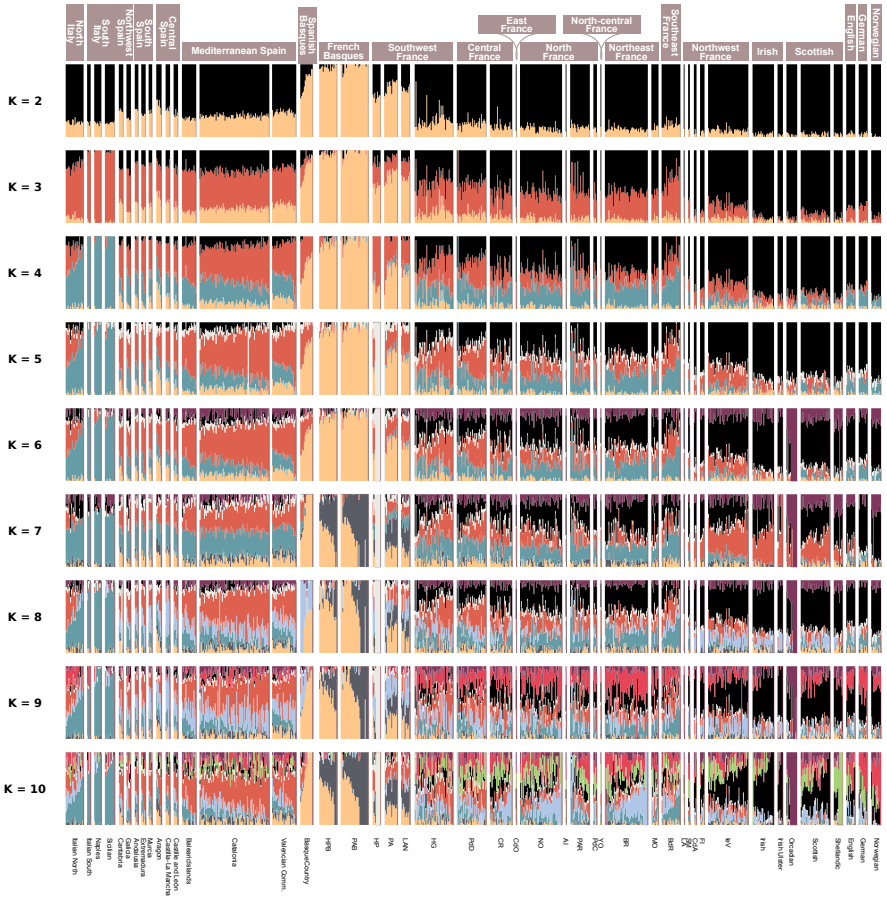
Supplementary Figure 3. Averaged Principal Component Analysis with **A)** Franco-Cantabrian samples, and **B)** without them. Color and symbol codes are the same as in main Figure 2. For each group, each averaged eigenvalue is represented along with standard deviation bars for the two PCs.



Supplementary Figure 4. On the left: heatmap and dendrogram based on F_{ST} matrices **A)** with the Franco-Cantabrian samples and **B)** without them. On the right: Multidimensional scaling (MDS) based on F_{ST} values **A)** with the Franco-Cantabrian samples ans **B)** without them.



Supplementary Figure 5. Mantel test of isolation by distance between the genetic (F_{ST}) and geographic (in Km) distances **A)** with the Franco-Cantabrian samples and **B)** without them. R^2 scores and p-values are within each figure.



Supplementary Figure 6. ADMIXTURE results from K=2 to K=10 for the 395 French samples divided in nine major groups, and 12 groups representing external sources from surrounding countries.

Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
BdR	Norwegian	Mediterranean_Spain	-0.000309	-2.926
BdR	Central_Spain	German	-0.00044	-2.893
BdR	Norwegian	Central_Spain	-0.000382	-2.661
BdR	German	Italian_North	-0.000299	-2.121
BdR	South_Spain	German	-0.00031	-2.075
BdR	Italian_North	Basque_Country	-0.000271	-1.991
BdR	German	Northwestern_Spain	-0.000279	-1.736
BdR	English	Italian_North	-0.000261	-1.702
BdR	German	Mediterranean_Spain	-0.000192	-1.692
BdR	South_Spain	English	-0.00026	-1.686
BdR	Northwestern_Spain	Scottish	-0.000211	-1.667
BdR	Central_Spain	English	-0.000254	-1.646
BdR	Italian_South	Mediterranean_Spain	-0.000151	-1.637
BdR	Mediterranean_Spain	Scottish	-0.000134	-1.563
BdR	Central_Spain	Scottish	-0.000175	-1.507
BdR	Central_Spain	Irish	-0.000189	-1.496
BdR	German	Basque_Country	-0.000231	-1.403
BdR	English	Basque_Country	-0.000229	-1.368
BdR	Italian_South	Northwestern_Spain	-0.000173	-1.183
BdR	English	Northwestern_Spain	-0.000195	-1.179
BdR	Northwestern_Spain	Irish	-0.000155	-1.121
BdR	Mediterranean_Spain	Irish	-9.6E-05	-1.041
BdR	English	Mediterranean_Spain	-5.7E-05	-0.473
BdR	Basque_Country	Scottish	-5.2E-05	-0.433
BdR	Norwegian	Northwestern_Spain	-4.8E-05	-0.295
BdR	Central_Spain	Italian_South	-2.9E-05	-0.24
BdR	Basque_Country	Irish	-2.4E-05	-0.182
BdR	Italian_North	Northwestern_Spain	-2.1E-05	-0.145
BR	Norwegian	Central_Spain	-0.000323	-2.801
BR	Northwestern_Spain	Irish	-0.000261	-2.697
BR	South_Spain	German	-0.000309	-2.647
BR	English	Northwestern_Spain	-0.00021	-1.57
BR	German	Basque_Country	-0.000189	-1.512
BR	Central_Spain	Irish	-0.00012	-1.47
BR	Norwegian	Northwestern_Spain	-0.000163	-1.191
BR	South_Spain	English	-0.000129	-1.106
BR	Central_Spain	English	-9.5E-05	-0.849
BR	Central_Spain	Scottish	-4.8E-05	-0.658
BR	English	Mediterranean_Spain	-3.4E-05	-0.469
BR	English	Basque_Country	-5.7E-05	-0.437
CdA	Norwegian	Central_Spain	-0.000127	-0.215
CdA	Norwegian	Basque_Country	-5.6E-05	-0.095
CdA	Central_Spain	Irish	-1.1E-05	-0.019
CR	Norwegian	Northwestern_Spain	-0.000396	-2.575
CR	German	Northwestern_Spain	-0.000357	-2.352
CR	Italian_South	Basque_Country	-0.000164	-1.225
FI	Norwegian	Basque_Country	-0.000794	-1.983
FI	Central_Spain	Irish	-0.000619	-1.727
FI	Norwegian	Central_Spain	-0.000668	-1.71
FI	German	Irish	-0.000594	-1.603
FI	South_Spain	Irish	-0.000522	-1.42
FI	South_Spain	Norwegian	-0.000542	-1.362

Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
FI	Basque_Country	Irish	-0.000485	-1.311
FI	Central_Spain	German	-0.000458	-1.195
FI	Norwegian	Mediterranean_Spain	-0.000408	-1.111
FI	Italian_South	Irish	-0.000393	-1.092
FI	Mediterranean_Spain	Irish	-0.000339	-0.998
FI	Italian_North	Irish	-0.000346	-0.966
FI	Central_Spain	Scottish	-0.000322	-0.909
FI	German	Basque_Country	-0.000279	-0.714
FI	Northwestern_Spain	Irish	-0.000269	-0.714
FI	Basque_Country	Scottish	-0.00023	-0.636
FI	English	Irish	-0.000224	-0.597
FI	South_Spain	Scottish	-0.000209	-0.581
FI	Norwegian	Italian_North	-0.000203	-0.518
FI	German	Scottish	-0.000173	-0.48
FI	English	Basque_Country	-0.00019	-0.467
FI	Central_Spain	English	-0.000184	-0.465
FI	Norwegian	Irish	-0.000117	-0.31
FI	Mediterranean_Spain	Scottish	-9.5E-05	-0.284
FI	South_Spain	German	-4.5E-05	-0.115
FI	Northwestern_Spain	Scottish	-4.3E-05	-0.114
FI	German	Mediterranean_Spain	-2.1E-05	-0.057
FI	Norwegian	Northwestern_Spain	-1.4E-05	-0.035
HG	German	Italian_South	-0.000338	-2.934
HG	English	Italian_North	-0.000305	-2.87
HG	German	Northwestern_Spain	-0.000349	-2.545
HG	German	Italian_North	-0.000193	-1.819
HG	Norwegian	Northwestern_Spain	-0.000118	-0.866
IeV	Northwestern_Spain	Irish	-0.000275	-2.923
IeV	Basque_Country	Irish	-0.000237	-2.676
IeV	Italian_North	Scottish	-0.000184	-2.61
IeV	South_Spain	Scottish	-0.000202	-2.559
IeV	Norwegian	Basque_Country	-0.00032	-2.537
IeV	Mediterranean_Spain	Scottish	-0.000105	-2.101
IeV	Italian_South	Scottish	-0.000149	-2.008
IeV	Central_Spain	Scottish	-0.00014	-1.952
IeV	South_Spain	Norwegian	-0.000217	-1.734
IeV	Northwestern_Spain	Scottish	-0.000141	-1.654
IeV	Norwegian	Central_Spain	-0.000168	-1.503
IeV	Norwegian	Mediterranean_Spain	-0.0001	-1.339
IeV	Norwegian	Italian_North	-0.000126	-1.163
IeV	Basque_Country	Scottish	-7.4E-05	-0.946
IeV	English	Basque_Country	-0.000111	-0.848
IeV	Central_Spain	English	-7.8E-05	-0.669
MO	English	Italian_South	-0.00073	-2.858
MO	South_Spain	Irish	-0.000653	-2.783
MO	Central_Spain	German	-0.000685	-2.737
MO	German	Mediterranean_Spain	-0.000588	-2.724
MO	Italian_North	Scottish	-0.000588	-2.661
MO	Italian_North	Irish	-0.000616	-2.658
MO	German	Northwestern_Spain	-0.000643	-2.594
MO	Norwegian	Central_Spain	-0.000642	-2.571
MO	South_Spain	English	-0.000622	-2.372
MO	German	Basque_Country	-0.000627	-2.363

Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
MO	German	Italian_North	-0.00051	-2.134
MO	Northwestern_Spain	Scottish	-0.000443	-1.987
MO	Mediterranean_Spain	Scottish	-0.000399	-1.977
MO	English	Basque_Country	-0.000483	-1.725
MO	Norwegian	Northwestern_Spain	-0.000426	-1.609
MO	English	Northwestern_Spain	-0.000417	-1.561
MO	Central_Spain	English	-0.000358	-1.353
MO	English	Mediterranean_Spain	-0.000311	-1.333
MO	Basque_Country	Scottish	-0.000316	-1.312
MO	Mediterranean_Spain	Irish	-0.000271	-1.309
MO	English	Italian_North	-0.000331	-1.286
MO	Central_Spain	Scottish	-0.000289	-1.266
MO	Northwestern_Spain	Irish	-0.000298	-1.253
MO	Central_Spain	Irish	-0.000214	-0.89
MO	Basque_Country	Irish	-0.000199	-0.803
MO	Italian_South	Basque_Country	-0.000146	-0.575
NO	German	Basque_Country	-0.000348	-2.806
NO	Basque_Country	Scottish	-0.000185	-2.294
NO	English	Northwestern_Spain	-0.000292	-2.234
NO	Norwegian	Northwestern_Spain	-0.000276	-2.088
NO	English	Basque_Country	-0.000269	-1.988
NO	Basque_Country	Irish	-0.000155	-1.74
PAR	Central_Spain	Irish	-0.000324	-2.734
PAR	Central_Spain	English	-0.000403	-2.731
PAR	Norwegian	Northwestern_Spain	-0.000438	-2.711
PAR	English	Basque_Country	-0.000251	-1.553
PAR	German	Basque_Country	-0.000219	-1.391
PAR	Basque_Country	Scottish	-7.8E-05	-0.672
PAR	Basque_Country	Irish	-3.3E-05	-0.25
PdC	Italian_South	Scottish	-0.001172	-2.752
PdC	Italian_North	Scottish	-0.001124	-2.712
PdC	Central_Spain	German	-0.001201	-2.709
PdC	Italian_South	Irish	-0.001116	-2.629
PdC	Italian_North	Irish	-0.001016	-2.429
PdC	Norwegian	Italian_North	-0.001026	-2.335
PdC	German	Italian_North	-0.000979	-2.205
PdC	German	Mediterranean_Spain	-0.000908	-2.141
PdC	Central_Spain	Scottish	-0.000872	-2.119
PdC	German	Northwestern_Spain	-0.000987	-2.091
PdC	German	Italian_South	-0.000946	-2.082
PdC	Mediterranean_Spain	Scottish	-0.000787	-1.973
PdC	Northwestern_Spain	Scottish	-0.000855	-1.958
PdC	German	Basque_Country	-0.000906	-1.932
PdC	Central_Spain	English	-0.000863	-1.928
PdC	Norwegian	Central_Spain	-0.000859	-1.923
PdC	Norwegian	Basque_Country	-0.000866	-1.839
PdC	English	Italian_North	-0.000789	-1.762
PdC	Norwegian	Mediterranean_Spain	-0.000742	-1.757
PdC	English	Italian_South	-0.000731	-1.604
PdC	English	Basque_Country	-0.000751	-1.594
PdC	Central_Spain	Irish	-0.00066	-1.588
PdC	English	Northwestern_Spain	-0.00075	-1.574
PdC	Basque_Country	Scottish	-0.000662	-1.536

Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
PdC	Norwegian	Italian_South	-0.000694	-1.519
PdC	English	Mediterranean_Spain	-0.000621	-1.461
PdC	South_Spain	Scottish	-0.00059	-1.384
PdC	South_Spain	German	-0.000624	-1.364
PdC	Mediterranean_Spain	Irish	-0.000523	-1.31
PdC	Northwestern_Spain	Irish	-0.000574	-1.297
PdC	South_Spain	Norwegian	-0.000566	-1.255
PdC	Norwegian	Northwestern_Spain	-0.000473	-1.002
PdC	Basque_Country	Irish	-0.000409	-0.932
PdC	South_Spain	Irish	-0.000395	-0.924
PdC	South_Spain	English	-0.000422	-0.923
PdC	German	Scottish	-0.00027	-0.611
PdC	German	Irish	-0.000183	-0.409
PdC	Italian_South	Basque_Country	-0.000111	-0.242
PdC	Italian_North	Basque_Country	-9.1E-05	-0.203
PdC	English	German	-2.6E-05	-0.055
PdD	Central_Spain	English	-0.000367	-2.93
PdD	Northwestern_Spain	Irish	-0.00031	-2.903
PdD	South_Spain	English	-0.000365	-2.857
PdD	English	Mediterranean_Spain	-0.000234	-2.813
PdD	Basque_Country	Irish	-0.000284	-2.751
PdD	English	Italian_North	-0.000321	-2.709
PdD	English	Northwestern_Spain	-0.000348	-2.523
PdD	Italian_North	Basque_Country	-0.000246	-2.207
PdD	Norwegian	Northwestern_Spain	-0.000179	-1.269
SM	Norwegian	Italian_North	-0.002166	-2.456
SM	Norwegian	Basque_Country	-0.002163	-2.408
SM	Norwegian	Mediterranean_Spain	-0.001996	-2.281
SM	Norwegian	Central_Spain	-0.002037	-2.279
SM	Italian_South	Irish	-0.00194	-2.263
SM	Norwegian	Italian_South	-0.001969	-2.202
SM	Italian_South	Scottish	-0.001848	-2.125
SM	South_Spain	Norwegian	-0.001844	-2.041
SM	Central_Spain	German	-0.00184	-2.039
SM	Italian_North	Irish	-0.001705	-1.986
SM	Italian_North	Scottish	-0.001665	-1.92
SM	German	Italian_South	-0.001682	-1.874
SM	German	Mediterranean_Spain	-0.001623	-1.844
SM	German	Basque_Country	-0.001663	-1.827
SM	German	Italian_North	-0.00158	-1.767
SM	Mediterranean_Spain	Scottish	-0.001441	-1.686
SM	English	Basque_Country	-0.001501	-1.664
SM	Central_Spain	English	-0.001492	-1.661
SM	Central_Spain	Scottish	-0.00145	-1.655
SM	English	Italian_South	-0.001456	-1.626
SM	Central_Spain	Irish	-0.001386	-1.584
SM	Mediterranean_Spain	Irish	-0.001325	-1.562
SM	Basque_Country	Scottish	-0.001362	-1.55
SM	English	Italian_North	-0.00138	-1.544
SM	English	Mediterranean_Spain	-0.001326	-1.512
SM	South_Spain	German	-0.001363	-1.507
SM	South_Spain	Scottish	-0.001271	-1.445
SM	Basque_Country	Irish	-0.001257	-1.44

Target	Source1	Source2	f_3(Target; Source1, Source2)	Z-score
SM	South_Spain	Irish	-0.001223	-1.406
SM	German	Irish	-0.001164	-1.325
SM	South_Spain	English	-0.001152	-1.279
SM	Norwegian	German	-0.001148	-1.253
SM	German	Scottish	-0.001104	-1.247
SM	Norwegian	Northwestern_Spain	-0.001061	-1.152
SM	German	Northwestern_Spain	-0.001037	-1.144
SM	Norwegian	English	-0.000996	-1.098
SM	English	German	-0.000909	-0.99
SM	Northwestern_Spain	Scottish	-0.000845	-0.96
SM	Norwegian	Scottish	-0.000822	-0.923
SM	English	Northwestern_Spain	-0.00079	-0.865
SM	English	Irish	-0.000722	-0.824
SM	English	Scottish	-0.000729	-0.823
SM	Northwestern_Spain	Irish	-0.000711	-0.808
SM	Italian_South	Basque_Country	-0.000709	-0.798
SM	Norwegian	Irish	-0.000681	-0.769
SM	Italian_North	Basque_Country	-0.000554	-0.621
SM	Scottish	Irish	-0.000211	-0.245
SM	Central_Spain	Italian_North	-0.000206	-0.23
SM	Italian_North	Mediterranean_Spain	-0.000183	-0.209

Supplementary Table 1. Values for all the Z-scores > -3 are reported for all negative values of the statistic f_3(Test; Source1, Source2). Test populations are the departments with at least 2 individuals, Sources are groups of external populations.

DISCUSSION

The Western Mediterranean: *eine Welt fur sich*

The Mediterranean is *eine Welt fur sich* (german expression for "a world unto itself"), and also *ein Welttheater* - a world-theater - in which every actor has a story that is hard to disconnect from everyone else's stories. Working with the Western Mediterranean populations means dealing with the continuum they represent. Even if we try to consider only a geographical part of it, or a specific historical moment we want to examine in depth, it is important to consider all the possible interactions with the surrounding landscape (Braudel, 1985).

Usually, a first Mediterranean identity is linked to the Neolithic (8,000-2,000 YBP), when the adoption of agricultural innovations allowed the establishment of first villages and the transition to a sedentary lifestyle (Braudel, 1985). In historical time, the Western Mediterranean has always acted as a ground of cultural and economic production; Greeks and Phoenicians settled trading colonies all around the basin, from the northern African shore up to the Italian, French, and Iberian coastlines. Together with Romans, Greeks and Phoenicians represented the main cultural, political, and economic powers in the Western Mediterranean basin. The Carthaginian and the Roman Empires contributed to the subjugation of most the cultures spread through the Western Mediterranean countries in North Africa, Spain, France (Celtic tribes), and Italy. Furthermore, the Arab expansion from the Arabic Peninsula

represented another strong influence on the Western Mediterranean basin during the 7th century CE (Ruiz, 2017).

Both prehistoric and historic events shaped the genetic variability of the present-day Western Mediterranean populations. Even if many genetic studies have shed light on different parts of the past relations and the processes of replacement and admixture of the different human groups, describing the complex variety of demographic processes that shaped the genetic structure of the Western Mediterranean populations, an overall image is still missing, possibly because of the lack of information about some parts of the big picture, and more insight is needed to overcome this unbalanced state.

In this chapter, I will discuss the results presented in the works on the island of Ibiza and France, highlighting the power of resolution different methods allow to achieve.

To avoid the ascertainment bias normally associated to SNP arrays designed for medical genetics, both these works have been carried out relying on the Affymetrix Human Origins Array which was specifically designed for population genetics studies. This array includes 13 different panels of SNPs ascertained in individuals of known ancestry from different populations.

Detecting the isolation of Ibiza

The variety of names Ibiza changed through time are a reflection of different moments of its history. It was *Ibossim* under the first Phoenician settlers in 654 BCE (Picornell *et al.*, 1996), *Ebusus* when the Carthaginians replaced the Phoenicians around the 550 BCE (Armstrong, 2004), inhabiting the island for about five centuries, then came *Yebisah* with the Umayyads of Cordoba that occupied the island around 902 CE, up to the Catalan *Eivissa* with the colonization occurred in 1235. With this extraordinary past, the outsider position Ibiza presented in different studies has been often interpreted as the result of the old Phoenician/Carthaginian legacy (Picornell *et al.*, 1996, 2005). However, a formal test has never been carried out.

In my work, I analysed the reasons behind the genetic diversity of Ibiza considering two different hypotheses: the possibility that the modern samples from Ibiza had something to share with the ancient Phoenician culture, and differentiation due to drift and the perpetuated practice of inbreeding.

To test the first hypothesis, I took advantage of the availability of an ancient sample, which proved to be an essential part of this work since it provided a direct information about the past of modern populations. More specifically, an ancient sample from the Phoenician Cas Molí site in Ibiza was used, and it did not appear to be an ancestral of modern Ibizans, thus rejecting previous

hypotheses often supported only by unprecise historical records. Moreover, the genetic discontinuity with modern Ibizans did not represented the only outcome of the work, but a major connection between the ancient individual and the modern samples from the old region of Canaan (Syria, Lebanon, Jordan, Israel) was also detected, possibly implying a signal of the continuity of modern Canaanites with the oldest Phoenician legacy (Biagini *et al.*, 2019).

On the other hand, the reduced genetic diversity, the extended identity by descent (IBD) sharing, and a higher number of runs of homozygosity (ROHs) pointed to the consequences of the effects of genetic drift (bottlenecks and founder effects), typical in environments that have experienced geographical and/or cultural isolation.

Runs of uninterrupted stretches of the genome (ROHs), inherited identically from both parents, are universally found with different length in human genomes even among outbred individuals (Ceballos et al.,2018). Compared to single-marker inbreeding coefficients, ROHs have a higher power of resolution in unveiling human demographic histories. Furthermore, the quality of ROH calling is influenced, among others, by the marker density, the quality of the genotype calling (especially the error rates), and the minor-allele frequencies (Ceballos *et al.*, 2018). Thus, compared to the whole-genome data, whose coverage, allele frequency, and the cost of a good sample size can be difficult to deal with, array data represent a better choice for this kind of analysis, since they provide

a high-density genome-wide scan data, low error rates, and allele frequencies greater than 5% (Ceballos *et al.*, 2018).

Long homozygous tracts of genotypes can be symptomatic of a more recent relatedness between subjects and are usually linked to inbreeding events. Compared to other Spanish groups, Ibiza presented the highest number of ROHs for the length category 1-2 Mb, while the rest of the groups did not show any outstanding value for this category (Biagini *et al.*, 2019). This result was consistent with the history of consanguineous unions on the island where marriages between relatives up to the 4th degree were very frequent, with records of unions between second-degree cousins dated back to the late 1960s (Claudio Alarco von Perfall, 1976). Furthermore, the analysis of IBD sharing allowed to retrieve information about the effective population size (N_e) evolution in the last 50 generations, identifying a collapse within the last 10-15 generations. This finding supports the genetic drift the island experienced, possibly corresponding to the bubonic plague occurred in 1652 (Biagini *et al.*, 2019).

Égalité: a matter of perspective

Overall, the existing literature about France tends to describe a homogeneous genetic landscape, mostly pointing to the main differentiation for the southwestern region, represented by the Basque-speaking groups from the Franco-Cantabrian area, and the north-western ones, represented by the region of Brittany. These

findings are consistent with the recognized distinct cultural entity these two groups represent; the genetic outlier position in the European landscape has been pointed out several times for the Basques (Rodríguez-Ezpeleta *et al.*, 2010), while the Celtic root of the Breton language has always been seen as the major connection between this area and the British Isles (Forster *et al.*, 2003).

According to mitochondrial DNA studies, the haplogroup composition of French people is homogeneous and not differentiated from the surrounding European countries, reflecting the usual variability of Western Europe (Dubut *et al.*, 2004; Richard *et al.*, 2007). In agreement with the homogeneous landscape defined by the mtDNA, studies on the Y-chromosome also described a lack of internal differentiation (Ramos-Luis *et al.*, 2009). Not so different is the panorama described by the only genome-wide study so far, that focused on the western regions, thus highlighting the only differentiation for the region of Brittany (Karakachoff *et al.*, 2015).

In the work on France I present in this thesis, I applied different methods of resolution in order to test whether the wide homogeneity described so far was the only possible landscape for the French territory, or whether relying on a higher resolution would have produced different results. Indeed, while with the allele frequency methods the only structure described was a general homogeneity with the only exception for the Franco-Cantabrian and the Breton

groups, the higher power of resolution reached with the haplotype-based methods allowed to unveil a more structured landscape.

Haplotypes are sets of SNPs in linkage disequilibrium and they allow a higher resolution compared to those methods that consider SNPs as separate units. This work is the first to explore the demographic landscape of France at the haplotype level, thus providing the first high-resolution genetic picture of this country. Furthermore, in the light of the f3-statistics analysis, a common background was found to all the tested groups, linking the modern French samples to two major sources of admixture that possibly date back to the first Celtic and Greek settlements. On the other hand, the internal pattern described with fineSTRUCTURE is a possible reflection of more recent events of differential gene flows circumscribed to specific areas (Biagini et al., in preparation).

However, a caveat about this work is related to the sampling and it needs to be discussed. Three different datasets have been joined in order to cover a wider area of France. The one made of samples from published data (Lazaridis *et al.*, 2016) presented inconsistencies between the location labels and the related geolocation information. To overcome this issue, the only two different possibilities were tested: trust the labels and change the coordinates or accept the provided coordinates and change the labels. The latter proved to be more reliable according to the pattern of haplotype distribution observed with fineSTRUCTURE. The matter raised by this situation is strongly connected to the need of a

good sharing of information in the scientific community, not only because of possible false positive results but even because the reliability of the data is fundamental for some analyses that rely on the geolocation information.

Future studies

In the wider context of the Western Mediterranean basin, different elements need to be taken into account for future studies. Today, the geographical coverage of the different countries facing into the Western Mediterranean basin is extensive, especially for genome-wide data, while whole-genome sequences have a scarcer distribution and an uneven sample size. However, based on genome-wide data, differences in SNP density, ascertainment methods applied, and the target of the different arrays (mostly designed for medical purposes) can affect the overlap between different datasets. The consequence is often represented by a severe reduction of the available variants with the following limitation of the analyses that can be performed. Indeed, haplotype-based methods need a higher density of variants in linkage disequilibrium in order to perform well, and an insufficient overlap between different dataset can affect the achievement of good results, or even prevent the application of such methods.

Haplotype-based methods have proved to be reliable tools for exploring the demographic history of different populations and, based on the fact that haplotypes are less affected by the fixation

index of derived alleles through drift, they proved to be more suitable for the study of populations that experienced recent demographic events such as founder effects and bottlenecks (Lawson *et al.*, 2012; Hellenthal *et al.*, 2014). This means that they can be valuable instruments for studying those populations that underwent processes of isolation.

Another limitation is represented by the lack of geolocation information for the samples from the majority of the public datasets. This represent an important matter that need to be solved since several tools that rely on the geographical information have been developed. For example, the EEMS software (Petkova *et al.*, 2016) allows to define patterns of migratory flows combining genetic data to geographical coordinates and its performance has a better resolution on a microscale level, meaning lower distances between the samples. Thus, the geolocation information should be provided not as an average value for entire groups, but preferably on a single sample level.

The density of the sampled areas is another important issue. Sometimes, there is the tendency to define an entire country based on a low representative sample size. It was, for example, the case of France, whose dataset so far was represented by few samples belonging to the same couples of areas, definitely insufficient to conduct demographic studies on the entire country, and also for contextualize France in the wider panorama of Europe or the Western Mediterranean basin.

In the context of the Western Mediterranean, today we can rely on a good sample size according to genome-wide data, even if the problem of the overlap between different arrays is still present. One of the natural consequences of my work will be related to build a comprehensive Western Mediterranean dataset combining samples from North Africa, Spain, France, and Italy. To overcome the overlap issue, there is a good representation of samples for each of those countries that have been genotyped using the Affymetrix Human Origins Array. Furthermore, it was my very personal interest to fill a gap in the genetic landscape of the southern Italian regions, represented by the lack of samples from Naples whose presence will certainly help to cover an important part of the demographic history of the Italian Peninsula, as also will be a valuable contribution to the wider study of the Western Mediterranean basin.

Finally, ancient DNA represents another fundamental resource. The Western Mediterranean has been vastly dominated by several civilizations that left different marks all over the countries facing into the basin. Providing more information on these past civilizations would enrich our knowledge about the different historical strata that characterized the complex history of this area. Today, an always higher number of ancient samples is offering the opportunity to explore our past, thus unveiling differential origins and patterns of ancient relations between populations. Therefore, it is also essential a good knowledge of the historical dynamics that defined the different moments of the past civilizations that lived

into the Western Mediterranean area. This is the reason I strongly support the importance of history as a complementary tool in genetics studies. As Fernand Braudel said, "*All history must be mobilized if one would understand the present*".

REFERENCES

Achilli, A. *et al.* (2004) 'The Molecular Dissection of mtDNA Haplogroup H Confirms That the Franco-Cantabrian Glacial Refuge Was a Major Source for the European Gene Pool', *The American Journal of Human Genetics*, 75(5), pp. 910–918. doi: 10.1086/425590.

Adams, S. M. *et al.* (2008) 'The Genetic Legacy of Religious Diversity and Intolerance: Paternal Lineages of Christians, Jews, and Muslims in the Iberian Peninsula', *American Journal of Human Genetics*, 83(6), pp. 725–736. doi: 10.1016/j.ajhg.2008.11.007.
Adovasio, J. M. and Hyland, D. C. (2000) 'The "Venus" Figurines', 41(4), pp. 511–537.

Aguirre, A., Vicario, A., Mazon, L. I., Estomba, A., Martínez de Pancorbo, M., Arrieta Picó, V., Pérez Elortondo, F. and Lostao, C. M. (1991) 'Are the Basques a Single and a Unique Population ?', *Am J Hum Genet*, 49(2), pp. 450–458.

Alexander, D. H., Novembre, J. and Lange, K. (2009) 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Research*, pp. 1655–1664. doi: 10.1101/gr.094052.109.vidual.

Arauna, L. R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhlaoui-Zid, K., Zalloua, P., Hellenthal, G. and Comas, D. (2017) 'Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa',

Molecular biology and evolution, 34(2), pp. 318–329. doi: 10.1093/molbev/msw218.

Arias, P. (1999) 'The Origins of the Neolithic Along the Atlantic Coast of Continental Europe: A Survey', *Journal of World Prehistory*, 13(4), pp. 403–464. Available at: d:%5CPromotion%5CCitavi%5CSicker Women at the Dawn of Agriculture%5CCitaviFiles%5CArias 1999 - The Origins of the Neolithic Along the Atlantic Coast of Continental Europe.pdf.

Armstrong, S. (2004) *The White Island*. Edited by Bantam.

Arredi, B., Poloni, E. S., Paracchini, S., Zerjal, T., Fathallah, D. M., Makrelouf, M., Pascali, V. L., Novelletto, A. and Tyler-Smith, C. (2004) 'A Predominantly Neolithic Origin for Y-Chromosomal DNA Variation in North Africa', *The American Journal of Human Genetics*, 75(2), pp. 338–345. doi: 10.1086/423147.

Barral-Arca, R., Pischedda, S., Gómez-Carballa, A., Pastoriza, A., Mosquera-Miguel, A., López-Soto, M., Martinón-Torres, F., Álvarez-Iglesias, V. and Salas, A. (2016) 'Meta-analysis of mitochondrial DNA variation in the Iberian Peninsula', *PLoS ONE*, 11(7), pp. 1–17. doi: 10.1371/journal.pone.0159735.

Bertranpetit, J. and Cavalli-Sforza, L. L. (1991) 'A genetic reconstruction of the history of the population of the Iberian Peninsula', *Annals of Human Genetics*, 55(1), pp. 51–67. doi:

10.1111/j.1469-1809.1991.tb00398.x.

Biagini, S. A., Solé-Morata, N., Matisoo-Smith, E., Zalloua, P., Comas, D. and Calafell, F. (2019) 'People from Ibiza: an unexpected isolate in the Western Mediterranean', *European Journal of Human Genetics*. Springer US. doi: 10.1038/s41431-019-0361-1.

Bittles, A. H. (2005) 'Endogamy, Consanguinity and Community Disease Profiles', (September), pp. 7–11. doi: 10.1159/000083332.

Boattini, A. *et al.* (2013) 'Uniparental Markers in Italy Reveal a Sex-Biased Genetic Structure and Different Historical Strata', *PLoS ONE*, 8(5). doi: 10.1371/journal.pone.0065441.

Bon, F. (1960) 'A brief overview of Aurignacian cultures in the context of the industries of the transition from the Middle to the Upper Paleolithic', (1913), pp. 133–144.

Bosch, E., Calafell, F., Comas, D., Oefner, P. J., Underhill, P. A. and Bertranpetit, J. (2001) 'High-Resolution Analysis of Human Y-Chromosome Variation Shows a Sharp Discontinuity and Limited Gene Flow between Northwestern Africa and the Iberian Peninsula', *The American Journal of Human Genetics*, 68(4), pp. 1019–1029. doi: 10.1086/319521.

Botigue, L. R., Henn, B. M., Gravel, S., Maples, B. K., Gignoux, C.

R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., Bertranpetit, J., Comas, D. and Bustamante, C. D. (2013) 'Gene flow from North Africa contributes to differential human genetic diversity in southern Europe', *Proceedings of the National Academy of Sciences*, 110(29), pp. 11791–11796. doi: 10.1073/pnas.1306223110.

Bourrillon, R. and White, R. (2015) 'Early Aurignacian Graphic Arts in the Vézère Valley: In Search of an Identity?', *P@lethnology*. doi: 10.4000/palethnologie.774.

Brandt, G., Szécsényi-Nagy, A., Roth, C., Alt, K. W. and Haak, W. (2015) 'Human paleogenetics of Europe - The known knowns and the known unknowns', *Journal of Human Evolution*, 79, pp. 73–92. doi: 10.1016/j.jhevol.2014.06.017.

Braudel, F. (1975) *The Mediterranean and the Mediterranean world in the age of Philip II: v. 1*. Edited by HarperCollins Distribution Services.

Braudel, F. (1985) *Une Leçon d'histoire*.

Browning, S. R., Browning, B. L., Daviglus, M. L., Durazo-arvizu, R. A., Schneiderman, N., Kaplan, R. C. and Laurie, C. C. (2018) 'Ancestry-specific recent effective population size in the Americas', *PLoS Genetics*, pp. 1–22.

Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C., Quintela-

García, I., Carracedo, Á., Donnelly, P. and Myers, S. (2019) ‘Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula’, *Nature Communications*, pp. 1–14. doi: 10.1101/250191.

Cambon-Thomsen, A. and Ohayon, E. (1988) ‘Practical Application of Population Genetics: The Genetic Survey “Provinces Françaises”’, in Mayr, W. (ed.) *Advances in Forensic Haemogenetics. Advances in Forensic Haemogenetics, vol2*. Berlin, Heidelberg: Springer.

Capocasa, M. *et al.* (2014) ‘Linguistic, geographic and genetic isolation: a collaborative study of Italian populations’, pp. 1–32. doi: 10.4436/JASS.92001.

Cavalli-Sforza, L. L. (1996) *Geni, popoli e lingue*. Edited by ADELPHI.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994) *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. and Wilson, J. F. (2018) ‘Runs of homozygosity: Windows into population history and trait architecture’, *Nature Reviews Genetics*. Nature Publishing Group, 19(4), pp. 220–234. doi: 10.1038/nrg.2017.109.

Cherni, L., Fernandes, V., Pereira, J. B., Costa, M. D., Goios, A., Frigi, S., Yacoubi-Loueslati, B., Amor, M. Ben, Slama, A., Amorim, A., El Gaaied, A. B. A. and Pereira, L. (2009) 'Post-last glacial maximum expansion from Iberia to North Africa revealed by fine characterization of mtDNA H haplogroup in Tunisia', *American Journal of Physical Anthropology*, 139(2), pp. 253–260. doi: 10.1002/ajpa.20979.

Claudio Alarco von Perfall (1976) 'Sobre los matrimonios consanguíneos en Ibiza', *Eivissa, num.8*, pp. 328–331.

Côrte-Real, H. B. S. M., Macaulay, V. A., Richards, M. B., Hariti, G., Issad, M. S., Cambon-Thomsen, A., Papiha, S., Bertranpetit, J. and Sykes, B. C. (1996) 'Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis', *Annals of Human Genetics*, 60(4), pp. 331–350. doi: 10.1111/j.1469-1809.1996.tb01196.x.

Denic, S., Nagelkerke, N. and Agarwal, M. M. (2011) 'On Some Novel Aspects of Consanguineous Marriages', pp. 162–168. doi: 10.1159/000321771.

Dubut, V., Chollet, L., Murail, P., Cartault, F., Béraud-Colomb, E., Serre, M. and Mogentale-Profizi, N. (2004) 'mtDNA polymorphisms in five French groups: Importance of regional sampling', *European Journal of Human Genetics*, 12(4), pp. 293–300. doi: 10.1038/sj.ejhg.5201145.

Fadhlaoui-Zid, K., Rodríguez-Botigué, L., Naoui, N., Benammar-Elgaaied, A., Calafell, F. and Comas, D. (2011) 'Mitochondrial DNA structure in North Africa reveals a genetic discontinuity in the Nile Valley', *American Journal of Physical Anthropology*, 145(1), pp. 107–117. doi: 10.1002/ajpa.21472.

Falcucci, A., Conard, N. J. and Peresani, M. (2017) *A critical assessment of the Protoaurignacian lithic technology at Fumane Cave and its implications for the definition of the earliest Aurignacian*.

De Fanti, S., Barbieri, C., Sarno, S., Sevini, F., Vianello, D., Tamm, E., Metspalu, E., Van Oven, M., Hübner, A., Sazzini, M., Franceschi, C., Pettener, D. and Luiselli, D. (2015) 'Fine dissection of human mitochondrial DNA haplogroup HV lineages reveals paleolithic signatures from European Glacial refugia', *PLoS ONE*, 10(12), pp. 1–19. doi: 10.1371/journal.pone.0144391.

Fiorito, G., Di Gaetano, C., Guarrera, S., Rosa, F., Feldman, M. W., Piazza, A. and Matullo, G. (2016) 'The Italian genome reflects the history of Europe and the Mediterranean basin', *European Journal of Human Genetics*. Nature Publishing Group, 24(7), pp. 1056–1062. doi: 10.1038/ejhg.2015.233.

Flores-bello, A., Mas-ponte, D., Rosu, M. E., Bosch, E., Calafell, F. and Comas, D. (2018) 'Sequence diversity of the Rh blood group system in Basques', *European Journal of Human Genetics*.

Springer US. doi: 10.1038/s41431-018-0232-1.

Flores, C., Maca-Meyer, N., González, A. M., Oefner, P. J., Shen, P., Pérez, J. A., Rojas, A., Larruga, J. M. and Underhill, P. A. (2004) 'Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: Implications for population demography', *European Journal of Human Genetics*, 12(10), pp. 855–863. doi: 10.1038/sj.ejhg.5201225.

Font-Porterias, N., Solé-Morata, N., Serra-Vidal, G., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., Calafell, F. and Comas, D. (2018) 'The genetic landscape of Mediterranean North African populations through complete mtDNA sequences', *Annals of Human Biology*. Informa UK Ltd., 45(1), pp. 98–104. doi: 10.1080/03014460.2017.1413133.

Forster, P. and Toth, A. (2003) 'Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European', *Proceedings of the National Academy of Sciences*, 100(15), pp. 9079–9084. doi: 10.1073/pnas.1331158100.

Frigi, S., Cherni, L., Fadhlaoui-Zid, K. and Benammar-Elgaaied, A. (2010) 'Ancient local evolution of African mtDNA haplogroups in Tunisian Berber populations.', *Human biology*, 82(4), pp. 367–384. doi: 10.3378/027.082.0402.

Di Gaetano, C., Fiorito, G., Ortu, M. F., Rosa, F., Guarrera, S.,

Pardini, B., CCusi, D., Frau, F., Barlassina, C., Troffa, C., Argiolas, G., Zaninello, R., Fresu, G., Glorioso, N., Piazza, A. and Matullo, G. (2014) 'Sardinians genetic background explained by runs of homozygosity and genomic regions under positive selection', *PLoS ONE*, 9(3), pp. 1–8. doi: 10.1371/journal.pone.0091237.

Di Gaetano, C., Voglino, F., Guarrera, S., Fiorito, G., Rosa, F., Di Blasio, A. M., Manzini, P., Dianzani, I., Betti, M., Cusi, D., Frau, F., Barlassina, C., Mirabelli, D., Magnani, C., Glorioso, N., Bonassi, S., Piazza, A. and Matullo, G. (2012) 'An Overview of the Genetic Structure within the Italian Population from Genome-Wide Data', *PLoS ONE*, 7(9). doi: 10.1371/journal.pone.0043759.

Gao, F. and Keinan, A. (2016) 'Explosive genetic evidence for explosive human population growth', *Current Opinion in Genetics & Development*. Elsevier Ltd, 41, pp. 130–139. doi: 10.1016/j.gde.2016.09.002.

Gardiner, R. M. (2001) 'The Human Genome Project: the next decade', *Genetics*, pp. 389–391.

Goldberg, A., Günther, T., Rosenberg, N. A. and Jakobsson, M. (2017) 'Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations', *Proceedings of the National Academy of Sciences*, 114(10), pp. 2657–2662. doi: 10.1073/pnas.1616392114.

Grimaldi, M.-C., Crouau-Roy, B., Amoros, J.-P., Cambon-Thomsen, A., Carcassi, C., Orru, S., Viader, C. and Contu, L. (2001) 'West Mediterranean islands (Corsica, Balearic islands, Sardinia) and the Basque population: contribution of HLA class I molecular markers to their evolutionary history', *Tissue Antigens*, (9), pp. 281–292.

Grugni, V., Raveane, A., Mattioli, F., Battaglia, V., Sala, C., Toniolo, D., Ferretti, L., Gardella, R., Achilli, A., Olivieri, A., Torroni, A., Passarino, G. and Semino, O. (2018) 'Reconstructing the genetic history of Italians: new insights from a male (Y-chromosome) perspective', *Annals of Human Biology*. Informa UK Ltd., 45(1), pp. 44–56. doi: 10.1080/03014460.2017.1409801.

Guillot, G., Leblois, R., Coulon, A. and Frants, A. C. (2009) 'Statistical methods in spatial genetics ´', *Molecular ecology*, pp. 4734–4756. doi: 10.1111/j.1365-294X.2009.04410.x.

Günther, T. *et al.* (2015) 'Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques', *PNAS*, 112(38), pp. 11917–11922. doi: 10.1073/pnas.1509851112.

Haak, W. *et al.* (2015) 'Massive migration from the steppe was a source for Indo-European languages in Europe', *Nature*, 522(7555), pp. 207–211. doi: 10.1038/nature14317.

Harris, E. E. (2017) 'Demic and cultural diffusion in prehistoric

Europe in the age of ancient genomes', *Evolutionary Anthropology*, 26(5), pp. 228–241. doi: 10.1002/evan.21545.

Hatzikotoulas, K., Gilly, A. and Zeggini, E. (2014) 'Using population isolates in genetic association studies', 13(5). doi: 10.1093/bfgp/elu022.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D. and Myers, S. (2014) 'A genetic atlas of human admixture history Garrett', *Science*, 343(6172), pp. 747–751. doi: 10.1126/science.1243518.A.

Henn, B. M., Botigué, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., Zalloua, P. A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C. D. and Comas, D. (2012) 'Genomic ancestry of North Africans supports back-to-Africa migrations', *PLoS Genetics*, 8(1). doi: 10.1371/journal.pgen.1002397.

Henn, B. M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S. and Bustamante, C. D. (2010) 'Fine-scale population structure and the era of next-generation sequencing', *Human Molecular Genetics*, 19(R2), pp. 221–226. doi: 10.1093/hmg/ddq403.

Hernández, C. L., Reales, G., Dugoujon, J. M., Novelletto, A., Rodríguez, J. N., Cuesta, P. and Calderón, R. (2014) 'Human maternal heritage in Andalusia (Spain): Its composition reveals high

internal complexity and distinctive influences of mtDNA haplogroups U6 and L in the western and eastern side of region', *BMC Genetics*, 15, pp. 1–16. doi: 10.1186/1471-2156-15-11.

Heutink, P. and Oostra, B. A. (2003) 'Gene finding in genetically isolated populations', 11(20), pp. 2507–2515.
Hoffecker, J. F. (2009) 'The spread of modern humans in Europe', 2009.

Hofmanová, Z. *et al.* (2016) 'Early farmers from across Europe directly descended from Neolithic Aegeans', *Proceedings of the National Academy of Sciences*, 113(25), pp. 6886–6891. doi: 10.1073/pnas.1523951113.

Husemann, M., Zachos, F. E., Paxton, R. J. and Habel, J. C. (2016) 'Effective population size in ecology and evolution', *Nature Publishing Group*. Nature Publishing Group, 117(4), pp. 191–192. doi: 10.1038/hdy.2016.75.

Jakkula, E., Rehnström, K., Varilo, T., Pietiläinen, O. P. H., Paunio, T., Pedersen, N. L., deFaire, U., Järvelin, M. R., Saharinen, J., Freimer, N., Ripatti, S., Purcell, S., Collins, A., Daly, M. J., Palotie, A. and Peltonen, L. (2008) 'The Genome-wide Patterns of Variation Expose Significant Substructure in a Founder Population', *American Journal of Human Genetics*, 83(6), pp. 787–794. doi: 10.1016/j.ajhg.2008.11.005.

Jakobsson, M. *et al.* (2008) ‘Genotype, haplotype and copy-number variation in worldwide human populations.’, *Nature*, 451(7181), pp. 998–1003. doi: 10.1038/nature06742.

Jobling, M., Hollox, E., Kivisild, T. and Tyler-Smith, C. (2014) *Human Evolutionary Genetics*. Garland Science.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. and Orlando, L. (2013) ‘MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters’, *Bioinformatics*, 29(13), pp. 1682–1684. doi: 10.1093/bioinformatics/btt193.

Kant, I. (1781) *Kritik der reinen Vernunft*.

Karakachoff, M. *et al.* (2015) ‘Fine-scale human genetic structure in Western France’, *European Journal of Human Genetics*, 23(6), pp. 831–836. doi: 10.1038/ejhg.2014.175.

Keller, A. *et al.* (2012) ‘New insights into the Tyrolean Iceman’s origin and phenotype as inferred by whole-genome sequencing’, *Nature Communications*. doi: 10.1038/ncomms1701.

Kherumian, R., Moullec, J. and Van Cong, N. (1967) ‘Groupes sanguins érythrocytaires A_1 , A_2 , BO, MN, Rh (CcDE) et sériques, Hp, Tf, Gm dans quatre régions militaires françaises.’, *Bulletins et Mémoires de la Société d’Anthropologie de Paris*, 1(XII), pp. 377–384. doi: DOI : https://doi.org/10.3406/bmsap.1967.1396.

Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M. and Wilson, J. F. (2010) 'Genomic runs of homozygosity record population history and consanguinity.', *PloS one*, 5(11), p. e13996. doi: 10.1371/journal.pone.0013996.

Klat, M. and Khudr, A. (1986) 'Religious Endogamy and Consanguinity in Marriage Patterns in Beirut, Lebanon', *Biodemography and Social Biology*, (December 2010), pp. 37–41. doi: 10.1080/19485565.1986.9988631.

Knapp, A. B. and Manning, S. W. (2016) 'Crisis in Context: The End of the Late Bronze Age in the Eastern Mediterranean'. doi: 10.3764/aja.120.1.0099.

Kulikowski, M. (2017) *L' età dell'oro dell'impero romano. Da Adriano a Costantino*. Newton Compton.

Kurki, M. I. *et al.* (2019) 'Contribution of rare and common variants to intellectual disability in a sub-isolate of Northern Finland', *Nature Communications*. Springer US, pp. 1–15. doi: 10.1038/s41467-018-08262-y.

Laframboise, T. (2009) 'SURVEY AND SUMMARY Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances', *Nucleic Acids Research*, 37(13), pp. 4181–4193. doi: 10.1093/nar/gkp552.

Latini, V., Sole, G., Doratiotto, S., Poddie, D., Memmi, M., Varesi, L., Vona, G., Cao, A. and Ristaldi, M. S. (2004) 'Genetic isolates in Corsica (France): linkage disequilibrium extension analysis on the Xq13 region', *European Journal of Human Genetics*, pp. 613–619. doi: 10.1038/sj.ejhg.5201205.

Latini, V., Vacca, L., Ristaldi, M. S., Marongiu, M. F., Memmì, M., Varesi, L. and Vona, G. (2003) 'ß-Globin Gene Cluster Haplotypes in the Corsican and Sardinian Populations', *Human Biology*, 75(6), pp. 855–871.

Lawson, D. J., Hellenthal, G., Myers, S. and Falush, D. (2012) 'Inference of Population Structure using Dense Haplotype Data', *PLoS Genetics*, 8(1), pp. 11–17. doi: 10.1371/journal.pgen.1002453.

Lazaridis, I. *et al.* (2014) 'Ancient human genomes suggest three ancestral populations for present-day Europeans', *Nature*, 513(7518), pp. 409–413. doi: 10.1038/nature13673.Ancient.

Lazaridis, I. *et al.* (2016) 'Genomic insights into the origin of farming in the ancient Near East', *Nature*. doi: 10.1038/nature19310.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp.

2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. and Myers, R. M. (2008) 'Worldwide human relationships inferred from genome-wide patterns of variation.', *Science*, 319(5866), pp. 1100–1104. doi: 10.1126/science.1153717.

Lindgreen, S. (2012) 'AdapterRemoval: easy cleaning of next-generation sequencing reads', *BMC Research Notes*.
Van De Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E. H., El Hajraoui, M. A., Amzazi, S., Hublin, J. J., Pääbo, S., Schiffels, S., Meyer, M., Haak, W., Jeong, C. and Krause, J. (2018) 'Pleistocene north african genomes link near eastern and sub-saharan african human populations', *Science*, 360(6388), pp. 548–552. doi: 10.1126/science.aar8380.

Maca-Meyer, N., González, A. M., Pestano, J., Flores, C., Larruga, J. M. and Cabrera, V. M. (2003) 'Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography', *BMC Genetics*, 4, pp. 1–11. doi: 10.1186/1471-2156-4-15.

Mellars, P. (2011) 'Palaeoanthropology: The earliest modern humans in Europe', *Nature*, 479(7374), pp. 483–485. doi: 10.1038/479483a.

Menozzi, P., Piazza, A. and L. Cavalli-Sforza, L. L. (1978) ‘Synthetic Maps ofHuman Gene Frequencies in Europeans’, *Science*, 201(September).

Mithen, S. (2006) *After the ice: a global human history, 20,000 - 5,000 BC*. Cambridge: Harvard University Press.

Moretti, E. and Cela, E. (2016) ‘A brief history of Mediterranean migration’, *Rivista Italiana di Economia Demografia e Statistica*, LXVIII(January 2014).

Myres, N. M. *et al.* (2011) ‘A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe’, *European Journal of Human Genetics*, 19(1), pp. 95–101. doi: 10.1038/ejhg.2010.146.

Omrak, A., Günther, T., Valdiosera, C., Svensson, E. M., Malmström, H., Kiesewetter, H., Aylward, W., Storå, J., Jakobsson, M. and Götherström, A. (2016) ‘Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool’, *Current Biology*, 26(2), pp. 270–275. doi: 10.1016/j.cub.2015.12.019.

Orlando, L., Gilbert, M. T. P. and Willerslev, E. (2015) ‘Reconstructing ancient genomes and epigenomes’, *Nature Publishing Group*. Nature Publishing Group, (June). doi: 10.1038/nrg3935.

Pereira, L., Franco-duarte, R., Fernandes, V., Pereira, J. B., Costa, M. D. and Macaulay, V. (2010) 'Population expansion in the North African Late Pleistocene signalled by mitochondrial DNA haplogroup U6', *BMC Evolutionary Biology*, 10, p. 390. doi: 10.1186/1471-2148-10-390.

Peter, B. M. (2016) 'Admixture, Population Structure, and F - Statistics', *Genetics*, 202(April), pp. 1485–1501. doi: 10.1534/genetics.115.183913.

Petkova, D., Novembre, J. and Stephens, M. (2016) 'Visualizing spatial population structure with estimated effective migration surfaces', *Nature Genetics*, 48(1), pp. 94–100. doi: 10.1038/ng.3464.Visualizing.

Picornell, A., Ana, M., Castro, J. A., Ramon, M. M., Arya, R. and Crawfor, M. H. (1996) 'Genetic Variation in the Population of Ibiza (Spain): Genetic Structure, Geography, and Language', *Human Biology*.

Picornell, A., Castro, J. A. and Ramon, M. (1997) 'Genetics of the Chuetas (Majorcan Jews): a comparative study', *Human Biology*, (July).

Picornell, A., Gómez-Barbeito, L., Tomàs, C., Castro, J. A. and Ramon, M. M. (2005) 'Mitochondrial DNA HVRI variation in Balearic populations', *American Journal of Physical Anthropology*,

128(1), pp. 119–130. doi: 10.1002/ajpa.10423.

Prevost, P., Busson, M. and Marcelli-Barge, A. (1984) 'Distribution of HLA-A,B alleles in 13 panels of blood donors in France', *Tissue Antigens*, 23(5), pp. 301–307. doi: 10.1111/j.1399-0039.1984.tb00049.x.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics*.
Ramos-Luis, E., Blanco-Verea, A., Brión, M., Van Huffel, V., Carracedo, A. and Sánchez-Diz, P. (2009) 'Phylogeography of French male lineages', *Forensic Science International: Genetics Supplement Series*, 2(1), pp. 439–441. doi: 10.1016/j.fsigss.2009.09.026.

Richard, C. *et al.* (2007) 'An mtDNA perspective of French genetic variation', *Annals of Human Biology*, 34(1), pp. 68–79. doi: 10.1080/03014460601076098.

Richards, M., Côrte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H.-J. and Sykes, B. (1996) 'Paleolithic and Neolithic Lineages in the European Mitochondrial Gene Pool', *American Journal of Human Genetics*.

Robledo, R., Corrias, L., Bachis, V. and Puddu, N. (2012) 'Analysis of a Genetic Isolate: The Case of Carloforte (Italy)', (December).

doi: 10.3378/027.084.0602.

Rodríguez-Ezpeleta, N., Álvarez-Busto, J., Imaz, L., Regueiro, M., Azcárate, M. N., Bilbao, R., Iriondo, M., Gil, A., Estonba, A. and Aransay, A. M. (2010) 'High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations', *Human Genetics*, 128(1), pp. 113–117. doi: 10.1007/s00439-010-0833-4.
Ruiz, T. F. (2017) *The Western Mediterranean and the World - 400 CE to the Present*. Wiley-Blackwell.

Salas, A., Comas, D., Lareu, M. V., Bertranpetit, J. and Carracedo, A. (1998) 'mtDNA analysis of the Galician population: A genetic edge of European variation', *European Journal of Human Genetics*, 6(4), pp. 365–375. doi: 10.1038/sj.ejhg.5200202.

Sarno, S., Boattini, A., Carta, M., Ferri, G., Alù, M., Yao, D. Y., Ciani, G., Pettener, D. and Luiselli, D. (2014) 'An ancient Mediterranean melting pot: Investigating the uniparental genetic structure and population history of Sicily and Southern Italy', *PLoS ONE*, 9(4). doi: 10.1371/journal.pone.0096074.

Sarno, S., Tofanelli, S., Fanti, S. De, Quagliariello, A., Bortolini, E., Ferri, G., Anagnostou, P., Brisighelli, F., Capelli, C., Tagarelli, G., Sineo, L., Luiselli, D., Boattini, A. and Pettener, D. (2015) 'Shared language, diverging genetic histories: high-resolution analysis of Y-chromosome variability in Calabrian and Sicilian Arbereshe',

European Journal of Human Genetics. Nature Publishing Group, 24(4), pp. 600–606. doi: 10.1038/ejhg.2015.138.

Sazzini M., Sarno S., L. D. (2014) 'The Mediterranean Human Population: An Anthropological Genetics Perspective.', in *Goffredo S., Dubinsky Z. (eds) The Mediterranean Sea.* Springer, Dordrecht, p. pp 529-551.

Schaefer, N. K., Shapiro, B. and Green, R. (2016) 'Detecting hybridization using ancient DNA', *Molecular ecology*, pp. 2398–2412. doi: 10.1111/mec.13556.

Schaefer, N. K., Shapiro, B. and Green, R. E. (2017) 'AD-LIBS: inferring ancestry across hybrid genomes using low-coverage sequence data', *BMC Bioinformatics*. BMC Bioinformatics, pp. 1–22. doi: 10.1186/s12859-017-1613-0.

Schraiber, J. G. and Akey, J. M. (2015) 'Methods and models for unravelling human evolutionary history', *Nature Publishing Group*. Nature Publishing Group, (November). doi: 10.1038/nrg4005.

Secher, B., Fregel, R., Larruga, J. M., Cabrera, V. M., Endicott, P., Pestano, J. J. and González, A. M. (2014) 'The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents', *BMC Evolutionary Biology*, 14(1), pp. 1–17. doi: 10.1186/1471-2148-14-109.

Seielstad, M. T., Minch, E. and Cavalli-sforza, L. L. (1998) ‘Genetic evidence for a higher female migration rate in humans’, *Nature Genetics*, 20(november), pp. 278–280. doi: 10.1038/3088.

Sikora, M. *et al.* (2014) ‘Population Genomic Analysis of Ancient and Modern Genomes Yields New Insights into the Genetic Ancestry of the Tyrolean Iceman and the Genetic Structure of Europe’, *PLoS Genetics*. Edited by R. E. Green, 10(5), p. e1004353. doi: 10.1371/journal.pgen.1004353.

Simón, M., Díaz, N., Solórzano, E., Montiel, R., Francalacci, P. and Malgosa, A. (2017) ‘Dissecting mitochondrial dna variability of balearic populations from the bronze age to the current era’, *American Journal of Human Biology*, 29(1). doi: 10.1002/ajhb.22883.

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M. T. P., Götherström, A. and Jakobsson, M. (2012) ‘Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe’, *Science*, 466. doi: 10.1126/science.1216304.

Solé-Morata, N., García-Fernández, C., Urasin, V., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., Comas, D. and Calafell, F. (2017) ‘Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81)’, *Scientific Reports*, 7(1), pp. 1–11. doi: 10.1038/s41598-

017-16271-y.

Solé-Morata, N., Villaescusa, P., García-Fernández, C., Font-Porterias, N., Illescas, M. J., Valverde, L., Tassi, F., Ghirotto, S., Férec, C., Rouault, K., Jiménez-Moreno, S., Martínez-Jarreta, B., Pinheiro, M. F., Zarrabeitia, M. T., Carracedo, Á., de Pancorbo, M. M. and Calafell, F. (2017) 'Analysis of the R1b-DF27 haplogroup shows that a large fraction of Iberian Y-chromosome lineages originated recently in situ', *Scientific Reports*, 7(1), pp. 1–13. doi: 10.1038/s41598-017-07710-x.

Svoboda, J., Novák, M. and Sázelová, S. (2014) 'Early Gravettian Occupations At Dolní Věstonice – Pavlov. Comments on the Gravettian Origin', *Mikulov Anthropology Meeting*, 20(2013), pp. 73–78.

Tang, H., Peng, J., Wang, P. and Risch, N. J. (2005) 'Estimation of Individual Admixture: Analytical and Study Design Considerations', *Genetic Epidemiology*, 301, pp. 289–301. doi: 10.1002/gepi.20064.

Terrenato, L. (2007) *Popolazioni e diversità genetica*. Edited by IlMulino. Bologna.

Teyssandier, N. (2006) 'QUESTIONING THE FIRST AURIGNACIAN: MONO OR MULTI CULTURAL PHENOMENON DURING THE FORMATION OF THE UPPER

PALEOLITHIC IN CENTRAL EUROPE AND THE BALKANS', *Anthtopologie*, pp. 9–29.

Tofanelli, S., Taglioli, L., Varesi, L. and Paolia, G. (2004) 'Genetic History of the Population of Corsica (Western Mediterranean) as Inferred from Autosomal STR Analysis', *Human Biology*.

Tomàs, C., Jiménez, G., Picornell, A., Castro, J. A. and Ramon, M. M. (2006) 'Differential maternal and paternal contributions to the genetic pool of Ibiza Island, Balearic Archipelago', *American Journal of Physical Anthropology*, 129(2), pp. 268–278. doi: 10.1002/ajpa.20273.

Tresset, A. and Vigne, J. D. (2011) 'Last hunter-gatherers and first farmers of Europe', *Comptes Rendus - Biologies*. Academie des sciences, 334(3), pp. 182–189. doi: 10.1016/j.crvi.2010.12.010.

Valverde, L., Illescas, M. J., Villaescusa, P., Gotor, A. M., Garc'a, A., Cardoso, S., Algorta, J., Catarino, S., Rouault, K., Férec, C., Hardiman, O., Zarrabeitia, M., Jiménez, S., Pinheiro, M. F. tim., Jarreta, B. M., Olofsson, J., Morling, N. and De Pancorbo, M. M. (2016) 'New clues to the evolutionary history of the main European paternal lineage M269: Dissection of the Y-SNP S116 in Atlantic Europe and Iberia', *European Journal of Human Genetics*, 24(3), pp. 437–441. doi: 10.1038/ejhg.2015.114.

Villaescusa, P., Illescas, M. J., Valverde, L., Baeta, M., Nuñez, C.,

Martínez-Jarreta, B., Zarrabeitia, M. T., Calafell, F. and de Pancorbo, M. M. (2017) 'Characterization of the Iberian Y chromosome haplogroup R-DF27 in Northern Spain', *Forensic Science International: Genetics*. Elsevier Ireland Ltd, 27, pp. 142–148. doi: 10.1016/j.fsigen.2016.12.013.

Vona, G., Moral, P., Memmì, M., Ghiani, M. E. and Varesi, L. (2003) 'Genetic structure and affinities of the Corsican population (France): Classical genetic markers analysis', *American Journal of Human Biology*, 15(2), pp. 151–163. doi: 10.1002/ajhb.10133.

Wangkumhang, P. and Hellenthal, G. (2018) 'Statistical methods for detecting admixture', *Current Opinion in Genetics & Development*. Elsevier Ltd, 53, pp. 121–127. doi: 10.1016/j.gde.2018.08.002.

Zalloua, P. A. *et al.* (2008) 'Identifying Genetic Traces of Historical Expansions: Phoenician Footprints in the Mediterranean', *American Journal of Human Genetics*, 83(5), pp. 633–642. doi: 10.1016/j.ajhg.2008.10.012.

Zalloua, P., Collins, C. J., Gosling, A., Biagini, S. A., Costa, B., Kardailsky, O., Nigro, L., Khalil, W., Calafell, F. and Matisoo-Smith, E. (2018) 'Ancient DNA of Phoenician remains indicates discontinuity in the settlement history of Ibiza', *Scientific Reports*, 8(1), p. 17567. doi: 10.1038/s41598-018-35667-y.

Zeggini, E. (2014) ‘Using genetically isolated populations to understand the genomic basis of disease’, pp. 1–3. doi: 10.1186/s13073-014-0083-5.

APPENDIX

ANCIENT DNA OF PHOENICIAN REMAINS INDICATES DISCONTINUITY IN THE SETTLEMENT HISTORY OF IBIZA

Zalloua P, Collins CJ, Gosling A, Biagini SA, Costa B, Kardailsky O, Nigro L, Khalil W, Calafell F, Matisoo-Smith E

Scientific Reports, 2018

Zalloua P, Collins CJ, Gosling A, Biagini SA, Costa B, Kardailsky O, et al. Ancient DNA of Phoenician remains indicates discontinuity in the settlement history of Ibiza. Sci Rep. 2018 Dec 1;8(1). DOI: 10.1038/s41598-018-35667-y