Analysis of co-ancestry links in modern and ancient human populations

Manuel Ferrando-Bernal

TESI DOCTORAL UPF / 2021

DIRECTOR DE LA TESI

Dr. Carles Lalueza-Fox

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

upf. Universitat Pompeu Fabra Barcelona

Acknowledgments

Gracias de corazón, a todos los que han contribuido a esta tesis.

Abstract

The ancient DNA field has allowed the merging of different disciplines such as Biology, Archaeology and History. This approach has disclosed some important events on the evolution of our species. Population genomic methods developed and applied in the ancient DNA field can uncover significant genetic differences and transitions between ancient and present-day individuals;nevertheless, samples from the last centuries can be genetically similar to present-day people, which can complicate the applicability of these methods. In addition, very few genomes have been sequenced so far from recent, historical samples. In this thesis we have applied methods developed to identify “Identical by Descent” (IBD) genomic blocks that determine co-ancestry links among individuals. We have applied this approach to an isolated population from the west coast of the African continent, and also to a large dataset of modern Europeans and a 700-years old high coverage genome from the Iberian peninsula. Our results show that IBD numbers can be shaped by the effects of recent demographic history of human populations, especially in those with increased endogamy. IBD analysis can also uncover individual links that could not be exposed with other, traditional methods. Moreover, we have explored these genetic connections with methods from graph theory. We expect this work will help to develop new strategies in the future, once more genomes from recent, historical times are retrieved, thus effectively linking the past with the present.

Resumen

El campo del ADN antiguo ha permitido la unión de disciplinas tan diferentes como la biología, la arqueología y la historia. Esta aproximación ha revelado algunos eventos importantes sobre la evolución de nuestra especie. Métodos de genética de poblaciones que han sido desarrollados y aplicados en el campo del ADN antiguo han permitido descubrir diferencias genéticas importantes entre individuos actuales y pasados; sin embargo, genomas de personas que vivieron hace pocos siglos pueden ser genéticamente similares a personas actuales, lo que puede complicar la aplicabilidad de estos métodos. Además, muy pocos genomas pertenecientes a épocas históricas han sido secuenciados. En esta tesis hemos aplicado métodos desarrollados para identificar haplotipos idénticos por descendencia que determinan enlaces de parentesco lejano entre diferentes individuos. Hemos utilizado esta aproximación en una población aislada de la costa oeste del continente africano y también en amplia base de datos de Europeos actuales y un genoma de hace 700 años de una mujer que vivió en la península ibérica. Nuestros resultados muestran que estas conexiones genéticas son moldeadas por patrones demográficos recientes, especialmente aquellos que generan una gran endogamia. Estos análisis pueden determinar enlaces de parentesco lejano que difícilmente podrían detectarse con otros métodos más tradicionales. Además, hemos explorado estas conexiones mediante métodos basados en la teoría de grafos. Esperamos que este trabajo pueda ayudar a desarrollar futuras estrategias cuando nuevos genomas de épocas históricas sean secuenciados, ayudando a conectar el pasado con el presente.

Preface

The Ancient DNA revolution started three decades ago and has allowed us to explore some of the most important events of the evolution of our species. However, their application in historical times has been underrated compared to other periods such as the Mesolithic, the Neolithic or the Bronze Age. This limitation is partly attributable to our genomic proximity to recent historical ancestors (i.e., those from the last centuries) that may complicate their study with most widely used population genomic tools.

However, this gap will likely be reduced in the coming years. New algorithms will need to be explored if we want to overcome this difficulty. In this thesis, I applied a methodology to identify Identity by Descent genomic blocks among present-day and ancient individuals to explore its potentiality to establish co-ancestry links in future studies with genomically-close populations.

Index

1. INTRODUCTION

In the first section I will summarize the origin, the characteristics, the history and some of the most important applications of the ancient DNA field (aDNA)(section 1.1.); afterwards, I will briefly explain some of the most commonly used genomic tools applied to aDNA data (section 1.2.), next, I will give some examples of the difficulties of studying recent genetic affinities in rather homogeneous populations (section 1.3.) and finally, the main methods for identifying identity by descent (IBD) blocks and how they can help in order to establish co-ancestral relationships among present-day populations and between present and ancient individuals (section 1.4.).

1.1. The field of ancient DNA

The recovery and analysis of DNA obtained from postmortem organisms, sometimes from individuals who died thousands of years ago (y.a.) is known as aDNA.

1.1.1. Characteristics and history of the ancient DNA

After the death of an organism, all the biochemistry processes within the cell stop working, including the proteins in charge of the DNA damage repair. The main consequence of these destructive processes -including the intracellular release of enzymes and the action of bacteria involved in the decomposition of the body- lead to the degradation of the DNA, with the initial molecules being spliced in short fragments. This reduces the ability to retrieve long DNA fragments from the original molecules (Pääbo, S. 1989; Allentoft, M. E. et al 2012). Typically, in an aADN study, most of the genomic fragments recovered measure no more than a few hundred base pairs (Glocke, I. & Meyer, M. 2017; Higgins, D. et al 2015; Dabney, J. et al 2013), while the length of a fragment of DNA recovered from current

individuals may reach several kilo-base pairs.

The rate at which the DNA molecule is degraded correlates with some environmental characteristics. For example, acid pH or high temperature conditions tend to increase the DNA degradation after an individual's death (Higgins, D. et al 2015; Wade, L. 2015). Only in some favorable conditions, such as frozen or dried tissues shortly after the body's death, the DNA damage is reduced and can prevent its complete destruction. However, even under those conditions, oxidation or hydrolytic actions can destroy the remaining DNA, limiting the surviving time for aDNA to be around a million years after the death of the organism, according to some theoretical estimates. Is for that reason, that these studies claiming the recovery of aDNA from organisms that lived several millions y.a., like dinosaurs or some insects enclosed in amber, have to be carefully revised (Dabney, J. et al. 2013).

The sensitivity of DNA molecules to participate in these chemical reactions depends to a large extent on the environmental conditions of the sample. For example, due to the axial inclination of the Earth, there are differences in average ambient temperature throughout the world, with latitudes near the north and south poles being colder than latitudes near the equator (Kottek, M. et al 2006). Because DNA molecules are best preserved at cold temperatures, the likelihood of recovering aDNA in the northern and southern latitudes of the Earth (like in Europe) is greater than recovering DNAs from equatorial areas, such as equatorial Africa. This explains, in part, why the number of aDNA studies from ancient European samples greatly exceeds the number of aDNA studies from equatorial areas (see Figure 1) (Kistler, L. et al 2017).

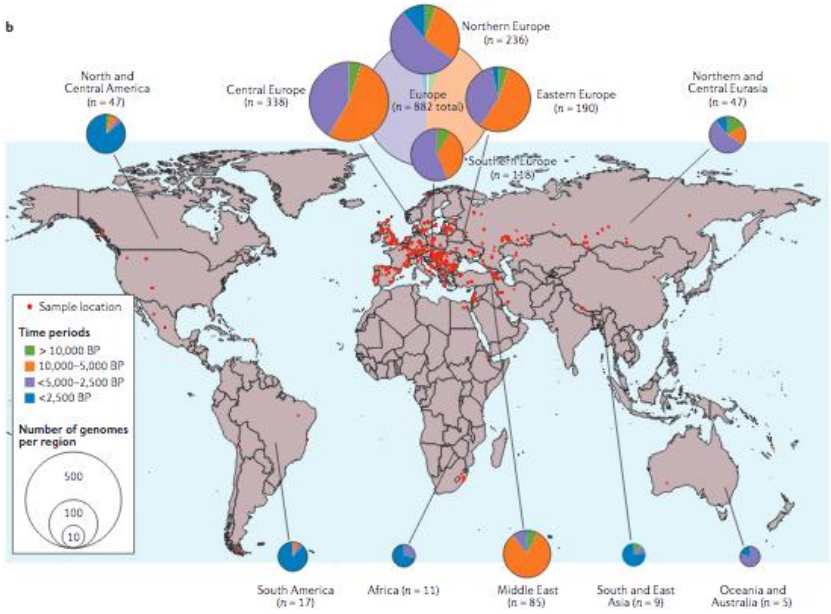


Figure 1. Geographical sites where ancient samples have been sequenced (the image dates from 2017). Bias toward a cooler environment is observed in aDNA preservation. This figure was taken from Marciniak, S. & Perry, G. H. (2017).

Apart from fragmentation of DNA molecules, there are other forms of damage (Dabney, J. et al 2013). These are classified according to the difficulties they cause during the sequencing process into miscoding and blocking lesions. Blocking lesions prevent polymerase-mediated synthesis and can manifest in two forms. The first is the modification of the original nucleotides. The best known example is the formation of hydantoins as a product of pyrimidine oxidation. Another example of nucleotide modification is the deamination of cytosine to uracil residues (Briggs, A.W. et al 2007). This usually leads to an increase in C to T substitutions, especially at the end of molecules, and has become one of the main characteristics for recognizing the authenticity of an aDNA sample. The second is caused by the link between the different DNA molecules between them or with other molecules. It has been suggested that these crosslinking between different strands of DNA may be due to furanones, furaldehyde or other products resulting from the biochemical process

following cell death. These bonds block access to the DNA molecule and therefore reduce its possibility to be sequenced.

Recovering the DNA of individuals who lived hundreds or thousands of years ago was once considered impossible because of how fast it degrades. Especially nuclear DNA, since this is made up of only two molecules of each chromosome in each cell. However, there are thousands of mitochondria in a single cell, each carrying its own DNA molecule, so it is more likely that some copies of mitochondrial DNA will survive the degradation processes longer. Accordingly, the first aDNA studies were able to recover mitochondrial DNA, and currently aDNA studies show a greater coverage of mitochondrial DNA than nuclear, confirming that the high number of mitochondrial DNA compared to nuclear in each cell facilitates its preservation.

The field of aDNA began in 1984 and 1985 with the publication of part of the genome of an extinct species, the quagga (Higuchi et al 1984) and part of a gene of a human Egyptian mummy (Pääblo, S. 1985) (which, in the end, turned out to be modern human pollution). As expected, the first studies recovered only a few short genomic molecules. Most belonged to mitochondrial DNA and sometimes to small parts of genes in the nuclear genome. With this scarce genomic material, most studies focused on characterizing the time of divergence between different clades (see Figure 2)(Higuchi, R. et al. 1984), or to identify functional mutations in specific genes, such as in the MCR1 gene in Neanderthals, which suggest a convergent evolution to clear skin in them and in modern Eurasia (Lalueza-Fox, C. et al 2007).

A B

Figure 2. A) The first cDNA study was able to recover part of the cytochrome oxidase I genome from an extinct animal, quagga. This gene is involved in the production of adenosine triphosphate (also known as ATP). B) The authors used mutations in this gene to calculate the time of divergence between this species and others. The figure was taken from Higuchi, R. et al (1984).

With the development of Second Generation Sequencing (SGS), it is possible to increase the sequencing of the nuclear and mitochondrial genome in current organisms and those that lived long ago. Moreover, the price of sequencing a human genome has been greatly reduced over the past two decades, leading to an increase in the number of available genomes (see Figure 3). In fact, nuclear genomic data for more than 3,500 ancient human samples are published today (see https://reich.hms.harvard.edu/). Most of these studies have been recovered by a specific matrix of single nucleotide polymorphisms (SNP) (about 1,200,000 SNPs) (Lazaridis, I. et al 2014) in collaboration with Dr. David Reich's laboratory at Harvard University (USA), which makes it easy to compare data from different studies.

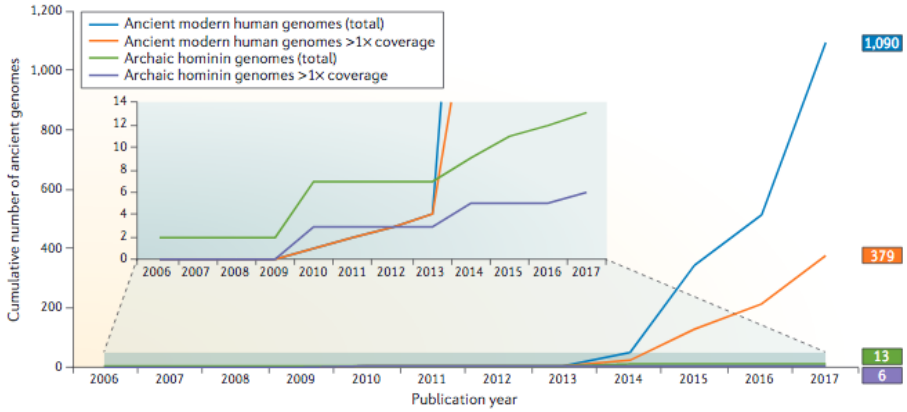


Figure 3. Number of aDNA samples sequenced has increased exponentially in the last few years. The data shown is until 2017. The Figure was taken from Marciniak, S. & Perry, G. H. (2017).

In a few cases, the endogenous DNA has been so well preserved, that it has been possible to recover the entire genome of the individuals. This process is carried out by a shotgun sequencing technology, where DNA fragments are sequenced and mapped several times. Whole-genome sequences are advantageous for many analyzes and provide new opportunities for the field. For example, it allows the ability to use new genomic tools, such as haplotype analysis, that show limitations when are used with low-density SNP matrices, or to further study new variants that are not included in standard SNP arrays. However, to date there are less than two tens of ancient samples where genomes have been fully sequenced at high coverage.

1.1.2 Discoveries of the aDNA field

With all of this aDNA data, scientists get deep into the study of *Homo sapiens* under population genomics, making it a promising field to understand some of the recent evolutionary adaptations and migratory

events of our species.

1.1.2.1 Past human adaptations

Some important discoveries have focused on the evolutionary adaptations of human populations, especially those of Europeans (probably due to the bias in the preservation of ancient individuals in Europe). One of the best examples is the determination of when Europeans developed light skin pigmentation. For decades, it was thought that it had happened for about 45,000 years, when the ancestors of Europeans reached the northern latitudes of Eurasia. However, aDNA studies show that early Europeans carried ancestral alleles related to skin pigmentation (so they probably had a darker skin color than current Europeans) and alleles related to the lighter skin pigmentation suffered positive selective pressure after the development of Agriculture (Olalde, I. et al 2014).

Another example is the evolution of lactose tolerance in European populations. This trait has evolved independently in different populations around the world. In Europeans, the mutation responsible for lactose tolerance prevents its inactivation after the weaning process, so adults can digest the lactose molecule. ADNA's studies of European individuals showed that this characteristic evolved, as expected, during the Neolithic period, when cattle were domesticated (Mathieson, I. et al. 2015).

Other aDNA's studies gave some light on other human adaptations that occurred more than 30,000 years ago. One of the most important examples of this field was the recovery of the DNA sequence of two extinct species of the genus *Homo*, *Homo neanderthalensis* (Green, R. et al 2007) and *Homo denisoviensis* (Meyer, M. et al 2012). Generally known as Neanderthals and Denisovans, respectively. These two species, native to Eurasia, exist for more than 700,000 years until they became extinct about 30,000 years ago. The comparisons of their genomes with those of

different populations of *Homo sapiens* showed that after the departure of the latter from Africa (for example, about 100,000 and 50,000 years a year), they intercrossed with the other two species (Reich, D. et al 2010). As a result of these interactions, all current populations, apart from South Africans, carry in their genomes some Neanderthal alleles and all Asian, Native American and Oceanic populations also show traces of denisovan alleles in their genomes (Sankararaman, S. et al 2014).

As these two species evolved in Eurasia, they were adapted to the different conditions of this supercontinent (for example, to local pathogens that were not in Africa). The interactions between these two species and the Homo sapiens groups facilitated the adaptation of the latter to new environmental conditions. For example, modern Europeans seem to be adapted to defend themselves against some viruses due to alleles inherited from Neanderthals (Enard, D. & Petrov, D.A. 2018). And this adaptive introgression of Neanderthal's alleles related to the immune system, could allow the *Homo sapiens* to expand through Eurasia in just a few thousand of years, eventually leading to the replacement and extinction of Neanderthals and Denisovans (Greenbaum, G. et al 2019).

1.1.2.2. Ancient demographic events of human populations

Most aDNA studies tend to focus on the migratory patterns and ancestry of the individuals analyzed. Due to the marked bias toward the conservation of ancient samples on the European continent, European populations and their migratory events are the most studied to date. However, the studies of aDNA have clarified part of the history of human colonization of the other continents. Here I will summarize the most important findings in each of them:

Africa:
Africa's geographical location makes it difficult to preserve cDNA because

of its high ambient temperature. In fact, to date, there are fewer than ten published genome-wide DNA studies of ancient African individuals (Vincent, M. & Schlebusch, C. M. 2020).

ADNA studies of individuals from North Africa showed that these populations have been in contact with Eurasian groups from the Middle East or the Iberian Peninsula since before the Holocene. On the other hand, the gene flow among North African populations with sub-Saharan populations was higher in the past than it has been in recent centuries. Since ancient individuals from Morocco show more affinities with southern African populations than present-day individuals. This different pattern is believed to be a consequence of the desertification of the Sahara over the past 5,000 years, which acted as a geographical barrier isolating human populations in North and South Africa. However, the aDNA of Egyptian mummies shows traces of genomic contact with South African populations during the Neolithic period, probably following the Nile route and influenced by the development of Egyptian culture (Vincent, M. & Schlebusch, C. M. 2020).

Studies of ancient DNA with individuals from South Africa indicate that before the invention of agriculture, hunter-gatherer groups were isolated from each other by distance. The independent development of agriculture in the Sahara/Sahel (about 7,000 y.a.), in the Ethiopian highlands (about 7,000-4,000 y.a.) and in West Africa (about 5,000-3,000 y.a.), was followed by human migration from these sites to the rest of the Africa continent (Vicente, M. & Schlebusch, C. M. 2020). During these migrations, first farmers interbreed with local groups of hunters-gatherers. One of the best examples of these migrations can be seen in the Bantu expansion. The Bantu people refer to the different Bantu language-speaking groups, and are descendants of a group of early West African farmers who expanded to East and South Africa after the development of agricultural practices (Patin, E. et al. 2017).

Asia:

There are about ten genome-wide DNA studies of samples from Eastern Eurasia (Zhang, M. & Fu, Q. 2020). Despite this small number, it has been possible to determine part of the complex genomic patterns in ancient and current Asian populations. For example, these analyzes enabled us to answer some intriguing questions, such as the date of division between European and Asian populations or the origins of Australasian populations.

As we saw, when the first *Homo sapiens* arrived in Asia they found the Neanderthals and the Denisovans. The sequences of some ancient remains of these species allowed them to determine that the current Asians carry traces of them in their genomes (Reich, D. et al 2010). As we saw with Europeans, this interaction with Neanderthals and Denisovans allowed them to adapt to the new conditions of the Asian continent. For example, current Tibetans may be able to survive low oxygen conditions at the high latitude of the Tibetan plateau due to the acquisition of an allele of the EPAS1 gene from the Denisovans (Huerta-Sanchez, E. et al 2014).

The oldest anatomically modern human sample sequenced corresponds to an individual known as Ust'-Ishim, who was found in Western Siberia (45,000 BC) (Fu, Q. et al 2014). Analyzes of its genome showed that it is equally related to East Asians as it was to the ancient European hunter-gatherers, suggesting that it belonged to the ancestral population of Europeans and Asians and that the division between these two populations was not older than 45,000 y.a. However, the genome of another individual found in a cave in Tianyuan, China, which dates back 40,000 y.a., showed that he is genetically closer to the populations from the East and South Asia, and North America, but not to Europeans, suggesting that the division between Asian populations and European populations occurred at some point between 45,000 y.a. and 40,000 y.a. (Zhang, M. & Fu, Q. 2020).

Other studies of aDNA showed that the former paleo-Siberian individuals are related to Native Americans (Sikora, M .et al 2019). In addition, paleo-Siberian ancestry is also found in some ancient Northeast Asians who suggest some connections between these two populations at least 8,000 y.a. (Zhang, M. & Fu, p. 2020)

With the development of agriculture, some human populations have spread from their local areas. For example, North East asian ancestry can be seen in some Southeast Asian populations, suggesting a possible route during the Neolithic period (Yang, M.A. et al 2020). Other neolithic movements have also been detected, for example, populations from Southeast Asia also moved to South Asia (about 4 k.y.a.) and Austronesia (about 3 k.y.a.), which later gave rise to most of today's oceanic populations.

Oceania:

Studies of aDNA from oceanic individuals show a bias toward Melanesia and Polynesia rather than mainland Australia, as there is only one ancient mainland Australian individual who has been sequenced (Rasmussen, M. et al. 2011). However, the sequence of this individual confirmed that the populations of Australia are the result of one of the oldest human migrations after the Out-of-Africa event. However, as this sample dates back to the last century, it cannot give much information about demographic patterns and human movements within Australia after its human settlement.

The rest of aDNAs' studies come from individuals from Polynesia and Melanesia who answered some questions that have intrigued archaeologists and linguists for decades. For example, the origin of the people of Polynesia, or the date of the first human migrations to the remote Oceania. These conclusions will be discussed in more detail in

section “1.3. Characteristics and difficulties in studying genetically homogeneous populations”.

America:

Ancient DNA studies with American individuals enabled us to answer some important questions about the origins, diversification, and cultural interactions among different human populations across the continent.

The most accepted hypothesis about the settlement of America says that the first Americans descend from the populations of Eastern Eurasia that moved toward the Bering land bridge to the Alaska peninsula during the LGM, and from there they moved towards the west and south.

Another example is the settlement on the Caribbean islands, which began about 8.000 years ago, making it one of the last places in America to be populated. Recently, analyzes of 93 former individuals (dating back to two of the most important historical periods in the Caribbean, the Archaic and Ceramics Ages) shed light on the history of the Caribbean's human population before the arrival of the first European colonies. Human populations arrived in the Caribbean in at least three different waves. One of them appears to have occurred before the division between the South American and Central American populations, and the last one corresponds to the human expansion during the Ceramics Era, originated in South America and probably reached the Caribbean through the Lesser Antilles (Nägele, K. et al 2020).

Other aDNA studies have been used to confirm human migrations suggested only from historical data. For example, in a particular case, researchers found that some ancient individuals from the Chincha Valley in southern Peru were genetically similar to former and current individuals from the northern coast of Peru (Bongers, J.L. et al 2020).These results are consistent with population movements during the Inca Empire,

something that was suggested by written sources from the Colonial Era. This is also supported by another study, based on the high genomic diversity of neighboring populations in the central areas of the Tiwanaku and Inca societies (Nakatsuka, N. et al 2020).

Europe:
Most of the ancient DNA samples published today come from the European continent, making the history of European settlement the best known to date.

As we have mentioned earlier, the first human groups populated the European continent shortly after the out of Africa event, inhabiting it with the Neanderthals for several thousand years. However, the initial proportion of Neanderthal ancestry has declined over time and current Europeans carry about 2% of it in their genome (Fu, Q. et al 2016).
.

For most of their history, Europeans had a lifestyle of hunter-gatherers. Different genomic studies confirm that the genetic flow between the different populations was mainly local, limited by European geography. For example, Jones, E.R. et al 2015, determined that hunter-gatherers from the Iberian Peninsula to Hungary separated from hunter-gatherer groups that inhabited the Caucasus shortly after their arrival in Europe (about 45,000 y.a.) and remained unrelated to each other for many thousands of years.

Despite this local restriction, there were some migratory events during paleolithic and mesolithic related to glacial events. During these periods, the ice expanded and was delayed, influencing the different ecosystems and consequently the lifestyle of the different human groups. As examples, 25.000 y.a., at the beginning of the LGM, the ancestors of the Anatolia populations separated from the hunter-gatherers of the Caucasus,

probably due to new geographical barriers (Jones, E.R. et al 2015). Or also, with the removal of the glacial ice sheet at the end of the LGM, about 11.000 y.a., humans were able to settle for the first time on the peninsula of Scandinavia (Mittnik, A. et al 2018).

Coinciding with the end of the LGM, some populations from the Middle East started to develop farming practices. Although some interlocal connections have been detected during mesolithic populations, the most important human migrations occurred during the Neolithic period as the first agricultural groups spread to Europe on different routes: From Anatolia they migrated west to Europe, from the Levant they moved east of Africa and from Iran they spread toward the Eurasian steppe (Lazaridis, I. et al 2017). ADNA's studies with early Neolithic individuals from different places of the European continent, showed that they were a mixture between two main populations: one from local European hunter-gatherers and the second derived from the first farmers of Anatolia. This results confirm that, during these expansions, the first farmer groups interbreed with local hunter-gatherer populations.

With the beginning of the Bronze Age, around 5.000 y.a., there was another important migration, related to the Yamnaya culture. They expanded across continental Europe from the Eurasian steppes (Haak, W. et al 2015). They were probably the descendants of the first farmers of Iran and the Caucasus, who had moved into the Euro-Asian steppes a few thousand years earlier (Jones, E.R. et al. 2015). Subsequently, several important cultures developed on the European continent, such as Bell Beaker ceramics or Corded Ware complex, both genetically close to the Yamnaya people (Haak, W. et al 2015; Olalde, I. et al 2018; Allentoft, M.E. et al 2015).

The increase in the European population and its subsequent migrations have greatly influenced the development and disappearance of new

cultures over the past 5 millennia. Archaeologists often found themselves wondering whether the appearance of similar cultural traits in different geographic cultures is the result of human migration or, instead, cultural diffusion. ADNA studies can be used to answer these questions, as they can determine genomic similarities between different populations. As an example, the expansion of the Bell Beaker culture arrived into Britain by human migration from the mainland Europe, as they were genetically similar to the Beaker individuals from the Lower Rhine region; meanwhile, it appears that the Bell Beaker culture reached the Iberian peninsula, mainly through cultural diffusion, since they were not genetically related to the Bell Beakers from Britain or Central Europe (Olalde, I. et al. 2018).

1.2. Some of the most commonly used tools for population genomics

Population genomics refers to the study of genomic differences between populations in order to understand the demographic, adaptive, phylogenomic, and ancestral patterns of a particular species.

The first methods were used to study a gene (or few genes) of particular interest in one or a few individuals. However, with the development of new sequencing and computer management technologies today, it is possible to study changes in thousands of genes from thousands of people at the same time. In order to analyze such a large amount of data, several softwares have been developed. A brief summary of some of the most commonly used in the field of the aDNA is listed below.

1.2.1. Principal Component Analysis

A database of hundreds or thousands of genomes can hide a lot of information. For example, if you make up individuals from different populations, there might be different patterns of fixed mutations. In order to

reduce the amount of information provided in these cases, the Principal Component Analysis (PCA) (a statistical method) tries to explain all the variance (all the different genomic patterns) in the given database, in only two or three factors, or components. The first component is the one that explains most of the variation in the database, the second component, the second that explains most of it, etc. The PCA has become one of the most used statistical methods in genomic population studies with human populations. For example, a PCA with current Europeans showed that their genomic patterns were closely correlated with geography, revealing that there is an important genomic substructure on the European continent (Lao, O. et al 2008; Novembre, J. et al. 2008)(see Figure 4).

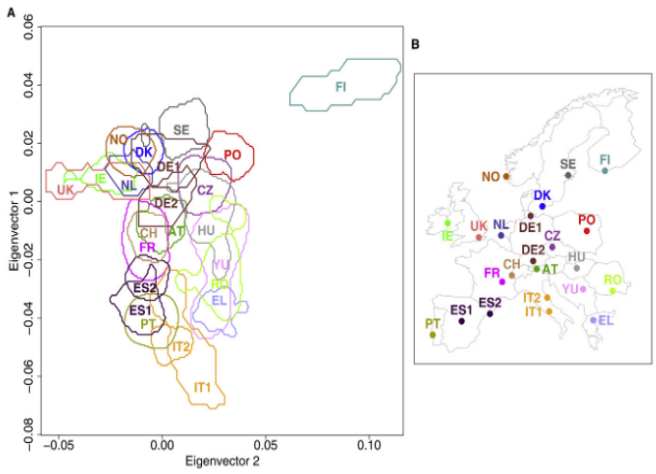


Figure 4. A) The Principal Component Analysis of 2,457 individuals from 23 European populations. B) The geographical position of these 23 populations. The image was taken from Lao, O. et al 2008.

1.2.2. ADMIXTURE analysis

ADMIXTURE is a software implemented to be used in population genomics studies to detect population structure. Specifically, it reconstructs the demographic history of individuals belonging to different

populations. It is able to detect differences in the ancestral composition between individuals in a given dataset.

The algorithms implemented in the ADMIXTURE program consider several scenarios where in each of them the data given can be explained by a particular number of ancestral populations. Each model explained by the ADMIXTURE that the individuals can be composed of different genetically different sources, up to K number. For example in Figure 5 the results indicate that the European population can be explained as a mixture of three different ancestral populations (k = 3), each of them represented by distinctive colors: i) In blue, the component of a native European population with a hunter-gatherer lifestyle, ii) in orange, the component of the first farmers entering Europe from the Anatolia peninsula and iii) in green, the component of a population related to the Yamnaya culture, which entered Europe about 5,000 y.a. (Lazaridis, I. et al 2014; Haak, W. et al 2015).

But how can we say which is the best scenario? The most used way to choose between the different scenarios follows was suggested by Alexander, D.H., et al. 2009. They propose a cross validation approach that runs the algorithm again with a particular K value and tests if the new results are consistent with the obtained previously. The scenario with the less variation in their repeated results is selected as the best one. In the case of Europeans, a K value of 3 shows less variations in the different runs.

However, this software is very sensitive to different parameters, for example, the results may vary depending on the size of the sample. It is important to note that although the software chooses the most parsimonious scenario capable of explaining the observed differences in the genomic composition of individuals. The complexity of the population relationships of our species makes the simplest scenario not need to be

exactly the real one. For example, ADMIXTURE assumes that each of the ancestral components is not the result of previous undetected admitted events.

Therefore, in the light of these limitations, it is important that the story explained by the results of the ADMIXTURE should be interpreted as a proxy of what already happened rather than taking it literally (for a tutorial on how to use and interpret the ADMIXTURE software go to Lawson, D.J. et al 2018).

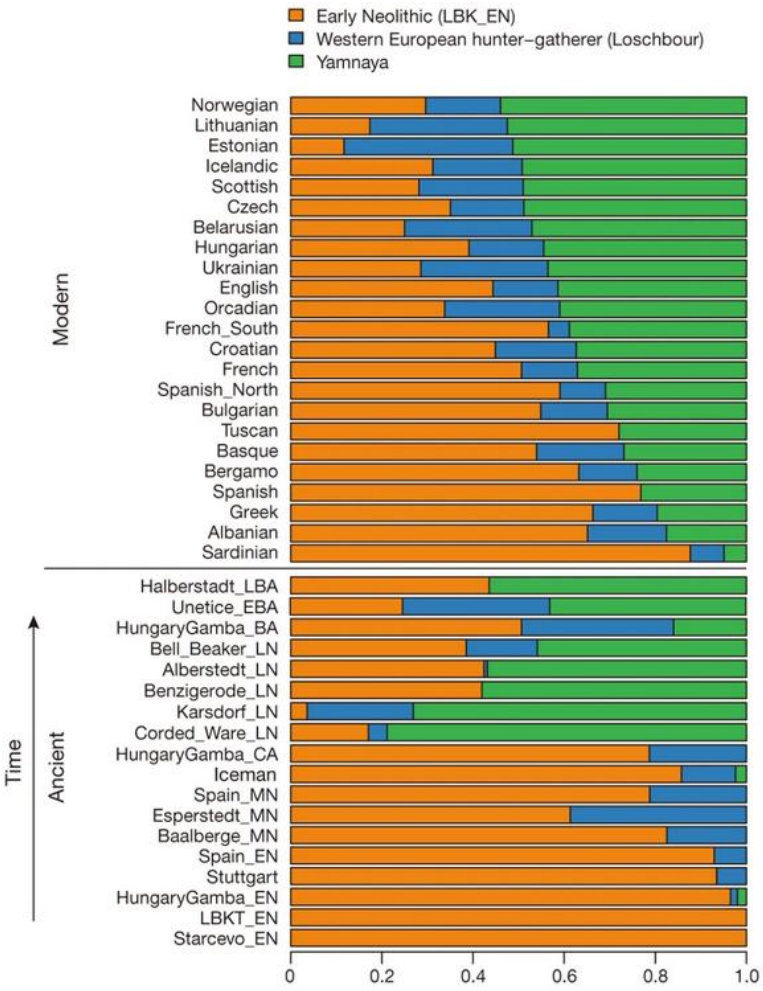


Figure 5. ADMIXTURE analysis with some current and old European individuals. This method has been able to explain the structure of the European population as a

combination of three mainly ancestral populations. The image was taken from Haak, W. et al 2015.

1.2.3. F3 and F4 statistics

F3 statistics have become a widely used genomic tool for studies with modern and old data since its publication in 2012 (Patterson, N. et al 2012). Its main use is to test the genomic relationships between three populations. This method is defined as the comparison of the differences between an out-of-group population (described as C) and population A and B (the focal population and the comparison population)):

F3=<(c-a)(c-b)>

Ultimately, it can be used in two scenarios: i) it can measure genetic drift between populations A and B compared to population C or ii) it can be used to check whether population C is a mixture of population A with population B. For example, In the first scenario, if the genetic deviation is statistically greater between population A and B than between these two with population C, it indicates that populations A and B have a recent common ancestor.

F4 statistics, introduced in Reich, D. et al 2009. In this case, the method compares the allelic differences between four populations and uses coalescing simulations to test the probability that the observed results can be explained by lineage sorting, or otherwise by introgression. With a model that calculates the differences in allele frequencies between population A and B, and C and D:

F4=(A-B)x(C-D)

In an incomplete lineage sorting scenario, the difference in allelic frequencies between population A and B should be independent of the differences in frequency between populations C and D. In this case the

values of the alleles frequencies of the different populations are not similar. However, this tends to result in similar values of the two calculations, leading to F4 value near zero.

Scenario 1 (no introgression, the values of the allele frequency of the populations are not similar):

(0.2-0.3)x(0.5-0.7)= -0.1 x -0.2 = 0.02

However, if there is an introgression from one population to another (such as from D into A), the values of the allele frequencies in these two populations would be similar, while the values in the calculations will not, leading the F4 values to be far from zero.

Scenario 2 (introgression from population D into A, their values are similar):
(0.7-0.3)x(0.5-0.7)= 0.4 x -0.3 = 0.12

1.2.4. Fst statistics

Pairwise Fst (Weir, B.S. & Cockerham, C.C. 1984) calculates the differences in populations based on allelic frequencies. High levels of FST indicate a large amount of genomic differentiation and vice versa. It is commonly applied in both the current and the old population. Because of positive selection actions on a particular set of genes, it is possible to use these statistics to detect selective events in particular populations: If the allele frequencies average indicate low FST levels, but particular genes show high levels, they may be caused by selective pressure in these regions. If applied to compare ancient and modern populations it allows to detect selective events occurring at some point between a determining past and the present.

All these genomic population tools work well in analyzes with heterogeneous populations. However, despite the power of these methods, their results are not accurate when analyzing populations with really recent common ancestors (as in homogeneous populations). In the next section I will summarize the main characteristics of homogeneous populations using two examples: present-day populations that were divided into the last millennia (such as the Polynesian people) and DNA analysis from individuals who lived only a few centuries ago.

1.3. Characteristics and difficulties studying genetically homogeneous populations

Homogeneous populations are populations whose genetic background is quite similar among them. When two or more populations are split over the last generations, it may not be enough time to accumulate genomic differences (fixed or near-fixed mutations) between them.

When population geneticists try to compare homogeneous populations, the most commonly used population genomic tools, such as PCA or MIXTURE analysis (Liu, CC. Et al 2020), are sometimes unable to differentiate between these populations based on their genomic characteristics. One of the most well-known examples of this are the Polynesian people.

The term Polynesia refers to one of the cultural regions of the Pacific Ocean (Jobling, M.A. et al 2014). It consists of some 1,000 islands spread over 30 million km2. The native inhabitants of this area are known as Polynesian, and can be classified into different populations. The human colonization of the Polynesian region is one of the last major expansions of our species (Jobling, M.A. et al 2014), starting about 3,000 years ago from the Taiwan peninsula (Chambers, G.K. 2013; Skoglund, P. et al. 2016). During the next two millennia, they moved from island to island. Reaching

places as remote as Papua New Guinea, the Hawaiian archipelago or the island of Rapanui, just 1,000 years ago (or even less). In other words, the different divisions among Polynesian populations occurred within the last 3 millennia (see Figure 6), and in some cases some populations departed from each other just in the last few centuries.

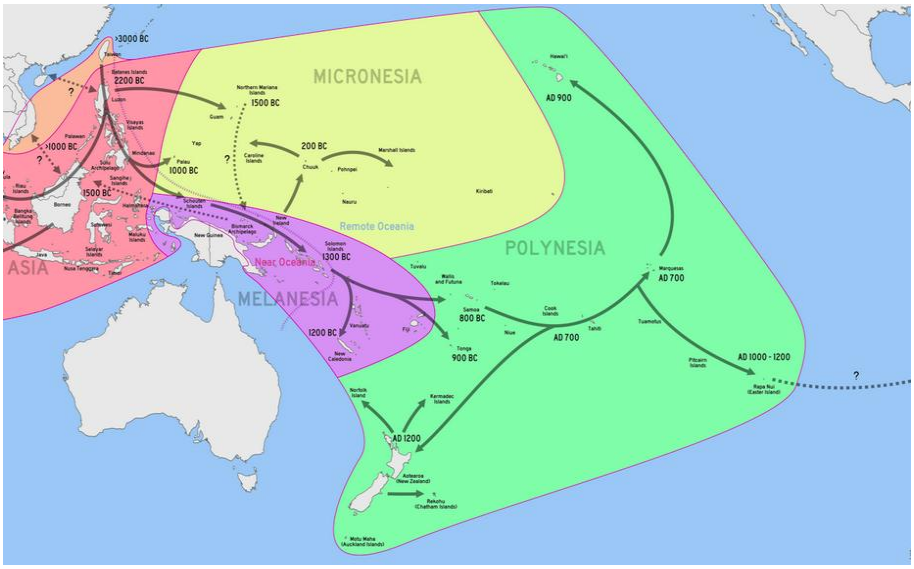


Figure 6. Map of the continent of Oceania. Polynesia is shown in green. The dates of colonization of the main islands are also shown. Most of the colonization of the Polynesian islands occurred in the last two millennia. Image of Chambers, G.K. 2013.

Genetic studies of Polynesian populations show high homogeneity among them (Hagelberg, E. et al 2008). This makes it difficult for some evolutionary questions to be resolved with the most common genomic tools. An example of this question is, what was the route that the different island-to-island populations followed across the Pacific Ocean? For example, the native Rapanui Island is expected to be genetically closer to the inhabitants of the original island from which it is sailed. However, due to the genomic homogeneity among Polynesian populations, the inhabitants of Rapanui Island, the New Guinea island, the Hawaiian

Archipelago, the Marquesas island, the Society Islands, etc. show similar patterns of genomic relationships between them all (see Figure 7).

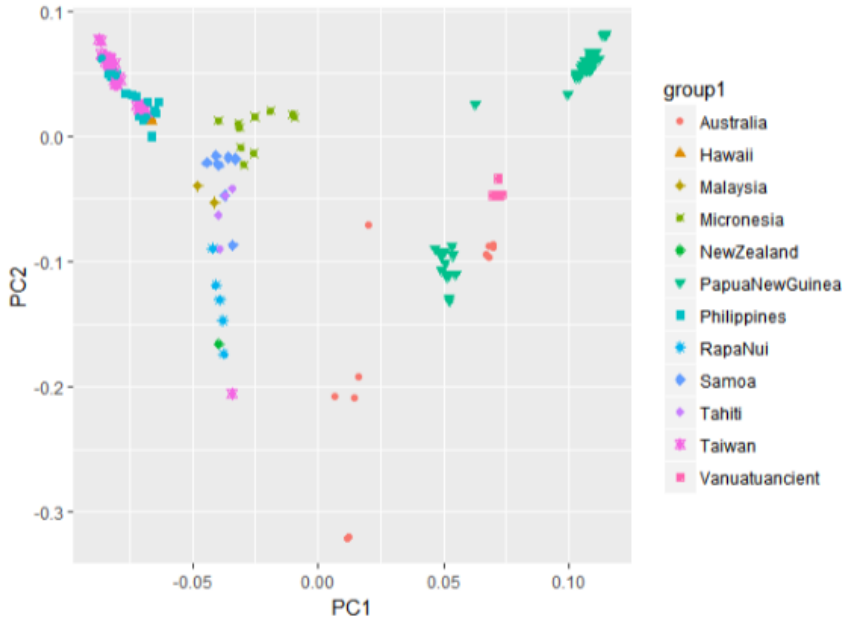


Figure 7. PCA analyses of the populations of Oceania and south Asia. Polynesian populations do not show a clear pattern. These analyses were done with the Oceanic populations from the Human Origins dataset.

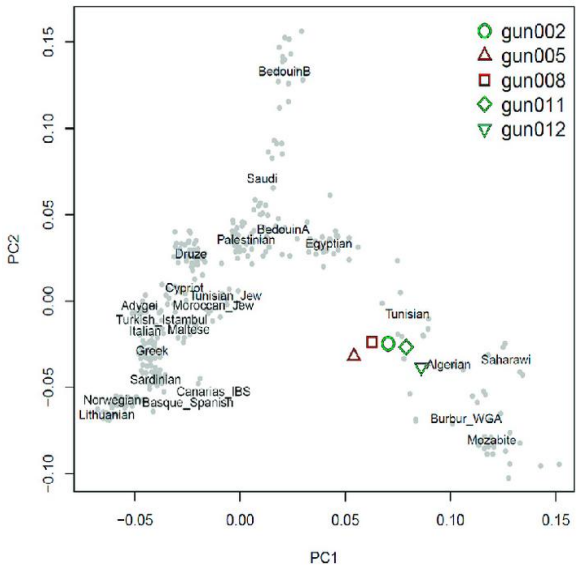
At present, scientists have tried to answer this question by comparing cultural, archaeological or linguistic similarities and differences between Polynesian populations, mainly because the evolution of these characteristics is generally faster than biological evolution.

Another example of this is the origin of the Guanche people. The Guanche people inhabited the Canary islands until the arrival of the first Europeans in the fifteenth century. The Guanches arrived in the Canary Islands about

1,500 years ago, so the division of this population with North Africa is relatively recent.

In Rodríguez Varela et. By 2017, they investigate their origins by recovering cDNA from some museum specimens and comparing them with current Europeans and populations in North Africa by using various genomic tools, such as PCA, f3, D statistics, analysis based on single-parent markers, ADMIXTURE analysis, etc.

In conclusion, the authors suggested that the Berbers are the most closely related population to the Guanches. However, the analyzes are not very accurate to determine which population is the one closest to the old specimens (the old specimens appear to be equally close to the Berbers of Algeria and Tunisia. See Figure 8). Some of the other analyzes are also inconsistent among them, for example, FST, single-parent markers, PCA, and ADMIXTURE point to one direction, while the D and f3 statistics point to another direction. It is quite possible that the main cause of this is that the short time from the division between these populations from North Africa to the Guanches is not enough to accumulate the need for genomic



variation between them to apply these genomic tools.

Figure 8. PCA analysis of five old samples (Guanches) and several modern populations from Europe and North Africa. The ancient individuals show similar patterns to several populations from North Africa. The image was taken from Rodríguez-Varela et. al. 2017.

Another example is the analysis of genomic ancestry by the French revolutionary Jean-Paul Marat (de-Dios, T. et al 2019). He was reading some newspapers at the time he was killed and his blood (from which his DNA has been obtained) stained them.

As Marat lived in the 18th century, its genomic background may be quite similar to today's Europeans. This is because in a period of three centuries there is not enough time to accumulate genomic differences. Assuming a human generational time of 25 years, in three centuries there must be 12 generations.

In the study, to explore Marat's ancestry, the authors applied a PCA and an ADMIXTURE analysis. Marat's historical records suggest that he had a combination of French and Italian ancestors. Although both the PCA and the ADMIXTURE analysis are consistent with it, their accuracy is a little low, and on the basis of these analyzes Marat could be: i) only French, ii) or equally related to French, Italian, Spanish, English , Romanian and Hungarian individuals (see Figure 9). Once again, it is quite possible that the difficulty in establishing Marat's ancestry may be due to the high genomic similarity between this individual, who lived 300 years ago, with the different European populations today.

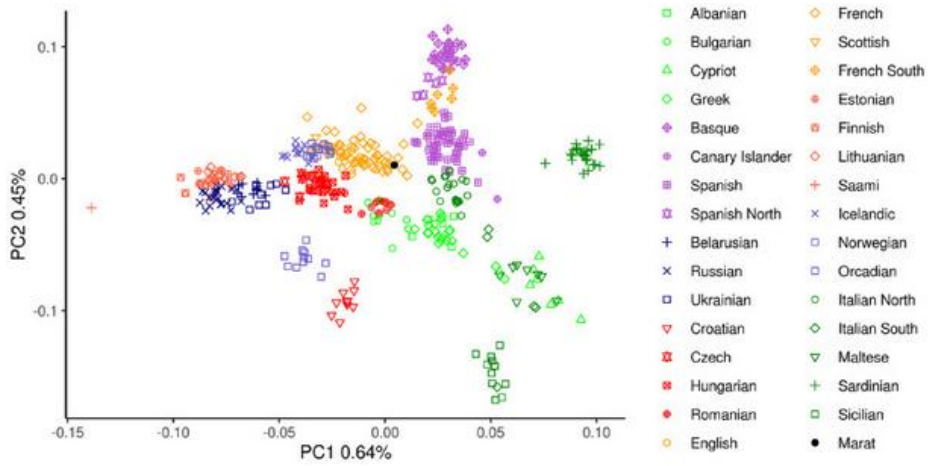


Figure 9. PCA analysis of the individual Jean-Paul Marat together with several current populations of the European continent. The old sample falls in the center of the graph, near many of the populations used in the analyzes. The image was taken from de-Dios, T. et al 2019).

The objective of this thesis will be to apply another methodology, based on shared haplotypes in two studies with homogeneous populations. The first, with modern individuals living on an island on the central west coast of Africa, and the second, comparing genomic ancestry between an individual who lived 700 years ago with current Europeans.

1.4. Identity by Descent fragments

To the date, genetic relationships between homogeneous populations have been based on analyses with genomic rare variants. A rare variant is an allele that is present in a given population in a minor allele frequency (MAF) up to 1%. Theoretically, these low frequency alleles are, on average, younger than high frequency ones, and then, can be used to detect recent shared ancestry. Schiffels, S. et al. (2016), use these approach to detect a recent migration of Anglo-Saxon into the British population (around 1.600 y.a.), and show that analyses based on rare variants can be more accurate than other methodologies (like PCA) when the two source population are genetically similar. To detect rare variants it is needed a large number of individuals with their whole genomes sequenced. However this is often methodologically impossible for ancient genomic studies. This is mainly due to (i) sometimes the DNA is not well preserved, so only a part of it can be recovered, and (ii) whole-genome sequencing is more expensive than SNP capture technologies and requires additional amounts of sample (Schiffels, S. et al. 2016).

The aim of this thesis is to develop an alternative to rare variants methodology that, despite is usually preferred to be applied on whole-genome, can also be used on densely-distributed SNP arrays. To do so, we applied Identity by Descent (IBD) analyses in present-day genetically close populations and between a “recent” ancient sample and modern Europeans. Instead of relying on single mutations that are in low frequency, IBD analyses are based on low frequent haplotypes in a given database, which can reflect, on average, relationships of recent co-ancestry between a pair of given individuals.

1.4.1. Characteristics of the Identity by Descent Fragments

When a genomic fragment is identical in all its base pairs in two or more individuals is called Identical by State (IBS). In some cases the same haplotype can be generated independently, for example if the same mutations originate in different individuals, or it can arise by recombination if the populations have little genetic variation. However, when an IBS is inherited in different individuals (at least in two) from the same genetic ancestor, it is said to be inherited “Identical by Descent” (IBD) (Figure 10). Those fragments are often referred to as IBD segments/fragments. The huge amount of genomic data available today allows us to explore genomic ancestry in different populations using IBDs analysis, but how can a genomic fragment be determined to be an IBD or not? In order to answer this question, we need to explore first what the IBD heritage is like.

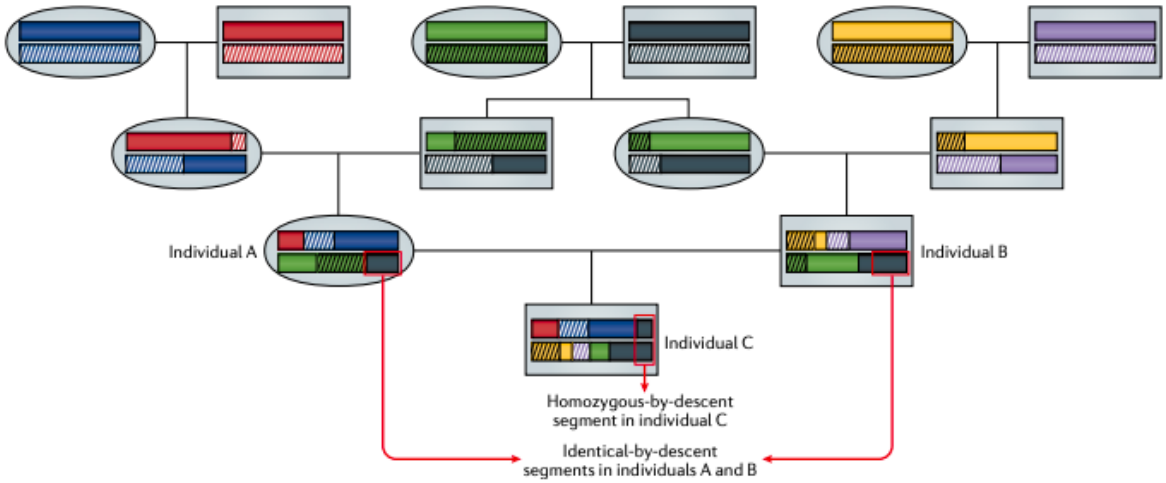


Figure 10. Schematic representation of the IBD sharing in two different individuals (Individuals A and B) which inherited it from a common ancestor. If the haplotype is inherited in the same individuals from both the maternal and paternal lineage, we call it a Run of Homozygosity (ROHs) (Individual C). Image taken from Racimo, F. et al. 2020.

1.4.1.1. Genomic Recombination

Sexual reproduction is widespread in nature. This type of reproductive mechanism is almost always linked to a process called recombination and in fact, it is difficult to understand one process without the other. Genetic recombination is a mechanism that combines parental genetic material that produces new haplotypes in offspring (Figure X). It needs the recombination of two homologous chromosomes, so it does not occur on the mitochondrial chromosome and only at the shared part of the sex chromosomes (as in the distal portion at the short arms of the X and Y chromosomes of mammals).

There are several hypotheses to explain the last causes of sexual reproduction and genetic recombination. The most accepted one says that this evolved as a mechanism that generates genetic variation, thus facilitating that a given population can increase the chances of adaptation to the new environmental variations. For example, it may be important for fighting parasites (see The Red Queen hypothesis for more).

It has also been suggested that recombination evolved because it avoids the accumulation of deleterious mutations in small populations. As mutations appear over generations, some lineages will begin to accumulate deleterious mutations as the number of generations increases. Recombination destroys haplotypes where some of these mutations may be linked, thus reducing the mutation load in new individuals (see Muller's ratchet hypothesis for more). Similarly, it would also allow several advantageous mutations to link together, thus increasing the adaptational fitness of new individuals.

Genetic recombination occurs mainly during meiosis, before the maturation of gametes. Therefore, sperm and eggs carry new haplotypes from the mixture between the parents. Genetic recombination is usually random. However, there are some regions of the genome, known as recombination hotspots, where the probability of a recombination event exceeds the average of the rest of the genome. More than 25,000 of these hotspots can be found in the human genome (Myers, S. et al 2005). In contrast, there are other regions in the genome where the recombination rate is very small, such as in the area of the centrosome or at the end of the chromosomes (Andy Choo, K.H. 1998).

1.4.1.2. The length and age of the IBD fragments

These differences in the likelihood of recombination of the different regions of the genome are used to determine the length of the IBD segments. Specifically, the length of an IBD is usually measured in centiMorgans (cM). The concept of cM is defined as the distance between two genetic positions for which the probability of being broken in a single recombination event (one generation) is 1%. However, this is sometimes difficult to measure. Therefore, and although due to these differences in recombination rates a cM is not equivalent to a real physical distance, some methods might consider, to simplify, that a cM is roughly equivalent to 1,000,000 base pairs.

Because IBD fragments break down due to recombination, large IBD fragments are generally inherited from a common ancestor (CA) that lived a few generations ago. By contrast, small fragments of IBD tend to come from CA that lived many generations ago. Some authors suggested that for relatives separated 10 generations ago we can expect to see 5 cM fragments (Han, L & Abney, M. 2013). However, there is no clear

consensus on the size of a fragment in a particular generation since the AC, as additional factors can influence the persistence of IBD in specific regions.

If we consider a human generation time of 25 years, 10 generations can correspond to 250 years. However, other authors, such as Ralph P. & Coop, G. 2013, who investigated the genealogies of 2,257 current Europeans within the last 3,000 years, have considered that 10 cM IBD would have originated within the last 500 years (corresponding to 20 generations), and that 4 cM IBD would come from AC from 20 and 60 generations ago. Again, some studies, such as in Browning, S.R. & Browning, B.L. 2010, or in Browning, S.R. AND Thompson, E.A. 2012, that used real and simulated populations to explore the possibility of identifying rare variants by IBD mapping (see section 1.4.2.2. “Other uses of IBD analyzes”), considered that IBDs originating 25 generations ago measured about 2 cM; In another study, where it is suggested that IBD analyses can be applied to detect geographic grouping or to detect relatives in GWAS studies, 2cM IBDs are expected to come from a CA that lived about 30 generations ago (Browning, B.L. & Browning, S.R. 2011).

To avoid all these discrepancies about the generation time corresponding to the length of a given IBD, the numbers suggested in Ralph, P. & Coop, G. 2013, have been used in this thesis, as it is the most influential work in which a relationship between the length (in cM) of an IBD and the number of generations since its origin is estimated.

1.4.2. Methods and uses of IBDs analyses

During the last two decades tens of softwares for detecting IBDs have been published. One of the main reasons to explain the development of so many different programs is the need to improve some characteristic over

previous programs, like reducing the computing time, getting lower values of false discovery rates, avoiding the need of phased genotypes or allowing the genomic markers to be in Linkage Disequilibrium (LD). Another reason is the development of methods that allow IBD analyses to be applied for new purposes, besides population genetics; some of these applications will be discussed in the next section.

1.4.2.1. IBDs analyses in population genomics

Positive selection

IBD analyses have been used to look for positive selection in a given population. Darwinian selection can reduce genomic variation around the target of selection, increasing homozygosity and/or decreasing intra-allelic recombination in the population (Albrechtsen, A., Moltke, I. & Nielsen, R. 2010). This little genomic variation allows detection of these selective events by IBD analysis, because the haplotype will be identical, or almost identical with few mutations arising among different individuals after the selective sweep. In addition, if the selection event is recent, it is expected to find less variation around the target area in a larger area, the probability of detecting it by IBD analysis increases. For example, Han, L. & Abney, M. 2013, found an increase in IBD segments around the human leukocyte antigen (HLA) region and the LCT gene in the Maasai population of Kenya. These loci were already known to be under recent selection probably due to the Maasai pastoralist lifestyle (Tishkoff, S.A. et al 2007). Other authors also detected recent detection events by IBD analysis around the HLA region in eleven HapMap 3 populations, including Asian, African, European, and Native Americans (Albrechtsen, A., Moltke, I. & Nielsen, R. 2010). This selection is probably explained by the role of the HLA region in the defense of the body against infectious diseases.

Recent gene flow

Long IBD fragments are likely to be inherited from a recent CA (see section 1.4.1.2. Length and age of IBD fragments). Following this logic, Botigué, L. R. et al. 2013 found an increase of long segments of IBD shared between populations of North Africa with populations of the Iberian Peninsula, but not with populations of Northern Europe, consistent with a gene flow between the Mediterranean populations and a subsequent isolation of Iberian individuals with the rest of Europe. Other authors dated, also through the analysis of IBD, that this gene flow stopped around 13,000 years (Harris, K. & Nielsen, R. 2013).

Ancient Admixture

As we saw the Neanderthals inhabited the Eurasian continent, where they interbreed with populations of *Homo sapiens* once they migrated from the Africa continent into Eurasia. At first, it was thought that there was basically no Neanderthal ancestry in sub-Saharan Africans. However, recent studies suggest some presence of Neanderthal ancestry in African individuals, probably due to back migrations from eurasians populations of *Homo sapiens*, who already carried the Neanderthal alleles. In fact, IBD analyzes have recently been used to explore this mixture with an innovative method that does not require human reference populations and revealed a greater amount of Neanderthal ancestry in African populations than was previously detected (Chen. L. and others 2020).

Population substructure

The likelihood of two individuals sharing a CA increases as we move into the past. From the last 1,500 y.a., a random pair of Europeans shares between 2 and 12 common ancestors (Ralph P. & Coop, G. 2013), and about 75% of the British share at least a CA among them (Saada, J.N. et al. 2020). Moreover, Weitz, J.S. et al. 2014, predicted that almost all the world's Jews have an ancestor from the Jewish population expelled from Spain, around 500 y.a.

In addition, the probability of finding a closer CA increases with geography. For example, in a studio with 2,257 Europeans, it was found that the recent relatedness is decreasing with geographical distance, showing that IBDs analysis can be used to reveal the genomic substructure of a given population. In fact, recently, Saada, J.N. et al. 2020, used IBD analysis with almost 500,000 British individuals to discover fine-scale structure in the British population. For example, they were able to determine that the average distance between the place of birth for a pair of 3rd degree cousins is 17 km, and for 2nd degree cousins reduces to only 5 km.

Kinship coefficients

The genomic material of any sexual organism is the combination of both the maternal and the parental genome and is expected to share half of its genome in IBD with each parent. As recombination breaks down the genomic molecules in each generation, an individual would share a quarter of his genome in IBD with each of his four grandparents, and so on. Therefore, the expected genome in IBD between two individuals may correlate with the number of recombination events from the CA. Based on this, some studies and computer programs use the proportion of the genome in IBD or the number and length of IBD segments between two individuals to estimate their family relationship (Guzev, A. et al. 2011). For example, although it will depend on the density of the SNP or the software used, in Saada, J.N. et al. 2020, 5th grade cousins are expected to share

about 3.5 cM in IBD across the genome, 3rd grade cousins about 56.6 cM, and 2nd grade cousins about 226.5 cM.

Phenotype-haplotype associations mapping

In some cases, genome-wide association studies (GWAS) are not able to correctly identify the causal variants of a particular disease, for example when the variants are too rare or are population-specific, thus missing their correct allocation from the reference databases. IBD analyses have overcome these difficulties as they can point to regions shared among case-pairs that are not shared with control individuals (Albrechtsen, A. et. al. 2009; Gusev, A. et al 2011; Browning, S.R. AND Thompson, E.A. 2012). For example, in Han, B. et al 2014, IBD-based analyses were able to map the HLA gene as the responsible for type 1 diabetes in the Wellcome Trust Control Consortium database.

1.4.3. Runs of Homozygosity

Runs of homozygosity (ROHs), are two segments of IBDs that are in homozygosity in an individual. In other words, the same loci on an individual's homologous chromosomes is identical because it has been inherited through the maternal and paternal lines from the same common ancestor (Dixit, S.P. et al. 2020) (See Figure X). ROHs are normally used to determine the level of inbreeding (or inbreeding coefficient) of a particular individual. For example, in small or isolated populations, the level of high inbreeding is expected to increase and can be measured by the proportion of ROH across the genome. Interestingly, analysis of the ROH of an older male allowed to determine the oldest known case of a first-order incestuous union, although they could not determine if his parents were full siblings or parent and offspring (Cassidy, L.M et al.

2020).

2. METHODS

In the following section I will summarize the process needed to be conducted previously and posterior to the IBD detection.

2.1. Previous analysis to the IBD detection

2.1.1. Phasing step

IBD detection requires a diverse dataset of individuals with high density of genomic positions to avoid false positive errors in IBD detection. There are some databases that can be used to merge with our data, such as the 1000 Genomes Project (The 1000 Genome Project Consortium), the Populations Reference Samples (POPRES) (Nelson, M.R. et al. 2008), the Simons Genomic Diversity Project (Mallick, S. et al. 2016) or the Human Origins Dataset (Patterson, N. et al. 2012).

Prior to IBD detections it is necessary to phase the haplotypes. The phasing step is necessary to make sure that two consecutive SNPs belong to the same chromosome and not one SNP to one chromosome and the other SNP to its homologue, as it could lead to both false positive and false negative errors in IBD detection (Browning, B.L. & Browning, S.R. et al. 2013). Nowadays, there are different softwares that allow to phase new samples using reference genomic maps.

Bubi analyses

As the main objective of the Bubi analyses was to understand their origins, we selected a dataset of 1,235 individuals from 35 West African populations (Patin, E. et al. 2017). We merged these individuals with the 13 sequenced Bubi individuals and phased them with SHAPEIT2 software (O'Connell, J. et al. 2014) using the HapMap phase 2 genetic map as a reference (International HapMap Consortium et al. 2007) .

Connections between ancient and modern Europeans

For analyses with modern Europeans and an ancient sample, we selected the Human Origins dataset, as there are data from several European populations with a significant number of individuals representative of each. For the phasing step we used BEAGLE 5.0 software (Browning, S.R. & Browning, S.L. 2007) to phase the ancient sample using also the HapMap phase 2 genetic map.

2.1.2. Genotyping control tests

Prior to the IBD analyses we performed different genotyping controls to minimize possible false positive and false negative errors.

2.1.2.1 Minor allele frequency

Minor allele frequency (MAF) refers to the frequency at which the second most common allele is found in a given population. If this allele has an extremely low frequency in a population, it can lead to false discovery errors (Tabangin, M.E. et al. 2009) in IBD detection if the allele is present in two unrelated individuals.

To avoid these situations, we use PLINK (Purcell, S. et al. 2007) software to discard all SNPs in which the second most common allele has a frequency of 0.05.

2.1.2.2. Missingness

Missingness refers to the lack of information at a particular locus in an individual. This can be due, for example, to sequencing errors. If the individual shows an increase of loci with no information across their genome, this could lead to false positive or false negative errors in IBD discovery. To avoid this situation we removed all individuals with more than 5% missing SNPs in the Human Origins database. In Figure 11 it can be seen that three individuals show high levels of missingness and heterozygosity. As this could be a consequence of sequencing errors we proceed to discard them.

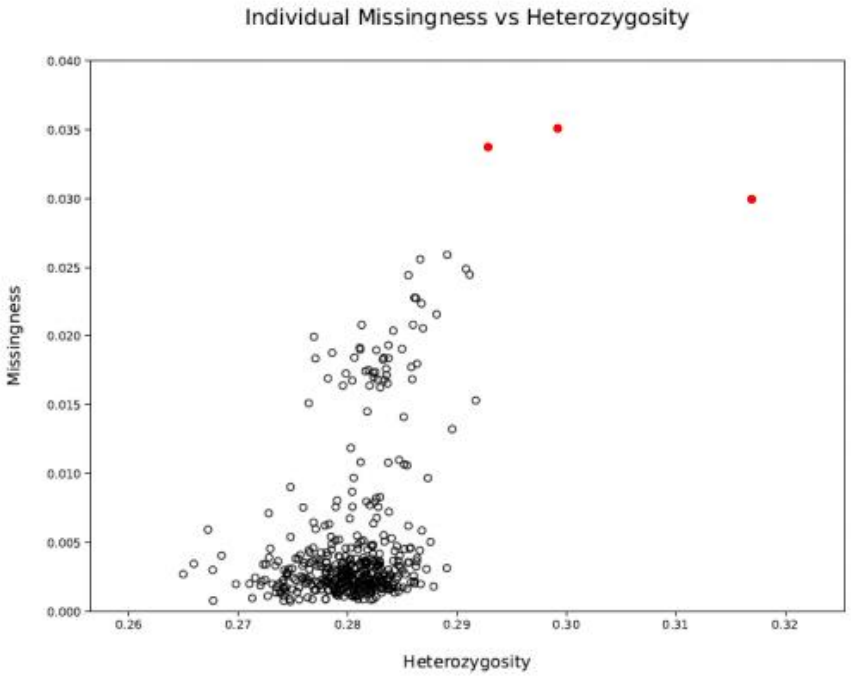


Figure 11. Missingness of modern European individuals plotted against their heterozygosity. The three individuals in red were removed from the database.

2.1.2.3. Individual relatedness

Closely related individuals will show an increase in the sharing of IBD fragments. This will lead to a bias in the IBD sharing in the population to which these individuals belong. We found a pair of related individuals in the Human Origins dataset with PLINK software and we eliminated one of these individuals.

2.1.2.4. Hardy-Weinberg disequilibrium

Hardy-Weinberg equilibrium refers to the maintenance of allele frequencies in a population when there are non evolutionary processes acting on it. Conversely, if processes such as selection, increased mutation rate or migration act on a given locus, changes in the original allele frequencies could occur (Wigginton, J.E. et al. 2005).

These situations could be reflected in false positive errors. To give an example, imagine a haplotype that has been in Hardy-Weinberg equilibrium in a population for many generations, and is recently selected. Selection will lead to an increase in the frequency of the haplotype and the different individuals carrying it will appear to be closely related to each other: even though this haplotype originated many generations ago, the IBD detection software would flag these individuals as close relatives, since the genomic similarity around these loci is higher than expected.

Since the evolutionary process constantly modulates the genomic set of populations, it is to be expected that many of the SNPs in Human Origins are not in Hardy-Weinberg equilibrium. Therefore, with PLINK software we

removed all these SNPs that are in Hardy-Weinberg disequilibrium in the database by filtering out all *p-values* below 1e-6.

2.1.2.5. Linkage Disequilibrium

Loci in Linkage Disequilibrium (LD) are those loci with low likelihood to segregate during the recombination, so they are usually inherited together through generations (Nordborg, M. & Tavaré, S. 2002). In these cases the haplotype would be very similar in different individuals even if they are not close relatives, leading to false positive errors in the IBD discovery.

To avoid this we used the PLINK software to remove all the SNPs in LD in our data.

2.2. Posterior analysis to the IBD detection

The IBD detection was conducted with the RefinedIBD software.

We used Ralph, P. & Coop, G. 2013 as reference for the relation between generation time and the expected IBD length. As the Bubi people are considered to colonise the Bioko island around 2,000 y.a., we considered IBDs of 2cM length as representatives of their relative connexions since this time. Regarding the modern European analyses, we selected IBDs higher than 6cM to explore the ancestral connexions inside the last few hundred years.

2.2.1. Abnormal score and length values

We plotted the score values against the length of each IBD and we removed those with odd values (Figure 12).

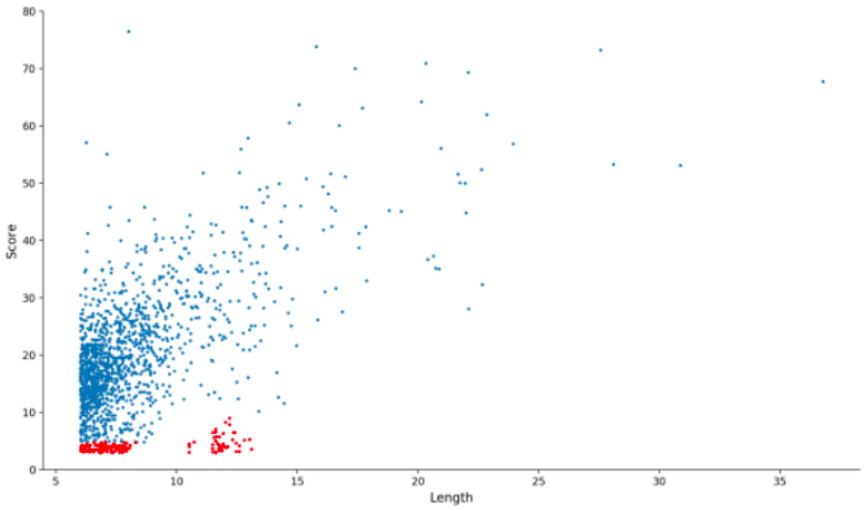


Figure 12. Plot of the score of each IBD segment against its length. Those points in red were considered to cluster abnormally and were removed.

2.2.2. Centromeric and telomeric regions

We remove IBD falling inside centromeric and telomeric regions because, exactly like haplotypes in Hardy-Weinberg disequilibrium, tend to have low recombination rates (as an example see Figure 13) Andy Choo, K.H. 1998; Shen, B. et al. 2018).

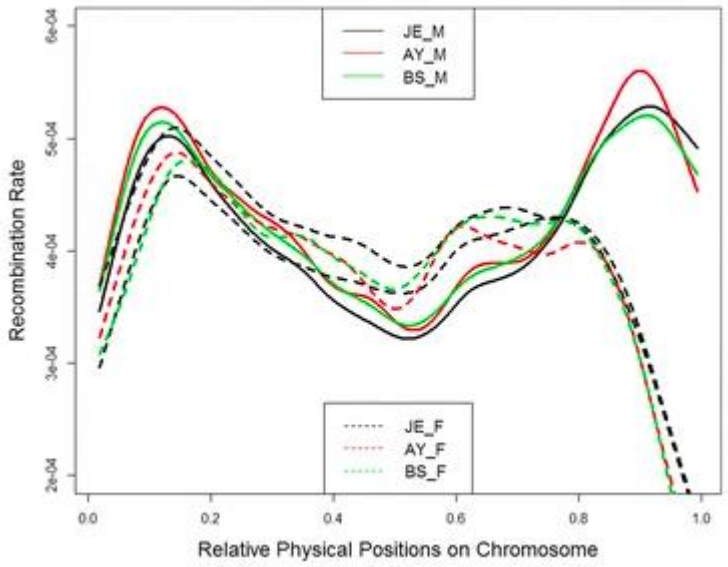


Figure 13. Variation of the recombination rates across chromosomes in different cattle breeds (Shen, B. et al. 2018).

2.2.3. Triangulation analysis

We detected 1,523 IBDs longer than 6cM between modern Europeans. In order to detect false positive errors in the IBD detection we looked for transitivity between trios of individuals. In other words, if the individual A shares an IBD with the individual B and the individual C, we expect to see the same IBD shared between the individuals B and C.

We detected 220 individuals sharing the same IBD with two other individuals. And expected, in most of these cases, the other two individuals also shared the IBD between them. In three cases did not detect this transitivity between the individuals involved. We discarded these three IBDs as they could be artifacts of the IBD detection.

3. OBJECTIVES

The main objective of this thesis is to explore the analysis based on the blocks of identity by descent (IBD) to unravel the genomic connections between genetically close populations (genetically homogeneous populations).

In particular:
a) Generate data on the entire genome of the current Bubi individuals, which are native to Bioko Island (Gulf of Guinea) -one of the few islands on the African continent- to:

-Identify which Bantu-speaking population of the African continent is more genetically close to the Bubi.
-Identify genetic signs of isolation by insularity from IBD and ROHs analysis, as well as possible recent links of the population to the continent

b) Estimate the large (>6cM) blocks of IBD among current Europeans using modern genomic data on a fine geographic scale to:

-Detect dense connections of co-ancestry occurred in the last hundreds of years that could be correlated with characteristics of the history of the population, such as inbreeding and isolation.
-Detect recent migration links that may be difficult to discover with the most widely used population genetic tools.
-Explore new methods for representing IBD blocks based on graph theory such as networks.

c) Generate a high coverage genome of an individual who lived in the late Middle Ages to:

-Apply an analysis of the IBD blocks between this individual and modern Europeans, to discover the recent co-ancestry connections between this individual and modern European populations

-Explore the potential for connecting the past and present in future genomic studies with the use of the IBDs blocks.

4. RESULTS

4.1. Genome-wide data from the bubi of Bioko Island clarifies the Atlantic fringe of the Bantu dispersal

Pere Gelabert, Manuel Ferrando-Bernal, Toni de-Dios, Benedetta Mattore, Elena Campoy, Amaya Gorostiza, Etienne Patin, Antonio González Martín & Carles Lalueza-Fox

BMC Genomics. 2019; 20 (179): 1-13

DOI: http://doi.org/10.1186/s12864-019-5529-0

Abstract

Background: Bioko is one of the few islands that exist around Africa, the most genetically diverse continent on the planet. The native Bantu-speaking inhabitants of Bioko, the Bubi, are believed to have colonized the island about 2000 years ago. Here, we sequenced the genome of thirteen Bubi individuals at high coverage and analysed their sequences in comparison to mainland populations from the Gulf of Guinea.

Results: We found that, genetically, the closest mainland population to the Bubi are Bantu-speaking groups from Angola instead the geographically closer groups from Cameroon. The Bubi possess a lower proportion of rainforest hunter-gatherer (RHG) ancestry than most other Bantu-speaking groups. However, their RHG component most likely came from the same source and could have reached them by gene flow from the mainland after island settlement. By studying identity by descent (IBD) genomic blocks and runs of homozygosity (ROHs), we found evidence for a significant level of genetic isolation among the Bubi, isolation that can be attributed to the island effect. Additionally, as this population is known to have one of the highest malaria incidence rates in the world we analysed their genome for malaria-resistant alleles. However, we were unable to detect any specific selective sweeps related to this disease.

Conclusions: By describing their dispersal to the Atlantic islands, the genomic characterization of the Bubi contributes to the understanding of the margins of the massive Bantu migration that shaped all Sub-Saharan African populations.

Background

The Gulf of Guinea, which covers a large portion of the African coast, is characterized by complex and rich ecosystems. Of the few islands located in Atlantic Africa, four of them are found within the Gulf of Guinea (Fig. 1a). Bioko is the largest of these islands, with a total area of 2017 km^2 (Fig. 1b). It is located 32 km offshore of Cameroon but constitutes in fact, the northernmost part of Equatorial Guinea, a former Spanish colony. The island is volcanic and very mountainous, with an abrupt coastline; despite its small size, its highest peak has an altitude of 3012 m. The indigenous population of Bioko, the Bubi, speak Bubi, a basal Bantu language [1]. The closest Bantu language on the mainland is most likely Galoa, which is spoken by Bantu-speaking groups of the Ogowe basin in Gabon [2]. The Bubi have a distinct and unique culture among Bantu-speaking people [3], including the belief that different spiritual beings reside in specific geographical locations along the island and the existence of well-defined matrilineal clans [4].

The origin of the Bubi people is controversial. Since the British explorer Richard Francis Burton visited the island (then called Fernando Poo) in 1874 [5], ethnographers generally considered the Bubi to be the original settlers of Bioko. However, it is currently accepted that the Bubi agriculturalists arrived from the mainland in dugout canoes about 2000 years ago during the Late Neolithic [6–8]. Ever since, the Bubi seem to have been isolated from mainland Bantu-speaking groups [9]. Bubi mythology explains that, upon their arrival to the island, they found other, more robust people living there, a population whom they called Balettérimo [1, 9].

In fact, some unsystematic archaeological prospects carried out by Spanish scholars have uncovered pre-Neolithic lithic tools of a typology

that has been called banapense, although this lithic typology does not currently have a clear chronological framework [9].

The expansion of the Bantu-speaking farming communities is probably one of the most important human movements that have taken place in recent African his tory [10]. This movement started approximately 4000 to 5000 years ago [11], likely from a source close to the present-day North Cameroonian [12] or Gabonese/An golan Bantu-speaking populations [13], depending on which model – “early-split” or “late-split” – is assumed. While the first model suggests that the Bantus made an early separation into western and eastern branches, the second model supports an initial movement south across the rainforest before splitting into two branches, one headed south and the other east. Whatever the route of dispersal, the Bantu-speaking migration triggered an expansion of agriculture and ironwork along with the spread of Bantu languages (part of the Niger-Congo family) across most of Central, South and East Africa [13–16].

During their extensive geographical migration, the Bantus encountered and admixed with local rainforest hunter-gatherer (RHG) tribes. Several Bantu-speaking groups have been studied from a genetic point of view over the years, especially using mitochondrial DNA (mtDNA) and Y chromosome markers [12, 17–20]. These studies have detected a substantial fraction of Pygmy mtDNA lineages within the Bantu speakers, but rarely the opposite [17, 18]. For example, traces (around 1%) of RHG Y chromosomes have been found in Bantu-speaking groups from Gabon and Cameroon [12] and signals of hunter-gatherer Khoisan Y chromosomes in Bantu-speaking groups from Mozambique [21].

Despite these efforts, the genetic patterns of variation within the Bantus remained largely unexplored on a genomic scale until fairly recently [14, 15, 22]. The analysis of a large and geographically diverse dataset of 35 Bantu groups has recently confirmed the existence of this RHG ancestry

as well as the acquisition of adaptive alleles from these local populations, especially at the human-leukocyte antigen (HLA) loci. By measuring the length of the introgressed genetic fragments, the admixture between western Bantu-speaking populations and RHG was estimated to occur about 800 years ago, mostly after Bantu-speaking populations began moving throughout Sub-Saharan Africa [13]. According to this study, while the most likely parental source population of Bantu ancestry in both eastern and southern Bantus was located in northern Angola, Bakoya of Gabon and Congo were the best parental source for RHG ancestry. Recent retrieval of ancient genomes from different Afri can localities, notably in the south and the east, could help elucidate these past admixture events [23, 24]. So far, however, no ancient genomes have been retrieved from the Gulf of Guinea.

There are hypotheses that could potentially be explored with genome-wide data from the Bubi. First, their relatively long period of isolation on the island is a potential way to test the age and extent of the Bantu ad mixture with RHG tribes, an event that supposedly took place among the coastal Bantu groups after isolation of Bubi ancestors on the island. Second, studying a tribe from one of the few islands around Africa could provide information about potential effects of endogamy and isolation that are less likely to occur in mainland tribes. Finally, genome-wide data from the Bubi can offer clues regarding the adaptation of this group to local conditions. For example, it is known that the population of Equatorial Guinea has one of the highest levels of malaria infection in the world [25] and ranks 13th in the list of malaria prevalence countries, representing the second highest cause of death in the country [26] Despite prevention efforts carried out since 2004, severe malaria prevalence in Bioko children remains high [27]. Thus, screening potential resistant variants can help us further understand the selective pressures faced by the Bubi during the last few hundred years.

We show in this work that the Bantu population of Bioko Island mirrors the genetic makeup of the extant, coastal Bantu-speaking groups, also clarifying the dynamics of this human expansion into the Atlantic islands of Africa.

Results

We sequenced 13 Bubi genomes obtaining a depth of coverage up to 21x-32x (Additional file 1: Table S5). All mtDNA haplogroups present in the Bubi are subclades of the L haplogroup (L1b, L2b, L3e, L3f, and L3e) and are common in other populations from the Gulf of Guinea. All male individuals of this dataset belong to subclades of the E1b1a1 haplogroup, the predominant Y-chromosome lineage in Western, Central and Southern Africa (Additional file 1: Table S6).

On a genome-wide scale, the first two components of the principal component analysis (PCA) separate the RHG from the Bantu speakers and the Western Africa non-Bantu populations (Fig. 1c). When PC3 and PC4 are plotted, Bantu-speaking and Western African populations cluster separately (Additional file 2: Figure S1). Three Bubi individuals that we named Bubi-subset1 (BBS014, BBS018, BBS023) fall within the Western Africa cluster showing that some individuals share a larger proportion of Western-Africa ancestry than others, while the rest (named Bubi-subset2) cluster within the Bantu diversity (Additional file 2: Figure S1). To examine if this clustering reveals the presence of population substructure, we have used the USCS liftover [28] to convert the chimpanzee reference sequence (Pan troglodytes 3.0 assembly, GCA_000001515.5) to human coordinates in 546,558 single nucleotide polymorphisms (SNPs) of our dataset. We have subsequently computed f_4 statistics in the form (Bubi-subset1, Bubi-subset2; X, chimpanzee), to examine the homogeneity of the Bubi population (Additional file 2: Figure S2, Additional file 1: Table S7). None of the tested populations showed elevated ($|Z| > 3$) values of Z-score; therefore we have treated the Bubi as a single population in subsequent analyses.

Moreover, ADMIXTURE analysis (K = 4) (Fig. 2 and Additional file 2: Figure S3) shows that the Bubi contain the same components as the other Bantu-speaking populations of the dataset but lower levels of the RHG component (in grey) than most of the mainland Bantu-speaking populations (with those from Angola being the exception). Some of the remaining ADMIXTURE plots (K = 2–15) (Additional file 2: Figure S4) also indicate potential substructuring among the Bubi; however, the sample size is too small to test if this is correlated with geography.

To determine which mainland populations share a higher level of genetic ancestry with the Bubi, we tested all neighbouring populations with the outgroup f_3 statistic (considering San as outgroup). Our results indicate that non-Bantu Western African populations such as Yoruba, Bariba, Fon and Azhizi, as well as Bantu-speaking groups from Angola such as Kongo, Ovimbundu and Kimbundu show more genetic affinities with the Bubi, apparently because they all show lower hunter-gatherer ancestry as compared to other groups (Fig. 3 and Additional file 1: Table S8). To explore the possible source of genetic admixture in the Bubi, we have also calculated the f_4 statistic for the combinations (Test, San; Bubi, Mbuti), (Test, San; Bubi, Baka), (Test, San; Bubi, Yoruba), (Test, San; Bubi, Fang) (Additional file 2: Figure S5A-D). This selection represents the four major genomic components in the Gulf of Guinea. We have found that the Bubi show the highest levels of shared ancestry with the Western African populations, which do not overlap with other populations when the comparison is established with RHG components. Angolan Bantu-speaking populations such as Ovimbundu, Kongo and Kimbundu also show elevated levels of shared genetic ancestry (Additional file 2: Figure S5). Bubi people seem to have very low levels of genetic admixture from the RHG populations; however, this small signal is absent in Western Africans.

Furthermore, owing to the colonial history of Bioko we have explored the possibility of some Iberian contribution to the Bubi ancestry by calculating the f_3 statistic in the form (Iberia, X; Mbuti). The Bubi are placed within the range of Western African and other Bantu-speaking populations (Additional file 2: Figure S6); therefore, no Iberian genetic affinities can be discerned within the current dataset.

Pairwise F_{st} is a statistic used to measure population differentiation based on allelic frequencies. We used this test to quantify the level of genetic differentiation between all combinations of populations in our dataset. Low levels of F_{st} statistics indicate that the tested populations share a large proportion of the genotypes, while high levels are indicators of genetic differentiation. Among all African samples, RHG tribes appear to have the highest levels of genetic differentiation, both among themselves and with the agriculturalist groups (Fig. 4). The lowest values of genome-wide F_{st} for the Bubi are again with Bantu-speaking Angolan groups (Kimbundu, F_{st} = 0.0045; Ovimbundu, F_{st} = 0.0045), and also with a Bantu-speaking group from Cameroon (Yaounde, F_{st} = 0.0045). On the other hand, we found that the Bubi dis played the highest F_{st} values when compared with RHG populations (Additional file 1: Table S9).

FineSTRUCTURE uses the coancestry matrix obtained from ChromoPainter to classify samples based on haplotype diversity. Using this approach, the matrix and the resulting dendrogram confirm that the closest populations to the Bubi are the Ovimbundu (Fig. 5 and Additional file 2: Figure S7), but also some Gabonese and Cameroonian Bantu-speaking populations. Interestingly, the Bubi are divided into two different clusters (Additional file 1: Table S7), one including only individuals from Bioko Island and the other shared with other Bantu-speaking groups. This, again, suggests some level of substructure in the population. Interestingly, the RHG show a higher level of haplotype differentiation compared to other populations. To additionally test for the presence of Iberian ancestry

we have repeated the fineSTRUCTURE analysis including 12 Iberian individuals (Additional file 2: Fig. S8), following the same methodology previously described and 244,897 phased SNPs.

To explore possible admixture events that could have led to the origins of the Bubi, we also performed GLOBETROTTER analysis (based on ChromoPainter); however, no admixing events could be identified. We also used fineSTRUCTURE to plot a PCA based on haplotype differentiation. This method clusters the populations into the same groups as those used when considering genotypes. In this analysis, the Bubi present a clear intermediate position between Western African and Bantu-speaking populations (Additional file 2: Figure S9). For computational convenience, we performed these analyses with a restricted dataset.

Furthermore, we also estimated identity by descent (IBD) tracks to help elucidate the recent co-ancestry links between the Bubi and mainland populations. We found the Bubi to share the highest number of IBD blocks longer than 2 cM (indicative of genealogical connections occurring during the last 2500 years [29]) with populations from Angola (Ovimbundu, Kongo), as well as some from Gabon (e.g., Obamba, Duma, Bateke and Bapunu) (Fig. 6 and Additional file 1: Table S10). Additionally, we estimated the average of runs of homozygosity (ROHs) in each population, as well as the fraction of the genome in homozygosity as signals of endogamy. We found that the Bubi are the second most endogamous Bantu-speaking group (Additional file 1: Table S11) after the Bekwil population, although the ROHs of RHG tribes were longer on average than those observed among the agriculturalist groups (Additional file 2: Figure S10, Additional file 1: Table S12).

We found that all Bubi individuals possessed the malaria-resistant allele of the ACKR1 [30] and CD36 genes [31], in addition to certain variations in other genes such as G6PD [32], ATP2B4 [33], GRK5 [34], and IL-10 [35].

Moreover, resistant variants are absent in ABO, HBB [36] and TIRAP [37] (Additional file 1: Table S13) [38]. However, considering these mutations are observed at low frequency among other African groups, it is likely they were simply not present in the ancestors of the Bubi (Additional file 1: Table S14). We compared the allelic frequencies of malaria SNPs in the Bubi with those from neighbouring populations of the Gulf of Guinea – such as Esan, Gambian, Mende, and Yoruba – for which genome-wide sequence data was available. We found statistically significant (Fisher's exact test) differences in some alleles, but nothing that indicated a unique trend in the Bubi (Additional file 1: Table S15). We subsequently conducted a genome-wide F_{st} scan between the Bubi and Yoruba, using all the variable positions with MAF > 0.05 and missing genotypes < 0.05, plotting the mean F_{st} values in 0.5 Mb windows. We set a threshold of significance in 0.25 [39]. We have failed to detect any signal of a selective event, including those regions related to immunity against malaria (Additional file 2: Figure S11).

Discussion

Our genomic study of the population of Bioko Island confirms that the Bantu-speaking migration that shaped most of the present-day human diversity in Sub-Saharan Africa [40] also had a significant impact on African islands of the Gulf of Guinea. The general components of ancestry found in the Bubi are not different from those found in mainland Bantu-speaking groups, although in the case of the Bubi, the RHG ancestry is lower than the amount detected in most Western Bantu-speaking groups. Moreover, we did not detect a significant difference in the origin of the Bubi RHG gen etic signal to the one observed in other Bantu populations. One potential explanation could be that an admixture event between the ancestors of the Bubi and the RHG tribes started about 2000 years ago and was brought to the island upon settlement, but continued to increase thereafter in most mainland Bantu-speaking groups. It is worth noting that the time of admixture can be underestimated when using methods based

on linkage-disequilibrium decay if continuous admixture events actually occurred [41]. Therefore, the current 800 years estimate [13] could in fact be the end of a long period of gene flow between mainland Bantu-speakers and RHG. This scenario could help explain the clustering of the Bubi with Western African groups in some analyses (the latter groups also show residual or no traces of RHG ancestry).

Due to a certain degree of heterogeneity detected within the Bubi that was evident from PCA, ADMIXTURE and fineSTRUCTURE analyses, the possibility that different populations from Bioko could harbour slightly different genetic histories existed. Notably, some Bubi show almost no signs of RHG; interestingly, one of these individuals is from Bariobé, a relatively isolated province in the mountainous interior of the island. An alternative possibility could be that the small fraction of RHG ancestry was acquired by gene flow from coastal regions after the ancestors of the Bubi settled in Bioko. For example, the presence of both Fang and Benga people in Bioko has been described in historical times, partly related to the slave trade. In fact, although the slave trade was not so important in Bioko, it was very active in other coastal centres of the Gulf of Guinea, es pecially in some of the minor islands such as Corisco and Annobón [42]. Nonetheless, due to the cultural particularities of the Bubi and the clear genetic signals of endogamy and isolation, it seems unlikely they would assimilate a significant number of foreign people. In addition, no signals of potential Iberian admixture have been detected among the Bubi.

Within the Bantu-speakers, the Bubi are more closely related to Angolan than to the geographically closer Cameroon groups (this is supported for instance by fineSTRUCTURE or f_3 statistics). Based on the evidence that Bantu expansion likely moved from Angola north wards [13], it is possible that Bantu-speaking groups from Cameroon experienced subsequent admixture events with neighbouring RHG populations.

The Bubi particularities are mirrored by their geographically induced genetic isolation as well as their linguistic differences with neighbouring, mainland populations. At the linguistic level, the Bubi language is basal to most Bantu languages [40, 43] and clusters together with northwest Bantu speakers. This correlates with archaeological findings from the region dated from 5000 to 2500 years ago and associated to the spread of Bantu languages [40]. This decoupling between language and genetics could be explained if the former was acquired by or imposed onto the Bubi mainland ancestors. Accordingly, there are some historical accounts that consider the Bubi to be an enslaved tribe that escaped to Bioko [44].

The Bubi seem to have experienced a certain history of isolation that left a mark in their genomes. Out of all the Bantu-speaking groups, for instance, we found that the Bubi have some of the highest levels of IBD tracks shared among members of the same population, a signal of low diversity that is compatible with endogamy. In fact, in the ROH analysis, the Bubi rank as the second most endogamous Bantu-speaking group, only after the Bekwil. Nevertheless, the fact that the Bubi do not show a large genetic differentiation from potential source populations along the coast also indicates that drift did not have time to operate at large scale and that colonization of the island did not occur a long time ago.

The Bubi, like other groups from the Gulf of Guinea, display a high frequency of some mutations associated with protection against malaria. Other mutations, however, are absent or segregating. The underlying mutation for the Duffy-negative phenotype (at the ACKR1 gene) that is known to protect against Plasmodium vivax and P. knowlesi, seems to be fixed, or at least is present at extremely high frequencies, in the Bubi population. In fact, this is a common trait in all Western Africa. At the beginning of the twenty-first century, malaria was responsible for a child mortality rate of 152 per every 1000 births in Bioko island, a figure that is only beginning to decrease thanks to recent malaria control projects [27].

However, in a genome-wide scan performed against the Yoruba, we were unable to identify genomic regions in the Bubi that appear to be shaped by natural selection, even despite their insular conditions.

However, due to the limited sampling size and restricted distribution within Bioko, our study has to be considered as a preliminary assessment of the current Bubi genetic diversity. Despite evidences that our sampling size can effectively estimate parameters of genetic diversity from a larger population (see Methods last section), additional sampling with a broader geographical distribution should be undertaken in the future.

Conclusions

In addition to the general population affinities of the Bubi, we have unveiled genetic evidence of a certain degree of isolation, which can be related to the insular conditions; this trait is quite unique in most of African mainland populations. Our study of the genomic com position of the Bubi not only adds further information to the current genetic diversity within Africa and its Atlantic islands, but also points to the importance of the genome-wide analyses in unravelling population affin ities, selective pressures and past migrations that can be correlated with linguistic and archaeological information. We conclude that the origins of the Bantu expansion still needs further research and that future retrieval of ancient genomes from Central and Western Africa could shed needed light on the cradle of the Bantu migrations.

Methods

Samples

All thirteen individuals analysed in this study are members of the Cultural Bubi Association of Fuenlabrada, Madrid (Spain). We obtained informed consent from all subjects. We discarded 25 of the interviewed individuals

because of admixed ancestry; many of them had a recent Fang ancestor from the mainland. Even though most of the individuals were not born in Bioko, we verified that the selected individuals had all grandparents born in the island; many of the volunteers' direct ancestors come from Malabo, Bariobé and Baney, which are located in the Northeast region of Bioko (Additional file 1: Table S1).

Extraction, sequencing and mapping

We isolated DNA from cotton swabs using all the available material and an organic-based DNA extraction method adapted to Amicon® Ultra 0.5-mL columns [45]. After extraction, we concentrated the DNA by centrifugation up to 50 µL and subjected samples to a quality control. To ensure there was a proper DNA concentration, 1 µL of sample was loaded in a 1% agarose gel and stained with ethidium bromide. Only a single band was observed. The samples were quantified with BioTek's Epoch and yielded values, on average, of 68.88 ng/µL.

Genomic DNA libraries were prepared using TruSeq DNA PCR-Free Library Preparation Kit (in accordance with the general settings of the preparation guide). The procedure produced a PCR-free library with 350 bp average insert size that requires 20 ng/ul (in 50 ul samples). DNA samples were randomly fragmented by Covaris system and sequenced in HiSeqX10 (Illumina) with hiseq2x150bp settings plus 65 bp paired-end adapters at Macrogen (South Korea).

We evaluated the paired-end sequenced reads with FASTQc to check their quality. The sequencing adapters were then removed using Adapter removal [46], reads shorter than 30 bp were removed, and the reads were mapped against the Human reference genome [National Center for Biotechnology Information (NCBI) 37, hg19] using Burrows-Wheeler Aligner (BWA) with default parameters [47]. Duplicated reads were removed using Picard Tools MarkDuplicates version 2.8.3 and low quality

mapping reads (< 30) were removed with SAMtools version 1.623 [48].

Genotyping and quality control tests

Unique aligned reads were processed with Base Quality Score Recalibration (BQSR) implemented in the GATK version 3.7 software [49]. Even if the plots did not show signals of systematic errors, we applied recalibration to all filtered reads. We used GATK HaplotypeCaller in GVCF mode for scalable variant calling (using the GRCh37 as a reference sequence). Individual variant calls were merged in a single VCF file using GATK genotypeGVCFs tool, and the variants were filtered using Variant Quality Score Recalibration (VQSR) with a filter level of 99%. We used QD, MQ ReadPosRankSum, FS, and SOR annotations in this step. We excluded any variant with less than 70% of the main depth coverage or more than 200%. We also removed those variants exhibiting qualities below 30. We removed those called variants with a minimum allele frequency below 0.05 and of Hardy-Weinberg disequilibrium p-value below 1e-6.

Population genetics dataset

We merged our filtered variants with 690,739 SNPs from 1235 genotyped individuals belonging to 35 Western Af rica populations. This dataset includes: Bantu-speaking populations, hunter-gatherers and Western African groups [13],using Plink 1.9 [50] (Additional file 1: Table S2). We excluded triallelic sites, A/T and C/G mutations and all sites with a minor allele frequency (MAF) below 0.05. We subsequently removed positions with > 10% missing data and those individuals with > 5% missing values. To ensure that genotypes were properly called after merging the dataset, the Yoruba SNP genotypes were compared against the 1000 Genomes Yoruba population. However, subsequent analyses were performed only with the Yoruba genotypes from Western Africa dataset [51]. Positions that exhibited > 0.2 values of pairwise F_{st} between both samples were also removed. Based on the colonial history of Bioko, we have assessed the

presence of potential genetic admixture of the Bubi with Spanish individuals, adding Iberian samples from 1000 Genomes [52] to the SNP dataset. After this procedure, we again removed positions with MAF below 0.05, missing data above 0.1 and Hardy-Weinberg disequilibrium p-values below 1e-10.

For most of the analyses, we have extracted a sub-dataset with representative populations from West ern and Central Africa. This reduced dataset includes 14 populations and 169 individuals (Additional file 1: Table S3). Some of the population genomics analysis require an unrelated outgroup to the tested populations. We have merged our genotypes with data of eleven San individuals [53] from the Human Origins array [54], followed with the same merging procedure previously detailed. The resulting African dataset –including the Bubi- comprises 130,647 SNPs present in 1259 individuals.

Mitochondrial (mt) DNA and Y-chromosome analysis Reads were mapped against the Revised Cambridge Reference Sequence (rCRS) of the human mtDNA [55]. After calling variants with GATK version 3.7 [49] as has been previously described, the mtDNA haplogroups were predicted using Haplogrep version 2 [56]. Y chromosome haplogroups were predicted by classifying the observed mutations according to the PhyloTree database [57].

Population genomic analyses

To situate the Bubi within the present diversity of the Gulf of Guinea and Western Africa, a principal components analysis (PCA) with the reduced dataset was generated using EIGENSOFT software [58], Results were plotted using R package ggplot2 ADMIXTURE plots were generated to estimate the proportions of K ancestral components on each individual genome [61] of the reduced dataset. As the analysis assumes linkage

disequilibrium (LD), we pruned the dataset. We used Plink 1.9 to remove SNPs with an LD > r^2 = 0.5 in windows of 50 SNPs. ADMIXTURE analyses were performed with K from 2 to 15 and were repeated five times. The ADMIXTURE iterations were consoli dated using CLUMPP with the large K greedy algorithm [62] and the results were plotted using R package pophelper [63].

Outgroup f_3 statistic is a useful test to determine the closest population to a target one using one outgroup population and measuring the amount of shared genetic drift with a test population. San were selected as out group, as they represent the most distant African population with genome-wide data, Bubi population was compared to all other populations in the dataset. The f_3 (San; Bubi, Test) statistic was calculated with popstats [64] and the results were again plotted using R. f statistics can also be implemented in order to determine which populations exhibited the highest genetic drift with the Bubi people, to do so, we used the popstats software to compute the f_4 statistic (Test, San; Bubi, Mbuti), (Test, San; Bubi, Baka), (Test, San; Bubi, Yor uba), (Test, San; Bubi, Fang). These combinations allow us to dissect the genetic admixture of the tested populations with the Bubi in relation to all the representative sources of genetic ancestry in Western Africa: Eastern RHG, Western RHG, Western-African populations and Bantu-speaking populations.

The fixation index (F_{st}) is a measure of population differentiation. We calculated the mean pairwise F_{st} values between all the populations present in the global dataset. All autosomal SNPs were included in this analysis using the approach of Cockerham and Weir integrated in Plink 1.9 [65]

The reduced dataset was phased with SHAPEIT2 [66], using 500 states, 50 MCMC main steps, 10 burn-in and 10 pruning steps; recombination maps were interpolated from the HapMap phase 2 genetic maps. After excluding all positions with at least one missing site, we ended up with a

dataset of 491,203 variable positions with no missing data.

We used CHROMOPAINTER to build a coancestry matrix based on haplotype data from the phased-reduced dataset. This software estimates the admixture proportions in recipient chromosomes by painting the proportion of each genetic component from the donor populations. We ran CHROMOPAINTER with linked data, estimating n and M parameters through an observation run with no prefixed parameters and including 30 randomly selected samples and three randomly selected chromosomes; fineSTRUCTURE analysis was performed with the counts obtained in CHROMOPAINTER and ran with 1000,000 Markov chain Monte Carlo (MCMC) iterations and output printed every 10,000 iterations. The best tree was calculated with 10,000 state attempts. We also generated a haplotype-based PCA with fineSTRUCTURE.

To identify any admixture events between Bubi ancestors and other populations during the last 4500 years, we used the GLOBETROTTER [41] software on the basis of the defined clusters from fineSTRUCTURE (Additional file 1: Table S4).

Identity by descent (IBD) analysis

Identity by descent (IBD) blocks are defined as identical chromosome fragments present in multiple individuals that have been inherited from the same ancestral chromosome [67]. We have used RefinedIBD software [68] setting “ibdcm” = 0.5, “ibdtrim” = 62, “ibdwindow” = 2478, and “overlap” = 413; the rest of the parameters were assigned by default. All IBD blocks longer than five centimorgans (cM) were kept and the statistical threshold marked by LOD (the base 10 log of the likelihood ratio of the IBD segments, which is a figure that will depend of the size of the database and the genetic diversity within it) was assigned by default (> 3). The number of SNPs used here was 685,382. We then filtered the IBD segments to keep only those that were shared by any Bubi and another

individual of the dataset (including the IBD fragments shared by two Bubi individuals). To reduce the impact that the population size could have on the global counts of IBD blocks per population, we corrected the value of the shared IBD fragments (IBDn) by the population size (t). In order to obtain the average of the IBD blocks shared by any Bubi with any other individual or population, we divided each number obtained in the previous step by the number of Bubi individuals, 13:ratioBubi_pop = (IBDn/t)/13.

Runs of homozygosity (ROHs) analysis

ROHs (> 1000 kb) were estimated with Plink software. First, we calculated the average (in kilobases) of the genome that it is in homozygosis for each population. Sec ond, we calculated the average of the number of genomic fragments that are in homozygosis for each population.

Malaria resistance

Relevant mutations associated with malaria resistance in 10 different genes (Additional file 1: Table S11) – as found in genome-wide association studies (GWAS) and other previous studies [31, 69] – were genotyped in Bubi and the 1000 Genomes African populations. Fisher's exact test was used to determine the statistical significance of the observed differences ($p < 0.001$).

Evaluation of the effects of limited sample size We have used a whole genome F_{st} approach to evaluate the effects of the small sample size used in this work. We have randomly grouped the 186 Yoruba individuals from 1000 Genomes in 14 subsamples of 13–17 individuals and we have estimated the mean pairwise F_{st} values among all population combinations. All autosomal SNPs were included in this analysis using the approach of Cockerham and Weir integrated in Plink 1.9 [65]. No comparison has shown values of mean pairwise F_{st} higher than 0.1, which

indicates that the sub samples do not show significant differences in terms of genetic diversity (Additional file 2: Figure S12). This result suggests that the limited Bubi sample size can be used to infer genetic diversity at a higher population level.

Abbreviations

BQSR: Base Quality Score Recalibration; Fst: Fixation index; GWAS: Genome wide association study; HLA: Human-leukocyte antigen; IBD: Identity by descent; LD: Linkage disequilibrium; MAF: Minor allele frequency; MCMC: Markov chain Monte Carlo; mtDNA: Mitochondrial DNA; NCBI: National Center for Biotechnology Information; PCA: Principal component analysis; RHG: Rainforest hunter-gatherer; ROHs: Runs of homozygosity; SNP: Single nucleotide polymorphism; VQSR: Variant Quality Score Recalibration

Acknowledgements

We are grateful to the Bubi people from the Asocuba Association of Fuenlabrada for their ongoing support along the study and to Justo Bolekia Boloka and Jose F. Gómez for their help.

Funding

This research was supported by a grant from FEDER and Ministry of Economy and Competitiveness (BFU2015–64699-P) of Spain to C.L-F. CLF is supported by Obra Social "La Caixa" and Secretaria d'Universitats i Recerca Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880).

Availability of data and materials

Sequenced reads of the 13 Bubi people analysed in the present study have been submitted to European Nucleotide Archive under the accession numbers: ERR2640217-ERR2640226.

Authors' contributions

A.G.-M. and C.L.-F. conceived and coordinated the study; E.C., A.G. and A.G.- M. collected the samples; A.G. extracted the DNA; E.C., P.G., B.M., T.d.-D., E.P. and M.F.-B. analysed data and interpreted results; A.G.-M., C.L.-F. and P.G. wrote the paper. All authors reviewed the manuscript. All authors have read and approved the manuscript.

Ethics approval and consent to participate

Written informed consent was provided by all participants. The confidentially and anonymity of all participants has been guaranteed. Experimental protocols and informed consent have been approved by the Clinical Research Ethics Committee from Institut Hospital del Mar d'Investigacions Mèdiques in Barcelona (CEIC-PSMAR)(2018/7845/I).

Consent for publication

The consent for publication has been given by all the participants in the study.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra),

Barcelona, Spain. [2]Laboratory of Anthropology, Department of Biology, University of Florence, Florence, Italy. [3]Department of Biodiversity, Ecology and Evolution, Complutense University of Madrid, Madrid, Spain. [4]Forensic Genetics Laboratory, GENOMICA S.A.U., Pharma Mar Group, Madrid, Spain. [5]Unit of Human Evolutionary Genetics, Department of Genomics & Genetics, Institut Pasteur, Paris, France. [6]CNRS UMR 2000, Paris, France. [7]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France.

Received: 6 November 2018 Accepted: 13 February 2019

References

1. Günter T. Los Bubis de Fernando Póo. Trujillo JR, Rodríguez B, editors. Madrid: SIAL ediciones; 2008. p. 270.
2. Bolekia BJ. Lingüística Bantú a través del Bubi. Salamanca: Universidad de Salamanca; 2008. p. 192.
3. Jeffreys MDW. Some notes on the Neolithic of West Africa. In: Desmond Clark J, Cole S, editors. Third pan-african congress on prehistory. Livingstone: London Chatto & Windus; 1951. p. 262–73.
4. Eteo Soriso JF. Los ritos de paso entre los Bubis. Madrid: SIAL- Casa de Àfrica; 2017.
5. Burton RF. A visit to Fernando Po Peak, and a Night in the Open. Alp J. 1872;VI(XXXVII).
6. Martín del Molino A. Tipología de la cerámica de Fernando Poo. Vol. 1, Estudios del Instituto Claretiano de Africanistas (Separata de la revista “La Guinea Española”). Santa Isabel: Instituto Claretiano de Africanistas; 1960. 1–36 p.
7. Martín del Molino A. Secuencia Cultural en el Neolítico de Fernando Póo. Madrid: Inst. Español de Prehistoria-C.S.I.C.; 1965.
8. Martín del Molino A. Etapas de la cultura Carboneras de Fernando Poo en el primer milenio de nuestra Era. Madrid: Inst. Español de Prehistoria-C.S.I.C.; 1968.
9. Martín del Molino A. Los Bubis. Ritos y Creencias. Madrid: Labrys 54, Ediciones; 1993.
10. Diamond J. Guns, germs and steel. Norton W., editor. A short history of everybody for the last 13,000 years. New York: W. W. Norton & Company; 1997. p. 480.
11. Nurse D, Philippson G. In: Nurse D, Philippson G, editors. The bantu languages. New York: Routledge, Taylor & Francis Group; 2003.
12. Berniell-Lee G, Bosch E, Bertranpetit J, Comas D. Y-chromosome diversity in bantu and pygmy populations from Central Africa. Int Congr Ser. 2006; 1288:234–6.
13. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al.

Dispersals and genetic adaptation of bantu-speaking populations in Africa and North America. Science. 2017;356(6337):543–6. https://doi.org/10.1126/ science.aal1988.

14. Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, et al. Admixture into and within sub-Saharan Africa. eLife 2016;5:15266. doi: https://doi.org/10.7554/eLife.15266.

15. de Filippo C, Bostoen K, Stoneking M, Pakendorf B. Bringing together linguistic and genetic evidence to test the bantu expansion. Proc R Soc B Biol Sci. 2012;279(1741):3256–63. https://doi.org/10.1098/rspb.2012.0318.

16. Tishkoff SA, Reed FA, Ranciaro A, Awomoyi AA, Bodo J, Doumbo O, et al. The genetic structure and history of Africans and African Americans. Science. 2009;324:1035–44. https://doi.org/10.1126/science.1172257.

17. Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, et al. Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol Phylogenet Evol. 2007; 43(2):635–44. https://doi.org/10.1016/j.ympev.2006.09.014.

18. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, et al. Maternal traces of deep common ancestry and asymmetric gene flow between pygmy hunter-gatherers and bantu-speaking farmers. Proc Natl Acad Sci U S A. 2008;105(5):1596–601. https://doi.org/10.1073/pnas. 0711467105.

19. Destro-Bisol G, Coia V, Boschi I, Verginelli F, Caglià A, Pascali V, et al. The analysis of variation of mtDNA hypervariable region 1 suggests that eastern and Western pygmies diverged before the bantu expansion. Am Nat. 2004; 163(2):212–26. https://doi.org/10.1086/381405.

20. Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A. The genetic legacy of western bantu migrations. Hum Genet. 2005;117(4):366–75. https://doi.org/ 10.1007/s00439-005-1290-3.

21. Rowold D, Garcia-Bertrand R, Calderon S, Rivera L, Benedico DP, Alfonso Sanchez MA, et al. At the southeast fringe of the bantu expansion: genetic diversity and phylogenetic relationships to other sub-Saharan

tribes. Meta Gene. 2014;2:670–85. https://doi.org/10.1016/j.mgene.2014.08.003.

22. Li S, Schlebusch C, Jakobsson M. Genetic variation reveals large-scale population expansion and migration during the expansion of bantu speaking peoples. Proc R Soc B Biol Sci. 2014;281(1793):20141448. https:// doi.org/10.1098/rspb.2014.1448.

23. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing Prehistoric African Population Structure. Cell. 2017; 171(1):59–71.e21. https://doi.org/10.1016/j.cell.2017.08.049.

24. Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 2017;358(6363):652–5. https://doi.org/10.1126/science.aao6266.

25. Cibulskis RE, Aregawi M, Williams R, Otten M, Dye C. Worldwide incidence of malaria in 2009: estimates, time trends, and a critique of methods. PLoS Med. 2011;8(12):e1001142. https://doi.org/10.1371/journal.pmed.1001142.

26. World Health Organisation. World Malaria Report 2017. 2017.

27. Rehman AM, Mann AG, Schwabe C, Reddy MR, Roncon Gomes I, Slotman MA, et al. Five years of malaria control in the continental region, Equatorial Guinea. Malar J. 2013;12(1):154. https://doi.org/10.1186/1475-2875-12-154.

28. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006. https://doi.org/10.1101/gr.229102.

29. Ralph P, Coop G. The Geography of Recent Genetic Ancestry across Europe. PLoS Biol. 2013;11(5):e1001555. doi.org/10.1371/journal.pbio.

30. Tournamille C, Colin Y, Cartron JP, Le Van Kim C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy– negative individuals. Nat Genet. 1995;10:224–8. https://doi.org/10.1038/ ng0695-224.

31. Rockett KA, Clarke GM, Fitzpatrick K, Hubbart C, Jeffreys AE,

Rowlands K, et al. Reappraisal of known malaria resistance loci in a large multicenter study. Nat Genet. 2014;46(11):1197–204. https://doi.org/10.1038/ng.3107.

32. Hirono A, Beutler E. Alternative splicing of human Glucose-6-phosphate dehydrogenase messenger RNA in different tissues. J Clin Invest. 1989;83: 343–6. https://doi.org/10.1172/JCI113881.

33. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature. 2012;489(7416):443–6. https://doi.org/10.1038/nature11334.

34. Gupta H, Jain A, Saadi AV, Vasudevan TG, Hande MH, D'Souza SC, et al. Categorical complexities of Plasmodium falciparum malaria in individuals is associated with genetic variations in ADORA2A and GRK5 genes. Infect Genet Evol. 2015;34:188–99. https://doi.org/10.1371/journal.pone.0175702.

35. Gibson AW, Edberg JC, Wu J, Westendorp RG, Huizinga TW, Kimberly RP. Novel single nucleotide polymorphisms in the distal IL-10 promoter affect IL-10 production and enhance the risk of systemic lupus erythematosus. J Immunol. 2001;166(6):3915–22.

36. Gouagna LC, Bancone G, Yao F, Yameogo B, Dabire KR, Costantini C, et al. Genetic variation in human HBB is associated with Plasmodium falciparum transmission. Nat Genet. 2010;42(4):328–31. https://doi.org/10.1038/ng.554.

37. Khor CC, Chapman SJ, Vannberg FO, Dunne A, Murphy C, Ling EY, et al. A mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. Nat Genet. 2007;39(4):523–8. https://doi.org/10.1038/ng1976.

38. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. 2012;487(7407):375–9. https://doi.org/10.1038/ nature11174.

39. Hartl DL, Clark AG. Principles of population genetics; 1997. p. 546.

40. Currie TE, Meade A, Guillon M, Mace R. Cultural phylogeography of the bantu languages of sub-Saharan Africa. Proc R Soc London B Biol Sci. 2013; 280:1762. https://doi.org/10.1098/rspb.2012.0318.

41. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343:747–51. https://doi.org/10.1126/ science.1243518.

42. Thomas H. In: Paperbacks S& S, editor. The Slave Trade: The story of the Atlantic slave trade: 1440–1870. New York: Simon & Schuster; 1999.

43. Guthrie M. Comparative bantu: an introduction to the comparative linguistics and prehistory of the bantu languages. Ltd GIP, editor. Farnborough: Ltd, Gregg International Publishers; 1971.

44. Aymemí A. Los Bubis en Fernando Poo: colección de los artículos publicados en la revista colonial La Guinea española. Madrid: Ed. G. Sáez; 1942.

45. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: a laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.

46. Lindgreen S. AdapterRemoval: easy cleaning of next generation sequencing reads. BMC Res Notes. 2012;5(1):337. https://doi.org/10.1186/1756-0500-5-337.

47. Li H, Durbin R. Fast and accurate short read alignment with burrows wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10. 1093/bioinformatics/btp324.

48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16): 2078–9. https://doi.org/10.1093/bioinformatics/btp352.

49. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next generation DNA sequencing data. Genome Res. 2010;20:1298–303. https:// doi.org/10.1101/gr.107524.110.

50. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-

based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10. 1086/519795.

51. Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, et al. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun. 2014;5:3163. https://doi. org/10.1038/ncomms4163.

52. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393.

53. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present day Europeans. Nature. 2014;513(7518):409–13 37.

54. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y. Ancient Admixture in Human History. Genetics. 2012;192:1065–93. https://doi.org/10. 1534/genetics.112.145037.

55. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 1999;23(2):147. https://doi.org/10. 1038/13779.

56. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016;44(W1): W58–63.

57. van Oven M. PhyloTree build 17: growing the human mitochondrial DNA tree. Forensic Sci Int Genet Suppl Ser. 2015;5:e392–4.

58. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):2074–93. https://doi.org/10.1371/journal.pgen.002019.

59. Ross I, Robert G, Ihaka R, Gentleman R. R: A language for data analysis and graphics. J Comput Graph Stat. 1996;5:299–314.

60. Gómez-Rubio V. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). J Stat Softw. 2017;77:2–5.

61. Alexander DH, Novembre J. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64. https://doi.org/10. 1101/gr.094052.109.

62. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007;23(14):1801–6. https://doi.org/10. 1093/bioinformatics/btm233.

63. Francis RM. Pophelper: an R package and web app to analyse and visualize population structure. Mol Ecol Resour. 2017;17(1):27–32. https://doi.org/10. 1111/1755-0998.12509.

64. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. Nature. 2015;525(7567):104–8. https://doi.org/10.1038/nature14895.

65. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution (N Y). 1984;38(6):1358–70.

66. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013;10(1):5–6. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x.

67. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97. https://doi.org/10.1086/521987.

68. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013;194(2):459–71. https://doi.org/10.1534/genetics.113.150029.

69. Gelabert P, Olalde I, De-Dios T, Civit S, Lalueza-Fox. Malaria was a weak selective force in ancient Europeans. Sci Rep. 2017;7(1):1377. https://doi.org/ 10.1038/s41598-017-01534-5.

Figures

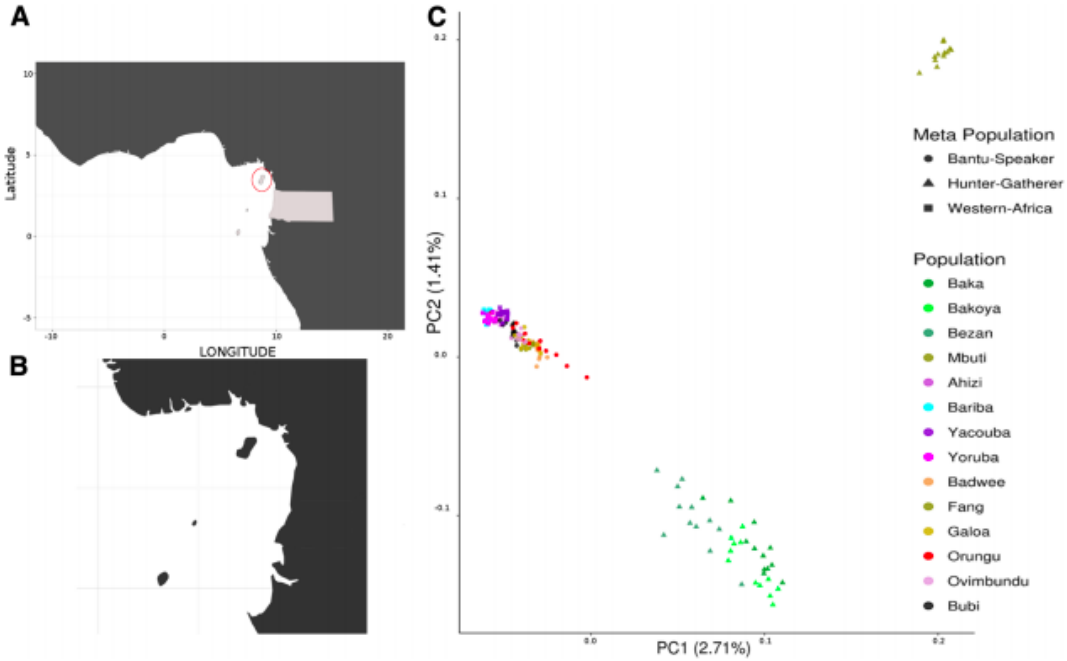


Figure 1. A Map of the Gulf of Guinea showing the location of Bioko Island and the neighbouring countries. B Map of the Bioko Island. C Genotype based principal components analysis (PCA) plot obtained with EIGENSOFT smartpca. A dataset of 168 individuals from 14 populations and 581,224 SNPs are included. The percentage of variance explained is written along the axes. The maps have been created with R software.

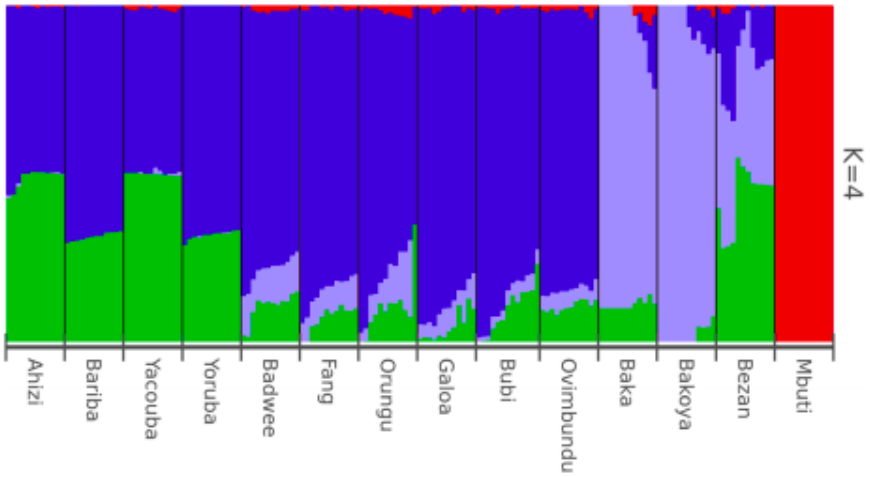


Figure 2. Graphical representation of shared genetic components performed using the ADMIXTURE (K = 4) software: 169 individuals from 13 populationsand 456,095 SNPs have been plotted. Although the cross-validation errors show the lowest value at K = 2, we have chosen K = 4 because it is the first plot where the crucial RHG, Bantu and Western African components are clearly identified.

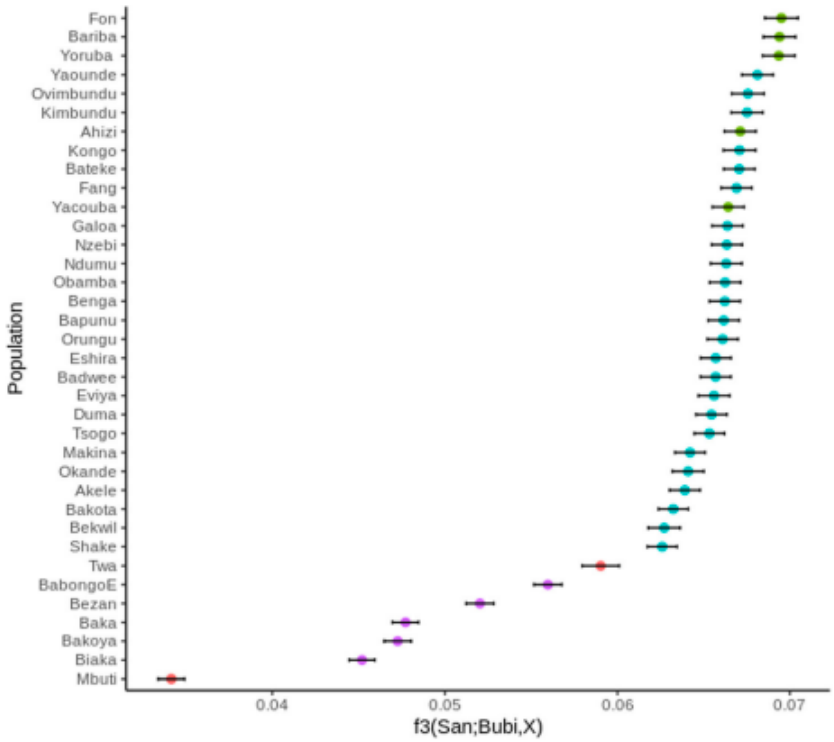


Figure 3. Statistic f3values obtained with popstats when taking San as outgroup. The statistic and one standard error deviation are presented for each combination test. Blue dots represent Bantu-speaking populations, Green dots represent Western-African populations, orange dots Represent Eastern Hunter-Gatherer populations and purple dots indicate Western Hunter-gatherer populations.

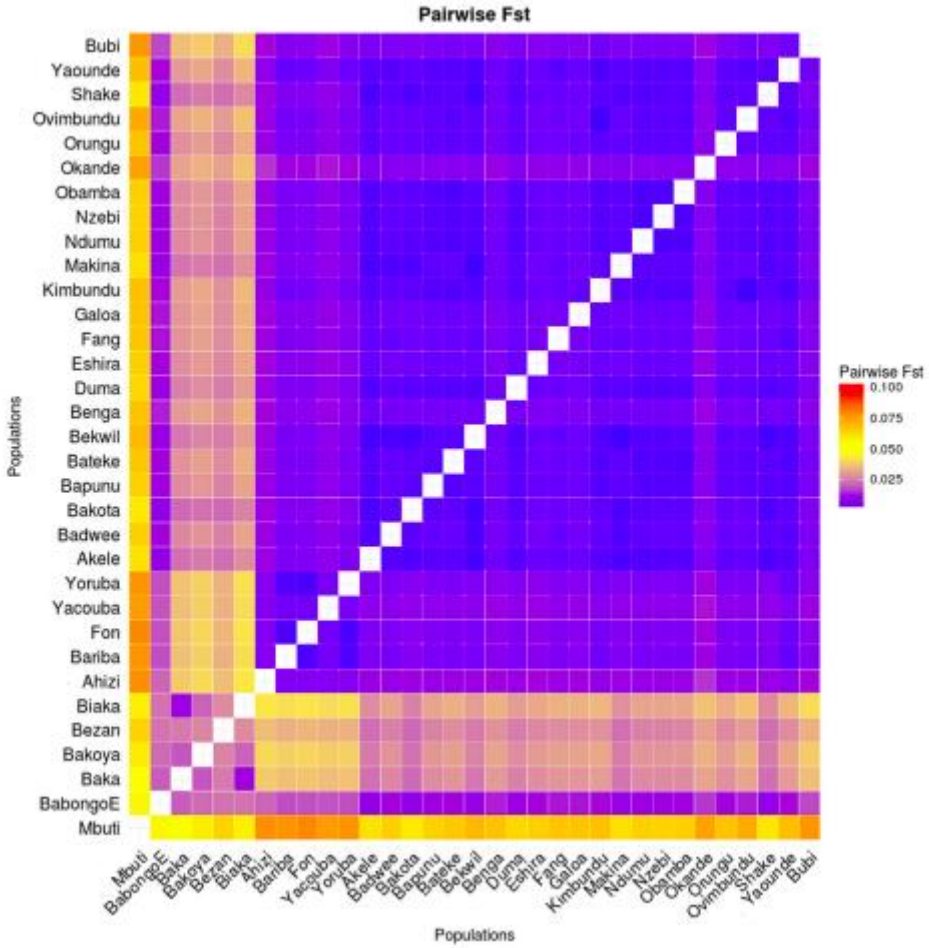


Figure 4. Matrix of pairwise Fst calculations. The Fst values have been calculated using plink and with a dataset of 592,395 SNPs.

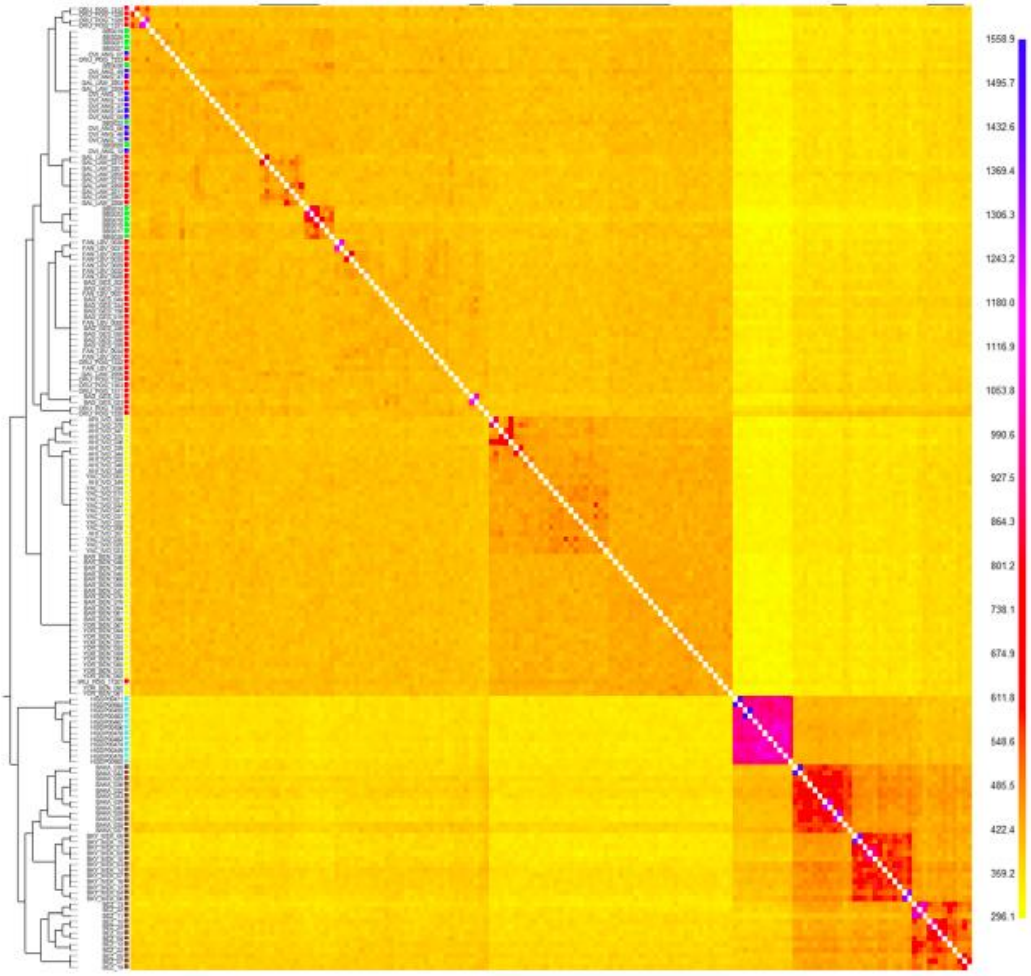


Figure 5. Matrix of shared counts of haplotypes obtained using fineSTRUCTURE. The dataset includes 491,203 SNPs and 169 individuals. The structure of the matrix has been adapted using the fineSTRUCTURE dendrogram. Green label represents the Bubi individuals, red label is used to represent the Gabon-Cameroon Bantu-speaking individuals, Blue label represents the Angolan Bantu-speaking individuals, Yellow label represents the Western-African individuals, Light blue label represents the Eastern Hunter-Gatherer individuals and Brown label represents the and Western Hunter-gatherer populations.

A B

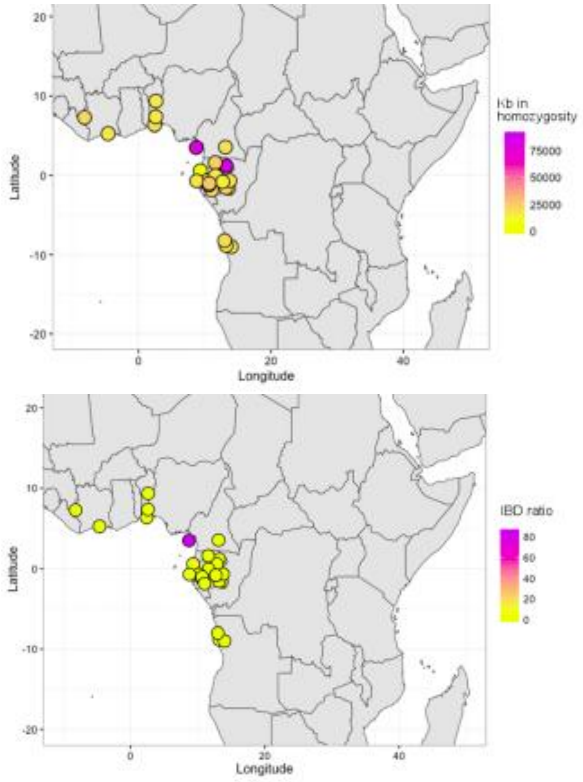


Figure 6. A Average of the genome in homozygosity (in kb) for the Bubi and mainland Bantu populations. B Average of the shared IBD genomic blocks between the Bubi and mainland Bantu populations (IBDs > 2 cM). The maps have been created with R software.

4.2. Mapping co-ancestry connections between the genome of a Medieval individual and modern Europeans

Manuel Ferrando-Bernal, Carlos Morcillo-Suarez, Toni de-Dios, Pere Gelabert, Sergi Civit, Antonia Díaz-Carvajal, Imma Ollich-Castanyer, Morten E. Allentoft, Sergi Valverde & Carles Lalueza-Fox.

Scientific Reports. 2020; 10 (6843)

DOI: https://doi.org/10.1038/s41598-020-64007-2

Abstract

Historical genetic links among similar populations can be difficult to establish. Identity by descent (IBD) analyses find genomic blocks that represent direct genealogical relationships among individuals. However, this method has rarely been applied to ancient genomes because IBD stretches are progressively fragmented by recombination and thus not recognizable after a few tens of generations. To explore such genealogical relationships, we estimated long IBD blocks among modern Europeans, generating networks to uncover the genetic structures. We found that Basques, Sardinians, Icelanders and Orcadians form, each of them, highly intraconnected sub-clusters in a European network, indicating dense genealogical links within small, isolated populations. We also exposed individual genealogical links -such as the connection between one Basque and one Icelander individual- that cannot be uncovered with other, widely used population genetics methods such as PCA or ADMIXTURE. Moreover, using ancient DNA technology we sequenced a Late Medieval individual (Barcelona, Spain) to high genomic coverage and identified IBD blocks shared between her and modern Europeans. The Medieval IBD blocks are statistically overrepresented only in modern Spaniards, which is the geographically closest population. This approach can be used to produce a fine-scale reflection of shared ancestry across different populations of the world, offering a direct genetic link from the past to the present.

Introduction

Many studies have demonstrated the existence of human population genetic structuring in Europe that correlates with geography; for instance, a two dimensional representation of the genetic variation with principal component analysis (PCA) essentially mirrors a geographical map of Europe (1,2). Several ancient DNA (aDNA) studies have shown that the overall genetic structure was shaped by three ancestral and over-imposed genomic components respectively deriving from the Mesolithic hunter-gatherers, the Early Neolithic farmers, and the steppe nomads that entered Europe from the East around 5,000 years ago (3–7). However, it is expected that the genetic homogenisation of the European populations during the last two millennia complicate our ability to discern subtle changes in ancestry by using the standard population genetic tools.

Complementary to these analyses, the distribution of so-called identity by descent (IBD) genomic stretches, which are co-inherited genetic segments delimited by recombination events, can provide information on more recent shared ancestry among individuals (8). Such genomic block characterization in current populations has demonstrated the presence of co-ancestry across geographically distant Europeans shared over the last few thousand years, and revealed recent co-ancestors living in neighboring populations (9). Nevertheless, most IBD blocks are not expected to be recognizable after a few hundreds of years because they are being broken by recombination during meiosis. Since the far majority of ancient human genomes sequenced to date are >2,000 years old and few of them are sequenced at high coverage, this analytical framework seems incompatible with the time scales offered by most published ancient DNA data.

To overcome this challenge, we used a reference dataset of genome-wide data from modern European individuals to explore recent genealogical IBD

structure among populations. We subsequently sequenced the genome of a 600 year-old Medieval skeleton from Barcelona (Spain) to high coverage, and used methods of graph theory to visualize the ancestry connections both among modern Europeans, and between modern Europeans and this Medieval individual. Networks can be used to analyze and visualize interactions between different types of elements (10) and have been used to represent different types of ecological and evolutionary interactions, e.g., between phages and their bacterial hosts (11). Here, we use a network framework to infer genomic similarity (and thus shared ancestry) between individuals and identify patterns of population demography. By combining ancient DNA data with the analysis of genomic blocks we establish, for the first time, the direct genealogical links between present-day people and an historical ancestor.

Results

IBD analysis in modern populations

To explore the general genetic structure of European co-ancestry, we filtered the Human Origins dataset (12) to analyze 429 individuals and 365,312 SNPs. We detected 1,249 genomic IBD blocks longer than 6 cM among pairs of European individuals (S2 Table).

Most IBD blocks are found within, and not between, populations (Fig. 1). Some inter population connections observed are plausible in the light of recent history (e.g., IBD tracks shared among Estonian, Russians, Lithuanian and Finnish individuals or IBD tracks shared among Icelanders, Orcadians and Norwegians).

We subsequently used a network representation of the IBD blocks distribution (Fig 2) similar to the approach previously applied for modern Europeans and African American exomes (13). Individuals in the plot

differ markedly in their connectivity, with some of them connected by IBD blocks to many individuals while others, the peripheral branches of the network, being connected to a single individual. This network displays community structure, i.e., the occurrence of groups of nodes (or modules) that are more densely connected internally than with the rest of the network. Individuals belonging to small and isolated populations such as Basques, Orcadians, Sardinians and Icelanders tend to constitute highly interconnected modules. Other potential modules such as here observed in Russians are likely explained by biased sampling because of high endogamy in a rather small and isolated community.

A total of 109 individuals are disconnected from the main network. Remarkably, all Sardinian individuals are among these and thus isolated from the rest of the continent, an observation that is in agreement with ancient genomic studies where Sardinians are shown to largely preserve the genetic legacy of early Neolithic farmers (4,5,14,15). Also, all Maltese individuals are in this situation, seven of them forming their own cluster along with a Sicilian individual.

Plotting only IBD connections involving selected regions we can see different co-ancestry patterns that might be indicative of differences in population's demography during the last hundreds of years. For instance, the Basque region shows a tight clustering of Basque individuals surrounded by a more disperse clustering of Spanish and French individuals (Fig. 3); the so-called Spanish-North, which derive from Alava - a region where Basque language was spoken in historical times- are located in an intermediate position between Basques and Spanish. Icelander, Norwegian and Orcadian individuals present almost exclusively intrapopulation connections with some connections between them, which is in agreement with what is known on the settlement of this island from the Atlantic North (16)(Fig. 4). By contrast, a region such as the south

Eastern Europe shows a much more mixed connectivity with a higher number of interpopulation connections involving quite a few different countries, which might be indicative of higher heterogeneity and more recent population movements (Fig. 5).

IBD analysis with a Medieval individual

To test for co-ancestry links back into the historical past we selected for genome sequencing a skeleton, T-145-2 (S1 Fig.), from the mid-XIVth century (17) excavated at the Medieval village of L'Esquerda near Roda de Ter (North of Barcelona, Catalonia) that was abandoned during the Plague epidemic (18,19). We extracted aDNA from the otic capsule of the petrous part of the temporal bone (20), constructed the library, and generated a total of 1,103,685,282 DNA reads. The putative endogenous human DNA accounted for 62.3% of the reads, after mapping them to the hg19 human reference genome. After removing duplicated reads and passing quality filtering steps, 543,173,362 unique reads remained, yielding a 11.3x depth of coverage for this genome. The postmortem damage at the 5' and 3' ends of the reads, which is a signal of DNA authenticity, was 26.4% and 17.3%, respectively. Contamination was estimated to be 1.8% by looking for discordant nucleotides at defining positions of the mtDNA K1C1 haplotype observed for the T-145-2 individual. With a method combining deamination patterns and fragment length distribution (21) we obtained a similarly low contamination estimate (0.5%-2.5%).

To our knowledge, only thirteen other ancient European genomes have been sequenced to a higher coverage: a Scandinavian Mesolithic individual (57.8x) (22), a Mesolithic individual from Loschbour (22x) and a Neolithic individual from Stuttgart (19x) (4), Hungarian Bronze Age and Neolithic individuals (21x and 22x, respectively) (20), a Copper Age individual from Spain (13x) (23), an Iron Age individual from Hinxton (11x)

(24), and six Longobards (12.86-14.48x)(25). With the current coverage on the Medieval L'Esquerda individual, it was possible to generate accurate genotypes calls (26), that were subsequently merged with SNPs from the Human Origins panel.

In a Principal Component Analysis (PCA) built with modern European individuals, our Late Medieval individual is placed in an undefined intermediate position in the representation of the first two axes, genetically close to modern day North Italians and also to Spanish and French populations (Fig. 6). In the ADMIXTURE analysis (K=4) four genomic components - represented by the ancestry of European hunter-gatherers, Early Neolithic farmers, Late Neolithic steppe nomads, and North African individuals - can be detected. We observe that the Late Medieval individual has more North African ancestry than the average in modern Iberians; the same trend is seen in other Middle Age individuals from the same region (7), suggesting there was a subsequent dilution of this component in more recent times (S2 Fig.).

We subsequently estimated the IBD blocks shared between our Medieval individual and the above-mentioned dataset of modern Europeans by using 280,690 single nucleotide polymorphisms (SNPs) that were shared between modern Europeans and medieval. We found a total of only 31 IBD tracks longer than 2 cM (S3 Table); 19 of them (61,3%) were shared with individuals from the Iberian Peninsula. As expected, a decreasing number of IBD blocks were found by raising the threshold: seven IBD blocks >3cM and only one >5cM (S3 Table); the latter was shared with a Catalan individual, thus coming from the same geographical region. No IBD blocks > 6cM were found. Although IBD blocks shared with our Medieval individual show a restricted geographical distribution (i.e. Spanish, Spanish North and Basques), there are still some that point to surprisingly long-distance connections to modern individuals. For instance,

there are three IBD segments with three Lithuanian individuals. However, the only significant overrepresentation of IBD blocks between the Medieval and modern population is with the Spanish (Fig. 7).

Discussion

Generally, European IBD blocks longer than 4 cM derive from common ancestors living 500-1500 years ago (9), which is roughly contemporaneous to our Late Medieval individual. The network based on >6cM genomic blocks in modern Europeans uncover some interesting genealogical features that are not evident in the commonly applied methods such as PCAs or ADMIXTURE. For example, these networks allow us to discriminate between small and possibly endogamous populations such as Icelanders or Orcadians versus large and highly dispersed populations that constitute the backbone of the network. A certain over-sampling of "interesting" populations can partially explain the modules, but down-sampling them does not change the pattern as these isolated populations are generally connected to the rest of the network by none or very few IBD links.

On the contrary, a more diffuse scattering of some populations across the network is informative of high genetic heterogeneity and more blurred co-ancestry links. For instance, the Greek and the Czech individuals are virtually dispersed along different branches, suggesting they come from regions where large population movements took place in the last hundreds of years.

The network also unravels individual connections that are unlikely to be exposed with other statistical approaches. For instance, in the European

network (Fig. 2) we found a cluster of seven Maltese individuals also connected to one Sicilian. This can be the product of the XIth century CE invasion of the island by Normands from Sicily. However, the Maltese cluster together in population genetic analyses such as PCA and Admixture because they share most of their overall ancestry (Fig 6). In another example, we found one Icelander that shares an IBD with a Basque individual (Fig 4). Basque whaling ships were common in the Icelandic Westfjords during the XVIIth century CE (27) and they even developed a Basque-Icelandic pidgin language for trading purposes (28). In principle it is not implausible that this is a signal that derives from a child conceived by an Icelander woman and a Basque sailor dating back to that period. Again, modern Icelanders, including this individual, cluster together in traditional analyses based on overall ancestry (Fig 6). In-depth genealogical and genetic analyses of modern Icelanders are required to confirm this finding but in any case it underlines the possibilities of the IBD approach to unravel genealogical connections across large geographical areas. Alternatively, the example of south Eastern Europe illustrates a more general population intermixing and a more complex demographic history in the last hundreds of years.

The Medieval individual shows a reduced number of IBD blocks with slightly shorter lengths than those shared between random modern European individuals from different populations (S8 and S9 Figs). The pattern of genomic block-sharing between the Medieval individual and modern European populations points to a restricted geographical clustering, with preferential co-ancestry links with present-day Iberians. This likely reflects a certain common ancestry and isolation among Iberians in the last hundreds of years. Interestingly, a similar observation was made with a different human genomic dataset (POPRES) (9). Moreover, the Late Medieval individual displays a limited but significant number of geographically distant relationships that point to a ubiquitous co-ancestry. This confirms previous inferences based on genomic data

from modern populations that showed presence of more diffuse co-ancestry links as we go back in time (9).

Both types of ancestry -ubiquitous and geographically-proximate - are present in the IBD results arising when using both the modern, contemporaneous European as reference and when using the Late Medieval genome as reference point. Going back in time, the IBD blocks shared with modern Europeans are shorter but seem to be also more geographically clustered compared to those found among present-day individuals, with the exception of the previously mentioned isolated populations. If additional Medieval genomes are available in the future, it is likely that more locally restricted IBD blocks could be identified across Europe.

The power of our approach to uncover individual genetic affinities in otherwise homogeneous populations is illustrated by comparing with a PCA analysis that includes our Medieval genome. In this standard analysis, it was not possible to attribute with certainty a geographic affinity of our individual that was occupying a central position between Iberians, French and North Italians (if anything, closer to the latter). However, the IBD analysis clearly placed it among modern Spanish. Remarkably, no IBD tracks are shared between the Medieval and North Italians.

With the current level of productivity in ancient genomic research, many more individuals from the recent historical past will be sequenced to high genomic coverage in the near future. This will allow us to extend this methodological framework to an increasingly large genomic population dataset of ancient and modern people. This approach will serve not only to uncover individual ancestry links, but will also unravel the origins and the spread of mutations subjected to positive selection, because this process should preserve longer genomic blocks than expected under a process of random recombination (29). In essence, such data will help visualize an

extended family tree with a vast, interconnected and complex network that will link the past genetic landscape with the present one.

Material and methods

The site

L'Esquerda is an archaeological site placed in an area of 12 hectares on cliffs overlooking a narrow meander of Ter river, near Roda de Ter (North of Barcelona). Due to its privileged geographical position it has been continuously occupied from the Late Bronze Age to the Middle Ages (18). The Visigothic settlement was temporarily abandoned during the Muslim occupation until the Carolingian times, when it was settled again, in parallel to the Frank conquest of Girona in 785 C.E. The Medieval village grew up around the church of Sant Pere de Roda, built in the XIth century over a previous, smaller church. A walled area around the church was destined for cemetery, where three main stratigraphic layers can be observed: a basal one from the Visigothic and Carolingian periods, an intermediate one with slab-stone burials dated between the XIth and the last XIIIth century and a superficial one from the final occupational period of the settlement. Despite L'Esquerda was destroyed and abandoned in 1314 for a new location close to the river, the final use of the burial area consists on communal graves dated to the first two-thirds of the XIVth century associated to the epidemics of 1348 and subsequent years (19).

The individual analyzed, labeled T-145-2, corresponds to a young (15-16 years-old) female that was excavated in the seasons 2009-2010 in one of the XIVth century's graves, along with another adolescent and an adult male that were buried simultaneously. A right petrous bone was selected for DNA extraction, due to better chances of DNA preservation (20).

DNA extraction and sequencing

The petrous portion of the temporal bone was sliced open using an electric diamond-coated cutting blade allowing us to remove and crush the otic capsule for DNA extraction (30). The DNA was extracted using a silica-in-solution method optimized for retaining short and degraded DNA molecules (6). First, a 15-min enzymatic pre-digestion step was implemented to reduce the amount of exogenous DNA (31). The samples were then incubated for 24 hours at 45ºC in 5 ml digestion buffer containing 4.7 ml 0.5 M EDTA, 50 µL Proteinase K (0.14-0.22 mg/ml, Roche), 250 µL 10% N-Laurylsarcosyl, and 50 µL TE buffer (100x). The solution was spun down and the supernatant transferred to a 50 ml tube, where it was mixed with 100 µl silica suspension and 40 ml binding buffer, prepared as in Allentoft et al. (2015) (6). After 1 hour of incubation, the supernatant was removed and the pelleted silica was re-suspended in 1 ml binding buffer, spun down and washed twice with 1 ml 80% cold ethanol. Finally, the DNA was eluted in 80 µl EB buffer (Qiagen). Extraction blanks were also included. Next, 2*20 µl of DNA extract was prepared as blunt-ended, double-stranded libraries using Illumina-specific adapters and the NEBNext DNA Sample Pre Master Mix Set 2 (E6070) kit, as described previously(6), except that we here used the KAPA HiFi HotStart Uracil + ReadyMix (KAPA Biosystems, Woburn, MA, USA) in the amplification step. The two index-amplified DNA libraries were purified and quantified on an Agilent Bioanalyzer 2100. The DNA extraction and library preparation (pre-amplification steps) were conducted using strict aDNA guidelines in a sterile clean lab at Centre for GeoGenetics at the Natural History Museum of Denmark. The libraries were sequenced (80 bp, single end sequencing) on an Illumina HiSeq 2500 platform at the Danish National High-throughput DNA Sequencing Centre.

Mapping Procedure

Sequencing-adapters were trimmed with Cutadapt 1.3 (32). The clipped reads were mapped against the human reference genome (hg19) and the revised Cambridge Reference Sequence (rCRS) (33) using BWA aln (34) setting no seeding, no read trimming, an edit distance of 0.01 and a gap open penalty of 2. Afterwards, duplicated reads were removed with Picard tools 2.18.6 (35). Finally, unique reads were filtered with SAMtools 1.6 (36) keeping only those with mapping qualities over 30. The mapped and filtered reads were analyzed with MapDamage 2.0.8 to determine the post-mortem aDNA damage pattern (37). Because the final sequences present a deamination percentage of 26.4% and 17.3% in the 5' and 3' ends respectively (S3 Fig), which could affect the variant calling, we trimmed 5 bases at each end using trimbam 1.0.13 (38).

Contamination estimates

The average level of DNA contamination was estimated by genotyping the mitochondrial DNA haplogroups with Haplogrep2 (39) and calculating the ratio of discordant reads with a homemade script. Modern mitochondrial contamination was estimated using *schmutzi (21)*.

Variant Calling

Genotypes were called with the Genome Analysis Tool Kit (GATK) v3.7, as previously described (40), using UnifiedGenotyper and a correction for the observed contamination (--contamination_fraction_to_filter 0.02232), --output_mode EMIT_ALL_CONFIDENT_SITES (this option is important to ascertain which nucleotide position displays the reference allele), the Human Genome 37 / 19 as the reference. We genotyped 616,938 positions present in the Human Origins dataset (12). Subsequently, variants with base qualities below 30 and genotype quality below 20 were discarded from the called bases by using VCFtools v0.1.14 (41). We

decided to filter out those variants displaying a lower coverage than the average depth of coverage of each chromosome. Despite the medium-high coverage of most of the genome, this procedure should in principle be enough for a confident variant calling. Filtered variants were merged with the Human Origins dataset using PLINK v1.9b (42). Only SNPs present in autosomal chromosomes were used in the analyses.

Mitochondrial haplogroup and molecular sex assignment

Mitochondrial genome variants were called with Genome Analyses Toolkit (GATK) UnifiedGenotyper (40), setting the same parameters used in the autosomal variant calling procedure. The mitochondrial haplogroup was assessed using Haplogrep 2 (39). Molecular sex was assigned using the methodology used in (43).

Datasets Preparation

From David Reich Lab datasets web page (12) we downloaded genotypes for 616,938 SNPs and 433 individuals belonging to 29 European populations (S1 Table). We removed a total of 121,699 SNPs that did not fulfill our quality control criteria (Genotype missingness < 5%, Hardy-Weinberg equilibrium test p-value > 1x10-6, MAF > 0). After plotting genotype missingness of the individuals against its heterozygosity (S4 Fig), three individuals with particularly extreme values were removed. Relatedness analysis of pairs of individuals discovered a family relationship. One of the members of this pair was also removed.The resulting filtered dataset (495,239 SNPs and 429 individuals) was used to study IBD relationships between modern Europeans.

A second dataset was generated adding the ancient individual and containing only those SNPs from the European Dataset that were also recovered from the ancient sample. It consisted of 369,859 SNPs and 430

individuals. The Medieval individual presented a high heterozygosity compared to modern individuals (S5 Fig). This dataset was used for studying IBD relationships between the Medieval individual and modern Europeans.

Population genetic analyses

A principal component analysis (PCA) was built, with 495,239 SNPs and present-day 429 European individuals from the previously described dataset, using Eigensoft (44,45). The resulting data was plotted using R package Ggplot2 (46,47).

An admixture analysis was performed with ADMIXTURE (48). We selected 881 individuals from West Eurasia and North Africa, and a dataset of 16 ancient individuals known to be representatives of the main ancestry components of modern Europeans, and our ancient individual T-145-2 (12). We then filtered the dataset by removing SNPs in linkage-disequilibrium (LD) using PLINK 1.9 flag --indep-pairwise with a windows size of 200 SNPs, advanced by 50 SNPs and establishing an r^2 threshold of 0.4(42). We performed the ADMIXTURE analysis, using 242,622 SNPs, with K ranging from 2 to 15 and performing 10 replicates for each run. We selected the K in accordance to the lowest cross-validation mean value (48) as well as the fact that we wanted to observe the four main European ancestry components: Western Hunter-Gatherers, Early European Farmers, Steppe Nomads and Northern Africans (12,48). We plotted ADMIXTURE results using package pophelper (49).

Haplotype Estimation

Phase of genotypes for both datasets was estimated without imputation of unknown genotypes with Beagle 5.0 software (50). The recombination map and the haplotype reference panel provided in the Beagle publication were used.

IBD Discovery in modern Europeans

We examined the data for IBD segments between all pair of individuals using the Refined IBD software (51) with the parameter "minimum length for reported IBD segments" set at 1 cM. A total of 289,561 IBD segments were identified with 1,523 being longer than 6 cM and were selected for further analysis.

Within the set of IBD segments longer than 6cM we studied the distribution of scores, the relationship between length and score and the coverage along the genome to discover potential false IBD blocks. We removed IBD segments that, after visual examination, clustered outside the main distribution with abnormal low scores considering their length. We also removed IBD segments overlapping centromeric and telomeric regions (S6 Fig). After quality control 1,249 IBD blocks longer than 6 cM remained.

Triangulation Analysis

To assess the validity of the IBD segments longer than 6cM among modern European individuals, we checked for the transitivity of the IBD relationships between trios of individuals. If individuals A and B share a IBD segment and individual B, in the same chromosome, shares the same IBD segment with individual C, we expect A and C to share it too (S7 Fig). We grouped overlapping IBD segments into clusters. Two IBD segments were considered to overlap when they shared at least one individual and were located in the same chromosome overlapping at least by one pair base. We obtained, as predicted, clusters consisting of trios of individuals

sharing IBD segments in a transitive way, together with clusters showing different arrangements (S7 Fig).

We plotted the chromosomes involved in some of the clusters to understand the origin of these clusters and defined some typologies (S7 Fig). We listed all the pairs of overlapping IBD segments longer than 6cM, classified them into the previously defined typologies and looked for the third member of the expected transitive triangulation Only 3 pairs out of 220 remained unexplained and point to a possible artifact in the IBD discovery process..

Graphical Representation of the Networks

To visualize the IBD relationships among the different populations of modern Europeans, we coded the estimated IBD segments previously obtained into a network structure where individuals correspond to nodes that are connected if at least one shared IBD region between them exists. We then plotted the resulting global network and some particular groups of populations using the NetworkX Python package, version 2.2 (52).

IBD Discovery in the Medieval Individual

The dataset containing the European plus the mMedieval individuals underwent the same process of IBD discovery described for the European dataset. A total of 242,158 IBD segments were generated with 1,472 longer than 6 cM. IBD blocks with score values lower than 4.8 and close to centromeric and telomeric regions were removed. From the 1,164 remaining, 31 IBD blocks between the Medieval and some other European individuals longer than 2 cM were selected for further study.

Population enrichment

To test if the Medieval individual presents a significant enrichment of IBD segments with particular populations, we assumed as a null hypothesis that the expected total number of IBD segments shared by the medieval with a given population should be proportional to the number of individuals of that population.

The tests performed with the Medieval are likely to increase type I error rates and also are not independent; however, the dependence among tests are expected to reduce the extent of alpha inflation (53). In similar cases of block-positive dependence among tests, it has been shown that a best option to control for false discovery rate (FDR) (54)is to use the two stage Benjamini-Hochberg (TSBH) procedure(55). We subsequently adjusted the p-values between observed and expected IBD blocks for the TSBH procedure; a nominal type I error rate (5%) was used to estimate the number of true null hypotheses in the two-stage TSBH with R multtest package (56).

Acknowledgements:

This research was supported by a grant from Obra Social "La Caixa", FEDER-MINECO (PGC2018-095931-B-100) of Spain to C.L.-F, by a grant from MINECO (FIS2016-77447-R) to S.C. and by 2017SGR 00622 grant from Generalitat de Catalunya's Agency (AGAUR) to S.C. Sequences from the Medieval genome are deposited at the European Nucleotide Archives under accession number PRJEB33120.

References:

1. Novembre, J., Johnson, T., Bryc, K., Kutalik, ZZ., Boyko, A.R., Auton, A., et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101 (2008).
2. Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., et al. Correlation between Genetic and Geographic Structure in Europe. *Curr Biol*. 18(16):1241–8. (2008):
3. Olalde, I., Allentoft, M.E., Sánchez-Quinto, F., Santpere, G., Chiang, C.W.K.K., DeGiorgio, M., et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, 507(7491):225–8 (2014).
4. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–13. (2014).
5. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–11 (2015).
6. Allentoft, M.E. Population genomics of Bronze Age Eurasia. *Nature*, 522:207–11 (2015).
7. Olalde, I., Mallick, S., Patterson, N., Rohland, N., Villalba-Mouco, V., Silva, M., et al. The genomic history of the Iberian Peninsula over the past 8000 years. *Science*, 363(6432):1230–4 (2019).
8. Ringbauer, H., Coop, G. & Barton, N.H. Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–51 (2017).
9. Ralph, P. & Coop, G. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biol*., 11(5):e1001555 (2013).
10. Newman, M.E.J. Networks : an introduction. Oxford; New York: Oxford University Press; 772 p. (2010).
11. Weitz, J.S., Poisot, T., Meyer, J.R., Flores, C.O., Valverde, S., Sullivan, M.B., et al. Phage-bacteria infection networks. *Trends Microbiol*., 21(2):82–91 (2013).

12. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., et al. Genotypes of ancient individuals and fully public Affymetrix Human Origins present-day individuals. Genomic insights into the origin of farming in the ancient Near East (2016).
13. Fu, W., Browning, S.R., Browning, B.L., Akey, J.M. Robust Inference of Identity by Descent from Exome-Sequencing Data. *Am J Hum Genet*. 99(5):1106–16 (2016).
14. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503 (2015).
15. Allentoft, M.E., Sikora, M., Sjögren, K.G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. Population genomics of Bronze Age Eurasia. *Nature*, 522(7555):167–72 (2015).
16. Ebenesersdottir, S.S., Sandoval-Velasco, M., Gunnarsdottir, E.D., Jagadeesan, A., Guethmundsdottir, V.B., Thordardottir, E.L., et al. Ancient genomes from Iceland reveal the making of a human population. *Science*, 360(6392):1028–32 (2018).
17. Ollich, I. & Mestres, J. Datació per Radiocarboni de material ossi d'origen humà procedent del sector medieval de l'Esquerda (Les Masies de Roda, Osona). In: L'Esquerda, àrea medieval Memòria de les Excavacions 2009-2010 a la necròpolis sud Inedit (2010).
18. Ollich, I., Ocaña, M, Ramisa, M. & Rocafiguera, M. A banda i banda del Ter, Història de Roda. Ajuntament de Roda de Ter, editor. Roda de Ter: Eumo Editorial; 271 p. (1995).
19. Ripoll, G., Molist, N. & Ollich i Castanyer, I. La necròpolis medieval de l'Esquerda (segles VIII-XIV dC). Cronologia i noves perspectives de recerca. In: Arqueologia funerària al nord-est peninsular (segles VI-XII), Monografies d'Olèrdola, 32 Museu d'Arqueologia de Catalunya, Barcelona. p. 275–86. (2012).
20. Gamba, C., Jones, E.R., Teasdale, M.D., McLaughlin, R.L., Gonzalez-Fortes, G., Mattiangeli, V., et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun*., 5:5257 (2014).

21. Renaud, G., Slon, V., Duggan, A.T. & Kelso J. Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol*. 16(1):1–18 (2015).

22. Günther, T., Malmström, H., Svensson, E.M., Omrak, A., Sánchez-Quinto, F., Kılınç, G.M., et al. Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol*., 16(1):e2003703 (2018).

23. Valdiosera, C., Günther, T., Vera-Rodríguez, J.C., Ureña, I., Iriarte, E., Rodríguez-Varela, R., et al. Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. *Proc Natl Acad Sci U S A*., 115(13):3428–33 (2018).

24. Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., et al. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat Commun*., 7:10408. (2016).

25. Amorim, C.E.G., Vai, S., Posth, C., Modi, A., Koncz, I., Hakenbeck, S., et al. Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat Commun*., 9(1):3547 (2018).

26. Bryc, K., Patterson, N. & Reich, D. A novel approach to estimating heterozygosity from low-coverage genome sequence. Genetics, 195(2):553–61(2013).

27. Huxley S. Los vascos en el marco Atlántico Norte: siglos XVI y XVII. In: Echebarria EA, editor. Volumen 3 de ITSASOA : El mar de Euskalerria La naturaleza, el hombre y su historia. Donostia-San Sebastián: ITSASOA; p. 1–336 (1988).

28. Deen, N.G.H. Glossaria duo vasco-islandica. H.J. Paris, editor. 119 p. (1937).

29. Albrechtsen, A., Moltke, I., Nielsen, R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308 (2010).

30. Fernandes, D., Sirak, K., Novak, M., Finarelli, J.A., Byrne, J., Connolly, E., et al. The Identification of a 1916 Irish Rebel: New Approach for Estimating Relatedness from Low Coverage Homozygous Genomes.

Sci Rep., 7:41529. (2017).

31. Damgaard. P.B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E. & Allentoft, M.E. Improving access to endogenous DNA in ancient bones and teeth. *Sci Rep*. 5:11184 (2015).

32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* 17(1):10–2 (2011).

33. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M. & Howell, N. Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nat Genet.,* 23(2):147 (1999).

34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60 (2009).

35. Broad Institute. Picard [Internet]. Available from: http://broadinstitute.github.io/picard/

36. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9 (2009).

37. Jónsson, H., Ginolhac, A.A., Schubert, M., Johnson, P.L.F.F., Orlando, L., Jonsson, H., et al. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. In: Bioinformatics. England, England; p. 1682–4. (2015).

38. Bam Util [Internet]. 2015. Available from: https://github.com/statgen/bamUtil

39. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 44(1):58–63 (2016).

40. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.,* 20(9):254–60 (2010).

41. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E.,

DePristo, M.A., et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–8 (2011).

42. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.,* 81(3):559–75 (2007).

43. Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J Archaeol Sci*. 40(12):4477–82 (2013).

44. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*., 38(8):904–9 (2006).

45. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.,* 2(12):e190 (2006).

46. Team R. R: A language and environment for statistical computing (Version 3.4. 2)[Computer software] [Internet]. Vienna, Austria: R Foundation for Statistical Computing. Vienna, Austria (2017). Available from: http://www.r-project.org/

47. Wickham, H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; (2016). Available from: http://ggplot2.org

48. Alexander, D.H. & Novembre, J. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res*., 19(9):1655–64 (2009).

49. Francis, R.M. pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour.*, 17(1):27–32 (2017).

50. Browning, S.R. & Browning, B.L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am J Hum Genet*. 81(5):1084–97 (2007).

51. Browning, B.L. & Browning, S.R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*., 194(2):459–71 (2013).

52. Hagberg, A., Swart, P. S. & Chult, D. Exploring network structure,

dynamics, and function using networkx. Available from: https://networkx.github.io/

53. Winer, B.J., Brown, D.R. & Michels, K.M. Statistical principles in experimental design. New York: McGraw-Hill; (1991).

54. Stevens, J.R., Al Masud, A. & Suyundikov, A. A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests. *PLoS One*, 12(4):1–12 (2017).

55. Benjamini, Y., Krieger, A.M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507 (2006).

56. Pollard, K.S., Dudoit, S. & van der Laan, M.J. Multiple Testing Procedures: the multtest Package and Applications to Genomics. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York, NY: Springer New York. p. 249–71 (2005).

Figures

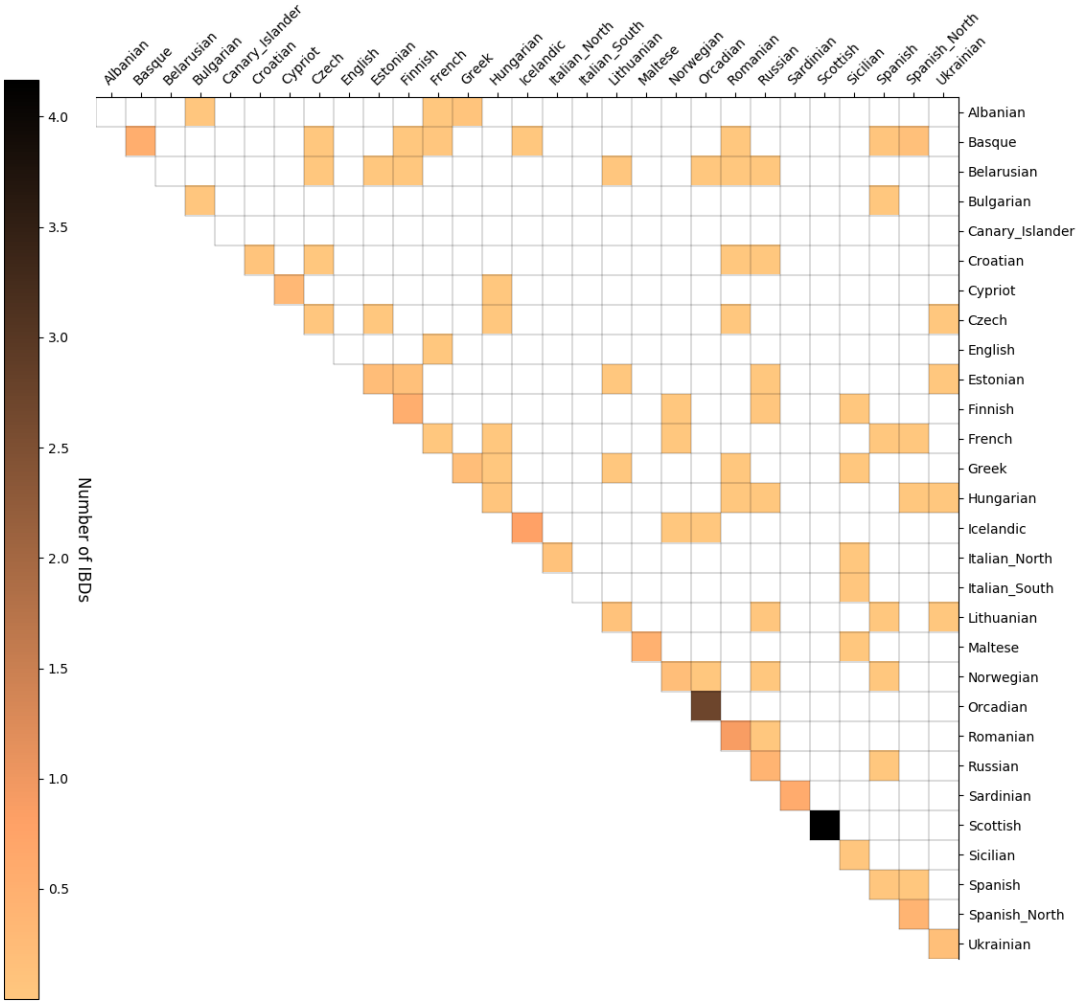


Fig 1. Mean number of IBD segments shared by a pair of individuals belonging to two given populations in the modern European dataset. It can be observed that most IBD blocks tend to fall within and not between populations, with some exceptions.

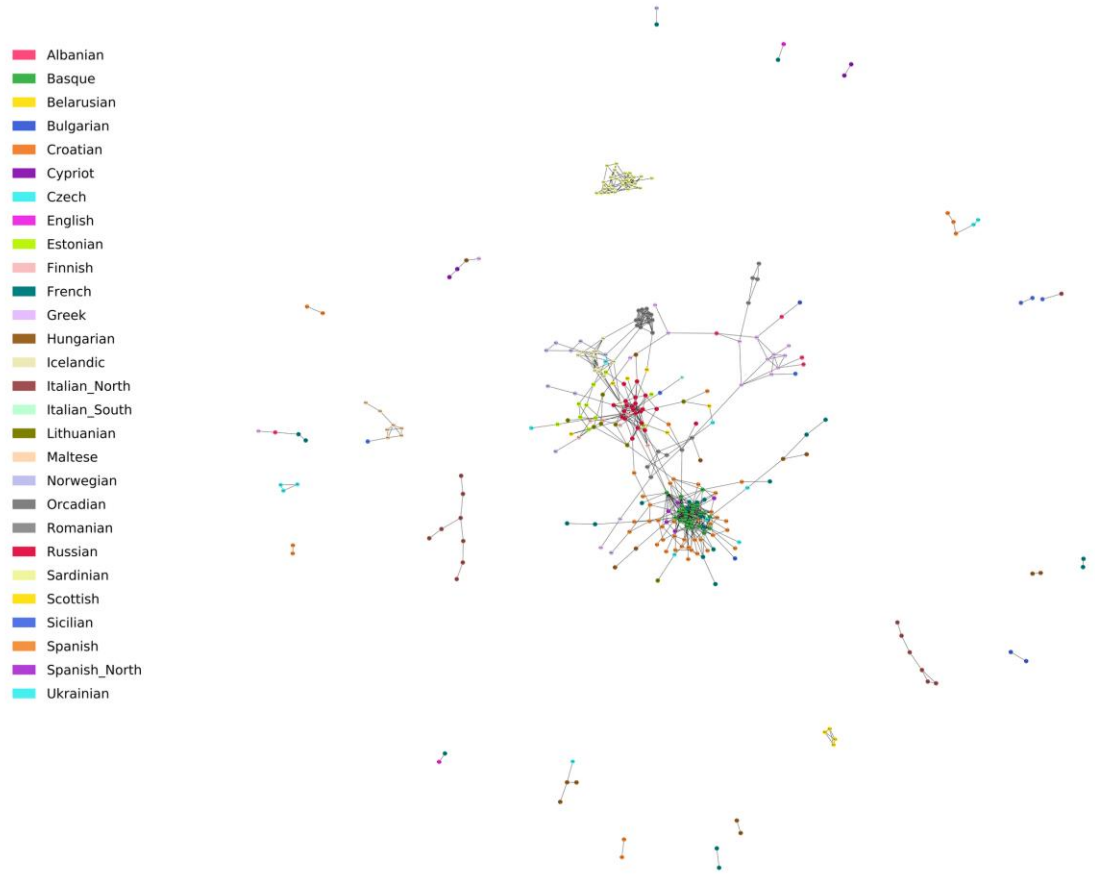


Fig 2. Network of identity by descent (IBD) blocks longer than 6 cM shared between present-day Europeans (nodes). There is a link between any pair of individuals sharing at least one IBD. Individuals are labeled according to the population of the Human Origins.

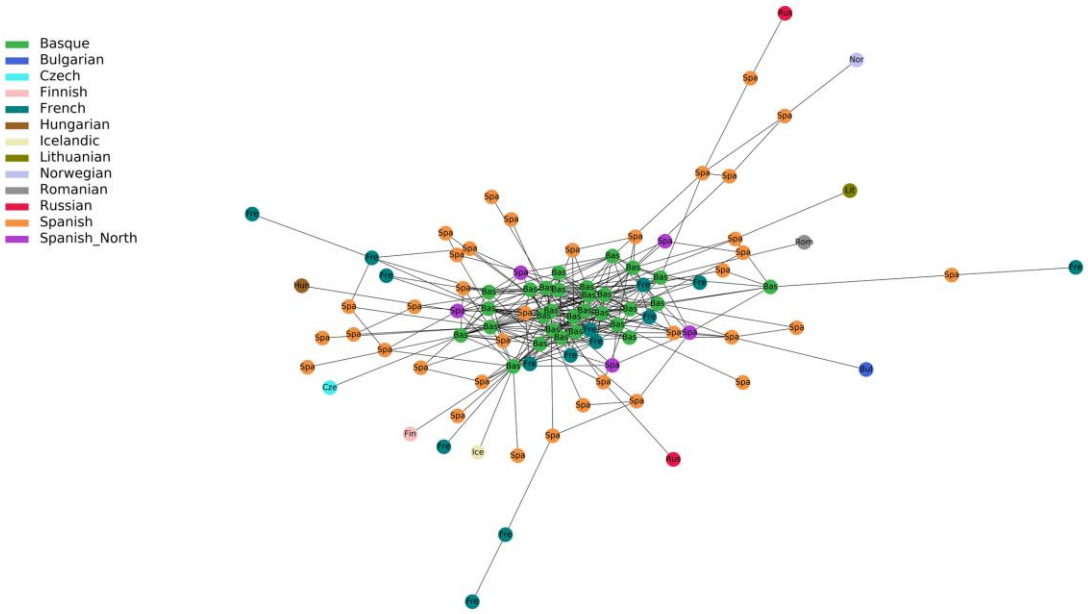


Fig 3. Regional network of present-day Baques. Includes Basques and individuals sharing IBD blocks with Basques.

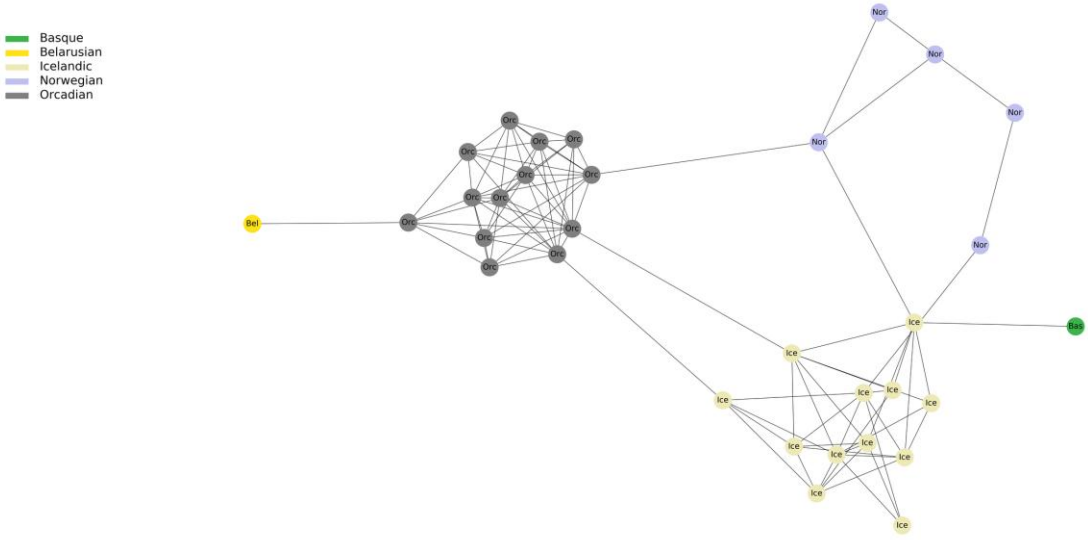


Fig 4. Regional network of present-day North Atlantic Europeans. Includes Icelandic, Norwegian and Orcadian and individuals sharing IBD blocks with those populations.

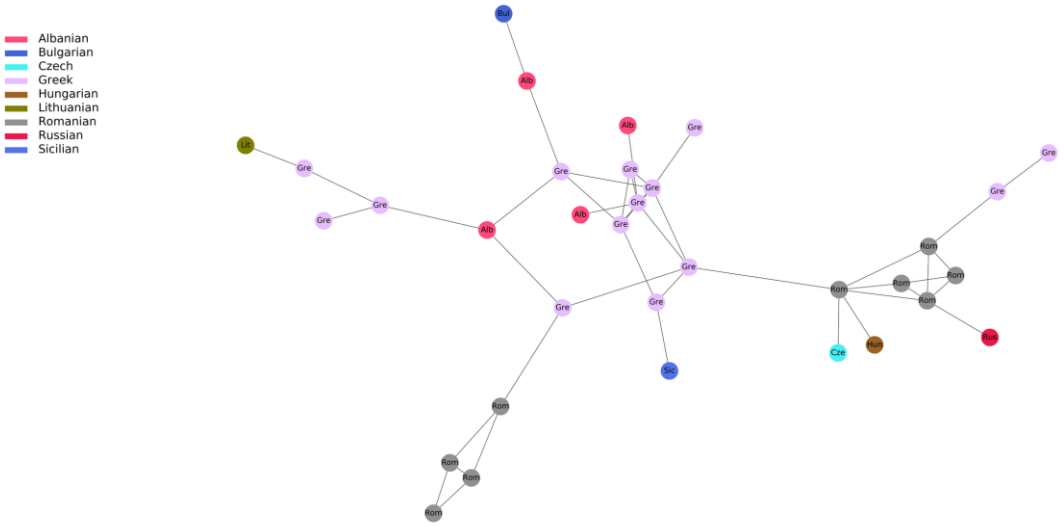


Fig 5. Regional network of present-day south Eastern Europeans. Includes Greeks, Romanians, Albanians and Bulgarians and individuals sharing IBD blocks with those populations.

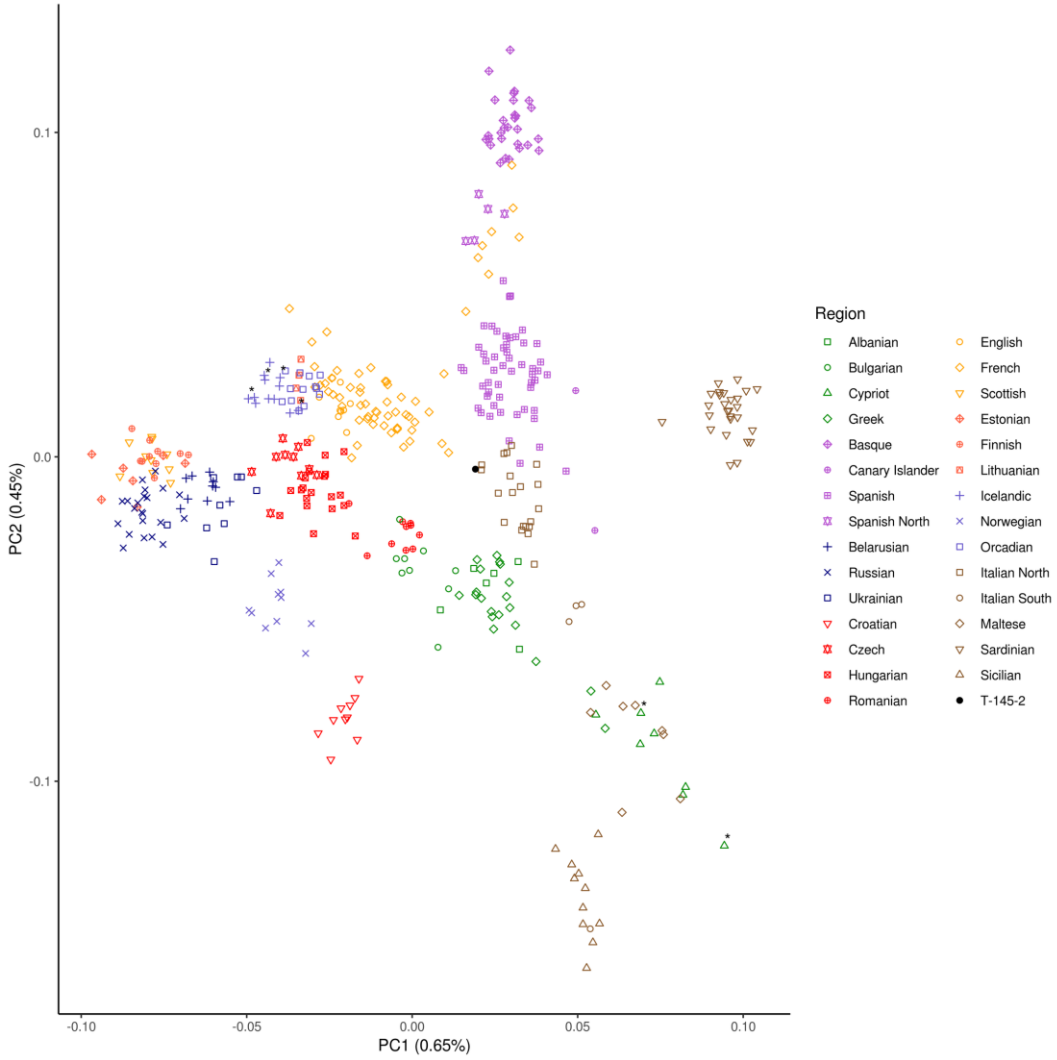


Fig 6. PCA with Human Origins' modern Europeans including the medieval individual.

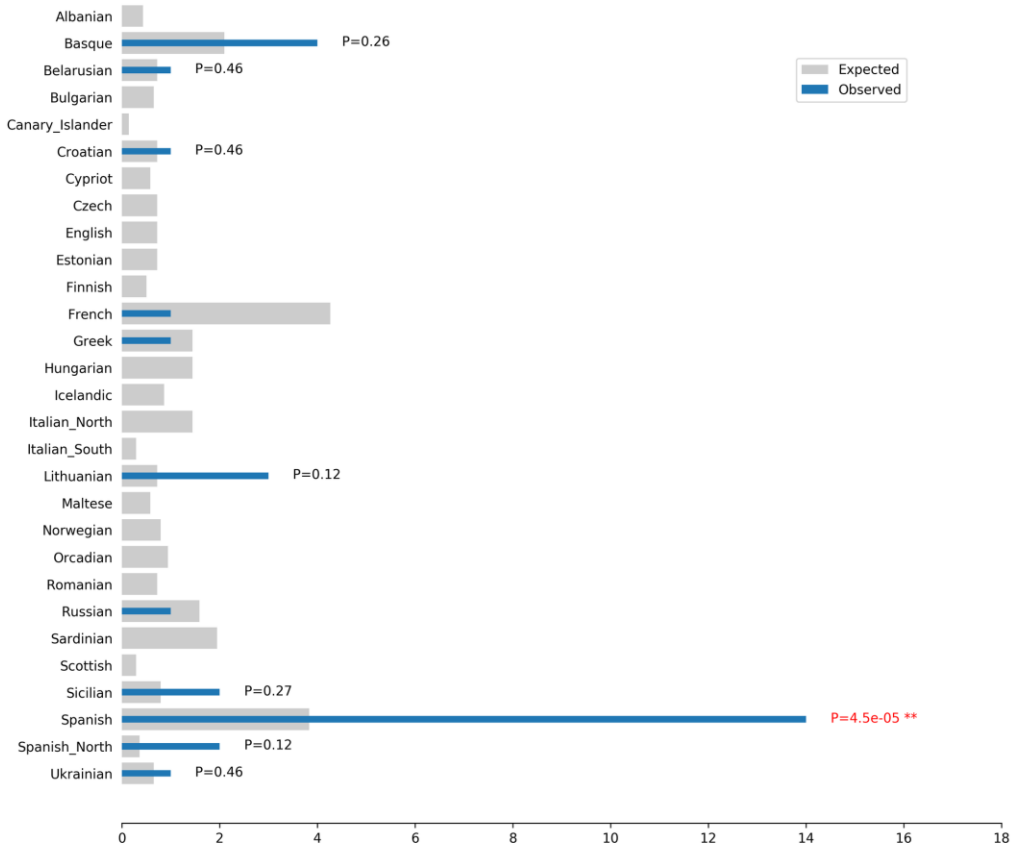


Fig 7. Observed and expected IBD segments between the medieval individual and individuals of each population. Expected values are generated assuming no correlation between the population to which the individuals belong and its probability of presenting an IBD segment with the medieval individual. The distribution of IBD segments for each population was assumed to follow a binomial distribution. The p-values were adjusted with a Benjamini-Hochberg procedure to control for false discovery rate.

5. DISCUSSION

In this section I will summarize the results of applying IBD analyses in order to find co-ancestral connections among homogeneous populations. Then I will discuss the future directions on the applicability of these methods on aDNA projects.

5.1. General overview

The aim of this thesis was to investigate recent connexions between human populations that are genetically close (genetically homogeneous populations).

Most commonly used genomic tools applied to aDNA field focused on samples older than 2,000 y.a. Through this time it is possible that these populations generate enough genetic variation to be able to be differentiated by these genomic tools. However, when we compare individuals separated a few generations ago (for example present-day samples with individuals that lived a few hundred y.a.) the results are not accurate or could be even contradictory.

The work for this thesis was made possible thanks to the development of genomic software able to detect shared haplotypes among apparently unrelated individuals (IBD methods) and the available genomic databases with a representative number of populations of the African and the European continents.

5.2. The origins of the Bubi population

The first work of this thesis focused on the origins and evolutionary history of the Bubi population from the Bioko island in the west coast of the African continent. In agreement with historical data, our results confirmed that the Bubi population derived from the Bantu group and a possible long standing isolation of the Bubis from the rest of the continental Bantu populations.

We applied IBDs analyses to detect the most genetically close Bantu population from the mainland African coast. We also applied IBDs and ROHs analyses to quantify the genomics imprints of a possible increase of endogamy due to the insular particularities of this population in comparison to the continental ones.

Our results indicated that the Bubi shared IBD longer than 2cM with populations from Angola and Gabon. The genomic similarity to the Angolan populations is also supported by other analyses (for example fineSTRUCTURE or f3 statistics) and suggest that the ancestral populations of the Bubi splitted from the ancestor of the Angolan Bantu populations. This is intriguing as the mainland populations geographically close to the Bioko island are the populations from Cameroon.

The Bubi is the Bantu population with the highest amount of IBD shared inside the same population. Also the Bubi is one of the Bantu populations with an increased rate of ROHs in its individuals. Both results are compatible with the expected amount of endogamy as a consequence of a long isolation.

Finally, we saw that there are no tracks of genetic drive in the Bubis after their split of the Bubi from the other bantu populations, suggesting that these 2,000 years of isolation may have not been time enough to generate important genomic differences between these populations. In consequence, and after detecting IBDs longer than 2 cM (which may came from the last 2,500 y.a.) among the Bubis and some mainland Bantu populations, our analyses confirm that the IBD studies can be used to detect genomic affinities and past migrations in homogeneous populations.

5.3. Unravelling ancestral connections among modern and ancient Europeans

The second work of this thesis aims to unravel genetic structures of modern European populations that cannot be detected with commonly used genomic softwares. IBDs analyses were used in this work to identify genealogical links among the European individuals dating from the last millenia. Additionally, these analyses were applied with a late medieval individual who lived a few centuries ago.

Here we highlight that by using a network representation of our results, we can differentiate the population substructure underlying the IBDs connexion among both present-day Europeans, and among them with the medieval sample. It allows us to detect clusters of individuals belonging to isolated populations, or some residual connexions among distant geographically individuals, as for example among a Basque and an Icelander individual. Something that could be difficult to detect with previous methodologies, as for example PCA or ADMIXTURE that foccus in the whole genome, whereas IBDs methodology split up the whole genome in different haplotypes blocks, and analyze them independently from each other.

Regarding the data, and in agreement with Ralph, P. & Coop, G. 2013, our results show that most of the European shared haplotypes can be found among individuals belonging to the same population or with individuals from geographically close populations. Evenmore, our results show that the effects of insulation can be detected in some European populations as Icelanders, Orcadians or Sardinians, or confirm other studies that reflect an important degree of isolation in the Basque populations compared to the surrounding ones. Additionally, and as we have seen, IBDs analyses can be used to find genealogical links among individuals belonging to geographically distant populations, that cannot be easily detected with other most widely used softwares.

On the other hand, IBDs analyses have been rarely used in aDNA samples due to they cannot be detected after some generations due to they are broken during the recombination events, and most of the aDNA sequenced samples are older than 2,000 y.a. In this work, after the sequencing of an individual that lived a few generations ago, we show that IBD can be used to find their genealogical connexions with modern individuals. This can be used to give in a higher scale detail the shared ancestry among ancient and present-day individuals and to link the past genomic landscape with the actual one.

In conclusion, our results confirm that IBDs analyses can be used in homogeneous populations, like modern Europeans and among modern and ancient samples, which can set some light about past human migrations that are difficult to be detected with most of the software commonly used in the aDNA field.

5.4. Future research and perspectives on the applicability of IBD methods in aDNA studies

As we saw, IBDs can be used to find co-ancestral relationships among homogeneous populations like populations that split a few generations ago. Evenmore, we show that they can be applied to detect genomic relationships among present-day individuals and ancient samples from the last few centuries. However, to date, there are less than two tens of ancient human samples sequenced to a coverage which may allow the applicability of IBD methods. And apart from it, most of these samples are too old that recombination must have broken all the possible IBD among them and present-day individuals.

Taking this into account, we think that future studies should focus on generating good quality genomes from ancient individuals that allows us to apply IBD softwares, specifically from the last few thousand of years. For example, Shotgun sequencing allows to recover most of the individuals sequence if there is enough genomic material. However, shotgun technology is still highly expensive, specifically for samples with low quantities of DNA, something that characterizes the aDNA material.

Is for that reason that genomic imputation could rise as an alternative solution. Genomic imputation refers to the statistical inference of the unknown parts of a particular genome based on the known haplotypes of its population. This would allow to increase the quality of low coverage genomes (which is the most of the aDNA sequenced to date). Fortunately, it is possible from SNP genotyping, which as a consequence of being less expensive than Shotgun sequencing technology, is the most worldwide commonly used option to sequence the ancient samples.

Once more aDNA samples are generated and publicly available it would allow the study of genomic relationships among populations that are hidden for most of the present-day genomic populations methods. And to conclude, we want to highlight three possible uses of applying IBDs methods to aDNA studies: first, it would allow the detection of past migrations among homogeneous populations. Second, it would allow us to get deep into the knowledge of the isolation effects of island populations or in highly endogamous populations, as they used to be in some past societies. And finally, it would allow a fine scale characterization of the genomic structure of some past populations by detecting sharing IBDs among ancient samples belonging to the sample age and populations, as for example, among the individuals that characterize the male substitution in European neolithic communities related to the arrival of the Pontic-Caspian steppe (Olalde, I. et al. 2019), or among the individuals belonging to endogamous lineages as for example the dominating class in some societies in Ireland (Cassidy, L.M. et al. 2020).

6. CONCLUSIONS

1. IBD detection is a powerful tool to discover co-ancestral relationships among apparently unrelated individuals that cannot be detected by the standard methods of population genomics.

2. IBD fragments are more commonly found among individuals belonging to the same populations, specifically in isolated ones with a higher amount of endogamy (for example, in islander populations such as the Bubi, the Sardinians, Orcadians, Icelanders or Malteses).

3. As IBD methods can be used to find ancestral connexions among individuals from geographically distant populations, it allows them to differentiate among ancient individual contacts or ancient migrations.

4. Good quality genome and a high SNPS density are needed for IBD analyses as they allow to reduce the false discovery rate. If more good coverage ancient genomes are generated, IBD methods would be used to uncover genomic connexions among present-day individuals with the people of our past.

Contribution to other publications

**Badiane, A., Carazo, P., Price-Rees, S.J., Ferrando-Bernal, M. & Whiting, M.J. 2018. Why blue tongue? A potential UV-based deimatic display in a lizard. Behav. Ecol. Sociobiol. 72, 104.*

**Olalde, I., Mallick, S., Patterson, N., Rohland, N.,Villalba-Mauco, V. et al. 2019. The genomic history of the Iberian Peninsula over the past 8000 years. Science 363, 6432.*

**Gelabert, P., Sandoval-Velasco, M., Serres, A., de Manuel, M., Renom, P. et al. 2020. Evolutionary History, Genomic Adaptation to Toxic Diet, and Extinction of the Carolina Parakeet. Current Biology 30, 1.*

**Gokhman, D., Nissim-Rafinia, M., Agranat-Tamir, L., Housman, G., García-Perez, R. et al. 2020. Differential DNA methylation of vocal and facial anatomy genes in modern humans. Nat. Commun. 11, 1189.*

**Ferrando-Bernal, M. & Lao, O.* 2021. *Obligate parthenogenesis in Reptiles and Genotypic Sex Determination Systems Ectotherms and Endotherms could be adaptations to extreme temperature environments. Submitted to Genes.*

Bibliography

Albrechtsen, A., Korneliussen, T.S., Moltke, I., van Overseem Hanse, T., Nielsen, F.C., et. al. 2009. Relatedness Mapping and Tracks of Relatedness for Genome-Wide Data in the Presence of Linkage Disequilibrium. *Genetic Epidemiology* 33: 266-274.

Albrechtsen, A., Mokte, I. & Nielsen, R. 2010. Natural Selection and the Distribution of identity-by-Descent in the Human Genome. *Genetics* 186: 295-308.

Alexander, D.H., Novembre, J., & Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19 (9): 1655-1664.

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L. et al. 2012. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. Biol. Sci.* 279 (1748): 4724-4733.

Allentoft, M.E., Sikora, M., Sjögren, K-G., Rasmussen, S., Rasmussen, M. et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* 522: 167-172.

Andy Choo, K.H. 1998. Why Is The Centromere So Cold? Genome Research 8: 81-82.

Bongers, J.L., Nakatsuka, N., O'Shead, C., Harper, T.K., Tantaleán, H. et al. 2020 Integration of ancient DNA with transdisciplinary dataset finds strong support for Inca resettlement in the south Peruvian coast. *PNAS* 117: 18359-68.

Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R. et al. 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *PNAS* 110(29): 11791-11796.

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J. et al. 2007. Patterns of damage in genomic DNA sequences from a Neanderthal. PNAS 104(37): 14616-14621.

Browning, S.R. & Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097

Browning, S.R. & Browning, B.L. 2010. High-Resolution detection of Identity by Descent in unrelated individuals. *Am. J. Hum. Genet.* 86(4): 526-539.

Browning, B.L. & Browning, S.R. 2011. A fast, powerful method for detecting Identity by Descent. *Am. J. Hum. Genet.* 88(2): 173-182.

Browning, S.R. & Thompson, E.A. 2012. Detecting rare variants associations by Identify-by-Descent Mapping in Case-Control Studies. *Genetics* 190:1521-1531.

Cassidy, L.M., Moaldúin, R.O., Kador, T.,Lynch, A., Jones, C. et al. 2020. A dynastic elite in monumental Neolithic society. *Nature* 582: 384-88.

Chambers GK. 2013. Genetics and the origins of the Polynesians. In: eLS. John Wiley and Sons Ltd: Chichester.

Chen , L., Wolf, A.B., Fu, W., Li, L. & Akey, J.M. 2020. Identify and Interpreting apparent Neanderthal Ancestry in African individuals. *Cell* 180: 1-11.

Dabney, J., Meyer, M. & Pääbo, S. 2013. Ancient DNA Damage. *Cold Spring Harb Perspect Biol* 5 (7).

Dixit, S.P., Singh, S., Ganguly, I., Bhatia, A.K., Sharma, A. et al. 2020. Genome-wide Runs of Homozygosity revealed selection signatures in *Bos indicus. Front. Genet.* 11:92.

de-Dios, T., van Dorp, L., Charlier, P., Morfopoulou, S., Lizano, E. et al. 2019. Metagenomic analysis of a blood stain from the French revolutionary Jean-Paul Marat (1743-1793). *Infection, Genetics and Evolution.* 89.

Enard, D. & Petrov, D. A. 2018. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* 175: 360-371.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., et al. 2014. The genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514 (7523): 445-449.

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S. et al. 2016. The genetic history of Ice Age Europe. *Nature* 534: 200-205.

Glocke, I. & Meyer, M. 2017. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res*. 27 (7): 1230-1237.

Green, R., Krause, J., Wriggs, A.W., Maricic, T.,Stenzel, U. et al. 2007. A Draft Sequence of the Neandertal Genome. *Science* 328 (5979): 710-722.

Greenbaum, G., Getz, W.M., Rosenberg, N.A., Feldman, M.W., Hovers, E. et al. 2019. Disease transmission and introgression can explain the long-lasting contact zone of modern humans and Neanderthals. *Nat commun* 10: 5003.

Gusev, A., Palamara, P.F., Aponte, G., Zhuang, Z., Darvasi, A. et al. 2011. The architecture of Long-Range Haplotypes Shared within and across Populations. Mo. Biol. Evol. 29(2): 473-486.

Gusev, A., Kenny, E.E., Lowe, J.K., Salit, J., Saxena, R. et al. 2011. DASH: A method for Identical-by-Descent haplotype mapping uncovers Association with Recent Variation. *J Hum Genet:* 706-717.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S. et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522: 207-211.

Hagelberg, E., Cox, M., Schiefenhövel, W. & Frame, I. 2008. A genetic perspective on the origins and dispersal of the Austronesians. Mitochondrial DNA variation from Madagascar to Easter Island. In: Sanchez-Mazas, A., Blench, R., Ross, M., Peiros, I. & Lin, M. (eds). Past human migrations in East Asia: matching archaeology, linguistics and genetics. London, New York: Routledge.

Han, L. & Abney, M. 2013. Using identity by descent estimation with dense genotype data to detect positive selection. *Am. J. Hum. Genet.* 21: 205-211.

Han, B., Yong Kang, E., Raychaudhuri, S., de Bakker, P.I.W., Eskin, E. et al. 2014. Fast pairwise IBD association testing in genome-wide association studies. Bioinformatics, 30 (2): 206-213.

Harris, K. & Nielsen, R. 2013. Inferring Demographic History from a Spectrum of Shared Haplotype Length. *PloS Genet. 9 (6).*

Higgins, D., Rohrlach, A.B., Kaidonis, J., Townsend, G. & Austin, J.J. 2015. Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies. *PloS One* 10 (5).

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O.A., Wilson, A.C. 1984. DNA sequences from the quagga, an extinct member of the horse family. *Nature* 312: 282-284.

Huerta-Sánchez, E., Jin, X., Asan, Biamba, Z., Peter, B. et al. 2014. Altitude adaptation in Tibet caused by introgression of Denisovan-like DNA. *Nature* 512: 194-197.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A. et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851-61.

Jobling, M.A., Hollox, E., Hurles, M., Kivisild, T. & Tyler-Smith, C. 2014. Into new-found lands. In: Human Evolutionary Genetics, 2th edition. New York: Garland Science.

Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., et al. 2015. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications* 6 (1): 1-8.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F. 2006. World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*. 15: 259-263.

Lalueza-Fox, C., Römpler, H., Caramelli, D., Stäubert, C., Catalano, G. et al. 2007. A Melanocortin 1 Receptor allele suggests varying pigmentation among Neanderthals. *Science* 318: 1453.

Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S. et al. 2008. Correlation Between Genetic and Geographic Structure in Europe. *Curr. Biol.* 18 (16): 1241-1248.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S. et al. et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513: 409.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D., Rohland, N. et al. 2017. Genomic insights into the origin of farming in the ancient Near East. *Nature* 236 (7617): 419-424.

Liu, CC., Shringarpure, S., Lange, K. & Novembre, J. 2020. Exploring Population Structure with Admixture Models and Principal Component Analysis. In: Dutheil J. (eds) Statistical Population Genomics. Methods in Molecular Biology, 2090. Humana, New York, NY.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M. et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538: 201–206.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N. et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528: 499-503.

Marciniak, S. & Perry, G.H. 2017. Harnessing ancient genomes to study the history of human adaptation. *Nat Rev Genet* 18: 659-674.

Meyer, M., Kircher, M., Gansauge, M-T., Li, H., Racimo, F. et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338 (6104): 222-226.

Mittnik, A., Wang, C-C., Pfrengle, S., Daubaras, M., Zarina, G. et al. 2018. The genetic prehistory of the Baltic Sea region. *Nature Communications* 9: 442.

Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. 2005. A Fine-Scale Map Recombination Rates and Hotspots Across the Human Genome. Science 310 (5746): 321-324.

Nakatsuka, N., Lazaridis, I., Barbieri,C., Skoglund, P., Rohland, N. et al. 2020. A Paleogenomic Reconstruction of the Deep Population History of the Andes. *Cell* 181: 1131-1145.

Nägele, K., Posth, C., Orbegozo, M.I., Chiniqie de Armas, Y., Hernández Godoy, S.T. Et al. 2020. Genomics insights into the early peopling of the Caribbean. *Science* 369 (6502): 456-460.

Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R. et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. hum. Genet.*: 347-58.

Nordborg, M. & Tavaré, S. 2002. Linkage disequilibrium: what history has to tell us? Trends in Genetics 18(2): 83-90.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R. et al. 2008. Genes mirror geography within Europe. *Nature* 456: 98-101.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S. et al. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* 10 (4).

Olalde, I., Allentoft, M.E., Sánchez-Quinto, F., Santpere, G., Chiang, C.W.K. et al. 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* 507 (7491): 225-228.

Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristianse, K., et al. 2018. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* 555: 190-196.

Olalde, I., Mallick, S., Patterson, N., Rohland, N., Villalba-Mouco, V., et al. 2019. The genomic history of the Iberian Peninsula over the past 8000 years. *Science*, 363(6432):1230–4.

Pääbo, S. 1985. Molecular cloning of Ancient Egyptian mummy DNA. *Nature* 314: 644-645.

Pääbo, S. 1989. Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci* 86: 1939-1943.

Patin, E., Lopez, M., Grollemund, R., Verdu, P, Harmant, C. et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356 (6337): 543-546.

Patterson, N. Moorjani, P., Luo, Y., Mallick, S., Rohland, N. et al. 2012. Ancient Admixture in Human History. *Genetics* 192 (3): 1065-1093.

Purcell, S. Neale, B., Todd-Brown, K., Thomas, L. Ferreira M. et al. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Gen* 81.

Racimo, F., Sikora, M., Vander Linden, M., Schroeder, H., & Lalueza-Fox, C. 2020. Beyond broad strokes: sociocultural insights from the study of ancient genomes. *Nat Rev Genet* 21(6): 355-366.

Ralph, P. & Coop, G. 2013. The Geography of Recent Genetic Ancestry across Europe. *PloS Biol.* 11(5).

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, A. et al. 2011. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science* 334 (6052): 94-98.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N. et al. 2010. Genetic history of an archaic hominid group from Denisova Cave in Siberia. *Nature* 468: 1053-1060.

Rimbauer, H., Steinrücken, M., Fehren-Schmitz, L. & Reich, D. 2020. Increased rate of close-kin unions of the central Andes in the half millennium before European contact. *Curr Biol* 30 (17): 980-981.

Rodríguez-Varela, R., Günther, T., Krzewinska, M., Stora, J., Gillingwater, T.H. et al. 2017. Genomic Analyses of Pre-European Conquest Human Remains From the Canary Island Reveal Close Affinity to Modern North Africans. *Curr. Biol.* 27 (21): 3396-3402.

Saada, J.N., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M. et al. 2020. Identity-by-Descent detection across 487,409 British samples reveals fine scale populations structure and ultra-rare variant associations. *Nat. Comm*. 11: 6130.

Sankararaman, S., Mallick, S., Dannemann, M. Prüfer, K., Kelso, J. et al. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507: 354-357.

Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E. et al. 2016. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications* 7, 10408.

Shen, B., Jiang, J., Seroussi, E., Liu, G. E. & Ma, L. 2018. Characterization of recombination features and the genetic basis in multiple cattle breeds. *BMC Genomics* 19:304.

Sikora, M., Pitulko, V.V., Sousa, V.C., Allentoft, M.E., Vinner, L. et al. 2019. The population history of northeastern Siberia since the Pleistocene. *Nature* 570: 182-188.

Skoglund, P., Posth, C. Sirak, K., Spriggs, M., Valentin, F. et al. 2016. Genomic insights into the peopling of the Southwest Pacific. *Nature* 538: 510-513.

Tabangin, M.E., Woo, J.G. & Martin, L.J. 2009. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc* 3: S41.

The 1000 Genome Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B., Babbitt, C.C. et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39(1): 31-40.

Vicente, M. & Schlebusch, C. M. 2020. African population history: an ancient DNA perspective. *Current Opinion in Genetics & Development*. 62: 8-15.

Wade, L. 2015. Breaking a tropical taboo. *Science* 349:370-371.

Weir, B.S. & Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution Int. J. Org. Evolution* 38: 1358-1370.

Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. 2005. A note on exact tests of Hardy-Weinberg Equilibrium. *Am J Hum Gen* 5: 887-893.

Yang, M.A., Fan, X., Sun, B., Chen, C., Lang, J. et al. 2020. Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 369: 282:288.

Zhang, M. & Fu, Q. 2020. Human evolutionary history in Eastern Eurasia using insights from ancient DNA. *Current Opinion in Genetics & Development*. 62: 78-84.

Weitz, J. 2014. Let my people go (home) to Spain: A genealogical model of Jewish identities since 1492. *PloS Gen*. 9 (1).