UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN COMPUTER ENGINEERING

Analysis of Human Genomic Data to Unveil pre Columbian Admixture Events in the Antilles

Relatore: Prof. Matteo Comin

Laureando: Pietro Toso

Correlatore: Prof. Luca Pagani

ANNO ACCADEMICO 2022 – 2023

Data di laurea 03/04/2023

To all the people, places and
moments that welcomed
my joy and sadness

To my family, to sincere friends
and to unexpected encounters

Thank you

Abstract

In this thesis the possibility of pre-Columbian contacts between the populations of the Antilles and of the Mediterranean area was explored through a series of genomic analyses aimed at better characterizing the interactions between these populations. These analyses focused both on ancient DNA coming from the Antilles, looking for traces of European DNA inside it, and on modern DNA of the same area, performing an ancestry deconvolution and analyzing the European component in detail through different methods. No particular signs supporting pre-Columbian contacts were found, but there remains space for more detailed analysis depending also on the available data.

Index

1 – Introduction

In this section the idea from which this thesis generated will be illustrated in detail, followed by a summary of the genetic history of the Caribbean area that is necessary to give context to this work.

1.1 - 'L'America dimenticata' and Tolomeo's mistake

The idea of this thesis comes from Prof. Lucio Russo's book '*L'America dimenticata'* (Russo, 2013), in which the author reflects upon the possibility of encounters between the 'new' and the 'old' world before Columbus' arrival in 1492.

1.1.1 - Russo's perspective on cultural evolution

In the first part of his essay Russo illustrates the contrast between neo-evolutionist and diffusionist theories within the realm of cultural evolution. The first ones support the idea that different civilizations evolved independently and that some stages of cultural evolution like writing, agriculture and social structures are inherent in mankind and can happen several times in different places in the same order. These theories try to describe the evolution of civilizations as particular cases of a general model governed by laws. In this perspective, the study of cultures that are in early stages of this evolution can represent a key tool to understand these laws. Diffusionist theories, instead, emphasize the importance of contacts between different civilizations and believe that key innovations are one-time events that happened in a specific culture and spread through the world reaching further and further populations with the passing of time. Russo shows in his book through several examples and argumentations that it is unlikely that the earliest civilizations of Nilo, Indo and Mesopotamia emerged independently, and that trades and cultural exchanges were much more influent than what is commonly thought. These contacts may have determined also the further development of cultures in Eurasia through time. One of the examples of cultural diffusion is the wheel, or rather wheeled carts, which emerged in Eurasia around the middle of the fourth millennial BC. This mean of transport emerged almost simultaneously in a really vast area, suggesting a one-time invention that quickly spread through different cultures. In addition to this, civilizations of sub-Saharan Africa and Australia were not able to re-invent them independently and did not know them until they later met Europeans or Arabs.

Diffusionist theories may have been obstructed in a way through time because they lead to the risk of being used to claim the superiority of certain cultures. This is not the case if one sees the development of cultures in the correct perspective. One could disregard, for example, Australian Aboriginal's technology as they could not come up with the idea of wheeled carts, but looking it the other way around all the other cultures were not able to invent the boomerang. The author argues that it is not a matter of superiority, but of number of interactions. In his perspective, the more interactions one culture has with other ones, the more its complexity will increase. To show this he points out the case of Tasmanian populations, which lived in a complete isolation for around 10000 years and did not move towards an increasing complexity but remained at a 'simple' level from a technological perspective.

There are many similarities between the path of evolution followed by cultures in Mesoamerica and cultures in Eurasia. The combination of this with the absence of contacts between these cultures is the strongest proof in support of neo-evolutionistic theories. Cultures of the so called 'old world' may have influenced each other though cultural exchanges but, under the assumption that pre-Columbus Mesoamerican never met them, diffusion cannot explain how cultures developed in the 'new world' in a form and a timeline that is quite similar to the other ones. Proving the fact that these two worlds came in contact with each other before and during the development of Mesoamerican cultures, which are more recent, could be a key discovery in favor of diffusionist theories according to the author. In this light, Russo firstly provides several anecdotical examples of possible evidence of pre-Columbus contacts between these cultures, like some metallurgy techniques and some games that resemble each other in the different cultures or some archeological findings, but all of them can be interpreted in ways that are somehow compatible with the absence of contacts.

1.1.2 – A mathematical geography riddle

The author then moves to more solid scientific grounds and focuses on an issue based on mathematical geography that constitutes the core of his book, showing that there are different elements in favor of the idea that, already in the second century BC, Greeks sources knew with a high degree of precision the position of the Lesser Antilles in the Caribbean area. This claim, if confirmed, would leave no doubts about the contacts between the two cultures. This theory is based on Tolomeo's *Geographia,* the only antique work of mathematical geography still available today in a usable form and dating to the end of first century BC, which, according to Russo, suffered from a serious and systematic bias that is at the core of his theory.

1.1.2.1 – Tolomeo's cultural context

To understand this possible bias some premises are needed. One key element in this narrative is the alleged cultural collapse that affected the Mediterranean world starting from the biennial 146-145 BC, following the change of foreign policies of the Roman Empire. According to the author this process is largely underestimated and its effects are much bigger than what it is thought. In those years Romans decided to extend the area they directly controlled, trying to eliminate any other autonomous entity in the Mediterranean world. This approach was carried out in different ways towards different populations. Carthage, for example, was completely wiped out and with that almost all its heritage was gone as well. The few remains available today, like the work of Magone in 28 books about agriculture, suggest that Carthaginians were more advanced than Romans in different fields, but Romans decided not to preserve their knowledge either because they did not understand it or because they were not interested in it. Romans' attitude towards Greece was different, since they did not actualize a complete destruction, but absorbed almost all of it under their direct control. With the Hellenistic states the approach was less invasive. With Egypt, for example, Romans decided not to take direct control, but to transform it into an undirect domain by interfering in the dynasty successions. As a consequence, Alessandria's library, probably the greatest cultural center of the time, underwent a huge decline caused by the lack of investments and interest of the Romans in this area. According to Russo, these events caused a general and sudden fracture in the process of cultural evolution, since most of the intellectuals found themselves without resources and just some of them were able to continue their activity by working for important Roman families, even if the interest in these cases was not towards scientific fields, but rather towards literature and history. Nonetheless these contacts generated a slow process of acculturation of a niche of the Roman society, even if the effects of the collapse are still evident in the intellectuals that lived just a couple of centuries after these events. Russo reports some passages in which, as far as geography is concerned, it is clear that there was a lack of methodology and comprehension of terminology with respect to the previous cultural world. Scientists and intellectuals in general kept using the same terms of their predecessors to give a sense of continuity, but without understanding them in many cases. Russo underlines how this process is generally underestimated and how historians tend to show this historical period as something continuous, without any important fracture. In his way of seeing it, instead, the gap in the way people understood the world around them was huge and this may have led to misunderstandings and mistakes by Latin scientists, as in the case of Tolomeo.

Another important point to clarify to understand the proof is the level of knowledge of geography reached in the Hellenistic world before the events of 146-145 BC and the impact that the cultural collapse had on it. Mathematical geography is a peculiar product of the Hellenistic culture and it consists in elaborating a mathematical model of Earth in order to make predictions on real measures and to be able to draw maps. A proper mathematical model for Earth presupposes the knowledge of Earth's sphericity, a discovery that was only made in the Hellenistic world. The process started in the sixth century BC with Anassimandro, who gave birth to the idea that objects do not fall downwards, but rather towards the earth. This idea was further developed through time and particularly in the third century BC with Eratostene. This thinker, who was also the director of Alessandria's library for 22 years, was particularly interested in geography, a term created by him in his work *Geographica*, and gave a great contribute to the development of mathematical geography. Indeed, in this work he already conceived each place on earth as a point on a spherical surface, identifiable through latitude and longitude. His most famous achievement is the measurement of earth's dimensions through the well-known method that exploits the ratio between the round angle and the angle formed by the sun's rays with respect to the vertical axis at midday of the summer solstice in a specific location. Once this ratio is calculated, it is sufficient to know the distance between that location (Alessandria in Egypt in this case) and the Tropic of Cancer along the same meridian to obtain the length of earth's circumference. The hardest part of this process is to obtain a precise measure of the distance along the meridian. The measure obtained by Eratostene for the circumference reported by almost all the sources is of 252 000 stadiums, corresponding to 700 stadiums per degree. The measure of the stadium used by Eratostene proposed by Russo is 157.5 meters, derived by a passage written by Plino in which the 'Eratosthenis ratio' is reported. This value, accepted by many experts, would give a measure of earth's circumference of 39 700 km with an error of 0.75%, an accuracy that suggests that the distance between Alessandria and the Tropic of Cancer was not just an estimate or an approximate measurement, but rather a value obtained through a campaign of detections carried out throughout Egypt. In addition to this, we have the words of Ipparco, reported by Strabone in his *Geographia*, who admits that he would not have been able to improve the value obtained by Eratostene. These words acquire great value if Ipparco's figure and his work are analyzed. Ipparco lived in the second century BC and reached a high level of knowledge in many fields, including astronomy and geography. He was able to continue his work after the events of 146-145 BC until 126 BC (the date of his last astronomic observation) thanks to his location, the isle of Rhodes, that managed to put off Roman's raid by remaining loyal to the Empire. Ipparco's level of precision can be understood from his measurement of the distance earth-moon, almost exact, but there are many other

examples. He was also the one who invented a procedure to obtain the difference of longitude between different places on the same parallel by exploiting the observations of the same eclipse event. Working with latitudes was much easier at that time with respect to longitudes, since there where different methods to evaluate them with accuracy like observing the maximum duration of light hours or measuring the angle formed by the vertical with the celestial north pole direction. To evaluate longitudes instead, or rather differences of longitudes since it is not an absolute value like latitude, it was necessary to measure the time interval between the observation of the same astronomic event in two different places along the same parallel. This was quite hard and Eratostene himself preferred not to give relative longitude coordinates in degrees, but rather to report the distances between places on the same parallel. Ipparco disapproved Eratostene's method and that is why he proposed his idea to calculate differences of longitude with high precision. Knowing all this and understanding Ipparco's high standards for precision in this field, the fact that he admitted that he would not have been able to improve Eratostene's value of earth's circumference suggests that the measurement made by the latter was likely obtained through means that were not available to the first one, like the campaign of detections mentioned before.

This mathematical approach towards geography and many of these notions and methods were lost after the cultural collapse. Already in the works of the first century BC one can witness the abandonment of the spherical coordinates and the return to a geography that is more descriptive than mathematical. In Strabone's *Geopgraphia*, one of the few works of the time available today, it is evident how the author struggles in following the reasoning of his sources and how often he does not understand them. Going on with time, Plinio's work *Naturalis Historia* is an example of how the concept of Parallel has been lost, it becoming a sort of physical zone that includes Tuscany and Puglia at the same time. According to Russo, not only the mathematical approach towards geography was lost, but also the knowledge of peoples and countries outside of the Mediterranean area, with the Roman Empire that ended up closing on himself. There are passages of Erodoto that talk about the hypothetical circumnavigation of Africa by the Phoenicians and other reports of their expeditions along Atlantic shores, while it is known that after the cultural collapse no one was able to repeat this challenge until Vasco de Gama in 1497. One can witness in general a change of attitude towards the 'outside' world and the ocean, seen by the Romans as something scary and full of monstruous creatures.

1.1.2.2 Tolomeo's bias

Now that the background situation is clearer, at least according to what Russo describes, it is possible to analyze the issue regarding Tolomeo's *Geographia*. Tolomeo, in the second century AD, is the first one that tries to recover the Hellenistic mathematical approach towards geography. The disconnection between Tolomeo and his sources is evident in his astronomic work *Almagesto*, in which there is a huge temporal gap in terms of observations between 126 BC (the last one made by Ipparco) and 92 AD. Tolomeo is often distant from his sources and their methods and he is not always able to fully understand them. This leads to two main mistakes that Russo analyzed: a wrong value assigned to the Earth's dimensions and a systematic deformation of longitude differences. The latitude values reported by Tolomeo, at least the one regarding well-known locations at that time, do not seem to be affected by relevant systematic errors. Longitude difference values, on the other hand, seem to be systematically dilated. This is evident when looking at the map of Italy in Figure 1 drawn according to his measures.

Figure 1: Map of Itay drawn according to Tolomeo's coordinates (Germanus)

Russo analyzes longitude values by considering only the ones related to 80 locations well known to the Hellenistic sources, in order to avoid errors of other nature, like identification of remote places or precise position, and to focus only on the systematic error. He then applies a linear regression trying to find the straight line that better approximates the points that represent the locations, each of which has coordinates x=real longitude starting from Greenwich meridian and y=longitude reported by Tolomeo starting from the 'Isole Fortunate', a group of islands in

the Atlantic Ocean.

The equation obtained is $y = 1{,}428\,x + 17{,}06$.

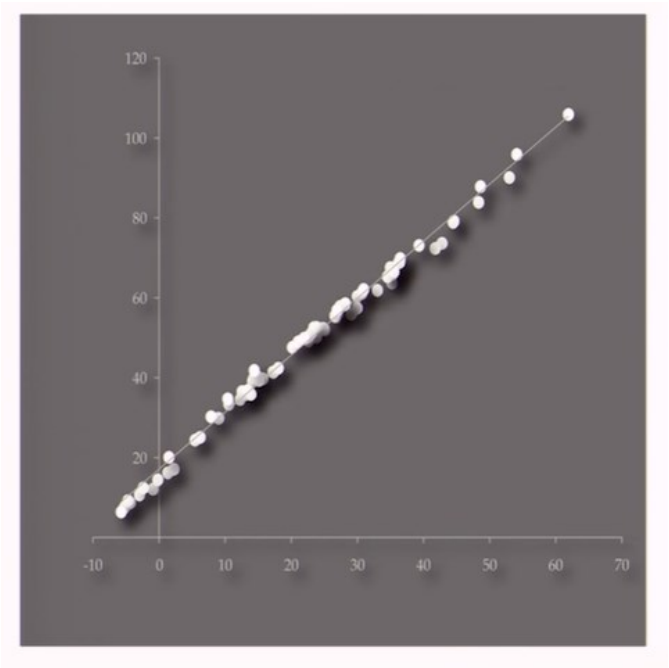


Figure 2: Linear regression plot (Russo, 2013)

1,428 is the slope of the line and it represents the average dilatation factor applied by Tolomeo. Each location with true longitude x will have a Tolemaic longitude y that is increased by a factor 1.428 (+ the constant 17,06 that represents Greenwich's longitude in Tolomeo's system, obtained by assigning the value 0 to x). The coefficient of determination R^2 for this model has a value of 0.9935. R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable in a regression model, so this value suggests that the model identified explains well the relation between the two sets of coordinates.

1.1.2.3- Potential causes and consequences of the bias

As it was explained before, determining differences of longitudes was much more difficult than working with latitudes. Tolomeo had access mostly to differences of longitudes expressed in distances along parallels by Eratostene, but he decided anyway to convert them in degrees of longitude. To do this he needed to know the exact length of each parallel, obtainable once one knows the dimensions of earth. Here comes the other mistake of Tolomeo, which is actually another side of the dilatation of longitude values: he assumed the earth's circumference was 180 000 stadiums long, in contrast with the value of 252 000 given by Eratostene. Russo shows how these two mistakes are connected between them, under the hypothesis that the value of the

stadium used by the two scientists is the same. This hypothesis is more or less proven by Russo, who demonstrates that according to the data available it is plausible that both used a stadium equal to 157.5 meters. The ratio between the value obtained by Eratostene for earth's circumference and the one of Tolomeo is of 1.4, close to the dilatation factor of 1.428 obtained through the regression procedure. This is a first hint of the fact that this aspect is connected with the longitude issue. By dilating the longitude values, in fact, Tolomeo was operating an implicit correction to the smaller value assumed for the earth's circumference, in order to keep everything in line with the distances upon which he based his longitude values and that he received from his Greek sources. This strong shrinkage of earth's dimensions appears strange, also because Tolomeo does not justify it at all, but just states that the value of 180 000 stadiums was largely accepted. Here comes into play the cultural collapse that was discussed above. It has been shown how the events that occurred in the Mediterranean area may have affected the passing on of knowledge and the interpretation of ancient sources. The case of earth's circumference value can be a good example. The value of 180 000 starts to appear in the first century AD and Russo shows how it may have been proposed. One key concept is the one of Ecumene, that is the fraction of the world that was believed to be the only one inhabitable on Earth. The Ecumene expanded from the 'Isole fortunate' in the West to some unclear location in the East, relatively close to China's capital city of the time. The longitude expansion of the Ecumene was equal to 180°, based on the fact that the same eclipse event was seen 12 hours later at the two extremities. This evidence, being an astronomic one, is probably traceable to Ipparco or to the Hellenistic culture in general, also according to Tolomeo's words. The point is that Tolomeo identifies the 'Isole fortunate' with the Canary Islands, in agreement with many authors that however were all subsequent to the cultural collapse. By assigning a longitude difference of 180° between the two ends of his Ecumene, Tolomeo is overestimating it by 40%. The method illustrated in his work through which he obtains this value seems coarse, as he starts from distance values, he applies qualitative corrections to those values and he converts them into degrees ending up exactly at 180°. It is much more plausible that Tolomeo was trying to adapt his measurement to the value of 180° that is traceable to his sources and that probably considered different ends for the Ecumene. This would explain the shrinkage of earth's dimensions by Tolomeo and the connected dilatation of longitude values. The position of China's capital city, to which Tolomeo assigns a longitude of 177° 15', should be outside of China to justify such an error, so the only possibility is that the misunderstanding was about the 'Isole fortunate' and their identification with the Canary Islands. The Canary Islands were known well before Tolomeo and if the coordinates that Tolomeo assigns them are analyze in further detail, something surprising emerges. First of all, there is a huge discrepancy between

their true latitude and the one given by Tolomeo. A distance of 15° that, to give an idea, separates Naples from Copenhagen. Then, according to Tolomeo's coordinates, the Canary Islands are aligned on the north-south axis, contrarily to what actually happens.

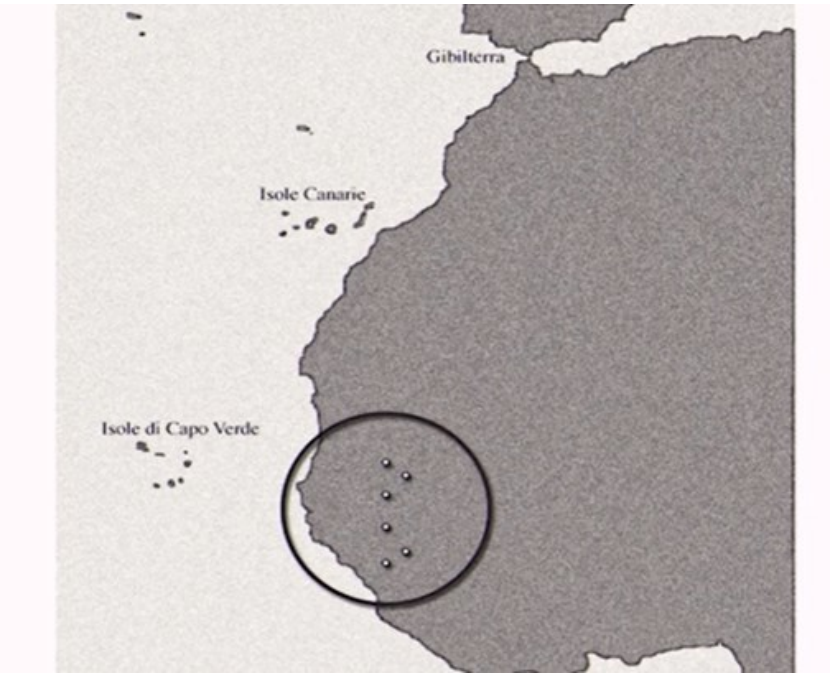


Figure 3: Isole Fortunate's location according to Tolomeo's coordinates (Russo, 2013)

Russo tries to identify the true corresponding of the islands by using the equation derived previously. He firslyt obtains the Greenwich longitude corresponding to Tolomeo's eastern extremity by resolving $180 = 1{,}428\,x + 17{,}06$ with x = 114° 6' E and then obtains the opposite meridian by subtracting 180° and obtaining a longitude value of 65° 64' W. The fact that the regression equation was used far from the range considered to determine it may lead to an error and, furthermore, it is not expected that coordinates of islands in the Atlantic Ocean were known with high precision. Still, it emerges from the map that the new coordinates of the 'Isole fortunate' are similar both in terms of longitude and latitude to the ones of the Lesser Antilles, an archipelago whose extension resembles the one of the 'Isole fortunate'. On top of that, Russo shows different passages of sources dating to the Hellenistic period that describe these islands and their vegetation in a way that reminds of tropical islands and not of the Canary Islands.

Figure 4: Isole fortunate's location after the correction for longitude dilatation (Russo, 2013)

All these elements, along with the fact that Tolomeo's Fortunate's Islands are 6 and have a vertical alignment as the Lesser Antilles do, whereas Canary Islands are 7 and have a horizonal one, makes it worth going into more detail and trying to see if a genetic trace is there to confirm the hypothesis of pre-Columbian encounters between the 'old' and the 'new' world, in particular in the area of the Lesser Antilles.

1.2 Summary of Caribbean history and genetics

It can be useful, at this point, to discuss and take an overlook to the history of the Lesser Antilles and of the area of the Caribbean in general. In this way it will be possible to get an idea of what to expect from the genomes of contemporary inhabitants of the islands and it will also be easier to interpret the samples of ancient genomes of this area. The summary presented will illustrate the canonical and acknowledged history of this area, in which pre-Columbus contacts are not present.

The Caribbean is a place with peculiar characteristics in terms of peopling and admixture. It was one of the last settled by humans in the American continent and the first to experience encounters between indigenous people and European colonizers. The history of the peopling of this area relied on historical and linguistic evidence until the last twenty years, when genomic research began to bring new information to the table. By the end of 2014, no complete genome from ancient Caribbean had been sequenced, but in the following six years that number increased to over 260 (Nieves-Colón, 2022). This allowed to better understand the dynamics of the peopling of this area, even if there are still parts that need further clarity.

Figure 5: The Caribbean area (Saylor Academy, 2013)

The first signs of human presence in the Caribbean trace back to around 5000 years ago (Nieves-Colón, 2022), but the route that brought people to this area is unclear. Different sources have been proposed in different studies: some think that it was a wave coming from central or south America, relying on archeological and linguistic sources, but others claim that a dispersal from north America was part of the peopling too. There is no agreement either on the number of dispersals involved in this first phase. In any case, this settlement initiated a period characterized by stone tool technology called Archaic Age, which lasted until approximately 500 BC. This date corresponds to the beginning of another wave, whose origins are instead known and locate in the northern part of south America (precisely in the area of present-day Venezuela and Guyana (Nieves-Colón, 2022)). The Arawak-speaking people coming from these territories started to move towards the Caribbean area and the earliest traces are found in Puerto Rico and the northern part of the Lesser Antilles. The route followed by this people is disputed, as some researchers suggest that ceramic people arrived in the Greater Antilles and

then expanded downwards, while other claim they passed through the Lesser Antilles. No clear signs of their presence have been found in the southern part of the Lesser Antilles before 1800 years ago (Fernandes, 2021), but some genetic studies support the second hypothesis as well as ceramic typology. Archaic genetic traces were largely erased by the arrival of Ceramic people, with some exceptions like in Cuba. This island is the main source of ancient DNA related to Archaic people and it kept hosting this culture probably until the arrival of Europeans in the 15th century (Fernandes, 2021). It is known that Archaic and Ceramic cultures coexisted in some of the islands, mostly in the Greater Antilles, even if the cases of admixture are extremely rare and the dynamics of the interactions between them are still currently unclear. What is sure is that they belonged to different waves that arrived in the Caribbean islands. The difference in the genetics of the two groups is evident both in the mtDNA lineages, molecules of DNA present in the mitochondria that are inherited from the mother and that are rarely subject to recombination or mutation, and in the autosomal studies (Nägele K, 2020). In Figure 6 it is possible to see some ancient DNA samples associated to the Archaic and Ceramic cultures projected on a PCA calculated from present-day indigenous genomes of the area. PCA is a technique that allows to reduce the dimensionality of data while keeping as much information as possible and it will be further explained later, but it is clear how the two groups fall into two different clusters.

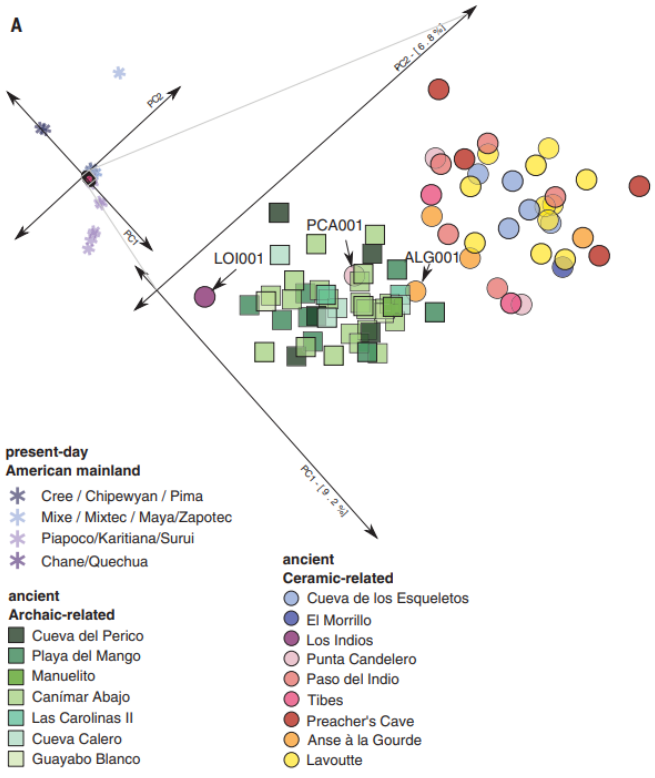


Figure 6: PCA projection of Archaic and Ceramic samples (Nägele K, 2020)

The dynamics of the peopling of the Caribbean area, especially the ones regarding the arrival of Ceramic people, are relevant to this study, since the centuries of their expansion overlap with the period in which Greeks, Carthaginians and Phoenicians may have been able to reach this area. If, for example, these Mediterranean populations encountered Archaic people and a Ceramic wave subsequent to 146-145 BC replaced the inhabitants of the Lesser Antilles, no traces of admixture would be visible today. The best scenario for this study is the one in which Ceramic people reached the Lesser Antilles before the hypothetical arrival of Mediterraneans, also because Archaic genetic traces are almost absent in today's inhabitants of these islands. It is also important to keep in mind that the dynamics are not necessarily the same for each island. Another thing to notice and to keep in mind for this work is that currently few ancient samples are available for the Lesser Antilles, therefore some information about what happened in that area may be missing.

No other major waves are detected until the arrival of European colonizers in the late 15th century, which dramatically changed the situation through the spread of diseases, the enslavement of indigenous people and the forced labor in plantations. All these things led to a drastic demographic decrease in the local populations, as reported by the census records kept by the European settlers (Nieves-Colón, 2022). By reading documents of the colonial period, it may be possible to have the sensation of an actual extinction of local people, but contrary to the narrative different communities were formed throughout the Caribbean Islands where native people opposed the colonizers and found shelter (Benn Torres J, 2019). Another consequence of the colonial period was the arrival in this area of millions of slaves coming from Africa and employed in plantations, further reshaping the local genetic landscape. Other migrations took place over time, especially after the abolition of slavery in the 19th century, but in most of the cases they can be considered marginal events at least from a genetic point of view. The first Europeans to arrive in the Caribbean area were the Spanish in the late 15th century. All the other populations, like French, Dutch and British, formally entered the area in the early 17th century and occupied most of the Lesser Antilles (Nieves-Colón, 2022). Each island has its own peculiar history in the colonial period in terms of impact of African slavery, formation of indigenous communities and European country that colonized it. For instance, genetic studies showed that the European source population for the genomes of contemporary inhabitants of this area is variable between different islands and sometimes even within the same island.

The idea and hope of genetic researchers was that contemporary inhabitants of the island could be a good source of pre-contact genetic diversity and this was confirmed in different studies. The first ones, that focused on mtDNA, found high proportions of indigenous lineages, especially in communities that were self-identified as descending from indigenous people. It may be possible that some of these genetic traces come from indigenous of mainland America that were brought to the islands during the colonial period, but there are proofs that support the claim that at least part of that is due to continuity and not to replacement. For example, some lineages of mtDNA found in samples of ancient DNA of the Caribbean area are present nowadays in modern inhabitants of the area and nowhere else (Nieves-Colón, 2022). There is also a good corresponding in the study of the autosomes between nowadays indigenous and ancient ones. It is important to notice that this corresponding is with Ceramic people, while traces of Archaic people are nearly absent. Different studies have shown that when plotting modern genomes of the area together with ancient ones in a PCA plot like the one seen before, the Archaic ones fall in an area that is outside of present-day indigenous, while there is overlapping with Ceramic genomes. Considering all this, the genomes of present-day indigenous can be considered a good reservoir of pre-contact genetic variation (at least the Ceramic one) and can thus be investigated to better understand ancient dynamics.

2 - Data

There are two different approaches that can be used to study the genetic variation of pre-Columbus Caribbean people in order to verify the possibility of encounters with Mediterraneans in the first centuries BC.

2.1 – Ancient DNA

The first and most obvious one is to directly analyze ancient DNA coming from the Caribbean area dating back to at least 550 years ago. In this way any trace of European genetic diversity found could not be assigned to European colonizers, but should be looked for in other sources. The field of ancient DNA (aDNA) started in the late 20th century and consists in extracting DNA from ancient specimens. It evolved a lot through time alongside the innovations of genomic fields, reaching in the late 2000s the first genome-wide results (Reich, 2018). Obtaining aDNA has different complications due to the fact that DNA molecules tend to degrade over time. The samples are shorter and they may also have gaps since nucleotides can physically drop. Furthermore, depending on the climatic area, the conservation of DNA molecules can be favored (typically cold environments help) or disfavored (like in tropical areas). Another problem regarding aDNA is that most of the material extracted from ancient specimens does not belong to the 'owner' of the bone. Usually it belongs either to microbes or other humans, like the ones who handled the sample or analyzed it (Slatkin M, 2016). For this reason, especially in the first years of this branch, obtaining sequences of the ancient humans was much more expensive than it was for modern ones. The problem of modern human's contamination in ancient samples was gradually solved by taking precautions in the process of handling the specimens, but there was still a lot of non-human material in the sample. The first method employed, developed by Svante Pääbo and his team, consisted in sequencing the entire sample and then focusing only on the fraction of human DNA, but this could be as low as 2% (Reich, 2018). The advantage of this method is that the data obtained are unbiased, since there is no pre-filtering of the positions in the genome that will be extracted. Later, David Reich and his team designed a new method, much cheaper than the previous one, that exploited the peculiar way of DNA of binding with other molecules. They designed a sort of 'bait' made of synthetized DNA sequences that attracted the human DNA in the specimen in specific positions, particularly in the ones that are known to vary within the human genome. This method ended up being very successful, allowing to resolve the problem of non-human material and to focus only on areas of interest of the human genome. In this way the value of efficiency in terms of

percentage of 'useful' DNA within the extracted one was much higher and the costs consequently lower, allowing to sequence many more individuals than what was previously feasible. The disadvantage of this process is the implicit bias in the results: the positions sequenced are always the same (around 1.2 million) and there is no chance of obtaining information regarding other positions that may have had mutations in the past and that could be highly informative.

The database used in this work for aDNA comes from David Reich lab and contains more than 10000 individuals gathered from different studies, with information referring to 1233033 SNPs (Allen Ancient DNA Resource, 2021). A SNP (single nucleotide polymorphism) is the substitution of a nucleotide in a specific position in the genome and in this field it is usually modeled as a one-time event and considered as biallelic. This means that if a position is associated with a SNP it will have only two possible variants (alleles) and it will not undergo a substitution process again in the future. There are just few samples in this database belonging to the Lesser Antilles, in particular to the islands of Guadeloupe and St. Lucia, while there are more belonging to the Caribbean area in general (Nägele K, 2020) (Fernandes, 2021) (Schroeder H, 2020). The colored map in Figure 7 reports in red the countries of the Lesser Antilles with available aDNA samples and in green the ones from the rest of the Caribbean area.

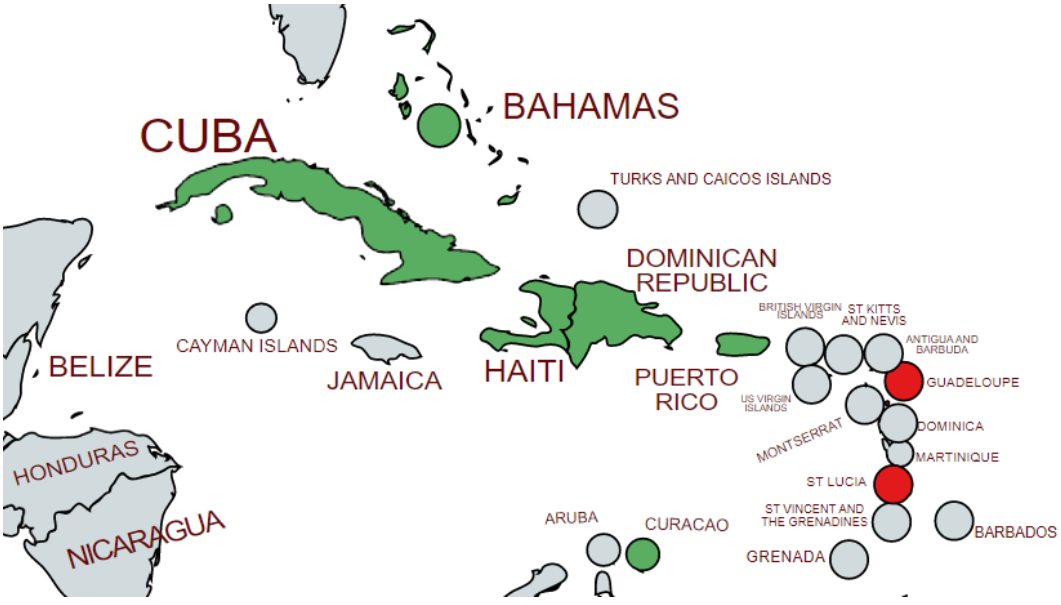


Figure 7: Countries of the Caribbean with aDNA samples used in this work

From this database also different populations belonging to the Mediterranean area were selected as proxies for the hypothetical populations that may have reached the Lesser Antilles (Lazaridis

I, 2017) (Feldman M, 2019) (Agranat-Tamir L, 2020) (Haber M, 2020). The Mediterranean samples were chosen verifying that their dating could be sufficiently compatible with the centuries in which the encounters may have happened and their location are shown in green in Figure 8. Unfortunately, no aDNA from Northern Africa was available for the period of interest.

Figure 8:Countries with aDNA samples used as possible Mediterranean sources

2.2 – Modern DNA

The second approach to investigate possible contacts consists in studying modern individuals to obtain information about the past. As it has been discussed before, present-day indigenous of the Caribbean Islands are considered a good reservoir of the genetic variation of ancient indigenous. It is therefore necessary to analyze modern population that have a decent amount of indigenous ancestry in their genome, hoping that at least part of it belongs to ancient inhabitant of the islands. The best way of doing it would be to analyze genomes of self-identified indigenous community, which are the ones that tend to have largest component of indigenous ancestry and that descend, at least according to their own judgement, from the original inhabitants of the island. There were two studies carried out on local communities in the Lesser Antilles, one in the island of Dominica (Keith MH, 2021) and the other one in the islands of Trinidad and St. Vincent (Benn Torres J, 2019). The data coming from the first one was not available for sharing, since the local community decided to maintain the right for privacy. The second study, instead, stated that upon reasonable request the data would be available and these is what was thought of as the main source of modern DNA at the beginning of this work. Through professor Pagani we got in touch with the authors of the paper, explaining what was the aim of the work and trying to reach an agreement with them. We were about to

receive the data when the local communities declared they were no longer happy to share their data. At this point an alternative was needed and it was identified in the data from Puerto Rico in the Greater Antilles that were publicly available through the 1000 Genomes Project Consortium. There was also the possibility of using data from Barbados Island, which is part of the Lesser Antilles, but the available genomes were made in large part by African component and the percentage of Native American genetic information was too small to carry out an analysis. The genomes from Puerto Rico, instead, have a sizeable part of indigenous component and even if the island is not part of the Lesser Antilles, its location is quite close to them and it is reasonable to consider it as a good proxy for indigenous genetic variation of that area. On top of that, as it was said before, Puerto Rico is one of the places where it is possible to see the earliest traces of Ceramic culture, so this could be good for the purposes of this work.

The two panels that were used as source for modern DNA in this work are the Human Genome Diversity Project - HGDP (Bergström A, 2020) and the 1000 genomes project (Consortium., 2015). The data used from the 1000 genomes project are the ones of phase 3, consisting in 2504 individuals from 26 different populations shown in Figure 9. This project used a low-coverage approach to sequence the genomes, which does not allow to identify all the variants, but if combined for all the individuals is enough to determine with accuracy the genotype of SNPs that have a frequency of at least 1% in the populations studied. These data are freely accessible (Consortium, 2013) and contain also the genomes that were used from Puerto Rico (PUR), consisting in 104 individuals.

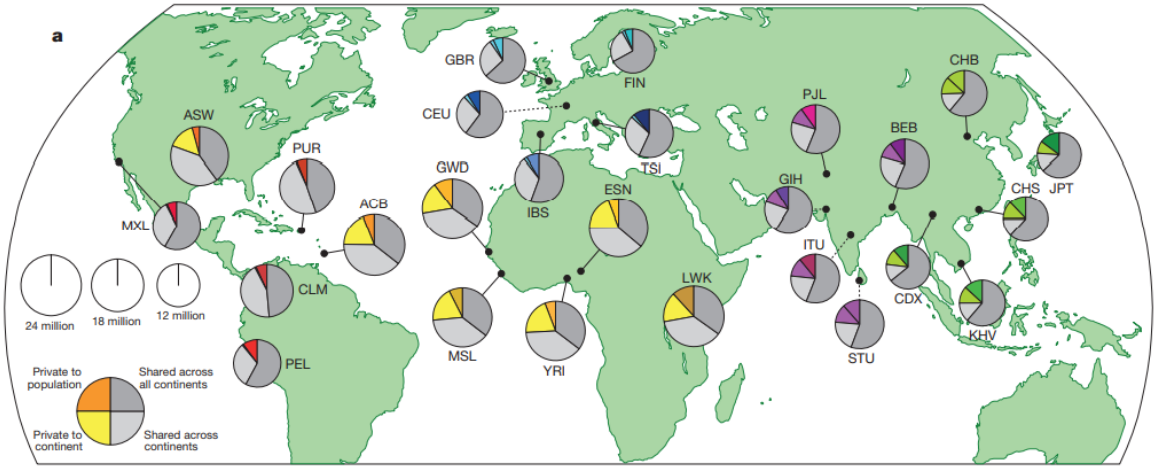


Figure 9: Population sampled in the 1000 genomes project (Consortium., 2015)

The HGDP project used a different approach with respect to the 1000 genomes one, sequencing fewer individuals for each population at high coverage and not only in large metropolitan populations like the other project did. This panel includes 929 genomes from 54 different populations, shown in Figure 10.

Figure 10: Population sampled in the HGDP (Bergström A, 2020)

3-Methods

In this chapter the main software and techniques used in this work are presented and explained in detail.

3.1-Plink

Plink is a software designed for management and analysis of position-based SNP-like data for large number of samples (Purcell, 2009). The basic file format used by Plink is the binary fileset, composed by three files with extensions: .bed, .bim, .fam.

The bed file contains the representations of the genotype calls at the different SNPs for the individuals of the dataset. It consists in a sequence of V blocks, where V is the number of SNPs, each of which contains the information relative to the correspondent SNP in the bim file.

The bim file contains the information regarding the different variants in the dataset. Each line of the file corresponds to a SNP and contains 6 fields: 1) Chromosome code or name, 2) Variant identifier, 3) Position in morgans or centimorgans (a centimorgan is defined as the distance between chromosome positions for which the expected average number of intervening chromosomal crossovers in a single generation is 0.01), 4) Base pair coordinate, 5) Allele 1, 6) Allele 2. An example of a line from one of the bim files used in this work is the following:

```
1	1_776546	0.020242	776546	G	A
```

The fam file contains information about the individuals of the dataset and it's composed by one line for each individual with 6 fields each: 1) Family ID, 2) Within family ID, 3) Within family ID of the father (0 if it is unknown, 4) Within family ID of the mother (0 if it is unknown), 5) Sex code (0=unknown, 1=male, 2=female), 6) Phenotype value (1=control, 2=case, -9 or 0 or non-numeric=missing). An example of a line from one of the fam files used in this work is the following:

```
Lebanon_IA2 SFI-55.SG 0 0 1 1
```

For some specific applications it was necessary to switch to another format, composed by two files with extensions .tfam and .tped. The tfam file is identical to the fam file previously described. The tped file, instead, contains the information for the bim and bed together. Each line refers to a variant and has 2N + 4 fields, where N is the number of samples. The first four fields are chromosome code, variant identifier, variant position in morgans or centimorgans and

base pair coordinate. Then there are two fields for each sample, representing the two alleles for that variant. An example of a line from one of the fam files used in this work is the following:

1 1_752566 0.02013 752566 A A A A G G … A A G G A A

The main plink functions used were the one to manage datasets. Plink was necessary to merge together different datasets, to filter them given a list of individuals or SNPs and to recode the files from the binary fileset to tepd-tfam one.

3.2-Principal Component Analysis

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of a dataset. Nowadays it is common to deal with datasets made of many observations of a large number of variables and it is usually difficult to interpret them. The goal of PCA is to reduce the dimensionality while preserving as much variability as possible (Jolliffe IT, 2015). PCA starts by identifying a new basis for the data made by orthonormal vectors ($< v_i, v_j > = 0$ when i≠j and $< v_i, v_i >= 1$), where the new variables are uncorrelated between each other and the vectors of the basis are ranked by fraction of explained variance. It is then possible to consider only the first p vectors to summarize the data. A typical choice for p is 2 or 3, so that it is possible to plot the data in 2 or 3-dimensional graph.

Here the explanation given in (Shlens, 2014) will be generally followed.
Considering a general dataset made of n observations of m variables it is possible to define the matrix X corresponding to the dataset as $X = [x_1 \ \dots \ x_n]$, where each x_i represents one of the observations. X will therefore be a $m\ x\ n$ matrix. The matrix Y, also $m\ x\ n$, will be the final representation of the dataset and is related to X by the linear transformation P ($m\ x\ m$) through $PX = Y$. By calling $p_1, p_2, \dots, p_m$ the rows of the matrix P it is possible to write

$$\mathbf{Y} = \begin{bmatrix} \mathbf{p_1 \cdot x_1} & \cdots & \mathbf{p_1 \cdot x_n} \\ \vdots & \ddots & \vdots \\ \mathbf{p_m \cdot x_1} & \cdots & \mathbf{p_m \cdot x_n} \end{bmatrix}$$

The equation $PX = Y$ represents a change of basis and thus the rows of P are a new set of basis vectors for expressing the columns of X. Indeed, each coefficient of the column y_i is the dot product of x_i with the corresponding row in P, namely the projection of x_i on a specific row of P. When the proper change of basis will be identified through the PCA process, the rows of P

will become the principal components of X. The issue is how to choose the best basis of vectors, the most meaningful one in expressing the data. It is easier to understand what this means through the example in Figure 11.

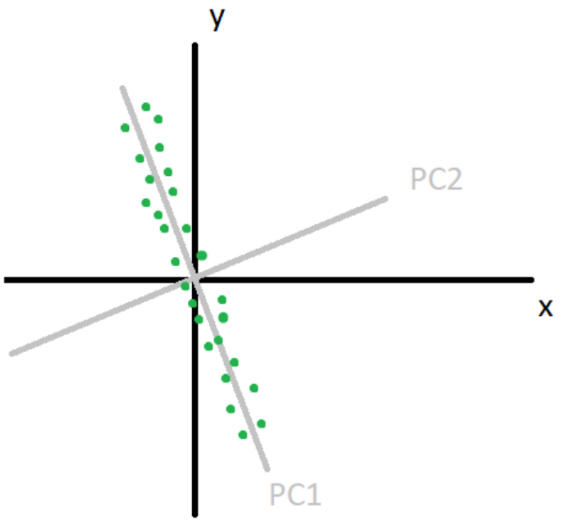


Figure 11: PCA on a 2-dimensional database (Yip, 2020)

It is possible to see that in this 2-dimensional dataset the original variables x,y are clearly correlated with each other. By identifying the principal components PC1 and PC2 and re-expressing the data according to these new vectors, it turns out that it could be possible to capture most of the data variability just by looking at the first dimension. This is the idea behind the process of PCA and dimensionality reduction. The process of finding the best basis starts with the standardization of the data with respect to the mean, in which the mean of each row is subtracted from the corresponding row of the data matrix so that each variable will have mean equal to zero across the dataset. The following step is to introduce the covariance matrix. If a single variable is considered in all its observations and this vector is called x_i (it will be $[x_{i1} \dots x_{in}]$), its variance σ_{x_i} will be $\frac{1}{n-1} x_i {x_i}^T$. If two different variables x_i and x_j are considered, their covariance $\sigma_{x_i x_j}$ will be $\frac{1}{n-1} x_i {x_j}^T$ and it will be equal to σ_{x_i} only if x_i=x_j, while it will be equal to 0 if x_i and x_j are entirely uncorrelated. By generalizing to the matrix X corresponding to the dataset, the covariance matrix will be $S_X = \frac{1}{n-1} X X^T$. S_X is a square symmetric matrix in which the diagonal terms are the variances of the different variables, while the other terms in position (i, j) with $i \neq j$ will be the covariances between variable i and variable j. This matrix is important as it captures the relationship between pairs of variables in the dataset and, since the goal is to reduce the redundancy of information as it was seen in the toy example in the figure, the idea is to manipulate the data matrix in a way that the different variables co-vary as little as possible. The total removal of redundancy would correspond to a case in which there is no correlation at all between different variables, that is the one in which

all the covariances are equal to 0. In this scenario the covariance matrix S_X is a diagonal matrix, so the goal of the PCA consists in the diagonalization of the covariance matrix. The final representation of the dataset is Y=PX, so P will have to be such that $S_Y = \frac{1}{n-1} YY^T$ is diagonal. It is possible to rewrite the relation in the following way:

$$\begin{aligned}
\mathbf{S_Y} &= \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T \\
&= \frac{1}{n-1}(\mathbf{PX})(\mathbf{PX})^T \\
&= \frac{1}{n-1}\mathbf{P}\mathbf{X}\mathbf{X}^T\mathbf{P}^T \\
&= \frac{1}{n-1}\mathbf{P}(\mathbf{X}\mathbf{X}^T)\mathbf{P}^T \\
\mathbf{S_Y} &= \frac{1}{n-1}\mathbf{P}\mathbf{A}\mathbf{P}^T
\end{aligned}$$

Where A is a m x m symmetric matrix, since the product of a matrix with its transposed is always symmetric. For a theorem of linear algebra, each symmetric matrix A can be written as $A = EDE^T$, where D is diagonal and E is composed by the eigenvector of A arranged as columns. If matrix A has rank $r < m$ it is possible to fill up matrix E with additional m-r orthonormal vectors that will not affect the final solution since their associated variance will be equal to 0. At this point P is chosen in a way that each row p_i is an eigenvector of A, so that $P = E^T$ and $A = P^T DP$. Knowing that the inverse of an orthogonal matrix is equal to its transposed, it is possible to write:

$$\begin{aligned}
\mathbf{S_Y} &= \frac{1}{n-1}\mathbf{P}\mathbf{A}\mathbf{P}^T \\
&= \frac{1}{n-1}\mathbf{P}(\mathbf{P}^T\mathbf{D}\mathbf{P})\mathbf{P}^T \\
&= \frac{1}{n-1}(\mathbf{P}\mathbf{P}^T)\mathbf{D}(\mathbf{P}\mathbf{P}^T) \\
&= \frac{1}{n-1}(\mathbf{P}\mathbf{P}^{-1})\mathbf{D}(\mathbf{P}\mathbf{P}^{-1}) \\
\mathbf{S_Y} &= \frac{1}{n-1}\mathbf{D}
\end{aligned}$$

This choice of P matches the goal, since S_Y is now a diagonal matrix. The principal components of X will be the rows p_i, which are the eigenvectors of XX^T, while the elements on the diagonal of S_Y will represent the variance of X along the direction given by p_i. This method allows also to order the different principal components obtained just by looking at matrix S_Y. It was stated that the initial goal was to keep as much variability as possible and, as one could grasp by looking at the simple example in Figure 11, the more the samples are spread along one direction, the more variability is kept if looking just at that direction while ignoring the other ones, since that direction will have the largest variance. Following this logic, it is possible to order the

elements on the diagonal of S_Y from largest to smallest and order the principal components consequently. It is then sufficient to choose the first p vectors to represent the data according to the number of dimensions p one wants to use to represent the dataset. It is also possible to know the percentage of total variance explained by each direction by applying the simple formula $\pi_i = \frac{S_{Y_{ii}}}{\sum_{j=1}^{m} S_{Y_{jj}}} = \frac{S_{Y_{ii}}}{tr(S_Y)}$. The larger the sum of the π_i's for the first p directions chosen, the smaller the loss of variability is suffered in representing the data.

To apply PCA to genomic data, the software smartpca (N Patterson, 2006) from the EIGENSTRAT 7.2.1 package (Harvard, 2010) was used. This software requires the EIGENSTRAT data format that contains the same information as the binary plink one, but organized in a different way. To operate the conversion the convertf program from EIGENSTRAT package was used. Smartpca outputs the principal components and the corresponding eigenvalues, but it also has several additional functionalities that allow the user to better control the PCA process for his purposes. The ones used in this work allowed to introduce the procedure of outlier removal and to project aDNA samples on the principal components calculated on the modern samples. The last one, least square projection, is particularly useful since aDNA can have several gaps along the sequenced SNPs and this could make the PCA process problematic. The standard procedure, in fact, is to obtain the coordinates of the projected sample as $c_i = s^T e_i$ where e_i is the i-th eigenvector and s is the sample data where each missing position is filled with the average of that allele's frequencies in the base populations used for PCA. If the quantity of missing SNPs is huge, this is not a good approach. Least square projection exploits the fact that the different c_i's are the ones minimizing the equation $\|s - \sum_i c_i e_i\|^2$ and considers only the positions in which the sample has valid data, finding the values of c_i that minimize

$$\sum_{j \in X} (s_j - \sum_i c_i \, e_{ij})^2$$

where X is the set of those positions. In Figure 12 it is possible to see an example of a plot of a PCA applied to a genomic dataset. In particular, this plot was obtained using the data from the HGDP.

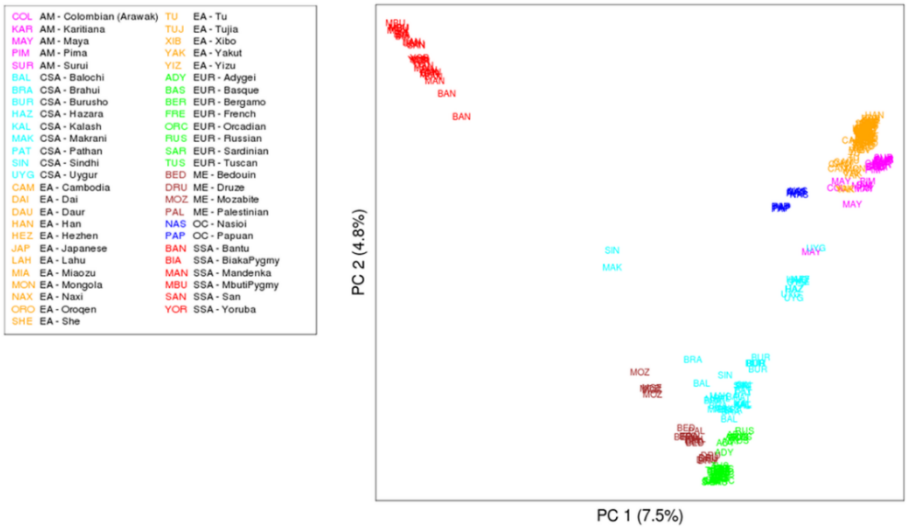


Figure 12: Worldwide PCA (López-Herráez D, 2009)

3.3-PCAdmix

PCAdmix (Brisbin A, 2012) is a software for estimating local ancestry. This process consists in trying to identify, for each tiny segment in the genome of an individual, the ancestral population from which the segment comes from. Individuals are actually diploids, meaning that each one of them has two set of chromosomes, one inherited from the father and one inherited from the mother. PCAdmix works by considering each individual as haploid, so from each sample will derive two 'individuals' that can be called individual_A and individual_B. This way of working allows also to obtain information about the ancestry contribution of an individual's parents. The separation of the two set of chromosomes coming from a single sample is made through a procedure called phasing.

Important features of PCAdmix are the possibility of working with more than two source populations, allowing to disentangle cases that do not involve just a mix of two populations (e.g. the case of Caribbean populations), and the possibility of capturing the non-independence of nearby SNPs by using windows of SNPs. The width of the windows is a parameter that can be controlled by the user to regulate the level of detail of the analysis. The first phase of the program consists in a quality-control filter that removes SNPs for which there is too much data missing or for which the frequency of the minor allele (e.g. the allele with the lowest frequency between the two) is too low, since these SNPs would not be informative. The SNPs are then filtered to correct for Linkage Disequilibrium (LD), namely the non-random association of

alleles at different positions in a population. Two positions in the genome are said to be in linkage disequilibrium if the frequency of association of their different alleles is higher or lower than what would be expected if the positions were independent (Slatkin, 2008). Given two SNPs x and y, respectively with alleles A/a and B/b, LD is usually calculated as $D = p_{AB} - p_A p_B$, where p_{AB} is the frequency of haplotype AB (an haplotype is a specific combination of alleles), p_A is the frequency of allele A and p_B is the frequency of allele B. The SNPs are said to be in linkage equilibrium if $D = 0$, so when $p_{AB} = p_A p_B$. When $D \neq 0$ the SNPs are in linkage disequilibrium, meaning that haplotypes occur with frequencies that are different from what would be expected by looking at the single alleles' frequencies, and the magnitude of D represents the degree of disequilibrium.

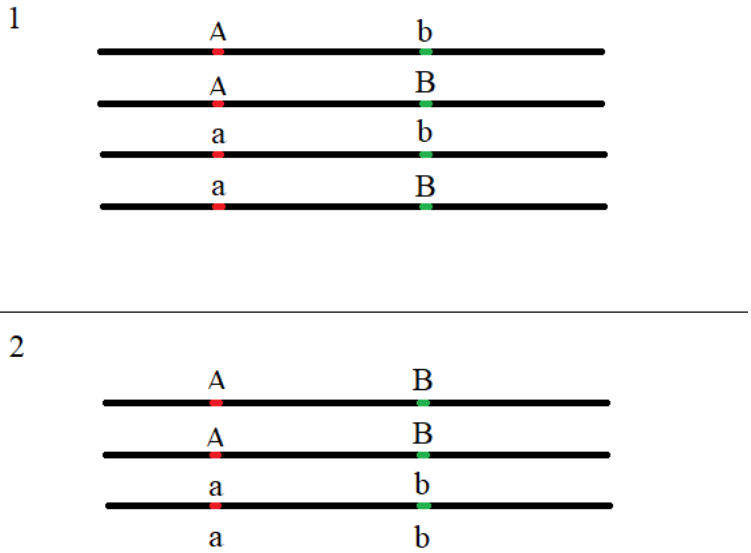


Figure 13: Example of a situation in linkage equilibrium(1) and linkage disequilibrium(2)

The example in Figure 13 shows two different situations, one in a situation of equilibrium and one in a situation of disequilibrium. In both cases $p_A = p_a = p_B = p_b = 0.5$, but in the first one $D = 0$, while in the second one $p_{AB} = 0.5$ while $p_A p_B = 0.25$, proving that haplotype AB is over-present with respect to expectations. Another way of measuring LD is the correlation coefficient $r^2 = \frac{D^2}{p_A p_a p_B p_b}$ that measures the independence between pairs of SNPs. PCAdmix removes one SNP of each pair for which r^2 is greater than 0.8 in any of the ancestral or admixed groups. By doing this the program is trying to avoid the presence of blocks of SNPs with a high degree of LD, since these could have an excessive influence on the estimate of local ancestry.

PCAdmix exploits PCA to make the estimates by assigning greater weight to SNPs that are more informative about ancestry estimation. Principal components may not be easy to interpret directly, but the point is not to extract information understanding the meaning of each PC, rather

to observe the position of admixed individuals relatively to the one of clusters of individuals belonging to ancestral populations. After having identified the PCs of the samples coming from ancestral populations, the phased genotypes of the admixed individuals are projected and ancestry scores are calculated. For each window of SNPs w, the vector S_{iw} is created, containing the ancestry scores for haplotype i across the first K-1 PCs, where K is the number of ancestral populations. S_{iw} is obtained as $L_w g_{iw}$, where L_w is a matrix that contains in each column the PC loadings (i.e. the coefficients of the linear combinations characterizing the principal components) of one SNP of the window for the first K-1 PCs and g_{iw} is a column vector of the haplotype's alleles in window w (each SNP is modelled as a discrete random variable) standardized by mean and standard deviations in the ancestral populations.

Once the ancestry scores are calculated, PCAdmix implements a Hidden Markov Model (HMM) to model the ancestry of each window probabilistically. HMM is a statistical model based on a Markov Chain, which is a model that describes a series of events in which the probability of each of them depends only on the previous one. A Markov Chain is defined by a set of N states Q ($q_1 \dots q_N$), a transition probability matrix A (NxN) in which each entry A_{ij} represents the probability of going from state i to state j, and an initial probability distribution π over the N states representing the probability of being in a state at the beginning of the series of events. Each event corresponds to the system being in a specific state.

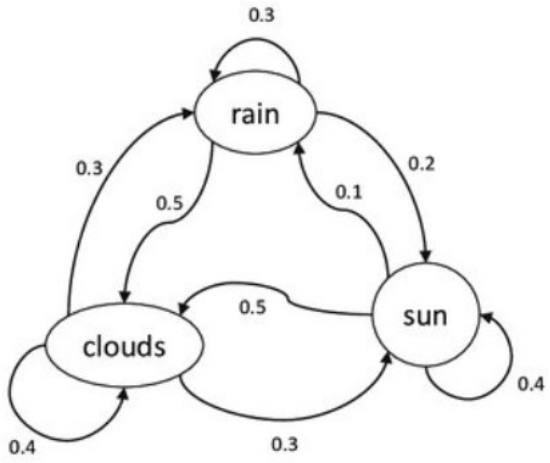


Figure 14: Example of a diagram of a Markov Chain (Seyr H, 2019)

The Markov Chain corresponding to the diagram in Figure 14 is characterized by three states (clouds, rain, sun) and the transition matrix A is the following:

$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.5 & 0.3 & 0.2 \\ 0.5 & 0.1 & 0.4 \end{bmatrix}$$

while the initial probability distribution is omitted.

In the HMM, contrarily to Markov Chains, the sequence of states corresponding to the events is not directly observable, but for each state there is an emission probability, which is the probability of observing a specific parameter when the system is in that state. Such parameter depends only on the state of the Markov Chain the system is in during the corresponding event. Starting from a sequence of T observations (o_1 ... o_T) of that parameter corresponding to T events it is possible to identify the hidden sequence of states with the highest probability of having generated that sequence of observations. Continuing on the example, the observed parameter could be the mood of one person (happy or sad), assuming it depends only on the weather. In this case an example of B could be:

$$B_{clouds} \begin{bmatrix} P(happy|clouds) = 0.5 \\ P(sad|clouds) = 0.5 \end{bmatrix}$$

$$B_{rain} \begin{bmatrix} P(happy|rain) = 0.3 \\ P(sad|rain) = 0.7 \end{bmatrix}$$

$$B_{sun} \begin{bmatrix} P(happy|sun) = 0.8 \\ P(sad|sun) = 0.2 \end{bmatrix}$$

Coming back to PCAdmix, the states of the HMM correspond to the K possible ancestries of each window and the transition probability is defined as (Brisbin A, 2012):

$$P(anc_{i,w}=j \mid anc_{i,w-1}=k) = \begin{cases} q_{i,j}\boldsymbol{\pi} & \text{if} k \neq j \\ q_{i,j}\boldsymbol{\pi} + (1-\boldsymbol{\pi}) & \text{if} k = j \end{cases}$$

where $anc_{i,w}$ is the ancestry of haplotype i in window w, $\boldsymbol{\pi}$ is the probability of recombination between windows and $q_{i,j}$ is the average ancestry proportion of population j in haplotype i estimated as $\frac{D_{i,j}}{\sum_k D_{i,k}}$. $D_{i,j}$ is an Euclidian distance in the PCs space and it's easier to understand it through the example of Figure 15.

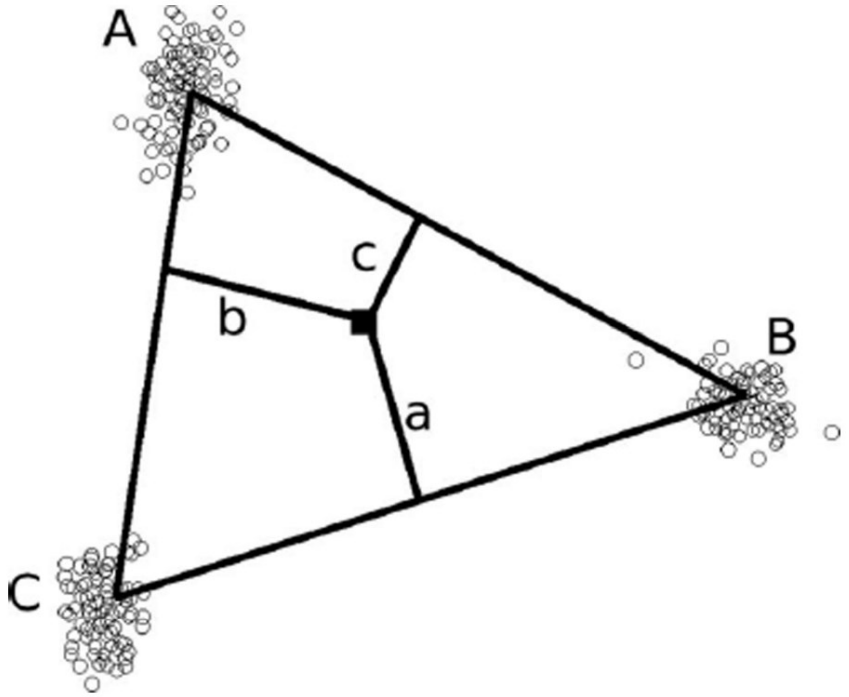


Figure 15: Example of calculation of $D_{i,j}$ (Brisbin A, 2012)

The black square represents haplotype i position in the PCs space, while A,B and C are the three ancestral populations and the circles are the different individuals. Lines that connect the mean point of each population's cluster are drawn, then in this case $D_{i,A}$ corresponds to a and $q_{i,j}$ is calculated as $\frac{a}{a+b+c}$. Intuitively a haplotype that is positioned close to the cluster of one population will have a high value for the corresponding D and consequently a high value for the corresponding q. The observed parameters for the HMM are the previously calculated ancestry scores of different windows and the emission probability is modeled as a multivariate normal distribution $S_{iw}|\,(anc_{i,w} = j) \sim N(\mu_{j,w}, \Sigma_{j,w})$, where $\mu_{j,w}$ is a vector containing the mean scores of ancestry j in window w over the first k-1 PCs and $\Sigma_{j,w}$ is the covariance matrix of the scores for window w among population j haplotypes. PCAdmix identifies the sequence of states with the highest probability of having generated the sequence of ancestry scores across the windows using a forward-backward algorithm, that is an algorithm designed to efficiently compute posterior probabilities for HMMs. PCAdmix outputs several files that report, for each 'individual', the probability of each ancestry for each window, the most probable ancestral population for each window, the SNPs composing each window, the overall ancestry proportions and the scores for the PCs.

3.4-Admixture

Admixture (Alexander DH, 2009) is a software for model-based estimation of ancestry in unrelated individuals. Being able to estimate the proportion of ancestral components in the genome of an individual can have multiple applications and different programs have been created to avoid relying on self-reported ancestry, which cannot be precise for obvious reasons. Admixture, with respect to his predecessor Structure (Pritchard JK, 2000), is much faster and allows to use more SNPs, resulting in a more precise estimate. Differently from PCAdmix, Admixture deals with global ancestry, that is the proportion of ancestry from the source populations as an average over the individual's genome. To get an idea of what this means looking at a typical plot of an Admixture output as the one in Figure 16 can help.

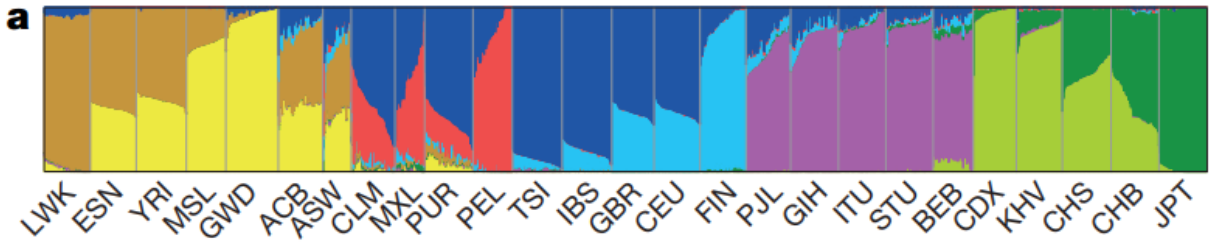


Figure 16: Admixture plot of the 1000 genomes dataset (Consortium., 2015)

Each tiny vertical line represents the genome of an individual and it is colored according to the estimated proportions of ancestral populations, each of which is represented by a different color. The number of ancestral populations K is a parameter given in input by the user, therefore it is helpful to know the history of the population under study in order to use a reasonable value for it. A typical approach is to perform a cross-validation in order to see the best value for k according to the data.

Admixture actually produces in output not only the estimate of ancestry proportions, but also the populations allele frequencies. The typical dataset for Admixture consists in a number I of individuals belonging to admixed populations and a series of SNPs J along the genome. A thing to consider is that the program assumes linkage equilibrium among the markers, therefore it may be useful to prune the markers before if the set is too dense, in order to obtain a more precise result. Each ancestral population k contributes to the genome of individual i with a fraction q_{ik}. The program works with SNPs that are considered to have only two alleles, therefore for each position it is sufficient to consider the frequency of one of the two alleles. This frequency in population k for SNP j is defined as f_{kj} and, as q_{ik}, is considered to be

unknown. Since the program considers the individuals as a random union of gametes, the probability of each genotype is as follows (Alexander DH, 2009):

$$\Pr(1/1 \text{ for } i \text{ at SNP } j) = \left[\sum_k q_{ik} f_{kj}\right]^2,$$

$$\Pr(1/2 \text{ for } i \text{ at SNP } j) = 2\left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1 - f_{kj})\right],$$

$$\Pr(2/2 \text{ for } i \text{ at SNP } j) = \left[\sum_k q_{ik}(1 - f_{kj})\right]^2.$$

where 1 and 2 are the two possible alleles for SNP j. The authors also define g_{ij} as the observed copies of allele 1 at SNP j for individual i. For what it's been said:

$$g_{ij} = \begin{cases} 2 \; if \; i \; is \; 1|1 \; at \; j \\ 1 \; if \; i \; is \; 1|2 \; at \; j \\ 0 \; if \; i \; is \; 2|2 \; at \; j \end{cases}$$

The parameters q_{ik} are enclosed in matrix Q, while f_{kj} are enclosed in matrix F. The optimization phase in which the values for q_{ik} and f_{kj} are found consist in maximizing the log-likelihood of the model which is expressed by the following (Alexander DH, 2009):

$$L(Q,F) = \sum_i \sum_j \left\{ g_{ij} \ln\left[\sum_k q_{ik} f_{kj}\right] + (2 - g_{ij}) \ln\left[\sum_k q_{ik}(1 - f_{kj})\right] \right\}$$

where an additive constant is missing but does not change the problem of maximization. The constraints on the parameters are $0 \leq f_{kj} \leq 1$, $q_{ik} \geq 0$ and $\sum_k q_{ik} = 1$. To solve this optimization problem the software exploits a block-relaxation algorithm that alternates the updates of matrices Q and F.

3.5-F statistics

F statistics (Reich D, 2009) are widely used in the field of population genetics, especially when one is working with admixture events. These statistics are based on allele frequencies and measure correlation between them, allowing to infer phylogenetic trees with the possibility of including admixture events. There are three types of f statistics: f2, f3 and f4. The last one is the most general one and is defined as: (Patterson N., 2012)

$$F_4(A,B;C,D) = E[(a' - b')(c' - d')]$$

where A, B, C and D are four populations and a', b', c' and d' are the respective allele frequencies. Since only biallelic SNPs are considered, the choice of the allele does not affect the value of the statistic, since changing allele would just flip the sign of both terms. The statistic is usually calculated among several markers and then the average is taken.
F2 and F3 can be expressed in terms of F4 as: (Lipson, 2020)

$$f_2(A,\ B) = f_4(A,\ B;\ A,\ B)$$

$$f_3(A;\ B,\ C) = f_4(A,\ B;\ A,\ C)$$

An intuitive way to interpret the F4 statistic, from a geometric point of view, is to consider its value as the intersection between the path from A to B and the one from C to D.

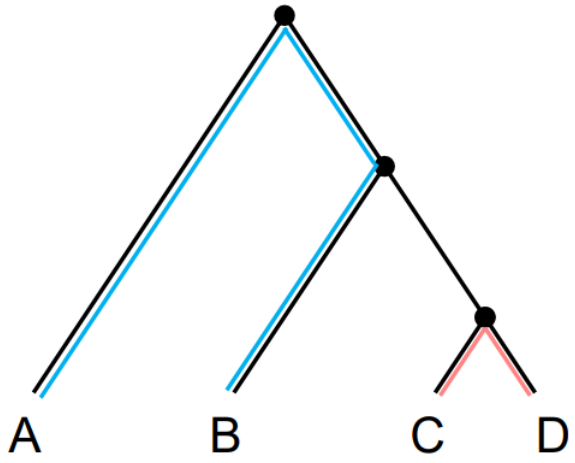


Figure 17: Geometrical interpretation of $f_4(A,B,C,D)$ (Lipson, 2020)

In the case of Figure 17 the value of $f_4(A,B,C,D)$ is 0, but if we consider the statistic $f_4(A,D,B,C)$ its value is equal to the quantity y in Figure 18.

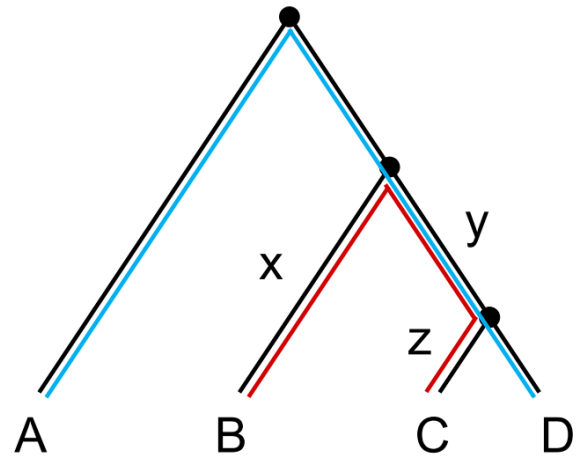


Figure 18: Geometrical interpretation of $f_4(A,D,B,C)$ (Lipson, 2020)

There are $4! = 24$ possible f4 statistics given four populations, but four permutations lead to identical values (e.g. $f_4(A,B,C,D) = f_4(C,D,A,B)$), leaving six unique values. These six can be grouped in three pairs that have same absolute value and opposite sign (e.g. $f_4(A,B,C,D) = -f_4(A,B,D,C)$), corresponding to the three possible tree topologies that one can draw given four populations that are not admixed within each other. Given these three topologies, one of them will lead to a f4 statistics equal to 0, corresponding to the correct topology for those four populations.

This does not hold if one considers the possibility of admixture between populations.

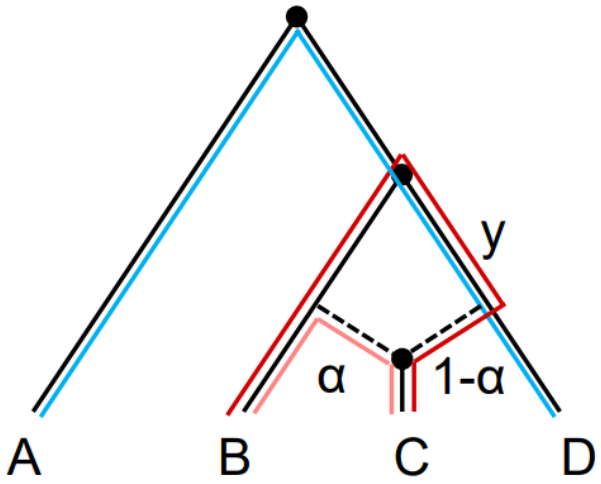


Figure 19: Geometrical interpretation of $f_4(A,D,B,C)$ with admixture. (Lipson, 2020)

Considering the tree in Figure 19, the value of $f_4(A,D,B,C)$ is given by the sum of the intersection between the two possible paths from B to C with the path from A to D, each of which is weighed with the respective coefficient of admixture. In this case $f_4(A,D,B,C) = (1-\alpha)y$. With a tree topology like this one, it is no longer possible to find a f4 statistic equal to 0, since each possible permutation leads to intersecting paths. This property represents a good way of testing the eventuality of admixture given four populations. In particular, if some previous knowledge is available regarding the populations under study, it's easier to choose the most informative statistics and to interpret the results. An example of this, that will be applied later in this work, consists in choosing the four populations in a proper way to study a specific possible admixture event. Choosing a population that is known to have a separate genetic history with respect to the other three, which is called outgroup population, facilitates the analysis, allowing to draw the topology of the tree in an easier way. Assuming one wants to investigate if one of the two populations A and B that share a recent genetic history is the result of an admixture event that involves a third population C that is known to have split from A and

B in a previous moment in history, calculating $f_4(A, B, C, D)$ with D represented by an outgroup is very informative.

Another type of f statistics that was used in this work is the so-called outgroup f3. As stated before $f_3(A, B, C)$=$f_4(A, B, A, C)$, which can be easily interpreted in a geometrical way as the path from C to the father node of A and B in Figure 20.

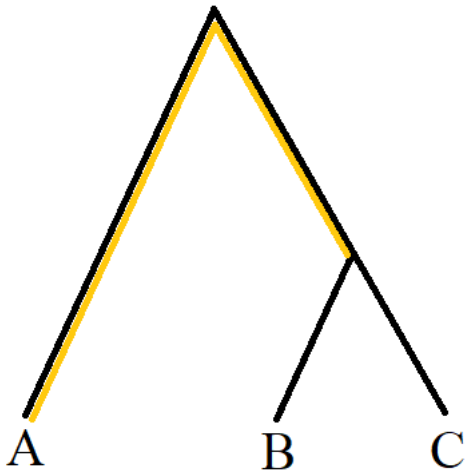


Figure 20: Geometrical interpretation of $f_3(A, B, C)$

By choosing A as a proper outgroup with respect to B and C, it is possible to fix either B or C as the population under study and then calculate this statistic for different Cs. The comparison of the different values obtained for the different choices of C gives an insight of the 'closeness' in terms of genetic history between B and these populations, given that the bigger is the value of the f3, the longer is the yellow branch and the closer are B and C.

3.6-Rolloff

Rolloff (Moorjani P, 2011) is a method that analyzes Linkage Disequilibrium in admixed populations to infer information about the date of the admixture event. An admixture event involving two previously separated populations creates in the admixed population an admixture LD (ALD) caused by association between SNPs inherited together from one of the two ancestral populations. After the event, over the course of generation, recombination breaks down the associations. During the process of meiosis, in which haploid cells called gametes are created starting from somatic cells of the individual that are diploids, recombination can take place through a mechanism called crossing-over. This mechanism consists in the exchange of portions of the couple of homologous chromosomes inside the cell, leading to the creation of gametes that will be made partly from the chromosome inherited from the mother and partly from the one inherited from the father. During reproduction a gamete will combine with a

gamete of the partner to form a new diploid cell that will generate a new individual. The process will repeat again through generations, causing subsequent cuts in the original chromosomes.

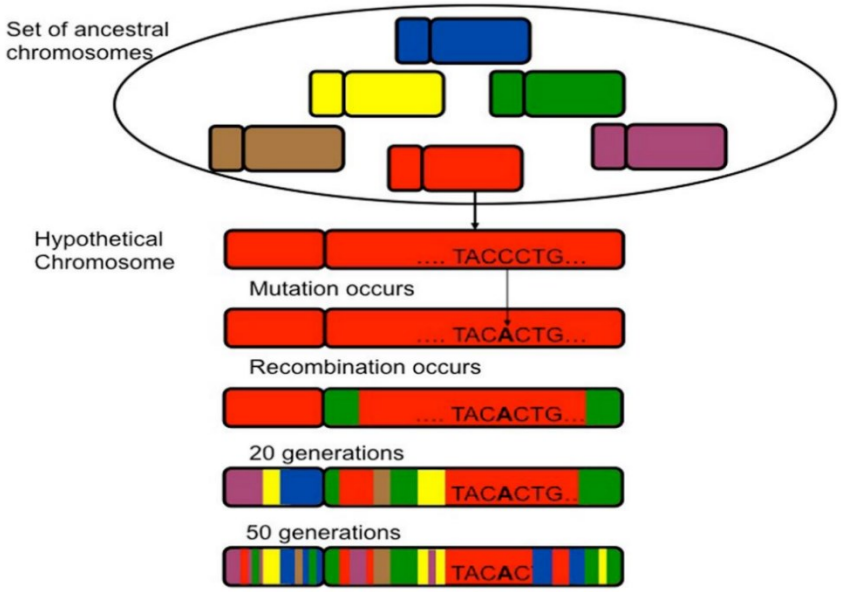


Figure 21: Example of effects of recombination

Due to this process the associations of SNPs are broken through time following an exponential decay and by studying the resulting LD in the admixed population it is possible to infer the amount of time necessary to reach the observable situation. The rate of the exponential decay gives the age of admixture.

Rolloff makes some assumptions that are necessary to simplify the situation. One of these is the choice to work only with pairs of SNPs when studying LD and not with multimarkers haplotypes. Another one is to consider the admixed population as homogeneous, a condition that is usually reached only some time after the admixture event. Finally, the admixture event is considered as a pulse, meaning that it occurred very quickly, and unique, meaning that no other admixture event happened afterwards with the source populations. Calling the two admixing population A and B, a weight function w(s) is required such that, for every SNP s, it gives a value that is positive when the frequency of the variant allele is higher in population A and negative otherwise. If modern populations that can represent good proxies for the admixing populations are available, a good weight function could be the difference of frequencies between A and B. Otherwise there are ways to obtain a good weight starting from admixed populations that partially contain the sources with known proportions. After having defined the weight function, the LD between SNPs is tracked by a score $z(s_1, s_2)$ defined as (Patterson N., 2012):

$$z = \frac{\sqrt{m-3}}{2} \log\left(\frac{1+\rho}{1-\rho}\right)$$

where m is the number of samples in which neither s_1 or s_2 have missing data and ρ is the Pearson correlation between the vectors v_1 and v_2 containing the genotype counts of the variant allele of the two SNPs. The weight function and the z-score are correlated along bins of SNPs that are defined, given a bin size x, as (Patterson N., 2012):

$$\mathcal{S}(d) = \{(s_1, s_2) | d - x < u_2 - u_1 \leq d\}$$

where u_1 and u_2 are the genetic positions of the two SNPs and d takes as values x, 2x, 3x, etc. The correlation is then (Patterson N., 2012):

$$A(d) = \frac{\sum_{s_1, s_2 \in \mathcal{S}(d)} w(s_1) w(s_2) z(s_1, s_2)}{\left[\sum_{s_1, s_2 \in \mathcal{S}(d)} (w(s_1) w(s_2))^2 \sum_{s_1, s_2 \in \mathcal{S}(d)} (z(s_1, s_2))^2\right]^{1/2}}.$$

If n is defined as the number of generations passed from the admixture event, it is possible to say that two alleles at distance d in an admixed individual have a probability e^{-nd} of having belonged to the same chromosome at the time of admixture n generations before. By fitting $A(d)$ to $A_0 e^{-nd}$ by least square it is possible to obtain the value of n.

3.7 – Masking

Masking is a process that exploits the results of local ancestry estimation and allows to address the different components of an individual's genome separately. An example that clarifies the masking process is the one in (Pagani L, 2015) that considers Ethiopians. The individuals of this population have a sizeable component of West Eurasian genetic ancestry in their genome and the authors were interested in isolating the unadmixed Ethiopian component to study it separately. Firstly, through PCAdmix local ancestry was estimated.

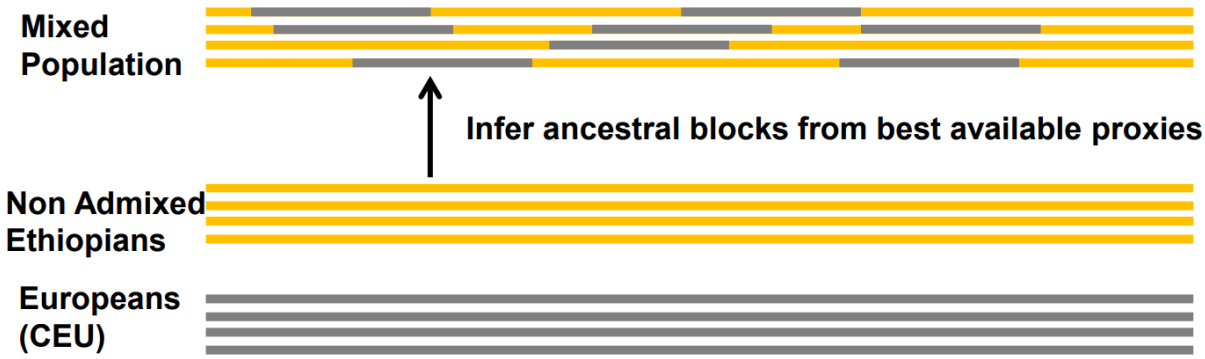


Figure 22: First step of masking

Then, a new set of individuals was created by masking blocks of SNPs that did not belong to the population of interest. This consists in treating the positions corresponding to these SNPs as if no data were available for them in the new individuals. The result is a set of individuals with 'gaps' inside their genome, made only by blocks of SNPs belonging to the population of interest.

Figure 23: Population resulting after the masking process

The original individual, depicted in the picture as a single line since it still holds the assumption discussed in the PCAdmix paragraph according to which we consider individuals as haploid, gets split into individual_Ethiopian_unadmixed and individual_European that can be analyzed independently from one another.

The general masking process starts with the creation ok K new individuals for each of the ones belonging to the admixed population, where K is the number of ancestral populations used in PCAdmix. Then the program reads the .vit file, one of the outputs of PCAdmix that contains for each window of SNPs the code corresponding to the most probable ancestry, and the .fbk file, containing the highest probability associated with that window. For each window, if the probability is higher than a threshold set by the user, the individual corresponding to the code in the .vit file keeps the information contained in the original individual, while the other ones have '0's in the SNPs of the corresponding window as if they were being masked. If the

probability is below the threshold, instead, all the new individuals are masked in that window. In this way the user is able to set a degree of confidence based on the desired level of precision.

Due to the reasoning that will be explained in section 4.2.2, the process of masking was slightly modified in this work in order to distinguish the two European sub-groups of PUR individuals. The first difference introduced in this work consists in creating K+1 individuals to take into account the second European ancestry. The second one consists in considering, for each window whose code in the .vit file corresponded to the European ancestry, the codes of the two neighboring window. If both windows are associated with the American ancestry, the code of the window is set to a new value that corresponds to the Ancient European ancestry. The same is done for blocks of consecutive windows of European ancestry that start and end with American ancestry. The threshold used for masking in this work was 0.9 and, to be even more cautious, a trimming process was operated by masking the two windows at the extremity of each streak belonging to one ancestry. In this way all the windows that are at the boundary between two different ancestry blocks and that may be ‘nosiy’ are ignored.

4-Results and Discussion

In this chapter the analyses performed using the methods previously described will be shown and the results will be interpreted. The main idea is to investigate, using different approaches, the possible presence of DNA segments coming from European populations and dating back before 1492 AD to see if there is genetic support to the theory of Lucio Russo. The first part will focus on aDNA of the Caribbean area, the second one on the analyses made on modern genomes.

4.1-Ancient DNA analysis

As it has been discussed before, the most direct way of studying pre-contact dynamics regarding the indigenous people of the Caribbean area is through samples of aDNA. The samples available from the Caribbean area and dating back to at least 550 years are 207, coming from 39 different sites in 8 different countries. Of these, only 16 come from the Lesser Antilles, 12 from St. Lucia and 4 from Guadeloupe.

The first analysis that was made consists in performing a series of f_4 tests of the type:

$$f_4(x, USA\_Anzick\_Realigned, French, MSL)$$

with different ancient Caribbean populations for x. USA_Anzick_realigned (Rasmussen M, 2014) is a sample belonging to an individual found in West Montana (USA) from more than 12000 years ago. This sample is the oldest available from the American continent and shows affinity with all the ancient samples found further south in the continent, meaning that he was probably part of the wave that settled it. His dating guarantees that any possible pre-Columbus contact between Caribbean indigenous and Mediterranean people happened after his death. It is thus possible to use Anzick as a reference for 'unadmixed' American, in the sense that he does not have European traces from the last 2500 years in his genome. French is a population sampled in HGDP and here serves as representative of European genome. MSL is a population from Sierra Leone in Africa sampled in the 1000 genomes project and serves as outgroup. The tree being tested with this f_4 test is the one in Figure 24.

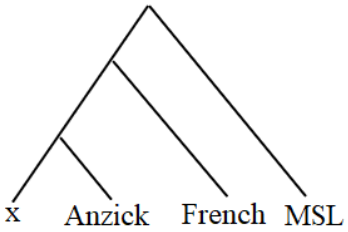


Figure 24: Tree tested with $f_4(x, USA\_Anzick\_Realigned, French, MSL)$

The idea behind this test is to try to see if there is some European component in the ancient samples of the Caribbean area that is absent in Anzick, meaning that it would have come from an admixture event with a European population. Since both archaic and ceramic population are being tested, it is also possible to see if there is a difference in terms of European traces between the two. Obtaining a value significantly different from zero would indicate that the tree structure being tested does not describe the situation properly and an admixture event would be required. In this specific case, the evidence of a European component would be supported by a significantly positive value for the f_4 test. A non-significant value would mean that Anzick and population x are equally non-European. A test with MXL as x was also performed, where MXL is a population of Mexicans sampled in the 1000 genomes project and is known to be admixed with European and American components. In Figure 25 and 26 the empty dots represent non-significant values of the f_4 test, namely values with a Z-score smaller than 3, while the full dots represent significant values.

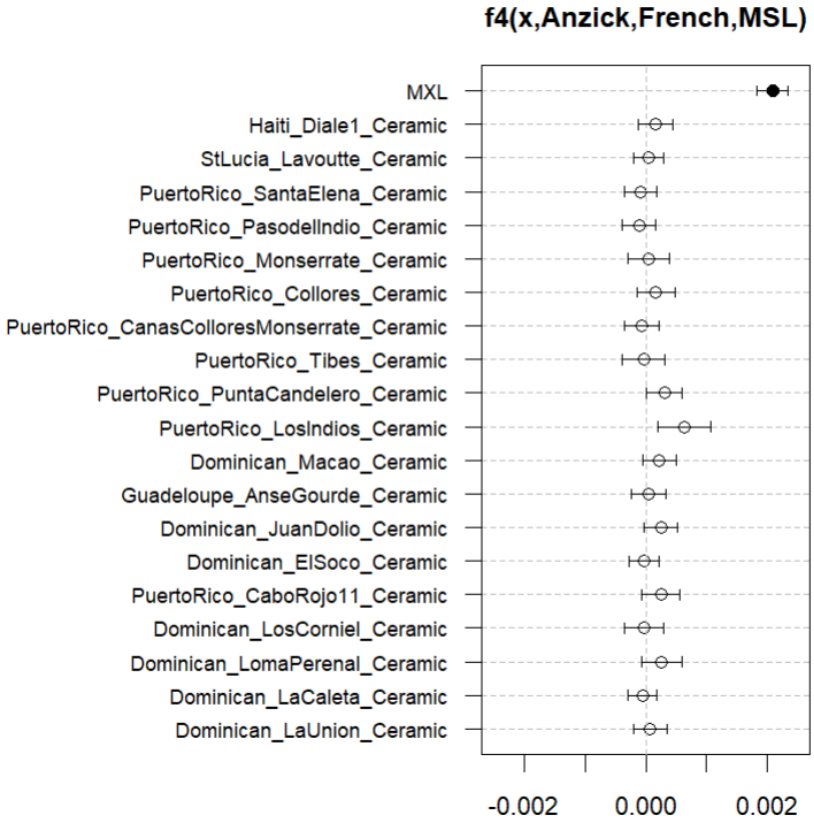


Figure 25: Results of f_4 (x, USA_Anzick_Realigned, French, MSL), part I

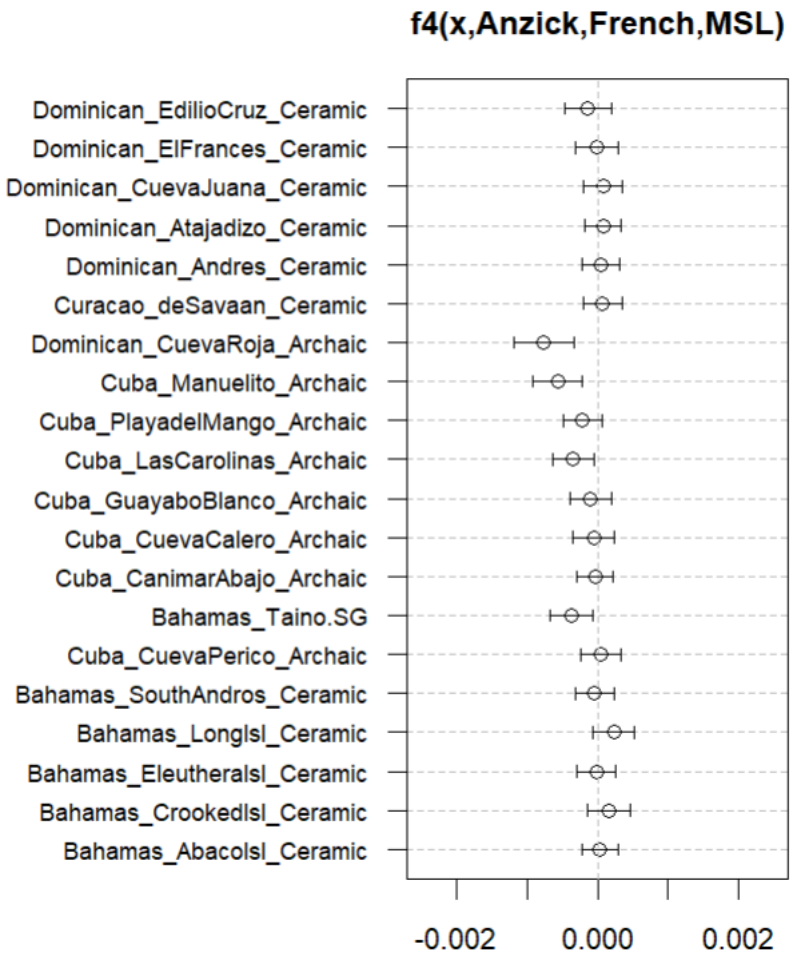


Figure 26: Results of f_4 (x, USA_Anzick_Realigned, French, MSL), part II

In this case, except for MXL, all the Z-scores were non-significant and in particular smaller than 2. These results suggest that none of the ancient populations tested show significant traces of European genome. The table T1 with the detailed results is consultable in the additional material section.

A further analysis was made on two of the populations from (Nägele K, 2020) that have samples that span on a time range that goes from more than 2700 years ago to 1500 years ago. These populations were split in two sub-population, one with the oldest samples dating back to what could have been before the contacts we are looking for and one with the most recent samples. These two populations, Canimar Abajo and Cueva del Perico in Cuba, and their sub-groups are observable in Figure 27.

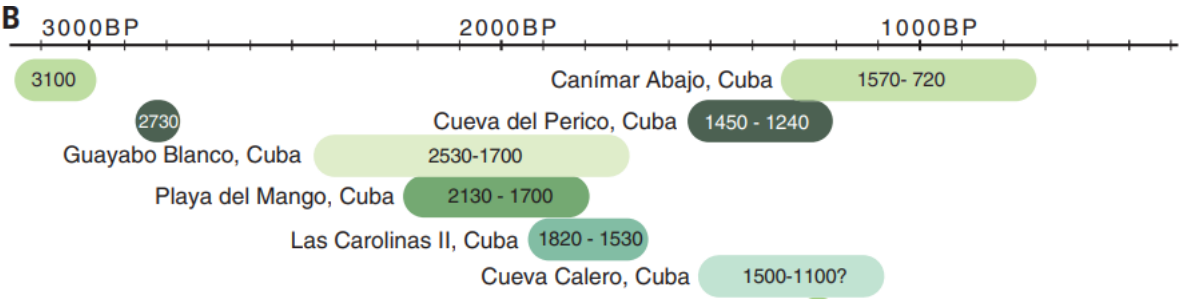


Figure 27: Dating of some of the samples in (Nägele K, 2020)

The test performed is of the type:

$$f_4(x\_after, x\_before, French, MSL)$$

with x_before that plays the same role played before by Anzick. A non-significant result would suggest that the most recent samples from this site do not have a greater affinity towards Europeans with respect to the oldest samples.

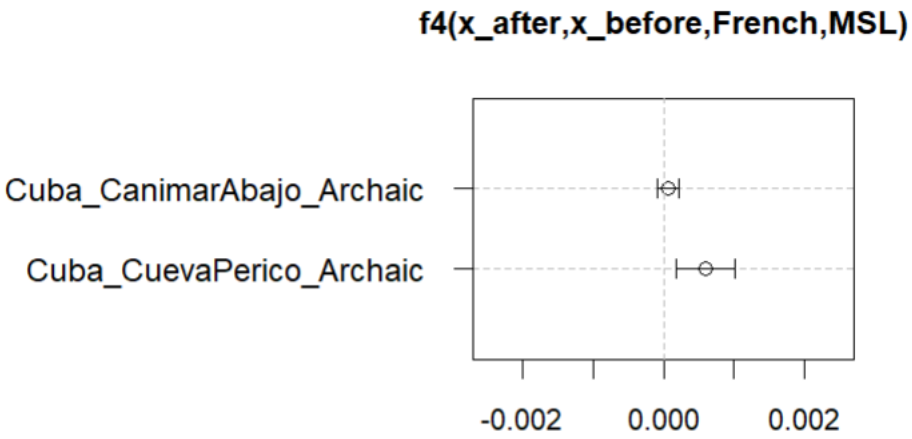


Figure 28: Results of f_4 (x_after,x_before,French,MSL)

No signs of particular affinity are shown by either of the two populations, with the Z-scores that are both smaller than 3. Detailed results are in table T2 in Additional Material.

Reading the paper of (Nägele K, 2020) it emerges a specific individual (PDI003) coming from Paso del Indio in Puerto Rico that was excluded from their analysis since it showed a significant European component in his genome, but it was not classified as contaminated.

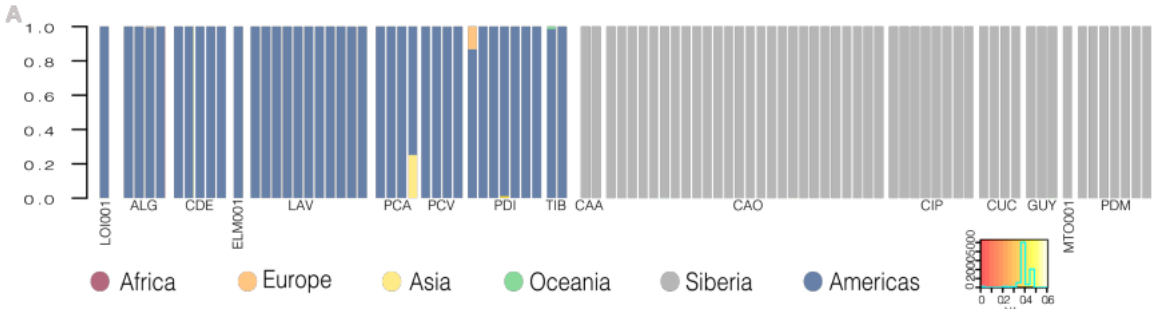


Figure 29:Model based estimation of ancestry components of the ancient Caribbeans sampled (Nägele K, 2020)

In Figure 29 it is possible to see that PDI003 (the leftmost vertical line in PDI group) has an approximate 10% of European component in his genome. A series of f_4 tests were performed of the type:

$$f_4(x, y, PDI003, MSL)$$

using for x different ancient Mediterranean populations and for y populations involved in colonialism in America, to investigate if PDI003 was particularly attracted by one of these two groups. Just for the sake of comprehension, in Figure 30 the modern Europeans have been placed on the left side and the ancient on the right side since a negative value of the f_4 tests would suggest a gene flow between population y and PDI003.

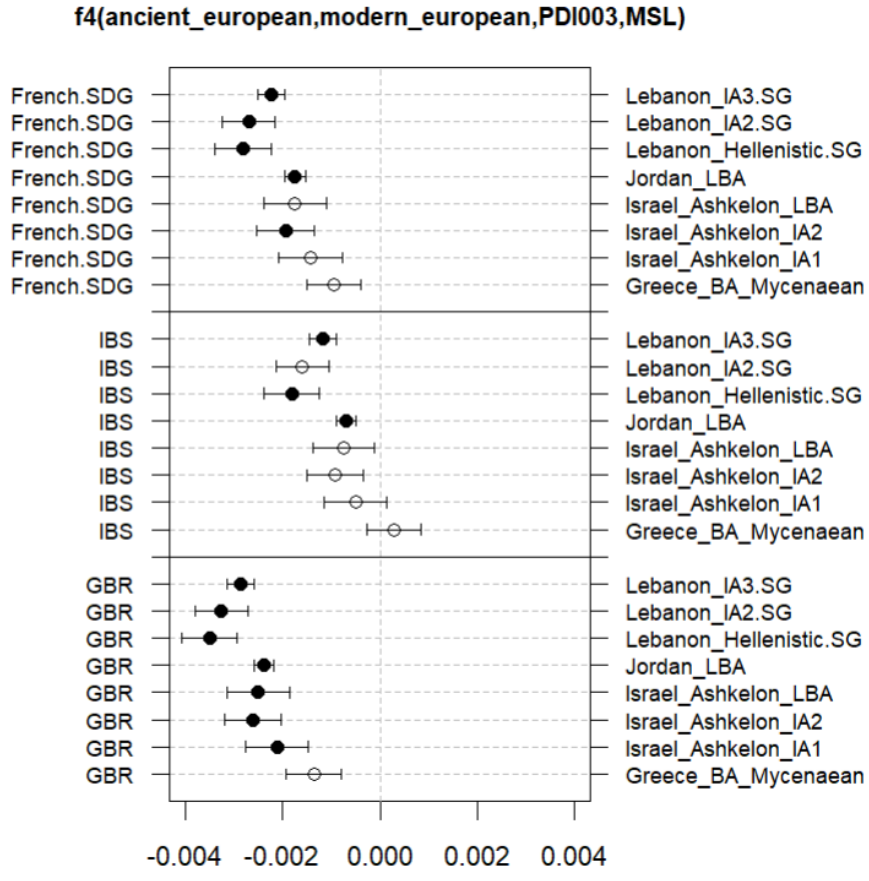


Figure 30: Results of f_4 (x, y, PDI003, MSL)

It is possible to notice that the tests tend to give negative values and in many cases with Z-score greater than 3, pointing to a higher affinity of PDI003 to Western contemporary Europeans than to ancient Mediterranean groups. Therefore, studying in more detail this sample is not in the interests of this work since it doesn't contain traces of ancient Mediterraneans in the genome. It is important to mention that none of the analysis is based on more than 164187 SNPs due to the fact that PDI003 has only information for 172819 SNPs and this makes the analysis less relevant. Detailed results are contained in table T3 in Additional Material.

A final analysis carried out on ancient Caribbean samples focused on two particular SNPs, rs16891982 and rs1426654. rs16891982 has alleles C and G and it is almost fixed in Europeans, with alle G having a frequency of 95.5%, while in the rest of the world it appears predominantly in his C version (NCBI, rs16891982, 2021). This SNP influences skin-pigmentation and hair color and allele G is typically associated with light skin. rs1426654 has alleles A and G and it is even more fixed in Europeans than the previous one, with allele A having a frequency of 99.6% and suggests a European ancestry in general (NCBI, rs1426654, 2021). This mutation too is involved in skin-pigmentation, with allele A that is typically associated with light skin. These markers were investigated in the same ancient Caribbean population used for the previous analysis, looking for 'G's in rs16891982 and 'A's in rs1426654. The corresponding lines from the .tped file are shown here:

5 rs16891982 0.517788 33951693 0 0 0 0 C C 0 0 C C C C C 0 0 C C C C C 0 0 0 0 0 0 C C C C 0 0 C C 0 0 C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C 0 0 C C C C C C C C 0 0 C C C C 0 0 C C C C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 C C C C C C C C C C C C C C C C 0 0 C C 0 0 0 0 C C C C 0 0 0 0 C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C 0 0 0 0 0 0 C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C 0 0 C C C C C C C C C C 0 0 C C C C C C C C C C C C C C 0 0 C C C C C C C C C C C C 0 0 C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C 0 0 C C C C C C C C C C C C C C 0 0 C C C C C C C C C C C C C C C C 0 0 0 0 0 0 C C C C C C C C C C C C C C 0 0 C C 0 0 C C 0 0 C C C C C C C C 0 0 0 0 0 0 C C C C C C 0 0 C C C C C C C C C C C C C C C C 0 0 0 0 0 0 0 0 0
0

15 rs1426654 0.608731 48426484 0 0 0 0 0 0 0 0 0 0 0 0 G G 0 0 0 0 0 0 0 0 0 0 0 0 G G G G 0 0 0 0 0 0 0 0 0 0 G G G G G G 0 0 0 0 G G G G G G G G 0 0 0 0 0 0 0 0 0 0 G G 0 0 G G G G G G 0 0 G G 0 0 0 0 0 0 0 0 0 0 0 0 0 0 G G G G 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 G G 0 0 0 0 G G 0 0 0 0 0 0 G G G G G G G G G G G G G G G G 0 0 G

G G G 0 0 0 0 0 0 0 0 0 0 G G 0 0 0 0 0 0 0 0 0 0 0 0 G G G G G G 0 0 0 0 0 0 G G 0 0 G G 0 0 G G G G 0 0 0 0 G G 0 0 0 0 0 0 0 0 0 0 0 0 0 0 G G 0 0 0 0 0 0 G G 0 0 G G 0 0 0 0 0 0 0 0 0 0 0 0 0 0 G G G G 0 0 G G G G G G G G G 0 0 G G 0 0 G G G G G G G G G 0 0 G G G G 0 0 0 0 0 0 0 0 G G 0 0 G G 0 0 0 0 0 0 G G 0 0 0 0 0 0 G G 0 0 0 0 0 0 G G 0 0 0 0 G G 0 0 0 0 G G 0 0 0 0 0 0 0 0 0 G G 0 0 0 0 0 0 0 0 0 0 0 0 G G 0 0 0 0 0 0 0 0 G G 0 0 0 0 0 0 0 G G 0 0 0 0 0 0 G G 0 0 0 0 0 0 0 G G G G 0 0 G G G G 0 0 0 0 0 0 0 0 0 0 0 0 0 0

For both SNPs there are no cases of alleles associated with European ancestry in the samples available. 0 means that there is no information available for the corresponding individual in that particular position of the genome.

All the tests that involved ancient Caribbean population showed no particular evidence of European traces. The next way of analyzing the situation consists in working with modern genomes from this area.

4.2-Modern DNA analysis

This part focuses on analyzing modern DNA, trying to get insights on past events regarding the Caribbean area from recent samples. Firstly, a specific analysis on Mayas sample was carried out to answer a specific doubt present in '*L'America dimenticata'*, then Puerto Ricans from the 1000 Genomes Project (PUR) data were analyzed in detail to investigate the Antilles' situation.

4.2.1 Mayas' admixture date estimation

Before starting to analyze some modern DNA to get more information about the population of Antilles, an analysis was carried out to answer a specific doubt raised by Russo in his book. The author, after having shown similarities between the culture of Mayas and the Mediterranean ones, is questioning why in genetic studies regarding native Americans populations only Mayas show clear European traces. He's referring to a paper (Hellenthal G, 2008) that worked on the HGDP dataset with methods that, given the year, were not as powerful as the one that are available today. The authors of that paper state that the European traces in Mayas' DNA are presumably due to post-Columbian Admixture, but don't prove it, so Russo asks himself if this could be due to pre-Columbian contacts. European traces are actually present in Mayas' DNA, as we can see in the admixture plot of Figure 31.

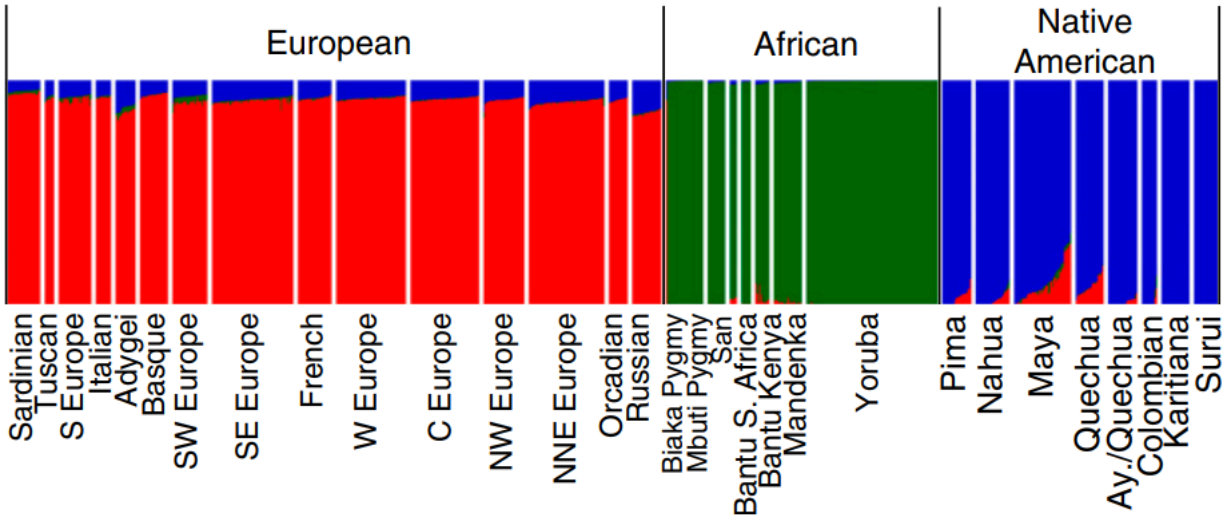


Figure 31: Admixture with k=3 on part of HGDP data (Bryc K, 2010)

To clarify the doubt, an analysis was carried out in this work, taking into account HGDP data of Mayas and running Rolloff using a European source and a 'pure' Native American one to trace the date of the admixture event that led to the observable situation. Karitiana was used as American source and French as European source, both coming from HGDP. The results are contained in Table 1.

CHR	SNPs	Estimated_date (generations)
1	49096	9.057736
2	53205	9.036638
3	44135	9.443648
4	39503	9.185237
5	40535	9.010204
6	42759	8.79816
7	35041	9.018672
8	36917	9.158331
9	30856	9.002671
10	34109	8.920502
11	31632	9.091138
12	31496	8.678555
13	24919	8.963737
14	21237	9.236256
15	19338	9.125371
16	19497	9.09157
17	16419	919127
18	19948	9.109103
19	10557	9.111869
20	16709	8.936515
21	9536	9.02067
22	9619	8.995627

Table 1: Rolloff results

For all chromosomes the estimated admixture date is close to nine generations. Considering that a generation is usually estimated in thirty years, these results leave no doubt about the fact that these European traces are post-Columbian.

4.2.2-Analysis of PUR

The first step to carry out analyses using modern genomes is to assign each segment of DNA to the ancestry it belongs to, in order to be able to analyze them separately. To do this, PCAdmix was applied to individuals belonging to the PUR population of the 1000 Genomes Project, the one that was identified as best proxy. According to the history of the Caribbean area that was discussed in section 1.2, the expectation is to find three different genetic components inside the genome of these individuals, the Native American one, the European one and the African One. This is confirmed by admixture tests that have been carried out on the 1000 genomes dataset in different studies (Martin AR, 2017) (Rustagi N, 2017).

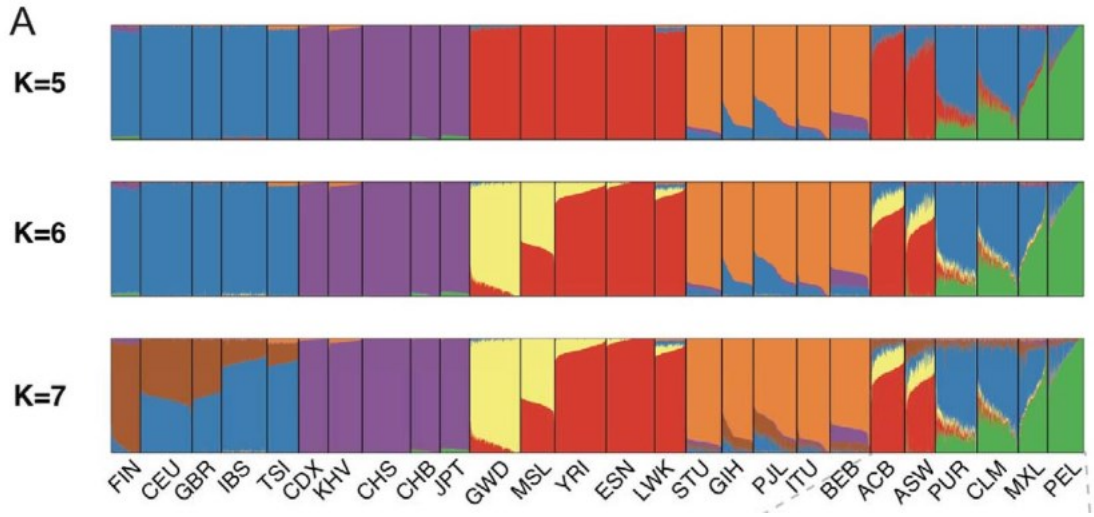


Figure 32: Admixture plot from (Martin AR, 2017) . Already with k=5 it is possible to see the three genetic components of PUR, with red=Sub Saharan African, blue=European, green=Native American

PCAdmix was therefore run with k=3, using as ancestral population Peruvian in Lima, Peru (PEL) for Native Americans, British from Great Britain (GBR) for Europeans and Yorubas from Ibadan, Nigeria (YRI) for Africans. PEL was chosen as source for the Native American component since, according to the same admixture tests previously mentioned, it is the American population in the dataset with the highest percentage of Native American ancestry. The choice for the European source population was instead based on a different reasoning. Instinctively the first choice may be to use IBS (Iberians in Spain), since Spanish were the most involved population in the colonization of central and south America, especially in the first

centuries after 1492. The problem with using IBS is that afterwards, when analyses will be carried out to study the European component of the admixed population, there may be a bias towards the population used to extract the component, since this will tend to be similar to the source used in PCAdmix due to the way the software works. Thus, it may be wise to use a population like GBR that is still a western European population, but also allows to eliminate potential biases towards IBS. PCAdmix was run on the 22 autosomal chromosomes considering the 1095838 SNPs resulting from the intersection between the 1000 Genomes dataset and the Allen Ancient DNA Resource dataset of aDNA (Allen Ancient DNA Resource, 2021), in order to be able to compare populations from both datasets afterwards.

Once PCAdmix was run, the technique of masking was ready to be applied. Before masking PUR individuals, however, it is necessary to understand what is expected to be obtained after the process. Reasoning under the hypothesis that considers pre-Columbus encounters, the European component obtainable through masking will be made mostly by blocks of SNPs belonging to modern Europeans, the ones involved in the colonization started in 1492, and just in a small fraction by blocks belonging to ancient Mediterraneans considering that, from the few information available about this eventuality, it was probably something marginal in case it really happened. Ideally, it would be necessary to perform a PCAdmix with k=4, considering also an ancient Mediterranean population, but aDNA and its poor quality does not allow to do this. Furthermore, the expected similarity between present day Europeans and ancient Mediterranean populations would decrease the accuracy of this analysis. It is therefore necessary to find another way to separate the two European components. The one adopted in this work is based on the fact that the time that passed from the first hypothetical encounters and the colonization can help to differentiate the two ancestries. Due to the process of recombination, blocks of SNPs belonging to the original chromosomes tend to become shorter and shorter with the passing of time, according to a theoretical exponential decay that stands under several simplifications (Pool JE, 2009).

With this assumption in mind and knowing that the process of crossing over that mixes chromosome between them occurs randomly across the genome, it’s possible to understand the following series of pictures that illustrates the two possible simplified scenarios, one with pre-Columbus contacts and one without them, and shows how it may be possible in the first case to distinguish the two different European components in the admixed genomes.

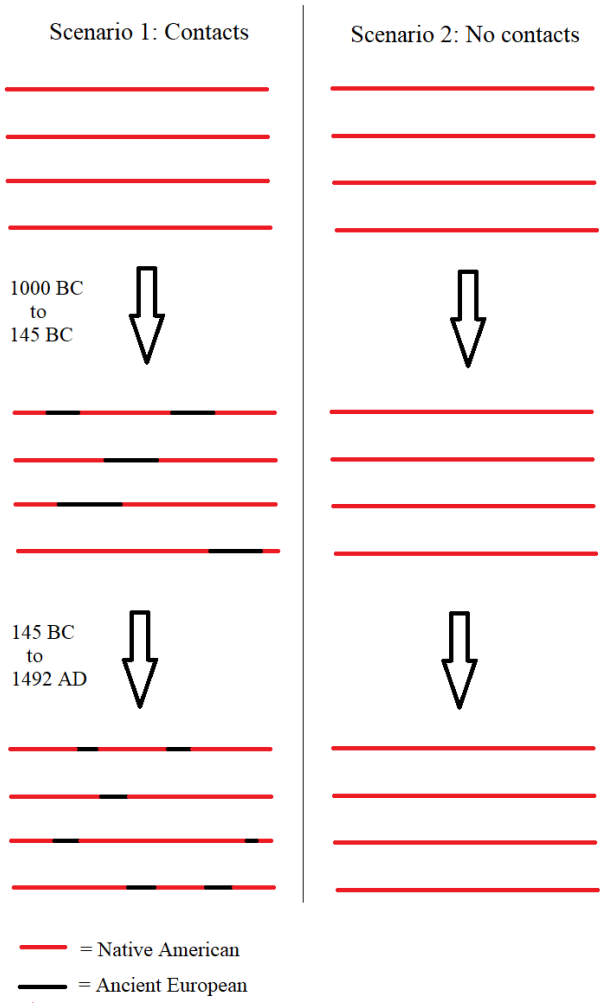


Figure 33: Simplified schema of the evolution of genomes of Antilles' population - part I

Figure 33 illustrates the two possible scenarios until right before the arrival of European colonizers. The right side does not change with the passing of time since it considers indigenous as isolated from the European world. The left side is characterized by two steps. The first one consists in the period in which contacts occur, from 1000 BC, which has been set as the possible initial date of encounters, until the cultural collapse in 145 BC. During these years blocks of SNPs belonging to European ancestry enter the genomes of inhabitants of the island. The second one consists in the period of momentaneous absence of contacts, from 145 BC to Columbus' arrival in 1492 AD. These years are characterized by the process of recombination, that involves admixed individuals and 'fully American ones', which causes European blocks to become shorter and to partially disappear.

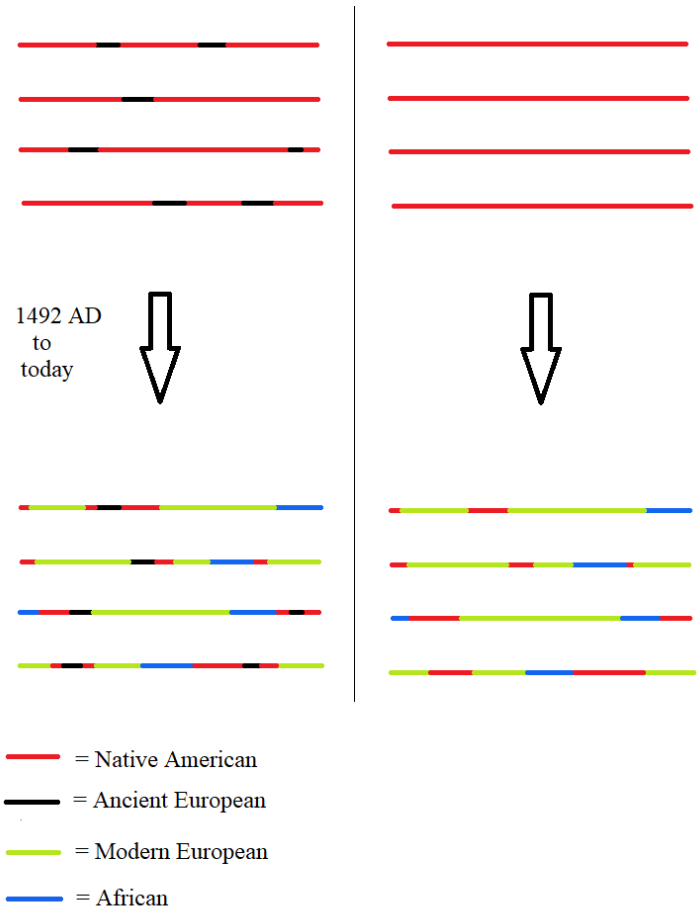


Figure 34: Simplified schema of the evolution of genomes of Antilles' population - part II

Figure 34 illustrates the entrance of two new genetic components consequently to colonialism and slave trade. The idea of this work is to exploit the distribution of the fragments along the genome, particularly the one of European components and their neighbors. The example shown is simplified, but intuitively the small blocks of Ancient European ancestry, that before 1492 AD are positioned between two blocks of Native American ancestry, have a higher probability of being in this same configuration with respect to Modern European ones. This because the only way for the black pieces to be surrounded by something different than red is if crossing over occurs right inside the small black portion. On the other hand, for a green piece to end up surrounded by two red pieces, it is necessary that two crossing over events occurs within this block, both mixing with chromosomes that are red in those areas. European blocks are thus separated on the basis of their neighbors, putting the ones in the configuration AMR-EUR-AMR (with an arbitrary number of EUR blocks) in group A and all the others in the group B. In this way, if the correct scenario is the first one, group A is expected to be enriched with Ancient European blocks, in the other case the two groups should look the same.

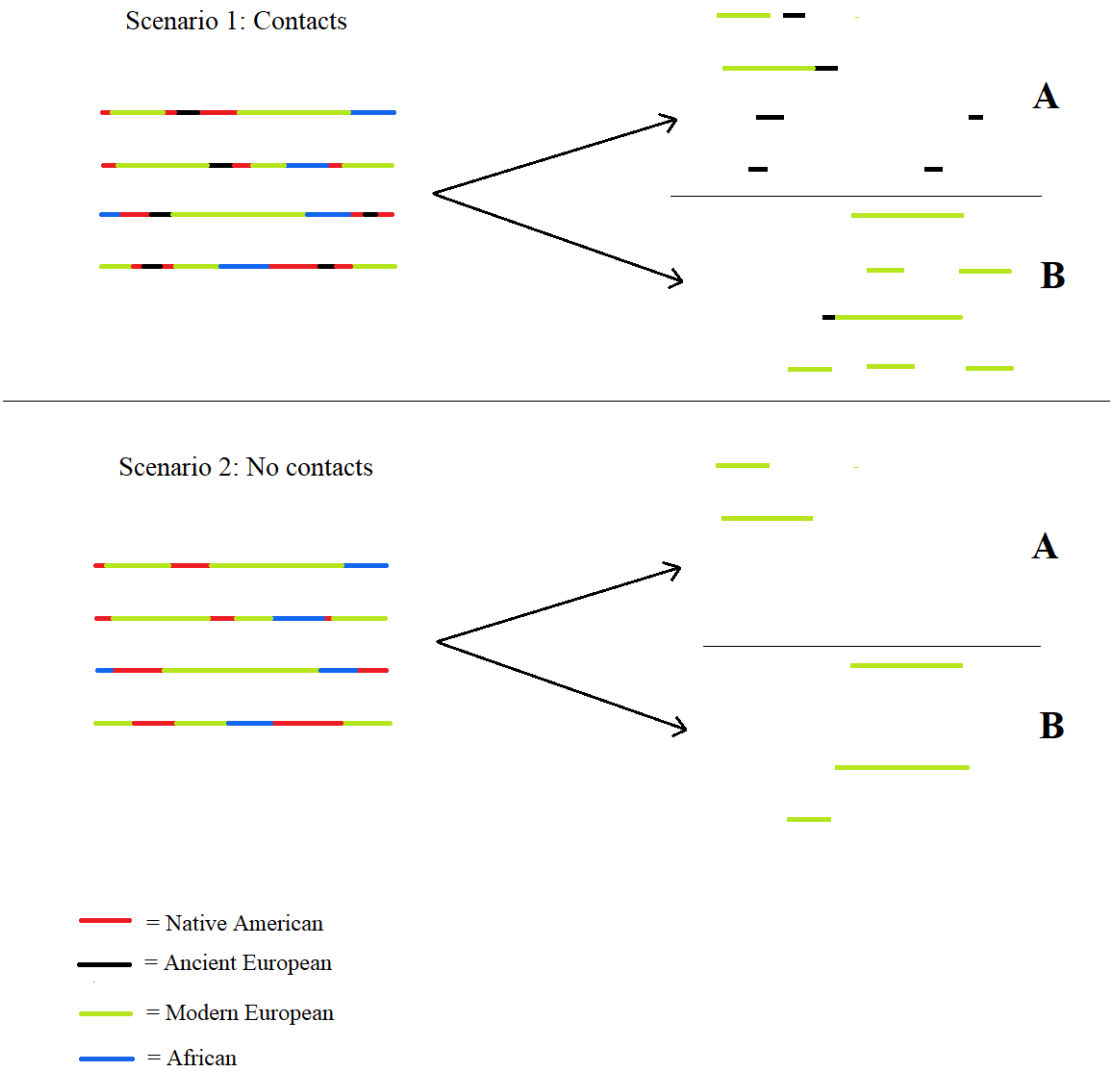


Figure 35: Resulting subgroups after the separation based on the configuration of blocks of SNPs. Group A contains the European blocks placed between native American blocks, Group B the other European blocks

With this idea in mind, the process of masking was slightly modified as described in section 3.7. The new individuals obtained through this process are called individual_AMR, individual_AFR, individual_mod_EUR, individual_anc_EUR. To check if the masking process was successful, PCA was applied.

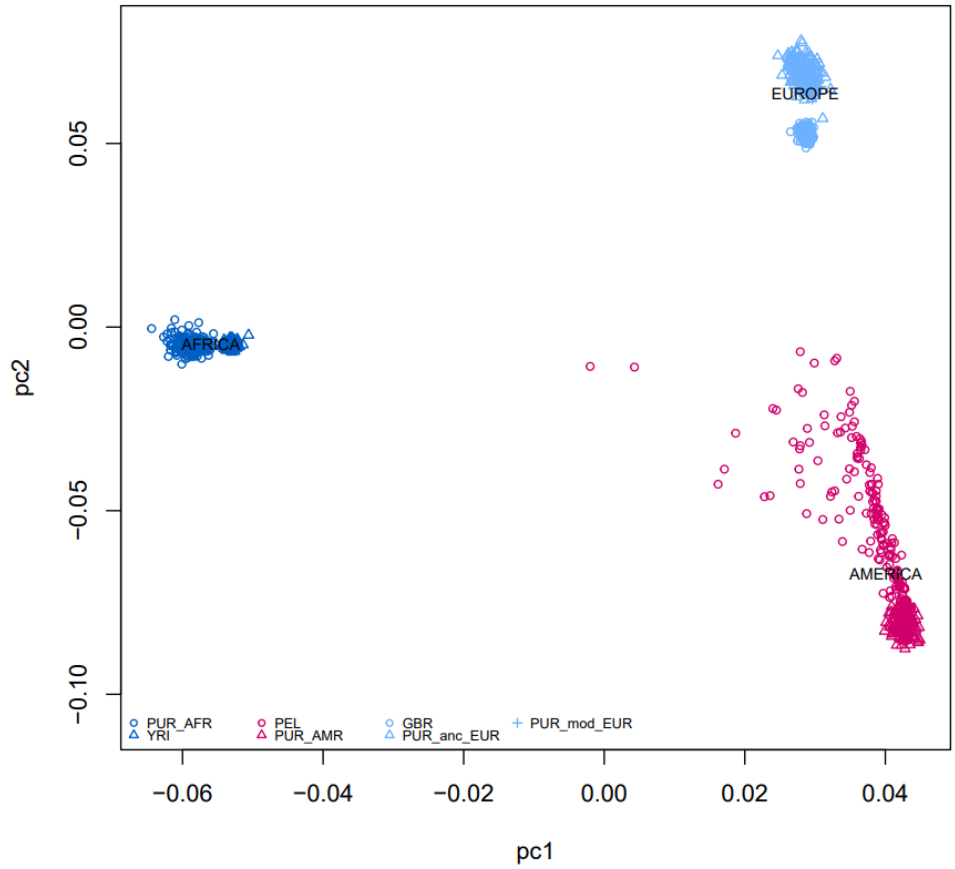


Figure 36: PCA plot of the masked population and the sources used

As it emerges from Figure 36, the position of the masked populations in the PCs space is coherent with the masking process that was carried out. The two European components almost overlap in the top light-blue cloud and appear to be more similar to each other than either of them to GBR. A zoom in the European area (Figure 37) gives a glimpse of some crosses representing Ancient Europeans in the bottom-right corner of the upper cluster.

Figure 37: Zoom on the European cluster of the PCA of Figure 36

After having separated the different components of the PUR population, a series of tests was carried out to investigate them in detail. A way to get a first general idea of the situation is by applying ADMIXTURE to the database and see how these new 'populations' compare to other available genomes.

ADMIXTURE was run with values of k from 2 to 10, with k=9 that is missing because of technical problems. The one that best suits the data is k=4 according to cross validation as shown in Figure 38.

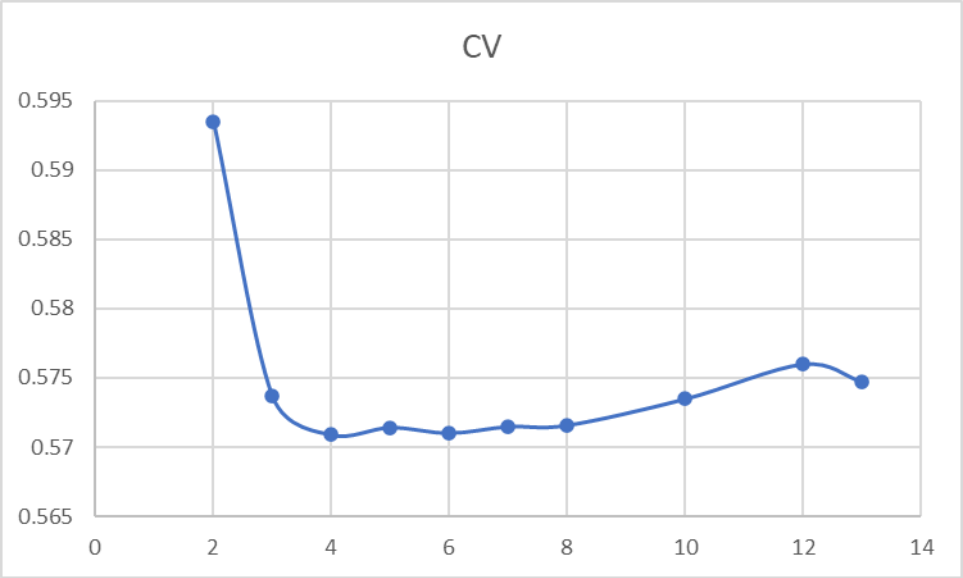


Figure 38: Cross validation for the different values of K

The estimates of the ancestral proportions for the different values of k are shown in Figure 39, which is reported in a bigger version in Figure A1 in the Additional Material section.

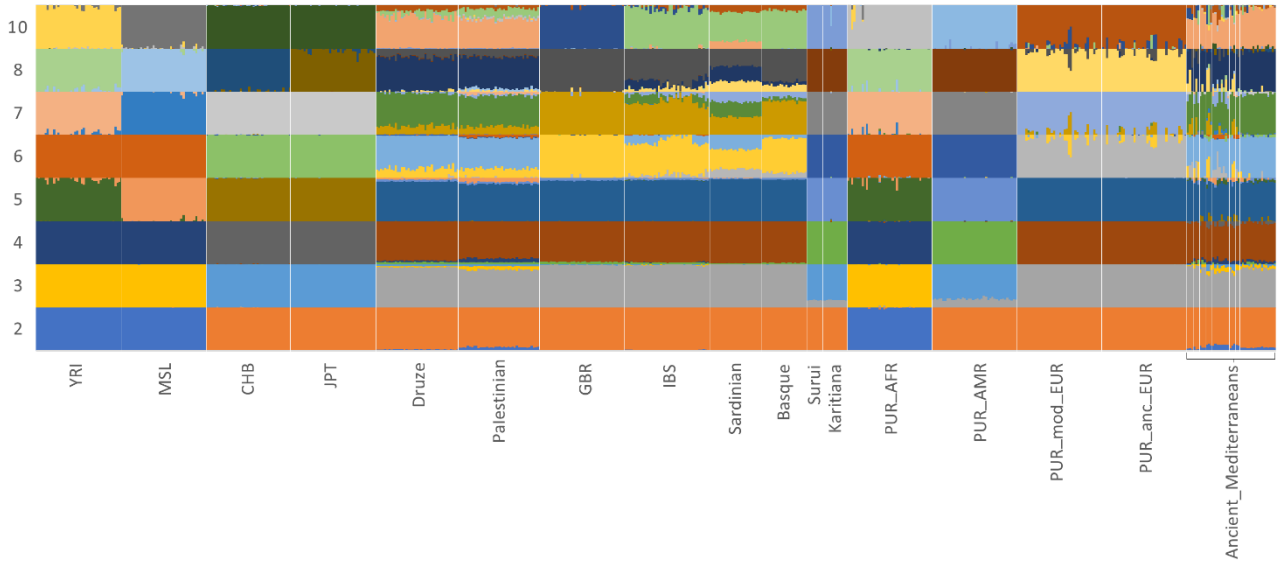


Figure 39: Admixture plot of the dataset used

With k=2 the program differentiates between African and non-African genetic ancestries. The African genetic component is present in YRI and MSL and slightly appears in near eastern populations like Druze, Palestinians and ancient Mediterraneans. PUR_AFR looks like a fully African population, confirming that the masking procedure for this component was successful. With k=3 the new component that appears is the one characterizing CHB and JPT, two populations from Asia (China and Japan respectively). This component is present also in Surui and Karitiana, two Native American populations, that appear as a mixture between this light blue component and the grey one that seems to represent Europe. This is due to the fact that, not having enough components to assign America one of its own, the best way to represent it is as a mixture of Asian and European components, coherently to the settling history of this area. PUR_AMR shows the same behavior as these two populations. With k=4 the new component is assigned to the American populations and to PUR_AMR confirming the successful masking, while the other populations are not affected. k=5 is not really interesting for the scope of this work, as it introduces a new African component that differentiates MSL and YRI, while non-African populations remain the same. With k=6 it is possible to start seeing some differentiation within European populations, with yellow that seems to be a northern feature as it characterizes GBR, while light blue seems to be associated with the Mediterranean area. It is interesting to see that PUR_mod_EUR and PUR_anc_EUR keep their own color, which is only slightly present in southern European populations like IBS, Sardinian and Basque and which may point to the founder event that brought genetically drifted European components into the PUR gene pool. They also show some trace of that yellow northern component, that may suggest a trace of the bias connected to PCAdmix that was discussed previously, but no trace of the green component that is associated with West Asian populations and partially with Mediterranean populations like IBS or Sardinian. The Ancient Mediterranean populations selected seem to behave similarly to Near Eastern one and don't show a particular affinity with PUR_anc_EUR. k=8 involves only Asian populations, while k=10 introduces further differentiations within Europeans population, but, also according to cross validation, it may start to overfit.

A first analysis carried out to investigate which is the population among the ones in the ancient panel that is closer to PUR_anc_EUR is a series of f_3 tests in outgroup mode of the type:

$$f_3(x, PUR\_anc\_EUR, MSL)$$

This type of test, with different ancient Mediterranean populations used as 'x' doesn't indicate that the closest one is necessarily similar to PUR_anc_EUR, but just that it is the one that shows greater affinity. Results are shown in Figure 40, with further details in table T4 in Additional Material.

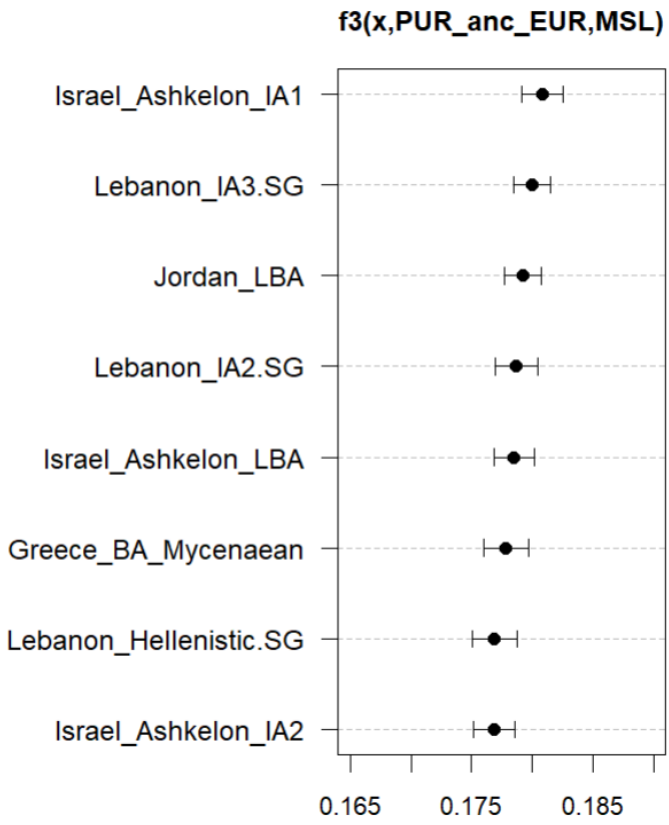


Figure 40: Results of $f_3(x, PUR\_anc\_EUR, MSL)$

The closest population appears to be Israel_Ashkelon_IA1, a group of individuals dating back to around 1200 BC. It's important to notice, however, that the population Israel_Ashkelon_IA2, found in the same site and dating to just a century later than the previous group, is the least similar one. This could mean that there is not a clear preference towards one of these populations. Lebanon_Hellenistic, that may have been the one with the highest expected value of f_3 because of its dating (around 150 BC) and its geographical collocation in what was once the Phoenixes' territory, is just above Israel_Ashkelon_IA2.

The same type of test can be carried out on the modern part to see the closest modern population to PUR_mod_EUR. The tests this time will be of the type:

$$f_3(x, PUR\_mod\_EUR, MSL)$$

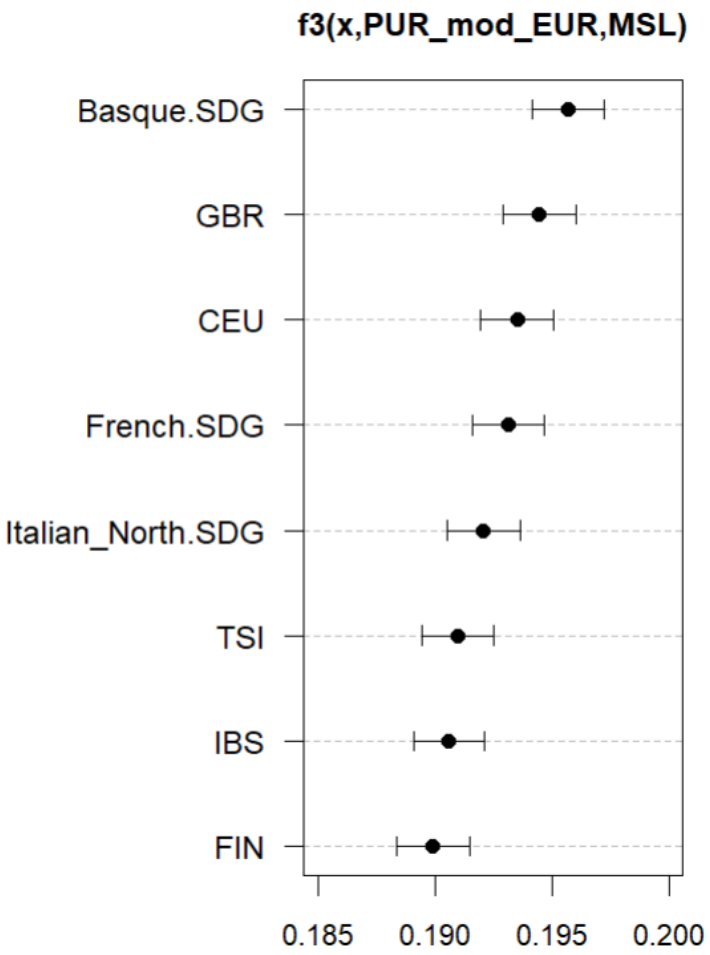


Figure 41: Results of f_3 (x,PUR_mod_EUR,MSL)

To clarify Figure 41, some of the population used were CEU (Central Europeans), TSI (Tuscans), FIN (Finnish). Detailed results are in table T5 in Additional Material. IBS is surprisingly low with respect to other European populations, considering that the Spanish population played a central role in the colonization of the island of Puerto Rico in the first centuries after 1492 AD. The high position of GBR may be caused by the bias previously described, which may have affected also the value of CEU. The Basque population from HGDP is at the first place and interestingly there are some traces of the fact that Basques had a significant impact on the Spanish rule of this island (Zubiri, 2009), but this is an issue that doesn't directly affect this work and may be investigated separately.

The previous f_3 tests give possible candidates for affinities with the subgroups obtained through the masking process, but are not an index of how similar they actually are. To explore this an informative analysis can be a f_4 tests of the type:

$$f_4(x\_anc, y\_mod, PUR\_anc\_EUR, MSL)$$

using for x_anc the Ancient European populations that gave the highest f_3 score and for y_mod the Modern ones following the same logic. A positive value, that would mean a greater affinity of PUR_anc_EUR with Ancient Europeans with respect to modern ones, is expected only if PUR_anc_EUR, following the logic discussed during the explanation of masking, is so enriched of ancient DNA to result closer to ancient than modern genetic components. For the same reason that was discussed early in this work, in the plot x_anc labels will be on the left and y_mod labels on the right. Detailed results are in table T6 in Additional Material.

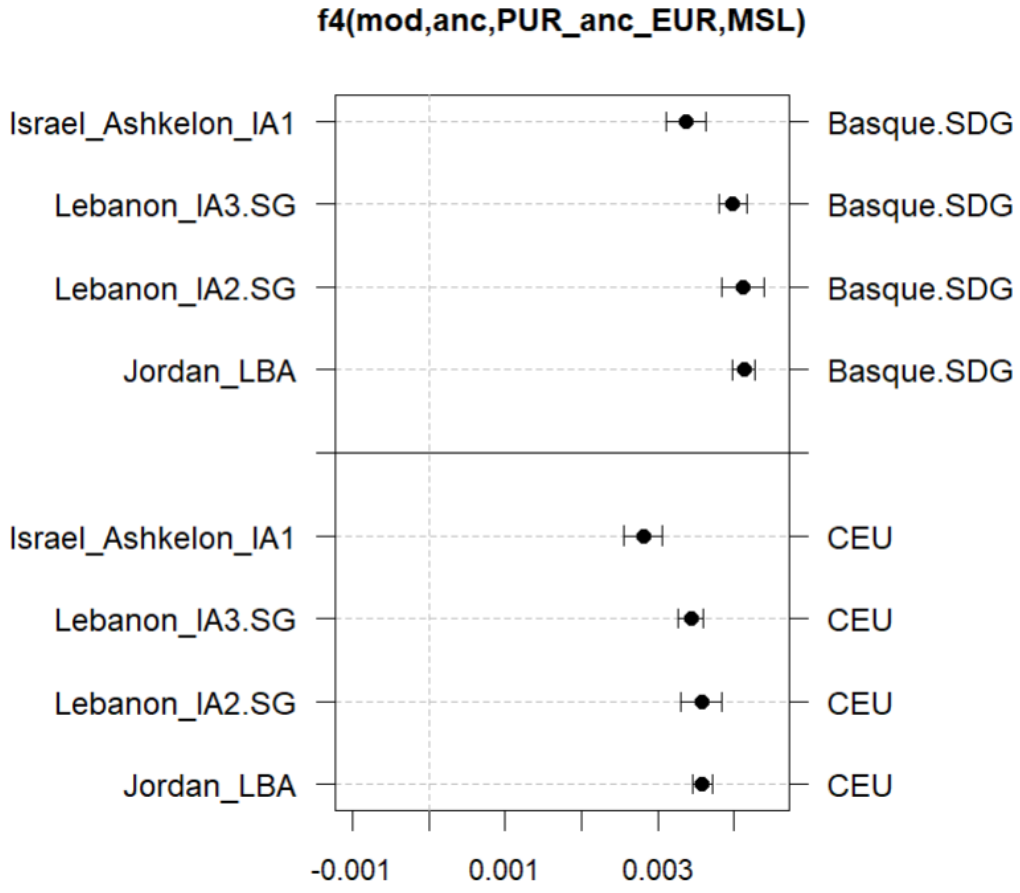


Figure 42: Results of f_4 (mod,anc,PUR_anc_EUR,MSL)

Figure 42 shows that PUR_anc_EUR are clearly more modern than ancient, but this was predictable since colonizers from Europe has left a huge trace in the population, while ancient contacts, if confirmed, would be something small in comparison.

For further confirmation, an analogous set of tests was performed substituting PUR_anc_EUR with PUR_mod_EUR in the form:

$$f_4(x\_anc, y\_mod, PUR\_mod\_EUR, MSL)$$

keeping the same populations for x and y. Detailed results are in table T7 in Additional Material and confirm the 'modernity' of these genomes.

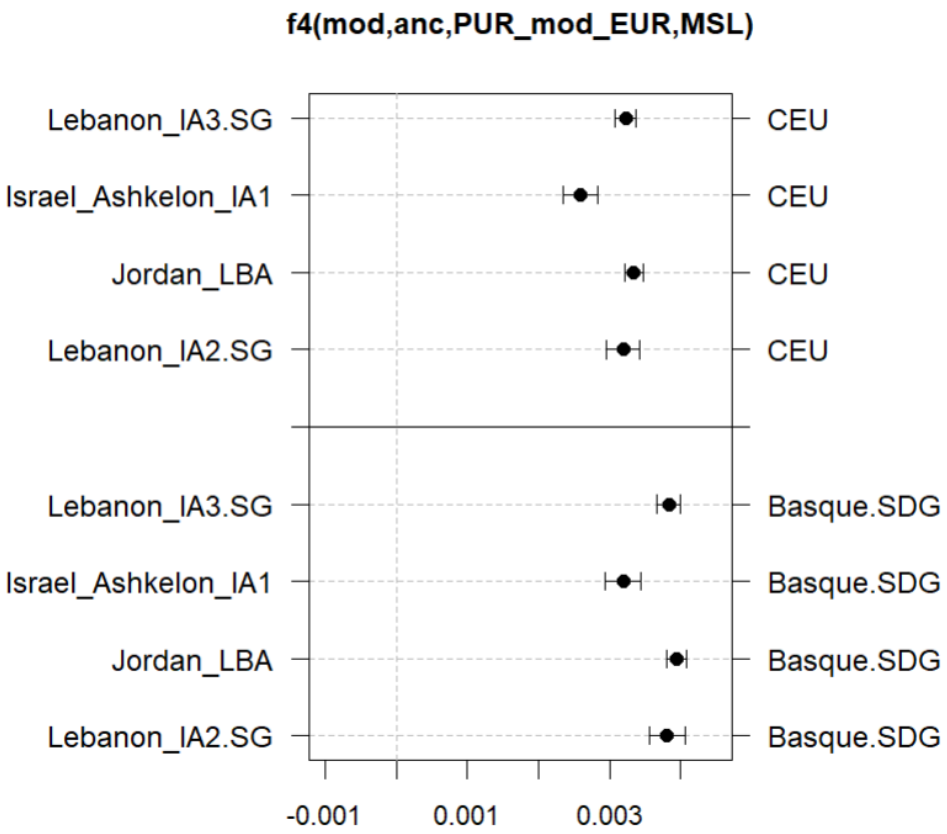


Figure 43: Results of f_4 (mod,anc,PUR_mod_EUR,MSL)

At this point, to see if PUR_anc_EUR is actually enriched with segments of ancient European genome, a set of f_4 tests of the type:

$$f_4(PUR\_anc\_EUR, PUR\_mod\_EUR, x, MSL)$$

was carried out, with x substituted by the different ancient European populations that were used throughout this work. The expectations are to obtain a significant positive value in case of enrichment or a non-significant value in case of absence of contacts.

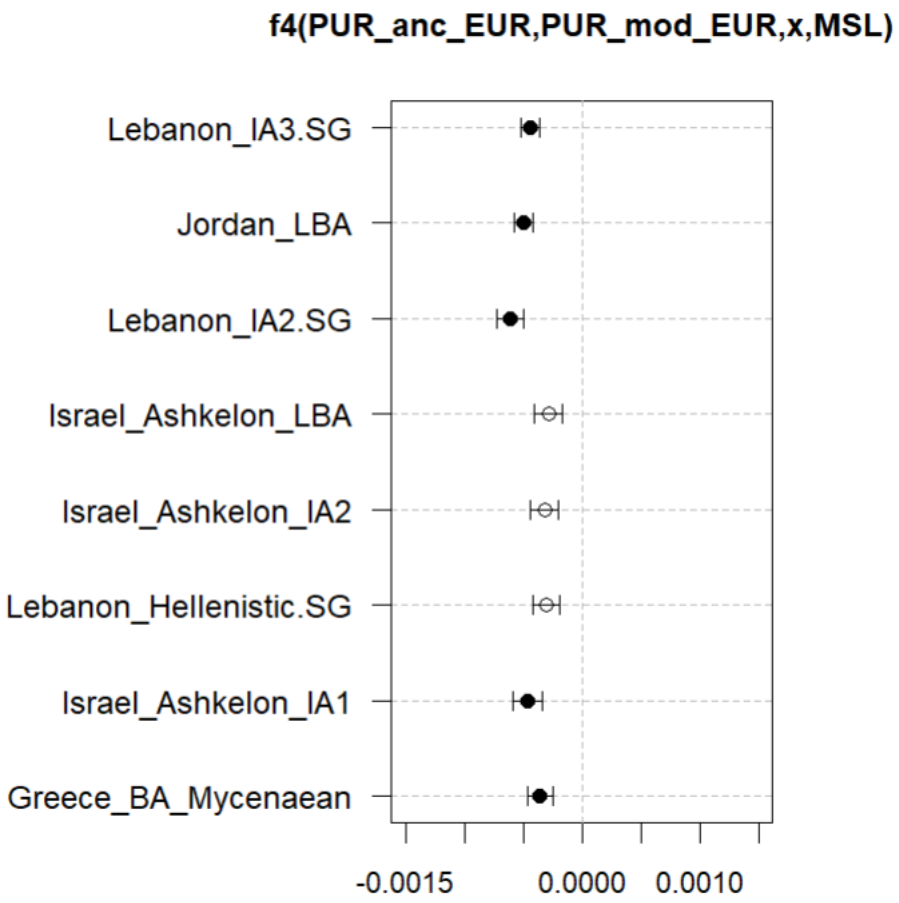


Figure 44:Results of f_4 (PUR_anc_EUR,PUR_mod_EUR,x,MSL)

The results shown in Figure 44, with detailed results in table T8 of Additional Material, are surprising considering the hypotheses, but before discussing them it may be useful to take a deeper look plot at the two European subgroups in the PCA plot.

To get a general idea of the structure of the two separated European components of PUR, PCs were calculated on a panel of European and West Asian modern populations, with PUR_anc_EUR and PUR_mod_EUR subsequently projected in the PCs space.

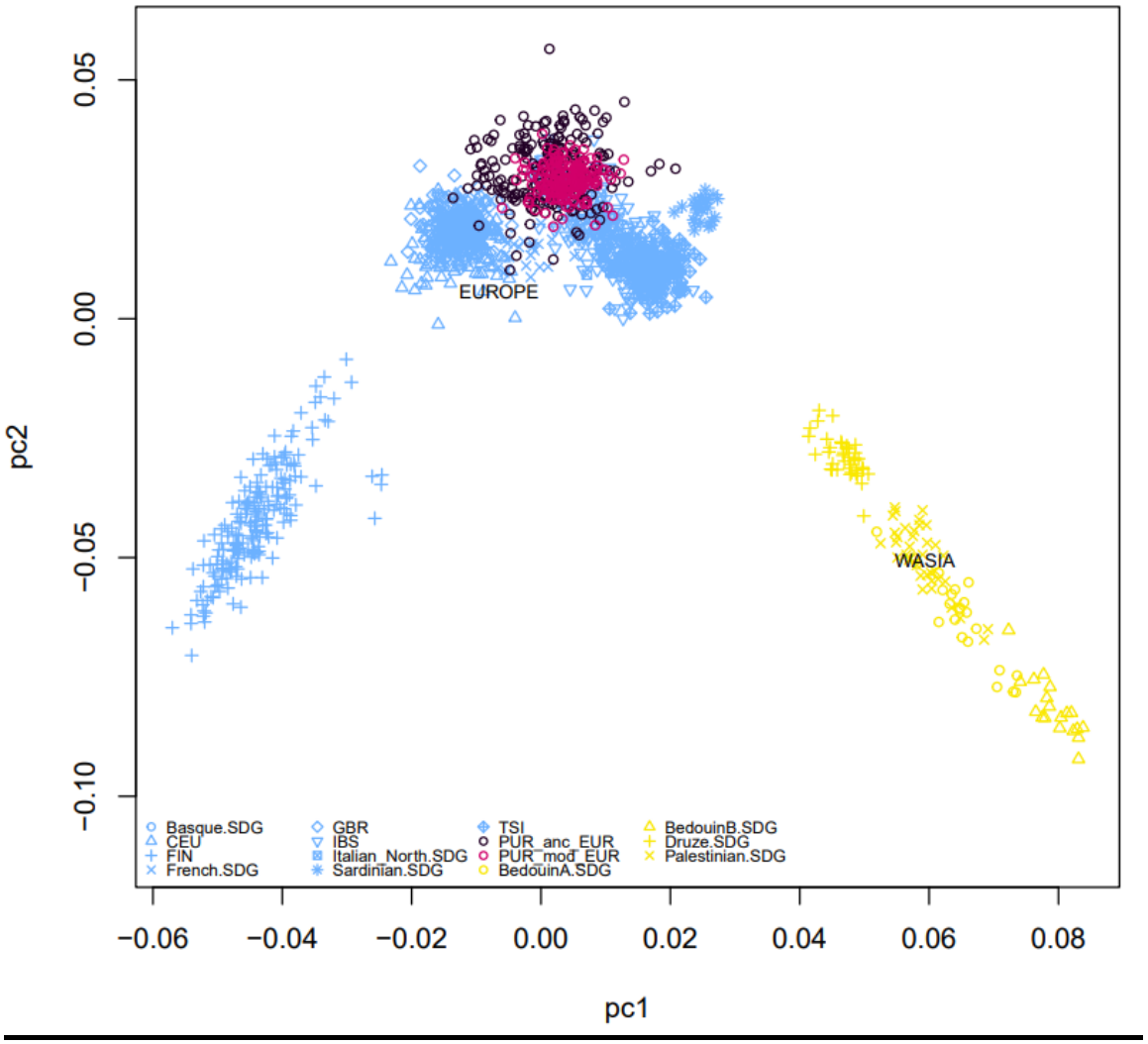


Figure 45: PCA plot with Europe, West Asia and the two subgroups

West Asian populations were used since the area of origin of some of the hypothetical Mediterranean populations responsible for earlier encounters overlaps with the one of these populations. From Figure 45 it emerges that pc1 separates populations on an ideal north-south axis, putting Finnish on one side and West Asian populations on the other one. The second component seems to distinguish a sort of East-West positioning. The result are three clusters, one for the Finnish, one for West Asians and one for continental Europe. Both European components of PUR locate at the vertex of the V-shaped configuration that emerges in the graph, in the middle of the continental European cluster. Neither of the two seems to be attracted by any of the other two clusters.

It can be interesting to project on the same PCs space the ancient Mediterranean populations used in this work. This projection is shown in Figure 46.

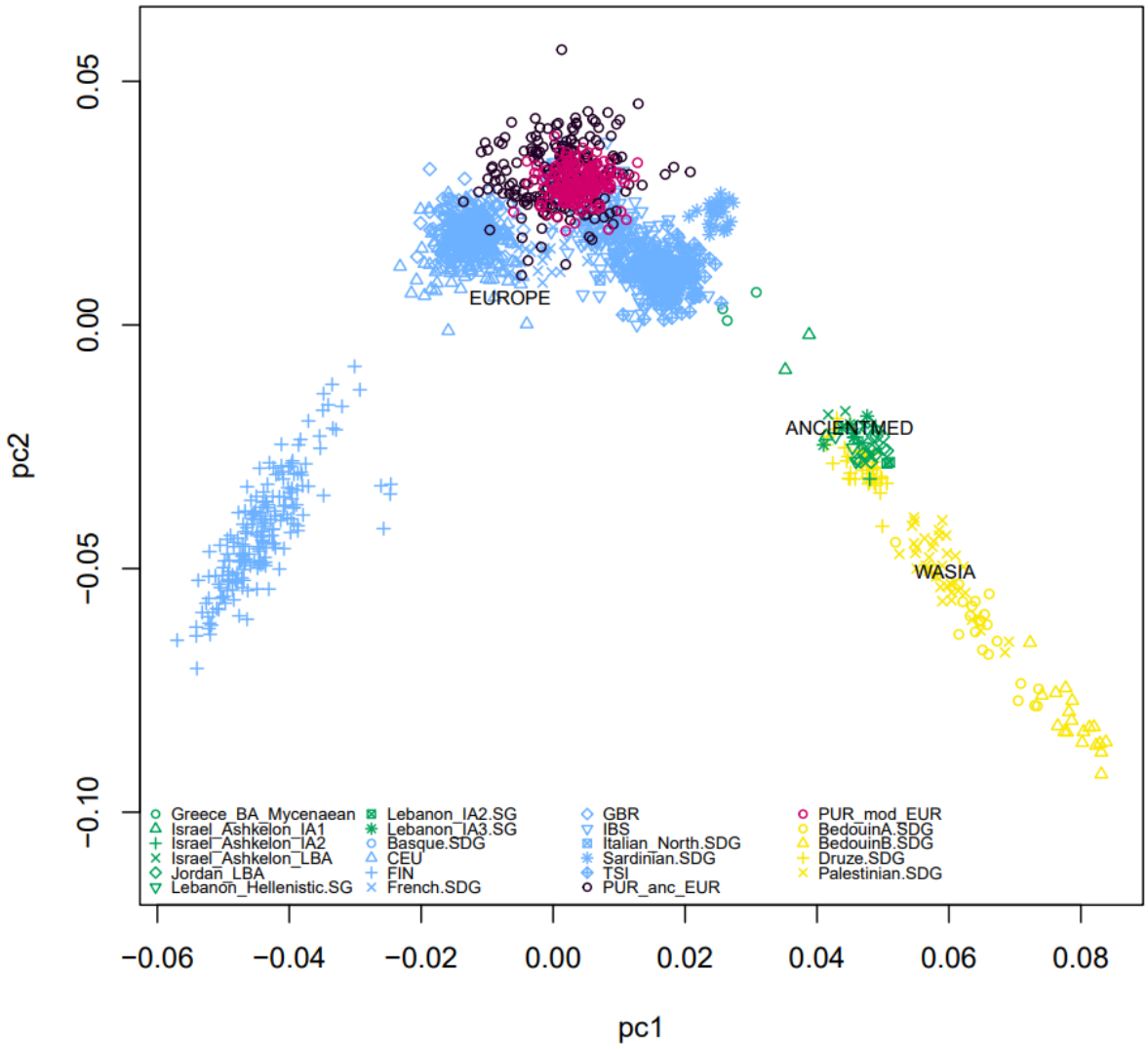


Figure 46: PCA plot with Europe, West Asia, ancient Mediterranean and the two subgroups

As it emerges from Figure 46, the ancient population of the Mediterranean area used in this work locate on the continental Europe – West Asia cline as it could be expected. This projection doesn't add much to our comprehension of the situation, since the position of the two European components of PUR seems to be neutral relative to these populations. To go into more detail, a PCA was made excluding FIN population and the four populations from West Asia, in order to clarify the positioning of the PUR components with respect to continental European populations.

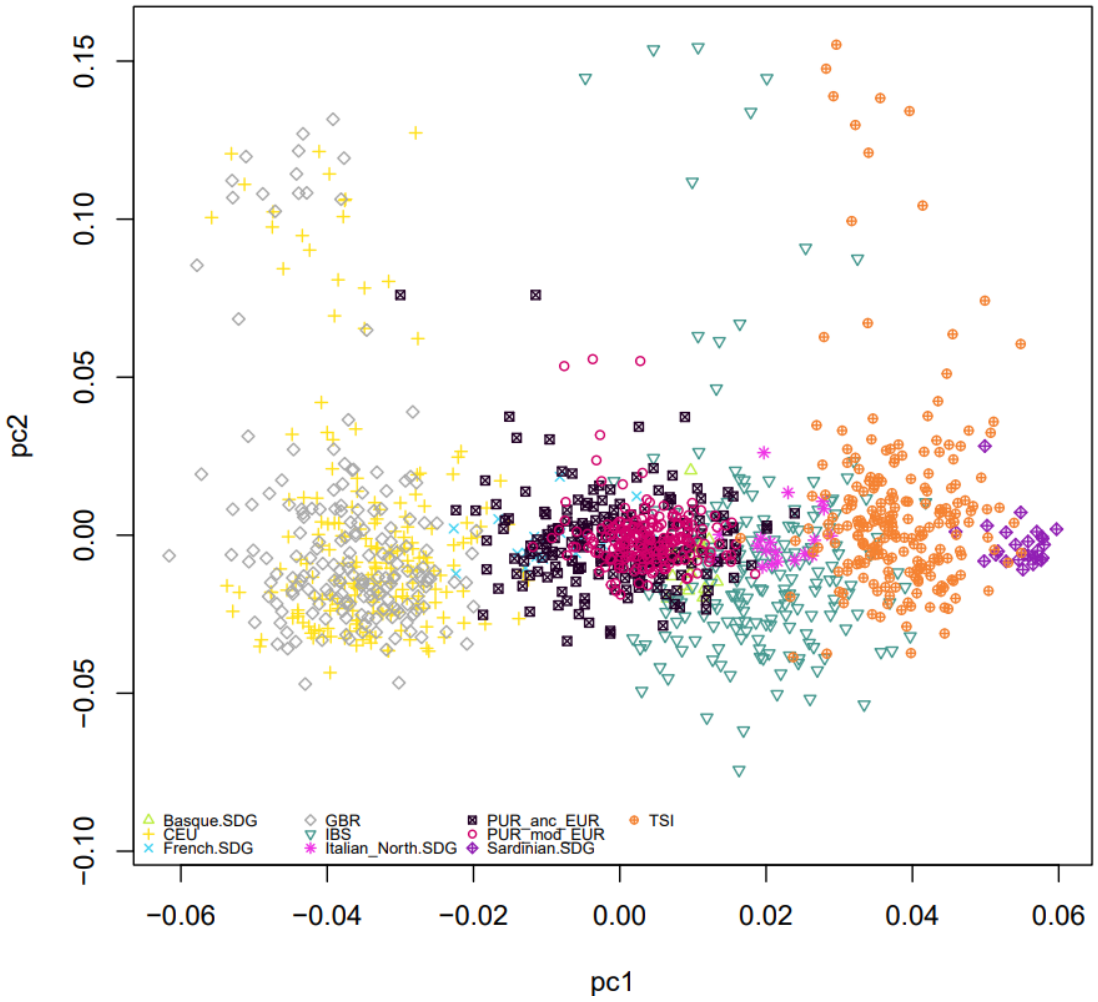


Figure 47: PCA plot of some European populations and the two subgroups

Figure 47 may seem confusing at a first glance, but it's actually intuitive and it is easier to understand it if a rotation is applied since pc1 distinguish the points on a north-south axis.

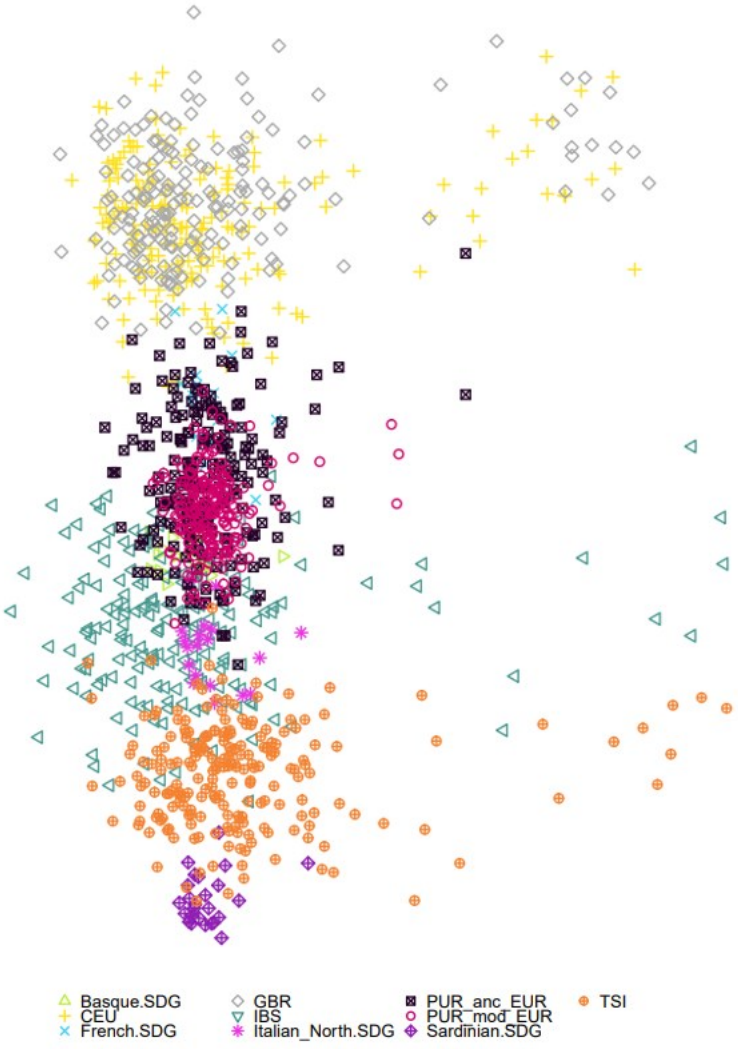


Figure 48: Figure 47 but rotated by 90°

By plotting the points in the this way it is possible to recognize northern populations like GBR at the top and southern populations like TSI and Sardinian at the bottom. The two PUR European components locate in the middle, in an area between IBS, Basque and French. According to this figure there is not a clear difference between PUR_anc_EUR and PUR_mod_EUR. The two groups mostly overlap, with the first one that is more scattered and exhibits a slight leaning towards the top. The scattering may be due to the fact that, on average, individuals of PUR_anc_EUR are missing information for 89% of the SNPs, while the ones of PUR_mod_EUR for 60% of them. This makes the projection less precise for the first group. The leaning towards the area occupied by northern populations, instead, may be due to the bias caused by the extraction made with GBR. Apart from these slight differences, they seem to belong to the same population and their positioning is coherent with the history of colonization.

With these considerations in mind, the surprising results of Figure 44 for which the modern subgroup appears significantly closer to ancient Mediterranean populations than the ancient subgroup may be interpreted in the following way. First of all, as it emerges from the PCA plot of Figure 46, the two subgroups appear quite distant and not attracted by the area in which the ancient Mediterranean populations are and also in the ADMIXTURE analysis both subgroups didn't show traces of the component that characterized these populations. At the same time, it is possible to see in Figure 48 that the modern subgroup is less scattered than the ancient one and this scattering seems to be leaning towards the northern European population. This characteristic, that may be responsible for the greater affinity of the modern subgroup with ancient Mediterraneans, may be due to an anomalous effect of the bias towards the population used as reference for the European component in PCAdmix (GBR). What is clear is that, according to these analyses, the ancient subgroup doesn't show any specific affinity with those populations that, according to the ideas illustrated in the introduction, may have had pre-Columbian contacts with Natives of the Antilles.

An alternative interpretation that arises is that the separation between 'ancient' and 'modern' European components may be actually referring to two different subgroups with respect to the ones that were considered in this work. Indeed, as it was said in the introduction section about the genetic history of the Caribbean area, two different moments of colonization can be distinguished and it may be possible that the two subgroups analyzed are the first colonizers (Spanish) and the second ones (English/French). To check this possibility, the following f statistic was tested:

$$f_4(PUR\_anc\_EUR, PUR\_mod\_EUR, IBS, MSL)$$

obtaining a value of -0.00032 and a Z-score of -4.90. This disproves what was just proposed, as the modern subgroup appears to be more similar to IBS than the ancient one, maybe again because of the same anomalous effect of the bias induced by GBR. On top of this, any reasoning on this hypothesis would be quite weak since the time that passed between the two colonization events is probably too small to create any concrete difference.

5 - Conclusions and Future Developments

The different analyses performed in this work, both regarding modern and ancient DNA, don't seem to support the theory proposed by Lucio Russo in his book, even though it is important to stress that the absence of genetic evidence of pre-Columbian contacts can only prove that there was no genetic exchange between the two populations. Given the nature and the 'size' of these hypothetical contacts, it may be reasonable to imagine a situation where the two populations met without admixing in any way or in such a small way that would not be detectable. This is just to say that, even though this works shows that no genetic traces of ancient Mediterraneans seem to be present in the Caribbean area, this cannot be used as a proof to completely refuse this event. On top of that, the analyses regarding modern DNA were carried out on Puerto Rican genomes, which are just a proxy for the area of the Lesser Antilles for which no modern DNA was available.

Of course there are several limitations that affected this work and that leave room for future developments and for more accurate results. For what regards the ancient DNA part, few samples of the Caribbean area dating back to the years of interests of this work are currently available, and just 16 of them come from the Lesser Antilles. The finding of more samples like these would help to investigate this hypothesis, considering that, even if contacts occurred, they probably left a small trace in the Antilles' genomes and a huge sample size would be necessary given the low probability of finding an admixed individual. Of course such finding would settle the debate, since a clear event of admixture in a pre-Columbian individual would leave no doubts about the correct scenario. The finding and sequencing of ancient DNA samples anyway, especially in an area like the Caribbean where the conditions are not favorable to the long-term preservation of DNA, is slow and costly and it is easier to work on modern DNA to try to infer information. On the side of modern DNA, in this work it was necessary to use the PUR population as a proxy for the Lesser Antilles populations because of lack of alternatives. As stated in the Data section, there are some studies that were carried out on modern Lesser Antilles populations, but the data are not publicly available and it was not possible to obtain permission to study them. The analysis of genomes from this area would clearly be more reliable, also because as stated at the beginning each island has its own particular genetic history. In the last years the costs of sequencing modern genomes are decreasing and they don't represent a difficulty as in the past, but it may still be complicated to obtain genomes suitable for this work as they would need to belong to local communities that are good representatives of the genetic history of that area and ethical issues may arise.

Regarding the analyses and the approach used in this work, a possible improvement could be to repeat all the analyses using a different population than GBR as source of European ancestry in PCAdmix when separating the different components of PUR. As it was discussed in the last section, there seems to be some influence given by the choice of GBR and maybe it could have been wiser to choose a population more easily distinguishable from continental European ones (maybe Finnish or Russians). On top of that, it may be useful to create a mathematical model for the intuition of the separation of the two subgroups based on the configuration of positions of the blocks of SNPs. This could help to understand how well the two subgroups have been divided and what to expect from the subsequent analyses. Finally, it may be useful to develop a simulation of a possible pre-Columbian admixture event and the consequent evolution of the genomes to have an idea of how those genomes would look like today, even if more information would be needed to generate trustful results.

6 - Additional Material

Table T1: Results of f_4(x, USA_Anzick_Realigned, French, MSL)

X	F4 VALUE	STANDARD ERROR	Z-SCORE	SNPS
BAHAMAS_ABACOISL_CERAMIC	2.96197571809e-05	0.000258816606289	0.114443032097	920667
BAHAMAS_CROOKEDISL_CERAMIC	0.000160824278176	0.000302118963826	0.532321030596	902260
BAHAMAS_ELEUTHERAISL_CERAMIC	-2.10456336818e-05	0.000270139698818	-0.0779064823641	973775
BAHAMAS_LONGISL_CERAMIC	0.000225562759707	0.000289418019519	0.779366675518	880990
BAHAMAS_SOUTHANDROS_CERAMIC	-4.71991813361e-05	0.000273796137789	-0.172388046512	931095
CUBA_CUEVAPERICO_ARCHAIC	4.05045678237e-05	0.000289997070069	0.139672334669	644525
BAHAMAS_TAINO.SG	-0.000375042104852	0.000307171704664	-1.22095264361	1080563
CUBA_CANIMARABAJO_ARCHAIC	-3.60284981451e-05	0.000258714242142	-0.139259817499	1073806
CUBA_CUEVACALERO_ARCHAIC	-5.98475065254e-05	0.000290942468224	-0.205702202538	941701
CUBA_GUAYABOBLANCO_ARCHAIC	-0.000101680969996	0.000291589615581	-0.348712589759	913840
CUBA_LASCAROLINAS_ARCHAIC	-0.000351850573876	0.000291374303463	-1.20755526378	871736
CUBA_PLAYADELMANGO_ARCHAIC	-0.000221575486509	0.000275097031689	-0.805444846672	917444
CUBA_MANUELITO_ARCHAIC	-0.000568993644818	0.000354052109613	-1.60709011292	510768
DOMINICAN_CUEVAROJA_ARCHAIC	-0.000762410496355	0.000429237754809	-1.77619626376	132550
CURACAO_DESAVAAN_CERAMIC	6.85966504767e-05	0.000275505468215	0.248984714972	883622
DOMINICAN_ANDRES_CERAMIC	4.53177942043e-05	0.000267173195 7	0.169619538688	949349
DOMINICAN_ATAJADIZO_CERAMIC	7.78867492274e-05	0.000256265527045	0.303929873539	994887
DOMINICAN_CUEVAJUANA_CERAMIC	7.24444302451e-05	0.000267419040491	0.270902289202	1001596
DOMINICAN_ELFRANCES_CERAMIC	-1.47132993099e-05	0.000296046909253	-0.049699216 0702	761233
DOMINICAN_EDILIOCRUZ_CERAMIC	-0.000137122345008	0.000336634195646	-0.407333380809	770077
DOMINICAN_LAUNION_CERAMIC	6.83781284813e-05	0.000272282558517	0.251129300582	970895
DOMINICAN_LACALETA_CERAMIC	-5.81074220957e-05	0.000243645262201	-0.238491902411	1065857
DOMINICAN_LOMAPERENAL_CERAMIC	0.000257694652674	0.000326527200808	0.789198118979	755477
DOMINICAN_LOSCORNIEL_CERAMIC	-3.15725324224e-05	0.000319567777912	-0.098797609 1602	750774
PUERTORICO_CABOROJO11_CERAMIC	0.000246879720671	0.000315437822 86	0.782657318746	750544
DOMINICAN_ELSOCO_CERAMIC	-3.37538244954e-05	0.000253805574458	-0.132990871329	1028408
DOMINICAN_JUANDOLIO_CERAMIC	0.000244466270924	0.000271877540848	0.899177880456	894100
GUADELOUPE_ANSEGOURDE_CERAMIC	4.3992013438e-05	0.000282368954828	0.155796211608	604223
DOMINICAN_MACAO_CERAMIC	0.000220308425013	0.000269182800119	0.818434257002	933936
PUERTORICO_LOSINDIOS_CERAMIC	0.000631806041037	0.000442806015974	1.426823525978	119327
PUERTORICO_PUNTACANDELERO_CERAMIC	0.000306422005655	0.000296290370044	1.034194954 12	559053

PUERTORICO_TIBES_CERAMIC	-3.95605197193e-05	0.000346001444561	-0.114336284837	344130
PUERTORICO_CANASCOLLORESMONSERRATE_CERAMIC	-7.19208314651e-05	0.000280271323846	-0.256611452353	912383
PUERTORICO_COLLORES_CERAMIC	0.000162925968136	0.000311082885311	0.523738128419	752545
PUERTORICO_MONSERRATE_CERAMIC	5.08915630033e-05	0.000341300270387	0.149110819472	705258
PUERTORICO_PASODELINDIO_CERAMIC	-0.000113721481581	0.000270154624241	-0.420949602104	1002798
PUERTORICO_SANTAELENA_CERAMIC	-9.07828296356e-05	0.000268317902981	-0.338340560931	930517
STLUCIA_LAVOUTTE_CERAMIC	3.86424298516e-05	0.000250342841445	0.154358038075	1051312
HAITI_DIALE1_CERAMIC	0.00016143675226	0.000286197955272	0.564073744365	979163
MXL	0.00208383272	0.000257171644968	8.10288677144	1081470

Table T2: Results of f_4(x_after, x_before, French, MSL)

X	F4 VALUE	STANDARD ERROR	Z-SCORE	SNPS
CUBA_CUEVAPERICO_ARCHAIC	0.000596222027695	0.000418490633036	1.42469623124	79323
CUBA_CANIMARABAJO_ARCHAIC	6.3304104261e-05	0.000153074662593	0.413550506588	963011

Table T3: Results of f_4(x,y, PDI003, MSL)

X	Y	F4 VALUE	STANDARD ERROR	Z-SCORE	SNPS
GREECE_BA_MYCENAEAN	FRENCH	-0.000948611476417	0.000562721514444	-1.68575654578	62195
ISRAEL_ASHKELON_IA1	FRENCH	-0.0014300377538	0.000657892246615	-2.17366561342	74421
LEBANON_IA3.SG	FRENCH	-0.0022294042059	0.000286319279676	-7.78642712578	163488
ISRAEL_ASHKELON_IA2	FRENCH	-0.00193816195554	0.000590150729357	-3.2841812424	103380
ISRAEL_ASHKELON_LBA	FRENCH	-0.00174514594307	0.000646861097728	-2.6978681346	86346
LEBANON_HELLENISTIC.SG	FRENCH	-0.00282145112512	0.000575884041692	-4.89933896559	110870
JORDAN_LBA	FRENCH	-0.00174599124094	0.000215953803607	-8.08502194348	163760
LEBANON_IA2.SG	FRENCH	-0.00269421716123	0.000539184621443	-4.99683606336	119613
ISRAEL_ASHKELON_LBA	IBS	-0.000749597826165	0.000625226167041	-1.198922671	86578

GREECE_BA_MYCENAEAN	IBS	0.000288954335166	0.000548051917272	0.527238982402	62301
JORDAN_LBA	IBS	-0.000686847042675	0.000198363357014	-3.46257016928	164187
ISRAEL_ASHKELON_IA1	IBS	-0.000497875571824	0.00064922871708	-0.766872380605	74620
LEBANON_IA2.SG	IBS	-0.00159256582824	0.000534836833359	-2.97766669927	119894
ISRAEL_ASHKELON_IA2	IBS	-0.000911970214402	0.000579242983073	-1.57441737069	103672
LEBANON_HELLENISTIC.SG	IBS	-0.00181611122455	0.000570974350302	-3.1807229582	111147
LEBANON_IA3.SG	IBS	-0.00117300301348	0.0002761826338	-4.24720047508	163908
ISRAEL_ASHKELON_IA1	GBR	-0.00211854795317	0.000651644076955	-3.25108142327	74620
GREECE_BA_MYCENAEAN	GBR	-0.00136215481262	0.000557742350461	-2.44226534259	62301
JORDAN_LBA	GBR	-0.0023811233386	0.000210686910982	-11.3017146034	164187
LEBANON_IA3.SG	GBR	-0.00286566675991	0.000287289663777	-9.97483418733	163908
ISRAEL_ASHKELON_IA2	GBR	-0.00261481298343	0.000587986670518	-4.44706166744	103672
ISRAEL_ASHKELON_LBA	GBR	-0.00250020782288	0.00063957516537	-3.90916964613	86578
LEBANON_HELLENISTIC.SG	GBR	-0.00350415629585	0.0005719018331	-6.12719892303	111147
LEBANON_IA2.SG	GBR	-0.00326254011988	0.000540790317834	-6.03291148582	119894

Table T4: Results of f_3(x, PUR_anc_EUR, MSL)

X	F3 VALUE	STANDARD ERROR	Z-SCORE	SNPS
ISRAEL_ASHKELON_IA2	0.176854419359	0.00170321075038	103.835898946	491971
LEBANON_HELLENISTIC.SG	0.176886072337	0.00182834862173	96.7463591103	714282
GREECE_BA_MYCENAEAN	0.177849924723	0.00184348868176	96.4746496589	404321
ISRAEL_ASHKELON_LBA	0.178522652869	0.00168844675708	105.731881755	394123
LEBANON_IA2.SG	0.178685082358	0.00174309451754	102.510265829	764209
JORDAN_LBA	0.179239482491	0.00151819552691	118.06086852	105599 5

LEBANON_IA3.SG	0.180004898883	0.00152118164818	118.332284049	1080630
ISRAEL_ASHKELON_IA1	0.180871983316	0.001712650467	105.609397131	293822

Table T5: Results of f_3(x, PUR_mod_EUR, MSL)

	F3 VALUE	STANDARD ERROR	Z-SCORE	SNPS
FIN	0.189923148052	0.00155241112934	122.340753981	1095818
IBS	0.190611121245	0.00150851572972	126.356734298	1095818
TSI	0.190977221692	0.00152298423408	125.396716143	1095818
ITALIAN_NORTH.SDG	0.192090341701	0.00155289240555	123.698423029	1092272
FRENCH.SDG	0.193135603423	0.00155339412494	124.331359519	1092267
CEU	0.193518647624	0.00154779861175	125.028311923	1095818
GBR	0.194476748454	0.00156108370641	124.578040021	1095818
BASQUE.SDG	0.195696637705	0.00154497511174	126.666530883	1092263

Table T6: Results of f_4(mod,anc, PUR_anc_EUR, MSL)

mod	anc	F4 VALUE	STANDARD ERROR	Z-SCORE	SNPS
CEU	JORDAN_LBA	0.003586592276 1	0.000132917516501	26.9835938145	1055995
CEU	LEBANON_IA2.SG	0.003573208691 8	0.000263569154868	13.5570062953	764209
CEU	LEBANON_IA3.SG	0.003441175474 87	0.00016455377230	20.9121640104	1080630
CEU	ISRAEL_ASHKELON_IA1	0.002809798974 02	0.000253709257786	11.0748775923	293822
BASQUE.SDG	JORDAN_LBA	0.004127640339 67	0.000153676970020	26.8591991313	1053464
BASQUE.SDG	LEBANON_IA2.SG	0.004116583010 89	0.000276040074161	14.9129905265	762519
BASQUE.SDG	LEBANON_IA3.SG	0.003984812222 74	0.000176590461428	22.5652744238	1078093

BASQUE.SDG	ISRAEL_ASHKELON_IA1	0.00336768929964	0.000261996828822	12.8539315334	293016

Table T7: Results of f_4(mod,anc, PUR_mod_EUR, MSL)

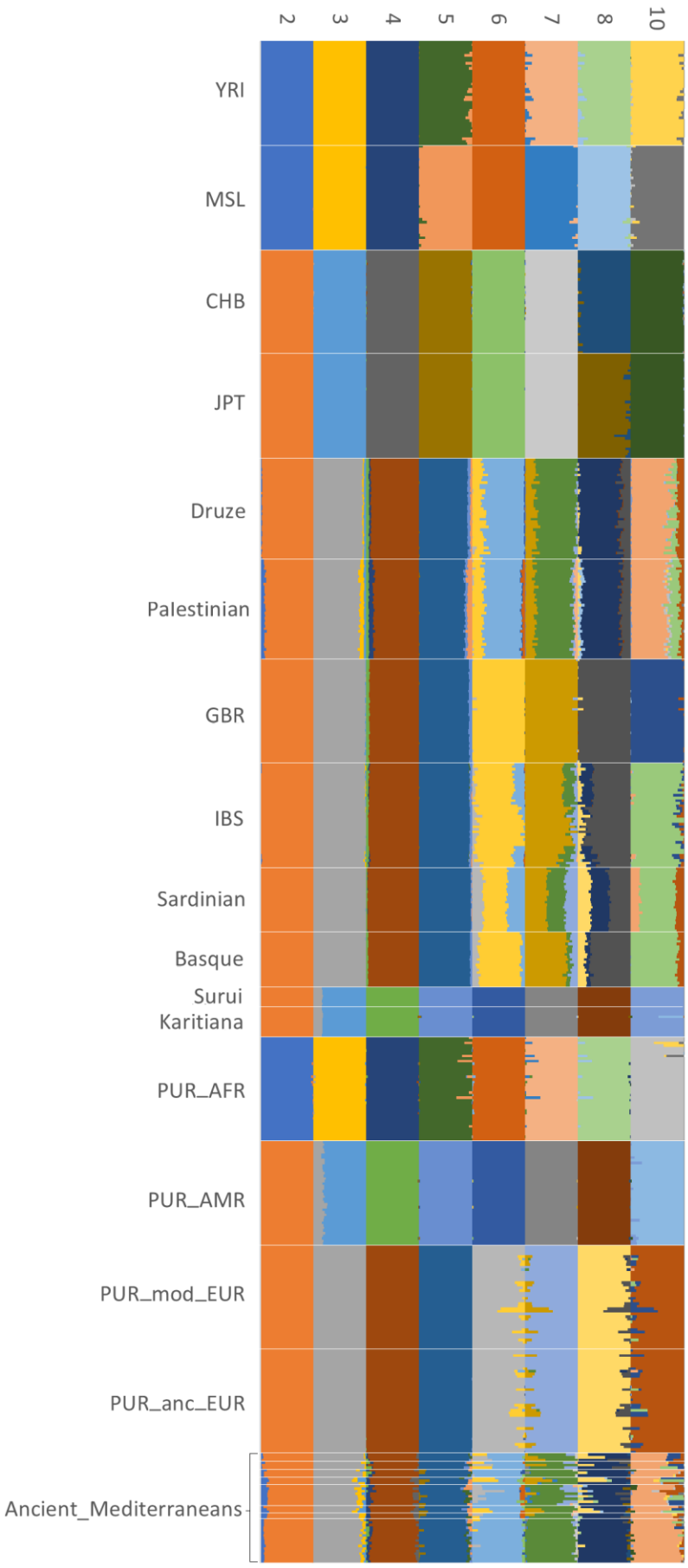
X	Y	F4 VALUE	STANDARD ERROR	Z-SCORE	SNPS
"BASQUE.SDG"	"LEBANON_IA2.SG"	0.00381018048532	0.00025043206268	15.2144275959	768614
"BASQUE.SDG"	"JORDAN_LBA"	0.00394482658845	0.00014299509678 2	27.5871458338	1061874
"BASQUE.SDG"	"ISRAEL_ASHKELON_IA1"	0.00319190208292	0.00025076928 99	12.7284408876	295945
"BASQUE.SDG"	"LEBANON_IA3.SG"	0.00383185864038	0.00016488457109	23.2396434369	1086615
"CEU"	"LEBANON_IA2.SG"	0.00319613688 94	0.00023602655162 8	13.5414294 17	770324
"CEU"	"JORDAN_LBA"	0.00334554876448	0.00012206175717 8	27.4086564196	1064442
"CEU"	"ISRAEL_ASHKELON_IA1"	0.00259326221191	0.00024170448474 2	10.7290612115	296769
"CEU"	"LEBANON_IA3.SG"	0.00322884041122	0.00015147500934 1	21.3159941384	1089188

Table T8: Results of f_4(PUR_anc_EUR, PUR_mod_EUR, x, MSL)

X	F4 VALUE	STANDARD ERROR	Z-SCORE	SNPS
GREECE_BA_MYCENAEAN	-0.00035708441427 4	0.00011000079081 3	-3.24619861035	404321
ISRAEL_ASHKELON_IA1	-0.00046139796102 9	0.00012246580035 8	-3.76756579945	293822
LEBANON_HELLENISTIC.SG	-0.00030289472692 7	0.00011451946026 6	-2.64491926719	714282
ISRAEL_ASHKELON_IA2	-0.00031854739094 2	0.00011986053121	-2.65765041858	491971

ISRAEL_ASHKELON_LBA	-0.000284467735528	0.000116414684879	-2.44357261134	394123
LEBANON_IA2.SG	-0.000611765275351	0.000109420149348	-5.59097459652	764209
JORDAN_LBA	-0.000499219695664	7.67527566814e-05	-6.50425753092	1055995
LEBANON_IA3.SG	-0.000442264806628	8.24076031998e-05	-5.36679613841	1080630

Figure A1: Admixture results for different values of K



7-Bibliography

Agranat-Tamir L, W. S. (2020). The Genomic History of the Bronze Age Southern Levant. *Cell*, 1146-1157.

Alexander DH, N. J. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 1655-64.

Allen Ancient DNA Resource. (2021). Obtido de https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data.

Benn Torres J, M. V. (2019). Analysis of biogeographic ancestry reveals complex genetic histories for indigenous communities of St. Vincent and Trinidad. *Am J Phys Anthropol.*, 482-497.

Bergström A, M. S. (20 de Mar de 2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science, 367*(6484).

Brisbin A, B. K. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol*, 343-64.

Bryc K, V. C. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *PNAS*.

Chakraborty R, S. P. (1988). Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. *PNAS*.

Consortium, T. 1. (2013). *http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/vcf_with_sample_level_annotation/*.

Consortium., T. 1. (2015). A global reference for human genetic variation. *Nature, 526*, 68-74.

Feldman M, M. D. (2019). Ancient DNA sheds light on the genetic origins of early Iron Age Philistines. *ScienceAdvances*.

Fernandes, D. S. (2021). A genetic history of the pre-contact Caribbean. *Nature, 590*, 103-110.

Fregel R, M. F. (2018). Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc Natl Acad Sci U S A*, 6774-6779.

Germanus, N. (s.d.).

Haber M, N. J. (2020). A Genetic History of the Near East from an aDNA Time Course Sampling Eight Points in the Past 4,000 Years. *Am J Hum Genet*, 149-157.

Harvard, U. (2010). *https://github.com/chrchang/eigensoft/tree/master/EIGENSTRAT*.

Hellenthal G, A. A. (2008). Inferring human colonization history using a copying model. *PLoS Genetics*.

Jolliffe IT, C. J. (2015). Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci.*

Keith MH, F. M. (2021). Genetic ancestry, admixture, and population structure in rural Dominica. *PLoS One*.

Lazaridis I, e. a. (2017). Genetic origins of the Minoans and Mycenaeans. *Nature, 548*(7666), 214-218.

Lipson, M. (2020). Applying f4-statistics and admixture graphs: Theory and examples. *Molecular Ecology Resources*.

Loh PR, L. M. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*.

López-Herráez D, B. M. (2009). Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One*.

Martin AR, e. a. (2017). Population genetic history and polygenic risk biases in 1000 Genomes populations. *Am J Hum Genet.*

Moorjani P, e. a. (2011). The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLOS Genetics*.

N Patterson, e. a. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*.

Nägele K, P. C. (24 de Lug de 2020). Genomic insights into the early peopling of the Caribbean. *Science, 369*(6502), 456-460.

NCBI. (2021). *rs1426654*. Obtido de https://www.ncbi.nlm.nih.gov/snp/rs1426654#frequency_tab.

NCBI. (2021). *rs16891982*. Obtido de https://www.ncbi.nlm.nih.gov/snp/rs16891982#frequency_tab.

Nieves-Colón, M. (2022). Anthropological genetic insights on Caribbean population history. *Evol Anthropol.*, 118-137.

Pagani L, e. a. (2015). Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am J Hum Genet.*

Patterson N., e. a. (2012). Ancient Admixture in Human History. *Genetics*.

Pool JE, N. R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*.

Pritchard JK, S. M. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 945-959.

Purcell S., C. C. (2009). *www.cog-genomics.org/plink/1.9/*.

Rasmussen M, e. a. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*.

Reich D, e. a. (2009). Reconstructing Indian population history. *Nature*.

Reich, D. (2018). *Who we are and how we got here.* Oxford.

Rodríguez-Varela R, G. T. (2017). Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. *Curr Biol.*, 3396-3402.

Russo, L. (2013). *L'America dimenticata.* Mondadori.

Rustagi N, e. a. (2017). Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. *BMC Genomics*.

Saylor Academy. (2013). *https://saylordotorg.github.io/text_world-regional-geography-people-places-and-globalization/s08-04-the-caribbean.html*.

Schroeder H, e. a. (2020). Origins and genetic legacies of the Caribbean Taino. *PNAS*.

Seyr H, M. M. (2019). Decision Support Models for Operations and Maintenance for Offshore Wind Farms: A Review. *Applied Sciences*.

Shlens, J. (2014). A Tutorial on Principal Component Analysis. *Int. J. Remote Sens.*, 1-12.

Slatkin M, R. F. (2016). Ancient DNA and human history. *Proc Natl Acad Sci U S A*, 6380-7.

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*.

The 1000 Genomes Project Consortium. (2013). *http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/vcf_with_sample_level_annotation/*.

Yip, M. (2020). *https://towardsdatascience.com/all-you-need-to-know-about-pca-part-1-29590dd9fb65*.

Zubiri, N. (2009). Obtido de https://euskalkazeta.com/basque-footprints-in-the-caribbean/.