There and Back Again

---

the genomic wanderings of human populations

Serena Aneli

University of Turin
PhD Program in Biomedical Sciences and Oncology
Human Genetics Curriculum
Academic year 2015/2019

UNIVERSITÀ DEGLI STUDI DI TORINO

Department of Medical Sciences
PhD Program in Biomedical Sciences and Oncology
XXXI cycle
Human Genetics Curriculum
Academic years 2015/2019

THERE AND BACK AGAIN

the genomic wanderings of human populations

Serena Aneli

Tutor: Prof. Giuseppe Matullo
PhD coordinator: Prof. Emilio Hirsch

Image credit: Peter Strain (www.peterstrain.co.uk).
Used by permission from the artist.

Contents

Ancient tales: Europeans wanderings

Modern tales: the genetic burden of our travels

Future tales: back home

Conclusion

Appendices

Abstract

The past lives of our ancestors are written in our genes. Such old stories of deaths and births, migrations and diseases left marks on their descendants' DNA. Recently, the molecular analysis of modern, ancient and archaic human DNA, in combination with bioinformatics analysis, enabled us to tell some of those stories, thus retracing our evolutionary history as modern humans. DNA whispers that modern humans evolved in Africa 300,000 years ago; that, sometime around 100,000 years ago, groups of men spread from Africa into Eurasia and that they encountered the Neanderthals. It also tells that the bodies of these different kinds of humans touched intimately and so did their genes. It tells that their wandering adventures did not end with that meeting: they travelled longer and further crossing deserts, overcoming mountains and reaching every corner of the world. The possible paths grew exponentially as they went further; thus, they separated into different groups and settled in their new home. However, humans have never been quiet: migratory flows went on for millennia, different populations met, touched and admixed, as well as their genes did. We still have those genes.

The reconstruction of the winding trails followed by our ancestors is possible thanks to the fact that the people who move leave prints. Time passes by, and some of those prints are masked in dust or covered by traces of the people who come after them, but many others are left untouched on the ground, in a cave and in our genome.

Throughout this thesis, I will tell you four different migration tales reconstructed by the genetic marks we found in the DNA of modern-day Italian and other European populations.

The first part, "archaic tales", deals with the episodes of migrations and admixture from about 100,000-50,000 years ago, between our modern human ancestors and Neanderthals, and the Neanderthal genetic legacy derived from these events. Here, we discovered that different European and even Italian modern populations show variable levels of Neanderthal introgression. Moreover, while many genetic variants whose frequency varies markedly among us are examples of adaptive variation, others seem to be involved in complex diseases susceptibility.

In the second part, "ancient tales", I describe the analyses we did on modern-day European populations in order to dig out traces of ancient migrations and admixtures from their genomes. The patterns of migrations which contributed to the peopling of Europe have been extensively studied, however there looks to be something missing from the current model proposed by those studies — a combination of early European foragers,

Neolithic farmers and Bronze Age nomadic pastoralists. In the genomes of modern-day Italians, we found some traces of one of those missing pieces: a fourth contribution coming from the East and related to an Anatolian Bronze Age sample, other southeastern European ancestries and, ultimately, to the Caucasus region.

The third part, "modern tales", is not really about the reconstruction of migration trails, but instead is about the genetic burden we are bearing in our travels. In particular, we analyse a large dataset of Italian whole-exome sequences and we explored the distribution of both common and rare variation. When we focused on rare variants to stratify Italy by effective population size, we discovered that they could be responsible for distinctive differences even at the regional level. Then, we explore the burden of putatively pathogenic variation in the healthy Italian exome, showing that almost 3% of the population carries protein-truncating variants in a list of medically actionable genes.

The last part, "future tales", talks about the most recent migratory flows: the European refugees crisis. In the past 20 years at least 33,000 men, women and children died in the Mediterranean Sea, thus becoming in the majority of the cases "faceless bodies". The main theme in this part is the return back home: our work aims to find the minimum number of genetic markers useful to understand where an individual comes, thus facilitating the family reunification.

In the last decade, thanks to the technical advances in molecular biology, we have witnessed an ever-increasing amount of genetic data becoming available to scientists. At the same time, due to the evolution of the bioinformatics approaches, we have been granted the opportunity to explore such huge amount of genetic data, digging for answers in our and our relatives' DNAs. There are ancestors in our genome, whispering their stories to the ones who are willing to listen.

Premise and acknowledgements

As well as my ancestors, I did a journey too. My PhD journey began in Turin, where I started to work on human genetic variability in the group of *"Genomic variation and translational research"*, directed by Prof. Giuseppe Matullo, whom I thank for having allowed me to completely dedicate my PhD years to the study of the human past.

I arrived to the first of my academic crossroads in the first year. At the beginning of that year, Prof Matullo started a collaboration with Prof Cristian Capelli, head of the "Human evolutionary genetics" group at the Department of Zoology, in Oxford. I spent eight months (from October 2016 to May 2017) in the group of Prof Capelli to generate a genome-wide geo-historical map of the Italian Peninsula through the analysis of single nucleotide polymorphisms (SNPs). There, I worked closely with Dr Alessandro Raveane and Dr Francesco Montinaro, my fellow travellers on this path. Our project was divided mainly into three parts: while Alessandro did the first one — a study on the genetic variability of modern-day Italians — I worked on the second and third — an inference of the ancient and archaic events which have shaped such variability. Francesco and Cristian were involved in all phases, patiently guiding us to results that made sense! We really shared the joys and sorrows of our journey towards publication: from awesome results about ancient populations to journal rejection, passing by the "Oh my God, I deleted the code!" moments. Without them and their work, our publication would not exist, my thesis would have been surely less interesting — as well as, unfortunately for the readers, shorter — and I would not have found three great friends.

Here, I will talk mainly about the two parts I was involved in, the analyses of the Neanderthal legacy in modern-day populations and the reconstruction of the ancestry profile of Italian and European populations (chapters "*The genetic legacy of Neanderthals in Italy and Europe*" and "*Population structure of modern-day Italians reveals patterns of ancient ancestries in Southern Europe*", respectively). Conversely, I will not explore the results Alessandro obtained on modern genetic structure. For that part, I refer you directly to the publication (Raveane *et al.,* 2019).

At the same time, my personal and working life reached another crossroad: at the beginning of 2016, Prof Matullo hired a new bioinformatician, Dr Giovanni Birolo. In the autumn of this same year, we fell in love and I left for Oxford — the timing was not so great, but if we could have made it through the distance, we would have gotten through anything (actually, we were right). I do not like to talk about my personal life, but, in this case, it

is necessary, because Giovanni and I worked together on some parts of this thesis. While I was in Oxford, he began a project aiming at the analysis of the Italian genetic variability from exome data, focusing mainly on the exploration of pathogenic genetic variation carried by healthy individuals. When I came back to Turin, I joined him in this work: while he was mainly on the technical part of the sequencing analyses and pathogenicity prediction, I used some of the competences acquired during my Oxfordian months to explore the genetic structure of Italians using coding variations. I will talk about this work in the chapter "*Protein-coding genetic variation within the Italian peninsula: genetic structure and functional implications*".

The last main project I devoted my time to during these four years is still ongoing and, again, I am writing codes and analysing results with Giovanni. The main aim of this work is to find the minimum number of genetic markers — ancestry informative markers (AIM) — to infer the area of origin of an individual. This is particularly useful in the context of mass disasters, such as the numerous shipwrecks in the Mediterranean Sea: the corpses rescued between the waves in the context of the present European refugees crisis are, mostly, "faceless bodies". If we could develop a strategy to restrict their area of origin, we could facilitate the returning of the remains back home. I will show you some preliminary results of this part in the chapter "*Novel strategies for AIMs selection*".

During my PhD journey, I worked to retrace the steps of our ancestors in order to understand how and to what extent human migrations have been shaping our genetic variability. In doing so, I realised that a huge portion of our past — and our human nature — involved migrations, so much that we "are" the travels of our ancestors. Thus, I decided to focus my PhD thesis on the reconstruction of those migrations organizing the writing in four "tales". In the *archaic* and *ancient* tales, I explore the episodes of migrations and admixtures that took place since the first modern humans were peeping out of Africa. In the *modern* tales, I talk about another important aspect of migration: the "burden" — a genetic burden — every human being brings through its life, in terms of genetic variants predisposing to or directly causing some clinical conditions or diseases. Finally, I have called the last part about the ancestry classification problem "*future* tales" because, even if the midst of the migrants crisis is right now, the solution to the family reunification problem is still far into the future. In fact, Italy, together with the other European countries, have not yet established standardised strategies for family reunification. Sadly, we are also still far from having acknowledged the importance and the need for such dispositions.

The works I will present here would not exist without the help and

support of the people working with me. For this reason, I sincerely thank my colleagues, collaborators and friends, who have allowed me to reach the end of this PhD journey with ideas, advice and samples. Among them, I especially thank Carlo Robino and Daniela Lacerenza, Alessandro Raveane, Francesco Montinaro and Cristian Capelli and all the people from my group in Turin. Finally, I am profoundly grateful to Giovanni, the best travelling companion in this and in many other journeys to come.

Declaration

I, Serena Aneli confirm that the work presented in this thesis is my own. Where external contributions have been provided, I confirm that this has been indicated in the thesis.
Publications arising from this thesis:

- (2019) A. Raveane*, S. Aneli*, F. Montinaro*, G. Athanasiadis, S. Barlera, G. Birolo, G. Boncoraglio, AM. Di Blasio, C. Di Gaetano, L. Pagani, S. Parolo, P. Paschou, A. Piazza, G. Stamatoyannopoulos, A. Angius, N. Brucato, F. Cucca, G. Hellenthal, A. Mulas, M. Peyret-Guzzon, M. Zoledziewska, A. Baali, C. Bycroft, M. Cherkaoui, C. Dina, JM. Dugoujon, P. Galan, J. Giemza, T. Kivisild, M. Melhaoui, M. Metspalu, S. Myers, LM. Pereira, FX. Ricaut, F. Brisighelli, I. Cardinali, V. Grugni, H. Lancioni, V. L. Pascali, A. Torroni, O. Semino, G. Matullo, A. Achilli, A. Olivieri, C. Capelli. ***Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe.*** Science Advances (2019 Impact Factor: 12.804).

Publications not directly related to this thesis:

- (2019) Margaret L. Antonio, Ziyue Gao, Hannah M. Moots, Michaela Lucci, Francesca Candilio, Susanna Sawyer, Victoria Oberreiter, Diego Calderon, Katharina Devitofranceschi, Rachael C. Aikens, Serena Aneli, Fulvio Bartoli, Alessandro Bedini, Daniel J. Cotter, Daniel M. Fernandes, Gabriella Gasperetti, Renata Grifoni, Alessandro Guidi, Francesco La Pastina, Ersilia Loreti, Daniele Manacorda, Giuseppe Matullo, Simona Morretta, Alessia Nava, Vincenzo Fiocchi Nicolai, Federico Nomi, Carlo Pavolini, Massimo Pentiricci, Philippe Pergola, Marina Piranomonte, Ryan Schmidt, Giandomenico Spinola, Alessandra Sperduti, Mauro Rubini, Luca Bondioli, Alfredo Coppa, Ron Pinhasi, Jonathan K. Pritchard. ***Ancient Rome: a genetic crossroads of Europe and the Mediterranean.*** Science (2018 Impact Factor: 41.04). PMID: 31699931

- (2019) H. Kumar, K. Haddish, D. Lacerenza, S. Aneli, C. Di Gaetano, G. Tewelemedhin, R. Manukonda, N. Futwi, V. A. Iglesias, M. de la Puente Vila, M. Fondevila, M. V. Lareu, C. Phillips, C. Robino. ***Characterization of ancestry informative markers in the Tigray population of Ethiopia: a contribution to the identification***

process of dead migrants in the Mediterranean Sea. Forensic Science International Genetics (2017 Impact Factor: 5.637). PMID: 31812100

- (2018) Robino C, Lacerenza D, Aneli S, Di Gaetano C, Matullo G, Robledo R, Calò C. ***Allele and haplotype diversity of 12 X-STRs in Sardinia.*** Forensic Science International Genetics (2017 Impact Factor: 5.637). PMID: 29221994

- (2017) Lacerenza D, Aneli S, Di Gaetano C, Critelli R, Piazza A, Matullo G, Culigioni C, Robledo R, Robino C, Calo' C. ***Investigation of extended Y chromosome STR haplotypes in Sardinia.*** Forensic Science International Genetics (2017 Impact Factor: 5.637). PMID: 28057510

Manuscript under revision:

- G. Birolo* , S. Aneli*, C. Di Gaetano, G. Cugliari, A. Russo, A. Allione, E. Casalone, E. M. Paraboschi, D. Ardissino, S. Duga, R. Asselta, G. Matullo. ***Protein-coding variation within 1686 Italian exomes: genetic structure and clinical implications.*** Under revision in Molecular Biology and Evolution (2018 Impact Factor: 14.797).

The images at pages 1 and 307 have been realized by Erika Rae Heins (© Erika Rae Heins) and are used here by permission from the artist.

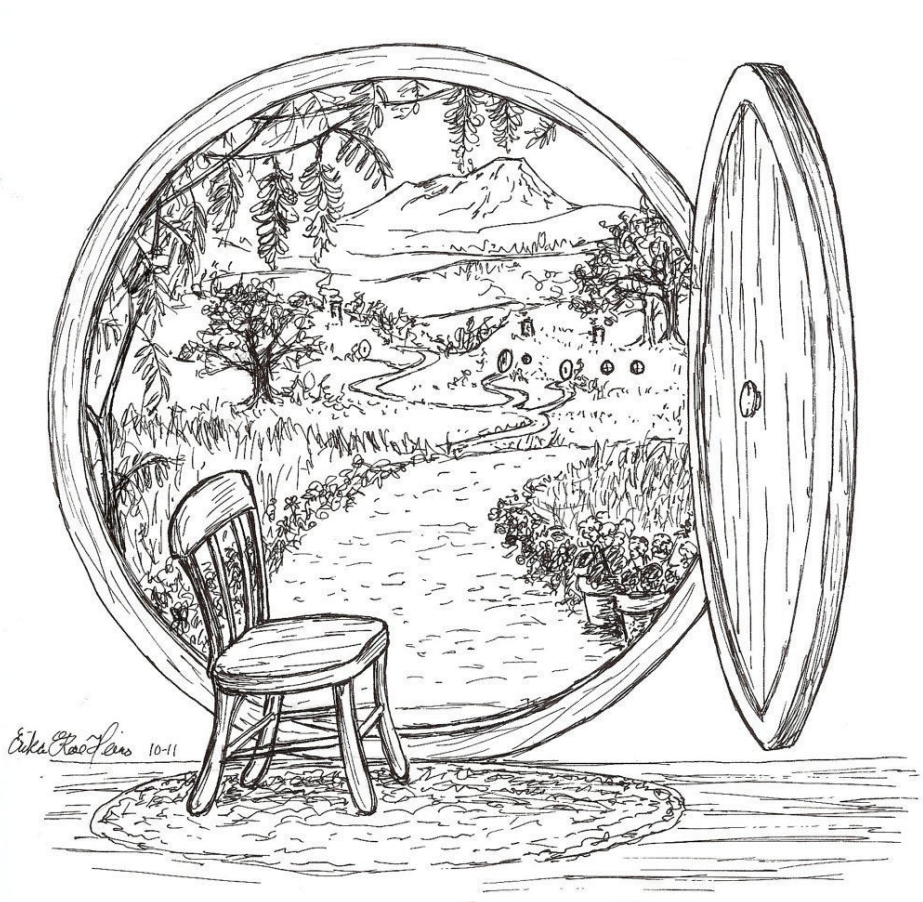
List of Figures

Future tales **351**

List of Tables

“ *Home is behind, the world ahead,*
And there are many paths to tread ”
Through shadows to the edge of night,
Until the stars are all alight.

J. R. R. Tolkien, *A Walking Song*



Introduction:
a moving people

An unexpected journey

On a branch of a tree there lived a creature. Not a nasty, dirty, uncomfortable tree, covered with grime and visited by creeping and scary animals. Actually, that particular branch had a very large base, where he could sit and rest. The wide umbrella-crown of leaves sheltered him from the scorching African sun, the lashing wind and the dangerous predators, also hiding him from prying eyes.

That had been his home his whole life, and there had lived his family before him. Other creatures like him lived on the neighbouring trees. His neighbours, warm and friendly, had always considered his family very respectable, not only because they were strong and thriving, but also because they never had any adventures or did anything unexpected.

For quite some time now, the weather had been changing, the great rainforests were moving back dried by the sun and his neighbours were thinking about moving. The trees were fewer and smaller and the open grasslands of the African savannah were beginning to beckon.

He let his gaze sweep over those vast lands until the horizon. Surely there will be other forests and other trees to live on. But what if we move and we don't like it?

He couldn't have imagined that in a few years, he would have gathered his courage, climbed down from his tree, wandered across those alluring lands, thus getting to Big Water.

That some of his descendants, a few million years later, would have walked for thousands of kilometres, reaching foreign lands, where everything — the weather, the trees, the animals, the soil — was different.

That, after taking different paths, his descendants would have become more and more diverse among themselves. Some would have been called Erectus, some Neanderthal, others Sapiens and so on. Some of them would have met again, fallen in love and lived together.

That the Sapiens would have found a way to make plants and even trees — like his house! — grow; other Sapiens would have tamed the big animals and used them to feed their man cubs; still others would have built cities and vehicles to move faster.

Definitely, he couldn't have imagined that his descendant would have walked longer and further, mixing between themselves and reaching every corner of what they would have discovered being a planet, a round planet, floating through the space. He could have never, ever imagined that they would have walked outside their planet, on the moon, that small whitish ball that he had many times seen sticking on the blue blanket over his head.

He couldn't have known anything about other Sapiens, who would have desperately run from his same lands, crossed the deserts and drowned in the Big Water, stretching their hands.

He didn't know anything about all those people who would have wandered through the vastness of the earth million of years later him. But the same feelings that would have pushed them to travel were starting to arise deep in his heart.

That first migration, so simple, so short, down from that tree, would have changed everything.

> *Then something Tookish woke up inside him,*
> *and he wished to go and see the great mountains,*
> *and hear the pine-trees and the waterfalls,*
> *and explore the caves,*
> *and wear a sword instead of a walking-stick.*
>
> J. R. R. Tolkien, *The Hobbit*

"Not all those who wander are lost."

J. R. R. Tolkien, *The Lord of the Rings*

An introduction to human migrations

MIGRATION has been a fundamental factor for all life on Earth. In fact, both animals and plants move, firstly because they share the same destiny of the lands where they are growing and secondly because they can actively spread from a land to another, permanently changing their home or following the migratory seasons.

And humans are no exception. Since around 6 million years ago, when that African hominin boldly ventured down of his tree, they have never stopped: they have wandered through the endless lands and open spaces of Africa. Then, when the first *Homo* appeared, by going after the horizon line, they went out of Africa over and over again in multiple waves, spreading to each continent of the planet.

During these ancestral journeys, migration has been an intrinsic element to human nature and an essential evolutionary factor. Our unstable condition on this planet, challenged by continuous climate change, makes migration a vital strategy of adaptation and flexibility (Pievani, T. and Calzolaio, V., 2016). Moreover, migrations have accelerated the human evolutionary trail from a biological and a cultural point of view: the bipedal locomotion has both deeply shaped our skeleton and freed our hands which now could be used for doing things. However, while moving, humans have not only adapted to the environments they have met, but they have also irreversibly altered the territories and the lands they have reached, forcing other humans, animals and plants to move away.

In the great circle of life, also migrations are evolving and changing in parallel with humans and their environments. Since the first unaware African wanderings, migrations have become faster and more focused, spanning from the moving of a family to the exodus of a whole population, by foot, wagons, ships and planes.

All this went on until the present days, when, even faster than people, ideas can migrate, spread and change so much our planet.

The concept of human migration and its evolution

"Migrate". *MIV* from Sanskrit and *MIG* from Latin simply mean the movement from a place to another. The Latin form derives from the putative word *migros*, whose Indo-European root indicates the social meaning of *proximity* and *alliance*. It also has a particular declination for *exchange*, *changing place* and *moving*, thus referring to the dual aspect of migration: travelling and changing place, but also exchanging gifts in order to turn the foreigner into a guest (Pievani, T. and Calzolaio, V., 2016).

In our current language, human migration denotes any movement of people, which ultimately changes the place of living, by going very far from home or moving around a little, for a brief period or forever.

During millennia, human species experienced different forms of migration, which have become more and more complex as the human civilizations evolved. Earlier hunter-gatherers did not follow structured migration paths, on the contrary, they were wandering in small areas primarily guided by environmental constraints more than by determination and awareness. Then, a new way of interacting with their territories came to life: a nomadic lifestyle where moving is a conscious strategy to survive in wider areas, following seasonal cycles of the climate, their preys and water.

With the agricultural revolution, human populations acquired even more consciousness of their necessity to migrate, thus spreading geographically, demographically and culturally (Pievani, T. and Calzolaio, V., 2016). However, with this transition, human migrations started to be forced by violence: people were migrating to conquer and invade new territories, while the resident ones had no other choice but to move elsewhere.

As civilizations became more sophisticated, human populations felt necessary to organize the migrations paths, to build roads, bridges and tunnels and to develop new tools for orienting and travelling. During and immediately after the agricultural revolution, there were a lot of migrants and, most of the time, they were welcome in their new territories.

Then, the rise of cities and civilizations brought along drawing imaginary and so meaningful lines — state boundaries — but human migrations naturally adapted to the new immaterial construction and kept on happening. In this very situation, migrant people had a key role: going on mingling cultures and languages, thus preventing boundaries from being freezed (Pievani, T. and Calzolaio, V., 2016).

In parallel with the evolution of human civilization, also the definition of migrant changed. In fact, a plethora of words can be used to indicate people who migrate: in addition to *migrants*, there are *emigrants* or *immigrants*, *economic migrants*, *climate migrants* or *refugees*.

The people who move are always migrants, however, the reason why they move defines the category they belong to, while their social status, wealth and education can alter the common perspective and influence their social acceptance.

Figure 1. Refugees crowd on board a boat some 25 kilometres from the Libyan coast. The picture has been taken in July 2014, prior the rescue by an Italian naval frigate. Image credit: Massimo Sestini

Especially in the recent past, being a migrant has an increasingly negative connotation, turning the prevailing view of migration flows into "inva-

sions", "catastrophes" and "emergencies" (Figure 1).

As we shall see later, since that first migration down from a tree, we have always been on the move, but now is not the best time to be a migrant.

Why people are moving and why they won't stop

Migration is a sea-change for the life of the people, as in many cases it is permanent or of uncertain duration. We then wonder which are the reasons that push or force a person, a family, or an entire population to move.

If we consider the evolution of all life forms and their migrations, we can see that the leading causes of moving were climate change and the distribution of the biodiversity (Pievani, T. and Calzolaio, V., 2016). The migration paths of the human species inhabiting our planet followed those same rules. Obviously, our ancestors preferred to live near water sources and in a warm climate, so, whenever these conditions strayed away, they left and followed them.

In particular, the climate has been, and it is still now, a crucial cause of migration. If we think of the multiple waves Out of Africa, we see that the effects of climate change on what is now the Sahara desert had played a major role, explained in the "Sahara pump theory" (Van Zinderen Bbarker, 1962). Cyclically, Sahara lands were transformed from a grassy and fertile prairie into a barren and hostile desert, thus firstly attracting animals, including us, towards Northern Africa and the Mediterranean Sea and then ejecting them Out of Africa.

All over the world ice ages and interglacial periods followed one another, affecting the air temperature and the sea depth and causing, consequently, the extinction of some species, the death of some individuals and the migration of others (Figure 2). But when the climate is benevolent, human species expand both geographically and culturally: during an interglacial period, between 8,000 and 5,000 years ago, the largest expansion of agriculture, commercial exchanges and culture occurred, leading to the foundation of the major civilizations. Climate itself was later responsible for their destruction.

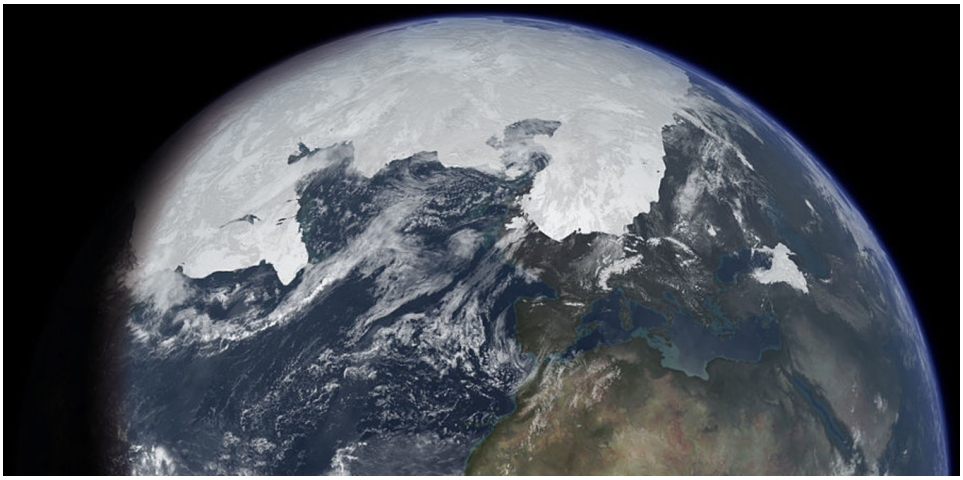


Figure 2. Artist's impression of an ice age. Image credit: Ittiz/wikimedia, CC BY-SA.

However, while climate changes are usually slow and progressive, thus leaving the opportunity to decide whether to stay or leave, extreme geophysical events are immediately crucial for life and death in many parts of the world, forcing people and animals to run away rapidly.

Climate change has always been a reason to migrate; however, the modern-day drive to move is even more urgent, as the climate is changing more quickly. Moreover, the failure to assess the environmental impact of human activities is exacerbating the amplitude, and so the consequences, of climatic and geophysics events, thus forcing more and more people to turn into "climatic" or "environmental" migrants.

Another old but still current reason to move is war. We have already said that since the Neolithic expansion and the construction of the first cities, social relationships became more and more complex, as many more people had to interact in small areas. Killing and waging wars were attractive options for people but bring deaths and people who migrate to run from death and slavery. Forced migration started to occur: slaves under the Roman Empire were deported to present-day Italy from many Roman provinces of the Mediterranean basin. Much more recently, in the 17th century, one of the biggest forced migration of history came about: around 10 million Sub-Saharan African individuals were deported to South America, and many others to North America, Indian Ocean countries and Mediterranean states.

Reconstructing human migrations

Migrations, in all their forms, occupied a huge portion of our time. However, not all human beings experienced or are experiencing migrations during their lives; some of them never changed physically or culturally their home. The time of migration, then, is fragmentary: it is composed of intervals of time which, only in the very long term, become real paths (Pievani, T. and Calzolaio, V., 2016).

Moreover, in most cases, migration is not a straight line, but is much more alike to a winding trail going on intermittently towards a new home and then back again, hindered by biological constraints first and environmental barriers then.

The reconstruction of those ancient trails is not an easy job; however, the people who move, consciously or not, leave prints. Time passes by, and some of those prints are masked in dust or covered by traces of people who come after, but others are left untouched on the ground, in a cave or in our genome.

Fortunately, in order to reconstruct the human past, science takes advantage of different disciplines, making thus possible to follow the footsteps of our ancestors with a shovel, a lens on an ancient book and a pipette.

First of all, archaeology can shed light on the past trails by observing material remains. The old patterns of migration can be deductively reconstructed by analysing, for instance, burial rites and their changing over time. The arrival of one or two wanderers should not affect the burial practise of the area: probably, if the newcomers died far from home, they would have been buried in the local manner. Conversely, when entire populations move, they usually bring with them their own traditions and burial rites (Manco, J, 2015). However, due to the complexity of human beings, these cannot be absolute rules: wanderers can change their traditional behaviours and rituals whey they adopt a different religion.

Archaeological culture

We find certain types of remains - pots, implements, ornaments, burial rites, house forms - constantly recurring together. Such a complex of regularly associated traits we shall term a "cultural group" or just a "culture". We assume that such a complex is the material expression of what today would be called a people.
V. Gordon Childe

A second tool we can use to study past migrations is pottery. Together with the above-mentioned burial rituals, ornaments and house forms, pottery is one of the traits which scientists use to define an archaeological culture. In this case, the sudden transition from a recurrent pottery style to a completely new one may suggest a migration. However, "pots are not people", says a common refrain and the diffusion of techniques or new religions may happen without necessarily the movement of people.

Innovative technologies, metallurgy and new lifestyles, such as farming and agriculture, may also be used to track the presence of a "culture" and its movements through the time. Moreover, the time spent to adopt a new

lifestyle could also suggest whether migration is involved or not: a sudden population growth which coincides with farming suggests the arrival of people completely familiar with that innovation, conversely, if the growth is slow, this may indicate that the new lifestyle was being gradually adopted by locals (Manco, J, 2015).

The introduction of radiocarbon dating techniques in 1947, working directly on fossils, sparked a revolution in the archaeological field. This technique measures the amount of the radioactively unstable ^{14}C carbon isotope in a sample and the rate of decay of the isotope itself (with a half-life of approximately 5,730 years) can be used to infer the age of the sample. Its precision in dating declines with the increasing age of the samples; however, we can expect reliable results up to 45,000 years ago (Mellars, 2006).

Another tool widely used for looking into the past is the study of modern and ancient languages, place-names or ethnonyms. Very briefly, these analyses are based on the assumption that if two different languages share some similarities, they have to be related and so may be the people speaking them.

Although these tools have been a great help in retracing ancient migrations, some questions are still unanswered. For instance, an open problem among archaeologists is discerning when cultural transitions are due to movements of people or to the spreading of ideas. In some cases, even a careful observation of the material expression of a culture cannot provide a clear solution to the dilemma. As we will see later, in some of these cases, such as the spreading of the farming lifestyle during Neolithic or the Bell Beaker expansion, scientists found an answer by taking a different path.

Actually, they just needed to look deep within themselves, in their DNA. By simply following the genetic footprints our ancestors left in their descendant's DNA, they could reconstruct the paths that the ancient populations took thousands and thousands of years ago (Figure 3). As a matter of fact, human migration may divide populations at times, but far more often migrations join and shuffle human populations, thus increasing the genetic variability (Pievani, T. and Calzolaio, V., 2016). Thus, by tracing the appearance and frequency of genetic markers in modern-day human populations, we can bring back to light the ancient trails our ancestors have drawn in our genome.

Figure 3. Map of Human Migration. Image credit: National Geographic, Genographic Project.

"I am terribly proud of
I was born in Cambridge in 1952
and my initials are D N A!"
Douglas Noel Adams, *The Salmon of Doubt*

Following DNA trails

How is it possible to dig out the ancient stories of the many wandering lives buried in our DNA?
First of all, DNA is the instruction book for replicating the cells and building the entire organisms, lying inside the cell nucleus of all living crea-tures. The book is easy to read, as only four letters have been used to write it; however, we are still a long way from understanding how this overwhelming simplicity can rule all life on Earth.

In the human genome, this elemental composition of four letters - C, G, A and T - forms a book 3 billion letters long grouped into 23 chapters: the chromosomes. As a human being inherits two sets of genomes (one from his father and one from his mother), each cell, with the exemption of gametes and anucleate cells, contains two sets of chromosomes, making a total of 6 billion base pairs of DNA per cell or, in other words, about 2 meters of DNA.

Another DNA molecule lives inside our cells and precisely, in the mitochondria: the mitochondrial DNA (mtDNA). Its size, 16,569 base pairs, is diminutive if compared with nuclear DNA, however, there are hundreds or thousands of mitochondria per cell and each mitochondrion contains 15 copies of mtDNA at most.

We can read the nuclear and the mitochondrial DNA books, by looking at the sequence of letters: the order of the bases itself determines the information available for life and it is accountable for the determination of

differences among us.

In fact, DNA of individuals belonging to the same species contain some differences, which are more when we compare different species and even more with distantly related species. This happens because genetic diversity is a function of a population's age; in other words, the earlier a species emerges, the more time it has to develop and accumulate differences.

As the human species is relatively young - emerging around 200,000-300,000 years ago - its genetic diversity is lower than that of many other species, including the chimpanzee, our nearest evolutionary relative. To give you a number, any two humans differ, on average, at about 1 in 1,000 DNA base pairs. This 0.1% is that interest population geneticist the most.

Genetic data and population geneticists

When Darwin first stated that "Light will be thrown on the origin of man and his history" near the end of the Origin of Species, the only two lines of evidence toward this end were comparative anatomy and palaeontology. But now, we are in an era of enormous potential for studies of our evolutionary past. As we will see, genomic analyses and the data that they produce allow us to examine genetic diversity in our genome among diverse populations of the world and compare it with those of our close primate relatives in order to answer age-old questions about where, when and how we evolved. And these are precisely the aim of population and human evolutionary genetics.

However, genetic diversity comes in many different forms. The substitution of C with T at position 478 of the gene *MC1R* is called a single nucleotide polymorphism (SNP) and individuals with this SNP will have red hair and freckles (Branicki *et al.,* 2007). Another type of genetic variation is the deletion or the insertion of a letter in the DNA sequence. This type of variant is called an InDel and one example is a three-base deletion in the coding sequence of the gene *CFTR* which, if homozygous, causes cystic fibrosis. Then, the presence of a variable number of repeat sequences in a stretch of DNA makes short tandem repeats (STRs) or microsatellites, while the different number of copies of a sequence is called copy number variation (CNV).

These many kinds of genetic differences arise because the DNA replication process, despite being very accurate, can make mistakes inserting a wrong nucleotide or sometimes adding or eliminating other others, at a rate of about 1 per every 100,000 nucleotides. Fortunately, most of these errors are fixed by the DNA repair machines; however, in some cases, DNA muta-

tions can escape those mechanisms, thus being passed down to the next generation.

If the mistakes during DNA replication create new mutations from scratch, the recombination process can shuffle these differences from one generation to another. The genetic recombination involves swapping portions of homologous chromosomes, thus creating new combinations of genes which are different from either parent's genetic material.

Once these differences arise, their frequency in the population may change, due to the operation of various other factors, including natural selection, environmental changes, random changes from one generation to the next, admixture events and the star of this thesis, migration.

Studying the pattern of these DNA changes and their effect on genetic characteristics in different populations allow us to take a trip back in time, thus retracing the events of the past accountable for those differences.

The evolution of evolutionary genetics

Notwithstanding the above, first attempts to infer human population history weren't made on DNA, but by using classical markers, such as ABO blood groups and protein allomorphs (Cavalli-Sforza *et al.,* 1994).

However, when the first methods for detecting DNA differences were developed, geneticists became immediately aware of having found a treasure trove. The first chapters of our genomic books to be read where the uniparentally inherited loci, mtDNA and the non-recombining portion of the Y chromosome (Table 1). Both markers have proven to be very informative about the human past, thus providing valuable insight about our maternal (mtDNA) and paternal (Y chromosome) lineages (Denaro *et al.,* 1981; Casanova *et al.,* 1985; Cann *et al.,* 1987; Hammer & Zegura, 2002; Underhill & Kivisild, 2007; Batini *et al.,* 2011). For instance, the discovery in 1987 of the African root of the human mtDNA tree, which provided first strong evidence of our African origins (Cann *et al.,* 1987), is due to mtDNA. However, due to their uniparental inheritance pattern, while they could highlight sex-biased migrations, they had limited power when it comes to the study of genetic diversity as a whole (Balloux, 2010).

uniparental loci

Genetic material in organisms with distinct sexes that is passed down to offspring through inheritance only from one sex; that is, mitochondrial DNA and the non-recombining portion of the Y chromosome *(Veeramah & Hammer, 2014).*

During the seventies, the techniques developed for the study of nuclear genetic variations allowed to analyse a limited number of markers. Nevertheless, in 1984, Ammerman and Cavalli-Sforza, by looking at the allele frequencies of only 34 loci in 16 chromosomes from different populations tried to solve the long-debated matter of the agriculture diffusion. According to

their data - analysed employing of a statistical technique on which we shall come back later - the European agricultural transition was mainly prompted by the diffusion of farming populations and not just by the spreading of ideas (Ammerman & Cavalli-Sforza, 1984).

Then, the development of DNA hybridization microarrays allowed to detect hundreds of thousands of genetic markers, up to 5 million polymorphisms nowadays (Table 1). This technology opened the floodgates for human population genetics studies, enabling researches of fine-scale population structure (Biswas *et al.,* 2009; Henn *et al.,* 2010; Lawson *et al.,* 2012; Leslie *et al.,* 2015; Fiorito *et al.,* 2016; Raveane *et al.,* 2019), positive selection and adaptation to specific environments (Coop *et al.,* 2009; Fumagalli *et al.,* 2015) and deepening our understanding of human mutation and genetic recombination rates (Hinch *et al.,* 2011; Sun *et al.,* 2012; Kong *et al.,* 2012). Moreover, this technical approach proved to be effective above all in medical genetics because, through Genome-Wide Association Studies (GWAS), it was possible to discover many variants associated with disease traits (Manolio & Collins, 2009).

However, despite the broad applications in human genetics, SNP genotyping microarrays suffer from ascertainment biases as, due to their basic operating principle - DNA hybridization - they can only interrogate genetic variants which have been already discovered. As a disproportionate number of SNPs have been identified by analysing European individuals, these platforms miss much diversity from other populations (Veeramah & Hammer, 2014; Lachance & Tishkoff, 2013).

In contrast, sequencing approaches (Metzker, 2009), while more expensive, have the valuable advantage of being able to discover new genetic variations. In fact, the automation of the first sequencing method — Sanger sequencing — made it possible to reconstruct the human reference genome sequence (Lander *et al.,* 2001). During the first 2000s, new sequencing technologies with increased throughput and an affordable price were developed, giving the possibility to produce an enormous volume of genetic data cheaply. These new methods were called Next Generation Sequencing (NGS, Table 1). Their operating chemistry, obtaining continuous DNA sequence from multiple chromosomes at thousands of loci, allows reducing ascertainment biases. In this way, it increases power to infer demographic events, regardless of whether genetic data are obtained from the entire genome of from large portions of it (Veeramah & Hammer, 2014).

Data type	Uses and advantages	Limitations
mtDNA	• The absence of recombination allows reconstruction of a gene tree • Smaller N_e than autosomal DNA, which allows better discrimination between populations • Samples from many thousands of individuals can be characterized at low cost • High copy number makes it amenable for ancient DNA extraction and analyses	• A single genealogy contains little information about the underlying population history • Likely to be subjected to the effects of natural selection • High uncertainty in mutation rates
NRY	• The absence of recombination allows reconstruction of a gene tree • Smaller N_e than autosomes, which allows better discrimination between populations • Samples from many thousands of individuals can be characterized at low cost	• A single genealogy contains little information about the underlying population history • Likely to be subjected to the effects of natural selection • High uncertainty in mutation rates and mutation model • Ascertainment bias results from the genotyping of specific SNPs
Autosomal STRs	• Hundreds of independent STRs can be genotyped in many individuals, which reduces the effect of evolutionary stochasticity • Their high mutation rate is useful for inferring recent demographic events and for distinguishing between closely related populations	• Limited inference of demographic events at deep timescales • High uncertainty in mutation rates and mutation model
SNP microarrays	• Hundreds of thousands of SNPs can be genotyped in a single experiment • Unprecedented resolution of population structure	• Large ascertainment bias results from the haphazard way by which SNPs were discovered • Less powerful for making inferences in populations that are diverged from those in which SNPs were discovered
Second-generation sequencing	• Massive amounts of relatively unbiased sequence data can be obtained from targeted regions or entire genomes compared with Sanger sequencing • High throughput and does not require a targeted pre-PCR step, which allows sequencing of ancient DNA • Lowest per-base cost of any current sequencing methodology	• Relatively error prone compared with Sanger sequencing • Biases may arise with regard to regions that are preferentially sequenced • Sequencing is through short reads (100–150 bp), which restricts the use of methods that require haplotype-phased data
Third-generation sequencing	• Can generate long sequence reads (>10 kb) • Some methods can sequence DNA from single cells, which is particularly useful for very ancient samples • Long reads may also allow *de novo* assembly and thus reduce reference biases	• Per-base cost is currently more expensive than second-generation sequencing • Bioinformatic tools have not yet been fully developed to cope with the increased read length

mtDNA, mitochondrial DNA; N_e, effective population size; NRY, non-recombining portion of the Y chromosome; SNP, single-nucleotide polymorphism; STR, short tandem repeat.

Table 1. The different types of genetic data used to study the human past. Image taken from Veeramah & Hammer, 2014.

The evolution of technologies for analysing and comparing human genetic variations enables the production of genetic data from different human populations at an unprecedented scale. Motivated by this fact, many different consortia and research institution put together their efforts to study human genetic variation at a broader scale. Since the assembly of a draft human reference genome in 2001 (Lander *et al.,* 2001), several human genetic datasets were collected and made available for researchers: HapMap

(Consortium, 2003; Consortium, 2005), the Human Genome Diversity Panel (HGDP, Cann *et al.,* 2002; Li *et al.,* 2008), whose samples have been recently sequenced with high-coverage (Bergström *et al.,* 2019), the 1000 Genomes Project (1KGP, Auton *et al.,* 2015), the African Genome Variation Project (AGVP, Gurdasani *et al.,* 2015), the Human Origins Array data (Lazaridis *et al.,* 2014; Lazaridis *et al.,* 2016), the Simons Genome Diversity Project (SGDP, Mallick *et al.,* 2016) and the Estonian Biocentre Human Genome Diversity Panel (EGDP, Pagani *et al.,* 2016).

Multidisciplinary teams

The unprecedented scale of genomic data being collected has revolutionized how and what we can learn about our origins. To keep up with the ever-growing body of genetic data, evolutionary genetics gradually turned into a multidisciplinary field of study.

The first visionary promoting this transformation was Luigi Luca Cavalli-Sforza, an Italian geneticist who spent the majority of his academic life at Stanford. In 1994, with the publication of *The History and Geography of Human Genes*, it was crystal clear that the study of our past wasn't just an archaeologist' and historians' matter any more. In that book, he synthesized what was known about the great migrations of our past from archaeology, linguistics, history and genetics (Cavalli-Sforza *et al.,* 1994). Moreover, he was also one of the first to analyse genetic data with the aid of statistics.

In the 1984 work (Ammerman & Cavalli-Sforza, 1984), Cavalli-Sforza and colleagues applied a principal component analysis (PCA) on genetic data from 95 alleles in an attempt to summarize the information coming from the allele frequencies across different populations.

PCA is a non-parametric dimensionality-reduction method used to visualise and identify spatial patterns of genetic variation. As a statistical framework to summarize genetic data, PCA would have proved to be effective and widely used among population geneticists (see page 150 for some details about how PCA works and Figures 70, 71 and 88 for some applications).

The great intuition of Cavalli-Sforza and colleagues was to superimpose the principal component coordinates of each sample on a geographic map. By plotting the first principal component onto a map of West Eurasia (Figure 4), he observed that it reached its maximum values in the Near East and then declined along a southeast-to-northwest gradient into Europe. He interpreted this cline as the genetic traces left by the migration of farmers into Europe from the Near East, concluding that the spreading of agricultural

lifestyle was also made by the movement of people.

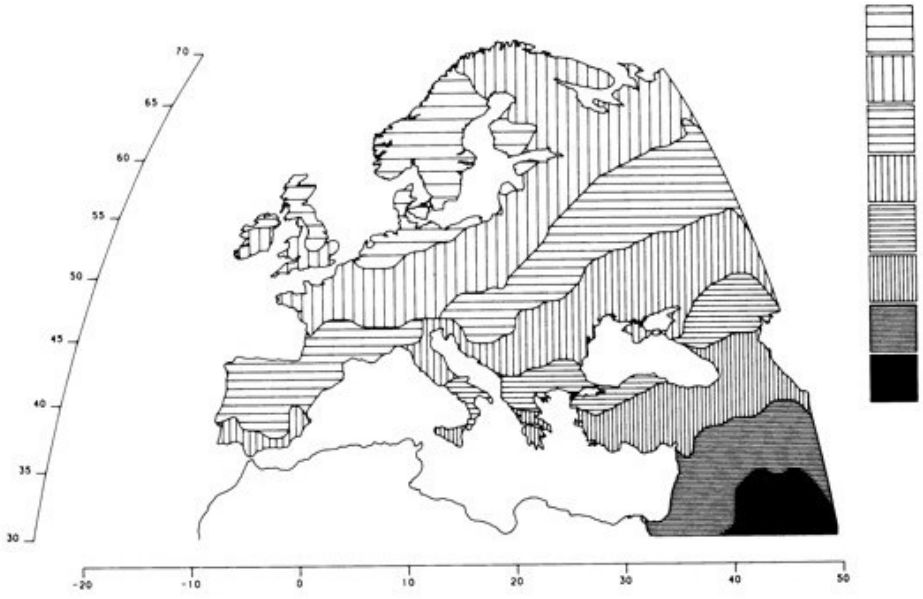


Figure 4. First principal component inferred from the genetic variation of 95 classical polymorphisms across Europe. This component is placed on a map of Europe and, according to the author, *"is almost superimposable to the archaeological dates of the spread of farming from the Middle East between 10,000 and 6,000 years ago.*" Image taken from Cavalli-Sforza, 1997.

The amazing idea was that the genetic footprints the ancient migrations left in our genome could be uncovered by directly looking at genetic frequency differences among modern-day populations. Cavalli-Sforza himself says:

> *A method that proved especially useful is a geographic study by principal components (PCs) or related techniques. It partitions the total variation into independent, additive components, ordered by their relative importance in determining the total variation observed. As for trees, many genes are necessary, and observations must be spread as regularly as possible over the area being analyzed; as for trees, the best check of the validity of the conclusions is their independence from the markers employed: that is, their reproducibility with different sets of markers.*

However, the models inferred from PCA patterns need to be interpreted with some caution. In 2008, a paper of John Novembre and colleagues demonstrated that the same gradual patterns could simply arise because there is a correlation between genetic and geographic distance, a phenomenon called "isolation by distance", as it happens in a PCA on European individuals (Figure 5). Consequently, the same patterns observed by Cavalli-Sforza could have also been formed without migration (Novembre & Stephens, 2008; Novembre *et al.,* 2008).

Other approaches developed for exploring and visualising genetic variations were, among the others, STRUCTURE (Pritchard *et al.,* 2000) and ADMIXTURE (Alexander *et al.,* 2009). These software tools implement an unsupervised analysis for inferring the likely ancestries for each target population, using a bayesian approach and a maximum likelihood framework, respectively. In both cases, the user defines a number of ancestral populations K. The software finds the most likely allele frequencies of the K populations and their proportion in each sample of the dataset. The accuracy of the inferred values is then checked by cross-validation. Examples of ADMIXTURE outputs can be seen in the following chapters (pages 291, 297 and 335).

While the results of these methods can be interpreted to infer past migrations and other demographic events, it can be difficult to understand whether some patterns are due to recent admixture between distinct populations or to shared ancestry. The common pitfalls of this process are examined in Lawson *et al.,* 2018.

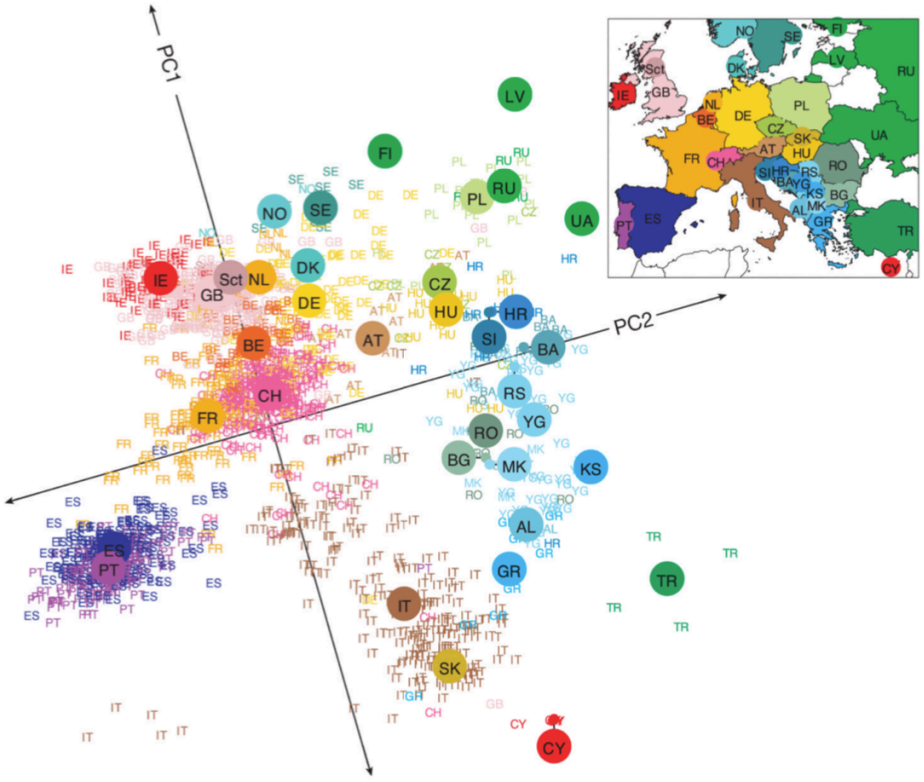


Figure 5. First two principal components inferred from genetic data of 1,387 Europeans. Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. Image taken from Novembre *et al.*, 2008.

Genetic drift

Change in the gene pool of a small population that takes place strictly by chance. Genetic drift can result in genetic traits being lost from a population or becoming widespread in a population without respect to the survival or reproductive value of the alleles involved.
Encyclopaedia Britannica.

For this reason, when the purpose is inferring whether admixture happened or not, it is usually better to avoid speculation and formally test it. In order to do so, a popular approach closely-related to the *f*-statistics (Wright, 1951) has been developed. This method, called F-statistics, measures the "shared genetic drift" between sets of two, three or four population, thus testing admixture hypothesis between populations. The underlying idea is that shared drift implies a shared evolutionary history. Some useful reviews about this framework are Patterson *et al.*, 2012 and Peter, 2016. I will more extensively discuss and utilize a related method — *D*-statistics — at pages 65 and 153.

In contrast to the above-mentioned techniques based on the analyses of *global* SNP allele frequencies, a series of haplotype-based *local* methods have

also been developed.

A haplotype is a group of SNPs that were inherited together from a single parent. During the meiotic events, the only recombination can break the groups of markers, thus preventing them from being inherited together. Uniparental markers, such as mtDNA, the non-recombining portion of the Y chromosome and the X chromosome in males, do not recombine and, thus, they are haplotypes on their whole.

Haplotype-based techniques, since they consider groups and not single markers, have been showed to be more powerful (Lawson *et al.,* 2012; Hellenthal *et al.,* 2014; Leslie *et al.,* 2015) and less affected by SNP ascertainment biases (Conrad *et al.,* 2006; Hellenthal *et al.,* 2008) than SNP-based methods.

Moreover, a major limitation of global methods is that they take into account only the physical position of SNPs along the DNA sequence, thus treating them as they were independent observation. However, we know that variants on the same chromosome are inherited together unless a recombination event happens between them. When we consider entire populations, this means that physically close markers are in linkage disequilibrium (LD), as a consequence of having a shared history of descent. However, this results in the invalidation of the assumption of independence. For this reason, many of the SNP-based methods are not LD-aware; hence it is usually necessary to perform some for of LD-based pruning before invoking them.

LD

Linkage disequilibrium refers to the non-random association of alleles at two or more loci in a general population. When alleles are in linkage disequilibrium, haplotypes do not occur at the expected frequencies. Linkage disequilibrium between two alleles is related to the time of the mutation events, genetic distance, and population history.
SpringerLink.

However, it is known that exploiting this correlation structure between genetic markers results in a study of the genetic structure at finer scales (Conrad *et al.,* 2006; Jakobsson *et al.,* 2008; Gattepaille & Jakobsson, 2012; Lawson *et al.,* 2012; Loh *et al.,* 2013; Leslie *et al.,* 2015).

For this reason, a long list of software and techniques exploiting haplotype information to reconstruct population structure at a fine-scale have been produced (Maples *et al.,* 2013; Brisbin *et al.,* 2012; Lawson *et al.,* 2012). Among them, CHROMOPAINTER (Lawson *et al.,* 2012) exploits a likelihood-based model, the *copying model* (similar to the approach developed by Li & Stephens, 2003), and uses a hidden Markov model (HMM). Basically, CHROMOPAINTER identifies at each location of each *recipient*'s two haploid genomes the best matching DNA segment from a set of sampled *donor* individuals. In this way, each recipient haploid genome is a mosaic of DNA "chunks", i.e., a DNA segment painted entirely from a single donor haplotype. For example, Figure 6 describes three *donor* haplotypes $h_1 - h_3$, while haplotype h^* is the *recipient*. The white and black circles represent two different alleles of N SNPs. The *recipient* haplotype h^* is a mosaic of the three *donor* haplotypes. In other words, it has a shared ancestry with

HMM

Hidden Markov Model is a statistical Markov model — a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event — in which the system being modelled is assumed to be a Markov process with unobserved (i.e. hidden) states.
Medium Website

the *donor* haplotype showing the best matching DNA segment. This "copying" approach is performed across all the genome, thus obtaining for each *recipient* a vector of *donors* — termed "copying vectors" or "painting profiles". Under the HMM approach underlying the CP algorithm, the observed states are the allelic information at each SNP, while the hidden states are the *donor* haplotypes the *recipient* h^* copies from (Lawson *et al.,* 2012). For details about the method, refer to Lawson *et al.,* 2012; for examples of the CP and the CP "unlinked" model, see pages 146 and 154, while the code for running CP "unlinked" can be found at page 322.

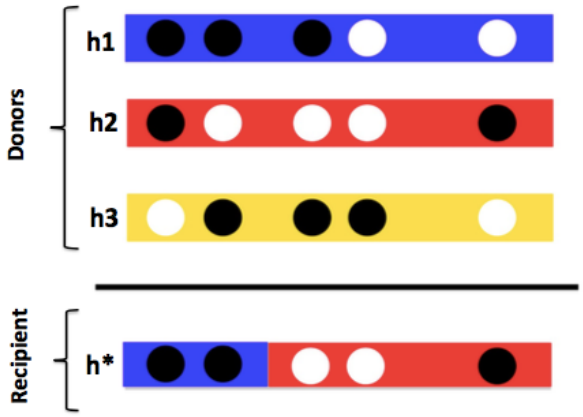


Figure 6. Example of *donor* and *recipient* haplotypes in the CP model. Each point is a SNP and its colour represents a specific allele. Image credit: Dr Lucy Van Dorp.

Ancient DNA: a direct window to the past

The molecular and analytical technologies we described are extensively used for reconstructing the great migrations and demographic events of the past from the genetic differences among present-day people. However, in the last few years, the possibility to extract and sequence DNA from fossils have opened an unprecedented window into the major demographic events contributing to human genetic variation (Figure 7).

It seems obvious that obtaining DNA directly from the bones of people living millennia ago could provide a more precise inference about their lives and movements. However, the DNA extracted from fossils, ancient DNA

(aDNA), raises some issues regarding its reliability: its main problems are fragmentation (Pääbo, 1989), post-mortem chemical modifications (Höss *et al.,* 1996) and contamination (Pääbo *et al.,* 2004).

Under favourable conditions, DNA can survive for thousands of years in the remains of dead organisms (Dabney *et al.,* 2013). However, as time passes by, fragmentation and chemical modifications inevitably alter the original DNA molecule, making it difficult to distinguish real mutations from post-mortem DNA damage.

In living cells, the chemical insults to DNA molecules are kept under control by repair mechanisms. After death, these mechanisms cease to function, allowing nucleases as well as microorganisms to degrade DNA molecules.

Moreover, in warm and wet climate conditions, these effects are exacerbated. Conversely, when the temperature is low, the salt concentration is high or the remains have been desiccated quickly after death, the processes mentioned above are slowed, preventing the destruction of all endogenous DNA (Dabney *et al.,* 2013). As a consequence, the age of a specimen does not necessarily correlate with its DNA quality (Hofreiter *et al.,* 2001a). However, other destructive factors, such as hydrolytic and oxidative processes, may limit the time that DNA can survive in a tissue (Dabney *et al.,* 2013).

A serious problem of aDNA is represented by the contamination of the bones from DNA belonging to researchers — archaeologists or molecular biologists — who handled the sample. It is even worse when working with an ancient human specimen because it is difficult to distinguish the remains' endogenous DNA from contaminants on the basis of the genetic variation patterns. However, this problem also occurs when working with distantly related hominin, such as Neanderthals. In this case, the typical Neanderthal DNA fragment is around forty letters long, while the rate of differences between Neanderthal and modern humans is about one per six hundred letters, thus making almost impossible to say if the DNA sequence comes from the sample or the scientist (Reich, 2018).

However, aDNA, being a direct witness from the past, is a resource too valuable to be replaced and forgotten. Thus, as these limitations were coming to light, always new solutions were developed, from laboratory precautions to minimize the likelihood of contaminations to variant callers capable of taking into account the DNA damage.

Using these innovations, in 2010 the Danish group led by professor Eske Willerslev produced the first high-coverage sequence of a Greenlander ancient human from 4,000 years ago (Rasmussen *et al.,* 2010). Since then, genetic data from thousands more ancient samples has been produced, analysed and published. These efforts, little by little, are filling the gaps in time

and space: scientists have now the possibility of directly accessing human genetic variation from all over the world and from almost every period, from Upper Palaeolithic until today.

And what is more, as we will see later, aDNA technologies produced the first full genomic sequence of an archaic extinct hominin — our Neanderthal cousin (see page 63) — finally allowing scientists to answer some age-old questions about our origins.

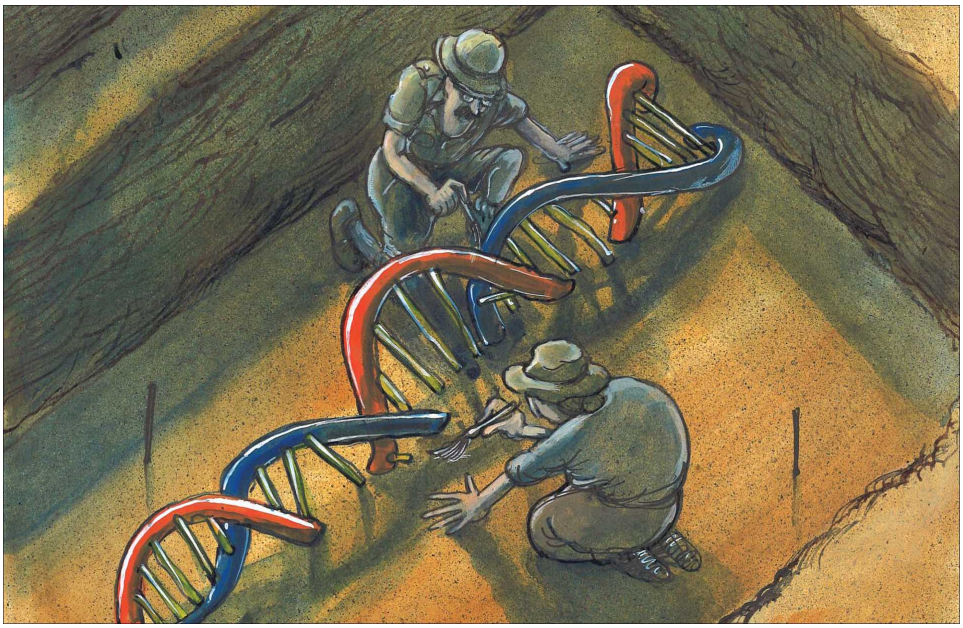


Figure 7. Ancient DNA illustration. Image credit: Martin Rowson.

"Humanisation begins with the feet."
André Leroi-Gourhan, *Le geste et la parole*

A brief history of everyone who ever moved

AFTER the creature left for an *unexpected journey*, climbed down of his tree and started walking, many others followed his example. Three of his descendants — probably *Australopithecus afarensis* — were walking upright through wet volcanic ashes around 3.6 million years ago when the volcano erupted again, covering their footprints with other ashes, thus preserving them. The traces they left are the oldest known footprints of early humans (Figure 8).

The stories of those people, as well as the story of the creature who was beginning to love new adventures, took place in the continent where we spent the majority of our time on Earth. Africa.

There, since around 10 million years ago, the formation of the 6,000 kilometres Great Rift Valley had hindered the Atlantic precipitations, thus leading to a gradual withering of the thick Eastern African rainforests into prairies and savannas. According to one of the oldest hypothesis — the Savanna theory — climate change was one of the compelling reasons behind the beginning of our bipedal adventures, even if in the last decades a series of studies suggested our ancestors' curiosity toward walking had started when there were still enough trees to live on (Senut *et al.,* 2018; Thorpe *et al.,* 2014).

Whatever was the reason for the origins of bipedalism, using just two feet to move had a series on unexpected consequences, many of which were "adaptive compromises" for the hominids that evolved them. For instance,

our ambitions for walking caused lower back and knee problems due to pressure on the spine, complicated childbirth resulting from a repositioned pelvis and the greater chance of choking on food, which is a consequence of a lowered voice box (Friedman, 2006).

Figure 8. Trail of Laetoli footprints. Footprints left, probably, by *A. afarensis*, around 3.6 million years ago in Laetoli (Tanzania). Image credit: John Reader.

Nevertheless, bipedalism was strongly selected in hominids, probably because it could provide some crucial advantages, such as the possibility

to stand up higher than the grassy prairies or to prevent overheating by exposing less of the body to the scorching sun.

> *If it be an advantage to man to stand firmly on his feet and to have his arms free ... then I can see no reason why it should not have been advantageous to the progenitors of man to have become more erect or bipedal.*
> *Charles Darwin, Descent of Man, and Selection in Relation to Sex, 1871*

Another great consequence came with walking upright: it left the hands free for doing things, building tools, writing poems or raising a glass. However, fossils show that bipedalism evolved a few million years before hominids started to build and use stone tools around 3.3 million years ago. Maybe in the first times of their upright wanderings, they had built only wood tools or they had waited a while before taking advantage of their free hands or, maybe, there was no one making good wine.

African roots and the descent of men

Our roots run deep in Africa, but when precisely did they start to grow?

The earliest populations of genus *Homo* emerged from a still unknown ancestral species between 3 and 2 million years ago (Figure 9, Spoor *et al.,* 2015; Villmoare *et al.,* 2015). In this temporal interval "humanity", or what was on the verge to be, spanned from the latest known occurrences of Australopithecus (*A. afarensis* in eastern Africa, *A. africanus* in southern Africa), to the earliest known records of two, perhaps three, species commonly attributed to the genus Homo (*H. habilis*, *H. rudolfensis* and *H. erectus*). In these million years, the earliest undisputed evidence for stone-tool manufacture was produced. Gradually, other defining characteristics of humanity, such as large brains and complex social behaviours, appeared and, with them, humanity was born.

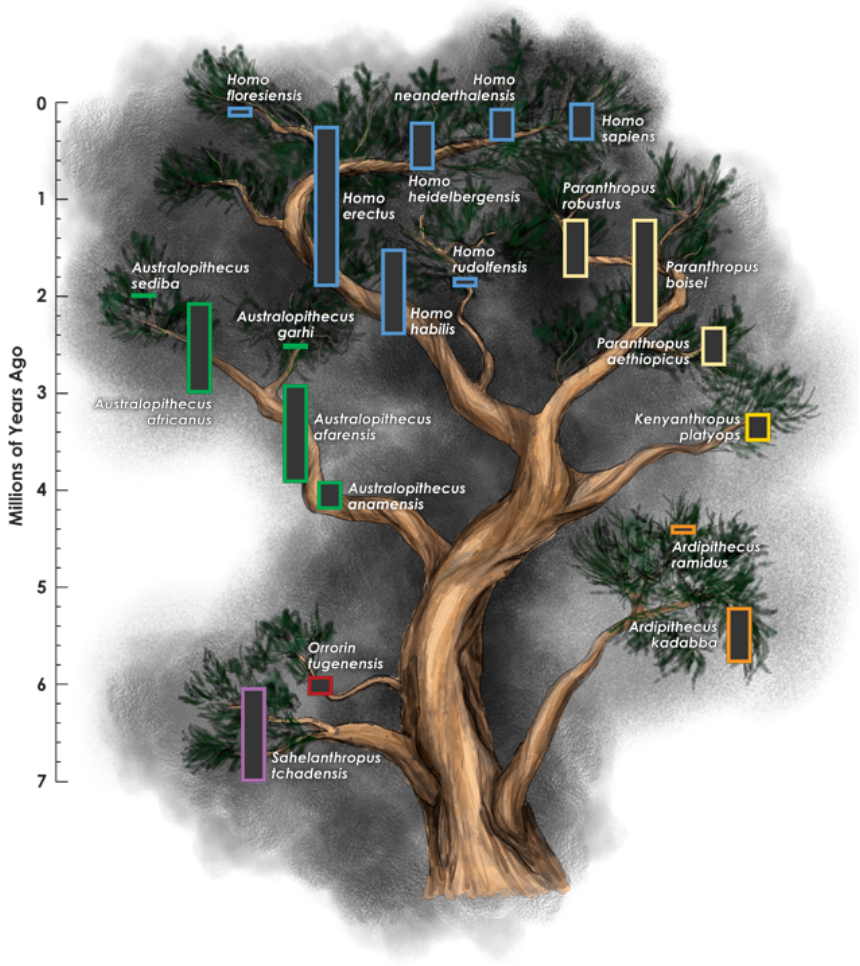


Figure 9. The hominin family tree. This picture does not contain *Homo naledi*, probably because it was drawn before its discovery in 2013, but, after all, we can try and imagine how many species and genera are not represented here. Image credit: K. Cantner, AGI.

The places where these "human" traits occurred were probably more than one. East Africa — where the famously bipedal ape Lucy (*A. afarensis*) lived and where the first *Homo* appeared — has been usually considered the cradle of humankind. However, some other Australopithecines had also been found in Southern Africa, and, nearby a place evocatively called "Cradle of Humankind", in 2013 cave explorers found more than 1,500 bones belonging

to at least 15 individuals belonging of a new species (Berger *et al.,* 2015; Dirks *et al.,* 2015). These individuals showed a miscellaneous of primitive and more advanced features, placing these people — called *Homo naledi* — somewhere between Australopithecus and early Homo species. However, the new dating proposed in 2017 pushed this species in a time very close to us — between 236,000 and 335,000 — meaning that while modern humans to be were already wandering in East Africa, in the South there was someone with a small brain (500 cc) still living on trees (Berger *et al.,* 2017; Dirks *et al.,* 2017; Hawks *et al.,* 2017).

Our first time Out of Africa

As soon as humans acquired all the needed physical features, they stood up, started to walk and never stopped.

Around 1.6 million years ago near Turkana Lake (in the Rift Valley and in what is now called Kenya) a boy, of around nine years old, died. That boy — who is known as the Turkana boy — was about 160 centimetres in tall and, in adulthood, he might have reached even 185 centimetres. His skeleton is incredibly similar to modern humans, indicating fully terrestrial bipedalism, but he has still a reduced brain capacity (880 cc). He was an *Homo ergaster* and the first great walker (Figure 10).

Homo ergaster started almost immediately to use their long legs. They wandered across East and Southern Africa, leaving their traces in Tanzania, Ethiopia, Eritrea and South Africa, and they went on toward East until reaching unexplored lands: they were the first to go out of Africa.

According to the fossil records, they managed to arrive 2 million years ago to the valleys of Lesser Caucasus, at Ubeidiya in the Middle East, in Southern Asia and at Riwat (Pakistan), and then reach China and Java island around 1.5 million years ago (Figure 11). It is worth saying that there could be some confusion about the name of this species, since sometimes they are called *Homo erectus*. The two terms have often been used interchangeably; however, the accepted explanation is that *ergaster* indicates the earlier African form, while *erectus* refers to a later Asian species. Nonetheless, Figure 11 shows the expansions of *Homo erectus*, but it could be considered also valid for the *ergaster* wanderings.

Figure 10. A reconstruction of the Turkana boy. Image credit: S. Entressangle, E. Daynes, Science Photo Library.

Traces of the wanderings of their descendants — probably *Homo antecessor* — got stuck in the mudflows of the Norfolk coast (England) around 850,000 years ago and represent the oldest footprints discovered outside Africa.

The term "diaspora" is often used to indicate this first movement out of Africa, however, as remarked by Telmo Pievani (Pievani, T. and Calzolaio, V., 2016), this is somewhat misleading. Actually, this migration wasn't a migration at all: it was not an intentional exodus of people, but rather a gradual, slow and irregular expansion due to the movements of their settlements, generation after generation. Surely, they were not driven by an

insatiable desire of exploration, but, more probably, by the climate change and the vegetation displacement.

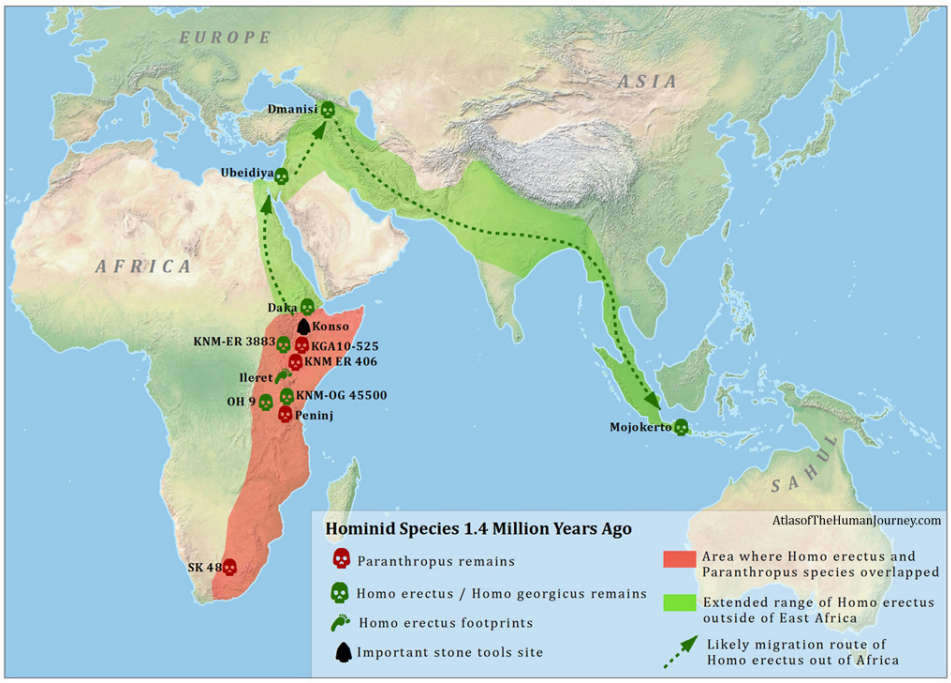


Figure 11. The first migration out of Africa. Image credit: AtlasOfTheHumanJourney.com

In the case of *ergaster*, as exemplified by the existence of *herectus*, movement means division and speciation. Since their expansion into Eurasia, the *ergaster* population become more and more fragmented, separating into many subgroups, which would never meet again. This process would have been called adaptive radiation.

Adaptive radiation

The evolution of an animal or plant group into a wide variety of types adapted to specialized modes of life. Adaptive radiations are best exemplified in closely related groups that have evolved in a relatively short time. *Encyclopaedia Britannica.*

Our second time Out of Africa

A million years passed and evolution went on both in our first home, Africa, and in our summer house, Eurasia; new species flourished, while others turned into dead branches. Among the former, in Africa, between 780,000 and 135,000 years ago, a new species of explorers appeared: *Homo heidelbergensis*. They were more large-brained (1200 cc), they used the "Acheulean" toolmaking technique and they travelled Out of Africa too.

Fossils belonging to *heidelbergensis* started to appear outside Africa since 500,000 years ago and the first discovered one was unearthed in a sand quarry at the Mauer site, nearby Heidelberg (Germany): the Mauer jaw (Figure 12, Schoetensack, 1908).

According to studies, this species was the protagonist of a second big population wave out of Africa (Figure 13). Then, as was the case with *ergaster*, due to the spreading of its population in an immense territory, *Homo heidelbergensis* started to diverge into truly separated species in the different regions of the world. They began to develop different physical features as a result of both environmental adaptation — Eurasian lands, being 10,000 kilometres large, crossed many different climatic regions — and genetic drift (see definition at page 23).

Figure 12. The Mauer Jaw. Fossil belonging to *Homo heidelbergensis* discovered at Mauer in 1907.

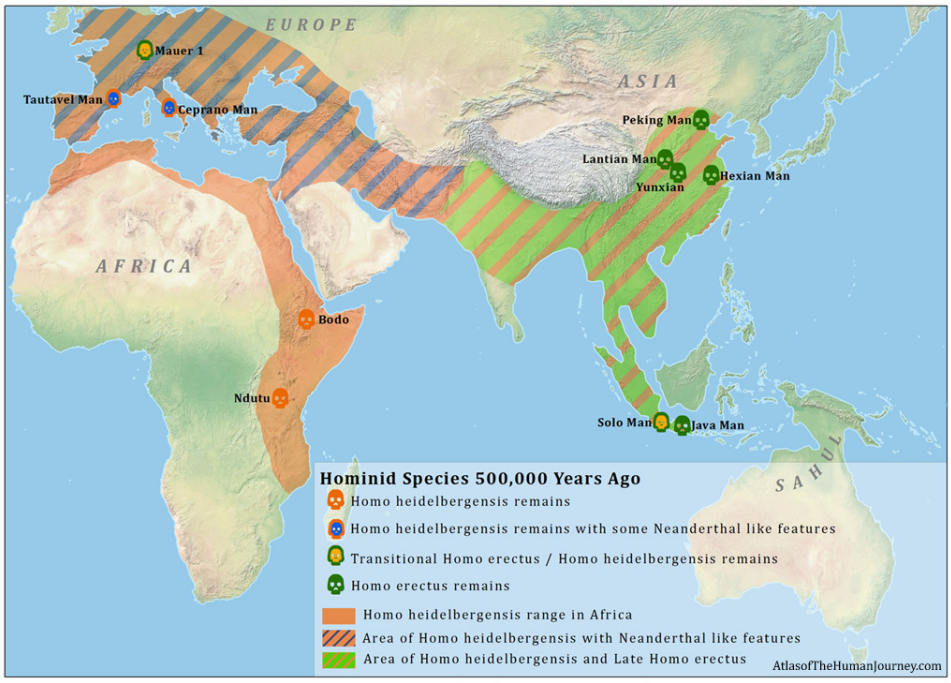


Figure 13. The second migration out of Africa. Image credit: AtlasOfTheHumanJourney.com

In the meantime, the tree of humankind was rapidly growing, ever closer to the modern human branch. In fact, while among European *Heidelbergensis* the "neanderthalization" had begun slightly before than 400,000 years ago (see the second part, at page 53), some characteristics which will be typical of modern humans, such as the big brain, were already started to appear already two hundred thousand years before in Africa. For instance, the Bodo Cranium, discovered in Ethiopia and dated back to 600,000 years ago, had already a brain size of about 88% of modern humans' (1,250 cc, Conroy *et al.,* 2000).

By about 150,000 years ago this "transition" process was already completed and three distinct species had emerged in the different regions of the world from *Homo heidelbergensis*: *Neanderthals* in Europe, *Denisovan* in East Asia and *Sapiens* in Africa.

In a few thousand years they would have travelled and met, but only one of them would have dominated the world.

Once again, African roots

The origin of modern humans has been and it is still now a matter of interest as great are the debates on it. Two main hypotheses about our origins dominating the second half of the 20th century were the Multiregional and the Out of Africa models. The Multiregional model proposed that modern humans evolved independently in different parts of the world, facilitated by a significant gene flow among the different subgroups of *erectus* (Weidenreich, 1940), so that modern humans have ancestral contributions from different hominin populations living in different parts of the globe (Lopez *et al.,* 2015). On the contrary, the Out of Africa hypothesis suggested a completely African origin of *Homo sapiens* and a subsequent dispersal across the world (Stringer, 2002; Stringer & Andrews, 1988; Ingman *et al.,* 2000; Relethford, 2008; Tattersall, 2009). The latter was originally proposed by Charles Darwin in the 19th century, by reasoning about the presence of chimpanzees and gorillas in Africa and the results of Huxley, who suggested that modern humans and apes shared a common ancestor (Darwin, 1896; Lopez *et al.,* 2015). For a detailed discussion about the models explaining our origins and evidence supporting them, see the review of (Lopez *et al.,* 2015).

In the last 150 years, the joint work of geneticists and archaeologists confirmed the African origin of the first modern humans — and now the political winds blowing in the last years need someone yelling it louder — but, at the same time, have unveiled a more complex scenario.

The first genetic evidence claiming an African origin for modern humans came from the analyses of mtDNA phylogenetic trees (Cann *et al.,* 1987; Ingman *et al.,* 2000). Then, both Y chromosome investigations and autosomal studies reached the same conclusion. Moreover, genome-wide data have demonstrated that genetic diversity decreases as a function of geographic distance from East or South Africa. This was shown, for example, by an approximately linear decrease in heterozygosity and increase in linkage disequilibrium (Prugnolle *et al.,* 2005; Ramachandran *et al.,* 2005). Moreover, genetic studies, by looking more deeply at the reduced genetic diversity in non-African populations, suggested that we — modern humans outside Africa — are the descendants of a small subgroup of the original African population. In other words, we are the result of a "bottleneck".

Archaeological evidence supports our African origins, too (Figure 14): the earliest known fossil representative of our species, dating back to around 315,000 years ago, was found in Africa, precisely in Marocco (at a site called Jebel Irhoud, Hublin *et al.,* 2017). The second oldest one is African too, dating back to 260,000 years ago and found in South Africa (at Florisbad,

Grün, R. and Brink, J. S. and Spooner, N. A. and Taylor, L. and Stringer, C. B. and Franciscus, R. G. and Murray, A. S., 1996).

For these reasons, our roots run deep, once again, in Africa, into a period as old as 300,000 years ago.

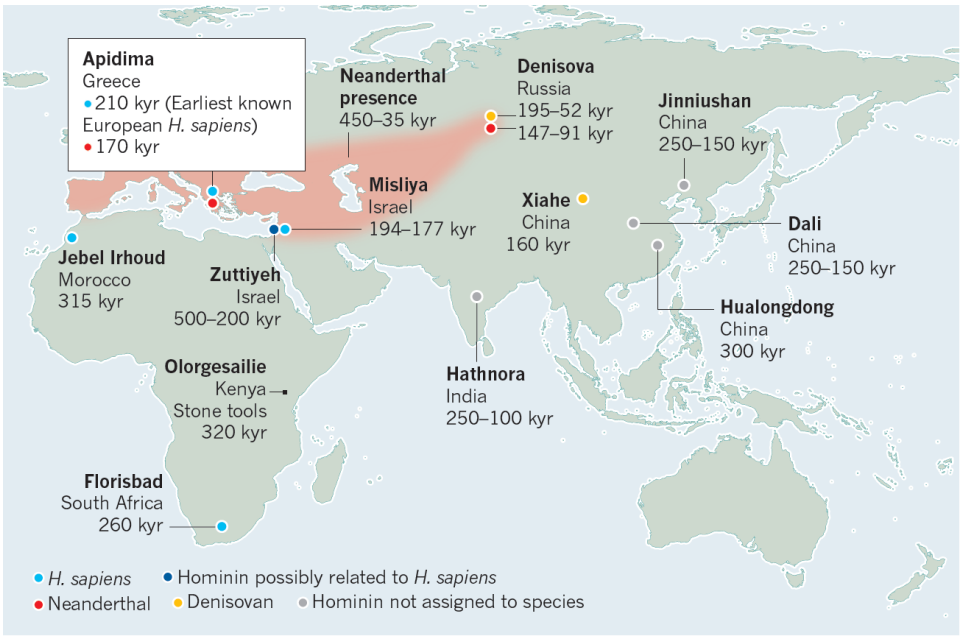


Figure 14. Some key early fossils of Homo sapiens and related species in Africa and Eurasia. Image taken from Delson, 2019.

However, these fossils do not show yet the entire physical kit of anatomically modern humans, in other words "us". On the contrary, they represent early forms of *Homo sapiens*. The first fossils to show the complete plethora of our features, *"a high neurocranium, rounded in lateral profile, a small face retracted under the frontal bone, a true chin even in infants, small discontinuous supraorbital tori, a lengthened post-natal growth period and life history, and a narrow trunk and pelvis with short superior pubic rami"* (Stringer, 2016) were found in Ethiopia. These archaeological remains of *Homo sapiens* in Africa are Omo Kibish 1, located near the Omo river, and Herto 1 and 2, found at Herto Bouri. The fossils, dated back to 195,000 years ago and between 160,000 and 154,000 years ago respectively, represent some of the earliest coalescences of most of the traits we associate with our extant species (Stringer, 2016).

Many other times Out of Africa

Africa, albeit our birthplace, started to appear too small for us. As soon as the climate changed around 160,000 years ago, turning many grasslands into inhospitable deserts, we run again.

Very recently, scientists unearthed some fossils from Apidima Cave, in southern Greece, testifying the presence of early modern humans at least 210,000 years ago, pushing back the time of our first wanderings out of Africa, as modern humans (Figure 14). Thus, the fossil — called Apidima 1 — is the earliest known example of *Homo sapiens* in Europe (Harvati *et al.,* 2019; Delson, 2019).

As we will more extensively see in the part "Archaic tales: Neanderthals and us" (specifically, at page 60), modern humans — both early and later forms — coexisted with Neanderthals at some Eurasian sites. And 210,000 years ago is a period fitting very well within the range of Neanderthals in Europe, thus suggesting that the two different species were probably living nearby each other. Also the Levantine material from Skhul and Qafzeh (dated back to 100,000 and 135,000 years ago and between 90,000 and 120,000 years ago, respectively, Grün, R., 2006) indicate that modern humans could have inhabited outside Africa very close to Neanderthals.

These coexistences at different times might suggest that the first wanderings out of Africa were many and tentative: the Apidima fossils show that modern humans, on more than one occasion, were pushing north and westwards from Africa and the Levant into Europe, but, probably, they were unable to compete successfully with Neanderthals (Delson, 2019).

As suggested by Eric Delson, these early modern human fossils could represent "failed" dispersal from Africa, as they reached the Middle East and southeastern Europe, but did not persist in these regions.

> *Rather than a single exit of hominins from Africa to populate Eurasia, there must have been several dispersals, some of which did not result in permanent occupations by these hominins and their descendants (Delson, 2019).*

With these timid dispersals Out of Africa, modern humans went all the way to China — a long-distance population movement in Central Asia has been recently confirmed around 45,000 years ago (Zwyns *et al.,* 2019) — while, after the last dispersal between 70,000 and 60,000 years ago (Underhill & Kivisild, 2007), they started their colonising adventure across the world.

Anyway, the many departures from Africa followed two main routes: the northern and the southern coastal route (Figure 15).

Figure 15. Possible dispersal routes of anatomically modern humans from Africa to Asia and Australia according to Forster & Matsumura, 2005. Image taken from Forster & Matsumura, 2005.

The northern route was probably used for the earlier expansions, as testified by the Skhul and Qafzeh Levantine fossils. These individuals — or their ancestors — should have travelled thousands of kilometres going up the Nile valley along the Red Sea coasts, and then went Out of Africa through the Levantine corridor. According to climate records, around 90,000 years ago, a new glacial period gripped that area (Pope & Terrell, 2008), sealing the fates of these earlier explorers. Also genetics comes in support of the use of such route. For instance, Pagani and colleagues, by analysing a genome-wide dataset of modern Ethiopian and Egyptian haplotypes and after having masked them for their recent West Eurasian components, found that non-African haplotypes were more similar to Egyptian haplotypes than to Ethiopian haplotypes, thus supporting an Egyptian route (Pagani *et al.,* 2015). Another genetic evidence favouring a northern route is the presence of Neanderthal ancestry in all non-African modern-day individuals (see page 66) and the fact that Neanderthals remains have been found in the Levant but not yet in the Arabian Peninsula.

Conversely, the Southern route across the Bab el Mandeb strait had been traditionally supported by mtDNA studies (Soares *et al.,* 2012; Quintana-

Murci *et al.,* 1999; Torroni *et al.,* 2006). From that strait, they could have reached rapidly South East Asia and Oceania following the coasts. This trail could have become the preferred one after 90,000 years ago, when the Levantine corridor would have probably been impassable due to the ice age. In fact, as reported in the first chapter, the climate has always played a crucial role in persuading our ancestors to move slowly or to run away. Other instances are the Saharan pump — the successions between "wet" and "deserted" Sahara — or the big geological catastrophes, such as the eruption of Mount Toba (Northern Sumatra).

The great colonising adventure

The discovery that genetic variability decreases as a function of geographic distance from Africa tell us that the movements of our ancestors were, as above said, tentative. They started from a little group in Africa and then spread all across the globe by breaking off, time after time, other small groups, in a gradual sequence of bottleneck events.

However, technical innovations — the sewing needle! — together with the development of complex social behaviour and the improved ability to build settlements made our ancestors less sensitive to the environmental constraints, allowing them to reach every strip of land across the globe, no matter how inhospitable (Figure 16).

The modern advances in the sequencing and analysis techniques of both modern and ancient humans, together with the finer analyses of paleoclimate and vegetation distribution (Timmermann & Friedrich, 2016), are providing more and more clues about our past, allowing to retrace backwards the migratory waves that have shaped the distribution of modern-day populations and their genetic diversity. However, many ancient trails remain still uncertain.

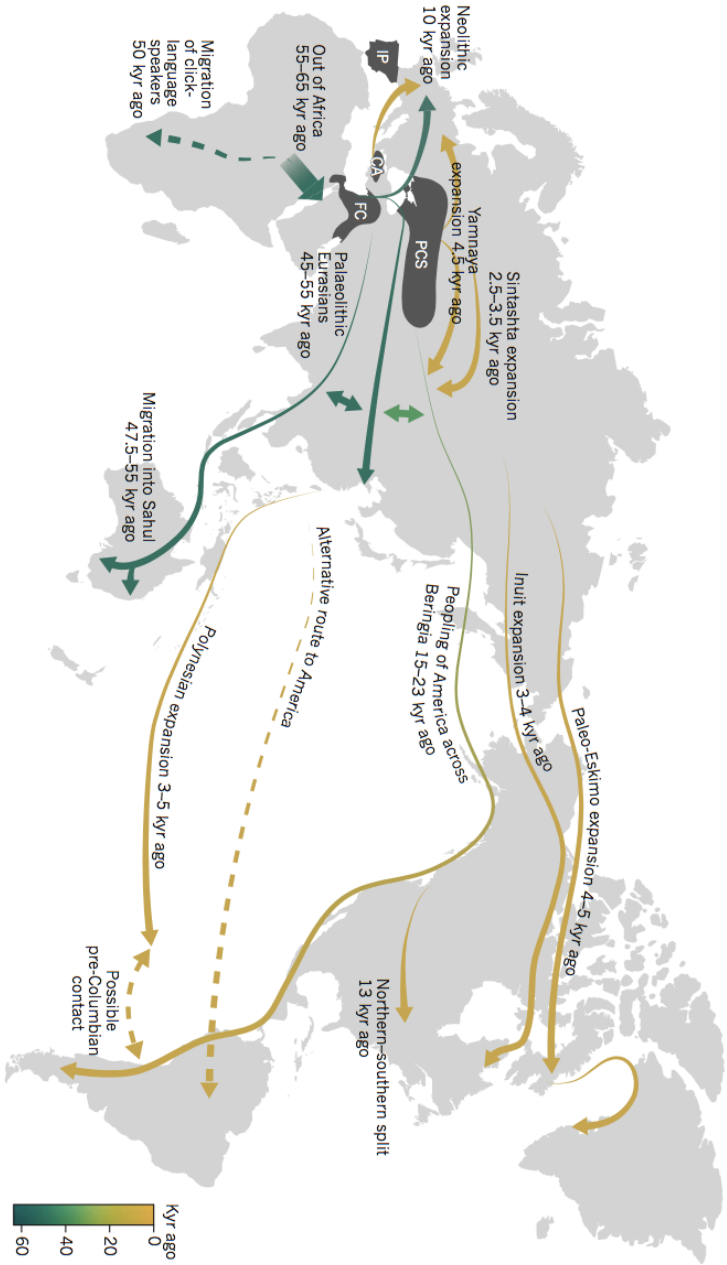


Figure 16. Major human migrations across the world inferred through the analyses of genomic data. Image taken from Nielsen *et al.,* 2017. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe.

For instance, it is still unclear whether the East — Asia and Oceania — was colonised through one or at least two waves of migration. Under the former hypothesis, East Asians could be the descendants of Southern migrants who moved toward the North, while Aboriginal Australians could have been diversified from these migrants. Conversely, the latter possibility could have happened through one wave which included the ancestors of Australasians and Papuan people. These individuals could have followed a Southern coastal route toward South Asia and Oceania during the Middle Pleistocene (from 50,000 to 100,000 years ago). The second wave could have been represented by the ancestors of East Asians, with possible admixtures between the two waves (Rasmussen *et al.,* 2011; Nielsen *et al.,* 2017). Interestingly, under the two-wave-model, some South Asian populations, such as the Negrito groups from the Andaman Islands, Philippines, Thailand and Indonesia, could carry the genetic traces of the first Southern dispersal (Figure 15). The recent discovery of 47 human teeth in the Fuyan Cave in Daoxian southern China), dated back to 80,000-120,000 years ago, seems to support an early migration toward the East. We can come to the same conclusion, thinking about the stone tools putatively associated with anatomically modern humans found in the volcanic ash layers left by the Toba eruption, thus meaning that humans could have reached South East Asia more than 74,000 years ago. Other archaeological evidence helping in tracing the first arrivals of anatomically modern humans in East Asia, as reported in (Lopez *et al.,* 2015), are the remains from Zhirendong (South China, Liu *et al.,* 2010), the teeth from Late Pleistocene Luna Cave (Bae *et al.,* 2014), the famous Southern Chinese Liujiang skeleton with seemingly anatomically modern features (Rosenberg, 2002), and the human foot bone of the Callao man discovered in Philippines, dating back to 67,000 years ago (Mijares *et al.,* 2010). Conversely, the archaeological remains from Oceania tell us that modern humans arrived there between 47,500 and 55,000 years ago (Clarkson *et al.,* 2015; O'Connell & Allen, 2015). One example is the modern human burial site at Lake Mungo (South East Australia, Figure 15) dating back to 40,000 years ago (Fitzsimmons *et al.,* 2014). For what concern the genetic evidence, Reich and colleagues analysed in 2011 the genetic variations across Asian and Oceanian populations. Their results, obtained through f-statistics, suggested a single wave Out of Africa followed by multiple dispersal waves peopling East Asia (Reich *et al.,* 2010). Another work, analysing a lock of hair from a 100-year-old Australian Aborigine and comparing its DNA with the worldwide genetic variation, pointed out toward an early split originating Australian Aborigines between 62,000 and 75,000 years ago.

However, once modern humans arrived at the Middle Eastern crossroad, not all of them followed the shores towards East. Some of them decided to go North, thus becoming the first European explorers. The first anatomically modern humans lived in Europe as early as 43,000 years ago. However, these people — called early Palaeolithic Europeans — are not our direct ancestors, in other words, they have genetically contributed very little to modern-day Europeans. This happened because Europe, in the following millennia, was crossed by many different populations coming from other far lands. A great contribution to our modern-day genetic diversity was given by the Neolithic people arriving around 10,000 years ago with their cattle and grain sacks and by herders coming from the Steppe during the Bronze Age. I will more extensively illustrate the migration waves peopling Europe at chapter "Who are the Europeans?".

The peopling of America had a completely different story, even if it is deeply connected with the Asian one. America was the last continent to be reached by modern humans between 23,000 and 15,000 years ago (Figure 16). I said that American and Asian fate were deeply connected because the only way to reach the "New World" was through Asia and, particularly, crossing the Bering land bridge. The first archaeological evidence for human settlements come from the Clovis culture, in a period after 13,000 years ago (Jenkins *et al.,* 2012). However, also in the case of America, climate condition had a huge role in guiding the movements of people and, consequently, the observation of its changes can help in disentangling the eventually different waves. Due to the presence of two large glaciers in present-day Canada and United State,s until 13,000 years ago, the gateway to the Southern parts of the continent was inaccessible. At the end of the glacial period, the ice melted, increasing the sea-level and covering the Bering passage. At the same time, the melting of the ice freed a corridor connecting the North to the unexplored Southern lands. Whatever was their trails, these "Paleo-Americans" started their descent through the Americas.

A very busy world

Over the last few millennia, we have grown so accustomed to be the only human species wandering through the globe, to think that that's the way it must have always been. But it was not.

When modern humans appeared in Africa, our planet was already inhabited by many different human species. Based only on archaeological evidence, in Africa there was *Homo naledi*, *Neanderthals* were living in Western

Eurasia, *Denisovans* in Eastern Eurasia, the last groups of *Homo erectus* were wandering in the Southern parts of Eurasia, while on a remote Indonesian island lived *"a little people, about half our height, and smaller than the bearded Dwarves"* [1] — *Homo floresiensis*, which due to its short stature, was nicknamed Hobbit. And who knows how many other species of which we have no physical remains might have been living close to our ancestors. We know, for sure, that on their way Out of Africa, modern humans have met some of these other human forms (Figure 17) and, for some of them, we also know what happened after having met.

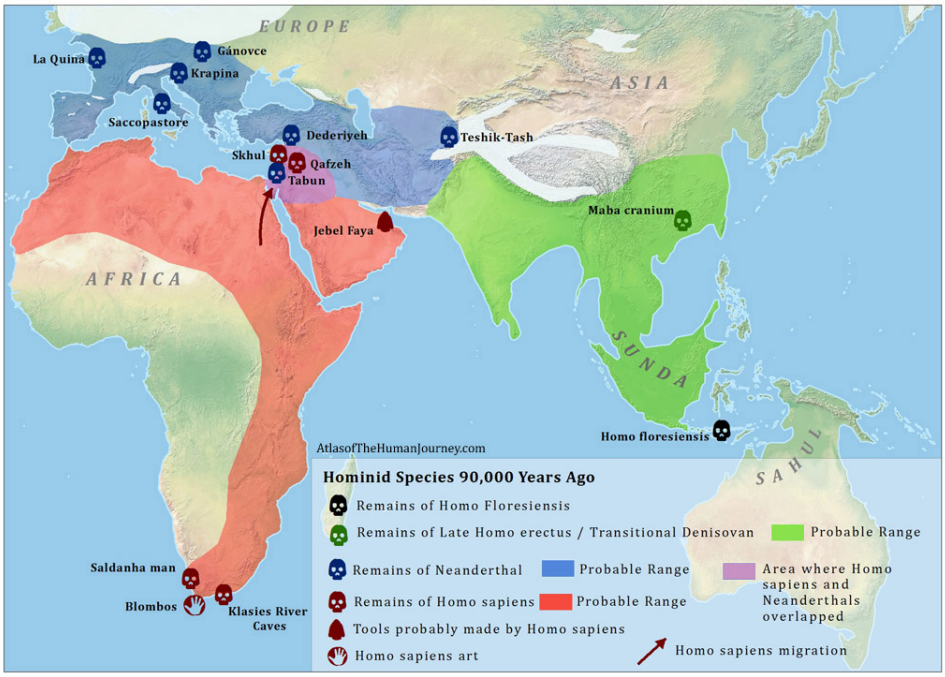


Figure 17. The different human species peopling Eurasia and Africa 90,000 years ago. Image credit: AtlasOfTheHumanJourney.com

Our modern human ancestors admixed with some of them. We are relatively sure about the archaic affairs they had with Neanderthals and Denisovans because both fossils and archaic DNA of these human species have been analysed (see chapter "Do Neanderthals exist today?" and, more specifically, page 66). Then, the archaic DNA was also compared with the DNA of modern-day populations (Green *et al.,* 2006; Prufer *et al.,* 2017;

[1] A description of the Hobbit people made by J.R.R. Tolkien in "The Hobbit".

Reich *et al.,* 2010), finally providing the principal piece of evidence towards admixture. However, in the last few years, it has become possible to detect archaic admixture even without the availability of an archaic reference genome (Browning *et al.,* 2018; Skov *et al.,* 2018). I will go much more into the details of some of the methods in the chapter "Do Neanderthals exist today?" (precisely, at page 72). However, we can just say that traces of admixtures with archaic "ghost" populations have been found in Eurasia and Africa, so far. As deeply explained in Santander *et al.,* 2019, the lack of archaic DNA is an urging problem above all in Africa, where the climatic conditions prevent the possibility to extract good quality DNA. However, thanks to these methods, in the absence of an archaeological site where to dig out aDNA from archaic bones, we can dig traces of admixtures from long ago in the genome of present-day people.

Whatever happened thousands of years ago, we are the sole survivors. Not long after our arrival to every corner of the world, the other human forms disappeared, leaving behind them a few bones and some DNA chunks in our genome. It is still unclear who is responsible for what happened to them; however, we surely know that our migrations have left permanent scars on our planet. A few thousand years after our arrival in Australia, twenty-three of the twenty-four Australian animal species weighing fifty kilograms or more became extinct (Miller *et al.,* 1999; Roberts *et al.,* 2001; Harari, 2015). The same happened in America: within 2000 years after the modern humans' arrival, North America lost thirty-four out of its forty-seven genera of large mammals, while South America lost fifty out of sixty (Harari, 2015). It is not difficult to think that we might also have driven the other human species to extinction as we came in contact with them (see page 80 for the possible reasons explaining Neanderthal extinction).

A synthesis for the descent of men

For a long time, the debate around our origins was focused on the opposition between multiregionalists and supporters of the Out of Africa model. Now we know that, as is often the case with science, none of them was right.

From the analyses of genetic data, it is crystal-clear that non-African modern-day populations had their ancestors in a group of African individuals who spread all around the world. However, it is also true that as these people continued on their wanderings, they met and admix with different human species at different rates. This means that, to simplify, while the ancestors of modern-day European populations admixed preferentially with

Neanderthals, the ancestors of East Asian, South-East Asian and Oceanian populations interbred with Denisovan. For this reason, none of us, modern humans, is a "pure" Sapiens, because while modern Europeans carry Neanderthal DNA portions, Asian and Oceanian people also have Denisovan DNA and, very probably, also modern Africans bear the genetic traces of the ancestral affairs with archaic human forms.

At this point, it is clear that neither the multiregional model, claiming that the origin of modern humans in a region strictly depends from the archaic forms inhabiting that same place, nor the Out of Africa, supporting a "pure" African origin for all modern humans, could describe the complexity of our past.

Thus, the new model — and the currently accepted one until new evidence will knock it down — is a mixture of the two and can be called, as proposed by David Reich (Reich, 2018), the "mostly out of Africa" model.

Migrations at historical times

The major challenge in bringing back to light the trails followed by our ancestors is that, once they crossed a river, overcame a mountain, reached a new continent, nothing prevented them or their descendants from going back again. Moreover, the same fate awaited their descendant populations: they could have come back or gone any further or both, thus covering the previous traces. And so many people walked over the same footprints!

This is true because, as we have extensively seen in this brief and necessarily incomplete reconstruction of ancient trails, people never stopped to move. They were continuing to migrate towards new territories, using new tools, by means of different vehicles. However, as they went further discovering always new territories, they became unaware of the lands and the people they left behind. It may seem trivial, but before the 16th century C.E., none of our ancestors knew how many continents there were on the globe. Europeans knew the existence of Africa, of the Middle East and suspected that somewhere in the East there were other endless and marvellous lands — Asia. They knew it and in an attempt to reach them, they ran into the Americas and Oceania.

After this moment, history repeated itself: Europeans arrived into the new marvellous lands — in this case, Americas — and rediscovered them, by leaving deep marks on those lands and the other Sapiens living there. This "Out of Europe", together with the slavery migrations, resulted in the fact that modern-day American populations carry DNA chunks inherited by

mainly three different ancestries: Native Americans, Europeans (for most Spanish) and Western Africans, reflecting the three major contributions arriving into Americas 15,000 the first and 500 years ago the second and the third.

In a few hundred years all continents would have been rediscovered again, so that since around 250 years ago it became almost impossible to find unexplored lands where to live lonely and far from other human beings. Human beings got everywhere and, as a consequence of this, linguistic or genetic isolation is no more possible (Pievani, T. and Calzolaio, V., 2016).

If the 16th century was characterised by enormous migratory waves forced by slavery from one side and desire to conquer from the other, 300 years later, during the industrialisation process, thousands of migrant workers were starting to move away from their countries to seek fortune far from home, reaching the nearest town or crossing the ocean. Labour migrations are happening still now, so much that entire continents such as Europe and America need immigrants from other continents to guarantee a sufficient number of residents and workers (Pievani, T. and Calzolaio, V., 2016).

In the 19th century, more and more people started to seek fortune outside their own countries, thus making frontiers and border management more and more relevant in politics. In an attempt to control and curb the migration flows, inevitably, they restricted individual freedom: the freedom of movement.

In the Global Village where we live now, the legal freedom of migration for everyone has been established, even if people aren't even aware of it. Article 13 of the United Nations' 1948 Universal Declaration of Human Rights states the following:

> *"Everyone has the right to freedom of movement and residence within the borders of each state. Everyone has the right to leave any country, including his own, and to return to his country."*

There couldn't have been a better way to end this introduction on human migrations.

Archaic tales: Neanderthals and us

"They left behind some bones, stone tools, a few genes in our DNA and a lot of unanswered questions. They also left behind us, Homo sapiens, the last human species."

Yuval Noah Harari, *Sapiens*

Do Neanderthals exist today?

For over 150 years, scientists from all over the world tried to answer the question "Do Neanderthals exist today?". After all this time searching for clues, it looks like we have our answer. The answer is: not exactly.

Actually, Neanderthals went extinct as recently as 30,000 years ago; however, everyone reading this thesis, unless of purely African descent, has a high probability of carrying Neanderthal DNA.

In order to understand this astonishing revelation, we need to look backwards in time retracing the routes that some archaic hominins took hundreds of thousands of years ago.

A son of Europe

It is not easy to precisely point out when Neanderthals emerged but we have no doubts in saying where: Europe.

Morphological evidence from fossils suggests the Neanderthal features were already present in Europe around 400,000 years ago (Weaver, 2009). The gradual emergence of the Neanderthal features indicates what is called "accretion model", in which the physical ingredients that made a Neanderthal man were added through a drawn-out process.

Why did the "neanderthalization" take place? Some features might be

the result of selective pressures, such as the shortening of the limbs as an adaptation to the cold European climate. However, it is really hard for us to explain other characteristics, such as their elongated skulls and protruding faces, as adaptation processes. In fact, these features might have a genetic drift explanation. When *heidelbergensis* arrived into Europe around 500,000 years ago, they stood in front of an enormous territory, barely inhabited by other humans. After some time, few thousands of individuals had spread through the entire area of the continent. Moreover, major climate changes during the Middle Pleistocene (between 781,000 and 126,000 years ago) prevented gene flows between Europe and Africa and between Europe and Asia, thus sentencing *heidelbergensis* to live in isolation for hundreds of thousands of years. In this situation, genetic drift took place, randomly sampling some features — some alleles — and eliminating others.

Whatever the explanation, for almost 300,000 years, Neanderthals characteristics were piling up, until 130,000 years ago, when they were fully established (Weaver, 2009).

As Silvana Condemi and François Savatier highlight in their book (Condemi & Savatier, 2016), the Neanderthal story is completely European for at least three reasons: i) pre-Neanderthal fossils have been found only in Europe; ii) the evolutionary process leading to Neanderthals took place only in Europe; iii) whereas Neanderthals also inhabited Asia, only in Europe a huge amount of Neanderthal remains have been found. Moreover, the absence of pre-Neanderthal fossils outside Europe suggests Neanderthal started to colonise Central Asia and Near East after the acquisition of their characterising features. However, it is true that this European monopoly could, in some way, be biased toward Western Europe. This area, although is only one-fifth the size of their further expansion, contains almost three-quarters of all Neanderthal sites. This is due both to the better preservation of fossils in limestone caves and rocks of France, Spain and Italy and to the bigger fieldworks done in this part of Europe (Roebroeks & Soressi, 2016).

Anyway, every way you look at it, Neanderthals were already present in Europe a long time before Sapiens arrived there and even before they appeared on Earth!

It is really amusing to think now, in this period of heated discussions about the rights to come in and remain into Europe, that a few thousand years ago it was actually us the black-skinned immigrants coming to the lands belonging to the first true Europeans: the Neanderthals.

Neanderthal home

Neanderthals were the undisputed ruler of large part of Europe and West Asia for almost 300,000 years, going as far as the Near East and Central Asia (Figure 18), thus covering an area of around 10 millions km^2, which is much larger than Australia (Roebroeks & Soressi, 2016).

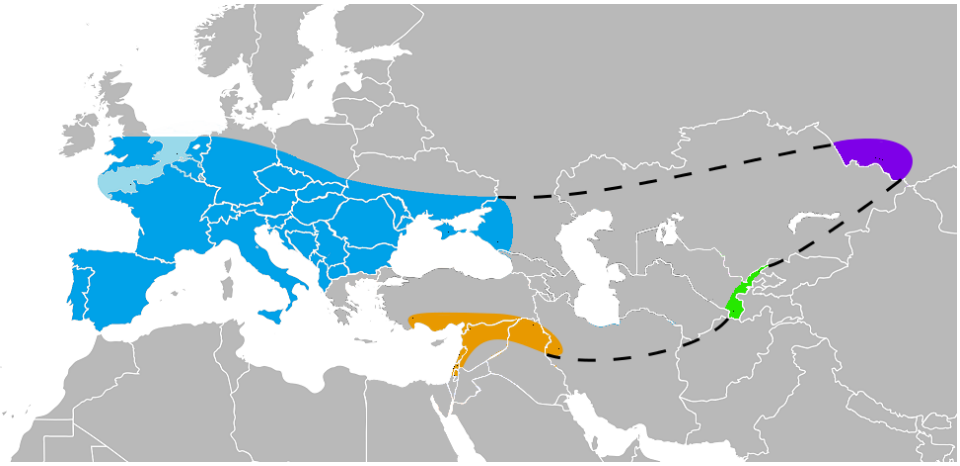


Figure 18. The known territory range of Neanderthals based on fossil records. Image credit: Nilenbert, Nicolas Perrault III.

As for their northern limits, no undoubted Neanderthal sites are located from areas above 53° in Eurasia. They also moved East, arriving into East Europe around 90,000 years ago, as demonstrated by the fossils found at Stajna, Polland (Urbanowski *et al.,* 2010) and going further to Central Asia. This spread happened during a temperate season (between 123,000 and 109,000 years ago): because of the heat, Caspian sea dried out, opening a gate toward the East, as shown by the remains of a young Neanderthal man found in Uzbekistan (at the Teshik-Tash site). Other fossils have been brought to light in Southern Siberia, at the Denisova cave, where we shall come back later in the story.

Their southern expansion stopped in the Near East, since no Neanderthal skeletal remains are known from Africa, and fossils from Near East, Iraq, Syria and Israel suggest that this expansion took place during the same previously mentioned interglacial period.

As for other humans, Neanderthal expansion was mainly guided by climate changes, which were so characteristic for Pleistocene in Eurasia. Glacial-interglacial cycles might have deeply shaped their demography, break-

ing up a population, according to paleo-demographist, of around 70,000 individuals (Bocquet-Appel & Degioanni, 2013).

They were few and isolated, but this would not last long: strangers were about to come.

Neanderthal looks: why the long face?

When some quarrymen in 1856 discovered in a cave in the Neander Valley (Düsseldorf) bones of a cranial vault (Figure 19), they thought them belonging to a cave bear. When some years later Rudols Virchow, a world-renowned German anatomy professor, saw those bones he thought they were the remains of a diseased man.

Figure 19. First Neanderthal fossil ever found. Original bones of *Homo neanderthalensis*, displayed at LVR-LandesMuseum Bonn.

Then, after some work and many discussions, these bones have been attributed to a primitive species other than *Homo sapiens*, which was called *Homo neanderthalensis*. Fun fact, the valley was named after Joachim Neander, a 17th century pastor and composer, whose name was simply a Greek translation of his real family name: Neumann, meaning "new man". Thus, this old extinct hominin is a fully-fledged "new man".

During the following years until recently, palaeoanthropological and genetics researches on his fossils have contributed in smoothing the rough portrait we unfairly did of him.

Actually, Neanderthals were not so physically different from us (Figure 20). They were large-bodied, muscular and robust people with short stature and a "barrel-shaped" chest, which could be a potential physical adaptation to the cold Eurasian environment and the so derived higher energetic needs (Heyes & MacDonald, 2015; Churchill, 2014).

The features which would make everybody turn around are mostly in the head area. First of all, they were large-brained hominins, with a long and flat skull, a strong midfacial prognathism and a pronounced brow ridge.

However, as the authors of "Néandertal mon frère" (Condemi & Savatier, 2016) point out, if we met a Neanderthal on the subway, we would not be shocked, as the 7 billion people of the human population seem enough to include the entire phenotypic variability of the Neanderthal population. Basically, we will never see a Neanderthal in one piece, but we surely have seen him apart!

Figure 20. Neanderthal woman. Portrait of a female Neanderthal from fossils reconstruction. Image credit: Atelier Daynès

Thinking like a Neanderthal

If we had some preconceptions about their hairy and beastly physical looks, these prejudices were even higher when we consider their culture and way of life. In fact, the most recent archaeological findings (Jaubert *et al.,* 2016; Beier *et al.,* 2018; Hoffmann *et al.,* 2018) are starting to recast the conventional thinking about Neanderthals, turning them from distantly related brutish creatures into more sophisticated cousins.

For a start, they probably could talk. Unfortunately, it is now impossible to demonstrate that they used an articulated language; however some studies are going in that direction. For example, Neanderthals had the hyoid bone (Bar-Yosef & Vandermeersch, 1991), which is the only bone in the vocal tract needed for mastication, swallowing and voice production. They possessed the Wernicke and Broca area, which are primarily involved in speech comprehension and production, respectively. They also carried the sapiens version of the *FOXP2* gene (Krause *et al.,* 2007), whose role is connected to the development of cerebral regions important for language learning (Hurst *et al.,* 1990). However, they were different in one position of a *FOXP2* intron, affecting a conserved binding site for the transcription factor POU3F2 (Figure 21). Functional assays have demonstrated that the Neanderthal variant makes the transcription factor less effective (Maricic *et al.,* 2013).

Hyoid bone

U-shaped bone situated at the root of the tongue in the front of the neck and between the lower jaw and the largest cartilage of the larynx, or voice box. The primary function of the hyoid bone is to serve as an anchoring structure for the tongue. It has no articulation with other bones. *Encyclopaedia Britannica*

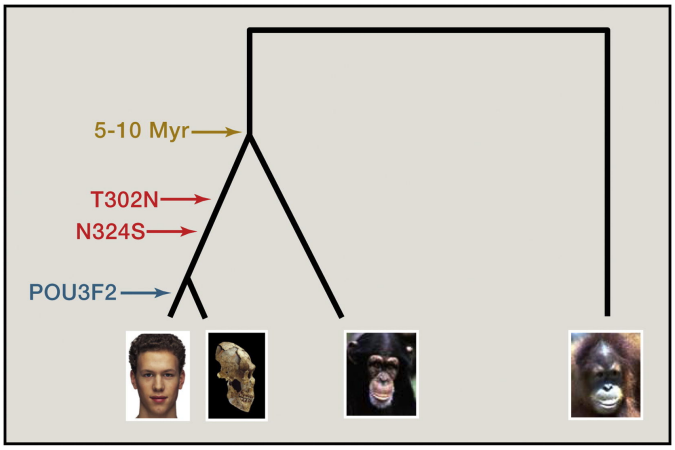


Figure 21. *FOXP2* mutations. Schematic illustration of the two amino-acid substitutions (red) and the change in a POU3F2-binding site (green) that have affected the *FOXP2* gene during human evolution. Figure taken from Pääbo, S., 2014.

Secondly, other evidence point out that Neanderthals had complex cognition: for example, they collected marine shellfish, they used iron oxides (Roebroeks *et al.,* 2012) as pigments for personal decoration as well as raptor claws (Figure 22, Finlayson *et al.,* 2012; Radovčić *et al.,* 2015) and bird feathers (Finlayson *et al.,* 2012; Peresani *et al.,* 2011), not to mention the first underground structures (Jaubert *et al.,* 2016) and the cave arts from Spain which U-Th (Uranium–thorium) dates revealed that they date back before the arrival of modern humans, thus implying a Neanderthal authorship (Hoffmann *et al.,* 2018).

Third, late Neanderthals in Europe buried their dead (Rendu *et al.,* 2014) already 10,000 years before the arrival of modern humans and took care of the weakest members in their groups. We need only think of the old man from La Chapelle-aux-Saints (Rendu *et al.,* 2014), who suffered serious degenerative diseases, a badly healed rib and a hip injury. Or the fifty years old Neanderthal man from Shanidar (Iraq): deaf, one-eyed, with so debilitating arthrosis that he could not walk and a withered arm. He would not be able to survive without help from his clan.

Figure 22. Eight eagle talons from a Neanderthal site. Eight eagle talons from a Neanderthal site (Krapina cave) in present-day Croatia (Radovčić *et al.,* 2015). This necklace has been produced by Neanderthals 120,000 years ago.

This evidence clearly suggests that they had a rich and complex culture, surely comparable to the sophistication of our modern humans' ancestors.

However, for around 250,000 years, they manufactured their cut stones in the same way: the Mousterian style.

Mousterian style

Tool culture traditionally associated with Neanderthal man in Europe, western Asia, and northern Africa during the early Fourth (Würm) Glacial Period (c. 40,000 BC). Mousterian implements disappeared abruptly from Europe with the passing of Neanderthal man.
Encyclopaedia Britannica

This relative lack of innovation in the lithic industry goes well with their rigid culture: during unfavourable weather periods, they tended to retreat into themselves, always using the same usages and customs, with small mating groups and avoiding large-scale migrations. According to Condemi and Savatier (Condemi & Savatier, 2016), this strict and traditional behaviour was one of their strongest points necessary for surviving for so many years in a Eurasia plagued by extreme climatic conditions. Every innovation could have been dangerous and could have been the last.

Wandering, learning and... meeting?

Genetics gives us solid proofs about the Palaeolithic meetings our Sapiens ancestors had with the Neanderthals and we will see more about this later on. However, there is also other evidence pointing out in that direction.

The most direct one is at Fumane, in northern Italy where around 44,000 years old Neanderthals were making stone tools that appear as the predecessor of modern human ones (Benazzi *et al.,* 2014). Moreover, in southwestern Europe (Grotte du Renne, France) tools typical of modern humans, made in the Châtelperronian style, have been found together with Neanderthal bones dating between 44,000 and 39,000 years ago, raising a tantalising scenario where Neanderthals learnt the modern human toolmaking. However, this debate is the subject of a still heated controversy (Gravina *et al.,* 2018; Bar-Yosef & Bordes, 2010; Higham *et al.,* 2010).

Châtelperronian style

The Châtelperronian material culture represents the earliest sign of the Upper Palaeolithic in Europe and its products span a period from about 45 to 40 ka. …The best have been found in the Grotte du Renne in eastern France.
wileyearthpages.wordpress

A possible meeting in the Near East is even more certain. In the Eastern Mediterranean regions both Neanderthal (Shanidar Cave, Iraq and the Amud, Kebara and Tabun sites in Israel) and modern human (Qafzeh and Skhul sites in Israel) remains have been found. The Neanderthal fossils at Tabun cave have been dated to a period when modern humans were already present in that area. In particular, a female Neanderthal from Tabun and other bones from there date back to 120,000 and 90,000 (Grün, R. and Stringer, C., 2000; Coppa *et al.,* 2007), while three burials of modern humans from Skhul and Qafzeh belonged to a period between 100,000 and 135,000 years ago and between 90,000 and 120,000 years ago, respectively (Grün, R., 2006). As Tabun and Skhul are neighbouring caves on the slopes of Mount Carmel, it seems unlikely that they did not meet, even just for a

coffee!

According to these findings, modern humans were already out of the scorching Africa 130,000 years ago; however, between 60,000 and 40,000 years ago, another wave of migration arrived from Africa in the Near East. At this point, something different happened: after 250,000 years of having been so traditional in their toolmaking techniques, Neanderthals suddenly adopted the sapiens' way. Isn't that a bit of a coincidence?

Wandering, learning, meeting and... kissing?

According to the findings described above, modern humans run into the Neanderthals during their Out-of-Africa wanderings. However, in all that chitchat, did they mix?

Answering to this question using only palaeoanthropological and archaeological evidence is hard work; however some clues have piled up.

In 2002 in the Peştera cu Oase (Romania) a modern human mandible, called Oase 1, was found and radiocarbon dating indicates it dates back to a period between 37,000 and 42,000 years (Trinkaus *et al.,* 2003). Interestingly, this mandible, as well as the other bones found in the same cave (individuals Oase 2 and Oase 3), exhibits a curious mix of modern and archaic features, as if it belonged to a Neanderthal-Sapiens hybrid.

These remains are not alone: a group of bones discovered in 1952 in the Cave of Peştera Muierii (Romania) and the skeleton of a young boy found in the Old Mill Rock Shelter (Lagar Velho, Portugal) show a miscellany of modern and archaic characteristics too.

However, the evidence coming from the bones could not demonstrate the relatedness of Neanderthals to us, since sometimes shared skeletal features could simply reflect adaptation to the same environment and not necessarily an inbreeding event (Reich, 2018). Moreover, resting solely on fossils, it is often not possible to recognise known hybrids among living animals, not to mention that we do not clearly know how the bones of a hybrid could appear (Harris, E. E., 2015).

They needed the smoking gun: DNA.

The Neanderthal genomic wave

The fascinating journey back in time surfing the Neanderthal genomic wave started a night in 1996. In his book — "Neanderthal man: in search of lost genome" (Pääbo, S., 2015) — Svante Pääbo described the moments of

excitement and doubts when a little piece of mitochondrial DNA, 61 nucleotides long, was extracted and analysed from a Neanderthal bone dating back to 40,000 years ago (the very first fossil found in 1856 in the Neander Valley, Figure 19, Krings *et al.,* 1997).

From then on, the techniques for working with ancient DNA have been improving and in 2008 Green and colleagues (Green *et al.,* 2008) obtained the entire mitochondrial genome of a 38,000-year-old Neanderthal from Vindija Cave (Croatia).

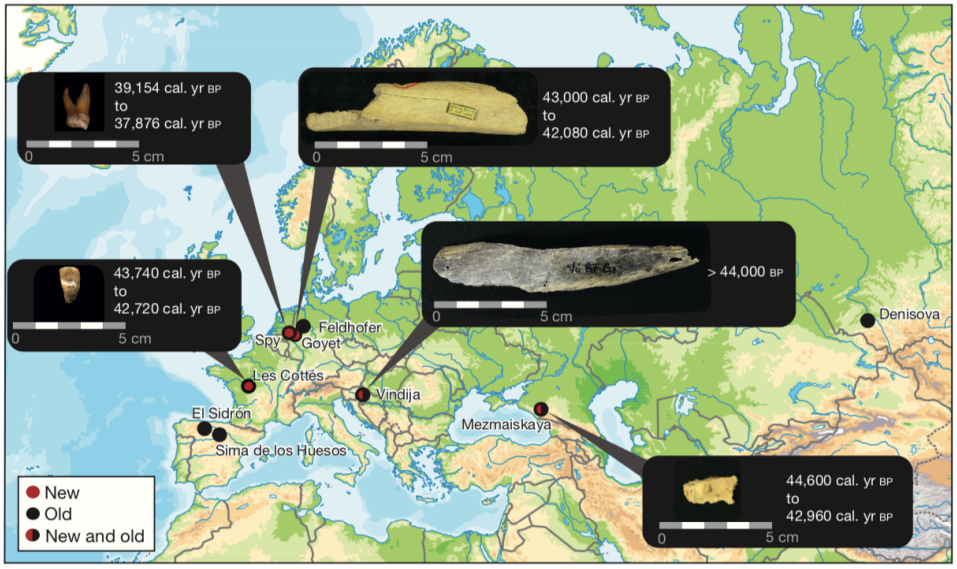


Figure 23. Five late Neanderthals analysed in Hajdinjak *et al.,* 2018. Location and age of the five late Neanderthal specimens analysed in that study (new), and other sites for which genome-wide data of Neanderthal samples have previously been published (old). Image taken from Hajdinjak *et al.,* 2018.

In the meantime, scientists were beginning to obtain also nuclear DNA from Neanderthals (Noonan *et al.,* 2006; Green *et al.,* 2006), culminating in the first draft of the entire genomes of even two extinct hominins: Neanderthals (Green *et al.,* 2010), coming from three Neanderthals from Vindija Cave (∼1.2-fold total coverage) and Denisovan (Reich *et al.,* 2010) from the Denisova Cave (∼1.9-fold coverage). However, these sequences were obtained through a slightly inefficient method, as it was based on double-stranded DNA. The years went by and the techniques improved: a new

method based on single strands was developed, allowing to increase the recovery of ancient DNA. Through this technical improvement, it was possible to obtain in 2012 a new version, of higher quality, of the Denisovan genome (Meyer *et al.,* 2012) and in 2014 a high-quality (∼50-fold coverage) genome of another Neanderthal individual, this time coming from the Altai mountains (Prufer *et al.,* 2014), and a Neanderthal genome from Mezmaiskaya Cave in the Caucasus (∼0.5-fold coverage).

As time goes by, new samples and new sequences have been produced: a ∼30-fold coverage DNA from Vindija Cave (Prufer *et al.,* 2017) and five late Neanderthals coming from different sites in Eurasia (Figure 23, Hajdinjak *et al.,* 2018), thus getting the number of complete Neanderthal genomes to go up to about ten. Moreover, exome sequences from a different individual from Vindija Cave and one from Sidron Cave in Spain have also been generated (Castellano *et al.,* 2014).

Figure 24. Sites were partial to complete nuclear Neanderthal genomes were retrieved. Image taken from Peyrégne *et al.,* 2019 reporting the sites from which partial to complete nuclear genomes from Neanderthals (or their ancestors in Sima de los Huesos) were retrieved. The origins of the two Neanderthals studied in Peyrégne *et al.,* 2019 are highlighted in purple and blue, respectively.

The latest addition to this list is the low-coverage nuclear DNA from two early German Neanderthal fossils (Peyrégne *et al.,* 2019), from which only mitochondrial DNA was previously recovered: HST, a femur from

Hohlenstein-Stadel Cave and Scladina and a maxillary bone from Scladina Cave (Figure 24).

This priceless amount of genetic data are allowing to exhaustively study the Neanderthal population across their temporal and geographic range and, above all, to start answering questions about the promiscuous affairs our ancestors had with them.

Mitochondrial attempts

Supporters of the Neanderthal-Sapiens admixture hypothesis were deeply disappointed by the results of the inquiries relied on mitochondrial DNA. The analyses of the first Neanderthal genetic sequence (Krings *et al.,* 1997), the hypervariable part of the mtDNA control region, lead to a fairly drastic conclusion: *"Neanderthals went extinct without contributing mtDNA to modern humans"* (Krings *et al.,* 1997). And it could not be otherwise, because when they compared the Neanderthal DNA with human mtDNAs, the Neanderthal sequence fell completely outside the range of present-day human variation. Actually, human sequences differ among themselves by an average of 8.0 ± 3.1 (range 1-24) substitutions, Neanderthal and modern human sequences by 27.2 ± 2.2 (range 22-36), while human and chimpanzee sequences show around 55.0 ± 3.0 (range 46–67) differences. Using these rates, they estimated that the common ancestor of Neanderthal and modern humans lived approximately 550,000 to 690,000 years ago, which is about four times older than the time when "Mitochondrial Eve" lived.

Since then, other Neanderthal mtDNA sequences have been released and analysed (Green *et al.,* 2008), pushing more recently the maternal-line ancestors the Neanderthals shared with us (470,000-360,000 years ago, Posth *et al.,* 2017), but all suggesting that modern humans replaced Neanderthals with no interbreeding.

Unfortunately, mtDNA is not the optimal starting material on which trying to answer such inbreeding questions, due to its maternal-only inheritance pattern. Even if modern humans sometimes interbred with Neanderthals, it is very likely that the women carrying Neanderthal mtDNA were not lucky enough to pass it down to the next generations through female descendants.

For this reason, a scenario where the eventual Neanderthal mtDNA gets lost during the millennia is not surprising at all: the mitochondrial data cannot be conclusive.

The long-awaited smoking gun

Thanks to the technical advancement in molecular biology and the introduction of sequencing machinery, a complete Neanderthal genome was already available in 2010 (Green *et al.,* 2010): scientists were finally holding their long-awaited smoking gun.

At this point, it was only a matter of developing a test to determine if archaic hominins, like the Neanderthals, interbred with modern humans. Researchers in the Reich group (Green *et al.,* 2010; Durand *et al.,* 2011) reasoned that if some human populations interbred with Neanderthals, these populations should share more DNA substitutions with Neanderthals compared to other populations that did not mix with them. Starting from this preamble, they developed the ABBA-BABA test, also called *D*-statistics (I will apply this method on a dataset comprising present-day and ancient human populations at page 153). The researchers examined only those DNA positions where individuals coming from different modern populations (e.g., African and non-African) carried different bases: one carried "A", while the other carried "B". In this case, "A" stands for Ancestral, meaning that it is shared with the chimpanzee. In other words, "A" has been inherited from our last common ancestor shared with the chimpanzee. Conversely, "B" refers to the Derived allele: this base is found in one of the two modern populations and in the archaic genome. At this point, we have a single position in four different populations (e.g., African, non-African, Neanderthal and Chimpanzee) and two possible configurations. The first, called ABBA, happens when the first population (let's say the African) carries the ancestral allele "A", as well as the outgroup population (chimpanzee), while the second population (non-African) shows the same derived allele "B" of Neanderthals. The second situation corresponds to the reverse BABA.

Using this model, Green and colleagues (Green *et al.,* 2010) defined a statistics computing the difference in the counts of ABBA and BABA configurations across the n DNA positions and normalizing by the total number of observations:

$$D(P_1, P_2, P_3, O) = \frac{\sum_{i=1}^{n} C_{ABBA}(i) - C_{BABA}(i)}{\sum_{i=1}^{n} C_{ABBA}(i) + C_{BABA}(i)}, \tag{1}$$

where P_1, P_2, P_3 and O indicate African, non-African, Neanderthal and Chimp individual, respectively.

If the Neanderthals did not interbreed with modern humans, then there will be an equal amount of ABBA and BABA sites. Conversely, if the ancestors of one of the two modern populations had some archaic affairs,

then that population will share more "B" with Neanderthals, resulting in different ABBA-BABA counts.

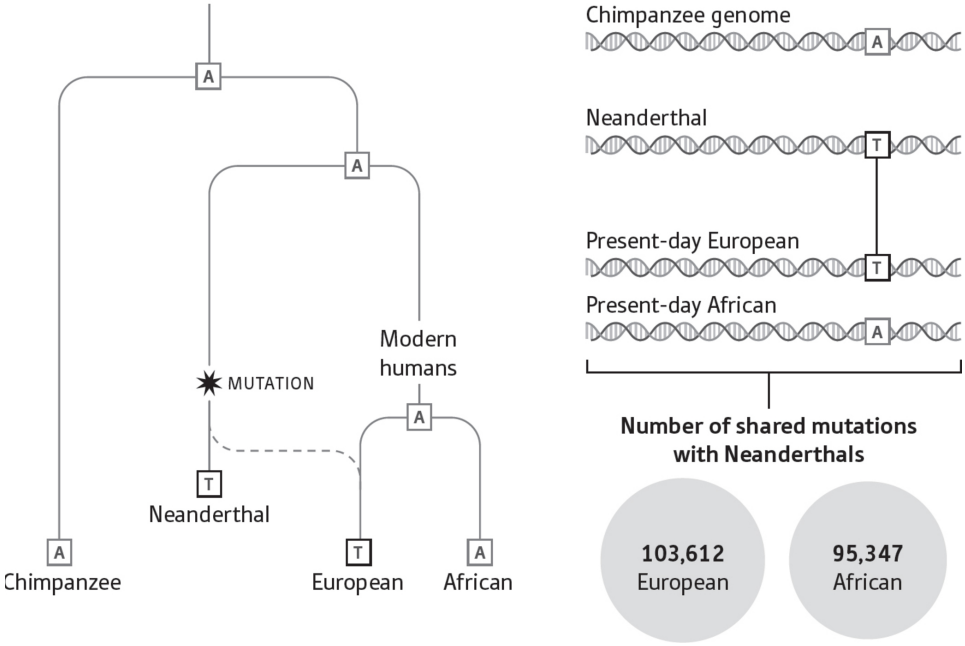


Figure 25. *D*-statistics tree. Image taken from Reich, 2018 reporting the ABBA-BABA test (also called *D*-statistics or four population test).

When Green and colleagues (Green *et al.,* 2010) applied this test on Papuan, French and Han individuals (Eurasians) and Yoruba and San individuals (Africans), they discovered that Eurasians shared more derived alleles with Neanderthals than African with Neanderthals themselves. In other words, the ABBA-BABA model highlighted a scenario where non-African individuals met and mixed with our archaic Eurasian hosts.

Moreover, they also estimated that up to 4% of the modern Eurasian genome has been inherited from Neanderthals, meaning that a little slice of some Sapiens genomes is not exactly Sapiens.

Digging for answers in the Neanderthal genome

The floodgate had been opened and, from then on, scientists from all over the world started riding the Neanderthal genomic wave. Thus, while Neander-

thal genetic data were growing in number and quality, it became possible to answer more and more questions about these ancient kisses.

In particular, the availability of a high-quality Neanderthal genome (Altai Neanderthal) allowed refining the estimate of Neanderthal ancestry in present-day populations to a value between 1.5% and 2.1% (Prufer *et al.,* 2014). However, contrary to what has been previously found, other studies demonstrated that some East Asian and South American populations contain Neanderthal ancestry proportions up to 24% more than Europeans (Meyer *et al.,* 2012; Wall *et al.,* 2013).

Then, when the high-coverage genome of the Vindija Neanderthal was analysed, they found that the European Neanderthals (Vindija, Prufer *et al.,* 2017 and Mezmaiskaya 1, Prufer *et al.,* 2014) were more closely related to the Neanderthal populations that mixed with modern humans, with respect to the Altai Neanderthal. Thus, the high-coverage European Neanderthal from Croatia allowed to further refine the proportions of Neanderthal introgression (Figure 26). Using that genome, Prufer and colleagues (Prufer *et al.,* 2017) found higher estimates than with the Altai in non-African populations outside Oceania, which carried between 1.8-2.6% Neanderthal DNA. They also found that East Asians showed more Neanderthal DNA (2.3-2.6%) than Western Eurasian populations (1.8-2.4%).

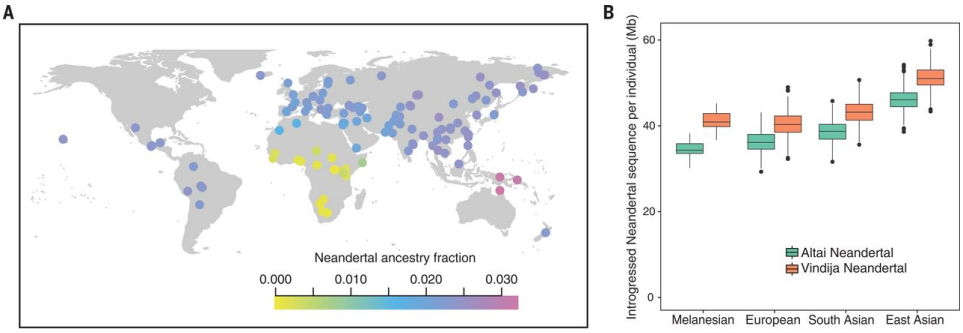


Figure 26. Fraction of Neanderthal DNA for present-day populations. Image taken from Prufer *et al.,* 2017 reporting the estimates of the fraction of Neanderthal DNA for present-day populations. **(A)** Colours indicate Neanderthal ancestry estimates. Oceanian populations show high estimates due to Denisovan ancestry that is difficult to distinguish from Neanderthal ancestry. **(B)** Amount of Neanderthal sequence in present-day Europeans, South Asians and East Asians.

The fact that Neanderthal ancestry is spread across all modern human

populations, with the exception of South Africans, raises some questions. Actually, a Neanderthal legacy is present in modern-day inhabitants of Asia, where a few Neanderthal fossils have been found and even in the people of Papua New Guinea, where definitely Neanderthal never arrived. A possible explanation of this pattern could be a scenario where our modern human ancestors met the Neanderthals before differentiating into Asians and Europeans, most probably in the Near East (Figure 27). In fact, both Neanderthal (Shanidar Cave in Iraq and the Amud, Kebara and Tabun sites in Israel) and modern human (Qafzeh and Skhul caves in Israel) fossils have been found in this region (see page 60). As the modern human fossils are as old as 100,000 years ago, this means that for about 50,000 years Neanderthals and Sapiens lived in the same place, potentially interbreeding.

Aurignacian style

Toolmaking industry and artistic tradition of Upper Palaeolithic Europe that followed the Mousterian industry. …The Aurignacian culture was marked by a great diversification and specialisation of tools, including the invention of the burin, or engraving tool, that made much of the art possible.
Encyclopaedia Britannica

Then, a second flux Out-of-Africa brought more and more Sapiens in that area: Svante Pääbo calls this second wave the "replacement crowd" (Pääbo, S., 2015). These individuals were in some ways more sophisticated: as an example, they had developed a new style of tool production, the Aurignacian.

Coming into the Near East, the replacing crowd may have absorbed the other Sapiens living there, which were already admixed with Neanderthals and, during their colonising adventure across the globe, they could have spread this archaic inheritance.

However, even if we accept an admixture in the Near East, how can we explain the discrepancy in Neanderthal ancestry between Europeans and East Asians?

Some explanations to this phenomenon have been proposed: it could be that admixture between Neanderthals and Sapiens happened with a single pulse, but demographic events, such as the admixture with the "Basal Eurasian" population or with northern African populations, could have diluted the Neanderthal ancestry in Europeans and above all in Southern Europe (Busby *et al.,* 2015). Alternatively, selective pressures could have subsequently shaped the frequencies of the Neanderthal inherited alleles, also depending on the ancestral effective population size: with this regard, Sankararaman and colleagues (Sankararaman *et al.,* 2014) reasoned that the lower effective population size of East Asian could have reduced the efficacy of purifying selection in purging deleterious alleles.

Basal Eurasians

A population that split off from other non-Africans before they split off from each other, around 101,000-67,000 years ago
(Lazaridis, 2018)

Alternatively, the admixtures could have occurred multiple times. In fact, the work of Villanea and Schraiber found strong support for a model of multiple admixture events (Villanea & Schraiber, 2019): in this model, the original pulse into the ancestral Eurasian population was followed by additional pulses to both European and East Asian populations, when they were

already split. Furthermore, also Vernot and colleagues found evidence of differential Neanderthal admixture between some European and East Asian populations (Vernot *et al.,* 2016).

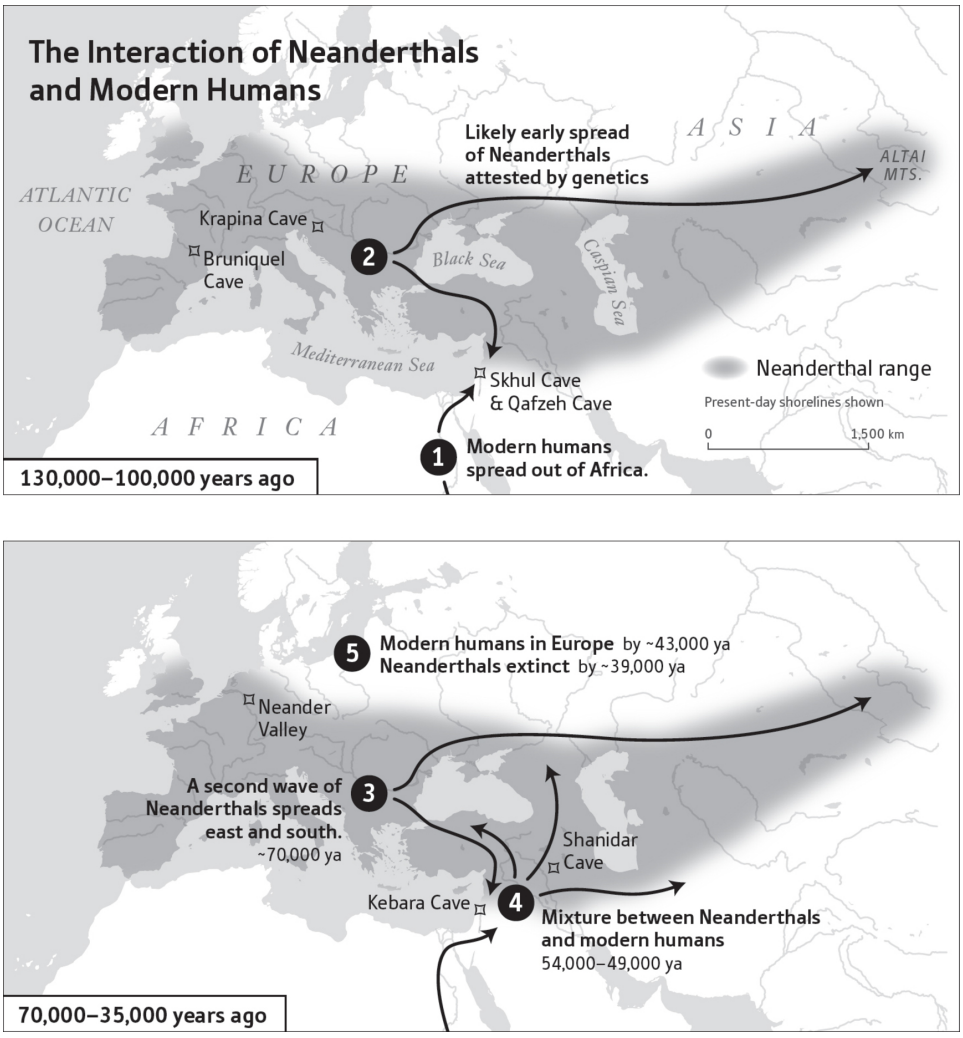


Figure 27. The meetings between Neanderthals and modern humans. Image taken from the book "Who we are and how we got here" (Reich, 2018) reporting the timeline of the at least two main meetings modern humans and Neanderthals had.

The availability of high-coverage archaic DNA allowed solving some ancient mysteries, such as European modern human fossils with also some

archaic features. If they could demonstrate that those fossils also contained Neanderthal inherited DNA, we might have an answer to another question: did interbreeding also happen in Europe?

When Pääbo's group sequenced DNA from the remains of Oase1, a modern human that Erik Trinkaus had interpreted as a Neanderthal-modern human hybrid, they discovered that its DNA contains a percentage of Neanderthal ancestry between 6 and 9 (Fu *et al.,* 2015). Moreover, its Neanderthal DNA legacy covers until a third of the length of his chromosomes, making them suggest that Oase1 had a Neanderthal ancestor no more than six generation before him.

This was an incredible discovery because it crushed all the objections against the episodes of interbreeding between different kinds of humans. However, it has also been demonstrated that Oase1 did not leave descendants among modern-day Europeans (Fu *et al.,* 2015). In other words, the population to which he belonged was a dead branch, thus, while Oase1 himself provides the actual smoking gun of interbreeding, he cannot demonstrate that the Neanderthal ancestry we see in non-African populations is actually derived also from European Neanderthals.

Do you remember when we first met?

The discoveries about the places where we met and mixed with Neanderthals beg another question: when did our ancestors admix with the Neanderthals?

Answering to this was important not also for the question per se, but also for the deep implication this question is bringing: at the first result of Green *et al.,* 2010 demonstrating the admixture, many scientists were not completely convinced about that conclusion. In fact, also a putative ancient population structure in Africa could explain the findings of Green *et al.,* 2010. The only way to get rid of this controversy was to date the interbreeding event: if this would be older than the time when first modern humans wandered out of Africa, then it cannot indicate an admixture event, but a signature of an "ancient African genetic structure".

In order to answer the questions, researchers took advantage of genetic recombination. Due to this process, the DNA segments inherited from Neanderthals have broken up into smaller pieces through the generations. Thus, it is possible to infer how many generations have passed since the admixture event, simply measuring the sizes of the Neanderthals inherited regions, which are DNA chunks matching the Neanderthal genome more than they do sub-Saharan African genomes. In other words, if two Neander-

thal alleles introgressed in modern humans at time t_{GF}, the probability that these two alleles will not be broken by the recombination process is proportional to $e^{-t_{GF}x}$, where x is the genetic distance separating the two alleles (Sankararaman *et al.,* 2012).

Using this strategy, Sankararaman and colleagues (Sankararaman *et al.,* 2012) found that some Neanderthal DNA entered the ancestors of modern-day non-African people around 37,000–86,000 years BP (Before Present). This date is too recent to be consistent with the "ancient African genetic structure" scenario, thus supporting the hypothesis of recent gene flow.

Then, this date has been refined thanks to the analysis of a modern human from Siberia lived around 45,000 years ago (Fu *et al.,* 2014; Moorjani *et al.,* 2016). Due to the high quality of this genome, it was possible to observe that the chunks of Neanderthal DNA were on average seven times larger than in modern-day individuals (Figure 28). This allowed estimating a more precise interval, between 50,000 and 60,000 years ago, which roughly corresponds to the major expansion of modern humans out of Africa and the Middle East. This finding also suggests that the bulk of the Neanderthal contribution to modern-day Eurasian people does not date back to the earlier times in the Middle East when Neanderthal and Sapiens were probably a stone's throw from each other.

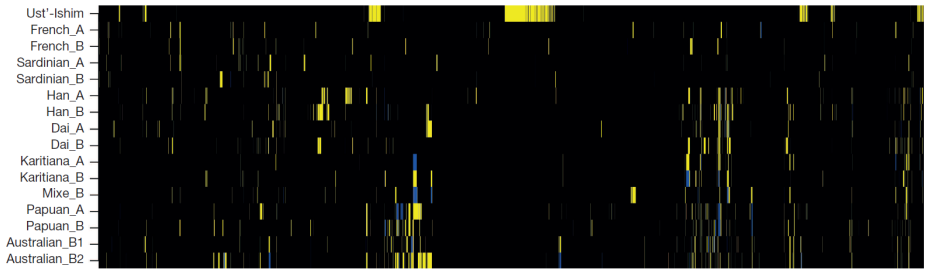


Figure 28. Neanderthal ancestry in chromosome 12 of the Ust'-Ishim individual. Image taken from Fu *et al.,* 2014 reporting the regions of Neanderthal ancestry on chromosome 12 in the Ust'-Ishim individual and fifteen present-day non-Africans.

Neanderthal DNA chunks and even more questions

As our understanding of the admixture dynamics between Neanderthals and modern humans increases, new questions arise: where and what is the genetic legacy our archaic cousins left inside us?

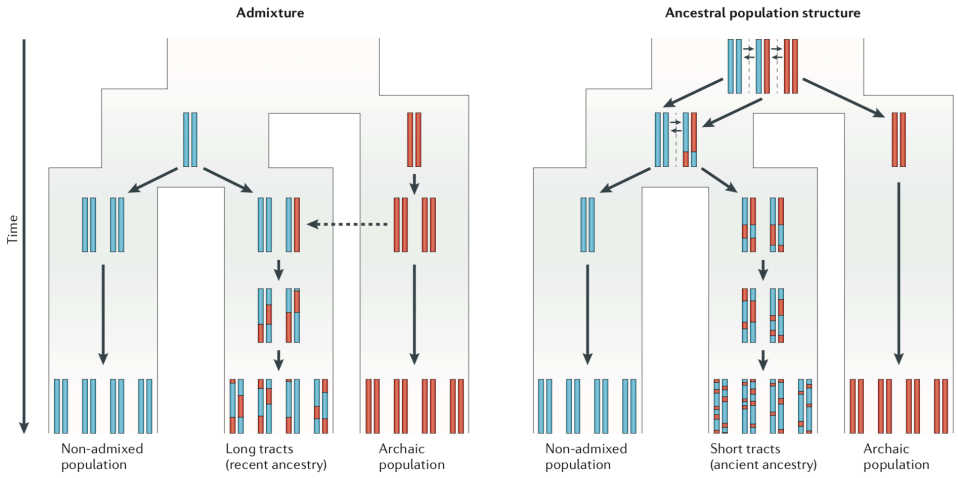


Figure 29. The expected length of archaic traits. Image taken from Racimo *et al.,* 2015 reporting the expected length of archaic tracts under admixture (left panel)and ancestral population structure scenarios (right panel). Because there is less time for recombination to break down the migrant tracts (red) in the admixed population, the expected tract length in this case will be longer than in the case of ancestral population structure.

ILS

A phenomenon whereby two or more lineages from different populations or species share a common ancestor more recently than their respective most recent common ancestor within populations, causing discordance between the population tree and a gene tree.
(Racimo *et al.,* 2015)

The first challenge in answering these questions is to identify the features of archaic introgressed regions, in order to properly distinguish them from shared ancestral genetic variation. In fact, in any two populations, there will always be some shared DNA segments originating in their common ancestors, thus two DNA chunks may share a most recent common ancestor (MRCA) more recently than other two DNA segments in the same population (Racimo *et al.,* 2015). This phenomenon, known as incomplete lineage sorting (ILS), needs to be differentiated from archaic admixture events (Figure 29).

During the years, many methods for identifying introgression have been developed. The most well known is the *D*-statistics (Green *et al.,* 2010; Durand *et al.,* 2011; Patterson *et al.,* 2012), which measures the sharing of derived alleles between two populations (sister ingroup), comparing them with an outgroup (see page 65). However, while this method was the first used for demonstrating the Neanderthal introgression, it cannot highlight DNA segments derived from an introgression event itself.

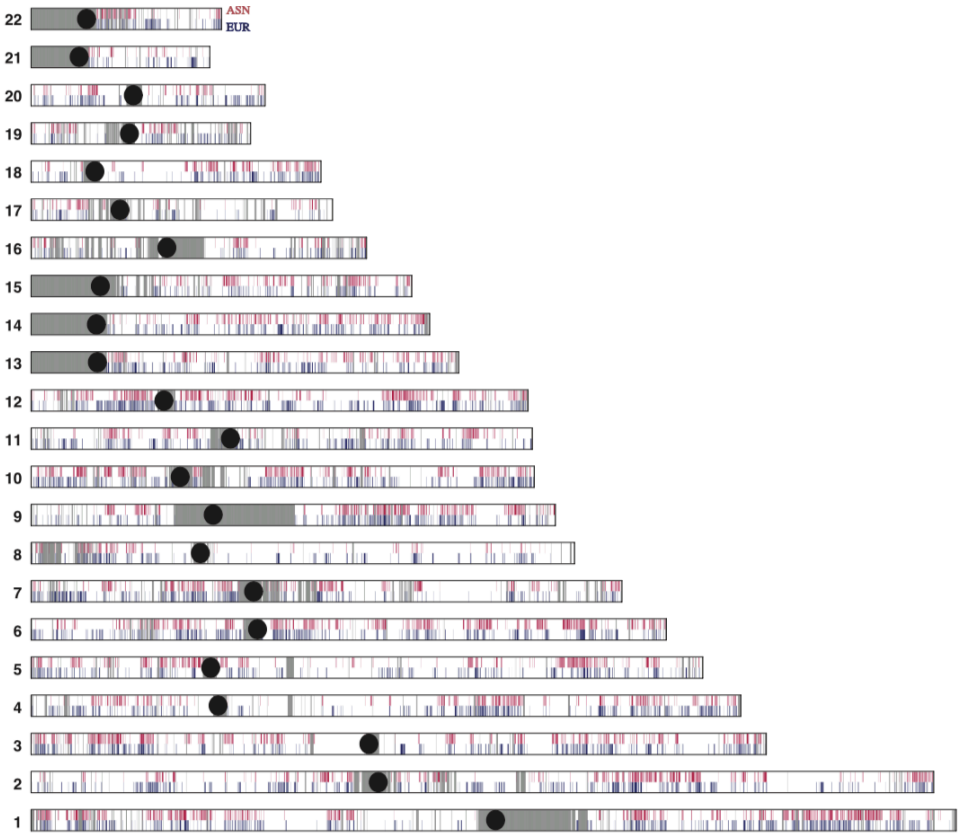


Figure 30. Neanderthal lineages in modern-day East Asians and Europeans. Image taken from Vernot & Akey, 2014 reporting the Neanderthal lineages identified in East Asians (ASN, red) and Europeans (EUR, blue). Gray shading denotes regions that did not pass filtering criteria, while black circles represent centromeres.

A characterising feature of introgressed segments is that they should be, on average, longer than ILS segments. Following this observation, Plagnol and Wall (Plagnol & Wall, 2006) exploited the LD pattern observed across the genome to pinpoint archaic inherited haplotypes: basically, they searched for derived alleles in high LD and they combined them into a score called S^*. Like the *D*-statistics, S^* can identify genome-wide evidence of introgression, but without any knowledge of the archaic donor population. A very good explanation of the logic underlying this method, as well as the other techniques for identifying archaic introgression, can be found in Santander *et al.,* 2019. In 2014, S^* was used by Vernot and colleagues for

identifying regions introgressed from Neanderthals into European and East Asian populations. Then, these regions have also been refined using the Neanderthal reference genome, by testing whether they matched it significantly more than expected by chance (Figure 30, Vernot & Akey, 2014).

The same two-step approach has been used two years later by the same group (Vernot *et al.,* 2016), in order to identify Neanderthal and Denisovan introgressed sequences in non-African modern-day populations (Figure 31), with a special focus on the Melanesian population.

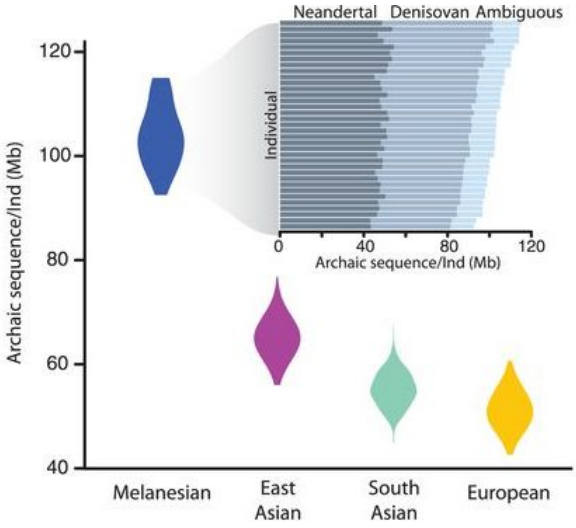


Figure 31. The amount of achaic introgressed sequences in modern-day populations. Image taken from Vernot *et al.,* 2016 reporting the amount of archaic introgressed sequences identified in each population. Inset, amount of Neanderthal, Denisovan, and ambiguous (Neanderthal or Denisovan) introgressed sequence for each Melanesian individual.

In 2018, an S^*-like method was introduced (Browning *et al.,* 2018). It is called Sprime and follows the same rationale of S^*, but avoiding windowing and thus focusing on the entire chromosomes. Moreover, the fact that through this new method they could analyse a large number of individuals simultaneously contributed in obtaining a much better trade-off of detection frequency to accuracy than the S^*. This improved performances allowed refining the archaic introgression proportions in modern-day populations (Figure 32) and to detect two distinct pulses of Denisovan admixture in East Asian populations.

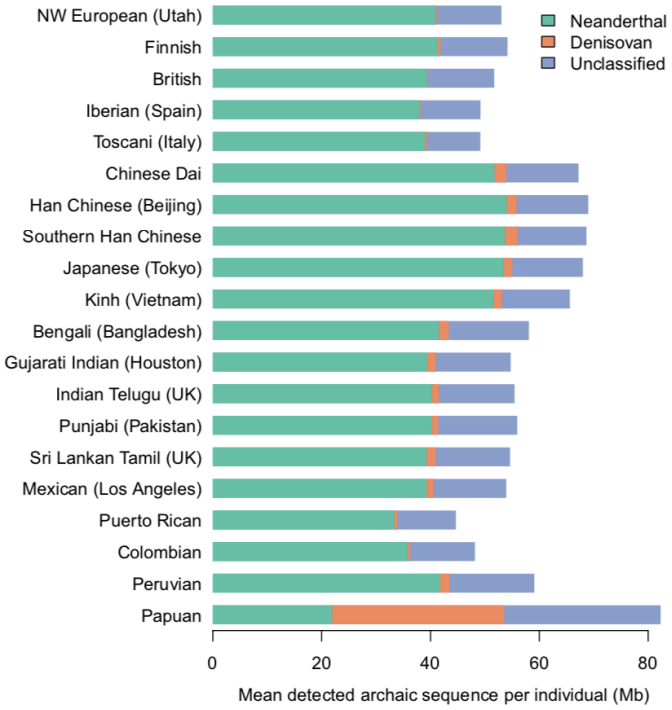


Figure 32. Amount of archaic introgressed sequences according to sprime. Image taken from Browning *et al.,* 2018 reporting the mean amounts of detected introgressed material per individual, classified by affinity to the Altai Neanderthal and Altai Denisovan genomes.

As an alternative to the analyses of LD patterns, mainly two probabilistic methods have been exploited for the detection of archaic regions: hidden Markov models (HMM, Prufer *et al.,* 2014; Seguin-Orlando *et al.,* 2014) and conditional random fields (CRF, Sankararaman *et al.,* 2014). Both methods start by some *a priori* parameters, which depend from demographic assumptions, and by an archaic reference genome. However, in 2018 Laurits Skov introduced a new HMM which does not rely on a reference, but on the removal of those variants shared with the outgroup (Skov *et al.,* 2018).

Being Neanderthal somewhere in our genome

Once scientists found the precise locations of Neanderthal introgressions, they noticed that these regions are not distributed evenly across the genome: while some portions of the genome are enriched in Neanderthals alleles, oth-

ers exhibited atypically low rates of introgression (Dannemann & Racimo, 2018). The latter is particularly attractive, as these regions represent events of restricted gene flow and, consequently, can be useful in understanding the molecular basis of reproductive isolation: examples are large sections on the X chromosome, genes near structural rearrangements and genes expressed in testes (Sankararaman *et al.,* 2012). Moreover, Neanderthal Y chromosomes and mtDNA have not been found in humans, despite the tens of thousands of samples tested so far. Their absence could be explained by random genetic drift, gender-biased gene flow and/or higher mutational load in the Neanderthal genomes. However, cross-species genetic incompatibilities could also have contributed to the genomic cleansing of Neanderthal ancestry observed in humans.

What does it mean to be Neanderthal in some spots of our genome? Why have some regions been purged by the archaic legacy, while other archaic variants have been maintained and, maybe, selected? What fate befalls those variants?

Since the release of the Neanderthal introgression maps, many studies have suggested that negative or purifying selection against archaic inherited regions had a big role in shaping the distribution of archaic DNA in modern-day human population (Figure 33, Harris & Nielsen, 2016; Juric *et al.,* 2016; Fu *et al.,* 2016; Sankararaman *et al.,* 2014). One of the striking evidence towards the action of purifying selection is the gradual decline in Neanderthal ancestry in modern humans over the past 45,000 years (Fu *et al.,* 2016), which cannot be explained only by a model of pure genetic drift. Even if the decline showed in Fu *et al.,* 2016, seems to be only the result of gene flow between modern human populations, the negative selection hypothesis against archaic DNA in conserved genomic regions remains stable (Petr *et al.,* 2019). Moreover, it has been also found that this depletion is stronger in regulatory and conserved noncoding regions and in the most conserved portion of protein-coding sequences (Petr *et al.,* 2019).

In contrast, other studies have suggested that the purifying selection on Neanderthal inherited regions was due to a higher deleterious burden in Neanderthals than in modern humans (Juric *et al.,* 2016). In particular, the analyses of high-coverage Neanderthal genome (Prufer *et al.,* 2014) and three Neanderthal exomes (Castellano *et al.,* 2014), suggest that around 0.5-1-0 million years ago their population become smaller, decreasing more and more for hundreds of thousands of years. In fact, their genetic diversity was almost a third of what has been estimated for modern humans (Castellano *et al.,* 2014). In this situation, Neanderthals accumulated many slightly deleterious variants, which, in their small population size, were neutral and

could not be removed by negative selection. In contrast, when these variants introgressed into the larger modern human population, they become more visible to the action of purifying selection.

However, many studies have also identified a large number of introgressed alleles which frequencies have increased in modern humans, thus representing examples of adaptive variation (Figure 33, Sankararaman *et al.,* 2014; Dannemann *et al.,* 2016; Gittelman *et al.,* 2016; Mendez *et al.,* 2013; Quach *et al.,* 2016; Racimo *et al.,* 2015; Sams *et al.,* 2016; Vernot & Akey, 2014). As Neanderthals and Denisovans were already living in Eurasia for at least 300,000 years before modern humans arrived, they were genetically adapted to the environmental conditions, nutrients and pathogens, which, in turn, were completely different from the African ones, to which our Sapiens ancestors were accustomed. Inheriting and carefully keeping those variants, meant new possibilities to cope the new world easily, a world that they were ready to conquer.

For this reason, a lot of studies have found the introgressed archaic human alleles to be involved in immunity, metabolism and the response to environmental conditions, like temperature, sunlight and altitude (Prufer *et al.,* 2014; Vernot & Akey, 2014; Sankararaman *et al.,* 2014; Mendez *et al.,* 2013; Sams *et al.,* 2016; Huerta-Sánchez *et al.,* 2014; Dannemann *et al.,* 2016; Gittelman *et al.,* 2016; Racimo *et al.,* 2017; Racimo *et al.,* 2015). Looks like we got ourselves a very sweet deal, when we admixed with our archaic cousins!

However, the majority of the archaic inherited alleles in the modern-day populations are not strongly adaptive, being present at frequencies around 2% (Dannemann & Kelso, 2017). Many studies have tried to investigate the biological meaning of these archaic variations, so far (Simonti *et al.,* 2016; Dannemann & Kelso, 2017), making astonishing discoveries. Simonti and colleagues searched the Neanderthal introgressed variants in a medical database, the Electronic Medical Records and Genomics (eMERGE) Network, which records both patients' genetic data and their diagnoses, providing a highly valuable resource for linking genes and conditions in almost 30,000 European individuals (Simonti *et al.,* 2016). Using this approach, they found that a large number of Neanderthal variants could influence many different traits, such as depression, precancerous skin lesions (called actinic keratoses), blood-clotting disorders, strokes, urinary tract disorders and tobacco addiction (Figure 34, Simonti *et al.,* 2016).

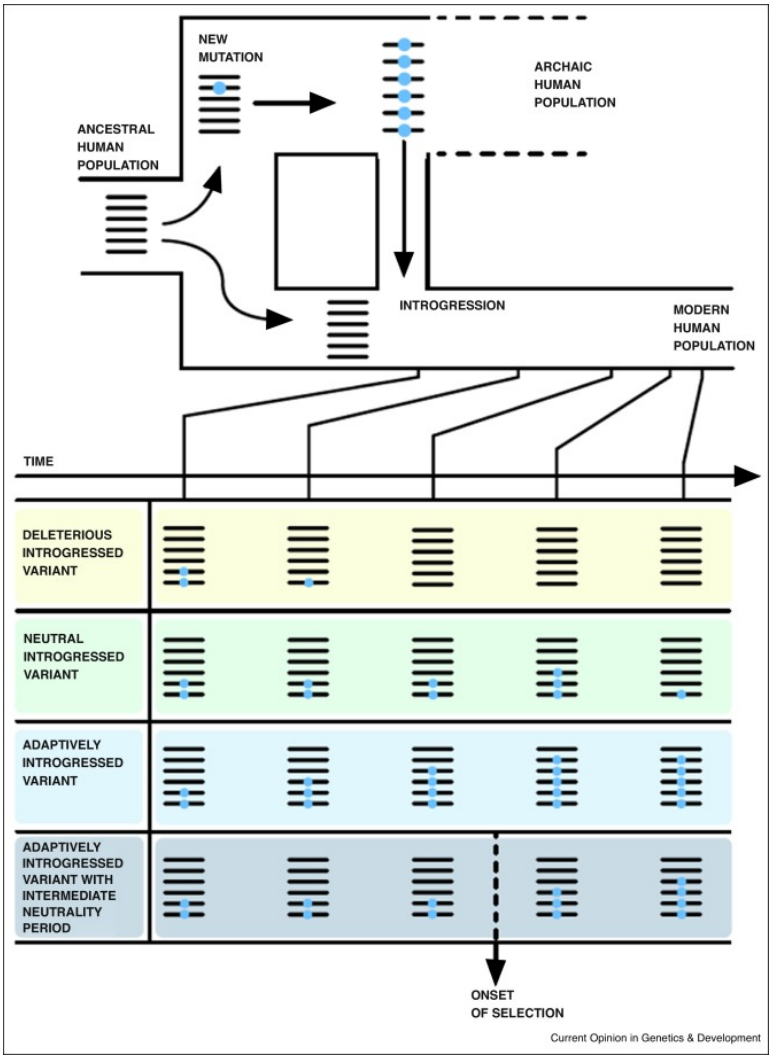


Figure 33. Possible fates of an archaic mutation in modern humans. Image taken from Dannemann & Racimo, 2018 reporting the possible fates of an introduced mutation into a modern human population via introgression from an archaic hominin group. The lines depict chromosomes in a population and the blue dot represents a mutation that first appeared and rose to high frequencies in the archaic human population, before being introgressed into the modern human population.

Michael Dannemann and Janet Kelso search, instead, in the UK Biobank, which includes traits related to physical appearance, diet, sun exposure, be-

haviour and diseases for more than 150,000 subjects (Dannemann & Kelso, 2017).

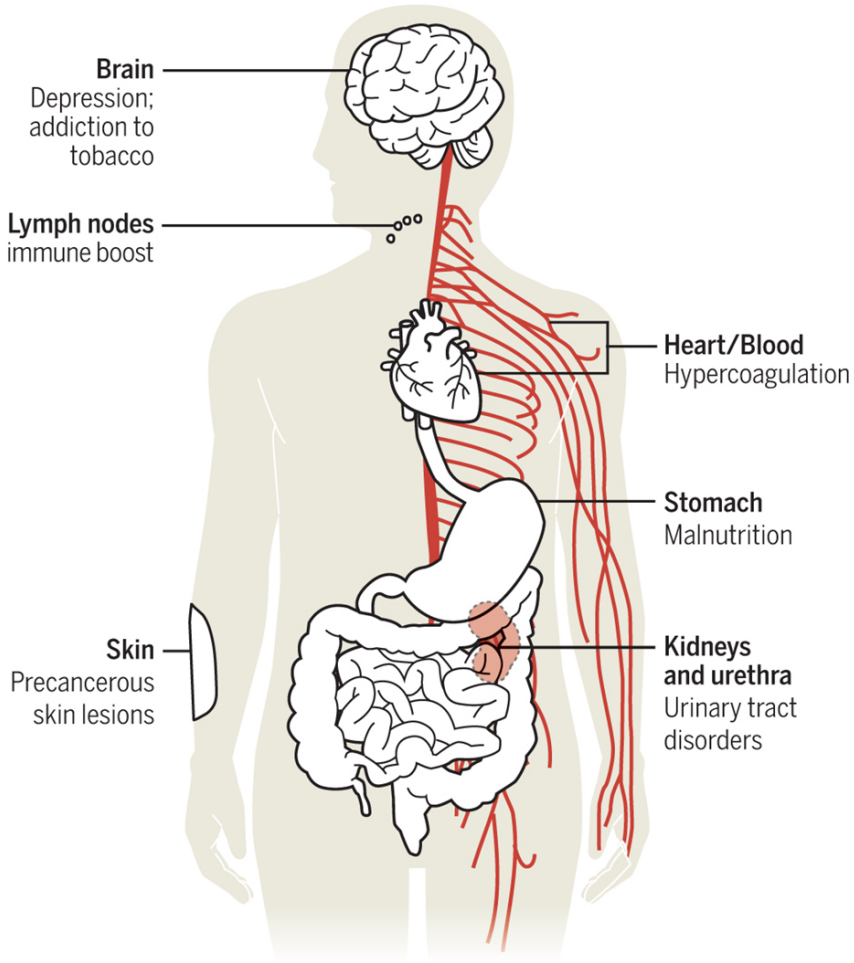


Figure 34. Phenotypes and medical conditions related to Neanderthal ancestry. Image taken from Gibbons, 2016 reporting the main phenotypes and medical conditions to which Neanderthals' hidden legacy has been related to.

In this study, they were able to test the contribution of archaic genetic variation to 136 common phenotypes in modern-day Europeans. Among them, they found that Neanderthal alleles contribute to four behavioural phenotypes, such as chronotype, loneliness or isolation, frequency of un-

enthusiasm or disinterest in the last two weeks, and smoking status. Moreover, they report that the same Neanderthal variant in *ASB1* significantly associated with the preference for evening activities, shows a correlation between its frequency and latitude, suggesting a relevant role for the differences in sunlight exposure. As a matter of fact, many phenotypes significantly influenced by Neanderthal introgression have some link to sunlight exposure (e.g., skin and hair colour, circadian rhythms, and mood, Dannemann & Kelso, 2017).

So in the end, do we have to thank or curse Neanderthals for having hidden a part of themselves in our genome? As in many other things, the truth lies somewhere in the middle. Take for example, the immune-boosting genes we have inherited from Neanderthals: as they were so beneficial in modern humans living in the Pleistocene, they have deleterious effects in the United States and Europe, where people face fewer parasites, thus causing autoimmunity, inflammation, and allergies.

Many genetic variants which have been crucial for the evolution and the adaptation of present-day human populations in the last 50,000 years, today, due to lifestyle changes and increased life expectancy, could influence the risk of complex diseases and other conditions.

On the Neanderthal demise and the way to avoid it

The thrilling discoveries that geneticists and palaeoanthropologists are doing about the past lives of our very old cousins have turned the overall feeling about them: the hairy and stupid beasts are gone, thus leaving only the human creatures they were.

Actually, we weren't so different and our paths crossed many times during the millennia. We share a common evolutionary history until around half a million years ago, but as our lineages separated, we had a different fate: while they were becoming less and scattered in little groups across Eurasia, we were increasing in number. When we met again before in the Middle East and later in Eurasian lands, we were more, we had better technology and superior social skills.

Around 40,000 years ago, while their time on Earth was running out, our civilization was about to flourish. In a few years, we would build cities, write poems and walk on the moon, while Neanderthals would remain buried under metres of dirt. They would have to wait patiently till 1856, in a cave and in our genomes, to be remembered again.

What did drive them to extinction? There are many hypotheses explain-

ing the Neanderthal demise. Maybe we had something to do with it: our ancestors could have filled the same ecological niche of Neanderthals, fighting for the same resources and the same space. Maybe violence and genocide could have happened. As Yuval Noah Harari notices, *"In modern times, a small difference in skin colour, dialect or religion has been enough to prompt one group of Sapiens to set about exterminating another group. Would ancient Sapiens have been more tolerant of an entirely different human species? They were to familiar to ignore, but to different to tolerate."*

However, even if this behaviour seems to fit our Sapiens attitude perfectly, it may not represent the most probable scenario. As Silvana Condemi points out in his book (Condemi & Savatier, 2016), very few Neanderthal fossils show signs of violent trauma. Moreover, in a genocide scenario, the Neanderthal demise would have occurred within a short period leaving a huge amount of fossils dating back that time, and not in at least five thousand years.

A new hypothesis, sadly still relevant for us today, is gaining acceptance: climate change. In particular, a recent work (Channell & Vigliotti, 2019) suggests that a weakening of Earth's magnetic field around 41,000 years ago - during the Laschamp event - could have reduced the protection of our planet from ultraviolet radiation. According to the study, Sapiens managed to survive thanks to a new version of the aryl hydrocarbon receptor (*AhR*), a receptor which can control the sensitivity to UV radiation. In the same vein, other climate-related hypotheses have been proposed, such as the eruption of the supervolcano Campi Flegrei, even if many researchers dispute this claim.

In the end, the answer to the first question of this chapter - *Do Neanderthal exist today?* - would seem crystal clear. However, Neanderthals had fought to survive, through the vast lands of Eurasia and the peaks of an Ice Age, for at least 300,000 years. They surely should have found other ways to survive.

Actually, they admixed with our ancestors and, consciously or not, they won their survival: as long as we are around on earth, they will live, more or less silently in our genome.

"In some spots of our genome, we are more Neanderthal than human."

Joshua Akey

The genetic legacy of Neanderthals in Italy and Europe

In this chapter I explore the Neanderthal genetic legacy still traceable in modern-day Italian and European people.

As previously said, the molecular analysis of Neanderthal remains provided evidence that the genomes of all people outside sub-Saharan Africa exhibit Neanderthal ancestry owing to interbreeding events. However, although it is known that differences in the amount of archaic introgression are present between human populations in Europe and Asia, the degree of variation in Neanderthal ancestry within Europe has not yet been explored. Due to the dual role of Neanderthal legacy, from one side an important source of adaptive variation in modern humans (Dannemann *et al.,* 2017), from the other a genetic burden influencing complex disease susceptibility risk and other conditions (Simonti *et al.,* 2016), it is worth investigating individual and populations variability.

Please, note that the work I will describe in this chapter is the first part of a bigger collaborative study between the Universities of Pavia and Oxford (see page XIII); for details about the project and other supplementary information, refer directly to the publication (Raveane *et al.,* 2019). In this work, in order to explore the distribution of the Neanderthal alleles and their effects in Italy and in Europe, we selected 7,164 tag-SNPs of introgressed regions from Neanderthals in a sample of 455 Italian individuals and 550

Europeans, African and Asian individuals.

This analysis revealed a genetic cline decreasing from the North to the South of Europe and Italy. By comparing allelic frequencies, we identified many Neanderthal introgressed variants whose frequency varies markedly among European populations, reflecting the existence of different demographic histories and selective pressures shaping European genetic diversity. We evaluated the biological functions and the phenotypic effects of the genes included in the putatively introgressed regions identified by the tag-SNPs, discovering that some of them appear to be involved in complex diseases such as neurodegenerative and metabolic disorders, while others contribute to individual variation in drug response.

We conclude recognizing that, certainly, archaic admixture has been useful for our modern human ancestors (adaptive variation). However, in many cases, the same variants so necessary during our first wanderings out of Africa could nowadays contribute in increasing the susceptibility risk to some complex diseases. For this main reason, the characterization of Neanderthal ancestry and the investigation of its functional consequences are essential to understand the genetic make-up of modern human population.

Methods

Dataset

In order to investigate the distribution of Neanderthal introgressed sequences in Italian and European populations, we focused on samples genotyped on the Illumina Infinium Omni2.5-8 BeadChip, which provided more than 2 million variants to be explored. Particularly, in this section, we used five European populations from 1000 Genomes Project (Auton *et al.,* 2015) (Utah Residents with Northern and Western European Ancestry-CEU, British in England and Scotland-GBR, Finnish in Finland-FIN, Iberian Population in Spain-IBS, Tuscans from Italy-TSI), 466 Italian samples grouped according to their geographical origin in four macro-areas (Italians from Northern Italy-ITN, Italians from Central Italy-ITC, Italians from Southern Italy-ITS, Italians from Sardinia-SAR) and finally one Asian (Han Chinese, CHB) and one African (Yoruba from Nigeria, YRI) for further comparisons. The Italian samples analysed in this part, except for TSI (Auton *et al.,* 2015), came from different sources. Part of these samples came from a study previously published by my group (Fiorito *et al.,* 2016). However, since this study did not analyse all Italian administrative regions, the groups of Oxford and Pavia had to newly genotype other Italian individuals on the Illumina Infinium Omni2.5-8 BeadChip, combining DNA samples from the Pavia DNA repository with that available from other collaborators (Dr. Brisighelli, Prof. Capelli, Prof. Lancioni, Dr. Montinaro, Prof. Pascali) for a final set of 166 Italian individuals (I refer to these samples in Table 2 with Raveane *et al.,* 2019").

We focused on autosomal SNPs and excluded insertions/deletions, multiallelic variants, polymorphisms with high missingness ($> 2\%$) and SNPs for which the strand could not be determined unambiguously. Moreover, we removed 179 SNPs whose minor allele frequency difference computed between Tuscans from Italy (TSI) genotyped as part of the 1000 Genomes Project and Central Italians (ITC) population was greater than 0.2, as these differences are likely due to mismatched annotations between different versions of the genotyping array. We also removed 11 individuals with global call rate under 98%.

The final dataset comprised 2,089,189 SNPs and 1,014 samples from Europe, Africa and Asia (Table 2), including 466 Italian samples (representing all of the 20 Italian administrative regions) whose four grandparents were born in the same Italian region. Italian samples were clustered as Northern, Central and Southern Italy according to the geographical position; the

POPULATION	SAMPLES	REFERENCE
Italian populations		
ITN (Northern Italy)	180	(Raveane *et al.*, 2019; Fiorito *et al.*, 2016)
ITC (Central Italy)	77	(Raveane *et al.*, 2019; Fiorito *et al.*, 2016)
ITS (Southern Italy)	82	(Raveane *et al.*, 2019; Fiorito *et al.*, 2016)
TSI (Tuscany in Italy)	99	(Auton *et al.*, 2015)
SAR (Sardinia)	28	(Raveane *et al.*, 2019; Fiorito *et al.*, 2016)
European populations		
CEU (Utah Residents (CEPH) with Northern and Western European Ancestry)	99	(Auton *et al.*, 2015)
FIN (Finnish in Finland)	100	(Auton *et al.*, 2015)
GBR (British in England and Scotland)	94	(Auton *et al.*, 2015)
IBS (Iberian Population in Spain)	100	(Auton *et al.*, 2015)
Extra-European populations		
YRI (Yoruba in Ibadan, Nigeria)	104	(Auton *et al.*, 2015)
CHB (Han Chinese in Bejing, China)	51	(Auton *et al.*, 2015)

Table 2. Human populations used in the study. List of populations available from the literature and newly genotyped with the Infinium Omni2.5-8 BeadChip kit (Illumina) used for investigating the contribution of Neanderthals into modern human genomes.

Tuscany (TSI) sample from the 1000 Genome Project was kept separate to maintain the size of each group comparable, and to provide an independent replication of a Central Italian population.

In order to assess the legacy of Neanderthal introgressed sequences in European populations we used the high-resolution introgression map of ∼6,000 Neanderthal haplotypes inferred by computing the S statistic and released by (Plagnol & Wall, 2006) and later refined by (Simonti *et al.,* 2016). They filtered out SNPs whose frequency significantly differed from the overall Neanderthal haplotype frequency and removed haplotypes with fewer than four likely Neanderthal-derived SNPs. In this way, ∼135,000 high-confidence "Neanderthal single-nucleotide polymorphisms" included in the introgressed haplotypes were retained (Simonti *et al.,* 2016). We intersected this list of high-confidence Neanderthal SNPs with our dataset and found 7,164 Neanderthal tag SNPs included in 3,281 Neanderthal introgressed regions present in the set of more than two million SNPs available to us.

The number of Neanderthal alleles in present-day human populations

Before computing the statistics on Neanderthal ancestry variation, we pruned the genotype data for linkage disequilibrium (LD), using a r^2 of 0.2, meaning that we removed one variant for each pairs of variants whose squared

correlation was greater than 0.2. We used a windows size of 500kb and a step size of 100 with the following command:

```
plink --bfile $FILE --indep-pairwise 500 100 0.2 --out $FILE
```

In this way, we obtained a dataset of 3,969 SNPs in putative linkage equilibrium.

At this point, we computed the counts of each Neanderthal allele for LD pruned introgressed sites across all samples. When we compare the Neanderthal allele with the minor allele in present-day populations, we noticed that, although they corresponded in most of the cases, in a few situations the Neanderthal introgressed allele was the most frequent at least in one modern-day population. For this reason, we had to identify the Neanderthal allele. We did it by analysing the $\sim$ 30-fold coverage sequence of the Neanderthal individual from the Altai Mountains in Siberia (Prufer *et al.,* 2014). It was straightforward to select the Neanderthal allele at homozygous positions, while the archaic allele at heterozygous positions was selected as the minor allele in Yoruba people (YRI).

The bash code at page 312 shows how I counted the number of Neanderthal alleles in heterozygosity and in homozygosity per individual in each population.

Then, we simply summed the number of Neanderthal alleles in each individual, multiplying by 2 the number of homozygous positions. In order to explore putative differences in the distribution of Neanderthal allele counts across populations, we performed a series of two-sample Wilcoxon rank sum tests and we used the Bonferroni adjustment for multiple testing correction. We also tested if the significant differences we found were maintained when we removed outlier individuals. Outliers were defined as samples with Neanderthal allele count lower than Q1–1.5 $\times$ IQR or greater than Q3 + 1.5 $\times$ IQR, where Q1 is the first quartile, Q3 is the third quartile and IQR is the interquartile range.

Basal Eurasian ancestry and Neanderthal contribution

We estimated the amount of Basal Eurasian and Neanderthal ancestry present in these European populations using the f_4-ratio implemented in the ADMIXTOOLS package (Patterson *et al.,* 2012). Specifically, *qpDstat* with f4mode: YES, and computed standard errors with a block jack-knife. We estimated the basal Eurasian and the Neanderthal fraction using the f_4-ratio in the forms:

$$\frac{f_4(Target, Loschbour, Ust\_Ishim, Kostenki14)}{f_4(Mbuti, Loschbour, Ust\_Ishim, Kostenki14)}$$

and

$$\frac{f_4(Mbuti, ChimpTarget, Altai)}{f_4(Mbuti, Chimp, Dinka, Altai)}$$

as in Lazaridis *et al.,* 2016 and Fu *et al.,* 2016.

Then, in order to increase the number of points available for correlation testing, we annotated the CP/fS cluster affiliation (see *Methods - Haplotype analysis (CHROMOPAINTER and fineSTRUCTURE)* on page 146) for each sample genotyped on the Illumina Omni 2.5 array. In this analysis, only clusters with at least 10 samples were included.

African ancestry and Neanderthal legacy

The impact of African contributions in shaping the amount of Neanderthal occurrence was evaluated by exploring how the removal of the clusters showing African gene-flow as detected by a GLOBETROTTER (GT) analysis performed by Dr. Alessandro Raveane in the context of the whole project (see page XIII) and how individuals belonging to these clusters affected the correlation between Basal Eurasian/Neanderthal estimates and the degree of population differentiation in the amount of Neanderthal alleles, respectively.

In particular, in relation to variation in Neanderthal signatures within each individual, we removed individuals belonging to clusters where the GT analysis identified signatures of African admixture (clusters SItaly1, SItaly2, Sicily1, Sardinia2, NWEurope3, e WEurope1, WEurope3 and WEurope4).

Following the same reasoning, we removed the clusters with signatures of African admixture (listed above) when testing the correlation between the amount of Basal Eurasian ancestry and Neanderthal ancestry. We additionally removed all the clusters with less than 10 individuals to minimise the impact of estimates based on small sample sizes (CAfrica3, EAsia2, Sardinia2, NCItaly3, Sicily1, EEurope5, Sicily2, Sardinia3, NWEurope1).

Comparison of Neanderthal allele frequencies across modern populations

For each SNP of the dataset, we computed the absolute "X" allele frequency differences between each of the possible pairs of the eleven populations (ΔXAF, where X is the minor allele for each SNP or the Neanderthal

allele when considering Neanderthal regions tag-SNPs), generating a total of 55 ΔXAF distributions. Then, we selected the Neanderthal-Tag SNPs corresponding to the Top 1% (NTT SNPs) of the ΔXAF genome-wide distributions of each of the 55 pairwise population comparisons.

The bash code at page 313 reports the pipeline I wrote in order to compute Neanderthal allele frequencies in each modern population, the ΔXAF values and the list of NTT SNPs. Specifically, see page 315 for the R script I wrote to compute the ΔXAF values among pairs of modern populations. R codes at pages 317 and 318 were used to compute the percentile thresholds and the list of Neanderthal SNPs in the 99th percentile (NTT SNPs), respectively.

The biological implications of Neanderthal introgression

Given the list of genes overlapping the Neanderthal introgressed regions harbouring the NTT SNPs and the list of genes directly harbouring the NTT SNPs, we performed different enrichment tests with the online tool EnrichR (Chen *et al.,* 2013; Kuleshov *et al.,* 2016). Particularly, we searched for significant enrichments compared to the human genome using the EnrichR collection of database, e.g. dbGaP (Mailman *et al.,* 2007; Tryka *et al.,* 2014), Panther 2016 (Mi *et al.,* 2017), HPO (Kohler *et al.,* 2017) and KEGG 2016 (Kanehisa *et al.,* 2016; Kanehisa *et al.,* 2017; Kanehisa & Goto, 2000). We then investigated known direct associations between the Neanderthal alleles of the NTT SNPs and phenotypes, by looking in the GWAS and PheWAS catalogues (Denny *et al.,* 2013; MacArthur *et al.,* 2017) and by applying the PheGenI tool (Ramos *et al.,* 2014). We used the circos representation as in Kanai *et al.,* 2018, to highlight different sets of NTT SNPs.

Results

Meeting the Neanderthals: how much kissing?

As previously reported, it is known from the literature (Vernot & Akey, 2014; Sankararaman *et al.,* 2012; Meyer *et al.,* 2012; Wall *et al.,* 2013; Prufer *et al.,* 2017) that the frequency of Neanderthal introgression is substantially higher ($\sim$ 30%) in East Asians than in Europeans, however the degree of variation in Neanderthal ancestry within European population has not yet been extensively explored.

By comparing the distributions of Neanderthal allele counts across populations, we confirmed the known higher amount of Neanderthal ancestry in Asian samples than in European samples. Moreover, we observed a noticeable genetic cline decreasing from the North to the South of Europe (Figure 35A), which was also appreciable alongside the Italian Peninsula.

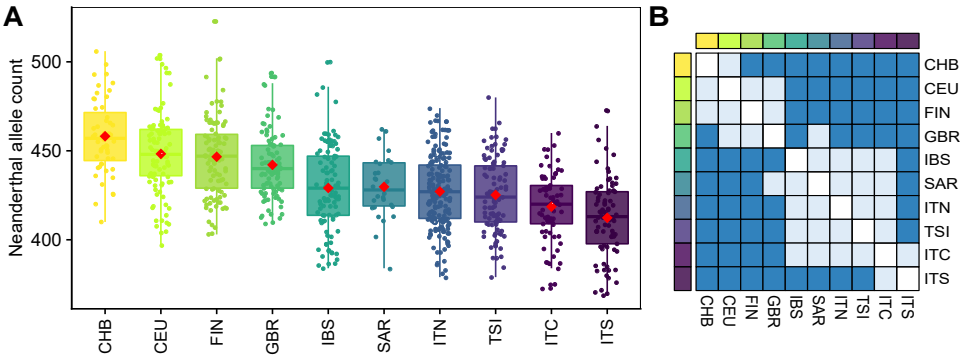


Figure 35. Neanderthal ancestry distribution in Eurasian populations. A) Neanderthal allele counts in individuals from Eurasian populations, sorted by median values on 3,969 LD-pruned Neanderthal tag-SNPs. CEU, Utah Residents with Northern and Western European ancestry; GBR, British in England and Scotland; FIN, Finnish in Finland; IBS, Iberian Population in Spain; TSI, Tuscans from Italy; ITN, Italians from North Italy; ITC, Italians from Central Italy; ITS, Italians from South Italy; SAR, Italians from Sardinia; CHB, Han Chinese. **B)** Matrix of significances based on Wilcoxon rank sum test between pairs of populations including (lower triangular matrix) and removing (upper) outliers (dark blue: adj p-value < 0.05; light blue: adj p-value > 0.05).

When we analysed the putative significant differences in the distributions, we found that the Asian sample was significantly different from the

distribution of all the other populations, with the exception of the CEU and FIN populations. Similarly, Northern European populations were significantly different from the Southern European populations. Interestingly, significant differences were found alongside the Italian Peninsula, specifically samples from Northern Italy showed higher Neanderthal allele count than Central and Southern Italian individuals (lower triangular matrix in Figure 35B). These significant differences between Italian and European population were mostly maintained also when we removed outlier individuals (upper triangular matrix in Figure 35B).

Footprints of a ghost: Basal Eurasian ancestry and Neanderthal contribution

Ancient samples have been reported to differ in the amount of Neanderthal DNA partially because of the variation in proportion of the "Basal Eurasian" lineage, which harbours only a negligible fraction of Neanderthal ancestry (Lazaridis *et al.,* 2016; Fu *et al.,* 2016).

In order to better understand the correlation between the amount of Neanderthal and Basal Eurasian ancestry across Europe, we estimated the basal Eurasian and the Neanderthal fraction using the f_4-ratio (see *Methods - Basal Eurasian ancestry and Neanderthal contribution*; Lazaridis *et al.,* 2016; Fu *et al.,* 2016).

We initially tested if the estimates of the amounts of Neanderthal admixture based on the f_4-ratio correlated both with the mean (Figure 36A) and the median (Figure 36B) of the Neanderthal allele counts in European populations in Figure 35A, and found these to be significantly correlated (r=0.94, p-value = $1.5e^{-4}$; r=0.96, p-value = $5.9e^{-5}$).

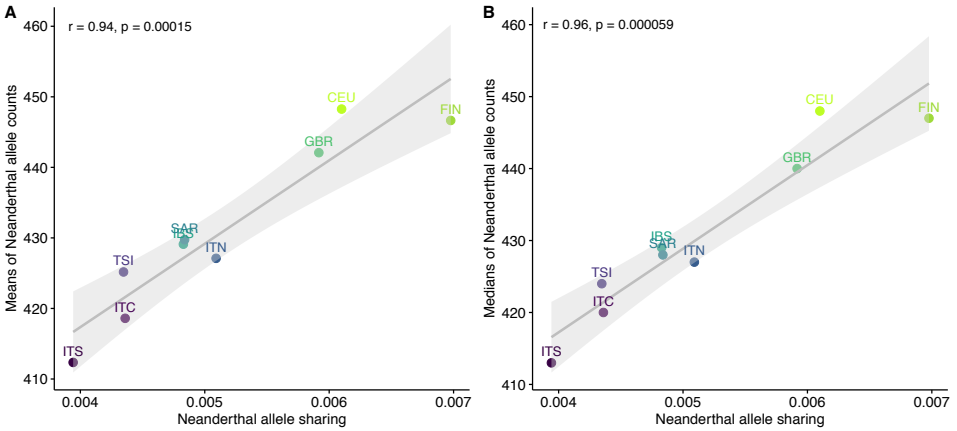


Figure 36. Correlation between the proportions of Neanderthal allele sharing computed with f_4-ratio and the counts per population of Neanderthal alleles in European populations. A) Correlation between the proportions of Neanderthal allele sharing computed with f_4-ratio and the means per population of Neanderthal allele counts. **B)** Correlation between the proportion of Neanderthal allele sharing computed with f_4-ratio and the medians per population of Neanderthal allele counts.

Then, we observed that the amount of Basal Eurasian component in European populations is inversely correlated with the f_4-ratio estimates of Neanderthal contribution, only when the Finnish cluster is not considered (cor=-0.7753605, p=0.02378; Figure 37A,B). Notably, the Finnish (FIN) population is enriched in Asian ancestry, a component shown to inflate the estimates of Basal Eurasian (Martin *et al.*, 2018).

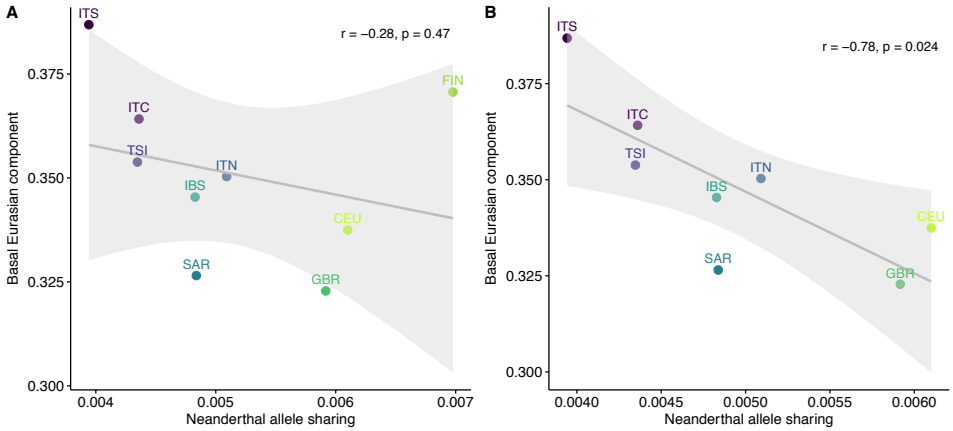


Figure 37. Correlation between the proportion of Neanderthal allele sharing and the amount of ancestry derived from a Basal Eurasian population in European populations. A) Correlation analysis including FIN (Finnish in Finland) population. **B)** Correlation analysis excluding FIN (Finnish in Finland) population.

When we increased the number of points for the correlation by considering modern-day European clusters, we found, again, the estimated amounts of Basal Eurasian and Neanderthal to be negatively correlated (Figure 38), consistent with (Lazaridis *et al.*, 2016).

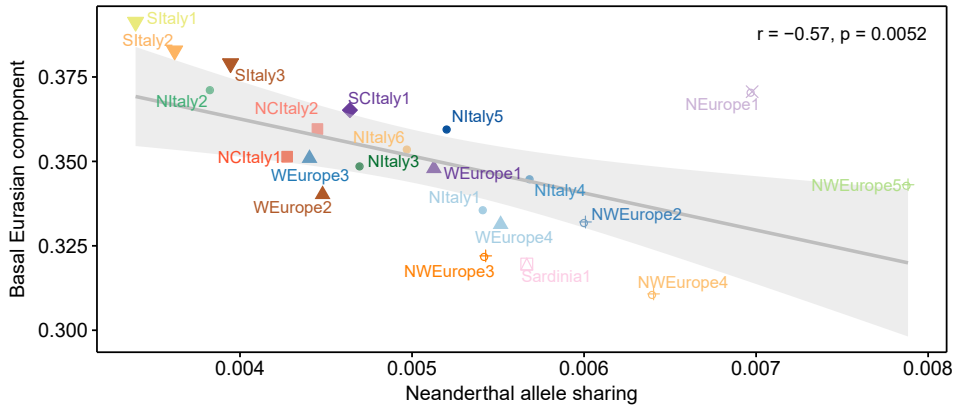


Figure 38. Basal Eurasian ancestry and Neanderthal contribution. Correlation between Neanderthal ancestry proportions and the amount of Basal Eurasian ancestry in modern-day European clusters.

Non-Eurasian contributions muddying the water

We reasoned that the contributions from African groups could have possibly influenced the differential patterns of Neanderthal introgression reported above, particularly in Southern European populations (Busby *et al.,* 2015). Figure A.1 in the Appendix shows which clusters exhibited Northern African contributions in the GLOBETROTTER and MALDER analyses (for further details see Raveane *et al.,* 2019). At the same time, also our estimation of Basal Eurasian ancestry might be affected by non-Eurasian admixtures.

When we removed individuals belonging to clusters showing African signals, we found that while some population-to-population comparisons were not significant anymore (GBR-IBS and TSI-ITS), the overall patterns of differences between Northern and Southern Europe as well variation within Italy were still present (Figure 39A,B; see *Methods - African ancestry and Neanderthal legacy*).

Then, we removed all clusters with African contributions and we tested the correlation between the amount of Basal Eurasian ancestry and Neanderthal ancestry again. We noted that once all these clusters are removed no significant correlation is present (Figure 39C; p-value = 0.091 vs p-value = 0.0052, with and without removing clusters with signatures of African admixture, respectively). However, the cluster NEurope1 is mostly characterised by Finnish samples, for whom an excess of both Basal Eurasian and Neanderthal components, possibly due to East Eurasian gene flow, has been previously reported (Lazaridis *et al.,* 2014). We therefore re-run the analysis excluding also NEurope1 and observed a significant correlation (p-value = 0.01, Figure 39D). We note here that the removal of NEurope1 in the initial correlation analysis (Figure 38) resulted in a stronger correlation (r=-0.64, p-value < 0.05).

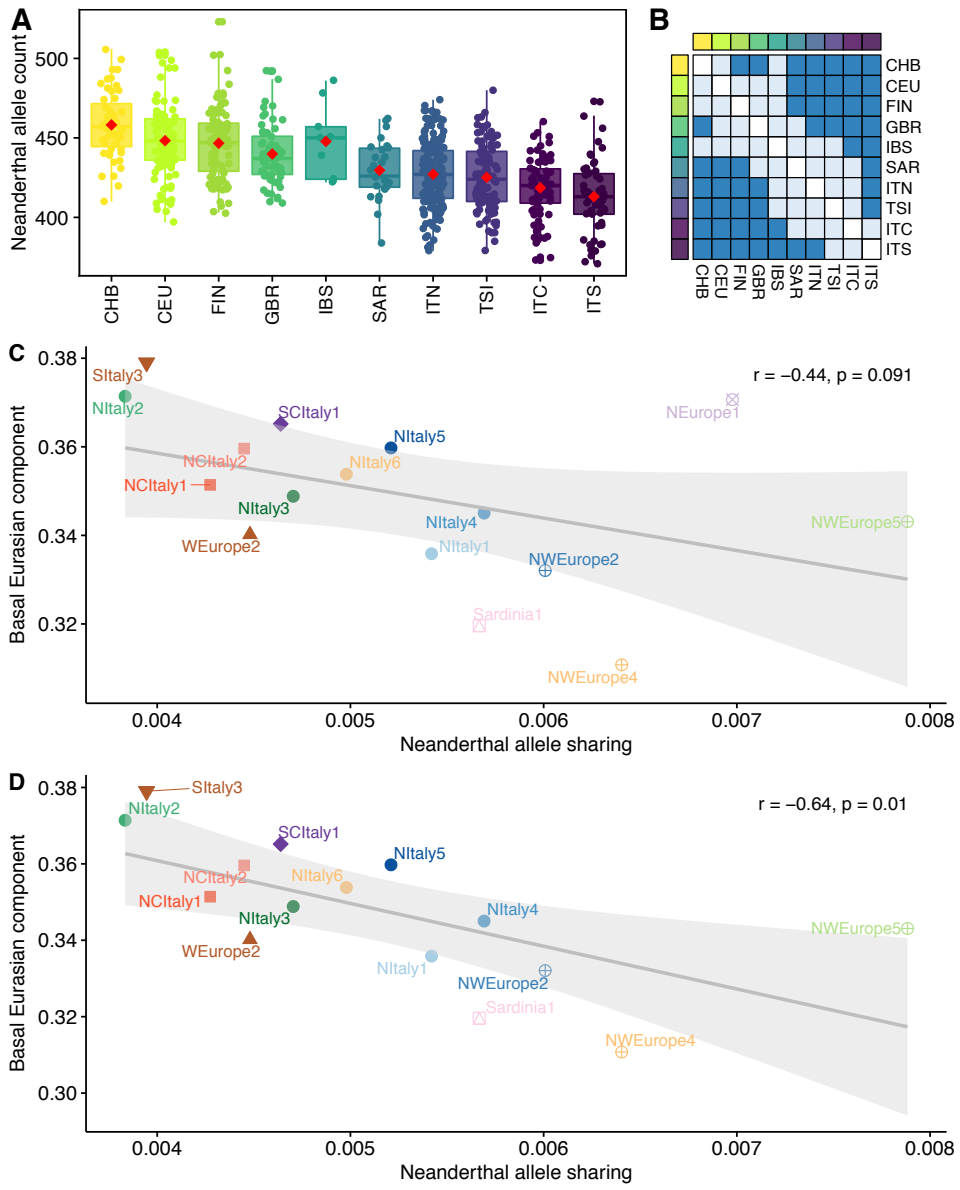


Figure 39. Exploring the relationship between Neanderthal ancestry and admixture with African sources. Same as in Figure 35A, B, and in Figure 37 but removing either the individuals belonging to clusters where the GT analysis identified signatures of African admixture (clusters SItaly1, SItaly2, Sicily1, Sardinia2, NWEurope3, WEurope1, WEurope3 and WEurope4) or the whole set of the clusters listed above.

Specifically: **A)** Neanderthal allele counts in individuals from Eurasian populations, on 3,969 LD-pruned Neanderthal tag-SNPs; **B)** Matrix of significances based on Wilcoxon rank sum test between pairs of populations including (lower triangular matrix) and removing (upper) outliers (dark blue: adj p-value < 0.05; light blue: adj p-value > 0.05). **C)** Correlation between Neanderthal ancestry proportions and the amount of Basal Eurasian ancestry in European clusters. **D)** Same as (C) but removing the cluster NEurope1. Clusters with less than 10 individuals were excluded in (C) and (D).

It's all about frequencies

Once confirmed that the genomes of people outside sub-Saharan Africa show subtle but discernible differences in the proportion of Neanderthal ancestry, we explored how such differences mapped at each locus with a Neanderthal allele. In other words, we identified significant differences in the frequencies of Neanderthal alleles across populations by computing the allele frequency differences (ΔXAF; see *Methods - Comparison of Neanderthal allele frequencies across modern populations*) for every SNPs for each of the possible pairs of the eleven populations in our dataset, thus obtaining 55 distributions (Figure 40).

Then, the NTT SNPs, i.e. the Neanderthal-Tag SNPs in the Top 1% of each distribution were selected (the supplementary file with the information about these SNPs can be found here).

In doing so, we identified 144 NTT SNPs, 128 of which present in comparisons between European populations. By computing the absolute minor/Neanderthal allele frequency differences for the entire set of 2,089,189 SNPs and by selecting only the Neanderthal variants in the top of the distributions, we can reasonably hypothesise that, at least for some of the SNPs, such high-frequency differences were possibly not caused by drift, but rather by different selective pressures.

We computed the odds ratio of being a Neanderthal allele in the top 1% of each comparison by a Fisher's exact test. We discovered that each comparison between pairs of populations was significantly depleted in SNPs in Neanderthal introgressed regions, with values ranging between 60 (OR: 0.0165 in the comparison between ITC and YRI) and 3 (OR: 0.2801 in the comparison between CEU and IBS) times the background values (Table A.1).

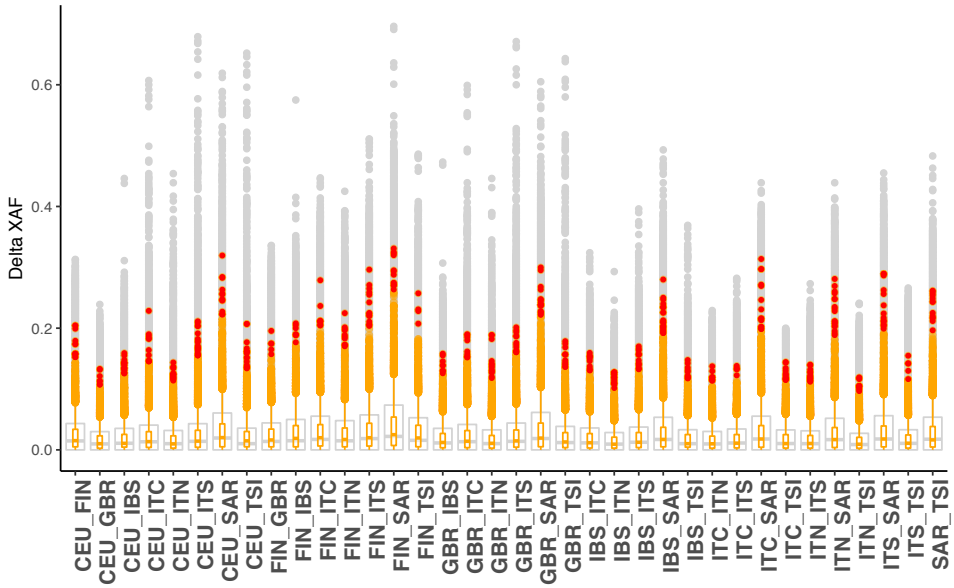


Figure 40. Absolute allele frequency differences (ΔXAF, where X is the minor allele for each SNP or the Neanderthal allele when considering Neanderthal regions tag-SNPs) for each pair of European populations. We reported in grey the boxplot representing the total distributions of the variants, and in orange the distribution of Neanderthal inherited variants. The red dots are the Neanderthal SNPs in the top 1% of the distributions.

The consequences of a promiscuous affair

The 144 NTT SNPs (Neanderthal alleles that showed frequency differences in the top 1% of the genome-wide distribution in comparisons between pairs of populations) mapped into 93 different Neanderthal introgressed regions, which hosted 141 genes.

The genomic regions spanned by the Neanderthal introgressed haplotypes harbour variants associated to several traits including body features, metabolites and electrolytes levels, neurodegenerative and neurological diseases, cardiovascular-related traits and environmental factors.

We performed some enrichment tests with the online tool EnrichR, by uploading i) the list of genes within Neanderthal introgressed regions harbouring the sets of 144 SNPs, and ii) the list of genes in which different sets of variants were placed.

The 141 genes were found significantly enriched, compared to the human genome, for "Body Weight" (dbGaP, adj p-value = 0.001963), and the metabolic pathways "Pentose phosphate pathway" (Panther 2016, adj p-value = 0.03531) and "Fructose galactose metabolism" (Panther 2016, adj p-value = 0.03531). Interestingly, the 50 genes harbouring the 144 variants proved to be significantly enriched in two Human Phenotype Ontology terms, such as "Facial shape deformation" (adj p-value = 0.01932) and "Potter's facies" (adj p-value = 0.01932), due to the *AGTR1* and *NPHP3* genes. In addition, the analysis revealed that these genes were enriched in the database of Genotypes and Phenotypes (dbGaP) for "Body Weight", "Body Height" and "Obesity" (adj p-value: 0.001187, 0.02872, 0.02395, respectively), "Glucose" and "Apolipoproteins B" (adj p-value=0.01465, 0.02395, respectively) and "Amyotrophic Lateral Sclerosis" (adj p-value = 0.01704).

We noticed that a subset of the NTT SNPs with the highest frequency differences followed peculiar patterns of frequency distributions (Figure 41). For example some alleles had high frequency in Europe and low frequency in CHB and YRI (Figure 41A), while others diverged between North and South Europe (Figure 41B).

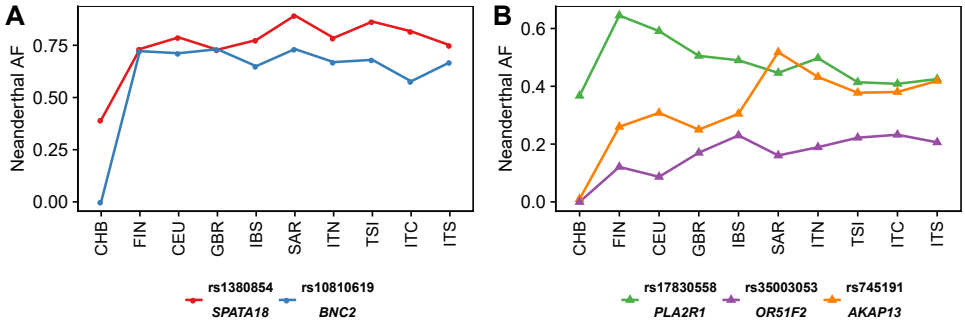


Figure 41. Neanderthal allele frequency (AF) for selected SNPs within the indicated genes. A) high frequency alleles in Europe; **B)** North-South Europe divergent alleles.

A single SNP (rs1380854) was unique to the comparisons between Eurasian and African populations, with the Neanderthal allele reaching frequencies above 70% in all the European groups (Figure 41A). This SNP mapped within the *SPATA18* (Spermatogenesis Associated 18) gene, which codes for a protein also known as "Mitochondria-eating protein". As the name suggests, the gene and its protein are key regulators of mitochondrial quality through the formation of intramitochondrial lysosome-like organelles that

eliminate the oxidized mitochondrial proteins and are involved in spermatogenesis (Bornstein *et al.,* 2011). Moreover, its Rattus norvegicus homolog, Spetex-1, was characterized as a spermatogenesis-related protein (Iida *et al.,* 2004). A recent study (Bornstein *et al.,* 2011) demonstrated that this gene is a testis-associated p53 target gene and it was proposed to be a structural component of the sperm flagella. The p53 (Hu, 2009) tumour suppressor protein is a sequence-specific transcription factor that mediates processes such as cell cycle arrest, cell death, DNA repair but also autophagy, differentiation, reproduction, metabolism, and aging. However, the finding that p53 orthologs are present also in organisms that do not develop cancer suggests that the "tumour suppressor" function was not its primordial role. In fact, it has been suggested that the role of the ancestral p53 is to control the fidelity of the reproduction and developmental processes (Hu, 2009).

When we explored the NTT SNPs showing differences within Eurasia, we found 14 SNPs whose frequencies were significantly different between Europe and Asia. In two of these variants, rs10810616 and rs10810619, the Neanderthal alleles have high frequency (57-73%) in Europeans, while being absent in the Han Chinese (CHB). These SNPs, on chromosome 9, mapped within the *BNC2* gene, which codes for the Zinc Finger Protein Basonuclin-2. This gene has been shown to contribute to skin colour saturation in Europeans (Jacobs *et al.,* 2013) and has been previously reported in relation to the phenotypic impact of Neanderthal alleles (Vernot & Akey, 2014; Dannemann *et al.,* 2017; Sankararaman *et al.,* 2012) (Figure 41A).

Eighty of the 144 NTT SNPs were found at least once in the top 1% of the genome-wide comparisons between Northern (Utah Residents (CEPH) with Northern and Western European Ancestry-CEU, British in England and Scotland-GBR and Finnish in Finland-FIN) and Southern European populations (Iberian Population in Spain-IBS and Italian groups).

Among these 80 SNPs, three (rs6759924, rs16844715, rs17830558) mapped to the Neanderthal introgressed haplotype hosting the *PLA2R1* gene, the archaic allele at these positions reaching frequencies of at least 43% in Northern European and at most of 35% in Southern European populations (Figures 41A,B). Ten SNPs showed an opposite frequency pattern, with the Neanderthal allele reaching the highest frequencies in Italian samples. Seven of these SNPs mapped to one Neanderthal introgressed region on chromosome 11 spanning the *OR51F1*, *OR51F2* and *OR52R1* genes (Figures 41B and 42), and other three identified regions hosting the *AKAP13* gene, within one of the high frequency European Neanderthal introgressed haplotypes recently reported (Gittelman *et al.,* 2016) (Figure 41B and 42).

NTT SNPs between Northern and Southern Italian populations included

variants located in the *WWOX*, *TNFRSF19*, *KDM2B* and *CCDC91* genes, while variants in the *RAI14*, *PRDM5* and *PRDM8* genes have been highlighted when comparing continental Italian and Sardinian populations (Figure 43). Interestingly, two variants in the *WWOX* gene have been associated to infectious diseases, hyperventilation and aphasia (Simonti *et al.,* 2016), while a SNP in the *CCDC91* gene has been associated to the "height" trait (Lango Allen *et al.,* 2010).

These 80 SNPs were included in 51 different Neanderthal introgressed regions and in 25 different genes. The 25 genes were found significantly enriched in the Human Phenotype Ontology database and dbGaP for "Peripheral demyelination" category (adj p-value = 0.02682, due to the presence of the *SH3TC2* and *MTTP* genes) and "Glucose" (adj p-value = 0.02543), respectively. Interestingly, the 76 genes mapping into the 51 introgressed regions were found enriched in the KEGG 2016 pathway "Butirosin and neomycin biosynthesis" (adj p-value = 0.01231). This signature is due to the *HKDC1* and *HK1* genes, that are hexokinases playing an important role in glucose metabolism. We found an enrichment in demyelination and in the decreased motor nerve conduction velocity categories according to the HPO database (adj p-value : 0.004903). Moreover, a significant enrichment (adj p-value : 0.000007695) of signals in cytogenetic band chr5q32 highlighted 5 genes (*RBM27*, *SH3RF2*, *SH3TC2*, *LARS* and *PLAC8L1*) in the "Chromosome Location" section of the EnrichR tool.

We then searched for known associations between all the 144 NTT SNPs with frequency differences among European populations and phenotypes by exploring GWAS and PheWAS catalogues and the online tool PheGenI (Phenotype-Genotype Integrator). A total of 34 SNPs were found to have at least one phenotypic association. The ancestral allele at SNP rs17830558 in the *PLA2R1* gene is one of the predominant risk loci for membranous nephropathy (Stanescu *et al.,* 2011; Sekula *et al.,* 2017), thus explaining the high frequency of the protective Neanderthal allele in present-day human populations (Figure 41B and 42). Neanderthal alleles resulted associated with increased gene expression in testis and in skin after sun exposure (SNP rs2281919 for *IP6K3* and *ITPR3* genes), susceptibility to cardiovascular and renal conditions (SNP rs5186 for the *AGTR1* gene), and "Brittle cornea syndrome" (SNP rs12499000 within the *PRDM5* gene, (Simonti *et al.,* 2016)), while rs745191, a missense variant within the AKAP13 gene (Figure 41B), was associated with several medical conditions in the Electronic Medical Records and Genomics (eMERGE) Network (Simonti *et al.,* 2016).

In order to best appreciate the frequency patterns present in the NTT SNPs, we started plotting NTT SNPs highlighted in comparisons between

Northern European and Italian populations (excluding Sardinia, Figure 42). Here we found again some genes reported above, such as *PLA2R1* and *AKAP13*.

Then, we plotted different sets of SNPs, according to the pair of populations in which they had different frequencies (e.g., Northern European populations vs Southern European populations NTT SNPs, European populations vs YRI NTT SNPs etc.) (Figures 42, 43, 44, 45, 46).

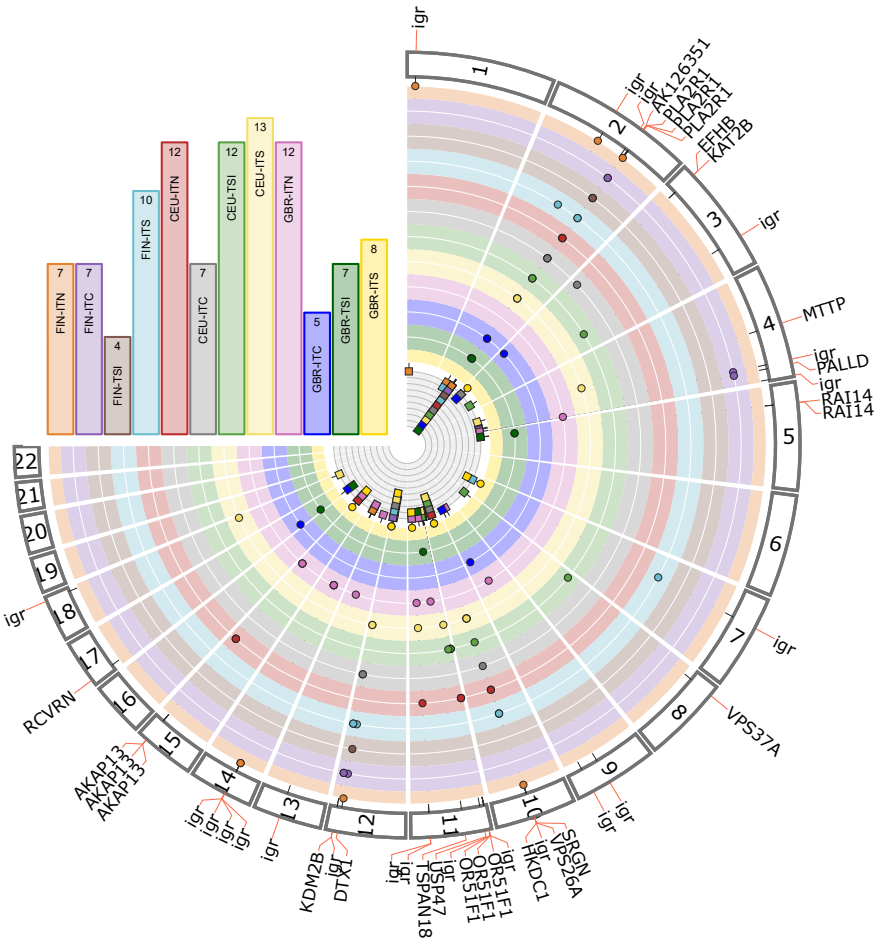


Figure 42. NTT SNPs highlighted in comparisons between Northern European and Italian populations (excluding Sardinia). Bars refer to comparison for reported pairs of populations; the number of NTT SNPs is reported within bars. Each section of the circos represents a tested chromosome; points refer to NTT SNPs. Colours, same as for bars; igr: intergenic region variant.

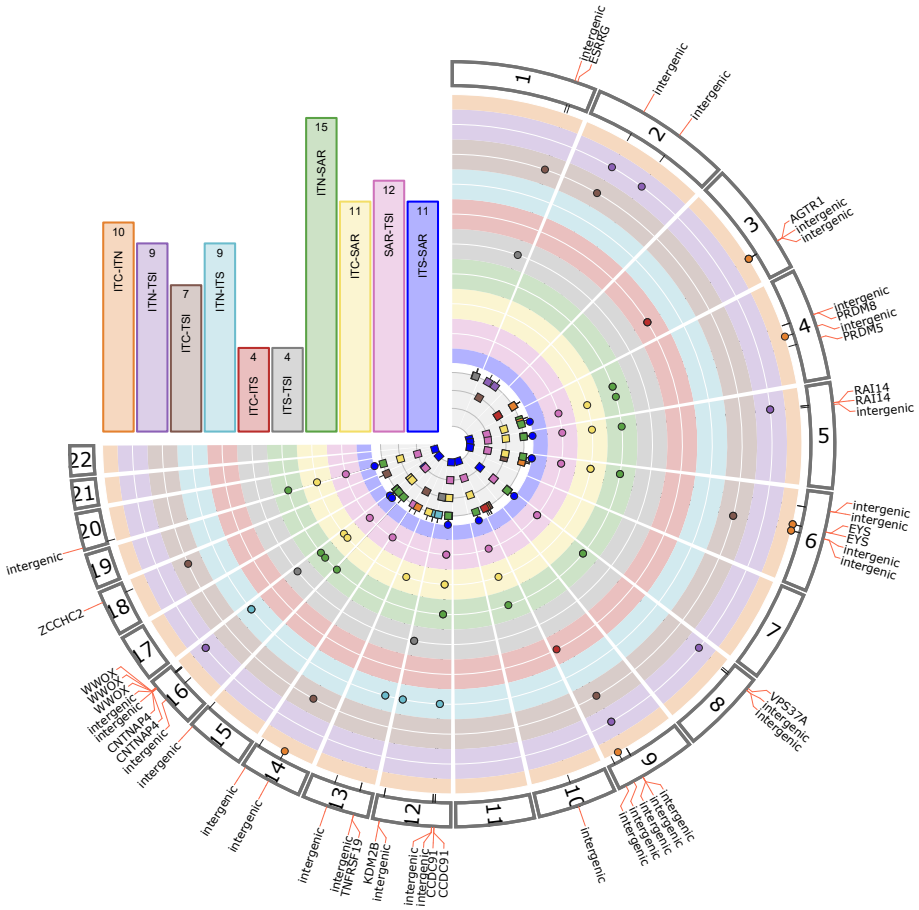


Figure 43. Circos representation of comparisons among Italian populations for Neanderthal alleles. Bars refer to comparison for reported pairs of populations; the number of NTT SNPs is reported within bars. Each section of the circos represents a tested chromosome; points refer to NTT SNPs. Colours, same as for bars; igr: intergenic region variant.

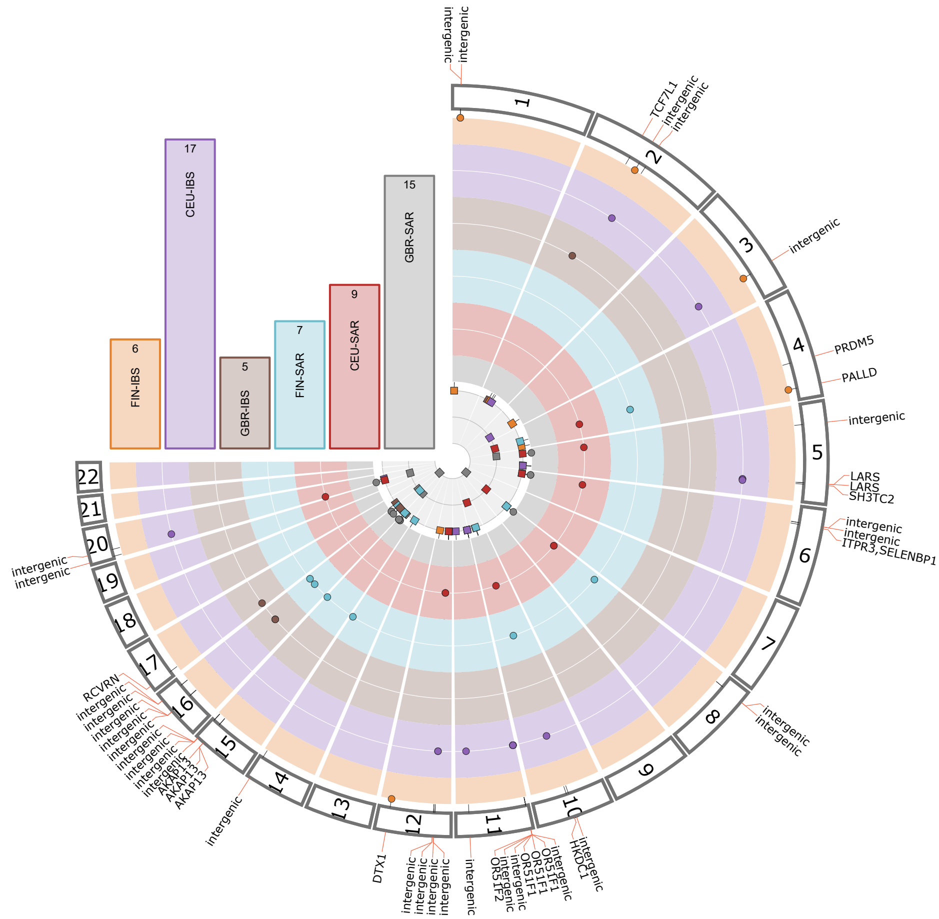


Figure 44. Circos representation of comparisons between Northern European and Sardinian and Iberian populations for Neanderthal alleles. Bars refer to comparison for reported pairs of populations; the number of NTT SNPs is reported within bars. Each section of the circos represents a tested chromosome; points refer to NTT SNPs. Colours, same as for bars; igr: intergenic region variant.

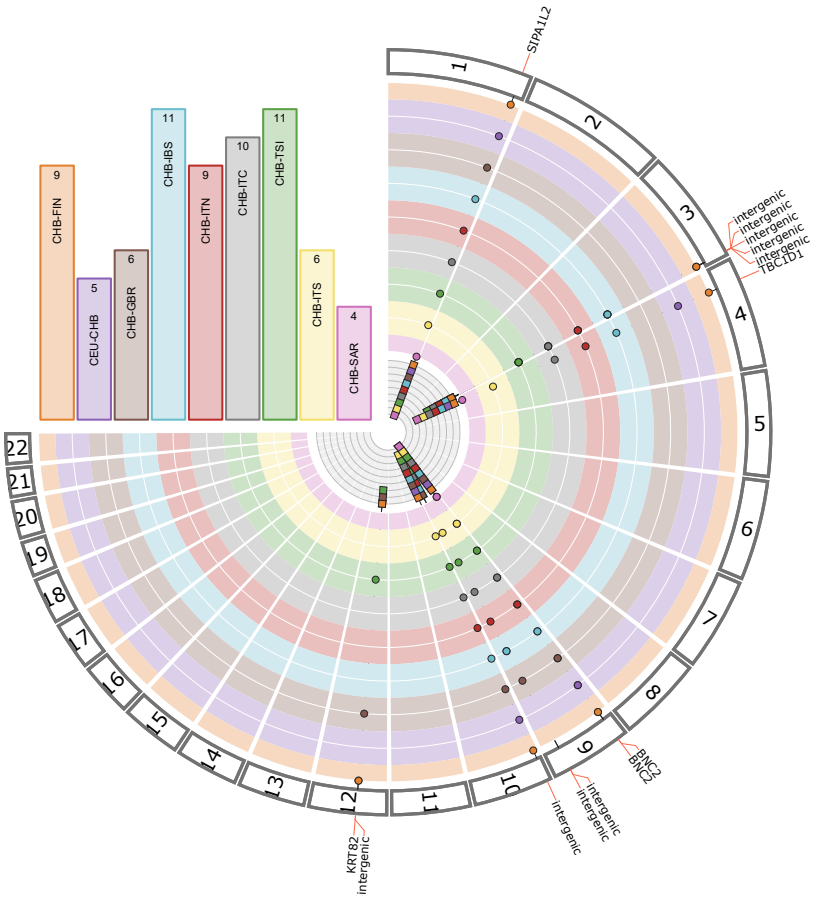


Figure 45. Circos representation of comparisons between European and Chinese (CHB) populations for Neanderthal alleles. Bars refer to comparison for reported pairs of populations; the number of NTT SNPs is reported within bars. Each section of the circos represents a tested chromosome; points refer to NTT SNPs. Colours, same as for bars; igr: intergenic region variant.

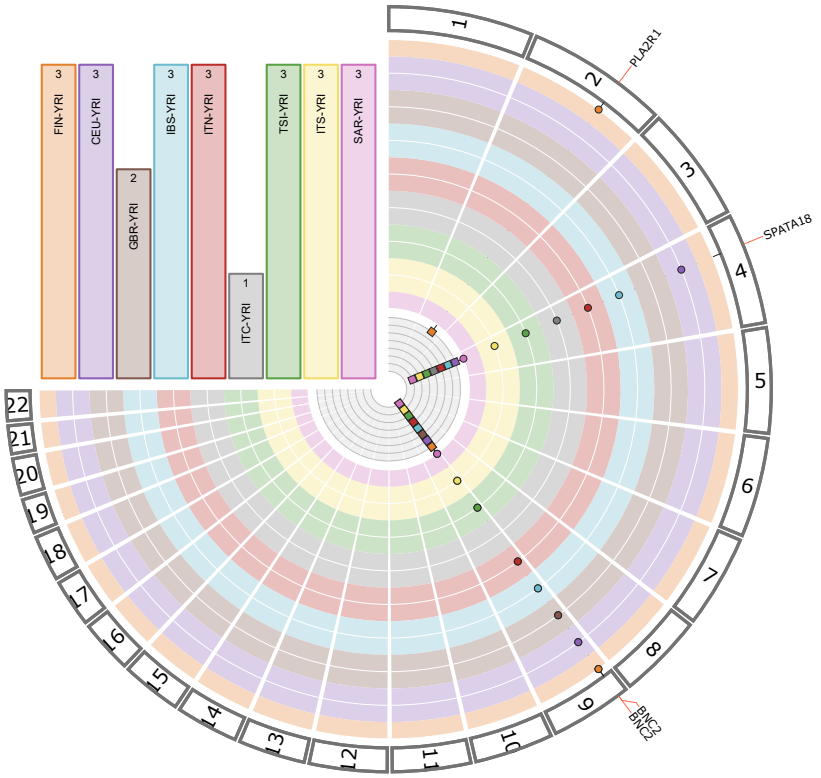


Figure 46. Circos representation of comparisons between European and African (YRI) populations for Neanderthal alleles. Bars refer to comparison for reported pairs of populations; the number of NTT SNPs is reported within bars. Each section of the circos represents a tested chromosome; points refer to NTT SNPs. Colours, same as for bars; igr: intergenic region variant.

Discussion

The availability of high-coverage Neanderthal genomes (Prufer *et al.,* 2014; Prufer *et al.,* 2017) has allowed to precisely quantify our "Neanderthal nature", thus discovering that not all of us are equally Neanderthal. Consistently with previous results (Meyer *et al.,* 2012; Sankararaman *et al.,* 2012; Vernot & Akey, 2014; Wall *et al.,* 2013; Prufer *et al.,* 2017), we found that East Asian populations have significantly more Neanderthal introgression than European population.

When we looked at Europe, we observed a noticeable genetic cline, meaning that the amount of Neanderthal ancestry decreased from the North to the South of Europe. This genetic cline, mirroring the latitude, is also appreciable alongside the Italian peninsula. The two-sample Wilcoxon tests revealed that Northern Europe populations were significantly different from the Southern European populations in their Neanderthal ancestry. Interestingly, significant differences were found alongside the Italian peninsula: specifically, samples from Northern Italy showed higher Neanderthal allele counts than Central and Southern Italian individuals.

It has been previously reported that variation in the effective population size might explain differences in the amount of Neanderthal DNA detected in European and Asian populations (Simonti *et al.,* 2016; Green *et al.,* 2010; Prufer *et al.,* 2014). Additional Neanderthal introgression events in Asia and gene-flow from populations with lower Neanderthal ancestry in Europe possibly provide further explanations for differences in Neanderthal occurrence across populations (Wolf & Akey, 2018). For instance, contributions from African groups could have influenced these patterns, particularly in Southern European populations (Busby *et al.,* 2015).

Moreover, the variation in the amount of Neanderthal DNA among populations could also depend on the differential Basal Eurasian legacy we retain in our genome (Lazaridis *et al.,* 2016; Fu *et al.,* 2016). This ghost population contributed about a quarter of the ancestry of modern-day Europeans and people from Near East. This happened because the early Europeans farmers, which left most of their DNA to modern-day Europeans' themselves, were descendant of an ancestral population that arose early from the main group peopling Eurasia. Interestingly, no ancient DNA from Basal Eurasian people exists, however footprints of this ghost population can be pointed up in our DNA, even when we are not following their tracks, but we are searching for other people: the Neanderthals.

A work from 2016 (Lazaridis *et al.,* 2016), analysing ancient samples living in Near East between fourteen thousand and ten thousand years ago,

discovered that these people harboured around 50% of Basal Eurasian ancestry and that the proportion of their Neanderthal ancestry decreased the more Basal Eurasian ancestry they had.

Consistent with (Lazaridis *et al.,* 2016), we found the estimated amounts of Basal Eurasian and Neanderthal to be negatively correlated across modern-day European clusters. This correlation was still present once populations with African and East Asian contributions were removed. These results suggest that the pattern of Neanderthal ancestry observed in Italy and across Europe has been shaped by different factors: variation in the amount of Basal Eurasian present in the ancient people whose shuffling gave rise to modern-day populations, variation in the ancient ancestry composition of modern samples and variation in the historical contribution from Africa and East Asia.

The spatial heterogeneity of Neanderthal legacy within Europe and Italy reported here appears as the result of ancient and historical events which, during thousands of years, brought together in different combinations groups harbouring different amounts of Neanderthal genetic material. While these events have shaped the overall continental distribution of Neanderthal DNA, locus-specific differences in the occurrence of Neanderthal alleles are also expected to reflect selective pressures acting on these variants since their introgression in the populations (Harris & Nielsen, 2016; Juric *et al.,* 2016; Gittelman *et al.,* 2016; Dannemann *et al.,* 2017).

Therefore, we moved on to an exploration of the Neanderthal genetic legacy in our genome, searching for special signals with high frequency differences across Europe. We focused on the subset of Neanderthal-tag SNPs showing the largest differences in allelic frequency between Eurasian and African populations (NTT SNPs). The variation in Neanderthal ancestry was also evident when SNP frequencies were evaluated and a total of 144 NTT SNPs were identified.

We observed that the top 1% of each distribution was significantly depleted in Neanderthal SNPs (Table A.1), in agreement with a scenario of Neanderthal mildly deleterious variants being removed more efficiently in human populations (Harris & Nielsen, 2016; Juric *et al.,* 2016; Dannemann & Racimo, 2018). In this context, recent studies revealed that Neanderthals had a genetic diversity corresponding to a third of what has been computed for present-day human populations (Castellano *et al.,* 2014), probably due to a sharp reduction of their effective population size, roughly ten times stronger than the out-of-Africa bottleneck. In this context, genetic drift is the main driver of genomic variability and the efficiency of purifying selection in eliminating deleterious variants is dramatically reduced, thus allowing the

persistence of mildly deleterious variants at low frequency (Castellano *et al.,* 2014). Alternatively, or in combination with, such variants might have been neutral in archaic hominins. However, after introgressing in modern human populations, they became exposed to negative selection and were partially removed, due to the larger effective population size of modern humans (Juric *et al.,* 2016). These scenarios of increased effectiveness of selection in modern humans could help to explain the depletion of significant differences in the frequency of Neanderthal inherited variants in present-day human populations.

Unfortunately, given the type of genetic data we worked on, i.e. genotypes from SNP array platforms, most of the variants were located in intronic and intergenic regions. For this reason, the majority of these polymorphisms are unlikely to be directly affecting phenotypes, even if they could be in linkage disequilibrium with the true causing variant (coding/regulatory). This is the main reason why we could not perform a proper annotation analysis reporting the genetic effects of the Neanderthal alleles. However, when we searched in the literature for known direct associations between the Neanderthal alleles of the NTT SNPs and phenotypes, we found 34 variants. Among them, two SNPs seems particularly interesting: within the *BNC2* gene, one of the genes considered to contribute to skin colour (Jacobs *et al.,* 2013), their Neanderthal alleles reach a very high frequency in European (57-73%), while they are absent in East Asian. This is a known example of "adaptive introgression": when modern humans migrated to Eurasia, archaic hominins already lived there for at least 200,000 years, thus carrying adaptive mutations useful to the evolutionary pressures typical of the Eurasian environment (e.g. adaptive introgressed loci with important role in hair and skin biology). After hybridization with archaic hominins, ancient humans acquired some of these mutations that were positively selected and reached high frequencies in modern humans. Another appreciable gift we received form our distant relatives is a specific allele at SNP rs17830558 in the *PLA2R1* gene, which can "protect" us from membranous nephropathy (Stanescu *et al.,* 2011; Sekula *et al.,* 2017).

Other genetic signals for which we found phenotypic association seem less enjoyable: we also found some Neanderthal alleles associated with diseases or clinical conditions, such as cardiovascular and renal disorders (*AGTR1*) and hypopotassemia (*AKAP13*). As said above the decreased effective population size of Neanderthals, together with inbreeding practices in the last Neanderthals, could have given rise to an accumulation of deleterious genetic effect, which, in different cases, could not have been purged in modern humans populations, thus reaching our time.

These results, together with many works in literature (Simonti *et al.,* 2016; Dannemann *et al.,* 2017; Gittelman *et al.,* 2016), demonstrate that our Neanderthal genetic legacy has had both positive and negative effects: several Neanderthal alleles in genes related to immunity, skin and hair pigmentation, and metabolism reached an unexpectedly high frequencies since they provided adaptive advantages; on the other hand, a large fraction of Neanderthal variants appear deleterious being involved in neurodegenerative and metabolic disorders.

In conclusion, archaic admixture has been an important source of adaptive variation in modern humans, making it easier to live and survive in the Eurasian lands of fifty thousand years ago. However, modern humans have, at the same time, inherited genetic variants that today could influence the susceptibility risk to complex disease, due to lifestyle changes and increased life expectancy.

Ancient tales: European wanderings

"Well, well! It might be worse, and then again it might be a good deal better. No ponies, and no food, and no knowing quite where we are, and hordes of angry goblins just behind! On we go!"

J. R. R. Tolkien, *The Hobbit*

Who are the Europeans?

It must have been frightening for the first African explorers to reach the new European territories and lands. They were so different from the endless grasslands and deserts of Africa; there were even big animals, new plants and other people, different from them — the Neanderthals. However, they decided to go further, travelling across those lands in search of food and better living conditions. Having made this choice, some of them became our ancestors, the first Europeans.

Where did these people come from? Who did arrive after them? What did they leave to us? An increasingly accurate genetic portrait of the European populations has been emerging in the last few years from the joint analyses of cultural transitions and the pattern of modern and ancient genetic variations. The gross edges of our genetic make-up have already been traced thanks to three major contributions, arriving in different periods and from different homelands: the first European hunter-gatherers, the Neolithic farmers and the Bronze Age herders from the Steppe.

However, the thinner lines of the European drawing have not emerged yet, thus leaving some unanswered questions.

The central theme of these questions, in the same way as other worldwide human populations, is migration. Europe is maybe the most remarkable example of a population that was defined by a history of repeated immigrations and internal movements. Who arrives first could leave practically nothing of himself in the individuals who would have lived years later in the same

place. Indeed the first lords of Europe — the Neanderthals — went extinct and have been replaced by other people coming from far away, Africa. Also, those people suffered the same fate, while those who would have left a major genetic footprint in European genomes would have come later from the East: Anatolia and Steppe regions.

Europe has not been made by Europeans only, but by travellers coming, almost always, from far and wide.

The first Europeans — if you don't count Neanderthals!

The first modern human travellers coming to Europe arrived as early as 46,000 years ago. The most likely route they followed was across the land bridge to the West of the Black Sea — which, at that time, was a lake — connecting Asia with Europe. The bravest ones could have also overcome the Caucasus mountains to the East of the Black Sea, and sites along the Don river valley (Russia), dating back to 40,000 years ago, could have been left by them.

Then, they started to spread across the European territories, dividing into different groups and leaving behind them some archaeological footprints. In Southern Italy (precisely, in the Grotta del Cavallo) some modern human teeth dating back to 45,000-43,000 years ago have been found (Benazzi *et al.,* 2011), in southern England (Kent's Cavern) a jaw belonging to a modern human has been recently dated to a period between 44,200 and 41,500 years ago (Higham, 2011), while, in Austria, Aurignacian assemblages dated to 43,000 years ago have been excavated (Nigst *et al.,* 2014) — the Aurignacian style is the characteristic lithic industry made by the early European settlers(see later in the same page and at page 68).

The peopling of Europe took place through multiple waves of hunter-gatherers coming ultimately from Africa, thus generating many different groups — in Russia, Georgia, Bulgaria, southern Europe and England — which were spatially and culturally distinct, but having in common the same style of life: they didn't grow crops or breed animals, they hunted and gathered. This lifestyle may seem quite remote from us, however for nearly the entire history of our species, we lived as foragers. Actually, the last 10,000 years, when we learnt more or less efficiently to control the resources of our planet through agriculture and breeding, are just the blink of an eye when compared to the hundreds of thousands of years spent hunting and gathering.

However, they did not live only to hunt, but they were also creative. Their first lithic industry is the Aurignacian (see page 68), the culture of the people inhabiting nearby the Kent's Cavern. Among the tools built with this style, there are the typical split-base antler point — which appear earliest in the Levant — or flutes made with mammoth and bird bone (Figure 47) and many different figurines representing animals. They made also cave paintings of animals, such as lions, bison and rhinoceros, which disappeared from Europe a long time ago.

Figure 47. 40,000 year old flute from the site of Geißenklösterle made from bird bones. Image credit: The University of Tübingen.

The first Europeans have produced many other lithic cultures, such as the Uluzzian — to which belonged the people living in Grotta del Cavallo around 44,000 years ago (Benazzi *et al.,* 2011). In a later period, between 28,000 and 23,000 years old, appeared the Gravettian culture, which was characterized by the pointed blades.

However, one of the most astonishing cultural expression was left in East Europe. There, around 25,000 years ago, a group of reindeer hunters camped near the city of Vladimir. 25,000 years later, archaeologists found their graves and clothes: thousands of mammoth ivory beads had been sewn to their fabric.

Gravettian style

designating or of an Upper Paleolithic culture, characterized by flint points resembling a pointed knife blade with a blunted back.
Collins dictionary

Winter is coming

As it was for the Neanderthals, hunter-gatherer populations were tiny, around 4,400-5,900 individuals (Bocquet-Appel *et al.,* 2005). Thus, when the devastating Ice Age came, the hunter-gatherers went almost extinct. As the glaciers advanced through Europe, plants, animals and men retreated.

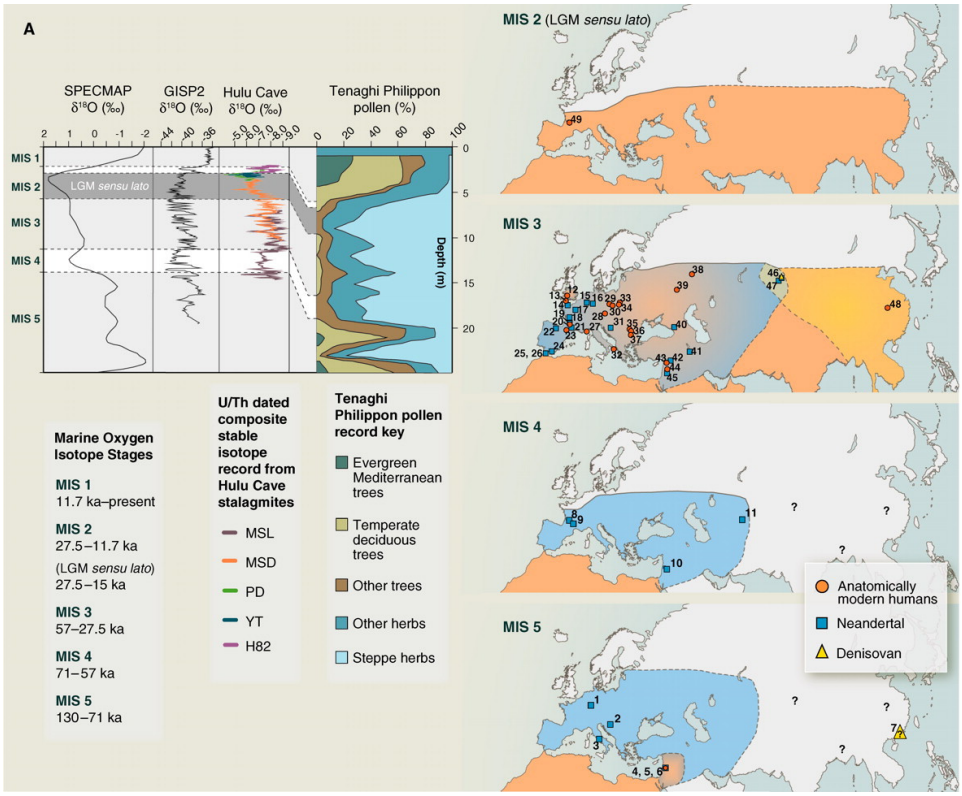


Figure 48. Human distribution in Eurasia between 130,000 and 15,000 years ago in relation to climate change. Image taken from Stewart & Stringer, 2012.

At the peak of this Ice Age, between 18,000 and 20,000 years ago, ice sheets miles thick covered most of northern Europe (Manco, J, 2015). The consequences of the drop in temperature were not exclusive to the poles and the Northern areas of the globe: for instance, the levels of rainfall dropped, thus expanding the deserts and reducing the forests. All around the world, people were pushed to refugia in Southern Europe, such as Spain

and Italy, or Asia Minor and Middle East (Figure 48), resulting, inevitably in population contractions (Stewart & Stringer, 2012).

Then, after the glaciers reached their maximum expansion around 20,000 years ago, they started to retire and the climate gradually improved, during the interstadial period called Bølling-Allerød. However, climate challenges for our hunter-gatherers' ancestors were not over: around 12,700 years ago (Brauer *et al.,* 2008), northern Europe climatic conditions went from temperate to glacial again and many didn't survive.

Foragers from long ago tell their stories

From archaeological evidence, we can understand how hunter-gatherers groups lived around 40,000 years ago, what they eat, how they thought and even who did survive the significant environmental challenges of Upper Palaeolithic. We can even spend a day in the life of these people, but from their remains, we cannot tell where they came from and where their descendants would have gone. DNA can. Thus, putting together the information coming from their burial sites, their tools and their bones, the lines of their portrait we are making, thousands of years later, are becoming more and more confident.

However, some of those lines are dead ends, because the people who started to draw them, tens of thousands of years ago, disappeared without leaving descendants. Oase1, of whom we spoke about Neanderthal admixture (see page 61 for the analyses on his bones and 70 for the DNA results), lived in Romania between 42,000 and 37,000 years ago — the oldest known modern human from Europe. His descendants left no genetic trace into modern-day Europeans, he was a dead branch. Neither Ust'-Ishim, a man who lived in Siberia around 45,000 years ago, showed any affinities with later Europeans (Fu *et al.,* 2016).

The hunter-gatherers' footprints we can find in our genome belong to individuals who lived after 37,000 years ago. One of these is the Kostënki man. Together with GoyetQ116-1 (who lived 35,000 years ago in Belgium), he is the first individual that clearly shared genetic ancestry with Europeans and not with East Asians, thus representing the first, genetically speaking, famous Europeans (Seguin-Orlando *et al.,* 2014; Fu *et al.,* 2016).

The story told by the Kostënki man is particularly interesting. North of the Black Sea, probably along the way taken by the first hunter-gatherers, nearby the village of Kostënki, there is the famous site where Kostënki 14 was found (Figure 49). He was on of those men: 20-25 years old, relatively

short (1.60 m), with a robust skull, a strongly developed supraorbital tori and a short face. As expected given the origin of his ancestors, his facial physiognomy is African. At some time between 38,700 and 36,200 years ago (Marom *et al.,* 2012), he died and was buried. His skeleton was found on its left side with the face to the north, the hands by the face and dark red pigment all around his skull (Seguin-Orlando *et al.,* 2014). In 2014 the DNA extracted from his bones was sequenced to a final 2.42-fold total coverage (Seguin-Orlando *et al.,* 2014) and Kostënki man could finally start to tell his stories. For a start, his genome told us that his ancestors interbred with Neanderthals. In fact, he lived close to the time of the archaic affairs, so much that the longest archaic inherited tract, on chromosome 6, was around 3Mb long!

Then, his DNA revealed that its most substantial fraction derives from a component which will be very abundant in later European hunter-gatherers (Mesolithic HG) and also in contemporary northern and eastern Europeans. However, he shared also genetic ancestry with early middle eastern farmers, suggesting that his ancestors admixed with Middle Eastern populations in their way to Europe. Moreover, its genome is also genetically similar to the genome of a boy who lived in Siberia around 26,000 years ago, called the "Mal'ta boy" (MA1). This boy, in turn, has a strong genetic affinity to modern Europeans and Native Americans. However, the fact that Kostënki shares alleles with European Neolithic farmers and contemporary people from the Middle East and the Caucasus, which are not found in MA1 and in western Mesolithic HG, indicates a gene flow between Kostënki and a Basal Eurasian population (see page 68) after the ancestors of MA1 and European MHGs had split (Seguin-Orlando *et al.,* 2014). However, when Fu and colleagues (Fu *et al.,* 2016) re-analysed this sample, they found no evidence for this putative linkage with the Basal Eurasian population. All these complicated genetic relationships, besides getting us a headache, are simply suggesting that the early modern human populations coming to Europe became very quickly structured and some relics of its genetic components can be traced in modern inhabitants of Europe.

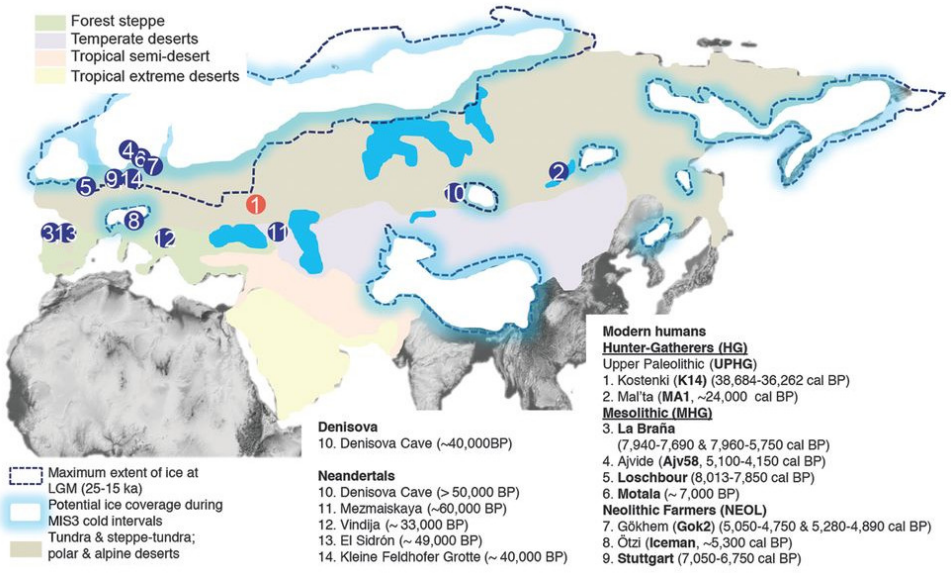


Figure 49. Location of Kostënki and the samples analyzed in the study of Seguin-Orlando *et al.*, 2014. Kostënki (K14) is shown in red, while comparative ancient samples are shown in blue.

In the work previously cited from Fu and colleagues (Fu *et al.,* 2016), they went deep into the putative genetic structure of the first inhabitants of Europe, by analysing other individuals coming from the Ice Age and thus providing a more detailed portrait of the European populations living in that period (Figure 50). By examining the pattern of genetic variations across 51 Eurasians from 45,000–7,000 years ago, they identified five clusters of individuals: the *Věstonice* cluster contained individuals who lived before the Ice Age (34,000–26,000 years ago) associated to the Gravettian culture; the *Mal'ta* cluster is formed by three individuals who lived between 24,000–17,000 years ago nearby the Lake Baikal (Siberia), a region where the Mal'ta–Buret' culture has been associated. The third cluster is *El Mirón* and is composed of seven individuals who lived after the Ice Age, between 19,000 and 14,000 years ago, associated with the Magdalenian culture. The *Villabruna* cluster is more recent (from 14,000–7,000 years ago), while the individuals belonging to the fourth cluster — *Satsurblia* cluster — lived in the Southern Caucasus between 13,000 and 10,000 years ago. Conversely, individuals such as Ust'-Ishim, Oase1, Kostënki 14 and GoyetQ116-1 represented distinct early lineages and were not assigned to any cluster.

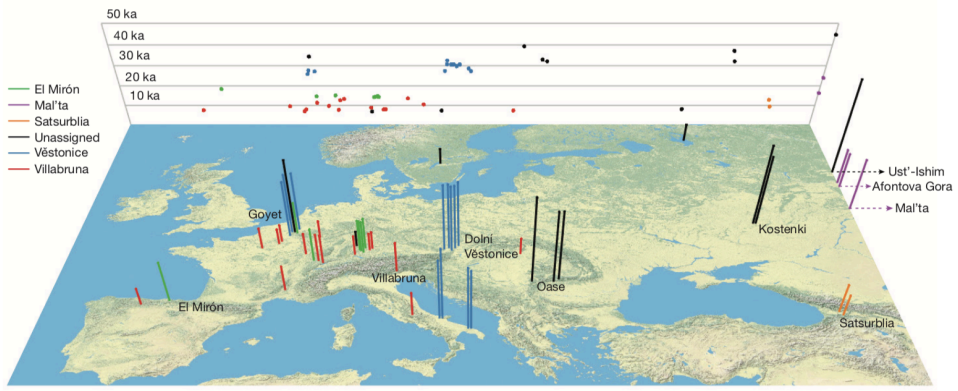


Figure 50. Location and age of the 51 ancient modern humans analysed in the study by Fu *et al.*, 2016. Image taken from Fu *et al.*, 2016.

By analysing the genetic relationship among these clusters and between ancient and modern individuals, they found that from 37,000 to 14,000 years ago — from Kostënki 14 to the Villabruna cluster — all individuals seem to share the same ancestral European component, without traces of gene flow from elsewhere.

However, not for all individuals, there was this kind of genetic continuity: for instance, GoyetQ116-1 is not associated with the Věstonice cluster, which will become predominant in Goyet between 34,000 and 26,000 years ago. While GoyetQ116-1 belongs to the Aurignacian culture, the Věstonice cluster is a Gravettian one, thus showing the spread of the Gravettian culture was also made by populations movements (Figure 51, bottom). However, the ancestry of GoyetQ116-1 did not disappear from Europe: it would have been found later — 19,000 years ago — in the Iberian El Mirón individual. This sample belonged to the Magdalenian culture, whose members over the following 5,000 years would have migrated towards the northeast of Europe (Figure 52, top).

Another interesting point is that after the appearance of the Villabruna cluster — which would have become one of the representative populations of Western Hunter Gatherers (WHG) — around 14,000 years ago, all Europeans analysed in the study show a genetic affinity with Near Eastern populations (Figure 52, bottom). As said in the previous section, this period corresponds to the Bølling-Allerød interstadial, a warm period where the Ice Age loosened its grip. At this time, the thick Alpine glacier which extended down until Nice melted down, thus finally freeing a gateway between Eastern

and Western Europe, a passage that had been closed for 10,000 years. Thus, this genetic similarity could be interpreted as the footprints left by south-eastern Europeans or Middle Eastern individuals who migrated to Europe as climate conditions became favourable. Probably, they brought with them their genes and also new cultures, as that period is also associated with some cultural transitions such as the appearance of the Epigravettian culture in southern Europe and the transition from the Magdalenian to Azilian in western Europe.

Conversely, Eastern Europeans hunter-gatherers (EHG) were a population of mixed WHG and Upper Palaeolithic Siberian ancestry, which appeared in European Russia later, around 8,000 years ago, and then contributed to the hunter-gatherers' population in many Northern European countries (Lazaridis, 2018).

Some beautiful pictures from the book of David Reich are enlightening about the major population transitions happened during the long period when hunter-gatherers populated Europe and wandered across it (Figure 51 and 52).

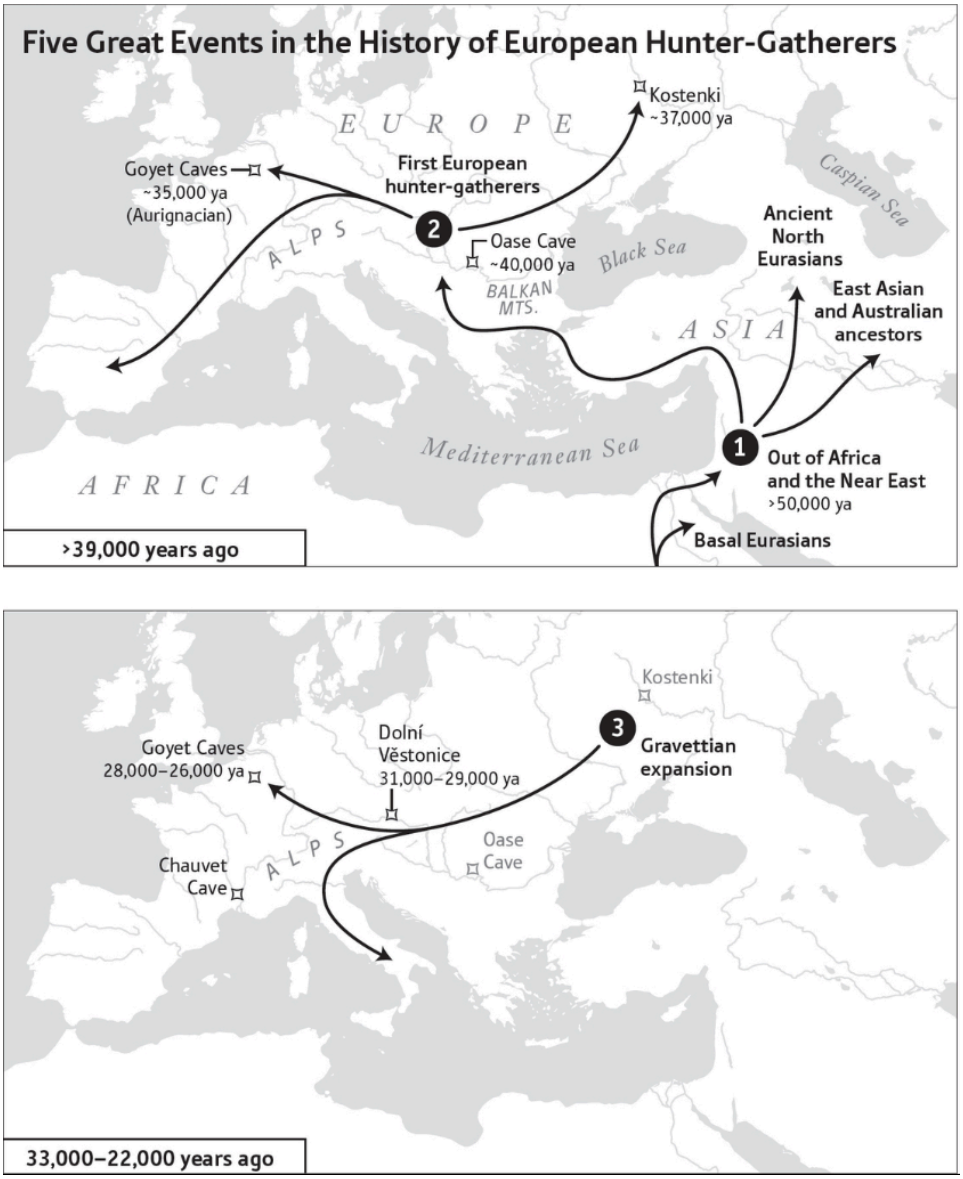


Figure 51. First two big population movements in the history of European Hunter Gatherers. Spreading of hunter-gatherers populations across Europe earlier than 39,000 years ago (top) and the expansion of the Gravettian culture (bottom). Image taken from Reich, 2018.

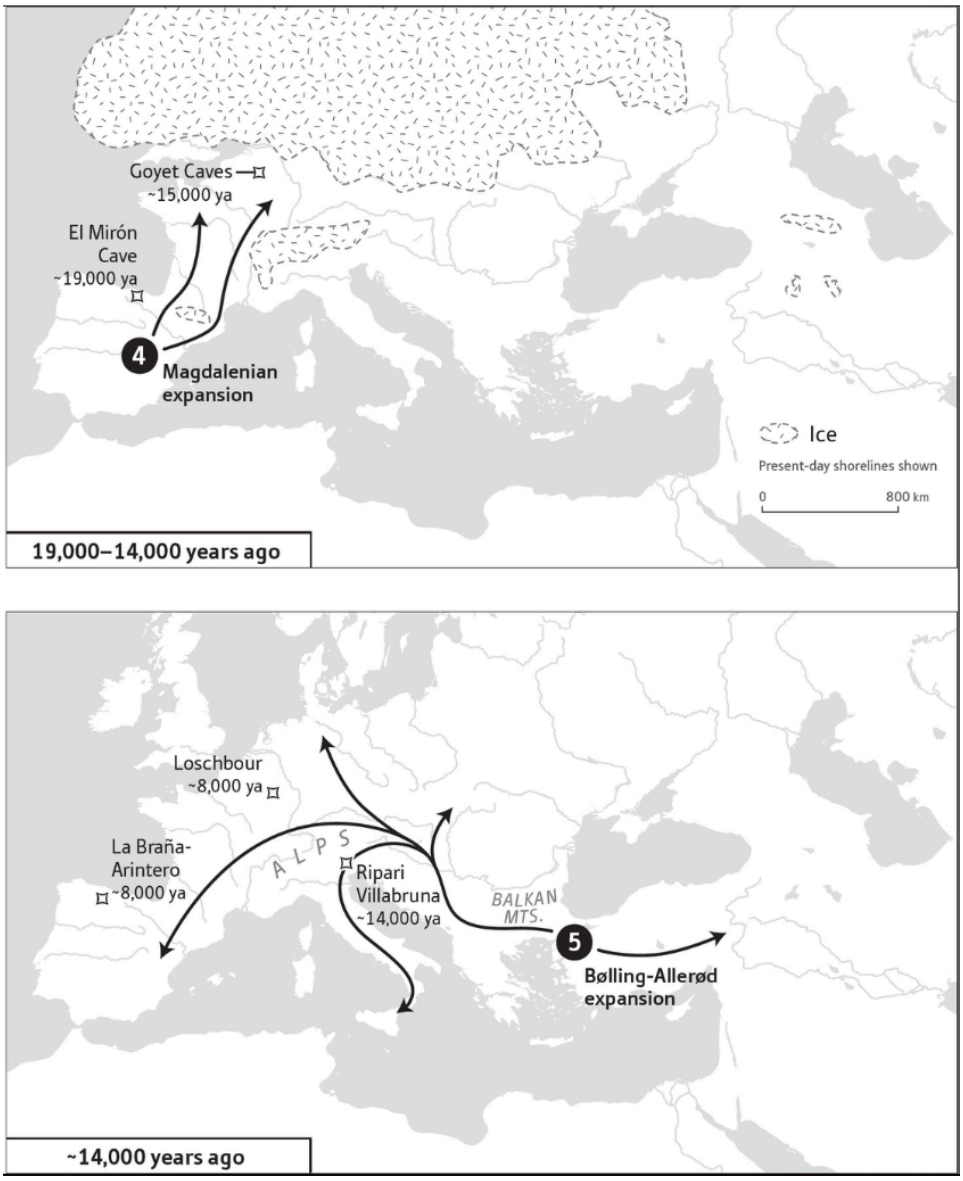


Figure 52. Population movements in the history of European Hunter Gatherers after 19,000 years ago. The Magdalenian expansion (top) and the movements of people after the end of the Ice Age from the East. Image taken from Reich, 2018.

Defrosting innovations

After the glaciers started to melt around 10,000 years ago, the ancient foragers roused from the movement lethargy in which they had fallen thousands of years earlier. We have already seen that a gradual increase of Near Eastern ancestry in European populations corresponds to the Bølling-Allerød interstadial, as a result of migrations back to Europe from the Eastern refugia. During this, from a migratory point of view, euphoric period, some foragers ranged northwards. They were still using stone tools, but with a slightly different style. For instance, they built tiny bladelets and put them on the top of harpoons. These characteristic tool kit allowed the archaeologists to recognise and classify the sites where these objects had been found as Mesolithic sites (Manco, J, 2015).

This warm period was also fruitful of human inventions, such as the regular use of boats in northern Europe: this would have led to new migratory trails across the world and opened the floodgates for sea exploration.

Figure 53. Carvings of people and a boat. Image credit: Dimit via Panoramio.

The first images of boats have been found on the coast of the Caspian Sea (Azerbaijan) and date back to 12,000-8,000 years ago (Figure 53).

Another great invention was the introduction of pottery in the Far East, which would have been very useful for collecting, storing and also cooking food. However, Europeans had to wait for the Neolithic period to benefit from it.

Around 10,000 years ago, European foragers were living carelessly hunting and fishing, but one of the biggest human revolutions was knocking on the door.

Wheat and goats

The change from foraging to farming has been one of the greatest revolutions of humanity. The possibility to control the food resources allowed to feed a higher number of individuals and the very next consequence of this was a population explosion. The population growth led, in turn, to technological and cultural innovations and, finally, the rise of the first civilizations. The world was beginning to look like the one we know today.

The first attempts at farming began around 12,000 years ago in the Near East and, precisely, in southeastern Turkey and northern Syria (Figure 54). In that region, where the Levant meets Anatolia, foragers were already using the plants that were naturally growing, such as wild wheat, barley and rye. Also, peas and lentils come from those same areas. The first evidence of cultivation comes from Cafer Höyük and Çayönü, near the headstreams of the Tigris and Euphrates, however, we will have to wait for the climate to improve, around 8,000 years ago, to have the actual crop domestication in that area. In the meantime, sheep and goats were domesticated in the Caucasian mountains from the northern Zagros to southeastern Anatolia.

The domestication of other plants and animals required more time: olive trees were domesticated by 7,000 years ago, horses by 6,000 years ago, while grapevines around 5,500 years ago.

Since farming began before the introduction of pottery, the earliest farming period is known as Pre-Pottery Neolithic A (PPNA). During this time, the domestication of plants and animals caused a rapid population growth, however, an inevitable consequence of this was the spreading of diseases, boosted both by the crowded villages and by the close contact with animals.

At the same time, to accommodate an ever growing population, villages turned into cities, while temples, megalithic monuments and other buildings

were erected, the first of which was unearthed at Göbekli Tepe, in Turkey.

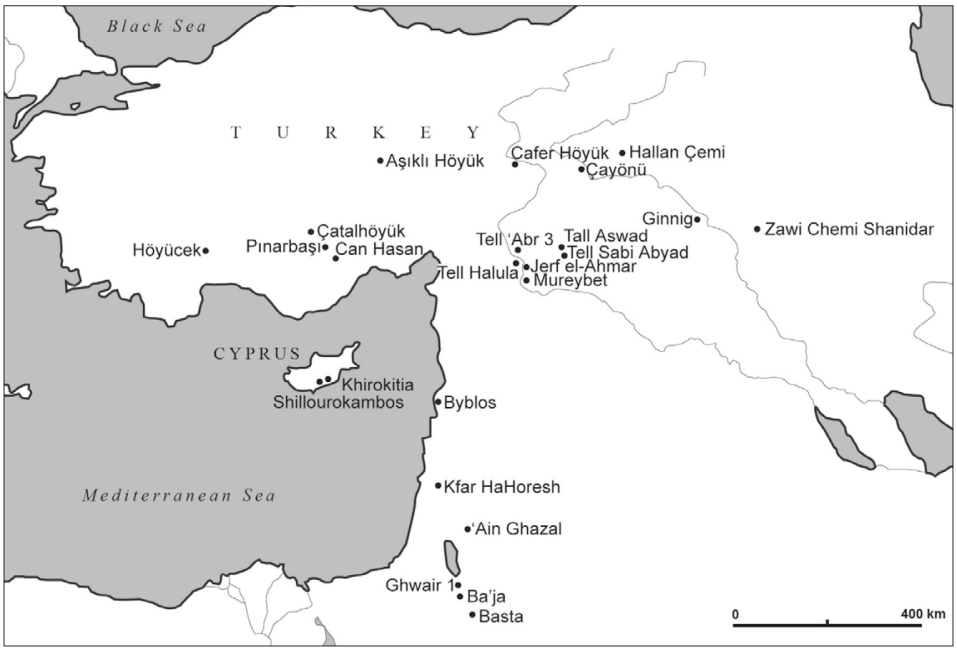


Figure 54. Map of Southwest Asia, indicating the important sites for the development of agriculture. Image taken from Russell *et al.,* 2009.

Farming began spreading around 11,000 years ago and almost immediately reached Cyprus (Vigne, 2011). After around 500 years began the period conventionally called Pre-Pottery Neolithic B (PPNB). In this case, as Cyprus had no goats, sheep, pigs or cattle, the farmers who would have settled there should have also brought stock and seed with them (Manco, J, 2015). For this reason, at least in this case, the spreading of farming was due to a migration of people.

After around 9,000 years ago farming started to spread far from its home: towards the west, farmers sailed from Cyprus to Crete, arriving at Greek mainland around 8,500 years ago. They also moved eastward, reaching the Indus Valley (Pakistan) slightly earlier.

However, climate changes were always lurking and hyper-arid conditions struck Europe, North Africa and southern Near East around 8,200 years ago. Farmers abandoned Cyprus as well as other farming sites in the Near East and started to look north.

Farmers in Europe

Farming arrived in Europe slightly after the climate crisis, leaving indelible marks on the European environment, on the pottery styles and in our genome.

At least four different trails had been used by "farming" to enter in Europe. One of these, brought to Corfu (western Greece), then to the Adriatic coasts and, along the Mediterranean coast, to Spain. The second followed the Danube Valley, towards Germany and, from Germany, it arrived in Scandinavia and the British Isles around 6,000 years ago. A third trail, tracked thanks to cattle DNA (Decker *et al.,* 2014), passed thought Northern Africa and then reached Iberia. The last route, instead, was through the sea.

As previously said, the pottery cultural transitions proved to be useful in retracing those ancient movements. In fact, the first two trails could be easily tracked thank to the pottery left on the way: Impressed and Linear (Figure 55).

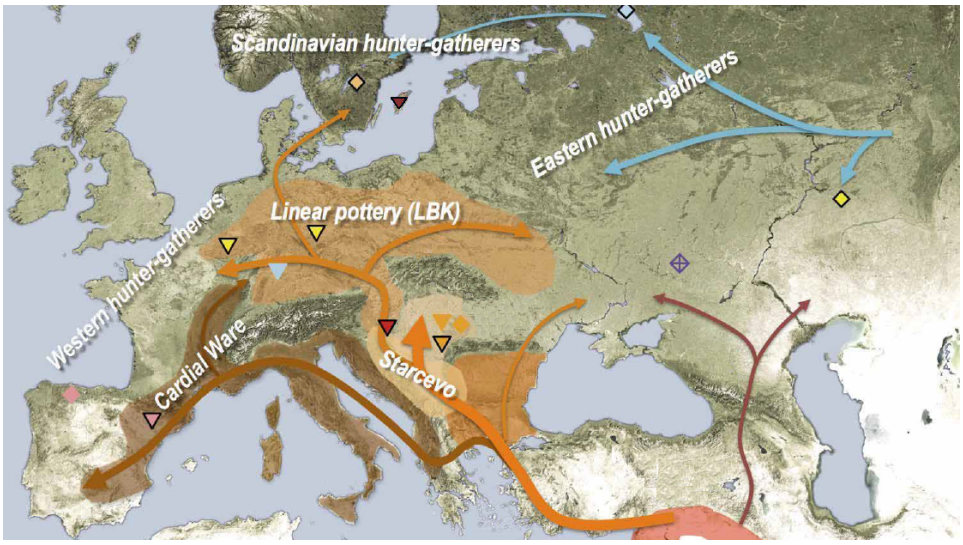


Figure 55. Proposed routes of migration by early farmers into Europe 9,000-7,000 years ago. Image taken from Haak *et al.,* 2015.

For instance, the Mediterranean route can be tracked by following the Impressed Ware, which is characterized by the patterns pressed onto the pots. In some regions, the culture producing this kind of pottery is known

as Cardial Ware, whose sites would have been found in Sardinia, Corsica, on the Ligurian and Tuscan coasts of Italy, France, Spain and Portugal.

The second route, through the Danube valley, could be followed by looking for Linear pottery, instead, which had then converged into the Linearbandkeramik (LBK) culture.

However, who rode the farming wave? People or ideas? The simplest explanation, since in Europe there weren't sheep or goats, is that someone moved to Europe bringing the cattle himself. However, migratory explanations fought for almost a century with anti-migrationism, until genetics came into play. We have already mentioned the pioneering work of Cavalli-Sforza (see page 21) and, after him, other papers had been piled up supporting the migration of farmers into Europe (Skoglund *et al.,* 2012; Skoglund *et al.,* 2014; Lazaridis *et al.,* 2014; Gamba *et al.,* 2014; Olalde *et al.,* 2015; Günther *et al.,* 2015; Cassidy *et al.,* 2016). Actually, these works demonstrated that early farmers from different parts of Europe showed a clear genetic differentiation from the foragers inhabiting the same areas: the difference among farmers and foragers were almost as marked as between populations from different continents (Skoglund *et al.,* 2014). Thus the farmers where, genetically speaking, strangers: they surely must come from distant lands.

A peculiarity of the incoming farmers is that, while Mesolithic Europeans showed some genetic similarities with modern-day northern and northeastern European populations, early farmers coming from the Middle East have higher genetic affinities with southwestern European populations, but not with modern-day Middle Eastern populations (Skoglund *et al.,* 2012; Skoglund *et al.,* 2014; Lazaridis *et al.,* 2014). Intriguingly, as told by the Tyrolean Iceman called Ötsi, the present-day individuals carrying on the Neolithic farmers' legacy are the Sardinians (Skoglund *et al.,* 2012; Keller *et al.,* 2012). Ötsi lived during the Late Neolithic-early Copper Age and, despite being found on the Ötztal Alps at the border between Austria and Italy, his closest living relatives can be found in Sardinia (Keller *et al.,* 2012).

However, where did the Neolithic component found in Sardinian and other southwestern Europeans come from? After 2016 it became possible to obtain high-quality DNA from human remains found in the Near East, thus allowing to start drawing a finer portrait of the ancient farmers living there. In these lands, the environmental conditions are not optimal for DNA preservation. Fortunately, some innovations saved the day: a new method for enriching DNA extracted from the bones (Fu *et al.,* 2013) and the discovery that it is possible to obtain a higher DNA density from the inner-ear part of the skull — the petrous bone — than from other skeletal parts (Pinhasi *et al.,* 2015). Some works by the group of David Reich and

Anders Götherström analysed the genetic variation of early farmers from Anatolia and the Levant, discovering that these populations were the source of the Neolithic component found in Europeans (Mathieson *et al.,* 2015; Omrak *et al.,* 2016; Lazaridis *et al.,* 2016; Kılınç *et al.,* 2016).

The work of Iosif Lazaridis and colleagues (Lazaridis *et al.,* 2016) was particularly eye-opening about the Near Eastern population from which farming was born and spread elsewhere. They found a high degree of genetic differentiation in the Near East. There lived the farmers of the western mountains of Iran, who derived from the hunter-gatherer populations living in the same region. Then, there were the farmers living in the Fertile Crescent, who genetically derived from the autochthonous hunter-gatherer populations, the Natufians. Besides sharing the same lifestyle — farming — these two groups were as genetically different as are present-day Europeans and East Asians, thus demonstrating that at least here the diffusion of farming happened by the spread of ideas (Lazaridis *et al.,* 2016). However, Lazaridis found that this high degree of genetic differentiation was not a Near Eastern peculiarity, but it was shared all across Western Eurasia. In fact, they found that at least four major distinct populations inhabited West Eurasia at that time: the farmers from Fertile Crescent, the farmers of Iran, the hunter-gatherers of central and western Europe and the hunter-gatherers of eastern Europe (Figure 56). Again, these four main groups were so genetically distinct among themselves as much as Europeans differ from East Asians today (Lazaridis *et al.,* 2016; Reich, 2018).

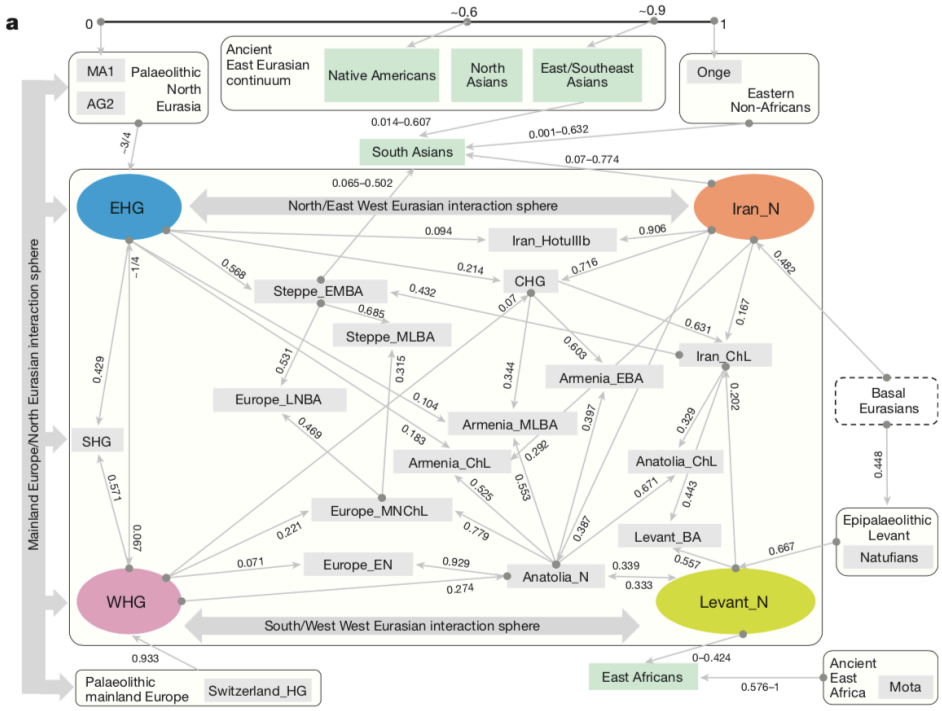


Figure 56. Model of the genetic components of ancient West Eurasians, East Africans, East Eurasians and South Asians. All the ancient populations can be modelled as mixtures of two or three other populations and up to four proximate sources (marked in colour). Image taken from Lazaridis *et al.,* 2016.

Moreover, they also found that the impact of Near Eastern farmers extended far beyond the Near East itself. For instance, the farmers of Anatolia spread into Europe, farmers of the Levant moved towards East Africa, Iranian farmers went northward into the Eurasian steppe, and some of the Iranian farmers and the pastoralist from the Eurasian steppe migrated towards South Asia (Lazaridis *et al.,* 2016).

A clash between cultures

The farmers arriving in Europe around 8,800 spread to Iberia and Germany without mixing, or minimally mixing, with the autochthonous hunter-gatherers living in those places. In fact, European Early Neolithic individuals retained at least 90% of their "original" DNA. This attitude changed in

the farmers living between 6,000 and 4,500 years ago: their DNA showed that they acquired around a 20% ancestry from the hunter-gatherers, thus suggesting that the two groups mixed together (Haak *et al.,* 2015; Lipson *et al.,* 2017). Little by little, as farming became the predominant style of life, the hunter-gatherers were gradually assimilated into the farmer groups, as you can see from Figure 57, where Late Neolithic individuals ("*_LN*") show the same blue component of hunter-gatherers ("*_HG*").

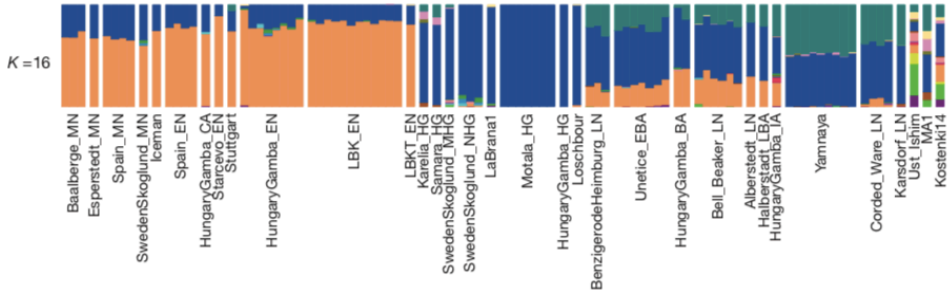


Figure 57. ADMIXTURE analyses of ancient individuals. Image taken from Haak *et al.,* 2015.

Coexistence and mixture were possible. An interesting example is Körös 1 (KO1), a blue-eyed male lived around 7,500 years ago in Hungary in one of the first farmer settlements. However, his DNA tells a different story: he was genetically similar to hunter-gatherers, so he was probably a first-generation hunter-gatherer migrating into a farmer settlement (Gamba *et al.,* 2014).

How could these two different cultures have coexisted? Together with KO1, other clues arrive from the Funnel Beaker culture. The name comes from its characteristic pottery — the funnel shaped clay vessels found in their graves (Figure 58). This culture appeared in Northern Europe around 6,100 years ago. Those lands were not immediately reached by farmers, probably because of the cold weather and the heavy soils of Northern Europe. While they gradually took some of the farmers' innovations from their southern neighbours (farmers belonging to the LBK culture as in Figure 55), such as the domestication of animals and crops, they kept also some of their hunter-gathering habits (Reich, 2018). However, around 6,000 and 5,000 years ago, farmers reached also those regions, they mixed with the residents and the mixing between the two cultures would have genetically characterized the Funnel Beaker populations. Actually, their DNA says that they are a mixture between a little amount of hunter-gatherer ancestry and a big portion of Anatolian farmer ancestry, while their culture says the opposite

— some farmer innovations, but many more hunter-gatherer activities.

Figure 58. A Funnel beaker clay pot from 6,000-4,800 years ago.

A beautiful map resuming the arrival of farming and the population upheavals that followed is represented in Figure 59.

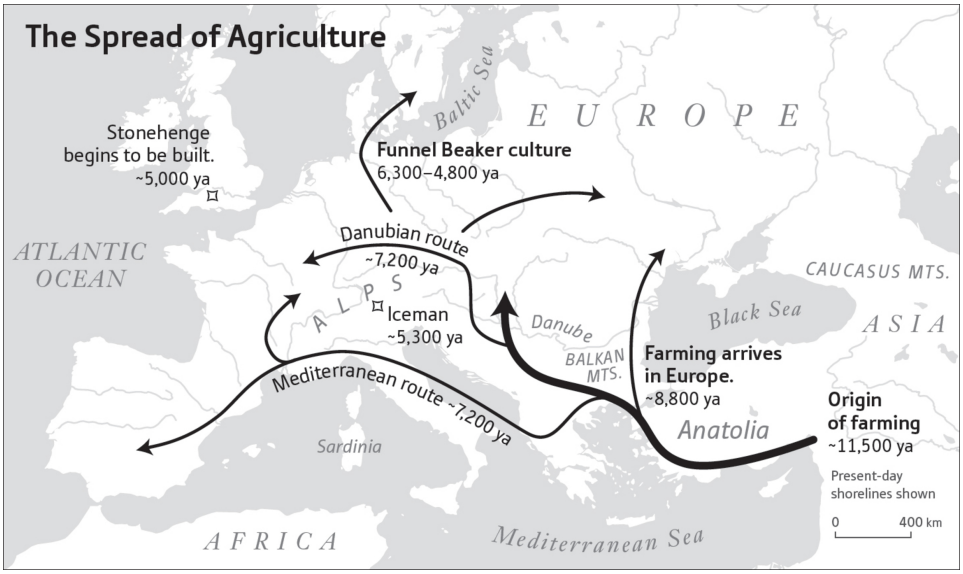


Figure 59. The spread of agriculture across Europe. Image taken from Reich, 2018.

Shining metals

The Bronze Age started around 5,000 years ago with the rise of the first cities in Mesopotamia. From that time on, copper alloys, such as bronze, had been used for building agricultural tools, weapons, personal ornaments, metal cauldron and cups for cooking, dining and drinking (Allentoft *et al.,* 2015). However, even before the introduction of smelting, people were using and working copper, attracted by its bright colours. Since 12,000 years ago, copper was used for making beads and ornaments, but the discovery that heat made copper-working easier was made only 10,000 years ago (Manco, J, 2015). Then, around 7,000 years ago, smelting copper objects started to appear and spread across Anatolia and Levant.

From urbanization, thus come trades, whose routes for selling copper-made objects extended towards the Caucasus Mountains. There, the commerce brought wealth and metallurgy: the result was the Maikop culture (5,700-5,100 years ago).

Figure 60. Copper axe of the Iceman. Image credit: Museo Archeologico dell'Alto Adige

In Europe, the first metal to be worked was copper in the Balkan region.

In fact, Balkans had deposits of copper and gold and the first attempts with bronze appeared as early as 6,500 years ago by using copper ores which naturally contained tin. Also, the Alps were rich in copper, so much that some copper age cultures arose in Italy, at Remedello and Rinaldone in the North and Gaudo in the South (Mallory & Adams, 1997). The objects of these metalworkers reached even Ötsi, the Iceman: he carried a copper axe of the Remedello type, made in northern Italy using the ores of Tuscany (Figure 60).

Horses, wheels and wagons from the East

In the meantime, nobody was wandering across the Pontic-Caspian steppes before 5,000 years ago, probably because there wasn't enough rain to support agriculture. But then, everything changed with the appearance of the Yamnaya. They were nomads, living in tents and moving by wagons. Through their economy based mainly on sheep and cattle herding, they could survive in a land so unsuitable for agriculture. This population emerged from the previous cultures living near the steppes, with substantial cultural and genetic influxes from the Maikop culture.

One of their distinctive elements was the round tumulus or barrow — the *kurgan*. From a cultural point of view, this typical burials represented an important step in the direction of modern civilization, because, for the first time, single rather that collective graves were introduced (Manco, J, 2015). The Yamnaya people wore woven clothes, gold or silver spiral hair rings and bone ornaments. Both men and women used hair binders on the ends of the braids, as the Trojan hero Euphorbus did in the Iliad (Manco, J, 2015).

Gradually, elements of their culture began to be found in Europe and Central Asia, up to the foothills of the Altai Mountains, thus testifying once again that migration was a building block of ancient civilizations. One of the inventions allowing Yamnaya to leave their footprints on such a big area was the the wheel. It is currently not known where wheel was invented: it appeared a few hundred years before the Yamnaya civilization, but as soon as it was introduced, it spread across Eurasia so fast that it is almost impossible to follow its route (Reich, 2018). Wheels were so useful when put together with wagons, which, probably, were a gift of the southern neighbour culture — the Maikop. These carts, together with the domestication of horses, made a new economy based on cattle herding and nomadism possible. Wheels, wagons and horses changed profoundly their lifestyle, leading to the

abandonment of the village life.

Their major technical advancements, such as the horse riding, wheeled transport and metalworking, spread rapidly across Europe and Asia, raising the idea-people dilemma again.

Another change in the climate conditions, around 5,200 years ago, could have been a boost for migratory processes from the steppe towards Europe, making the debate leaning towards the "people" hypothesis. At this point, we should have learnt which is the winning horse: genetics.

In fact, the DNAs of those people showed that they were just the missing piece of the genetic mosaic that makes the European people (Haak *et al.,* 2015; Allentoft *et al.,* 2015).

Corded Ware and Bell Beaker — pots vs people?

The already mentioned work from Haak and colleagues (Haak *et al.,* 2015) showed that the Yamnaya population were a mixture of different ancestries: eastern European hunter-gatherers (EHG) and an Iranian-related population. The latter component could derive from the southern population which, together with some genes, give them also the wagons — the Maikop.

However, this strange jumble of ancestries appeared also in Central Europe with the Corded Ware culture (Haak *et al.,* 2015; Allentoft *et al.,* 2015). This culture was named after their characteristic pottery decorated with cords and twine impressed on soft clay, which started to appear and spread in a vast zone, from Switzerland to European Russia, around 4,900 years ago (Figure 61). They shared with the Yamnaya culture many traits, such as the large burial mounds, the intensive use of horse and herding, a male-centred culture and the finely executed copper axes (Reich, 2018; Haak *et al.,* 2015). However, given that they were different for many other cultural expressions — first of all, the pottery style — the "ideas-people" debate was raised again with regard to the relationship between Corded Ware and Yamnaya. The group of David Reich analysed DNAs belonging to people buried in Germany with Corded Ware pots. They discovered that about three-quarters of their genome derive from Yamnaya-related populations, while the remaining part has been inherited from the local farmers. Figure 57 shows clearly the similarity between Yamnaya (called "*Yamnaya*" in the admixture plot) and the Corded Ware (called "*Corded_Ware_LN*", where "*LN*" stands for "Late Neolithic") individuals (Figure 57). Moreover, they discovered that with the Corded Ware culture, European people at that time

were beginning to look like modern Europeans, genetically speaking (Haak *et al.,* 2015; Allentoft *et al.,* 2015).

Figure 61. Corded ware pottery. Image credit: Getty Images/DeAgostini

Around 4,700 years ago, another culture spread across Europe, probably starting from Iberia — the Bell Beaker culture — and then covered a vast area. Bell Beaker sites were located in Poland, northern Morocco, Scotland, northern Denmark and even the southern part of Norway (Manco, J, 2015), and, after 4,500 years ago, also in Britain. As for other archaeological cultures, the name comes from their pottery: the bell-shaped drinking vessels (Figure 62). Other characteristic traits of this culture were the plough, wheeled vehicles, woolly sheep and the use of the horse, together with decorated buttons and archers' wristguards (Manco, J, 2015; Reich, 2018).

The "ideas-people" dilemma did not spare even them. According to the isotopic composition of teeth, it seems that some people of the Bell Beaker culture had moved hundreds of kilometres from their places of birth (Reich, 2018; Fitzpatrick, 2011). However, while the Corded Ware people were all genetically similar across Europe — recent studies have found this comparing Corded Ware people from the Baltics with those from Central Europe (Mittnik *et al.,* 2018; Saag *et al.,* 2017) — Bell Beaker were heterogeneous.

The analysis of Iberian Bell Beaker told that those individuals were genetically indistinguishable from the people who lived there earlier (Olalde *et al.,* 2015). Conversely, Bell Beakers from Central Europe showed a considerable amount of their ancestry deriving from the steppe populations. Thus, DNA suggests that the first spreading of Bell Beaker culture from Iberia eastwards was uniquely mediated by the movement of ideas and not by the movement of people. However, once Bell Beaker culture reached Central Europe, the subsequent waves had been driven by migration: after 4,500 years ago, in England, all individuals buried in a Bell Beaker context showed steppe ancestry and no genetic affinities with the Iberian Bell Beaker (Figure 63, Olalde *et al.,* 2015).

The mosaic of the European DNA was apparently completed in 2018 with the discovery that the Yamnaya, Corded Ware and Bell Beaker migrations contributed together in spreading the high level of steppe ancestry we see in modern-day Europeans, especially in Northern and Central Europe. However, even if we have certainly found the corner pieces and the tesserae composing the main figure, we still lack some fine details and we are far from putting together the blue sky part.

Figure 62. Beaker pottery vessel (2,500–2,150 BC, Oxfordshire). Image credit: The Ashmolean Museum, University of Oxford.

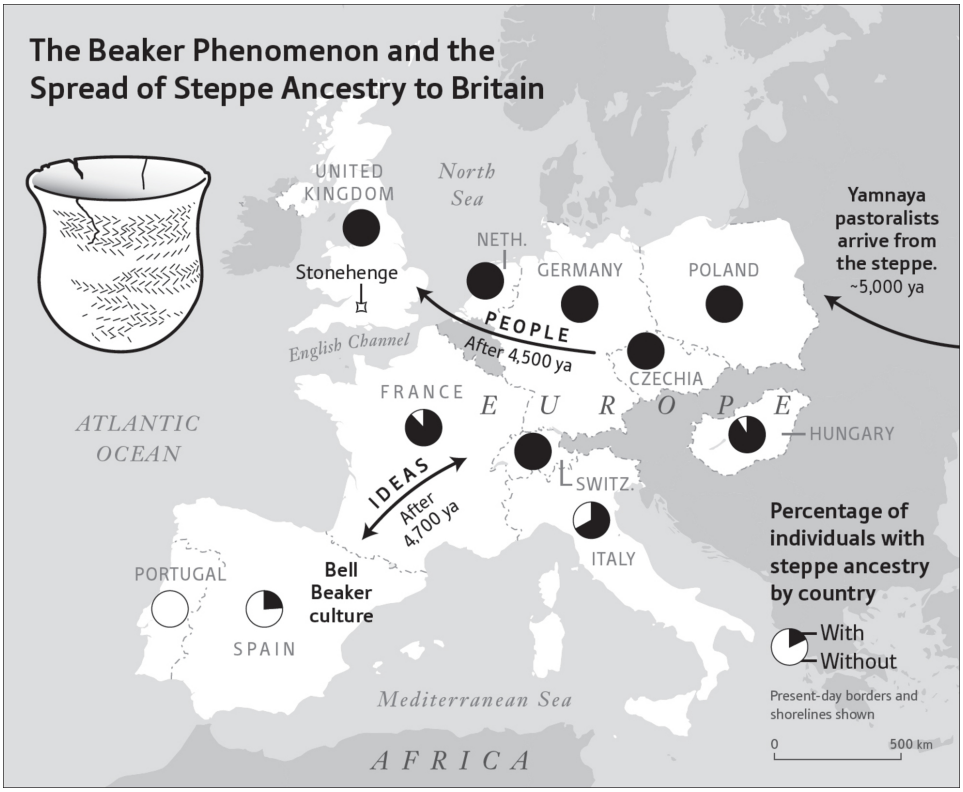


Figure 63. The spread of the Bell Beaker culture across Europe. Image taken from Reich, 2018.

Something more in Southern Europe

Southeastern Europe was the first European region to welcome the steppe ancestry, with sporadic individuals showing such component in Bulgaria as early as 6,700-6,500 years ago (Mathieson *et al.,* 2018). During the Bronze Age (5,400-3,100 years ago), almost everyone there harboured around 30% of steppe ancestry.

Moving further south, the steppe component was present in the Aegean during the Mycenaean period, that was around 3,500 years ago, in a substantial amount (∼15%), however, it was absent from the genetically similar Minoan culture (Lazaridis *et al.,* 2017).

However, before going in the genetical matters, it is worthwhile to spend a few words on these two civilizations. The Minoan culture of Crete was far more sophisticated than the Early Bronze Age people from the continent

— the ancestors of the Hellenes. They left writing dating from 4,100 to 3,600 years ago (Ferrara, 2010; Manco, J, 2015), built decorated communal buildings and used the wheel-thrown, a sign of mass-production pottery. However, where did they come from? If they were the descendants of the earlier farmers inhabiting Crete, they would have been the first civilization with European roots (Manco, J, 2015)! Neither their language, not the archaeological remains, could provide a final answer. Indeed, even if their script has never been fully deciphered, some scientists think that it is not Greek (Ferrara, 2010). The archaeological evidence seem to be more inclined to say something: around 5,000 years ago, when Neolithic meets the Bronze Age, the archaeological culture of Crete went through a period of big changes, suggesting an influx of immigrants from Anatolia on the island. However, it was not a complete population replacement because, at least in the interior of the island, it is evident the continuity from the Neolithic (Legarra Herrero, 2009).

Moving on mainland Greece the Mycenaean society, Bronze Age people from the northeast arrived in those regions, as testified by the presence of their ancestry in ancient Mycenaean people (Lazaridis *et al.,* 2017) and by the founding of Yamnaya stelae connecting the steppe populations with the ancestors of the Mycenaean Greeks. However, no other evidence has turned up, thus keeping the mystery alive.

By 3,900 years ago, the Minoan civilization reached its maximum cultural development, but a catastrophe was about to happen: a volcanic eruption dated back to 3,600 years ago on the Santorini island buried part of Crete. At that time, exploiting the Minoan weakness, the Myceneans started to control Crete. However, the situation would not be better. Around 3,200 years ago, the main Greek centres were destroyed and abandoned, there was a sharp reduction in the population, the writing ceased and a return to the village life (Manco, J, 2015). The responsible for that demise was probably, once again, the climate change, in the form of a long arid period, which lasted until the Romans.

The recovery of the Greek population and agriculture started around 2,900-2,800 years ago, along with the adoption of iron (Manco, J, 2015). It was at this stage that the vast expansion of Greece to southern Italy, Sicily, Corsica, Provence, Asia Minor and the Black Sea took place (Figure 64), through the which they left, together with cultural and monumental traces, a great genetic footprint, too.

Returning to genetics, Iosif Lazaridis analysed 19 Bronze Age individuals, including Minoan from Crete, Mycenaean from mainland Greece and southwestern Anatolia (Lazaridis *et al.,* 2017). They discovered that Minoan

and Mycenaean individuals were genetically similar, carrying at least three-quarters of their ancestry from the first Neolithic farmers inhabiting those areas, while the remnant part was ultimately connected with ancient populations from Caucasus and Iran. However, while Mycenaean individuals also harboured an EHG-related component, Minoans do not show this ancestry. For this reason, at least in the Minoans, the Caucasus/Iran-related component seems strange in the light of the three-way split European genetic composition. In fact, the Caucasus/Iran-related component in the Minoans could not be traced back to the steppe — even if the Yamnaya populations do contain an Iranian-related signal (see page 135) — because Minoans lacked the other ancestry forming the steppe legacy: the EHG-related ancestry.

Figure 64. The extent of the Greek colonization. Image taken from World History Website.

For this reason, at least in Southern Europe, the arrival of the Caucasus/Iran-related component raises other questions about who brought it into Europe and when.

In the next chapter, I describe the second part of the work I did together with Francesco Montinaro (University of Tartu) and Cristian Capelli (University of Oxford) in order to shed light on the ancestry composition of Italians and Europeans in general, and then discover that the three-way split model is not enough to explain the genetic variability of Southern Italians. A few months after we put our work on Biorxiv (Raveane *et al.,* 2019),

another work was released as a preprint (Fernandes *et al.,* 2019). They analysed ancient Bronze Age samples from the Islands of the Western Mediterranean (Balearic Islands, Sicily and Sardinia), confirming with ancient samples what we have found (see Chapter "*Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe*"): in Sicily, the Iranian-related component arrived by the Middle Bronze Age, thus earlier than the classical period of Greek expansion in Southern Italy.

What happened next

The genetic portrait of European populations from 4,000 years ago is pretty similar to the one of the modern-day populations. At that time, the three dominant colours used for painting our genome had already been added (Lazaridis *et al.,* 2014; Allentoft *et al.,* 2015). This does not mean that European migratory phenomenon would have come to an end after the Bronze Age, but just that the incoming populations would have not been so highly differentiated as they were before. Moreover, it could also be that the growing European people made the later migrations less effective in deeply shuffling our genes (Günther & Jakobsson, 2016). However, many other migrants from different population would have arrived in the following millennia, thus mixing the dominant colours and creating slightly different tones.

The peopling of Europe during the last 3,000 years has been extensively studied using both modern and ancient DNA (Patterson *et al.,* 2012; Ralph & Coop, 2013; Hellenthal *et al.,* 2014; Leslie *et al.,* 2015; Busby *et al.,* 2015; Martiniano *et al.,* 2016; Schiffels *et al.,* 2016). As well synthesized in Günther & Jakobsson, 2016, several small and large scale migrations reached Europe and, together with "isolation by distance", shaped and reworked the European genetic variability painting. A lot of different people from outside Europe came here, mixed with the locals and, in a few years, become Europeans themselves, thus rendering the words "autochthonous" completely meaningless. A lot of Europeans moved within and beyond the mobile borders of an ever-changing "Europe", thus leaving their footprints in the genomes of worldwide populations.

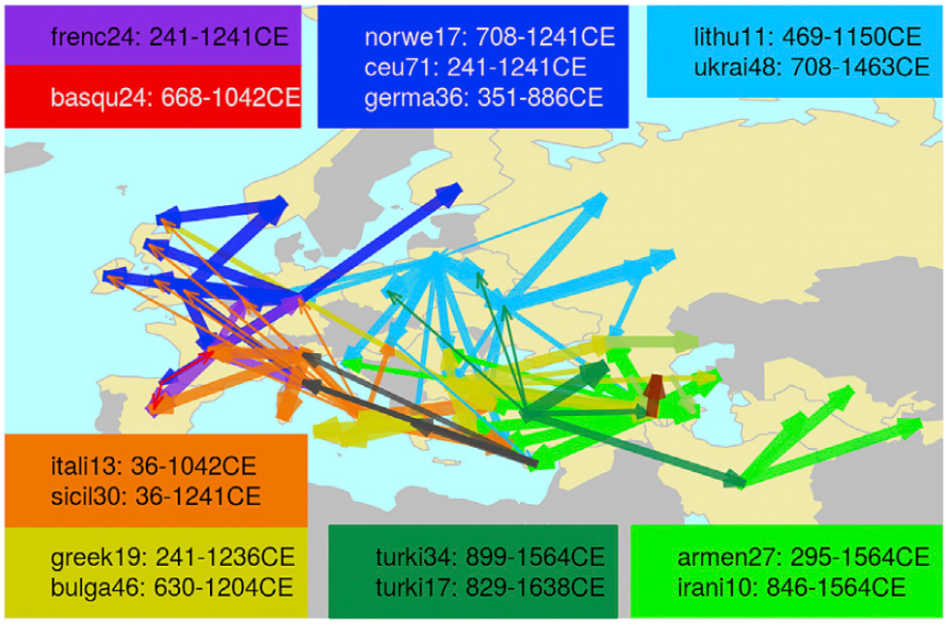


Figure 65. Gene flow within West Eurasian populations. Linecolors represent the regional identity of the donor group, and line thickness represents the proportion of DNA coming from the donor group. Image is taken from Busby *et al.,* 2015.

Etruscans, Romans, Vikings, Longobards, Saxons, Normans, Arabs and so on bathed in the European genetic pool, thus creating the modern-day European populations as we know them. But the European wanderings did not stop with them: many others will keep flowing towards Europe, shaping and enriching our genetic pool without a single lifeguard stopping it (Figure 65).

"The stars are veiled.
Something stirs in the East."

J. R. R. Tolkien, *Lord of the Rings*

Population structure of modern-day Italians reveals patterns of ancient ancestries in Southern Europe

In this chapter, I describe the analyses we did on modern-day European populations in order to dig out traces of ancient migrations and admixtures from their genome.

As reported in the previous chapter, the study of the genetic make-up of modern-day European populations has been boosted by the availability of ancient DNA. However, it is still unclear how the combination of early European foragers, Neolithic farmers and Bronze Age nomadic pastoralists can fully explain all the genetic variation across Europe.

In this context, Italy could really help in filling the gap. In fact, due to its position in the centre of the Mediterranean basin, is an ideal population to recover the genetic signatures of past European demographic events, offering the opportunity to complement and enrich the scenario depicted by ancient DNA studies.

In this work, we assembled and analyzed a genome-wide SNP dataset composed by 5,192 modern-day samples of which 1,616 Italians, to which we added already published genomic data of ancient individuals. Then, we investigated the ancestry composition of our modern samples by testing different combination of ancient samples.

We found interesting north-south differences across Italy: a Steppe Bronze Age signal — the known contribution from the Pontic-Caspian steppes — was higher in Northern Italy, while another Bronze Age contribution coming from the East (related to an Anatolian Bronze Age sample and other south-eastern European ancestries) showed an opposite distribution, representing a substantial component of the ancestry profile of populations in the south of Italy.

At the end of the work, this signature will remain still uncharacterised in terms of precise dates and origin. However, we hope that additional ancient Italian samples could help in shedding light on this contribution and, judging from one of the latest works (Fernandes *et al.,* 2019), we are on track and going in the right direction. For the moment, we only know that there was something stirring in the East around five thousand years ago. Then, this population arrived into Southern Europe and Italy at least in the Bronze Age, mixed with autochthonous people and left remarkable traces in their genomes, which have come, down the ages, to us.

Methods

Analyses on modern-day populations

In this first part of Methods, I will report briefly some of the analyses performed by Dr Alessandro Raveane, since they represent the starting point for the following investigations with ancient samples. For more details, please refer to the Supplementary Materials of Raveane *et al.,* 2019.

Modern datasets

He compiled a dataset comprising 1,616 both novel and previously published Italian individuals genotyped for different Illumina arrays (Table B.2). Then, in order to compare the genetic variability of the Italian population with the worldwide variation, he built a larger dataset composed by European (including six new samples from Albania newly genotyped and 16 Corsicans, 10 Portuguese and five French genotyped in Tamm *et al.,* 2019 for the first time) and a set of representative Asian and African populations (Table B.2). In order to avoid SNP-loss due to the partial incompatibility between the first and second generation of Illumina arrays, he assembled two different worldwide datasets. The Full Modern Dataset (FMD) included 4,852 samples (Busby *et al.,* 2015; Fiorito *et al.,* 2016; Hellenthal *et al.,* 2014; Li *et al.,* 2008; Yunusbayev *et al.,* 2015; Behar *et al.,* 2013; Rasmussen *et al.,* 2010; Raghavan *et al.,* 2014; Behar *et al.,* 2010; Kovacevic *et al.,* 2014; Kushniarevich *et al.,* 2015; Auton *et al.,* 2015; Metspalu *et al.,* 2011; Pagani *et al.,* 2015; Parolo *et al.,* 2015; Hodoğlugil & Mahley, 2012; Haber *et al.,* 2013; Haber *et al.,* 2016; Chaubey *et al.,* 2012) (1,589 Italians) and 218,725 SNPs genotyped with Illumina arrays; the High Density Dataset (HDD) contained 1,651 samples (Fiorito *et al.,* 2016; Yunusbayev *et al.,* 2015; Behar *et al.,* 2013; Raghavan *et al.,* 2014; Auton *et al.,* 2015; Pagani *et al.,* 2015; Haber *et al.,* 2013; Haber *et al.,* 2016) (524 Italians) and 591,217 SNPs genotyped with the Illumina Omni array.

Before merging, each dataset was cleaned excluding all SNPs and individuals with more than 2% missing data, using PLINK 1.9. When needed, the genetic position of the markers was lifted to the UCSC b37 genetic reference, using the released data for each chip, downloaded from the Illumina website. Thus, each dataset was singularly merged, and inconsistencies due to different strand were resolved flipping the inconsistent SNPs after the removal of ambiguous C/G and A/T and tri-allelic markers. The identity by descent *pi-hat* metrics were estimated using the *––genome* flag in PLINK

1.9. To avoid hidden relatedness among individuals, one random individual was excluded from each pair showing a *pi-hat* value higher than 0.2.

To avoid the impact of SNPs in linkage disequilibrium (LD) in the analyses based on SNP allele frequencies, a recursively removal of SNPs with a threshold of r^2 (squared correlation coefficient) higher than 0.2 and using 50 kb sliding windows (using flag *–-indep-pairwise 50 5 0.2*) was carried out. This led to a final number of 89,672 and 135,628 SNPs for the FMD and the HDD, respectively.

Haplotype analysis (CHROMOPAINTER and fineSTRUCTURE)

In order to exploit the higher degree of information about genetic variation provided by SNP-based haplotypes, Alessandro phased the dataset using SHAPEIT 2 (Delaneau *et al.,* 2013) and the HapMap b37 genetic map.

CP was employed to generate a matrix of recipient individuals "painted" as a combination of donor samples (copying vector). The *Ne* ("recombination scaling constant") and θ (per site mutation rate) parameters were inferred performing 10 iterations on four randomly selected chromosomes [2,5,13,17] (Busby *et al.,* 2015; Martin *et al.,* 2018) and 382 individuals from nine representative populations (Han Chinese, Finnish, British, Spanish, Italians, Sardinians, Moroccans, Turkish and Yoruba from Nigeria). The resulting estimates were then averaged taking into account the different recombination patterns of each analysed chromosome. The resulting parameters ($Ne = 365$, $\theta = 0.00051$) were then used in the painting runs as input values for *-n* and *-M* flag, respectively.

Then, Alessandro ran CP three times for each dataset generating three different outputs:

- a matrix of all the individuals "painted" as a combination of all the individuals, for cluster identification and GT analysis;
- a matrix of all Italians as a combination of all Italians, for F_{ST} analysis;
- a matrix of all the samples as a combination of all the other samples but excluding Italians, for "noItaly" GT analysis.

However, only the first CP run is needed for the analyses on ancient ancestries I will describe later on.

Then, the painted chromosomes generated by CP were combined using the ChromoCombine utility.

The chunkcount coancestry matrix was then used to obtain a hierarchical clustering tree based on the Bayesian Markov Chain Monte Carlo (MCMC)

algorithm implemented in fineSTRUCTURE (fS) (Lawson *et al.,* 2012). fS was run for both the datasets identified as HDD and FMD (see "Datasets" section) using the same pipeline (for details, refer to the Materials and Methods and the Supplementary Materials of Raveane *et al.,* 2019).

By visually inspecting the tree, he grouped together clusters whose samples origin was less relevant to the characterization of the genetic structure and admixture history of the Italian population (Busby *et al.,* 2015; Montinaro *et al.,* 2015). For this task *cuttree()* and *cut.dendrogram()* functions of R Package *dendextend* (Galili, 2015) were used. Then, he assessed the robustness of the final set of clusters by generating the MCMC pairwise coincidence matrix with the flag "*-e meancoincidence*" in fS. This matrix gives a measure of the number of MCMC samples in which two individuals fall together (Lawson *et al.,* 2012). Using this matrix, he calculated the mean of these estimates for all the samples falling in the same macro-cluster and observed an average co-occurrence not below 95%. The final tree for FMD is reported in Figure B.1.

Detecting recently admixed individuals ("Cluster Self-Copy" analysis)

Recently admixed individuals were identified as those copying from members of the cluster they belong less than the amount of cluster self-copying for samples with all the four grandparents from the same geographic region (for details, refer to Raveane *et al.,* 2019). A total of 195 putative recently admixed samples (13% of the initial dataset considered in this analysis) were discarded from the subsequent CP and GT analyses of FMD, of these, 11 were also present in the HDD and were similarly discarded.

ADMIXTURE analysis

Alessandro was almost finishing the analyses described above and I joined the group in Oxford in October 2016. Paralleling to the exploration of Neanderthal legacy (see chapter *"The genetic legacy of Neanderthals in Italy and Europe"*), I started to see some of the analyses I would have done on ancient samples.

In this context, we run the ADMIXTURE software (ADMIXTURE 1.3.0, Alexander *et al.,* 2009) on the FMD dataset. We used the cross-validation approach to define the most supported number of ancestral components K describing the modern samples included in the FMD. We performed 10 different ADMIXTURE runs using random seed on the LD-pruned dataset of

4,606 modern individuals. Then, we merged the resulting ancestry composition matrix with CLUMPP (Jakobsson & Rosenberg, 2007) using the largeKGreedy algorithm and the random input orders with 10,000 repeats. We applied the cluster visualisation program distruct for many K's implemented in CLUMPAK (Kopelman *et al.,* 2015; Rosenberg, 2004) in order to find the best alignment of the CLUMPP results. Finally, we visualised the ancestral composition of fS clusters as bar plots for each individual or for cluster means, using R statistical software. We identified $K = 15$ as the number of ancestries with the minimum cross-validation error.

Dataset

At this point, I started to play with ancient samples. The aim of the study was to reconstruct the ancestry profiles of modern populations included in the FMD, focusing on European and, specifically, on Italian individuals.

For this purpose, we selected 63 ancient samples from recent studies (Lazaridis *et al.,* 2016; Fu *et al.,* 2016; Lazaridis *et al.,* 2017; Olalde *et al.,* 2018; Mathieson *et al.,* 2018; Hofmanova *et al.,* 2016; Broushaki *et al.,* 2016; Mathieson *et al.,* 2015) covering a period from $\sim 14,000$ to 1,400 BCE (Late Palaeolithic to Iron Age) and representing several different archaeological assemblages (Table 3). For each interesting cultural assemblage or literature dataset, I tested which combination of samples was the best in maximizing the number of SNPs, by doing multiple intersections, iteratively. In the end, the chosen samples were characterised by a high number of SNPs and provided an informative representation of Western Eurasian cultural variation across time.

Finally, the 63 ancient samples were merged with the FMD using the *--bmerge* command in PLINK 1.9. The merged dataset is stored in the binary data format handled by PLINK: three files with the suffices ".bed", ".bim" and ".fam", containing the genotypes, the variant information and the individual information, respectively.

Sample name	Data source	PCA and ADMIXTURE labels	CP/NNLS labels	LAT	LONG	AGE	Culture
I9030	Lazaridis *et al.*, 2016	Villabruna		46.15	12.21	12230-11830 calBCE	HG
Bichon	Fu *et al.*, 2016	Switzerland_HG	WHG	47.1	6.87	11820-11610 calBCE	HG
SATP	Fu *et al.*, 2016	CHG	SATP	42.38	42.59	11430-11180 calBCE	HG
I1072	Lazaridis *et al.*, 2016	Natufian		32.65	35.07	11840-9760 BCE	Natufian
KK1	Fu *et al.*, 2016	CHG	CHG	42.28	43.28	7940-7600 calBCE	HG
WC1	Broushaki *et al.*, 2016	Iran_N	IN	34.06	46.65	7455-7082 cal BCE	Neolithic
I0867	Lazaridis *et al.*, 2016	Levant_N		31.79	35.17	7300-6750 BCE	Neolithic
I0061	Fu *et al.*, 2016	EHG	EHG	61.65	35.65	6850-6000 BCE	HG
I0707	Lazaridis *et al.*, 2016	Anatolia_N		40.3	29.57	6500-6200 BCE	Neolithic
I0746	Lazaridis *et al.*, 2016	Anatolia_N		40.3	29.57	6500-6200 BCE	Neolithic
I0745	Lazaridis *et al.*, 2016	Anatolia_N		40.3	29.57	6500-6200 BCE	Neolithic
Bar31	Hofmanova *et al.*, 2016	Anatolia_N		40.3	29.61	6419-6238	Neolithic
Bar8	Hofmanova *et al.*, 2016	Anatolia_N	AN	40.3	29.61	6212-6030	Neolithic
Loschbour	Fu *et al.*, 2016	WHG		49.81	6.4	6210-5990 calBCE	HG
LaBrana1	Fu *et al.*, 2016	WHG		42.91	-5.38	5983-5747 calBCE	HG
I0585	Lazaridis *et al.*, 2016	WHG		42.91	-5.38	5983-5747 calBCE	HG
I0412	Lazaridis *et al.*, 2016	Europe_EN		42.5	0.5	5308-5080 calBCE	Neolithic
Stuttgart	Fu *et al.*, 2016	Europe_EN	EEN	48.78	9.18	5310-5070 calBCE	Neolithic
I0100	Lazaridis *et al.*, 2016	Europe_EN		51.89	11.04	5202-4852 calBCE	Neolithic
I1407	Lazaridis *et al.*, 2016	Armenia_ChL		39.73	45.2	4350-3700 BCE	Copper/Bronze Age
I1584	Lazaridis *et al.*, 2016	Anatolia_ChL		40.3	29.57	3943-3708 calBCE	Copper/Bronze Age
Matojo	Lazaridis *et al.*, 2016	Europe_MNChL		42.35	-3.52	3010-3879 calBCE	Neolithic
RISE487	Lazaridis *et al.*, 2016	Europe_MNChL		45.26	10.38	3483-3107 calBCE	Neolithic
Iceman	Mathieson *et al.*, 2015	Iceman	Iceman	46.78	10.83	3359-3105 BCE	Neolithic
I1658	Lazaridis *et al.*, 2016	Armenia_EBA		40.38	43.87	3347-3092 calBCE	Copper/Bronze Age
I0443	Lazaridis *et al.*, 2016	Steppe_EMBA		53.38	50.38	3300-2700 BCE	Copper/Bronze Age
I0231	Lazaridis *et al.*, 2016	Steppe_EMBA	SBA	52.42	48.24	2921-2762 calBCE	Copper/Bronze Age
ATP2	Lazaridis *et al.*, 2016	Europe_MNChL		42.35	-3.52	2899-2678 calBCE	Neolithic
RISE489	Lazaridis *et al.*, 2016	Europe_MNChL	Remedello	45.26	10.38	2908-2578 calBCE	Neolithic
I2499	Lazaridis *et al.*, 2017	Anatolia_BA		37.92	30.71	2836-2472 BCE	Copper/Bronze Age
I0103	Lazaridis *et al.*, 2016	Europe_LNBA		51.42	11.68	2578-2468 calBCE	Copper/Bronze Age
I1633	Lazaridis *et al.*, 2016	Armenia_EBA		40.65	45.12	2619-2410 calBCE	Copper/Bronze Age
RISE552	Lazaridis *et al.*, 2016	Steppe_EMBA		46.62	43.33	2849-2143 calBCE	Copper/Bronze Age
I2495	Lazaridis *et al.*, 2017	Anatolia_BA		37.92	30.71	2558-2295 BCE	Copper/Bronze Age
I1706	Lazaridis *et al.*, 2016	Levant_BA		31.99	35.98	2490-2300 BCE	Copper/Bronze Age
I1730	Lazaridis *et al.*, 2016	Levant_BA		31.99	35.98	2489-2299 calBCE	Copper/Bronze Age
I0112	Lazaridis *et al.*, 2016	Europe_LNBA		51.79	11.14	2457-2142 calBCE	Copper/Bronze Age
I2683	Lazaridis *et al.*, 2017	Anatolia_BA	ABA	37.92	30.71	2500-1800 BCE	Copper/Bronze Age
I1705	Lazaridis *et al.*, 2016	Levant_BA		31.99	35.98	2198-1966 calBCE	Copper/Bronze Age
I0070	Lazaridis *et al.*, 2017	Minoan_Lasithi		35.08	25.83	2000-1700 BCE	Minoan
I0071	Lazaridis *et al.*, 2017	Minoan_Lasithi	MIN	35.08	25.83	2000-1700 BCE	Minoan
I0073	Lazaridis *et al.*, 2017	Minoan_Lasithi		35.08	25.83	2000-1700 BCE	Minoan
I0074	Lazaridis *et al.*, 2017	Minoan_Lasithi		35.08	25.83	2000-1700 BCE	Minoan
I9005	Lazaridis *et al.*, 2017	Minoan_Lasithi		35.08	25.83	2000-1700 BCE	Minoan
RISE505	Lazaridis *et al.*, 2016	Steppe_MLBA		53.46	85.45	1746-1626 calBCE	Copper/Bronze Age
RISE500	Lazaridis *et al.*, 2016	Steppe_MLBA		53.46	85.45	1700-1500 BCE	Copper/Bronze Age
I1656	Lazaridis *et al.*, 2016	Armenia_MLBA		40.38	43.94	1501-1402 calBCE	Copper/Bronze Age
I9010	Lazaridis *et al.*, 2017	Mycenaean		37.5	23.45	1700-1200 BCE	Mycenaean
I9041	Lazaridis *et al.*, 2017	Mycenaean	MYC	37.5	23.45	1700-1200 BCE	Mycenaean
I9033	Lazaridis *et al.*, 2017	Mycenaean		36.92	21.7	1416-1280 BCE	Mycenaean
I9006	Lazaridis *et al.*, 2017	Mycenaean		37.97	23.5	1411-1262 BCE	Mycenaean
I0247	Lazaridis *et al.*, 2016	Steppe_IA		52.43	51.16	375-203 calBCE	Iron Age
I2478	Olalde *et al.*, 2018	Beaker_Northern_Italy	I2478_ITN_Beaker	44.8	10.33	2200-1900 BC	Copper/Bronze Age
I2477	Olalde *et al.*, 2018	Beaker_Northern_Italy	I2477_ITN_Beaker	44.8	10.33	2200-1900 BC	Copper/Bronze Age
I1979	Olalde *et al.*, 2018	Beaker_Northern_Italy	I1979_ITN_Beaker	44.8	10.33	2200-1900 BC	Copper/Bronze Age
I4930	Olalde *et al.*, 2018	Beaker_Sicily	Sicily_Beaker	37.73	12.89	2500-1900 BC	Copper/Bronze Age
I1108	Mathieson *et al.*, 2018	Balkans_MP_Neolithic		43.98	26.4	5800-5400 BCE	Neolithic
I1499	Lazaridis *et al.*, 2016	Europe_EN		48.52	21.17	5210-5010 calBCE	Neolithic
I5427	Mathieson *et al.*, 2018	Peloponnese_N		36.64	22.38	6005-5879 calBCE	Neolithic
I3708	Mathieson *et al.*, 2018	Peloponnese_N		36.64	22.38	5500-3700 BCE	Neolithic
I2318	Mathieson *et al.*, 2018	Peloponnese_N		37.42	23.13	4043-3947 calBCE	Neolithic
I3709	Mathieson *et al.*, 2018	Peloponnese_N		36.64	22.38	3990-3804 calBCE	Neolithic
I3920	Mathieson *et al.*, 2018	Peloponnese_N	PN	36.64	22.38	3933-3706 calBCE	Neolithic

Table 3. Ancient samples used in this study. List of the 63 ancient samples used in the study, including the location of the sites where they had been found, their estimated antiquity and the labels used in each analysis.

Principal Component Analysis (PCA)

A first step was a simple PCA to explore the patterns of genetic variations of modern populations with respect to ancient individuals. First of all, a few details on PCA. PCA is a *non-parametric* — it do not involve any distributional assumptions — *unsupervised* — the data do not need to be labelled — *dimensionality-reduction* technique, which can detect the "directions" of major variation in the data. Thus, when applied to genetic data, it can easily detect evidence of population structure, relatedness among samples and technical variability, such as genotyping errors.

In brief, the first step is to go from genetic data to a relationship matrix where each entry is a measure of the average genetic similarity between two individuals. Then, the PCA is performed to obtain an "*eigendecomposition*" of the matrix. In other words, it consists of picking the directions in the data along which the variance is maximised. The principal components (PCs), eigenvectors, are orthogonal axes of variations that best explain the observed patterns of genetic similarity (Figure 66). At the end, the lots of starting dimensions are flattened to the first two or three, which can be plotted and visualised. To summarise, the underlying idea is simple — reduce the number of variables, while preserving as much information as possible.

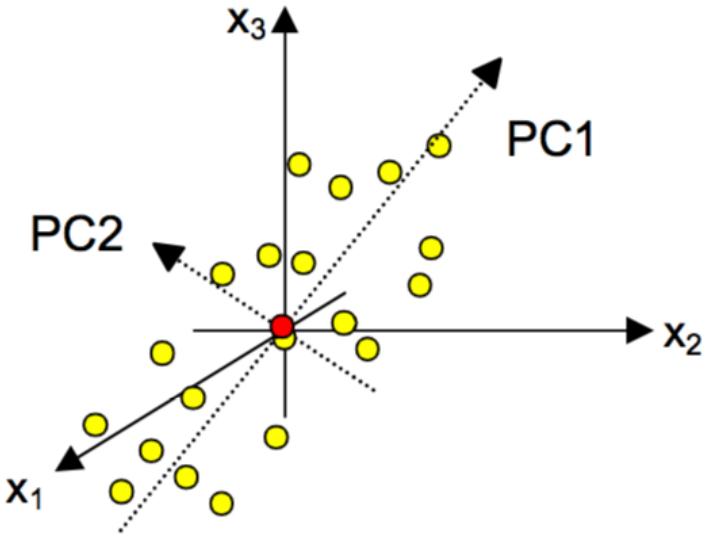


Figure 66. Example of PC1 and PC2 extraction. Image taken from Umetrics Suite Blog.

Moreover, the interpretation is straightforward, even if we need to be careful to not over-infer past demographic scenarios (see page 22 and the paper by Novembre & Stephens, 2008). In general, we assume that individuals with similar values for a particular top principal component will have similar ancestry for that axis. To cite an example, the PCA of European populations in Novembre *et al.,* 2008 was able to geographically map their country of origin quite accurately (Figure 5).

Returning to our work and PCAs, first, we pruned the dataset for LD (r^2>0.2) and we cleaned it for SNPs with high missingness, thus obtaining 92,172 variants. We did this with the EIGENSOFT (Patterson *et al.,* 2006) *smartpca* software, projecting 63 ancient individuals onto the components inferred from two different sets of modern individuals by using the *lsqproject* and *shrinkmode* options. In particular, we projected the ancient samples on a PCA based on 3,282 modern individuals assigned, through the CP/fS analysis, to European, West Asian or Caucasian clusters. We repeated the same procedure a second time using the 2,469 modern individuals assigned to European clusters only. An overview of the modern-day samples included in each cluster is presented in Figure B.1, while the ancient samples are presented in Table 3.

In the same way, we also projected a selection of the ancient samples used as donors in the CP/*NNLS* approach (see below) on the components inferred from modern European individuals (Figure 88).

Then, as one reviewer requested, we explored the impact of DNA damage in the polymorphisms obtained from ancient samples. We repeated the PCA with the 63 ancient samples and modern European, Caucasian and West Asian samples by removing transition polymorphisms (for a final number of 16,999 SNPs). In fact, the most observed lesions in ancient DNA are the transitions adenine → guanine (A → G), cytosine → thymine (C → T), G → A, and T → C (Binladen *et al.,* 2006; Hansen *et al.,* 2001; Hofreiter *et al.,* 2001b; Gilbert *et al.,* 2003). Thus, it could be possible that the patterns we saw in the PCAs were mainly guided by DNA damage genetic variations. To control for that, we listed the transition polymorphisms (TC|CT|GA|AG), we excluded them from the PCA computation using the option *badsnpname* and we did the PCA again. Finally, we compared the PCAs with and without transition polymorphisms: we recorded significant correlations for the localisation of ancient samples along PC1 and PC2 between the two PCAs, thus demonstrating that the main patterns of genetic variations are not biased by aDNA damage (Figure 67, PC1: Pearson r = -0.995, p-value < $2.2e^{-16}$; PC2: Pearson r=0.989, p-value < $2.2e^{-16}$).

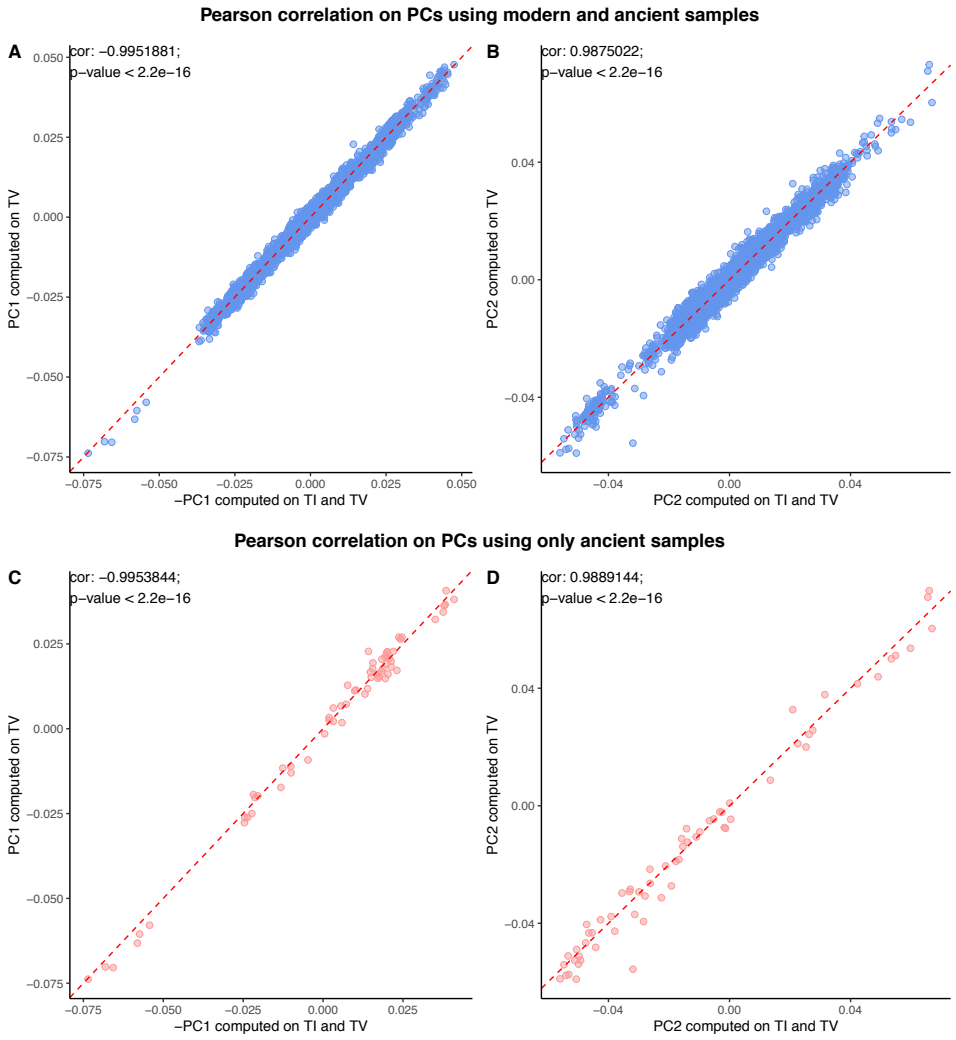


Figure 67. Correlation of the PCs of modern and ancient samples estimated including and excluding transition polymorphisms. A) Correlation of the PC1 values of modern and ancient samples. **B)** Correlation of the PC2 values of modern and ancient samples **C)** Correlation of the PC1 values of ancient samples. **D)** Correlation of the PC2 values of ancient samples.

ADMIXTURE analysis

We used the ancestral allele frequencies inferred from ten different ADMIXTURE runs on 4,606 modern samples in order to "project" the inferred

components onto the ancient samples with the option *-P*. As for the analysis with modern samples only, we merged the resulting ancestry composition matrix with CLUMPP (Jakobsson & Rosenberg, 2007) and visualised the ancestral composition with the program distruct for different values of K as implemented in CLUMPAK (Kopelman *et al.,* 2015; Rosenberg, 2004).

D-STATISTICS

We used the *D*-statistics method implemented in the *qpDstat* program in the package ADMIXTOOLS v4.2 (Patterson *et al.,* 2012) in order to test for admixture events. I have already explained the theory underlying this test at page 65 when we were talking about Neanderthals-modern humans admixture; however, I will report here a few more details on the way this test works and on its interpretation. Considering three different populations (W, X and Y) and an outgroup (Z), the *D*-statistics is defined as the difference in the counts of ABBA and BABA sites, where A and B are the two allelic forms, normalized by the total number of observations. When the *D* significantly deviates from zero, the configuration tested is not supported by a tree-like scenario. If the *D* is positive (negative), W and Y are more (less) related than X and Y, possibly as the results of gene flow between the two populations. We consider only tests with more than 30,000 SNPs and significance was assessed for absolute values of the Z score greater than 3.

Given A=Ancient, M=Modern, O=Outgroup, we performed the following *D*-statistic analyses permuting all the possible pairs of Modern (M) populations or Ancient (A) individuals/groups: AAMO, MMAO, AMAO, AAAO and MAMO. Here we analysed 63 ancient samples (A) considering them both as individuals and as groups, 66 modern clusters (M) from Europe, West Asia, Northern Africa and Caucasus and three Mbuti individuals from the Democratic Republic of Congo as outliers (O). Then, we considered all the 84 worldwide modern clusters in the combination MMMC. In these worldwide tests, we used Chimp as outgroup instead of the Mbuti individuals, as some African populations were also tested. Through the Results section, I will show some of the most interesting *D*-statistics tests, but I will not report in this thesis the complete results (we also removed them from the last version of the manuscript by Raveane *et al.,* 2019). Moreover, I will focus particularly on some *D*-statistic analyses evaluating the relationship of Italian clusters with AN, ABA and SBA. In details, we performed the *D*-statistics in the form:

$$D(\text{Ita1}, \text{Ita2}, \text{AN/ABA/SBA}, \text{Mbuti})$$

where Ita1 and Ita2 are different clusters composed mainly by Italian individuals as inferred by fS analysis.

CHROMOPAINTER and Non-Negative Least Squares analysis

In order to disentangle the main components of genetic variation in modern Europeans as a combination of ancient samples, we applied CHROMOPAINTER (CP) (Lawson *et al.,* 2012; Li & Stephens, 2003) using the "unlinked" mode (Hofmanova *et al.,* 2016). In fact, due to the limited number of ancient samples and their high rate of missingness, we could not phase them and we went on with the unlinked mode. However, this mode does not allow missing genetic position; for this reason we selected sets of ancient representative samples with the least amount of missing genotypes (Table 4). Moreover, when the genetic data of ancient samples were obtained through sequencing, we preferred samples with high coverage (Table 4). However, due to the low quality of some aDNA genomes used in the following analyses, we considered SNPs as independent of each other (i.e. "unlinked") as in Hofmanova *et al.,* 2016 — in other words, we did not the LD pruning step. Under this model, the result is simply the sum of the matching chunk counts each *recipient* individual receives by each *donor* group.

We first removed all variants with even one missing SNP in an ancient individual with the plink command *–-geno 0*. Then, we merged the ancient samples with the phased FMD (the modern phased dataset does not contain missing position, because all missing genotypes have been imputed during the phasing step).

In the CP runs, we used the same *Ne* and θ parameters obtained from the modern dataset (see *Haplotype analysis (CHROMOPAINTER and fineSTRUCTURE)* on page 146 and page 322 for the code used for running CP). All modern and ancient individuals were painted, using only modern samples as donors, by following the same approach as described in Hofmanova *et al.,* 2016 and Broushaki *et al.,* 2016. Then, we used the ChromoCombine program to collapse the painted chromosome generated by CP into one file (see page 323 for the command used for this step).

Subsequently, we applied the mixture fit *NNLS*-based analysis (Non-Negative Least Squares analysis, Leslie *et al.,* 2015; Lawson & Hanson, 1995; Montinaro *et al.,* 2015), which deconstructs the "painting profile" inferred by CP as the combination of defined sources. In details, we reconstructed the profiles of the modern samples as the combination of the ancient ones.

In order to do so, we used an adaptation of the non-negative-least-squares

(*nnls*) function implemented in R. The problem of non-negative least squares is a constrained least squares problem where the coefficients are not allowed to become negative (see the equation below). Given a matrix A and a vector of dependent variables b, the goal is to solve:

$$\begin{aligned} minimize \quad & ||Ax - b||, \\ subject \quad to \quad & x \geq 0 \end{aligned}$$

Where $x \geq 0$ is the constraint requiring that each component of the vector of the solutions x need to be positive. Precisely from this constraint, it derives the name "non-negative". $|| \cdot ||$ indicates the Euclidean norm. Concerning our ancestry problem, NNLS could be used as an approach for segment assignment given the "painting profile" inferred by CHROMOPAINTER. In fact, thanks to the combination with CHROMOPAINTER, it can "paint" each haploid genome of a recipient using the donor groups from which ancestry is to be assigned (Figure 68).

In particular, our matrix A is built by taking the rows corresponding to the ancients and all the columns in the *.chunkcounts.out* file and then by transposing the matrix. The vector b is the vector of one modern sample containing the SNPs copied from all modern samples. The vector x is the vector of the unknowns, i.e., the ancient components of the modern sample in b. However, if we consider the nnls problem (the above equation), we obtain an overdetermined system, in which there are more equations — the number of modern samples — than unknowns, the number of ancient samples. Due to this characteristic, this system cannot have an exact solution. However, we can compute the solution that best approximate an exact solution, in other words, the solution that minimize the error. The error is computed as the norm $||Ax - b||$, that is, the sum of the squared differences, and is called residual.

We analysed the modern samples (4,852 modern samples from the FMD with CP and 4,606 of them with *NNLS*) and different sets of ancient samples, which can be classified into two main groups. The first set (*ultimate* sources) included samples more ancient than the second (*proximate* sources). Specifically, the *ultimate* set contain samples older than 6,000 calBCE, while *proximate* sources are younger than 6,000 calBCE. We performed different *NNLS* analyses on the results of the CP runs from *ultimate* and *proximate* sources in order to explore the relationships between samples from different locations and of different ages. Then, we expressed the ancestry proportion

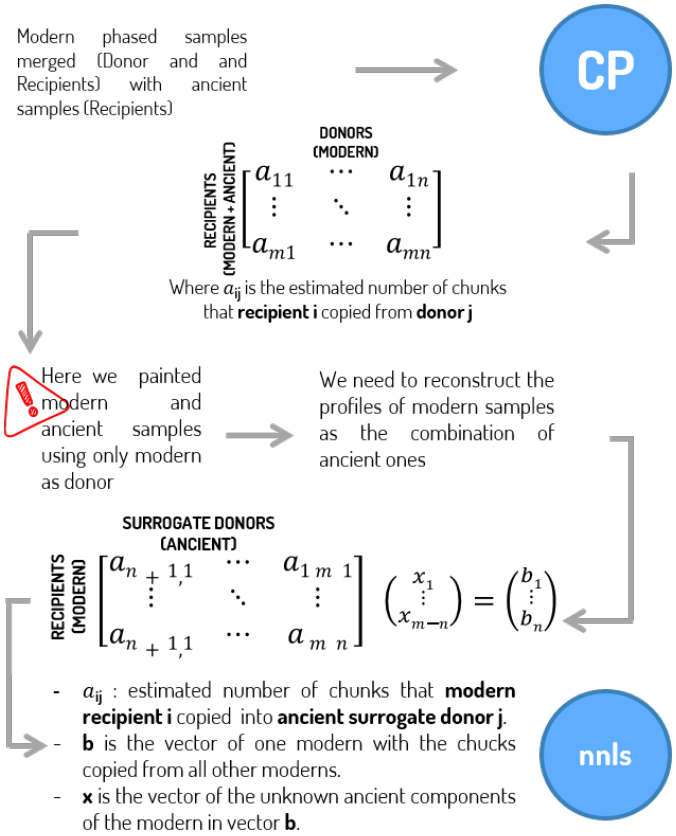


Figure 68. The CP/NNLS approach.

of each recipient population as a combination of each donor ancient sample or modern cluster components and we visualized these proportions through bar charts using R statistical software. Finally, we evaluated the goodness of fit of the *NNLS* performed on *proximate* sources by including/excluding ABA and SBA samples one at the time, through the computation of the sum of the squared residuals produced by the *NNLS*.

Sample name	Data source	CP/NNLS labels	Mean Coverage
Bichon	Fu et al. 2016	WHG	8.119
KK1	Fu et al. 2016	CHG	12.157
WC1	Broushaki et al. 2016	IN	10
I0061	Fu et al. 2016	EHG	1.952
Bar8	Hofmanova et al. 2016	AN	7
Stuttgart	Fu et al. 2016	EEN	19
Iceman	Mathieson et al. 2015	Iceman	8
I0231	Lazaridis et al. 2016	SBA	3
RISE489	Lazaridis et al. 2016	Remedello	0.5
I2683	Lazaridis et al. 2017	ABA	3.695
I0071	Lazaridis et al. 2017	MIN	7.312
I9041	Lazaridis et al. 2017	MYC	1.558
I2478	Olalde et al. 2018	I2478_ITN_Beaker	1240k capture (622,069 SNPs)
I2477	Olalde et al. 2018	I2477_ITN_Beaker	1240k capture (782,739 SNPs)
I1979	Olalde et al. 2018	I1979_ITN_Beaker	1240k capture (375,725 SNPs)
I4930	Olalde et al. 2018	Sicily_Beaker	1240k capture (40,252 SNPs)
I3920	Mathieson et al. 2018	PN	3.144
SATP	Fu et al. 2016	SATP	1.195

Table 4. Ancient samples coverage. The sequencing coverage of each ancient sample used in CP/*NNLS* analyses.

Ultimate sources analyses

We assembled the set of ultimate sources by including one representative ancient sample for Western Hunter Gatherers (WHG), Caucasus Hunter Gatherers (CHG), Eastern Hunter Gatherers (EHG), Anatolian Neolithics (AN) and Iranian Neolithics (IN) and we investigated the composition of the genetic clusters in relation to them.

After merging ancient and modern samples, a total of 173,006 variants were available for this set. We included modern North African and Asian clusters as sources in the *NNLS* analyses in order to consider recent contributions from non-Eurasian countries (Busby *et al.,* 2015).

Proximate sources analyses

Then, we investigated the genetic cluster composition in relation to *proximate* sources, which included European Early Neolithic (EEN), Steppe Bronze Age (SBA), and Anatolian Bronze Age (ABA) samples, with the latter as a representative of the recently characterised South Eastern European Bronze Age populations (SEE, Lazaridis *et al.,* 2017; Mathieson *et al.,* 2018). Since variation in the admixture dynamics between farmers and hunter-gatherers

across Europe has been reported (Lipson *et al.,* 2017), we added WHG as source in the *proximate* sources analyses.

After merging ancient and modern samples, a total of 135,022 SNPs were available for the *proximate* analyses.

We repeated the *proximate* analysis replacing the ABA sample with two alternative proxies for SEE populations: I9041 (Lazaridis *et al.,* 2017), a representative of the Mycenaean culture, and I0071 (Lazaridis *et al.,* 2017), associated with the Minoan cultural assemblage. These analyses retained a total of 59,477 and 163,744 SNPs, respectively. As before, we included modern day North African and Asian clusters as sources in the *NNLS* analyses (Lazaridis *et al.,* 2017; Busby *et al.,* 2015). We also considered as alternative SEE source the recently released Peloponnese Neolithic (PN) samples (Mathieson *et al.,* 2018), by selecting the best sample in terms of the number of non-missing variants when coupled with the ancient samples in the *proximate* sources set (I3920). We added I3920 to the set of ancient *proximate* sources, together with EEN, SBA and WHG obtaining a dataset composed by 146,526 SNPs on which we performed CP and *NNLS*.

Post-Neolithic Italian samples

We were also interested in the signatures of post-Neolithic events in ancient Italian samples. For this reason, we repeated the *proximate* analysis adding some Italian post-Neolithic samples as recipients.

A few post-Neolithic samples from Italy have been sequenced so far. One is the famous Iceman (Keller *et al.,* 2012), who dates to 3,359-3,105 BCE (Copper Age, around 5,000 years ago) and was discovered on the Tisenjoch Pass in the Italian part of the Ötztal Alps. Other three belong to the Remedello culture (Allentoft *et al.,* 2015; Lazaridis *et al.,* 2016) dating around 3,000 and 2,000 BCE (RISE487: 3,483-3,107 calBCE, RISE486: 2,134-1,773 calBCE; RISE489: 2,908-2,578 calBCE) and were found in the region of Lombardy in Northern Italy. For only two of these samples (Iceman and Remedello RISE489) a relatively large number of SNPs is available; thus we focused on these two in our analyses. When both the ancient Italian samples were added to the modern and ancient dataset, the total number of SNPs was 74,622 for *ultimate* sources. For *proximate* sources, when both the ancient Italian samples and the WHG sample were added, we obtained 57,582, 70,434 and 25,998 SNPs for the analyses with ABA, MIN and MYC, respectively.

A set of Italian samples related to the Bell Beaker cultural assemblage have been recently published in the work of Olalde *et al.,* 2018, which in-

cluded three individuals from Northern Italy and three from Sicily. The first set is from Emilia-Romagna, a region in the Po plain of Northern Italy and dated to an advanced Bell Beaker phase (2,200–1,930 BCE). The three Sicilian samples came from the Bell Beaker layer of burials of a small artificial cave and dated between 2,500 and 1,900 BCE. While all the Northern Italian Bell Beaker samples were used for our study, we could use only one (I4930) of the Sicilian individuals, due to the very low number of SNPs of the other two samples. Even so, the total number of available genetic variants of this Sicilian sample is still low (40,252), raising some issues about the confidence of the CP/*NNLS* inferences.

We used this Sicilian Bell Beaker individual as a recipient in a *proximate* sources CP run (EEN, SBA, ABA and PN) for a total of 6,119 SNPs and in another one where we replaced ABA and PN with MIN (6,877 SNPs). Then, we performed *NNLS* analyses to paint modern samples with the following sets of sources: EEN, SBA, ABA and PN; EEN, SBA and ABA; EEN, SBA and MIN; EEN, SBA and PN; all the analyses were run by including and excluding WHG as an additional source. We note here that we did not extend this set of analyses to the third SEE sample (Mycenaean, MYC) as the overall final number of SNPs was too small (2,284).

Then, we repeated the first CP run (EEN, SBA, ABA and PN) and replaced the Sicilian Bell Beaker sample, with the three Northern Italian Bell Beaker samples as recipients (39,006 SNPs). In a separate run, we included only the best of these samples as a recipient (I2477; 112,089 SNPs).

Finally, we also run one *ultimate* sources analysis, with 5,754 SNPs, using the Sicilian and the best Northern Italy Bell Beaker individual together as recipients and, in a separate run, all the three Northern Italian Bell Beaker samples as recipients (37,890 SNPs).

Geography and ancestry proportions

We explored the relationship between ancestry composition and geography, by performing Spearman's rank correlation tests between geographic locations of the European samples and the ancestry proportions inferred by means of the *NNLS* analyses described above. As a matter of fact, the majority of the ancestry proportions and latitude distributions were not normally distributed, according to the Shapiro-Wilk Normality Test. This is the reason why we used the Spearman's rank correlation test.

In details, we focused on West Eurasian populations, by keeping only those groups localised between -10 and 55 degrees of longitude and between 28 and 65 degrees of latitude for which a geographic location could be as-

signed. Among the Italian individuals, we considered only those samples whose grandparents were born in the same Italian administrative region. In tables B.3 and B.4, we have reported the modern populations that we have considered in the CP/*NNLS* analyses, the ancestry proportions and the standard errors computed applying a weighted jackknife bootstrap, for *ultimate* and *proximate* analyses, respectively.

qpAdm analysis

Due to the request of one of our reviewers, we decided to try to replicate the results obtained with the CP/NNLS pipeline, with the software *qpAdm*, which has recently become one of the reference methods for ancestral reconstruction. Thus, Dr Montinaro did the following analyses with *qpAdm*.

qpAdm harness different relationships of populations related to a set of outgroups, as in the following example:

$$f_4[Target, O1, O2, O3]$$

In details, for each tested cluster of the FMD and HDD, we have evaluated all the possible combinations of N "left" sources with $N = \{2, \ldots, 5\}$, and one set of right/left Outgroups. Specifically, the following left and right populations were used: left = Anatolia_BA, Anatolia_N, Steppe_EMBA, WHG, EHG, Iran_N, Minoan_Lasithi, Minoan_Odigitria, Mycenaean, NAfrica1, Levant_N, CHG, Europe_EN, Europe_LNBA and Europe_MNChL; right = AfontovaGora3, EHG, El Mirón, GoyetQ116-1, Iran_N, Kostenki14, Levant_N, MA1, Mota, Natufian, Ust'-Ishim, Vestonice16 and CHG. For each of the tested combinations, we used *qpWave* to evaluate if the set of chosen outgroups is able to:

- discriminate the combinations of sources,
- find if the target may be explained by the sources.

We used a p-value threshold of 0.01. Finally, we used *qpAdm* to infer the admixture proportions. We performed the same analysis on ancient Italian samples (Iceman, Remedello and Bell Beaker individuals from Sicily and North Italy).

Results

In this work, we have explored how past demographic event have shaped the genetic diversity of Italian and other European populations by combining the FMD comprising around 300 modern worldwide populations with 63 representative ancient samples (Figure 69).

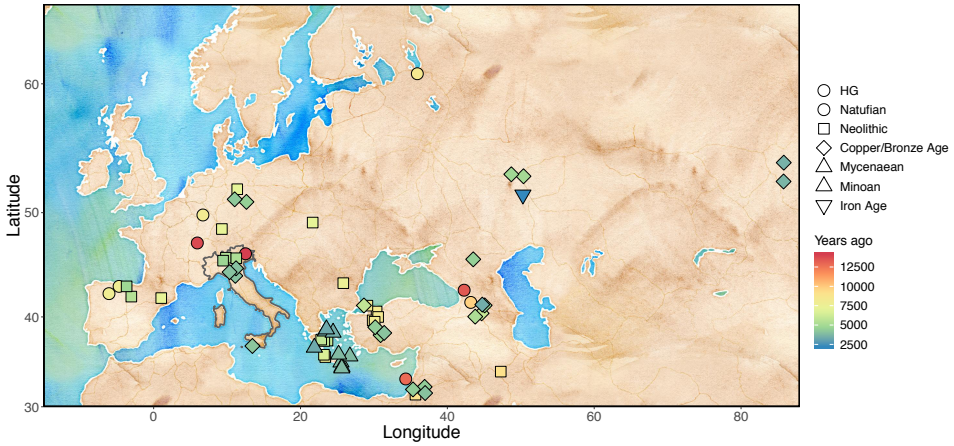


Figure 69. Geographic location of the 63 ancient samples used in this work. Samples information are reported in Table 3.

Firstly, we projected the 63 ancient samples onto the components inferred from modern European, West Asian or Caucasian modern-day individuals (Figure 70) and from European populations only (Figure 71). Both PCAs placed Western Hunter Gatherers (WHG), Anatolian Neolithic (AN) and Steppes Bronze Age (SBA) around Italian samples, in agreement with previous observations for the entirety of Europe (Lazaridis *et al.,* 2017; Lazaridis *et al.,* 2016; Fu *et al.,* 2016; Hofmanova *et al.,* 2016; Broushaki *et al.,* 2016). The Italian samples, with the exception of Sardinians, appeared also close to post-Neolithic samples from Anatolia and Caucasus (Bronze Age and Chalcolithic Anatolia, Lazaridis *et al.,* 2017) and the recently characterised Minoan and Mycenaean populations (Lazaridis *et al.,* 2017).

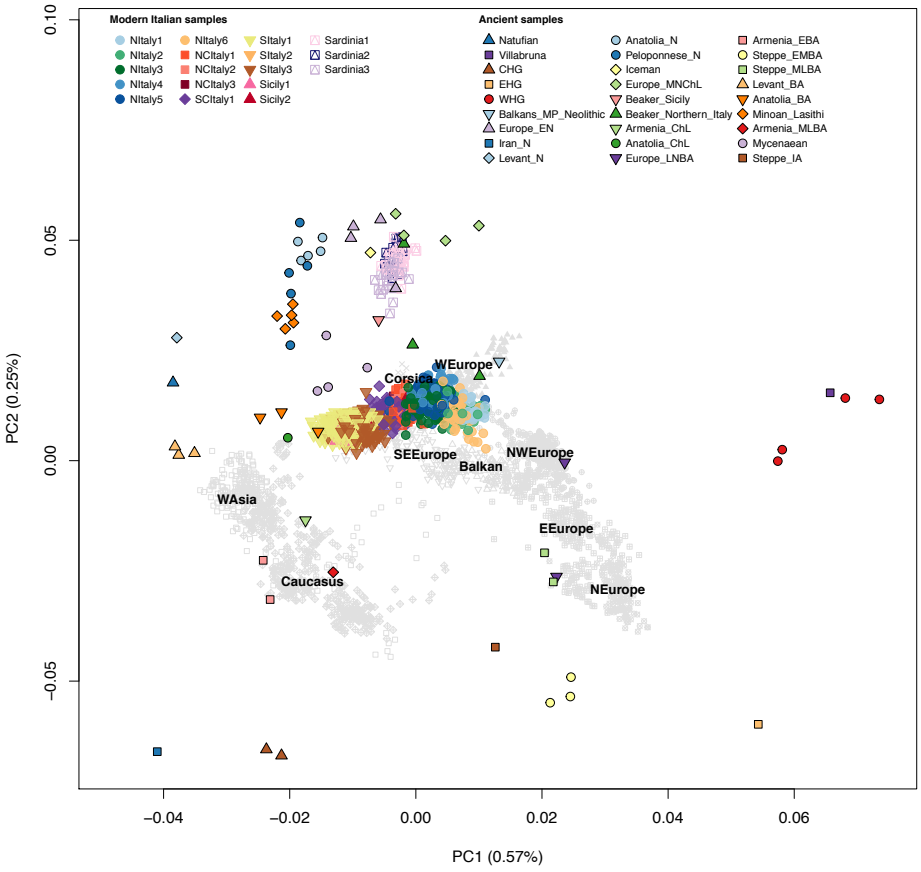


Figure 70. Principal component analysis projecting 63 ancient individuals onto the components inferred from modern individuals. Principal component analysis projecting 63 ancient individuals onto the components inferred from 3,282 modern individuals assigned, through a CP/fS analysis, to European, West Asian and Caucasian clusters. The labels are placed at the centroid of the macroarea. The centroids are calculated by computing the means of the coordinates of individuals in modern clusters within each macroarea.

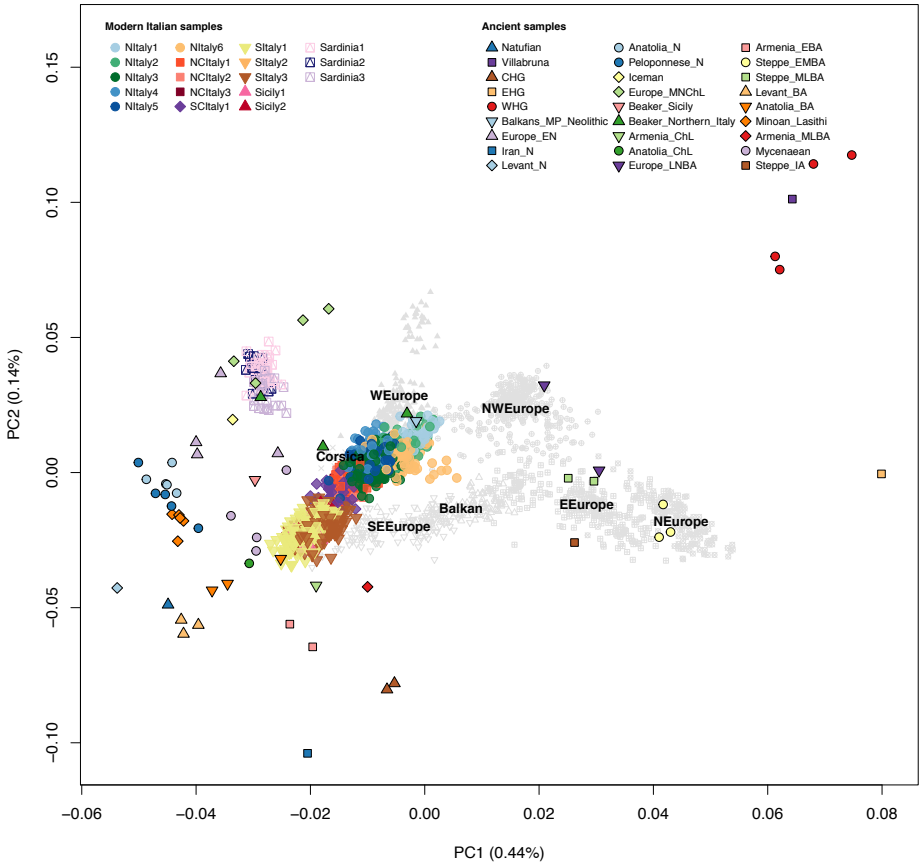


Figure 71. Principal component analysis projecting 63 ancient individuals onto the components inferred from modern individuals. Principal component analysis projecting 63 ancient individuals onto the components inferred from 2,469 modern individuals assigned, through a CP/fS analysis, to European clusters. The labels are placed at the centroid of the macroarea. The centroids are calculated by computing the means of the coordinates of individuals in modern clusters within each macroarea.

Similar patterns than in the PCA were also shown by the ADMIXTURE analysis on both ancient (Figure 72) and modern samples (Figures B.2 and B.3). We show also the $K = 15$ results for both modern and ancient samples averaged by clusters and ancient groups, respectively (Figure 73A, B).

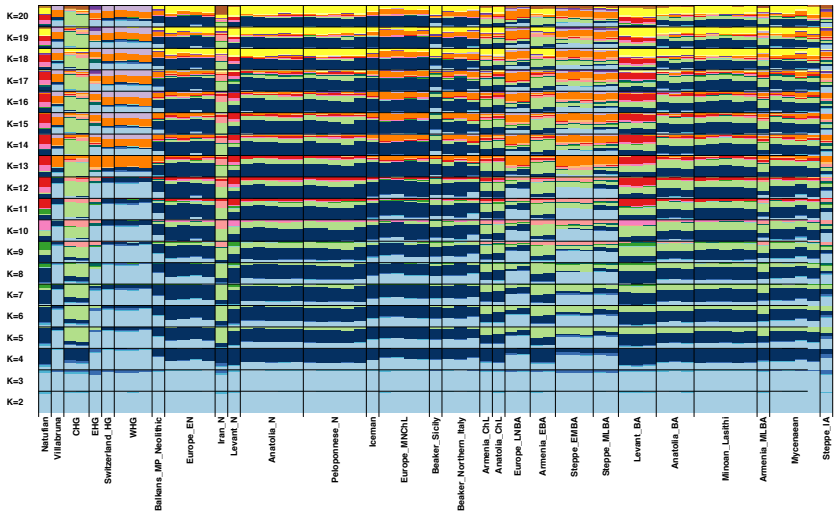


Figure 72. ADMIXTURE analysis of 63 ancient samples. Ancestral allele frequencies were inferred from ten different ADMIXTURE runs on 4,606 modern samples and projected onto the ancient samples. Each bar represents an individual grouped into ancient groups.

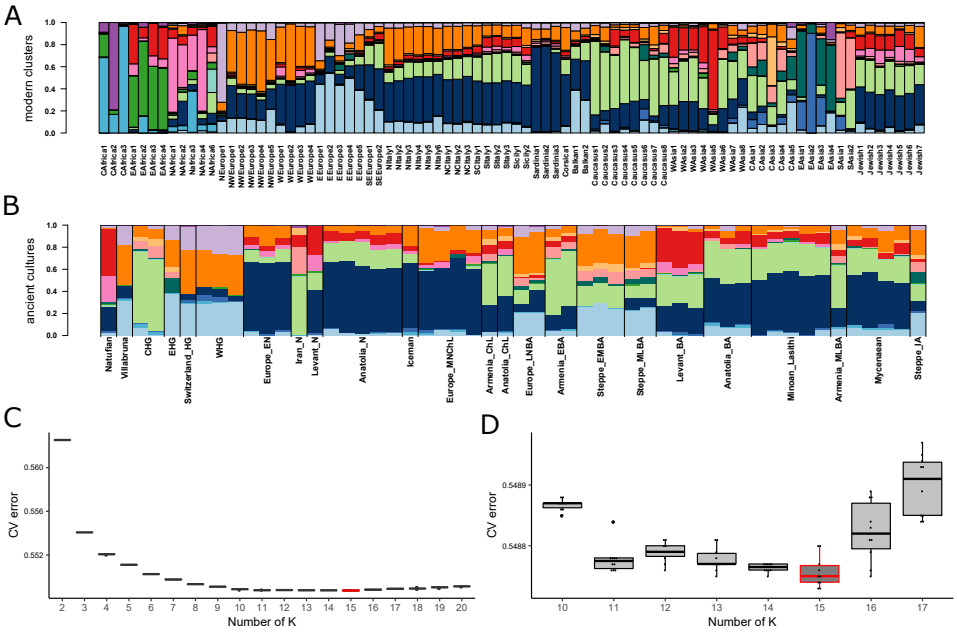


Figure 73. ADMIXTURE analysis of 63 ancient samples and 4,606 modern samples for K=15. A-B) Results of the ADMIXTURE analysis as in Figures 72 and B.2 for $K = 15$ including both modern **(A)** and ancient samples **(B)**. **C)** Box plots of the ten CV-errors for $K = \{2, \ldots, 20\}$. **D)** Detailed box plots for the ten CV-errors for $K = \{10, \ldots, 17\}$.

The colours that made us

We tested different combination of ancient putative sources using the "unlinked" mode implemented in CP and a "mixture fit" approach (non-negative least square algorithm, NNLS; Leslie *et al.,* 2015; Montinaro *et al.,* 2015; Lawson & Hanson, 1995). We want to point out that the "unlinked" approach makes CP similar to other routinely performed analyses based on genotype data, such as qpAdm and ADMIXTURE; however, the strength of the pipeline itself lays inside the subsequent mathematical solution to the NNLS problem (see page 155).

In order to investigate how and to what extent the ancient demographic events have shaped the modern-day Italian and European genetic variations, we built the *ultimate* and *proximate* sets of ancient samples and we applied the CP/*NNLS*approach.

Notably, in both analyses, Southern Italian clusters showed contributions from Northern Africa (reaching a maximum of 6.6% and 11% in Sicily2, for

ultimate and *proximate* analysis, respectively), confirming the historical admixture signatures which Dr. Raveane detected by GT/MALDER (Raveane *et al.,* 2019), and as previously reported (Busby *et al.,* 2015; Hellenthal *et al.,* 2014; Sazzini *et al.,* 2016; Capelli *et al.,* 2007; Cerezo *et al.,* 2012; Fiorito *et al.,* 2016).

Ultimate sources: Neolithic rules

We assembled the set of *ultimate* sources by including one representative ancient sample for Western Hunter Gatherers (WHG), Caucasus Hunter Gatherers (CHG), Eastern Hunter Gatherers (EHG), Anatolian Neolithics (AN) and Iranian Neolithics (IN) and we applied the CP/*NNLS* approach (Figure 74).

When *ultimate* sources were considered, we observed the pervasive presence of AN ancestry across all European populations. In particular, all the Italian clusters were characterised by relatively high amounts of Anatolian Neolithic (AN), ranging between 56% (SItaly1) and 72% (NItaly4), distributed along a North-South cline (Spearman $\rho = 0.52$, p-value < 0.05; Figures 74, Figures 75, and Table B.3), with Sardinians showing values above 80%, as previously suggested (Sarno *et al.,* 2017; Haak *et al.,* 2015).

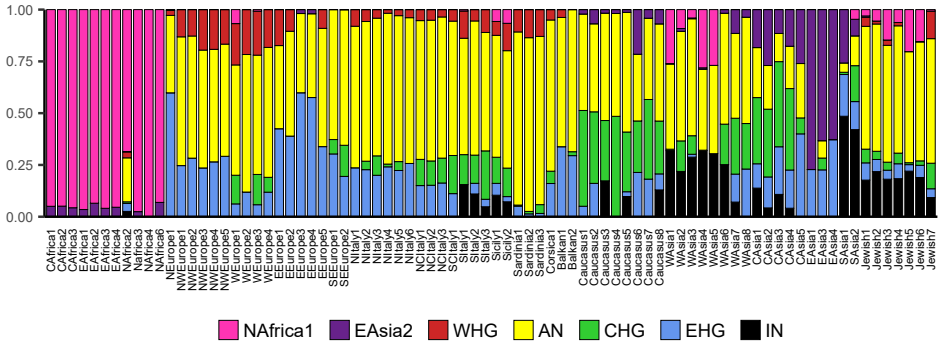


Figure 74. CP/*NNLS* results for *ultimate* sources reporting all modern Eurasian and African clusters.

The remaining ancestry was mainly assigned to WHG (Western Hunter-Gatherer), CHG and EHG. In particular, the first two components were more present in populations from the South of Italy (Student's t-Test p-value < 0.05), while the latter was higher in Northern Italian clusters (Student's t-Test p-value < 0.05). These observations suggest the existence of different

secondary sources contributions to the two edges of the peninsula, with the North affected more by EHG-related populations and the South affected more by CHG-related groups. Iran Neolithic (IN) ancestry was detected in Europe only in Southern Italy (Figures 74 and 75E).

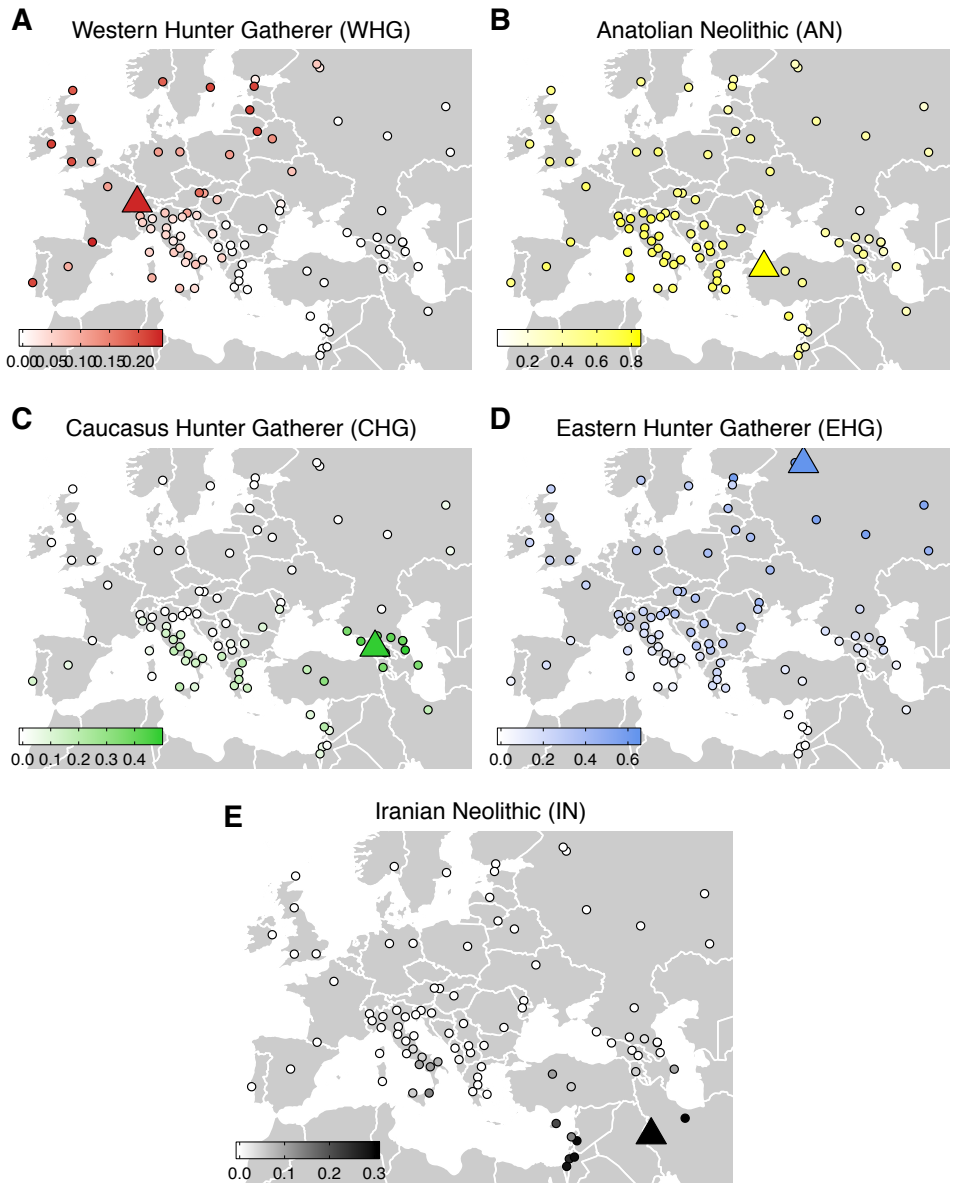


Figure 75. CP/*NNLS* results for *ultimate* sources in Western Eurasia. (A) WHG (Western Hunter Gatherer, Bichon), **(B)** AN (Anatolian Neolithic, Bar8), **(C)** CHG (Caucasus Hunter Gatherer, KK1), **(D)** EHG (Eastern Hunter Gatherer, I0061) **(E)** and IN (Iranian Neolithic, WC1) contributions in Western Eurasia, as inferred in Figure 74.

In the *ultimate* sources, all the scrutinised ancestries resulted significantly associated with longitude and latitude across Europe, except for EHG, that did not correlate with longitude (Figure 76A, B). Interestingly, the distributions of the Anatolian and Iranian Neolithic (IN) ancestries were opposite to those of some HG groups. Particularly, the Neolithic contributions (from both Anatolia and Iran) and CHG ancestry were negatively correlated with latitude (AN: $\rho = -0.25$, p-value $= 0.018$; IN: $\rho = -0.64$, p-value $= 1.9e^{-11}$; CHG: $\rho = -0.71$, p-value $= 6.4e^{-15}$). Conversely, Western and Eastern Hunter Gatherer ancestries resulted positively correlated with latitude (WHG: $\rho = 0.51$, p-value $= 4.5e^{-07}$; EHG: $\rho = 0.84$, p-value $< 2.2e^{-16}$). The strong positive correlation exhibited by the EHG contribution could be a proxy for the "steppe-related" ancestry, which arrived in Western Europe between the Late Neolithic and the Bronze Age period. Caucasus and Western Hunter Gatherer ancestries followed opposite correlation patterns with longitude (WHG: $\rho = -0.70$, p-value $= 1.4e^{-14}$; CHG: $\rho = 0.40$, p-value $= 9.6e^{-05}$), as well as Anatolian and Iranian Neolithic ancestries (AN: $\rho = -0.68$; p-value $= 1.5e^{-13}$; IN: $\rho = 0.26$, p-value $= 0.014$). When considering the Italian groups, all the ancestries were associated with latitude and longitude, except for WHG (Figure 76C, D). Notably, the EHG contribution is strongly correlated with latitude, decreasing while moving South along the Italian Peninsula (EHG: $\rho = 0.97$, p-value $= 4e^{-12}$), in contrast to the Iranian Neolithic ancestry, which is present mainly in Southern Italy (IN: $\rho = -0.74$, p-value $= 0.00017$). As expected, given the characteristic shape of the Italian country (mostly distributed in a relatively narrow strip along a North-South axis), the correlation between ancestry and longitude was weak, but still significant for the IN contribution: the closer the samples were to the Middle East, the higher the Iranian Neolithic signature was observed (IN: $\rho = 0.77$, p-value $= 6.5e^{-05}$, Figure 76D).

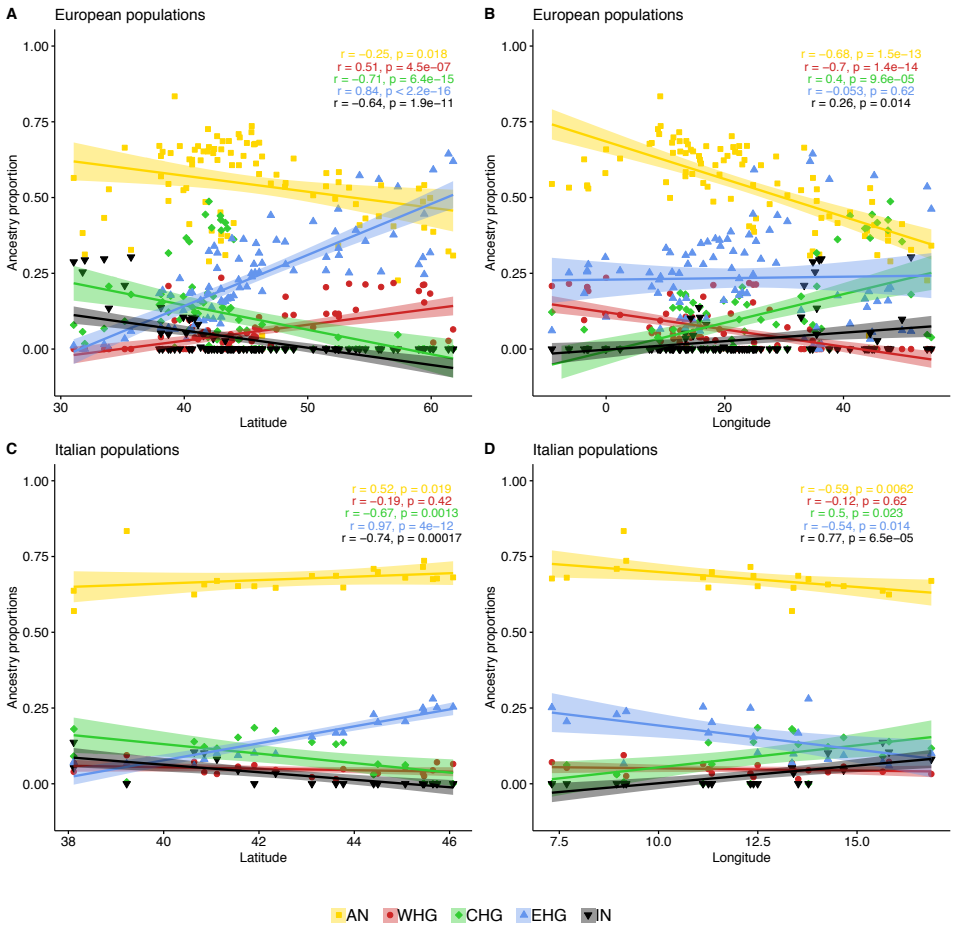


Figure 76. Spearman correlations between *ultimate* sources ancestry components, estimated with the CP/*NNLS* analysis across European population and geography. **(A)** Correlations between *ultimate* sources ancestry in European populations and latitude. **(B)** Correlations between *ultimate* sources ancestry in European populations and longitude. **(C)** Correlations between *ultimate* sources ancestry in Italian populations and latitude. **(D)** Correlations between *ultimate* sources ancestry in Italian populations and longitude.

D-statistic results are in line with the CP/*NNLS* findings. In particular, *D*-statistic on *ultimate* sources showed a higher affinity of most Italian clusters with Neolithic farmers from Anatolia than with Palaeolithic for-

agers from Western Europe as demonstrated by $D(AN, WHG, Y, Mbuti)$ (Figure 77A). This held for all but three Northern Italian clusters (NItaly1, NItaly2, NItaly6). Furthermore, the D values increased along the North-South axis, suggesting a geography-correlated asymmetric relationship to Western European foragers and farmers, in line with *NNLS* results (Figure 74A). When the $D(AN, CHG, Y, Mbuti)$ was considered, all the Italian clusters were significantly closer to the Neolithic farmers (Figure 77B). Interestingly, clusters from Southern Italy showed smaller positive values (with the exception of one Northern Italian group), suggesting a closer relationship between populations of the Central and Southern part of the peninsula and populations related to CHG. On the other hand, Sardinians showed the highest D values, which pointed to a smaller impact by CHG related populations. Similar results were obtained when IN was used instead of CHG. The opposite pattern emerged when the $D(AN, EHG, Y, Mbuti)$ was tested, which showed smaller positive values associated with Northern Italy (Figure 77C). All the Sardinian clusters showed WHG signatures (at least 11%), as reflected in the significant negative values of the D-statistic $D(CHG/EHG, WHG, Sardinian, Mbuti)$.

Other evidence reporting the closer affinity of Northern Italian than Southern Italian clusters to AN were the significantly negative values in the *D(South Italy, North Italy, AN, Mbuti)* in Figures B.4 and 86.

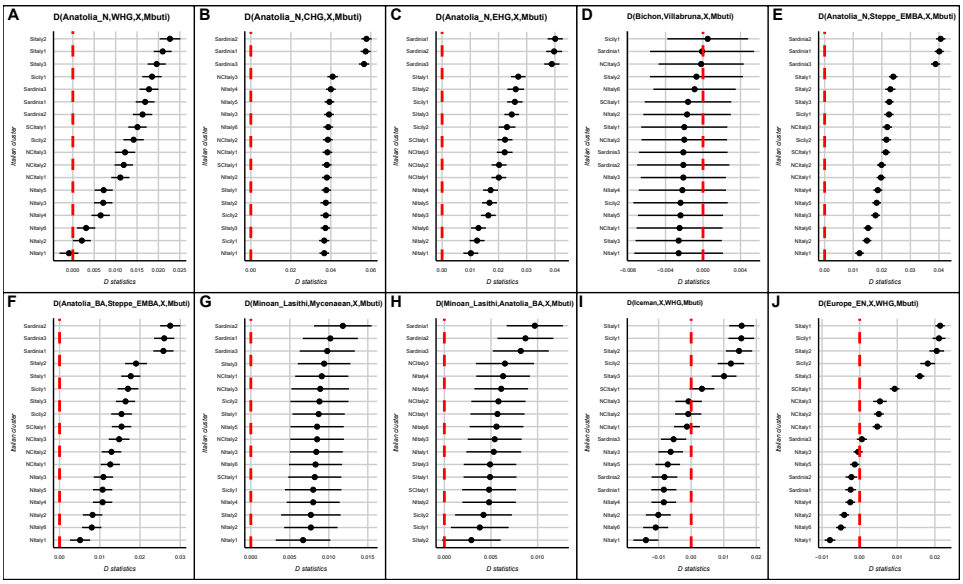


Figure 77. A selection of *D*-statistics using Italian clusters as *X*.

Proximate sources: something stirs in the East

Then, we built a second set of ancient sources (*proximate*) by considering European Neolithic (EEN), Steppe Bronze Age (SBA) and Anatolia Bronze Age (ABA) specimens. As before, we performed a CP run and the *NNLS* solution (Figure 78). This second set is chronologically more recent than the *ultimate* dataset and is expected to provide an indication for the correspondence between modern and post-Neolithic genetic variation.

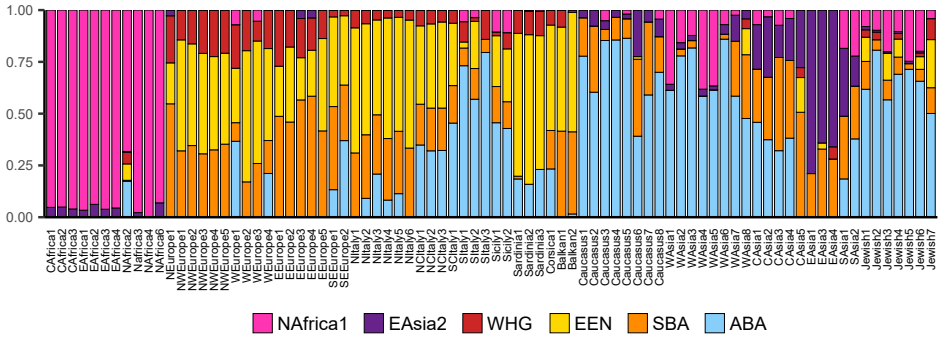


Figure 78. CP/*NNLS* results for *proximate* sources reporting all modern Eurasian and African clusters.

A notable result is that, as expected from literature, SBA contribution appeared clearly across Europe and the Italian peninsula (Figure 78). However, setting itself apart from other European regions, Italy harboured a strong signal related to ABA, which was paralleled by a substantial reduction in EEN ancestry in Southern Italy (Figures 78 and 79B, D).

As did the *ultimate* sources, also the *proximate* analyses highlighted north-south differences across the Italian population. As a matter of fact, the SBA component was higher in Northern than in Southern Italy (Wilcoxon rank sum test p-value = 0.00040), while ABA followed the opposite pattern with significantly higher contributions in Southern Italy (Student's t-Test p-value = 0.00044).

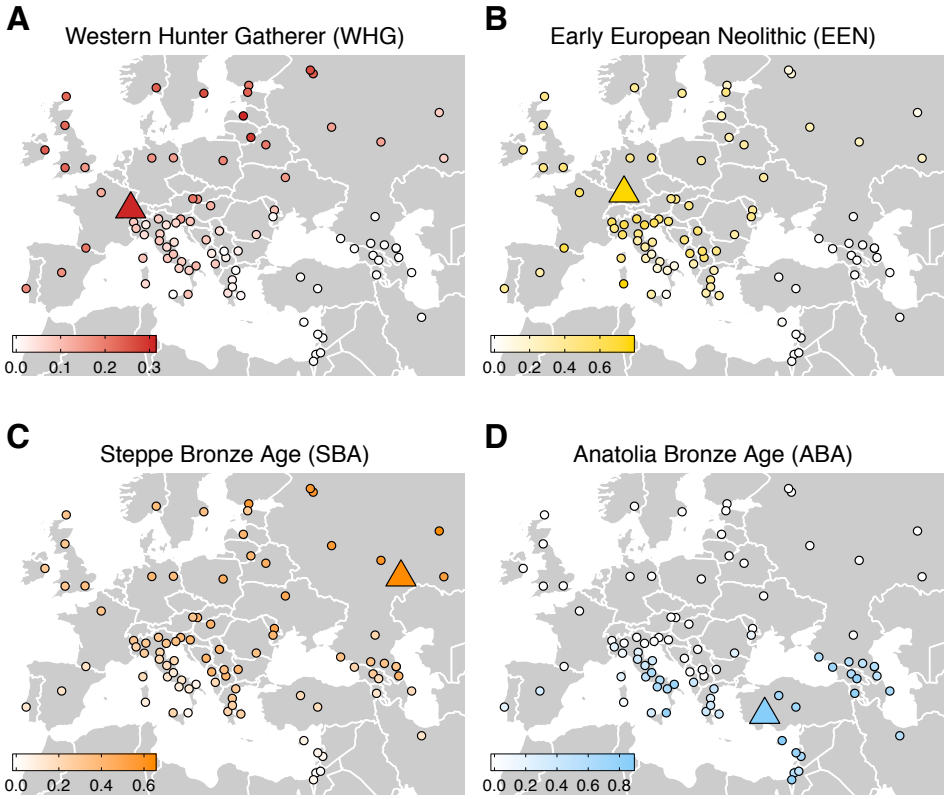


Figure 79. CP/*NNLS* results for *proximate* sources in Western Eurasia. (A) WHG (Western Hunter Gatherer, Bichon) **(B)** EEN (Early European Neolithic, Stuttgart), **(C)** SBA (Steppe Bronze Age, I0231), **(D)** ABA (Anatolian Bronze Age, I2683) contributions in Western Eurasia, as inferred in Figure 78.

Interestingly, the distribution of ABA and SBA mirror the CHG and EHG patterns, respectively (Figure 74). Indeed, contrary to previous reports (Lazaridis *et al.,* 2016), the occurrence of CHG as detected by our CP/*NNLS* analysis did not mirror the presence of SBA. When we compared this analysis and the one using a different CHG sample (SATP, Fu *et al.,* 2016), the two were highly correlated (Spearman $\rho = 0.972$, p-value $< 2.2e^{-16}$; Figure 80). We therefore speculate that our approach might in general underestimate the presence of CHG across the continent; however, we note that even considering this scenario, the excess of Caucasus related ancestry detected in the South of the European continent, and in Southern Italy in particular,

is striking and unexplained by currently proposed models for the peopling of the continent.

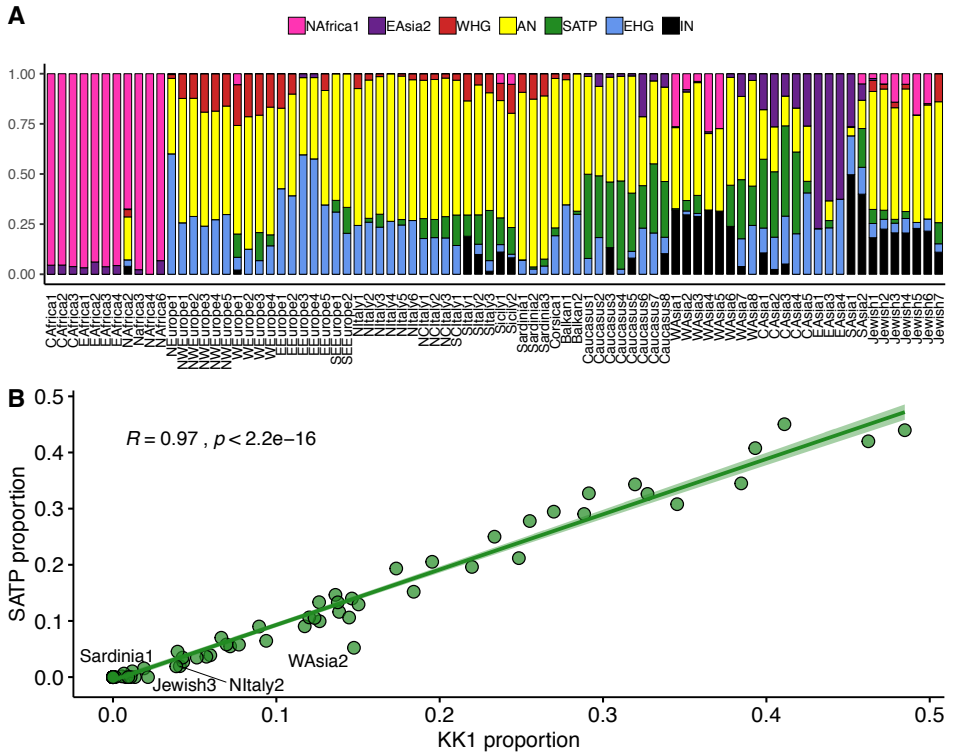


Figure 80. CP/*NNLS* results for *ultimate* sources for all modern clusters, replacing KK1 with SATP as CHG source. (A) *Ultimate* sources analysis reporting all modern Eurasian and African clusters and replacing KK1 with SATP as a CHG source. **(B)** Correlation between KK1 and SATP proportion of ancestry inferred for all modern clusters. Labelled points represent modern clusters where $|\log_{10}(\frac{KK1}{SATP})| > 0.3$.

The different impact of ABA and SBA across Italy was additionally supported by the reduced fit of the *NNLS* (sum of the squared residuals) when the *proximate* analysis was run excluding one of these two sources. The residuals were almost up to twice as much for Southern Italians when ABA was not included as a source, while a substantial increase in the residual values was observed for Northern Italians when SBA was removed from the panel of *proximate* sources (Figure 81).

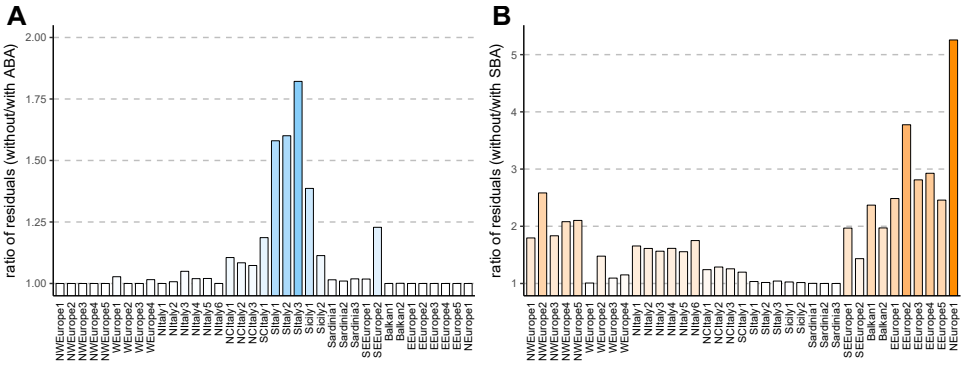


Figure 81. Residuals of the *NNLS* analysis for all the Italian and European clusters. (A) Ratio of the residuals in the *NNLS* analysis for all the Italian and European clusters when ABA was excluded and included in the set of *proximate* sources; **(B)** as in **(A)**, but excluding/including SBA instead of ABA.

Other evidence of the different affinities of Southern and Northern Italians for ABA and SBA were also highlighted in the PCA and ADMIXTURE analysis on ancient and modern samples (Figures 88, 70, 71, 72, B.2).

However, this peculiar genetic affinity between Southern Italy and ABA, did not appear to be driven by the presence of AN ancestry: significantly negative values were observed in the *D*-statistic D(AN, ABA, CHG/IN, Mbuti) and Northern Italian samples were closer to AN than Southern samples (negative values of the *D*-statistic *D(Southern Italian cluster, Northern Italian cluster, AN, Mbuti)*).

Moreover, the pattern of Southern Italy was robust to the use of alternative SEE sources (Mycenaean - MYC and Minoan - MIN) and Peloponnese Neolithic (PN), suggesting a broad link with this region and chronology (Lipson *et al.,* 2017, Figure 82). Indeed, as for ABA source, MIN, MYC and PN ancestries resulted more prevalent in Southern than in Northern Italian clusters (Wilcoxon rank sum test p-value = 0.02109, Student's t-Test p-value = 0.002631 and Student's t-Test p-value = 0.01279, respectively). Signatures related to SEE and PN sources were less evident in clusters from the Balkan Peninsula (Figures 78, 79D and 82). As above-said for ABA, also these contributions mirrored the IN/CHG ancestries observed in Southern Italian populations and could explain their occurrence.

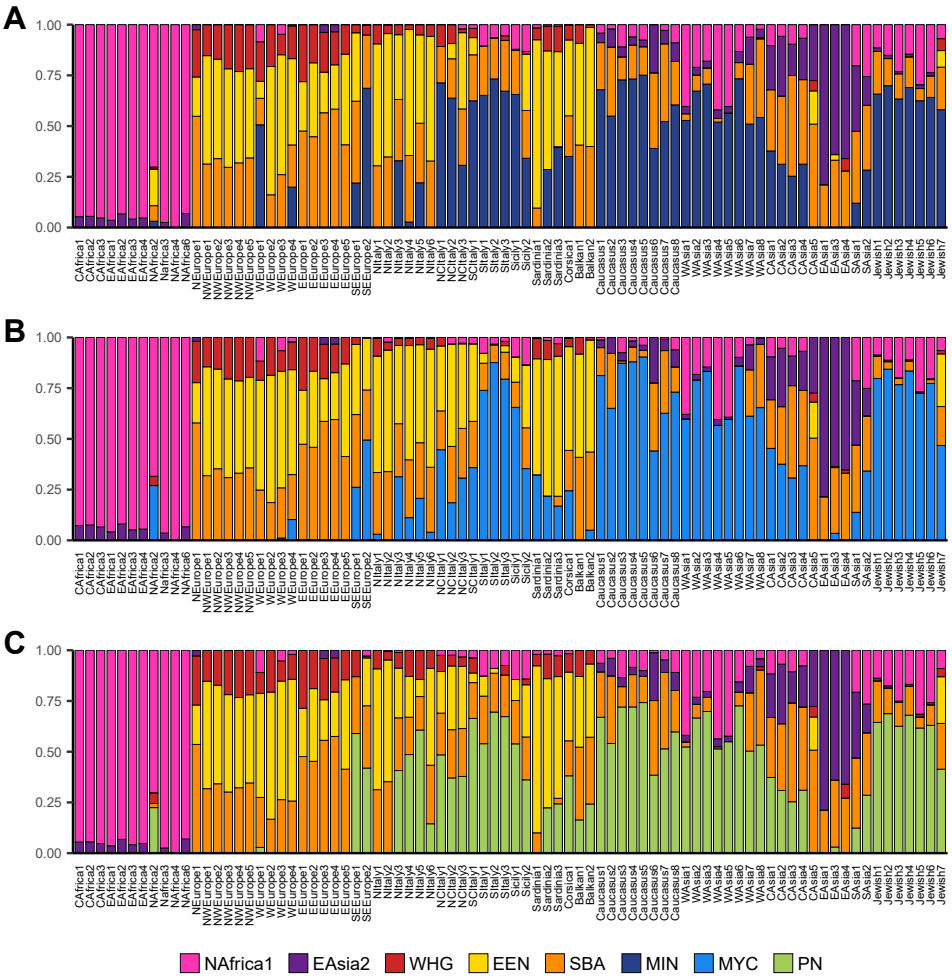


Figure 82. CP/*NNLS* results for *proximate* sources for all modern clusters using alternative SEE sources. *Proximate* sources analysis replacing ABA with alternative SEE sources: **(A)** Minoan, MIN, **(B)** Mycenaean, MYC, **(C)** Peloponnese Neolithic, PN.

Then, as we did for *ultimate* sources proportions, we explored the relationship between the estimates for *proximate* sources (Table B.4 and Figure 78) in relation to latitude across European populations (Figure 83A): while EEN was present with minimal variation across European countries (EEN: $\rho = 0.25$, p-value $= 0.019$), the two Bronze Age contributions (SBA and ABA) followed opposite correlation patterns (SBA: $\rho = 0.90$, p-value $< 2.2e^{-16}$; ABA: $\rho = -0.74$, p-value $< 2.2e^{-16}$). When we analysed

the longitude correlations (Figure 83B), we observed that, while the EEN ancestry strengthened its correlation with geography (EEN: $\rho = -0.83$, p-value $< 2.2e^{-16}$), the Steppe Bronze Age contribution showed no correlation with longitude ($\rho = -0.091$, p-value $= 0.4$) and the correlation for ABA became less evident (ABA: $\rho = 0.49$, p-value $= 1.1e^{-06}$). When we considered only the Italian populations, both latitude (Figure 83C) and longitude (Figure 83D) resulted more strongly correlated with the three proximate ancestries (Latitude: EEN: $\rho = 0.67$, p-value $= 0.0013$; SBA: $\rho = 0.96$, p-value $= 3.4e^{-11}$; ABA: $\rho = -0.77$, p-value $= 6.5e^{-5}$; Longitude: EEN: $\rho = -0.82$, p-value $= 8.5e^{-06}$; SBA: $\rho = -0.57$, p-value $= 0.0087$; ABA: $\rho = 0.8$, p-value $= 2e^{-05}$).

Finally, when we replaced the ABA source with the recently characterised Minoan (Figure 84) and Mycenaean populations (Figure 85), the results mimicked the significant trends shown by ABA (Figure 83) for the correlations with latitude and longitude. As reported in the main text, we decided to add WHG to the *proximate* sources in order to take into consideration the admixture dynamics between farmers and hunter-gatherers. The addition of the hunter-gatherer component did not alter the general patterns of correlation found in the *proximate* sources ABA, MIN and MYC. However, the inclusion of WHG slightly weakened some correlations when other more recent ancestries, such as ABA, MIN and MYC, were considered (Figures B.5, B.6, B.7).

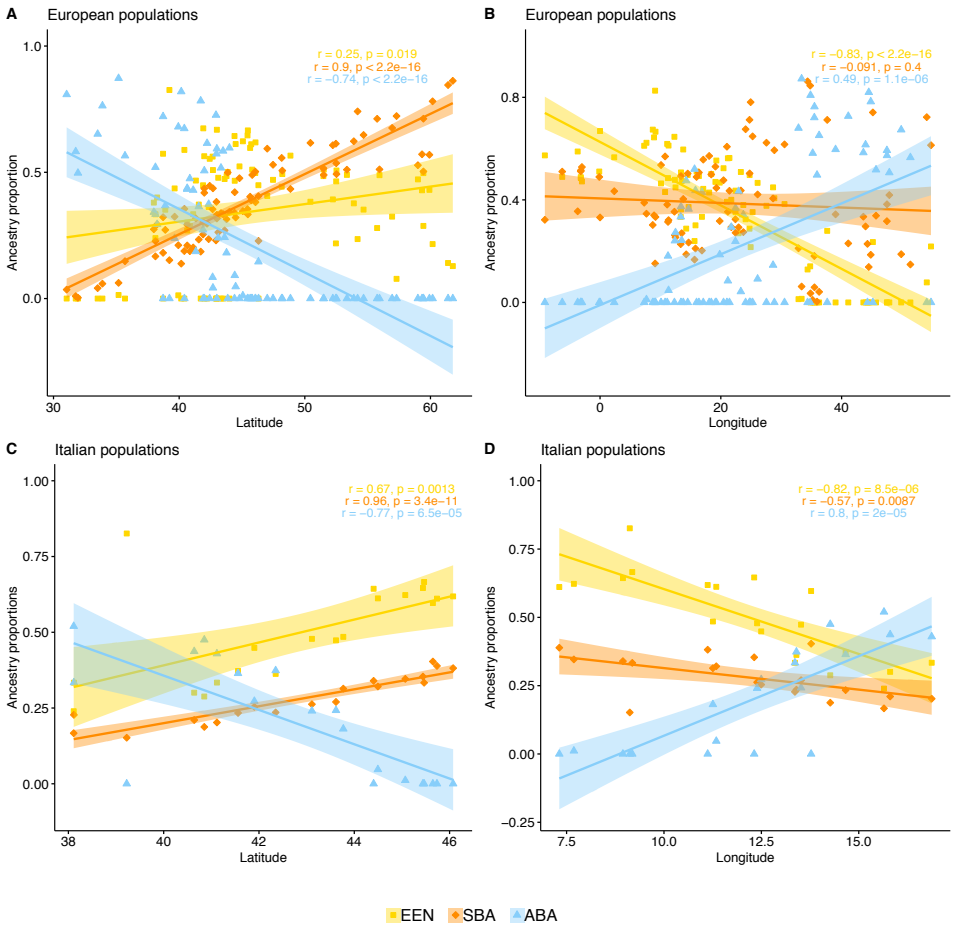


Figure 83. Spearman correlations between *proximate* sources ancestry components (considering ABA as SEE source), estimated with the CP/*NNLS* analysis and geography. (A) Correlations between *proximate* sources ancestry (ABA as SEE source) in European populations and latitude. **(B)** Correlations between *proximate* sources ancestry (ABA as SEE source) in European populations and longitude. **(C)** Correlations between *proximate* sources ancestry (ABA as SEE source) in Italian populations and latitude. **(D)** Correlations between *proximate* sources ancestry (ABA as SEE source) in Italian populations and longitude.

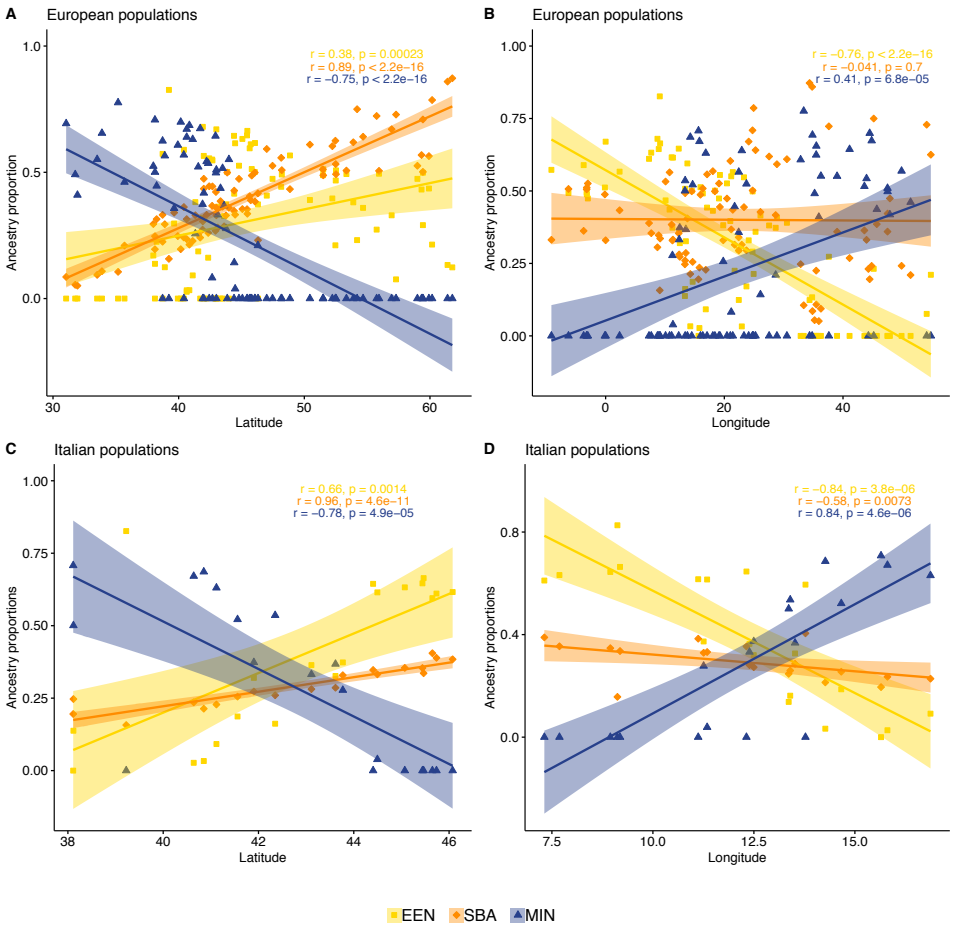


Figure 84. Spearman correlations between *proximate* sources ancestry components (considering MIN as SEE source), estimated with the CP/*NNLS* analysis and geography. (A) Correlations between *proximate* sources ancestry (MIN as SEE source) in European populations and latitude. **(B)** Correlations between *proximate* sources ancestry (MIN as SEE source) in European populations and longitude. **(C)** Correlations between *proximate* sources ancestry (MIN as SEE source) in Italian populations and latitude. **(D)** Correlations between *proximate* sources ancestry (MIN as SEE source) in Italian populations and longitude.

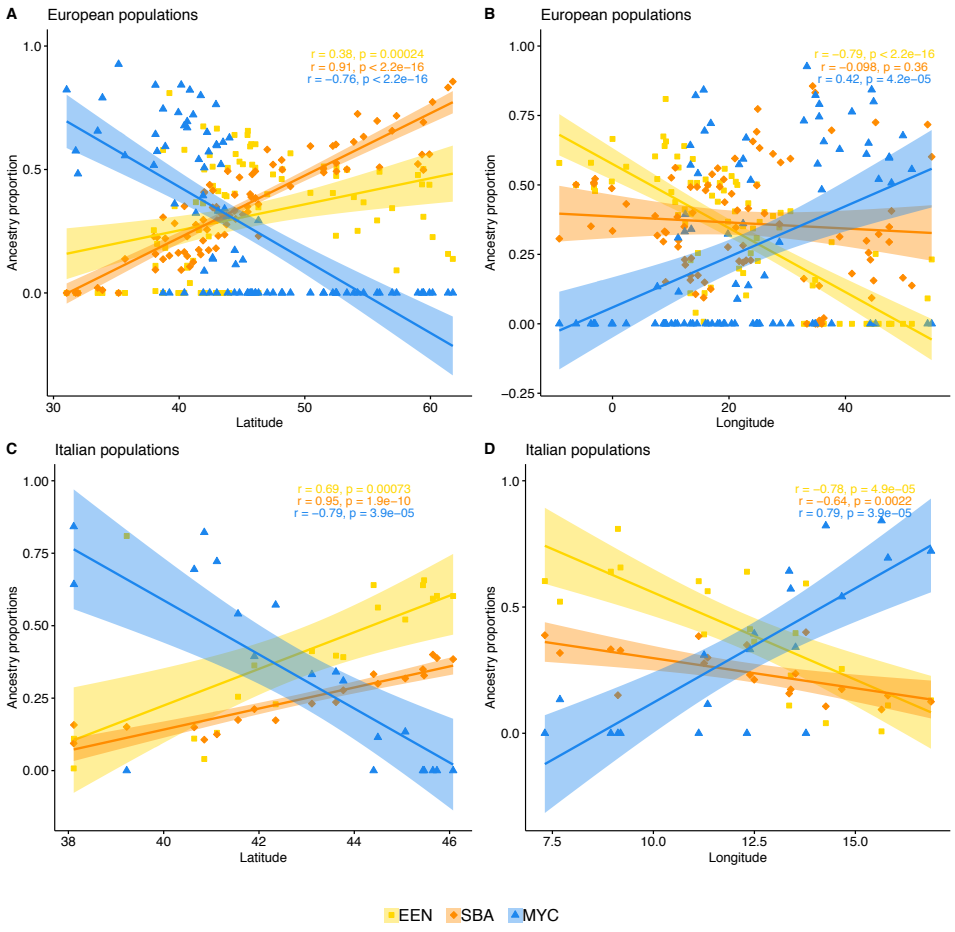


Figure 85. Spearman correlations between *proximate* sources ancestry components (considering MYC as SEE source), estimated with the CP/*NNLS* analysis and geography. (A) Correlations between *proximate* sources ancestry (MYC as SEE source) in European populations and latitude. **(B)** Correlations between *proximate* sources ancestry (MYC as SEE source) in European populations and longitude. **(C)** Correlations between *proximate* sources ancestry (MYC as SEE source) in Italian populations and latitude. **(D)** Correlations between *proximate* sources ancestry (MYC as SEE source) in Italian populations and longitude.

Then, we performed *D*-statistics analysis in the form *D(AN, X, Bronze Age, Mbuti)*. When Early/Middle Bronze Age individuals from the steppe were used, they showed a strong relationship to all the Northern Italian groups, but not Southern Italian groups (except for the SCItaly1 cluster, $Z < -3$), suggesting a stronger impact of steppe ancestry in the North. For the Italian clusters, all the *D*-statistics in the form *D(ABA, SBA Y, Mbuti)* showed positive significant results, highlighting a higher affinity between ABA and the Italian groups, with larger positive values for Sardinia and Southern Italy (Figure 77G). This result may reflect the high proportion of AN characterizing both ABA and present-day Italians. However, as reported above, the CP/*NNLS* analysis and *D*-statistics in Figure B.4 showed that Northern Italian clusters had similar or slightly higher AN than the Southern ones. Moreover, when we directly tested the affinities of NItaly/Sardinia and SItaly/Sicily clusters to AN and ABA, we observed a closer affinity of the latter to ABA (Figure 86).

Thus, Southern Italian groups have a high affinity with populations related to ABA, which could explain at least in part the CHG/IN ancestry observed in these populations in the *ultimate* CP/NNLS analysis. In fact, CHG and IN were more related to ABA than AN, as revealed by the significant negative values of the *D*(AN, ABA, CHG/IN, Mbuti).

Furthermore, Minoans showed a higher affinity to Italians than any other Bronze Age individual tested, including Mycenaeans and ABA (Figure 77G, H). The differences between the two Hellenic samples might be driven by the Steppe signal identified in Mycenaeans (Lazaridis *et al.,* 2017). It may therefore be possible that Italian groups came into contact with a population either more related to the Minoans or to an outgroup to both Minoans and Mycenaeans with no significant contributions from the steppe.

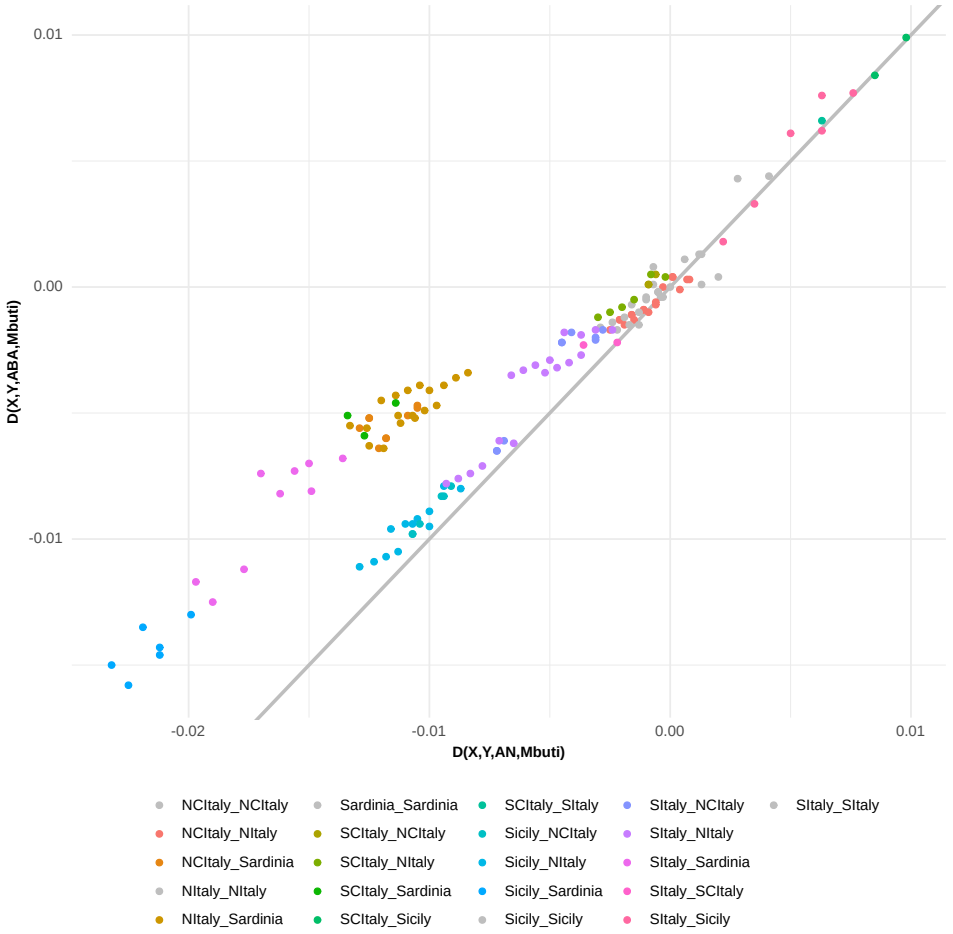


Figure 86. Comparison of AN and ABA affinity to Italian clusters using *D*-statistics. Scatter plot of D(Ita1, Ita2, AN, Mbuti) and D(Ita1, Ita2, ABA, Mbuti) for all the Italian clusters. Points for pairs of clusters from the same (grey points) or closely related geographic location fall in proximity of the grey line, reflecting a similar affinity to AN (x-axis) and ABA (y-axis). Comparisons of clusters from NItaly/Sardinia and SItaly/Sicily fall above the grey line, reflecting a closer affinity of the latter to ABA.

Interestingly, these interesting results about the Bronze Age contribution in Italy were also confirmed by the qpAdm analyses performed by Dr Montinaro, where all the analysed Italian clusters could be modelled as a combination of ABA, SBA and European Middle-Neolithic/Chalcolithic,

their contributions mirroring the pattern observed in the CP/NNLS analysis (Figure 87). Sardinian clusters were consistently modeled as AN + WHG + CHG/IN across runs, with the inclusion of North Africa and SBA when different number of sources were considered. The qpAdm analyses of Italian HDD clusters generated similar results.

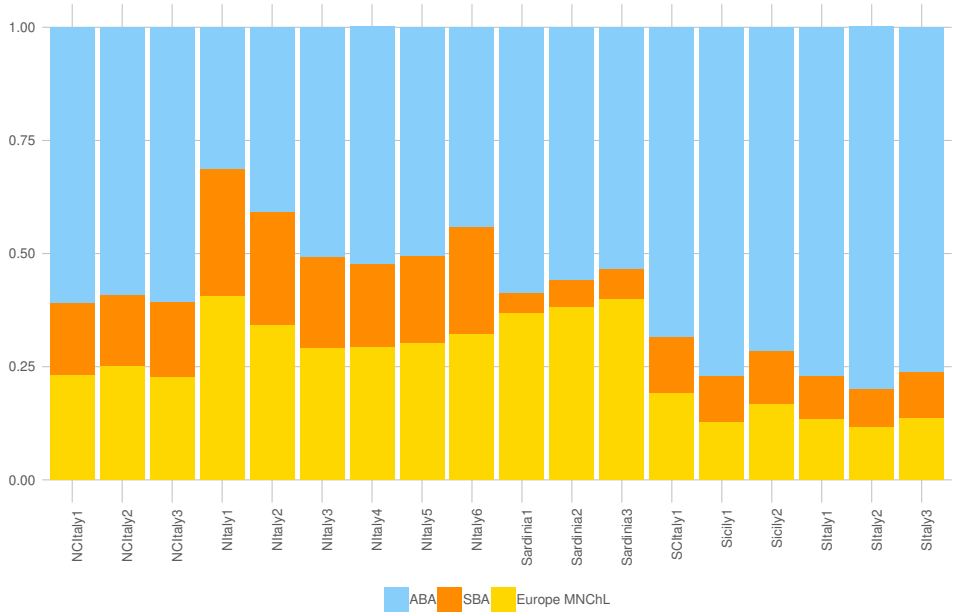


Figure 87. Mixture proportions on modern Italian clusters inferred by qpAdm as a combination of ABA, SBA and European Middle-Neolithic/Chalcolithic. For each tested cluster, we have evaluated all the possible combinations of N "left" sources with $N = \{2, \ldots, 5\}$, and one set of right/left Outgroups.

Ancestry composition of ancient Italian samples

In order to obtain temporal insights on the emergence of the differences between Northern and Southern Italy in relation to SBA and ABA ancestries, we performed the CP/*NNLS* and *qpAdm* analyses on post-Neolithic/ Bronze Age Italian individuals. In Figure 88 we report a PCA performed with these ancient Italian samples and other ancient CP sources.

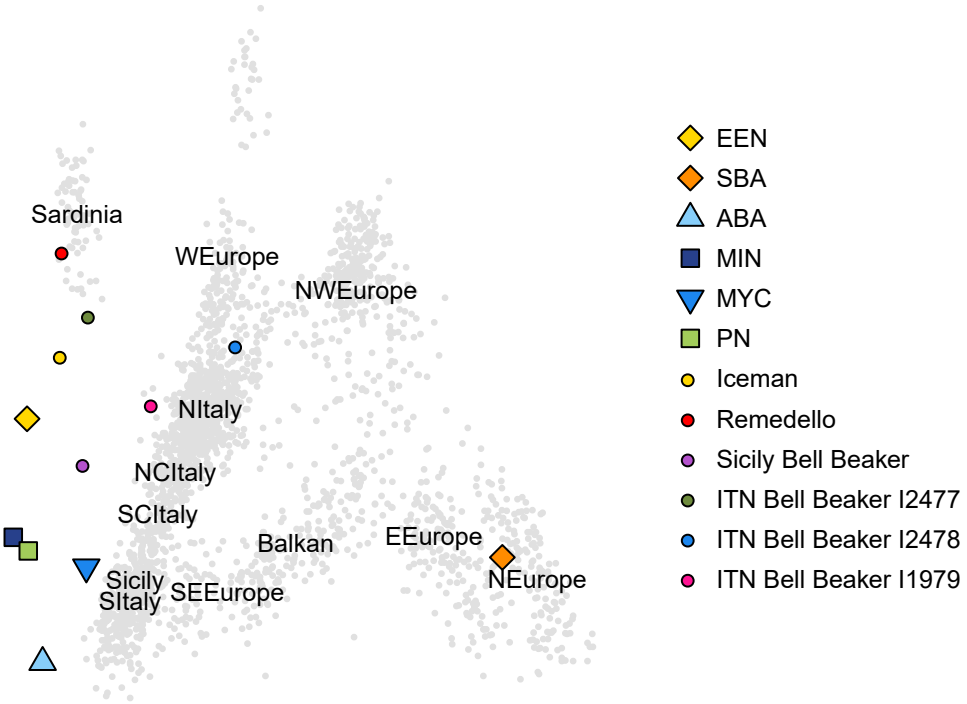


Figure 88. PCA of modern European individuals and ancient Italian and other selected ancient samples. Ancient Italian and other selected ancient samples projected on the components inferred from modern European individuals. Labels are placed at the centroid of the individuals belonging to the indicated clusters.

Northern Italian Copper Age samples dating to 5,000 years ago, Iceman (Keller *et al.,* 2012) and Remedello (Allentoft *et al.,* 2015; Lazaridis *et al.,* 2016), did not consistently show Bronze Age signatures (Figures 89 and B.8).

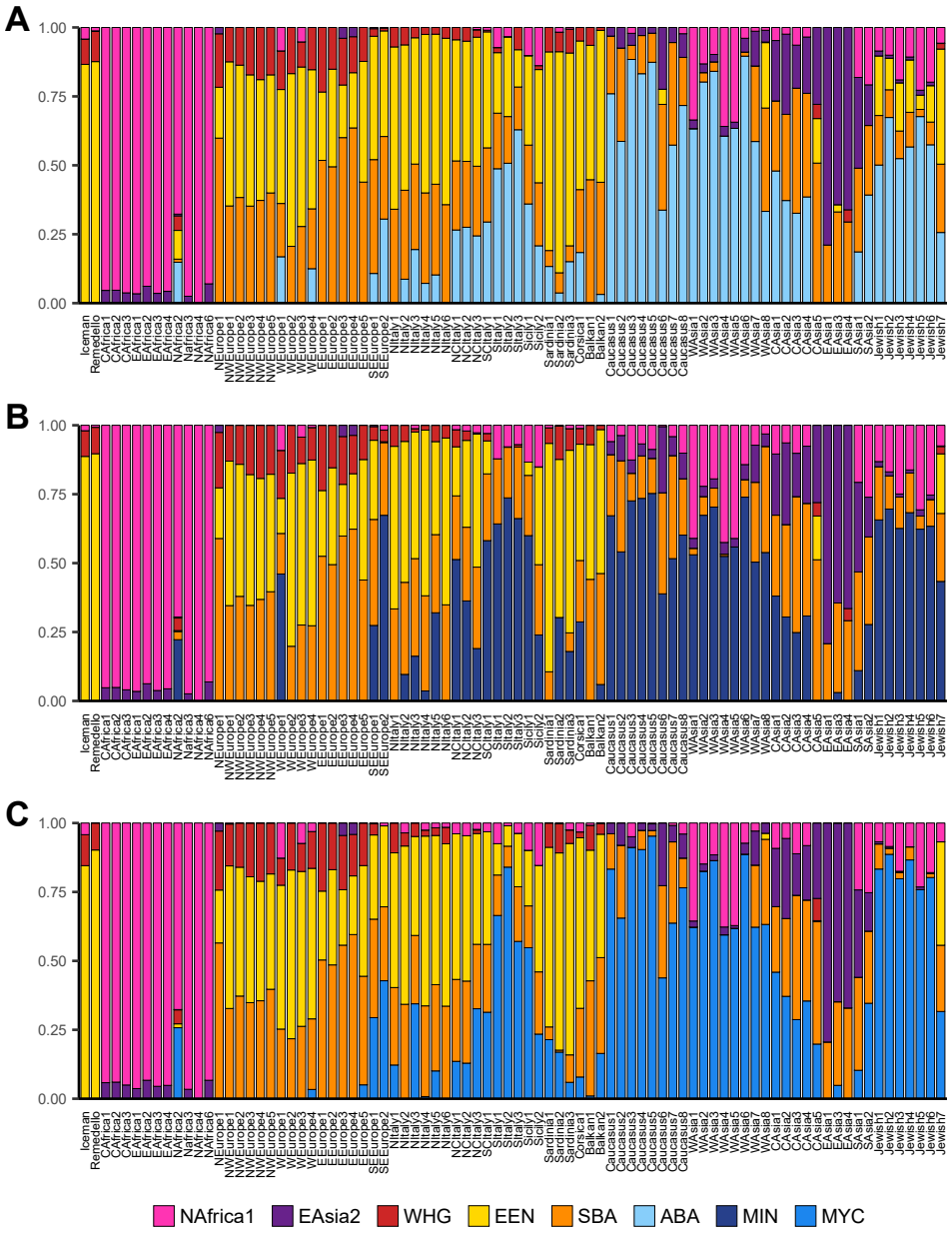


Figure 89. CP/*NNLS* results for *proximate* sources, including the Iceman and a Remedello samples as recipients. (A) *Proximate* sources analysis considering ABA as SEE source and including WHG as a source. **(B)** *Proximate* sources analysis considering Minoan as SEE source and including WHG as a source. **(C)** *Proximate* sources analysis considering Mycenaean as SEE source and including WHG as a source.

On the contrary, SBA ancestry was detected in all but one (I2477) of the Bronze Age Bell Beaker samples from North Italy (ITN Bell Beaker), associated with the spread of this cultural assemblage from Central Europe, as previously reported (Olalde *et al.,* 2018, 2,200-1,900 BCE; Figures 91 and B.9). Sample ITN Bell Beaker I2477 showed instead an ancestry profile similar to Iceman and Remedello (Figures 89 and B.8). These three individuals all clustered close to each other when ancient samples were projected on the PCA of modern-day data, near to Sardinian clusters, while the two ITN Bell Beaker samples with SBA ancestry were closer to modern Northern Italians (Figure 88). These observations suggested a non-homogenous presence of SBA ancestry at the time of the appearance of the Bell Beaker complex in Italy and some degree of continuity between Copper and Bronze age communities.

On the other hand, we consistently identified SEE/PN ancestries (15-43%; Figures 92 and 93) in a Sicilian Bell Beaker sample dating to 2,500-1,900 BCE (Olalde *et al.,* 2018) which suggested the presence in Italy of this additional ancestry from the East already 4,000 years ago, more than a thousand years prior to the establishment of Hellenic colonies in Southern Italy and unrelated to historical contributions from North Africa (Cunliffe, 2015).

In the PCA this sample mapped between Sardinia, Iceman and EEN on one side and Southern Italian clusters, SEE and PN samples on the other (Figures 88, 70 and 71). Overall, the ancestry profiles of ancient Italian samples supported the presence in Italy of two different continental signals already in the Bronze Age, each influencing mostly either Northern or Southern Italy. We also noted that in the Balkan peninsula signatures related to SEE/PN sources were less evident across modern-day populations, possibly masked by historical contributions from Central Europe (Busby *et al.,* 2015; Hellenthal *et al.,* 2014).

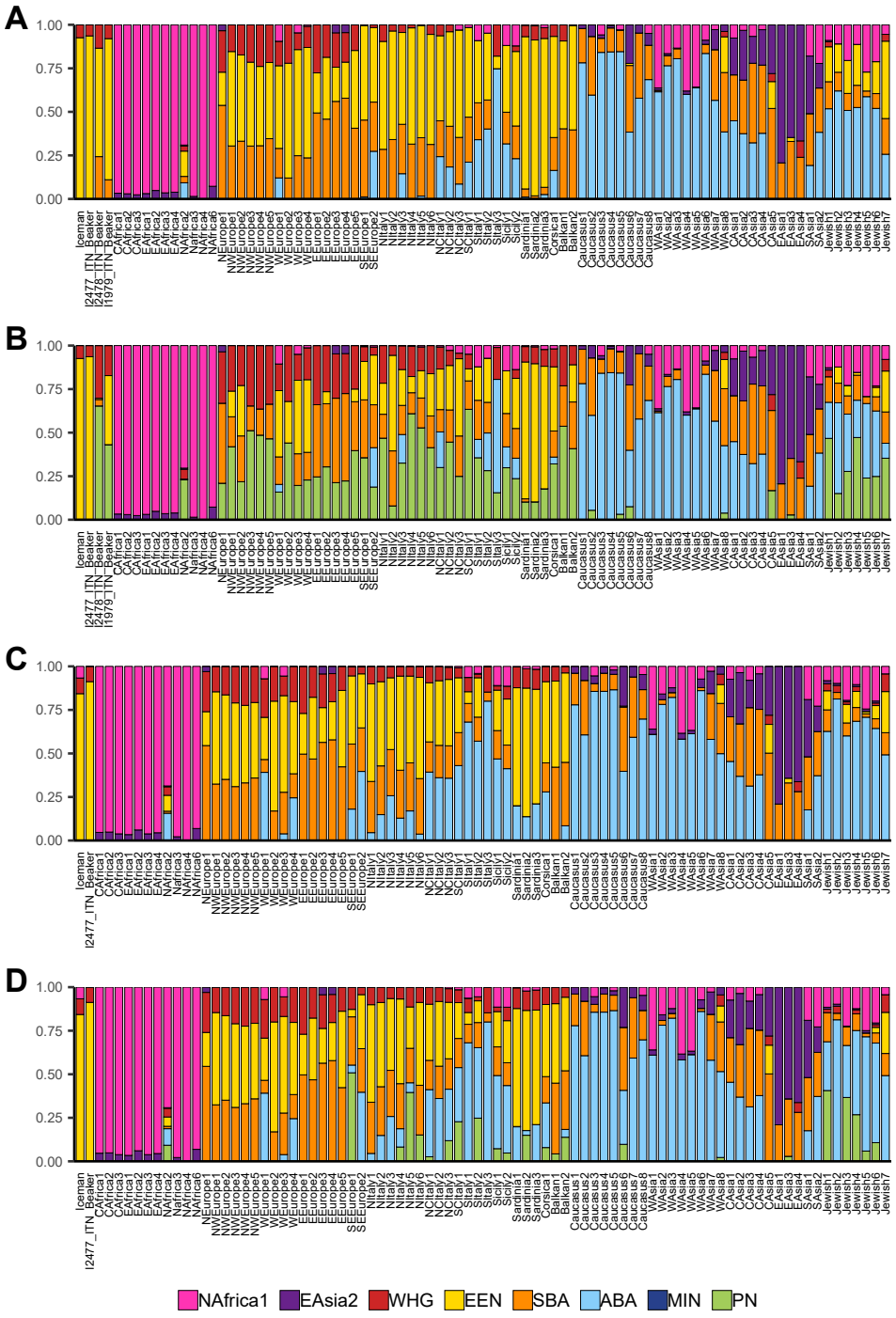


Figure 90. *Legend on next page.*

Figure 91. CP/*NNLS* results for *proximate* sources plus a PN sample, including modern-day clusters and the Iceman and ITN Bell Beaker samples as recipients. (A) *Proximate* sources analysis considering ABA as SEE source, including WHG as source and three ITN Bell Beaker samples as recipients. (B) *Proximate* sources analysis considering ABA and PN as SEE sources, including WHG as source and three ITN Bell Beaker samples as recipients. (C) *Proximate* sources analysis considering ABA as SEE source, including WHG as source and a ITN Bell Beaker sample (I2477) as recipient. (D) *Proximate* sources analysis considering ABA and PN as SEE sources, including WHG as source and a ITN Bell Beaker sample (I2477) as recipient.

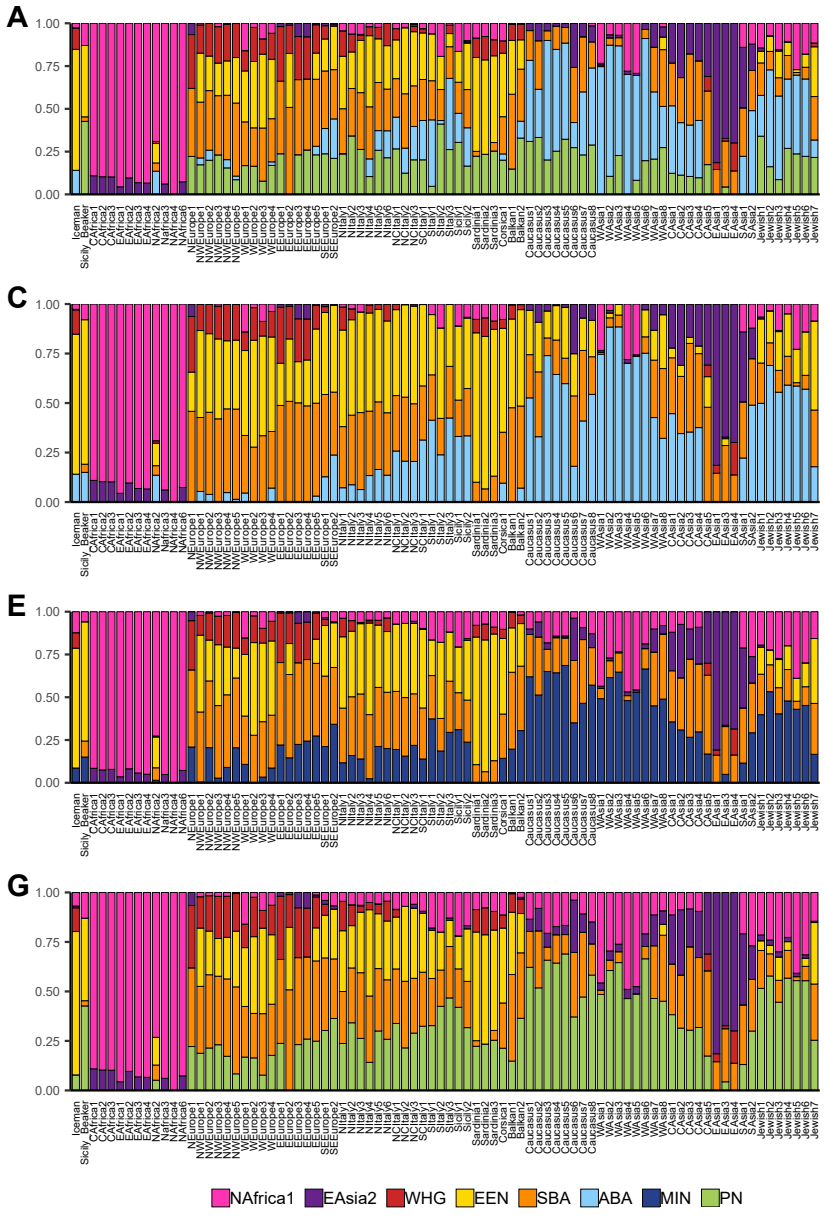


Figure 92. CP/*NNLS* results for *proximate* sources (EEN, SBA and WHG) and PN sample, including the Iceman and the Sicilian Bell Beaker samples, as recipients. *Proximate* sources analysis considering ABA and PN **(A)**, ABA **(B)**, MIN **(C)** and PN **(D)** as SEE sources, including WHG as source and a Sicilian Bell Beaker sample (I4930) as recipient.

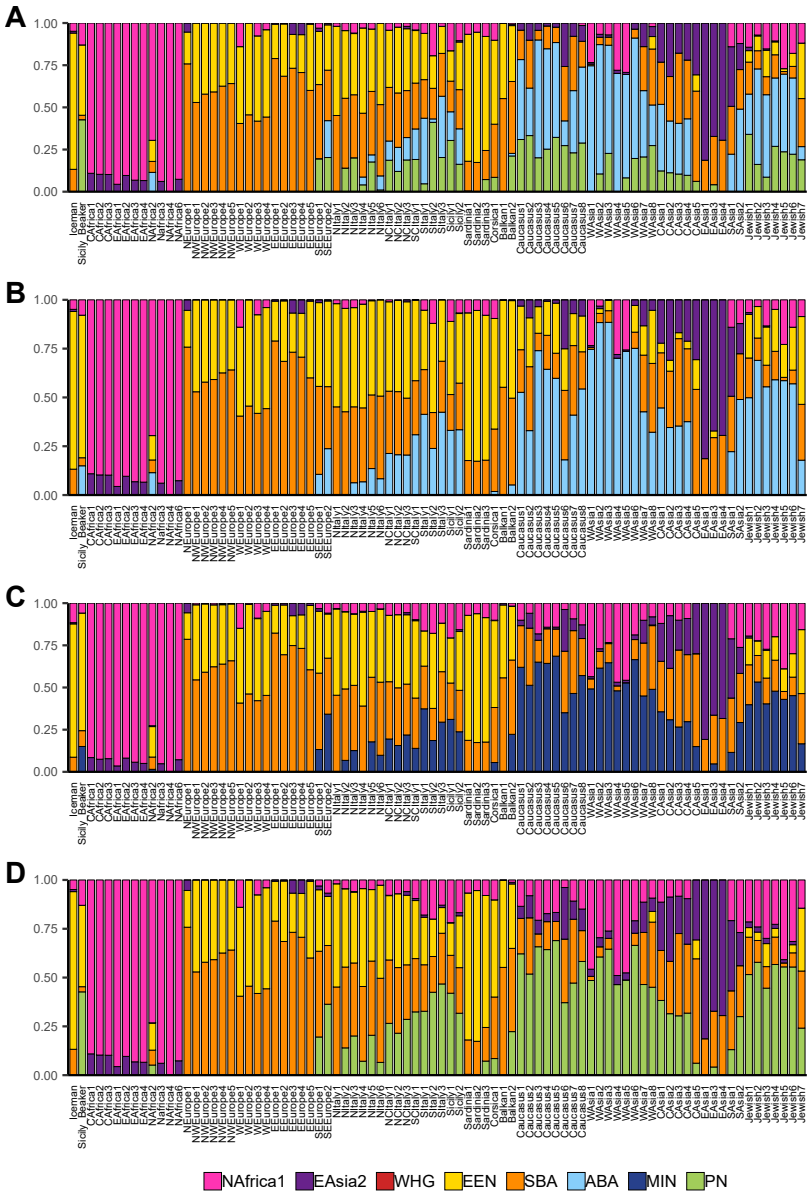


Figure 93. CP/*NNLS* results for *proximate* sources (EEN, SBA and WHG) and PN sample, including the Iceman and the Sicilian Bell Beaker samples, as recipients. *Proximate* sources analysis considering ABA and PN **(A)**, ABA **(B)**, MIN **(C)** and PN **(D)** as SEE sources, excluding WHG as source and a Sicilian Bell Beaker sample (I4930) as recipient.

Our CP/*NNLS* analyses suggest that this ancient component coming from the East predates the Hellenic colonization in Southern Italy, since this component was already present in the Sicilian Beaker sample from 4,000 years ago. However, the analyses on this sample have been done considering only ∼6,000 SNPs, thus raising some issues about the reliability of this result. Moreover, while in Figure 93 we see that only the Sicilian Beaker harbour this "exotic" component, this signature is weirdly present also in the Iceman sample when we consider WHG as a *proximate* source (Figure 92).

For this reason, we decided to model the ancestry composition of ancient Italian samples using *qpAdm*. As found in the CP/*NNLS* analysis, Iceman and Remedello, the oldest Italian samples here included (3,400-2,800 Before Common Era, BCE), were composed by high proportions of AN (74% and 85%, respectively, Table 6). The Bell Beaker samples of Northern Italy (2,200-1,930 BCE) were modelled as ABA and AN + SBA and WHG (Table 6 and 7). Although ABA estimates in these samples were characterised by large standard errors, the detection of Steppe ancestry, at approximately 14%, was more robust (Table 7). In contrast, Bell Beaker samples from Sicily 2,500-1,900 BCE) were modelled almost exclusively as ABA, with less than 5% SBA (Table 5). Despite the fact that the small number of SNPs and prehistoric individuals tested prevents the formulation of conclusive results, differences in the occurrence of AN ancestry, and possibly also Bronze Age related contributions, are suggested to be present between ancient samples from North and South Italy. As above-reported, differences across ancient Italian samples were also supported by their projections on the PCA of modern-day data (Figure 88). Remedello and Iceman clustered with European Early Neolithic samples, together with one of the three Bell Beaker individuals from North Italy, as previously reported (Olalde *et al.,* 2018), and modern-day Sardinians. The other two Bronze Age North Italian samples clustered with modern North Italians, while the Bell Beaker sample from Sicily was projected in between European Early Neolithic, Bronze Age Southern European and modern-day Southern Italian samples (Figure 88). These results suggest a differentiation in ancient ancestry composition between different areas of Italy dating at least in part back to the Bronze Age.

Target	**P1**	**P2**	**Prop1**	**Prop2**	**se1**	**se2**	**feasible**	**P-value**
ITN_Beaker	Anatolia_BA	WHG	0.904	0.096	0.016	0.016	Yes	0.026171
Beaker_Sicily	Anatolia_BA	Steppe_EMBA	0.985	0.015	0.125	0.125	Yes	0.077208
Beaker_Sicily	Anatolia_BA	EHG	0.941	0.059	0.11	0.11	Yes	0.071396

Table 5. qpAdm modelling of aDNA samples. For each tested cluster we have evaluated all the possible combinations of N "left" sources with N=2, and one set of right/left Outgroups.

Target	**P1**	**P2**	**P3**	**Prop1**	**Prop2**	**Prop3**	**se1**	**se2**	**se3**	**feasible**	**P-value**
ITN_Beaker	Anatolia_BA	Anatolia_N	WHG	0.658	0.221	0.12	0.268	0.235	0.038	Yes	0.045629
ITN_Beaker	Anatolia_BA	Anatolia_N	EHG	0.176	0.676	0.148	0.142	0.127	0.027	Yes	0.0172505
Iceman	Anatolia_BA	Anatolia_N	WHG	0.189	0.739	0.072	0.416	0.388	0.043	Yes	0.0238291
Remedello	Anatolia_BA	Anatolia_N	WHG	0.02	0.847	0.132	0.322	0.297	0.047	Yes	0.159551

Table 6. qpAdm modelling of aDNA samples. For each tested cluster we have evaluated all the possible combinations of N "left" sources with N=3, and one set of right/left Outgroups.

Target	P1	P2	P3	P4	Prop1	Prop2	Prop3	Prop4	se1	se2	se3	se4	feasible	P-value
ITN_Beaker	Anatolia_BA	Anatolia_N	Steppe_EMBA	WHG	0.129	0.602	0.139	0.13	0.216	0.171	0.053	0.028	Yes	0.31302
Iceman	Anatolia_BA	Anatolia_N	Steppe_EMBA	WHG	0.001	0.893	0.039	0.067	1.227	1.021	0.236	0.048	Yes	0.0148758

Table 7. qpAdm modelling of aDNA samples. For each tested cluster we have evaluated all the possible combinations of N "left" sources with N=4, and one set of right/left Outgroups.

Discussion

The genetic variation of European populations can be considered as the combination of ancestries related to the Mesolithic hunter-gatherers who inhabited Europe early on, the Anatolian farmers entering Europe around 10,000 years ago and the Bronze Age nomadic herders from the Pontic-Caspian steppes (Günther & Jakobsson, 2016; Nielsen *et al.,* 2017). The Bronze Age Steppe ancestry has been reported as the combination of ancestries related to hunter-gatherers from the East of Europe and the Caucasus (Lazaridis *et al.,* 2016; Jones *et al.,* 2015; Kılınç *et al.,* 2016). The major signatures of Neolithic farmers reported so far have been traced back to the Anatolian farmers (Lazaridis *et al.,* 2016; Kılınç *et al.,* 2016). However, a recent investigation suggested that an additional source related to hunter-gatherers in the Caucasus and Neolithic farmers in Iran contributed to the variation present in Southern Europe (Lazaridis *et al.,* 2017).

While genetic studies on ancient Europe are piling up, it is still unclear how the canonical combination of foragers, farmers and pastoralists can fully explain a pretty good dent of the modern European genetic variation.

Thus, in the hope of obtaining a European time transect of ancient migrations and admixture, we analysed the genetic affinity of modern Europeans with two chronologically different sets of ancient samples, *ultimate* and *proximate* sources.

In both analyses, we discovered that Europe and, particularly, Italy have been deeply shaped by different patterns of ancient admixtures. If we focus on the Italian population we observe that while the North is affected more by EHG-related populations, the South show an higher contribution from CHG-related people. In line with this result, also SBA and ABA appeared to have different distribution patterns in Italy and reflect those present in Europe: more common in North Italy and across the continent the former, more localised in the South of Europe and Italy the latter. Similar results were obtained when other Southern European ancient sources replaced ABA in the *proximate* analysis.

This is an interesting result because it adds another piece to the puzzle of our European past. As a matter of fact, SBA is the well-characterised Steppe signature associated with nomadic groups from the Pontic-Caspian steppes, known as one of the main three genetic contributors to modern-day Europeans. In contrast, the other ancient source (ABA, Anatolia Bronze Age) is still uncharacterised in terms of precise dates and origin, however we know that it is ultimately associated with CHG ancestry, as previously suggested (Sarno *et al.,* 2017), and that, surprisingly, affected predominantly

Southern Italy where it represents a substantial component of the ancestry profile of local modern populations.

The analyses of ancient Italian samples help in placing geographical and chronological boundaries for the appearance of these various ancestries in Italy (Figure 94).

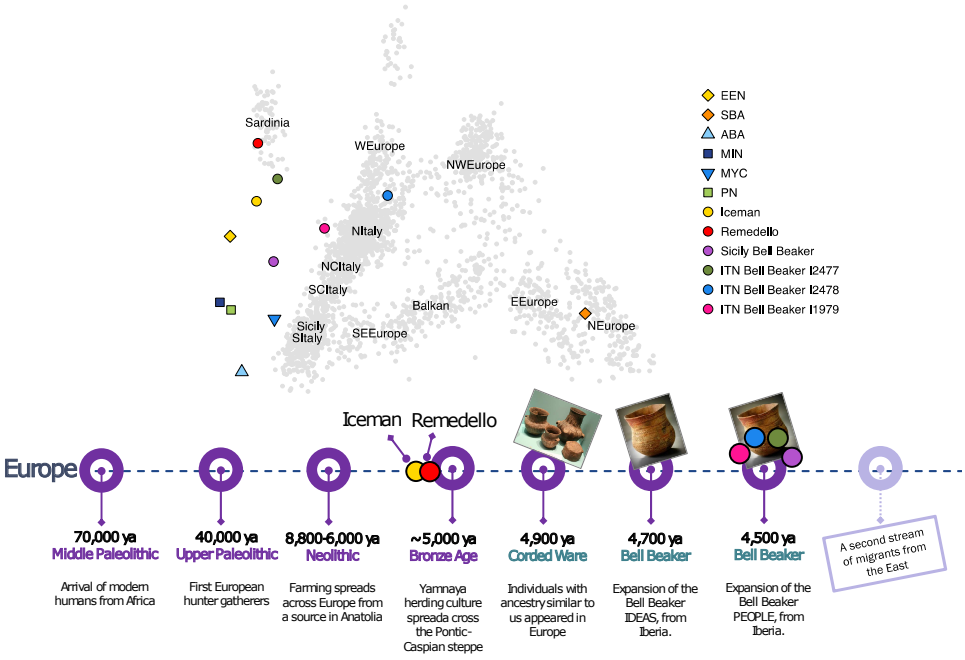


Figure 94. PCA of modern European individuals, ancient Italian and other selected ancient samples. Same PCA as in Figure 69. The timeline at the bottom of the figure reports the archaeological and cultural assemblage of ancient Italian samples.

A plausible model for the peopling of this region suggests an initial admixture between farmers and indigenous hunter-gatherers (Lazaridis *et al.*, 2014; Allentoft *et al.*, 2015; Haak *et al.*, 2015; Olalde *et al.*, 2018; Günther & Jakobsson, 2016). This was then followed by at least two, possibly post-Neolithic, contributions from the East, which brought to Italy ancestries related to Iran Neolithic, Caucasus and Eastern hunter-gatherers. Of these, one is the Steppe Bronze Age signal: this component arrived from mainland Europe, entered the peninsula at least in the Bronze Age, as suggested by its presence in Bell Beaker samples from North Italy and continuing up until historical times and affected mainly Northern Italy. The second is the

Anatolian Bronze Age contribution, which, as suggested by our analyses on the Sicilian Bell Beaker sample, was present as early as the Bronze Age in Southern Italy, possibly arriving via the Southern part of the Balkan Peninsula.

The aforementioned genomic time transect of the ancient migration and admixture events coming to Italy, and more generally Southern Europe, is relevant because no major structure has been highlighted so far in pre-Neolithic Italian samples (Jones *et al.,* 2015). An arrival of the CHG-related component in Southern Italy from the Southern part of the Balkan Peninsula, including the Peloponnesus, is compatible with the identification of genetic corridors linking the two regions and the presence of Southern European ancient signatures in Italy. The temporal appearance of CHG signatures in Anatolia and Southern East Europe in the Late Neolithic/Bronze Age suggests its relevance for post-Neolithic contributions (Mathieson *et al.,* 2018).

The very low presence of CHG signatures in Sardinia and in older Italian samples (Remedello and Iceman) but the occurrence in modern-day Southern Italians might be explained by different scenarios, not mutually exclusive: 1) population structure among early foraging groups across Italy, reflecting different affinities to CHG; 2) the presence in Italy of different Neolithic contributions, characterised by different proportion of CHG-related ancestry; 3) the combination of a post-Neolithic, prehistoric CHG-enriched contribution with a previous AN-related Neolithic layer; 4) a substantial historical contribution from Southeastern Europe across the whole of Southern Italy.

In conclusion, our results suggest contributions from ancestries additional to the three "canonical" considered so far in the literature, which are Western hunter-gatherers, Anatolian Neolithic farmers and Bronze Age herders coming from the Steppe region (referred in our analyses as WHG, AN and SBA, respectively). And what's more important, the differential distribution of these ancestries contributed to the unique degree of population structure present within Italy and the differentiation observed between Northern and Southern Italian clusters.

We hope that in a near future additional aDNA samples from around this time in Italy will be available, thus clarifying what ancient scenario might best support the ancestry composition here presented, and ultimately shed light upon the routes followed by our migrant ancestors.

Modern tales: the genetic burden of our travels

"We are all pilgrims who seek Italy."
Goethe, *Italian Journey*

Speaking of Italians, rare variants and clinical genetics.

Till here, our journey through the human genetic variability has taken us along the hidden roads where our ancestors walked, hundreds of thousands of years ago. However, the genomic journey has also the power to make us look forward, directly into our future.

Our DNA contains not only the genomic footprints of ancestral migrations and admixtures but also some information about our future, mainly in the form of genetic predisposition to certain diseases.

Recently, the genomic revolution, led by the technical innovation in DNA sequencing, has opened the floodgates for more and more ambitious studies aiming at better understanding the genetic composition of human populations and how this variability could interact with public health. Undoubtedly, studying the pattern of these differences and their phenotypical effects is of fundamental interest to many scientific fields, such as medical, forensic and anthropological sciences.

Before the sequencing revolution, these fields have usually been explored analysing common genetic variations. However, a great portion of rare and common diseases, as well as recent demographic events, are related to rare mutations. Actually, the vast majority of coding variation, which is predicted to harbour most of the gene-based disease-causing variants, is rare

and thus unknown if large-scale sequencing genomic datasets are not available. Such information turned out to be essential for ultimately enabling a proper genetic personalized medicine, identifying the genetic causes of Mendelian disorders and, through gene-based burden testing approaches, understanding the complex genetic bases of common diseases.

Especially in the case of personalized medicine and clinical genetic diagnosis, comprehensive knowledge of the healthy population the patient originates from is mandatory. But, what does it mean a "knowledge of the healthy population"?

Here we return to one of the central themes of this thesis: our past. We have extensively said that each human population is the "genetic" result of a few factors: natural selection, genetic drift and, of course, migration and admixture. For this reason, comprehensive knowledge of the "genetic" past of the test population can make a difference for many fields of study, e.g., genetic diagnosis. For instance, a variant may be relatively more common in a specific subpopulation, but below the frequency threshold used to define potential candidate pathogenic variants from the general population database. This is particularly important for countries with a high degree of genetic structure: in these contexts, the possibility to consult geographically stratified databases could be of help in understanding the functional role of genetic variations.

Bearing these concepts in mind, I will report here the case of a country where thousands and thousands of migrations, movements and admixture have contributed in the formation of a relatively high degree of genetic structure, with respect to other countries. Besides being a land rich in history, it is also the place where my most recent ancestors walked and crossed their paths: Italy.

The lure of Italy

Italy, with its position in the middle of the Mediterranean Sea and its complex history, makes a rich study subject for geneticists.

Running through the previous chapters about human migrations, we should see that Italy had always been present in our ancestors' stories. In fact, since Upper Palaeolithic, modern humans have inhabited there. We could recall the archaeological sites of Grotta di Fumane, at the foot of the Venetian Pre-Alps in the western Lessini Mountains, where both Neanderthal and modern human remains had been found (Benazzi *et al.,* 2014), or Grotta del Cavallo, in Southern Italy, where archaeologists found some mod-

ern human teeth dating back to 45,000-43,000 years (Benazzi *et al.,* 2011). Thus, both Neanderthals and the early hunter-gatherers had discovered, thousands of years ago, the beauties of Italy! The lure of Italy became even more pressing during the peak of the Ice Age — between 18,000 and 20,000 years ago — when Italy and the other Southern Mediterranean lands were used as refugia from the frozen north.

During the Neolithic period, the Italian peninsula played a major role in spreading the farming lifestyle. At least two diffusion routes crossed Italy: one started from Apulia and, following the eastern coast, reached the north, while the other started from East Sicily and travelled up along the Tyrrhenian coast (Pessina & Tiné, 2008). During this period and the thousands of years later, the Mediterranean Sea itself contributed in shortening the distances: acting for millennia as a barrier separating the African and the European continent, turned into a bridge as first Bronze Age seafarers started to travel in open water.

Both Anatolian Neolithic and Bronze Age steppe people left deep footprints in the genome of Italian populations. Since the beginning of the Bronze Age and the first attempts with metal smelting, some centres in northern Italy, such as Remedello, became a reference point for metalworkers (Mallory & Adams, 1997). In the meantime, a complex trade network was growing, connecting the sites where the raw materials were extracted, such as Tuscany, to the production centres (see page 134).

In Italy the post-Neolithic settlements were spread all over the country: the Nuraghe stones in Sardinia, the Polada culture in Southern Piedmont and the Camuni, from whom the Valcamonica takes its name, the Terramare and the following Villanovan Culture, until we get to Etruscans and the long period of Roman Empire. A consolidated net of immigrants characterised the Roman domination, thus enriching the Italian — if we can already call it that — gene pool.

However, we have to wait for the Barbarian Invasion (300 CE - 800 CE) and the Arabic rule (827 CE - 1091 CE) to see a significant impact on the genetic variability of modern-day Italians.

Fascinating genes

Italy, due to his fascinating history, raised almost immediately the curiosity of the first scientists studying the genetic variability of human populations. In fact, due to its central position, the Italian genes were thought to resume the main demographic events of, at least, Southern European people. It

took almost fifty years of studies to show the truth of this hypothesis.

The first studies were performed by the group of Alberto Piazza, a collaborator of Cavalli-Sforza. He used the same statistical techniques of Cavalli-Sforza — the synthetic images realised by principal component analyses (see pages 20 and 150) — to represent the genetic structure of Italy, the country "whose unity of people and cultures was quite a recent event" (Piazza *et al.,* 1988). In the study, the analyses of only 34 alleles, among which the AB0 blood groups, was sufficient to observe some peculiar patterns for the Italian population (Figure 95).

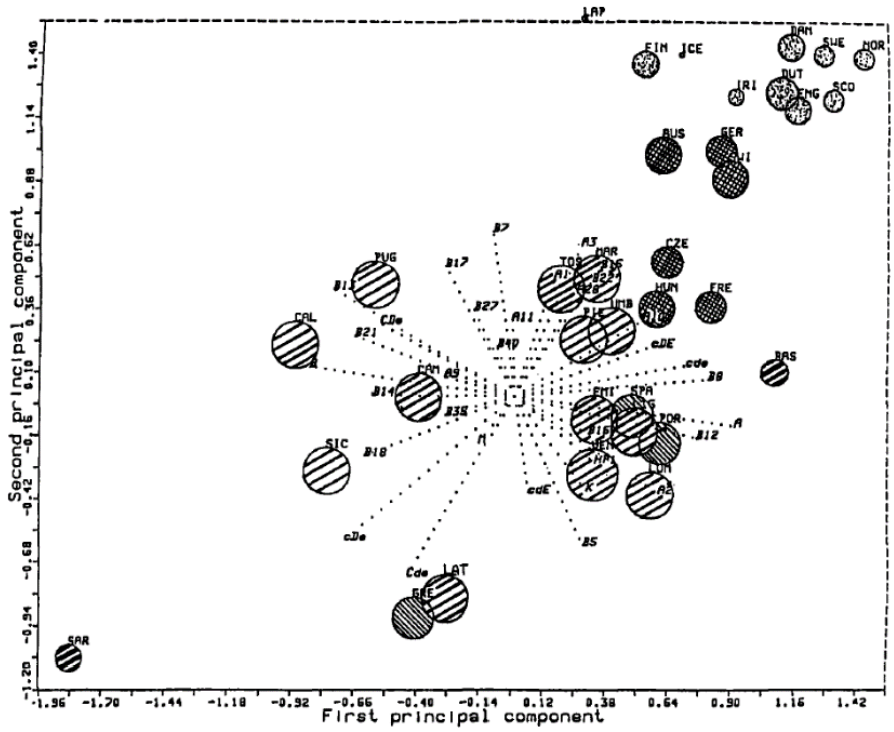


Figure 95. PCA of Italian and European samples. Image taken from Piazza *et al.,* 1988.

In this plot, Sardinia is located far away from the other Italian and European people, thus representing a genetic outlier among them. However, when they excluded Sardinia, the rest of Italy appeared stretched by a north-south genetic gradient: samples from Northern Italy were closer to Central European individuals, while samples from Southern Italy to some

other Mediterranean populations, as Greeks.

In that work, Piazza and colleagues tried to interpret the synthetic maps in the light of historical events, thanks also to the linguistic records. These were the first attempts to integrate different kinds of information to reconstruct the demographic events of the past. Then, the authors concluded that "the genetic structure of Italy still reflects the ethnic stratification of pre-Roman times".

Also the first studies using uniparental markers reached the same conclusion (Barbujani *et al.,* 1995; Capelli *et al.,* 2007; Destro Bisol *et al.,* 2008; Giacomo *et al.,* 2003; Boattini *et al.,* 2013; Brisighelli *et al.,* 2012b; Brisighelli *et al.,* 2012a).

During the last years, several studies investigating the Italian genetic structure also using autosomes are piling up (Fiorito *et al.,* 2016; Sarno *et al.,* 2017; Sazzini *et al.,* 2016; Parolo *et al.,* 2015).

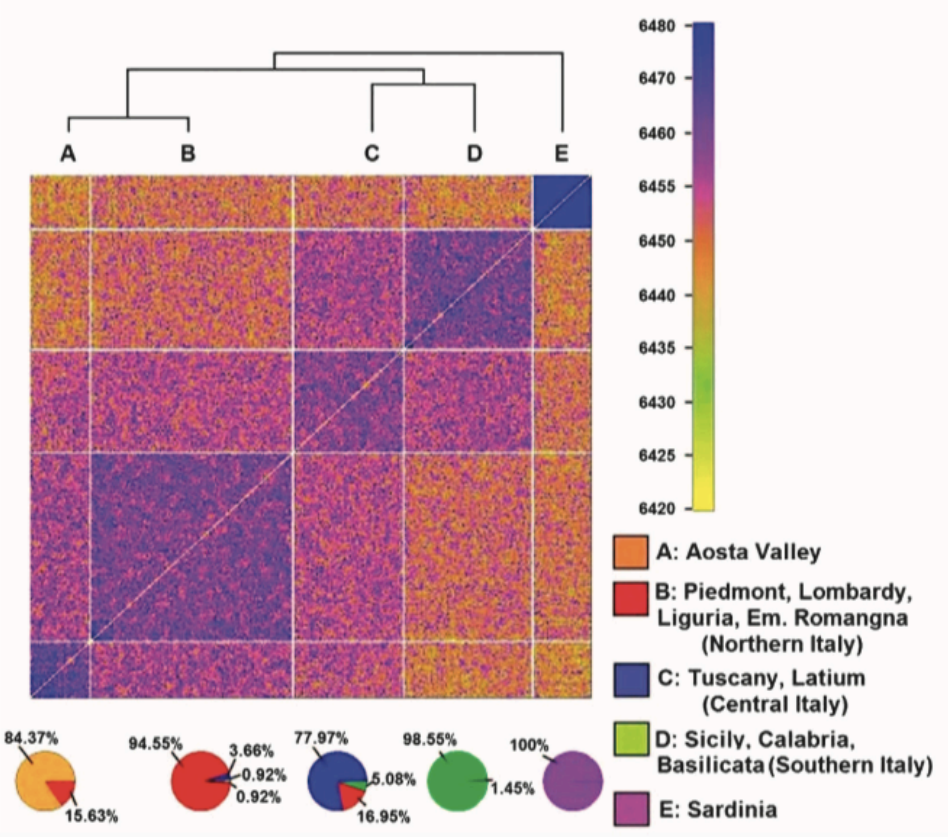


Figure 96. Five major genetic groups in Italy. Heatmap representing the coancestry matrix indicating the number of genomic segments inherited from the same ancestral populations for each pair of samples: dendrogram based on hierarchical clustering (at the top), and pie charts representing the overlap between inferred and self-reported origin of the Italian individuals. Image taken from Fiorito *et al.,* 2016.

These works gave a great contribution to the knowledge of demographic processes which had shaped the Italian genetic structure; however, none of them had samples coming from all the Italian administrative regions. Both Parolo *et al.,* 2015 and Sazzini *et al.,* 2016 aimed at explaining the north-south genetic differences detected using global method, in the light of local adaptation and the different susceptibility to complex diseases. Conversely, Fiorito *et al.,* 2016 and Sarno *et al.,* 2017 analysed the genetic structure of the country by applying local ancestry pipeline. In particular, they used CHROMOPAINTER and fineSTRUCTURE to perform a fine genetic clus-

tering of their samples, making some interesting discoveries. Fiorito *et al.,* 2016 found that the Italian peninsula could be divided into five main genetic clusters (Figure 96), while Sarno *et al.,* 2017 detected two Post-Neolithic ancestries related to Caucasian and Levantin populations in Southern Italians.

However, both works had some limitations in uncovering the main demographic events experienced by the entire Italian population: in the case of Fiorito *et al.,* 2016, they did not cover some of the Italian regions, while in the case of Sarno *et al.,* 2017, they focused only on the Southern part of Italy.

The first part of the work by (Raveane *et al.,* 2019) tried to overcome some of these limitations. Dr Alessandro Raveane, supervised by Dr Francesco Montinaro and Dr Cristian Capelli, made a great work collecting genome-wide data from a large number of Italian individuals. He assembled and analysed two different datasets: LDD (Low Density Dataset, comprising more than 1500 Italians genotyped for around 220,000 SNPs) and HDD (High Density Dataset, with 700 Italians and around 600,000 SNPs). Thanks to the fact that the samples came from all the Italian administrative regions, he was able to perform a finer dissection of the population genetic structure with respect to previous works.

He applied the CHROMOPAINTER/fineSTRUCTURE pipeline and discovered that Italian clusters separated into three main groups (Figure 97A-C): Sardinia, Northern (North / Central-North Italy) and Southern Italy (South / Central-South Italy and Sicily). When compared to other worldwide populations, he found that while the first two were close to populations coming from Western Europe, the last was closer to Middle Eastern groups.

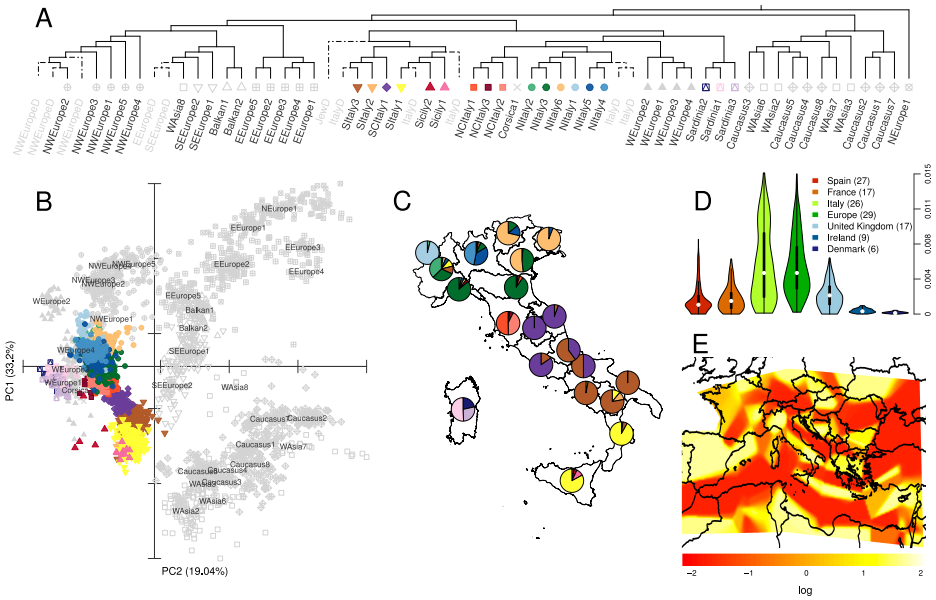


Figure 97. Genetic structure of the Italian populations. (A) Simplified dendrogram of 3,057 Eurasian samples clustered by the fS algorithm. **(B)** PCA based on CP chunkcount matrix. **(C)** Pie charts summarising the relative proportions of inferred fS genetic clusters for all the 20 Italian administrative regions. **(D)** Between-clusters F_{ST} estimates within European groups. **(E)** Estimated Effective Migration Surfaces (EEMS) analysis in Southern Europe; colours represent the $\log_{10}$ scale of the effective migration rate, from low (red) to high (yellow).

One of the most interesting results was related to the Italian genetic structure itself: the differences in ancestry composition between Northern and Southern Italy (see chapter *"Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe"*) have generated the largest degree of population structure detected so far in the continent. As a matter of fact, he compared the degree of variation among genetic clusters in Italy with those in several European countries and across the whole of Europe (Figure 97D). By doing so, he discovered that the *among-clusters* variability in Italy was significantly higher than in any other country examined but comparable with the estimates across European clusters. For some insights regarding the clustering methods, see page 145. The detailed results of the analyses performed by Dr Raveane can be found in Raveane *et al.*, 2019.

Common vs rare variations

The genetic structure of the Italian population has been extensively investigated, as reported above, using SNP array data. However, these platforms can genotype only those variants which have been previously detected in other individuals, thus being unable to discover new genetic variation (see page 18). For the same reason, SNP arrays are mainly comprised of genome-wide common markers — which, usually, refer to variants with a minor allele frequency (MAF) >1%.

This is not a problem at all for population genetics. In fact, many of the approaches previously described rely on common and neutral genetic variations, often requiring the filter of rare or monomorphic variants (e.g., PCA and *f*-statistics). Actually, all the results presented in this thesis, so far, have been produced using mainly common SNPs.

However, also rare genetic variations can tell their own stories. Due to the simple fact that they are rare in the general population, many of these variants are thought to be more recent than common variants. For this reason, they could provide new insight into more recent demographic events. Moreover, they are typically geographically localised (Tennessen *et al.,* 2012) and could contribute to a finer scale geographic structure.

In the last decade, these variants have acquired more and more relevance in genetic research, mainly because, the technology that truly can detect them — sequencing techniques and their evolutions — has become ever more diffuse only in recent times. In fact, sequencing does not rely on previous knowledge of genetic variants but can detect all variants in a genomic region. Moreover, as rare genetic variants are often population-specific or even individual-specific (singleton), it couldn't be possible to build a new SNP-array platform once new rare variants are discovered. Conversely, sequencing the entire genome across a large number of samples — the larger the sample size, the more recent the epoch it probes (Keinan & Clark, 2012) — is the best strategy to go into a more recent past. Of course, there are some technical challenges, such as the difficulty in distinguishing singleton from sequencing errors, especially in low-coverage data. However, due to the high interest in the method, new tools able to overcome these challenges are continuously introduced.

As the first sequencing projects on human populations terminated (Tennessen *et al.,* 2012; Auton *et al.,* 2015), the allele frequency spectrum (AFS) from different populations became rapidly skewed toward rare genetic variations.

However, where do these variants came from? The high load of rare vari-

AFS

The allele frequency spectrum is the distribution of the allele frequencies of a given set of loci (often SNPs) in a population or sample. *Wikipedia*

ation, due to recent mutations, in human populations is due to rapid recent population growth (Keinan & Clark, 2012). In fact, the human population has experienced a rapid growth of more than three orders of magnitude within fewer than 400 generations. This growth started around 10,000 years ago — the Neolithic period (see pages 125-125 for some details about Neolithic in Western Eurasia) — when the new lifestyle of farmers brought more abundant and regular sources of food; this was followed by another accelerated growth starting around 100 generations ago (Figure 98, Keinan & Clark, 2012).

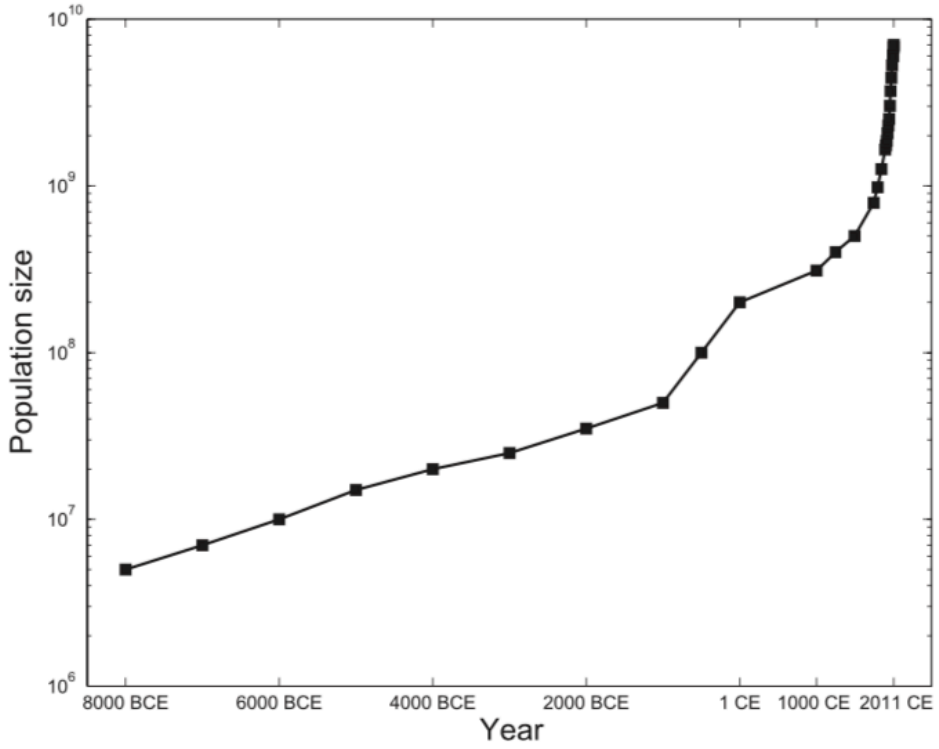


Figure 98. Census (rather than effective) population size over the past 10,000 years (until 2011). The population size is presented on a logarithm scale. Image taken from Keinan & Clark, 2012.

However, why rapid human population growth is of interest for genetic researchers? Because the high load of rare variants derived from it may play an important role in disease risk Keinan & Clark, 2012. We know that purifying selection makes sure that deleterious mutations cannot reach

high frequencies in the population and, in many cases, it even removes them. From this, it follows that disease-promoting variants should not be common. Actually, the skewness towards rare variations is more pronounced for non-synonymous mutations — variants causing an amino acid substitution — thus reflecting once again, the action of purifying selection (Gibson, 2012).

From these considerations, it emerges that rare variations, especially those harboured in coding regions, could be crucial for understanding the genetic bases of disease risk and, given their usually localised distribution, also the differential disease susceptibility among different populations. For this reason, national and international projects aiming at exploring the genetic frequency spectrum across a large number of individuals through sequencing could be of great help.

To date, in addition to the well-known studies exploring human genetic variation worldwide from whole-exome data (Auton *et al.,* 2015; Tennessen *et al.,* 2012; Lek *et al.,* 2016; Karczewski *et al.,* 2019), many groups (Dopazo *et al.,* 2016; Kwak *et al.,* 2017; Van Hout *et al.,* 2019) worked on national sequencing projects with three main goals in mind:

- studying the genetic structure of a population by exploiting also lower frequency variants,
- understanding the distribution of putative pathogenic variation in healthy cohorts,
- generating a catalogue of local variability.

For what concerns Italy, previous sequencing-based studies on its population include the 1000 Genomes Project (with 107 Tuscans) and more recent studies (Xue *et al.,* 2017; Nutile *et al.,* 2019) that focused on specific isolates. These datasets are not well suited to represent and explore the whole Italian genetic variability since that was not the primary goal of those studies. However, in the following section, we will see how a national dataset describing the comprehensive genetic variation within a population could help for medical genetics.

Clinical genetics in a nutshell

NGS technologies have revolutionised genetics, by leading to a throughput several order of magnitude higher than Sanger sequencing. In particular, these methods have rapidly become routine tools in the diagnosis of patients with different genetic conditions and even tumor profiling. In this context,

the main aim is to simultaneously query the diagnostically relevant genes of a patient in order to formulate a clinical decision and this is exactly what these technologies are great at doing. In fact, they can be used to sequence a specific panel of genes (targeted sequencing, TS), only the coding exons (whole-exome sequencing, WES) or the entire genome of an individual (whole-genome sequencing, WGS).

However, every individual's genome contains millions of variants, but only a few of them — in some cases, one of them — could be linked to the disease. NGS technologies, while allowing the detection of a sheer number of variants, they also raise new challenges in filtering the non-informative genetic variants.

In 2017, the American College of Medical Genetics and Genomics (ACMG) published the guidelines for the interpretation of sequence variants (Kalia *et al.,* 2017), which nowadays have been widely adopted into clinical practice. The process to detect which one, among the millions of genetic variants (in the case of a WGS), is clinically relevant is reported in Figure 99.

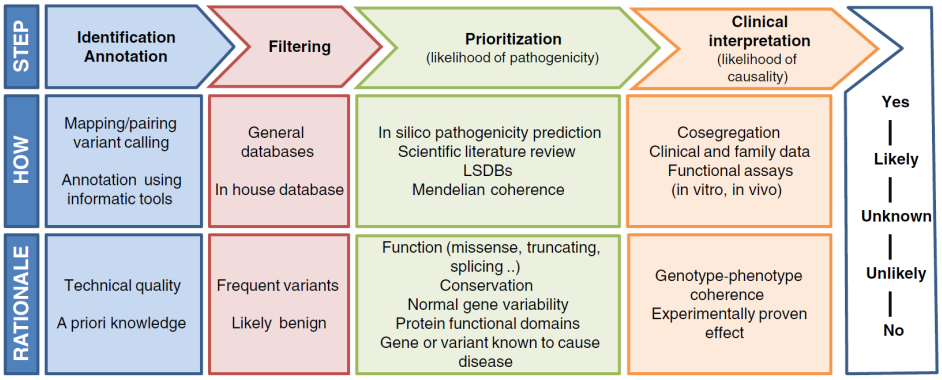


Figure 99. Stepwise evidence pipeline for clinical interpretation genetic variants. Image taken from Quintans *et al.,* 2014.

The first step is the **identification** of genetic variants through variant calling software, which considers parameters such as coverage and quality to determine the presence of a variant. Then, it comes the **annotation**, i.e., the process of describing the position, the nature and the effect of each genetic variant. This step could be performed by several bioinformatics tools, such as ANNOVAR (Wang *et al.,* 2010), VEP (McLaren *et al.,* 2016), SnpEff (Cingolani *et al.,* 2012) and so on. The Sequence Ontology (Eilbeck *et al.,* 2005) provides a standardised terminology for variant annotation,

describing the variant in terms of its "sequence alteration". However, one of the challenges of this step is the existence of different reference sequences, genomic builds and transcripts. The latter is one of the hardest to handle: actually, a missense variant in a transcript could be intronic if annotated on another transcript where that particular exon was not present (Quintans *et al.,* 2014). Since it is usually unknown which transcript is expressed in a particular organ or at a particular stage relevant for the disease, a solution is to provide an annotation for all the possible transcripts. Sometimes, the longest transcript — called, the *canonical* — is used or the most biologically relevant, when it is known.

At this point, we should have millions of variants, their positions, their functional classification but still the urgent need to find the causative ones. In order to ease this situation, a proper **filtering** step can help to reduce the huge number of variants, by excluding the likely benign ones. One of the variant proprieties used for filtering is its frequency, under the assumption that a variant present in random individuals above a certain threshold is likely to be benign (Quintans *et al.,* 2014). For this purpose, recent catalogues of genetic variations in human populations could provide a powerful tool (Table 8). Among the others, the 1000 Genome Project contains genome and exome sequence data from 2,504 individuals belonging to 26 worldwide populations, the Exome Variant Server (EVS) contains frequency information of 6,503 exomes of European and African ancestry, the Exome Aggregation Consortium (ExAC) has 60,706 whole-exome sequences, while the Genome Aggregation Database (gnomAD) includes 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals. These large-scale reference datasets are a valuable resource for clinical geneticists; however, the fact that many worldwide populations are not represented motivates scientists to go on collecting, producing and sharing genetic data for more and more individuals. Another limitation related to allele frequency filters is the population stratification, which can confound the assumption that rare variants are more likely to be damaging than common variants. While all individual genotypes in the 1000 Genomes Project are available, ExAC and gnomAD can only be consulted in an aggregated manner (with some stratification possible among macro-populations), thus making it impossible to access country-specific genetic variation or individual genotypes. This is, of course, a major inconvenience when population structure is present. The example reported in Figure 100 is a good illustration of this problem: a variant that is rare in European ancestry may be more common in African ancestry populations (Eilbeck *et al.,* 2017). This effect could also be present, even if with a lower impact, in populations with a

Population stratification

The difference in allele frequencies across sub-populations.
(Eilbeck et al., *2017)*

high degree of population structure. However, the frequency filter is strictly dependent on the target disease: its penetrance, its time of onset and the clinical consequences for the patients are, together, factors that can influence the rarity of its causative variants. For example, the variant causing a Mendelian disease with complete penetrance, an early-onset and serious disease manifestations could be even absent in the databases mentioned above, which, in the majority of the cases, report healthy individuals. Conversely, complex diseases with incomplete penetrance, a late-onset and clinical conditions that do not alter the individual fitness could be caused or show an increased susceptibility by variants with higher frequencies in the healthy population (even if below the 1%).

Database	Characteristics	URL
Population frequency of variants in whole-exome data		
Exome Aggregation Consortium	- 60,706 unrelated individuals	http://exac.broadinstitute.org/
Exome Variant Server	- 6,500 exomes of European and African American ancestry - Includes healthy individuals as well as those with different diseases	http://evs.gs.washington.edu/EVS/
Population frequency of variants in whole-genome data		
1000 Genomes Project	- Final dataset contains data for 2,504 individuals from 26 populations	https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/
Genome Aggregation Database	- 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals	http://gnomad.broadinstitute.org/
Korean Reference Genome Database	- Whole genome sequencing project for 1,722 Korean individuals	http://152.99.75.168/KRGDB/
Computational and predictive data		
dbNSFP	- Developed for functional prediction and annotation of all potential nonsynonymous single-nucleotide variants	https://sites.google.com/site/jpopgen/dbNSFP (freely downloadable)
	- Compiles prediction scores from 20 prediction algorithms (SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster2, MutationAssessor, FATHMM, MetaSVM, MetaLR, CADD, VEST3, PROVEAN, FATHMM-MKL coding, fitCons, DANN, Geno, Canyon, Eigen coding, Eigen-PC, M-CAP, REVEL, MutPred), six conservation scores (PhyloP×2, phastCons×2, GERP++, and SiPhy), and other related information	
dbscSNV	- Splice site prediction that scores the likelihood that the variant affects splicing	https://sites.google.com/site/jpopgen/dbNSFP (freely downloadable)
	- Includes all potential human single-nucleotide variants within splicing consensus regions (-3 to +8 at the 5′ splice site and -12 to +2 at the 3′ splice site)	
Variant Effect Predictor	- Prediction toolset using a population database and prediction algorithms	https://www.ensembl.org/info/docs/tools/vep/index.html
Variant type and gene-specific information data		
ClinVar	- Freely accessible, public archive of reports of relationships among human variations and phenotypes, with supporting evidence	https://www.ncbi.nlm.nih.gov/clinvar/
Human Gene Mutation Database	- Provides comprehensive annotation for all published inherited disease mutations	http://www.hgmd.org
ClinGen	- Gene-disease validity, gene dosage sensitivity	http://www.clinicalgenome.org

Table 8. Useful databases for variant classification. Image taken from Kim *et al.,* 2019.

After filtering likely benign variants by relying on frequencies in the healthy population, there is the **prioritization**, which refers to the "process of ranking the variants observed in an individual genome on the basis of

factors such as the predicted consequence of each variant and the observed frequency in a population" (Eilbeck *et al.,* 2017). The simplest approach to do that is considering the Sequence Ontology terms recovered during the annotation step: for example, a variant annotated as *"frameshift_variant"* or *"stop_gained"* can obviously cause the truncation of the resulting protein (protein-truncating variant, PTV), thus being relevant in the clinical interpretation (see page 221, for an example of the annotation step and the classification of PTVs). However, using only this criterion is not so effective, since also healthy individuals harbour some PTVs: a recent survey on the UK Biobank individuals found a median PTV number per individual of 15 (in the next chapter *"Protein-coding genetic variation within the Italian peninsula: genetic structure and functional implications"*, we computed the mean and median numbers of PTVs in the Italian population). Another limitation is that the impact of a mutation also depends on the gene function, for example, the truncation of a protein in a poorly conserved genes or in a pseudogene may be more tolerated than a missense variant in a highly conserved gene (see page 247 for the example of a PTV in a olfactory receptor).

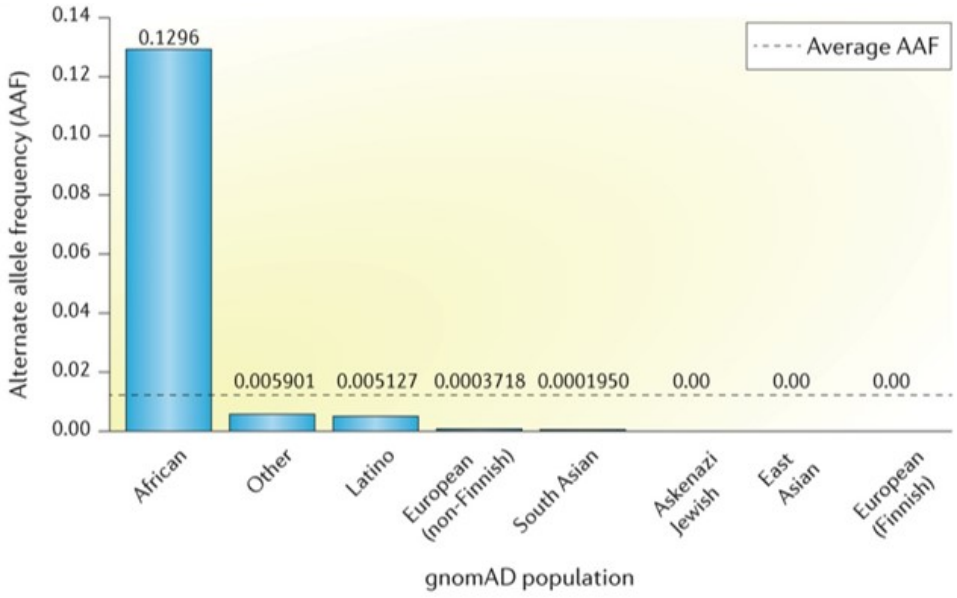


Figure 100. Population stratification as a problem in variant interpretation. Image taken from Eilbeck *et al.,* 2017.

For this reason, is always advisable to go further in the prioritization step by choosing some tools in the plethora of bioinformatic solutions and databases for variant interpretation. For example, using literature information (ncbi.nlm.nih.gov/pubmed) and databases such as OMIM (omim.org, Amberger *et al.,* 2015), ClinVar (ncbi.nlm.nih.gov/clinvar, Landrum *et al.,* 2016) and HGMD (portal.biobase-international.com, Stenson *et al.,* 2003), it is possible to detect genes and variants which are specifically relevant for the target disease. Then, genetic variants could be prioritized based, as above, on their population frequency (Table 8), the phylogenetic conservation across different species — e.g., PhastCons (Siepel *et al.,* 2005), GERP++ (Davydov *et al.,* 2010) and PhyloP (Siepel *et al.,* 2005) — or their effect on protein function or structure, such as SIFT (Kumar *et al.,* 2009), PolyPhen2 (Adzhubei *et al.,* 2013), MutationTaster2 (Schwarz *et al.,* 2014). A list of the commonly used software for variant interpretation (updated to 2017) is reported in Table 9.

Tool	Category	Coding (missense only)	Indel	Non-coding	Method summary
SIFT	Missense prediction	Y (Y)	N	N	The degree of protein sequence conservation is used to predict the impact of a missense variant
PolyPhen2	Missense prediction	Y (Y)	N	N	Uses protein sequence and structure to predict the impact of a missense variant
FATHMM	Missense prediction	Y (Y)	N	N	Uses protein sequence homology identified with HMMER3 to predict the impact of a missense variant
PROVEAN	Missense and indel prediction	Y (N)	Y	N	The degree of protein sequence conservation is used to predict the impact of an amino acid change or an indel
REVEL	Missense prediction (ensemble method)	Y (Y)	N	N	Incorporates 18 individual scores from 13 different tools to produce an ensemble 'pathogenicity' score for missense variants
PhastCons	Sequence conservation	Y (N)	Y	Y	Uses multiple sequence alignments from diverse species to identify conserved elements
PhyloP	Sequence conservation	Y (N)	Y	Y	Uses multiple sequence alignments from diverse species to assign per-base *P* values of conservation
GERP++	Sequence conservation	Y (N)	Y	Y	Measures sequence conservation in the human genome through alignments to 43 other vertebrate genomes
MutationTaster2	Multi-data integration	Y (N)	Y	Y (intronic)	Integrates sequence conservation, as well as data from the 1000 Genomes Project, ENCODE and ClinVar, to predict the consequence of variants within a gene model
VAAST	Multi-data integration	Y (N)	Y	Y	Integrates variant frequency data with phylogenetic conservation for variant prioritization and burden testing
CADD	Multi-data integration	Y (N)	Y (short indels)	Y	Integration of conservation metrics, functional data (for example, DNase I hypersensitivity and transcription factor binding) and scores such as SIFT and PolyPhen2 to predict the deleteriousness of nucleotide or short indel change in the genome
FitCons	Multi-data integration	Y (N)	Y (short indels)	Y	Integrates functional genomic data from the ENCODE project to cluster genomic regions and to predict the probability of a fitness consequence based on sequence conservation and the degree of regional polymorphism in the human genome

CADD, Combined Annotation-Dependent Depletion; ENCODE, Encyclopedia of DNA Elements; FATHMM, Functional Analysis Through Hidden Markov Models; FitCons, fitness consequence; GERP, Genomic Evolutionary Rate Profiling; HMMER3, a tool based on a hidden Markov Model (HMM) for searching sequence databases for homologues of protein or DNA sequences; indel, small insertion or deletion; PolyPhen2, polymorphism phenotyping version 2; PROVEAN, Protein Variation Effect Analyzer; REVEL, rare exome variant ensemble learner; SIFT, Sorts Intolerant From Tolerant; VAAST, Variant Annotation, Analysis and Search Tool.

Table 9. Commonly used software for assessing variant impact. Table taken from Eilbeck *et al.,* 2017.

The last step is the **clinical interpretation**. Here, all the lines of evidence collected through the methods mentioned above are put together with some patient-specific information, such as clinical and family data, in order to analyse the segregation pattern, or with functional assays, thus directly testing the effect of that variant.

I will not go into the details of clinical interpretation of variants. In addiction to being a very complex matter, it is not directly related to the main topic of my work. However, it is interesting how the demographic and biological events a population experiences in its past could influence many other, apparently distant, topics.

In the next chapter, I will describe a study, Dr Giovanni Birolo and I worked on, which puts together the three main themes of this introduction: the Italian population, the rare variants hidden in its genomes and the burden of more or less pathogenic variants Italians are carrying around.

"The problem with the gene pool is that there's no lifeguard."

David Gerrold

Protein-coding genetic variation within the Italian peninsula: genetic structure and functional implications

We have already seen how the Italian genes could be so fascinating for geneticists. However, the literature on the Italian population is far from being exhaustive, because most of previous works on the Italian population are based on SNP array data, while those employing a sequencing approach focused mainly on isolated populations.

For this reason, we collected a large dataset of Italian whole-exome sequences, and to the best of our knowledge, this is the first study to investigate the general Italian population by whole-exome sequencing on a such large sample with well-defined ancestry. In order to reconstruct the Italian genetic structure and its implication for phenotypes and diseases, we analyzed 1,686 healthy Italian individuals and 669,718 variants, exploring both common and rare variation.

We show how the Italian genetic structure is apparent also from exome sequencing data and it is consistent with the one inferred from SNP-array based studies. We find strong frequency differences in coding variants between Northern and Southern Italy, some of which have not been reported before. By a novel analysis, we take advantage of rare variation to stratify Italy by effective population size, observing distinctive differences even at

the regional level.

Moving to a more clinical perspective, we use our dataset to quantify how the availability of population-specific databases of allele frequencies can increase accuracy in the assessment of pathogenic variants. In this context, due to the distinct genetic identity of Italy, we show how a large national database composed by individuals with well-defined ancestry improve the identification of potentially pathogenic variants. Then, we explore the burden of putatively pathogenic variation in the healthy Italian exome, showing that almost 3% of the population carries protein-truncating variants in the ACMG medically actionable genes.

Believing in the utility of large population-specific databases, especially for a complex population like the Italian one, the natural conclusion of the work is that we make the aggregated variant frequencies from our dataset publicly available as a tool for both clinicians and researchers.

Methods

Sample study

We obtained variant calls from whole exome sequencing (WES) data of 1,751 healthy individuals enrolled in the Italian Genetic Study on early-onset myocardial infarction (ATAVBIS, 2003) and as part of the "Myocardial Infarction Genetics Consortium" (Migen, Do *et al.,* 2015). All participants in the study provided written informed consent for genetic studies. The institutional review boards at the Broad Institute and each participating institution approved the study protocol. The ancestry information comprises the place of birth of an individual, of their parents and grandparents: 1,235 individuals had complete information, 174 lacked the birthplace of their grandparents, 339 lacked both parents and grandparents and 3 had no birthplace themselves.

Sequencing

The cluster amplification, sequencing, read-mapping and variant calling were performed by the Broad Institute, as described in (Do *et al.,* 2015). Samples were kept when the read depth was 20X or greater on at least 80% of the exome target.

Data cleaning

For our analysis, we needed a dataset of high-quality genotype calls of unrelated individuals with reasonably certain Italian ancestry. In order to produce this dataset, we removed individuals with unclear ancestry and low-quality variants and genotypes from the original multisample VCF, comprising 1,373,696 variants and 1,751 individuals.

As first-pass cleaning step on the individuals, we removed 26 individuals with reportedly non-Italian ancestry (even partial), leaving 1,725 individuals. Then, individual sex from the database was cross-checked with sex inferred from variant calls (with *bcftools +guess-ploidy*, v1.5) and individuals with discordant sex were removed (since they were likely misreported in the database), leaving 1,717 individuals.

For what concerns genotypes, low quality genotype calls were set to missing, specifically calls with low read depth (DP < 10), calls with very high read depth (DP > 180, which is three standard deviations more than the mean depth of 60) and calls with low confidence (GQ < 20). Genotype calls in non-autosomal regions in males were converted to hemizygous.

Then, we applied a second low quality variant soft filtering in which variants not satisfying our criteria were marked (but not removed). Approximately 93% of variants (1,187,119) passed the filters. The criteria for exclusion were: missing genotypes for more than 10% of individuals (88,492 variants); extreme deviations (p-value smaller than 10-10) from Hardy–Weinberg equilibrium (2,811 variants); location in low complexity regions of the genome as described in Li & Kahveci, 2006 (4,613 variants).

A second-pass filtering step on individuals was applied using the pass variants: none of the individuals had more than 5% of missing genotypes, but related samples (up to 2nd degree) were detected from genotypes. In this case, only one relative was kept, leaving 1,688 individuals. The last filtering step on individuals was the outlier removal according to PCA. PCA was performed with PLINK 1.9 (Purcell *et al.,* 2007) using pass variants with major allele frequency at least 0.2% and pruning by LD at 0.2 r^2. Outliers defined as samples with Euclidean distance (computed from the first two PCs) greater than 10 standard deviations from the mean position of all samples were removed iteratively, re-computing the PCA until no more outliers were detected. Two samples were removed in one iteration as genetic outliers.

Finally, variants where all individuals were homozygous for the reference allele were removed, leaving a dataset of 669,718 variants for 1,686 unrelated individuals of Italian ancestry. This dataset was used in all analyses looking at the Italian population as a whole.

In order to perform comparisons within Italy, we selected individuals with a well-defined ancestry at the macro-area level, removing individuals with uncertain or likely misreported ancestry. From the previous dataset of 1,686 individuals, we selected the 1,197 individuals who had information about all four grandparents' birthplace available. We assigned them to a macro-area if all their grandparents were born there. We performed further iterative PCA-based outlier removal to exclude individuals who did not cluster within the others from their macro-area. We dropped samples with more than 3.5 standard deviations from the center of their cluster, removing 43 individuals in 7 iterations, leaving 1,154 individuals. This dataset was used in all analyses comparing different macro-areas. When comparing regions, we furthermore selected those individuals whose grandparents were born in the same region.

Variant annotation and interpretation

Functional annotation was performed with SnpEff (Cingolani *et al.,* 2012) with respect to the canonical RefSeq transcript, except for ACMG SF v2.0 actionable genes, where we used the transcript that occurred most frequently in the ClinVar annotations of pathogenic variants. Variants were labeled as protein-truncating variants (PTVs) when their allele frequency in the whole dataset was under 5% and their reported effect in the SnpEff annotation was one of frameshift_variant, splice_acceptor_variant, splice_donor_variant or stop_gained. Missense variants were evaluated with seven pathogenicity predictors: MutPred, VEST 3, REVEL, fathmm from dbNSFP with rankscore at least 0.73, M-CAP and MetaSVM from dbNSFP and CADD with score at least 25. Missense variants were considered damaging (DMG) when at least five predictors supported this conclusion and their allele frequency in the whole dataset was under 5%. Variants which were annotated "pathogenic" or "likely pathogenic" without any other conflicting annotation and whose allele frequency in the whole dataset was under 5%, were labeled CLNPAT variants. To highlight variants with a putative role in pharmacogenetic we annotated them with a specialized public repository, e.g., PharmGKB (Whirl-Carrillo *et al.,* 2012).

We calculated the cumulative number of PTV and DMG variations with allele frequency lower than 1% across the 20,445 genes and the 293 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. Then, in order to normalize these counts, we divided them for the total number of variants (with a frequency lower than 1%) in each gene and each KEGG category.

Exploring the genetic structure with coding variants

PCAs on the whole dataset and on macro-areas and using only individuals with a "well-defined" macro-area were performed with PLINK 1.9 (Purcell *et al.,* 2007) with major allele frequency at least 0.2% and pruning by LD at 0.2 r^2. On the same dataset, we inferred pairwise F_{ST} estimates among macro-areas and among Italian administrative regions using the *smartpca* software implemented in the EINGESOFT package (Patterson *et al.,* 2006), which computes the Hudson's F_{ST} estimator.

In order to investigate demographic events from rare variations in Italian macro-areas, we computed the allele frequency spectrum in each Italian macro-area and administrative region and we tested for differences. In order to avoid bias caused by the different sizes of our regional subpopulations, we performed random sub-sampling without replacement on the individu-

als, producing one thousand subsamples of ten unrelated individuals for each subpopulation with at least ten samples. Allele counts were computed separately in each subsample, for the variants that were observed in that subsample, thus producing counts ranging from one to ten (since we counted the minor allele for subsamples of ten individuals). This process yielded one thousand estimates of the allele frequency spectrum (with ten frequency bins) for each subpopulation. Each frequency bin was analyzed separately, using the one thousand subsamples to estimate the distribution of values for each subpopulation. This method has the nice property of producing estimates whose median is independent of the size of the subpopulation. However, the smaller subpopulations showed a reduced variance, since the subsamples have high overlap and are thus not independent enough. Thus, the allele frequency spectrum was computed independently for each subsample of a subpopulation. We then compared the distribution of these uniformly sized subsamples for each allele count using a Wilcoxon rank sum test and Bonferroni's correction in R programming language environment. Note that this is not a bootstrap: we are subsampling individuals and not variants (as one would normally do to estimate a distribution of variants) and we are subsampling without replacement, since sampling individuals more than once would entail having related individuals in the samples, which would in turn produce completely skewed allele counts. In particular, sampling individuals with replacement would cause very rare variants occurring only in one individual to be counted more than once. This would happen more when subsampling from smaller subpopulations, producing a very strong bias where rare variants were shifted towards higher frequency bins, completely skewing the allele frequency spectrum distribution and making comparisons impossible.

Genetic comparison within Italy

Differences of allele frequency between macro-areas were tested with Fisher's Exact Test performed by PLINK 1.9. We only tested Northern versus Southern Italy (622 and 305 individuals, respectively): we excluded both Sardinia and Central Italy because of their reduced sample size of 20 and 76 individuals respectively and, for Central Italy, also because of its intermediate position in the North-South cline. We only tested variants with allele frequency greater than 1% in the whole dataset since with our sample sizes we did not have power to test lower frequency variants. We considered significant all variants passing the 0.01 significance threshold after Bonferroni's multiple test correction. Then, we computed single-locus F_{st} estimates to

confirm the genetics signals retrieved with the Fisher's Exact Test.

We considered the top 1% variants with the lowest p-values in the following enrichment analyses. Using the online tool EnrichR (Chen *et al.,* 2013; Kuleshov *et al.,* 2016) we searched for significant enrichment by looking at EnrichR databases, e.g. dbGaP (Mailman *et al.,* 2007; Tryka *et al.,* 2014), GWAS catalog 2019 (Buniello *et al.,* 2019) and "Jensen disease" category. Then, we searched for known associations between the top 1% list of variants and phenotypes, by directly looking in the GWAS catalog (downloaded in March 2019).

Genetic comparison between Italy and Europe

We tested for allele frequency differences between the Italian dataset and non-Finnish Europeans from gnomAD, using Fisher's Exact Test. Here, due to the increase sample size, we considered all genetic variants and we considered significant all genetic variants at p-value < 0.01 (Bonferroni corrected). As above, we searched for significant enrichment using EnrichR and checked for known associations in the GWAS catalog and in ClinVar.

Results

We analyzed our whole-exome sequencing dataset of healthy Italians with the dual goal of exploring the genetic patterns of the Italian population and their importance in clinical genetics. After quality control procedures (ee *Methods - Data cleaning* on page 219), the dataset contained 1,686 unrelated individuals, with 669,718 observed variants. As expected for a sequencing dataset with this sample size, most variants are low-frequency and more than 60% are singletons, i.e. only observed in a single sample. Functionally, most variants are missense, followed by intronic and synonymous variants, while protein-truncating variants accounted for 2% of the total variants (Table 10).

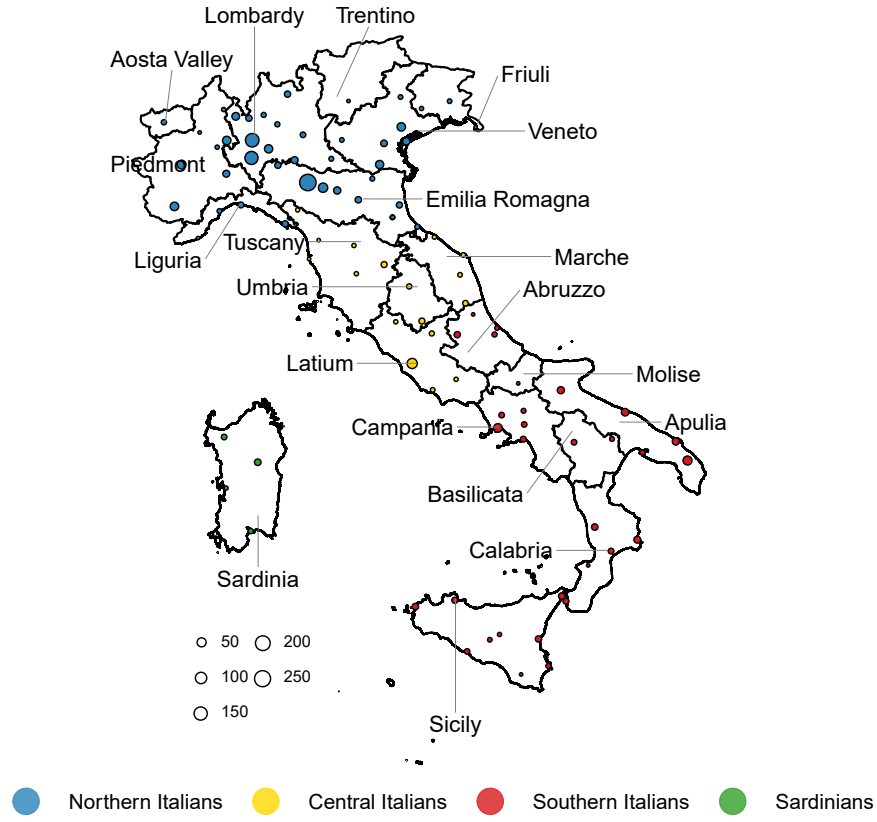


Figure 101. Italian samples per macro-area and province. The province of origin of the individuals is reported. The size of points denotes the number of individuals from that province.

EFFECT	VARIANT COUNT	VARIANT PER INDIVIDUAL	SINGLETON COUNT	SINGLETON PER INDIVIDUAL
missense	237,336	6,344	154,729	91.8
intron variant	197,857	10,962	112,704	66.8
synonymous	145,639	7,495	82,881	49.2
downstream gene variant	17,949	893	10,275	6.1
upstream gene variant	16,868	812	9,779	5.8
other	13,817	544	8,383	5
3' UTR variant	13,539	713	7,785	4.6
5' UTR variant	7,833	370	4,561	2.7
frameshift indel	5,756	58	4,624	2.7
stop gained	4,731	31	3,653	2.2
intergenic variant	3,039	223	1,670	1
inframe indel	2,779	45	2,040	1.2
essential splice variant	2,575	48	1,952	1.2
TOTAL	669,718	28,538	405,036	240.3

Table 10. Number of variants in the dataset. The columns are, in order, the number of variants in the total dataset, the mean count of variants observed in one individual, as well as the same information for the subset of singleton variants, split by their functional effect.

Following previous works (Sazzini *et al.,* 2016; Fiorito *et al.,* 2016), we split the twenty Italian administrative regions into four macro-areas (Figure 101): Northern Italy, Central Italy, Southern Italy and Sardinia. We assigned individuals to a macro-area or to a region only when it was the shared birthplace of all four grandparents. Additionally, we removed those who did not cluster well with their macro-area in Principal Component Analysis (PCA, see *Methods - Data cleaning* on page 219). The final sample sizes of macro-areas and regions are in Table 11.

This dataset has enough power to detect 96% of variants with allele frequency greater than 0.1% and virtually all of the variants with frequency greater than 0.2%. By subsampling the Italian dataset, we observed a different trend between high and low-frequency variations in subsamples of increasing sizes. Common variants increase rapidly with the first few individuals, but their number rapidly stops growing 102. On the other hand, the number of low-frequency variants continues to increase with greater sample sizes, making up the vast majority of variants in large samples.

MACRO-AREAS	NUMBER OF SAMPLES
ITN	622
ITC	76
ITS	305
SAR	20

REGIONS	NUMBER OF SAMPLES
PIEDMONT	78
LOMBARDY	227
LIGURIA	6
EMILIA ROMAGNA	124
VENETO	78
FRIULI	3
TUSCANY	11
MARCHE	18
UMBRIA	10
MOLISE	1
LATIUM	25
ABRUZZO	19
CAMPANIA	50
BASILICATA	7
APULIA	106
CALABRIA	41
SICILY	57
SARDINIA	20

Table 11. Number of samples in the dataset. Number of samples in each Italian macro-area and in each Italian administrative region.

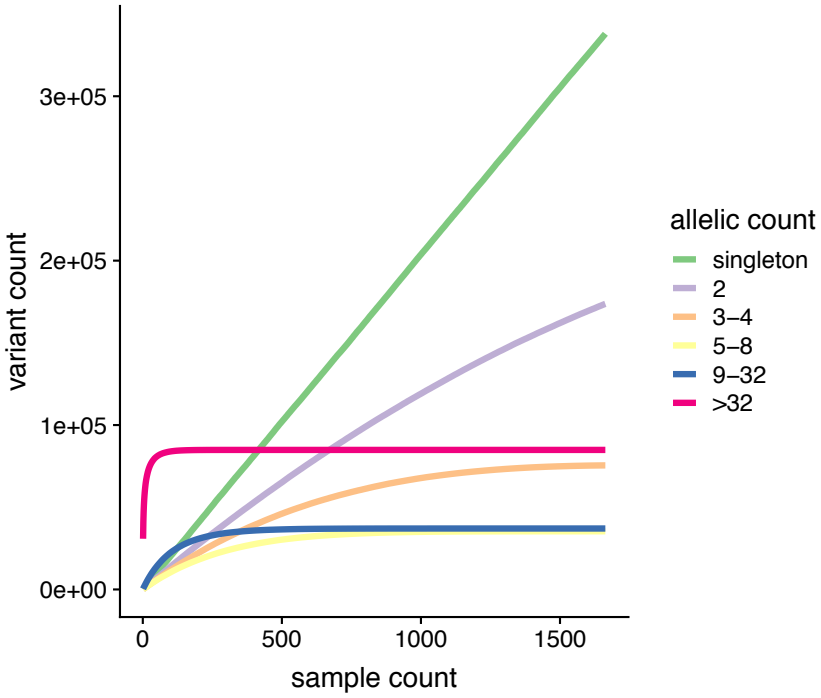


Figure 102. Number of variants observed by subsampling our dataset. We plot the number of variants observed in subsamples of increasing sizes from of our dataset, grouped by the frequency of their alternative allele in the full dataset.

Exploring the genetic structure with exome variants

The genetic structure of Italy appears clearly from PCA (Figure 103). The main visible features are the North-South cline and the Sardinian isolate, thus confirming the known genetic profile shown by high-density arrays. Some individuals cluster in a different macro-area than the reported one. Of these individuals, 43 have all four grandparents from the same macro-area, which could be due to either misreporting or recent migrations. We excluded these individuals in the subsequent analyses involving macro-areas and regions.

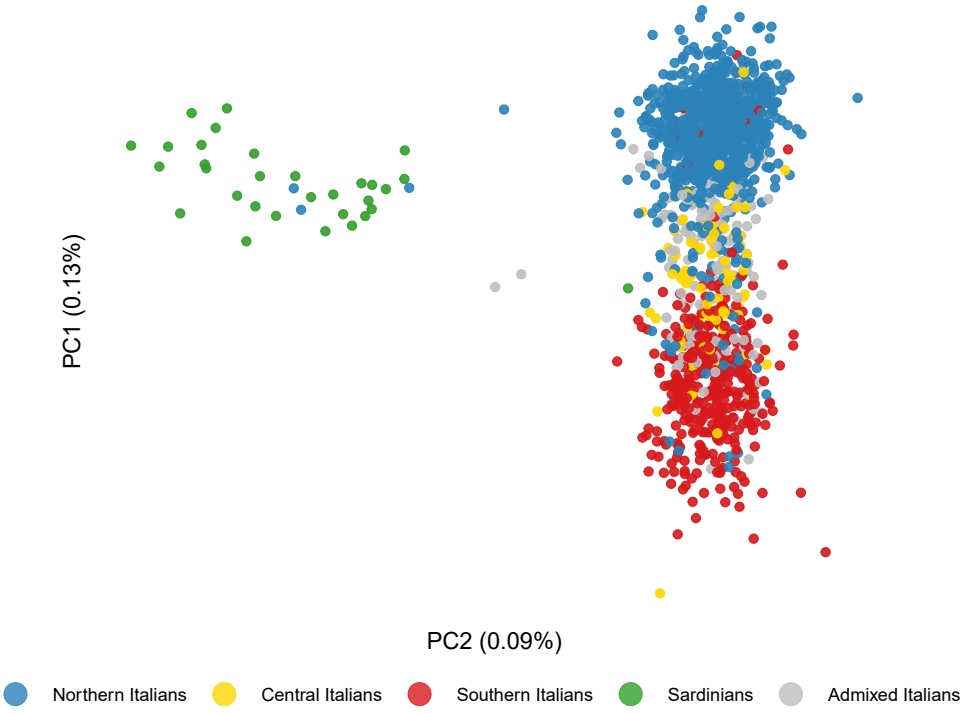


Figure 103. Genetic structure of the Italian population from PCA. Plot of the first two components of genotype PCA.

While we cannot discern any inner structure in macro-areas from the scatter plots, by plotting values of the first principal component for each region, we observed some stratification even at the regional level (Figure 104A and Table 12).

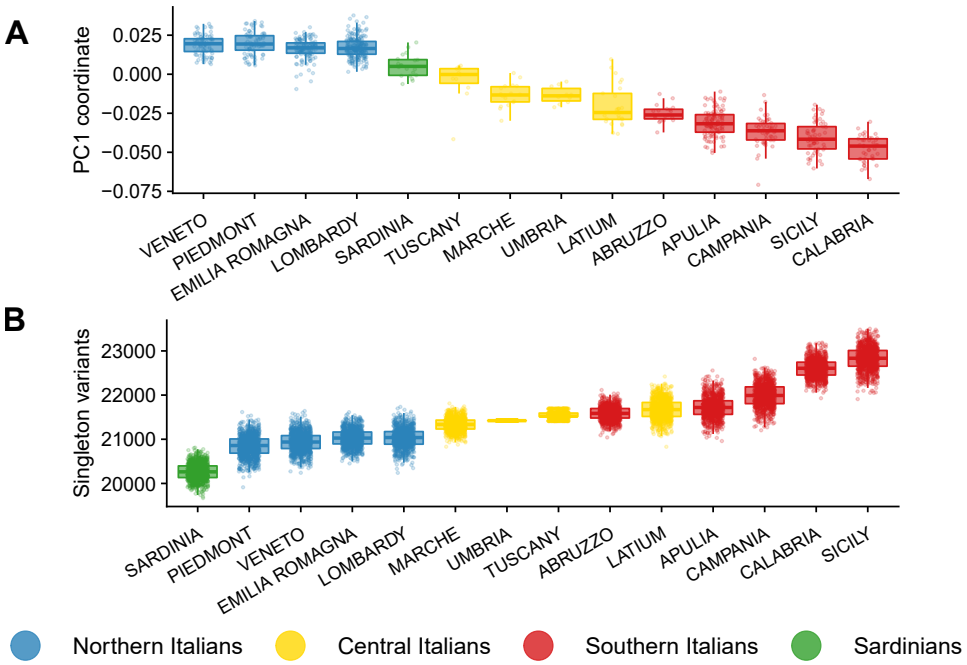


Figure 104. Regional stratification of Italian samples. **(A)** Strip and boxplot of the first principal component of each individual grouped by Italian regions. **(B)** Strip and boxplot of the number of low-frequency variants in one thousand resampling of ten individuals from each region. Only regions with at least ten individuals have been considered.

	ABRUZZO	CALABRIA	CAMPANIA	EMILIA ROMAGNA	LATIUM	LOMBARDY	MARCHE	PIEDMONT	APULIA	SARDINIA	SICILY	TUSCANY	UMBRIA	VENETO
ABRUZZO	1.00E+00	1.17E-14	9.76E-07	2.51E-12	5.57E-01	4.62E-13	8.53E-06	1.69E-11	1.19E-03	2.90E-11	1.59E-07	7.20E-05	1.94E-05	1.69E-11
CALABRIA	1.17E-14	1.00E+00	5.45E-08	9.38E-22	2.65E-16	2.27E-24	3.09E-15	3.94E-19	2.84E-15	3.21E-16	8.11E-04	6.46E-09	1.57E-10	3.94E-19
CAMPANIA	9.76E-07	9.50E-01	1.00E+00	6.49E-25	1.49E-07	1.83E-28	2.80E-09	1.71E-21	4.53E-04	8.37E-11	2.40E-02	7.67E-06	1.67E-06	1.71E-21
EMILIA ROMAGNA	5.45E-08	4.97E-16	9.89E-14	1.00E+00	1.22E-14	9.50E-01	1.27E-11	7.70E-04	5.27E-39	1.44E-07	3.75E-27	2.27E-07	1.79E-07	9.75E-03
LATIUM	2.51E-12	8.53E-06	1.10E-03	2.90E-11	1.00E+00	4.97E-16	1.64E-02	9.89E-14	1.68E-04	3.34E-08	1.06E-08	3.26E-03	3.76E-02	1.05E-13
LOMBARDY	9.38E-22	3.09E-15	4.55E-11	3.21E-16	8.11E-04	1.00E+00	1.70E-12	1.10E-03	6.39E-49	1.12E-08	1.82E-31	3.14E-08	8.95E-08	1.21E-02
MARCHE	6.49E-25	2.80E-09	1.19E-03	8.37E-11	2.40E-02	7.20E-05	1.00E+00	4.55E-11	1.92E-09	5.78E-09	5.34E-10	2.62E-03	9.44E-01	4.55E-11
PIEDMONT	5.57E-01	1.27E-11	2.84E-15	1.44E-07	3.75E-27	6.46E-09	5.53E-06	1.00E+00	5.35E-31	9.81E-09	4.12E-23	9.21E-08	3.04E-07	3.94E-01
APULIA	2.65E-16	1.64E-02	4.53E-04	3.34E-08	1.06E-08	7.67E-06	2.55E-02	3.76E-02	1.00E+00	1.51E-12	8.60E-08	5.53E-06	1.15E-06	5.35E-31
SARDINIA	1.49E-07	1.70E-12	5.27E-39	1.12E-08	1.82E-31	2.27E-07	2.07E-06	8.95E-08	6.09E-07	1.00E+00	3.69E-11	2.55E-02	1.33E-07	9.81E-09
SICILY	1.22E-14	1.69E-11	1.68E-04	5.78E-09	5.34E-10	3.26E-03	1.94E-05	9.44E-01	4.79E-03	9.75E-03	1.00E+00	2.07E-06	6.09E-07	4.12E-23
TUSCANY	4.62E-13	3.94E-19	6.39E-49	9.81E-09	4.12E-23	3.14E-08	1.57E-10	3.04E-07	1.69E-11	1.05E-13	3.94E-01	1.00E+00	4.79E-03	9.21E-08
UMBRIA	2.27E-24	1.71E-21	1.92E-09	1.51E-12	8.60E-08	2.62E-03	1.67E-06	1.15E-06	3.94E-19	1.21E-02	5.35E-31	4.12E-23	1.00E+00	3.04E-07
VENETO	1.83E-28	7.70E-04	5.35E-31	1.59E-07	3.69E-11	9.21E-08	1.79E-07	1.33E-07	1.71E-21	4.55E-11	9.81E-09	9.21E-08	3.04E-07	1.00E+00

Table 12. Raw two-sample Wilcoxon tests p-values for each pairwise region comparison of PC1 coordinates.

In order to provide an additional measure of population differentiation, we estimated the pairwise fixation index (F_{ST}, see *Methods - Exploring the genetic structure with coding variants* on page 221) between regions and macro-areas (Figures 105 and C.1, respectively). Again, we observed a strong differentiation between macro-areas. However, a finer dissection

of the Italian population was not possible, due to the low genetic distance between regions in the same macro-area combined with the relatively high standard errors of the estimates, especially for regions with small sample size (Figure C.1 and Tables 11, C.1, C.2).

In order to infer demographic events experienced by populations in Italian macro-areas, we tested for differences in the allele frequency spectrum, i.e. the distribution of the allele frequency of the variants in the subpopulations (see *Methods - Exploring the genetic structure with coding variants* on page 221). Significant differences were detected between all regions (except Lombardy and Emilia Romagna) in the low-frequency end of the spectrum, with Southern Italy having the most low-frequency variants (regions are plotted in Figure 104, macro-areas in Figure C.2 and nominal *p-values* are reported in Table 13).

	ABRUZZO	CALABRIA	CAMPANIA	EMILIA ROMAGNA	LATIUM	LOMBARDY	MARCHE	PIEDMONT	APULIA	SARDINIA	SICILY	TUSCANY	UMBRIA	VENETO
ABRUZZO	1	0	7.83E-217	0	2.73E-22	1.93E-302	2.02E-184	0	1.49E-45	0	0	4.35E-17	5.18E-180	0
CALABRIA	0	1	1.94E-288	0	0	0	0	0	0	0	6.37E-83	0	0	0
CAMPANIA	7.83E-217	0.7744130895	1	0	2.19E-129	0	0	0	1.78E-98	0	0	7.97E-268	0	0
EMILIA ROMAGNA	1.94E-288	4.43E-303	0	1	0	0.7744130895	5.79E-198	2.25E-64	0	0	0	0	0	7.69E-22
LATIUM	0	2.02E-184	3.23E-62	0	1	4.43E-303	2.30E-206	0	2.77E-05	0	0	4.97E-60	2.67E-202	0
LOMBARDY	0	0	7.75E-277	0	6.37E-83	1	8.20E-185	3.23E-62	5.74E-308	0	0	0	0	1.12E-21
MARCHE	0	0	1.49E-45	0	0	4.35E-17	1	7.75E-277	1.37E-228	0	0	1.21E-206	6.28E-76	6.15E-243
PIEDMONT	2.73E-22	5.79E-198	0	0	0	0	4.24E-97	1	0	8.39E-289	0	0	0	1.43E-15
APULIA	0	2.30E-206	1.78E-98	0	0	7.97E-268	0	2.67E-202	1	0	0	4.24E-97	4.33E-238	0
SARDINIA	2.19E-129	8.20E-185	0	0	0	0	0	0	0	1	0	0	0	0
SICILY	0	0	2.77E-05	0	0	4.97E-60	5.18E-180	6.28E-76	2.71E-249	7.69E-22	1	0	0	0
TUSCANY	1.93E-302	0	5.74E-308	8.39E-289	0	0	0	0	0	0	1.43E-15	1	2.71E-249	0
UMBRIA	0	0	1.37E-228	0	0	1.21E-206	0	4.33E-238	0	1.12E-21	0	0	1	0
VENETO	0	2.25E-64	0	0	0	0	0	0	0	6.15E-243	0	0	0	1

Table 13. Raw two-sample Wilcoxon tests *p-values* for each pairwise region comparison of allele frequency spectrum. Only regions with at least ten individuals have been considered.

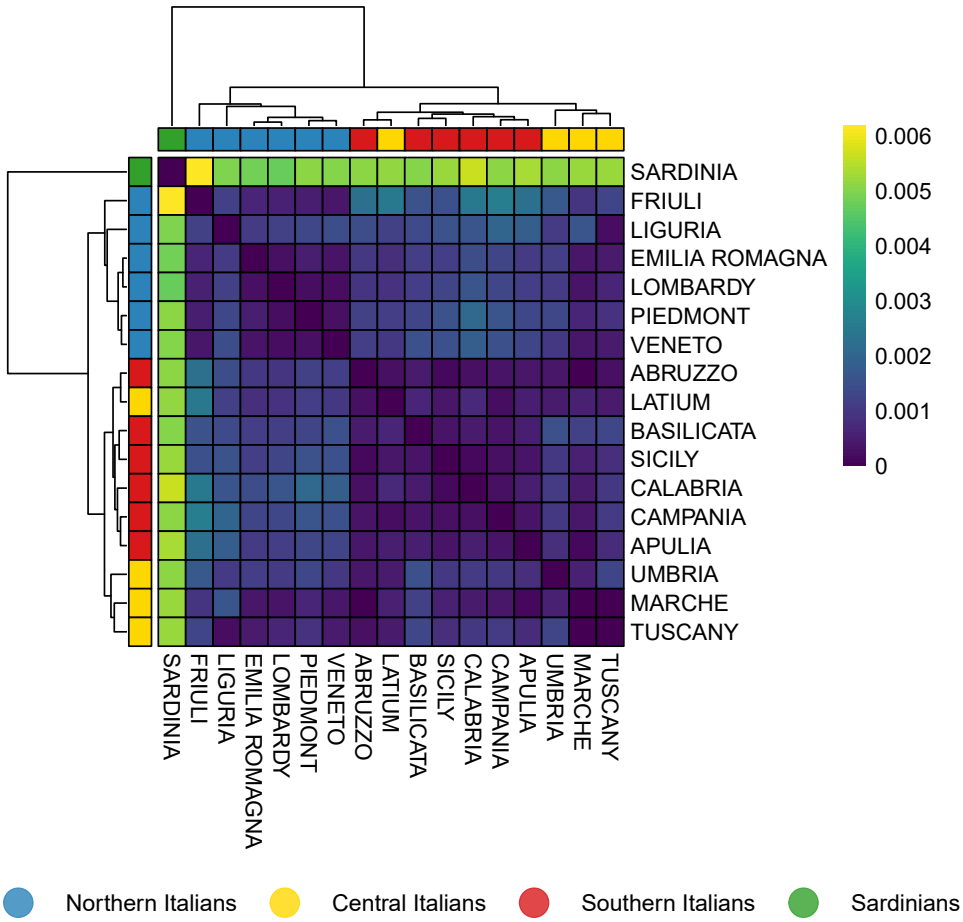


Figure 105. Genetic structure of the Italian population from F_{ST} values. Heatmap representing the F_{ST} values computed for each pairwise comparison between Italian regions except Molise, for which just one individual was available.

Allele frequency differences between Northern and Southern Italy

After evaluating the genetic structure of Italy from a global point of view, we delved into details examining which variants and genes present the highest degree of differentiation between Northern and Southern Italy (622 and 305 individuals, respectively), focusing on the possible phenotypic implications of the structure we observed.

Allele frequencies in the North and South are highly correlated (Pearson

r=0.998, p-value $< 2.2e^{-16}$, Figure 106), with a maximum difference of 17%. We listed the variants with the most significant allele frequency differences that are statistically significant with p-value under 0.01 after correction for multiple testing (Table 14). They are also the six variants with the highest single-locus F_{ST} estimate. From the test results, we observed a genomic inflation factor of 1.77 (Figure 107), as expected when comparing groups with population stratification.

CHR	SNP	REF	ALT	Northern Italy	Southern Italy	Nominal p-value	Gene symbol
15	rs1129038	C	T	51.10%	34.20%	$5.76e^{-12}$	*HERC2*
11	rs2053116	A	C	37.40%	53.40%	$7.14e^{-11}$	*OR52R1*
11	rs6578533	T	A	37.50%	53.40%	$1.04e^{-10}$	*OR52R1*
11	rs7941731	A	G	37.50%	52.80%	$5.28e^{-10}$	*OR52R1*
4	rs1229984	T	C	93.11%	84.40%	$1.28e^{-08}$	*ADH1B*
5	rs256438	T	G	38.10%	25.40%	$5.19e^{-08}$	*THBS4*

Table 14. Significantly different variants between Northern and Southern Italy (Bonferroni's multiple test correction with p-value < 0.01).

The strongest signal is rs1129038, located in the 3' UTR of the *HERC2* gene. The derived allele T is enriched in Northern Italy and it is associated with eye and hair color and skin pigmentation (Morgan *et al.,* 2018). Its homozygous occurrence is highly predictive of blue eye color (Eiberg *et al.,* 2008), but it also associated with pigmentation-related diseases like melanoma and vitiligo (Jin *et al.,* 2012).

The second strongest signal comprises three very close (less than 400bp apart) and previously unreported missense variants in the *OR52R1* gene, an olfactory receptor. RefSeq reports *OR52R1* as a segregating pseudogene.

The third signal is rs1229984, a missense variant in the *ADH1B* gene, encoding the beta subunit of class I ADH (alcohol dehydrogenase), which is involved in ethanol metabolism. The derived allele T (which is the reference allele) protects against alcoholism by metabolizing alcohol to acetaldehyde more efficiently than the ancestral allele C, leading to elevated acetaldehyde levels that make drinking unpleasant. It is also associated with alcohol-related diseases, alcohol dependence (Sanchez-Roige *et al.,* 2019), cancers (Lesseur *et al.,* 2016) and, not surprisingly, to the trait "Regular attendance at a pub or social club" (Day *et al.,* 2018).

The fourth signal, located in the intronic SNP rs256438 of *THBS4* gene, is closer to the background (Figure 107). The variant was inconclusively as-

sociated with serum TSH (Thyroid-stimulating hormone) levels in a GWAS study (Malinowski *et al.,* 2014).

We also examined the rest of the top 1% of the most significant variants. We found some interesting and significant associations in the GWAS catalog (Denny *et al.,* 2013; MacArthur *et al.,* 2017): rs16891982 in the *SLC45A2* gene was associated with pigmentation, rs16903574 in the *OTULINL* gene with allergic diseases (asthma, hay fever, eczema) and rs3131379 in the *MSH5* gene with systemic lupus erythematosus. Many other variants were associated with various cardiovascular traits.

Moreover, using the 540 genes harboring the top 1% most significant variants, we performed gene enrichment analysis with EnrichR (Chen *et al.,* 2013; Kuleshov *et al.,* 2016). Significant results included morphological traits (e.g., skin and hair color), light-related reactions ("Skin sensitivity to sun", "Low tan response", "Melanoma" and "Skin cancer"), immune terms (e.g., "Type 1 diabetes and autoimmune thyroid diseases", "HIV-1 control", "Asthma", "Psoriasis", "Systemic_scleroderma", "Rheumatoid_arthritis" and "Multiple_sclerosis") and complex diseases, such as cardiovascular diseases and cancer. Additionally, nine variants in the top 1% were in the *LCT* gene, coding for the lactase enzyme.

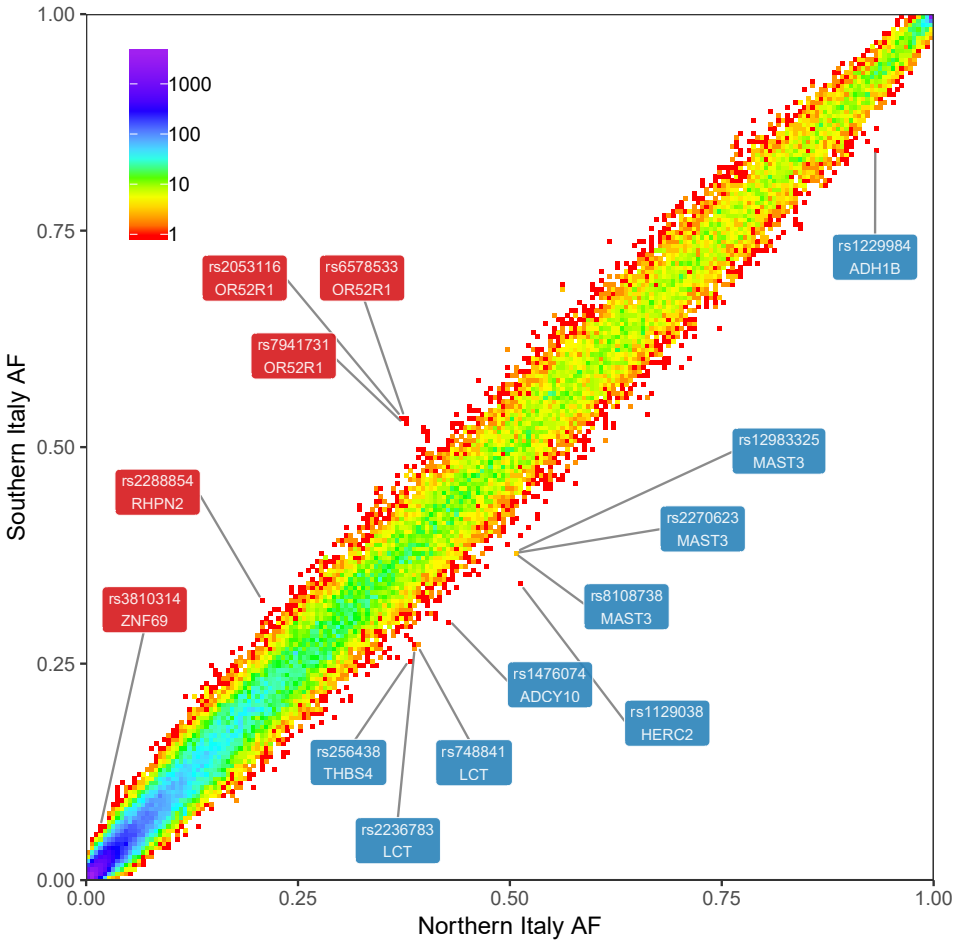


Figure 106. Allele frequency (AF) differences between Northern and Southern Italy. Density plot of all variants plotted by their allele frequency in Northern and Southern Italy. Labels with the variant ID and gene are shown for the 14 variants with p-value under 0.04 after multiple test correction. Label color denotes the higher frequency of the allele in the Northern or Southern Italian population.

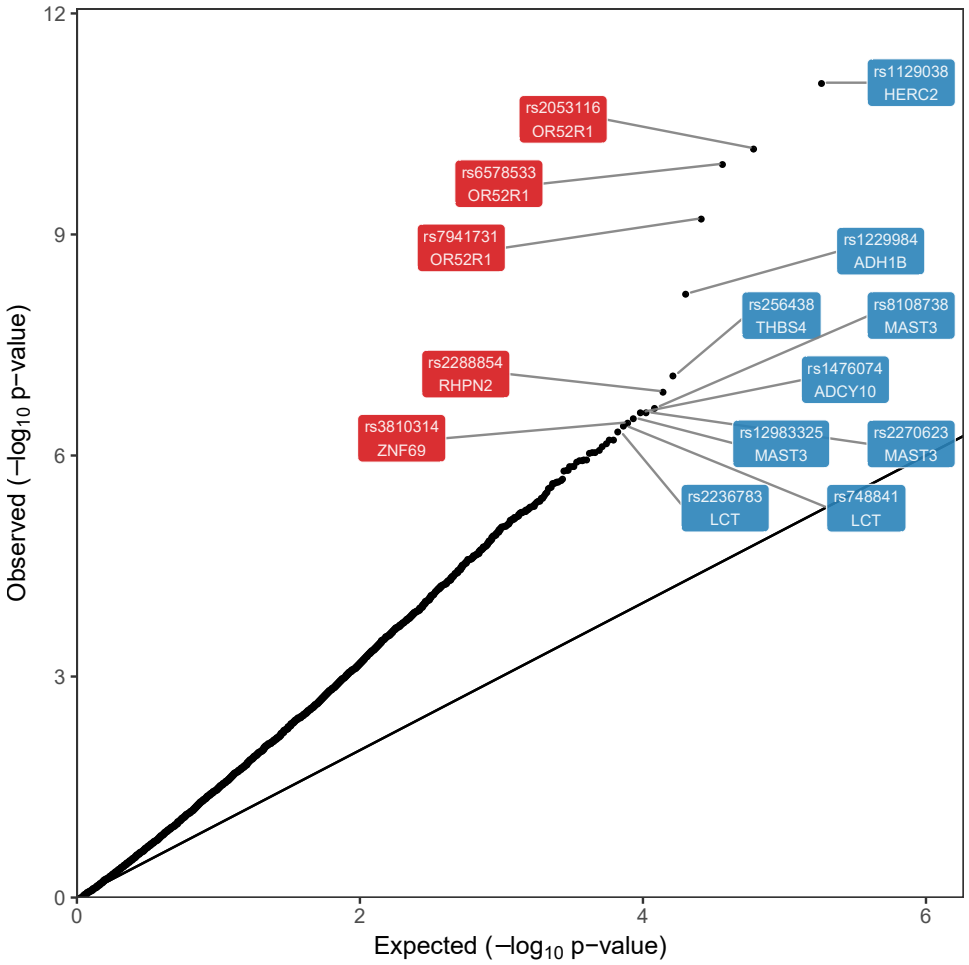


Figure 107. Quantile-quantile plot of Fisher's exact test p-values for the comparison between Northern and Southern allele frequencies. Labels with the variant ID and gene are shown for the 14 variants with p-value under 0.04 after multiple test correction. Label color denotes the higher frequency of the allele in the Northern or Southern Italian population.

Allele frequency differences between Italy and Europe

The allele frequency differences we observed along the Italian peninsula are part of larger North-South European clines reported in gnomAD. In the case of *HERC2* and *OR52R1*, these clines extend to Africa. On the other hand, for *ADH1B* the derived allele is enriched only in Southern Europe, while the ancestral allele is almost fixed in both Northern Europe and Africa. Only

East Asia sports a (much) higher frequency of the derived allele (Table 15).

gnomAD AF	***HERC2*** **rs1129038**	***OR52R1*** **rs6578533**	***ADH1B*** **rs1229984**	***THBS4*** **rs256438**
North-western European	0.7722	0.3365	0.9745	0.3654
Southern European	0.4392	0.4431	0.9067	0.3506
African	0.1194	0.5149	0.9884	0.2266
East Asian	0.0008046	0.01782	0.2623	0.2199

Table 15. gnomAD allele frequency in Europe, Africa and East Asia of mostly different variants in Italy.

We tested all the variants for significant allele frequency differences between the whole Italian population from our dataset and gnomAD non-Finnish European population. We found 19,561 variants passing the 0.01 significance threshold after Bonferroni's multiple test correction. We found 988 of these variants in the GWAS catalog with highly significant associations (p-value $< 5.00e^{-08}$. We observed features mostly related to the different environment, such as "hair color", "bitter taste perception" and "low tan response", others involving the immune system (e.g., "Rheumatoid arthritis", "Type 1 diabetes" and "Myeloid white cell count") and many others related to the blood protein, triglycerides and LDL cholesterol levels. Interestingly, 15 variants resulted associated with cardiovascular traits, such as hypertension, systolic and diastolic blood pressure, pulse and mean arterial pressure.

Eight variants, out of the significantly different ones between Italian and European non-Finnish population, have been reported as pathogenic or likely pathogenic in the ClinVar database (Landrum *et al.,* 2016, with no conflicting interpretation of benign/likely benign/uncertain significance; Table 16). Among the variants reported in Table 16, we observed a stop-gained variant with higher frequency in gnomAD European non-Finnish population in the gene *FLG* (coding for profilaggrin). This variant has been described as pathogenic for atopic dermatitis (Richards *et al.,* 2015) and ichthyosis vulgaris diseases (Smith *et al.,* 2006). Another stop-gained was present in the beta-globin gene (*HBB*) causing thalassemia (Richards *et al.,* 2015): interestingly, the frequency of this variation was higher in the Italian population. We also found a splicing variant in *BRCA1*, with higher frequency in Italy, reported pathogenic for hereditary breast and ovarian cancer syndrome, causing the skipping of exon 11 (Bonatti *et al.,* 2006).

However, during 2018 this variant was reclassified with the flag "Uncertain significance" (Nykamp *et al.,* 2017).

It might also be noted that 169 variants were annotated in the PharmGKB database, suggesting that the Italian population might respond differently to some drugs. Finally, we performed enrichment analysis with EnrichR on the 9,158 genes harboring the 19,561 variants, where traits related to morphological phenotypes (e.g., the hair color), immune and other complex diseases (e.g., "Cardiovascular system disease") emerged.

Variant ID	Gene	Variant effect	ITALY AF	ITN AF	ITC AF	ITS AF	SAR AF	NFE AF	nominal p-value	ClinVar condition
rs142063461	*TSHR*	Missense variant	0.45%	0.16%	0.66%	0.66%	0.00%	0.05%	$6.05e^{-19}$	Congenital non goitrous hypothyroidism
rs11549407	*HBB*	Stop gained	0.47%	0.24%	0.00%	0.66%	5.00%	0.07%	$1.50e^{-15}$	Thalassemia
rs769409705	*SLC34A1*	Missense variant	0.12%	0.24%	0.00%	0.16%	0.00%	0.00%	$1.60e^{-10}$	Hypercalcemia infantile
rs61816761	*FLG*	Stop gained	0.24%	0.33%	1.32%	0.17%	0.00%	1.67%	$2.60e^{-10}$	atopic dermatitis and ichthyosis vulgaris
rs28730837	*MPO*	Missense variant	0.36%	0.40%	0.00%	0.33%	0.00%	1.80%	$5.15e^{-10}$	Myeloperoxidase deficiency
rs151206295	*TYR*	Missense variant	0.21%	0.48%	0.00%	0.00%	0.00%	0.02%	$5.94e^{-09}$	Tyrosinase-negative oculocutaneous albinism
rs80358178	*BRCA1*	Splice donor variant	0.09%	0.00%	0.66%	0.00%	0.00%	0.00%	$1.91e^{-08}$	Hereditary breast and ovarian cancer syndrome
rs28929474	*SERPINA1*	Missense variant	0.53%	0.64%	0.00%	0.33%	0.00%	1.82%	$4.20e^{-08}$	Alpha-1-antitrypsin deficiency

Table 16. Significantly different variants between Italian and European non-Finnish population reported "pathogenic" or "likely pathogenic" in the ClinVar database (Landrum *et al.,* 2016). *P*-value computed with Fisher's exact test is reported. AF: allele frequency; ITN: Northern Italy; ITC: Central Italy; ITS: Southern Italy; SAR: Sardinia; NFE: Non-Finnish European from gnomAD.

Importance of the reference population for assessing pathogenicity

When assessing pathogenicity of variants, it is often appropriate to assume that variants that occur frequently in healthy individuals are not pathogenic, at least not with high penetrance. This is normally done taking advantage of frequencies reported in public databases. We explored the risk of misassessing the pathogenicity of a variant by using frequencies estimated from either too small a sample or a population that is not a perfect match for the individual. We compared the allele frequencies estimated from our Italian sample with those estimated from the Tuscans and Non-Finnish European in the 1000 Genomes Project and from the Non-Finnish European in gnomAD, as shown in Table 17.

For instance, if we select an allele frequency threshold of $\geq 1\%$ for assessing variants as non-pathogenic, there are 3,782 false pathogenic candidates variants whose frequency is above the threshold in our sample but under it in gnomAD. While most of them are very close to the threshold in both datasets, a few are quite different: 590 are under 0.5% and 33 under 0.1% in gnomAD, while still being above 1% in our sample. Frequency discrepancies as large as these in variants satisfying other pathogenicity criteria could cause an incorrect assessment of pathogenicity even under careful scrutiny.

DB	HIGH	LOW	FALSE_POSITIVE	TRUE_NEGATIVE	TOTAL
KGP_TSI	0.01	0.01	5889	5338	267339
KGP_NFE	0.01	0.01	3854	4627	267339
GND_NFE	0.01	0.01	3782	5167	669718
KGP_TSI	0.011	0.009	2373	4663	267339
KGP_NFE	0.011	0.009	2264	3832	267339
GND_NFE	0.011	0.009	2264	3383	669718
KGP_TSI	0.012	0.008	1893	3850	267339
KGP_NFE	0.012	0.008	1189	2019	267339
GND_NFE	0.012	0.008	1182	1959	669718
KGP_TSI	0.015	0.005	963	523	267339
KGP_NFE	0.015	0.005	140	249	267339
GND_NFE	0.015	0.005	80	241	669718
KGP_TSI	0.005	0.005	11290	8214	267339
KGP_NFE	0.005	0.005	7031	6726	267339
GND_NFE	0.005	0.005	6267	7463	669718
KGP_TSI	0.0055	0.0045	3647	7735	267339
KGP_NFE	0.0055	0.0045	3683	6104	267339
GND_NFE	0.0055	0.0045	4051	5528	669718
KGP_TSI	0.006	0.004	2922	6916	267339
KGP_NFE	0.006	0.004	2808	5073	267339
GND_NFE	0.006	0.004	2382	3615	669718
KGP_TSI	0.0075	0.0025	1757	4188	267339
KGP_NFE	0.0075	0.0025	751	690	267339
GND_NFE	0.0075	0.0025	218	664	669718
KGP_TSI	0.001	0.001	38191	15872	267339
KGP_NFE	0.001	0.001	13431	43203	267339
GND_NFE	0.001	0.001	31064	15818	669718
KGP_TSI	0.0011	0.0009	37875	15514	267339
KGP_NFE	0.0011	0.0009	13230	42304	267339
GND_NFE	0.0011	0.0009	28193	13608	669718
KGP_TSI	0.0012	0.0008	30411	12506	267339
KGP_NFE	0.0012	0.0008	9331	34414	267339
GND_NFE	0.0012	0.0008	15717	8513	669718
KGP_TSI	0.0015	0.0005	24443	7682	267339
KGP_NFE	0.0015	0.0005	6536	3991	267339
GND_NFE	0.0015	0.0005	4569	2712	669718

Table 17. Number of variants in our dataset that could be incorrectly classified by their allele frequency using an external reference. "False positive" column indicates the number of variants with high frequency in our Italian dataset and low frequency in the external dataset. "True negative" column reports the number of variants with low frequency in our Italian dataset and high frequency in the external dataset. We used the following external datasets: KGP TSI are the Tuscans from the 1000 Genomes Project, KGP NFE are the non-Finnish Europeans - TSI, IBS, CEU, GBR - from the 1000 Genomes Project and GND NFE are the non-Finnish Europeans from gnomAD.

Putatively pathogenic variants

We examined how other common pathogenicity assessment criteria behave on the healthy Italian cohort. We employed three different and complement-

ary methods for assessing the pathogenicity of a variant: protein-truncating variants (PTV), missense variants predicted to be damaging (DMG) (Materials and Methods) and variants diagnosed pathogenic in the ClinVar database (CLNPAT, Landrum *et al.,* 2016). After excluding variants with allele frequency greater than 5% in our dataset, we obtained 12,851 PTV, 23,682 DMG and 1,149 CLNPAT variants in the whole dataset. We refer to these variants as putatively pathogenic (PP) variants. We related the functional effect and pathogenicity of variants to their frequency, measured as the ratio of singleton variants to the total number of variants in each category (Figure 108A). We also investigated the pathogenic burden in our healthy individuals in Figure 108B, C and D, showing that the mean number of PP variants in an individual is 16.5 for PTV, 23.2 for DMG and 2.4 for CLNPAT variants.

We counted and normalized the number of PTV and DMG variants with allele frequency lower than 1% across the genes and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, thus sorting genes and pathways according to their tendency to accumulate PTV and DMG variations (Figures 109 and 110, respectively). We found that chemical carcinogenesis, drug and xenobiotics metabolism, other metabolic pathways and olfactory transduction pathway accumulate the highest number of functionally disrupting variants (Figure 109). Among the pathways with the lowest tendency to accumulate PTVs we found the highly conserved "Hippo signaling" and other pathways with essential cell functions (Figure 109). Other metabolic pathways and the "ABC transporters" class resulted the most prone to accumulate DMG variations, while, as for the PTVs, categories participating to basic function of the cell (e.g., "Protein export", "Spliceosome" and "mRNA surveillance pathway") were depleted for DMG variants (Figure 110).

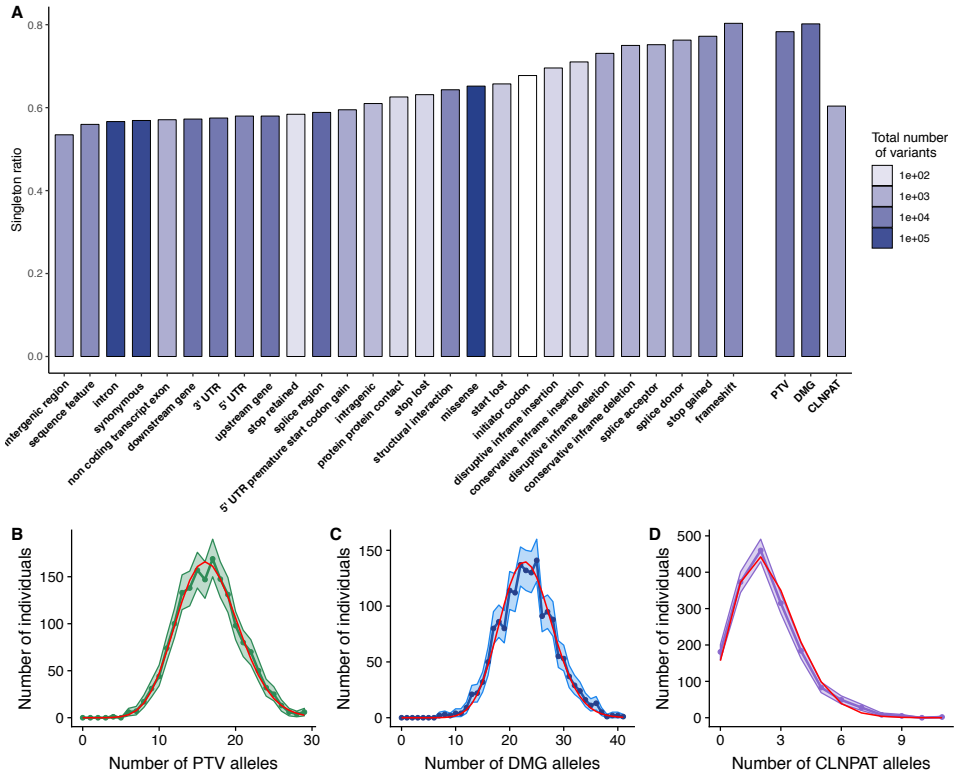


Figure 108. Evaluation of pathogenic variation. **(A)** Ratio of singleton to total number of variants for each effect and pathogenic category. **(B)**, **(C)** and **(D)** Distribution of the burden of each pathogenic category per individual, counted as the number of variants observed. Shaded area shows the 5-95% confidence interval estimated by bootstrap. The red line represents the theoretical Poisson distribution as reference.

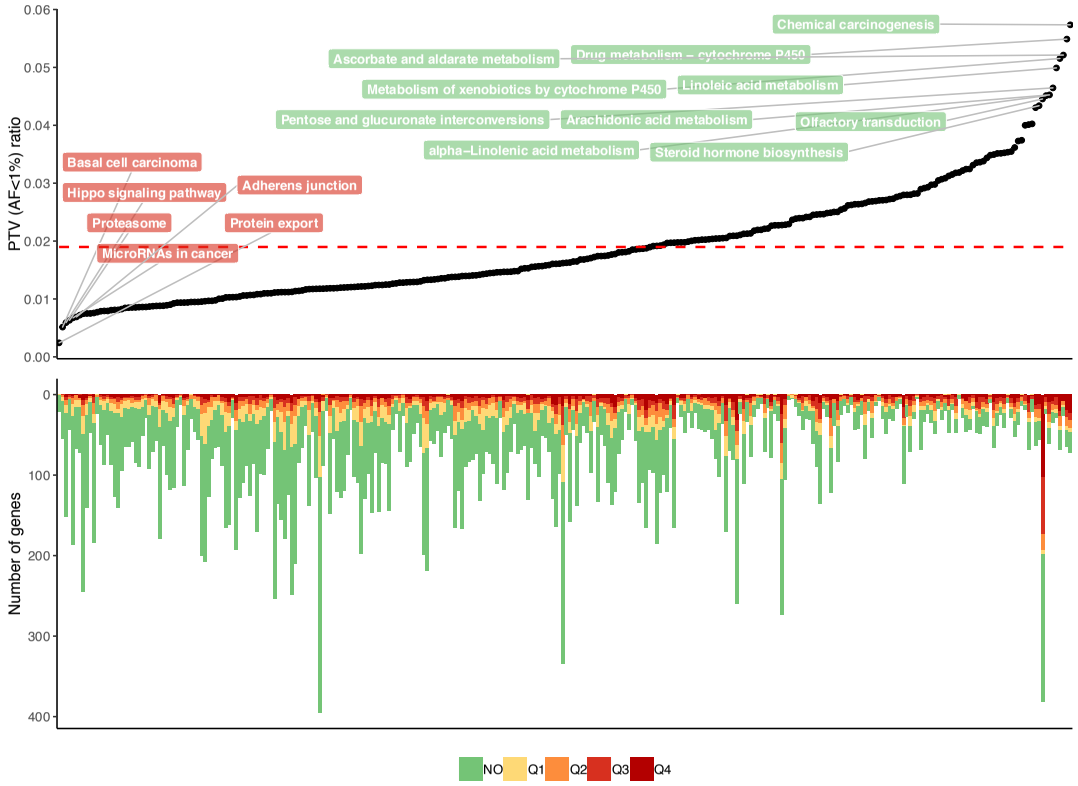


Figure 109. PTV Burden in KEGG pathways. The top figure shows the count of PTVs (with allelic frequency lower than 1%) per KEGG pathway divided by the total number of variants (lower than 1%) in that pathway (PTVs ratio). The dashed line shows the mean value across all KEGG pathways. The bottom figure is a stacked barplot representing the number of genes in each KEGG pathway. Colors refer to the quantile of PTVs ratio distribution in which each gene falls.

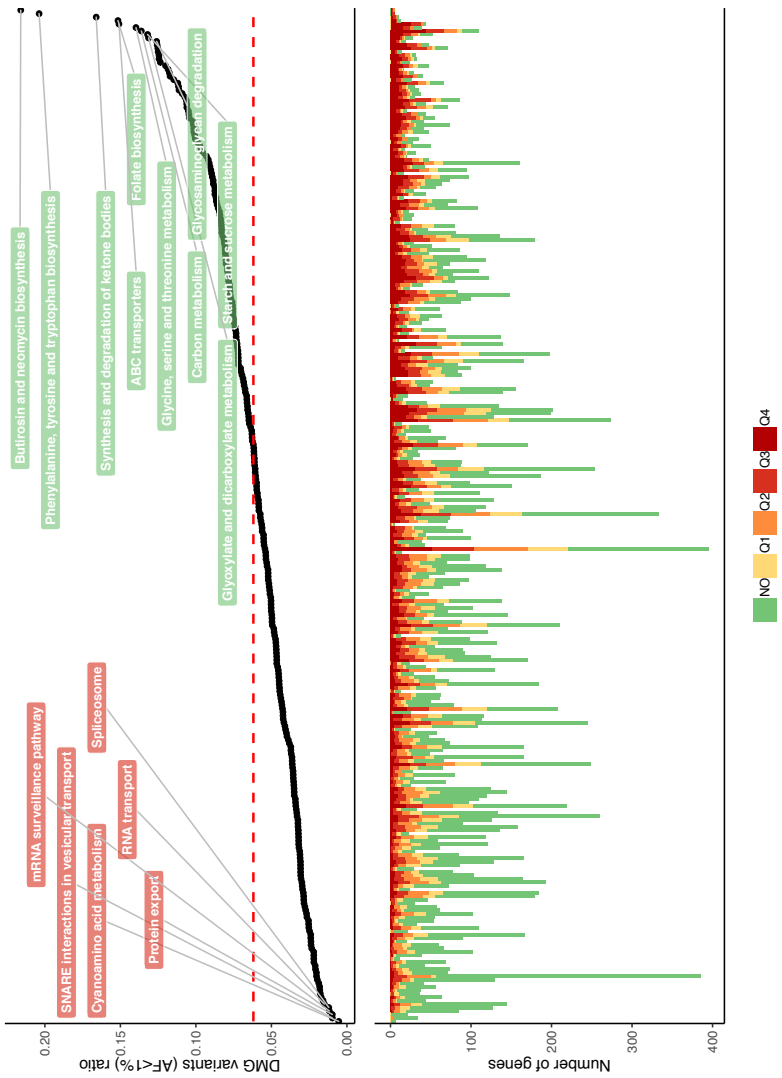


Figure 110. DMG Burden in KEGG pathways. The top figure shows the count of DMG (with allelic frequency lower than 1%) per KEGG pathway divided by the total number of variants (lower than 1%) in that pathway (DMG variants ratio). The dashed line shows the mean value across all KEGG pathways. The bottom figure is a stacked barplot representing the number of genes in each KEGG pathway. Colors refer to the quantile of DMG variants ratio distribution in which each gene fall.

Putatively pathogenic variants in ACMG SF genes

Previous studies (Karczewski *et al.,* 2019; Lek *et al.,* 2016) showed that many genes are quite tolerant of variants causing loss of function. As a consequence, PP variants in those genes are unlikely to be actually pathogenic, thus explaining part of the burden of PP variants that we observed in our and other healthy cohorts. In order to explore PP variants in a clinically relevant context, we focused on the 59 medically actionable genes recommended by the American College of Medical Genetics and Genomics for reporting of incidental findings (ACMG SF v2.0; Kalia *et al.,* 2017).

Comparing the amount of PP variants found in the whole exome versus those in the ACMG SF genes (Table 18), we observed that the proportion of PTVs in the ACMG genes is half of the proportion in other genes. On the other hand, DMG and CLNPAT variants in ACMG genes account for a much higher part of the variants than in the rest of the exome. As a comparison we see that missense variants are only slightly less abundant.

	Variant count in exome	Variant ratio in exome	Variant count in ACMG	Variant ratio in ACMG	Variant ratio enrichment in ACMG
PTV	12,852	1.92%	43	1.00%	52.24%
DMG	23,682	3.54%	334	7.79%	220.22%
CLNPAT	1,149	0.17%	49	1.14%	665.90%
MISSENSE	225,817	33.72%	1,384	32.27%	95.70%

Table 18. PP variants in whole exome and in the ACMG genes. Comparison of the number of PP variants in the whole exome versus those in the ACMG SF genes(referred in the table as ACMG).

We observed also the interaction between PTV/DMG variants in the ACMG genes and their annotation in ClinVar (Table 19).

	PTV	DMG	OTHER
BENIGN	0	9	1,097
BENIGN,UNCERTAIN_SIGNIFICANCE	0	32	414
UNCERTAIN_SIGNIFICANCE	1	84	326
UNCERTAIN_SIGNIFICANCE,PATHOGENIC	1	28	14
PATHOGENIC	19	20	10
CONFLICTING	1	12	31
Not in ClinVar	21	150	2019

Table 19. ClinVar annotation of PTV/DMG variants in the ACMG genes. We collapsed the likely benign and likely pathogenic classes into the benign and pathogenic classes, respectively.

Almost half of the variants in the ACMG genes were annotated in ClinVar. Unsurprisingly, almost all annotated PTVs were classified as pathogenic. The classification of DMG variants is less foregone: most are of uncertain significance and the rest is evenly split between the benign and pathogenic classes.

We reported the prevalence of PP variants in our healthy Italian population (Table 20). PTV and DMG variants were further restricted removing those that were reported benign or likely benign in ClinVar at least once:

	INDIVIDUALS	RATIO
PTV	82	4.90%
PTV without benign	49	2.90%
DMG	548	32.50%
DMG without benign	327	19.40%
CLNPAT	72	4.30%

Table 20. PP variants in the ACMG genes. Prevalence of PP variants in ACMG genes in healthy Italian population.

We found two variants in two ACMG genes in the top 1% of genetic differences between Northern and Southern Italy. For the reasons outlined above, we expected no serious genetic consequent for variants in the top 1%: indeed, the variants in the genes *PCSK9* and *MSH6* were reported to be intronic and synonymous, respectively, and were classified as benign in ClinVar.

Discussion

The genetic structure of the Italian population has already been investigated, usually using SNP array data mainly comprised of genome-wide common genetic markers (Raveane *et al.,* 2019; Sazzini *et al.,* 2016; Fiorito *et al.,* 2016; Parolo *et al.,* 2015; Di Gaetano *et al.,* 2012). Other well-known studies explored human genetic variation worldwide, from whole-exome data (ExAC/gnomAD). However, this is the first study to explore the Italian genetic structure from whole-exome data for a sizable number of individuals.

We corroborate the previously observed genetic structure of Italy, with north-south cline of the mainland and the outlying Sardinian isolate. This shows that whole-exome sequencing data provides enough information to highlight the Italian macro-areas, both in PCA, F_{ST} and in the allele frequency spectrum for rare variants.

The most evident structure is the split between Sardinia and the other continental macro-areas. Note that this split is associated with the second and not the first principal component because the Sardinian sample size is small. The genetic isolation of Sardinia is clearly shown in 104B since it has the lowest amount of low-frequency variants, which denotes a lower effective population size (Lao *et al.,* 2008; Marth *et al.,* 2004).

The second most important substructure (as shown in F_{ST}) that we observed is the split between the more genetically homogeneous Northern Italy and the Central and Southern macro-areas, which show more substructure, both in the first principal component and in the amount of low-frequency variants. The amount of low-frequency variants increases going from North to South, peaking in Sicily and nearby Calabria, denoting a greater effective population size in Southern Italy. Our results on the effective population size in Italy confirm previous results inferred by Identity By Descent (IBD) analysis (Fiorito *et al.,* 2016). They are also suggestive in light of the historical domination of Sicily by the Moors and the Normans. Similarly to previous results (Raveane *et al.,* 2019; Sazzini *et al.,* 2016), our analyses do not show a clear genetic distinction between Central and Southern Italy. This evident genetic structure in the Italian population drives us to explore the differences in the allele frequencies leading the peculiar genetic makeup of the Italian population.

Differences in allele frequencies between populations are due to many causes: evolutionary forces, different demographic history (i.e., different contributions of admixture and ancestral migrations) and genetic drift. Since Northern and Southern Italy are genetically separated only by a moderate distance (confirmed by the high correlation in allele frequencies), we can

reasonably rule out a substantial contribution of drift in producing the topmost difference of frequencies that we observe. As reported in literature (Raveane *et al.,* 2019; Fiorito *et al.,* 2016; Sazzini *et al.,* 2016; Sarno *et al.,* 2017; Parolo *et al.,* 2015), the Italian populations experienced, since prehistoric times, a complex history of migrations and admixtures, which left some remarkable genetic signals.

Many of the differences in frequency between Northern and Southern Italy occur in genes involved in greater European latitudinal clines like skin/hair pigmentation (*HERC2*, *SLC45A2*) and lactose tolerance (*LCT*), where also the associated phenotypes are known to follow the same cline. Pigmentation and skin diseases caused by UV light exposure were also a recurrent theme in GWAS direct associations and GWAS gene enrichments both in the Northern-Southern Italy comparison and in the Italy-Europe comparison. The rs1229984 variant in the *ADH1B* gene follows a similar yet different distribution: the derived allele is almost absent in Northern Europe but more frequent in Spain and Italy, especially in Southern Italy. However, it is rare also in Africa and worldwide, except in Eastern Asia, where it is the major allele and shows signs of recent positive selection (Polimanti & Gelernter, 2018). Unfortunately, we can only speculate on the cause of this similarity between such distant places.

Other differences we observed are harder to link to a phenotype, like the differences in olfactory receptor genes. The olfactory receptors represent the largest gene family and, in mammals, modifications and gain/loss of these genes are typically related to environmental adaptation (Hayden *et al.,* 2010). However, while approximately 400 are functional genes, pseudogenes account for the same number in the human genome (Glusman *et al.,* 2001). Olfactory receptors are generally considered to be under low selective pressure for their tolerance of loss of function variants (Karczewski *et al.,* 2019) and thus quite variable. This makes the signal we found in the *OR52R1* gene, while quite strong, also hard to interpret.

Investigating the variants with high-frequency differences between Northern and Southern Italy and between Italy and Europe, both by direct association in the GWAS catalog and by gene enrichments, produced a vast number of phenotypes and diseases. We can group most of them in five broad categories: pigmentation (already discussed), cardiovascular (both as susceptibility to diseases and related phenotypes), immune diseases, cancer and neurological disorders. All of these are complex phenotypes and it is not possible to give a unified phenotypic interpretation of the many variants associated with each one. We can argue that environmental differences, acting either where European populations live or in the lands from where the

different ancestral populations contributing to present-day Europeans came from, could have shaped the frequencies of some of these variants. The most striking example is represented by the immune system: the very fact that it is called upon to protect the body from harmful intrusion from the environment makes this system genetically malleable and, thus, involving genes with great frequency differences in human population.

Knowledge of the genetic makeup of the healthy population is crucial for the study of pathogenic variation in disease affected individuals. One of the main goals in medical genetics is finding associations between variants and disease occurrence, both in research and in diagnosis. A common scenario in the field is the diagnosis of a genetic disease, where the researcher/clinician looks for high-penetrance disease-causing variants.

In this setting, an often-employed criterion for assessing pathogenicity is excluding variants that have a high allele frequency in the healthy population on the assumption that purifying selection should have curbed the frequency of high-penetrance pathogenic variants (Richards *et al.,* 2015). While the validity of the assumption and the appropriate frequency threshold depend on the specific disease, in many cases discarding variants with high frequency is a powerful tool to rapidly sift through long lists of candidate pathogenic variants. Of course, low frequency alone is not evidence of pathogenicity and other criteria must be satisfied for a variant to be diagnosed as such.

One weakness of this strategy is that, unless the study has an appropriate control population from which the allele frequency can be reliably estimated, it must be taken from a public database instead. While this approach is generally sensible, it entails a risk when the population sampled in the public database is not fully representative of the population the study sample originates from or when patients with several diseases are included in the database.

The risk is assessing a variant as potentially pathogenic when it is rare in the chosen public database but it is relatively common in the actual population the sample comes from. We will call these variants false pathogenic candidates. Note that the reverse (rare in the actual population and common in the reference one) is not as problematic since a variant that is common in any population is unlikely to be pathogenic with high-penetrance.

As we have seen in our results, the choice of a database of allele frequencies plays a role in the assessment of pathogenic variation. The bigger and more closely related the reference sample is, the most accurate is the assessment. What does this entail for Italy when publicly available European reference database are employed? We observed that a large sample size seems to be the most important factor: the 107 Tuscans from KGP3, while

the most closely related to our sample, are too few to produce accurate frequencies for low-frequency variants. The Non-Finnish European populations from KGP3 (∼400 individuals) and gnomAD (∼56,000) produce more accurate frequency estimates for our Italian sample. However, even with gnomAD, which provides the best results, we found a few false pathogenic candidate variants.

Note that our dataset is part of gnomAD and while this is undoubtedly a bias, its effect should not be very significant as our Italian sample amounts to just 3% of the entire Non-Finnish European gnomAD sample. This is also the reason why we did not compare our sample to the Southern Europe gnomAD subpopulation, which has been recently made available and where the overlap would be much greater.

In conclusion, a dataset such as gnomAD provides a very good frequency estimate for Italian individuals and is indeed a precious resource. However, the researcher/clinician should be aware of the small likelihood that variants could be misclassified by relying on it or any other aggregated database that is not population specific. When possible, we recommend that the use of gnomAD be complemented with a more specific national or regional database of allele frequencies estimated from at least several hundred individuals. For this purpose, whole-exome aggregated frequency data from our Italian sample, together with other Italian genomic data, will be publicly available from the website of the Italian partnership called Network for Italian Genomes (http://www.nig.cineca.it/). We then investigated the potential functional role of variants in our dataset, producing three classes of putatively pathogenic variants: PTV, DMG and CLNPAT. We called these variants putatively pathogenic since they satisfy commonly used criteria in diagnosing pathogenicity, but in the absence of a known disease/phenotype it is impossible to confirm their pathogenicity.

We saw that categories of variants understood to have greater effect and likely to be enriched in pathogenic variants have higher ratios, especially those included in the PTV category. The high ratio of PTV and DMG categories is consistent with our interpretation. On the other hand, CLNPAT sports a much lower ratio but this can be explained by the fact that low-frequency variants are less likely to be found in ClinVar since they may not have been observed or diagnosed before.

We compared the amount of PTVs in our Italian individuals to UK individuals from UK Biobank (Hout *et al.,* 2019). Their definition of PTV is slightly different than the one we used. The authors used a lower allele frequency threshold of 1% instead of our 5% and distinguished two classes of PTV: a more restrictive one of variants that were PTVs in all transcripts

and a wider one of those that were PTVs in at least one transcript. Our PTV class falls in between their classes since we only looked at the canonical transcript. The median amount of PTVs in our sample was 12 considering the canonical transcript, versus 15 and 24 in the UK sample when considering all transcripts and any transcript, respectively (we only considered PTVs with allele frequency under 1% to match their choice). Thus, Italian individuals seem to carry a lower amount of PTVs with respect to UK individuals. This observation is suggestively mirrored in (Karczewski *et al.,* 2019) where they found the European population to be enriched in PTVs with respect to other continents, especially considering that Southern European are underrepresented in the gnomAD European population.

When looking at the ACMG genes, our three classes of putatively pathogenic variants behave quite differently. The lower proportion of PTVs found in ACMG genes is readily explained by the fact that ACMG genes are more intolerant to loss of function than most genes, while the enrichment of DMG and CLNPAT seems counter-intuitive. In the case of DMG variants, this is likely due to the fact that the gene or some related feature (e.g., genetic position, conservation score, etc.) is considered in the prediction, producing a positive bias in clinically relevant genes (note that missense variants themselves are not enriched in ACMG genes). In the case of CLNPAT, the likely explanation is a greater interest of the clinical community in the AMCG genes and thus a greater representation in ClinVar. We also see that the amount of DMG variants in ACMG genes is much higher than that of PTV and CLNPAT variants, suggesting that DMG variants have lower impact than PTVs in ACMG genes.

A common theme underlying the results we have presented is that the great number of rare variants produced by sequencing studies of large cohorts can be a resource in many different research and clinical contexts. Rare variation is often localized and gives information on the fine structure of a population: estimating the number of rare variants has proven a very sensitive approach in detecting the subtle demographic differences between the Italian regions. High-penetrance variants are the primary target in the study of most Mendelian diseases and thus is very import to be able to accurately estimate their frequency in the population the patient comes from. We have elaborated at length on the usefulness of population-specific frequency databases, especially when estimated from large samples in order to be more accurate in the low end of the frequency spectrum. Moreover, when working on complex diseases or phenotypes, where common variation plays an important role, sequencing data and burden tests have proven the importance of rare variants.

We believe large whole-exome or even whole-genome sequencing datasets to be very relevant to many fields in genetics, especially for highly structured populations like the Italian one. They are also instrumental to a more comprehensive approach to clinical genetics that uses population genetics as a lens to better understand the interplay between polymorphisms, genetic susceptibility and pathogenic variation. A great amount of whole-exome sequencing data is routinely produced in Italy both for clinical and research purposes. Collecting it in a comprehensive database would be a valuable resource in many research and clinical contexts.

Future tales:
back home

"Apriti mare
E lasciali passare
Non hanno fatto niente
Niente di male.

Mannarino, *Apriti cielo*

Their dead like our dead

The reader will be happy to know that we are almost at the end of this thesis. This part will be the shortest, maybe the least epic and adventurous, but undoubtedly the most human. It will be the shortest because it is about a work that Dr Carlo Robino, Dr Giovanni Birolo and I have been thinking for quite a long time, but we are now at the beginning and a great deal of work is still to be done. It will be less epic because we will not use DNA to go back in time until the dawn of men but to try and help people alive today.

From a certain point of view, the main theme here is not migration: this part will not deal with people who travelled a long way, settled down in a new country and admixed with the people living there, thus leaving their genomic footprints wherever they moved. This is a story of some people who could not make it: they left, travelled and died far from home. This is also the story of their families whose voices are echoing from afar — Eritrea, Bangladesh, Syria — claiming the bodies of their loved ones. We will not talk about people who migrate, but about people who come back home.

As you may imagine, family reunification is not easy and I will briefly talk about its winding road later on. There are many tools we can use to identify a corpse, trace its country and, ultimately, the family from which he came. One of these tools is DNA and, particularly, some specific positions in the genome whose frequencies are extremely variable among different continents, regions or even countries. Thus, finding the minimum number

of SNPs useful to understand where an individual comes from is the primary purpose of our work (which I will talk about in the next chapter).

At this time when immigrants are at the heart of political, economic and social debates, not many are thinking about the people who could not become immigrants, only because death found them before reaching their intended destination. There is no clear law compelling a country to carry out the family reunification process and this, together with the difficulties in handling such a complex phenomenon — a continuous stream of deaths — means that a functional protocol for dead identification does not yet exist. However, even though there are no legal obligations, some people are taking this problem to heart. This is precisely why this part will also be the most human.

Desperate journeys in numbers

Never before the topic of migration — and above all "immigration" — has become an all-pervasive issue for the European people. Nowadays, "migration" is foremost in all debates: from the politician who tries to win votes to the family where the "migrants are taking our jobs" refrain is a way to justify their own jobless sons. However, the volumes of talk about this matter are so much tainted by political, economic and social reasons that it is hard to think rationally about it. However, this thesis is not the proper place to discuss immigrants rights and rational thinking. I truly feel that talking about such a complex phenomenon without a deep knowledge of its historical, political, economic and sociological facets would do nothing but fuel the bar talks we are usually hearing in far more official debates. I will simply say that if we rationally think about this matter, we would discover that migratory flows are anything but a novelty for European countries (see the chapter *"Who are the Europeans?"*). Indeed, no one currently living in Europe descends exclusively from the first humans arriving there and, in turn, the first European settlers were not European at all: they were African wanderers.

That, however, does not mean that we should deny the importance and the seriousness of the migratory phenomenon and does not mean it is not a problem, either. Nevertheless, in-depth knowledge of our origins could help in better understanding what has been happening in the last decade.

Actually, the problem is real and, more than a problem, it is a real tragedy. Over the last four years, almost 800,000 migrants arrived in Europe, running from torture, abuse, persecution, misery or simply pushed by the

desire for a better life. Focusing specifically on Italy, over 300,000 migrants arrived between 2016 and 2017, while in the last two years total numbers were lower (around 30,000 migrants, *migration.iom.int,* Figure 111).

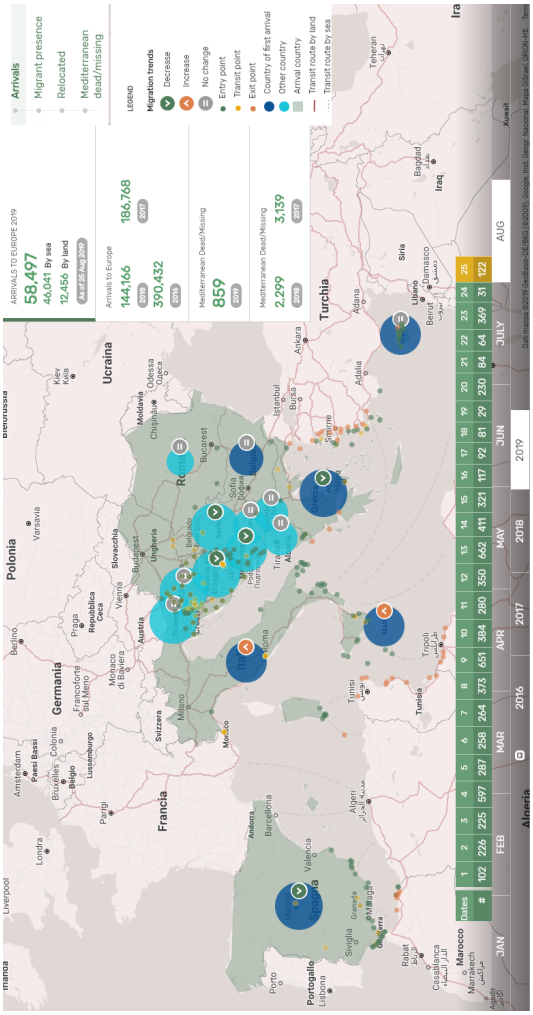


Figure 111. Migration flows to Europe registered on the 25th of August 2019. Image modified from the International Organization for Migration website.

What is worse is that many die on their way toward Europe and the number of deaths is so high that the Mediterranean Sea routes have the sad primacy all over the world (Figure 112 and Figure C.3). In the past

20 years at least 33,000 men, women and children died in Mare Nostrum — "Our Sea" was a Roman name for the Mediterranean Sea — while in just six years, from 2013 to 2019, the total number of dead amounts to 19,311 (Figure C.4). On page 355 you can find the UNHCR (Office of the United Nations High Commissioner for Refugees) report about the numbers of dead and missing migrants in the Mediterranean Sea from January to July of 2019. The majority of the remains rescued from the sea have been buried in cemeteries in Italy, Malta, Greece and Spain (Piscitelli *et al.,* 2016), but judging by the enormous portion of missing individuals (Figure C.4 at page 353), the larger cemetery is the Mediterranean Sea itself. According to this same Figure (page 353), the Central Mediterranean route (from Libya to Italy across the Straits of Sicily) represents the most dangerous migration route: between 2014 and 2017, 208 sunken ship incidents, with an average number of 44.8 fatalities per incident, have been recorded (Missing Migrants Project, 2017). We, as Europeans and Italians, stand witness to one of the greatest humanitarian tragedies of recent times.

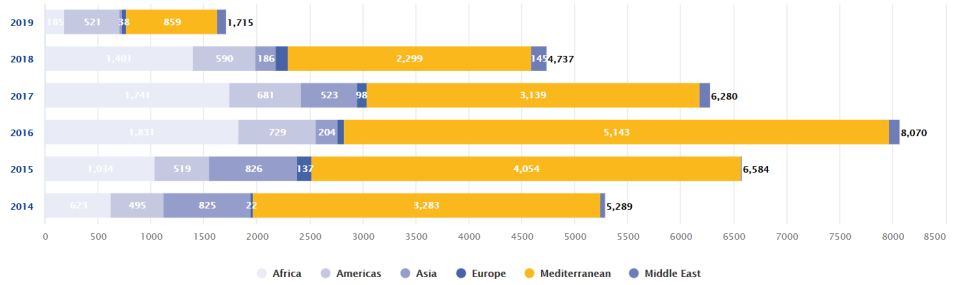


Figure 112. Recorded migrants deaths by region according to the Missing Migrants Project. Image taken from the Missing Migrants Project website.

Nameless dead

The need to identify dead people and its consequences in humanitarian, administrative and judicial contexts are universally recognised values enshrined by national and international law, including the four Geneva Conventions of 1949 (International Committee of the Red Cross, 1949) and their Additional Protocols of 1977 (International Committee of the Red Cross, 1977). Even though these obligations should be respected without discrimination, for hundreds of men, women and children, who die in Our Sea trying to

reach Europe, this fundamental right remains unfulfilled (Cattaneo *et al.,* 2015). Indeed, around the 60% of dead migrants remain without a name, with serious consequences also for their families: uncertain if their loved one is dead or alive, they could suffer from severe psychological distress or administrative, civil and social repercussions (Piscitelli *et al.,* 2016). They, faceless bodies, are buried in communal cemeteries across Southern Mediterranean countries.

As noted by Dr. Cristina Cattaneo in her book (Cattaneo, 2018), while the complex machinery of personal identification had been set in motion in a lot of casework from mass disasters, the same machinery could hardly be applied to the Mediterranean shipwrecks, without anybody objecting. What would you say if some Italian passengers died in a plane crash in another continent and the local government did not do anything to identify the corpse? Undoubtedly, and with good reason, there would be a great deal of public indignation! However, these same feelings do not arise when the one who dies is a (black-skinned) migrant (Figure 113).

Figure 113. The artwork entitled "Raft of Lampedusa", 2016. Artist: Jason deCaires Taylor.

The routine for corpses identification is usually based on a simple prin-

ciple: the comparison between *post mortem* information (PM) recovered from the corpse, and *ante mortem* information (AN) acquired from the relatives of the victims. Usually, the preferred disciplines for identification are *dactyloscopy*, *genetics* and *odontology*, which are based on comparison of fingerprints, genetic variants and dental characteristics. However, whatever data may be useful to perform personal identifications is registered so that PM data could be personal descriptors, dental data, clothing, personal belongings and DNA samples while AM data could be objects belonging to the victims which may contain DNA, clinical information, pictures of the subject and so on. Finally, a direct comparison between the genetic profiles of the victims and their putative relatives can easily detect family relationships, thus ultimately promoting the reunification.

In mass disasters, such as plane crashes, it is easy to recover information from the victims' relatives because there are passenger lists; however, in other situations, such as explosions in train stations or other mass disasters in open spaces — think, for example, about the Phuket tsunami or the Twin Towers tragedy — it is more difficult to know which people are involved. Nevertheless, thanks to the collaboration with the authorities from the involved countries it is usually possible to retrieve both AM and PM data, thus giving a name to all the victims.

On the contrary, AM or PM or both data are often missing during the identification of dead migrants. First of all, there are no official data about who was on the boats, thus making it difficult to know who to contact among the different national authorities. For this reason, it could be almost impossible to collect AM data from migrants' relatives: indeed, they could be spread all over the world, both in their countries of origin (African and Asian countries) and in the countries where the victims wanted to go. Moreover, if the corpses remain in the water for too long — in some cases, years passed between the shipwreck and the recovery — some of the identifying traits easily registered during an autopsy — think, for instance, about tattoos — are no more present. In many cases, it is barely possible to recover high-quality DNA. Unfortunately, even if just one of the two kinds of data — PM or AM — is missing, the identification process can not be successful.

The large scale of these disasters would require more forensic personnel and financial resources; however, governments of southern Europe, which are the most involved in corpses recovery and identification, are under pressure to provide aid to living migrants, thus making it difficult to justify the use of public resources on the identification of dead migrants (Piscitelli *et al.,* 2016).

The involvement of financial resources from national states and Europe

had to be preceded by the awareness of the humanitarian, social and political importance of migrant identification. This awareness began to take hold in 2014, after the 2013 disaster of Lampedusa, where 368 migrants died. From that moment on, as told by Dr Cattaneo in her book (Cattaneo, 2018), we started thinking about their dead as they were ours. The Italian Government and the Italian Academia made the first initiatives to identify these people, through the organisation of standardised procedure for data collection.

On the 18th of April 2015, another mass disaster of exceptional proportions happened in the Straits of Sicily (Figure 114). In this incident — called "the Melilli case" from the military base at Melilli where the wreck was transported — almost one thousand migrants died and over 500 bodies were recovered. Among them, Dr Cattaneo also found a 17-year-old boy who had a report card in his pocket with his grades in chemistry and physics. Many volunteers arrived there to help. They could not give them their lives back. At least, they could give them their names.

Figure 114. Picture of the ship sunken on the 18th of April 2015. Image credit: Salvatore Cavalli.

Where did they come from?

In the very frequent case where the migrants do not have their documents, one of the first essential questions is to understand where do they come from.

We know, from up-to-date statistics which are the most common nationalities of sea arrivals during 2019 (Table 21, data refer to refugees) and previous years (Figure 115), thus we can reasonably assume that these data

could also be applied to the victims of shipwrecks. Due to the evolution of the socio-economic and political situations in different countries, the nationalities of the migrants could shift from one year to another. However, their macro-areas of origin do not change, as they span a wide area: from Bangladesh and Pakistan in Western Asia to Ivory Coast in Western Africa, encompassing the Middle East, Eastern and Northern Africa.

Country of origin	Source	Data date	Population	
Tunisia		31 Jul 2019	22.2%	858
Pakistan		31 Jul 2019	16.0%	620
Côte d'Ivoire		31 Jul 2019	10.9%	421
Others		31 Jul 2019	9.5%	367
Algeria		31 Jul 2019	8.8%	339
Iraq		31 Jul 2019	8.0%	310
Bangladesh		31 Jul 2019	4.9%	190
Sudan		31 Jul 2019	4.9%	188
Guinea		31 Jul 2019	2.4%	91
Morocco		31 Jul 2019	1.8%	71
Cameroon		31 Jul 2019	1.8%	68
Egypt		31 Jul 2019	1.6%	62
Senegal		31 Jul 2019	1.6%	61
Mali		31 Jul 2019	1.3%	52
Nigeria		31 Jul 2019	1.0%	37
Ghana		31 Jul 2019	0.8%	32
Libya		31 Jul 2019	0.7%	28
Somalia		31 Jul 2019	0.7%	28
Gambia		31 Jul 2019	0.5%	20
Eritrea		31 Jul 2019	0.4%	15
Syrian Arab Rep.		31 Jul 2019	0.2%	8
Sierra Leone		31 Jul 2019	0.0%	1

Table 21. Most common nationalities of sea arrivals in Italy (since 1 January 2019). Table taken from the UNHCR website.

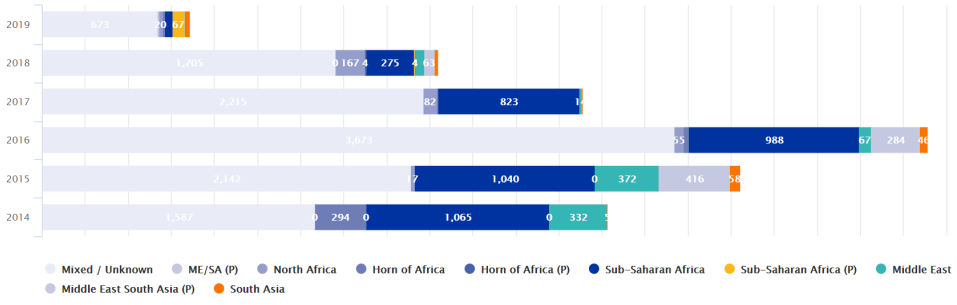


Figure 115. Deaths by region of origin over the last five years according to the Missing Migrants Project. Image taken from Missing Migrants Project website.

The geographic area from which migrants arrived in Italy following the Central Mediterranean route is too large to make contact with the families who ask about their loved ones. Thus, it is necessary to narrow it down.

Usually, forensic scientists make use of specific databases reporting the variability — in terms of frequencies — of *morphological* (skeletal biometrics, such as facial and dental markers) and *genetic* traits (Ancestral Informative Markers), which will be useful to the estimation of ancestry. I will not talk here about the morphological traits currently analysed to infer the origin of an individual, but I will focus only on DNA.

Ancestral Informative Markers

Ancestral Informative Markers (AIM) are DNA polymorphisms whose alleles have different frequencies among different populations. Thanks to this, AIMs can be used to build relatively small-scale PCR-based assays — with a consequent low cost — to estimate the geographical origins of the ancestors of an individual.

Going more in-depth in ancestry inference, Christopher Phillips, one of the leading expert on ancestry inference in the forensic field, wrote in his review (Phillips, 2015) that "*ancestry can be described as the genetic inheritance each individual carries from their ancestors, in the immediate past from their kinship, over longer periods from population members that have occupied the same place of origin*". At this point, the main principles underlying the AIM functioning should be clear. Human populations are not fully interbred, since geographic distances and geophysical barriers limit the application of a pure *random mating* model. At the same time, human populations tend to interbred with neighbour population, causing allele frequency differences among populations increase with geographic distance in a gradual way. However, some sharp differences remain: these discontinuities form the genetic clusters — detectable by software such as ADMIXTURE (see page 22 and Figure B.2) — which can be exploited to track the ancestral origin of an individual.

The wandering nature of humanity has determined the patterns of genetic differences among modern populations, with important consequences for ancestry inference. The first studies of worldwide genetic variability based on genome-wide data (Rosenberg *et al.,* 2002; Wang *et al.,* 2007; Li *et al.,* 2008) found from five to seven genetic clusters corresponding to continental division and reflecting the main migratory flows peopling the world. Due to the reasons mentioned above, such clusters are obviously not neat:

for instance, no Middle East populations showed exclusive membership to one cluster (i.e., 100% of correct assignation), thus reflecting the transit nature of that region (Li *et al.,* 2008).

Another characteristic, which we should be aware of in choosing an AIM set, is that the degree of genetic variability is not on the same scale when we compare different human populations. As a matter of fact, due to the successive bottleneck events starting from Africa and the peopling of the world, the genetic variation decreases along the Africa-Asia/Eurasia-Oceania-America chain (Wang *et al.,* 2007). In other words, the paths followed by our ancestors led to a strong divergence between African and other populations, followed by that between Eurasians and others, with East Asians showing the lowest divergence with Oceanians and Americans, due to recent founding events in these regions (Phillips, 2015). As a consequence, if we consider only the rate of genetic differences, we will find many more African-informative loci than for other group comparisons. If the goal is to distinguish Africa, Europe and East Asia, it will be harder to find markers differentiating the last two. Figure 116 shows five examples of AIM SNPs. Among them, the first two SNPs (rs4988235 and rs12075) show contrasting population specific divergences between Europe and the other two populations. Conversely, the last three are the most informative as they have fixed alleles in each group. However, the case of fixed genetic variation is the exception rather than the rule, because soft sweeps are more common than hard sweeps. For this reason, the final set of AIM SNPs must not be simply informative for a pair of populations, but also balanced so that it can distinguish with high accuracy all target populations thanks to their frequency differences.

However, as seen in the previous chapters, humans have never been static: migratory flows went on for millennia and, together with admixture, selection and genetic drift, have been muddying the water for millennia, thus making the inference of origins more challenging.

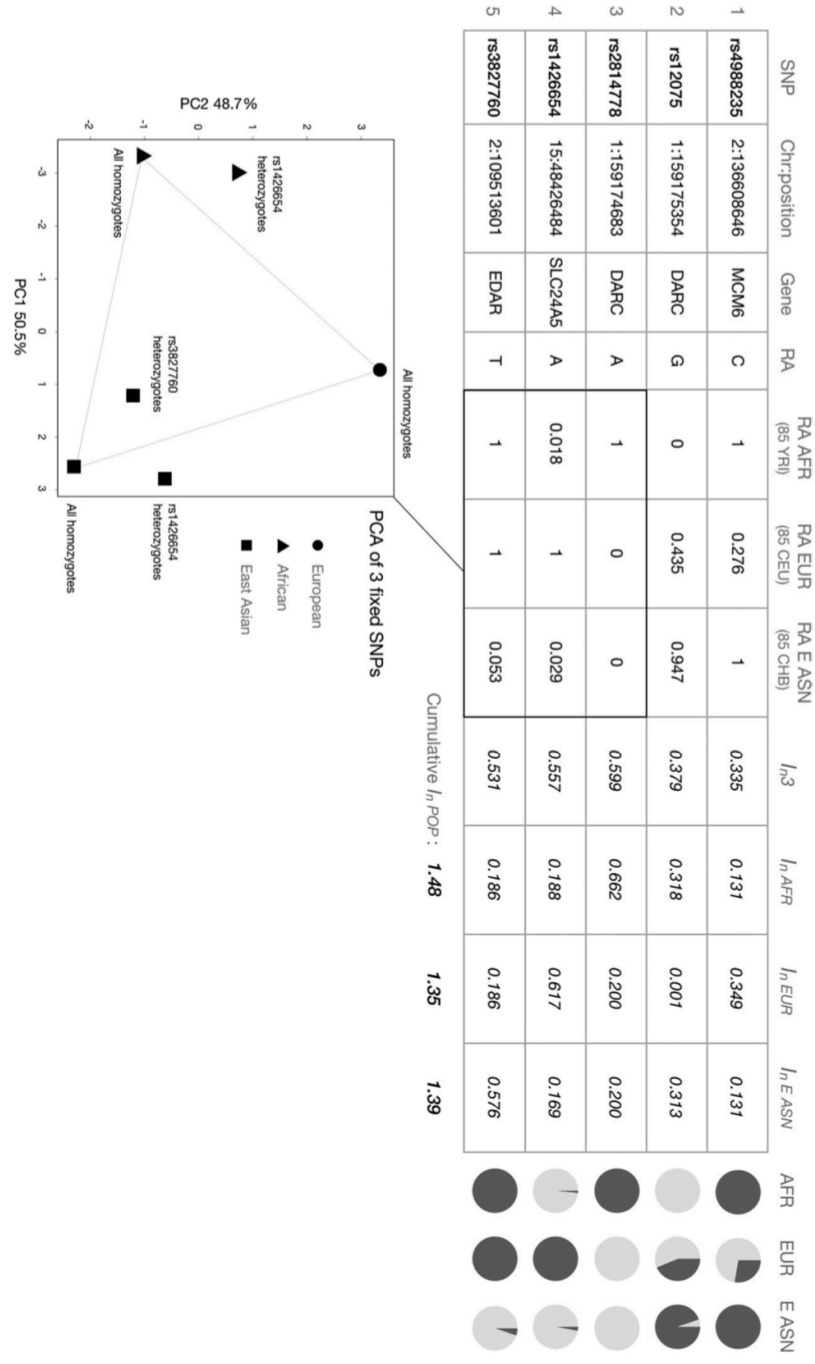


Figure 116. Examples of AIM SNPs. Image taken from Phillips, 2015

AIM for migrants

The inference of ancestry in the forensic field has proved to be useful in many applications (Phillips, 2015):

- substitute eyewitness testimony when descriptions are uncertain, unavailable or may misdirect investigators;
- aiding cold case reviews with additional data on linked profiles;
- achieving more complete identifications of missing persons or disaster victims;
- confirming donor's self-declared ancestry and therefore maintaining the accuracy of databases for STRs, Y-markers and mtDNA variation;
- refining familial search strategies highly dependent on STR allele frequency assumptions made prior to searching (Rohlfs *et al.,* 2012);
- assessing atypical combinations of physical characteristics in individuals with admixed parentage (Freire-Aradas *et al.,* 2014);
- enhancing genetic studies where forensic sensitivity is necessary, e.g., testing medical archive material or archaeological DNA (Bouakaze *et al.,* 2009).

What interests us the most is the employment of AIM in the context of mass disasters in order to restrict the area of origin of an individual. In this situation, AIMs can guide the choice of the genotype and haplotype frequencies, which will be used in case of kinship calculations (Prinz *et al.,* 2007; Rohlfs *et al.,* 2012). Then, when the victims come from poorly-studied regions, the detection of genetically homogeneous clusters allows to use the groups themselves as genotype or haplotype reference datasets for matching probability estimates (Olivieri *et al.,* 2018). Finally, when coupled with other PM data available, AIM can be used to adjust prior probabilities in likelihood ratio (LR) calculations (Goodwin, 2017).

However, in order to properly choose a set of markers to distinguish the main macro-areas of origin of migrants, we need genetic data from their origin populations. Unfortunately, many of these populations have been poorly studied (e.g., Eastern Africa) and the main public dataset of worldwide genetic variations lack such regions.

For this reason, Dr Carlo Robino started a project aiming at characterising the genetic variability of some of the Eastern African populations,

which make up a great part of the migrant arrivals and deaths. He started from the Tigray population, the fourth largest ethnic group in Ethiopia, reaching up to 4.5 million and representing over 95% of the population in the regional state of Tigray, in Northern Ethiopia (*Census-2007 Report*). Moreover, Tigray is also the major ethnic group of neighbouring Eritrea (*European Asylum Support Office*), which is one of the most common countries of origin for sub-Saharan African refugees crossing the Mediterranean Sea (Figure 115). We genotyped 252 Tigray individual for a total of 77 AIMs (46 Indel and 31 SNPs) and we compared them with the same genetic positions of sub-Saharan Africans (AFR), East Asians (EAS), Europeans (EUR) and South Asians (SAS) from 1000 Genomes Project (Auton *et al.,* 2015), and with Middle Eastern populations (MEA) from the Human Genome Diversity Panel (HGDP, Cann *et al.,* 2002; Li *et al.,* 2008). Figure 117 shows PCA and STRUCTURE analyses performed on the dataset.

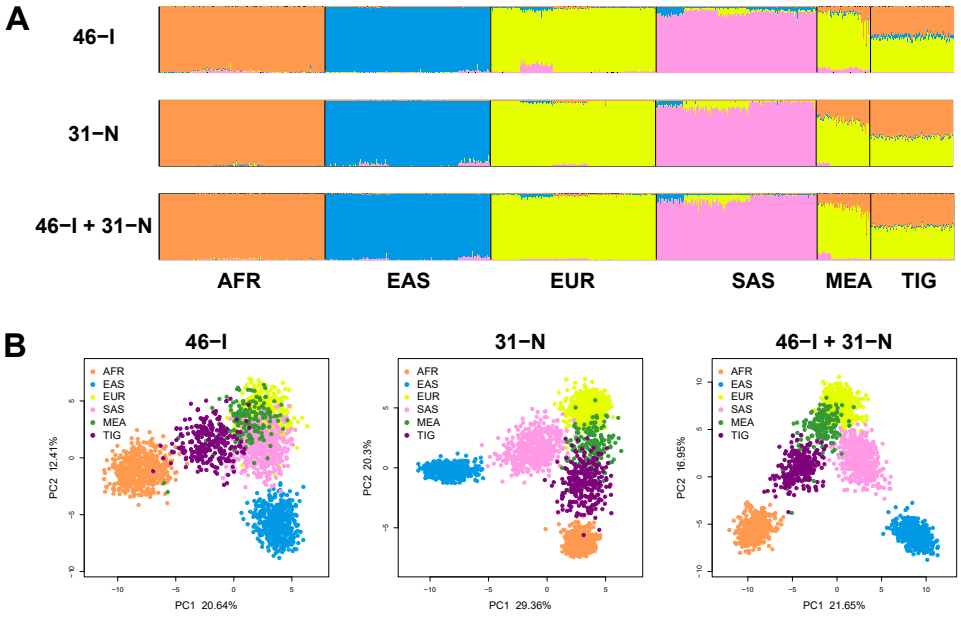


Figure 117. STRUCTURE and PCA analyses of 77 AIM. (A) STRUCTURE ancestry analysis of the Tigray study sample (TIG) and reference population groups (K:4). **(B)** Principal component analysis (PC1 vs PC2) of the Tigray study sample (TIG) and reference population groups.

The 77 AIMs come from a set which was primarily designed to separate continental ancestries, e.g., sub-Saharan Africa, East Asia, Europe and

Native America (Pereira *et al.,* 2012; de la Puente *et al.,* 2016). When we analysed their informativeness in distinguishing Eastern African populations, such as the Tigray sample, they proved to effectively discriminate both between Tigray and non-African populations (Europeans and East Asians), and between Tigray and other sub-Saharan African populations. Unfortunately, they show less power in distinguishing Tigray from Middle Eastern populations. The data I am referring to is described in a paper, which I co-authored, currently under revision in the journal *Forensic Science International: Genetics*.

In the context of European refugees crisis, the discrimination between Eastern African and Middle Eastern populations can also be supported by other genetic data, such as Y chromosome (D'Atanasio *et al.,* 2019; Iacovacci *et al.,* 2017) and mtDNA variations (Boattini *et al.,* 2013), as well as additional non-genetic data as cranial morphometric analysis (Hefner & Ousley, 2014).

However, it may be worth trying to refine the accuracy of the classification between similar and "overlapping" ancestries (e.g., Middle East, Northern and Eastern Africa), taking advantage of the high-density SNP datasets in literature (Pagani *et al.,* 2012; Henn *et al.,* 2012). For this purpose, we investigated new methods to analyse the informativeness of genetic variations for population classification.

In the next chapter, I will briefly introduce a new study we are currently working on and which will primarily focus on the main macro-areas of origin in the context of the present European refugees crisis.

"In a sea of human beings, it is difficult, at times even impossible, to see the human as being."

Aysha Taryam

Novel strategies for AIM selection

NCESTRY informative markers (AIMs) are genetic markers whose frequencies differ between populations, thus making them useful in characterizing the ancestry of individuals from their genotypes. Genotyping many thousands of AIMs allows very precise assessment of an individual's ancestry. On the other hand, smaller sets of AIMs provide less discerning power. However it has been shown that genotyping just one dozen SNPs is enough to distinguish continental ancestry: this happens because the more diverging the frequencies in the different populations, the fewer markers are needed to assess ancestry accurately.

A reason to prefer a smaller AIM set is to reduce the time and costs involved in genotyping. Moreover, the use of reduced sets of AIMs is especially appealing in forensics, where quality and availability of DNA is often compromised.

However, markers that are informative for distinguishing populations A and B may not be informative for populations B and C. Thus, the selection of markers must be tailored to the populations of interest, especially when a smaller set of markers is desired. For these reasons, we are interested in methods for selecting AIM sets that are the smallest possible providing a required amount of discerning power for a given set of populations.

The set of populations we are primarily interested in are Northern and Eastern Africa, Middle East and South Asia. For this purpose, I collected a dataset, to the which we refer as the "*migrants*" dataset, comprising liter-

ature genotypic data of these populations (Behar *et al.,* 2010; Behar *et al.,* 2013; Pagani *et al.,* 2012; Busby *et al.,* 2015; Cann *et al.,* 2002; Auton *et al.,* 2015). However, before working directly on this dataset, we decided to search for, develop and apply new strategies for AIM selection on an easier set of populations from different continents in the well-known 1000 Genomes Project (Auton *et al.,* 2015). We will refer to this dataset as the "*continent*" dataset. Thus, throughout the Methods and Results section, I will first describe the set up of the method by testing its accuracy in distinguishing five continents, then I will show its classification performances on the migrants macro-areas of origin.

Methods

Dataset

In the first part of our work on AIM selection, we wanted to start in the best position possible: classifying five different populations from different continents. In order to do so, we applied our approach on whole-genome sequencing data of BEB (Bengali from Bangladesh), CHB (Han Chinese in Beijing), GBR (British in England and Scotland), PEL (Peruvians from Lima) and YRI (Yoruba in Ibadan) populations from the 1000 Genomes Project Phase 3 (Auton *et al.,* 2015) for a total of 473 samples: the "*continents*" dataset (Table 22).

Population Code	Population Description	Super Population Code	number of samples
BEB	Bengali from Bangladesh	SAS (South Asian)	86
CHB	Han Chinese in Beijing, China	EAS (East Asian)	103
GBR	British in England and Scotland	EUR (European)	91
PEL	Peruvians from Lima, Peru	AMR (Ad Mixed American)	85
YRI	Yoruba in Ibadan, Nigeria	AFR (African)	108

Table 22. The *continents* dataset: five populations from different continents (Auton *et al.,* 2015) to be classified.

In the second part of the work, we applied the same AIM selection pipeline on a dataset built from literature (Behar *et al.,* 2010; Behar *et al.,* 2013; Pagani *et al.,* 2012; Busby *et al.,* 2015; Cann *et al.,* 2002; Auton *et al.,* 2015) and focusing specifically on four migrants'macro-areas of origin: Eastern Africa, Middle East, Northern Africa and Southern Asia: the "*migrants*" dataset. Table 23 shows the number of samples per macro-area, while table C.4 also shows sub-populations and literature references.

macro-area	samples
EAST_AFRICA	166
MIDDLE_EAST	498
NORTH_AFRICA	154
SOUTH_ASIA	809

Table 23. The *migrants* dataset: four migrants'macro-areas of origin to be classified.

Cleaning steps and descriptive analyses

Firstly, we filtered the biallelic genotypes by minor allele frequency (MAF < 0.01) and LD (using flag *–-indep-pairwise 500 50 0.2*) and we extracted the populations of interest. In order to remove all missing positions — our machine learning implementation does not allow missing values — we used the flag *–-geno 0* on the *migrants* dataset.

After these filters, we obtained 760,246 and 43,944 variants on the *continents* and the *migrants* dataset, respectively.

We created two kinds of plink files:

- *bed, bim, fam* files, which will be used to perform preliminary descriptive analyses (PCA and ADMIXTURE)
- *traw* file, which reports the allelic dosages (0/1/2/'NA' for diploid variants, 0/2/'NA' for haploid), very useful for analysing frequency differences.

As you will see from the lines of code I have reported in the Appendix section (see page 356), we organised the analyses on the dataset, as well as the pipeline for AIM selection, using Snakemake (Koster & Rahmann, 2018), a Python-based workflow engine useful for pipeline development and execution. In this environment, the workflows are specified as Snakefiles, which are collections of rules describing "recipes" on how to create output files from input files. Throughout the Methods section, I will refer to specific lines of code by using their *rule* name (I made the name of the rules in this section clickable as you can easily access the specific code in the appendix).

We started by some descriptive analyses: the visualisation of the genotypic variation in our dataset by PCAs and ADMIXTURE analyses. We filtered badly genotyped and low-quality samples, checked for Hardy-Weinberg equilibrium discrepancies and performed PCA on autosome genetic variations (*rule pca*) using PLINK 1.9 (Purcell *et al.,* 2007).

Then, we computed the ancestral allele frequencies of our populations with the ADMIXTURE software (ADMIXTURE 1.3.0, *rule admixture*), Alexander *et al.,* 2009) using $K = \{2, \ldots, 10\}$.

AIM pipeline

The most straightforward way of choosing an AIM set is to select the markers with the highest informativeness estimated by "scores", that are computed from their allelic frequencies in the populations of interest. Common

choices for these scores are the fixation index (F_{ST}, Kidd *et al.,* 2011) and Rosenberg's Informativeness Index (In, Cheung *et al.,* 2019). Other methods consist of the assessment of which genetic variants contribute the most to PCA or ADMIXTURE "clusters".

Once an AIM set is selected, a Naïve Bayes classifier is usually used to perform ancestry assignment of genotyped individuals. It is a natural choice, given that the informativeness of markers is defined in terms of their frequencies, which are the very parameters of the Naïve Bayes model. This makes the action of the classifier trivially interpretable and its training very fast, which allows us to test the accuracy of many different AIM sets easily.

Taking advantage of this, we tried a machine learning inspired approach to solve the problem of selecting an optimal AIM set for a given set of populations (Figure 118).

We are now living in an era when the amount of information produced is growing exponentially day by day, thus requiring even more powerful technique to analyse such "big data". Unfortunately, most of this data are neither understandable nor treatable by humans and or standard analytical methods. In the second half of the twentieth century, machine learning techniques evolved as a subfield of artificial intelligence and their main objective was to gain knowledge from that data in order to make predictions. In other words, machine learning algorithms enable computers to understand, capture and learn information and patterns from big data. The so-acquired knowledge can be used to build predictive models on new data.

The human genome, with its 3 billion base pairs and the multiple layers of information, is one of the most striking examples of big data. Actually, the biological questions investigating the complexity and sheer amount of genomic information could really benefit from machine learning methods (Zou *et al.,* 2019). Therefore, given the potential of machine learning techniques in genomic fields, we think that these methods could represent one of the best choices to face the problem of AIM selection.

We will try to solve the problem of AIM selection for classification using *supervised learning*, which is a category of machine learning methods that learn a model from labelled training data and then makes predictions about new unseen data. The term *supervised* refers to the fact that the set of samples used for training the model (train set) is labeled.

Classification is a subcategory of supervised learning: here, the aim is to predict a label, instead of a quantity (*regression*). When the categorical class labels to predict are more than two, we are dealing with a *multi-class classification* problem.

Therefore, we will develop a machine learning-based strategy to solve a

supervised multi-class classification problem: the AIM selection to classify individuals into different human populations. The direct acyclic graph of our pipeline is shown in Figure 118, which will be explored in the next section.

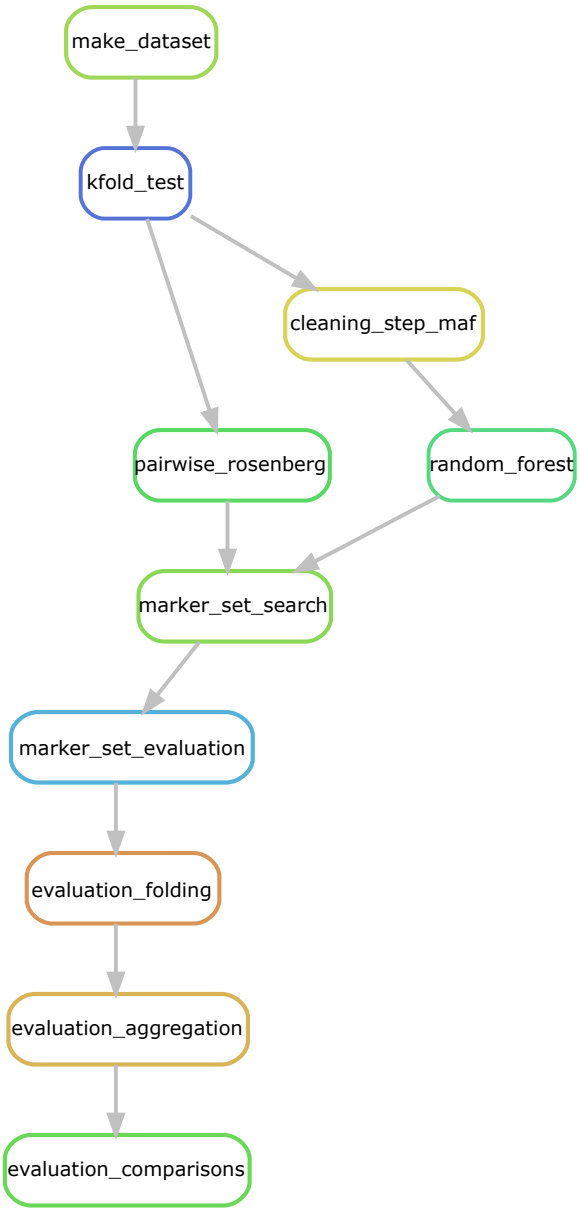


Figure 118. Graph of the AIM selection pipeline.

Cross-validation

Even if machine learning methods are powerful, there they are not magic: they submit to the main law of data — garbage in, garbage out — and need to be explored with some cautions. One of is the need for a **cross-validation** strategy in order to prevent *overfitting*. In fact, it is always advisable to validate a machine learning model to make sure that it will be able to generalize well to unseen data — overfitting means that the method works accurately in the training data, but performs badly when applied on unseen data (Figure 119). To check for this possibility, we could split the dataset in a *training* and a *test* set. In this case, overfitting is easy to detect: it the model performs much better on the training set than on the test set, probably, we are overfitting.

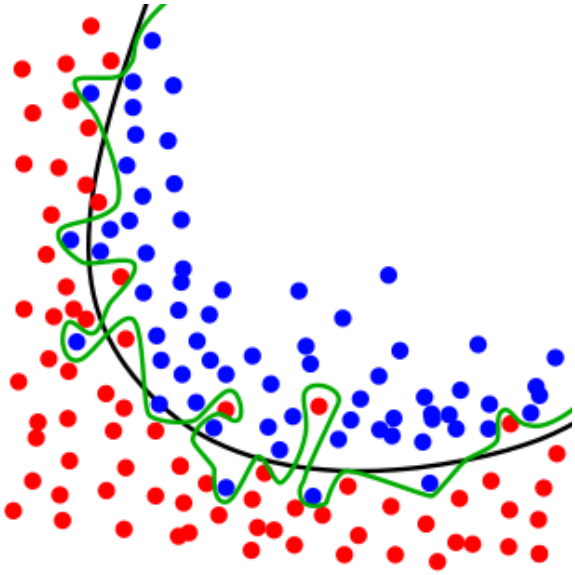


Figure 119. Overfitting example. Overfitting happens when, for instance, a model has learned the noise instead of the signal. In this case, it is considered "overfit" because it fits the training dataset but has poor fit with unseen data. In this image, the black line fits the data well, while the green line is overfit.

However, a first train-test split is not sufficient to control the stability of the model. In fact, when we need to choose between different settings of a model — the "hyperparameters" — or among many models, there could be still the risk of overfitting on the test set because the parameters or the

different model could perform well on the test simply due to the intrinsic characteristics of the test set. For this reason, another part of the dataset has to be left out, the *validation* set: we will test the hyperparameters and the different models on this part and then we will do a final evaluation of the best model's performances on the test set.

On the other hand, if we leave out both the test and the validations sets often there is not enough data to train the model. To avoid this inconvenience, a k-fold cross-validation can be employed. This technique consists of randomly splitting the training set into k folds without replacement, where $k-1$ folds are used for the model training and one fold is used for testing. Then, this splitting is repeated k times, thus obtaining k models and k performance estimates.

The k-fold cross-validation is the first step of our analysis (node *kfold_test* in Figure 118 and Figure 120A).

In this phase, we split the dataset into $k = 5$ folds, shuffling the samples 10 times and thus obtaining $5 * 10$ different reshuffling dataset. Since we needed to evaluate the classification performances in each category (i.e., population), we used the scikit-learn function *StratifiedKFold*, which makes the folds by preserving the percentage of samples for each class. Then, we used the same function again to split the $k - 1$ sets into $k = 4$ folds: 3 folds will be used as training (285 samples), while 1 fold will be used as validation (95 samples). The code used for the stratified k-fold is in the rule *kfold_test*.

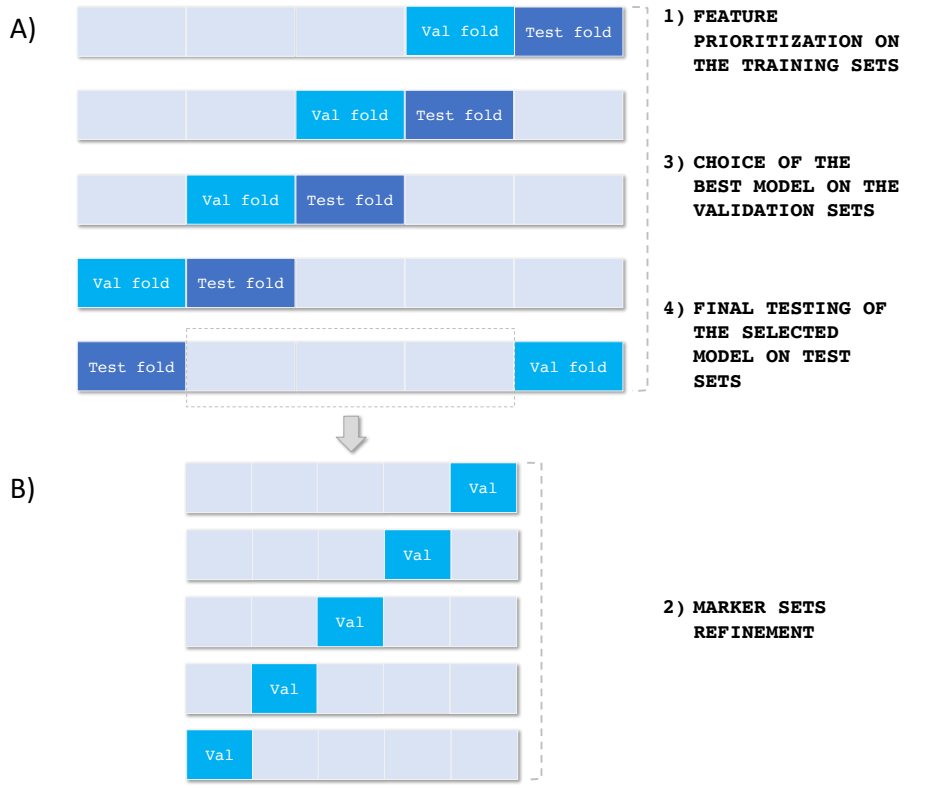


Figure 120. The cross-validation strategy of our work. (A) The initial 5-fold cross-validation generate 5 different training, validation and test sets (since it was performed shuffling the samples in each fold 10 times, we have 50 different groupings). Then, the training sets are used for feature prioritization (1). **(B)** A new 5-fold cross-validation is applied on the training set generating 5 groups, of which one is a new validation. This step is needed for marker set evaluation (2). The validation set left out in step A is used to compare the performances of the different models and markers sets (3), while on the test set we will finally evaluate the classification accuracy of the best model and the best marker set (4).

Feature prioritization

In the second step of our model, we used two different ways of understanding which features — the genetic markers are the features of our model — are contributing most to our prediction variable: Rosenberg's informativeness

for assignment (Rosenberg *et al.,* 2003) and the feature importance from a random forest (Breiman, 2001). These methods are reported in Figure 118 at nodes called *pairwise_rosenberg* and *random_forest*, respectively.

Rosenberg's informativeness for assignment

The **Rosenberg's informativeness for assignment** (I_n) is a metric introduced by Noah Rosenberg and colleagues in 2003 (Rosenberg *et al.,* 2003) in order to compute the amount of information that genetic markers can provide about the ancestry of an individual. Table C.3 shows some metrics commonly used for understanding the genetic markers' informativeness for ancestry inference.

The Rosenberg's informativeness for assignment with K populations is defined as:

$$I_n = \sum_i (-p_{i,j} \ln p_{i,j} + \frac{1}{K} \sum_k (p_{i,j,k} \ln p_{i,j,k}))$$

where $p_{i,j}$ is the frequency of the ith allele at the ith locus in all populations considered and $p_{i,j,k}$ represents the same value in the kth population. For the reasons discussed above, a multiple populations I_n computation, when one of the target is African, will inevitably detect much more African-informative loci than for the other populations. For this reason, a better strategy is to compute a pairwise I_n index. In this case, it is defined by:

$$I_n = \sum_i (-p_{i,j} \ln p_{i,j} + \frac{1}{2} (p_{i,j,k} \ln p_{i,j,k} + q_{i,j,k} \ln q_{i,j,k}))$$

where $q_{i,j,k}$ is the frequency of the same ith allele at the same ith locus in another population.

We applied this formula (node *pairwise_rosenberg* in Figure 118) to computed the pairwise I_n for all the $\frac{K(K-1)}{2}$ (10) population comparisons in the training set (Figure 120B). Then, we sorted each I_n distribution in descending order (the more a marker is informative, the lower will be its index) and we iteratively took the first marker in each of the 10 distribution until a set of 200 unique markers was collected (repeated markers were added only once).

The code implementing this computation is reported in the rule *pairwise_rosenberg*.

Random forest

The second method for feature prioritization is the feature importance computed from a **random forest**, which is an ensemble machine learning technique. The advantage of an ensemble technique is that it combines multiple individual learning models to produce an aggregate model, which is much more powerful than its parts. In our case, the learning model is a *decision tree* and when many decision trees are combined together operating as an ensemble, we obtain, like its name implies, a random forest. Decision trees classifier are attractive models because their results are easy to interpret. Each tree uses a tree-like model of decisions, which breaks down the data by making decisions based on asking a series of questions (Raschka, 2015). In our case, many features — the SNPs — are evaluated at each node by looking at the genotype frequencies across all the individuals (Figure 121), then the feature which classifies best is chosen at that node. Thus, based on the train set, the decision tree learns a series of SNPs to explore in order to make a classification and then it can infer the class labels of new samples. However, decision trees are prone to overfitting: in fact, it may happen that they choose and learn very specific questions, which work well only on the training dataset, but do not work well on unseen data. However, by combining different trees into a forest, this effect is mitigated: a forest can average out their individual mistakes to reduce the risk of overfitting while maintaining strong prediction performances.

Random forests work through four steps (Raschka, 2015):

1. randomly choose n samples from the training set without replacement (bootstrap sample);
2. train a decision tree from the bootstrap samples and at each node:
   (a) randomly sample d features (the SNPs) without replacement;
   (b) use the feature providing the best split to split the node (the selection of the best features at each node is made evaluating, for instance, the maximization of the information gain.)
3. repeat the steps 1 and 2 t times.
4. combine the prediction of each tree in order to assign the class label by majority vote.

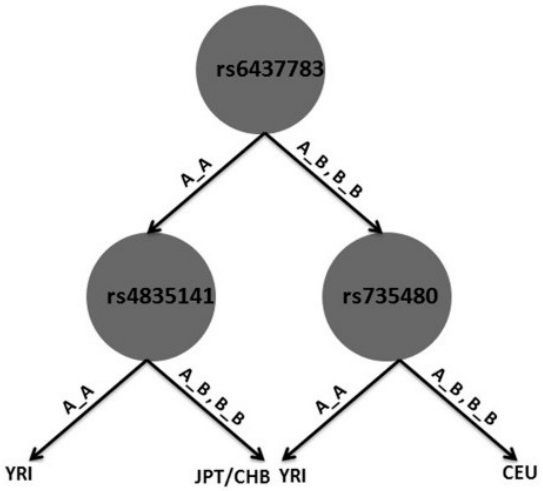


Figure 121. Example of a decision tree working with genetic variation data. Image taken from Hajiloo *et al.,* 2013.

In this phase, we are not interested in the classification power of random forest but instead in its ability to infer the feature informativeness. In fact, at each node, the trees of the forest divide the dataset into k groups by choosing the feature which allows the "purest" groups possible. The random forest can average the "importance" of each feature across each tree simply by looking at the so-formed groups "purity": at the end of this step, each feature will have an "importance" score. Seen in these terms, the more important the feature is, the more that feature will decrease the impurity at the node.

As previously said, we are not interested in the random forest as a classifier and, for this reason, we did not search for the best hyperparameters as usually done with a random and a grid search. Instead, our primary aim was to obtain an importance score for each feature. For this reason, we wanted to build a classifier able to explore all of our features — the SNPs — and compute a score for all of them (the code for building, training and fitting the random forest classifier is in the rule called *random_forest*).

We used the scikit-learn class *RandomForestClassifier* to build a random forest classifier (node *random_forest* in Figure 118) with 10,000 trees (*n_estimators parameter*) and a maximum depth of 10 for each tree (*max_depth parameter*). In order to speed up the process, we "helped" the random forest in its work of finding the most informative features, by removing the surely uninformative genetic variants, i.e., the variants with a MAF (minor allele

frequency) too low for being shared across all individuals of a population and thus for effectively contributing to a population classification. Even though this consideration is kind of obvious, we demonstrated it by plotting the feature importances across three MAF bin. In Figure 122, we see that among the features with the highest importances there are no low-frequency variants (with MAF < 0.1).

Thus we iteratively MAF-filtered the 50 training sets, retaining only those SNPs whose maximum MAF value was greater than our MAF thresholds (0, 0.20) in at least one population (node *cleaning_step_maf* in Figure 118). Table 24 shows the means and standard deviations of the remaining SNPs with different MAF thresholds, across the 50 groupings.

In conclusion, we trained a random forest on the training sets filtered with 0 and 0.20 MAF, thus obtaining $5 * 2 * 10$ list of features sorted by their informativeness scores (5 because we have a 5-fold cross-validation as in Figure 120A, 2 because we have 2 MAF thresholds and 10 because the groupings were repeated shuffling the samples 10 times).

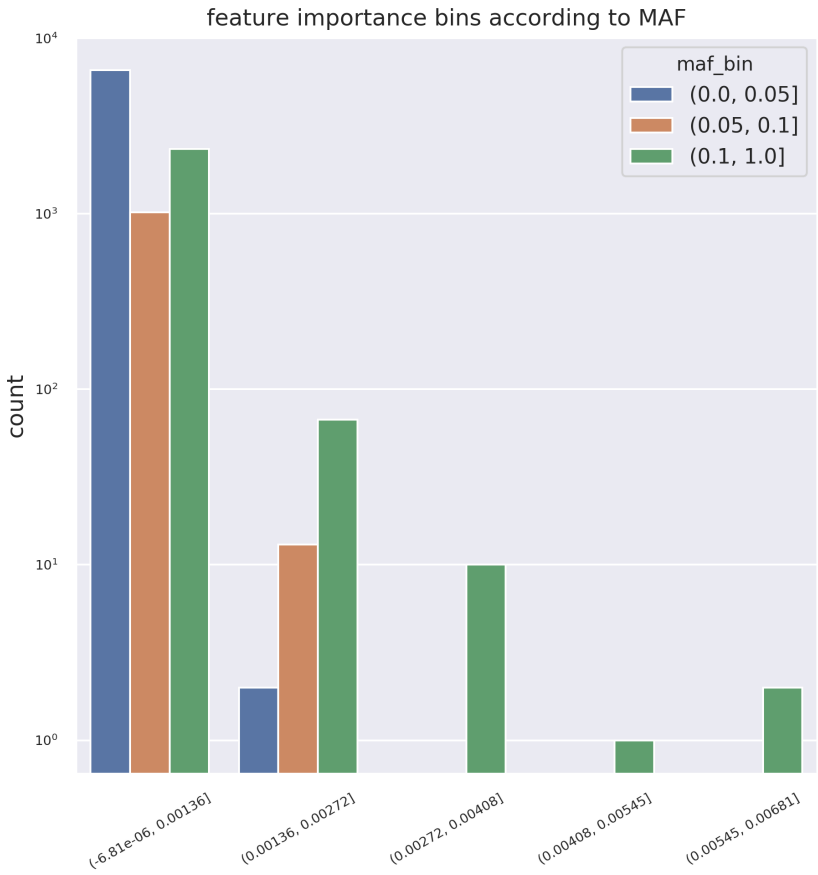


Figure 122. Feature importance bins according to three MAF subdivisions. The feature importance was computed on 10,000 SNPs and selecting the best random forest classifier with the function *GridSearchCV*.

MAF	mean	sd
0	760,181	0
0.01	759,304.54	36.44
0.02	750,301.62	203.09
0.03	723,752	535.60
0.04	667,557.64	1205.50
0.05	598,767.26	1657.05
0.06	541,080.96	1814.28
0.07	483,435.12	5819.97
0.08	433,172.06	1600.64
0.09	397,225.9	1173.61
0.1	366,895.74	1394.15
0.11	336,027.78	1298.59
0.12	316,925.74	1245.10
0.13	296,386.9	358.67
0.14	278,906.48	1555.99
0.15	263,643.28	1054.97
0.16	252,628.66	1014.05
0.17	240,553.66	1607.49
0.18	229,381	453.89
0.19	220,458.68	909.38
0.2	212,829.4	887.06

Table 24. Mean and median of the number of SNPs left after MAF filtering across the training sets. MAF thresholds range from 0 (all SNPs retained) to 0.2 (only SNPs whose maximum MAF value in a population was greater than the MAF thresholds were retained).

Marker set refinement

Now we have come to the main aim of the work: discover a method able to find sets of genetic markers for population classification that are both accurate and small. We performed the **marker set evaluation** step using a 5-fold cross-validation strategy (Figure 120B). Basically, we started from the informativeness-sorted set of variants obtained from the training set in Figure 120A (one set of variants is a result of the I_{n} computation and the other derived from the random forest), then we selected many sets of markers and we trained a classifier on 5 different training sets and finally we evaluated the performances on the union of the validation sets (Figure

120B).

Thus, in order to achieve these tasks, we need three ingredients:

- a metric measuring the performances of the model,
- different sets of genetic markers to compare,
- one or more classifiers to test the marker sets.

Concerning the first point — a performance metric — the simplest possible one is the *classification accuracy*, which is the ratio of the number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

One limitation of this metric is that it works well only when all classes to predict are balanced in terms of number of samples per class. However, as shown in table 22 and 23 we are not in this situation. Thus we used the *balanced accuracy* implemented in *sklearn*, which accounts for classes that are unevenly represented in the dataset.

The second ingredient of our method is the different sets of genetic markers. However, even though we are interested only in "small" marker sets of length, the possible number of sets is too high for being able to explore them all. For instance, the total number of sets of $k = \{1, \ldots, 15\}$ elements (a suitable number for the continent dataset) selected from N genetic markers is defined as:

$$\sum_{k=1}^{15} \binom{N}{k} = \sum_{k=1}^{15} \frac{N!}{N!(N-k)!}$$

For this reason, we tried a strategy which creates growing sets of markers, by starting from the first variant in the informativeness-sorted lists and adding one marker at every step, until we reach the 0.99 accuracy value (node *marker_set_search* in Figure 118).

Specifically, the steps are:

1. take the first marker of the list — the most informative according to the prioritization strategy (I_n and random forest feature importance) — and train a classifier using only this marker. Then test its accuracy on the union of the 5 validation sets. Let the accuracy of the first marker be n.

2. Add the second marker to the list, train and test the classifier on the validations. If the accuracy increases of more than a fifth the deviation from the final value, thus reaching a value of $n+\frac{0.99-n}{5}$, then the second marker is added to the list, otherwise, the search goes on to the next marker.

3. Add the third marker to the list and repeat from step 2 until the requested accuracy value (0.99) is reached.

Importantly, given that the accuracy when a new marker is added depends on the other markers already present in the list, it is possible to add a marker that was previously evaluated and discarded in a previous iteration. The rule for marker sets refinement is *marker_set_search*

A limitation of this method is that each marker set necessarily contains the first marker in the initial sorted list. Even though this variant has a high classification power, does not mean that the second variant, when combined with another markers set, could exceed the performance of the first. On the other hand, the constraint about the improvement of the accuracy is beneficial to avoid adding a marker which allows a negligible increase of the performances (for instance, it could happen when you add markers in LD).

However, this is only one of the many methods for efficiently finding subsets and we plan to explore other strategies for exploring the huge space of possible marker sets.

The third ingredient is a classifier, which will be necessary to ultimately discriminate the individuals among the five populations, using the random sets of marker. At this stage of the work we tried three different classifiers:

- Naïve Bayes
- HWE Naïve Bayes (i.e., assuming Hardy-Weinberg Equilibrium)
- Random Forest

We have already said what is a random forest and how it works at page 279. While before we used this machine learning method to evaluate the classification informativeness of the variants, now we use it as a classifier.

The second classifier is called "Naïve Bayes" — because it is based on Bayes' theorem — and it is a probabilist machine learning method useful for solving classification problems. It is pretty simple and it is outlined as follows.

Given a vector n of features $X = (x_1, \ldots, x_n)$ and a class C_k, if we want to know $P(C_k|x_1, \ldots, x_n)$, which is the probability to be in the class C_k

given the features $x_1, \ldots, x_n$, we can find this probability thanks to Bayes' theorem with:

$$P(C_k|x) = \frac{P(C_k, x)}{P(x)}$$

In our case, n is the number of different classes (i.e., populations) and X is an individual taken from a population $P_i \in \{P_1, \ldots, P_n\}$. We do not know to which populations the individual X belongs to, but we know the allelic and the genotypic frequencies of all the populations. Thus, we only have to observe the genotypes of the individual X in his genetic markers $M_1, \ldots, M_k$ (Table 25). Here, f_{ijAA} is the frequency of the homozygous genotype AA in the population i at the marker j.

	M_1			M_2			...			M_k		
	AA	Aa	aa	AA	Aa	aa	...	...	...	AA	Aa	aa
Pop_1	$f_{1,1,AA}$	$f_{1,1,Aa}$	$f_{1,1,aa}$	$f_{1,2,AA}$	$f_{1,2,Aa}$	$f_{1,2,aa}$	...	...	...	$f_{1,k,AA}$	$f_{1,k,Aa}$	$f_{1,k,aa}$
Pop_2	$f_{2,1,AA}$	$f_{2,1,Aa}$	$f_{2,1,aa}$	$f_{2,2,AA}$	$f_{2,2,Aa}$	$f_{2,2,aa}$	...	...	...	$f_{2,k,AA}$	$f_{2,k,Aa}$	$f_{2,k,aa}$
...	...	...	...	...	...	...	...	...	...	...	...	...
Pop_n	$f_{n,1,AA}$	$f_{n,1,Aa}$	$f_{n,1,aa}$	$f_{n,2,AA}$	$f_{n,2,Aa}$	$f_{n,2,aa}$	...	...	...	$f_{n,k,AA}$	$f_{n,k,Aa}$	$f_{n,k,aa}$

Table 25. Examples of genotypic frequencies.

The probability we want to solve — $P(X \in P_1|M_1, \ldots, M_k)$ — can be written, thanks to Bayes' theorem, as:

$$P(X \in P_1|M_1, \ldots, M_k) = \frac{P(X \in P_1|M_1, \ldots, M_k)}{P(M_1, \ldots, M_k)}$$

Markers are assumed to be conditionally independent (we LD-pruned our dataset), thus we have:

$$P(X \in P_i|M_1, \ldots, M_k) = \frac{P(X \in P_i)P(M_1|X \in P_i) \ldots P(M_k|X \in P_1)}{\sum_{j=1}^{n} P(X \in P_j)P(M_1, \ldots, M_k|X \in P_j)}$$

We know the quantity $P(M_j|X \in P_i)\ \forall j$, since:

$$\begin{aligned} P(M_j|X \in P_i) &= f_{i,j,AA} \quad if M_j = AA \\ P(M_j|X \in P_i) &= f_{i,j,Aa} \quad if M_j = Aa \\ P(M_j|X \in P_i) &= f_{i,j,aa} \quad if M_j = aa \end{aligned}$$

Moreover, all populations have a uniform priori probability, since they get the same chance of being the population of origin of the individual X:

$$P(X \in P_i) = \frac{1}{n} \ \forall i \in \{1, \ldots, n\}$$

At this point, we have all the components to solve the Naïve Bayes classification problem. However, we introduced two little modifications, in order to adapt the computation to our genotypic environment. The first is the Laplace smoothing correction. Basically, it consists of adding a quantity to the observed allele or genotypic frequency in order to avoid the worst case where none of the individuals in the training set shows that particular allele or genotype, i.e., the allele or the genotypic frequency is 0. Let n be the number of individuals, thus the allele frequency for the allele k is:

$$\frac{k}{2n}$$

With the Laplace smoothing correction we have:

$$\frac{k+1}{2n+2}$$

The second modification has been added only in the case of a Naïve Bayes classifier assuming Hardy-Weinberg Equilibrium (HWE Naïve Bayes). In this case, instead of using the genotypic frequency as they appear in the dataset, we compute the allelic frequencies from the observed genotypes and then estimate the genotype frequencies assuming the Hardy-Weinberg Equilibrium:

$$f(AA) = f(A)^2, \quad f(Aa) = 2f(A)(1 - f(A)), \quad f(aa) = (1 - f(A))^2.$$

In conclusion, we used the Naïve Bayes classifier both in its standard form and with the Hardy-Weinberg Equilibrium assumption. For clarity, the steps in the "normal" Naïve Bayes classifier are:

1. genotypic frequencies computation,
2. Laplace smoothing correction,
3. final genotypic frequencies.

While the steps in the HWE Naïve Bayes classifier are:

1. genotypic frequencies computation,

2. allele frequencies computation,
3. Laplace smoothing correction,
4. Hardy-Weinberg Equilibrium transformation,
5. final genotypic frequencies

In conclusion, we tested three different classifiers, training them on a training set and evaluating their performances on the correspondent validation set. At the end, their accuracy will be computed on the union of the validations.

Final testing of the models

At this point of the analyses, we needed to compare the different models we built between each other. For the sake of clarity, it could be useful to summarize the different combinations of methods we employed:

- **prioritization strategy**
  - I_{n}
  - Random Forest with 4 MAF filters (0, 0.05, 0.10, 0.20)
- **classifier**
  - Naïve Bayes
  - HWE Naïve Bayes (i.e., assuming Hardy-Weinberg Equilibrium)
  - Random Forest

We tried all the combinations between prioritization and classification strategies 50 times (because we did a 5-fold cross-validation by shuffling the samples 10 times).

Here we truly realized that "if your strategy is too complicated to be easily explained and understood, maybe it is not a good strategy" or, maybe, I could simply improve my English skills!

However, we tested each marker set found with a specific pipeline (e.g., prioritization made with a random forest with a MAF filter of 0.05 and a classification with a HWE Naïve Bayes), by fitting its classifier on its corresponding training set and markers and then predicting the labels of the corresponding validation and test sets (rule *marker_set_evaluation* and node *marker_set_evaluation* in Figure 118). Again, in order to compute accuracy values on the highest number of samples possible, we put together the

validations (rule *evaluation_folding* and node *evaluation_folding* in Figure 118), while the rule called *evaluation_aggregation* is used to combined the 10 different sample shuffling (node *evaluation_aggregation* in Figure 118).

Then, we selected the best models, in terms of genetic variants and method, and we finally tested on the corresponding test set.

Results

The aim of our work is to find small (minimal) set(s) of genetic markers that can classify individuals into one of multiple ancestries. As mentioned above, the need of finding minimal sets is a technical one, in order to face the often low-quality DNA material available in these cases (e.g., mass disasters) and reduce genotyping time and costs. The more the target populations are diverse, the less markers will be necessary to correctly classify their individuals.

Throughout this section, I will first describe the analyses performed on the "easier" problem of continents classification (the *continents* dataset), and then move on to the more challenging problem of distinguish the main macro-areas of origin of the current European refugees crisis (the *migrants* dataset).

The *continents* dataset

Starting from the continent separation problem on 1000 Genomes populations, we did some descriptive analyses of the samples. A first visual inspection of the dataset shows that its populations are relatively well separated. In the PCA (Figure 123), all continents appear distinct with the exception of the American population PEL (Peruvians from Lima) and BEB (Bengali from Bangladesh). However, a better separation is obtained when the third PC is considered.

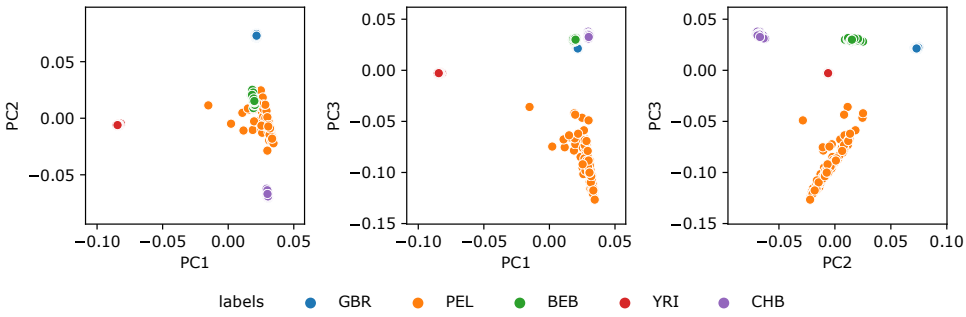


Figure 123. PCA analyses of 473 samples from the *continents* dataset. The samples belong to five different populations: GBR (British in England and Scotland), PEL (Peruvians from Lima), BEB (Bengali from Bangladesh), YRI (Yoruba in Ibadan), CHB (Han Chinese in Beijing).

Figure 124 shows that the ancestral allele frequencies computed with

the ADMIXTURE software have little overlap among the five populations. A further confirmation of this is that $K = 5$ is the number of ancestries with the minimum cross-validation error. It appears that we are in the best situation possible for population classification.

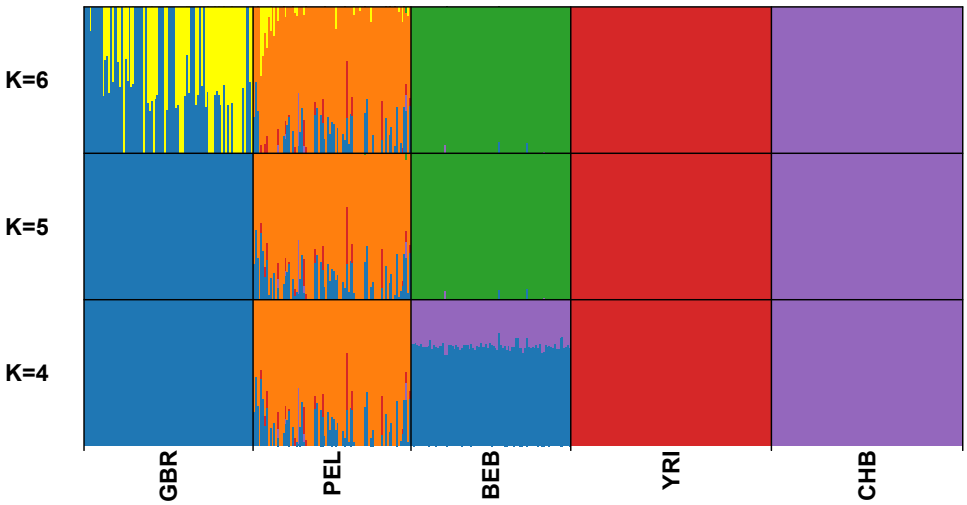


Figure 124. ADMIXTURE analysis on 473 samples from the *continents* dataset. The samples belong to five different populations: GBR (British in England and Scotland), PEL (Peruvians from Lima), BEB (Bengali from Bangladesh), YRI (Yoruba in Ibadan), CHB (Han Chinese in Beijing).

At this point we applied our AIM selection pipeline and we compared the classification performances with sets of up to 15 markers: these sets were selected with the three strategies for feature prioritization (I_{n} and the random forest without/with 0.20 MAF filters) and classified with three methods for classification (Naïve Bayes, HWE Naïve Bayes and random forest). Results are shown in Figures 125, and C.6.

The first important result is that the feature prioritization with I_{n} informativeness is better for population classification. In fact, I_{n} allows to obtain the 100% of correct assignations on the training sets and accuracy values up to 95% on both the validation and test sets with 11 markers, while exceeding 96% with 15 markers (Figure 125A). Moreover, the accuracy values grow faster, thus reaching higher performances with less markers, which is precisely our aim (Figure 125A).

Conversely, the performances using marker sets sorted with the random

forest feature importance are lower and we can not appreciate differences according to the MAF filters applied (Figures 125B and C.6).

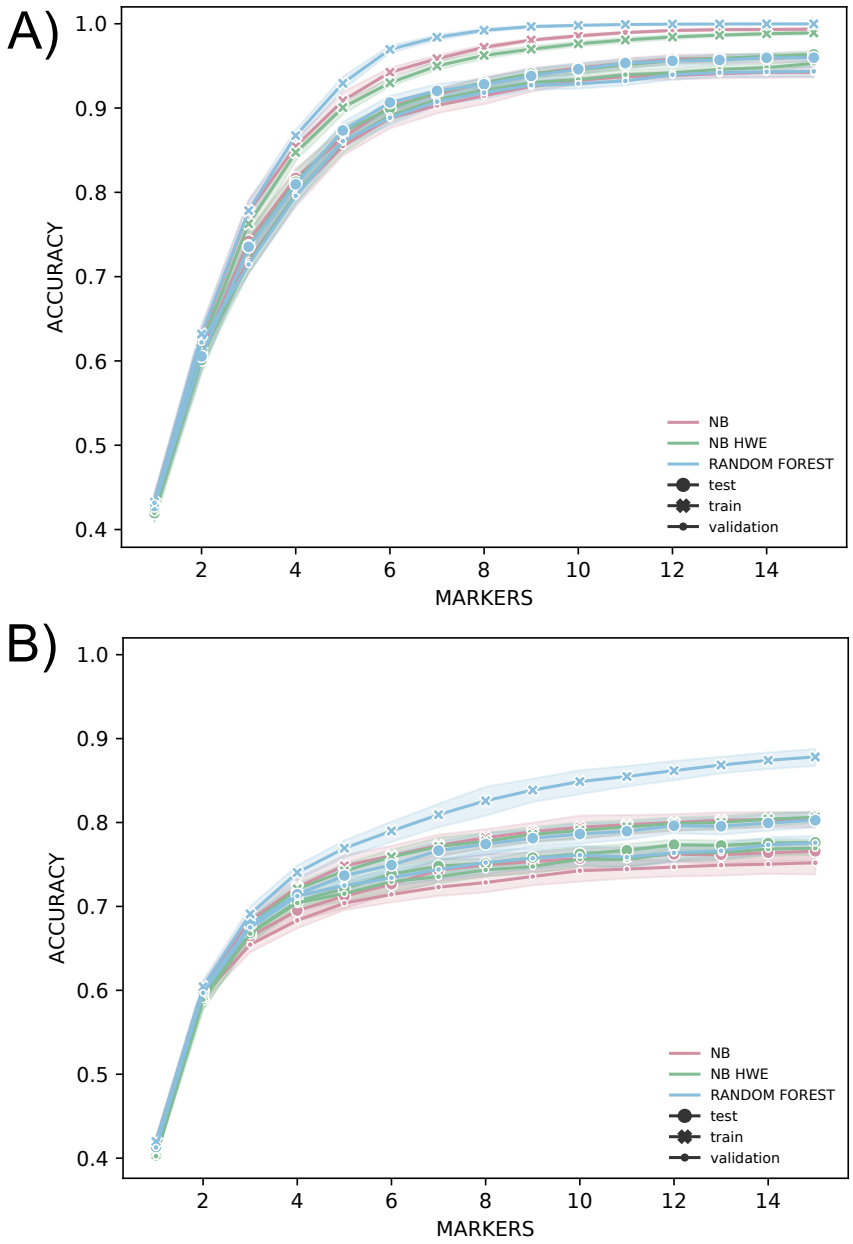


Figure 125. Accuracy comparison between different feature prioritization strategies on the *continents* dataset. Classification accuracy using I_{n} **(A)** and random forest feature importance without MAF filtering **(B)** as prioritization strategies according to three different classifiers.

When we come to the evaluation of the three methods for classification (Naïve Bayes, HWE Naïve Bayes and random forest), we see that all classifiers show some overfitting, i.e., their performances are higher on the training than on the validation and test sets (Figure 125), however the plot B shows that this phenomenon is more pronounced for random forest classifier.

According to these results, the combination I_{n} prioritization - HWE Naïve Bayes shows the higher classification accuracy. Thus, we computed its the confusion matrix, which shows the correspondence between real and predicted assignations for each pair of populations (Table 26). As expected, the most misclassified population is PEL (Peruvian from Lima), where some individuals are predicted to be Bangladeshi (BEB) or British (GBR); however, the error rate remains low.

	BEB	**CHB**	**GBR**	**PEL**	**YRI**
BEB	89.0	0.0	0.0	0.0	0.0
CHB	0.0	104.5	0.4	0.1	0.0
GBR	0.1	0.1	93.7	0.1	0.0
PEL	10.8	2.5	5.9	65.7	0.1
YRI	0.0	0.0	0.0	0.0	110.0

Table 26. Confusion matrix for 15 markers chosen with I_{n} prioritization and classified with HWE Naïve Bayes on the *continents* dataset. This matrix corresponds to the classification results of the green curve on validation sets in Figure 125A.

From this preliminary results on the "easy" problem — the continents classification — we can conclude that I_{n} metric is better in detecting those markers useful for population classification, while we do not see appreciable differences among the three classifiers' performances.

The *migrants* dataset

As for *continents*, we begin the analyses on the *migrants* dataset with PCA and ADMIXTURE. In this case, making out the four macro-areas is not as straightforward (Figures 126 and 127). Due to geographical proximity and historical reasons, some of them share, in variable proportions, the same ancestral components and this is a challenge for classification. Examples of this fact are the green portion in African areas (Northern and Eastern Africa) and in the Middle East or the purple one shared by Middle Eastern populations and some groups from Southern Asia, such as the Pakistani individuals.

Moreover, from the PCA we can see that the macro-areas in the *migrant* dataset are not well separated, with the some individuals labelled in one class and clustering in another (Figure 126).

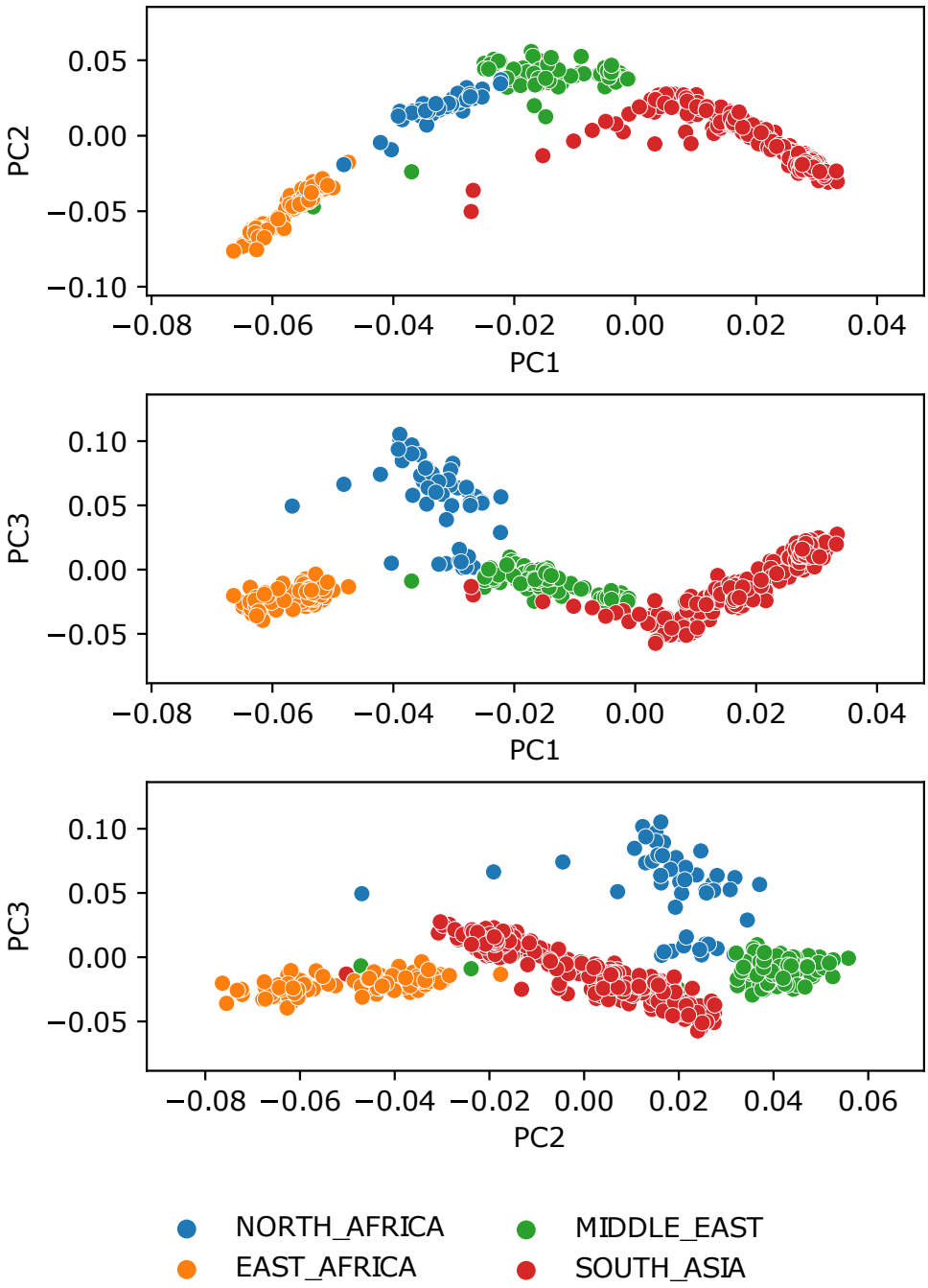


Figure 126. PCA analyses of 1088 samples from the *migrants* dataset. The samples belong to five different macro-areas: Northern Africa, Eastern Africa, Middle East and Southern Asia.

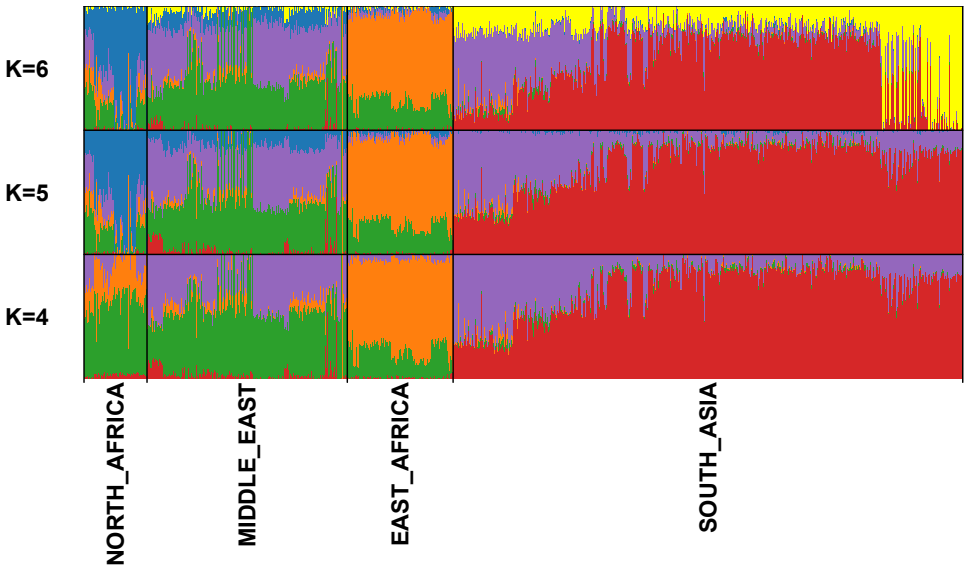


Figure 127. ADMIXTURE analysis of 1088 samples from the *migrants* dataset. The samples belong to five different macro-areas: Northern Africa, Eastern Africa, Middle East and Southern Asia.

From multiple points of view, this new classification problem is different from the previous one, showing some challenges we did not face before. For this reason, we needed to design a proper classification strategy suitable for the new macro-areas.

Before applying our AIM selection pipeline on our samples and genotypes, we explored the parameter space of the model, in order to choose the best ones.

As a first step, we tested the performances of the classifiers (Naïve Bayes and random forest) on the first ∼200 genetic markers sorted by their I_n, by tuning some parameters. We compared the performances of Naïve Bayes and HWE Naïve Bayes and varying the Laplace smoothing parameters ("hwe=False/True" and "smoothing" in Figure C.7). Then, we ranged the maximum depth of each tree in the Random Forest from 5 to 20 ("max_depth") and the number of the trees ("n_estimators") from 200 to 800 (Figure C.8). From the evaluation of the results, we concluded that the HWE assumption and a smoothing parameter greater than 0 are necessary to reach satisfactory performances in a Naïve Bayes classifier (Figure C.7); for what concerns the random forest, we can exclude a "max_depth" of 5, while 10 and 20 are similar.

For these reasons, we run the AIM selection pipeline using 1 as Laplace smoothing parameter in the Naïve Bayes classifier and "max_depth"=10 and "n_estimators"=200 in the random forest. Then, we compared the performances of the selected marker sets composed by up to 30 markers, when the prioritization strategy was the I_{n} and the feature importance of the random forest, while the classifiers where Naïve Bayes, Naïve Bayes HWE and the random forest. As you may notice, for the *migrants* problem, we considered bigger marker sets: due to the genetic similarity of some of the classes, we will need more feature to classify them.

As you will see, the analyses performed on the *migrants* dataset go in the same direction of the previous dataset; moreover, they offer to us more elements to discriminate among the various strategies in order to seek out the more suitable ones.

The plots do not clearly show which prioritization strategy allows to reach the best performances (Figure 128); however marker sets chosen thanks to I_{n} reach an accuracy of 86% with 30 markers, while feature importance-prioritization of markers stops at 82%.

As it was for the *continents* dataset, there is not appreciable difference due to MAF filtering (Figures 128 and C.9).

Unlike in the "continents" dataset, the classifier performances are clearly different. In both cases, the Naïve Bayes without HWE assumption performs very badly. Moreover, while both Naïve Bayes HWE and random forest tend to overfit, this is much more pronounced for the random forest, falling from a very high accuracy on the train set where it outperforms HWE NB, to a much lower accuracy on the validation and test sets, where it is soundly outperformed by HWE NB (Figure 128A,B).

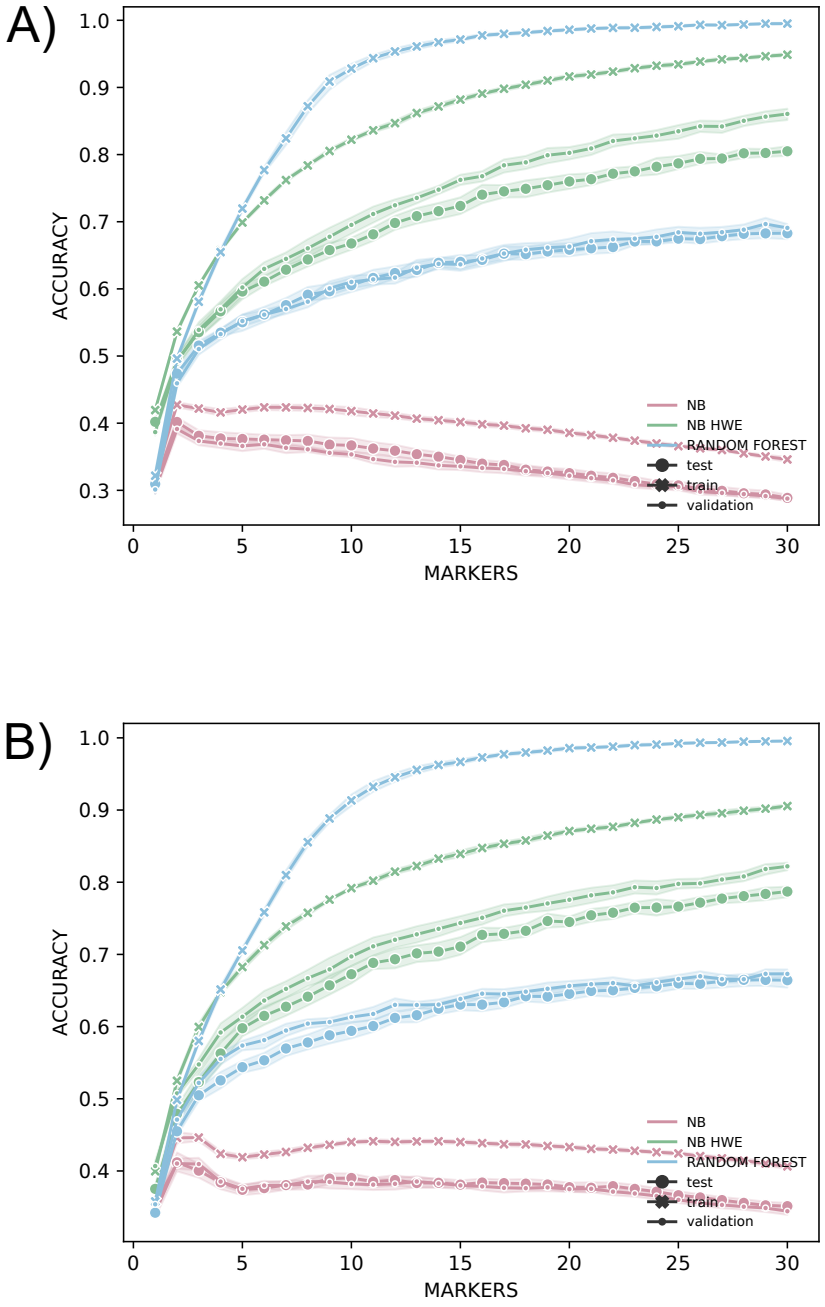


Figure 128. Accuracy comparison between different feature prioritization strategies on the *migrants* dataset. Classification accuracy using I_n **(A)** and random forest feature importance without MAF filtering **(B)** as prioritization strategies according to three different classifiers.

Interestingly, the analyses of the confusion matrix for the best marker set confirms what the ADMIXTURE analyses (Figure 127) had already suggested: Middle Eastern samples are being mistaken for Northern Africans, as well as South Asians and Middle Eastern individuals (Table 27).

	EAST_AFRICA	MIDDLE_EAST	NORTH_AFRICA	SOUTH_ASIA
EAST_AFRICA	120.3	2.7	11.0	0.0
MIDDLE_EAST	3.1	194.7	30.0	22.2
NORTH_AFRICA	2.5	15.1	61.9	0.5
SOUTH_ASIA	0.6	4.2	0.3	628.9

Table 27. Confusion matrix for 15 markers chosen with I_n prioritization and classified with HWE Naïve Bayes on the *migrants* dataset. This matrix corresponds to the classification results of the green curve on validation sets in Figure 128A.

Interpretation and what's next

As you may have understood from the results presented here, this study is work in progress and we are still a bit confused about how to improve our AIM selection pipeline. However, we have already got some insights about the classification problem. While the *continents* dataset proved to be really useful to understand the main characteristics of the population classification problem and to develop a method for doing it, is on the *migrants* dataset that we are most interested in for contributing to the problem of the personal identification of dead migrants.

Both in the *continents* and in the *migrants* dataset, the classification performances are higher on the training than on the validation and test sets. This is a sign of the presence of "variance", which is represented by the classification errors made on the validation or test set and it is ultimately a consequence of overfitting. This effect is more evident in the *migrants* dataset (Figure 128) and, specifically, when the classifier is a random forest.

Conversely, when there is a classification error on the training set — i.e., the maximum accuracy value on the training is not 1 — we are talking about "bias". In our plots, we could see a small amount of bias when we use the Naïve Bayes HWE classifier: in fact, the green line corresponding to the training, never reaches 100% of accuracy, meaning that the model can not correctly classify all the samples in the training set (Figure 128). From the other hand, Naïve Bayes HWE shows a mouch lower variance than random forest, thus confirming itself as the top classifier.

The presence of bias in a machine learning model depends mainly on

the fact that the method used is not completely suitable for the data or the dataset is too little and inhomogeneous. For example, a high bias could be seen when we try to classify a non-linearly separable dataset with a linear classifier. However, in our case we are probably in the situation where the classes in the dataset are not homogeneous, as ADMIXTURE plot could show (Figure 127). Thus, one possibility we can follow to improve the Naïve Bayes HWE classification performances on the training is to consider more homogeneous classes in the *migrants* problem or increase the size of the dataset.

However, when we compare the performances of the various models between validation and test sets — note that they are completely independent replica between each other — we see that they are similar and for the most part overlapped, meaning that the performances predictions are really stable. As a consequence, there is no reason to think that the performances will drop when new samples will be analysed.

Another point to consider is the source of the genotypes: in the case of the *continents* dataset they come from whole-genome sequences, even if at low coverage. Conversely, the *migrants* dataset has been merged from many different published SNP-array datasets. Given the ascertainment biases shown by these platforms with respect to non-European samples, we can reasonably assume that we are missing a great deal of the African and Asian genetic variation.

In order to fix the problems we have encountered, we can work both on our methods and on our datasets. We will try new classifiers, possibly combining them together, other prioritization and marker set exploration strategies. As for the dataset, we could try relabeling or cleaning our *migrant* dataset, but another possiblity is replacing with sequenced data. In fact, the HGDP samples, comprising many Northern African and Middle Eastern samples, have recently been sequenced (Bergström *et al.,* 2019), while a great effort have being made to sequence (Pagani *et al.,* 2015) and genotype (López *et al.,* 2019) more Eastern African samples (Pagani *et al.,* 2015). Having a good dataset is very important in machine learning, and with a bigger and less biased selection of variants we hope to see strong improvements in our results.

Conclusion

"Roads go ever ever on
Under cloud and under star,
Yet feet that wandering have gone
Turn at last to home afar"

J. R. R. Tolkien, *The Hobbit*

Migration tales

THE only way we can deeply understand our human nature is by looking back. Today, it is now possible. The scientific revolution we are experiencing is leading us in a real genomic voyage back in time. Following the footsteps of our ancestors in our genome, we are witnessing new and often astonishing aspects of human populations history, such as the intimate relationship that they had with Neanderthals several thousands of years ago, as well as the complex patterns of admixture and migrations, which gave rise to the modern human cultures.

However, the more we dig in our archaic, ancient and more recent past, the more we realize that there always have been a common thread connecting the human beings across the millennia. It is migration.

Throughout this thesis, its four parts narrate different "stories" about human populations and their movements.

The *"archaic tales"* are impressive and magnificent examples of the power of past migrations in deeply shaping our human nature. Even if due to a small portion of our genome, we - Eurasians - are not "pure" Sapiens, but instead, we hide something else: the genetic legacy of an extinct human group, the Neanderthals. When we consider what makes up the core of human identity only from a genetic point of view, it is pretty clear what we are: a miscellany of at least two distinct human groups, of which one no longer wanders this world. If our Sapiens ancestors hadn't armed themselves with courage and left their African home, they couldn't have met

and mixed with the Neanderthals. In that case, we would not exist in our present form, but instead we would be something else. It would have been, perhaps, neither better nor worse, but surely different: maybe, without some Neanderthal genes, our colonizing adventure would have been slowed down by the Eurasian environment, hostile for our African adapted bodies.

But this also applies to what happened next, in the *"ancient tales"*: the peopling of the world and, specifically for what concerns this thesis, the peopling of Europe have been taken place by multiple migratory waves, each of them, almost completely replacing or mixing with the previous settlers. This phenomenon went on for millennia and it is still happening, so much that it is not easy to answer the question "who are the Europeans?".

When we come to the *"modern tales"*, we talk about a slightly different matter. Here, the main theme is not migration, but instead the consequences that migration and admixture, together with selection and genetic drift for the great part, had on our genome. In particular, even healthy human beings are carrying more or less deleterious variants and the frequency of their alleles could be highly variable among different populations, as a consequence of the factors mentioned above. This means that the genetic structure of a population could determine the entity of a genetic burden the individuals bear during their wanderings.

Obviously, as treated in the last part, *"future tales"*, migrations are not a relic of the past. Around 800,000 migrants arrived in Europe over the last four years running from torture, abuse, misery or simply seeking a better life. They come from Eastern Africa (Somalia and Eritrea), Southern Asia (Bangladesh and Pakistan) and from the war-torn countries of the Middle East. Many of them fail to reach our shore and drown in the Mediterranean Sea. At this point, they have to face another journey: the winding path of personal identification and, in case of success, the way back home.

In my opinion, telling the tales of our past and the migrations who gave rise to the modern human populations have a great value today. In the current historical period, we have fond the concept of cultural identity, mainly because it could be very useful to draw boundaries: we know who we are by understanding whom is different from us. However, what would you say if I told you that around 80,000 years ago we were the African black-skinned migrants arriving in Europe? From a genetic point of view, the concept of "identity" is fluid: we always moved and always mixed with the people we met, always forgetting some "identities" while forming others.

Going more and more back in our past, the cultural, social, religious and national identities, which today seem so static and immutable, begin to crumble. At this point, there's just one identity left: the human one.

“ *Home is behind, the world ahead,*
And there are many paths to tread
Through shadows to the edge of night, ”
Until the stars are all alight.
The world behind and home ahead,
We'll wander back to home and bed.

J. R. R. Tolkien, *A Walking Song*

Appendices

Archaic tales

Neanderthal allele count

Below you can find the bash code for counting the number of Neanderthal alleles in heterozygosity and in homozygosity per individual in each population.

First, I created a raw file with a line per individual and a column per SNP, reporting 0 if the individual had two non-Neanderthal allele at that particular SNP, 1 if he shows an heterozygosity and 2 is the sample has a Neanderthal homozygous genotype. The options *--recode A --recode-allele* allow to customize the allele to be analysed, in this case, the Neanderthal allele.

After listing the populations, I computed the number of "1", "2", "0" and missing values per individual per population.

```
#############
# raw file #
#############

plink --bfile $INPUT --recode A --recode-allele nea.allele --out
    $INPUT.recode

#########
# count #
#########

cat pops.txt | while read line
do
grep -w ${line} $INPUT.raw > ${line}.raw
done

cat pops.txt | while read line
do
cut -d' ' -f7-7170 ${line}.raw | sed 's/[^1]//g' | awk '{ print
    length }' > ${line}.1
cut -d' ' -f7-7170 ${line}.raw | sed 's/[^2]//g' | awk '{ print
    length }' > ${line}.2
cut -d' ' -f7-7170 ${line}.raw | sed 's/[^0]//g' | awk '{ print
    length }' > ${line}.0
cut -d' ' -f7-7170 ${line}.raw | sed 's/[^NA]//g' | awk '{ print
    length }' > ${line}.NA
paste -d' ' <(cut -d' ' -f2 ${line}.raw) ${line}.1 ${line}.2 ${
    line}.0 ${line}.NA > ${line}.count
done
```

Neanderthal allele frequency differences

Here I report the bash code written to detect the most variable SNPs across pairs of populations: by calling three R scripts (see below) it computes the Neanderthal allele frequencies in each modern population, the ΔXAF values and the list of NTT SNPs (see page 315, 317 and 318, respectively).

```
####################
# create frq file #
####################

echo ``Creating .frq.strat file through plink ...''

plink --bfile $FILE --freq --family --out $FILE

###############
# update frq #
###############

# update frequency values with neanderthal inconsistencies

Rscript --vanilla update.nea.frequency.R

##############
# pops freq #
##############

# create frq file for each population if you will use R
echo ``Creating .freq file for each population ...''

for pop in CEU CHB FIN GBR IBS TSI YRI SAR ITC ITN ITS; do grep ${
    pop} $FILE.correct.frq.strat > ${pop}.freq; done

#######################
# compute Delta XAF #
#######################

# compute absolute difference of frequencies for each pair of
    population
echo ``Computing delta_freq values through delta_freq.R ...''

Rscript --vanilla delta_freq.R $OUTPUT2.bim

######################
# compute quantiles #
######################
```

```
# compute quantiles for each distribution of absolute delta
    frequencies for each pair of population
echo ``Computing quantiles through quantile.R ...''

Rscript --vanilla quantile.R

###############
# top1% SNPs #
###############

### search for the Neanderthal SNPs in 99th and 95th percentiles
echo ``Searching for the Neanderthal SNPs in 99th and 95th
    percentiles ...''

Rscript --vanilla intersection.95_99.R $NEA_SNP
```

R code to compute the ΔXAF values among pairs of modern populations.

```
#####################
### Delta_freq.R ###
#####################

args = commandArgs(trailingOnly=TRUE)
if (length(args)==0) {
stop(``At least one argument must be supplied (input file).n",
    call.=FALSE)
}

input_bim <- args[1]

### calculate absolute allele frequency differences (delta) for
    each SNPs of each pair of populations

files = list.files(pattern = "\\.freq$")

bim <- read.table(input_bim, nrows = 1000)
c <- sapply(bim, class)
bim <- read.table(input_bim, colClasses=c)

print(``Reading .freq file and creating data.frame for SNPs
    frequencies...")
print(paste0(``data.frame dimensions --> nrow: ", dim(bim)[1], ";
    ncol: ", length(files)))

m <- matrix(NA, nrow = dim(bim)[1], ncol = length(files))
m <- as.data.frame(m)
rownames(m) <- bim$V2

for (i in 1:length(files)){
a <- read.table(files[i])
m[,i] <- a$V6
names(m)[i] <- substr(files[i], 1, 3)
}

print("Creating vector containing all population pair combinations
    ...")
x <- vector('character')
for (pop in 1:(length(files)-1)){
for (plus in 1:(length(files) - pop)){
x <- append(x, paste0(files[pop],"_", files[pop+plus]))
}}
x <- gsub(".freq", "", x)

print("Creating data.frame for absolute delta frequencies...")
print(paste0("data.frame dimensions --> nrow: ", dim(bim)[1], ";
```

```
    ncol: ", length(x)))

n <- matrix(NA, nrow = dim(bim)[1], ncol = length(x))
n <- as.data.frame(n)
colnames(n) <- x
rownames(n) <- a$V2

for (i in 1:ncol(n)){
n[,i] <- abs(subset(m, select=substr(x[i],1,3)) - subset(m, select
   =substr(x[i],5,7)))
}

write.table(round(n, digits = 8), "delta.freq.tsv", row.names = T,
     col.names = T, quote = F, sep = "\t")
```

R code written to compute the percentile thresholds across the SNPs distributions.

```
####################
### quantile.R ###
####################

print(``Reading delta.freq.tsv file...")
n <- read.table("delta.freq.tsv", hea=T)

print(``Creating quantile table for each distribution of absolute
    delta frequencies..")
d <- matrix(NA, nrow = 55, ncol = 101)
d <- as.data.frame(d)
colnames(d) <- names(quantile(n$CEU_CHB, prob = seq(0, 1, length =
     101), type = 7))
rownames(d) <- names(n)
for (i in 1:55){
for (j in 1:101){
d[i,j] <- unname(quantile(n[,i], prob=seq(0, 1, length = 101),
    type = 7)[j])
}}

write.table(d, "quantile.tsv", row.names=T, col.names=T, quote=F,
    sep="\t")
```

R code used to find the NTT SNPs (Neanderthal SNPs in the 99th percentile) in each distribution.

```
##############################
### intersection.95_99.R ###
##############################

args = commandArgs(trailingOnly=TRUE)
if (length(args)==0) {
stop("At least one argument must be supplied (input file).n", call
    .=FALSE)
}

n <- read.table("delta.freq.tsv", hea=T)
nea <- read.table(args[1], hea=F)

### function that save a list to file
fnlist <- function(x, fil){ z <- deparse(substitute(x))
cat(z, "\n", file=fil)
nams=names(x)
for (i in seq_along(x) ){ cat(nams[i], "\t",  x[[i]], "\n",
file=fil, append=TRUE) }
}
### Neanderthal SNPs list in 99th percentile
print("Searching for the Neanderthal SNPs in 99th percentile...")
mylist.names <- names(n)
mylist_99 <- vector("list", length(mylist.names))
names(mylist_99) <- mylist.names
for (i in 1:55){
q <- unname(quantile(n[,i], prob=.99, type=7))
mylist_99[[i]] <- as.vector(intersect(rownames(n[which(n[,i] >= q)
    ,]),nea$V10))
}
fnlist(mylist_99, "mylist_nea_99.txt")
```

pops	OR	P-VALUE	1/OR
CHB_YRI	0	8.08E-25	
ITC_YRI	0.01647256	7.88E-25	60.70702002
GBR_YRI	0.032522102	9.32E-24	30.74831981
FIN_YRI	0.04703441	3.19E-23	21.26102976
FIN_GBR	0.047159442	2.82E-23	21.20466138
ITN_YRI	0.047293569	4.66E-23	21.14452382
CEU_YRI	0.048300311	1.29E-22	20.70380025
TSI_YRI	0.048740748	1.80E-22	20.5167143
IBS_YRI	0.048749399	1.80E-22	20.51307355
ITS_YRI	0.049012706	2.59E-22	20.40287258
SAR_YRI	0.057187783	1.04E-18	17.48625226
CHB_SAR	0.059398348	1.93E-23	16.83548503
FIN_TSI	0.062702271	4.81E-22	15.94838557
ITC_ITS	0.065314988	3.94E-21	15.3104215
ITS_TSI	0.065550264	5.82E-21	15.25546868
CEU_CHB	0.069488702	2.53E-24	14.3908286
GBR_IBS	0.081702311	5.34E-20	12.2395558
IBS_TSI	0.081776382	5.32E-20	12.22846953
GBR_ITC	0.082239299	7.55E-20	12.15963665
CEU_GBR	0.082882313	1.07E-19	12.06530033
CHB_GBR	0.08350411	3.34E-23	11.97545845
CHB_ITS	0.083968988	4.78E-23	11.90915862
CEU_FIN	0.094885927	1.16E-19	10.53897068
FIN_IBS	0.098598978	9.30E-19	10.14209298
FIN_ITN	0.107901428	2.13E-19	9.267717961
FIN_ITC	0.109654412	6.25E-19	9.119560067
FIN_SAR	0.11131815	1.18E-18	8.983261062
CEU_ITC	0.113984204	4.83E-18	8.773145414
GBR_TSI	0.1143967	4.64E-18	8.741510913
IBS_ITC	0.114494304	4.67E-18	8.734058955
ITC_TSI	0.11490028	6.66E-18	8.703198959
CHB_ITN	0.124889379	2.10E-20	8.007085984
CHB_FIN	0.126443572	6.07E-20	7.908666169
GBR_ITS	0.130656111	4.69E-17	7.65367951
CHB_ITC	0.140418739	3.81E-19	7.121556639
ITN_TSI	0.143999948	1.05E-16	6.944446976
ITN_ITS	0.144549641	1.01E-16	6.918038618
IBS_ITS	0.147987034	3.93E-16	6.757348748
CEU_SAR	0.150826054	1.07E-15	6.630154219
CHB_TSI	0.153708911	1.60E-18	6.505803684
FIN_ITS	0.154024538	8.17E-17	6.492471998
CHB_IBS	0.154026584	2.36E-18	6.492385762
ITC_ITN	0.161301678	8.20E-16	6.199563519
ITS_SAR	0.186198725	6.21E-14	5.370606051
ITC_SAR	0.187029262	5.93E-14	5.346756921
CEU_ITN	0.19271563	2.32E-14	5.188992702
GBR_ITN	0.192945436	2.28E-14	5.182812405
SAR_TSI	0.201660976	2.20E-13	4.958817623
CEU_TSI	0.207476815	7.70E-13	4.819815654
CEU_ITS	0.210152282	1.51E-13	4.758454158
ITN_SAR	0.245171428	3.83E-12	4.078778702
IBS_SAR	0.249291547	9.76E-12	4.011367469
GBR_SAR	0.253751506	2.52E-11	3.940863304
IBS_ITN	0.256021241	6.33E-12	3.905925912
CEU_IBS	0.280114697	8.00E-11	3.569966202

Table A.1. Odds ratio of NTT SNPs. Odds ratios for the presence of SNPs in Neanderthal introgressed regions in the top 1% of each population pair distribution of differences in Neanderthal allele frequency.

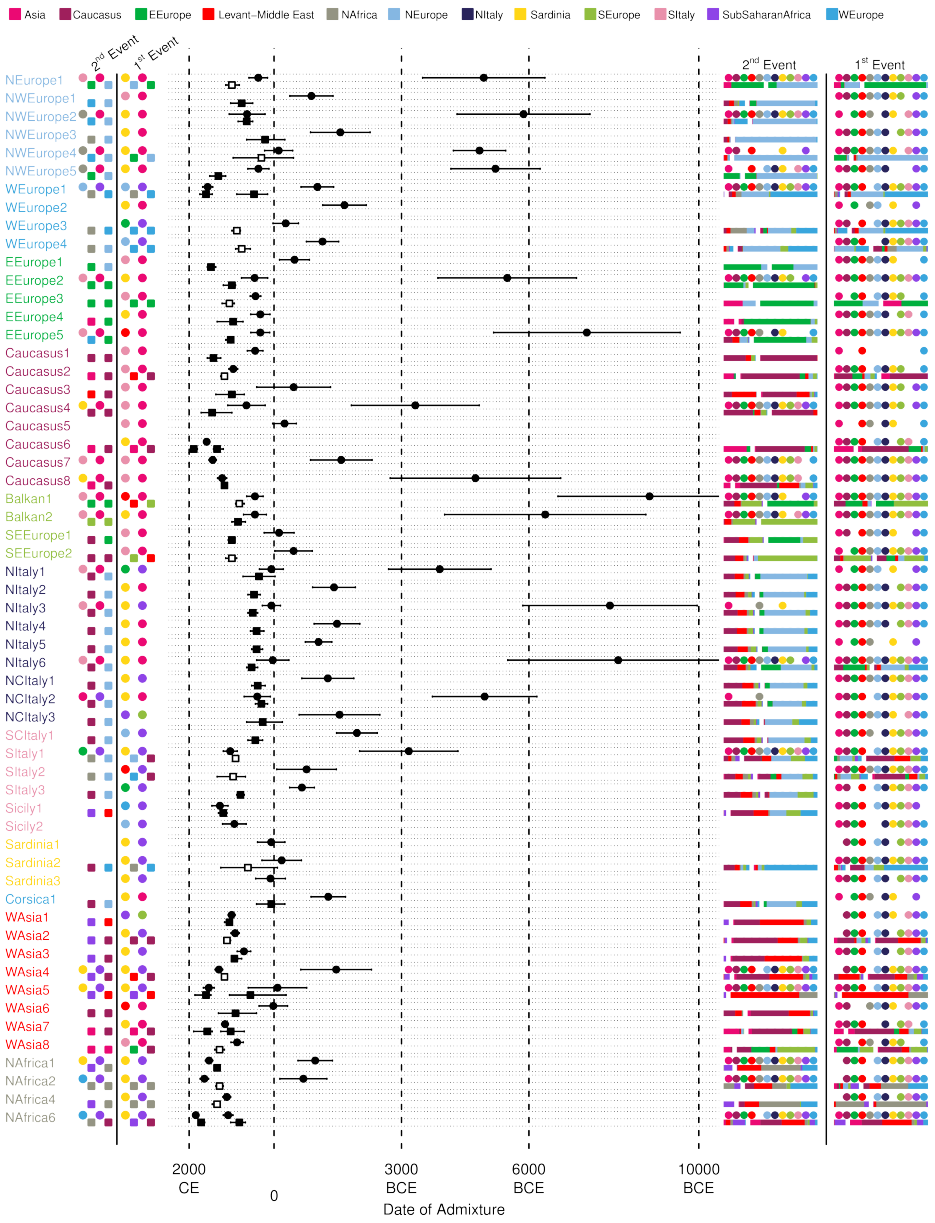


Figure A.1. GT and MALDER analyses for all the Eurasian and North African clusters. Dates of the events inferred by "noItaly" GT (squares) and MALDER (circles) for population clusters are reported in the central part of the plot; lines encompassed the 95% CI for GT and ±1 Standard Error for MALDER. For further details see Raveane *et al.*, 2019.

Ancient tales

CHROMOPAINTER (CP)

Below you can find the bash and R code used for running CP in parallel (the code was used to run CP on *ultimate* sources). As you can see from the code, in the "unliked" mode three input files are used:

- An haplotype file ("*-g*") which contains the genetic information for the *donor* and *recipient* chromosome in *.phase* format.
- A donor file ("*-f*") which provides the information on donor individuals or populations.
- A file reporting all individuals or populations participating in the painting step ("*-t*").

The *gsub.py* is simply the code I used to generate job files running in parallel on Marconi server (CINECA) with a Slurm system (code not shown).

```
#!/bin/bash

for i in {1..22};
do
A=1
B=50
for j in {1..98};
do
./gsub2.py --time=16 --mem=8 --log $WDIR/results.cp.ultimate/
    results_combinedLD_chr.all.max.ultimate.clean.chr${i}.${j}.
    joblog --name "cpv2.-$i-$j" "ChromoPainterv2 \
-u \
-g $WDIR/results_combinedLD_chr.all.max.ultimate.clean.chr${i}.
    phase \
-t $WDIR/label_infile.ultimate \
-f $WDIR/pops_list_infile.ultimate $A $B \
-o $WDIR/results.cp.ultimate/results_combinedLD_chr.all.max.
    ultimate.clean.chr${i}.${j} \
-M 0.00051 \
-n 365" --test
A=$((A+50))
B=$((B+50))
done
done | sed 's/$/ \&/' | split --lines=16 --suffix-length=3 --
    numeric-suffixes - job_list_chunk
```

```
files <- list.files(pattern = "job_list_chunk")
for (i in 1:length(files)){
a <- read.delim(files[i], sep=" ", hea=F)
a$V18 <- a$V17
a$V17 <- paste0("> log_file", sub(".*chr", "", a$V12), " 2> err_
    file", sub(".*chr", "", a$V12))
write.table(a, paste0("new.", files[i]), na="", row.names=F, col.
    names=F, quote=F)
}
```

```
for j in new.job_list_chunk*; do echo 'wait' >> ${j}; done

for j in {000..134};
do ./gsub2.py --log new.job_list_chunk${j}.joblog --name "cp.$j"
    --cores=16 --mem=60 --time=24 --queue gll_usr_prod --account
    account_name "bash new.job_list_chunk${j}";
done
```

Here is the command used for combining the CP results across all chromosomes and sets of individuals in a single output.

```
fs combine -o cp.combine.ultimate results.cp.ultimate/
    results_combinedLD_chr.all.max.ultimate.clean.chr*chunkcounts.
    out -v
```

Population	Geographic Region	Continent	Country	FMD	Platform	Source
Abkhasian	Caucasus	Eurasia	Georgia	23	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2013; Yunusbayev *et al.*, 2015
Adygey	Caucasus	Eurasia	Russia	17	OmniExpress 610 – 660	Li *et al.*, 2008
Albanian	Balkan	Europe	Albania	6	Omni 2.5M	Raveane *et al.*, 2019
Algerian	NorthAfrica	Africa	Algeria	5	OmniExpress 610 – 660	Behar *et al.*, 2013
Altaian	Siberia	Asia	Altai Republic	19	OmniExpress 610 – 660 / Omni 1M	Rasmussen *et al.*, 2010; Raghavan *et al.*, 2014 Yunusbayev *et al.*, 2015
Armenian	Caucasus	Eurasia	Armenia	35	OmniExpress 610 – 660	Yunusbayev *et al.*, 2012; Behar *et al.*, 2010
AshkenaziJew	North-East Europe	Europe	NA	20	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
Azerbaijani	Caucasus	Eurasia	Daghestan	2	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
Azerbaijani	Caucasus	Eurasia	Iran	21	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2015
AzerbaijaniJew	Caucasus	Eurasia	Azerbajan	7	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
Balkar	Caucasus	Eurasia	Kabardino-Balkaria	22	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2015
Balochi	CentralAsia	Asia	Pakistan	24	OmniExpress 610 – 660	Li *et al.*, 2008
BantuKenya	EastAfrica	Africa	Kenya	10	OmniExpress 610 – 660	Li *et al.*, 2008
BantuSouthAfrica	SouthAfrica	Africa	South Africa	8	OmniExpress 610 – 660	Li *et al.*, 2008
Bashkir	CentralAsia	Asia	Russia	23	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2015
Basque	West Europe	Europe	Spain	24	OmniExpress 610 – 660	Li *et al.*, 2008
Bedouin	NorthAfrica	Africa	Egypt	46	OmniExpress 610 – 660	Li *et al.*, 2008
Belarusian	EastEurope	Europe	Belarus	16	OmniExpress 610 – 660	Behar *et al.*, 2010; Behar *et al.*, 2013
BiakaPygmy	EastAfrica	Africa	Central African Republic	20	OmniExpress 610 – 660	Li *et al.*, 2008
Bosnian	Balkan	Europe	Bosnia and Herzegovina	15	OmniExpress 610 – 660	Kovacevic *et al.*, 2014
Brahui	SouthAsia	Asia	Pakistan	25	OmniExpress 610 – 660	Li *et al.*, 2008
British	North-West Europe	Europe	United Kingdom	95	Omni 2.5M	Auton *et al.*, 2015
Bulgarian	EastEurope	Europe	Bulgaria	31	OmniExpress 610 – 660	Yunusbayev *et al.*, 2012; Hellenthal *et al.*, 2014
Burusho	SouthAsia	Asia	Pakistan	25	OmniExpress 610 – 660	Li *et al.*, 2008
Utah Residents (CEPH) with Northern and Western European Ancestry	North-West Europe	Europe	Utah	99	Omni 2.5M	Auton *et al.*, 2015
Han Chinese in Beijing	CentralAsia	Asia	Cina	51	Omni 2.5M	Auton *et al.*, 2015
Chechen	Caucasus	Eurasia	Chechen Republic	24	OmniExpress 610 – 660	Metspalu *et al.*, 2011; Chaubey *et al.*, 2012 Yunusbayev *et al.*, 2012
Chukchin	Siberia	Asia	Chukotka Autonomous Okrug (Russia)	10	OmniExpress 610 – 660 / Omni 1M	Rasmussen *et al.*, 2010; Yunusbayev *et al.*, 2015
CochinJewish	MiddleEast	Asia	Israel	3	Omni 1M	Behar *et al.*, 2013
Corsican	West Europe	Europe	Corsica	16	OmniExpress 610 – 660	Tamm *et al.*, 2019

Population	Geographic Region	Continent	Country	FMD	Platform	Source
Croatian	SouthEastEurope	Europe	Croatia	43	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2013; Hellenthal *et al.*, 2014
Cypriot	SouthEastEurope	Europe	Cyprus	12	OmniExpress 610 – 660	Behar *et al.*, 2010
Druze	MiddleEast	Asia	Lebanon	44	OmniExpress 610 – 660 / Omni 1M	Li *et al.*, 2008; Behar *et al.*, 2013
Egyptian	NorthAfrica	Africa	Egypt	24	OmniExpress 610- 660 / Omni 2.5M	Behar *et al.*, 2010; Pagani *et al.*, 2012
English	North-West Europe	Europe	England	8	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
Estonian	NorthEurope	Europe	Estonia	21	OmniExpress 610 – 660	Raghavan *et al.*, 2014; Kushniarevich *et al.*, 2015
Ethiopian	EastAfrica	Africa	Ethiopia	120	OmniExpress 610- 660 / Omni 2.5M	Behar *et al.*, 2010; Pagani *et al.*, 2012
EthiopianJew	EastAfrica	Africa	Ethiopia	15	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
Finnish	NorthEurope	Europe	Finland	102	OmniExpress 610- 660 / Omni 2.5M	Hellenthal *et al.*, 2014; Auton *et al.*, 2015
French	North-West Europe	Europe	France	28	OmniExpress 610 – 660	Li *et al.*, 2008
French	North-West Europe	Europe	Provance (France)	5	OmniExpress 610 – 660	Tamm *et al.*, 2019
FrenchJew	North-West Europe	Europe	France	6	OmniExpress 610 – 660	Behar *et al.*, 2013
Gagauz	EastEurope	Europe	Moldova	12	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
GeorgianJew	Caucasus	Eurasia	Georgia	11	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2013
Georgian	Caucasus	Eurasia	Georgia	30	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
German	North-West Europe	Europe	Germany	44	OmniExpress 610 – 660	Hellenthal *et al.*, 2014; Parolo *et al.*, 2015 Yunusbayev *et al.*, 2015
GermanAustrian	North-West Europe	Europe	Germany	4	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
Gond	SouthAsia	Asia	India	4	OmniExpress 610 – 660	Metspalu *et al.*, 2011
Greek	SouthEastEurope	Europe	Greek	20	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
GreekCentral	SouthEastEurope	Europe	Greek	10	OmniExpress 610 – 660	Behar *et al.*, 2013
GreekMacedonian	SouthEastEurope	Europe	Greek	7	OmniExpress 610 – 660	Kushniarevich *et al.*, 2015
GreekPeloponesian	SouthEastEurope	Europe	Greek	9	OmniExpress 610 – 660	Kushniarevich *et al.*, 2015
GreekThessalyian	SouthEastEurope	Europe	Greek	10	OmniExpress 610 – 660	Behar *et al.*, 2013
Hungarian	EastEurope	Europe	Hungaria	19	OmniExpress 610 – 660	Behar *et al.*, 2010
IranianJew	MiddleEast	Asia	Iran	11	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
Iranian	MiddleEast	Asia	Iran	19	OmniExpress 610 – 660	Behar *et al.*, 2010
IraqiJew	MiddleEast	Asia	Iraq	13	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
Irish	North-West Europe	Europe	Irland	7	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
IT_U	SouthEurope	Europe	Italy	138	OmniExpress 610 – 660	Behar *et al.*, 2013; Parolo *et al.*, 2015
ITC-ABR_B	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-ABR_G	SouthEurope	Europe	Italy	19	OmniExpress 610- 660 / Omni 2.5M	Raveane *et al.*, 2019; Behar *et al.*, 2013
ITC-ABR_P	SouthEurope	Europe	Italy	5	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-LAZ_B	SouthEurope	Europe	Italy	5	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-LAZ_F-ITS-PUG_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-LAZ_G	SouthEurope	Europe	Italy	32	OmniExpress 610 - 660 / Omni 2.5M / Omni 1M	Raveane *et al.*, 2019; Fiorito *et al.*, 2016

Population	Geographic Region	Continent	Country	FMD	Platform	Source
ITC-LAZ_P	SouthEurope	Europe	Italy	3	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-MAR_B	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-MAR_F-ITN-LOM_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-MAR_G	SouthEurope	Europe	Italy	16	Omni 2.5M	Raveane *et al.*, 2019
ITC-MAR_P	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-TSI_B	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-TSI_F-ITC-ABR_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-TSI_F-ITN-LOM_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-TSI_F-ITN-VEN_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-TSI_F-ITS-SIC_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-TSI_G	SouthEurope	Europe	Italy	135	OmniExpress 610 - 660 / Omni 2.5M / Omni 1M	Auton *et al.*, 2015; Fiorito *et al.*, 2016 Li *et al.*, 2008; Tamm *et al.*, 2019 Raveane *et al.*, 2019
ITC-TSI_P	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-UMB_B	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITC-UMB_G	SouthEurope	Europe	Italy	9	Omni 2.5M	Raveane *et al.*, 2019
ITN-EMI_B	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-EMI_F-ITC-ABR_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-EMI_F-ITN-LOM_M	SouthEurope	Europe	Italy	4	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-EMI_F-ITN-VEN_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-EMI_G	SouthEurope	Europe	Italy	30	Omni 2.5M / Omni 1M	Fiorito *et al.*, 2016
ITN-EMI_P	SouthEurope	Europe	Italy	9	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-FRI_B	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-FRI_F-ITC-EMI_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-FRI_F-ITN-LOM_M	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-FRI_F-ITN-PIE_M	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-FRI_F-ITS-CAL_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-FRI_G	SouthEurope	Europe	Italy	15	Omni 2.5M	Raveane *et al.*, 2019
ITN-FRI_P	SouthEurope	Europe	Italy	4	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LIG_B	SouthEurope	Europe	Italy	6	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LIG_F-ITN-LOM_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LIG_G	SouthEurope	Europe	Italy	25	Omni 2.5M / Omni 1M	Raveane *et al.*, 2019; Fiorito *et al.*, 2016
ITN-LIG_P	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_B	SouthEurope	Europe	Italy	76	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITC-EMI_M	SouthEurope	Europe	Italy	6	OmniExpress 610 – 660	Parolo *et al.*, 2015

Population	Geographic Region	Continent	Country	FMD	Platform	Source
ITN-LOM_F-ITC-LAZ_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITC-TSI_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITC-UMB_M	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITN-ALT_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITN-FRI_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITN-LIG_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITN-PIE_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITN-TRE_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITN-VEN_M	SouthEurope	Europe	Italy	12	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_F-ITS-SIC_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-LOM_G	SouthEurope	Europe	Italy	52	OmniExpress 610 - 660 / Omni 2.5M / Omni 1M	Li *et al.*, 2008; Raveane *et al.*, 2019 Fiorito *et al.*, 2016
ITN-LOM_P	SouthEurope	Europe	Italy	147	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-PIE_B	SouthEurope	Europe	Italy	259	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-PIE_F-ITN-LOM_M	SouthEurope	Europe	Italy	5	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-PIE_F-ITN-VDA_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-PIE_G	SouthEurope	Europe	Italy	40	OmniExpress 610 - 660 / Omni 2.5M / Omni 1M	Fiorito *et al.*, 2016; Tamm *et al.*, 2019
ITN-PIE_P	SouthEurope	Europe	Italy	3	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-TRE_B	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-TRE_F-ITN-LOM_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-TRE_F-ITN-PIE_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-TRE_F-ITN-VEN_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-TRE_G	SouthEurope	Europe	Italy	8	Omni 2.5M	Raveane *et al.*, 2019
ITN-VDA_G	SouthEurope	Europe	Italy	29	Omni 2.5M / Omni 1M	Raveane *et al.*, 2019; Fiorito *et al.*, 2016
ITN-VEN_B	SouthEurope	Europe	Italy	10	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-VEN_F-ITC-EMI_M	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-VEN_F-ITN-LOM_M	SouthEurope	Europe	Italy	9	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-VEN_F-ITN-PIE_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITN-VEN_F-ITS-SAR_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015

Population	Geographic Region	Continent	Country	FMD	Platform	Source
ITN-VEN_G	SouthEurope	Europe	Italy	16	Omni 2.5M	Raveane *et al.*, 2019
ITN-VEN_P	SouthEurope	Europe	Italy	40	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-BAS_B	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-BAS_G	SouthEurope	Europe	Italy	37	Omni 2.5M / Omni 1M	Raveane *et al.*, 2019; Fiorito *et al.*, 2016
ITS-BAS_P	SouthEurope	Europe	Italy	9	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAL_B	SouthEurope	Europe	Italy	7	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAL_F-ITS-SIC_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAL_G	SouthEurope	Europe	Italy	32	OmniExpress 610 - 660 / Omni 2.5M / Omni 1M	Hellenthal *et al.*, 2014; Fiorito *et al.*, 2016 Raveane *et al.*, 2019
ITS-CAL_P	SouthEurope	Europe	Italy	12	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAM_B	SouthEurope	Europe	Italy	5	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAM_F-ITC-LAZ_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAM_F-ITN-LOM_M	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAM_F-ITN-VEN_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-CAM_G	SouthEurope	Europe	Italy	14	Omni 2.5M	Raveane *et al.*, 2019
ITS-CAM_P	SouthEurope	Europe	Italy	17	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-MOL_G	SouthEurope	Europe	Italy	2	Omni 2.5M	Raveane *et al.*, 2019
ITS-PUG_B	SouthEurope	Europe	Italy	15	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-PUG_F-ITN-LOM_M	SouthEurope	Europe	Italy	3	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-PUG_F-ITN-VEN_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-PUG_F-ITS-CAM_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-PUG_F-ITS-SIC_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-PUG_G	SouthEurope	Europe	Italy	14	Omni 2.5M	Raveane *et al.*, 2019
ITS-PUG_P	SouthEurope	Europe	Italy	27	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SAR_B	SouthEurope	Europe	Italy	5	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SAR_F-ITS-CAM_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SAR_G	SouthEurope	Europe	Italy	58	OmniExpress 610 - 660 / Omni 2.5M / Omni 1M	Li *et al.*, 2008; Fiorito *et al.*, 2016
ITS-SAR_P	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SIC_B	SouthEurope	Europe	Italy	11	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SIC_F-ITC-EMI_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SIC_F-ITN-FRI_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SIC_F-ITN-LOM_M	SouthEurope	Europe	Italy	2	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SIC_F-ITN-VEN_M	SouthEurope	Europe	Italy	1	OmniExpress 610 – 660	Parolo *et al.*, 2015
ITS-SIC_G	SouthEurope	Europe	Italy	64	OmniExpress 610 - 660 / Omni 2.5M / Omni 1M	Behar *et al.*, 2013; Hellenthal *et al.*, 2014 Fiorito *et al.*, 2016

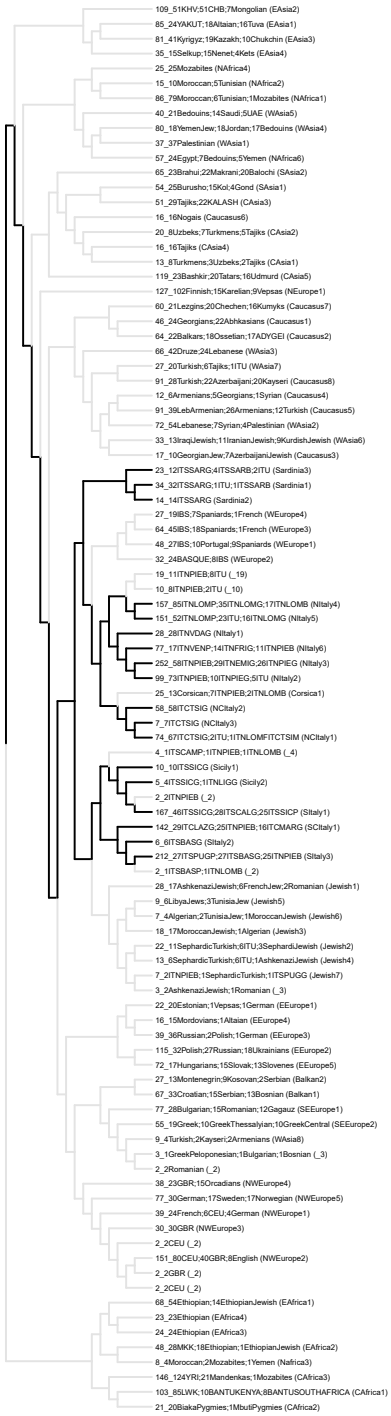
Population	Geographic Region	Continent	Country	FMD	Platform	Source
ITS-SIC_P	SouthEurope	Europe	Italy	25	OmniExpress 610 – 660	Parolo *et al.*, 2015
Jordanian	MiddleEast	Asia	Jordan	20	OmniExpress 610 – 660	Behar *et al.*, 2010
Kabardian	Caucasus	Eurasia	Kabardino-Balkaria	3	Omni 1M	Yunusbayev *et al.*, 2015
Kalash	SouthAsia	Asia	Pakistan	22	OmniExpress 610 – 660	Li *et al.*, 2008
Kalmyk	Caucasus	Eurasia	Kalmykia	14	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
Karakalpak	CentralAsia	Asia	Uzbekistan	10	Omni 1M	Yunusbayev *et al.*, 2015
Karelian	NorthAsia	Asia	Republic of Karelia	15	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
Kayseri	WestAsia	Eurasia	Turkey	23	OmniExpress 610 – 660	Hodoğlugil & Mahley, 2012
Kazakh	CentralAsia	Asia	Kazakhstan	20	OmniExpress 610 – 660 / Omni 1M	Raghavan *et al.*, 2014 Yunusbayev *et al.*, 2015
Maasai in Kinyawa	EastAfrica	Africa	Kenya	28	Omni 2.5M	Auton *et al.*, 2015
Ket	Siberia	Asia	Krasnoyarsk Krai (Russia)	4	OmniExpress 610 – 660 / Omni 1M	Rasmussen *et al.*, 2010 Yunusbayev *et al.*, 2015
Kol	SouthAsia	Asia	India	15	OmniExpress 610 – 660	Metspalu *et al.*, 2011
Komi	NorthAsia	Asia	Komi Republic	16	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
Kosovan	Balkan	Europe	Serbia	9	OmniExpress 610 – 660	Kovacevic *et al.*, 2014
Kryashen	CentralAsia	Asia	Kazakhstan	3	Omni 1M	Yunusbayev *et al.*, 2015
Kumyk	WestAsia	Asia	Kalmykia	17	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2012 Yunusbayev *et al.*, 2015
Kurd	MiddleEast	Asia	Turkey	6	OmniExpress 610 – 660	Yunusbayev *et al.*, 2012
KurdishJew	MiddleEast	Asia	Turkey	9	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2013
Kyrigyz	CentralAsia	Asia	Kyrgyzistan	41	OmniExpress 610 – 660 / Omni 1M	Hodoğlugil & Mahley, 2012 Raghavan *et al.*, 2014 Yunusbayev *et al.*, 2015
Latvian	EastEurope	Europe	Latvia	6	OmniExpress 610 – 660	Kushniarevich *et al.*, 2015
Lebanese	MiddleEast	Asia	Lebanon	79	OmniExpress 610 – 660	Behar *et al.*, 2010; Haber *et al.*, 2013
LebArmenian	MiddleEast	Asia	Lebanon	39	OmniExpress 610 – 660	Haber *et al.*, 2016
Lezgin	Caucasus	Eurasia	Dagestan	21	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
LibyaJew	NorthAfrica	Africa	Lybia	6	OmniExpress 610 – 660	Behar *et al.*, 2013
Lithuanian	EastEurope	Europe	Lithuania	10	OmniExpress 610 – 660	Behar *et al.*, 2010
Luhya in Webuye	EastAfrica	Africa	Kenya	85	Omni 2.5M	Auton *et al.*, 2015
Macedonian	Balkan	Europe	Macedonia	14	OmniExpress 610 – 660	Kovacevic *et al.*, 2014
Makrani	SouthAsia	Asia	Pakistan	25	OmniExpress 610 – 660	Li *et al.*, 2008
Mandenka	WestAfrica	Africa	Senegal	21	OmniExpress 610 – 660	Li *et al.*, 2008
Mari	CentralAsia	Asia	Mari El Republic (Russia)	15	OmniExpress 610 – 660	Raghavan *et al.*, 2014
MbutiPygmy	EastAfrica	Africa	Democratic Republic of Congo	1	OmniExpress 610 – 660	Li *et al.*, 2008
Moldovian	EastEurope	Europe	Moldova	7	Omni 1M	Behar *et al.*, 2013
Mongolian	EastAsia	Asia	Mongolia	21	OmniExpress 610 – 660 / Omni 1M	Li *et al.*, 2008; Rasmussen *et al.*, 2010 Yunusbayev *et al.*, 2015
Montenegrin	Balkan	Europe	Montenegro	14	OmniExpress 610 – 660	Kovacevic *et al.*, 2014
Mordovian	CentralAsia	Asia	Mordovia	15	OmniExpress 610 – 660	Yunusbayev *et al.*, 2012
Moroccan	NorthAfrica	Africa	Morocco	93	OmniExpress 610 - 660 / Omni 2.5M	Behar *et al.*, 2010; Hellenthal *et al.*, 2014
MoroccanJew	NorthAfrica	Africa	Morocco	18	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013

Population	Geographic Region	Continent	Country	FMD	Platform	Source
Mozabite	NorthAfrica	Africa	Algeria	29	OmniExpress 610 – 660	Li *et al.*, 2008
Nenet	Siberia	Asia	Russia	15	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2015
Nogai	CentralAsia	Asia	Russia	16	OmniExpress 610 – 660	Yunusbayev *et al.*, 2012
Norwegian	NorthEurope	Europe	Norwey	18	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
Orcadian	North-West Europe	Europe	Orkney Islands	15	OmniExpress 610 – 660	Li *et al.*, 2008
Ossetian	CentralAsia	Asia	Ossetia	18	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2012; Behar *et al.*, 2013
Palestinian	MiddleEast	Asia	Palestine	52	OmniExpress 610 – 660 / Omni 1M	Li *et al.*, 2008; Behar *et al.*, 2013
Polish	EastEurope	Europe	Poland	36	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2013; Hellenthal *et al.*, 2014
Portuguese	West Europe	Europe	Portugal	10	OmniExpress 610 – 660	Tamm *et al.*, 2019
Romanian	EastEurope	Europe	Romania	20	OmniExpress 610 – 660	Parolo *et al.*, 2015; Behar *et al.*, 2010
Russian	CentralAsia	Asia	Pinega	4	OmniExpress 610 – 660	Li *et al.*, 2008
Russian	CentralAsia	Asia	Russia	59	OmniExpress 610 – 660 / Omni 1M	Li *et al.*, 2008; Behar *et al.*, 2013 Yunusbayev *et al.*, 2015 Kushniarevich *et al.*, 2015
Samaritian	MiddleEast	Asia	Israel	2	OmniExpress 610 – 660	Behar *et al.*, 2010
Saudi	MiddleEast	Asia	Saudi Arabian	20	OmniExpress 610 – 660	Behar *et al.*, 2010
Scottish	North-West Europe	Europe	Scottland	6	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
Selkup	CentralAsia	Asia	Siberia	17	OmniExpress 610 – 660	Raghavan *et al.*, 2014 Rasmussen *et al.*, 2010
SephardicJew	MiddleEast	Asia	Israel	3	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2013
SephardicTurkish	MiddleEast	Asia	Israel	19	OmniExpress 610 – 660	Behar *et al.*, 2010
Serbian	Balkan	Europe	Serbia	18	OmniExpress 610 – 660	Kovacevic *et al.*, 2014
Slovakian	EastEurope	Europe	Slovakia	15	OmniExpress 610 – 660	Kushniarevich *et al.*, 2015
Slovenian	Balkan	Europe	Slovenia	15	OmniExpress 610 – 660	Kushniarevich *et al.*, 2015
Spanish	West Europe	Europe	Spain	34	OmniExpress 610 – 660	Behar *et al.*, 2010; Hellenthal *et al.*, 2014
Spanish	West Europe	Europe	Spain	100	Omni 2.5M	Auton *et al.*, 2015
Swedish	NorthEurope	Europe	Sweden	17	OmniExpress 610 – 660	Behar *et al.*, 2013
SyriaJew	MiddleEast	Asia	Syria	2	OmniExpress 610 – 660	Behar *et al.*, 2013
Syrian	MiddleEast	Asia	Syria	16	OmniExpress 610 – 660	Behar *et al.*, 2010
Tabassaran	Caucasus	Eurasia	Dagestan (Russia)	2	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2013
Tajik	CentralAsia	Asia	Tajikistan	60	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2012
Tatar	CentralAsia	Asia	Tatar	20	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
Tunisian	NorthAfrica	Africa	Tunisia	12	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
TunisiaJew	NorthAfrica	Africa	Tunisia	6	OmniExpress 610 – 660	Behar *et al.*, 2013
Turkish	MiddleEast	Asia	Turkey	49	OmniExpress 610 - 660 / Omni 2.5M	Behar *et al.*, 2010; Hodoğlugil & Mahley, 2012 Paschou *et al.*, 2014
Turkish	MiddleEast	Asia	Turkey	20	OmniExpress 610 – 660	Hodoğlugil & Mahley, 2012
Turkmen	CentralAsia	Asia	Turkmenistan	23	OmniExpress 610 – 660 / Omni 1M	Yunusbayev *et al.*, 2012 Yunusbayev *et al.*, 2015 Di Cristofaro *et al.*, 2013

Population	Geographic Region	Continent	Country	FMD	Platform	Source
Tuva	CentralAsia	Asia	Tuva	16	OmniExpress 610 – 660 / Omni 1M	Rasmussen *et al.*, 2010 Yunusbayev *et al.*, 2015
Emirates	MiddleEast	Asia	uae	14	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
Udmurt	CentralAsia	Asia	Kirov Oblast	16	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
Ukrainian	EastEurope	Europe	Ukraine	20	OmniExpress 610 – 660	Yunusbayev *et al.*, 2012
Uzbek	CentralAsia	Asia	Uzbekistan	24	OmniExpress 610 – 660	Behar *et al.*, 2010; Raghavan *et al.*, 2014 Di Cristofaro *et al.*, 2013
Vepsian	CentralAsia	Asia	Republic of Karelia	10	OmniExpress 610 – 660	Yunusbayev *et al.*, 2015
Vietnamese	EastAsia	Asia	Vietnam	51	Omni 2.5M	Auton *et al.*, 2015
Welsh	North-West Europe	Europe	Wales	4	OmniExpress 610 – 660	Hellenthal *et al.*, 2014
Yakut	Siberia	Asia	Yakutia	25	OmniExpress 610 – 660 / Omni 1M	Li *et al.*, 2008; Yunusbayev *et al.*, 2015
Yemeni	MiddleEast	Asia	Yemen	8	OmniExpress 610 – 660	Behar *et al.*, 2010
YemenJew	MiddleEast	Asia	Yemen	18	OmniExpress 610 – 660 / Omni 1M	Behar *et al.*, 2010; Behar *et al.*, 2013
Yoruba	EastAfrica	Africa	Nigeria	124	OmniExpress 610 - 660 / Omni 2.5M	Li *et al.*, 2008; Auton *et al.*, 2015

Legend on next page.

Table B.2. *Worldwide populations included in the FMD.* 1,589 individuals representing all of the 20 Italian administrative regions and data from 140 world-wide reference populations were included in the FMD, for a final dataset of 4,852 (FMD) and 1,651 (HDD) samples. Only Illumina genotyped samples are included and the database/publication from which they have been gathered is indicated.



Legend on next page.

Figure B.1. fineStructure dendrogram of FMD (4,852 individuals). Each tip of the dendrograms represents a group of individuals with similar copying vectors. The first number of each tip label refers to the total number of individuals in the cluster. This value is followed by "_" and the name of the three most representative geographically-assigned populations, each with its number of samples. Within brackets, a summary name arbitrarily assigned to the cluster is reported. Thick lines in black refer to the Italian clusters.

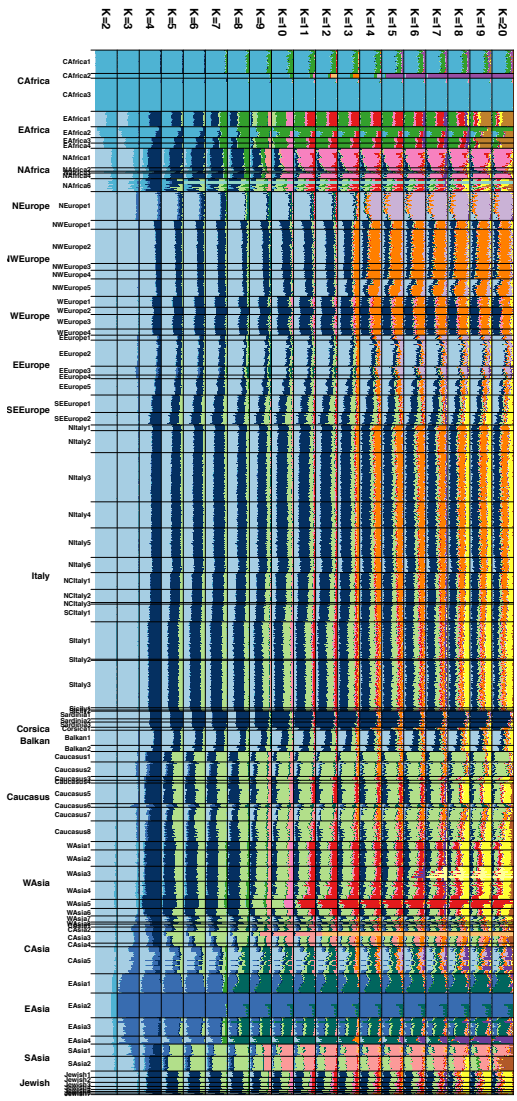


Figure B.2. Individual-level ADMIXTURE analysis of modern samples. Samples are grouped according to the genetic clusters inferred by the CP/fS pipeline.

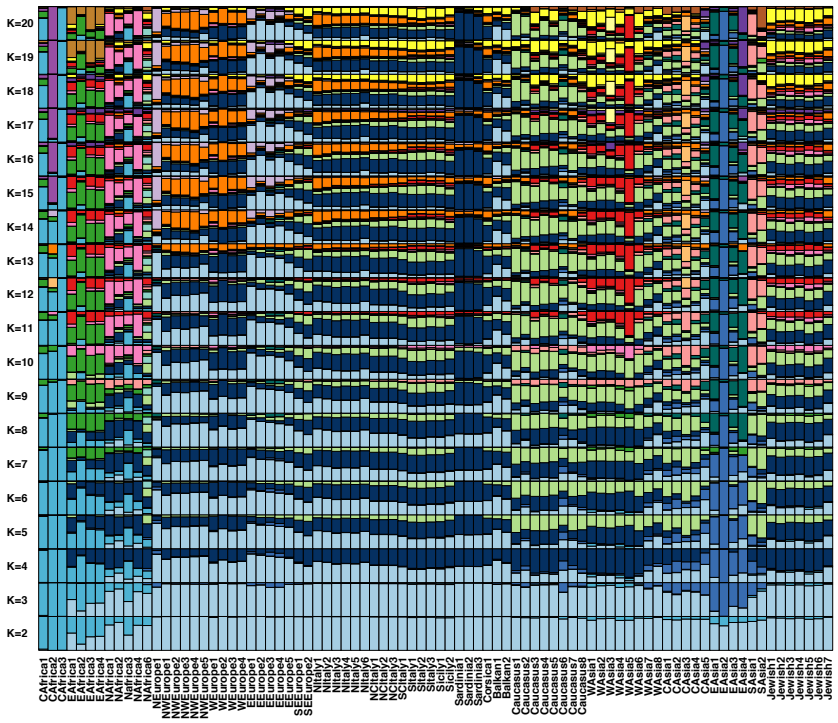


Figure B.3. ADMIXTURE analysis of modern samples averaged by cluster. Samples are grouped according to the genetic cluster inferred by the CP/fS pipeline.

	NAfrica1	NAfrica1.SE	EAsia2	EAsia2.SE	AN	AN.SE	IN	IN.SE	WHG	WHG.SE	EHG	EHG.SE	CHG	CHG.SE
Caucasus5	0.0000	0.0000	0.0129	0.0087	0.5751	0.0093	0.0979	0.0547	0.0000	0.0000	0.0229	0.0145	0.2912	0.0593
WAsia2	0.0907	0.0175	0.0134	0.0068	0.5264	0.0167	0.2204	0.0385	0.0000	0.0000	0.0007	0.0100	0.1484	0.0416
WAsia3	0.0400	0.0182	0.0030	0.0061	0.5621	0.0174	0.2892	0.0323	0.0000	0.0000	0.0145	0.0107	0.0912	0.0373
SAsia1	0.0000	0.0000	0.2575	0.0059	0.0457	0.0111	0.4853	0.0508	0.0000	0.0027	0.2018	0.0108	0.0097	0.0514
WAsia8	0.0000	0.0000	0.0364	0.0074	0.5109	0.0129	0.0000	0.0338	0.0000	0.0000	0.2304	0.0157	0.2222	0.0322
WAsia7	0.0000	0.0000	0.1146	0.0073	0.4082	0.0067	0.0707	0.0504	0.0000	0.0000	0.1350	0.0118	0.2715	0.0547
Caucasus8	0.0000	0.0000	0.0685	0.0076	0.4663	0.0087	0.1295	0.0466	0.0000	0.0000	0.0774	0.0109	0.2582	0.0487
SEEurope1	0.0000	0.0000	0.0000	0.0000	0.6261	0.0272	0.0000	0.0000	0.0004	0.0271	0.3030	0.0404	0.0706	0.0374
Caucasus2	0.0000	0.0000	0.0680	0.0054	0.4217	0.0142	0.0000	0.0000	0.0000	0.0000	0.1609	0.0163	0.3493	0.0276
EAfrica1	0.9548	0.0097	0.0452	0.0097	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EAfrica3	0.9443	0.0139	0.0557	0.0139	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EAfrica2	0.9109	0.0228	0.0891	0.0228	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EAfrica4	0.9398	0.0139	0.0602	0.0139	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SItaly3	0.0000	0.0000	0.0000	0.0000	0.5747	0.0235	0.0486	0.0330	0.1119	0.0382	0.0352	0.0455	0.2296	0.0454
Sardinia3	0.0000	0.0000	0.0000	0.0035	0.8121	0.0404	0.0000	0.0000	0.1312	0.0499	0.0147	0.0632	0.0420	0.0524
NItaly4	0.0000	0.0000	0.0000	0.0000	0.7216	0.0363	0.0000	0.0000	0.0191	0.0475	0.2374	0.0598	0.0219	0.0465
NItaly5	0.0000	0.0000	0.0000	0.0000	0.7026	0.0370	0.0000	0.0000	0.0320	0.0473	0.2203	0.0599	0.0451	0.0482
NItaly3	0.0000	0.0000	0.0000	0.0000	0.6614	0.0360	0.0000	0.0000	0.0434	0.0474	0.1965	0.0593	0.0987	0.0465
SCItaly1	0.0000	0.0000	0.0000	0.0000	0.6243	0.0365	0.0000	0.0000	0.0742	0.0465	0.1054	0.0580	0.1961	0.0470
SItaly1	0.0058	0.0222	0.0000	0.0000	0.5645	0.0153	0.1566	0.0452	0.1327	0.0131	0.0000	0.0000	0.1404	0.0443
NItaly2	0.0000	0.0000	0.0000	0.0000	0.6776	0.0364	0.0000	0.0000	0.0565	0.0474	0.2290	0.0602	0.0370	0.0474
NItaly6	0.0000	0.0000	0.0000	0.0000	0.7038	0.0129	0.0000	0.0000	0.0375	0.0228	0.2587	0.0276	0.0000	0.0102
NCItaly1	0.0000	0.0000	0.0000	0.0000	0.6706	0.0378	0.0000	0.0000	0.0543	0.0478	0.1496	0.0598	0.1254	0.0484
Corsica1	0.0000	0.0000	0.0000	0.0000	0.7172	0.0406	0.0000	0.0000	0.0796	0.0516	0.1128	0.0648	0.0904	0.0520
EEurope1	0.0000	0.0000	0.0000	0.0000	0.4012	0.0136	0.0000	0.0000	0.1729	0.0343	0.4259	0.0324	0.0000	0.0000
Caucasus1	0.0000	0.0000	0.0210	0.0055	0.4621	0.0152	0.0000	0.0000	0.0000	0.0000	0.0504	0.0156	0.4666	0.0262
SEEurope2	0.0000	0.0000	0.0000	0.0000	0.6535	0.0280	0.0000	0.0000	0.0013	0.0311	0.1941	0.0428	0.1512	0.0386
Balkan2	0.0000	0.0000	0.0000	0.0000	0.6840	0.0237	0.0000	0.0000	0.0000	0.0111	0.2949	0.0269	0.0211	0.0338
Jewish6	0.1536	0.0217	0.0000	0.0006	0.5707	0.0260	0.1888	0.0449	0.0040	0.0220	0.0571	0.0333	0.0259	0.0592
Jewish3	0.1464	0.0194	0.0000	0.0002	0.5610	0.0201	0.1809	0.0420	0.0274	0.0287	0.0415	0.0388	0.0429	0.0572
EEurope5	0.0000	0.0000	0.0000	0.0000	0.5703	0.0118	0.0000	0.0000	0.0901	0.0251	0.3395	0.0268	0.0000	0.0000
NItaly1	0.0000	0.0000	0.0000	0.0000	0.6840	0.0141	0.0000	0.0000	0.0827	0.0241	0.2333	0.0287	0.0000	0.0130
EAsia1	0.0000	0.0000	0.7720	0.0040	0.0003	0.0038	0.0000	0.0000	0.0000	0.0000	0.2277	0.0062	0.0000	0.0000
EEurope4	0.0000	0.0000	0.0200	0.0079	0.4036	0.0140	0.0000	0.0000	0.0000	0.0000	0.5760	0.0169	0.0004	0.0202
Caucasus4	0.0000	0.0000	0.0184	0.0060	0.4931	0.0180	0.0000	0.0000	0.0000	0.0000	0.0000	0.0074	0.4884	0.0249
Jewish1	0.0296	0.0195	0.0065	0.0076	0.5897	0.0168	0.1790	0.0317	0.0452	0.0236	0.0821	0.0262	0.0679	0.0329
Jewish4	0.0614	0.0228	0.0000	0.0010	0.6114	0.0256	0.1843	0.0396	0.0186	0.0271	0.0694	0.0370	0.0550	0.0522
Caucasus3	0.0000	0.0000	0.0167	0.0075	0.5147	0.0113	0.1730	0.0432	0.0000	0.0000	0.0000	0.0000	0.2955	0.0396
Caucasus7	0.0000	0.0000	0.0412	0.0044	0.3887	0.0143	0.0000	0.0000	0.0000	0.0000	0.1814	0.0154	0.3887	0.0268
CAsia5	0.0000	0.0000	0.2599	0.0048	0.2597	0.0133	0.0000	0.0000	0.0000	0.0000	0.4002	0.0131	0.0802	0.0230
EEurope2	0.0000	0.0000	0.0000	0.0000	0.5056	0.0122	0.0000	0.0000	0.1050	0.0306	0.3895	0.0316	0.0000	0.0000
Balkan1	0.0000	0.0000	0.0000	0.0000	0.6235	0.0111	0.0000	0.0000	0.0375	0.0235	0.3390	0.0270	0.0000	0.0000
EAsia3	0.0000	0.0000	0.6326	0.0036	0.0820	0.0099	0.0000	0.0000	0.0000	0.0000	0.2261	0.0082	0.0594	0.0156
NWEurope1	0.0000	0.0000	0.0000	0.0000	0.6213	0.0114	0.0000	0.0000	0.1330	0.0209	0.2457	0.0229	0.0000	0.0000
Sardinia1	0.0000	0.0000	0.0000	0.0022	0.8365	0.0345	0.0000	0.0000	0.1098	0.0443	0.0495	0.0522	0.0042	0.0406
Jewish2	0.0546	0.0228	0.0000	0.0023	0.6109	0.0247	0.2180	0.0353	0.0165	0.0303	0.0562	0.0381	0.0437	0.0464
NWEurope2	0.0000	0.0000	0.0000	0.0000	0.5894	0.0119	0.0000	0.0000	0.1315	0.0280	0.2791	0.0272	0.0000	0.0000
WEurope1	0.0672	0.0143	0.0000	0.0000	0.5324	0.0188	0.0000	0.0132	0.2011	0.0297	0.0608	0.0369	0.1384	0.0303
Jewish7	0.0063	0.0204	0.0000	0.0000	0.5987	0.0250	0.0932	0.0414	0.1359	0.0265	0.0399	0.0384	0.1260	0.0552
NAfrica6	0.9289	0.0056	0.0711	0.0056	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Nafrica3	0.9626	0.0124	0.0374	0.0124	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
NEurope1	0.0000	0.0000	0.0039	0.0075	0.3736	0.0126	0.0000	0.0000	0.0249	0.0253	0.5976	0.0308	0.0000	0.0000
EAsia4	0.0000	0.0000	0.6280	0.0069	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3720	0.0069	0.0000	0.0000
WAsia4	0.2807	0.0130	0.0063	0.0052	0.3885	0.0132	0.3207	0.0206	0.0000	0.0000	0.0000	0.0000	0.0038	0.0226
NWEurope5	0.0000	0.0000	0.0000	0.0000	0.5413	0.0129	0.0000	0.0000	0.1688	0.0299	0.2899	0.0285	0.0000	0.0000
EEurope3	0.0000	0.0000	0.0190	0.0069	0.3822	0.0137	0.0000	0.0000	0.0000	0.0000	0.5988	0.0154	0.0000	0.0000
WAsia6	0.0000	0.0000	0.0171	0.0079	0.5309	0.0103	0.2515	0.0482	0.0000	0.0000	0.0000	0.0000	0.2006	0.0452
NWEurope4	0.0000	0.0000	0.0000	0.0000	0.5426	0.0124	0.0000	0.0000	0.1938	0.0258	0.2635	0.0254	0.0000	0.0000
NWEurope3	0.0000	0.0000	0.0000	0.0000	0.5687	0.0120	0.0000	0.0000	0.1973	0.0258	0.2340	0.0261	0.0000	0.0000
WEurope3	0.0074	0.0134	0.0000	0.0000	0.5762	0.0216	0.0000	0.0000	0.2140	0.0369	0.0568	0.0442	0.1456	0.0278
WEurope2	0.0000	0.0000	0.0000	0.0000	0.6638	0.0135	0.0000	0.0000	0.2209	0.0254	0.1153	0.0249	0.0000	0.0000
WEurope4	0.0000	0.0000	0.0000	0.0000	0.6266	0.0362	0.0000	0.0000	0.1839	0.0499	0.1176	0.0622	0.0719	0.0482
SAsia2	0.0466	0.0167	0.0799	0.0045	0.1426	0.0118	0.4207	0.0257	0.0000	0.0000	0.1353	0.0092	0.1748	0.0252
CAsia3	0.0000	0.0000	0.1150	0.0062	0.1351	0.0109	0.1076	0.0397	0.0000	0.0000	0.2298	0.0082	0.4126	0.0405
CAfrica2	0.9168	0.0317	0.0832	0.0317	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WAsia5	0.2678	0.0187	0.0000	0.0000	0.4233	0.0126	0.3089	0.0201	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WAsia1	0.2603	0.0162	0.0019	0.0050	0.4082	0.0140	0.3193	0.0238	0.0000	0.0000	0.0000	0.0000	0.0104	0.0265
CAfrica3	0.9248	0.0305	0.0752	0.0305	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CAfrica1	0.9193	0.0290	0.0807	0.0290	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sardinia2	0.0000	0.0000	0.0000	0.0051	0.8381	0.0334	0.0000	0.0071	0.1381	0.0388	0.0122	0.0464	0.0116	0.0371
NCItaly3	0.0000	0.0000	0.0000	0.0000	0.6802	0.0483	0.0000	0.0000	0.0369	0.0579	0.1618	0.0746	0.1211	0.0632
NAfrica4	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sicily1	0.0573	0.0239	0.0000	0.0000	0.5871	0.0228	0.1034	0.0422	0.0673	0.0240	0.0566	0.0348	0.1283	0.0592
Sicily2	0.0655	0.0238	0.0000	0.0000	0.5660	0.0318	0.0730	0.0500	0.1339	0.0343	0.0236	0.0446	0.1379	0.0681
SItaly2	0.0000	0.0000	0.0000	0.0000	0.6457	0.0281	0.1164	0.0351	0.0544	0.0447	0.0546	0.0565	0.1290	0.0478
Jewish5	0.2031	0.0194	0.0000	0.0020	0.5330	0.0160	0.2183	0.0296	0.0000	0.0000	0.0310	0.0106	0.0146	0.0335
NAfrica2	0.6851	0.0152	0.0022	0.0038	0.2115	0.0151	0.0255	0.0206	0.0281	0.0215	0.0385	0.0267	0.0091	0.0318
NCItaly2	0.0000	0.0000	0.0000	0.0000	0.6793	0.0389	0.0000	0.0000	0.0525	0.0479	0.1511	0.0600	0.1171	0.0496
Caucasus6	0.0000	0.0000	0.2152	0.0062	0.3202	0.0122	0.0000	0.0000	0.0000	0.0000	0.2138	0.0145	0.2508	0.0263
CAsia4	0.0000	0.0000	0.1766	0.0061	0.2036	0.0082	0.0408	0.0427	0.0000	0.0000	0.1842	0.0096	0.3947	0.0405
CAsia1	0.0000	0.0000	0.1831	0.0072	0.2401	0.0092	0.1377	0.0512	0.0000	0.0000	0.1172	0.0114	0.3219	0.0530
CAsia2	0.0000	0.0000	0.2697	0.0080	0.2096	0.0088	0.0422	0.0370	0.0000	0.0000	0.1492	0.0082	0.3293	0.0371

Table B.3. CP/*NNLS* results from *ultimate* sources on all modern clusters using weighted jackknife bootstraps (one replicate per chromosome) along modern cluster for *ultimate* source. For each putative source, the mean value and the Standard error (SE) is reported.

	NAfrica1	NAfrica1.SE	EAsia2	EAsia2.SE	SBA	SBA.SE	ABA	ABA.SE	WHG	WHG.SE	EEN	EEN.SE
Caucasus5	0.0197	0.0227	0.0229	0.0096	0.0953	0.0189	0.8621	0.0248	0.0000	0.0000	0.0000	0.0000
WAsia2	0.1583	0.0194	0.0306	0.0085	0.0350	0.0163	0.7761	0.0199	0.0000	0.0000	0.0000	0.0000
WAsia3	0.1235	0.0217	0.0248	0.0088	0.0376	0.0171	0.8141	0.0213	0.0000	0.0000	0.0000	0.0000
SAsia1	0.1871	0.0101	0.3279	0.0037	0.3076	0.0146	0.1773	0.0205	0.0000	0.0000	0.0000	0.0000
WAsia8	0.0000	0.0029	0.0422	0.0067	0.3083	0.0571	0.4763	0.1222	0.0470	0.0418	0.1262	0.1079
WAsia7	0.0250	0.0183	0.1257	0.0071	0.2656	0.0140	0.5837	0.0207	0.0000	0.0000	0.0000	0.0000
Caucasus8	0.0443	0.0187	0.0849	0.0078	0.1736	0.0156	0.6971	0.0210	0.0000	0.0000	0.0000	0.0000
SEEurope1	0.0000	0.0000	0.0005	0.0031	0.4007	0.0495	0.1313	0.1042	0.0324	0.0367	0.4352	0.0916
Caucasus2	0.0000	0.0000	0.0773	0.0093	0.3199	0.0240	0.6028	0.0193	0.0000	0.0000	0.0000	0.0000
EAfrica1	0.9558	0.0096	0.0442	0.0096	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EAfrica3	0.9466	0.0144	0.0534	0.0144	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EAfrica2	0.9142	0.0234	0.0858	0.0234	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EAfrica4	0.9412	0.0140	0.0588	0.0140	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SItaly3	0.0000	0.0082	0.0000	0.0034	0.0548	0.0438	0.7984	0.0522	0.1468	0.0239	0.0000	0.0405
Sardinia3	0.0065	0.0159	0.0000	0.0000	0.0000	0.0052	0.2260	0.0506	0.1163	0.0140	0.6511	0.0509
NItaly4	0.0000	0.0000	0.0000	0.0000	0.2972	0.0502	0.0929	0.1027	0.0370	0.0375	0.5729	0.0917
NItaly5	0.0000	0.0000	0.0000	0.0000	0.2952	0.0519	0.1229	0.1068	0.0387	0.0378	0.5433	0.0944
NItaly3	0.0000	0.0000	0.0000	0.0000	0.2877	0.0482	0.2088	0.0994	0.0450	0.0358	0.4584	0.0892
SCItaly1	0.0172	0.0239	0.0000	0.0000	0.2441	0.0718	0.3198	0.1788	0.0315	0.0569	0.3874	0.1451
SItaly1	0.0551	0.0313	0.0001	0.0046	0.0808	0.0899	0.7292	0.1995	0.1008	0.0677	0.0340	0.1526
NItaly2	0.0000	0.0000	0.0000	0.0000	0.3061	0.0535	0.0843	0.1067	0.0652	0.0387	0.5444	0.0940
NItaly6	0.0000	0.0000	0.0000	0.0000	0.3342	0.0175	0.0000	0.0083	0.0468	0.0148	0.6190	0.0234
NCItaly1	0.0000	0.0000	0.0000	0.0000	0.1951	0.0467	0.3445	0.0991	0.0764	0.0345	0.3840	0.0896
Corsica1	0.0000	0.0000	0.0000	0.0000	0.1374	0.0530	0.2970	0.1086	0.0914	0.0371	0.4742	0.0952
EEurope1	0.0000	0.0000	0.0000	0.0000	0.4864	0.0236	0.0000	0.0000	0.2710	0.0210	0.2426	0.0171
Caucasus1	0.0000	0.0000	0.0375	0.0106	0.1868	0.0265	0.7758	0.0201	0.0000	0.0000	0.0000	0.0000
SEEurope2	0.0000	0.0000	0.0000	0.0023	0.2678	0.0487	0.3681	0.1058	0.0272	0.0365	0.3370	0.0959
Balkan2	0.0000	0.0000	0.0000	0.0000	0.3968	0.0447	0.0164	0.0945	0.0103	0.0331	0.5765	0.0857
Jewish6	0.1995	0.0190	0.0081	0.0080	0.0617	0.0612	0.6493	0.1206	0.0153	0.0433	0.0661	0.0916
Jewish3	0.2008	0.0155	0.0064	0.0072	0.0979	0.0314	0.5638	0.0612	0.0004	0.0223	0.1306	0.0479
EEurope5	0.0000	0.0000	0.0000	0.0000	0.4156	0.0184	0.0000	0.0000	0.1377	0.0152	0.4468	0.0182
NItaly1	0.0000	0.0000	0.0000	0.0000	0.3057	0.0158	0.0000	0.0051	0.0869	0.0150	0.6075	0.0207
EAsia1	0.0000	0.0000	0.7892	0.0034	0.2108	0.0034	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EEurope4	0.0000	0.0000	0.0378	0.0064	0.5820	0.0211	0.0000	0.0000	0.1570	0.0193	0.2232	0.0158
Caucasus4	0.0012	0.0154	0.0350	0.0096	0.1121	0.0220	0.8517	0.0176	0.0000	0.0000	0.0000	0.0000
Jewish1	0.0800	0.0250	0.0156	0.0079	0.1373	0.0708	0.6144	0.1427	0.0341	0.0499	0.1186	0.1055
Jewish4	0.1051	0.0231	0.0092	0.0079	0.0881	0.0649	0.6790	0.1329	0.0235	0.0439	0.0951	0.1032
Caucasus3	0.0530	0.0210	0.0424	0.0092	0.0567	0.0181	0.8479	0.0238	0.0000	0.0000	0.0000	0.0000
Caucasus7	0.0000	0.0007	0.0574	0.0093	0.3536	0.0223	0.5890	0.0176	0.0000	0.0000	0.0000	0.0000
CAsia5	0.0000	0.0000	0.2773	0.0043	0.5074	0.0162	0.0000	0.0000	0.0479	0.0158	0.1673	0.0141
EEurope2	0.0000	0.0000	0.0000	0.0000	0.4577	0.0225	0.0000	0.0000	0.1790	0.0173	0.3633	0.0187
Balkan1	0.0000	0.0000	0.0000	0.0000	0.4146	0.0201	0.0000	0.0000	0.0823	0.0154	0.5031	0.0185
EAsia3	0.0000	0.0000	0.6411	0.0036	0.3293	0.0092	0.0000	0.0000	0.0000	0.0000	0.0296	0.0095
NWEurope1	0.0000	0.0000	0.0000	0.0000	0.3177	0.0178	0.0000	0.0000	0.1450	0.0147	0.5373	0.0161
Sardinia1	0.0000	0.0101	0.0000	0.0000	0.0115	0.0566	0.1817	0.0933	0.1127	0.0375	0.6941	0.0761
Jewish2	0.0975	0.0241	0.0093	0.0094	0.0520	0.0384	0.8054	0.0734	0.0358	0.0253	0.0000	0.0483
NWEurope2	0.0000	0.0000	0.0000	0.0000	0.3401	0.0190	0.0000	0.0000	0.1661	0.0157	0.4938	0.0155
WEurope1	0.0710	0.0187	0.0000	0.0000	0.0904	0.0527	0.3602	0.1115	0.2096	0.0360	0.2689	0.0832
Jewish7	0.0420	0.0206	0.0000	0.0000	0.1274	0.0457	0.4930	0.0882	0.0991	0.0287	0.2385	0.0657
NAfrica6	0.9289	0.0055	0.0711	0.0055	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Nafrica3	0.9647	0.0126	0.0353	0.0126	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
NEurope1	0.0000	0.0000	0.0276	0.0058	0.5440	0.0233	0.0000	0.0000	0.2290	0.0226	0.1994	0.0153
EAsia4	0.0000	0.0000	0.6605	0.0063	0.2830	0.0351	0.0000	0.0000	0.0566	0.0315	0.0000	0.0000
WAsia4	0.3826	0.0167	0.0337	0.0059	0.0000	0.0000	0.5836	0.0170	0.0000	0.0000	0.0000	0.0000
NWEurope5	0.0000	0.0000	0.0000	0.0000	0.3492	0.0211	0.0000	0.0000	0.2100	0.0177	0.4408	0.0161
EEurope3	0.0000	0.0000	0.0406	0.0060	0.5643	0.0234	0.0000	0.0000	0.1900	0.0226	0.2051	0.0160
WAsia6	0.0719	0.0232	0.0468	0.0093	0.0274	0.0196	0.8538	0.0237	0.0000	0.0000	0.0000	0.0000
NWEurope4	0.0000	0.0000	0.0000	0.0000	0.3220	0.0219	0.0000	0.0000	0.2261	0.0167	0.4519	0.0142
NWEurope3	0.0000	0.0000	0.0000	0.0000	0.3032	0.0205	0.0000	0.0000	0.2102	0.0150	0.4865	0.0152
WEurope3	0.0533	0.0137	0.0000	0.0000	0.2571	0.0367	0.0000	0.0794	0.0967	0.0251	0.5929	0.0642
WEurope2	0.0000	0.0000	0.0000	0.0000	0.1676	0.0211	0.0000	0.0000	0.1975	0.0174	0.6350	0.0165
WEurope4	0.0000	0.0000	0.0000	0.0000	0.1582	0.0514	0.2093	0.0961	0.1859	0.0375	0.4465	0.0845
SAsia2	0.2249	0.0131	0.1456	0.0043	0.2605	0.0160	0.3690	0.0209	0.0000	0.0000	0.0000	0.0000
CAsia3	0.0750	0.0142	0.1558	0.0059	0.4539	0.0168	0.3154	0.0230	0.0000	0.0000	0.0000	0.0000
CAfrica2	0.9190	0.0328	0.0810	0.0328	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WAsia5	0.3672	0.0211	0.0202	0.0065	0.0000	0.0000	0.6126	0.0199	0.0000	0.0000	0.0000	0.0000
WAsia1	0.3589	0.0199	0.0300	0.0061	0.0000	0.0022	0.6111	0.0189	0.0000	0.0000	0.0000	0.0000
CAfrica3	0.9285	0.0312	0.0715	0.0312	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CAfrica1	0.9223	0.0295	0.0777	0.0295	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sardinia2	0.0053	0.0177	0.0000	0.0000	0.0000	0.0112	0.1544	0.0590	0.1132	0.0161	0.7271	0.0546
NCItaly3	0.0017	0.0154	0.0000	0.0000	0.2042	0.0421	0.3208	0.0996	0.0559	0.0302	0.4173	0.0880
NAfrica4	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sicily1	0.1057	0.0160	0.0000	0.0030	0.1745	0.0383	0.4561	0.0598	0.0177	0.0250	0.2460	0.0448
Sicily2	0.1088	0.0217	0.0000	0.0007	0.1278	0.0591	0.4294	0.1004	0.0789	0.0361	0.2551	0.0719
SItaly2	0.0408	0.0187	0.0018	0.0063	0.1595	0.0439	0.5457	0.0821	0.0123	0.0269	0.2399	0.0618
Jewish5	0.2476	0.0221	0.0102	0.0074	0.0297	0.0206	0.7125	0.0301	0.0000	0.0070	0.0000	0.0182
NAfrica2	0.6822	0.0133	0.0041	0.0038	0.0041	0.0390	0.1761	0.0863	0.0563	0.0288	0.0772	0.0691
NCItaly2	0.0000	0.0000	0.0000	0.0000	0.2048	0.0448	0.3175	0.0952	0.0669	0.0326	0.4108	0.0859
Caucasus6	0.0000	0.0000	0.2243	0.0081	0.3719	0.0171	0.3918	0.0640	0.0000	0.0000	0.0120	0.0557
CAsia4	0.0418	0.0132	0.2027	0.0047	0.3776	0.0168	0.3780	0.0236	0.0000	0.0000	0.0000	0.0000
CAsia1	0.0717	0.0160	0.2151	0.0064	0.2598	0.0145	0.4534	0.0214	0.0000	0.0000	0.0000	0.0000
CAsia2	0.0334	0.0131	0.2918	0.0060	0.3035	0.0115	0.3713	0.0171	0.0000	0.0000	0.0000	0.0000

Legend on next page.

Table B.4. CP/*NNLS* results from *proximate* sources on all modern clusters using weighted jackknife bootstraps (one replicate per chromosome) along modern cluster for *proximate* source. For each putative source, the mean value and the Standard error (SE) is reported.

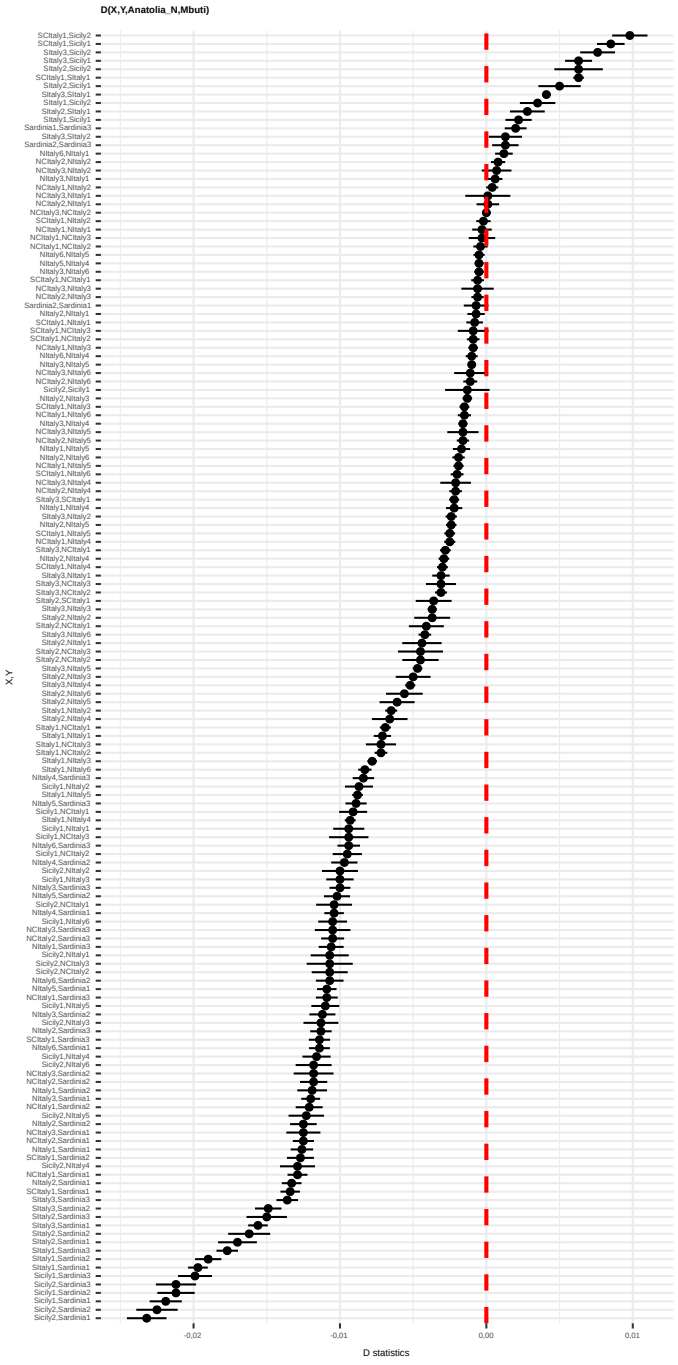


Figure B.4. ***D*-statistics in the form *D*(X,Y, AN,Mbuti) for all the possible pairs of Italian clusters.**

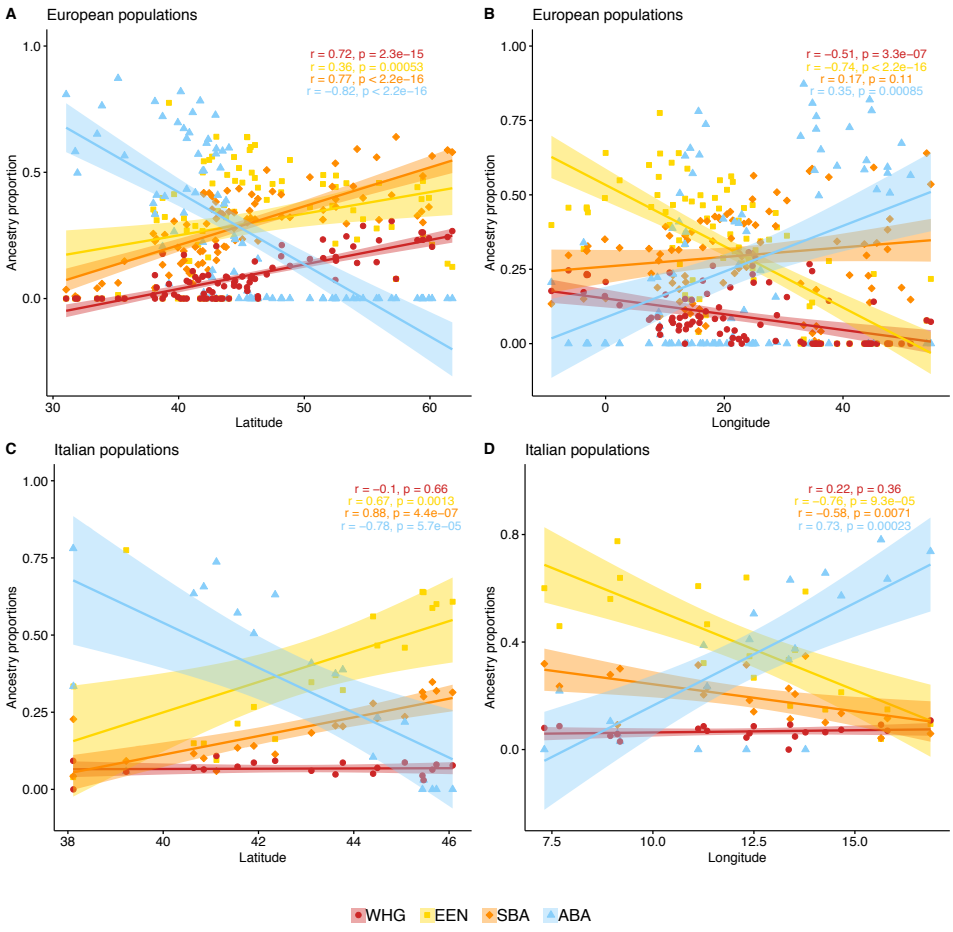


Figure B.5. Spearman correlations between *proximate* sources ancestry components including WHG and considering ABA as SEE source, estimated with the CP/*NNLS* analysis and geography. (A) Correlations between *proximate* sources ancestry (ABA as SEE source) in European populations and latitude, including WHG as source. **(B)** Correlations between *proximate* sources ancestry (ABA as SEE source) in European populations and longitude, including WHG as source. **(C)** Correlations between *proximate* sources ancestry (ABA as SEE source) in Italian populations and latitude, including WHG as source. **(D)** Correlations between *proximate* sources ancestry (ABA as SEE source) in Italian populations and longitude, including WHG as source.

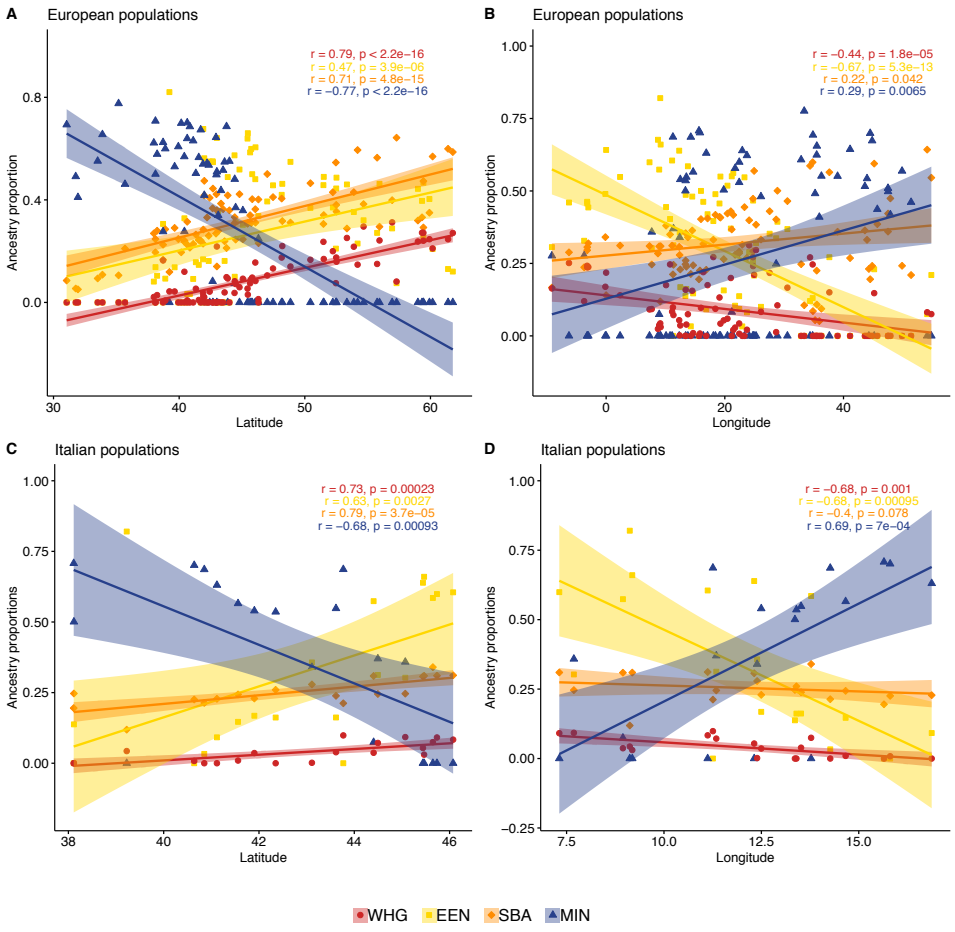


Figure B.6. Spearman correlations between *proximate* sources ancestry components including WHG and considering MIN as SEE source, estimated with the CP/*NNLS* analysis and geography. (A) Correlations between *proximate* sources ancestry (MIN as SEE source) in European populations and latitude, including WHG as source. **(B)** Correlations between *proximate* sources ancestry (MIN as SEE source) in European populations and longitude, including WHG as source. **(C)** Correlations between *proximate* sources ancestry (MIN as SEE source) in Italian populations and latitude, including WHG as source. **(D)** Correlations between *proximate* sources ancestry (MIN as SEE source) in Italian populations and longitude, including WHG as source.

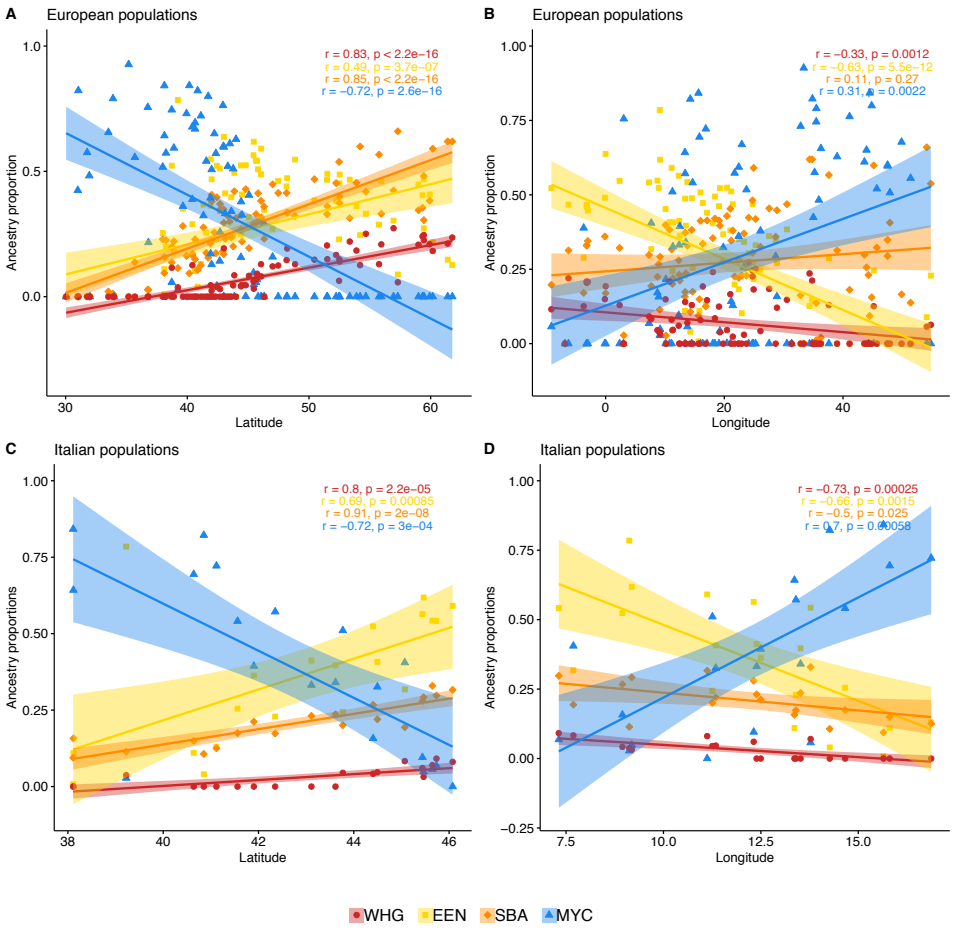


Figure B.7. Spearman correlations between *proximate* sources ancestry components including WHG and considering MYC as SEE source, estimated with the CP/*NNLS* analysis and geography. (A) Correlations between *proximate* sources ancestry (MYC as SEE source) in European populations and latitude, including WHG as source. **(B)** Correlations between *proximate* sources ancestry (MYC as SEE source) in European populations and longitude, including WHG as source. **(C)** Correlations between *proximate* sources ancestry (MYC as SEE source) in Italian populations and latitude, including WHG as source. **(D)** Correlations between *proximate* sources ancestry (MYC as SEE source) in Italian populations and longitude, including WHG as source.

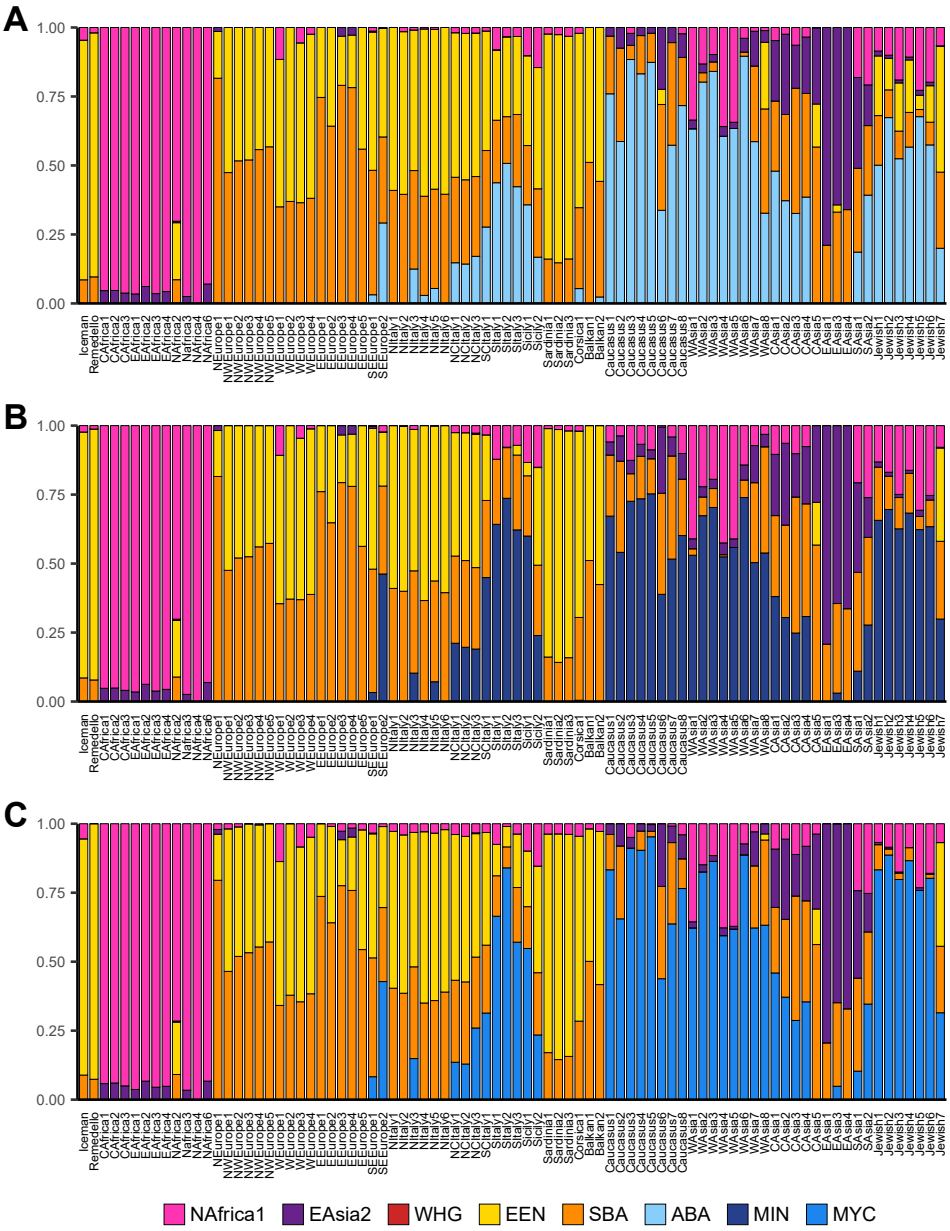
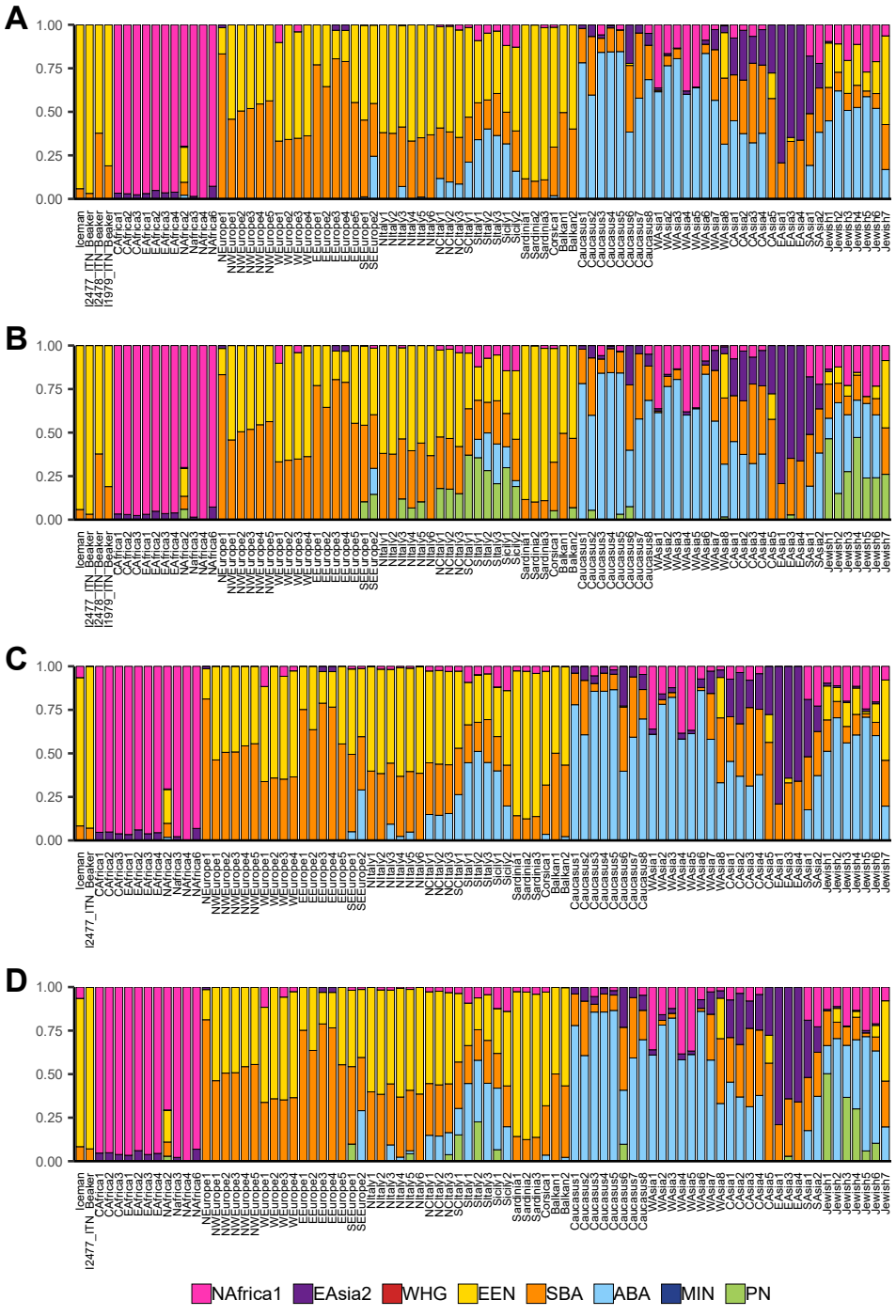


Figure B.8. CP/*NNLS* results for *proximate* sources, including the Iceman and a Remedello samples as recipients. **(A)** *Proximate* sources analysis considering ABA as SEE source and excluding WHG from sources. **(B)** *Proximate* sources analysis considering Minoan as SEE source and excluding WHG from sources. **(C)** *Proximate* sources analysis considering Mycenaean as SEE source and excluding WHG from sources.



Legend on next page.

Figure B.9. CP/*NNLS* results for *proximate* sources plus a PN sample, including modern-day clusters and the Iceman and ITN Bell Beaker samples as recipients. (A) *Proximate* sources analysis considering ABA as SEE source, excluding WHG as source and three ITN Bell Beaker samples as recipients. (B) *Proximate* sources analysis considering ABA and PN as SEE sources, excluding WHG as source and three ITN Bell Beaker samples as recipients. (C) *Proximate* sources analysis considering ABA as SEE source, excluding WHG as source and a ITN Bell Beaker sample (I2477) as recipient. (D) *Proximate* sources analysis considering ABA and PN as SEE sources, excluding WHG as source and a ITN Bell Beaker sample (I2477) as recipient.

Modern tales

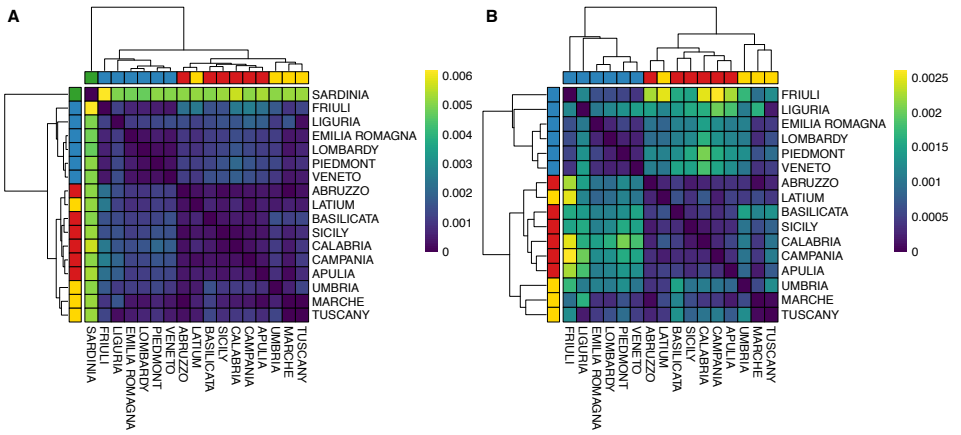


Figure C.1. F_{ST} **values among Italian regions.** Heatmap representing the F_{ST} values computed for each pairwise comparison between Italian regions, including **(A)** and excluding **(B)** Sardinia. We excluded the region Molise as it was represented by just one individual.

	ABRUZZO	BASILICATA	CALABRIA	CAMPANIA	EMILIA ROMAGNA	FRIULI	LATIUM	LIGURIA	LOMBARDY	MARCHE	MOLISE	PIEDMONT	APULIA	SARDINIA	SICILY	TUSCANY	UMBRIA	VENETO
ABRUZZO	1	0.000447	0.000284	0.000339	0.000989	0.002254	0.000279	0.001473	0.000928	5.60E-05	0.006741	0.001226	0.000388	0.005698	0.000185	0.000273	0.000383	0.001112
BASILICATA	0.000447	1	0.000489	0.000327	0.001145	0.001579	0.000661	0.001416	0.001131	0.001193	0.007533	0.00129	0.000511	0.005057	0.000314	0.001392	0.001485	0.001525
CALABRIA	0.000284	0.00072	1	0.000301	0.001475	0.002481	0.00072	0.001627	0.001649	0.000438	0.00842	0.002067	0.000522	0.005641	0.000169	0.001013	0.001078	0.001843
CAMPANIA	0.000489	0.000212	0.000617	1	0.001254	0.002598	0.000212	0.002017	0.001315	0.000427	0.008907	0.001606	0.000349	0.00512	0.000286	0.001086	0.001009	0.001532
EMILIA ROMAGNA	0.000339	0.000848	0.00091	0.008097	1	0.00063	0.000848	0.001064	0.000265	0.000387	0.008759	0.00053	0.001103	0.004862	0.001136	0.000467	0.001138	0.000346
FRIULI	0.000337	0.00249	0.001208	0.008759	0.001322	1	0.00249	0.00122	0.000617	0.000972	0.011074	0.000542	0.002242	0.006174	0.001509	0.001248	0.001716	0.000399
LATIUM	0.000301	0.001473	5.60E-05	0.011074	0.000245	0.000183	1	0.001189	0.00091	0.000578	0.009064	0.001166	0.000501	0.005126	0.000373	0.000434	0.000461	0.001023
LIGURIA	0.000989	0.001416	0.001193	0.009064	0.000652	0.008388	0.005211	1	0.001208	0.001653	0.009996	0.001322	0.001843	0.004983	0.001575	0.000196	0.001087	0.001442
LOMBARDY	0.001145	0.001627	0.000438	0.009996	0.009326	0.001353	0.013203	0.001575	1	0.000336	0.008932	0.000245	0.00117	0.004724	0.001236	0.000634	0.001084	0.000237
MARCHE	0.001475	0.002017	0.000427	0.008932	0.000388	0.005098	0.005099	0.001236	0.001086	1	0.005964	0.000652	0.000183	0.005211	0.000606	5.60E-05	0.000573	0.000386
MOLISE	0.001254	0.001064	0.000387	0.005964	0.000511	0.005057	0.00532	0.000606	0.000467	0.000744	1	0.000326	0.000388	0.012203	0.000797	0.000345	0.000525	0.000316
PIEDMONT	0.002254	0.00122	0.000972	0.001226	0.000522	0.005641	0.000185	0.008797	0.001248	0.005206	0.001716	1	0.001353	0.005099	0.001566	0.00091	0.001297	0.000269
APULIA	0.001579	0.001189	0.000578	0.00129	0.000349	0.00512	0.000314	0.001566	0.000434	0.00081	0.000461	0.000861	1	0.00532	0.00032	0.000744	0.000861	0.001285
SARDINIA	0.002481	0.000928	0.001653	0.002067	0.001103	0.004862	0.000169	0.00032	0.000196	0.000383	0.001087	0.005091	0.001843	1	0.002241	0.005206	0.005091	0.005005
SICILY	0.002598	0.001131	0.000336	0.001606	0.002242	0.006174	0.000286	0.005241	0.000634	0.001485	0.001084	0.001	0.001532	0.001442	1	0.00081	0.001	0.001497
TUSCANY	0.00063	0.001649	0.006741	0.00053	0.009501	0.005126	0.001136	0.000273	5.60E-05	0.001078	0.000573	0.001249	0.000346	0.009237	0.000269	1	0.001249	0.000489
UMBRIA	0.000279	0.001315	0.007533	0.000542	0.001843	0.004983	0.001509	0.001302	0.000345	0.001009	0.008525	0.001112	0.000399	0.000386	0.001285	0.001497	1	0.00103
VENETO	0.000661	0.000265	0.00842	0.001166	0.00117	0.004724	0.000373	0.001013	0.00091	0.001138	0.001297	0.001525	0.001023	0.008016	0.005005	0.000489	0.00103	1

Table C.1. Pairwise F_{ST} between Italian administrative regions except Molise, for which just one individual was available.

	ABRUZZO	BASILICATA	CALABRIA	CAMPANIA	EMILIA ROMAGNA	FRIULI	LATIUM	LIGURIA	LOMBARDY	MARCHE	MOLISE	PIEDMONT	APULIA	SARDINIA	SICILY	TUSCANY	UMBRIA	VENETO
ABRUZZO	1	0.000538	0.000211	0.000201	0.000171	0.00107	0.000238	0.000614	0.000167	0.000283	0.004518	0.000181	0.000171	0.000371	0.000184	0.000404	0.00042	0.000188
BASILICATA	0.000538	1	0.000456	0.000452	0.000431	0.00132	0.000514	0.000954	0.00043	0.000539	0.00464	0.000454	0.00042	0.000607	0.000443	0.000661	0.000679	0.00045
CALABRIA	0.000211	0.000187	1	0.000126	0.000105	0.001006	0.000187	0.000539	9.70E-05	0.000203	0.004371	0.00013	9.50E-05	0.00273	0.000115	0.000333	0.000344	0.000125
CAMPANIA	0.000456	0.000157	0.000943	1	9.20E-05	0.000998	0.000157	0.000523	8.30E-05	0.000189	0.004429	0.000111	8.70E-05	0.00028	9.50E-05	0.000307	0.000343	0.000106
EMILIA ROMAGNA	0.000201	0.000136	0.000129	0.004429	1	0.000955	0.000136	0.000492	3.50E-05	0.000172	0.004413	6.30E-05	5.50E-05	0.00023	7.60E-05	0.00028	0.000307	6.20E-05
FRIULI	0.000452	0.001089	0.000486	0.004413	0.000505	1	0.001089	0.001356	0.000943	0.001112	0.004694	0.000961	0.000989	0.001123	0.000993	0.001167	0.00124	0.000966
LATIUM	0.000126	0.000614	0.000283	0.004694	4.90E-05	0.00017	1	0.000575	0.000129	0.000248	0.004357	0.000153	0.000138	0.000304	0.000158	0.000369	0.000383	0.00015
LIGURIA	0.000171	0.000954	0.000539	0.004357	0.000187	0.004372	0.000352	1	0.000486	0.000612	0.004599	0.000505	0.000496	0.0006	0.00051	0.000702	0.000736	0.000516
LOMBARDY	0.000431	0.000539	0.000203	0.004599	0.004422	7.60E-05	0.004484	0.00051	1	0.000161	0.004362	4.90E-05	5.20E-05	0.002225	6.80E-05	0.000273	0.000296	5.10E-05
MARCHE	0.000105	0.000523	0.000189	0.004362	0.000171	0.000371	0.000252	6.80E-05	0.000307	1	0.004438	0.000187	0.00017	0.000352	0.000188	0.000403	0.000425	0.000182
MOLISE	9.20E-05	0.000492	0.000172	0.004438	0.00042	0.000607	0.000246	0.000188	0.000284	0.000284	1	0.004422	0.004372	0.004484	0.004349	0.004518	0.004619	0.004401
PIEDMONT	0.00107	0.001356	0.001112	0.000181	9.50E-05	0.000273	0.000184	0.004349	0.001167	0.000428	0.00124	1	7.60E-05	0.002252	9.40E-05	0.000291	0.000327	7.30E-05
APULIA	0.00132	0.000575	0.000248	0.000454	8.70E-05	0.00028	0.000443	9.40E-05	0.000369	0.000299	0.000383	0.000305	1	0.000246	7.20E-05	0.000284	0.000305	7.40E-05
SARDINIA	0.001006	0.000167	0.000612	0.00013	5.50E-05	0.00023	0.000115	7.20E-05	0.000702	0.00042	0.00076	0.000469	0.000125	1	0.000272	0.000428	0.000469	0.000256
SICILY	0.000998	0.00043	0.000161	0.000111	0.000989	0.001123	9.50E-05	0.000272	0.000273	0.000679	0.000296	0.000334	0.000106	0.000516	1	0.000299	0.000334	9.90E-05
TUSCANY	0.000955	9.70E-05	0.004518	6.30E-05	0.000138	0.000304	7.60E-05	0.000404	0.000403	0.000344	0.000425	0.000546	6.20E-05	5.10E-05	7.30E-05	1	0.000546	0.000292
UMBRIA	0.000238	8.30E-05	0.00464	0.000961	0.000496	0.00066	0.000993	0.000661	0.004518	0.000343	0.004619	0.000188	0.000966	0.000182	7.40E-05	9.90E-05	1	0.000328
VENETO	0.000514	3.50E-05	0.004371	0.000153	5.20E-05	0.000225	0.000158	0.000333	0.000291	0.000307	0.000327	0.00045	0.00015	0.004401	0.000256	0.000292	0.000328	1

Table C.2. Standard errors of the F_{ST} estimates between Italian administrative regions except Molise, for which just one individual was available.

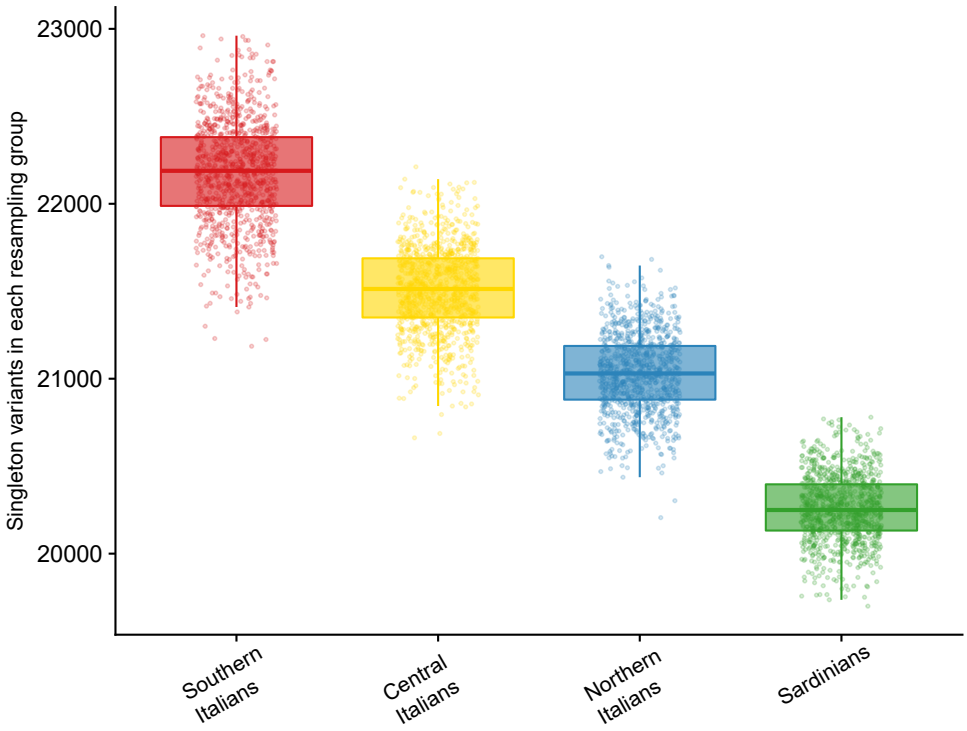


Figure C.2. Strip and boxplot of the amount of low-frequency variants in a thousand resampling of ten individuals from each macro-area.

Future tales

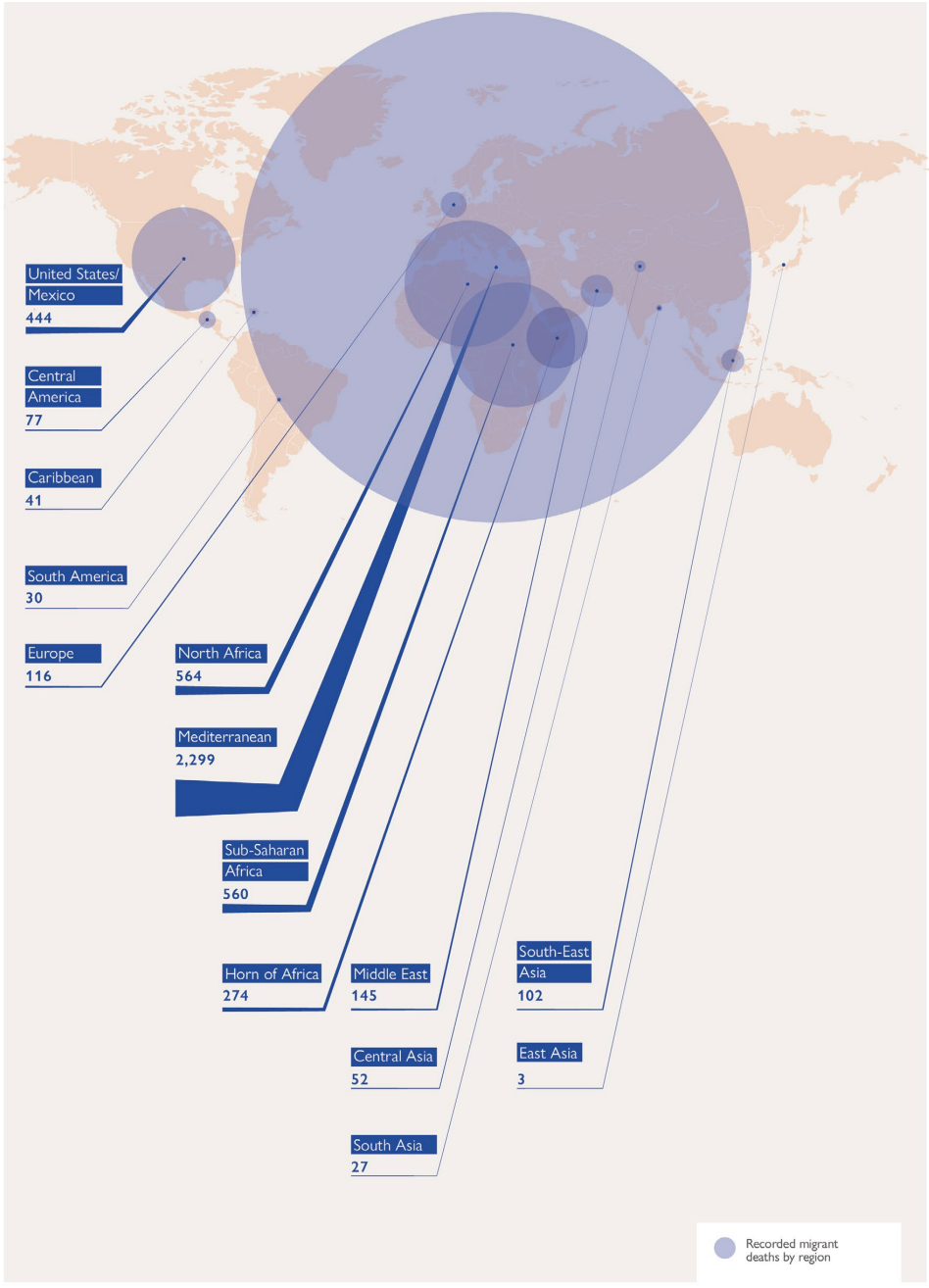


Figure C.3. Migrant deaths recorded worldwide in 2018. Image taken from Missing Migrants Project, 2019.

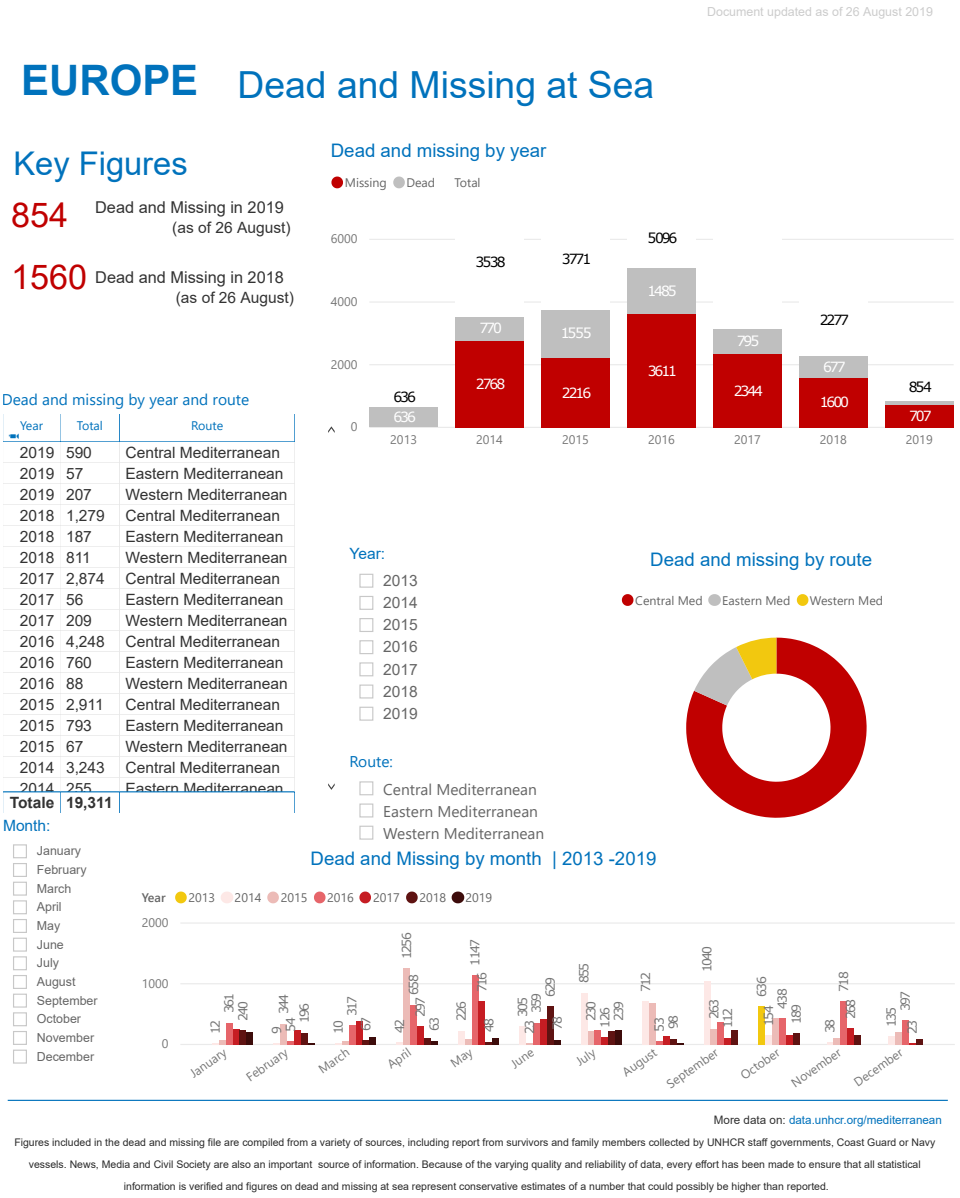


Figure C.4. UNHCR (Office of the United Nations High Commissioner for Refugees) report about dead and missing migrants in the Mediterranean Sea from 2013 to 2019.

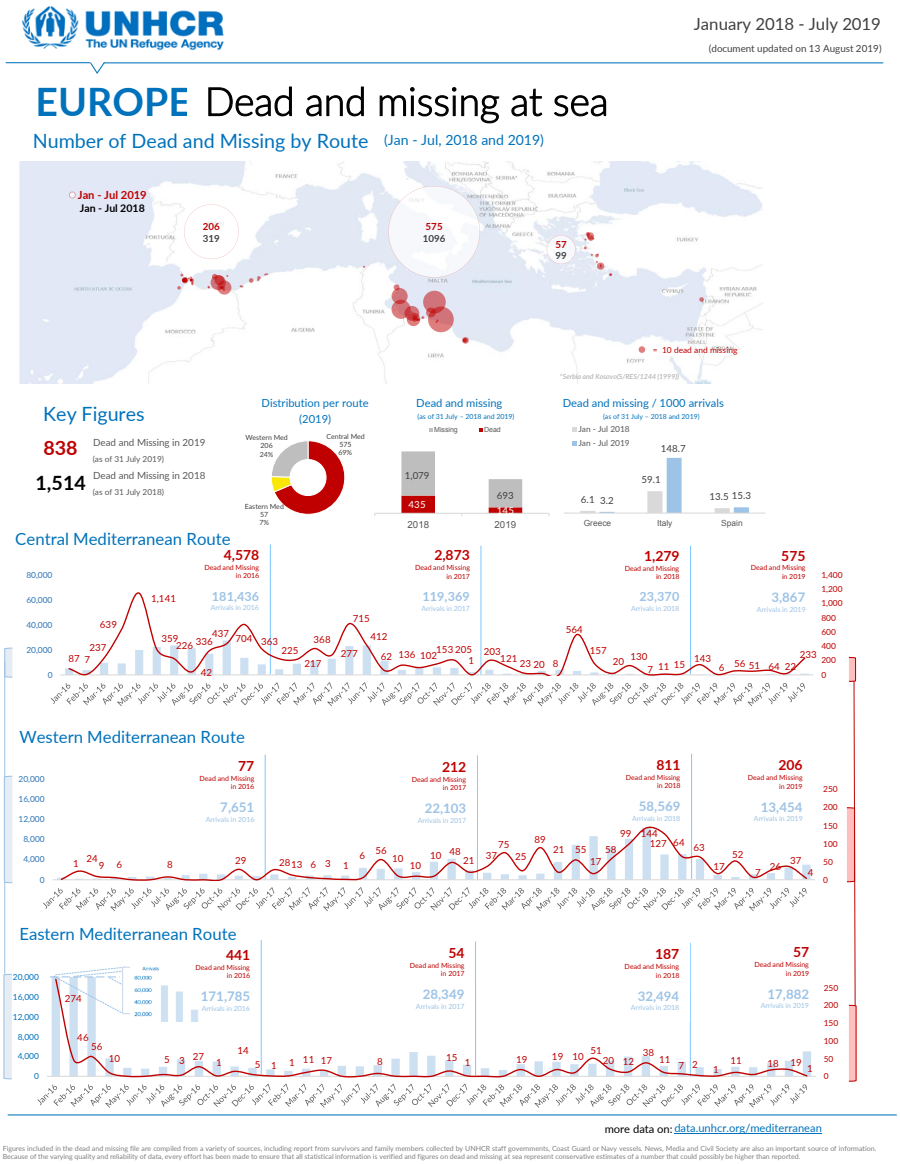


Figure C.5. January 2018-July 2019 UNHCR report about dead and missing migrants in the Mediterranean Sea.

Measures of Marker Information Content

Criterion	Description	Features	Limitations
Absolute allele frequency difference (δ)	$\lvert p_{1i} - p_{2i} \rvert$	Is related to amount of linkage disequilibrium in an admixture model (Chakraborty and Weiss 1988); is related to probability of correct assignment in a multilocus no-admixture model (Riech et al. 2002); is related to Fisher information curvature criterion for $K = 2$ (eq. [18]); is related to *ORCA* for $K = 2$ (eq. [11])	Requires that only two populations be possible sources; does not take into account all available information about allele frequencies (Stephens et al. 1999; Campbell et al. 2003); statistical features do not apply to the Shriver et al. (1997) multiallelic extension of δ
F_{st}	Excess in probability of identity of alleles from the same population compared with randomly chosen alleles (Excoffier 2001, for example)	Is related, for biallelic markers, to the quotient of expected posterior and prior variance of ancestry in a population equally admixed from two sources[a] (McKeigue 1998; Molokhia et al. 2003)	Performs only slightly better than random markers (Rosenberg et al. 2001)
Expected heterozygosity[b]	$1 - \sum_{i=1}^{N} p_i^2$ (bias correction can be applied in estimation from data)	Performs better than random markers (Rosenberg et al. (2001) and fig. 3)	Measures the amount of variation but not the differences across populations
Number of alleles[b]	N	Performs better than random markers (Rosenberg et al. 2001)	Measures the amount of variation but not the differences across populations; is useful only for multiallelic markers that have variation in number of alleles
Fisher information curvature criterion	Reciprocal of the largest eigenvalue of the information matrix for maximum-likelihood estimation of ancestry coefficients (Gomulkiewicz et al. 1990; Millar 1991)	Enables predictions about approximate variances of ancestry estimates (see "Number of Markers" subsection of the "Theory" section); information matrix is additive across loci that are independent within populations	Depends on unknown ancestry coefficients and requires computation for many possible parameter values; largest eigenvalue gives an upper bound that might not be generally applicable across the parameter space
Pairwise Kullback-Leibler divergence[c]	$\sum_{i=1}^{N} p_{1i} \log \frac{p_{1i}}{p_{2i}} + p_{2i} \log \frac{p_{2i}}{p_{1i}}$ Bremner 1998; Smith et al. 2001; Anderson and Thompson 2002)	Provides a natural measure, for $K = 2$, of average potential for assignment of an allele to one population compared with the other; has a natural multilocus extension; enables measurement of contributions of specific alleles	Requires that only two populations be possible sources; has upwardly biased estimates in small samples
Informativeness for assignment (I_n)	Equation (4)	Provides a natural measure of potential for assignment of an allele to one population compared with the "average" population; has a natural multilocus extension; enables measurement of contributions of specific alleles or populations; performs better than random or highly heterozygous markers (fig. 3)	Has upwardly biased estimates in small samples
Informativeness for ancestry coefficients (I_a)	Equation (14)	Provides a natural measure of potential for assignment of an allele to a point on the set of all possible ancestry coefficient vectors; has a natural multilocus extension; enables measurement of contributions of specific alleles	Has upwardly biased estimates in small samples; is difficult to compute in samples with populations of equal sample size
Optimal rate of correct assignment (ORCA)	Equation (10)	Gives the probability of correct assignment of an allele using the decision rule with lowest risk; has a natural multilocus extension (eq. [12]); enables measurement of contributions of specific alleles	Has upwardly biased estimates in small samples

NOTE.—Notation is defined in the "Theory" section. All criteria apply to multiallelic loci in any number of populations, except where specified.

[a] A multiallelic statistic related to this ratio was suggested by Molokhia et al. (2003).

[b] Can also be calculated by using average values of the statistic across populations rather than by using values for the whole collection of populations.

[c] A similar statistic based on genotype frequencies was suggested by Shriver et al. (1997). Some authors multiply by a factor of 1/2 in the formula for this statistic.

Table C.3. Approaches used for measuring the genetic markers contribution to ancestry inference. Image taken from Rosenberg *et al.*, 2003.

AIMs selection pipeline

Cleaning steps, descriptive analyses and cross-validation

```
import pandas
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt
import numpy
import pickle
from dataset import *
from freqs import *
import gzip
from sklearn.model_selection import StratifiedKFold
import seaborn as sns
from matplotlib.backends.backend_pdf import PdfPages
from sklearn.ensemble import RandomForestClassifier
import itertools
from tree_search import any_first_search

FILES = ["kgp/kgp.pruned_500_50_0.2.BEB_CHB_PEL_GBR_YRI"]

search_methods = dict(
first_good=any_first_search,
)
classifiers = dict(
rf=RandomForestClassifier,
nb=GenotypicNB,
)

kfold = 5
mafs = [0, 5, 10, 20]
mafs = [10]
datasets = ['kgp/kgp.pruned_500_50_0.2.BEB_CHB_PEL_GBR_YRI']
prioritization_targets = ['fs_pIn'] + expand('maxmaf_{maf}.fs_rf',
     maf=mafs)
classifier_targets = ['cl=rf(max_depth=10,n_estimators=200)', 'cl=
    nb(hwe=True)']

search_input_stem = 'ms_search={search}({search_params})_cl={cl}({
    cl_params})'
search_output_stem = 'ms_search={search,' + '|'.join(
    search_methods) + '}({search_params})_cl={cl,' + '|'.join(
    classifiers) + '}({cl_params})'

aim_suffixes = expand('{prio}.ms_search={search}({search_params})_
    {cl}.pickle.gz', prio=prioritization_targets, search='
    first_good', search_params='min_score=0.99,max_markers=15', cl
```

```
    =classifier_targets)

def bfile(path):
  return path + '.bed', path + '.bim', path + '.fam'

rule pca:
  input:
    bfile("{file}")
  output:
    "{file}.eigenvec"
  params:
    bfile=lambda wildcards, input: input[0][:-4],
    ofile=lambda wildcards, output: output[0][:-9],
  shell:
    "plink --bfile {params.bfile} --chr 1-22 --geno 0.05 --mind
    0.05 --hwe 1e-10 --pca --out {params.ofile}"

rule pca_plot:
  input: "{file}.eigenvec", "/scratch/shared/AIM/AIM_BITS/kgp3.
    pheno"
  output: "{file}.pca.pdf"
  run:
    pca = pandas.read_csv(input[0], sep=" ", index_col=1)
    labels = pandas.read_csv(input[1], sep="\t", index_col=0)
    pca['labels'] = labels[1]
    with PdfPages(output[0]) as pdf:
      ax = sns.scatterplot(x=2, y=3, hue="labels", data=pca)
      ax.set(xlabel='PC1', ylabel='PC2')
      pdf.savefig()
      plt.close()
      ax = sns.scatterplot(x=2, y=4, hue="labels", data=pca)
      ax.set(xlabel='PC1', ylabel='PC3')
      pdf.savefig()
      plt.close()
      ax = sns.scatterplot(x=3, y=4, hue="labels", data=pca)
      ax.set(xlabel='PC2', ylabel='PC3')
      pdf.savefig()
      plt.close()

rule admixture:
  input: "{file}.bed"
  output: expand("{{file}}.admixture_K{{K,\d+}}.{ext}", ext=['P',
    'Q'])
  threads: 32
  log: "{file}.admixture_K{K}.log"
  #shadow: "shallow"
  params:
    baseout=lambda w: os.path.basename('{file}.{K}'.format(**w)),
    destdir=lambda w: os.path.dirname('{file}.{K}'.format(**w)),
```

```
  shell: '''admixture -j{threads} --seed=0 --cv {input} {wildcards
    .K} 2>&1 | tee {log} && mkdir -p {params.destdir} && mv "{
    params.baseout}.P" {output[0]} && mv "{params.baseout}.Q" {
    output[1]} '''


def save_pickle(path, *objects):
  with gzip.open(path, 'wb') as f:
    for o in objects:
      pickle.dump(o, f)

def load_pickle(path, n):
  with gzip.open(path, 'rb') as f:
    if n == 1:
      return pickle.load(f)
    else:
      return [pickle.load(f) for i in range(n)]

rule load_dataset:
        input: "{file}.traw.gz", "/scratch/shared/AIM/AIM_BITS/
    kgp3.pheno"
        output: "{file}.pickle.gz"
        run:
                dataset, labels = load_dataset(input[0],input[1])
                save_pickle(output[0], dataset.astype(numpy.uint8)
    , labels)

rule kfold_test:
        input: '{file}.pickle.gz'
        output: expand('{{file}}.cv_seed{{X,\d+}}_fold{f}.
    train_val_test.pickle.gz', f=range(kfold))
        run:
                geno, labels = load_pickle(input[0], 2)
                assert (geno.index == labels.index).all()
                skf = StratifiedKFold(n_splits=kfold, shuffle=True
    , random_state=int(wildcards.X))
                for path, (train_validation, test) in zip(output,
    skf.split(geno, labels)):
                        skf1 = StratifiedKFold(n_splits=kfold - 1)
                        geno_tv, labels_tv = geno.iloc[
    train_validation], labels.iloc[train_validation]
                        for train, validation in skf1.split(
    geno_tv, labels_tv):
                                print('Saving')
                                save_pickle(path, geno_tv.iloc[
    train], labels_tv.iloc[train], geno_tv.iloc[validation],
    labels_tv.iloc[validation], geno.iloc[test], labels.iloc[test
    ])
                                break
```

```
rule cleaning_step_maf:
        input: '{file}.pickle.gz'
        output: '{file}.maxmaf_{maf,\d+}.pickle.gz'
        run:
                maf = int(wildcards.maf)/100
                assert maf >= 0 and maf < 0.5
                geno_train, labels_train = load_pickle(input[0],
    2)
                assert compute_allelic_freqs(geno_train, alpha=0).
    max() < 1 - maf
                db_freq = pandas.DataFrame({pop:
    compute_allelic_freqs(geno_train.loc[labels_train[labels_train
     == pop].index], alpha=0) for pop in labels_train.unique()})
                db_freq['max'] = db_freq.max(axis=1)
                save_pickle(output[0], geno_train.loc[:,db_freq[
    db_freq['max'] > maf].index], labels_train)
```

Feature prioritization

```
############################
# MARKER PRIORITIZATION #
############################

rule random_forest:
        input: '{file}.pickle.gz'
        output: '{file}.fs_rf.pickle.gz'
        threads: 32
        run:
                geno_train, labels_train = load_pickle(input[0],
    2)
                assert all(geno_train.index == labels_train.index)
                clf = RandomForestClassifier(max_depth=10,
    n_estimators=10000, random_state=0, n_jobs=threads)
                print('train start')
                clf.fit(geno_train, labels_train)
                print('train done')
                fi = pandas.Series(clf.feature_importances_, index
    =geno_train.columns) columns=['importance']).sort_values('
    importance',ascending=False)
                index_fi = list(fi.sort_values(ascending=False).
    index)
                save_pickle(output[0], index_fi, fi, clf)

rule pairwise_rosenberg:
        input: '{file}.pickle.gz'
        output: '{file}.fs_pIn.pickle.gz'
        run:
                geno_train, labels_train = load_pickle(input[0],
    2)
                db = pandas.DataFrame({p: rosenberg(geno_train.loc
    [labels_train.isin(p)], labels_train[labels_train.isin(p)])
    for p in itertools.combinations(labels_train.unique(),2)},
    index=geno_train.columns)
                seen_markers = set()
                sorted_markers = []
                for marker_row in zip(*(db[pops].sort_values(
    ascending=False).index for pops in db)):
                        for m in marker_row:
                                if m not in seen_markers:
                                        seen_markers.add(m)
                                        sorted_markers.append(m)
                save_pickle(output[0], sorted_markers, db)
```

Marker set refinement

```
#########################
# MARKER SET REFINEMENT #
#########################

from sklearn.model_selection import RepeatedStratifiedKFold

def parse_params(s):
        pd = {}
        for par in s.split(','):
                key, sep, val_str = par.partition('=')
                assert sep
                try:
                        val = val_str
                        val = float(val_str)
                        val = int(val_str)
                except ValueError:
                        pass
                pd[key] = val
        return pd

search_input_stem = 'ms_search={search}({search_params})_cl={cl}({
    cl_params})'
search_output_stem = 'ms_search={search,' + '|'.join(
    search_methods) + '}({search_params})_cl={cl,' + '|'.join(
    classifiers) + '}({cl_params})'


rule marker_set_search:
        input: '{file}.fs_{m}.pickle.gz', '{file}.pickle.gz'
        output: '{file}.fs_{m,[^.]+}.' + search_output_stem + '.
    pickle.gz'
        run:
                search_params = parse_params(wildcards.
    search_params)
                cl_params = parse_params(wildcards.cl_params)
                cl = classifiers[wildcards.cl](**cl_params)
                sorted_markers = load_pickle(input[0], 1)
                geno, labels = load_pickle(input[1], 2)
                assert all(geno.index == labels.index)
                X, y = geno[sorted_markers].values, labels.values
                with open(output[0] + '.log', 'wt') as f:
                        marker_set = search_methods[wildcards.
    search](X, y, cl, f, **search_params)
                selected_markers = [sorted_markers[i] for i in
    marker_set]
                save_pickle(output[0], selected_markers)
```

Final testing

```
##################
# FINAL TESTING #
##################


rule dev:
        input:
                'kgp/kgp.pruned_500_50_0.2.BEB_CHB_PEL_GBR_YRI.
    cv_seeds10_folded5.train_val_test.maxmaf_10.fs_rf.ms_search=
    first_good(min_score=0.99,max_markers=15)_cl=nb(hwe=True).
    evaluation.pickle.gz',
                'kgp/kgp.pruned_500_50_0.2.BEB_CHB_PEL_GBR_YRI.
    cv_seed0_folded5.train_val_test.maxmaf_10.fs_rf.ms_search=
    first_good(min_score=0.99,max_markers=15)_cl=nb(hwe=True).
    evaluation.pickle.gz',
                'kgp/kgp.pruned_500_50_0.2.BEB_CHB_PEL_GBR_YRI.
    cv_seed0_fold0.train_val_test.maxmaf_10.fs_rf.ms_search=
    first_good(min_score=0.99,max_markers=15)_cl=nb(hwe=True).
    evaluation.pickle.gz',

rule marker_set_evaluation:
        input: '{file}.train_val_test.pickle.gz', '{file}.
    train_val_test.{prio}.' + search_input_stem + '.pickle.gz'
        output: '{file}.train_val_test.{prio}.' +
    search_output_stem + '.evaluation.pickle.gz'
        run:
                geno_train, labels_train, geno_validation,
    labels_validation, geno_test, labels_test = load_pickle(input
    [0], 6)
                markers = load_pickle(input[1], 1)
                train = pandas.DataFrame({'TARGET': labels_train})
                validation = pandas.DataFrame({'TARGET':
    labels_validation})
                test = pandas.DataFrame({'TARGET': labels_test})
                for i in range(len(markers)):
                        mi = markers[:i + 1]
                        cl = classifiers[wildcards.cl](**
    parse_params(wildcards.cl_params))
                        cl.fit(geno_train[mi], labels_train)
                        train[i + 1] = cl.predict(geno_train[mi])
                        validation[i + 1] = cl.predict(
    geno_validation[mi])
                        test[i + 1] = cl.predict(geno_test[mi])
                save_pickle(output[0], markers, train, validation,
     test)
```

```
rule evaluation_folding:
        input: lambda w: expand('{dataset}.cv_seed{seed}_fold{fold
    }.{params}.evaluation.pickle.gz', fold=range(int(w.folds)), **
    w)
        output: '{dataset}.cv_seed{seed,\d+}_folded{folds,\d+}.{
    params}.evaluation.pickle.gz'
        run:
                validations = []
                tests = []
                for path in input:
                        markers, train_pred, validation_pred,
    test_pred = load_pickle(path, 4)
                        validations.append(validation_pred)
                        tests.append(test_pred)
                save_pickle(output[0], *map(pandas.concat, [
    validations, tests]))

rule evaluation_aggregation:
        input:
                lambda w: expand('{dataset}.cv_seed{seed}_folded{
    folds}.{params}.evaluation.pickle.gz', seed=range(int(w.seeds)
    ), **w)
        output: '{dataset}.cv_seeds{seeds,\d+}_folded{folds,\d+}.{
    params}.evaluation.pickle.gz'
        run:
                from sklearn.metrics import
    balanced_accuracy_score
                outputs = []
                for path in input:
                        balanced_accuracy = []
                        for predictions in load_pickle(path, 2):
                                p = predictions.fillna(method='
    ffill', axis=1)
                                balanced_accuracy.append({n:
    balanced_accuracy_score(p['TARGET'], p[n]) for n in p.columns
    [1:]})
                        score_df = pandas.DataFrame(
    balanced_accuracy)
                        score_df.index.name = 'SEED'
                        score_df.columns.name = 'MARKERS'
                        outputs.append(score_df)
                save_pickle(output[0], *outputs)
```

The *continents* dataset

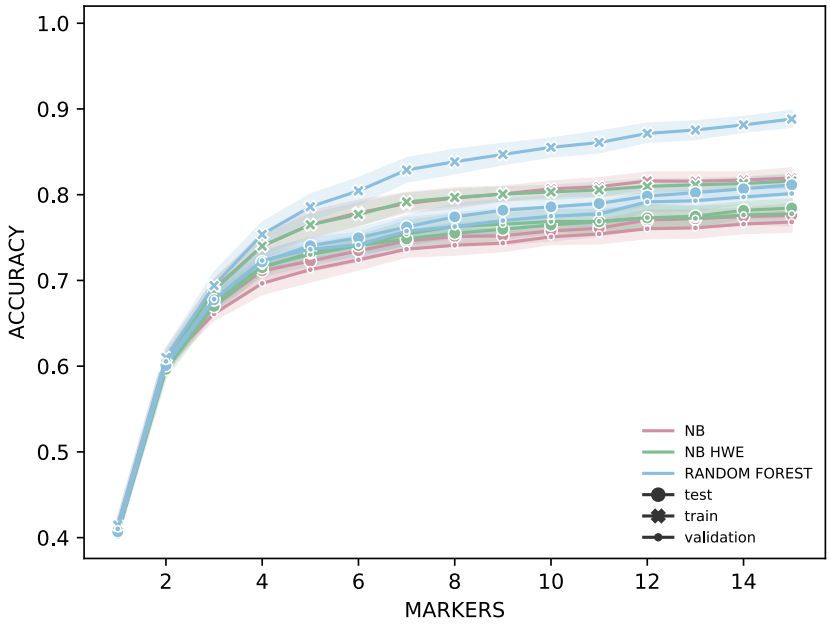


Figure C.6. Accuracy comparison of random forest feature importance with 0.20 MAF filtering as prioritization strategy on the *continents* dataset. Three different classifiers (Naïve Bayes, HWE Naïve Bayes and random forest) have been used.

The *migrants* dataset

population_label	country	study	samples	macroarea
Ethiopia	Ethiopia	Behar *et al.*, 2010	19	EAST_AFRICA
ethiopiana	Ethiopia	Busby *et al.*, 2015	7	EAST_AFRICA
ethiopiano	Ethiopia	Busby *et al.*, 2015	7	EAST_AFRICA
ethiopiant	Ethiopia	Busby *et al.*, 2015	5	EAST_AFRICA
AFAR	Ethiopia	Pagani *et al.*, 2012	12	EAST_AFRICA
AMHARA	Ethiopia	Pagani *et al.*, 2012	26	EAST_AFRICA
ESOMALI	Ethiopia	Pagani *et al.*, 2012	17	EAST_AFRICA
OROMO	Ethiopia	Pagani *et al.*, 2012	21	EAST_AFRICA
SOMALI	Ethiopia	Pagani *et al.*, 2012	23	EAST_AFRICA
TYGRAY	Ethiopia	Pagani *et al.*, 2012	21	EAST_AFRICA
WOLAYTA	Ethiopia	Pagani *et al.*, 2012	8	EAST_AFRICA
Iran	Iran	Behar *et al.*, 2010	20	MIDDLE_EAST
Israel	Israel	Behar *et al.*, 2010	3	MIDDLE_EAST
Jordania	Jordan	Behar *et al.*, 2010	20	MIDDLE_EAST
Lebanon	Lebanon	Behar *et al.*, 2010	7	MIDDLE_EAST
Saudi	Saudi_Arabia	Behar *et al.*, 2010	20	MIDDLE_EAST
Syria	Syria	Behar *et al.*, 2010	16	MIDDLE_EAST
Yemen	Yemen	Behar *et al.*, 2010	10	MIDDLE_EAST
Lebanon	Lebanon	Behar *et al.*, 2013	1	MIDDLE_EAST
bedouin	Israel (Negev)	Busby *et al.*, 2015	45	MIDDLE_EAST
druze	Israel (Carmel)	Busby *et al.*, 2015	42	MIDDLE_EAST
iranian	Iran	Busby *et al.*, 2015	20	MIDDLE_EAST
jordanian	Jordan	Busby *et al.*, 2015	20	MIDDLE_EAST
kurd	Kurd	Busby *et al.*, 2015	6	MIDDLE_EAST
lebanese	Kurd	Busby *et al.*, 2015	5	MIDDLE_EAST
palestinian	Israel (Central)	Busby *et al.*, 2015	46	MIDDLE_EAST
saudi	Saudi Arabia	Busby *et al.*, 2015	19	MIDDLE_EAST
syrian	Syria	Busby *et al.*, 2015	16	MIDDLE_EAST
uae	United Arab Emirates	Busby *et al.*, 2015	27	MIDDLE_EAST
yemeni	Yemen	Busby *et al.*, 2015	9	MIDDLE_EAST
Bedouin	Israel (Negev)	Cann *et al.*, 2002	48	MIDDLE_EAST
Druze	Israel (Carmel)	Cann *et al.*, 2002	47	MIDDLE_EAST
Palestinian	Israel (Central)	Cann *et al.*, 2002	51	MIDDLE_EAST
Egypt	Egypt	Behar *et al.*, 2010	12	NORTH_AFRICA
Morocco	Morocco	Behar *et al.*, 2010	10	NORTH_AFRICA
egyptian	Egypt	Busby *et al.*, 2015	12	NORTH_AFRICA
moroccan	Morocco	Busby *et al.*, 2015	40	NORTH_AFRICA
mozabite	Algeria (Mzab)	Busby *et al.*, 2015	29	NORTH_AFRICA
tunisian	Tunisia	Busby *et al.*, 2015	21	NORTH_AFRICA
Mozabite	Algeria (Mzab)	Cann *et al.*, 2002	30	NORTH_AFRICA
Balochi	Pakistan	Cann *et al.*, 2002	25	SOUTH_ASIA
Brahui	Pakistan	Cann *et al.*, 2002	25	SOUTH_ASIA
Burusho	Pakistan	Cann *et al.*, 2002	25	SOUTH_ASIA
Makrani	Pakistan	Cann *et al.*, 2002	25	SOUTH_ASIA
Pathan	Pakistan	Cann *et al.*, 2002	23	SOUTH_ASIA
Sindhi	Pakistan	Cann *et al.*, 2002	25	SOUTH_ASIA
BEB	Bangladesh	Auton *et al.*, 2015	144	SOUTH_ASIA
GIH	Indian Telugu from the UK	Auton *et al.*, 2015	113	SOUTH_ASIA
ITU	Indian Telugu from the UK	Auton *et al.*, 2015	118	SOUTH_ASIA
PJL	Pakistan	Auton *et al.*, 2015	158	SOUTH_ASIA
STU	Sri_Lanka	Auton *et al.*, 2015	128	SOUTH_ASIA

Table C.4. Populations included in the study of AIMs on the migrants' macro-areas of origin.

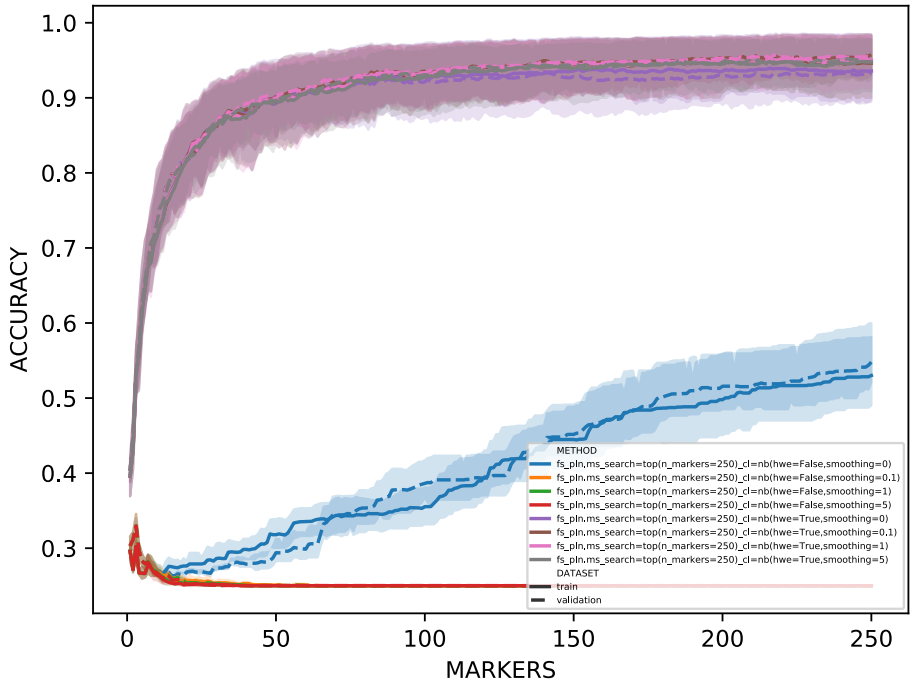


Figure C.7. Comparison of classification performances of Naïve Bayes classifier with different parameters on the first 250 markers selected by I_n. The accuracy values have been evaluated by varying the assumption of Hardy Weinberg Equilibrium ("hwe=False/True") and the parameter of Laplace smoothing ("smoothing").

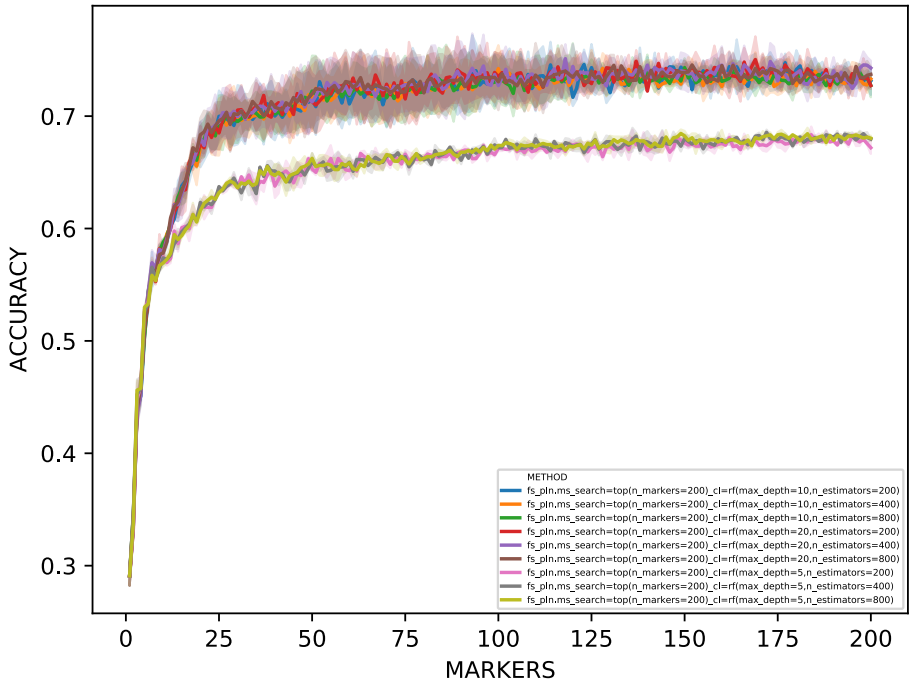


Figure C.8. Comparison of classification performances of the Random Forest classifier with different parameters on the first 200 markers selected by I_n. The accuracy values have been evaluated by varying the maximum depth of each tree ("max_depth") and the number of trees ("n_estimators").

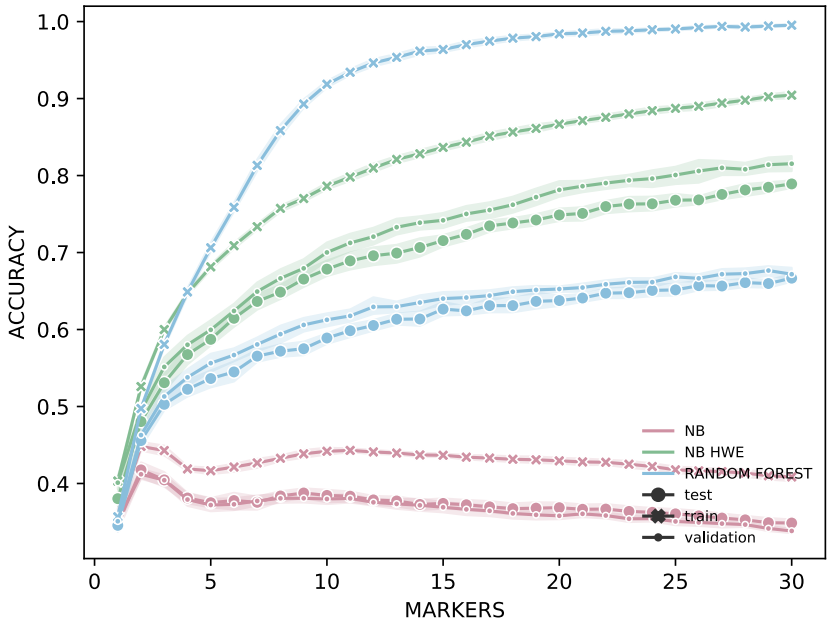


Figure C.9. Accuracy comparison of random forest feature importance with 0.20 MAF filtering as prioritization strategy on the *migrants* dataset. Three different classifiers (Naïve Bayes, HWE Naïve Bayes and random forest) have been used.

Bibliography

1. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7,** Unit7.20 (Jan. 2013).
2. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (Sept. 2009).
3. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172 (June 2015).
4. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43,** D789–798 (Jan. 2015).
5. Ammerman, A. J. & Cavalli-Sforza, L. L. *The Neolithic Transition and the Genetics of Populations in Europe.* (Princeton University Press, 1984).
6. ATAVBIS, G. No evidence of association between prothrombotic gene polymorphisms and the development of acute myocardial infarction at a young age. *Circulation* **107,** 1117–1122 (Mar. 2003).
7. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (Oct. 2015).

8. Bae, C. J. *et al.* Modern human teeth from Late Pleistocene Luna Cave (Guangxi, China). *Quaternary International* **354.** Multidisciplinary Perspectives on the Gigantopithecus Fauna and Quaternary Biostratigraphy in East Asia, 169 –183. ISSN: 1040-6182 (2014).
9. Balloux, F. The worm in the fruit of the mitochondrial DNA tree. *Heredity (Edinb)* **104,** 419–420 (May 2010).
10. Bar-Yosef, O. & Bordes, J. G. Who were the makers of the Châtelperronian culture? *J. Hum. Evol.* **59,** 586–593 (Nov. 2010).
11. Bar-Yosef, O. & Vandermeersch, B. Le Squelette Moustérien de Kébara 2. *Cahiers de Paléoanthropologie* (1991).
12. Barbujani, G., Sokal, R. R. & Oden, N. L. Indo-European origins: a computer-simulation test of five hypotheses. *Am. J. Phys. Anthropol.* **96,** 109–132 (Feb. 1995).
13. Batini, C. *et al.* Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol. Biol. Evol.* **28,** 2603–2613 (Sept. 2011).
14. Behar, D. M. *et al.* No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. *Hum. Biol.* **85,** 859–900 (Dec. 2013).
15. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (July 2010).
16. Beier, J., Anthes, N., Wahl, J. & Harvati, K. Similar cranial trauma prevalence among Neanderthals and Upper Palaeolithic modern humans. *Nature* **563,** 686–690 (Nov. 2018).
17. Benazzi, S. *et al.* Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature* **479,** 525–528 (Nov. 2011).
18. Benazzi, S. *et al.* Middle Paleolithic and Uluzzian human remains from Fumane Cave, Italy. *J. Hum. Evol.* **70,** 61–68 (May 2014).
19. Berger, L. R. *et al.* Homo naledi, a new species of the genus Homo from the Dinaledi Chamber, South Africa. *Elife* **4** (Sept. 2015).
20. Berger, L. R., Hawks, J., Dirks, P. H., Elliott, M. & Roberts, E. M. Homo naledi and Pleistocene hominin evolution in subequatorial Africa. *Elife* **6** (May 2017).
21. Bergström, A. *et al. bioRxiv.* doi:10.1101/674986 (2019).
22. Binladen, J. *et al.* Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics* **172,** 733–741 (Feb. 2006).

23. Biswas, S., Scheinfeldt, L. B. & Akey, J. M. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am. J. Hum. Genet.* **84,** 641–650 (May 2009).

24. Boattini, A. *et al.* mtDNA variation in East Africa unravels the history of Afro-Asiatic groups. *Am. J. Phys. Anthropol.* **150,** 375–385 (Mar. 2013).

25. Bocquet-Appel, J.-P. & Degioanni, A. Neanderthal Demographic Estimates. *Current Anthropology* **54,** S202–S213 (2013).

26. Bocquet-Appel, J.-P., Demars, P.-Y., Noiret, L. & Dobrowsky, D. Estimates of Upper Palaeolithic meta-population size in Europe from archaeological data. *Journal of Archaeological Science* **32,** 1656 –1668. ISSN: 0305-4403 (2005).

27. Bonatti, F. *et al.* RNA-based analysis of BRCA1 and BRCA2 gene alterations. *Cancer Genet. Cytogenet.* **170,** 93–101 (Oct. 2006).

28. Bornstein, C. *et al.* SPATA18, a spermatogenesis-associated gene, is a novel transcriptional target of p53 and p63. *Mol. Cell. Biol.* **31,** 1679–1689 (Apr. 2011).

29. Bouakaze, C., Keyser, C., Crubezy, E., Montagnon, D. & Ludes, B. Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis. *Int. J. Legal Med.* **123,** 315–325 (July 2009).

30. Branicki, W., Brudnik, U., Kupiec, T., Wolanska-Nowak, P. & Wojas-Pelc, A. Determination of phenotype associated SNPs in the MC1R gene. *J. Forensic Sci.* **52,** 349–354 (Mar. 2007).

31. Brauer, A., Haug, G. H., Dulski, P., Sigman, D. M. & Negendank, J. F. An abrupt wind shift in western Europe at the onset of the Younger Dryas cold period. en. *Nature Geoscience* **1.** Received 14 September 2007. accepted 30 May 2008. published 1 August 2008., 520 –523 (2008).

32. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. ISSN: 1573-0565 (2001).

33. Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* **84,** 343–364 (Aug. 2012).

34. Brisighelli, F. *et al.* Patterns of Y-STR variation in Italy. *Forensic Sci Int Genet* **6,** 834–839 (Dec. 2012).

35. Brisighelli, F. *et al.* Uniparental Markers of Contemporary Italian Population Reveals Details on Its Pre-Roman Heritage. *PLOS ONE* **7,** 1–15 (Dec. 2012).

36. Broushaki, F. *et al.* Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353,** 499–503 (July 2016).

37. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173,** 53–61 (Mar. 2018).

38. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47,** D1005–D1012 (Jan. 2019).

39. Busby, G. B. *et al.* The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Curr. Biol.* **25,** 2518–2526 (Oct. 2015).

40. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–262 (Apr. 2002).

41. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325,** 31–36 (1987).

42. Capelli, C. *et al.* Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. *Mol. Phylogenet. Evol.* **44,** 228–239 (July 2007).

43. Casanova, M. *et al.* A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* **230,** 1403–1406 (Dec. 1985).

44. Cassidy, L. M. *et al.* Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 368–373 (Jan. 2016).

45. Castellano, S. *et al.* Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 6666–6671 (May 2014).

46. Cattaneo, C. *et al.* The forgotten tragedy of unidentified dead in the Mediterranean. *Forensic Sci. Int.* **250,** 1–2 (2015).

47. Cattaneo, C. *Naufraghi senza volto* Italian. ISBN: 9788832850574 (Raffaello Cortina Editore, 2018).

48. Cavalli-Sforza, L. L. Genes, peoples, and languages. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 7719–7724 (July 1997).

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

50. *Census-2007 Report* http://www.csa.gov.et/census-report/complete-report/census-2007.

51. Cerezo, M. *et al.* Reconstructing ancient mitochondrial DNA links between Africa and Europe. *Genome Res.* **22,** 821–826 (May 2012).

52. Channell, J. & Vigliotti, L. The role of geomagnetic field intensity in late Quaternary evolution of humans and large mammals. *Reviews of Geophysics* (2019).

53. Chaubey, S. *et al.* Six-year survival of a patient with pulmonary artery angiosarcoma. *Asian Cardiovasc Thorac Ann* **20,** 728–730 (Dec. 2012).

54. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128 (Apr. 2013).

55. Cheung, E. Y. Y., Phillips, C., Eduardoff, M., Lareu, M. V. & McNevin, D. Performance of ancestry-informative SNP and microhaplotype markers. *Forensic Sci Int Genet* **43,** 102141 (Aug. 2019).

56. Churchill, S. E. *Thin on the Ground: Neandertal Biology, Archeology, and Ecology* (Wiley Online Library, 2014).

57. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).

58. Clarkson, C. *et al.* The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *J. Hum. Evol.* **83,** 46–64 (June 2015).

59. Condemi, S. & Savatier, F. *N andertal mon fr re. 300 000 ans d'histoire de l' homme* (Flammarion, 2016).

60. Conrad, D. F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* **38,** 1251–1260 (Nov. 2006).

61. Conroy, G. C. *et al.* Endocranial capacity of the bodo cranium determined from three-dimensional computed tomography. *Am. J. Phys. Anthropol.* **113,** 111–118 (Sept. 2000).

62. Consortium, I. H. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (Oct. 2005).

63. Consortium, I. H. The International HapMap Project. *Nature* **426,** 789–796 (Dec. 2003).

64. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5,** e1000500 (June 2009).

65. Coppa, A., Manni, F., Stringer, C., Vargiu, R. & Vecchi, F. Evidence for new Neanderthal teeth in Tabun Cave (Israel) by the application of self-organizing maps (SOMs). *J. Hum. Evol.* **52,** 601–613 (June 2007).

66. Cunliffe, B. *By Steppe, Desert, and Ocean: The Birth of Eurasia* (Oxford University Press, 2015).

67. Dabney, J., Meyer, M. & Paabo, S. Ancient DNA damage. *Cold Spring Harb Perspect Biol* **5** (July 2013).

68. Dannemann, M. & Kelso, J. The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. *Am. J. Hum. Genet.* **101,** 578–589 (Oct. 2017).

69. Dannemann, M. & Racimo, F. Something old, something borrowed: admixture and adaptation in human evolution. *Curr. Opin. Genet. Dev.* **53,** 1–8 (Dec. 2018).

70. Dannemann, M., Prufer, K. & Kelso, J. Functional implications of Neandertal introgression in modern humans. *Genome Biol.* **18,** 61 (Apr. 2017).

71. Dannemann, M., Andres, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am. J. Hum. Genet.* **98,** 22–33 (Jan. 2016).

72. Darwin, C. *The descent of man and selection in relation to sex* (D. Appleton, 1896).

73. D'Atanasio, E. *et al.* Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa. *Forensic Sci Int Genet* **38,** 185–194 (Jan. 2019).

74. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (Dec. 2010).

75. Day, F. R., Ong, K. K. & Perry, J. R. B. Elucidating the genetic basis of social interaction and isolation. *Nat Commun* **9,** 2457 (July 2018).

76. De la Puente, M. *et al.* The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic Sci Int Genet* **22,** 81–88 (May 2016).

77. Decker, J. E. *et al.* Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.* **10,** e1004254 (Mar. 2014).

78. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (Jan. 2013).

79. Delson, E. An early dispersal of modern humans from Africa to Greece. *Nature* **571,** 487–488 (July 2019).

80. Denaro, M. *et al.* Ethnic variation in Hpa 1 endonuclease cleavage patterns of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.* **78,** 5768–5772 (Sept. 1981).

81. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31,** 1102–1110 (Dec. 2013).

82. Destro Bisol, G. *et al.* Italian isolates today: geographic and linguistic factors shaping human biodiversity. *J Anthropol Sci* **86,** 179–188 (2008).

83. Di Cristofaro, J. *et al.* Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. *PLoS ONE* **8,** e76748 (2013).

84. Di Gaetano, C. *et al.* An overview of the genetic structure within the Italian population from genome-wide data. *PLoS ONE* **7,** e43759 (2012).

85. Dirks, P. H. *et al.* Geological and taphonomic context for the new hominin species Homo naledi from the Dinaledi Chamber, South Africa. *Elife* **4** (Sept. 2015).

86. Dirks, P. H. *et al.* The age of Homo naledi and associated sediments in the Rising Star Cave, South Africa. *Elife* **6** (May 2017).

87. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518,** 102–106 (Feb. 2015).

88. Dopazo, J. *et al.* 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Mol. Biol. Evol.* **33,** 1205–1218 (May 2016).

89. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28,** 2239–2252 (Aug. 2011).

90. Eiberg, H. *et al.* Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **123,** 177–187 (Mar. 2008).

91. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18,** 599–612 (Oct. 2017).

92. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6,** R44 (2005).

93. *European Asylum Support Office* https://www.easo.europa.eu/sites/default/files/public/Eritrea-Report-Final.pdf.

94. Fernandes, D. M. *et al.* The Arrival of Steppe and Iranian Related Ancestry in the Islands of the Western Mediterranean. *Biorxiv* (Mar. 2019).

95. Ferrara, S. in *A Companion to the Ancient Greek Language* 9–24 (John Wiley and Sons, 2010).

96. Finlayson, C. *et al.* Birds of a feather: Neanderthal exploitation of raptors and corvids. *PLoS ONE* **7,** e45927 (2012).

97. Fiorito, G. *et al.* The Italian genome reflects the history of Europe and the Mediterranean basin. *Eur. J. Hum. Genet.* **24,** 1056–1062 (July 2016).

98. Fitzpatrick, A. P. *The Amesbury Archer and the Boscombe Bowmen: Bell Beaker Burials on Boscombe down, Amesbury, Wiltshire.* (Wessex Archaeology, 2011).

99. Fitzsimmons, K. E., Stern, N. & Murray-Wallace, C. V. Depositional history and archaeology of the central Lake Mungo lunette, Willandra Lakes, southeast Australia. *Journal of Archaeological Science* **41,** 349–364. ISSN: 0305-4403 (2014).

100. Forster, P. & Matsumura, S. Did Early Humans Go North or South? *Science* **308,** 965–966. ISSN: 0036-8075 (2005).

101. Freire-Aradas, A. *et al.* Exploring iris colour prediction and ancestry inference in admixed populations of South America. *Forensic Sci Int Genet* **13,** 3–9 (Nov. 2014).

102. Friedman, M. J. The Evolution of Hominid Bipedalism (2006).

103. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524,** 216–219 (Aug. 2015).

104. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 2223–2227 (Feb. 2013).

105. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (Oct. 2014).

106. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534,** 200–205 (June 2016).

107. Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349,** 1343–1347 (Sept. 2015).

108. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31,** 3718–3720 (Nov. 2015).

109. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun* **5,** 5257 (Oct. 2014).

110. Gattepaille, L. M. & Jakobsson, M. Combining markers into haplotypes can improve population structure inference. *Genetics* **190,** 159–174 (Jan. 2012).

111. Giacomo, F. D. *et al.* Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. *Molecular Phylogenetics and Evolution* **28,** 387 –395. ISSN: 1055-7903 (2003).

112. Gibbons, A. HUMAN EVOLUTION. Neandertal genes linked to modern diseases. *Science* **351,** 648–649 (Feb. 2016).

113. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (Jan. 2012).

114. Gilbert, M. T. *et al.* Characterization of genetic miscoding lesions caused by postmortem damage. *Am. J. Hum. Genet.* **72,** 48–61 (Jan. 2003).

115. Gittelman, R. M. *et al.* Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. *Curr. Biol.* **26,** 3375–3382 (Dec. 2016).

116. Glusman, G., Yanai, I., Rubin, I. & Lancet, D. The complete human olfactory subgenome. *Genome Res.* **11,** 685–702 (May 2001).

117. Goodwin, W. H. The use of forensic DNA analysis in humanitarian forensic action: The development of a set of international standards. *Forensic Sci. Int.* **278,** 221–227 (Sept. 2017).

118. Gravina, B. *et al.* No Reliable Evidence for a Neanderthal-Châtelperronian Association at La Roche-à-Pierrot, Saint-Césaire. *Sci Rep* **8,** 15134 (Oct. 2018).

119. Green, R. E. *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134,** 416–426 (Aug. 2008).

120. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (May 2010).

121. Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444,** 330–336 (Nov. 2006).

122. Grün, R. Direct dating of human fossils. *Am. J. Phys. Anthropol.* **Suppl 43,** 2–48 (2006).

123. Grün, R. and Brink, J. S. and Spooner, N. A. and Taylor, L. and Stringer, C. B. and Franciscus, R. G. and Murray, A. S. Direct dating of Florisbad hominid. *Nature* **382,** 500–501 (Aug. 1996).

124. Grün, R. and Stringer, C. Tabun revisited: revised ESR chronology and new ESR and U-series analyses of dental material from Tabun C1. *J. Hum. Evol.* **39,** 601–612 (Dec. 2000).

125. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517,** 327–332 (Jan. 2015).

126. Günther, T. & Jakobsson, M. Genes mirror migrations and cultures in prehistoric Europe-a population genomic perspective. *Curr. Opin. Genet. Dev.* **41,** 115–123 (Dec. 2016).

127. Günther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 11917–11922 (Sept. 2015).

128. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (June 2015).

129. Haber, M. *et al.* Chad Genetic Diversity Reveals an African History Marked by Multiple Holocene Eurasian Migrations. *Am. J. Hum. Genet.* **99,** 1316–1324 (Dec. 2016).

130. Haber, M. *et al.* Genome-wide diversity in the levant reveals recent structuring by culture. *PLoS Genet.* **9,** e1003316 (2013).

131. Hajdinjak, M. *et al.* Reconstructing the genetic history of late Neanderthals. *Nature* **555,** 652–656 (Mar. 2018).

132. Hajiloo, M. *et al.* ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. *BMC Bioinformatics* **14,** 61 (2013).

133. Hammer, M. F. & Zegura, S. L. The Human Y Chromosome Haplogroup Tree: Nomenclature and Phylogeography of Its Major Divisions. *Annual Review of Anthropology* **31,** 303–321 (2002).

134. Hansen, A., Willerslev, E., Wiuf, C., Mourier, T. & Arctander, P. Statistical evidence for miscoding lesions in ancient DNA templates. *Mol. Biol. Evol.* **18,** 262–265 (Feb. 2001).

135. Harari, Y. *Sapiens: A Brief History of Humankind* ISBN: 9780062316103 (Harper, 2015).

136. Harris, K. & Nielsen, R. The Genetic Cost of Neanderthal Introgression. *Genetics* **203,** 881–891 (June 2016).

137. Harris, E. E. *Ancestors in Our Genome: The New Science of Human Evolution* (Oxford University Press, 2015).

138. Harvati, K. *et al.* Apidima Cave fossils provide earliest evidence of Homo sapiens in Eurasia. *Nature* **571,** 500–504 (July 2019).

139. Hawks, J. *et al.* New fossil remains of Homo naledi from the Lesedi Chamber, South Africa. *Elife* **6** (May 2017).

140. Hayden, S. *et al.* Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res.* **20,** 1–9 (Jan. 2010).

141. Hefner, J. T. & Ousley, S. D. Statistical classification methods for estimating ancestry using morphoscopic traits. *J. Forensic Sci.* **59,** 883–890 (July 2014).

142. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343,** 747–751 (Feb. 2014).

143. Hellenthal, G., Auton, A. & Falush, D. Inferring human colonization history using a copying model. *PLoS Genet.* **4,** e1000078 (May 2008).

144. Henn, B. M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S. & Bustamante, C. D. Fine-scale population structure and the era of next-generation sequencing. *Hum. Mol. Genet.* **19,** R221–226 (Oct. 2010).

145. Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* **8,** e1002397 (Jan. 2012).

146. Heyes, P. & MacDonald, K. Neandertal energetics: Uncertainty in body mass estimation limits comparisons with Homo sapiens. *J. Hum. Evol.* **85,** 193–197 (Aug. 2015).

147. Higham, T. *et al.* Chronology of the Grotte du Renne (France) and implications for the context of ornaments and human remains within the Châtelperronian. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 20234–20239 (Nov. 2010).

148. Higham, T. European Middle and Upper Palaeolithic radiocarbon dates are often older than they look: problems with previous dates and some remedies. *Antiquity* **85,** 235–249 (2011).

149. Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476,** 170–175 (July 2011).

150. Hodoğlugil, U. & Mahley, R. W. Turkish population structure and genetic ancestry reveal relatedness among Eurasian populations. *Ann. Hum. Genet.* **76,** 128–141 (Mar. 2012).

151. Hoffmann, D. L. *et al.* U-Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art. *Science* **359,** 912–915 (Feb. 2018).

152. Hofmanova, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 6886–6891 (June 2016).

153. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Paabo, S. Ancient DNA. *Nat. Rev. Genet.* **2,** 353–359 (May 2001).

154. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. & Paabo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29,** 4793–4799 (Dec. 2001).

155. Höss, M., Jaruga, P., Zastawny, T. H., Dizdaroglu, M. & Paabo, S. DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Res.* **24,** 1304–1307 (Apr. 1996).

156. Hout, C. V. V. *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *Biorxiv* (2019).

157. Hu, W. The role of p53 gene family in reproduction. *Cold Spring Harb Perspect Biol* **1,** a001073 (Dec. 2009).

158. Hublin, J. J. *et al.* New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* **546,** 289–292 (June 2017).

159. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **194,** 512 (July 2014).

160. Hurst, J. A., Baraitser, M., Auger, E., Graham, F. & Norell, S. An extended family with a dominantly inherited speech disorder. *Dev Med Child Neurol* **32,** 352–355 (Apr. 1990).

161. Iacovacci, G. *et al.* Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries. *Forensic Sci Int Genet* **27,** 123–131 (Mar. 2017).

162. Iida, H., Ichinose, J., Kaneko, T., Mōri, T. & Shibata, Y. Complementary DNA cloning of rat spetex-1, a spermatid-expressing gene-1, encoding a 63 kDa cytoplasmic protein of elongate spermatids. *Mol. Reprod. Dev.* **68,** 385–393 (Aug. 2004).

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

164. International Committee of the Red Cross. *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)* (1977).

165. International Committee of the Red Cross. *The Geneva Conventions of August 12, 1949* (1949).

166. Jacobs, L. C. *et al.* Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. *Hum. Genet.* **132,** 147–158 (Feb. 2013).

167. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23,** 1801–1806 (July 2007).

168. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451,** 998–1003 (Feb. 2008).

169. Jaubert, J. *et al.* Early Neanderthal constructions deep in Bruniquel Cave in southwestern France. *Nature* **534,** 111–114 (June 2016).

170. Jenkins, D. L. *et al.* Clovis age Western Stemmed projectile points and human coprolites at the Paisley Caves. *Science* **337,** 223–228 (July 2012).

171. Jin, Y. *et al.* Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nat. Genet.* **44,** 676–680 (May 2012).

172. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat Commun* **6,** 8912 (Nov. 2015).

173. Juric, I., Aeschbacher, S. & Coop, G. The Strength of Selection against Neanderthal Introgression. *PLoS Genet.* **12,** e1006340 (Nov. 2016).

174. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19,** 249–255 (Feb. 2017).

175. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50,** 390–400 (Mar. 2018).

176. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (Jan. 2000).

177. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–462 (Jan. 2016).

178. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (Jan. 2017).

179. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *Biorxiv* (2019).

180. Keinan, A. & Clark, A. G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336,** 740–743. ISSN: 0036-8075 (2012).

181. Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun* **3,** 698 (Feb. 2012).

182. Kidd, J. R. *et al.* Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics* **2,** 3600 (Jan. 2011).

183. Kim, Y. E., Ki, C. S. & Jang, M. A. Challenges and Considerations in Sequence Variant Interpretation for Mendelian Disorders. *Ann Lab Med* **39,** 421–429 (Sept. 2019).

184. Kohler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45,** D865–D876 (Jan. 2017).

185. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (Aug. 2012).

186. Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour* **15,** 1179–1191 (Sept. 2015).

187. Koster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34,** 3600 (Oct. 2018).

188. Kovacevic, L. *et al.* Standing at the gateway to Europe–the genetic structure of Western balkan populations based on autosomal and haploid markers. *PLoS ONE* **9,** e105090 (2014).

189. Krause, J. *et al.* The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr. Biol.* **17,** 1908–1912 (Nov. 2007).

190. Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90,** 19–30 (July 1997).

191. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–97 (July 2016).

192. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4,** 1073–1081 (2009).

193. Kushniarevich, A. *et al.* Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. *PLoS ONE* **10,** e0135820 (2015).

194. Kwak, S. H. *et al.* Findings of a 1303 Korean whole-exome sequencing study. *Exp. Mol. Med.* **49,** e356 (July 2017).

195. Kılınç, G. M. *et al.* The Demographic Development of the First Farmers in Anatolia. *Curr. Biol.* **26,** 2659–2666 (Oct. 2016).

196. Lachance, J. & Tishkoff, S. A. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* **35,** 780–786 (Sept. 2013).

197. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 2001).

198. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–868 (Jan. 2016).

199. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (Oct. 2010).

200. Lao, O. *et al.* Correlation between genetic and geographic structure in Europe. *Curr. Biol.* **18,** 1241–1248 (Aug. 2008).

201. Lawson, C. L. & Hanson, R. J. *Solving least squares problems* Revised reprint of the 1974 original, xii+337. ISBN: 0-89871-356-0 (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995).

202. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun* **9,** 3258 (Aug. 2018).

203. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8,** e1002453 (Jan. 2012).

204. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (Sept. 2014).

205. Lazaridis, I. *et al.* Genetic origins of the Minoans and Mycenaeans. *Nature* **548,** 214–218 (Aug. 2017).

206. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536,** 419–424 (Aug. 2016).

207. Lazaridis, I. The evolutionary history of human populations in Europe. *Current Opinion in Genetics & Development* **53.** Genetics of Human Origins, 21 –27. ISSN: 0959-437X (2018).

208. Legarra Herrero, B. The Minoan fallacy: cultural diversity and mortuary behaviour on Crete at the beginning od the Bronze Age. *Oxford Journal of Archaeology* **28,** 29–57 (2009).

209. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (Aug. 2016).

210. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519,** 309–314 (Mar. 2015).

211. Lesseur, C. *et al.* Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat. Genet.* **48,** 1544–1550 (Dec. 2016).

212. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (Feb. 2008).

213. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165,** 2213–2233 (Dec. 2003).

214. Li, X. & Kahveci, T. A Novel algorithm for identifying low-complexity regions in a protein sequence. *Bioinformatics* **22,** 2980–2987 (Dec. 2006).

215. Lipson, M. *et al.* Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551,** 368–372 (Nov. 2017).

216. Liu, W. *et al.* Human remains from Zhirendong, South China, and modern human emergence in East Asia. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 19201–19206 (Nov. 2010).

217. Loh, P. R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193,** 1233–1254 (Apr. 2013).

218. Lopez, S., van Dorp, L. & Hellenthal, G. Human Dispersal Out of Africa: A Lasting Debate. *Evol. Bioinform. Online* **11,** 57–68 (2015).

219. López, S. *et al.* The genetic landscape of Ethiopia: diversity, intermixing and the association with culture. *bioRxiv.* doi:10.1101/756536 (2019).

220. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (Jan. 2017).

221. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39,** 1181–1186 (Oct. 2007).

222. Malinowski, J. R. *et al.* Genetic variants associated with serum thyroid stimulating hormone (TSH) levels in European Americans and African Americans from the eMERGE Network. *PLoS ONE* **9,** e111301 (2014).

223. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538,** 201–206 (Oct. 2016).

224. Mallory, J. P. & Adams, D. Q. *Encyclopedia Of Indo-European Culture* (Fitzroy Dearborn, 1997).

225. Manco, J. *Ancestral journeys* (Thames & Hudson, 2015).

226. Manolio, T. A. & Collins, F. S. The HapMap and genome-wide association studies in diagnosis and therapy. *Annu. Rev. Med.* **60,** 443–456 (2009).

227. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93,** 278–288 (Aug. 2013).

228. Maricic, T. *et al.* A recent evolutionary change affects a regulatory element in the human FOXP2 gene. *Mol. Biol. Evol.* **30,** 844–852 (Apr. 2013).

229. Marom, A., McCullagh, J. S., Higham, T. F., Sinitsyn, A. A. & Hedges, R. E. Single amino acid radiocarbon dating of Upper Paleolithic modern humans. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 6878–6881 (May 2012).

230. Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166,** 351–372 (Jan. 2004).

231. Martin, A. R. *et al.* Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. *Am. J. Hum. Genet.* **102,** 760–775 (May 2018).

232. Martiniano, R. *et al.* Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat Commun* **7,** 10326 (Jan. 2016).

233. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528,** 499–503 (Dec. 2015).

234. Mathieson, I. *et al.* The genomic history of southeastern Europe. *Nature* **555,** 197–203 (Mar. 2018).

235. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (June 2016).

236. Mellars, P. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature* **439,** 931–935 (Feb. 2006).

237. Mendez, F. L., Watkins, J. C. & Hammer, M. F. Neandertal origin of genetic variation at the cluster of OAS immunity genes. *Mol. Biol. Evol.* **30,** 798–801 (Apr. 2013).

238. Metspalu, M. *et al.* Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am. J. Hum. Genet.* **89,** 731–744 (Dec. 2011).

239. Metzker, M. L. Sequencing technologies —the next generation. *Nature Reviews Genetics* **11,** 31 EP – (Dec. 2009).

240. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (Oct. 2012).

241. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **45,** D183–D189 (Jan. 2017).

242. *migration.iom.int* https://migration.iom.int/europe?type=missing. Updated August 25, 2019.

243. Mijares, A. S. *et al.* New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines. *J. Hum. Evol.* **59,** 123–132 (July 2010).

244. Miller, G. H. *et al.* Pleistocene extinction of genyornis newtoni: human impact on australian megafauna. *Science* **283,** 205–208 (Jan. 1999).

245. Missing Migrants Project. *Fatal Journeys 4: Missing Migrant Children* English. ISBN: 978-92-9068-786-3 (IOM, 2019).

246. Missing Migrants Project. *Fatal Journeys Volume 3 Part 1: Improving Data on Missing Migrants* (IOM, 2017).

247. Mittnik, A. *et al.* The genetic prehistory of the Baltic Sea region. *Nat Commun* **9,** 442 (Jan. 2018).

248. Montinaro, F. *et al.* Unravelling the hidden ancestry of American admixed populations. *Nat Commun* **6,** 6596 (Mar. 2015).

249. Moorjani, P. *et al.* A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 5652–5657 (May 2016).

250. Morgan, M. D. *et al.* Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat Commun* **9,** 5271 (Dec. 2018).

251. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541,** 302–310 (Jan. 2017).

252. Nigst, P. R. *et al.* Early modern human settlement of Europe north of the Alps occurred 43,500 years ago in a cold steppe-type environment. *Proceedings of the National Academy of Sciences* **111,** 14394–14399. ISSN: 0027-8424 (2014).

253. Noonan, J. P. *et al.* Sequencing and analysis of Neanderthal genomic DNA. *Science* **314,** 1113–1118 (Nov. 2006).

254. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40,** 646–649 (May 2008).

255. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (Nov. 2008).

256. Nutile, T. *et al.* Whole-Exome Sequencing in the Isolated Populations of Cilento from South Italy. *Sci Rep* **9,** 4059 (Mar. 2019).

257. Nykamp, K. *et al.* Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet. Med.* **19,** 1105–1117 (Oct. 2017).

258. O'Connell, J. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *Journal of Archaeological Science* **56.** Scoping the Future of Archaeological Science: Papers in Honour of Richard Klein, 73 –84. ISSN: 0305-4403 (2015).

259. Olalde, I. *et al.* A Common Genetic Origin for Early Farmers from Mediterranean Cardial and Central European LBK Cultures. *Mol. Biol. Evol.* **32,** 3132–3142 (Dec. 2015).

260. Olalde, I. *et al.* The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555,** 190–196 (Mar. 2018).

261. Olivieri, L. *et al.* Challenges in the identification of dead migrants in the Mediterranean: The case study of the Lampedusa shipwreck of October 3rd 2013. *Forensic Sci. Int.* **285,** 121–128 (Apr. 2018).

262. Omrak, A. *et al.* Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. *Curr. Biol.* **26,** 270–275 (Jan. 2016).

263. Pääbo, S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. U.S.A.* **86,** 1939–1943 (Mar. 1989).

264. Pääbo, S. *et al.* Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38,** 645–679 (2004).

265. Pääbo, S. *Neanderthal man: in search of lost genome* (BASIC BOOKS, 2015).

266. Pääbo, S. The human condition-a molecular approach. *Cell* **157,** 216–226 (Mar. 2014).

267. Pagani, L. *et al.* Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* **91,** 83–96 (July 2012).

268. Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538,** 238–242 (Oct. 2016).

269. Pagani, L. *et al.* Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96,** 986–991 (June 2015).

270. Parolo, S. *et al.* Characterization of the biological processes shaping the genetic structure of the Italian population. *BMC Genet.* **16,** 132 (Nov. 2015).

271. Paschou, P. *et al.* Maritime route of colonization of Europe. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 9211–9216 (June 2014).

272. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (Nov. 2012).

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

274. Pereira, R. *et al.* Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS ONE* **7,** e29684 (2012).

275. Peresani, M., Fiore, I., Gala, M., Romandini, M. & Tagliacozzo, A. Late Neandertals and the intentional removal of feathers as evidenced from bird bone taphonomy at Fumane Cave 44 ky B.P., Italy. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 3888–3893 (Mar. 2011).

276. Pessina, A. & Tiné, V. *Archeologia del neolitico: l'Italia tra il VI e IV millennio a.C.* ISBN: 9788843045853 (Carocci, 2008).

277. Peter, B. M. Admixture, Population Structure, and F-Statistics. *Genetics* **202,** 1485–1501 (Apr. 2016).

278. Petr, M., Paabo, S., Kelso, J. & Vernot, B. Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U.S.A.* **116,** 1639–1644 (Jan. 2019).

279. Peyrégne, S. *et al.* Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. *Science Advances* **5.** doi:10.1126/sciadv.aaw5873 (2019).

280. Phillips, C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci Int Genet* **18,** 49–65 (Sept. 2015).

281. Piazza, A., Cappello, N., Olivetti, E. & Rendine, S. A genetic history of Italy. *Annals of Human Genetics* **52,** 203–213 (1988).

282. Pievani, T. and Calzolaio, V. *Libertá di migrare* (Einaudi, 2016).

283. Pinhasi, R. *et al.* Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. *PLoS ONE* **10,** e0129102 (2015).

284. Piscitelli, V., Iadicicco, A., De Angelis, D., Porta, D. & Cattaneo, C. Italy's battle to identify dead migrants. *Lancet Glob Health* **4,** e512–513 (Aug. 2016).

285. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2,** e105 (July 2006).

286. Polimanti, R. & Gelernter, J. ADH1B: From alcoholism, natural selection, and cancer to the human phenome. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **177,** 113–125 (Mar. 2018).

287. Pope, K. O. & Terrell, J. E. Environmental setting of human migrations in the circum-Pacific region. *Journal of Biogeography* **35,** 1–21 (2008).

288. Posth, C. *et al.* Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat Commun* **8,** 16046 (July 2017).

289. Prinz, M. *et al.* DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Sci Int Genet* **1,** 3–12 (Mar. 2007).

290. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (June 2000).

291. Prufer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358,** 655–658 (Nov. 2017).

292. Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (Jan. 2014).

293. Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15,** R159–160 (Mar. 2005).

294. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (Sept. 2007).

295. Quach, H. *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* **167,** 643–656 (Oct. 2016).

296. Quintana-Murci, L. *et al.* Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat. Genet.* **23,** 437–441 (Dec. 1999).

297. Quintans, B., Ordonez-Ugalde, A., Cacheiro, P., Carracedo, A. & Sobrido, M. J. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Appl Transl Genom* **3,** 60–67 (Sept. 2014).

298. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sanchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16,** 359–371 (June 2015).

299. Racimo, F., Marnetto, D. & Huerta-Sanchez, E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Mol. Biol. Evol.* **34,** 296–317 (Feb. 2017).

300. Radovčić, D., Sršen, A. O., Radovčić, J. & Frayer, D. W. Evidence for Neandertal jewelry: modified white-tailed eagle claws at Krapina. *PLoS ONE* **10,** e0119802 (2015).

301. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505,** 87–91 (Jan. 2014).

302. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol.* **11,** e1001555 (2013).

303. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15942–15947 (Nov. 2005).

304. Ramos, E. M. *et al.* Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **22,** 144–147 (Jan. 2014).

305. Raschka, S. *Python Machine Learning* ISBN: 1783555130, 9781783555130 (Packt Publishing, 2015).

306. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334,** 94–98 (Oct. 2011).

307. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (Feb. 2010).

308. Raveane, A. *et al.* Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Science Advances* **5.** doi:10.1126/sciadv.aaw3492. eprint: https://advances.sciencemag.org/content/5/9/eaaw3492.full.pdf. <https://advances.sciencemag.org/content/5/9/eaaw3492> (Sept. 2019).

309. Reich, D. *Who We Are and How We Got Here* (Oxford University Press, 2018).

310. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (Dec. 2010).

311. Relethford, J. H. Genetic evidence and the modern human origins debate. *Heredity (Edinb)* **100,** 555–563 (June 2008).

312. Rendu, W. *et al.* Evidence supporting an intentional Neandertal burial at La Chapelle-aux-Saints. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 81–86 (Jan. 2014).

313. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–424 (May 2015).

314. Roberts, R. G. *et al.* New ages for the last Australian megafauna: continent-wide extinction about 46,000 years ago. *Science* **292,** 1888–1892 (June 2001).

315. Roebroeks, W. & Soressi, M. Neandertals revised. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 6372–6379 (June 2016).

316. Roebroeks, W. *et al.* Use of red ochre by early Neandertals. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 1889–1894 (Feb. 2012).

317. Rohlfs, R. V., Fullerton, S. M. & Weir, B. S. Familial identification: population structure and relationship distinguishability. *PLoS Genet.* **8,** e1002469 (Feb. 2012).

318. Rosenberg, K. R. A late pleistocene human skeleton from liujiang, china suggests regional population variation in sexual dimorphism in the human pelvis. (2002).

319. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298,** 2381–2385 (Dec. 2002).

320. Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73,** 1402–1422 (Dec. 2003).

321. Rosenberg, N. A. distruct: a program for the graphical display of population structure. *Molecular Ecology Notes* **4,** 137–138 (2004).

322. Russell, N., Martin, L. & K., T. C. Building memories: commemorative deposits at atalh y k. *Anthropozoologica* **44,** 103 –125 –23 (2009).

323. Saag, L. *et al.* Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. *Curr. Biol.* **27,** 2185–2193 (July 2017).

324. Sams, A. J. *et al.* Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.* **17,** 246 (Nov. 2016).

325. Sanchez-Roige, S. *et al.* Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am J Psychiatry* **176,** 107–118 (Feb. 2019).

326. Sankararaman, S., Patterson, N., Li, H., Paabo, S. & Reich, D. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8,** e1002947 (2012).

327. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507,** 354–357 (Mar. 2014).

328. Santander, C., Montinaro, F. & Capelli, C. Searching for archaic contribution in Africa. *Ann. Hum. Biol.* 1–11 (June 2019).

329. Sarno, S. *et al.* Ancient and recent admixture layers in Sicily and Southern Italy trace multiple migration routes along the Mediterranean. *Sci Rep* **7,** 1984 (May 2017).

330. Sazzini, M. *et al.* Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula. *Sci Rep* **6,** 32513 (Sept. 2016).

331. Schiffels, S. *et al.* Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat Commun* **7,** 10408 (Jan. 2016).

332. Schoetensack, O. *Der Unterkiefer des Homo Heidelbergensis aus den Sanden von Mauer bei Heidelberg* (Leipzig, 1908).

333. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11,** 361–362 (Apr. 2014).

334. Seguin-Orlando, A. *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science* **346,** 1113–1118 (2014).

335. Sekula, P. *et al.* Genetic risk variants for membranous nephropathy: extension of and association with other chronic kidney disease aetiologies. *Nephrol. Dial. Transplant.* **32,** 325–332 (Feb. 2017).

336. Senut, B., Pickford, M., Gommery, D. & Ségalen, L. Palaeoenvironments and the origin of hominid bipedalism. *Historical Biology* **30,** 284–296 (2018).

337. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (Aug. 2005).

338. Simonti, C. N. *et al.* The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351,** 737–741 (Feb. 2016).

339. Skoglund, P. *et al.* Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344,** 747–750 (May 2014).

340. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336,** 466–469 (Apr. 2012).

341. Skov, L. *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* **14,** e1007641 (Sept. 2018).

342. Smith, F. J. *et al.* Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. *Nat. Genet.* **38,** 337–342 (Mar. 2006).

343. Soares, P. *et al.* The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* **29,** 915–927 (Mar. 2012).

344. Spoor, F. *et al.* Reconstructed Homo habilis type OH 7 suggests deep-rooted species diversity in early Homo. *Nature* **519,** 83–86 (Mar. 2015).

345. Stanescu, H. C. *et al.* Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N. Engl. J. Med.* **364,** 616–626 (Feb. 2011).

346. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21,** 577–581 (June 2003).

347. Stewart, J. R. & Stringer, C. B. Human evolution out of Africa: the role of refugia and climate change. *Science* **335,** 1317–1321 (Mar. 2012).

348. Stringer, C. Modern human origins: progress and prospects. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **357,** 563–579 (Apr. 2002).

349. Stringer, C. The origin and evolution of Homo sapiens. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **371** (July 2016).

350. Stringer, C. B. & Andrews, P. Genetic and fossil evidence for the origin of modern humans. *Science* **239,** 1263–1268 (Mar. 1988).

351. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44,** 1161–1165 (Oct. 2012).

352. Tamm, E. *et al.* Genome-wide analysis of Corsican population reveals a close affinity with Northern and Central Italy. *bioRxiv.* doi:10.1101/722165 (2019).

353. Tattersall, I. Human origins: Out of Africa. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 16018–16021 (Sept. 2009).

354. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (July 2012).

355. Thorpe, S. K., McClymont, J. M. & Crompton, R. H. The arboreal origins of human bipedalism. *Antiquity* **88,** 906–914 (2014).

356. Timmermann, A. & Friedrich, T. Late Pleistocene climate drivers of early human migration. *Nature* **538,** 92–95 (Oct. 2016).

357. Torroni, A., Achilli, A., Macaulay, V., Richards, M. & Bandelt, H. J. Harvesting the fruit of the human mtDNA tree. *Trends Genet.* **22,** 339–345 (June 2006).

358. Trinkaus, E. *et al.* An early modern human from the Peştera cu Oase, Romania. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 11231–11236 (Sept. 2003).

359. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* **42,** D975–979 (Jan. 2014).

360. Underhill, P. A. & Kivisild, T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* **41,** 539–564 (2007).

361. Urbanowski, M. *et al.* The first Neanderthal tooth found North of the Carpathian Mountains. *Naturwissenschaften* **97,** 411–415 (Apr. 2010).

362. Van Hout, C. V. *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv.* doi:10.1101/572347. eprint: https://www.biorxiv.org/content/early/2019/03/09/572347.full.pdf. <https://www.biorxiv.org/content/early/2019/03/09/572347> (2019).

363. Van Zinderen Bbarker, E. M. A Late-Glacial and Post-Glacial Climatic Correlation between East Africa and Europe. *Nature* **194,** 201–203 (Apr. 1962).

364. Veeramah, K. R. & Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat. Rev. Genet.* **15,** 149–162 (Mar. 2014).

365. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343,** 1017–1021 (Feb. 2014).

366. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352,** 235–239 (Apr. 2016).

367. Vigne, J. D. The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. *C. R. Biol.* **334,** 171–181 (Mar. 2011).

368. Villanea, F. A. & Schraiber, J. G. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nat Ecol Evol* **3,** 39–44 (Jan. 2019).

369. Villmoare, B. *et al.* Paleoanthropology. Early Homo at 2.8 Ma from Ledi-Geraru, Afar, Ethiopia. *Science* **347,** 1352–1355 (Mar. 2015).

370. Wall, J. D. *et al.* Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194,** 199–209 (May 2013).

371. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (Sept. 2010).

372. Wang, S. *et al.* Genetic variation and population structure in native Americans. *PLoS Genet.* **3,** e185 (Nov. 2007).

373. Weaver, T. D. Out of Africa: modern human origins special feature: the meaning of neandertal skeletal morphology. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 16028–16033 (Sept. 2009).

374. Weidenreich, F. Some Problems Dealing with Ancient Man. *American Anthropologist* **42,** 375–383 (1940).

375. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92,** 414–417 (Oct. 2012).

376. Wolf, A. B. & Akey, J. M. Outstanding questions in the study of archaic hominin admixture. *PLoS Genet.* **14,** e1007349 (May 2018).

377. Wright, S. The genetical structure of populations. *Ann Eugen* **15,** 323–354 (Mar. 1951).

378. Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun* **8,** 15927 (June 2017).

379. Yunusbayev, B. *et al.* The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* **29,** 359–365 (Jan. 2012).

380. Yunusbayev, B. *et al.* The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet.* **11,** e1005068 (Apr. 2015).

381. Zou, J. *et al.* A primer on deep learning in genomics. *Nat. Genet.* **51,** 12–18 (Jan. 2019).

382. Zwyns, N. *et al.* The Northern Route for Human dispersal in Central and Northeast Asia: New evidence from the site of Tolbor-16, Mongolia. *Sci Rep* **9,** 11759 (Aug. 2019).