UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Joanna GIEMZA

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : Biologie Santé

Spécialité : Génétique, Génomique, Bioinformatique
Unité de recherche : l'institut du thorax
Laboratoire des Sciences
du Numériques de Nantes (LS2N)

Soutenue le 7 Mai 2019

Fine-scale genetic population structure in France

JURY

Président :	**M. Michael BLUM**, Directeur de recherche, Université Grenoble Alpes
Rapporteurs :	**M^{me} Evelyne HEYER**, Professeur des universités, Musée de l'Homme **M. James F. WILSON**, Professeur Etranger, University of Edinburgh
Co-encadrants de thèse :	**M^{me} Géraldine JEAN**, Maîtresse de conférences, Université de Nantes **M. Christian DINA**, Ingénieur de Recherche, CNRS
Directeur de thèse :	**M. Richard REDON**, Professeur des universités, Université de Nantes

Remerciements

This work has been financed by the French Regional Council of Pays de la Loire (Project GRIOTE, set up by Jérémie BOURDON, Richard REDON, Dominique TESSIER) and the Agence Nationale de la Recherche (Project ID: ANR-14-CE10-0001, coordinated by Richard REDON).

First of all, I would like to thank my supervisors: lead supervisor Richard REDON and co-supervisors Christian DINA and Géraldine JEAN. I feel lucky that I had an opportunity to work with these three wonderful people. I especially thank Christian DINA for his support, ideas and enthusiasm, Richard REDON for his vast experience and individual approach, Géraldine JEAN for her professional attitude. Thanks to them I learned a lot during 4 years of working together.

To Isabel ALVES who generously shared her expertise in population genetics since she joined our lab in April 2018. Discussions with her and her literature recommendations have been invaluable.

Special thanks to Evelyne HEYER, James F. WILSON and Michael BLUM who accepted to be my examiners.

I would like to thank Michael BLUM and Masha NIKOLSKY who accepted to be in my *comité de suivi de thèse* and monitored the progress of this thesis. Your feedback at the end of each year was very helpful.

Special thanks to all French volunteers who participated in SU.VI.MAX and VaCaRMe / PREGO projects. It has been exciting for me, as someone who is not at all French, to have the opportunity to spend the last 4 years of my life trying to disentangle links between your genetics and history, geography, linguistics. Again, to Christian DINA, head of population genetics project, who hired me even though I hadn't had any population genetics experience back in 2015. To Stephanie CHATEL who was a coordinator of VaCaRMe project and to all collaborators. I acknowledge also Serge HERCBERG and Pilar GALAN who conceived and led the SU.VI.MAX study. To Aude SAINT-PIERRE and Emmanuelle GÉNIN, my collaborators on SU.VI.MAX and 3C project. I also thank historians Martial MONTEIL for preparing historical maps of France and Yves COATIVY for help in interpreting results from the PREGO dataset.

I acknowledge Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest), whose cluster I used for my computational analyses. I thank Audrey BIHOUÉE, Eric CHARPENTIER and particularly Jean-François GUILLAUME for technical support. The latter helped me tens of times in situations I felt helpless.

I thank Pierre LINDENBAUM, Eric CHARPENTIER, Matilde KARAKACHOFF and Adrien FOUCAL for their help and tips in programming.

I thank Floriane SIMONET, Eric CHARPENTIER and Matilde KARAKACHOFF for initial treatment of PREGO data.

Special thanks to Charlotte BERTHELIER for her M2 project linked to my research.

To all members of Team 1 Cardiovascular Genetics, in particular its leader Jean-Jacques SCHOTT, and to all members of GenoBiRD platform. Special thanks to statistics group MAGES (leader Christian and Matilde, Floriane, Elodie, Sidwell, Clément, Pierre-François) and POPGEN group (Christian, Isabel and Charlotte).

To the secretary of the team Ophélie TINDILIÈRE and general secretary Isabel RIVAUD.

To all people from l'institut du thorax not cited yet, who contributed to make this workplace so pleasant.

I would like to thank especially PhD students Elodie, Wafa, Lindzy and Marine, for their friendship and support. The PhD journey has been by far better in their company.

Finally I thank my family and my friends, for their unconditional support and believing in me.

Publications arising from this thesis

— Aude Saint-Pierre*, Joanna Giemza*, Céline Bellenguez, Luc Letenneur, Claudine Berr, Carole Dufouil, Philippe Amouye, Martial Monteil1, Serge Hercberg, Pilar Galan, Richard Redon, Emmanuelle Génin*, Christian Dina*. 2019. **The Genetic History of France**, *submitted*.
*These authors contributed equally

— Joanna Giemza et al. 2019. **The genetic structure of historical population in Northwestern France**, *in preparation*.

— Alessandro Raveane, Serena Aneli, Francesco Montinaro, Georgios Athanasiadis, Simona Barlera, Giovanni Birolo, Giorgio Boncoraglio, Anna Maria Di Blasio, Cornelia Di Gaetano, Luca Pagani, Silvia Parolo, Peristera Paschou, Alberto Piazza, George Stamatoyannopoulos, Andrea Angius, Nicolas Brucato, Francesco Cucca, Garrett Hellenthal, Antonella Mulas, Marine Peyret-Guzzon, Madzia Zoledziewska, Abdellatif Baali, Clare Bycroft, Mohammed Cherkaoui, Christian Dina, Jean-Michel Dugoujon, Pilar Galan, Joanna Giemza, Toomas Kivisild, Mohammed Melhaoui, Mait Metspalu, Simon Myers, Luisa Mesquita Pereira, Francois-Xavier Ricaut, Francesca Brisighelli, Irene Cardinali, Viola Grugni, Hovirag Lancioni, Vincenzo Lorenzo Pascali, Antonio Torroni, Ornella Semino, Giuseppe Matullo, Alessandro Achilli, Anna Olivieri, Cristian Capelli. 2018. **Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe**, *in press*, doi: https://doi.org/10.1101/494898

Contents

List of Tables

List of Figures

1

Introduction

1.1 Historical perspective of human population genetics

1.1.1 Luigi Luca Cavalli-Sforza, the pioneer of the field

On 31st August 2018, by coincidence the day I obtained a key result of my thesis, Luigi Luca Cavalli-Sforza died at age 96. He was a pioneer of the field of human population genetics. A few days later a blog article "The man who tried to catalogue humanity" [74] was published by paleoanthropologist John Hawks. I begin my thesis manuscript with the description of the career of Cavalli-Sforza as his legacy constitutes the foundations for my project. His work is best summarised in his encyclopedia-sized book "The history and geography of human genes" [33], but the blog post mentioned above is a more objective source that also points out pitfalls in the approach of Cavalli-Sforza (like every human and scientist, many of his ideas turned out to be wrong in the details). The blog post is more up-to-date too and I write this subsection on its basis.

Cavalli-Sforza entered the field in the 1950s. He developed ways to test people for invisible traits such as blood types (Figure 1.1). These traits became known as "classical markers". Classical markers remained state-of-the art evidence of human genetic variation until 1980s.

Cavalli-Sforza first studied the population in the Parma Valley in his native Italy. He tried to understand how inbreeding within this region was connected to the slight differences in frequency of blood groups, tracing church records of marriages and births. He was able to show that consanguineous marriages were the main drivers of genetic variation between small towns of Parma Valley. By doing so, he provided one of the earliest evidence that humans were still being affected by *genetic drift*, the random change in *allele* frequencies that impacts particularly smaller populations.

Cavalli-Sforza realized that genetic drift might have affected humanity over a much deeper past. He thought that, as genetic drift is a force that drives populations slowly apart over long period of time, a tree would be an appropriate illustration of the human evolutionary history.

The tree looked different from the tree obtained by his lab over 40 years later (Figure 1.2). Nowadays it is known that African populations are the most diverse populations, while in a tree published by Cavalli-Sforza in 1966 they represented a minor twig. Cavalli-Sforza's early trees turned out to be wrong due to a phenomenon called "*ascertainment bias*". Blood groups were first discovered and studied in people of European descent. As a consequence, African variation was not fully included in the set of traits that vary in Europe. Recently, using mitochondrial DNA (mtDNA), Y chromosome and genome-wide datasets, it was shown that the deepest branches of human population trees are African. On the other hand, the early evolutionary trees illustrate one of the limitation of classical markers. Their small number did not provide

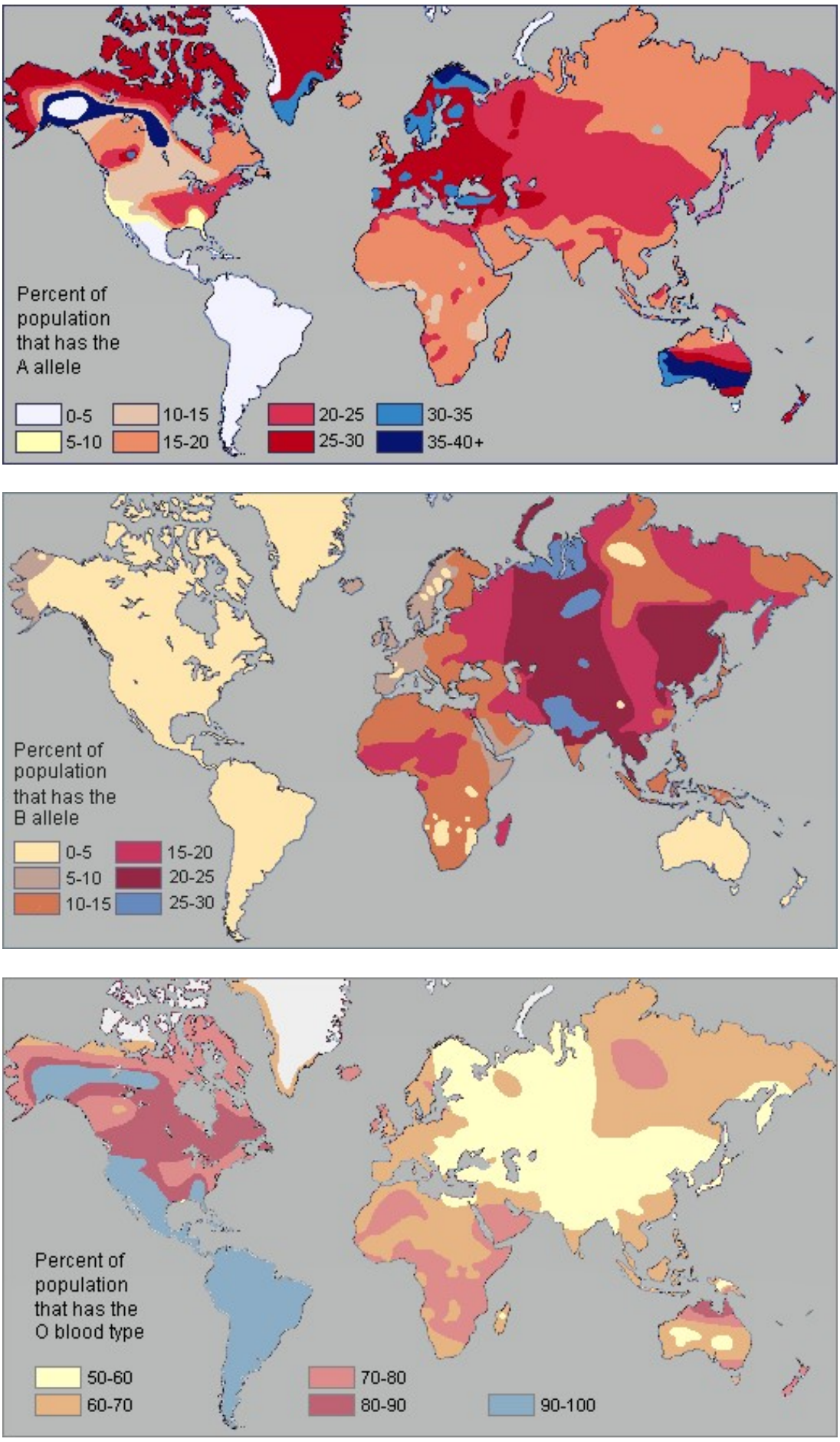


Figure 1.1 – Distribution of blood types in native populations of the world. Source of images: [180], however they were reproduced from [146].

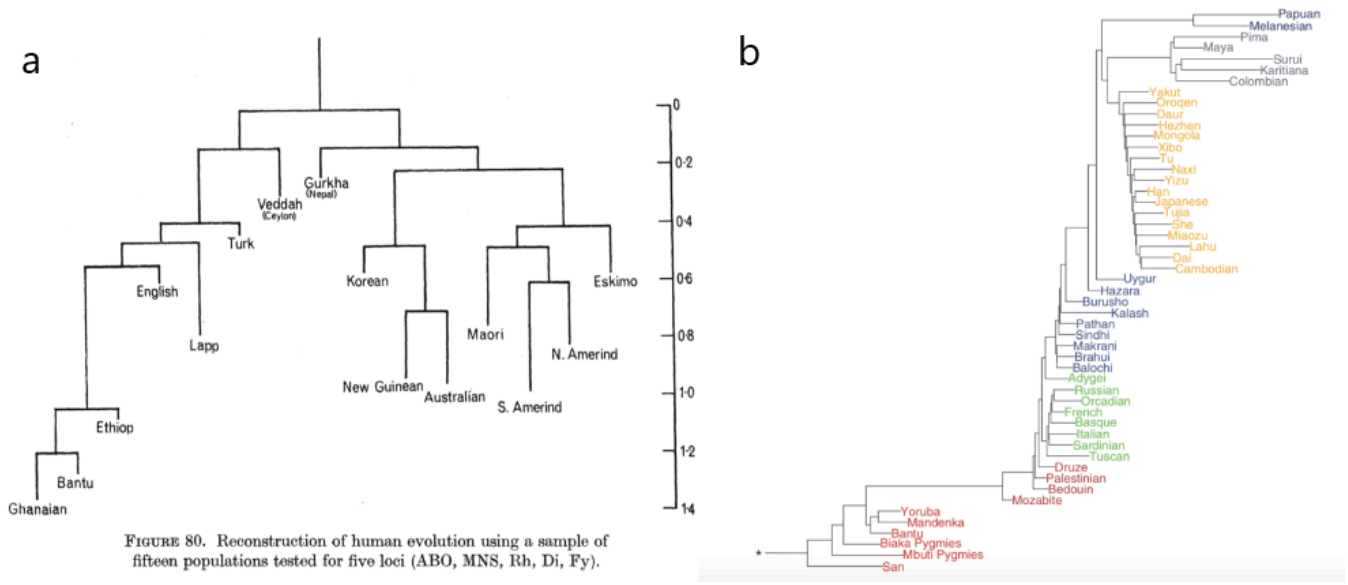


Figure 1.2 – Cavalli-Sforza's trees, a) published in 1966 [30], b) published in 2008 [118]. The substantial difference is in the placement of the root.

enough information to draw the right conclusions. Cavalli-Sforza was willing to assume that migration and *admixture* were rare. The main evolutionary force in his model was genetic drift, underestimating other forces, such as *gene flow*, which in today's understanding play a considerable role.

Subsequently, Cavalli-Sforza used an early computer to run principal component analysis on the genetic differences between populations. In this way, they could show how the correlations of many gene frequencies, not only single *loci*, changed on a map (Figure 1.3). In Cavalli-Sforza's vision, a gradient across all the classical markers could show the migration routes in the past. What the maps with principal components of genetic variation could not show was how and why the migrations had happened [153, 53].

To disentangle this problem, Cavalli-Sforza turned out his attention to Neolithic, the time when agriculture first spread from Near East into Europe. He believed this event should have been the most powerful to sculpt gene frequencies across Europe. By the early 1970s, the radiocarbon dating of the archaeological sites with signs of Neolithic traditions revealed a pattern. The Neolithic had spread slowly, around one kilometer a year, from southeast to northwest. It had a clear implication for Cavalli-Sforza: farming spread as farming populations gradually increased in size, bringing with them a new way of life. He claimed a diffusion of culture together with genes and named it *demic diffusion*. It was a dominant model of demographic change for the Neolithic in the 1980s and 1990s. His interpretations of PCA with respect to Neolithic turned out to be wrong as pointed in 2008 by Novembre and Stephens [153].

One of his further initiatives was the Human Genome Diversity Project (HGDP) [32] that aimed to collect genetic samples from over one thousand people from small-scale societies around the world to capture more closely the breadth of human diversity. The main idea of HGDP was to use a novel technique to transform samples into immortal cell lines. Unique genetic variation of indigenous people from isolated populations was preserved in this way and cell lines were made available for distribution to research laboratories. Cavalli-Sforza saw the HGDP as a necessary corrective for Human Genome Project, which at time was spending billions of dollars on a first draft of the sequences of all 23 pairs of human chromosomes.

I think a reader is now familiar with the seminal ideas of Luigi Luca Cavalli-Sforza. Even though his ideas and models were not always right, his work helped to set up foundations of our current knowledge of human variation across the world. Cavalli-Sforza and his colleagues led human population genetics from the birth of the domain in the mid-1950s to 1990s. In the next subsection I follow a topic of historical perspective of population genetics, presenting state-of-the-art insights on human evolution.

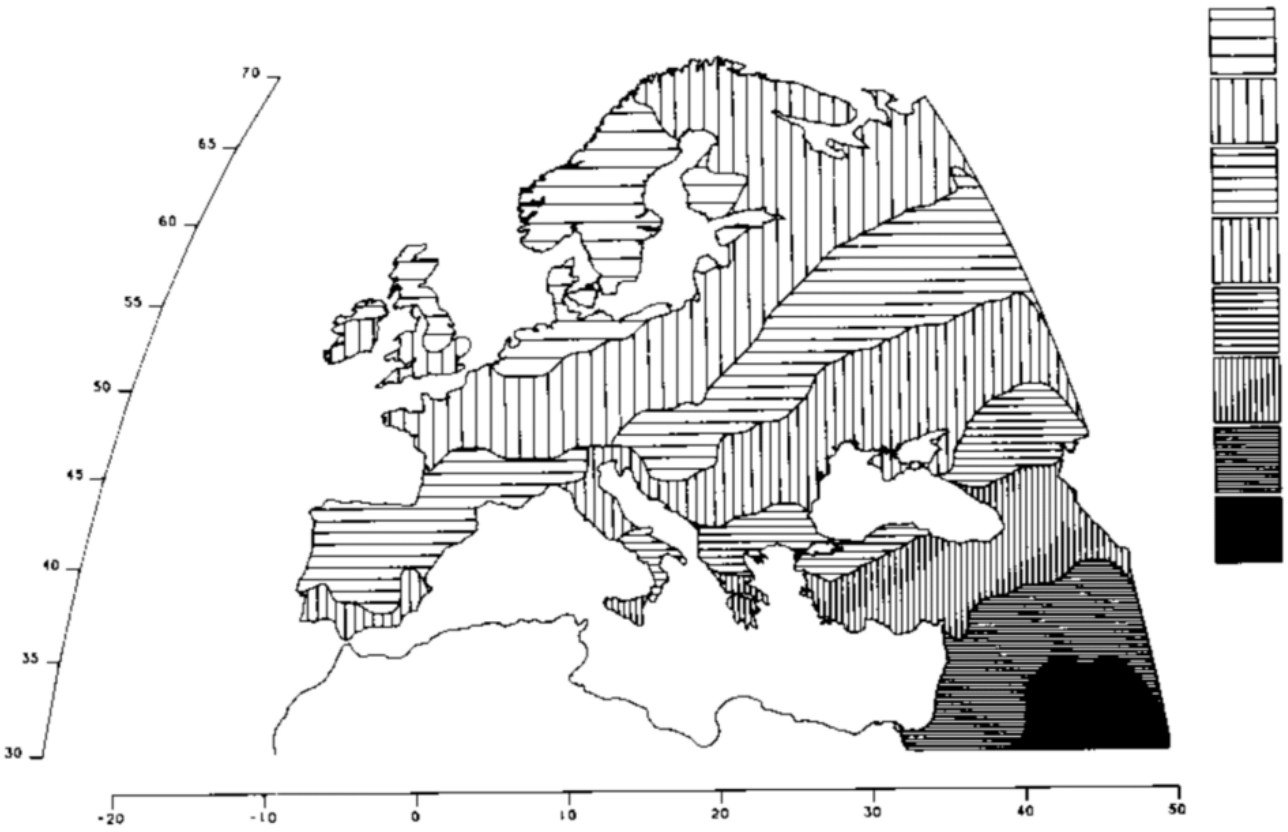


Figure 1.3 – The first principal component of genetic variation across Europe, from [31].

1.1.2 Human evolution: insights from human population genetics

Population genetics is a domain that studies the evolutionary forces responsible for generating and shaping genetic variability within and between populations over time and space. Here, I provide a brief overview on the current knowledge of the evolutionary history of our species, presenting the most important discoveries, conclusions and questions remaining in the field.

Human evolution used to be studied with archaeological and palaeontological data. These data facilitated inference about the events that led to the emergence and spread of anatomically modern humans (AMH). However, they may be insufficient to determine the genetic relationships between different individuals or groups of people. Importantly, interpretation of the evolutionary history and adaptation of humans is being transformed by analyses of new genomic data [148]. The power of genomic data from modern or ancient people is a possibility of direct determination of the genealogical relationships between humans as well as the elucidation of migration routes, diversification events and genetic admixture among various groups.

Human population genetics in the 1980s, 1990s and 2000s - uniparental markers

Since 1980s, a popular tool of genomic research was mtDNA. Thanks to it the out-of-Africa hypothesis, proposing that modern humans originated in Africa, from where they expanded later outwards, was accepted [218]. Widely rejected alternative hypothesis was the multiregional model, which suggests that evolution of anatomically modern humans took place simultaneously in multiple locations and was facilitated by gene flow. Nevertheless, there are drawbacks of the mtDNA approach. First of all, it reflects only female inheritance. Moreover, mtDNA does not recombine, thus it has the information content of only a single genetic marker. As a consequence, there is a considerable chance that a phylogenetic tree inferred from only

one marker is not representative of the overall genomic pattern and history of human evolution. Analysis of the nuclear genome is thus essential. Subsequently, the study of Y chromosome gained attention, but it suffers from similar drawbacks to mtDNA.

Genomic analyses conducted in the 1980s, 1990s and 2000s have left several missing puzzles in the history of human evolution. It has been left unclear if gene flow had occurred between anatomically modern humans and other hominins [149]. The origins of Native Americans have not been yet deciphered with confidence as well [143]. Lastly, the relative importance of the movement of people and ideas in cultural revolutions, such as for example agriculture [5], often remained unclear. These open questions about human history and evolution have been addressed recently thanks to considerable advances in DNA sequencing and methods for the enrichment and extraction of ancient DNA. Year 2010 was a milestone when the first genomes from ancient humans [177] as well as other hominins [66] were retrieved. Ancient DNA genomes led to numerous scientific breakthroughs in our understanding of human evolutionary history. A discovery of interbreeding between AMH and extinct hominins, Neanderthals, Denisovans and possibly others, is probably the most important of them [66, 168, 55, 179, 198]. Also, we have learned with increasing detail and complexity about the dispersal of modern humans out of Africa and their population expansion worldwide [148]. In the following years numerous ancient DNA studies have been completed. Overall, the addition of both temporal and geographic aspects to genome sequencing, by including samples from a wide range of historical times and locations, has provided fresh insights into human evolutionary history, some of which I summarise further in this subsection.

Brief summary of human evolutionary history

Anatomically modern humans originated in Africa. The earliest evidence comes from fossils located in Morocco that can be dated to about 300 thousand years (kyr) ago [89]. Beyond Africa, the eldest fossils date back as early as about 100 kyr ago in the Middle East [67] and about 80 kyr ago in southern China [122]. However, other hominins, such as Neanderthals, which disappeared from the fossil record about 40 kyr ago [80], have been found throughout Eurasia as far back as 400 kyr ago. A recent review [148] proposes a cartoon with a simplified model of human evolutionary history (Figure 1.4). On the other hand, Figure 1.5, also from [148], illustrates major human migrations across the world inferred through analyses of genomic data. Expansion outside of Africa gradually led to colonisation of the rest of the globe, which was accompanied with a serial founder effect [9, 175]. Divergence times remain under debate, the deepest split between population of humans is captured by genetic lineages of click-language-speaking San populations of southern Africa that diverged around 160-110 kyr ago [166, 192, 68, 193, 216], the out of Africa bottleneck happenned 50-100 kyr ago [68, 127, 65, 117, 72, 190, 189] and Europeans diverged from Asians 36-45 kyr ago [127, 173, 56, 196].

An exciting topic of research is human-hominin interactions. The ancestors of all contemporary non-Africans encountered, and admixed with, Neanderthals. A non-African individual contains on average 2% Neanderthal ancestry, with higher levels occurring in East-Asian individuals than in Europeans [221, 217, 184]. It suggests that admixture mostly occurred shortly after the dispersal of anatomically modern humans from Africa and is consistent with a single–dispersal-based out-of-Africa model [127]. The date of hybridization has been estimated to be approximately 50–65 kyr ago [185]. At least one other type of archaic human — the enigmatic Denisovans — lived in Eurasia when the first modern humans started to appear in the continent. Little is known about the morphology and the geographical dispersal of Denisovans, who are known only from the genome sequences of a finger bone and three teeth that were found in the Denisova Cave in Siberia [179, 188, 204]. Last year an extraordinary discovery of half Neanderthal and half Denisovan genome took place [201]. It was the first time a first generation offspring of interbreeding hominins was found.

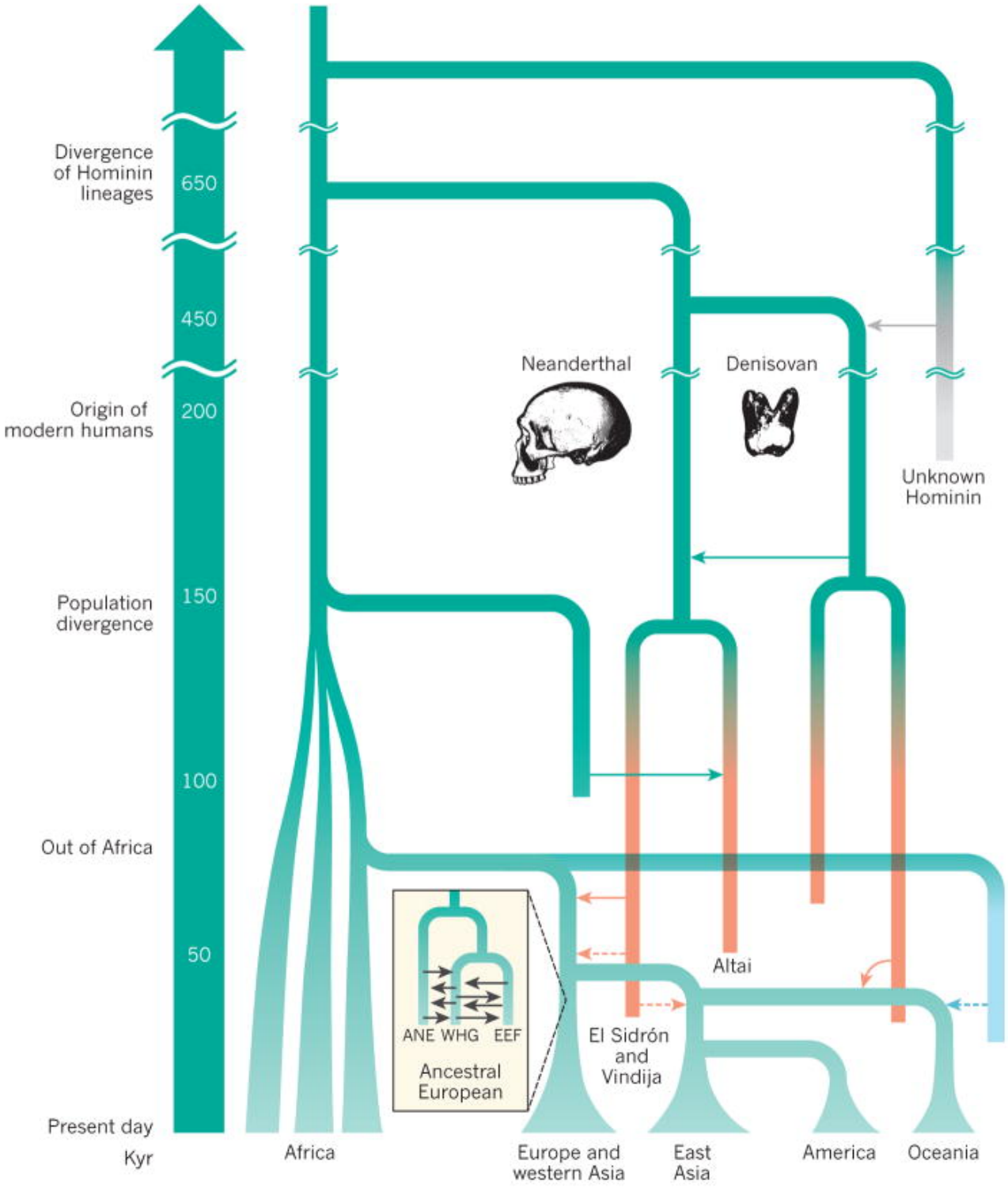


Figure 1.4 – Simplified model of human evolutionary history, from [148]. Relationships between contemporary populations and the approximate times at which they diverged are shown. These include important well established (solid lines) and tentative (dashed lines) admixture events between groups of modern humans and between modern and archaic humans. The model also shows the potential small proportion of ancestry in Oceanic populations that is derived from an early out-of-Africa migration (turquoise). Studies of ancient DNA can provide high-resolution insights into the history of populations and have revealed that present-day Europeans comprise admixture between three ancestral groups [112] (inset). ANE, ancient north Eurasian; EEF, early European farmer; WHG, west European hunter-gatherer.

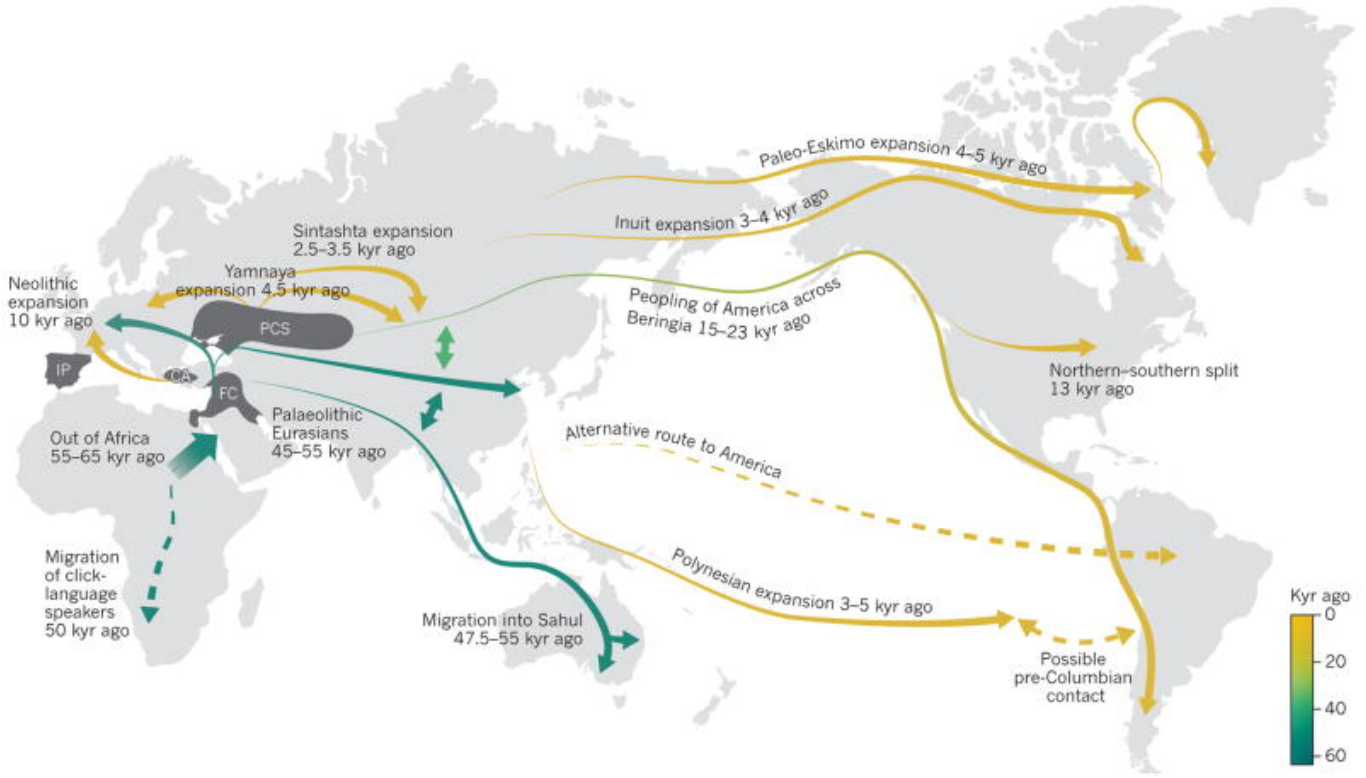


Figure 1.5 – Major human migrations across the world inferred through analyses of genomic data , from [148]. Some migration routes remain under debate. For example, there is still some uncertainty regarding the migration routes used to people the Americas. Genomic data are limited in their resolution to determine paths of migration because further population movements, subsequent to the initial migrations, may obscure the geographic patterns that can be discerned from the genomic data. Proposed routes of migration that remain controversial are indicated by dashed lines. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe.

The peopling of Europe

I focus on the peopling of European continent as it is the most relevant to the context of the origins of the French population. Human evolutionary history in Europe has been studied more extensively than in other continents.

Palaeolithic The first anatomically modern humans lived in Europe as early as 43 kyr ago [80, 13, 55]. These early Palaeolithic Europeans have probably made little genetic contribution to the European people of today [64]. Climate oscillations related to the Last Glacial Maximum (LGM) likely caused the turnover in the genetic composition of Europeans [57] — although the exact contributions from early Europeans is still under debate [64].

Mesolithic European populations are likely to be composed of at least three genetic components that entered Europe at different times [200, 112, 71, 59, 4, 64] (see again Figures 1.4 and 1.5). Western European hunters-gatherers (WHG) that appeared about 15 kyr ago [57] and Eastern European hunters-gatherers (EHG), attested in European Russia about 8 kyr ago [71, 137], were considerable differentiated. Hunter-gatherers had low *effective population size* [112]. WHG dominated across much of the mainland Europe until 7th millennium BC.

Neolithic Around 11 kyr ago, after the LGM had passed, a new way of life based on animal husbandry, agriculture and sedentism appeared, called a Neolithic lifestyle [6]. It initiated in the region of Anatolia /

Near East. Neolithic farmers began to dominate in number over WHG since the 7th millenium BC [137, 85]. They arrived via southeastern Europe and propagated their ancestry as far as Scandinavia [199, 200] and Iberia [71]. The WHG populations were, nevertheless, persistent, with individuals with predominant WHG ancestry found as late as 4th millenium BC [120, 18].

Neolithic populations had larger effective population size than hunter-gatherers [199], suggesting that the Neolithic lifestyle helped to increase the size of populations. Interestingly, current genetic variation in Europe shows a gradient of decreasing diversity with increasingly northern latitudes, no matter what diversity measure is used [9], which can be linked to the fact that Neolithic arrived from southeastern Europe and to the difference in effective population size between WHG and Neolithic population. The Neolithic populations have larger contributions in modern southern European population and lesser in the North. For instance, the genetic make-up of present day Sardinians contains high levels of the Neolithic genetic component [200, 112, 199]. Conversely, hunther-gatherers contributed more to present day northern Europeans rather than southern Europeans.

Neanderthal ancestry in Europe seems to have been reduced over the Ice Age [57] by natural selection against Neanderthal variants [98, 184]. Further reduction of Neanderthal ancestry took place in the Neolithic period by migration from the near east. Migrants descended from a postulated group of "Basal Eurasians" that have no Neanderthal ancestry at all [112, 111]. The origins of Basal Eurasians are uncertain. Both Anatolian farmers and Caucasus hunter-gatherers had ancestry from this population [96] and through them so do later Europeans.

Steppe related ancestry The third European genetic component is linked to Yamnaya culture. The occurrence of this component dates back to the late Neolithic period and the early Bronze Age. Yamnaya culture was brought by herders from the western Eurasian steppe that migrated to central Europe about 4.5 kyr ago. [71, 4]. The herders themselves were a mix of at least two elements [71], the EHG and a southern population element related to present-day Armenians [71], ancient Caucasus hunther-gatherers [96] and farmers from Iran [111]. This migration is being linked to conquests and technological innovations such as horseback riding, and possibly caused the spread of Indo-European languages to Europe [71, 4] (although certain linguistics researchers debate that these languages were already spoken by Neolithic farmers [19]).

Southeastern Europe received steppe-related ancestry before any other population in Europe outside the steppe itself, as early as 6.7-6.5 kyr ago in Bulgaria [136]. However, the presence of steppe ancestry in this part of Europe was low-level, around 30% during Bronze Age, about 5.4-3.1 kyr ago [136], and around 15% in the Aegean during the Mycenaean Period about 3.5 kyr ago, but absent in Minoan culture of Crete [110], which represents the most recent sampled European population without steppe ancestry [109]. The steppe migrants made a massive impact in central and northern Europe [71, 4]. Late Neolithic people from the Corded Ware material culture from Germany traced 75% of their ancestry to the Yamnaya [71]. People from the Corded Ware culture in the Baltics were genetically similar to those from Central Europe [144, 182]. During the Middle Bronze Age about 3.5-3 kyr ago a nearly complete turnover of ancestry happened in Britain, local Neolithic individuals were 90% replaced by a population that lived in continental Europe before 2450 BC and bore steppe ancestry [157]. Steppe ancestry was also present in Bronze Age Ireland [29] and Iron-Age and Anglo-Saxon England [191]. Steppe ancestry arrived in Iberia too, [135], but had a lesser impact there [157]. Today, steppe related ancestry is lower in southern Europe and higher in northern Europe [71].

Perspective In the next years, ancient DNA research will hopefully address questions of later European history, such as how did the elements of the populations of Europe combine during Antiquity, the Migration Period and the following centuries to form the genetic diversity of present-day Europeans.

1.2 Brief history and geography of France

1.2.1 Modern France description

The territory of modern France is located at the northwestern end of the European platform mostly surrounded by natural borders: the Atlantic Ocean on the western side, the English Channel on the northern, the Pyrénées and the Alps on the southwestern and eastern sides, respectively, as well as the Mediterranean Sea on the southernmost edge. In spite of these natural borders, France has no major mountain barriers along its eastern border, making it not only the final goal of a large number of migrations (Celts, "Invasions Barbares") but also a place of transit either to the North (British Isles) or the South of Europe (Iberian peninsula) and North Africa. Modern French territory also hosts many sites reporting early human peopling and the Aurignacian rock shelter of Cro-Magnon gave the name of the species which was later to be called European Early Modern Humans. As for modern migration, France's position was also at the crossing of the assumed expansion era of early Western Hunter-Gatherers (15 kyr ago), Neolithic farmers (7 kyr ago) and later steppe Eneolithic Age populations [109, 112].

1.2.2 Geographic scope and the people of Pre-Roman Gaul (IInd-Ist century BC)

Gallia est omnis divisa in partes tres [27] was one of the earliest demographic description of antique France (known as Gaul). These three parts were Aquitania, in the South West, with Garonne and the Pyrénnées mountains as borders; Belgia in North East, following the Seine as Southern border; and finally what we know as Celtic Gaul, which spans from the Atlantic Ocean to the Rhine River and Alps (Figure 1.6). A fourth part of the present-day French territory, already part of Romanized territories at this time, is Gallia Narbonensis, which is a strip of lands from Italy to Iberia, with Alps and Cevennes mountains as northern border. Finally, the Cisalpine Gaul was already a Roman province, where populations were a mix of ancient Celtic tribes, although some of them – like the Boii in 191 BC – were slaughtered, and Roman citizens (Italians) colonies.

The territory of modern France seems to have been only a part of a larger Celtic territory, spanning to the Hungarian plains and where differentiation with the so-called German tribes was not as obvious as the official Roman history reports it. Finally, the southwestern population was identified as different from the Celtic population, for instance by Strabo [203], who defined them as Aquitani and close to Vascones, ancestors of present-day Basques.

The part of Gaul which was not initially part of the Roman influence (Gallia Comata or Long-Haired Gaul), seems to have been organized as a juxtaposition of civitates or pagi, which were the land of Gaul tribes or nations. Although Figure 1.7 is representing the Roman provinces from the Ist century AD, the ancient civitates are also displayed. In terms of peopling, it is thought that Belgae hosted a mixture of Celtic and Germanic populations.

A long period of incorporation into the Roman Empire led to the flourishing of a Gallo-Roman culture. Demographically, Roman elements but also elements from the whole Empire (mainly around Mediterranean Sea) could have been directed to what was Gaul, for instance through the settlement of veterans from Roman legions. Although such settlement existed, like in Lugdunum (Lyon) [206, 130], it does not seem to have occurred at a large scale.

1.2.3 Great Invasions (IIIrd-VIth century AD)

Later on, started the Great Invasions which initially consisted in infiltrations of so-called Germanic tribes. These infiltrations, with various degrees of violence, started around the IIIrd century AD.

Subsequently, the "immigration" of Germanic tribes was organized by the Romans, in order to create federate kingdoms that could, in turn, take part into the Empire's defence. Visigoths were settling the South West, Franks, which were to give their name to the country, in the North East and Burgunds in the South East. These peoples were soon able to establish independent kingdoms, even before the formal end of

Figure 1.6 – Map of Gaul in the times of Caesar (Ist century BC).

the Roman Empire. The Franks extended their northern territory to the land between Seine and Loire, the Visigoths established a vast kingdom spanning from Strait of Gibraltar up to the Loire River; to have a common border with the Franks. The South East was the land of the Burgunds (and would later become Burgundy). The extreme south was disputed with the Ostrogoths installed in Italy.

Other tribes, like the Sueves and the Vandals, just crossed Gaul *en route* to Spain and North Africa.

Although it does not seem that Germanic tribal migrations involved very large groups [76] of people, they still represented migrations of whole tribes – males and females. Therefore, the impact of such migration on the French demography remains to be evaluated. It is worth pointing out that non-Germanic populations, like steppe Alans [11] and Huns also immigrated in the same lands, along with the previously described Germanic tribes.

Figure 1.7 – Map of provinces of Roman Gaul and Germanie at the end of Ist century AD.

1.2.4 Barbarian kingdoms to Merovingians (VIth century – 987 AD)

At the end of the VIth century (See Figure 1.8), what was to be modern France is divided into two main Barbarian kingdoms: the Northern Neustria and Austrasia. They were two parts of the initial Clovis Frankish kingdom. Aquitaine, which was previously part of the Visigoth kingdom – and lost to Franks in Vouillé battle (507) - was at that time mainly part of the Frankish kingdoms (mainly Neustria). Within Aquitaine, Novempopulania (Latin for "country of the nine peoples") corresponded to the original Aquitaine, inhabited by basque related tribes. In the South, Septimania represented the remains of the Visigoth kingdom and was even later to be part of the medieval Muslim territory al-Andalus. The Provence was also a territory of borders (*Marches*) and a buffer to Saracen raids and Italian Ostrogoth invasions.

Interestingly, the Brittany peninsula, in the North-West, remained independent from the Barbarian Kingdoms. It has, moreover, received migration from the British Islands from which Celtic populations that were fleeing Saxon invasions (and Gaelic raids). These populations mixed with the Romanized Gauls from Armorica.

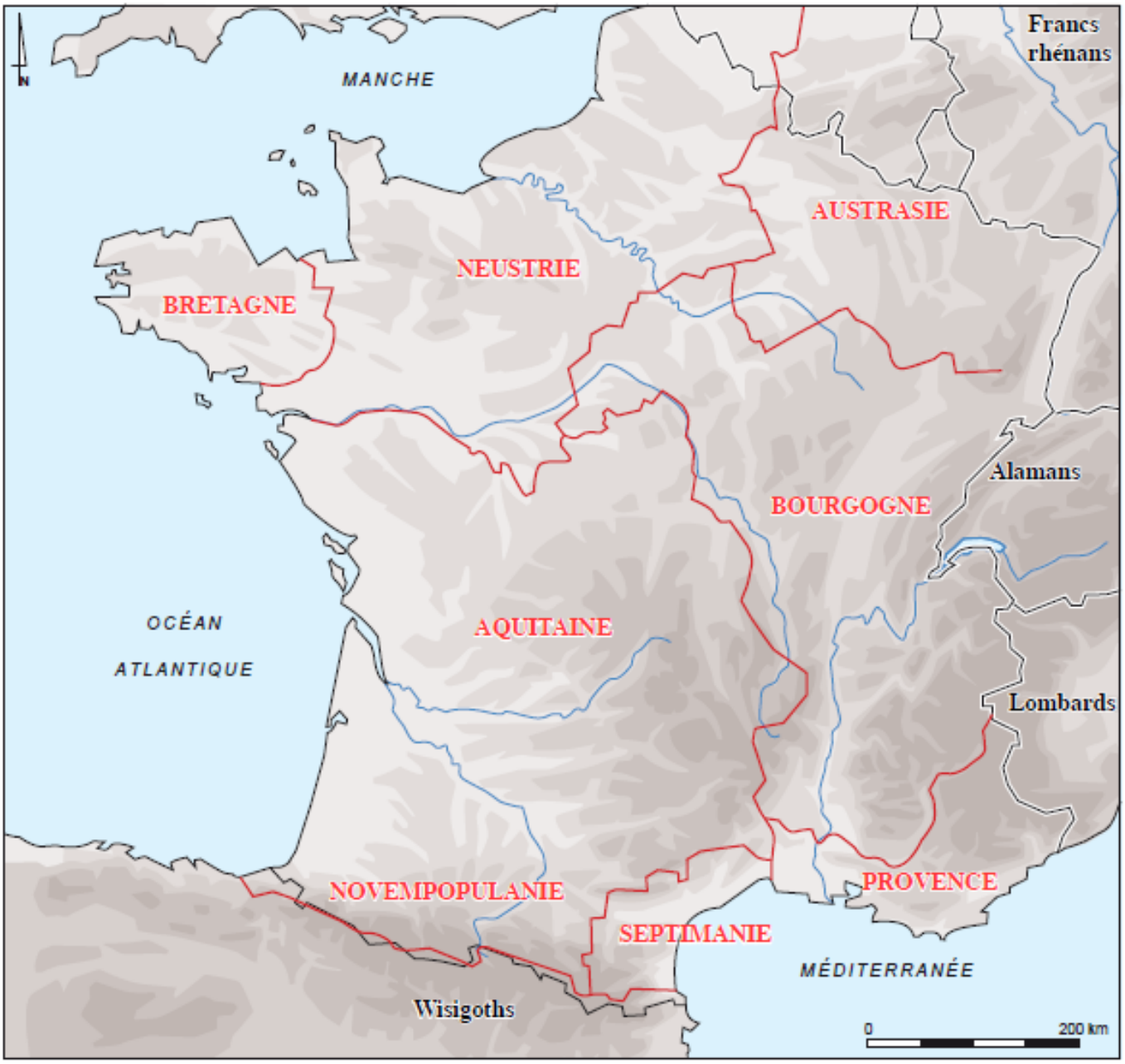


Figure 1.8 – Map of Barbarian Kingdoms in the VIIth century AD.

The Merovingian kings of the Frankish kingdom were soon replaced by Carolingian dynasty along which Charlemagne imprinted the biggest political changes. He was able to create an empire unifying territories of Modern France and Germany. He reorganized the administration into a system of "counties", headed by officials appointed by him bearing the title of counts. He, among others, prepared the future feudal system which prevailed the following centuries.

At Charlemagne's son's death, his empire was divided into three large kingdoms, through the Verdun treaty in 843 [104]. The eastern part covered a part of Germany, the central part, Lothringen, was a long corridor spanning from Netherlands down to Provence and Lombardia. The western part, or Occidental Francia is covering the West of the modern French territory, without Provence and the Rhône valley and without the Rhine valley. Brittany was also still independent during this whole period. The central part (Lothringen) was to disappear, disputed by the kingdoms that would evolve from the West and East Francia (France and Germany, respectively). Modern France would inherit the East banks of the Rhône, the Alp Mountains, the Rhine valley and Burgundy, Alsace and Lorraine and part of the North – although these regions would show in the meantime a fierce spirit of independence (see the Duchy of Burgundy). The

central authority progressively collapsed with the last Carolingians kings and counties progressively gained a quasi-independence.

1.2.5 Middle Ages and feudality (987-1643 AD)

When the Capetian dynasty finally replaced the last Carolingian, this process was irreversible. While the King remained the ultimate overlord, he had limited power over his vassals' territories, out of the king's own land (Domaine Royal).

A few political entities were cohesive and important, like the Viking-created duchy of Normandy and the Duchies of Anjou, Burgundy, Aquitaine (Figure 1.9).

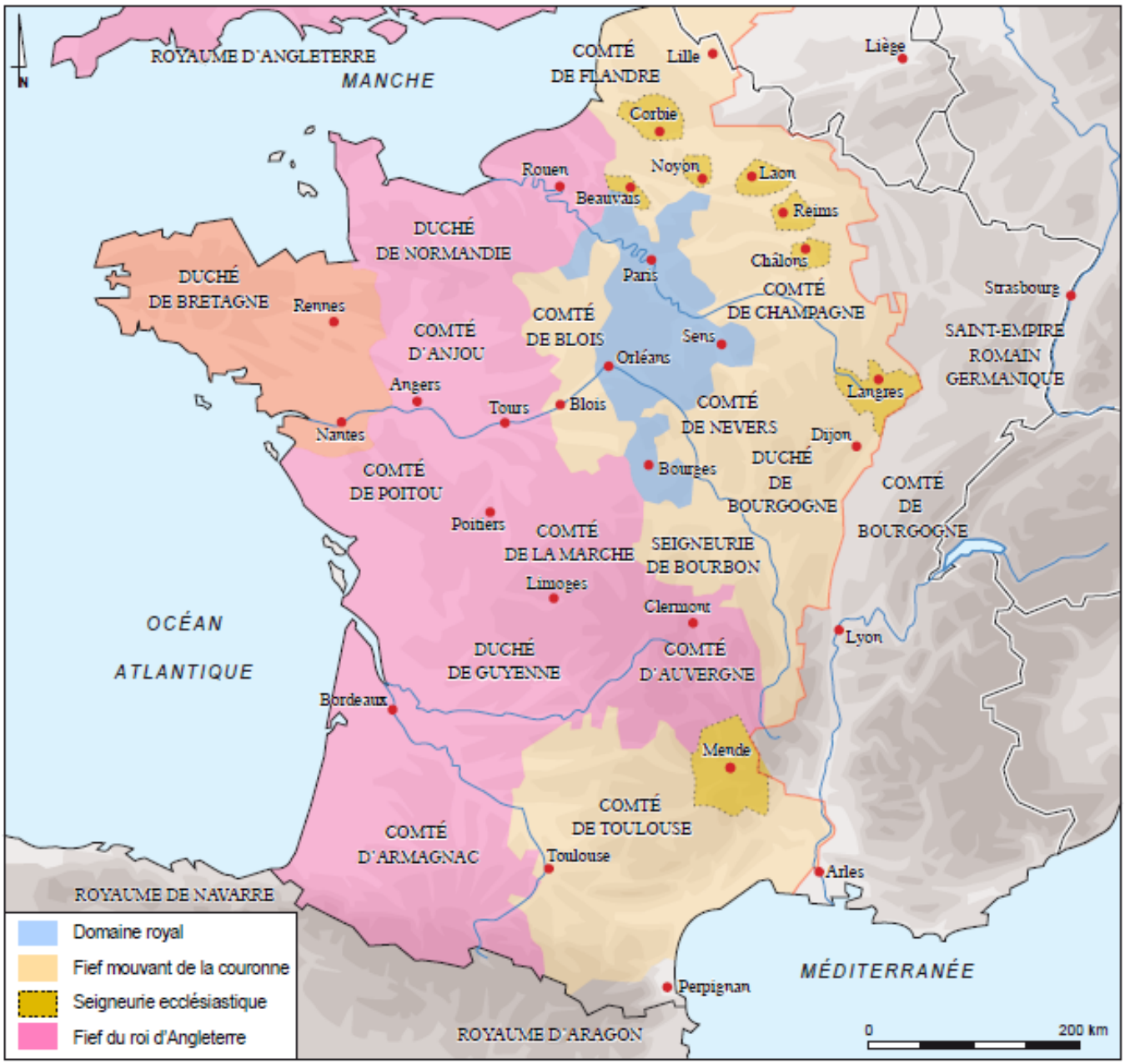


Figure 1.9 – Map of France in the XIIth century AD.

While the French kings, until Philippe Augustus, succeeded in gaining political control back, a new crisis, the Anglo-French Hundred Years war, created a new division. A large part of modern France became vassal to the English King who was also Duke of Aquitaine. While there was no trans-English Channel

migration which could have significantly changed the composition of a local population, this political pattern did certainly not favour population mixing. Some of the Gascon lords, for instance, serving the king of England did not speak a single word of French [42].

The end of the Anglo-French Hundred Years war, approximately in 1453, brought this division to an end. The King of France gained power over the whole territory.

Still, this royal authority was not uniform over all the entities, which differed in feudal rights and local institutions, laws and rules until the French Revolution (1789). During this period, language also varied a lot according to the Provinces (Figure 1.10) [35]. In spite of political unity, a developed road system and absence of large political obstacles, different language and cultural borders could have shaped the genetic structure of modern France. At this point, the extreme North (Artois), Alsace-Lorraine and the Rhine borders as well as south-east parts were not yet included in the Kingdom.

Figure 1.10 – Map of languages in France at the end of Hundred Years war. Source: [35]

1.2.6 Completion of Modern France (1643-1918 AD)

Louis XIV started expanding the French kingdom to reach defensible borders - notably the Rhine border. His wars allowed recovering Artois, Flanders and nearly the whole Alsace and Lorraine and reached the Rhine – although this Rhine border was returned to Germany for 48 years (1870-1918). During Wars of Revolution the French Empire, in spite of a final defeat – following one of the best military campaigns of Napoleon, in French homeland - secured the Rhine and the North border. Finally, Napoleon III, during the war for Italian liberation, managed to add Savoy and Nice to what was to be the final draft of modern continental France.

1.3 The importance of understanding population structure

1.3.1 The importance of population structure from an evolutionary perspective

Humans have long been preoccupied with their past. Out of fascination with their origins they have created numerous complementary disciplines such as history, linguistics and archaeology. Genetics play an increasingly important role too in the study of the human past, especially through characterizations of population structure, just as it was envisioned by Cavalli-Sforza and his colleagues.

Strictly, population structure is a term for non-random patterns of genetic similarity between individuals from the same species. It manifests itself in systematic differences in allele frequencies. Population structure would not be present if mating were random throughout the entire population history. Random mating, though it is often assumed in theoretical models, is not a case for most populations. One of the causes is physical separation, by geographical barriers such as oceans separating continents, high mountain chains or simply long distance. Moreover, for human populations specifically, lack of a common language might be an obstacle to mating and language patterns often coincide with genetic similarity patterns. Population structure is also influenced by forces such as migration, admixture, natural selection and genetic drift.

Novembre and Peter [151] state that ignoring population structure would result in only a 5–15% average error when predicting the proportion of observed heterozygotes at a locus. This example illustrates how little evolutionary time has passed since the common origin of all humans in Africa. In contrast, the evolutionary time has been much longer for chimpanzees. Their groups within central Africa are more different genetically than humans living on different continents [20]. The departure from random mating predictions due to population differentiation has a classic quantitative measure, called F_{ST}, which approximately takes on values of 5–15% for common variants in global samples of human populations (see subsection 1.5.1 for more detailed description of F_{ST}).

F_{ST} is a dissimilarity measure, thus if we move from a global scale to within continental regions, F_{ST} is even lower. It regularly takes values below 1%, which is defined in [151] as a threshold for "fine-scale structure". The study of fine-scale structure is an exciting frontier of contemporary population genetics and an extensive progress has been made in the last years. Recent works commonly focus on large genome-wide datasets with samples from a territory of a single country or even a region [115, 210, 134, 101, 63, 36, 99, 26, 7] (these studies are described in section 1.4). Within such populations, average F_{ST} may be lower than 0,01 % (e.g. 0,007% between 30 collection sites in [115]). I informally define such threshold as "extremely fine scale studies".

In fact, human population structure is fractal-like. It is illustrated in Figure 1.11. Fascinating patterns arise at all scales. The coarsest structure is a tree of global human diversity. If we continue to zoom in, we see respectively continental and local scale of human population structure, and finally the fine detail of a family.

Surveying structured levels of genetic similarity gives us a scientific perspective on human origins, informs knowledge on evolutionary processes that shape both human adaptation and disease. It is thus not only important for fulfilling the curiosity about our past, but also essential for carrying the mission of medical genetics.

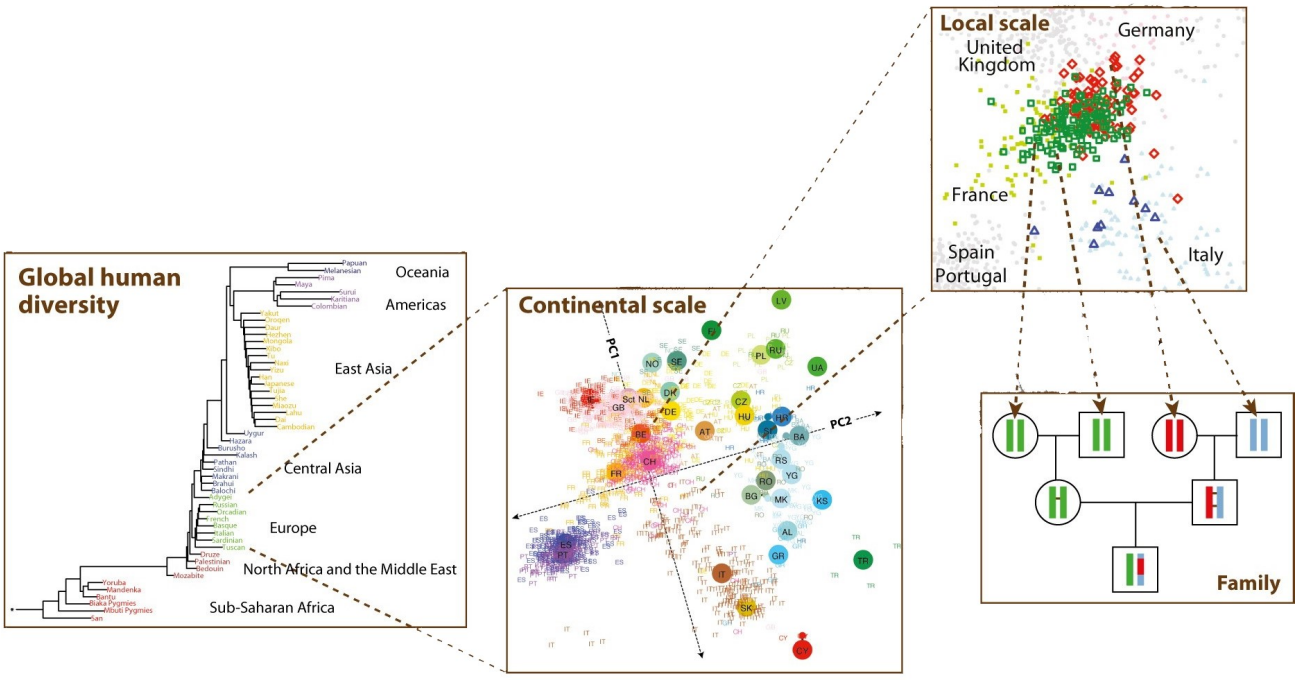


Figure 1.11 – The fractal-like nature of human population structure, adapted from [152], which reproduced images from [118] – global human diversity, [150] – both continental and local scale.

1.3.2 The importance of population structure from medical genetics perspective

Direct implications of not taking into account population structure

The presence of population structure challenges disease association studies of medical genetics. It is a problem known since many years [212, 227, 132].

The most common formulation of the population structure problem is that undetected population structure can mimic the signal of association and in this way cause either false positive results or lead to missed real effects. Figure 1.12 illustrates the concern at a *single-nucleotide polymorphism* (SNP) locus and a case-control study from two populations. The two populations are genetically different and the disease prevalence differs substantially across them too. In a such scenario, the number of cases and controls recruited for a study from each subpopulation tends to differ too, and so does allele and genotype frequency at any locus. As a consequence, significant differences in allele and genotype frequencies arise by chance, mimicking a real signal of association with the disease. The opposite situation when a real effect is masked is possible too. What is more, it is challenging to distinguish a real signal from abundant random fluctuations when both demonstrate statistical significance.

Importantly, differences of disease prevalence of a factor of two or more are not uncommon, they appear even within countries [212, 145]. They are likely linked to environmental or geographical risk factors.

In 2004, Marchini et al. [132] examined how this problem affects large-scale studies. Thanks to a good fit of theoretical model to data, they were able to simulate thousands of samples, which substantially exceeded the typical sample size in those times and anticipated the scale of future studies. They showed that with the increase of sample sizes, the confounding effects of population structure become more severe. Even small levels of population structure can lead to false positive results, undermining a disease association study. Interestingly, since the publication of this article sample sizes increased enormously, surpassing a million participants [94]. Genome-wide association studies (GWASs) have mostly targeted common variants, but as sample sizes grew larger, they have become better powered to detect rare variant associations. However, rare variants show stronger geographic clustering than common variants [138, 65, 211]. Rare disease-associated variants are therefore expected to be population specific or to reflect recent population demography.

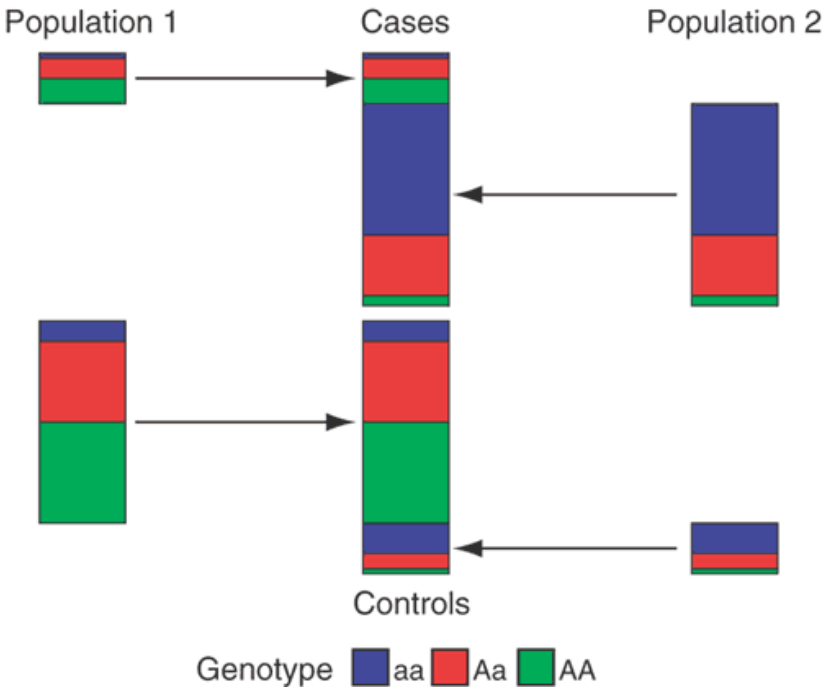


Figure 1.12 – The effects of population structure at a SNP locus, from [132]. If the study population consists of subpopulations that differ genetically, and if disease prevalence also differs across these subpopulations, then the proportions of cases and controls sampled from each subpopulation will tend to differ, as will allele or genotype frequencies between cases and controls at any locus at which the subpopulations differ. The figure shows an example of this scenario with two populations in which the cases have an excess of individuals from population 2 and population 2 has a lower frequency of allele A than population 1. In this example, the structure mimics the signal of association in that there is a significant difference in allele and genotype frequencies between cases and controls.

Indirect implications of not taking into account population structure

Nowadays the field of medical genetics faces additional open questions. One of them is transferability of genome-wide association studies (GWAS) results to a global population. To date, GWASs yielded tens of thousands significant associations between common genetic variants and phenotypes. However, the vast majority of GWASs was performed in individuals of European ancestry (more than 80% of studied individuals). Even though the majority of the results have replicated in different ethnic groups [28, 223, 82], the applicability of associations to the underrepresented populations is increasingly debated [133, 81]. Portability of GWAS findings to new populations is dependent on factors such as allele frequency, linkage disequilibrium and trait genetic architecture [133].

Serious misinterpretations may happen and one had already happened with hypertrophic cardiomyopathy [131]. Multiple African Americans have received misdiagnosis because they carried a variant identified as pathogenic in a study of mostly European cohorts. It would have been prevented if even a small number of samples of African ancestry was included, in which the variant has later turned out to be prevalent and thus unlikely disease-causing. The view on population structure in medical genomics is changing in a sense that population structure can no longer be approached as a nuisance factor, but a factor that facilitates a deeper understanding of a disease. The article [133] illustrates this shift in perception and describes the challenge. The authors focused on admixed populations of the Americas. The study of admixed populations have proven fruitful in identifying risk loci for diseases that are stratified across diverse ancestral origins [169, 161, 49, 37, 54, 15]. This study modelled subcontinental ancestry and found that heterozygosity is greatest in European ancestry tracts in two variant reference databases (the GWAS catalog and ClinVar pathogenic and likely pathogenic sites), while the results across the genome (and theoretical expectations) are that heterozygosity is the highest in African ancestry tracts (see Figure 1.13). It provides one more

evidence for strong bias toward Europeans, but this time highlighting imbalance in genome interpretability across local ancestry tracts in recently admixed populations. Moreover, it shows the advantage of analysing variants jointly with local ancestry tracts over genome-wide ancestry estimates alone. Additionally, comparison of imputation accuracy between local ancestry tract diplotypes revealed lower scores for diplotypes containing at least one Native American ancestry tract than for tracts with European and/or African ancestry only.

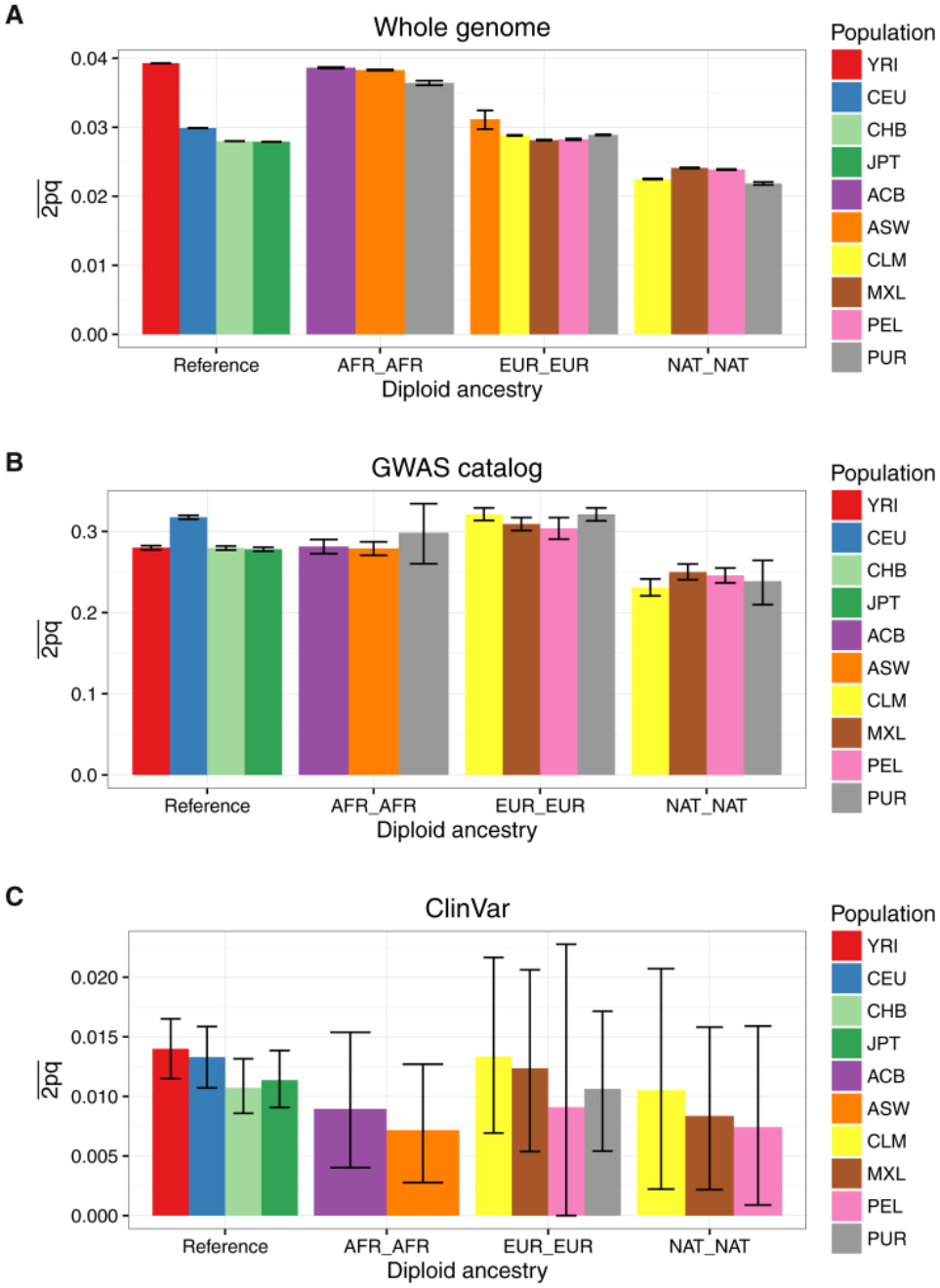


Figure 1.13 – Heterozygosity by Continental and Diploid Local Ancestry, from [133]. Heterozygosity, estimated here as 2pq, is calculated in admixed populations stratified by diploid local ancestry in (A) the whole genome, (B) sites from the GWAS catalog, and (C) sites from ClinVar classified as "pathogenic" or "likely pathogenic." The mean and 95% confidence intervals were calculated by bootstrapping 1,000 times.

Another issue with GWASs is that peaks meeting genome-wide significance often explain only a small fraction of the phenotypic variance. Scientist thus attempt to better understand genetic architecture of complex traits. Improvements of methodology include among others polygenic risk scores that take into account multiple variants. However, polygenic risk scores are sensitive to population structure too. One study has

shown that prediction accuracy of polygenic risk scores for schizophrenia in East Asians and Africans based on summary statistics derived from European cohort was reduced by more than 50% in comparison to Europeans [219]. The article [133] continues the topic of polygenic risk score transferability for several well-studied traits including height and schizophrenia and reveals biased predictions. African populations are predicted to be considerably shorter than all Europeans and minimally taller than East Asians, which is contradictory to robust anthropological evidence (excluding indigenous pygmy populations). Furthermore, schizophrenia shows considerably decreased scores in Africans compared to other populations, while it is known to have similar prevalence across all populations where it has been studied. Finally, [133] completed their study of GWAS findings transferability with coalescent simulations. The correlation between true risk scores and those inferred from summary statistics from single-ancestry (European) GWAS is generally low. It is thus imperative to include more diverse individuals in medical genomics studies and critical to understand the demographic histories of populations, including the multiethnic ones. It has been shown that the effect size estimates from diverse cohorts tend to be more precise and robust than from single-ancestry cohorts [28]. Moreover, the resolution of causal variant fine-mapping is considerably improved too [40], mainly for the reason of differential linkage disequilibrium (LD) between Europeans and Africans. If diversity will not be prioritized, genetics may contribute to health disparities [181].

Overall, understanding the genetic structure of human populations is of fundamental interest to medical and anthropological sciences. In the next section, I present selected recent studies of population structure in Europe.

1.4 Recent studies of population structure in Europe

In recent years, studies of fine-scale population structure are often conducted on genome-wide data of several hundreds or thousands of individuals. In this section I summarise findings of such studies in Europe, starting from the first continental studies using genome-wide single nucleotide polymorphism (SNP) data, continuing with recent country-level studies and finally describing what is known about population structure in France.

In the year 2008, several groups used principal component analysis (PCA)-based approaches to obtain a visual summary of the observed genetic variation [75, 150, 107, 213]. Two dimensional PCA projections from these studies resemble a geographic map of Europe (example from Population Refence Study (POPRES) [150] in Figure 1.14). These results highlight that the genetic structure of European populations is predicted by geography, with population differentiation following an isolation-by-distance pattern across the continent. Nevertheless, cultural-driven genetic structure is evident at small scales, for example Novembre et al. [150] shows structure within Switzerland, distinguishing between German-, Italian- and French-speaking populations. An exception to the pattern of genetic structure resembling the geographical map is Finland – the distance between Finnish population and the remaining European populations suggests they are an isolate [107].

Recently, several studies were performed at the scale of a single country or even a part of a country. Before, a population within a territory of typical European country was often considered homogeneous and unstructured, but thanks to novel statistical methods, large sample sizes and geo-referenced data, fine-scale structure has been revealed within the United Kingdom [115] (Figure 1.15 a), the Netherlands [210], Western France [99] (Figure 1.17), Ireland [63](Figure 1.15 b), Finland [134, 101], Spain [26] (Figure 1.15 c) and Sardinia [36]. Overall these studies have shown that genetic structure corresponds to geography, but differentiation is not always uniform with respect to physical distance: large southern and central part of England remains homogeneous [115], whereas differentiation in Spain is along east-west axis, while south-north direction presents genetic similarity [26] and in the Netherlands the structure appears along north-south gradient [210]. Spatial assessment of gene flow, effective migration surfaces were also estimated in many of these studies, giving results consistent with population structure identified with other methods and additionally low levels at country borders [63, 134] (Figure 1.16). Using patterns of ancestry sharing,

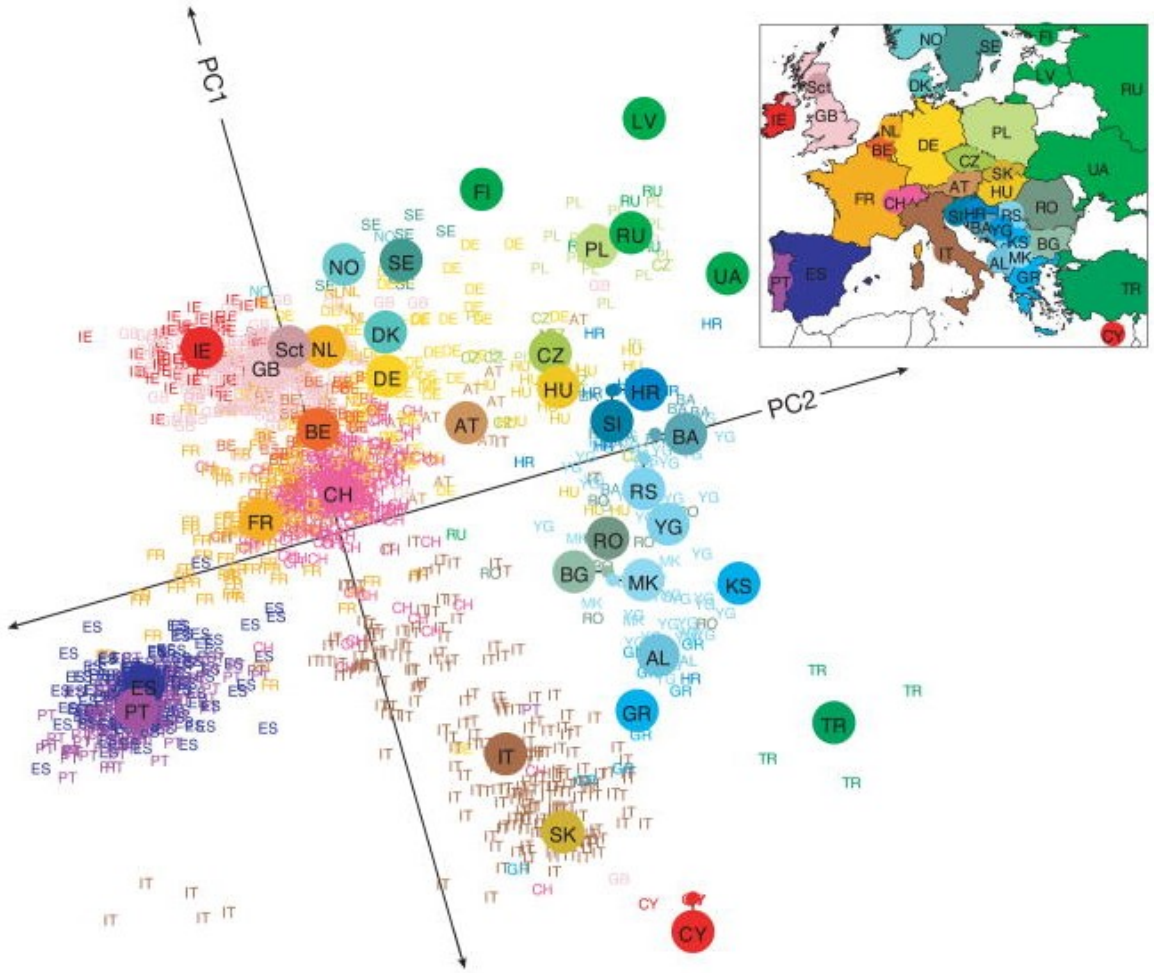


Figure 1.14 – Population structure within Europe, from [150]. A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe. AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia.

Bycroft et al. [26] shows that identified clusters are driven by genetic drift with exception to one cluster that reflects admixture from a group related to Basque population.

Many studies estimate ancestry profiles and detect differential input within a country from other European groups, for example [115, 26]. Ireland turned out to have distinct ancestry profile compared to British Isles [63]. The study of Sardinia revealed elevated shared ancestry levels with the Basque population and subtle variation of ancestry proportions with ancient DNA samples [36]. Presence of recent admixture was detected, for example admixture from Norway in Orkney dated to 1100 CE linked to Norse vikings [115] or from North African population in Spain between 860-1200 CE [26].

The country-specific studies enriched also the understanding of European demographic history. Athanasiadis et al. [7] reveals that Denmark have a disparate demographic history compared to other Scandinavian countries. The Genome of the Netherlands study [210] claims south to north gradient of decreasing ancestral population size and increasing *homozygosity*. Excess *haplotype* sharing indicates more severe bottlenecks in

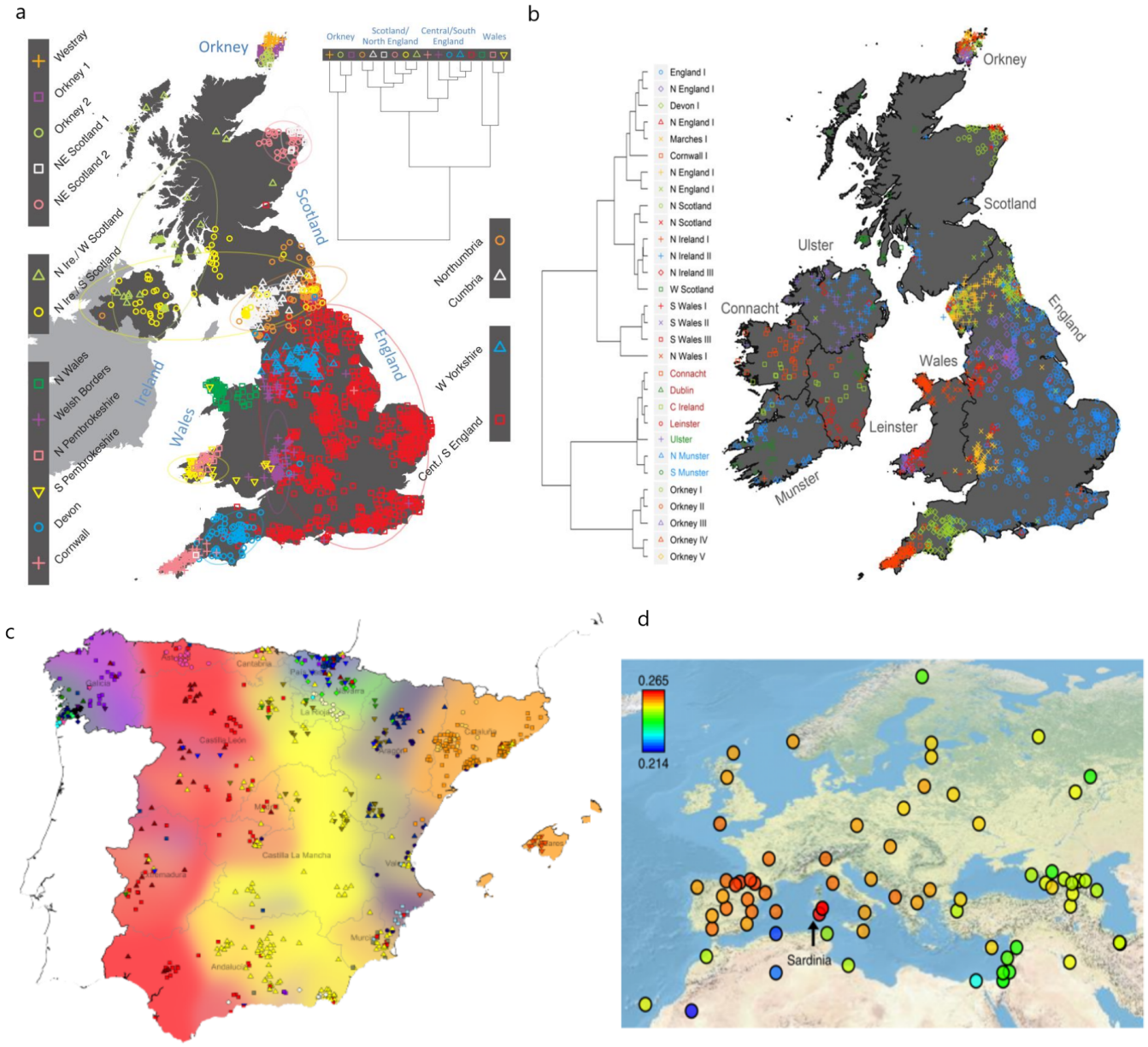


Figure 1.15 – Results from recent European studies of population structure. (a) Clustering of the 2,039 UK individuals into 17 clusters based only on genetic data, from [115]. (b) 30 clusters of individuals with Irish and British ancestry based solely on genetics, from a study that focused on Ireland [63]. (c) Spanish individuals grouped into clusters using genetic data only, from [26]. (d) Similarity of ancient samples to populations across Europe, including Sardinia, indicated by outgroup f3 statistics. From [36].

Eastern Finland than in South-Western Finland [134]. Chiang et al. [36] found a relatively deep divergence from mainland European populations which is evidence for Sardinia's isolation.

France was a part of European datasets published in 2008 [75, 150, 118, 107], but precise birthplaces of individuals were unknown, which prevented the description of regional differentiation. These studies placed the French population between its neighbours on the genetic map of Europe and revealed a gradient of variation (see Figure 1.14). Karakachoff et al. [99] was the first study of genetics of French population with geo-referenced data. It reveals that fine-scale population structure occurs in Western France (Figure 1.17). Brittany Region, situated on a peninsula in the North-West of France, is relatively highly differentiated from the neighbouring regions. Also linkage-disequilibrium is higher in Brittany, suggesting a lower effective population size. The departments with the highest local genetic differentiation measured in F_{ST} per 30 km are the departments of Brittany region and additionally the department Vendée. Last but not least, the

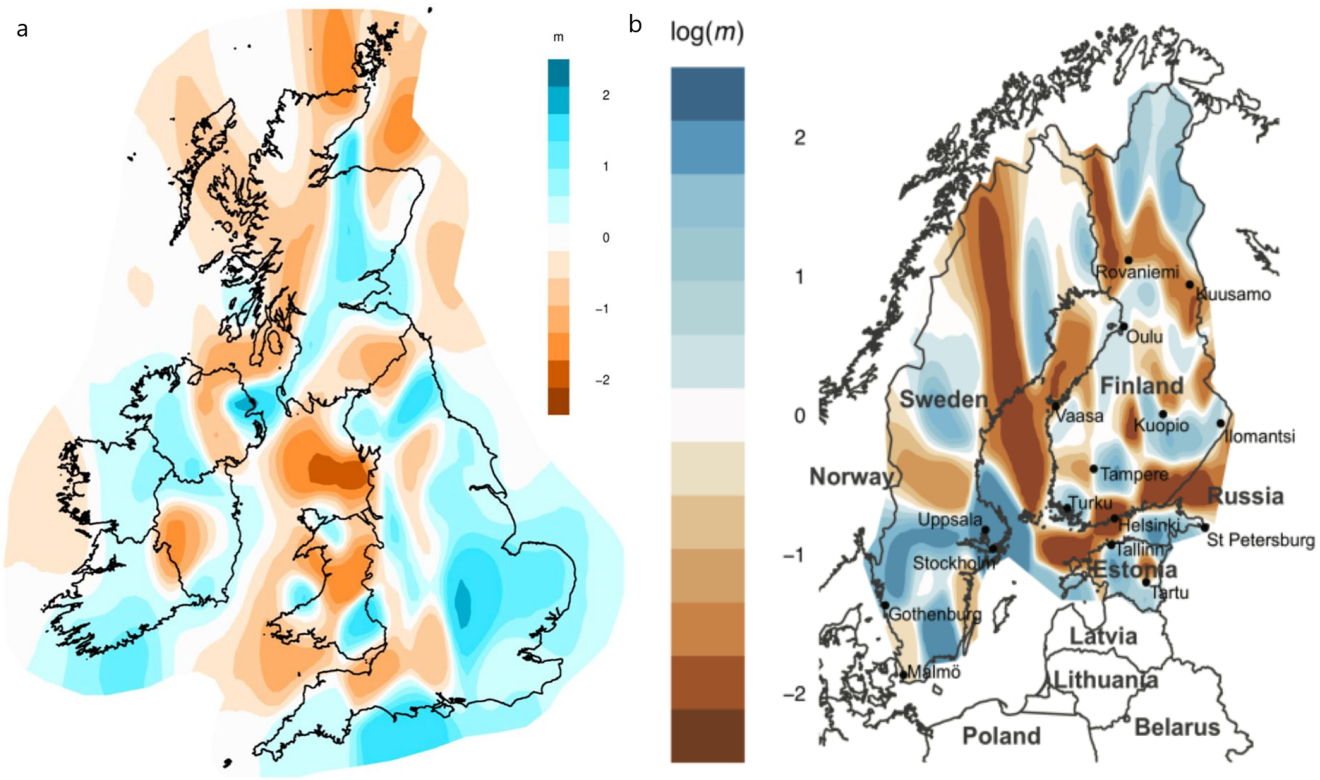


Figure 1.16 – The estimated effective migration surfaces from recent European studies of population structure. (a) of Ireland and Britain, from [63]. (b) of Finland, Sweden, Estonia and St. Petersburg, Russia; from [134].

authors also found genetic proximity between Bretons and Irish, both at the genome-wide level and at the two loci informative about Breton origins, lactase and HLA.

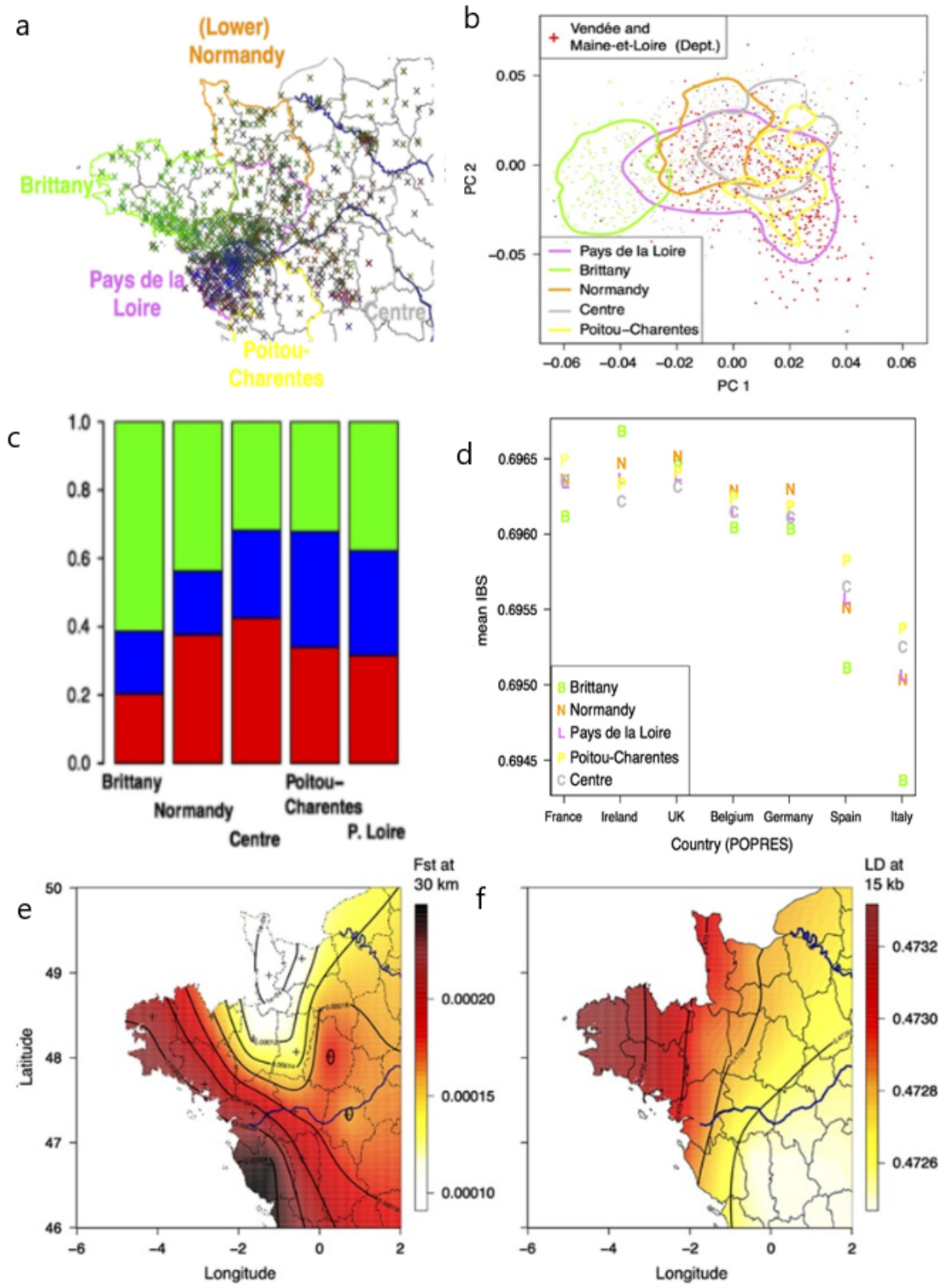


Figure 1.17 – Population structure in Western France, adapted from [99]. a) Sampling of the study. b) PCA indicating differentiation of Brittany from neighboring regions. c) Mean ancestry coefficients grouped by regions of Western France d) Mean IBS statistics between individuals from the French regions and individuals from the neighboring countries of France. e) Spatial variation of local genetic differentiation (Fst at 30 km) and f) of LD (at 15 kb).

1.5 How can we detect population structure?

The results presented in the previous section as well as the results of this thesis would not be obtained without powerful statistical methods. The majority of the methods I describe in this section can be divided into two categories: allele-frequency-based and haplotype-based methods. The difference between these two lies in the way the information contained in dense SNP datasets is used. The first approach relies on differences in allele frequencies. It does not use all the information contained in dense SNP dataset because they do not take *linkage disequilibrium* (LD) into account. It is necessary to prune the data, removing linked markers. In contrast, haplotype-based approaches model recombination and in this way take account of LD. These approaches are viewed as more comprehensive as it provides additional information about relatedness beyond allele frequency patterns. The possible drawback of haplotype-based methods is increased computational cost required to run analysis compared to allele-frequency approaches.

This section is not an exhaustive review of all existing methods. I familiarise a reader with methods that I used in my work: F_{ST}, PCA, ADMIXTURE, EEMS, CHROMOPAINTER, fineSTRUCTURE (FS) and IBDNe. Table 1.1, adapted from a review [194], contains a longer list of approaches (softwares) for population structure and demographic inference. Certain methods are not applicable in my project, as they require sequencing data.

1.5.1 Allele-frequency based approaches

For decades, the most common approach to describe population structure and quantify genetic distance between populations was Wright's F_{ST} [128, 225], a member of the family of measurements known as fixation indices or F-statistics, originally defined as ratio of genetic variation between subpopulations to genetic variation of the total population. F_{ST} became widespread thanks to interpretation that it is equivalent to the proportion of genetic diversity that can be explained by the differences in allele frequencies among populations [46, 88].

Another allele frequency-based approach, also in use since only a handful of markers was available, is principal component analysis (PCA). Briefly, PCA is a multivariate technique that reduces high-dimensional dataset of correlated variables into a smaller set of uncorrelated variables that explain most of the variation in the data [93]. These uncorrelated variables can be plotted. First few PCs ideally separate samples according to their recent ancestry: with more similar genotypes on average will cluster closer in PC space, whereas more distantly related individuals will lie further apart. This method thus serves to explore and summarize visually population structure (see Figure 1.18 A for schematics for the output). PCA is often performed as initial exploratory data analysis, because it can also help to identify potential unwanted sources of variability [194]. A large number of markers in the dataset is crucial to reveal subtle fine-scale structure [151] (Figure 1.19), which nowadays does not pose a problem anymore with the availability of relatively low cost genome-wide SNP arrays. One should not limit the exploration of PCA space to the first two axes, because population structure patterns are sometimes more readily observed in lower-ranked PCs than in higher-ranked PCs [16]. Care must be taken to not over-interpret PCA results: the disadvantage of the method is that the results are affected by sampling density [142, 153]. Moreover, in populations with recent, large-scale shifts in demography (such as humans), the directions of highest variability may be counterintuitive [194]. Although it has been proposed that PCs can be used to understand demographic processes underlying the data [142], it has also been shown that mathematical artifacts can arise from spatial data, thus the historical inferences made from these results need to be taken with caution [153].

The next approach developed for allele frequency data is model-based clustering, with the most prominent example being STRUCTURE [171]. These approaches assign individuals into populations and provide an estimate of admixture proportions: the fraction of each individual's genome that comes from each population (see Figure 1.18 B for schematics for the output). In principle, they look for groups of individuals that share common underlying allele frequencies and that are mutually in *Hardy–Weinberg equilibrium* (HWE). STRUCTURE is Bayesian and uses Markov chain Monte Carlo (MCMC) to average over underlying allele

Name	Inference	References
STRUCTURE	Population structure, admixture	[171]
FRAPPE	Population structure, admixture	[207]
ADMIXTURE	Population structure, admixture	[3]
fastSTRUCTURE	Population structure, admixture	[174]
Structurama	Population structure, admixture	[91]
HAPMIX	Chromosome painting	[169]
fineSTRUCTURE	Population structure, admixture, chromosome painting	[108]
GLOBETROTTER	Population structure, admixture, chromosome painting	[77]
LAMP	Chromosome painting	[186]
PCAdmix	Chromosome painting, population structure	[21]
dadi	Demographic history	[70]
Fastsimcoal2	Demographic history	[47, 48]
Treemix	Admixture graph	[167]
fastNeutrino	Demographic history	[14]
DoRIS	Demographic history	[159, 160]
IBS tract inference	Demographic	[72]
PSMC	Demographic history	[117]
MSMC	Demographic history	[190]
CoalHMM	Demographic history	[83, 45, 125, 84, 126]
diCal	Demographic history	[197, 202]
LAMARC	Demographic history	[105]
BEAST	Species trees, effective population sizes	[41]
MCMCcoal	Divergence times between populations	[176]
G-PhoCS	Demographic history	[68]
Exact likelihoods using generating functions	Demographic history	[124, 123]

Table 1.1 – Software for demographic inferences, table adapted from [194].

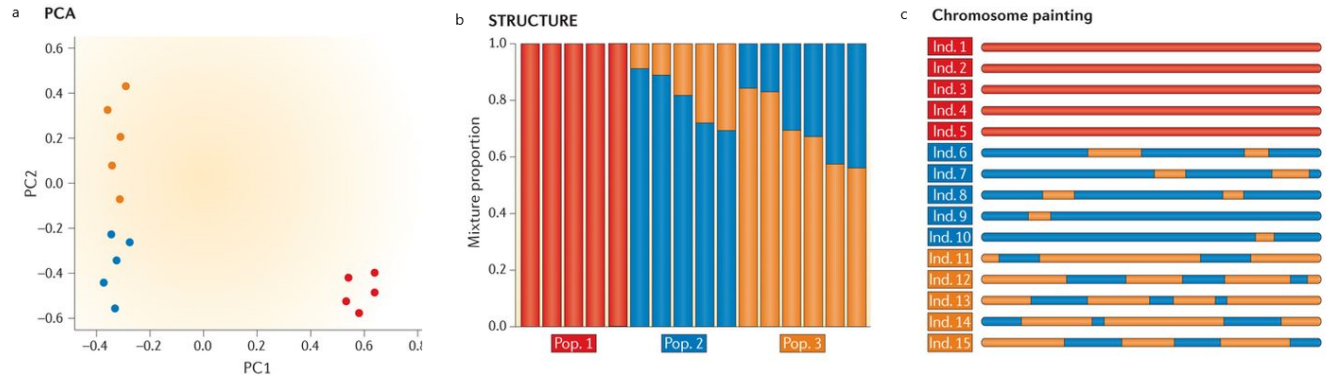


Figure 1.18 – Schematics for the output of three approaches discussed in the text are illustrated: (A) PCA, (B) STRUCTURE-like approaches, (C) CHROMOPAINTER. Adapted from [194].

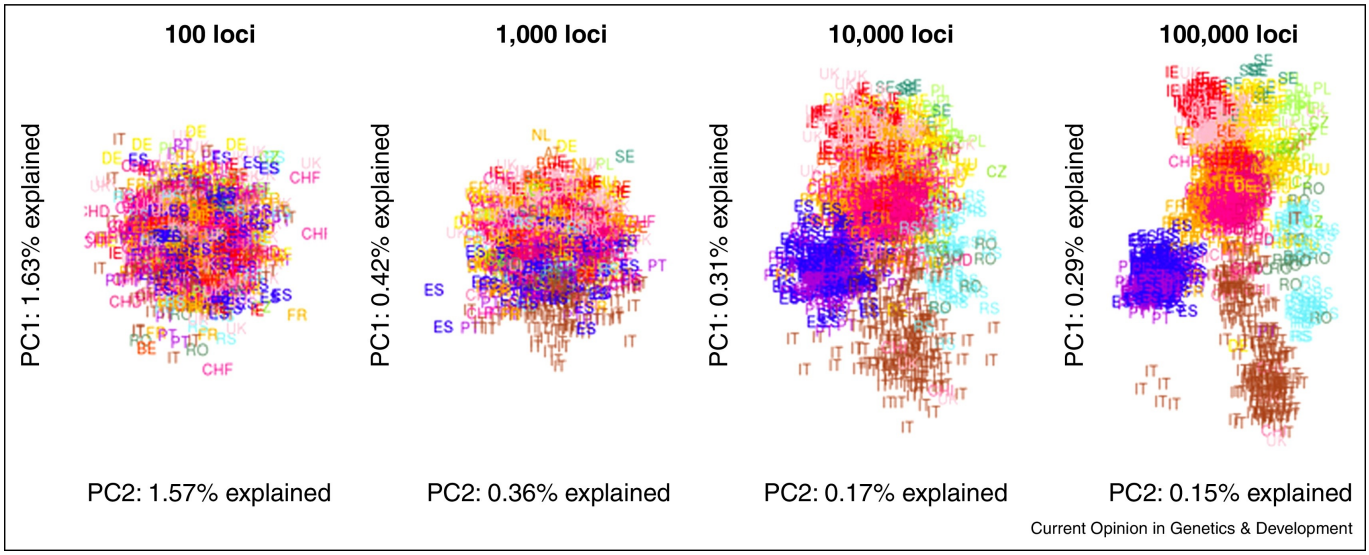


Figure 1.19 – A large number of loci is required to reveal fine-scale population structure using PCA, from [151]. Four subsamples with an increasing number of random loci were taken from the [150] dataset. Using 100 loci, Europe appears panmictic, whereas 1.000 loci are sufficient to establish a North-South cline. With 10 000 and 100 000 loci, fine-scale details are revealed.

frequencies and assignments, and thus is too slow for modern genomic data sets [194]. Faster implementation based on likelihoods, e.g. ADMIXTURE [3], is recommended for large datasets. In STRUCTURE-like methods the likelihood of the data will always improve by adding more parameters (that is, more populations) and at certain point those extra populations are overfitting the noise in the data. Therefore, the choice of the number of populations (K) that best fits allele frequency differences in the data is problematic. The most popular solution to this problem is a cross-validation procedure that measures the consistency between different runs at a particular K [2].

EEMS (Estimated Effective Migration Surfaces) [165] is a method for analysing population structure from geo-referenced individuals and identifying potential barriers to gene flow. EEMS uses pairwise genetic dissimilarities and provides a visual representation of spatial patterns in genetic variation (see example output from recent European studies in Figure 1.16). It highlights regions where deviations of exact isolation by distance occur, regions of higher-than average and lower-than average historic gene flow, which were difficult to discern using previous methods such as PCA and STRUCTURE-like approaches that ignore sampling locations even if they are known. The term "effective" comes from the fact that EEMS assumes equilibrium in time. The resulting migration surfaces should be best interpreted as a tool for visualising patterns of genetic differentiation relative to geographic distance [151].

1.5.2 Haplotype-based approaches

Many approaches relating genetic variation to LD are based on the coalescent theory [102] and its generalization that includes recombination [90]. Although this framework has been useful for simulations, its computational complexity is still too high for statistical inference. An efficient alternative approximation was developed in 2003 by Li and Stephens [119]. It captures the essential properties of the coalescent process and it overcomes the limitations of previous approaches as it relates patterns of LD directly to the underlying recombination process, considers all loci simultaneously, rather than pairwise and avoids the assumption that LD necessarily has a "block-like" structure. In Li and Stephens model, every haplotype of an individual is represented as a mosaic of the haplotypes that are copied from the reference panel. A possible challenge in this approach is accurate modelling of recombination across the genome and the ascertainment bias introduced by selecting SNPs based on physical distance [152].

The key information utilised by haplopype-based approaches is related to the fact that recombination breaks up chromosomes progressively in each transmission of genetic material from parent to offspring. The length of the haplotype segment shared between two individuals is thus correlated with the time to their recent ancestor. Longer length of shared haplotypes typically reflects more recent common ancestry (Figure 1.18 C). Based on this approach, CHROMOPAINTER computes a coancestry matrix, i.e. the estimates of the proportion of the genome of each individual that is most closely related to every other individual in the matrix [108]. Individuals are in turn donors and recipients of haplotype chunks and a set of donors does not necessarily have to be the same set as the set of recipients. Summing up, coancestry matrices contain rich information about population relationships.

Coancestry matrices generated by CHROMOPAINTER can serve as input to other methods. One of them is fineSTRUCTURE [108], an MCMC clustering approach that aims to partition individuals into groups with indistinguishable haplotype similarity profiles (see example output from recent European studies in Figure 1.15 a,b,c). After fine subdivision is obtained, fineSTRUCTURE builds a tree (FS-tree) that aims to capture historical relationships amongst the inferred populations. Thanks to this tree clustering becomes hierarchical: at each level two populations with the highest posterior probability among all merges are joined. One area of concern is that FS-tree results depend significantly on sample size [108]. To overcome this limitation, Kerminen et al. [101] proposed an alternative tree building approach based on *total variation distance* (TVD) that produces more consistent results across different sample sizes.

Total variation distance (TVD) was defined in Leslie et al [115]. It is calculated based on coancestry matrix. The coancestry matrix contains copying profiles for each individual i: $C_i = (c_{i,1}, c_{i,2}, \ldots, c_{i,n})$, $i \in 1, \ldots, n$ (for simplicity, we assume a set of donors identical to a set of recipients). Those values are then averaged across individuals in a given cluster $k \in 1, \ldots, K$ (K is a number of clusters) to get cluster copying vector $X_k = (x_1, x_2, \ldots, x_n)$ as follows:

$$x_j = \frac{\sum_{l \in k} c_{j,l}}{\sum_{m=1}^{n} c_{j,m}}, j \in 1, \ldots, n.$$

Given a pair of inferred clusters A and B, with copying vectors $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$, the total variational distance between the pair is computed as following:

$$TVD(A, B) = 0.5 * \sum_{i=1}^{n} |a_i - b_i|.$$

This value can be interpreted as a measure of the difference between the two clusters. TVD matrix is a square symmetric matrix containing TVD values for each pair of clusters. Tree built based on TVD matrix thus uses a measure of genetic differentiation, contrary to FS-tree based on posterior probabilities.

The IBDNe [23] approach makes inference about demography. It is a non-parametric method for estimating recent effective population size. Estimates of historical effective population size reveal such demographic features as bottleneck events and rates of growth. Similarly to CHROMOPAINTER, the theoretical

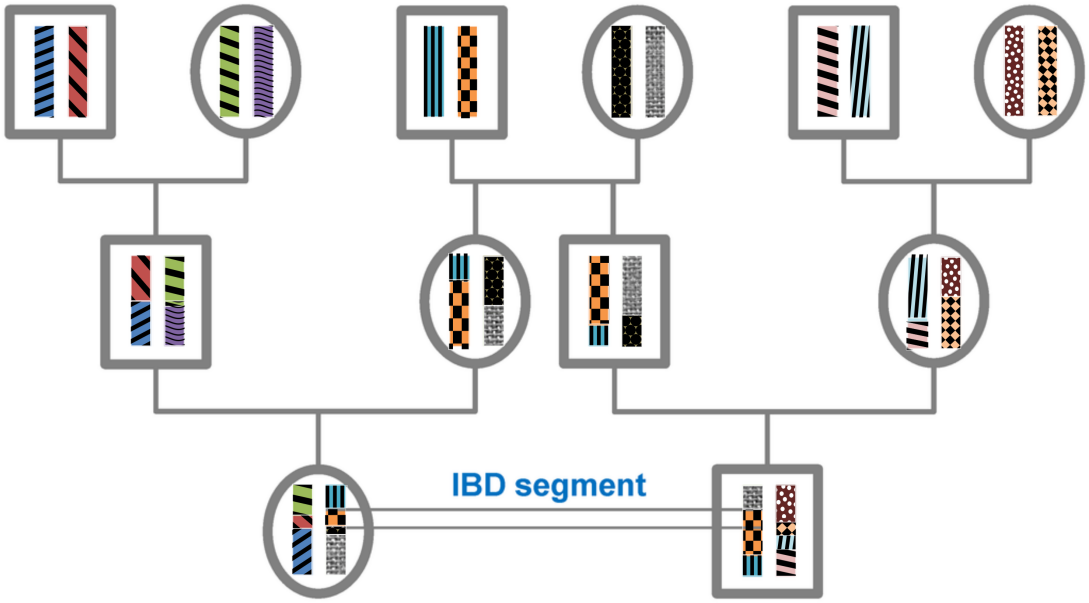


Figure 1.20 – The origin of IBD segments is depicted via a pedigree. Source: Wikipedia.

foundations of IBDNe rely on the correlation between shared haplotype length and the time to the recent common ancestor. However, it uses inferred identity-by-descent (IBD) segments length rather than the Li and Stephens model. A segment is identical-by-descent in two or more individuals if its nucleotide sequence shows high levels of similarity and it was inherited from a common ancestor without recombination (Figure 1.20). In a finite population all individuals are related if traced back long enough in time. Therefore, in their genomes there are segments that are IBD. During meiosis, segments of IBD are broken up by recombination. Therefore, the expected length of an IBD segment is negatively correlated with the number of generations since the most common recent ancestor.

IBD-based estimates for effective population size changes are relatively accurate for recent time intervals (from around 4 generations to around 50 generations ago) for SNP array data [23]. An advantage of IBDNe is a non-parametric approach that addresses the limitations of parametric methods to model complex or unanticipated features of population-size history that result from the necessity to pre-specify the class of models that are considered as well as computational and statistical constraints linked to the number of parameters considered.

1.6 Thesis aims and structure

My thesis aim is to provide a thorough analysis of the genetic structure of the continental French population, in particular Western France, and shed light on the historical, demographic and cultural events that have shaped it. Along my thesis I address the following questions: 1) is there population structure within France? and if so, what are the levels of population structure? 2) Does genetic differentiation coincide with past political or linguistic borders? 3) how population differentiation within France compares to that described in other European studies? 4) Can we find any evidence of migrations that shaped present population genetic structure? 5) Do the patterns of diversity reflect differences in effective population size? To achieve these aims, I had to identify, implement and apply a comprehensive set of standard statistical methods and techniques.

To answer the aforementioned questions, I first employed a genotype marker dataset SU.VI.MAX. This dataset often serves as control datasets in disease association studies. The geographical range of SU.VI.MAX dataset is the whole territory of continental France, thus this dataset enabled me to describe the genetic structure and differentiation patterns on the level of a country. I corroborated my results by analogous analysis on second, independent dataset, 3C, performed by collaborators. This dataset is also

a genotype marker dataset, it has a similar geographical scope and is also often used as control dataset in disease association studies. Later I focused on a particular subpopulation, North-Western France, to study it in more detail. It was feasible thanks to the PREGO dataset, which was designed to serve as best as possible the purpose of extremely-fine scale study of genetic structure.

This manuscript is organised as follows:

— Chapter 2 contains an article summarising the results of the study of a fine-scale structure of the whole continental France with the use of SU.VI.MAX and 3C datasets, the latter one as a collaboration.

— In Chapter 3 I explore an extremely fine population structure and demographic history of the part of French territory – Northwestern France with the use of PREGO dataset.

— In the last chapter, Chapter 4, I discuss the relation of the results of the two studies to history, demography, liguistics and culture, as well as analysis of rare variants as a perspective of the field and consequences of presence of a population structure for medical studies.

2

The genetic history of France

The Genetic History of France

Aude Saint Pierre[1*], Joanna Giemza[2*], Matilde Karakachoff[2], Philippe Amouyel[3+], Jean-François Dartigues[4+], Christophe Tzourio[4+], Martial Monteil[5], Pilar Galan[6], Serge Hercberg[6], Richard Redon[2], Emmanuelle Génin[1*], Christian Dina[2*]

[1]*Univ Brest, Inserm, EFS, CHU Brest, UMR 1078, GGB, F-29200 Brest, France*

[2]*l'institut du thorax, INSERM, CNRS, Univ Nantes, CHU Nantes, Nantes, France*

[3]*Univ. Lille, Inserm, CHU Lille University Hospital, Institut Pasteur de Lille, LabEx DISTALZ-UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related, F-59000 Lille, France*

[4]*Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, CHU Bordeaux, F-33000 Bordeaux, France*

[5]*Université de Nantes, UMR 6566 CReAAH, LARA, Nantes, France*

[6]*Université Paris 13, Equipe de Recherche en Epidémiologie Nutritionnelle, Centre de Recherche en Epidémiologie et Statistiques, Inserm (U1153), Inra (U1125), Cnam, COMUE Sorbonne Paris Cité, F-93017, Bobigny, France*

* These authors contributed equally;

[+] On behalf of the 3C study

Correspondence to:

Christian Dina, christian.dina@univ-nantes.fr

Aude Saint Pierre, aude.saintpierre@univ-brest.fr

Introduction

Gallia est omnis divisa in partes tres [*commentarii de bello gallico*[1]] was one of the earliest demographic description of antique France (known as Gaul). These three parts were Aquitania, in South West, with Garonne and the Pyrenees mountains as borders; Belgia in North West, following the Seine as Southern border; and finally what we know as Celtic Gaul, that spanned from the Atlantic Ocean to the Rhine River and Alps. A fourth part of the present-day French territory, already part of Romanized territories at this time, was Gallia Transalpina, a strip of lands from Italy to Iberia, with Alps and Cevennes mountains as northern border.

The area that was to be modern France was subject to successive population migrations: Western Hunter-Gatherers (15 kya), Neolithic farmers (7 kya) and later steppe Enolithic Age populations[2,3], Celtic expansion, integration in Roman empire, Barbarian Great migrations, whose demographical importance remains to be assessed. France's position in Europe, at the edge of the Eurasian peninsula, has made it not only the final goal of a large number of, potentially massive, migrations but also a place of transit either to the North (British Isles) or the South of Europe (Iberian Peninsula) and North Africa, as well as an important crossroad for trade and exchanges.

Before France became a single political entity, its territory was divided into various kingdoms and later provinces, which often displayed fierce independence spirit towards the central power. The pre-Roman Gaul was divided into politically independent territories. After the fall of Roman Empire, the modern French territory was divided into Barbarian Germanic kingdoms (Franks, Wisigoths and Burgunds). After a short period of reunification and extension into the Carolingian Empire (VIIth century), the weakening of the central power led to the reduction of the Occidental France at its western part and the rise of local warlords gaining high independence within the Kingdom itself. The feudality period created provinces that were close to independence, although nominally linked through the oath of allegiance to the King of France (Figure S1).

During centuries, in spite of important backlashes such as the Hundred Years War, the French Kings managed to slowly integrate the Eastern lands as well as Britanny, enforcing in parallel the central power until the French Revolution. However, every province kept displaying political, cultural and linguistic differences, which could have left imprints in the genetic structure of modern French populations.

Geographically, modern France is a continental country surrounded by natural borders: the Atlantic Ocean on the West side, the Channel Sea up North, mountains (Pyrenees and Alps) closing the South-West and East/South-East borders, as well as the Mediterranean Sea on the South side. The Eastern side has the Rhine as a natural border on less than 500 kilometers while the Northeastern borders shows no notable obstacle and exhibits a continuum with Germany and Belgium.

The study of the genetic structure of human populations is of major interest in many different fields. It informs on the demographic history of populations and how they have formed and expanded in the past with some consequences on the distribution of traits. Genetic differences

between populations can give insights on genetic variants likely to play a major role on different phenotypes, including disease phenotypes[4]. This explains the growing interest of geneticist for human population studies that aim at describing the genetic diversity and are now facilitated by the rich genetic information available over the entire genome. In the last decades, several studies were performed using genome-wide SNP data often collected for genome-wide association studies. These studies have first shown that there exist allele frequency differences at all geographic scales and that these differences increase with geographic distances. Indeed, the first studies have shown differences between individuals of different continental origins[5,6,7] and then, as more data were collected and marker density increased, these differences were found within continents and especially within Europe[8,9]. Several studies have also been performed at the scale of a single country and have shown that differences also exist within country. This was for instance observed in Sweden, where Humphreys et al.[10] reported strong differences between the far northern and the remaining counties, partly explained by remote Finnish or Norwegian ancestry. Fine-scale stratification was also detected in Western France (mainly *Pays de le Loi*re and Brittany) where a strong correlation between genetics and geography could be observed[11]. More recent studies have shown structure in the Netherlands[12], Ireland[13], UK[14] or Iberian peninsula[15].

In this paper, we used data from two independent cohorts, 3C and SUVIMAX with more than 2,000 individuals whose birthplace covered continental France and genotyped at the genome-wide level, to assess the genetic structure of the French population and draw inferences on the demographic history.

Material and Methods

Data SU.VI.MAX & 3C

Genetic data were obtained from two French studies, SU.VI.MAX[16] and the Three-Cities study[17] (3C). For every individuals, information on places of birth was available, either the exact location (3C study) or the "*département*" (SU.VI.MAX). *Départements* are the smallest administrative subdivisions of France. There are a total of 101 French *départements* and 94 of them are located in continental France. These units were created in year 1789, during French Revolution, partly based on historical counties.

3C Study: The Three-City Study was designed to study the relationship between vascular diseases and dementia in 9,294 persons aged 65 years and over. For more details on the study, see http://www.three-city-study.com/the-three-city-study.php. Analyses were performed on individuals who were free of dementia or cognitive impairment by the time their blood sample was taken and who were previously genotyped[18]. The geographical locations of individuals were defined according to the latitude and longitude of their place of birth, declared at enrolment. Individuals with missing place of birth or born outside continental France were excluded. A total of 4,659 individuals were included in the present study.

SU.VI.MAX: The study was initiated in 1994 with the aim of collecting information on food consumption and health status of French people. A subset of 2,276 individuals born in any of the 94 continental French *départements* was included in this study. The geographic coordinates

of each *départements* were approximated based on the coordinates of the corresponding main city.

Quality control

Quality control of the genotypes was performed using the software PLINK version 1.9[19,20].

3C: raw genotype data were generated in the context of a previous study[18] on Illumina Human610-Quad BeadChip. Following the recommendations from Anderson et al.[21], individuals were removed if they had a call rate < 99%, heterozygosity level ± 3 standard deviations (SD) from the mean or if they were related to another individual from the sample with an IBD proportion of 0.1875 or above (only one individual was kept from each pair). As a final quality control to exclude outlier individuals from populations, we performed principle component analysis (PCA) using the smartpca software from the EIGENSOFT package version 6.0.1[22] and removed outliers across the first 10 eigenvectors. The default procedure was used for outlier removals with up to 5 iterative PCA runs and, at each run, removing of individuals with any of the top 10 PCs departing from the mean by more than 6 standard deviations. SNPs in strong linkage disequilibrium (LD) were pruned out with PLINK 1.9 (described in PCA section). Outlier individuals were removed prior to performing further analyses. Applying all these QC filters led to the removal of 226 individuals. To avoid redundant information from individuals born at the same place, we randomly selected only one individual from each place of birth. A total of 770 individuals covering the 94 continental French "*départements*" were included. All samples failing sample-level QC were removed prior to performing SNPs QC. Markers were removed if they had a genotype-missing rate > 1%, a minor allele frequency < 1% or departed from Hardy–Weinberg proportion ($P \leq 10^{-7}$). After QC, there were 770 individuals and 490,217 autosomal SNPs.

SU.VI.MAX: Genotype data of the 2,834 samples were available from previous studies using different SNP chips: 1,978 with Illumina 300k/300kduo and 856 with Illumina 660W. Individuals with an unknown birthplace or a birthplace outside of continental France were removed, 1,416 samples were left. Two individuals were removed because of a call rate < 95%. IBD statistic, calculated in PLINK version 1.9, didn't identify any related samples with a threshold of 0.1875. SNPs were removed if they had a genotype-missing rate > 2%, a minor allele frequency < 10 % or departed from Hardy–Weinberg proportion ($P \leq 10^{-5}$). After QC, there were 1,414 individuals and 271, 886 autosomal SNPs.

Population structure within France

1) Chromopainter/FineSTRUCTURE analysis

For investigating fine-scale population structure, we used Chromopainter version 2 and FineSTRUCTURE version 2.0.7[23]. Data were phased with SHAPEIT v2.r790[24] using the 1000 genomes dataset as a reference panel. In the 3C dataset, we removed 932 of these SNPs because of strand issues prior to phasing. Files were then converted to Chromopainter format using the 'impute2chromopainter2.pl' script. Chromopainter outputs from the different chromosomes

were combined with chromocombine to generate a final coancestry matrix of chunck counts for FineSTRUCTURE. For the FineSTRUCTURE run we sampled values after successive series of 10,000 iterations for 1 million MCMC iterations following 10 million "burn-in" iterations. Starting from the MCMC sample with the highest posterior probability among all samples, FineSTRUCTURE performed 100,000 additional hill-climbing moves to reach its final inferred state (See[14] for details). The final tree was visualized in R with the help of FineSTRUCTURE and 'dendextend' libraries.

2) Ancestry profiles of the French population & Spatial pattern of genetic structure EEMS

We used ADMIXTURE v1.3[25] to estimate mixing coefficients of each individual. We performed runs for values of K between 2 and 10, with 5-fold cross-validation using the set of pruned SNPs, as described in the PCA analyses. To identify if cluster differences existed, we performed a one-way analysis of variance (ANOVA) on the admixture components, followed by *post hoc* pairwise comparisons.

We estimated an effective migration surface using the software EEMS. We run EEMS with slightly different grids to investigate how/whether these changes affected the results. Plots were generated in R using the "rEEMSplots" package according to instructions in the manual. For both datasets the full set of SNPs was included. For more information on the specific pipeline, see Supplementary Data.

3) IBD-estimated population size

We estimated the recent effective population size with IBDNe[26]. IBDNe was run with the default parameters and a minimum IBD segment length of 4 cM (mincm=4). We used the default settings to filter IBD segments from IBDseq v. r1206 software package[27]. Breaks and short gaps in IBD segments were removed with the merge-ibd-segments utility program. For IBD detection, we varied the minimum IBD segment length in centiMorgan units by the mincm parameter (mincm argument) from the default value, 2 cM, to 8 cM. IBDNe analysis was applied on the whole SU.VI.MAX and 3C datasets as well as on the major subpopulations from fineSTRUCTURE clustering. Growth rates were calculated with the formula $\frac{end\ value - start\ value}{start\ value}$. We assumed a generation time of 30 years, as assumed in the original paper.

4) Principal Component Analysis (PCA) and F_{ST}

Both PCA and F_{ST} analyses were carried out on a pruned set of SNPs in each dataset independently and using the smartpca tool in the EIGENSOFT program (v6.1.1) [22]. The pairwise F_{ST} matrix were estimated using the option 'fsthiprecision=YES' in smartpca. We calculated the mean F_{ST} between clusters inferred by FineSTRUCTURE as group labels. In each dataset, SNPs in strong LD were pruned out with PLINK in a two-step procedure. SNPs located in known regions of long range linkage disequilibrium (LD) in European populations were excluded from the analysis[31]. Then, SNPs in strong LD were pruned out using the 'indep-pairwise' command in PLINK. The command was run with a linkage disequilibrium r^2=0.2, a

window size of 50 SNPs and 5 SNPs to shift the window at each step. This led to a subset of 100,973 SNPs and 83,246 SNPs in the 3C and SU.VI.MAX datasets respectively. To evaluate the geographic relevance of PCs, we tested for the significance of association between the latitude and longitude of each *département* and PCs coordinates ('cor.test' function in R) using a Spearman's rank correlation coefficient.

Relation with neighbor European populations: 1000G and HGDP

We assembled SNP data matching either the SU.VI.MAX or the 3C genotype data (after quality control) with the European individuals from the 1000G phase 3 reference panel and from the Human Genome Diversity Panel data (HGDP), to generate four genome-wide SNP datasets analyzed independently.

The 1000G reference panel served as donor populations when estimating ancestry proportions. First, in order to define a set of donor groups from 1000G Europe (EUR), we used the subset of unrelated and outbred individuals generated in the study of Gazal et al.[32]. Four European populations were considered: British heritage (GBR, n=85 and CEU, n=94), Spain (IBS, n=107) and Italy (TSI, n=104). These 390 Europeans individuals were then combined with individuals from both datasets independently resulting in a set of 484,874 common SNPs with 3C and a set of 232,148 common SNPs with SU.VI.MAX. The filtered datasets (after pruning) included 1,160 individuals genotyped on 100,851 SNPs in the 3C Study and 1,804 individuals genotyped on 64,653 SNPs in SU.VI.MAX. We inferred European ancestry contributions in France using the new haplotype-based estimation of ancestry implemented in SOURCEFIND[33]. To define homogenized donor groups, we ran FineSTRUCTURE on the four European populations defined above and selected the level of clustering describing the main features of the donor populations. These European donor groups served as reference in SOURCEFIND. We performed analysis of variance (ANOVA) on French admixture component per cluster group to identify whether cluster differences existed.

Additional analyses combining the European participants of the HGDP panel were carried out in order to estimate the contribution of Basque population of our South West clusters. A total of 160 European HGDP participants were included from 8 populations: Adygei (n=17), French-Basque (n=24), French (n=29), Italian (n=13), Italian from Tuscany (n=8), Sardinian (n=28), Orcadian (n= 16) and Russian (n=25). Using the same procedure for merging panels, the filtered datasets (after pruning) included 930 individuals and 93,938 SNPs in the 3C Study and 1,574 individuals and 57,775 SNPs in SU.VI.MAX.

Results

Chromopainter/FineSTRUCTURE analysis

We performed a FineSTRUCTURE analysis on each of the two datasets separately. Results reveal fine-scale population patterns within France at a very fine level. Figure 1 shows the map for a limited number of clusters while Figure S2 depicts maps at the finest level of genetic differentiation. FineSTRUCTURE identified respectively 17 and 27 clusters in 3C and

SU.VI.M.AX, demonstrating local population structure (Figure S2). Both analyses show very concordant partitions with a broad correlation between clusters and geographic coordinates. The major axis of genetic differentiation runs from the south to the north of France.

In both datasets, the coarsest level of genetic differentiation (i.e. the assignment into two clusters) separates the south-western regions from the rest of France (Figure S3 and S4). Next levels of tree structures slightly differ between the two datasets but converge into a common geographic partitions at k=6 clusters in 3C and k=7 in SU.VI.MAX (Figure 2). The clusters are geographically stratified and were assigned labels to reflect geographic origin: the South-West SW for the dark-red cluster, the South (SO) for the orange cluster, the Centre (CTR) for the yellow cluster, the North-West (NW) for the pink cluster, the North (NO) for the blue cluster and the South-East (SE) for the cyan cluster. In each dataset, one cluster (labelled "*Others*" and coloured in red) included individuals geographically dispersed over France. Furthermore, one cluster identified in SU.VI.MAX included only one individual and was removed in further analysis so that k=7 also resumed to 6 clusters in SU.VI.MAX. At this tree level of 6 clusters, individuals from the NO, NW and CTR clusters are clearly separated in the two datasets. The SW cluster and part of the SO cluster in 3C match geographically the SW cluster identified in SU.VI.MAX while the SE subgroup was not detected in the 3C. This might be explained by differences in the geographic coverage between the two studies especially in the south of France. Indeed, SU.VI.MAX has a better coverage of the south-east whereas 3C lacks data from this region and the reverse is true for the south-west. In the two datasets, two large clusters (CTR and NO) are found that cover most of the central and northern France. Notably, even at the finest level of differentiation (17 and 27 clusters in 3C and SU.VI.MAX respectively), these clusters remain largely intact.

The broad-scale genetic structure of France in six clusters strikingly aligns with two major rivers of France, "La Garonne" and "La Loire" (Figure 2). At a finer-scale, the "Adour" river partition the SW to the SO cluster in the 3C dataset. The mean F_{ST} between clusters inferred by FineSTRUCTURE (Table S1 and S2) are small, confirming subtle differentiation. In both datasets, the strongest differentiation is between the SW cluster and all other regions. These F_{ST} values vary from 0.0016 with the SO cluster to 0.004 with the NW cluster in the 3C dataset and from 0.0009 with the CTR cluster to 0.0019 with the NW cluster in SU.VI.MAX. Finally, besides this subtle division, genetic differentiation within France is also due to isolation by distance as shown by the gradient exhibited on the values of the 1st component of the PCA (Figure S5).

Ancestry profiles by ADMIXTURE & Spatial genetic structure by EEMS

Results obtained by using ADMIXTURE corroborate the FineSTRUCTURE analysis with the SW cluster been the most different from the other groups (Figure S6 and S7). At k=2, the SW cluster shows a light blue component that is significantly less frequent in the other groups (ANOVA *post-hoc* tests, p-value<10^{-6}) (Figure S8). In the 3C dataset, the proportion of light blue tends to decrease gradually from the south-western (SW) part of France to the centre of France (CTR) to finally remain similar in the north of France (NO, NW and *Others*). In SU.VI.MAX, the proportion of light blue component tends to discriminate the north from the

south of France (Figure S8). For k=3, a third major component can be defined, the light green ancestry. In the 3C Study this component is predominant in the north of France (NW and NO clusters) and almost absent in the SW while in SU.VI.MAX this component is predominant in the SE and minimal in the extreme west of France (NW and SW). At k=6, both datasets highlight the differentiation of the SW and the NW cluster from the others clusters.

We performed EEMS analysis in order to identify gene flow barriers within France; i.e; areas of low migrations. We varied the number of demes from 150 to 300 demes and selected a grid of 250 demes showing good concordance between datasets (Figure S9). In both datasets, we identified a genetic barrier around the south-west region (Figure 3). This barrier mirrors the first division in the FineSTRUCTURE. The plots also reveal a gene barrier around Bretagne in the North-West and along the Loire River, which covers the separation of the North cluster. Finally, another barrier is also present on the South-East side that roughly corresponds to the location of the Alp Mountains at the border with Northern Italy.

IBD- estimated population size

Demographic inferences based on IBD patterns in the two datasets were also very concordant. We observed a very rapid increase of the effective population size, Ne, in the last 150 generations (Figure S10). The growth rate was 121% (0.8% per generation) in SUVIMAX and 100% (0.7% per generation) in 3C. This is in accordance with previous observations[34] which reports a very rapid increase of human population size in Europe in the 150 last generations. However, the increase of Ne was not constant over time and a decrease of Ne is observed in both datasets in-between generations 12-22 (from ~1300 to 1700). The growth rates in the period preceding and the period following this decrease were rather different. These growth rates were respectively of 3.1% and 2.4% per generation in 3C and SU.VI.MAX in the last 12 generations and of 0.4% and 0.6% per generation in the first period (22 to 150 generations ago). In-between these two periods, a bottleneck could be detected that could reflect the devastating Black Death. This decrease in Ne seems to affect mainly the Northern part of France (Figure S11). However, this result was not robust to change in parameters: the bottleneck effect was no longer seen when longer IBD chunks were used for Ne estimation (Figure S12).

Relation with neighbor European populations

To study the relationship between the genetic clusters observed in France and neighbor European populations, we combined our two datasets with the 1000G European dataset. As a first step, we run FineSTRUCTURE on the 1000G European populations excluding Finland and found that they could be divided into 3 donor groups as CEU and GBR clustered together (British heritage) (Figure S13). We estimated European ancestry contributions in France with SOURCEFIND[33] and reported the total levels of ancestry proportions for each individual grouped by cluster (Figure 4). We observed similar patterns of admixture between datasets. The proportion of each admixture component from neighboring European countries was significantly different between the six FineSTRUCTURE clusters in both the 3C and SU.VI.MAX datasets (ANOVA, p-value$<10^{-16}$). As expected, the British heritage was more marked in the north than in the south of France where, instead, the contribution from southern

Europe was stronger. TSI was contributing to the SE cluster while IBS was mainly contributing to the SW cluster, which again was very coherent with the geographic places of birth of individuals. In both dataset, SW had the highest proportions of IBS component. Part of this IBS component could in fact reflect a Basque origin as shown on the PCA plot obtained when combining 3C, SU.VI.MAX and HGDP European dataset (Figure S14). This trend is even more pronounced in the 3C where few individuals are grouped together with Basque individuals in the first three dimensions. This SW region also corresponds to the "Aquitaine" region described by Julius Caesar in his "Commentarii de Bello Gallico"[1] (Figure S1).

Discussion

Modern France is located at the western-most part of Europe and is on the way of migrations from and towards South and North of Europe.

In this paper, we have studied the genetic structure of France using data from two independent cohorts of individuals born in different regions of France and whose places of birth could be geolocalized. The French genomes, at the European level, are mapping at their expected position in between Nordic (British and CEU), Italian and Spanish genomes from 1000 genomes project.

Within France, we report correlation between genetic data and geographical information on the individual's place of birth. Although the relation seems linear – reflecting isolation by distance process – we also observed that, based on genotype patterns, the population can be divided into subgroups, which match geographical regions and are very consistent across the two datasets.

An important division separates Northern from Southern France. It may coincide with the von Wartburg line, which divides France into "*Langue d'Oïl*" part (influenced by Germanic speaking) and "*Langue d'Oc*" part (closer to Roman speaking) – Figure S15. This border has changed through centuries and our North-South limit is close to the limit as it was estimated in the IXth century[35,36]. This border is also following the Loire River, which has long been a political and cultural border between kingdoms/counties in the North and in the South. The Aquitaine duchy for instance, spanning the South West of the French Territory, has long represented a civilization on its own.

Secondly, Brittany is identified in both datasets. It may not be surprising as it is both positioned at the end of the continent as a peninsula and, following the political maps, has been an independent political entity (Kingdom and, later, duchy of *Bretagne*), with stable borders, for a long time[37].

The extreme South-West regions show the highest differentiation to neighbor clusters. This is particularly strong in 3C dataset, where we even observe an additional cluster. This cluster is likely due to a higher proportion of possibly Basque individuals in 3C, which overlap with HGDP Basque defined individuals. The F_{ST} between the south-west and the other French clusters were markedly higher than the F_{ST} between remaining French clusters. In 3C these values are comparable to what we observed between the Italian and the British heritage clusters (F_{ST}=0.0035). Similar trends are observed in SU.VI.MAX even though the level of differentiation with the SW was weaker.

We also observe that the broad-scale genetic structure of France strikingly aligns with two major rivers of France "La Garonne" and "La Loire" (Figure 3). At a finer-scale, the "Adour" river partition the SW to the SO cluster in the 3C dataset.

While historical, cultural and political borders seem to have shaped the genetic structure of modern-days France, exhibiting visible clusters, the population is quite homogeneous with low F_{ST} values between-clusters ranging from 2.10^{-4} up to 3.10^{-3}. We find that each cluster is genetically close to the closest neighbor European country, which is in line with a continuous gene flow at the European level. However, we observe that Brittany is substantially closer to British Isles population than North of France, in spite of both being equally geographically close. Migration of Britons in what was at the time Armorica (and is now Brittany) may explain this closeness. These migrations may have been quite constant during centuries although a two wave model is generally assumed. A first wave would have occurred in the X^{th} century when soldiers from British Isles were sent to Armorica whereas the second wave consisted of Britons escaping the Anglo-Saxon invasions[38]. Additional analyses, on larger datasets may be required to discriminate between these various models.

Studying the evolution of French population size based on genetic data, we observe a very rapid increase in the last generations. This observation is in line with what has been seen in European populations[39]. We also observe, in most cases, a depression during a period spanning from 12 to 22 generations ago. This may correspond to a period spanning from 1,300 to 1,700. Indeed, this period was characterized by a deep depression in population size due to a long series of plague events. While the population size in kingdom of France was estimated to be 20 Million in 1348, it dropped down to 12 Million in 1400, followed by an uneven trajectory to recover the 20 Million at the end of Louis XIV^{th} reign (1715)[40].

However, the decrease we observe in the genetic data does seem to affect mainly the Northern part of France, and for instance is mainly observed in the NO cluster. We see no reason for this trend based on historical records (Figure S16) except perhaps the last plague epidemics in 1666-1670 that was limited to the North of France. Alternatively, a more spread population in the South (which is in general hilly or mountainous) may explain a lower impact of these dramatic episodes. Plague is expected to have had a very strong impact on the population demography in the past as some epidemics led to substantial reduction in the population sizes [41]. However, we could not detect in our data any footprint of the Justinian plague (541-767 PC) although, according to historical records it had a major impact on the population at that time. This may be due to difficulty to estimate population changes in ancient times, deeper than 50-100 generations, especially in presence of more recent bottleneck and given our sample size and IBD resolution power. We expect that increasing sample size will help getting more detailed information farther in the past. Finally, we cannot rule out the effect of undetected admixture, which can mimic a bottleneck. This plateau is less and less visible when we restrict the analysis to the largest IBD, and hence most informative, segments. Therefore, this observation remains to be confirmed in independent datasets and in a more comprehensive study that could account for possible confounding events.

Identification of genetic structure is also important to guide future studies of association both for common, but more importantly, for rare variants[42]. In the near future, interrogating the demographical history of France from genetic data will bring more precise results thanks to

whole genome sequencing. This new data, along with new methods will allow testing formal models of demographic inference.

References

1. C. Julius Caesar, De bello Gallico, COMMENTARIUS PRIMUS, chapter 1, section 1. Available at: http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus%3Atext%3A1999.02.0002%3Abook%3D1%3Achapter%3D1%3Asection%3D1. (Accessed: 4th September 2018)
2. Lazaridis, I. The evolutionary history of human populations in Europe. *Curr. Opin. Genet. Dev.* **53**, 21–27 (2018).
3. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
4. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
5. Rosenberg, N. A. *et al.* Genetic Structure of Human Populations. *Science* **298**, 2381–2385 (2002).
6. Jorde, L. B. *et al.* The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data. *Am. J. Hum. Genet.* **66**, 979–988 (2000).
7. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
8. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
9. Heath, S. C. *et al.* Investigation of the fine structure of European populations with applications to disease association studies. *Eur. J. Hum. Genet. EJHG* **16**, 1413–1429 (2008).

10. Humphreys, K. *et al.* The genetic structure of the Swedish population. *PloS One* **6**, e22547 (2011).

11. Karakachoff, M. *et al.* Fine-scale human genetic structure in Western France. *Eur. J. Hum. Genet.* (2014). doi:10.1038/ejhg.2014.175

12. Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet. EJHG* (2013). doi:10.1038/ejhg.2013.48

13. The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland - s41598-017-17124-4.pdf. Available at: https://www-nature-com.insb.bib.cnrs.fr/articles/s41598-017-17124-4.pdf. (Accessed: 20th March 2018)

14. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).

15. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat. Commun.* **10**, (2019).

16. Hercberg, S. *et al.* The SU.VI.MAX Study: A Randomized, Placebo-Controlled Trial of the Health Effects of Antioxidant Vitamins and Minerals. *Arch. Intern. Med.* **164**, 2335 (2004).

17. 3C Study Group. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology* **22**, 316–325 (2003).

18. Lambert, J.-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094–1099 (2009).

19. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

20. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

21. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).

22. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

23. Lawson, D., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet* **8**, e1002453 (2012).

24. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).

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

26. Browning, S. R. & Browning, B. L. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).

27. Browning, B. L. & Browning, S. R. Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. *Am. J. Hum. Genet.* **93**, 840–851 (2013).

28. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194**, 459–471 (2013).

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

30. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).

31. Price, A. L. *et al.* Long-Range LD Can Confound Genome Scans in Admixed Populations. *Am. J. Hum. Genet.* **83**, 132–135 (2008).

32. Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E. & Leutenegger, A.-L. High level of inbreeding in final phase of 1000 Genomes Project. *Sci. Rep.* **5**, 17453 (2015).

33. Chacon-Duque, J. C. *et al.* Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. *bioRxiv* 252155 (2018). doi:10.1101/252155

34. Keinan, A. & Clark, A. G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**, 740–743 (2012).

35. Wartburg, W. von. *Les origines des peuples romans*. (Presses universitaires de France, 1941).

36. Fabula, É. de recherche. J. Chaurand (dir.), Nouvelle histoire de la langue française. *https://www.fabula.org* Available at: http://www.fabula.org/actualites/j-chaurand-dir-nouvelle-histoire-de-la-langue-francaise_104.php. (Accessed: 29th November 2018)

37. Leprohon, R. *Vie et mort des Bretons sous Louis XIV*. (Bibliophiles de Bretagne, 1984).

38. Fleuriot, L. *Les origines de la Bretagne: l'émigration*. (Payot, 1980).

39. Keinan, A. & Clark, A. G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**, 740–743 (2012).

40. *Histoire de la population française Coffret 4 volumes : Volume 1, Des origines à la Renaissance. Volume 2, De la Renaissance à 1789. Volume 3, De 1789 à 1914. Volume 4, De 1914 à nos jours - Jacques Dupâquier*.

41. Le temps de la Guerre de Cent ans (1328-1453). *Belin Editeur* (2016). Available at: https://www.belin-editeur.com/le-temps-de-la-guerre-de-cent-ans-1328-1453. (Accessed: 13th December 2018)

42. Persyn, E., Redon, R., Bellanger, L. & Dina, C. The impact of a fine-scale population stratification on rare variant association test results. *PLOS ONE* **13**, e0207677 (2018).

FIGURES

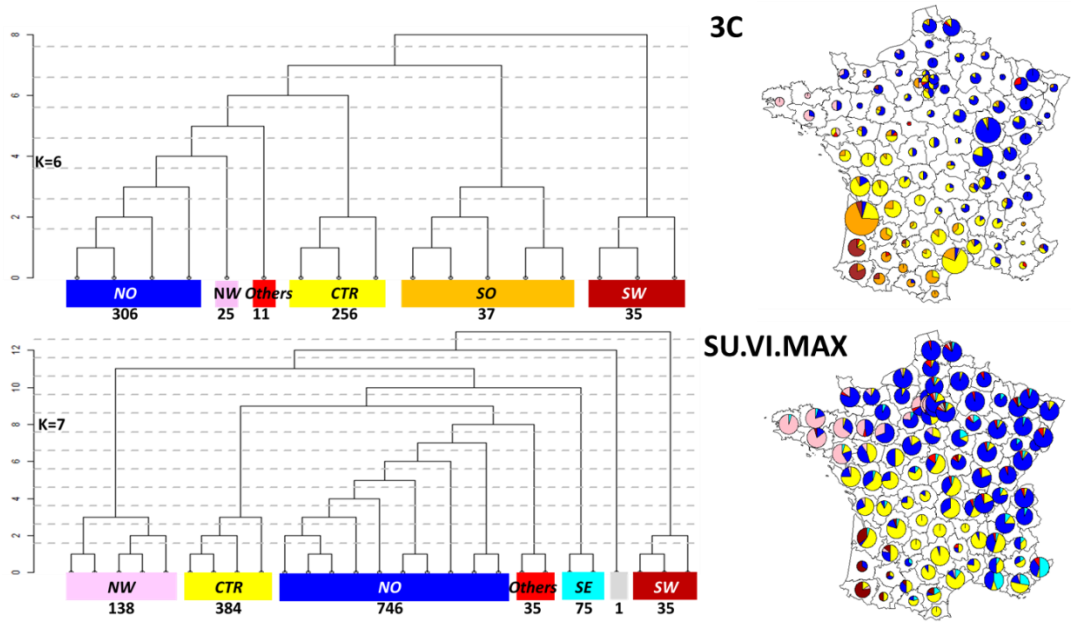


Figure 1: FineSTRUCTURE clustering of the 3C Study (770 individuals) and SU.VI.MAX (1,414 individuals) pooled in 6 and 7 clusters respectively. Left side shows the tree structure and right side shows by *département* pie charts indicating to which of the six clusters the individuals belong to.

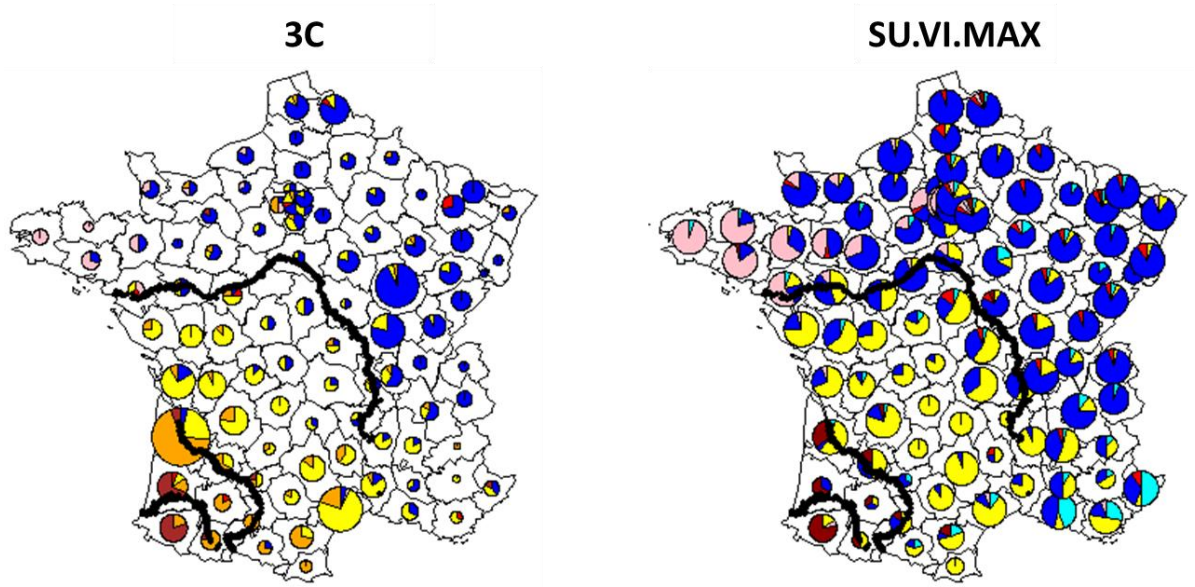


Figure 2: Pie charts indicating the proportion of individuals from the different “départements” assigned to each cluster. Results are reported for the partition in 6 clusters obtained by running FineSTRUCTURE in the 3C dataset (left) and in SU.VI.MAX (right) independently. Geographic coordinates of three rivers of France are drawn in black: Loire, Garonne and Adour from north to south.

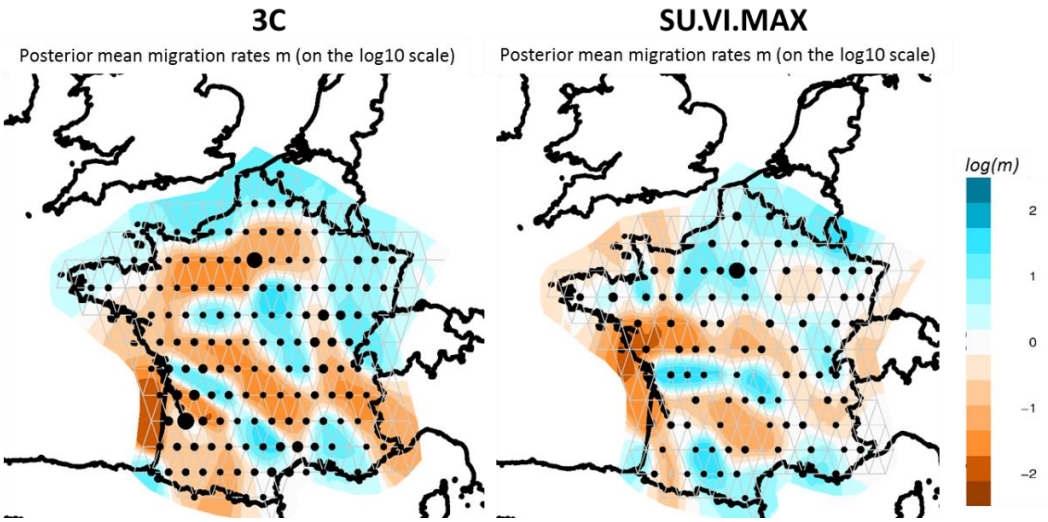


Figure 3: Estimated effective migration surfaces of France obtained from EEMS on the 3C (left) and SU.VI.MAX (right) datasets. The colour scale reveals low (blue) to high (orange) genetic barriers between populations localized on a grid of 250 demes. Each dot is proportional to the number of populations included.

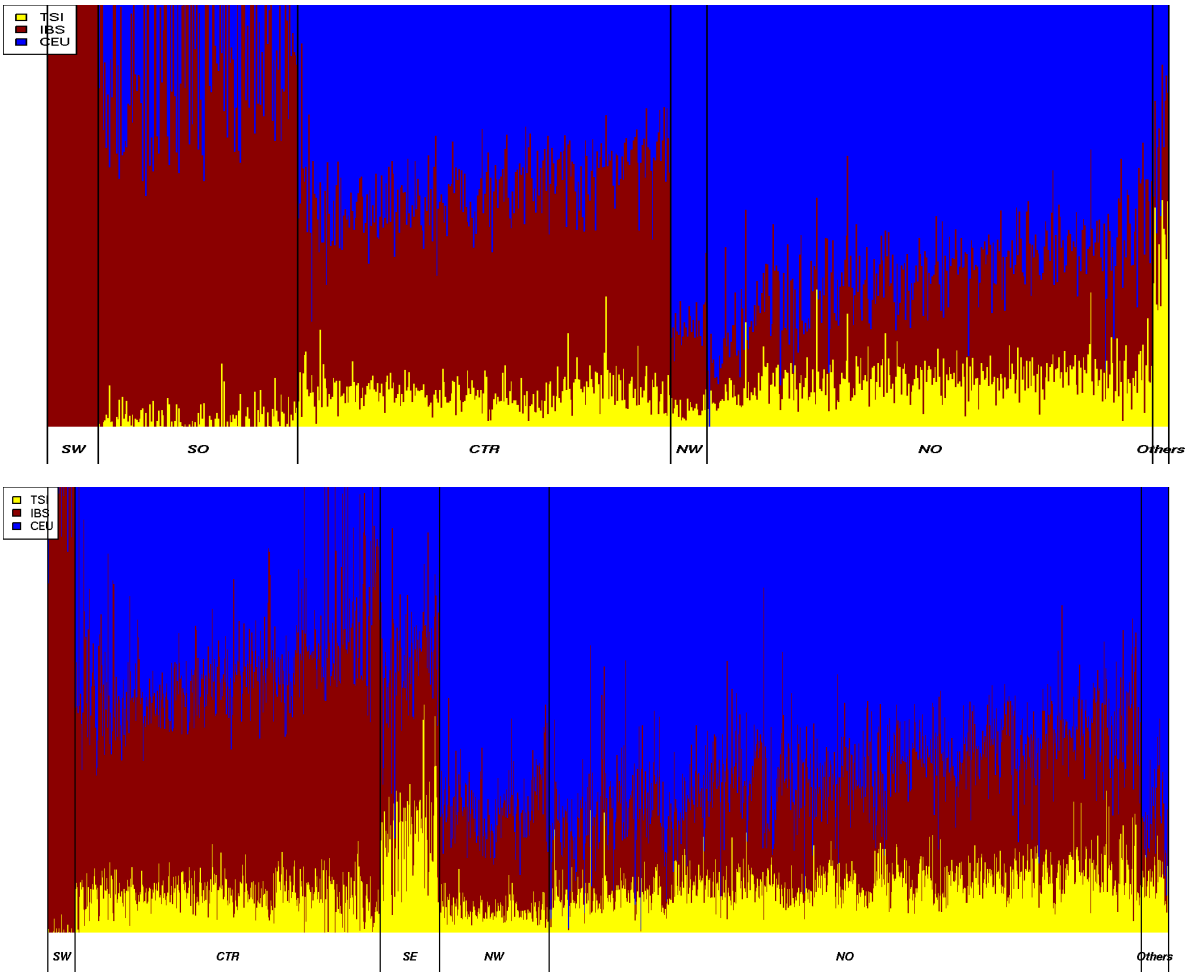


Figure 4: Ancestry profiles from the three neighbouring European populations inferred by SOURCEFIND in the French 3C (top) and SU.VI.MAX individuals (bottom) datasets. In each cluster, individuals are ordered according to the latitude of their reported birth place.

SUPPLEMENTARY DATA

Effective migration surface EEMS

We draw polygons around France with the online Google Maps tool (http://www.birdtheme.org/useful/v3tool.html). Depending on their location, several populations can be included in one deme, whose size increases accordingly. We ran the Markov chain Monte Carlo 5 times with different random seeds, each time with 9.9 million burn-in and 10 million regular iterations, thinning every two hundred iterations. For each deme, we chose the chain with the highest final log-likelihood, and started a second round of EEMS chains using this chain as a starting point and 1,000,000 additional sampling iterations thinning every 9,999 iterations. The dissimilarity between observed versus fitted deme pairs show a general trend with some deviation and the log-posterior trace of the replicate MCMC chains (Figure S9) show convergence of the independent EEMS runs.

SUPPLEMENTARY TABLES

F_{ST} in 3C	SO	CTR	NW	NO	IBS	TSI	CEU+GBR
SW	0.0016	0.0029	0.0040	0.0035	0.0022	0.0049	0.0047
SO		0.0002	0.0013	0.0006	0.0003	0.0019	0.0016
CTR			0.0009	0.0002	0.0006	0.0016	0.0012
NW				0.0006	0.0019	0.0033	0.0005
NO					0.0010	0.0017	0.0006
IBS						0.0014	0.0023
TSI							0.0035

Table S1: F_{ST} table between the French 3C clusters and the 1000G European clusters inferred by FineSTRUCTURE. Mean F_{ST} statistics are estimated using EIGENSOFT.

F_{ST} in SU.VI.MAX	CTR	SE	NW	NO	IBS	TSI	CEU+GBR
SW	0.0009	0.0015	0.0019	0.0014	0.0007	0.0028	0.0026
CTR		0.0004	0.0007	0.0002	0.0005	0.0016	0.0012
SE			0.0013	0.0004	0.0006	0.0007	0.0017
NW				0.0004	0.0017	0.0030	0.0005
NO					0.0010	0.0017	0.0006
IBS						0.0015	0.0023
TSI							0.0035

Table S2: F_{ST} table between the French SU.VI.MAX clusters and the 1000G European clusters inferred by FineSTRUCTURE. Mean F_{ST} statistics are estimated using EIGENSOFT.

SUPPLEMENTARY FIGURES

a) Pre-Roman division

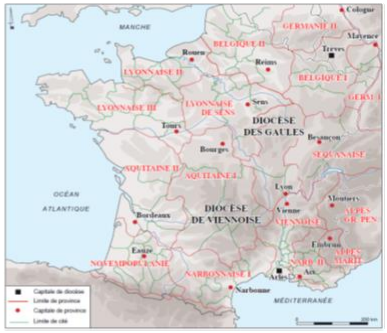
Roman conquest -52

b) Roman Provinces I[st] Century

Franks – *Saliens* – installed in Gaul

Caracalla Edict 323

c) Roman Provinces IV Century



Wisigoths conquer South West Gaul 410

d) VIIth century
German Barbarian Kingdoms

Frank Kingdom of Clovis
Charlemagne
Traité de Verdun 509

e) XIIth century Feudality

Hundred Years War

Figure S1: a/ Pre-Roman Gaul as described by Julius Caesar (map reproduced from Caesar's footprints: Journeys to Roman Gaul from Bijan Omrani); b/ Ist Century Roman provinces. The *pagi* represent also ancient Gaul people territories. c/ Fourth Century Roman Provinces. d/ VIIth century (Franks, Wisigoths and Burgunds) e/ Duchies and Counties of early French feudality.

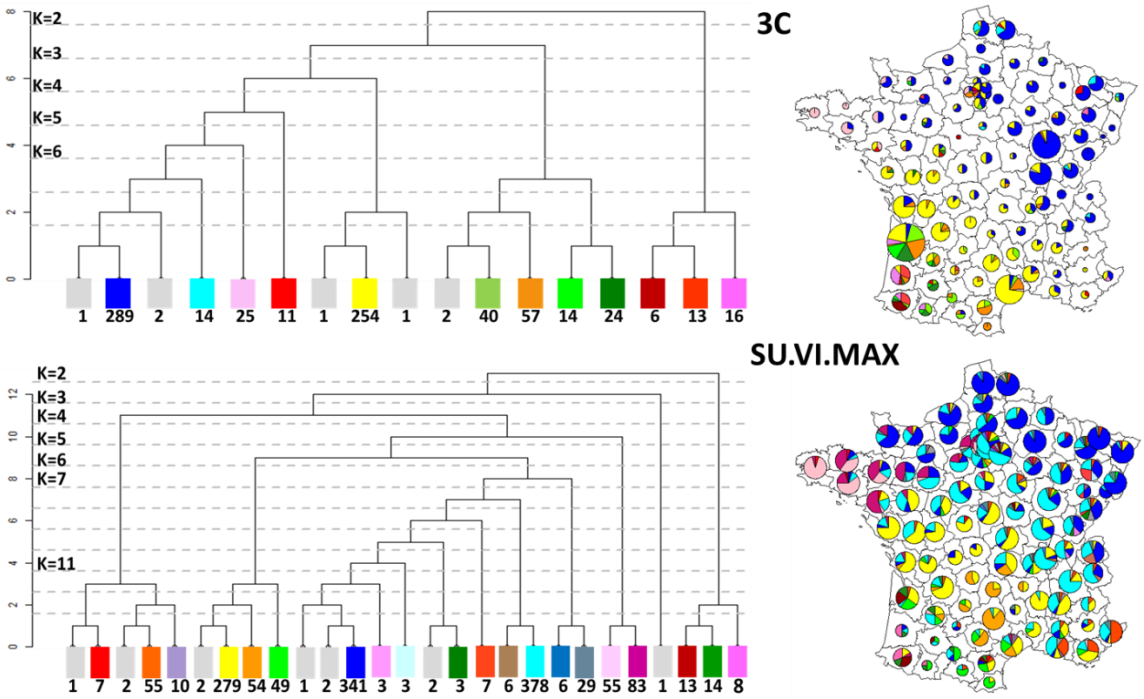


Figure S2: FineSTRUCTURE clustering of the 3C individuals (770 individuals, top) pooled in 17 clusters and SU.VI.MAX individuals (1,414 individuals, bottom) pooled in 27 clusters. Left side: tree structure. Right side: birth location of individuals coloured according to their assigned cluster. Clusters with less than 3 individuals are coloured in grey.

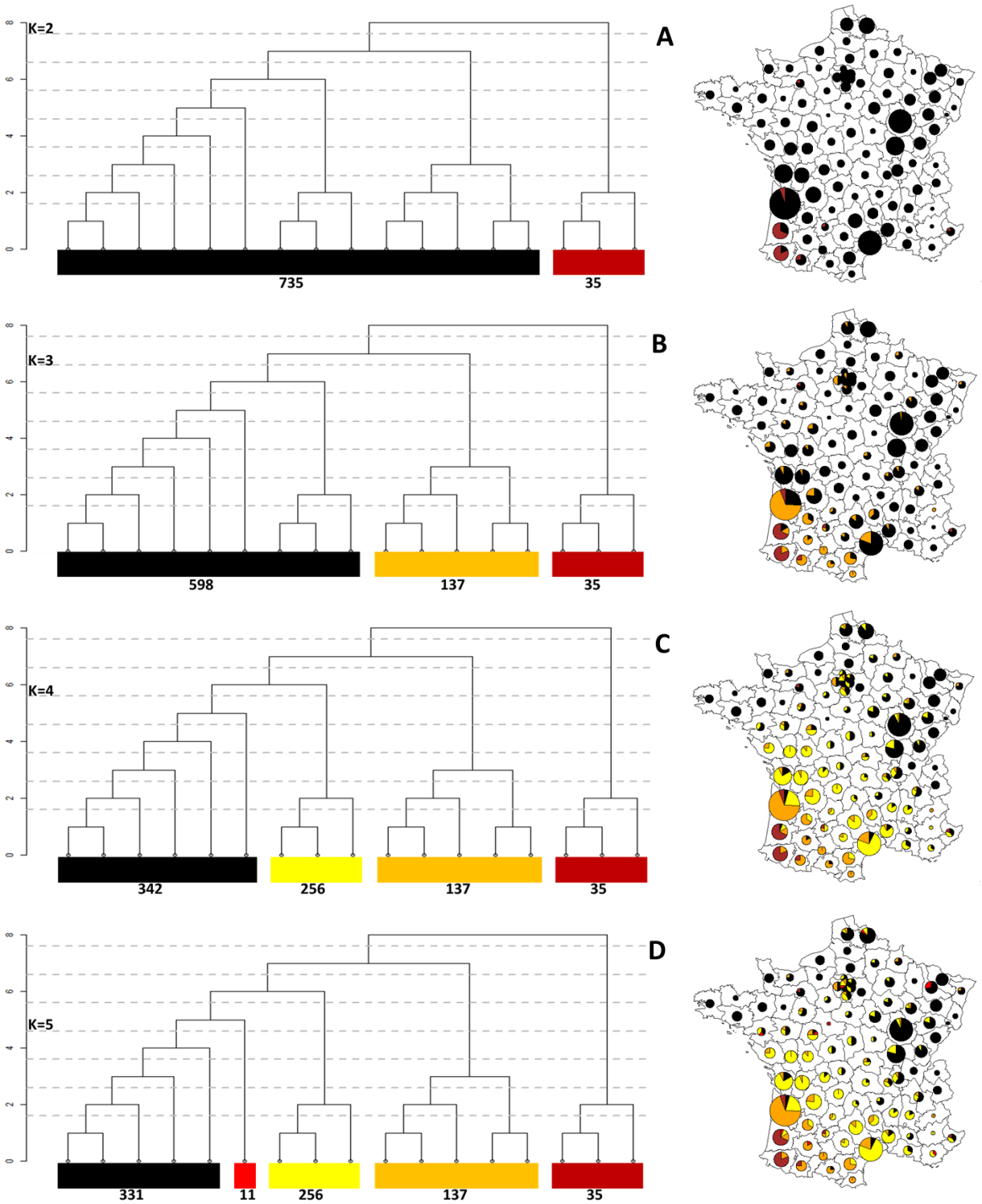


Figure S3: Genetic clusters in the 3C data inferred by the FineSTRUCTURE analysis at levels of the hierarchical clustering varying from k=2 (A), k=3 (B), k=4 (C) and k=5 (D).

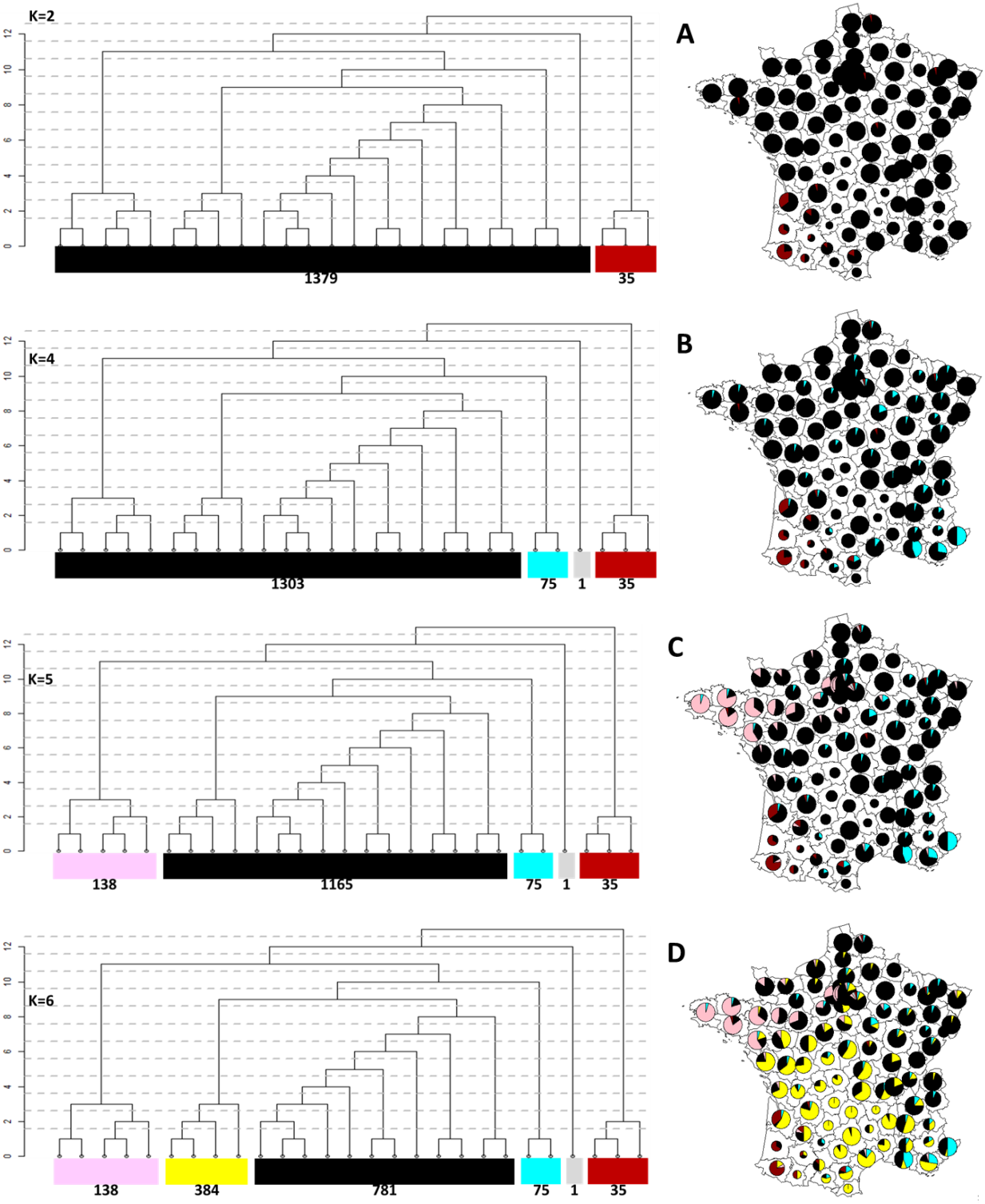


Figure S4: Genetic clusters in the SU.VI.MAX data inferred by the FineSTRUCTURE analysis at levels of the hierarchical clustering varying from k=2 (A), k=4 (B), k=5 (C) and k=6 (D).

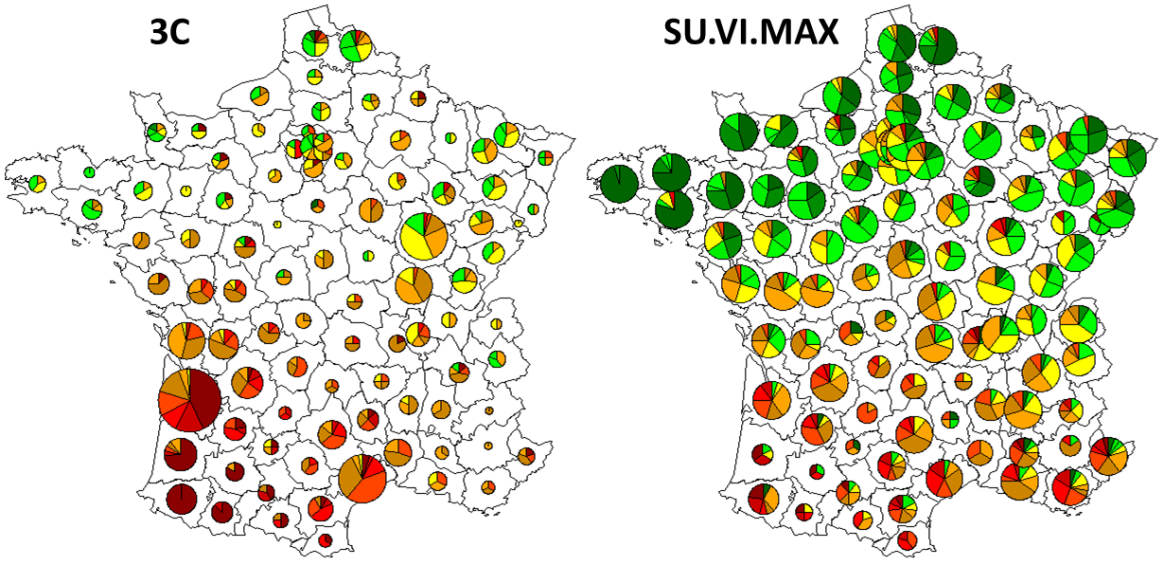


Figure S5: Distribution of PC1 values per *département* in the 3 Cities study and SUVIMAX. Colour of the points indicates the range of PC values. Red colours indicate negative values while green colours indicate positive values.

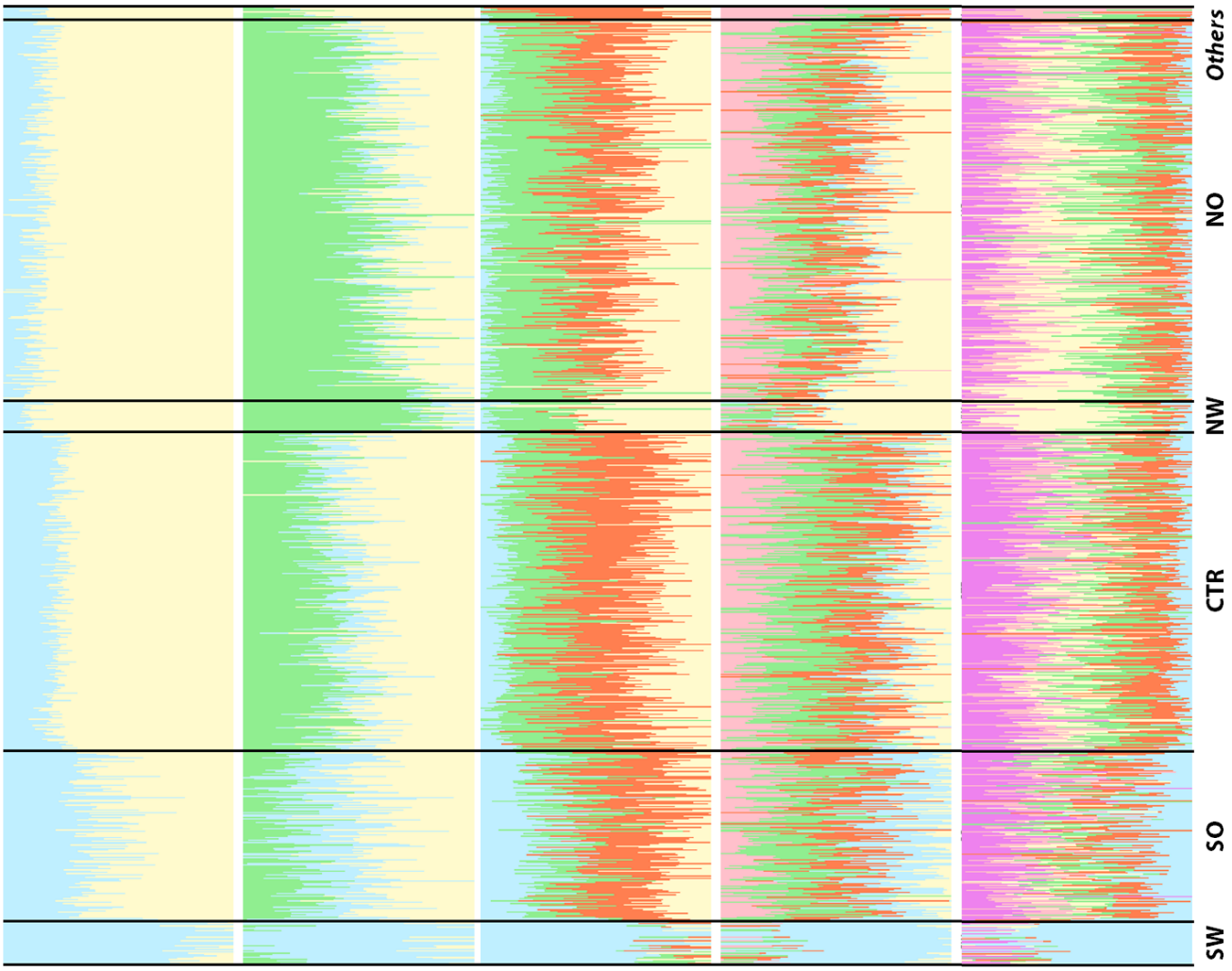


Figure S6: Results of ADMIXTURE assuming k=2 to k=6 in 3C. Each vertical line represents an individual and the different colors represent various ancestry components. In each cluster, individuals are sorted according to their geographical latitude.

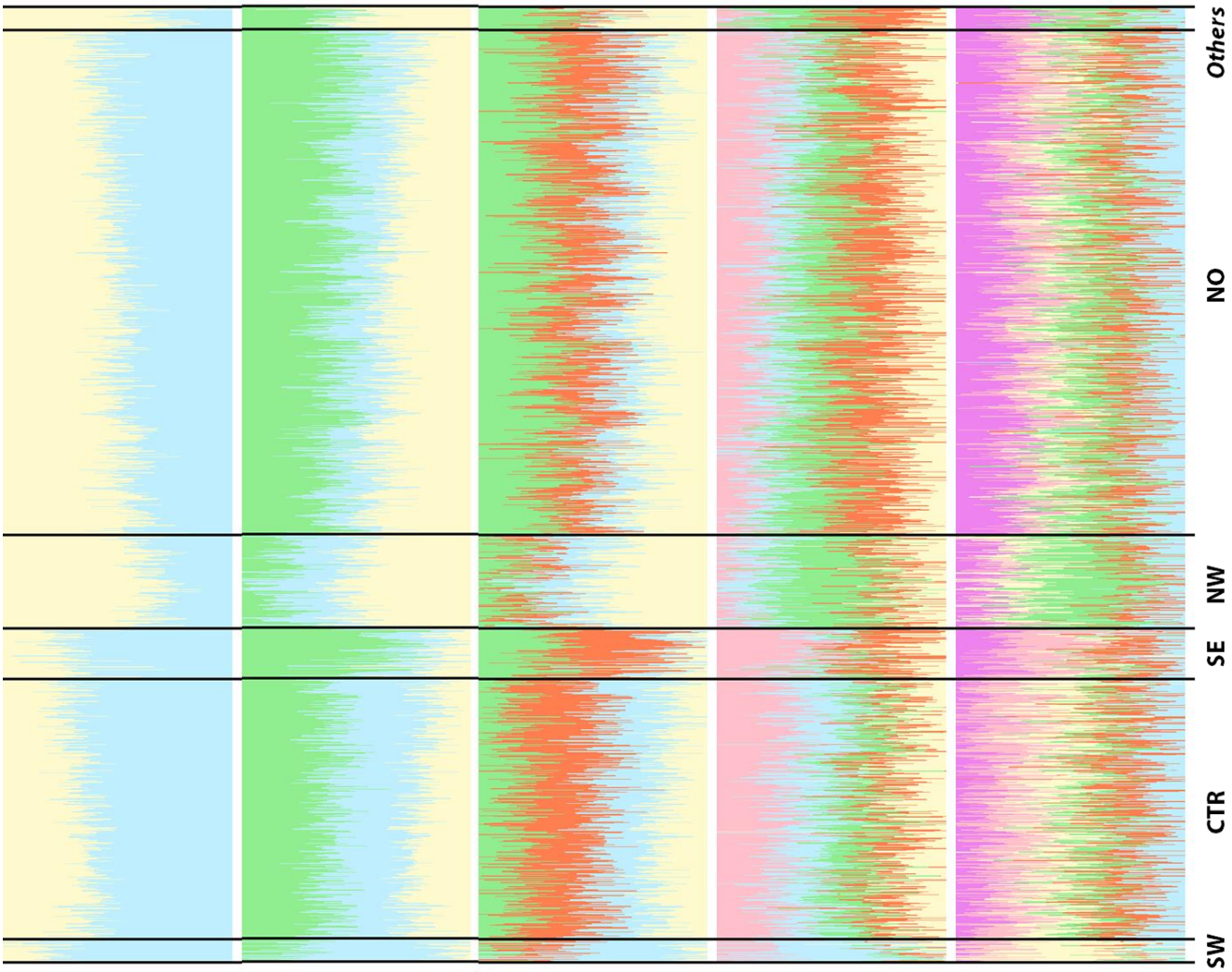


Figure S7: Results of ADMIXTURE assuming k=2 to k=6 in SU.VI.MAX. Each vertical line represents an individual and the different colors represent various ancestry components. In each cluster, individuals are sorted according to their geographical latitude.

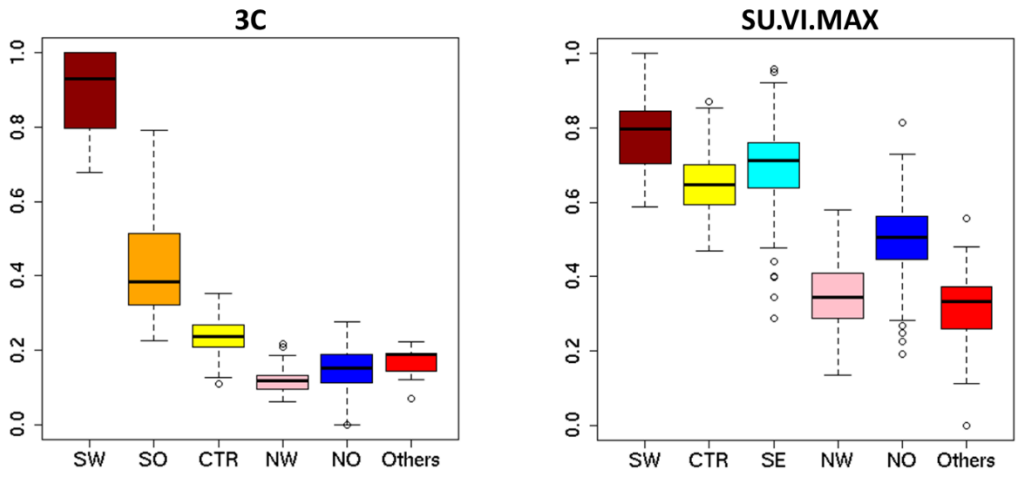


Figure S8: Proportion of the light blue ancestry at k=2 from ADMIXTURE in the 3 Cities study (left) and SU.VI.MAX (right)

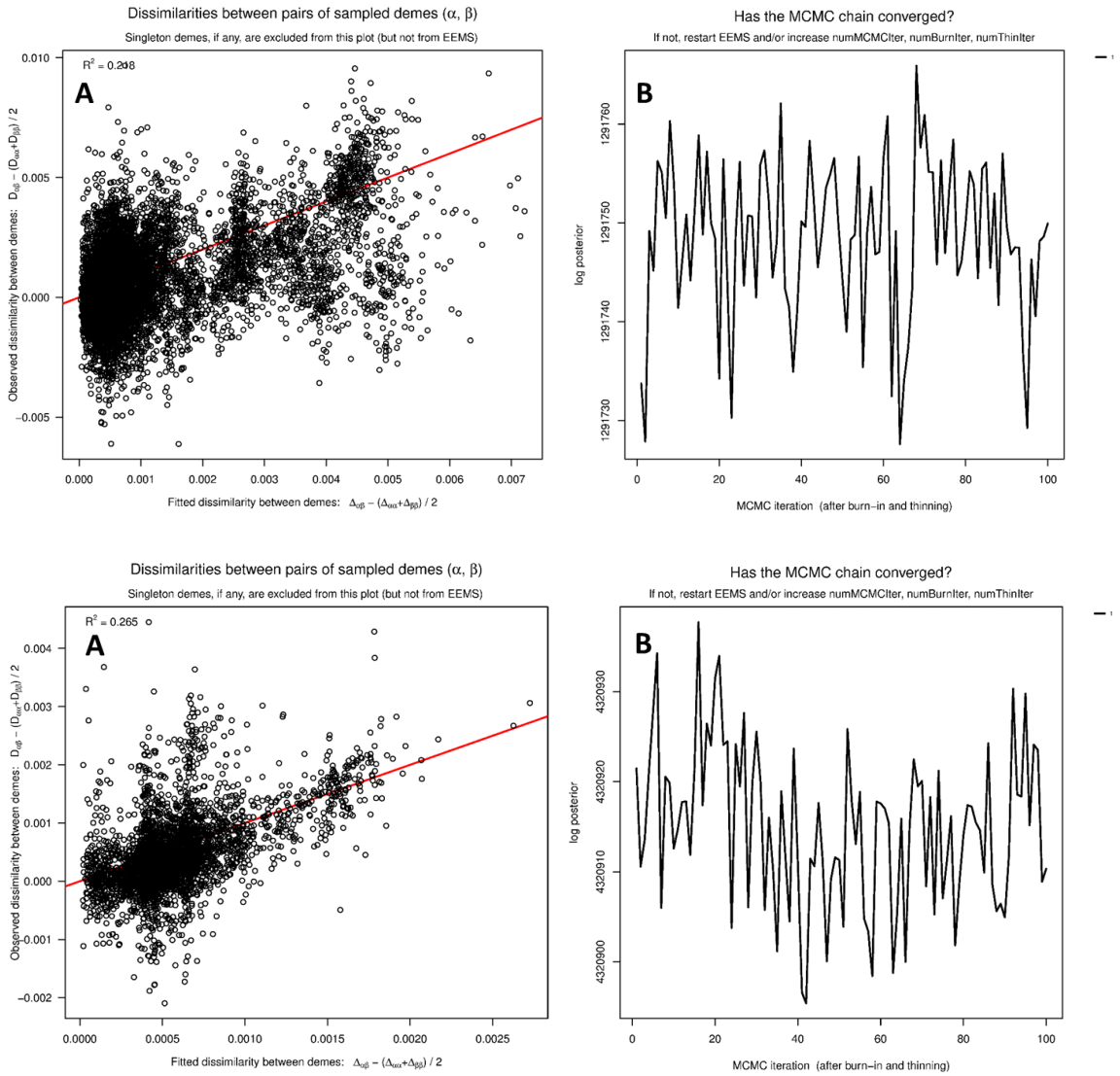


Figure S9: Estimated Effective Migration Surface Diagnostic Plots from EEMS performed on a grid of 250 demes with 770 3C (top) and 1,414 SU.VI.MAX individuals (bottom). (A) The observed versus expect dissimilarity between pairs of demes. Strong deviations from the fitted line (red) indicate pairs of demes much more genetically distant than expected. (B) The posterior probability log of the EEMS run, indicating whether the MCMC chains have converged.

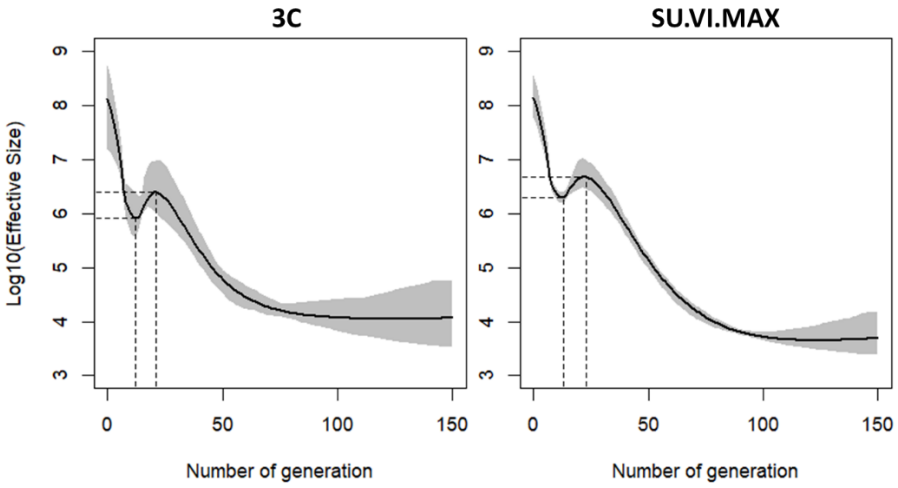


Fig S10: Evolution of effective size in 3C and in SU.VI.MAX.

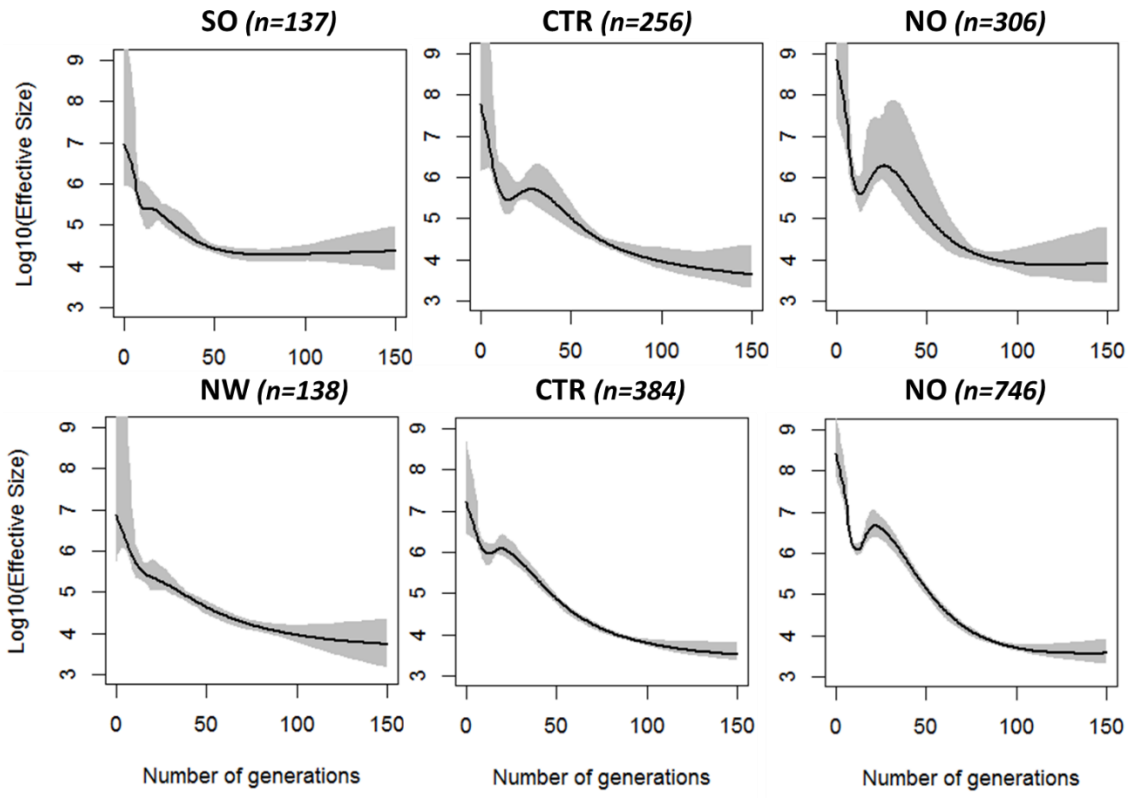


Figure S11: Evolution of effective size within cluster of France higher than 100 individuals in 3C (top: SO, CTR and NO) and in SU.VI.MAX (bottom: NW, CTR and NO).

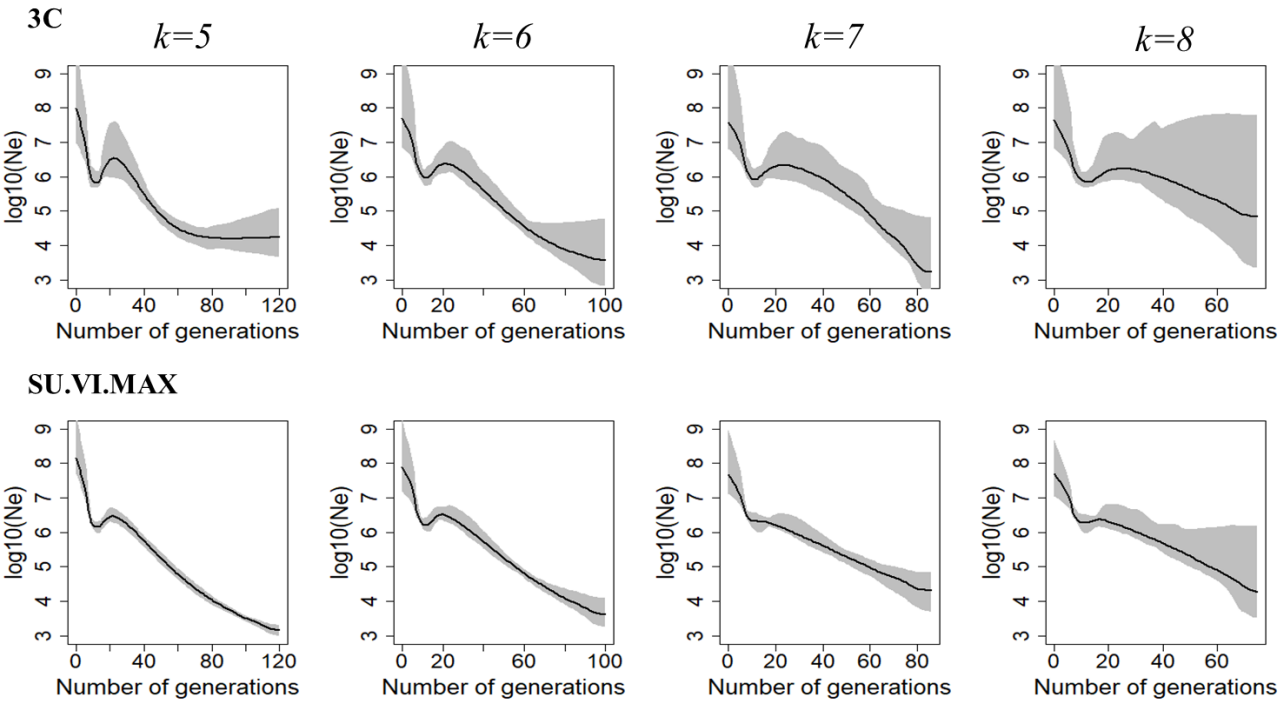


Figure S12: Evolution of effective size in the NO cluster with various mincm (k=5, 6, 7, 8) parameter in 3C (top) and in SU.VI.MAX (bottom).

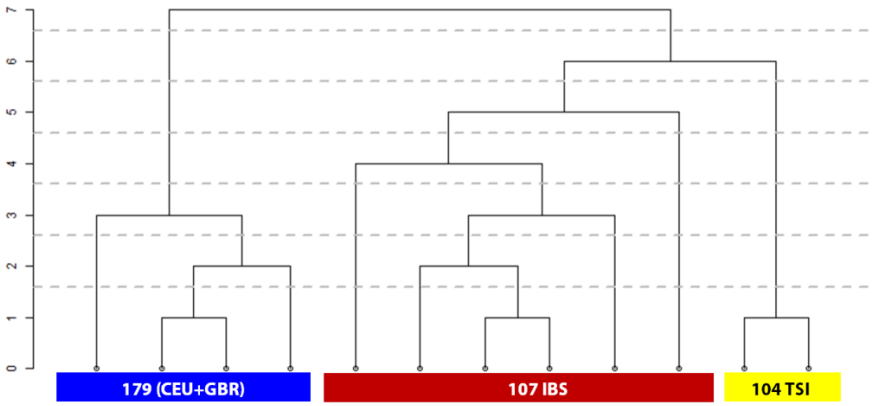


Figure S13: FineSTRUCTURE clustering of the European 1000G populations (390 individuals). The set of common SNPs with the 3C sample was used in this analysis.

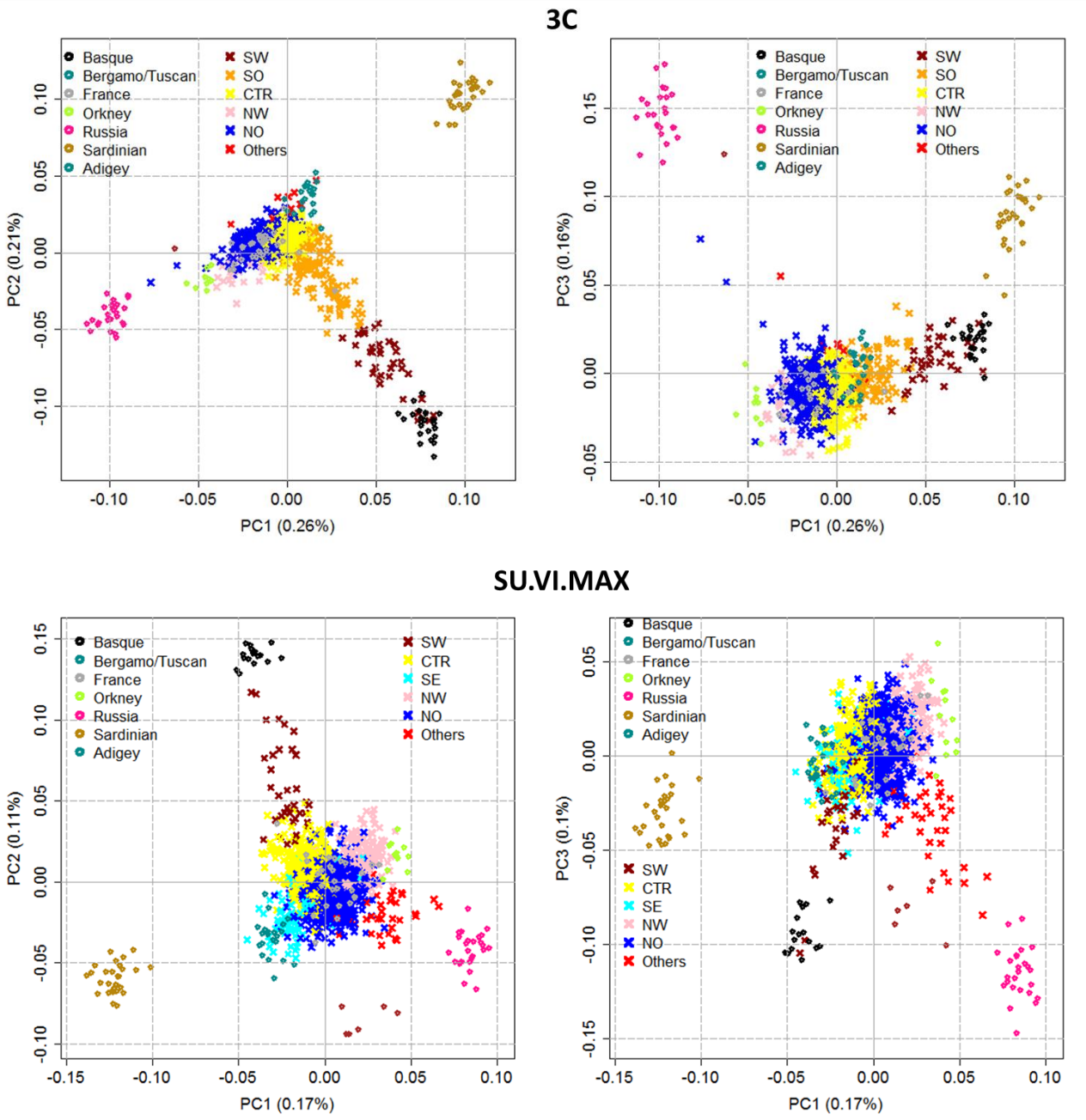


Fig S14: The scatter plot of the first three PCs from PCA performed on individuals from the 3 Cities study (top) or SU.VI.MAX (bottom) coloured according to the FineSTRUCTURE clustering and combined with the Europeans populations from the HGDP panel.

Fig S15: Limits of French languages. After Chaurand

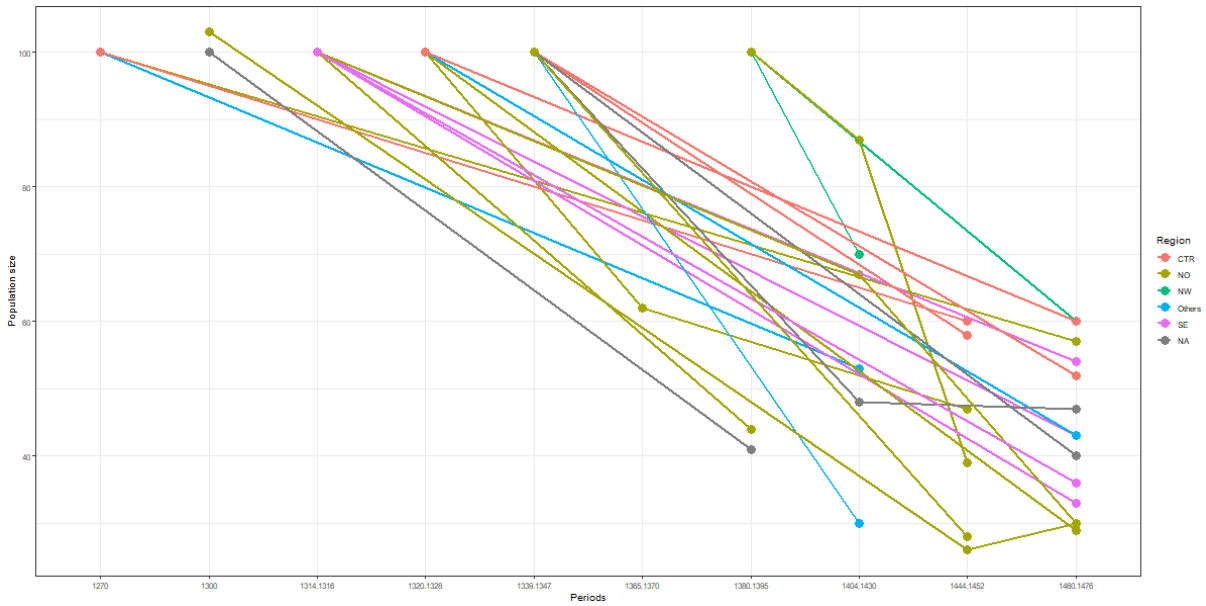


Figure S16: Change in population size in France between 1270 and 1476.

The table compiles results from more than 20 studies of local demography and is published in “Boris Bove, La France de la guerre de Cent Ans, Paris, Belin, 2009, rééd. 2013, p. 277-309 (chap. 8 - Les épidémies et la saignée démographique (XIVe-XVe siècles)".

Each trajectory represents the decay in population size of a French province. Lines are coloured according to the geographical cluster where each province belongs to cluster:

NO : Artois, Cambrésis, Normandie orientale, Île-de-France, Champagne méridionale, Verdunois

NW : Britanny

CTR : Limousin, Forez, Lyonnais, Dauphinois

SE : Basse Provence centrale, Basse Provence occidentale, Provence orientale, Haute Provence

SW : Languedoc, Bigorre, Navarre

2.1 Additional information on the SU.VI.MAX study

2.1.1 Materials and methods

The SU.VI.MAX study

The study SU.VI.MAX (Supplémentation en vitamines et minéraux antioxydants) was initiated in 1994 with the aim to collect information on alimentary consumption of French people and their health state. It was led by Serge Hercberg, an epidemiologist and a specialist in nutrition. The results of this clinical study, together with a detailed description of the recruitment of 12,741 participants and the trial, were published in 2004 [79].

Samples, genotyping and quality control

2712 participants were genotyped with SNP arrays: 1,856 with Illumina 300k/300kduo and 856 with Illumina 660W. Out of them, 2276 individuals contributed data to my analysis. The reason for removal of certain samples was insufficient quality (2 individuals had a call rate below 95%) or an unknown birthplace or a birthplace outside of continental France (434 individuals). Birthplace information was not very precise - only the department (French administrative unit) of birth was known. This still represents an interesting grid because it often overlaps historical counties and is relatively fine-grained. All 94 continental French departments are covered. However, some departments are overrepresented while other are underrepresented. Thus, to obtain a more regular coverage, we considered a subset of 1414 individuals for analysis, with the number of individual per department not exceeding 20. Exclusion of individuals from overrepresented departments was random.

SNPs were removed if they had a genotype-missing rate > 2%, a minor allele frequency < 1% or departed from Hardy–Weinberg proportion (p-value $< 10^{-5}$). To make the study more comparable with Three Cities (3C) cohort, SNPs not present in the 3C dataset were removed. This led to 263, 284 autosomal SNPs.

Data were obtained with two different SNP chips, thus we tested for presence of batch effect, in a following way: we selected a subset of the individuals, so that in each department the number of individuals genotyped with Illumina 300k/300kduo was the same as the number of individuals genotyped with Illumina 660W. In this way, we obtained equal spatial distribution for both SNP chips. This subset counted 1176 individuls (588 for each SNP chip). A two sample t-test was applied to values of PC1, PC2, PC3, PC4 and PC5. In each case, the mean values for two SNP chips were not significantly different. We concluded that there is no batch effect caused by the use of two different SNP chips in our dataset.

Principal component analysis

PCA was carried out using the smartpca software from the EIGENSOFT package version 6.0.1 [162]. No outliers were detected with the default outlier removal procedure. SNPs in strong LD were pruned out with PLINK 1.9 [34, 172] in a two-step manner. First, the `indep-pairwise` command in PLINK was run with $r^2 = 0.2$, a window size of 50 SNPs and 10 SNPs to shift the window at each step. Second, in order to deal with the remaining long-range LD, SNPs were pruned again for r^2 above 0.2, computed between variants at increased 5Mb distance and in windows of 50 variants. 82,362 SNPs were left for analysis after LD pruning.

To evaluate the geographic relevance of PCs, I tested for the significance of association between the latitude and longitude of each department and PCs coordinates (`cor.test` function in R) using a Spearman's rank correlation coefficient.

CHROMOPAINTER and FineSTRUCTURE analyses

The genotype data (after QC) were phased jointly for all individuals with SHAPEIT v2.r790 [39], without reference panel, with the genetic map build 37 provided with that software. Phased genotype files were

converted into CHROMOPAINTER format using the `impute2chromopainter2.pl` script. Global Ne and mutation rates were estimated with CHROMOPAINTER version 2 on chromosomes 1, 4, 10 and 15 and 140 individuals (10% of the total sample). Then CHROMOPAINTER was run on full data with estimated parameter values. Coancestry matrix estimates the proportion of the genome of each individual that is most closely related to every other individual in the matrix, in particular chunkcounts matrix is based on the number of copied haplotype chunks (alternative matrix is based on lengths of the copied chunks). On the chunkcounts matrix, FineSTRUCTURE version 2.1.3 was run with 10,000,000 burn-in iterations, 1,000,000 MCMC iterations from which every 10,000th iteration was recorded, keeping default values of the other options. Tree was built using 100,000 tree comparisons and 10,000,000 additional optimisation steps. MCMC convergence was assessed by comparing the assignment to clusters in a second independent chain.

EEMS

I estimated an effective migration surface using the software EEMS. The matrix of average pairwise genetic dissimilarities was generated for 82,362 SNPs (pruned dataset) and 1414 individuals using the `bed2diffs` software included in the EEMS package. Samples were assigned to the nearest of 300 demes (but fewer than 300 demes were observed). Grids of 200 and 250 demes were tested too and presented consistent results (data not shown). We run ten independent MCMC chains, each with a random seed, for 10,000,000 iterations, including 9,900,000 burn-in iterations, thinning every 200 iterations. Next, I chose the chain with the highest final log-likelihood, and started a second round of ten EEMS chains using this chain as a starting point for 1,000,000 additional sampling iterations, thinning every 9,999 iterations. Log-posterior trace of ten replicate MCMC chains from the last round shows mixing and convergence of the independent EEMS runs (Figure 2.1). Plots were generated in R using the `rEEMSplots` package.

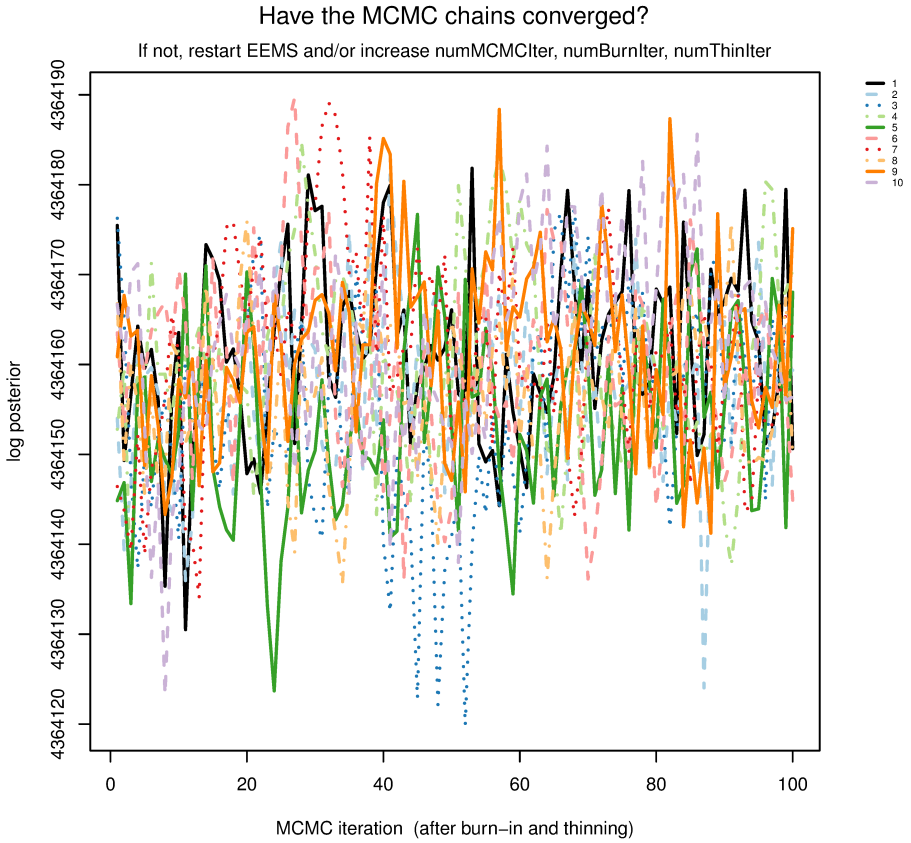


Figure 2.1 – Log-posterior trace of ten MCMC chains from the last round of EEMS analysis. The diagram indicates mixing and convergence of the chains.

IBDNe

I estimated the trajectories of effective population size with IBDNe (version released on 7th May 2018) [23]. I detected IBD segments with IBDseq (version r1206, with default settings). I performed scan for genomic regions with excess IBD as described in [23] to avoid potential bias. Excess IBD was present on the part of the major histocompatibility complex (MHC) on chromosome 6. We decided to exclude MHC and the large inversion on chromosome 8 (known of low recombination) from final analysis, using coordinates of corresponding extended long-range LD regions from [170]. Thus, we split chromosome 6 into two continuous parts (0-25.5 Mb and 33.5Mb-172Mb and considered them as separate chromosomes, and trimmed one side of chromosome 8 (0-12Mb). Discarding not only the inversion but the preceeding part was necessary, because continuous chromosome intervals of length substantially shorter than 50 cM may lead to high variability of confidence interval bootstrap estimation. Additionally, we performed two robustness analyses, either leaving one chromosome out or keeping excess IBD regions. The impact on the IBDNe results was inconsiderable (data not shown). We varied mincm parameter (minimum length of IBD segments considered) to find appropriate value for the SNP density on our SNP arrays. We decided that 6 cM is the most appropriate value for SU.VI.MAX dataset that minimizes potential technical artifacts.

Genetic signature of natural selection: Integrated Haplotype Score (iHS)

The extended haplotype homozygosity (EHH) test [183] searches for regions that display an excess of long homozygous haplotypes. Integrated Haplotype Score (iHS) is a measure of the amount of EHH at a given SNP along the ancestral allele relative to the derived allele [220]. This statistic identifies signatures of positive selection in the genome. We calculated iHS with R library "rehh" [61, 60] with 271,886 phased SNPs and 1414 individuals in SU.VI.MAX.

Testing association between SNPs and fineSTRUCTURE clusters

The Cochran-Mantel-Haenszel (CMH) test implemented in PLINK 1.07 (option `-mh2`) [172] was used on 271,886 SNPs to scan for SNPs that vary between 6 main fineSTRUCTURE clusters. Manhattan plot was generated in R.

2.1.2 Additional results

Correlation between principal components and geographical coordinates

Several principal components are significantly correlated with longitude and latitude (Table 2.1).

Genetic selection

First, we used the iHS statistic to identify genomic signatures of positive selection in the SU.VI.MAX dataset. Large peak on chromosome 6 - HLA region was identified (Figure 2.2) and smaller peak, but still reaching statistical significance level on the intergenic region of chromosome 11. The latter is not easy to interpret but HLA was identified in several genome-wide scans of selection and linked to adaptation to pathogenic environment. It is interesting to see that HLA signal is easily detectable at the scale of France.

Second, we employed Cochran-Mantel-Haenszel test to scan for SNPs that vary between 6 main clusters identified with fineSTRUCTURE. The largest peak corresponds to lactase locus linked to avoidance of lactose intolerance (Figure 2.3). This locus was also identified in several genome-wide scans, including a study of Western France [99].

	longitude	p-value	latitude	p-value
PC1	0,198 *	6,09E-14	-0,621 *	1,12E-151
PC2	0,545 *	2,98E-110	0,111 *	2,85E-05
PC3	0,031	2,44E-01	0,000	9,86E-01
PC4	-0,008	7,58E-01	0,040	1,28E-01
PC5	0,075 *	4,57E-03	-0,010	7,01E-01
PC6	0,016	5,38E-01	-0,018	4,98E-01
PC7	-0,024	3,73E-01	0,023	3,93E-01
PC8	0,040	1,28E-01	-0,011	6,87E-01
PC9	-0,001	9,80E-01	-0,033	2,14E-01
PC10	0,017	5,16E-01	0,015	5,79E-01
PC11	0,047	7,83E-02	0,079 *	2,79E-03
PC12	-0,005	8,48E-01	-0,009	7,42E-01
PC13	-0,003	8,97E-01	0,053 *	4,75E-02
PC14	-0,015	5,69E-01	-0,027	3,15E-01
PC15	0,000	9,89E-01	-0,012	6,57E-01
PC16	0,094 *	4,25E-04	-0,006	8,28E-01
PC17	-0,021	4,31E-01	0,000	9,90E-01
PC18	-0,021	4,33E-01	0,000	9,96E-01
PC19	-0,026	3,31E-01	0,016	5,41E-01
PC20	-0,014	6,09E-01	0,060 *	2,50E-02

Table 2.1 – Correlation coefficient between PCs and latitude and longitude. Statistically significant values are marked with a star.

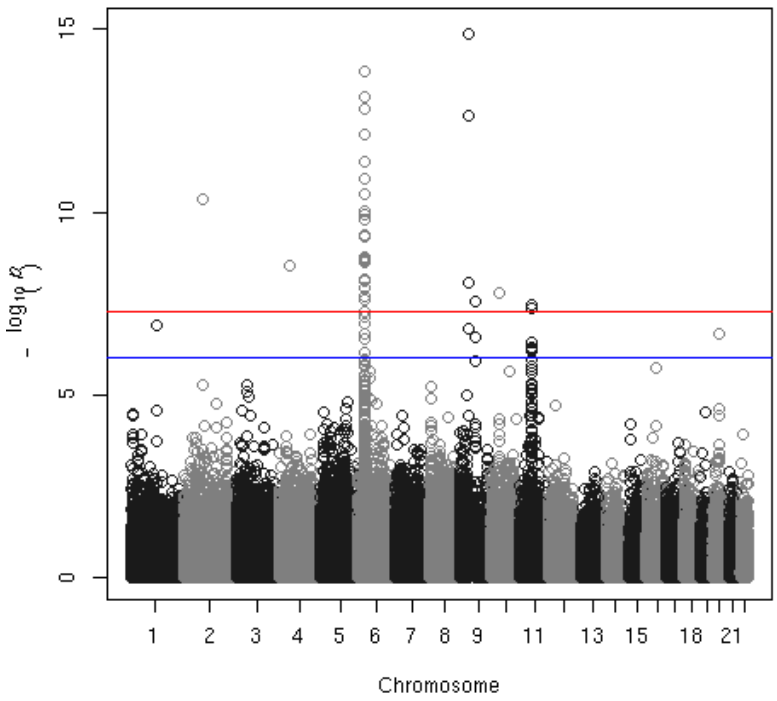


Figure 2.2 – Manhattan plot of iHS statistic on SU.VI.MAX dataset

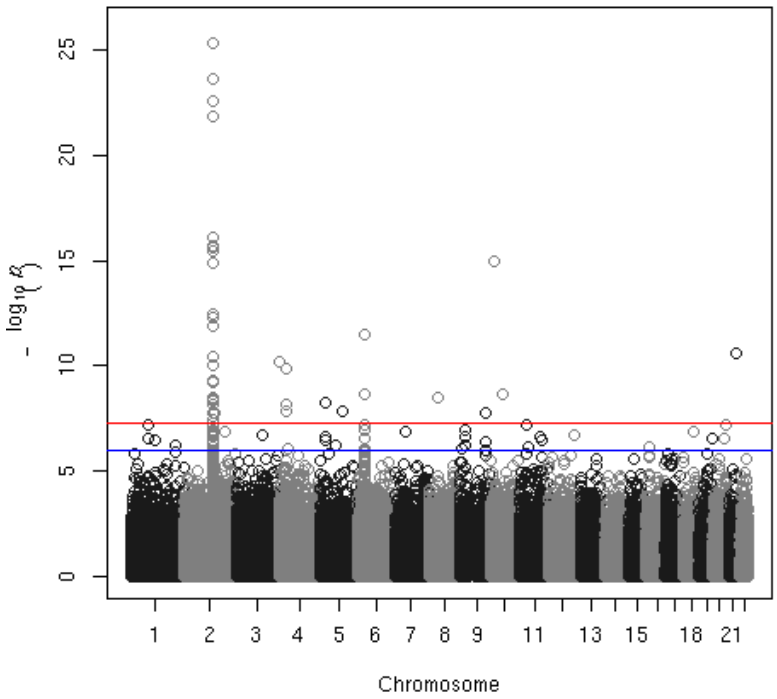


Figure 2.3 – Manhattan plot of Cochran-Mantel-Haenszel test on SU.VI.MAX dataset revealing SNPs that vary between 6 main fineSTRUCTURE clusters

3

The genetic structure of historical populations in Northwestern France

3.1 Introduction

A large part of Western France is a peninsula, delimited by the English Channel to the north, Atlantic Sea to the West and Biscay Bay to the south. Its end is named Finistère (Finis terrae) or Penn Ar Bed (head/end of the world in Breton language) which reflects its position at the northwestern end of European platform. Despite the relative isolation due to its long coast along the north- and western regions, east- and southwards this territory lacks substantial geographical obstacles apart from the longest French river, Loire, which has its estuary in the region. The Armorican Massif covers the territory, but it has been eroded and consists of a low plateaus, with its highest summit being only 417 m above sea level.

Historically, while it has been home of tribes who resisted Caesar Roman conquest, like the Veneti and the Namnetes, this region was finally well Romanised and named Armorica, which is a Latinisation of a Celtic word meaning coastal region. Since IVth century AD, Britons from British Isles pushed by the simultaneous pressure of Saxons and Gaels, immigrated into the Romanised Armorica [52], which then became known as the Little Britain or Brittany. Several kingdoms were assembled in what was to be the Duchy of Brittany. The borders of Brittany, either as independent kingdom or as a French duchy were remarkably constant through the centuries. The Duchy of Brittany was surrounded by the County of Maine and the Duchy of Anjou in the east and County of Poitou in the south.

This large Peninsula was also divided into several languages, with Breton, a Celtic language, being quite different while the others were variations of the old French. The Loire River was, around Xth century, a division between the Langue d'Oïl (influenced by Germanic speaking) and Langue d'Oc (closer to Roman speaking) [22].

Karakachoff et al. have already studied northwestern France and shown that fine-scale genetic structure occurs in the region. The study revealed that Brittany Region is relatively highly differentiated from the neighbouring regions. Also linkage-disequilibrium is higher in Brittany, suggesting a lower effective population size. The departments with the highest local genetic differentiation measured in F_{ST} per 30 km are the departments of Brittany region and additionally the department Vendée. Last but not least, the authors also found genetic proximity between Bretons and Irish, both at the genome-wide level and at the two loci informative about Breton origin, lactase and HLA. However, the study had uneven sampling, ancestry information on the last generation only and didn't use state-of-the-art haplotype-based methods. We address these limitations in this study.

To explore fine-scale structure in northwestern France, we assembled a dataset of 3,234 individuals whose grandparents were born in proximity within a region. We first calculated population diversity and differentiation measures. We examined the relationship between genetic structure and the geographical origin of individuals' grandparents. We investigated also effective migration patterns and effective population size trajectories. To evaluate the impact of neighbouring European countries, we analysed our dataset together with samples from the 1000 Genomes project. Our results provide novel insights into the historical, demographic and cultural events that have shaped the genetic landscape of Western France.

3.2 Materials and Methods

3.2.1 Biocollection PREGO

Project PREGO (*Population de référence du Grand Ouest)*) collected the DNA of 5,707 healthy persons originating from Western France. Individuals were recruited during 295 blood drives organized by the French Blood Service (EFS in French) carried out between February 2014 and March 2017, with a mean of 19 donors per blood drive. Blood drives were spatially and temporally sampled in order to obtain a coverage as homogeneous as possible of 9 departments included in the study. Priority was given to blood drives taking place in rural areas. Participants should be native of western France; inclusion criteria were 4 grandparents born in western France and preferably within a radius of 30 km. Venous blood samples (6ml) were collected from recruited individuals by venipuncture into Vacutainer tubes.

Participants filled out a questionnaire providing grandparents, parents and own birthplaces, residence, age, sex and information about previous participation to the study (of individual itself or another member of the family). Neither phenotypic nor clinical data was collected in present study. Declaration and ethical approval process was achieved in 2013 and involved the Ministry of Research, the Committee of protection of persons (CNN in French), the Advisory Committee on Information Processing for Health Research (CCTIRS in French), and the National Commission on Informatics and Liberty (CNIL in French). Participants signed a written informed consent for participation to the study, inclusion in bioresource and personal data processing.

3.2.2 Samples, genotyping and quality control

At the moment of present study, out of 5,707 collected samples, 3,385 were genotyped on the Axiom™ Precision Medicine Research Array (around 800,000 markers). After removal of related individuals, identified with an PI_HAT ($P(IBD = 2) + \frac{1}{2} * P(IBD = 1)$, probability of two alleles being IBD plus half of probability of one allele being IBD) of 0.08 or above, there were 3,234 samples left for analysis. Their characteristics are shown in Table 3.1 and their spatial distribution in Figure 3.1.

Genotyping was conducted in three batches of respectively 971, 1266, 997 individuals. Batches can create bias in estimation of correlation of differentiation and geographical distance. To investigate potential batch effect, linear regression of batch indicator variable on the first five principal components (see subsection 3.2.5) was employed. No significant association was found at the significance level p-value < 0.05, thus I assume lack of batch effect with respect to relation with geography.

Individuals were assigned the geographical coordinates based on the birthplace of their 4 grandparents (the most common pair of coordinates if possible, otherwise a pair drawn randomly).

3.2.3 LD decay

LD decay was computed using the PopLDdecay software [226]. Default parameter values were used for LD decay calculation, meaning that SNPs with MAF of less than 0.005, or missing sample rate of over 0.25, or heterozygosis ratio of over 0.8 were removed and the maximum distance between two SNPs was 300 kb. LD decay plot was based on the r^2 statistic, which was smoothed in following bin sizes: 100 for

	M	F	T
n	1816	1418	3234
%	56.2	43.8	
Mean age (in 2017)	53.3	49.3	51.5
Age class (in 2017) %			
<35	9.1	16.1	12.2
35-54	38.5	43.7	40.8
>=55	52.4	40.3	47.1
Mean distance between grandparental birthplace			
Median (km)	6.12	6.82	6.38
25% quartile (km)	3.14	3.37	3.25
75% quartile (km)	11.66	13.59	12.33

Table 3.1 – Characteristics and numbers of 3,234 genotyped individuals from PREGO database: age, sex and mean distance between grandparental places of birth.

distance between two SNPs below 100 and 1000 for larger distances. After filtering, 199,528 SNPs were left. To avoid bias that could have been introduced by unequal sample size of the subgroups, I downsampled the data into 1185 individuals in each of the two subgroups Pays de la Loire and Brittany in the analysis per administrative region.

3.2.4 Homozygosity by descent and identity by descent segments

Homozygosity by descent (HBD) and identity by descent (IBD) segments were calculated using RefinedIBD (version from 23rd December 2017) with genetic map build 37. Sum of HBD segments lengths in centiMorgans (cM) per individual was considered. Average values per administrative units were visualized. Two linear regression models were performed in R, in order to examine relation between length of runs of homozygosity (ROH) and geographical coordinates as well as distance to the westernmost commune of Brittany. One covariate in these models is mean distance between grandparents' birthplaces which controls for higher probability of relatedness and thus longer ROH of grandparents born in proximity. IBD segments were divided into 3 subgroups by their length: 1-2 cM, 2-7 cM and longer than 7 cM, corresponding to different periods of demographical history (Figure 3.2). Counts of IBD segments per pair of 32 administrative units (arrondissements) normalised by number of individuals per unit, were used in the analysis. They were visualised as heatmap, with columns and rows reordered using complete linkage clustering. Later counts of IBD segments were visualised as heatmap also for 18 clusters identified with fineSTRUCTURE, however this time without reordering to facilitate comparison with coancestry patterns (section 3.2.6).

3.2.5 Principal component analysis and F_{ST}

PCA was carried out using the smartpca software from the EIGENSOFT package version 6.0.1 [162]. No outliers were detected with the default outlier removal procedure. SNPs in strong LD were pruned out with PLINK 1.9 [34, 172] in a two-step manner. First, the `indep-pairwise` command in PLINK was run with $r^2 = 0.2$, a window size of 50 SNPs and 10 SNPs to shift the window at each step. Second, in order to deal with the remaining long-range LD, SNPs were pruned again for r^2 above 0.2, computed between variants at increased 5Mb distance and in windows of 50 variants. 66,374 SNPs were left for analysis after LD pruning.

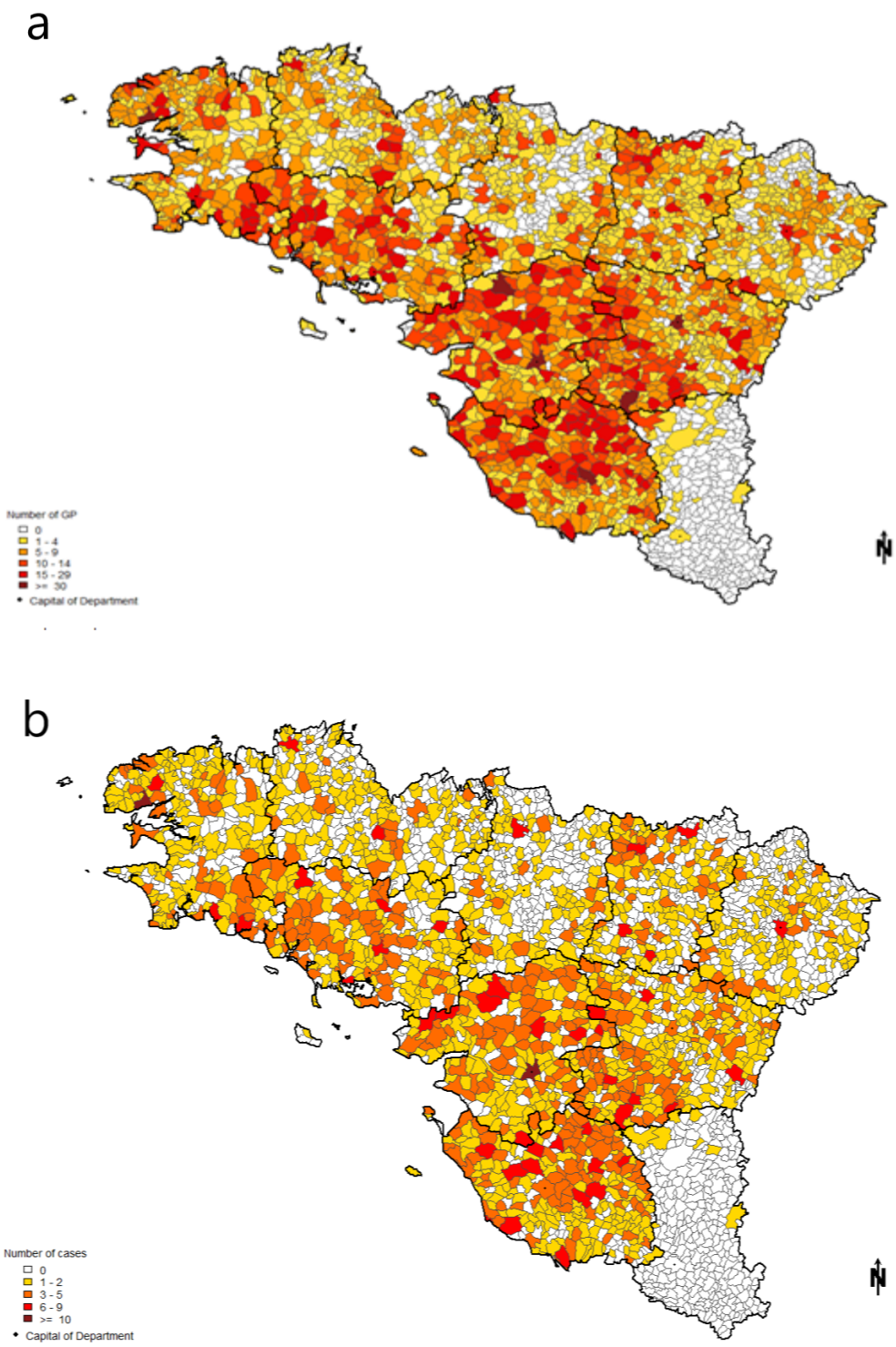


Figure 3.1 – Commune of birth of grandparents of 3,234 PREGO individuals, (a) for all 4 grandparents (b) for each individual, principal (most common) grandparental commune of birth is shown.

To evaluate the geographic relevance of PCs, I tested for the significance of association between the latitude and longitude of each department and PCs coordinates (`cor.test` function in R) using a Spearman's rank correlation coefficient.

F_{ST} values were obtained from smartpca software with option `fsthiprecision`.

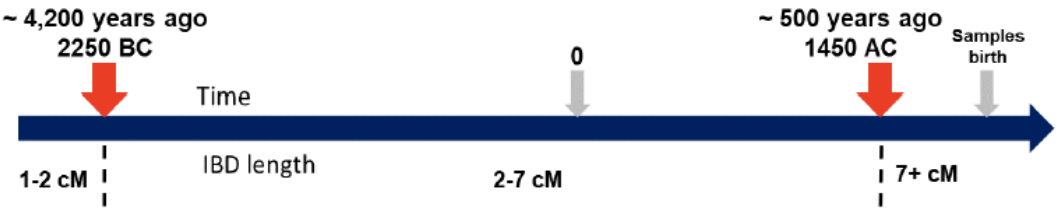


Figure 3.2 – Timeline with estimates for the time to recent common ancestor of an IBD segment of a certain length range. Estimates come from [210] and were obtained with DoRIS software, assuming a generation length of 25 years.

3.2.6 CHROMOPAINTER and FineSTRUCTURE analyses

The genotype data (after quality control (QC)) were phased jointly for all individuals with SHAPEIT v2.r790 [39], without reference panel, with the genetic map build 37 provided with the software. Phased genotype files were converted into CHROMOPAINTER format using the `impute2chromopainter2.pl` script. Global Ne and mutation rates were estimated with CHROMOPAINTER version 2 on chromosomes 1, 4, 10 and 15 and 330 individuals (around 10% of the total sample). Then CHROMOPAINTER was run on full data with estimated parameter values. PCA on its chunkcounts output (coancestry matrix) was performed in R. Coancestry matrix estimates the proportion of the genome of each individual that is most closely related to every other individual in the matrix. In particular, chunkcounts matrix is based on the number of copied haplotype chunks (alternative matrix is based on lengths of the copied chunks). On the same coancestry matrix, FineSTRUCTURE version 2.1.3 was ran with 10,000,000 burn-in iterations, 1,000,000 MCMC iterations from which every 10,000th iteration was recorded, keeping default values of the other options. Tree was built using 100,000 tree comparisons and 10,000,000 additional optimisation steps. MCMC convergence was assessed by comparing the assignment to clusters in a second independent chain.

Confidence of cluster assignment measure and visualisation of coancestry matrix were implemented with the help of the R library `FinestructureLibrary` provided with the software. As an alternative to the tree provided by FineSTRUCTURE software, based on posterior probabilities of population configuration and not a measure of population differentiation, we used a TVD matrix. TVD is a square matrix with number of rows and column equal to number of clusters and its values can be interpreted as a measure of the difference between the two clusters [115]. It was obtained with a script from [101]. TVD matrix was calculated on the final fineSTRUCTURE partition (finest level of FS-tree), then a tree (TVD-based tree) was obtained with complete linkage hierarchical clustering.

3.2.7 EEMS

We estimated an effective migration surface using the software EEMS. The matrix of average pairwise genetic dissimilarities was generated for 66,374 SNPs (pruned dataset) and 3,234 individuals using the `bed2diffs` software included in the EEMS package. Samples were assigned to the nearest of 750 demes (but fewer than 750 demes were observed). Grid of 500 demes was tested too and presented consistent results (data not shown). We run ten independent MCMC chains, each with a random seed, for 1,000,000 iterations, including 990,000 burn-in iterations, thinning every 50 iterations. Next, I chose the chain with the highest final log-likelihood, and started a second round of ten EEMS chains using this chain as a starting point for 1,000,000 additional sampling iterations, thinning every 9,999 iterations. This procedure was repeated three more times, thus the final results origin from 5,000,000 MCMC iterations. Log-posterior trace of ten replicate MCMC chains from the last round shows mixing and convergence of the independent EEMS runs (Figure 3.3). Plots were generated in R using the `rEEMSplots` package.

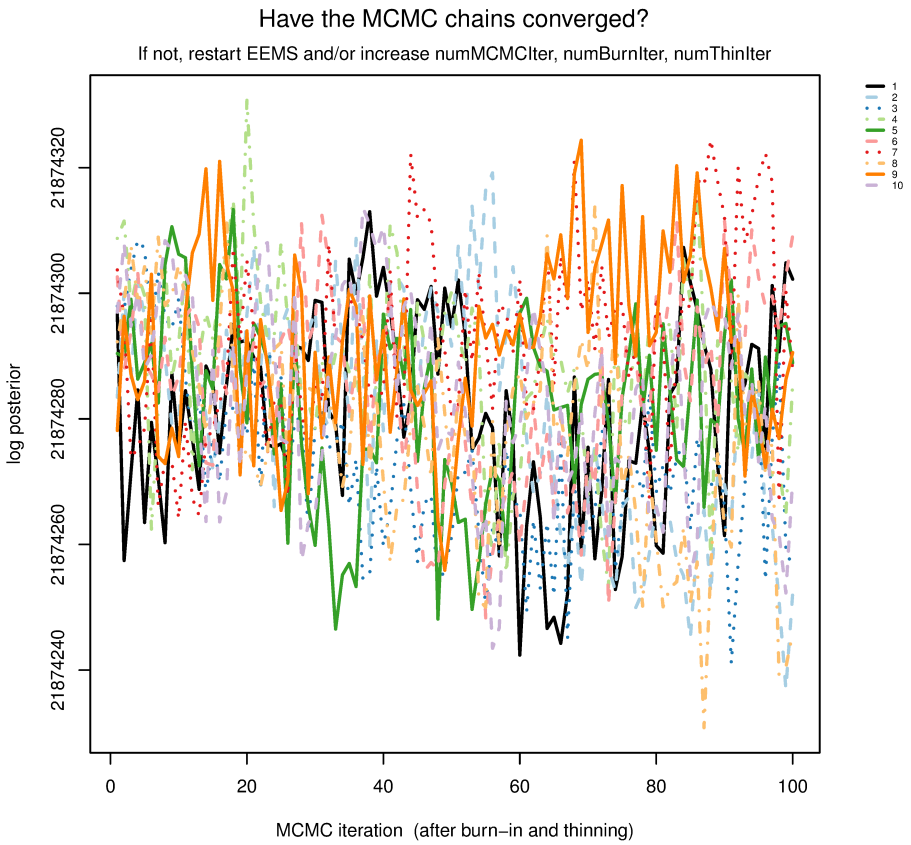


Figure 3.3 – Log-posterior trace of ten MCMC chains from the last round of EEMS analysis. The diagram indicates mixing and convergence of the chains.

3.2.8 IBDNe

We estimated the trajectories of effective population size with IBDNe (version released on 7th May 2018) [23]. IBDseq was used to detect IBD segments (version r1206, with default settings). To identify genomic regions with excess of IBD we followed [23] procedures. Excess of IBD was present on the part of the major histocompatibility complex (MHC) on chromosome 6 (26291527-33464061). This region was thus excluded, resulting in a chromosome 6 split into two continuous parts.

To test the impact of minimum length of IBD segments used by IBDNe on the effective size change trajectories varying `mincm` parameter were tested. Value of `mincm` should be chosen considering SNP density on the SNP array.

A generation time of 30 years was assumed to convert generations into years [190, 50, 140].

A simulation study was performed in the lab to test the ability of IBDNe to retrieve a recent and short bottleneck and test the solutions across varying parameters. 22 chromosomes imitating human genome sequences were simulated under the evolutionary model (Figure 3.4) with ARGON [158]. Chromosome sizes were taken from the human genome GRCh37/hg19 and the genetic maps from 1000G were used. 10 independent simulations were performed in order to test variability of the approach. As in the study on the real data, IBD segments were detected with IBDSeq and effective population size was estimated with IBDNe.

3.2.9 ADMIXTURE with European samples from 1000G

To explore the relationship between our samples and the neighbouring European populations, we merged PREGO data with European samples from Phase 3 1000 Genomes Project (1000G): Finnish (FIN, n = 99), Utah Residents with Northern and Western European Ancestry (CEU, n = 99), British in England and Scot-

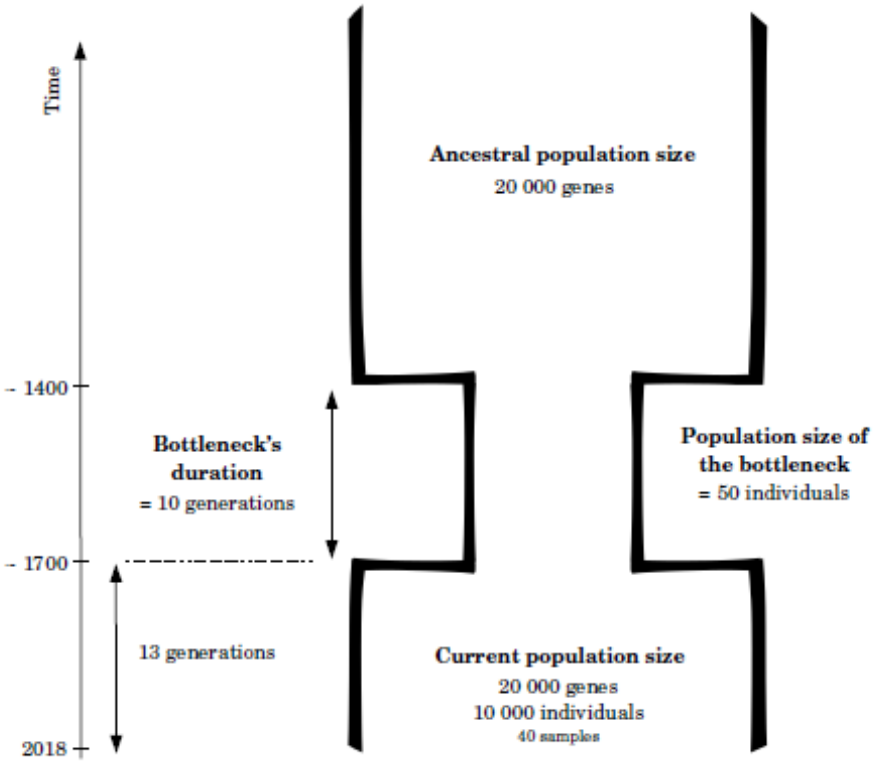


Figure 3.4 – A demographic model of a recent short bottleneck used in the simulation study. Realised by Charlotte BERTHELIER.

land (GBR, n = 91), Iberian Population in Spain (IBS, n = 107), Toscani in Italia (TSI, n = 107). CEU have likely ancestry from British Isles, they have been shown to overlap substantially with GBR on PCA plot [75].

After removal of multiallelic SNPs, SNPs with genotyping rate below 95%, SNPs with minor allele frequency (MAF) below 10%, A/T or C/G SNPs and SNPs not in Hardy–Weinberg equilibrium (HWE) (p-value < 10^{-5}), there were 188,809 SNPs left in merged PREGO and European 1000G dataset. LD pruning was done with analogical procedure as for PREGO dataset alone (subsection 3.2.5) and resulted in final European dataset of 64,317 SNPs which we analysed with ADMIXTURE version 1.23 [3]. We used unsupervised mode with assumed number of ancestral populations K ranging from 2 to 6. For each K, we ran ADMIXTURE 20 times with different seeds to ensure our results are not local optima. For K values 2-5, all 20 runs converged to the same result, while for K = 6, 5 out of 20 runs converged to the other optimum. Ancestry proportions for each K (averaged across multiple runs) were visualised with software pong [12].

3.3 Results

3.3.1 Distribution of diversity across populations

To investigate fine-scale structure in Western France, I assembled genotype data for 3,234 individuals whose grandparents were born within two administrative regions, Brittany and Pays de la Loire. To understand the distribution of population diversity in Western France , we computed measures such as LD decay and length of ROH across the different regions.

LD decay was calculated for two regions (Figure 3.5). Like in [99], LD is higher in Brittany Region than in Pays de la Loire. I corroborated LD decay results with length of runs of homozygosity. Length of runs of homozygosity (ROH) gradually increases towards the end of Brittany Peninsula. We observe that visually at the departmental level (Figure 3.6 a) as well as in the linear regression models. We regressed longitude,

latitude and average distance between grandparents' birthplace on the length of ROH of each individual (Table 3.2). Estimated coefficients suggest that length of ROH and longitude are negatively correlated, thus length of ROH increase westwards. Next we regressed distance between individual birthplace and the westernmost commune in Brittany (Le Conquet) and average distance between grandparents' birthplace on the length of ROH (Table 3.3). Estimated coefficients reveal that the length of ROH increases when the distance to westernmost point of Brittany decreases. Both LD and ROH results suggest smaller effective population size in Brittany. Finer scale of arrondissements for ROH allows to notice local phenomena, in particular near Brest, St. Malo and Cholet (Figure 3.6 b).

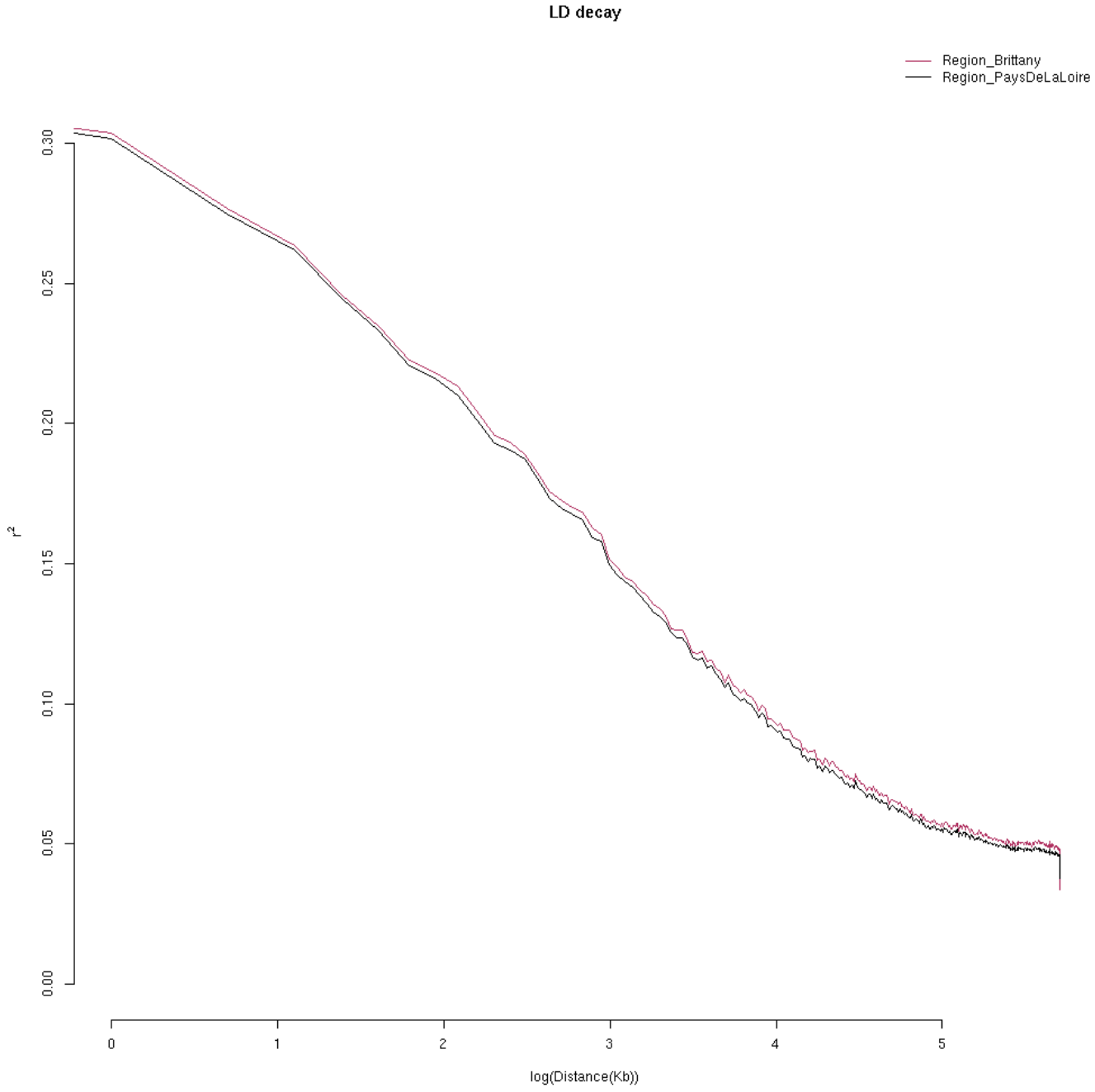


Figure 3.5 – LD decay for two administrative regions of Western France reveals higher LD values in Brittany Region than in Pays de la Loire.

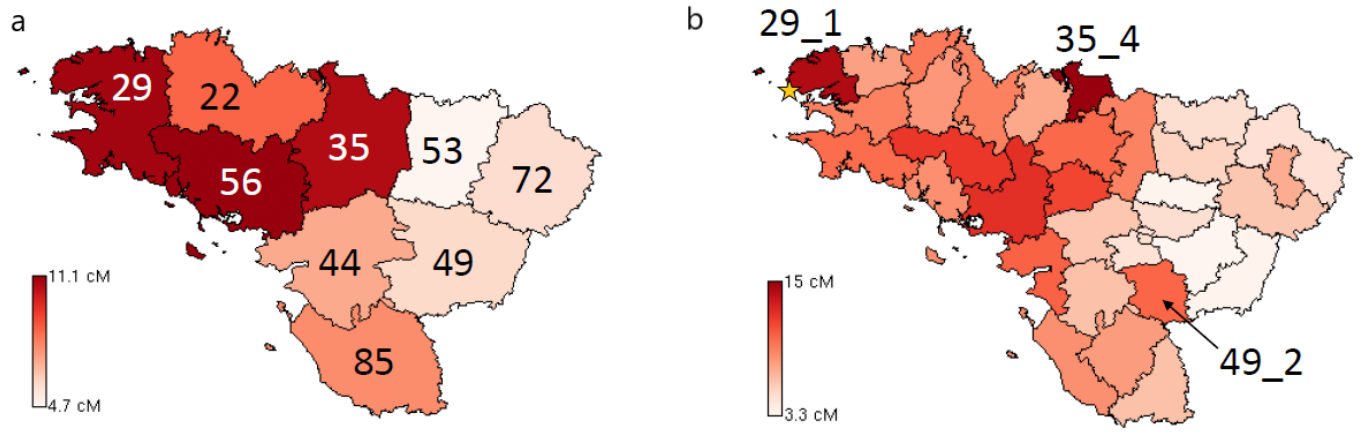


Figure 3.6 – Mean length of runs of homozygosity per (a) department, (b) arrondissement, visualised on the map of Western France. All departments and three arrondissement with the most distinct values are labelled with their administrative numbers. Commune Le Conquet, used in linear regression model, is marked with a star.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.868e+01	4.893e+01	0.791	0.429279
longitude	-2.032e-05	5.331e-06	-3.811	0.000142 ***
latitude	-3.069e-06	7.127e-06	-0.431	0.666830
mean dist GP	-8.500e-05	2.255e-05	-3.769	0.000168 ***

Table 3.2 – Coefficients from linear regression model length ROH $\sim$ longitude + latitude + average distance between grandparents' birthplaces. The length of ROH increases towards West. Mean distance between grandparents' birthplaces was added to regression to control for higher probability of relatedness of grandparents born in proximity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.571e+01	1.380e+00	11.388	<2e-16 ***
dist west brittany	-1.569e-05	4.431e-06	-3.542	0.000405 ***
mean dist GP	-8.535e-05	2.256e-05	-3.783	0.000159 ***

Table 3.3 – Coefficients from linear regression model length ROH $\sim$ distance to westernmost commune of Brittany (Le Conquet) + average distance between grandparents' birthplaces. The length of ROH increases when the distance to westernmost point of Brittany decreases. Mean distance between grandparents' birthplaces was added to regression to control for higher probability of relatedness of grandparents born in proximity.

3.3.2 Measures of population differentiation

Next, population differentiation was assessed by means of F_{ST} as well as IBD sharing. Pairwise F_{ST} values between nine departments are shown in Table 3.7, the average value equals to 0.000484, with a minimum of 0.00018 between Sarthe and Maine-et-Loire and Sarthe and Mayenne and a maximum of 0.00132 between Vendée and Finistère. This indicates that population structure within the region is subtle. On a more granular level the differentiation appears slightly higher: average pairwise F_{ST} between 32 administrative districts (arrondissements) equals to 0.0007, with a minimum of 0.000003 between arrondissements Le Mans and Mamers in Sarthe department and a maximum of 0.001808 between Brest in Finistère and La Roche sur Yon in Vendée. Consistent with an isolation-by-distance pattern, pairwise F_{ST} values increase with physical distance between departments.

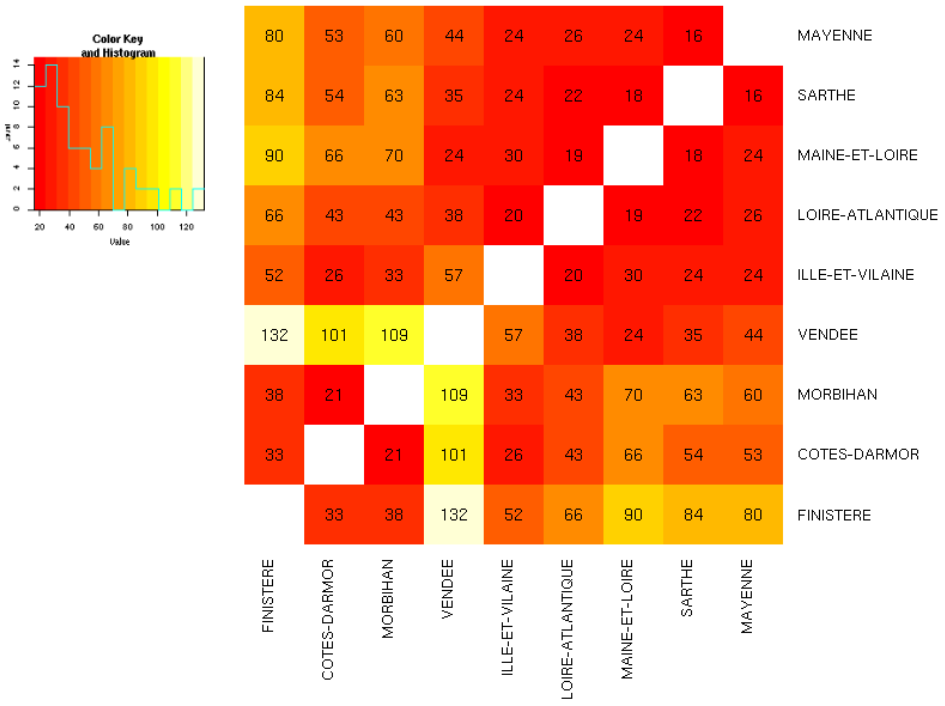


Figure 3.7 – Pairwise F_{ST} values multiplied by 100,000 between 9 departments in Western France

IBD sharing is stronger among the arrondissements of Brittany (Figure 3.8). The area of cluster of high IBD sharing, indicated by first split of hierarchical clustering, decreases with the increase of IBD segment sizes and thus with the more recent period of demographical history (see Figure 3.2). It is worth to note the high value within Cholet arrondissement (49_2) for the most recent period. High levels of IBD sharing and large ROH point to strong bottleneck and/or relatedness in this region.

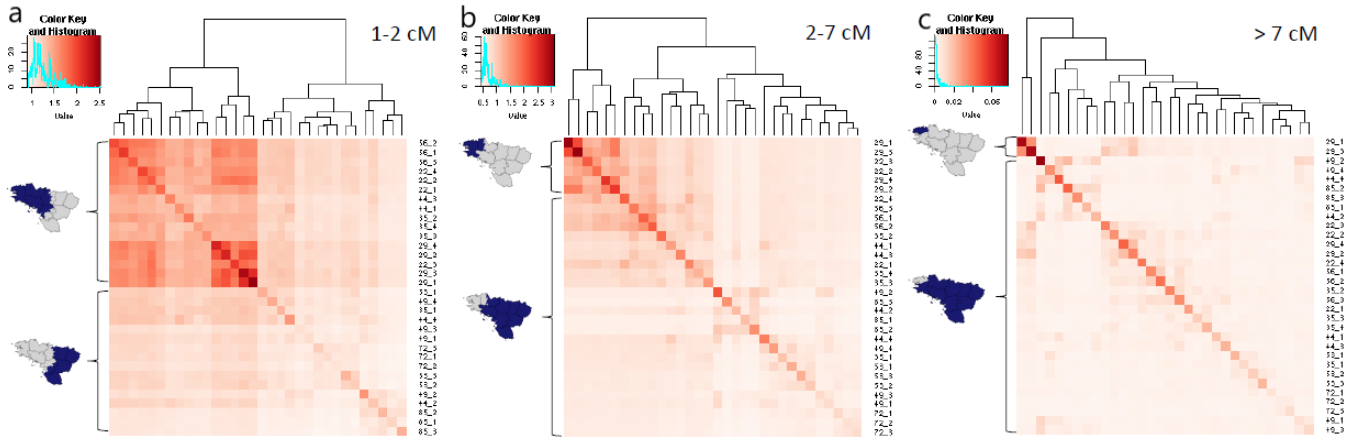


Figure 3.8 – Heatmaps and hierarchical clustering of average number of shared IBD segments. IBD segments were divided into 3 subgroups ((a) 1-2 cM, (b) 2-7 cM, (c) > 7 cM), corresponding to different periods of demographical history (Figure 3.2). Maps highlight the division into “high” and “low” IBD sharing clusters (first split of hierarchical clustering). Map with labelled arrondissements in Figure A.1.

3.3.3 Population differentiation visualization with principal component analysis

I performed PCA analysis on all individuals to reveal gradients of genetic differentiation. PC1, the axis with the largest variance, differentiates between individuals from the northwesternmost department Finistère and the individuals from southern departments, Vendée and Maine-et-Loire (Figure 3.9). Structure

is visible in lower-ranked PCs too (Figure 3.10), in particular Maine et Loire individuals (colored orange) cluster together in PC4. The visual pattern is consistent with pairwise F_{ST} values reported in Figure 3.7 in previous subsection. Genetic variation closely mirrors geography, which is even more visible on the profiles of single principal components on the map (Figure 3.11). Several principal components are significantly correlated with longitude and latitude (Table 3.4).

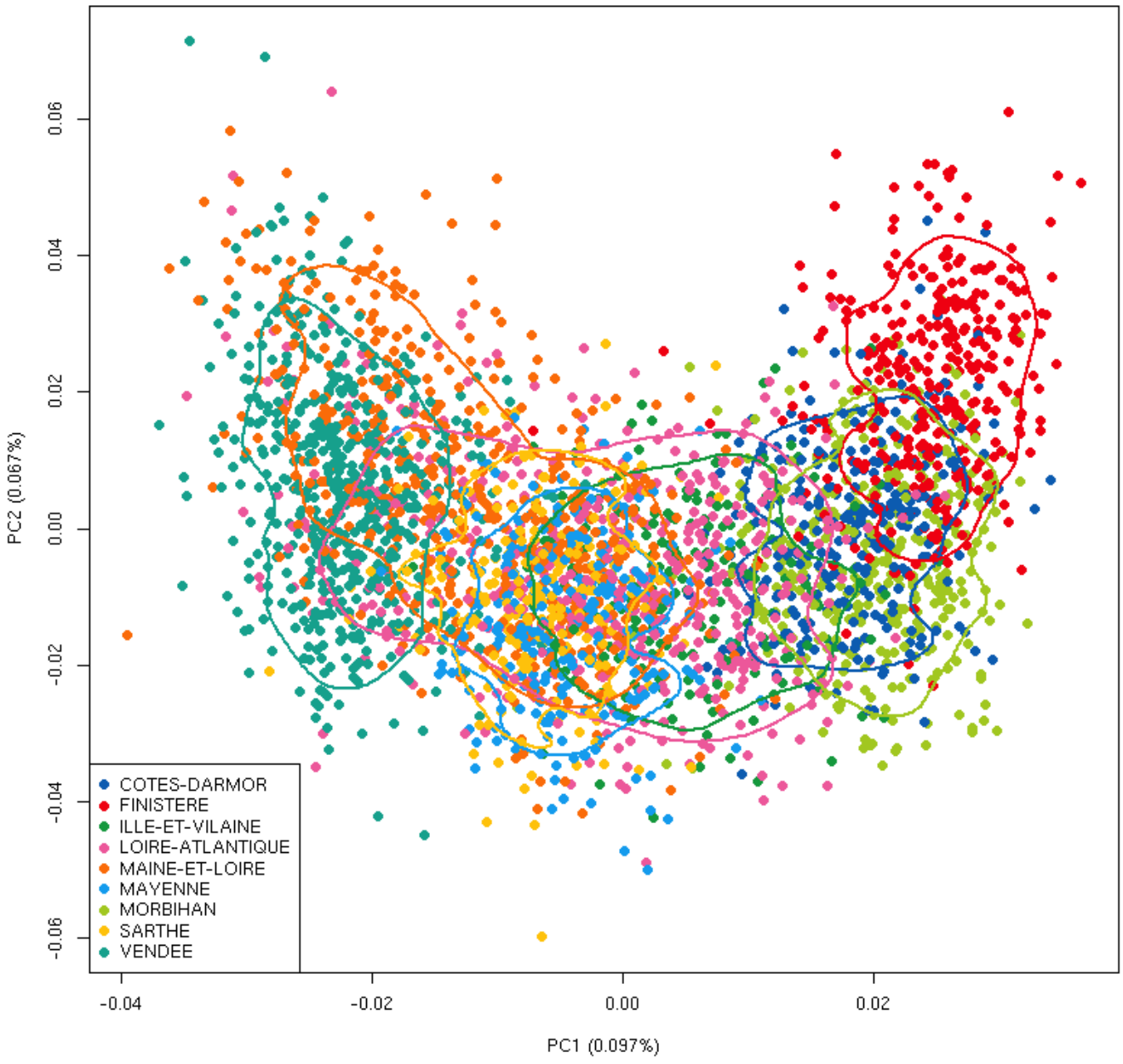


Figure 3.9 – First two PCs from the PREGO dataset reveal population differentiation. Points represent individuals, color-coded by the department of their grandparents's birth. Lines encircle 75% of the individuals from a given department.

	cor longitude	p-value	cor latitude	p-value
PC1	-0,717 *	0	0,762 *	0
PC2	-0,313 *	2,1625E-74	-0,035	0,04382246
PC3	-0,027	0,1287241	-0,030	0,08479562
PC4	-0,104 *	3,0486E-09	0,047	0,0076099
PC5	-0,127 *	4,1818E-13	-0,165 *	3,6102E-21
PC6	0,071 *	5,3411E-05	0,085 *	1,154E-06
PC7	0,025	0,1527141	0,015	0,39215
PC8	0,014	0,413567	-0,005	0,7622401
PC9	0,003	0,8745379	0,017	0,3416038
PC10	0,052	0,00286645	0,047	0,007292
PC11	-0,026	0,1452975	-0,067 *	0,00013864
PC12	-0,006	0,7440231	-0,004	0,8157205
PC13	-0,019	0,2797328	-0,049	0,00551954
PC14	0,011	0,5485341	-0,007	0,7016348
PC15	0,001	0,9431302	0,015	0,4025327
PC16	-0,013	0,4623118	-0,007	0,7012743
PC17	-0,046	0,00923946	-0,050	0,00457652
PC18	0,032	0,07129243	-0,019	0,2881187
PC19	-0,049	0,00578493	-0,030	0,09198998
PC20	-0,026	0,1443339	0,025	0,1520481

Table 3.4 – Correlation coefficient between PCs and latitude and longitude. Statistically significant values are marked with a star, significance level after Bonferroni correction for multiple hypothesis testing is 0.00125.

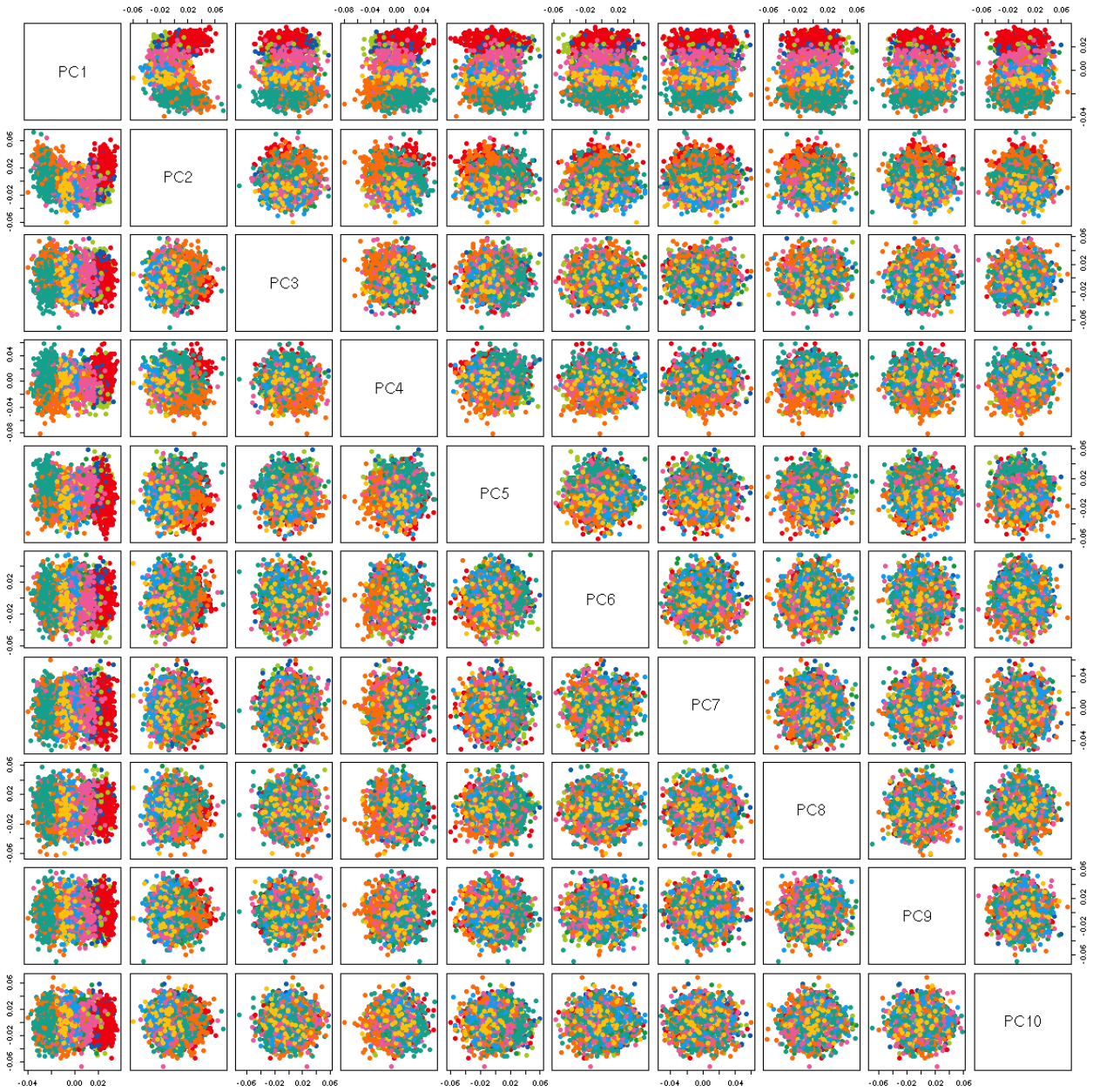


Figure 3.10 – First 10 PCs from the PREGO dataset. Structure is visible in lower-ranked PCs too, in particular Maine et Loire individuals (colored orange) cluster together in PC4. Points represent individuals, color-coded by the department of their grandparents's birth, analogically to Figure 3.9.

3.3.4 Principal components analysis on the coancestry matrix from CHROMOPAINTER

Next, we verified if PCA based on a coancestry matrix performs better than allele frequency-based PCA (Figures 3.12, 3.13). The coancestry matrix contains information not only from allele frequencies but also from SNPs in LD and this is likely the reason behind an increased resolution in the genetic differentiation observed. While profiles of independent SNP-based single PCs present a pattern uncorrelated with geography for lower-ranked PCs (PC3, PC6), all coancestry-based PCs I attempted to plot mirror geography (PC1-PC6 in Figure 3.14, PC7-PC12 in Figure A.2). Many PCs are significantly correlated with longitude and latitude (Table 3.5). Given these results, I anticipate that haplotype-based fineSTRUCTURE method will reveal a very fine-detailed population structure of the region.

	cor longitude	p-value	cor latitude	p-value
PC1	0.481 *	3.612306E-187	0.119 *	1.214851E-011
PC2	0.683 *	0	-0.804 *	0
PC3	0.331 *	1.118419E-083	-0.011	0.5225068
PC4	0.143 *	2.513834E-016	0.265 *	3.699855E-053
PC5	0.007	0.7023866	0.034	0.0560626
PC6	-0.042	0.01584871	0.081 *	3.604801E-006
PC7	0.086 *	9.342522E-007	0.121 *	4.45043E-012
PC8	-0.084 *	1.75649E-006	-0.263 *	2.452582E-052
PC9	0.051	0.004021888	-0.083 *	2.236557E-006
PC10	0.054	0.002050244	-0.156 *	5.905616E-019
PC11	-0.018	0.3017439	-0.169 *	4.229465E-022
PC12	0.003	0.8838014	0.032	0.0655449
PC13	-0.022	0.2113882	-0.045	0.01093632
PC14	0.016	0.3607097	0.011	0.5286654
PC15	0.040	0.02146651	-0.024	0.1652255
PC16	0.027	0.1211359	0.043	0.01340376
PC17	0.042	0.01597168	-0.012	0.4882627
PC18	-0.060 *	0.0006593648	-0.078 *	9.84081E-006
PC19	0.043	0.01374093	0.046	0.008753665
PC20	0.096 *	4.978608E-008	-0.024	0.1657228

Table 3.5 – Correlation coefficient between PCs on the coancestry matrix from CHROMOPAINTER and latitude and longitude. Statistically significant values are marked with a star, significance level after Bonferroni correction for multiple hypothesis testing is 0.00125

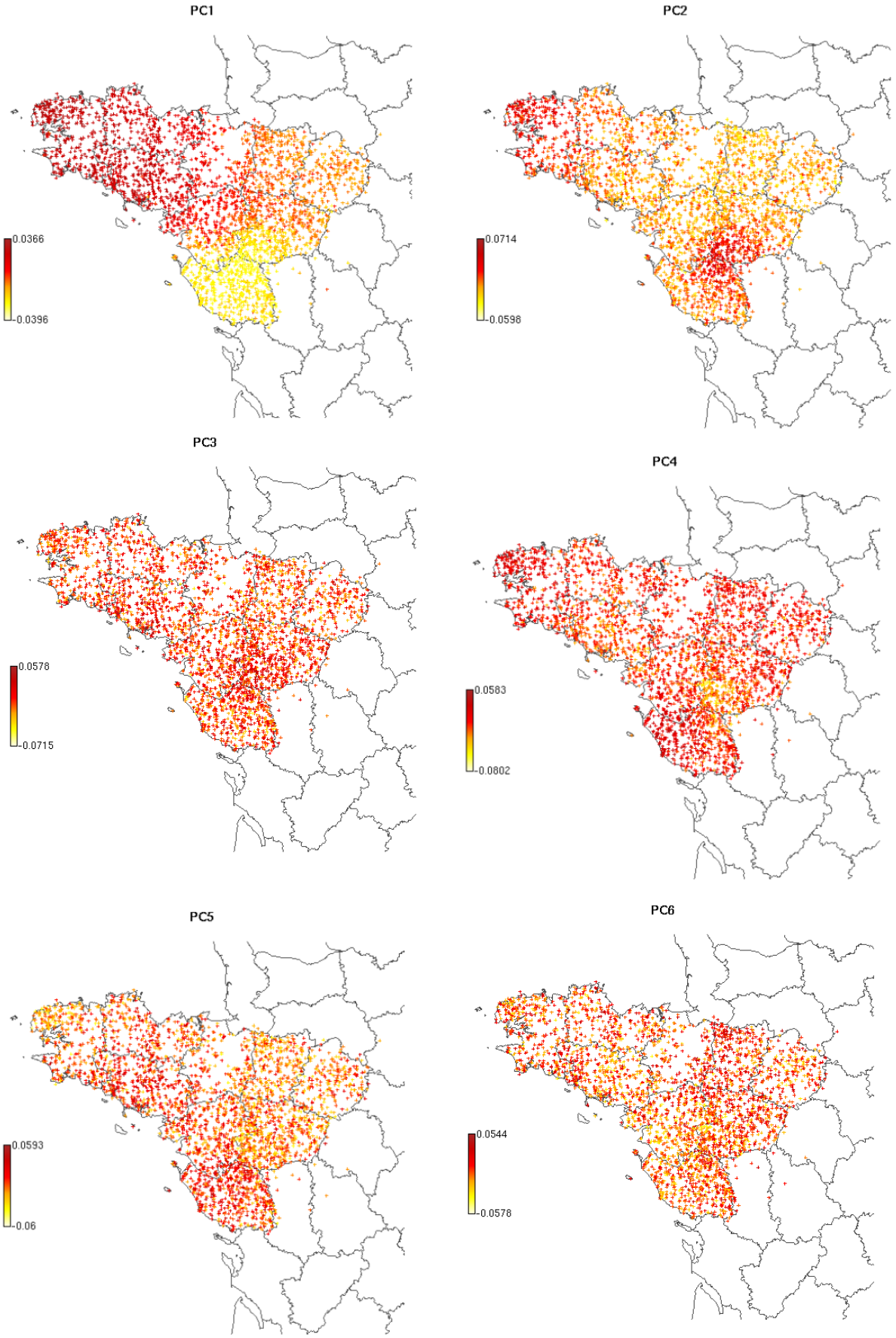


Figure 3.11 – Profiles of first six PCs on the map of Western France reveal link between genetic variation and geography. PC3 is not correlated with geographical coordinates. Coordinates of individuals correspond to their grandparents' birthplaces

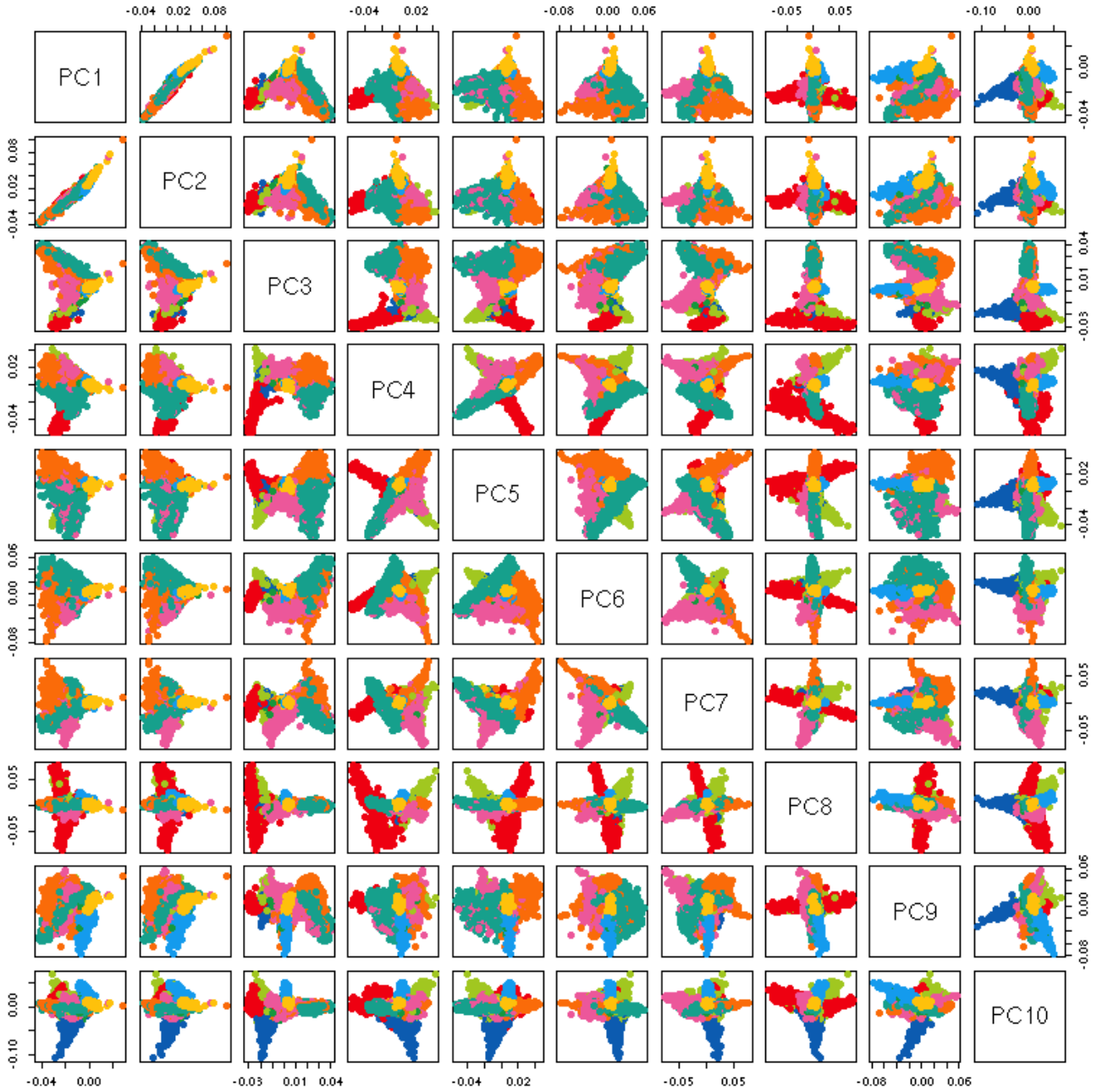


Figure 3.12 – First 10 PCs from the coancestry matrix obtained from the PREGO dataset. Patterns of genetic differentiation are captured better than with the independent SNP-based PCA. Points represent individuals, color-coded by the department of their grandparents's birth, analogous to Figure 3.9.

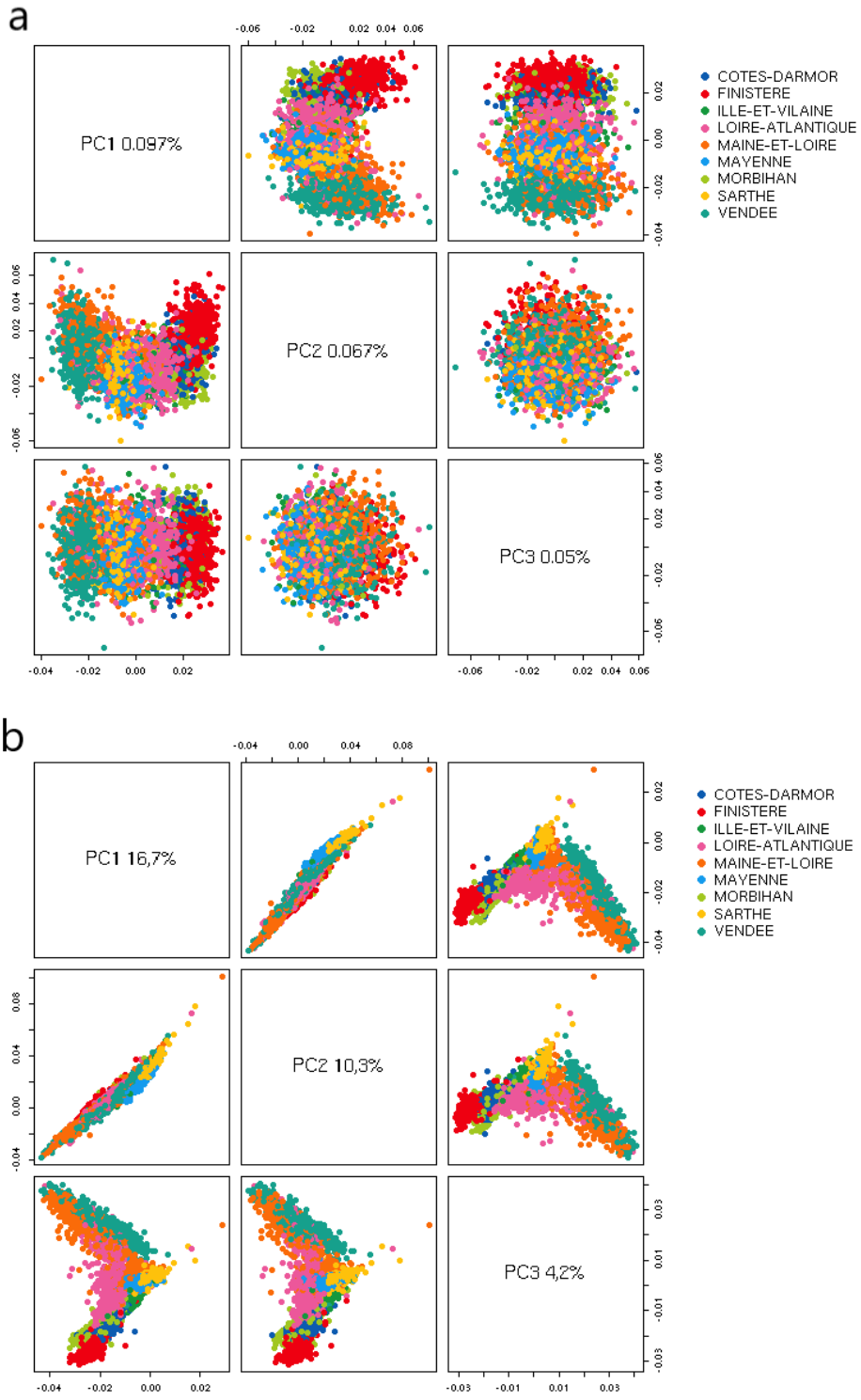


Figure 3.13 – Comparison of first 3 PCs between (a) allele frequency-based PCA and (b) coancestry matrix-based PCA. Coancestry matrix-based PCA captures genetic differentiation better than with the allele frequency-based PCA. Points represent individuals, color-coded by the department of their grandparents's birth, analogous to Figure 3.9.

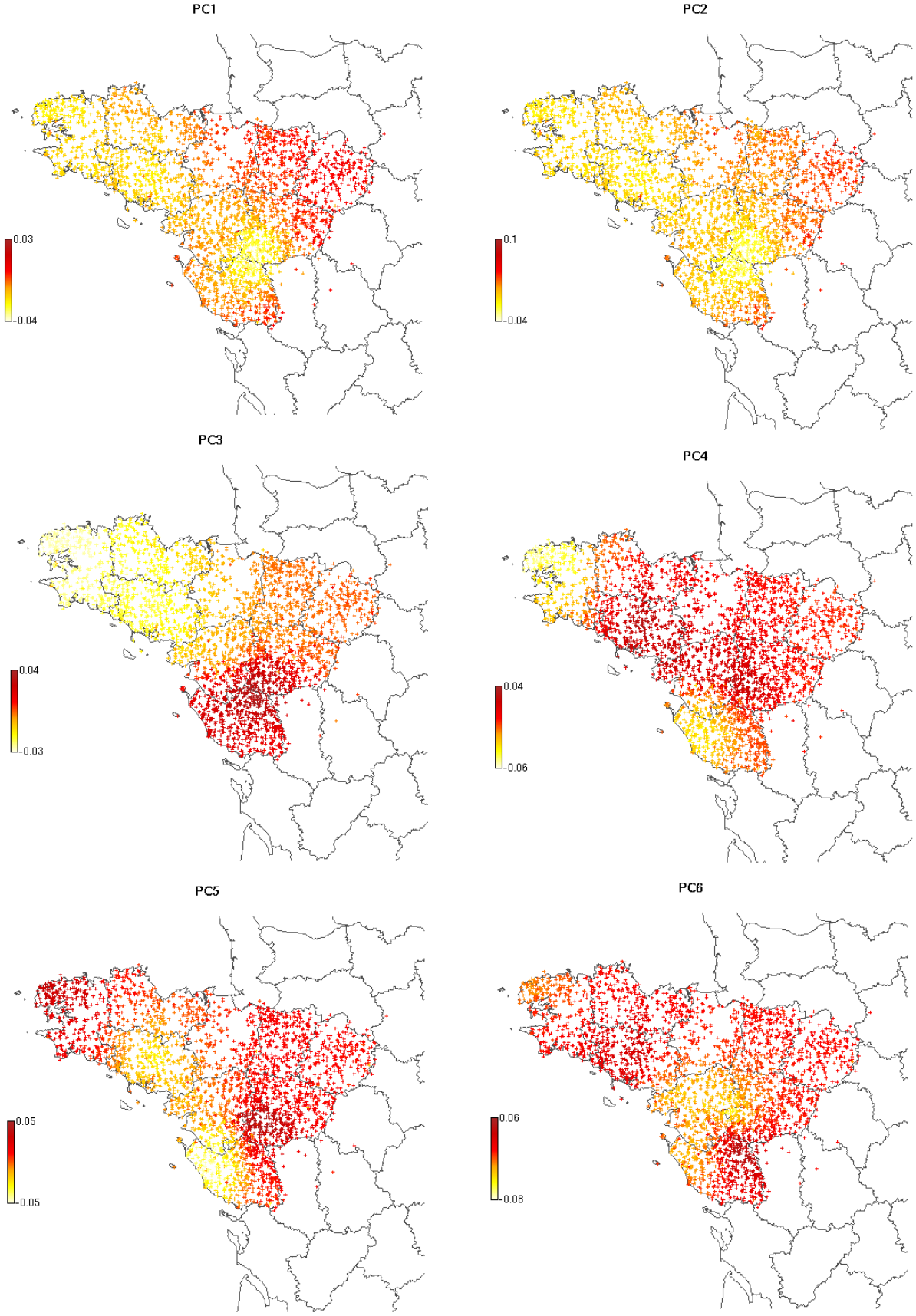


Figure 3.14 – Profiles of coancestry-based PC1-PC6 on the map of Western France. Coordinates of individuals correspond to their grandparents' birthplaces

3.3.5 Fine-scale population structure in western France

To explore in more detail the fine-scale structure in western France, I used state of the art haplotype-based method fineSTRUCTURE. It divided 3,234 samples into 154 clusters at the finest level of its clustering (Figure 3.15). However, here I focus on the coarser levels of 3 and 18 clusters (Figures 3.16 and 3.17). I use these two divisions for downstream analysis. The choice of 18 clusters was not arbitrary, it is explained in subsection 3.3.6. Figure 3.17 shows the hierarchical relationship between the 18 clusters inferred by fineSTRUCTURE(Figure 3.17). Clusters correspond to geography in a striking manner: they are highly localised and regions of overlap are relatively small. As the dataset effectively contains information about the individuals' grandparents that were born in early 1900s, the spatial distribution of genetic structure would reflect that of western France around that time.

The first split of the tree of hierarchical relationship between populations coincides closely with a geographical feature, the Loire River, separating individuals located on the south of it from the rest. The next level separates the majority of the Brittany peninsula (Figure 3.16) and the resulting limit is close to the Loth line (Figure 3.18), the limit associated with Breton toponymy, as well as limit between spoken Breton and Gallo dialects. At the level of 18 clusters (Figure 3.17), clusters *Leon*, *Cornouaille* and *Vannes* align with Breton dialects (Figure 3.19). Individuals located in north-eastern part of the regions (departments: Sarthe, Mayenne and northern part of Maine-et-Loire) remain in one relatively large cluster *Maine-Anjou*, whereas there are numerous smaller clusters in the southern part of the studied region. In particular, the strongest sub-structure is observed in the region Mauges, between the Loire and the Sèvre Nantaise rivers. This differential pattern might be related to demographical history, it corroborates the results of runs of homozygosity analysis (Figure 3.6), in particular the relatively long runs of homozygosity in Cholet arrondissement in Mauges. The aforementioned clusters will be investigated further with patterns in the coancestry matrix in subsection 3.3.8.

Average pairwise F_{ST} between 18 identified clusters equals to 0.00095, with a minumum of 0.00022 between clusters *Maine-Anjou* and *Sèvres* and a maximum of 0.00239 between clusters *Leon* and *Mauges* (all pairwise values presented in Figure 3.20), it is thus larger than F_{ST} between administrative districts, indicating that the obtained clustering reflects population differentiation better than administrative division. The continental France study (Chapter 2) found larger average pairwise F_{ST} between French clusters in 3C dataset (0.00154), but slightly smaller in SU.VI.MAX dataset (0.0009). Difference between clusters can be alternatively measured with TVD, the matrix of pairwise TVD values is presented in Figure 3.21.

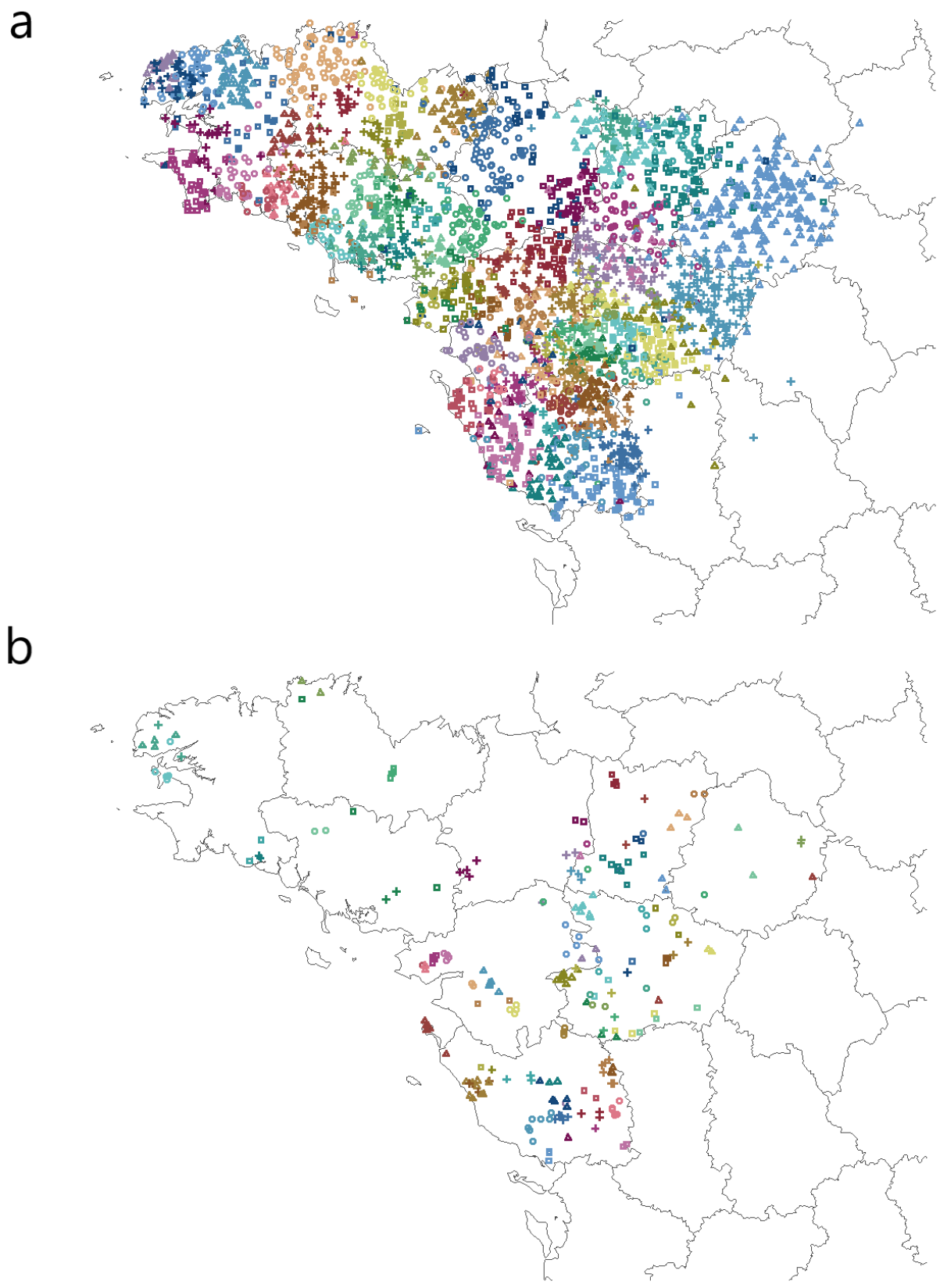


Figure 3.15 – All 154 clusters in Western France identified with fineSTRUCTURE. (a) 78 clusters with large sample sizes (at least 12 individuals), (b) remaining 76 clusters with low sample sizes (maximum 10 individuals). Coordinates of individuals correspond to their grandparents' birthplaces

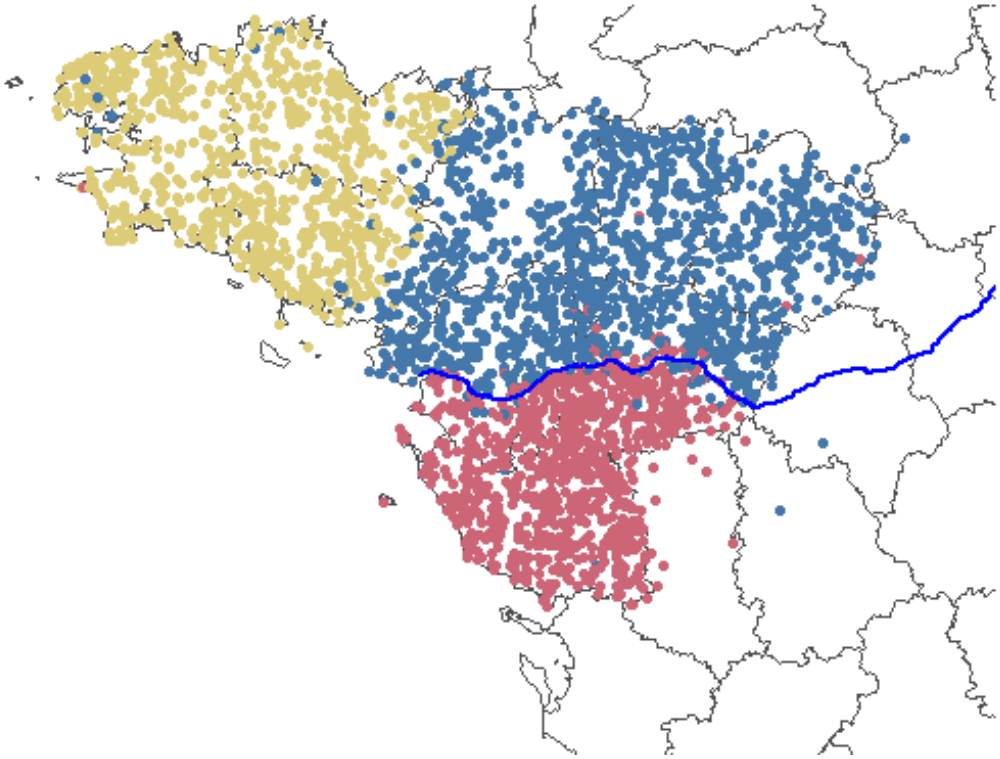


Figure 3.16 – 3 main clusters in Western France identified with fineSTRUCTURE. Coordinates of individuals correspond to their grandparents' birthplaces

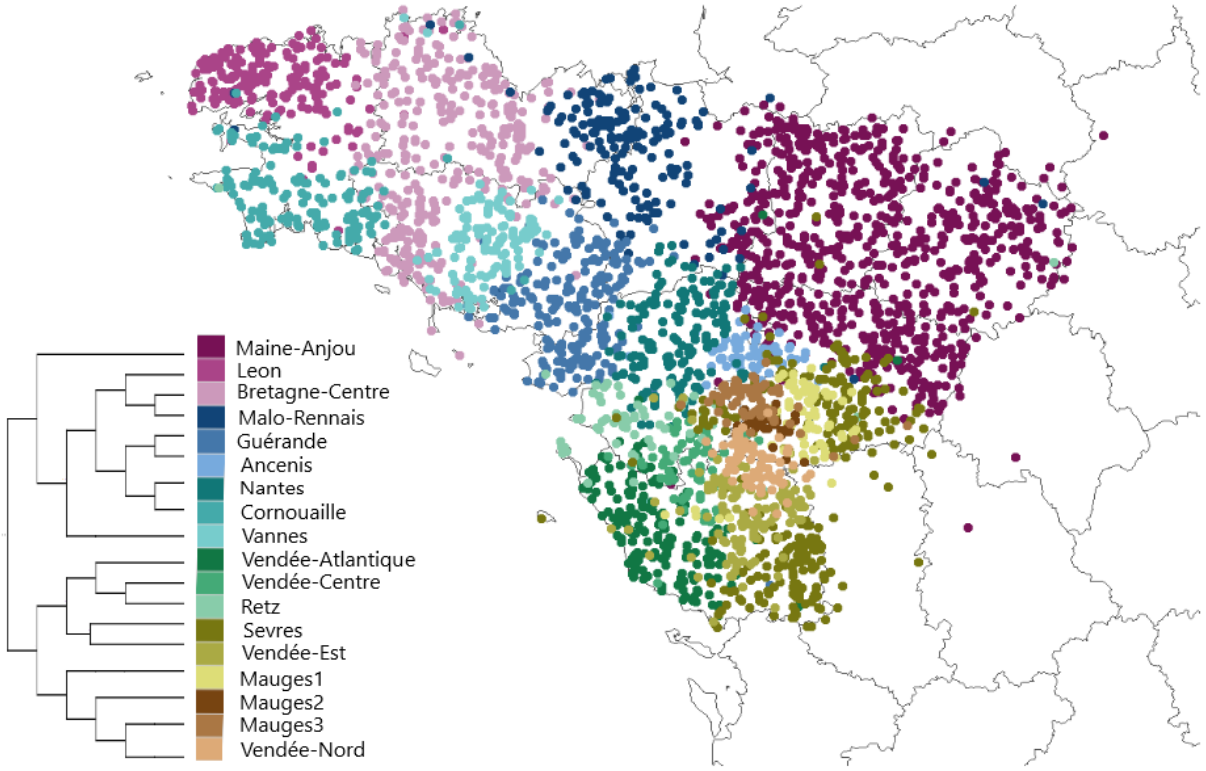


Figure 3.17 – 18 clusters in Western France identified with fineSTRUCTURE. Coordinates of individuals correspond to their grandparents' birthplaces. Tree illustrates hierarchical reliationships between clusters.

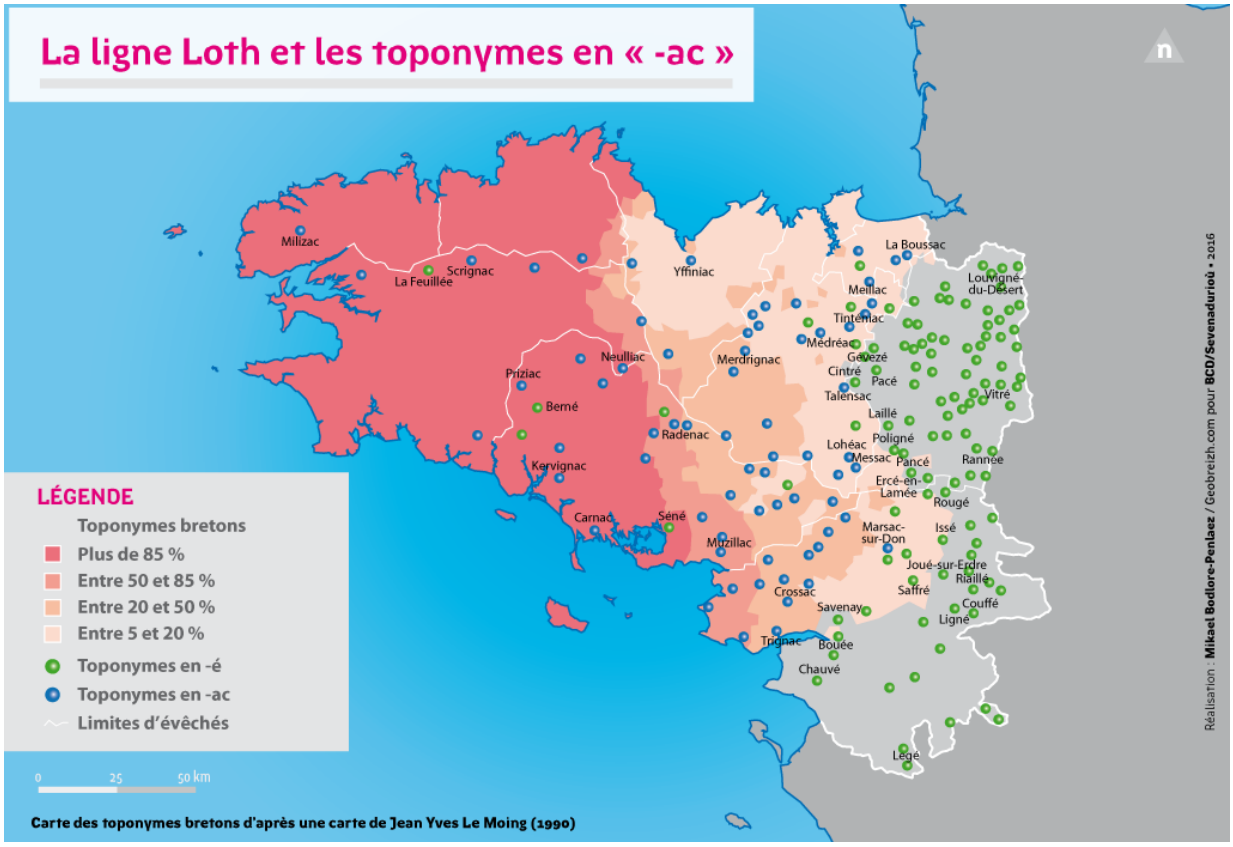


Figure 3.18 – A map of Breton toponymy and Loth line. Loth line separates "-ac" toponyms from "-é" toponyms. Source: [106], but based on the map from [113].

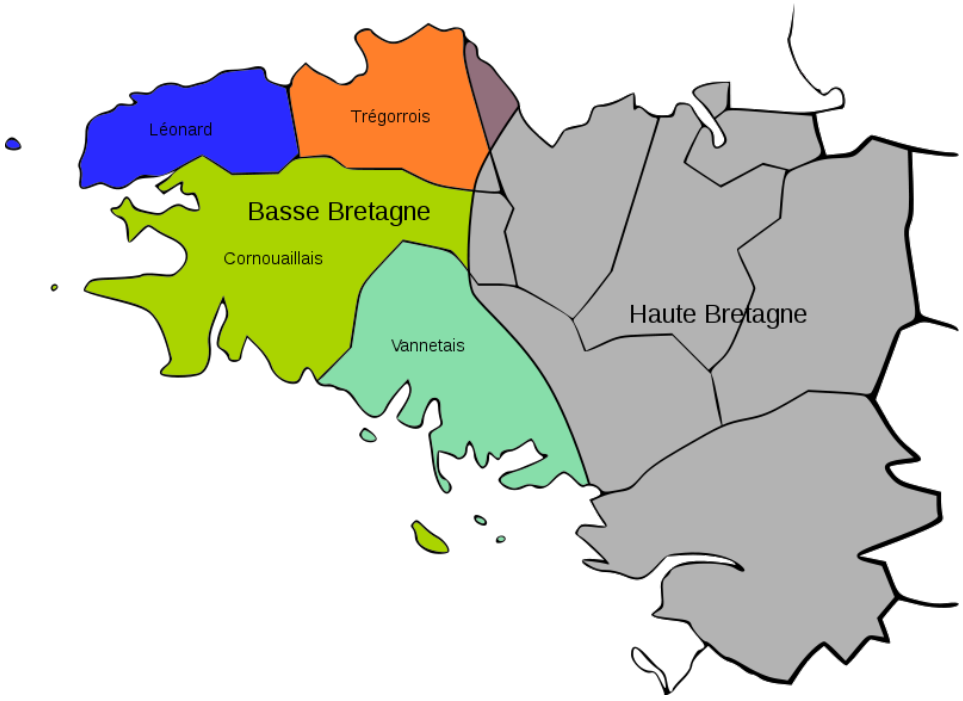


Figure 3.19 – A map of dialects of Breton language. Source: Wikipedia.

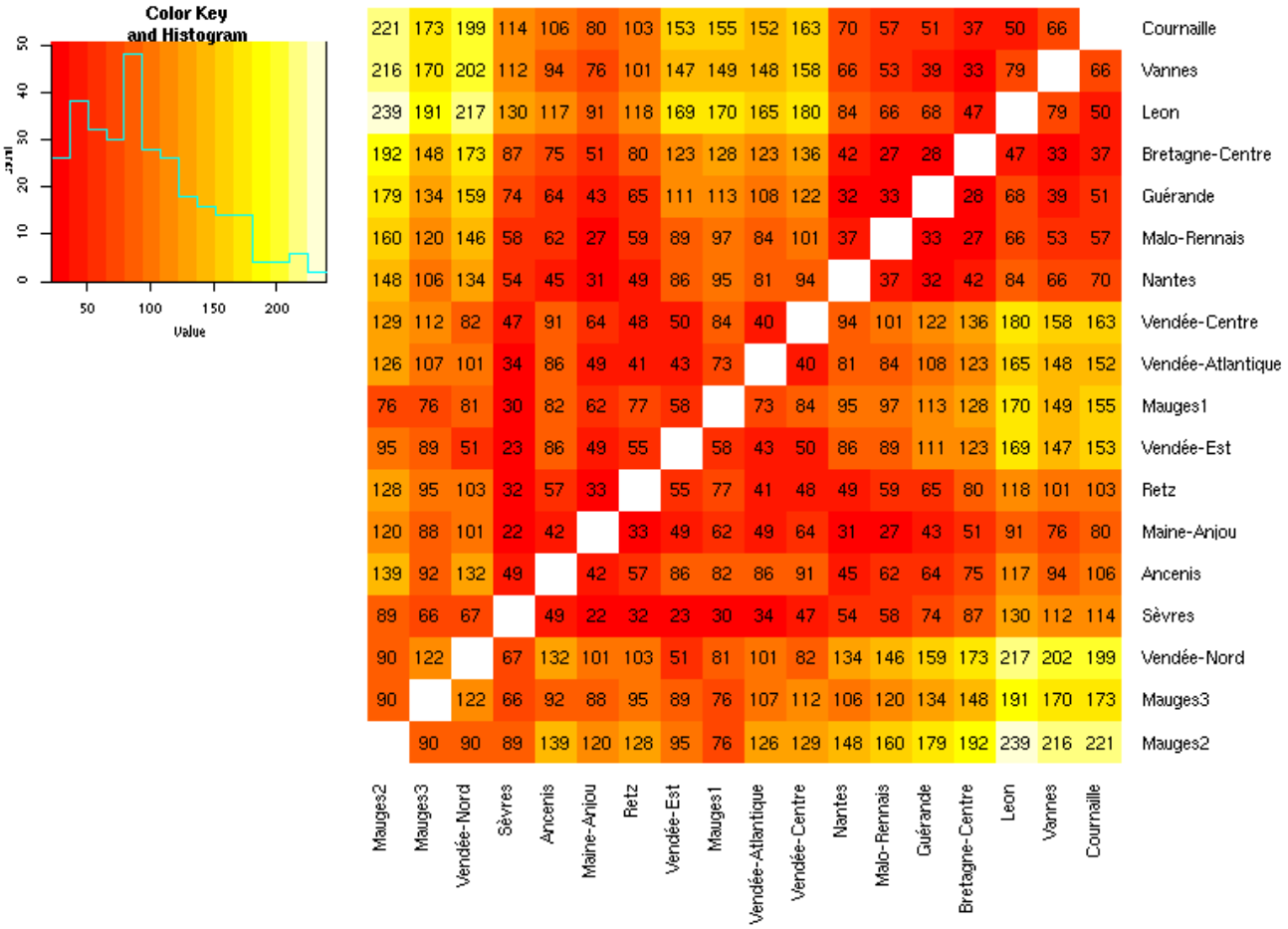


Figure 3.20 – Pairwise F_{ST} values multiplied by 100,000 between 18 clusters in Western France identified with fineSTRUCTURE.

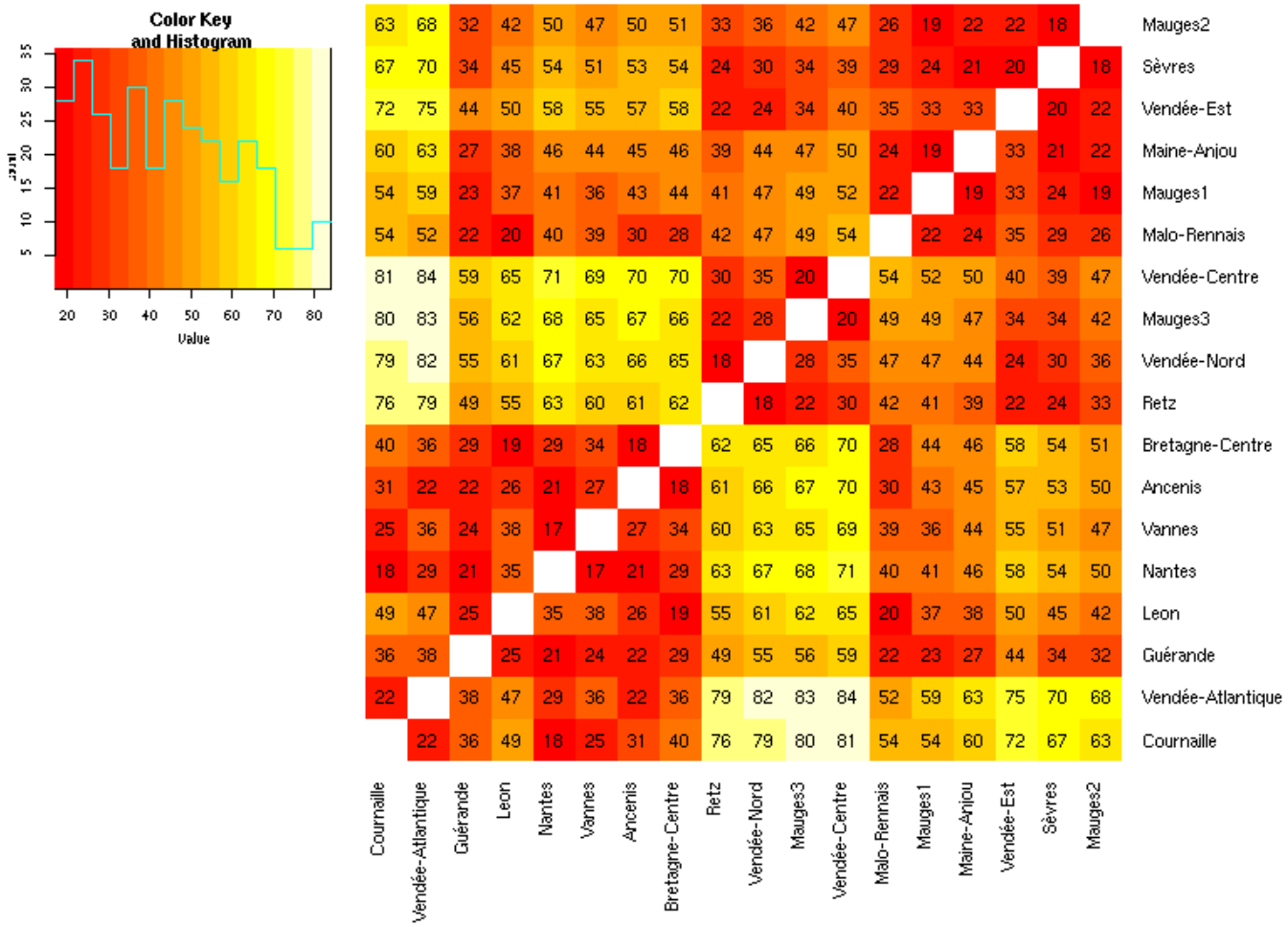


Figure 3.21 – Pairwise TVD values multiplied by 1,000 between 18 clusters in Western France identified with fineSTRUCTURE.

3.3.6 The choice of 18 clusters

The hierarchical tree returned by fineSTRUCTURE (FS-tree) is based on posterior probabilities of population configuration, not a measure of genetic differentiation, and it has been shown to depend significantly on the sample size [108]. I have tested an alternative approach based on total variation distance (TVD-based tree) that has been shown to provide more consistent results [101]. I assessed the performance of these two tree building approaches on the PREGO dataset by measuring the confidence of cluster assignment according to [115]. The value of confidence of cluster assignment decreases with the number of clusters. I found that FS-tree has a lower confidence of cluster assignment than TVD-based tree for the same numbers of clusters (Figure 3.22), for example the confidence of cluster assignment is similar for 12 clusters in FS-tree and 39 in TVD-based tree, equal to 0.934, whereas for 12 clusters in TVD-based tree confidence of cluster assignment is as high as 0.962. It indicates better performance.

However, it is important to mention one aspect by which trees resulting from these two approaches differ. While FS-trees separate relatively evenly sized groups at the top of the hierarchy, top splits of TVD-based trees regularly disconnect single, seemingly outlier individuals from the rest. I chose the level of 39 clusters in the TVD-based tree and decided to examine the outlier individual clusters. There were 21 of them, I proceeded as follows: I either joined an outlier individual to the closest cluster of normal size (17 individuals) or removed it if it was not possible to determine the closest cluster (4 individuals). In this way I dealt with the problem of single-individual clusters and obtained a TVD-based tree with 18 clusters, still performing better than FS-tree which achieves similar value of confidence of cluster assignment for only 12 clusters.

Following the approach described in [115], I tested whether inferred clusters capture significant differences in ancestry. I randomly reassigned the individuals to the clusters, maintaining the cluster sizes to obtain a new division into clusters and to test the hypothesis that, given the cluster sizes, the individuals are assigned randomly to each cluster. I repeated the process 1000 times to obtain p-values from the permutation test. P-values are the ratio of the number of permutations where random assignment resulted in higher value of TVD between any pair of clusters to the total number of permutations. For the level of 18 clusters, p-values were < 0.001 for all pairs of clusters.

3.3.7 The resulting clustering is not an artefact of the sampling scheme

I performed analyses to test whether the clustering could be an artefact of our sampling scheme. I compared the distribution of pairwise IBD statistic within the finest level of 154 clusters to its distribution across the whole sample (Figure 3.23) as well as on the coarser level of clustering (Figure 3.24 for 18 clusters and Figure 3.25 for 3 clusters). This statistic reflects shared relatedness in the recent generations in the past. Although the distribution within clusters are slightly shifted towards higher IBD values compared to the whole sample, those shifts are relatively small for reasonably-sized clusters, indicating that observed clusters are not an artefact of a sampling scheme. Out of 78 clusters with 12 or more individuals, only five have median pairwise IBD statistics exceeding the third quartile from the distribution of the statistic across the whole sample (Figure 3.23 a). For the coarser levels of 18 clusters the differences between distributions for clusters and across whole sample are even smaller, with median pairwise IBD statistics not exceeding third quartile from the distribution of the statistic across the whole sample. Interestingly, at the lever of granularity of 3 clusters I observe a slight shift for cluster associated with Brittany Peninsula which may be linked to smaller effective population size in that area. The distributions of the remaining two populations, South and North-East, have their interquartile ranges visually identical to the ones from the distribution from the whole sample.

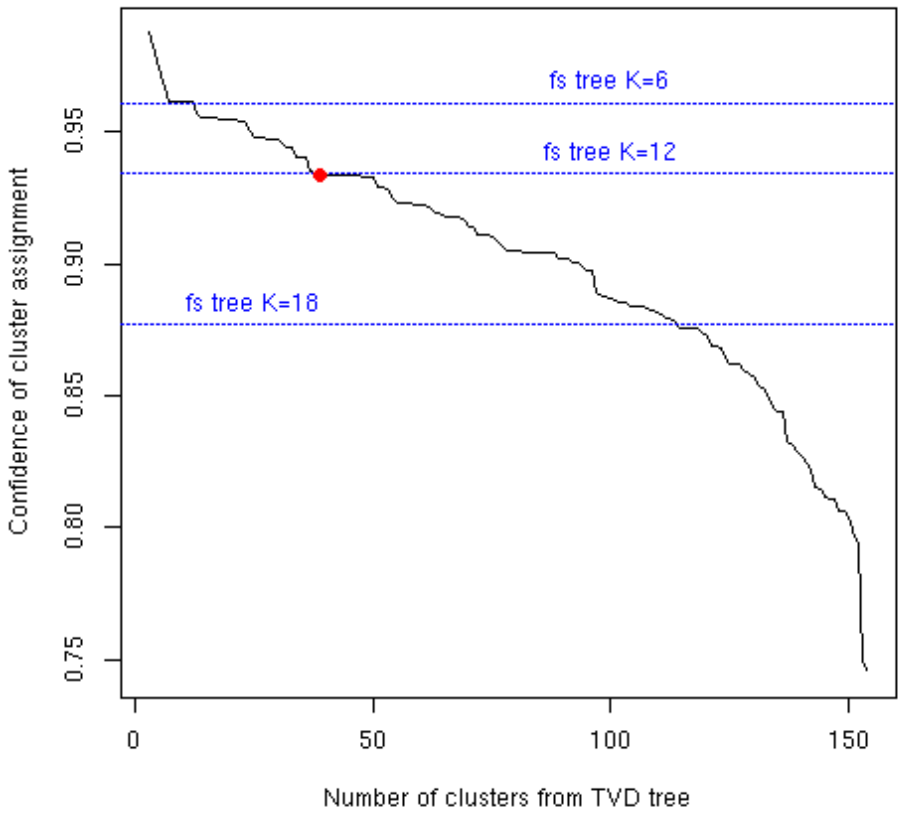


Figure 3.22 – Comparison of-FS tree and TVD-based tree performance measured by confidence of cluster assignment. While all clustering levels for TVD-based tree are shown with a black line, only three levels (K=6,12,18) are highlighted in blue for FS-tree. FS-tree has lower confidence of cluster assignment than TVD-based tree for the same levels of number of clusters, for example confidence of cluster assignment is similar for 12 clusters in FS-tree and 39 clusters in TVD-based tree (red point).

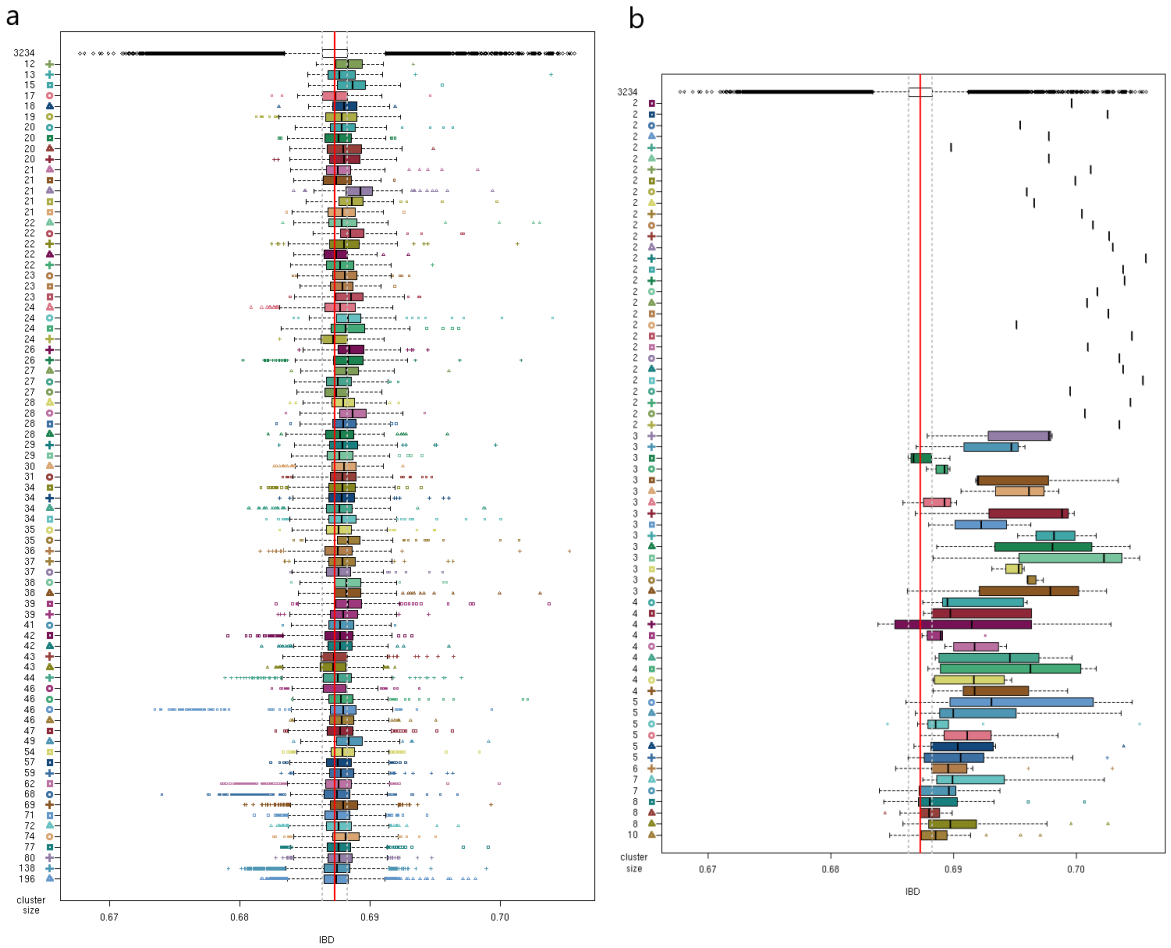


Figure 3.23 – Distributions of IBD statistics within 154 clusters. (a) Clusters with at least 12 individuals. (b) Clusters with low sample sizes. Clusters are color-coded as in Figure 3.15.

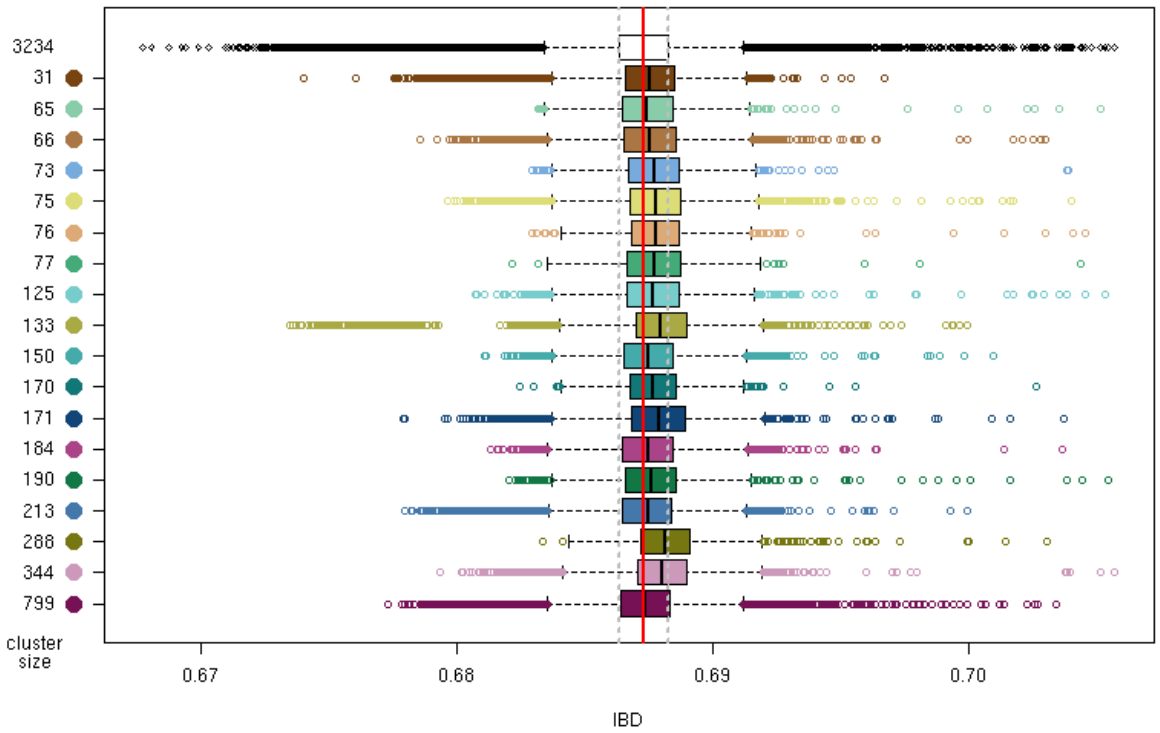


Figure 3.24 – Distributions of IBD statistic within 18 clusters that are color-coded analogous to Figure 3.17

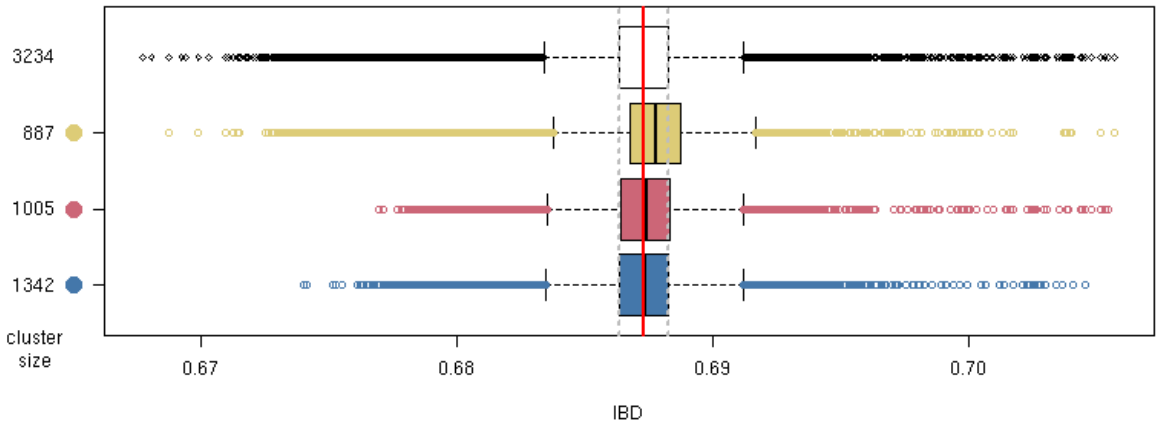


Figure 3.25 – Distributions of IBD statistic within 3 clusters that are color-coded analogous to Figure 3.16

3.3.8 Patterns in the coancestry matrix

To further understand the relationships between the clusters inferred by fineSTRUCTURE, I examined patterns in the coancestry matrix. The coancestry matrix between each pair of 3,230 individuals is presented in Figure 3.26. To better observe patterns characteristic of entire clusters, I also look at the averaged values over 18 clusters (Figure 3.27). Coancestry between individuals within a cluster is higher than between individuals in different clusters, reflecting genetic drift unique to each cluster. I observe differential patterns of ancestry sharing. Cluster *Maine-Anjou* and to a slightly smaller degree *Sèvres, Malo-Rennais and Retz* have relatively high level of contributions from all clusters, while the inverse pattern is found in clusters *Leon, Cornouaille, Vannes, Mauges1, Mauges2, Mauges3, Vendée-Nord*, indicating isolation of these regions. These results are consistent with previously described diversity patterns, in particular with runs of homozygosity.

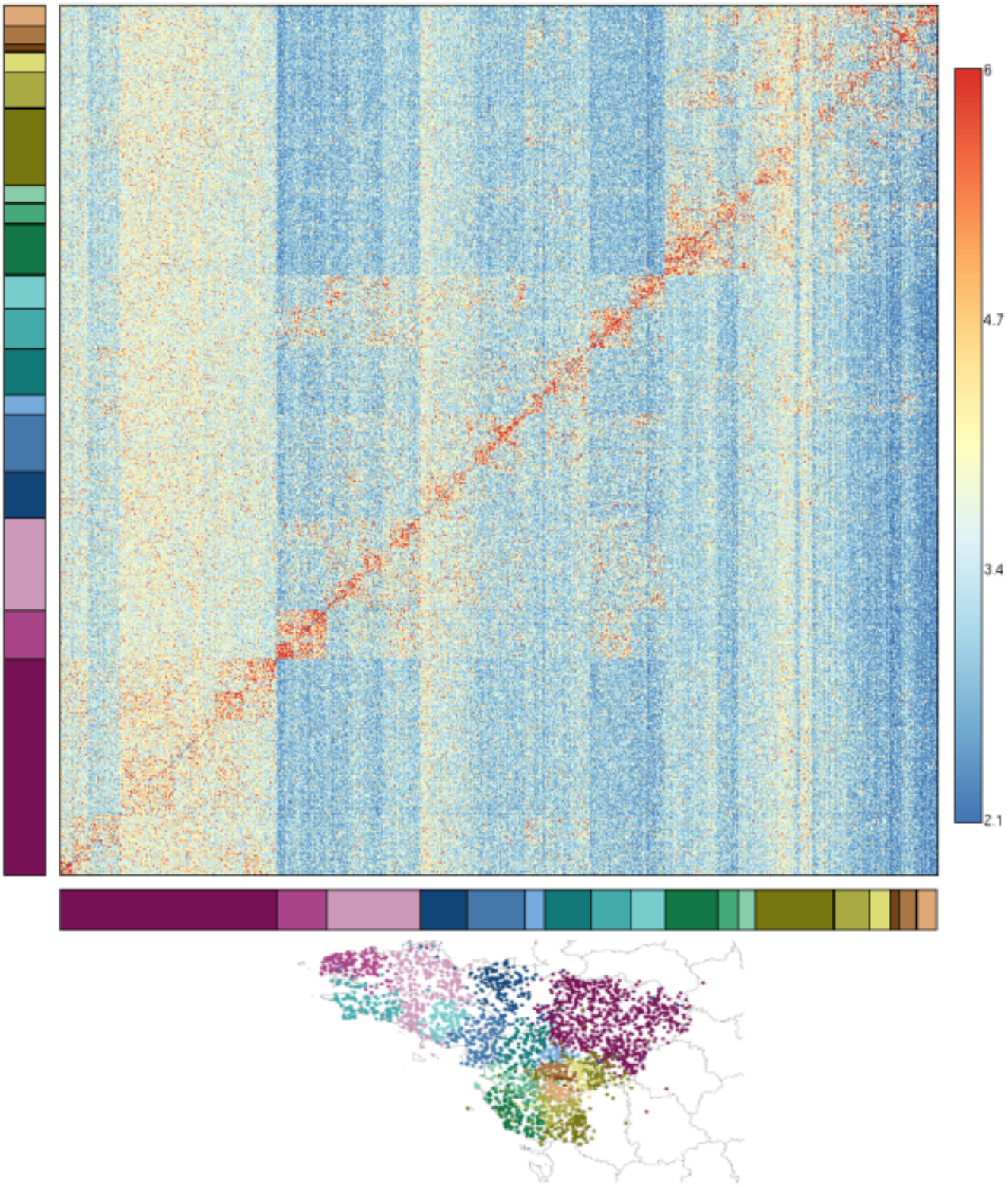


Figure 3.26 – Individual-level coancestry matrix from CHROMOPAINTER. For each individual contribution of ancestry from other individuals is shown in columns, while rows represent contribution of ancestry from an individual to others. In order to visualise the bulk of the variation, only values between 1-99 percentile were considered and those exceedind the range were colored according to the closer 1 or 99 percentile value. Warmer colors indicate higher levels of ancestry sharing. 18 clusters are color-coded analogous to Figure 3.17, present in smaller version here at the bottom of the figure.

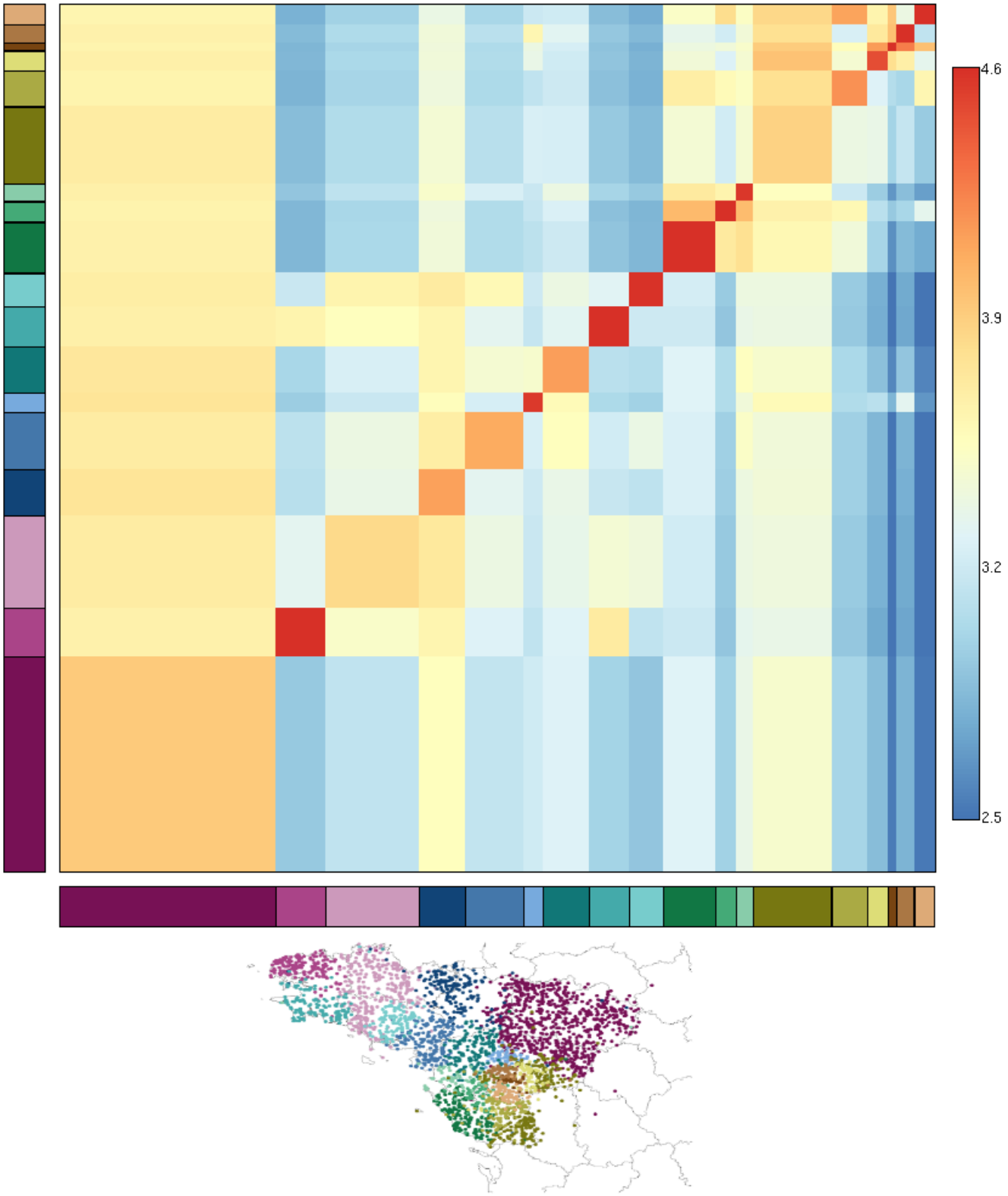


Figure 3.27 – Population-level coancestry matrix from CHROMOPAINTER for 18 clusters. For each cluster contribution of ancestry from other clusters is shown in columns, while rows represent contribution of ancestry from a cluster to others. In order to visualise the bulk of the variation, only values between 1-99 percentile were considered and those exceeding the range were coloured according to the closer 1 or 99 percentile value. Warmer colours indicate higher levels of ancestry sharing. Clusters are color-coded analogous to Figure 3.17, present in smaller version here at the bottom of the figure.

3.3.9 IBD segment counts for clusters identified with fineSTRUCTURE rather than for administrative units

Clusters identified by fineSTRUCTURE are supposed to reflect population differentiation better than administrative clustering. For this reason I recalculated IBD segment counts with the new division into 18 clusters. The resulting heatmap draws attention to Mauges region, especially to cluster *Mauges2* (Figure 3.28). Contrary to results on administrative units, clusters in Brittany have relatively lower levels of IBD sharing than clusters in Mauges, particularly for IBD segments longer than 2 cM, but it is important to note that the absolute values increased with the new division. This analysis helps to characterise Mauges clusters in terms of their particularly high IBD sharing levels, suggesting low effective population size.

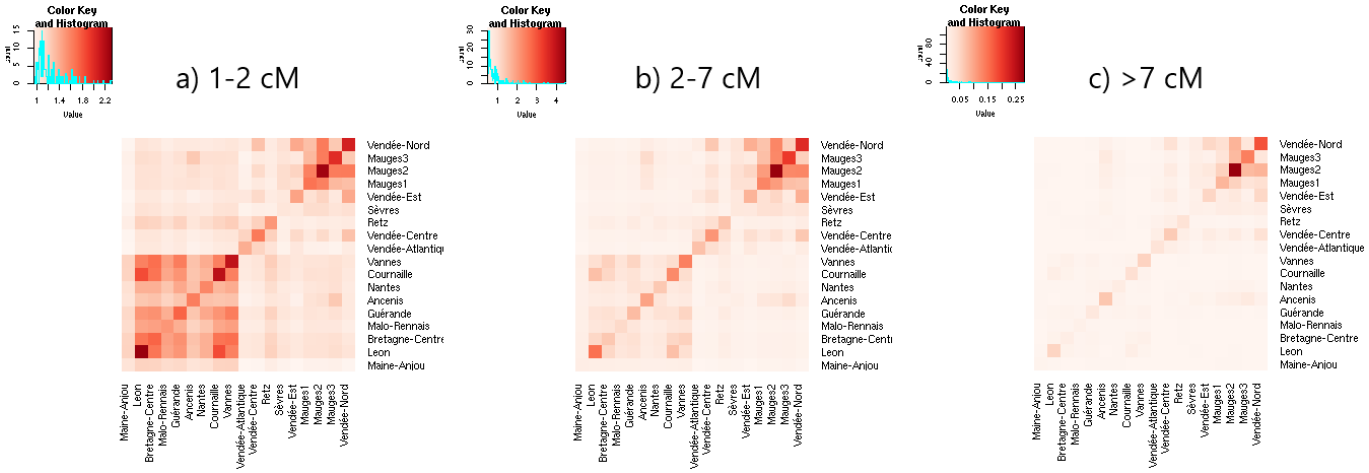


Figure 3.28 – Heatmaps of average number of shared IBD segments for 18 clusters identified with fineSTRUCTURE. IBD segments were divided into 3 subgroups ((a) 1-2 cM, (b) 2-7 cM, (c) > 7 cM), corresponding to different periods of demographical history (Figure 3.2).

3.3.10 Effective migration surfaces

With the EEMS method I inferred locations with corridors of gene flow (blue) and barriers to gene flow (orange). I found variable effective migration rates across Western France (Figure 3.29), many of which are consistent with the findings of other methods and prior historical knowledge. For example, the strongest barrier to gene flow is latitudinal and coincides with the Loire River, the limit of the first split of fineSTRUCTURE tree and with the strongest substructure in Mauges region. There is a correspondence with coancestry and diversity patterns: barriers to gene flow dominate in isolated Brittany and in southern part of the region, while the north-eastern part (area of cluster *Maine-Anjou*) is associated with corridors of gene flow.

3.3.11 Regional effective population size changes over time

Figure 3.30 shows that the three main clusters identified by fineSTRUCTURE differ in terms of historical effective population size (Ne) patterns estimated with IBDNe software for the last 1500 years (50 generations). (Figure 3.30). I chose `mincm` value (parameter that is responsible for minimum length of IBD segments included in the estimation) equal to 6 cM for interpretion (Figure 3.30 a), for which estimated trajectories are the most conservative (in a sense that they exhibit the simplest pattern). Current effective population size is estimated as $10^{5.75}$ for northeastern cluster, $10^{5.66}$ for Brittany Peninsula cluster and $10^{5.62}$ for southern cluster. However, the estimates for the last two generations (0 and 1) are extrapolated from earlier growth rates. Effective population size estimates for generation 2 (around year 1890 AD, assuming that a generation lasts 30 years) are equal to $10^{5.68}$ for the northeastern cluster, $10^{5.57}$ for Brittany Peninsula

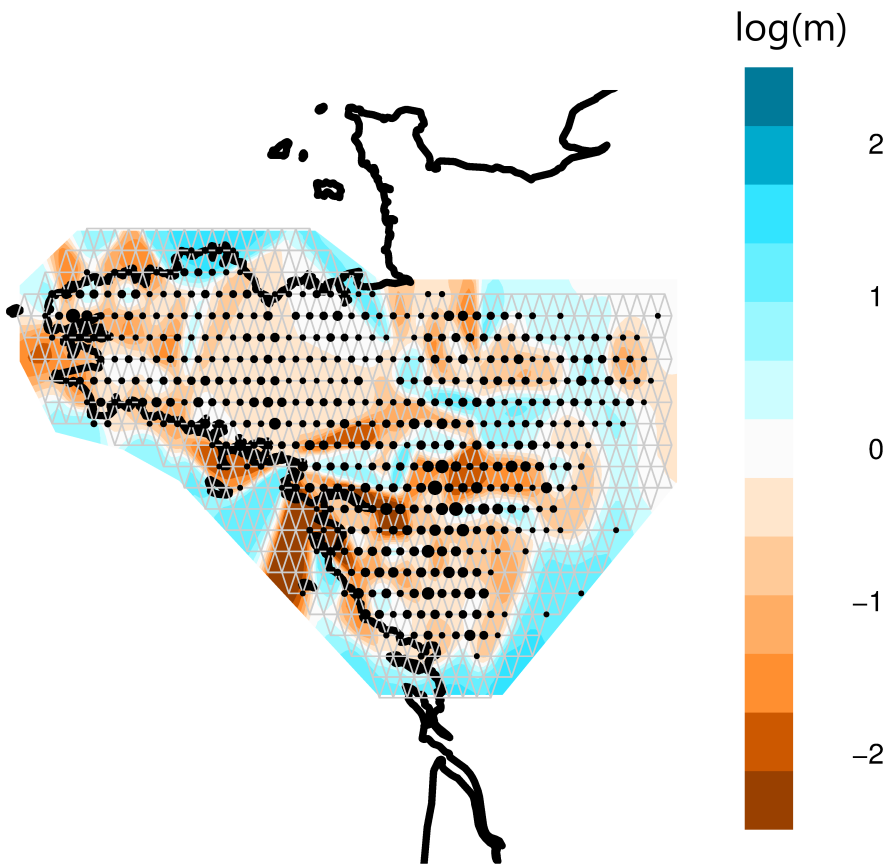


Figure 3.29 – Effective migration surface inferred with EEMS

cluster and $10^{5.46}$ for southern cluster, preserving the relative order from generation 0. Further in the past, estimated patterns diverge substantially between subpopulations. The pattern associated with cluster encompassing the Brittany Peninsula is a population growth at nearly steady exponential rate. The confidence interval of the Ne trajectory for this subpopulation is relatively narrow, indicating consistency within the data. The narrow confidence interval for the Brittany cluster is not a consequence of a larger sample size, because actually the sample size is the smallest for Brittany. Also the trajectory for the northeastern cluster shows a population growth, however the growth rate is higher for generations 0-9 (years 1950-1680 AD). The effective population size of this subpopulation is the highest consistently across time. The confidence interval widens around generation 17 (year 1440 AD), suggesting that effective population size might have been higher in this generation and might even present a decreasing pattern until generation 11 (year 1620 AD), a bottleneck spanning 6 generations. The South cluster undergoes a slight decline in effective population size between 50 and 25 generations ago (years 1200-450 AD). Interestingly, the relative order between subpopulations changes at generation 29 (around year 1080 AD), where Brittany Peninsula and South have both estimates of effective population size equal to $10^{4.71}$ and North-East remains the largest with $10^{5.25}$. At the last interpreted generation, generation 50 (around year 450 AD), the northeastern cluster has effective population size of $10^{5.02}$, the southern $10^{4.81}$ and the Brittany Peninsula cluster shows the smallest Ne $10^{4.51}$. Confidence interval for the southern population widens around year 1470, 1170 and towards the year 450 AD, indicating high levels of heterogeneity and leading to overlap with confidence interval of northeastern cluster and thus similar effective population sizes for these two populations.

Given the fact that the size of the segments may have impact on the results [23], I investigated IBDNe trajectories for `mincm` values between 3 and 8 cM. Consistently with the size of confidence intervals for $\texttt{mincm} = 6$ described above, patterns for southern and northeastern subpopulations are more variable than

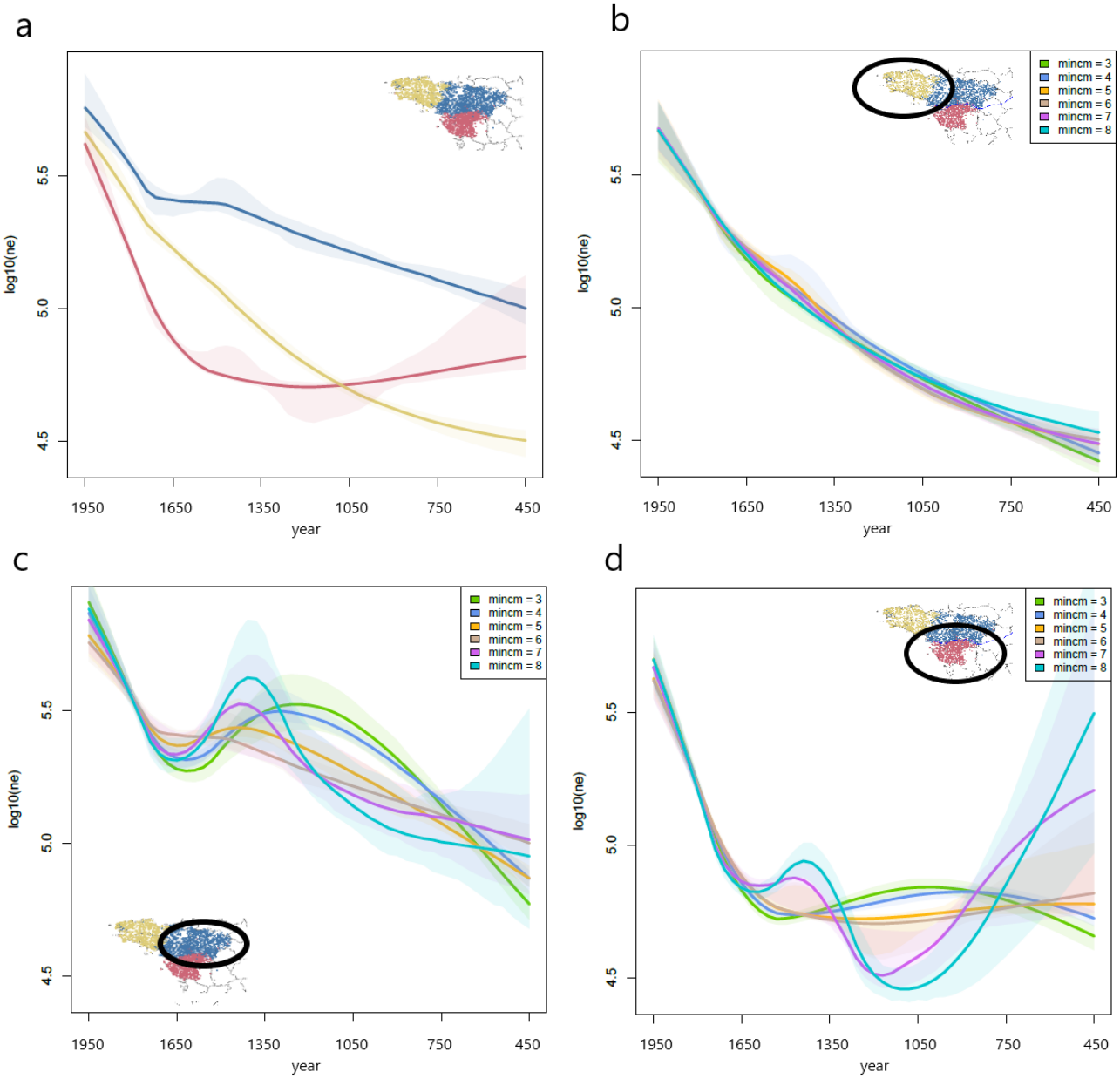


Figure 3.30 – Effective population size trajectories inferred with IBDNe. (a) Trajectories for 3 clusters identified with fineSTRUCTURE, for `mincm` $= 6$, color-coded analogous to Figure 3.16 (red - South, yellow - North-West, blue - North-East). (b-d) Trajectories for different values of `mincm` parameter for (b) northwestern, (c) northeastern and (d) southern cluster.

for the Brittany Peninsula cluster. All six trajectories for the latter are very close to each other. For the southern and northeastern cluster, different `mincm` parameter values give different pictures of population demography. Trajectories for the northeastern cluster oscillate since the generation 9 (year 1680 AD). For high `mincm` of 7 and 8, there is a peak at generation 18 (year 1410 AD), for `mincm` $= 5$ a smaller peak at generation 17 (year 1440 AD). Oscillations for `mincm` $= 3$ and 4 peak only at generations 23 (year 1260 AD) and 22 (year 1290 AD) respectively. Those peaks are followed by declines in effective population sizes. The trajectory for `mincm` $= 6$ is the only one with constant population expansion. For the southern cluster, there are three types of patterns. For longer segments (`mincm` $= 7$ and 8), there is a series of two bottlenecks. Effective population size at generation 50 (year 450 AD) is relatively large, followed by population contraction until generation 26 (year 1170 AD) and 28 (year 1110 AD), respectively, then re-expansion until generation 16 (year 1470 AD) and 17 (year 1440 AD), respectively,

followed by contraction until generation 12. When intermediate-length segments are included (mincm = 5 and 6), there is a mild population contraction between generations 50 (year 450 AD) and 25 (year 1200 AD). Trajectories for mincm = 5 and 6 indicate one bottleneck that starts at generation 30 (year 1050 AD) with minimum effective population size achieved at generation 14 (year 1530 AD). Ne oscillations taking place around 8-20 generations ago (years 1350-1710 AD, such as the ones observed in our data, should be consistently detected regardless of the minimum segment size used. Overall, for two out of three analysed subpopulations the results for different mincm parameter values were divergent. However, such inconsistency suggests we might have limited power to correctly infer IBD segment sizes for particular lengths. Browning and Browning [23, 24] emphasize in their original work that such small-scale oscillations can be artifacts and should not be overinterpreted.

Difficulties in the interpretation of the oscillating IBDNe results inspired a simulation study in our lab. Its goal was to test the ability of IBDNe to detect a recent and short bottleneck (spanning 10 generations, 13-23 generations ago) and how the different mincm parameter values may affect the estimates. The study revealed that IBDNe detects a bottleneck for each of 10 sets of simulated data (Figure 3.31), moreover it detects correct size of the bottleneck. However, IBDNe does not detect the true shape of the bottleneck. It reports a progressive bottleneck rather than a sudden one. The choice of mincm parameter value has only a slight impact on the results at generation 15 and further in the past, the trajectories remain similar. The presence of bottleneck is detected no matter what the mincm parameter value is (Figure 3.32), which is not the case on the real PREGO data (Figure 3.30 b,c,d).

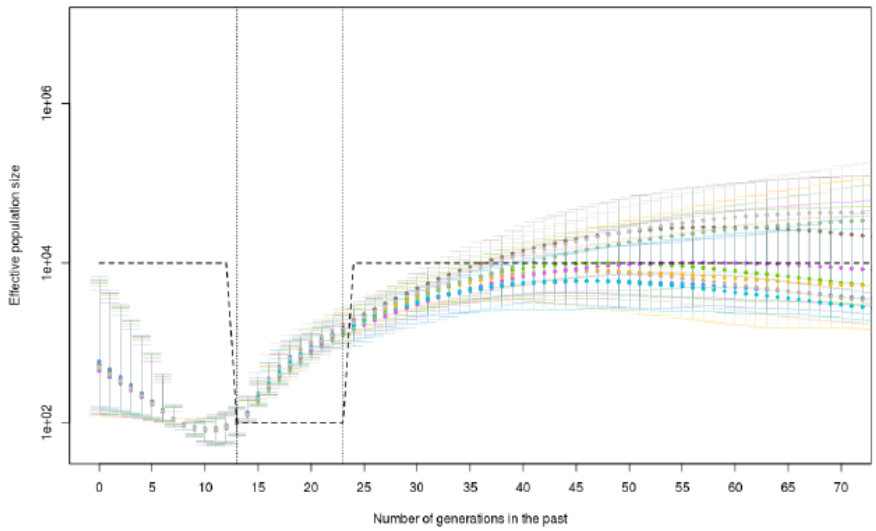


Figure 3.31 – IBDNe results on 10 sets of simulated data, distinguished by different colours. Dashed line reflects effective population size of the bottleneck model under which the data were simulated. The IBDNe detects a progressive change of effective population size rather than a sudden change as in the bottleneck model used to generate the data. Realised by Charlotte BERTHELIER.

3.3.12 PREGO dataset in relation to neighbors in Europe (1000G)

We merged PREGO data with European samples from 1000G (FIN, CEU, GBR, IBS, TSI) to explore gene flow from other populations surrounding France. We estimated ancestry proportions with ADMIXTURE under assumption of different numbers of ancestral populations ($K = 2 - 6$). At the level $K = 2$, the proportions differ between subpopulations in the dataset. The ancestral components at this level seem to be northern and southern Europe, with FIN having on average 86.4% northern ancestry and TSI 76.6% southern ancestry (Figure 3.33). PREGO individuals from Brittany peninsula cluster are closer to GBR and

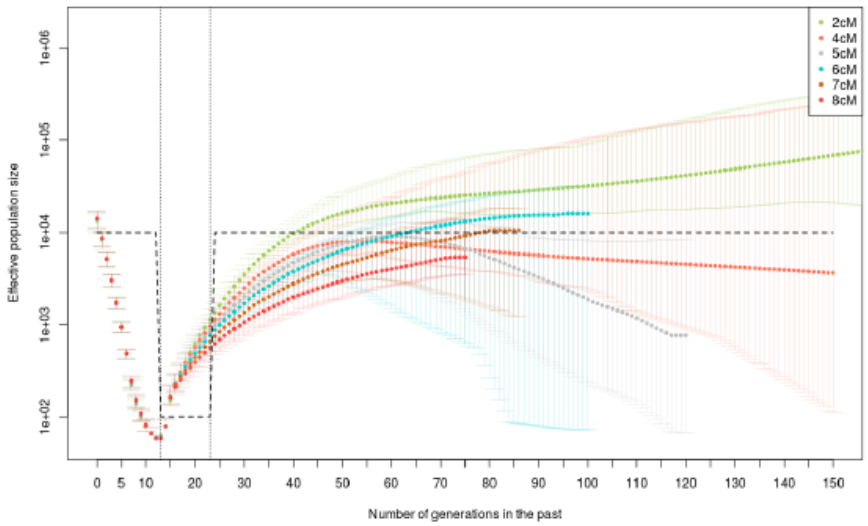


Figure 3.32 – IBDNe results for different `mincm` values (between 2 and 8 cM) on simulated data. Dashed line reflects effective population size of the bottleneck model under which the data were simulated. The choice of `mincm` parameter value has an impact on the results, but the presence of a bottleneck is detected no matter what the `mincm` parameter value is. Realised by Charlotte BERTHELIER.

CEU than to the other two clusters with more southern ancestry. Level $K = 3$ reveals even more differences between three subpopulations in Western France. The southern cluster has a similar profile to Iberian samples from 1000G. Level $K = 4$ separates IBS and TSI from the southern cluster. Care should be taken when results are interpreted because the model underlying the program ADMIXTURE might not represent the nature of the data well. For further levels, $K = 5$ and $K = 6$, certain ancestral components present more variability within subgroups than between them, which may indicate that K value is too high.

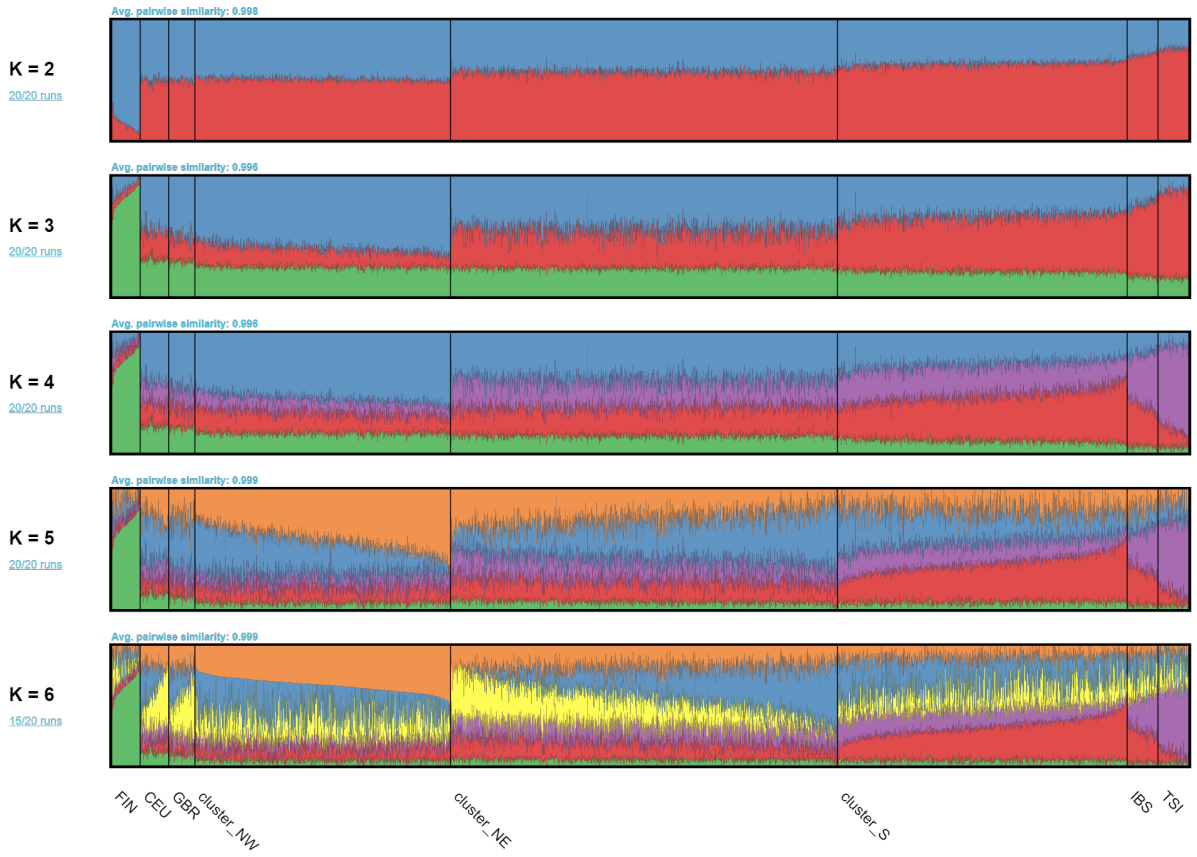


Figure 3.33 – ADMIXTURE results on merged PREGO and 1000G datasets. PREGO individuals are grouped into 3 clusters identified by FineSTRUCTUREfineSTRUCTURE. FIN - Finland, CEU and GBR - Great Britain, IBS - Spain, TSI - Italy.

3.4 Discussion

In this study we investigated fine patterns of population structure and diversity within Western France by using high-density genome-wide SNP array data of 3,234 individuals with three generations of ancestry from the region. This study is, to date, the most comprehensive genetic study on the population structure of Western France revealing substantial levels of population stratification likely caused by different demographic histories across the region. Moreover, it broadens our understanding of the local levels of genetic stratification across European populations.

Population differentiation measure F_{ST} revealed subtle levels of differentiation increasing with geographical distance, in line with isolation by distance model. The lowest F_{ST} values occur in the pairwise comparisons within the department of Sarthe while the largest values are found between Finistère and Vendée department. These results corroborate local genetic differentiation patterns from Karakachoff et al. [99], where the largest values were obtained in the three westernmost departments of Brittany peninsula and in Vendée.

In our study we observe larger average F_{ST} ($95 * 10^{-5}$ for 18 clusters) than previously reported values in Ireland ($30 * 10^{-5}$ [63]) and Denmark ($20 * 10^{-5}$ [7]); and smaller than the average F_{ST} found in Finland ($250 * 10^{-5}$ [101]) and the British Isles ($200 * 10^{-5}$ [115]).

More differentiated populations than North-Western France, like Finland and the British Isles, occur across substantially larger areas. When we compare North-Western France with the British Isles regions of similar area, northern part of the United Kingdom has still higher average F_{ST} ($180 * 10^{-5}$ for 4 clusters from Scotland and Northern Ireland, Orkney Islands excluded) than North-Western France, but southeastern part has smaller average F_{ST} ($< 50 * 10^{-5}$ for South-East, Welsh Border and W.Yorkshire clusters). Also

Ireland occupied a larger surface than Western France, but is less differentiated. It is worth noting that North-Western France is not an island like Ireland or British Isles, but a part of continental platform, where people can move without obstacles. This fact should decrease the levels of differentiation.

Despite the general pattern of isolation by distance, discrete population structure was captured with haplotype-based fineSTRUCTURE clustering. fineSTRUCTURE identified an unprecedented number of subpopulations on its finest level, 154. Identified clusters reveal a rich pattern of subtle fine-scale genetic differentiation with many details not captured previously. Haplotype-based methods increase resolution for detecting population structure as we have shown visually examining allele-frequency-based and coancestry matrix-based PCA profiles on the map (Figure 3.13) and comparing number of PCs correlated with geographical coordinates and as has been suggested before [101, 108, 115]

Several clusters match geo-political and linguistic boundaries. For example, the first split in the clustering coincides with the Loire River as well as the political border between Roman provinces Aquitaine and Lyonnaise which was stable for centuries (Figures 3.16 and 1.7) . The second split is close to Loth line (Figure 3.18), the limit associated with Breton toponymy, as well as the limit between spoken Breton and Gallo dialects. After these two splits we obtain division into 3 subgroups, Brittany peninsula, North-East and South, that serves us for several downstream analysis.

On the finer level of 18 clusters (Figure 3.17), we see for example clear division between *Leon* and *Cornouaille* within Finistère, which can be explained by different Breton dialects, presence of the Elorn River or historic tradition of opposition of this two groups. Cluster *Vannes* is characterised by particular Breton dialect, substantially different from others. Cluster *Retz*, though located in the south of Loire River, covers the region which is culturally under Breton influence (the southernmost Breton cluster), but, interestingly, at the coarser division it belongs to the southern rather than the northern clusters. Thus, this cluster likely constitutes a transition area.

The choice of number of clusters is a constant problem in fineSTRUCTURE approach for which no guidelines have been proposed yet. The decision is arbitrary, often motivated by getting rid of clusters with a small sample size. Our idea to make a decision less arbitrary was to compare confidence of cluster assignment between FS-tree [115] and TVD-based tree. Unlike FS-tree, TVD-based tree uses measure of difference between clusters. TVD was defined in Leslie et al. [115] and it was used for the tree building purposes in Kerminen et al. [101], where it showed better properties compared to FS-tree in terms of sample size consistency and analysis context, but the confidence of cluster assignment was not analysed. We observed on PREGO data that confidence of cluster assignment for TVD-based tree is higher than in FS-tree given the same number of clusters (Figure 3.22). If this observation is repeated on other datasets, it will add argument for preferential use of coarser levels of TVD-based tree rather than FS-tree.

Genetic differentiation appears not uniform with respect to geographical distance. Whilst in the North-East we observe a relatively large cluster *Maine-Anjou* containing nearly 25% of all samples, in the south-eastern region Mauges there is a dense substructure. Large unstructured clusters have already been observed in Leslie et al. [115] with possible explanation of local migration over many generations. Such an explanation fits to our data as well, because *Maine-Anjou* is associated with corridors of gene flow in the EEMS results (Figure 3.29) and in the coancestry matrix it has relatively high levels of copying from other clusters (Figure 3.27). This region is an open area lacking major geographical barriers to human movement. On the other hand, dense substructure in Mauges resembles the one from the province of Pontevedra in Galicia, Spain from the study of Bycroft et al. [26] where some clusters have ranges of less than 10 km and align with hills and river valleys.

It is possible that differential density of structure is linked to the enhanced effect of genetic drift in populations with lower effective population size. IBDNe results suggest larger effective population size across last 50 generations in the North-East, while in the South the effective population size is the smallest in the most recent generations. Mauges subpopulation is a part of southern cluster, but we do not know to which extent it is affecting the estimate for the entire South population. Mauges has relatively long average runs of homozygosity also pointing to lower effective population size. We hypothesize that the particular landscape might have been partially responsible for the low effective population size in the southern cluster,

in particular in Mauges. It is a hilly bocage, which is a terrain of mixed woodland and pasture, with fields and winding country lanes sunken between narrow low ridges and banks surmounted by tall thick hedgerows. Such landscape is hardly accessible, slows down marching, thus might have caused relative isolation. Other possible link is a peasant counter-revolutionary uprising, Wars of the Vendée, that initiated in Cholet in Mauges in 1793. About 170,000 military and civilians were killed in the insurrection [92], out of a population around 800,000. Another possibility is that cryptic relatedness occurs in our sample which can be a confounding factor. We cannot exclude that our sample contains distant relatives but our analysis of pairwise IBD statistics distribution within clusters shows that it is not a common issue (Figure 3.23, 3.24, 3.25). All methods consistently point to low long-term effective population size in the Peninsula of Brittany. Such observations are in line with Karakachoff et al. [99]. Low effective population size can be a consequence of geography causing isolation from migration, as Brittany is only accessible through land from the east and limited by the Atlantic Ocean and the English Channel from the other directions. Additionally, Brittany owes its name to migrations of Celtic populations from the British Isles fleeing the Saxon invasions in the Vth and VIth century. These migrations might have been followed by a founder effect likely reducing the effective population size. Furthermore, Brittany remained in a specific political context for centuries that might have contributed to isolation: as the edge of Roman Gaul, later independent territory from Barbarian kingdoms, independent Brittany Duchy until the union with France in 1532 in which it conserves many privileges. Moreover, linguistic barriers may have posed problems of communication and contributed to isolation. In Brittany two main languages have been traditionally spoken: 1) Breton, which is a Celtic language and spoken at the end of the Brittany Peninsula and Gallo, and Oïl language spoken in the eastern part of the peninsula.

Larger effective population size estimate in the North East part of the region might be as well a consequence of migrations into this region (it contains *Maine-Anjou* cluster). Contrary to Brittany, geographical location and dialects were not obstacles to migration here, resulting in a relatively less isolated region. I cannot distinguish the effects of effective population size and migration in my analysis, because the approach I use does not explicitly model migration. It could be done with other approach, for example with whole genome sequencing (WGS) data and fastsimcoal2 [47] or Approximate Bayesian Computation.

Given the fact that the size of the segments may have impact on the IBDNe results [23], we varied `mincm` parameter that is responsible for minimum length of IBD segments included in the estimation (Figure 3.30 b,c,d). We noticed that the decision on the value of this parameter should be taken with care. For two out of three analysed subpopulations, South and North-East, results for different parameter values were divergent and giving different picture of demographic history, in particular suggesting a bottleneck. These divergent patterns of the IBDNe method may be technical artifacts that could be explained by potentially inaccurate estimates of IBD segment lengths in PREGO data. Authors of the IBDNe method, Browning and Browning [23] warn that sometimes artifactual oscillatory behaviour appears over short timescales of the IBDNe trajectory that should not be overinterpreted. Moreover, the IBD segments cannot localize large changes in population size to a single generation. The IBDNe trajectory may hence be more smooth and miss a sharp turn [23]. It is the case in our simulation study of recent sudden bottleneck (Figure 3.31). This simulation study also showed that the presence of bottleneck is detected no matter the value of `mincm` parameter (Figure 3.32).

Athanasiadis et al. combined trajectories from IBDNe with epidemics events and found impact on effective population size of Black Death Plague in Scandinavian countries. We thought of events that could have impacted the demography of our population. Substantial changes in census size happened in the middle of VIth century and in IV-XIXth century. In 536 an Icelandic volcano eruption spewed ash across the northern hemisphere, limiting sunlight and causing $1.5 - 2.5°\mathrm{C}$ drop of temperatures in summer, failure of the crops and starvation [62]. It was followed by the Plague of Justinian in 541, which killed 35-55% of the population [62]. Second pandemic, the Great Plague or Black Death, struck in the middle of XIVth century. While the population size in the kingdom of France was estimated to be 20 million in 1348, it dropped down to 12 million in 1400, followed an uneven trajectory to recover the 20 million at the end of Louis XIVth reign (1715) [44]. A period of temperature cooling and altered climate conditions known

as little ice age took place between XVth and XIXth century [129]. However, IBDNe does not detect any decrease in effective population size in corresponding generations in subpopulation of Brittany, for which the trajectories consistently show the absence of a bottleneck for every `mincm` parameter value. This suggests that above events might have only affected census population size and not effective population size in Brittany. It is possible that oscillations in southern and northwestern subpopulations are related to above events.

It would be very interesting to have WGS data, on which we could investigate effective population size changes with approaches such as MSMC [190] and SMC++ [209].

Karakachoff et al. [99] shows genetic proximity between Bretons and Irish. Second closest population to Bretons in this study (measured by the mean IBS statistic) is UK. In our study we chose 1000G as a panel with neighboring populations. Although this decision leaves us only with GBR and CEU populations as a proxy for Ireland, 1000G as a WGS study maximises the number of SNPs common between datasets. Inclusion of isolated population FIN that have been through bottlenecks can help to contextualise results for PREGO samples. It may also serve as a proxy for northern component that polarises results between north and south. Similar profiles of ADMIXTURE proportions that we obtained for Breton and GBR and CEU populations (Figure 3.33) corroborate the result of Karakachoff et al. [99]. Our Southern cluster is closer to populations from southern Europe, IBS and TSI. Historically the territory of the Southern cluster was a meeting place of southern and northern cultures [156].

A strong genetic clustering and localised genetic drift, as we see in the peninsula of Brittany, might have contributed to the increase in frequency of deleterious alleles [1]. An important though non-deleterious example is the lactase persistence allele (SNP rs4988235). It has the allele frequency above 70% for 3 out of 4 departments in Brittany whereas the allele frequency decreases to between 45 and 55% in the easternmost departments of Western France [99]. A deleterious example consists of three mutated alleles in CFTR gene linked to cystic fibrosis: G551D, 1078delT, W846X [195, 51]. These mutations are rare in France, but have high frequency in Breton population, respectively 3.8%, 3.7% and 1.1%, and are particularly observed in the western part of the peninsula. The incidence of cystic fibrosis in the Finistére department is around 1/2000 compared to general incidence in Caucasian population of 1/3500. Our results imply that the phenomenon of increased frequency of possibly pathogenic alleles will likely happen in specific areas such as Brittany.

A previous study [99] has shown that structure does occur at the scale of Western France. However, that study had several limitations that we manage to address in this study. First, our sampling is much more even with respect to geography, with only one out of nine departments, Ille-et-Vilaine, being slightly less densely covered. Second, our sample size doubles the sample size of that study, although being restricted to a smaller geographical area. It facilitates more detailed description of population structure, especially for regions underrepresented in Karakachoff et al. [99] study, for example Brittany. Third, the PREGO dataset was designed to serve as best as possible the purpose of extremely fine-scale study of genetic structure. Recruited individuals have their 4 grandparents born in proximity within two regions of northwestern France. We thus effectively studied the structure of population at the times when those grandparents were born. It was around the beginning of XXth century, thus before large-scale long-range migrations of this century happened. In spite of such individual ascertainment being an advantage for population genetics study, one has to keep in mind that the patterns we observed might thus reflect other population than general present-day French population. Last but not least, we extended Karakachoff et al. study by employment of recent methods, in particular haplotype-based approaches.

In summary, we have presented the first (to our knowledge) extremely fine-scale study of genetic diversity, genetic differentiation and demographic history within a region of a country, by using careful sampling, genomic data and powerful statistical methods. This study demonstrates the utility of building reference panels with detailed information about geographic origin of three generations. The resulting genetic clusters and the characterisation of their effective population size and their ancestry proportions compared to other European groups provide important and novel insights into the peopling of Western France and potential explanation for different disease prevalence within this region.

4

Discussion

4.1 Relation to history, demography, linguistics and culture

In this thesis we report two studies (SU.VI.MAX/3C and PREGO) that together form a very thorough analysis of fine-scale patterns of population structure and diversity in France. We demonstrate that population structure does occur at the countrywide level as well as within Western France. Although the studied population is relatively homogenous like other mainland continental European populations, the levels of stratification are significant and may have impact on disease association studies. This stratification might have been shaped by historical, demographic and cultural events and we attempted to identify the relations between these events and genetic patterns.

In both studies, we find correlation between genetic data and geographical information on places of birth of individuals in France. Individuals that were born in proximity tend to be more genetically similar than individuals born in distant regions, which is in line with isolation by distance principle. However, we also identify discrete patterns, subpopulations that match geographical regions. Remarkably, we distinguish 6-7 subpopulations that are highly concordant between two independent studies of whole continental France, 3C and SU.VI.MAX. The PREGO study can be thought of as a zoom-in on a smaller territory, North-Western France, which encompasses the territory of Brittany and parts of central and northern subpopulations from 3C and SU.VI.MAX (but not the same individuals). Three main subgroups of this zoom seem to correspond relatively precisely to the subpopulations from 3C and SU.VI.MAX. The granularity of the data is much finer in PREGO than in 3C and SU.VI.MAX, we assign to individuals the smallest administrative unit – commune, versus a coarser level of department. Thus, in PREGO we also distinguish and interpret the finer level of 18 subpopulations.

The most differentiated subpopulation in the whole continental France studies is South-West. This region is neighbouring with Spanish border and encompassing a French part of the territory called the Basque Country. It is thus likely that Southwestern subpopulations received gene flow from Spanish and Basque populations. With 3C and HGDP datasets we tested the hypothesis if individuals from South-Western group in 3C colocalize with French Basques from HGDP in the principal component space and several individuals indeed do. Basques are an indigenous population that speak the non-Indo-European Basque language. The present-day Basque population is located on both sides of the western Pyrenees, but according to toponymical data, ancient Basques occupied region stretching from Ebro River in the South (in modern Spain) to the Garonne River in the North. The limit of the South-West population aligns with the Garonne River, thus coinciding with the limit of territory of the ancient Basques. This limit may correspond to the times of Ist century AD, when the Roman historian Strabo differentiated the population

living between the Garonne River and Pyrenees from Celtic population and called them Aquitani.

The second important division, concordant in both whole continental France studies, also aligns with a river and a linguistic border. It separates roughly Northern from Southern France. The division seems to coincide with the longest French river, the Loire. However, it is unsure if this river could be an obstacle to gene flow, especially in its upper course. Although an argument that bridges were rare before XIXth century speaks for it, there are counterarguments that several hundreds of fords existed or the river could be traversed by a ferry. Boat transport could in theory even accelerate gene flow. Political borders, though variable in time, often coincided with the Loire River too and might be reflected in this division. It aligns also with the von Wartburg line which is a linguistic limit between "Langue d'Oïl" part (influenced by Germanic speaking) and "Langue d'Oc" part (closer to Roman speaking) (Figure 1.10). This limit was located on the Loire River in the Xth century. Later on, it moved towards the South, for example in XIIIth century Oïl was spoken in Poitou and Saintonge. This linguistic expansion might be a result of cultural change, it was accompanied by diffusion of new agrarian system [22]. With the granularity of our data, the division limit may lay slightly south of the Loire too. This division may correspond as well to St. Malo - Genève line (Figure 4.1) which was introduced by Charles Dupin in 1827 [43]. This imaginary line is related to the socioeconomic context. It represents a duality between industrial and urban North-Eastern France and agricultural and rural South-Eastern France.

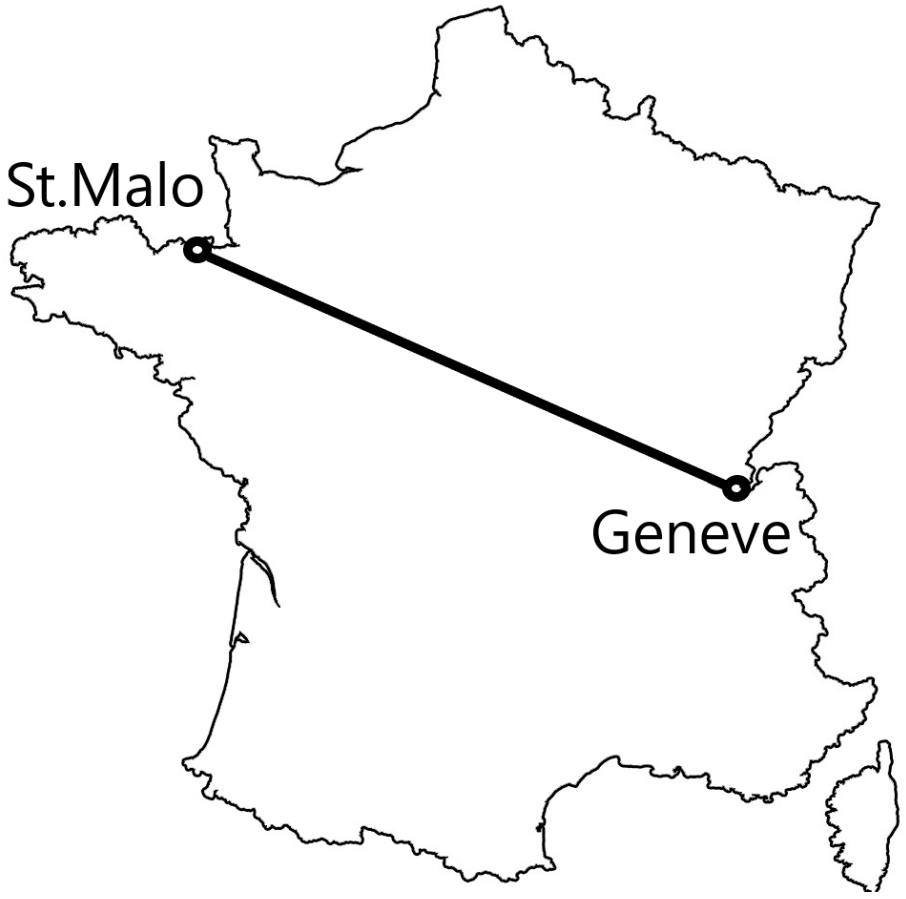


Figure 4.1 – A map of France with St.Malo-Genève line.

Another French subpopulation concordantly identified in both datasets is Brittany. There are geographical, linguistic and historical facts associated with that. Brittany is a peninsula, limited by the ocean from three sides and only accessible from land from the East. Bretons had their own languages, with two main groups Breton and Gallo. While Gallo is one of Oïl languages, with similarities to neighboring French dialects, Breton, spoken at the end of Brittany Peninsula, is a Celtic language, making communication with French a challenge. Historically, Brittany had a long tradition of independence. In the early Middle Ages it was independent from Barbarian kingdoms, then a Brittany Duchy was formed and remained indepen-

dent until 1532. In this year a union with France took place, but nevertheless Brittany conserved many privileges. However, likely the most important historical fact are migrations of Celtic people from British Isles, that settled in Brittany fleeing from Saxon invasions, giving a new name to the region previously called Armorica. We can see a signature of this migratory movement in the ancestry profiles. Brittany has the highest proportion of ancestry from British Isles and is substantially closer to them than the Northern subpopulation, despite both being equally close geographically. A strong Celtic background distinguishes Brittany from other parts of France.

The PREGO dataset enables us a zoom into Brittany population. Finer granularity of the data allows us to notice striking alignment of a limit between two genetic subpopulations and lower course and estuary of Loire River (Figure 3.16). This is concordant with results on SU.VI.MAX and 3C, but adds precise location of the limit. Loire River was not only a physical obstacle, but in certain times also a political border, for example between Roman provinces Aquitaine and Lyonnaise. PThe REGO dataset allows us to see population splits not identified in the whole continental France studies. On the coarse level of 3 subpopulations in North-Western France we identify a limit close to Loth line (Figure 3.18), associated with Breton toponymy, as well as the limit between spoken Breton and Gallo dialects.

It is possible that we see relatively high levels of genetic stratification in North-Western France thanks to high stability of the rural population. On the one hand, rural areas lacked long distance migration, thus they were relatively isolated and static [86]. On the other hand, rural populations are characteristic for high endogamy and high percentage of marriages occurring within a short distance, at least until XIXth century. Nearly 70% of marriages were between people from the same village in the years 1740-1829 [17]. According to the analysis done by N. Pellen [163] in Kerlouan, a Western French county, about 90% of mating was from the same village during the 17th and 18th centuries. R. Leprohon [114] found that a high percentage of marriages took place between people living within a radius limited to 5-10 km.

In conclusion, we were able to correlate our results with numerous geographical, linguistic, cultural features and historical events. What is important, our results are convergent through datasets, despite different individual and SNP ascertainment. Moreover, many different methods show results that are concordant and not contradicting each other. This gives us more certainty that the observed signal is real and it is not a random fluctuation.

4.2 Perspective of the field - analysis of rare variants

Our analyses on SU.VI.MAX, 3C and PREGO data were conducted on genome-wide SNP arrays, thus on fairly common variants. On such arrays, rare variants are incompletely represented and impute poorly with current reference panels [211]. However, with the advent of whole genome sequencing, we anticipate more focus on population genetics studies conducted on rare variants. With studies such as 1000G [9], GoNL [210], UK10K [211], rare variant analysis becomes a standard. Rare variants tend to be younger and more geographically restricted compared to common variants [65, 208]. Rare variants are highly differentiated between populations [25] and provide information about both demographic history and fine-scale population structure [65, 138].

Humans harbor an excess of rare variation, primarily due to recent dramatic increases in population size [100, 147, 208, 58]. Even though the information content of a single rare variant is often smaller than a single common variant, their abundance provides cumulative information which is a powerful tool for testing hypotheses about fine-scale population structure [154]. There are methods developed particularly for WGS and rare variants. One of them is analysis of the rarest shared variants known as f_2 variants or doubletons. Mathieson and McVean developed a method to estimate age (time to most recent common ancestor) of f_2 haplotypes [139]. The median age of f_2 haplotypes in Europe was estimated between 50 to 160 generations. As another example, Schiffels et al. [191] investigated relative rare allele sharing patterns between ancient and modern samples to estimate percentage of the ancestry of modern Britons that was contributed by Anglo-Saxon immigrants and proposed a population history model from rare variants with

a new approach rarecoal. UK10K examined excess of rare allele sharing as a function of geographical distance [211]. We did the same on the part of the PREGO cohort whole-genome sequenced together with samples from other parts of France (Figure 4.2). Rare genetic variants showed excess allele sharing at distances smaller than about 100 km, reduced allele sharing for distances about 200-300 km, and allele sharing close to the expected value for more than 400 km, except for doubletons (AC=2) that presented reduced sharing in this interval. This implies population structure that can confound disease association studies with the cohort. This is an example of what can be done with WGS and rare variants coming from a follow up to my PhD thesis project.

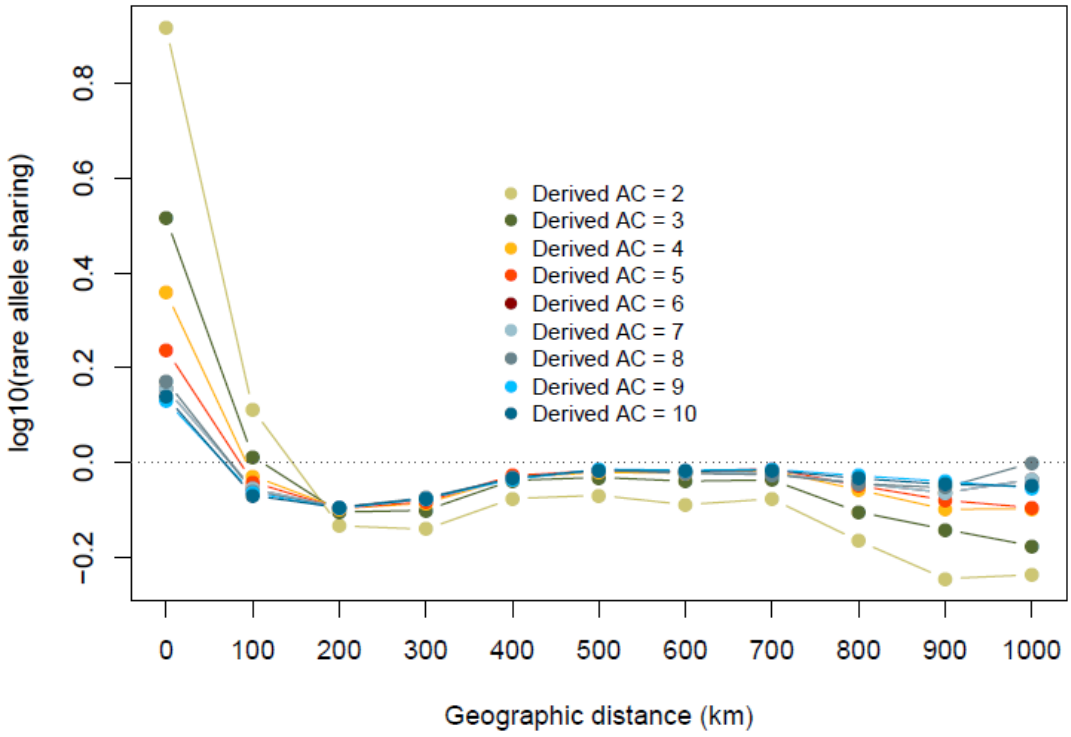


Figure 4.2 – The excess of rare variants as a function of geographical distance for the WGS PREGO dataset. Courtesy of Isabel ALVES.

While SNP arrays provide a more economical way to obtain genome-wide data than WGS, it is important to determine what type of scientific questions can be addressed with rare variants from WGS rather than with SNP array data and LD-based approaches. For questions of population structure and admixture inference, haplotype-based approaches perform reasonably well. SNP arrays are however biased with ascertainment that affects many approaches of demographic inference. Most sequence data used to design SNP arrays developed so far were from people of European ancestry and thus may not interrogate rare variants in other populations very well [8]. The clearest advantage of WGS is the access to the whole allele frequency distribution that allows a reliable demographic inference analysis.

Certain scientific questions may require assembling datasets from different sources. However, this potentially introduces batch effects. While batch effects are well studied in the context of genotyping arrays and can be addressed using quality control (QC) measures, they are less understood and more complex in WGS. When merging multiple datasets, repeating the process of variant calling is necessary which is computationally costly and time-consuming. This constitutes a drawback of rare variant analysis.

4.3 Consequences of presence of a population structure for medical studies

The PREGO dataset has a particular individual ascertainment, with recruited individuals having their 4 grandparents born within a restricted region and relatively homogeneous coverage of the region. Such individual ascertainment in fact maximises the level of population differentiation in a cohort. If we took randomly two samples from the dataset, we would likely obtain non-negligible levels of population structure. As a consequence, the PREGO dataset, while being an excellent resource for population genetics studies, is not a good control group in genetic association studies. Patient samples would not have the same level of population structure, which would likely cause false positive signal mimicking disease association.

An example of cryptic population stratification confounding a genome-wide association study comes from a study of autism [222, 141]. The SNP most significantly associated with the disease was rs4307059. The ancestral allele T was identified as risk variant, with allele frequencies of 0.65 among cases and 0.61 among controls (odds ratio 1.19, p-value $3.4 * 10^{-8}$). However, the frequency of risk variant varies from 0.21 to 0.77 across European populations [38] which is 14-fold more than the difference between cases and controls. The difference in allele frequencies attributed to autism phenotype could thus be explained by subtle difference in the ancestries within Europe. Strikingly, the frequency of the T allele is 1 in the African population. A subsequent analysis did not replicate the association between rs4307059 and autism [224], it has even found the opposite trend, with the minor allele being more frequent among affected persons.

Another example comes from a sample of Native Americans tribes, Pima and Papago, in which there is an association between a Gm haplotype and type 2 diabetes [103]. One might conclude that the absence of this haplotype is a causal risk factor for the disease. However, this haplotype is a marker for European admixture, more often present in the controls, and it does not play a role in type 2 diabetes susceptibility.

Perfect ancestral matching of cases and controls almost never happen. As a resolution to the issue, replication studies are generally suggested. An alternative strategy to address this problem is using family designs to compare genotypes of cases to their healthy relatives.

Methods correcting for population stratification in association studies have been developed (reviewed in [215]), however with a focus on common variants (MAF > 5%). Several studies have shown that population stratification is a confounding factor in rare association studies too and more difficult to account for [214, 138, 10, 95, 121, 155]. For instance, Babron et al [10] showed that applying PCA-based correction method calculated on variants in low frequency category performed poorly. Persyn et al. [164] studied through simulation the impact of fine-scale population structure on rare variant association tests. The PCA-based correction did not totally reduce the inflation of p-values in the presence of very fine structure as seen for example in two close French regions. Adjusting for potential bias must thus be done carefully and there is a need of new methodological developments devoted specifically to rare variant stratification issues in association tests.

On the other hand, studies of population structure may reveal a population isolate that has undergone a bottleneck as we suspect in the case of Mauges region. Such populations have several characteristics that are interesting to a geneticist. Because of reduced genetic diversity due to increased genetic drift, population isolates often show a non-concordant profile of allele frequencies with other non-isolated populations [8, 73]. Some deleterious alleles that are rare in the parent population may drift to higher frequency in the isolate, which empowers the identification of these variants with smaller sample sizes [73]. An association study in an isolate can often be motivated by a suspected higher prevalence of a trait or disease in that particular population. Furthermore, population isolates often demonstrate environmental and cultural homogeneity, resulting in a lack of phenotypic variability which is potentially advantageous for an association study [73]. Iceland is an example of an isolate which has an extensive genealogical and disease history database that contributed to identification of low- and rare-frequency variants associated with sick sinus syndrome [87], gout [205], Alzheimer's disease [97] and prostate cancer [69]. Perhaps the most well-known examples of risk variants found in population isolate are the BRCA1 and BRCA2 mutations which occur at high frequency in the Ashkenazi Jewish population and are associated with risk for breast and

ovarian cancer [116]. Sometimes identified variants are private to a population, for instance a rare variant in LDLR gene (MAF=0.5%) associated with LDL-cholesterol and specific to Sardinia [187]. Also non-isolated populations, but displaying high levels of population structure, likely present high occurrence of rare alleles [78] and may be thus appropriate for gene mapping. It is the case with lactase persistence allele and variants in CFTR gene in Brittany and we predict more instances of increased frequency of possibly pathogenic alleles will tend to happen in areas such as the South-West of France, the Peninsula of Brittany or Mauges.

Appendices

A

Additional figures

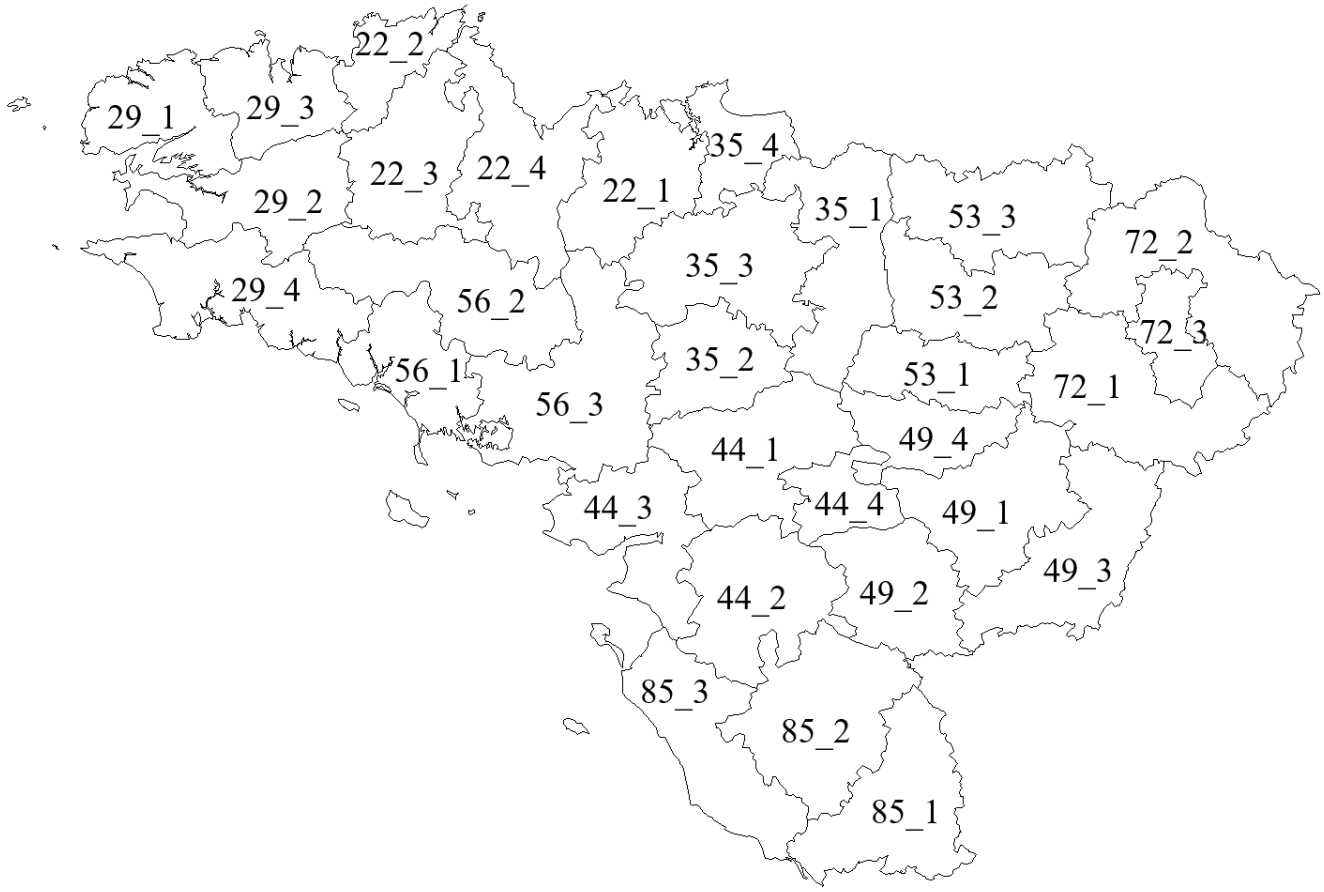


Figure A.1 – Arrondisements of Brittany and Pays de la Loire

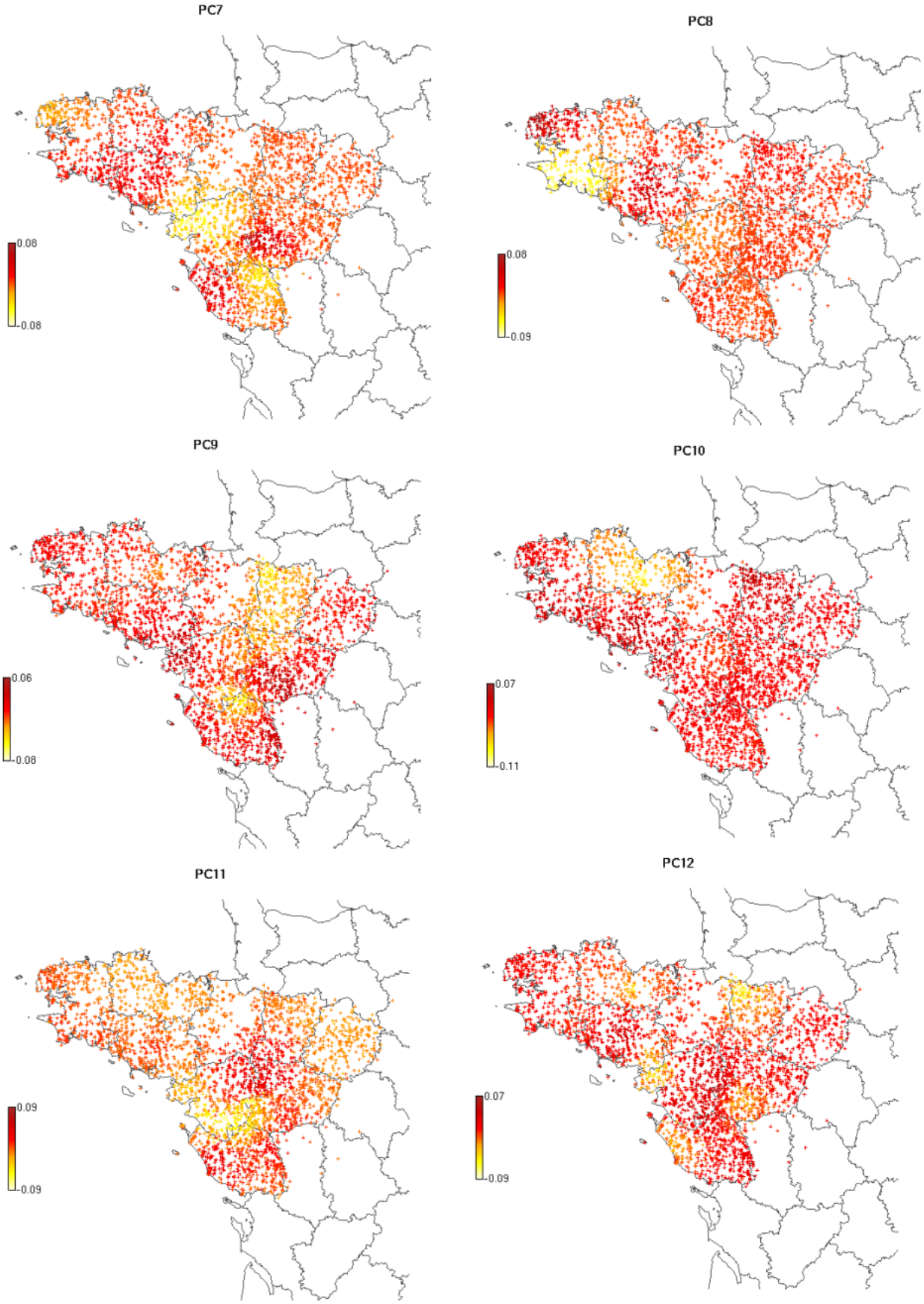


Figure A.2 – Profiles of coancestry-based PC7-PC12 on the map of Western France. Coordinates of individuals correspond to their grandparents' birthplaces.

B

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

In addition to two articles as a first (co-) author, I contributed to publication Raveane et al. [178], which is contained in this appendix. I provided pairwise F_{ST} values among clusters in France (from SU.VI.MAX dataset).

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

A. Raveane[1,2]*†, S. Aneli[2,3,4]*†, F. Montinaro[2,5]*†, G. Athanasiadis[6], S. Barlera[7], G. Birolo[3,4], G. Boncoraglio[8,9], AM. Di Blasio[10], C. Di Gaetano[3,4], L. Pagani[5,11], S. Parolo[12], P. Paschou[13], A. Piazza[3,14], G. Stamatoyannopoulos[15], A. Angius[16], N. Brucato[17], F. Cucca[16], G. Hellenthal[18], A. Mulas[19], M. Peyret-Guzzon[20], M. Zoledziewska[16], A. Baali[21], C. Bycroft[20], M. Cherkaoui[21], C. Dina[22], JM. Dugoujon[17], P. Galan[23], J. Giemza[22], T. Kivisild[5,24], M. Melhaoui[25], M. Metspalu[5], S. Myers[20], LM. Pereira[26], FX. Ricaut[17], F. Brisighelli[27], I. Cardinali[28], V. Grugni[1], H. Lancioni[28], V. L. Pascali[27], A. Torroni[1], O. Semino[1], G. Matullo[3,4]‡, A. Achilli[1]‡, A. Olivieri[1]‡, C. Capelli[2]*‡

1 Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy.

2 Department of Zoology, University of Oxford, Oxford, UK.

3 Department of Medical Sciences, University of Turin, Turin, Italy.

4 IIGM (Italian Institute for Genomic Medicine), Turin.

5 Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Estonia.

6 Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark.

7 Department of Cardiovascular Research, Istituto di Ricovero e Cura a Carattere Scientifico–Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy.

8 Department of Cerebrovascular Diseases, IRCCS Istituto Neurologico Carlo Besta, Milan, Italy.

9 PhD Program in Neuroscience, University Milano-Bicocca, Monza, Italy.

10 Center for Biomedical Research & Technologies, Italian Auxologic Institute IRCCS, Milan, Italy.

11 APE lab, Department of Biology, University of Padua, Padua, Italy.

12 Computational Biology Unit, Institute of Molecular Genetics, National Research Council, Pavia, Italy.

13 Department of Biological Sciences, Purdue University, USA; 14 Academy of Sciences, Turin, Italy.

15 Department of Medicine and Genome Sciences, University of Washington, Seattle, WA.

16 Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy.

17 Evolutionary Medicine Group, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse, Centre National de la Recherche Scientifique (CNRS), Université de Toulouse, Toulouse, France.

18 University College London Genetics Institute (UGI), University College London, London, UK.

19 Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Lanusei, Italy.

20 The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.

21 Faculté des Sciences Semlalia de Marrakech (FSSM), Université Cadi Ayyad, Marrakech, Morocco.

22 l'institut du thorax, INSERM, CNRS, University of Nantes, Nantes, France.

23 Equipe de Recherche en Epidémiologie Nutritionnelle (EREN), Centre de Recherche en Epidémiologie et Statistiques, Université Paris 13/Inserm U1153/Inra U1125/ Cnam, COMUE Sorbonne Paris Cité, F-93017 Bobigny, France.

24 Division of Biological Anthropology, University of Cambridge, Cambridge, UK.

25 Faculté des Sciences, Université Mohammed Premier, Oujda, Morocco.

26 Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal.

27 Section of Legal Medicine, Institute of Public Health, Catholic University of the Sacred Heart, Rome, Italy.

28 Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy.

Author List Footnotes

* corresponding author

† These authors contributed equally to this work

‡ Co-senior authors

Contact Info

* Correspondence: alessandro.raveane01@universitadipavia.it (A.R.), serena.aneli@gmail.com (S.A.), francesco.montinaro@gmail.com (F.M.) and cristian.capelli@zoo.ox.ac.uk (C.C.).

One sentence summary. Ancient and historical admixture events shaped the genetic structure of modern-day Italians, the ancestry profile of Southern European populations and the continental distribution of Neanderthal legacy.

Abstract

European populations display low genetic diversity as the result of long term blending of the small number of ancient founding ancestries. However it is still unclear how the combination of ancient ancestries related to early European foragers, Neolithic farmers and Bronze Age nomadic pastoralists can fully explain genetic variation across Europe. Populations in natural crossroads like the Italian peninsula are expected to recapitulate the overall continental diversity, but to date have been systematically understudied. Here we characterised the ancestry profiles of modern-day Italian populations using a genome-wide dataset representative of modern and ancient samples from across Italy, Europe and the rest of the world. Italian genomes captured several ancient signatures, including a non-steppe related substantial ancestry contribution ultimately from the Caucasus.

Differences in ancestry composition as the result of migration and admixture generated in Italy the largest degree of population structure detected so far in the continent and shaped the amount of Neanderthal DNA present in modern-day populations.

Introduction

Our understanding of the events that shaped European genetic variation has been redefined by the availability of ancient DNA (aDNA). In particular, it has emerged that, in addition to the contributions of early hunter-gatherer populations, major genetic components can be traced back to Neolithic (*1*–*4*) and Bronze Age expansions (*3*, *5*).

The arrival of farming in Europe from Anatolia led to a partial replacement via admixture of autochthonous and geographically structured hunter-gatherers, a process that generated individuals genetically close to present-day Sardinians (*2*, *4*, *6*, *7*). During the Bronze Age the dispersal of a population related to the pastoralist nomadic Yamnaya from the Pontic-Caspian steppe area dramatically impacted the genetic landscape of the continent, particularly of Northern and Central Europe (*3*, *5*, *8*). This migration, supported by archaeological and genetic data, has also been putatively linked to the spread of the Indo-European languages in Europe and the introduction of several technological innovations in peninsular Eurasia (*9*). Genetically, ancient steppe populations have been described as a combination of Eastern and Caucasus Hunter Gatherer/Iran Neolithic ancestries (EHG and CHG/IN) (*6*), whose genetic signatures in the population of Central and Northern Europe were introduced via admixture. However, the analysis of aDNA from Southern East Europe identified the existence of additional contributions ultimately from the Caucasus (*10*, *11*) and suggested a more complex ancient ancestry composition for Europeans (*6*).

The geographic location of Italy, enclosed between continental Europe and the Mediterranean Sea, makes the Italian people relevant for the investigation of continent-wide demographic events, to complement and enrich the information provided by aDNA studies. In order to characterise the ancestry profile of modern-day populations and test the validity of the three-ancestries model

across Europe (related to early European foragers, Neolithic farmers and Bronze Age nomadic pastoralists), we characterised the genetic variability of present-day Italians and other Europeans in terms of their ancient ancestry composition as the result of migration and admixture. In doing so, we assembled and analyzed a comprehensive genome-wide SNP dataset composed by 1,616 individuals from all the 20 Italian administrative regions and more than 140 worldwide reference populations, for a total of 5,192 modern-day samples (fig. S1, table S1), to which we added genomic data available for ancient individuals (data file S1).

Results

Distinctive genetic structure in Italy

We initially investigated patterns of genetic differentiation in Italy and surrounding regions by exploring the information embedded in SNP-based haplotypes of modern samples (Full Modern Dataset, FMD, including 218,725 SNPs). The phased genome-wide dataset was analysed using the CHROMOPAINTER (CP) and fineSTRUCTURE (fS) pipeline (*12*, *13*) (Supplementary materials) to generate a tree of groups of individuals with similar "copying vectors" (clusters, Fig. 1A). The fraction of pairs of individuals placed in the same cluster across multiple runs was on average 0.95 for Italian clusters and 0.96 across the whole set of clusters (see Materials and Methods, Supplementary materials). Related non-European clusters were merged into larger groups in subsequent analyses (see Materials and Methods, Supplementary materials).

Italian clusters separated into three main groups: Sardinia, Northern (North/Central-North Italy) and Southern Italy (South/Central-South Italy and Sicily); the former two were close to populations originally from Western Europe, while the latter was in proximity of Middle East groups (Fig. 1A, fig. S2, data file S2). The cluster-composition of the administrative regions of Italy provided further evidence for geographic structuring (Fig. 1B) with the separation between Northern and Southern areas being shifted North along the peninsula; the affinity to Western and Middle Eastern populations was also evident in the haplotype-based PCA (Fig. 1C), allele frequency PCA (fig. S3) and the ADMIXTURE analysis (fig. S4).

These observations were replicated using a subset of the dataset genotyped for a larger number of SNPs (High Density Dataset, HDD, including 591,217 SNPs; see Materials and Methods, Supplementary materials, Fig. 1B, table S1). Recent migrants and admixed individuals, as identified on the basis of their copying vectors (fig. S5, fig. S6, table S2), were removed in subsequent CP/fS analyses (see Supplementary materials).

We explored the degree of within-country differentiation by comparing the distribution of F_{ST} values among fS genetic clusters in Italy with the ones in several European countries (*13–16*) and across the whole of Europe. Clusters within Italy were significantly more different from each other than within any other country here included (median Italy: 0.004, data file S3; range medians for listed countries 0.0001-0.002) and showed differences comparable with estimates across European clusters (median European clusters: 0.004, Fig. 1D, see Materials and Methods, Supplementary materials). The analysis of the migration surfaces (EEMS) (*17*) highlighted several barriers to gene flow within and around Italy but also suggested the existence of migration corridors in the southern part of the Adriatic and Ionian Sea, and between Sardinia, Corsica and continental Italy (Fig. 1E; fig. S7) (*11*).

Multiple ancient ancestries in Italian clusters

We investigated the ancestry composition of modern clusters by testing different combination of ancient samples using the CP/NNLS pipeline, a previously implemented analysis that reconstructs the profiles of modern populations as the combination of the "painted" profiles of different ancient samples by using a "mixture fit" approach based on a non-negative least square algorithm (NNLS) (*13*, *18*, *19*). We applied this approach to ancient samples using the unlinked mode implemented in CP, similarly to other routinely performed analyses based on unlinked markers or allele frequency, such as qpAdm and ADMIXTURE. In addition, data from modern individuals (FMD) were harnessed as donor populations (see Materials and Methods, Supplementary materials). Following Lazaridis et. al 2017 (*10*), we performed two separate CP/NNLS analyses, *"Ultimate"* and *"Proximate"*, referring to the least and the most recent putative sources, respectively (Fig. 2, fig. S8, fig. S9). In the *Ultimate* analysis, 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; Fig. 2A-C, fig. S8A), with Sardinians showing values above 80%. A closer affinity of Northern Italian than Southern

Italian clusters to AN was also supported by D-statistics (fig. S10). 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 (higher estimates in SItaly1 ~13% and SItaly3 ~ 24% for WHG and CHG respectively), while the latter was more common in Northern clusters (NItaly6 = 15%). These observations suggest the existence of different secondary sources contributions to the two edges of the peninsulas, 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.

North-South differences across Italy were also detected in the *Proximate* analysis. When *Proximate* sources were evaluated, SBA contribution ranged between 33% in the North and 6% in the South of Italy, while ABA (Anatolia Bronze Age) showed an opposite distribution (Fig. 2D-F, fig. S9), in line with the results based on the D statistics (fig. S10, fig. S11), and mirroring the EHG and CHG patterns, respectively. Contrary to previous reports, the occurrence of CHG as detected by the CP/NNLS analysis did not mirror the presence of Steppe Bronze Age (SBA), with several populations testing positive for the latter but not for the former ((*6*), Fig. 2, fig. S8). 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.

Interestingly, clusters belonging to the North had more EEN (European Early Neolithic) than Southern ones, which in turn were composed by an higher fraction of ABA, although the high AN-related component in both these ancient groups might have affected the exact source identification. The relevance of ABA in Italy was additionally supported by the reduced fit of the NNLS (sum of the squared residuals; Materials and Methods, Supplementary materials) when the *Proximate* analysis was run excluding ABA. Results were similar to the full *Proximate* analysis for most of

the European clusters, but not for Southern European groups, where the residuals were almost up to twice as much when ABA was not included as a source (Fig. 2G). A similar behaviour, but for Northern Italian and most of the European clusters, was observed when SBA was removed from the panel of *Proximate* sources (Fig. 2H). The closer affinity of the Southern Italian clusters to ABA was also highlighted by the PCA and ADMIXTURE analysis on ancient and modern samples (Fig. 2I, fig. S12, fig. S13, fig. S14) and significantly higher ABA ancestry in Southern than Northern Italy, as estimated by NNLS analysis (Fig. 2D, Student's t-Test p-value < 0.05, Supplementary materials). We also noted that in the Balkan peninsula signatures related to ABA were present but less evident than in Southern Italy across modern-day populations, possibly masked by historical contributions from Central Europe (*20*, *21*) (Fig. 2, Fig. 3, fig. S8B). Overall, SBA and ABA appear to have very different distribution patterns in Europe: continent-wide the former, more localised (in the South) the latter. Similar results were obtained when other Southern European ancient sources replaced ABA in the Proximate analysis (fig. S9, Materials and Methods, Supplementary materials). These results were confirmed by qpAdm analysis. When two sources were evaluated, a large AN contribution was supported only in one cluster (SItaly2), while the vast majority of supported models included ABA, Minoan or Mycenaean and one of the hunter-gatherer groups or SBA (table S3, table S4). When three possible sources were allowed, AN was supported for all the Southern Italian clusters, mostly in association with EHG/WHG/SBA and CHG/IN. Nevertheless, all the analysed 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 (fig. S15, table S3, table S4). North African contributions, ranging between 3.8% (SCItaly1) to 14.5% (SItaly1) became evident when combinations of five sources were tested. 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 (Materials and Methods, Supplementary materials, table S4). In

order to obtain insights about the relationship between ancient and modern groups, we performed the same qpAdm analysis on post-Neolithic/Bronze Age Italian individuals (fig. S15, table S5). Iceman and Remedello, the oldest Italian samples here included (3,400-2,800 BCE, Before Current Era), were composed by high proportions of AN (74 and 85%, respectively). The Bell Beaker samples of Northern Italy (2,200-1,930 BCE) were modelled as ABA and AN + SBA and WHG, although ABA was characterised by large standard errors but the detection of Steppe ancestry, at 14%, was more robust. On the other hand Bell Beaker samples from Sicily (2,500-1,900 BCE) were modelled almost exclusively as ABA, with less than 5% SBA. 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. Differences across ancient Italian samples were also supported by their projections on the PCA of modern-day data (Fig. 2I). Remedello and Iceman clustered with European Early Neolithic samples, together with one of the three Bell Beaker individuals from North Italy, as previously reported (*22*), 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 Italian samples (Fig. 2I).

Historical admixture

In order to investigate the role of historical admixture events in shaping the modern distribution of ancient ancestries, we generated the admixture profiles of Italian and European populations using GLOBETROTTER (GT, (*21*)) (Fig. 3, fig. S16, table S6, table S7).

We discussed here the results based on the full modern dataset (FMD) as it provided a wider coverage at population level.

We run the analysis excluding the Italians as donors in order to reduce copying between highly similar groups (GT "noItaly" analysis; Fig. 3). The events detected in Italy occurred mostly between

1,000 and 2,000 years ago (ya), and extended to 2,500ya in the rest of Europe (Fig. 3A and fig. S16). Clusters from Caucasus and North-West Europe were identified all across Italy as best-proxies for the admixing sources, while Middle Eastern and African clusters were identified as best proxies only in Southern Italian clusters and Sardinia (Fig. 3B, C). We noted that when we extended the search for the best-proxies to include also Italian clusters, these were as good as or better proxies than clusters from the Caucasus and the Middle East. On the other hand, North-West European and African clusters were usually still better proxies than groups from any other area (Fig. 3B, C). Notably, Eastern and Middle Eastern clusters were not detected as best proxies when we run the GT analysis including all clusters as donors, contrary to African, European and Italian groups ("GTall" analysis; table S6). Overall these results supported a scenario in which gene flow mostly occurred between resident Italian sources and non-Italian sources. SBA and ABA ancestries were detected in Italian and non-Italian best-proxies (Fig. 2D, Fig. 3, table S6, table S7), which suggests that part of these ancestries arrived from outside Italy in historical times, but also that these components were already present in Italian groups at the time of these admixture events. Episodes of gene flow were also detected in Sardinia, combining signals from both the African continent and North West Europe. MALDER results for the more recent episodes replicated the admixture pattern identified by GT (fig. S16, table S8).

The Neanderthal legacy across Italy and Europe

The variation in ancestry composition reported across Italy and Europe is expected to influence other aspects of the genetic profiles of European populations, including the presence of archaic genetic material (*6*). We investigated the degree of Neanderthal ancestry in Italian and other Eurasian populations by focusing on SNPs tagging Neanderthal introgressed regions (*23*, *24*). SNPs were pruned for LD and a final set of 3,969 SNPs was used to estimate the number of Neanderthal alleles in samples genotyped for the Infinium Omni2.5-8 Illumina beadchip. Asian and Northern European populations had significantly more Neanderthal alleles than European and Southern

European groups respectively, as previously reported (*25–28*), with significant differences also highlighted within Italy (Fig. 4A, B). Contributions from African groups possibly influenced these patterns, particularly in Southern European populations (*20*) (Fig. 2, Fig. 3). However differences within Europe and Italy were still present once individuals belonging to clusters with African contributions were removed (fig. S17, see Materials and Methods, Supplementary methods). Ancient samples have been reported to differ in the amount of Neanderthal DNA due to variation in the presence of a so-called "Basal Eurasian" lineage, stemming from non-Africans before the separation of Eurasian groups and harbouring only a negligible fraction of Neanderthal ancestry (*6*). Consistent with this (*6*), we found the estimated amounts of Basal Eurasian and Neanderthal to be negatively correlated across modern day European clusters (Fig. 4C, fig. S18, fig. S19), irrespective of the removal of all the clusters admixed with African sources (see Materials and Methods, Supplementary materials; fig. S17).

The variation in Neanderthal ancestry was also reflected at specific loci. A total of 144 SNPs were identified among the Neanderthal-tag SNPs showing the largest differences in allelic frequency in genome-wide comparisons across Eurasian and African populations (see Materials and Methods, Supplementary materials - Neanderthal-Tag SNPs within the Top 1% of the genome-wide distributions of each of the 55 pairwise population comparisons - NTT SNPs; fig. S20). The top 1% of each distribution was significantly depleted in Neanderthal SNPs (see Materials and Methods, Supplementary materials, table S9), in agreement with a scenario of Neanderthal mildly deleterious variants being removed more efficiently in human populations (*29–31*).

The 50 genes containing NTT SNPs were enriched for phenotypes related to facial morphology, body size, metabolism and muscular diseases (see Materials and Methods, Supplementary materials, data file S4). A total of 34 NTT SNPs were found to have at least one known phenotypic association (*32*, *33*) (data file S4). Among these, we found Neanderthal alleles associated with increased gene expression in testis and in skin after sun exposure (SNPs within the *IP6K3* and

ITPR3 genes), susceptibility to cardiovascular and renal conditions (*AGTR1*), and Brittle cornea syndrome (*PRDM5*) (*24*). NTT SNPs between European and Asian/African populations included previously reported variants in *BNC2* and *SPATA18* genes (*23*, *34*, *35*) (see Materials and Methods, Supplementary materials, Fig. 4D), while 80 NTT SNPs were involved in at least one comparison between Northern (CEU, GBR and FIN) and Southern European populations (IBS and Italian groups). Among these SNPs, three 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 (Fig. 4E, F). Ten SNPs showed an opposite frequency gradient: seven mapped to one Neanderthal introgressed region spanning the *OR51F1*, *OR51F2* and *OR52R1* genes (Fig. 4E, F), and the other three identified regions hosting the *AKAP13* gene, within one of the high frequency European Neanderthal introgressed haplotypes recently reported (*36*) (Fig. 4E, F).

Discussion

The pattern of variation reported across Italian groups appears geographically structured in three main regions: Southern and Northern Italy and Sardinia. The North-South division in particular appeared as shaped by the distribution of Bronze Age ancestries with signatures of different continental hunter-gatherer groups. The results of the analyses of both modern and ancient data suggest that ancestries related to Caucasus and Eastern hunter-gatherers were possibly initially brought in Italy by at least two different contributions from the East. Of these, one is the well-characterised SBA signature ultimately associated with the nomadic groups from the Pontic-Caspian steppes. This component entered Italy from mainland Europe and was present in the peninsula in the Bronze Age, as suggested by its presence in Bell Beaker samples from North Italy (table S5). SBA ancestry continued to arrive from the continent up until historical times (Fig. 3). The other contribution is ultimately associated with CHG ancestry and affected predominantly the South of Italy, where it now represents a substantial component of the ancestry profile of local

304 populations. This signature is still uncharacterised in terms of precise dates and origin; however
305 such ancestry was possibly already present during the Bronze Age in Southern Italy (table S5) and
306 was further supplemented by historical events (Fig. 3).
307 The very low presence of CHG signatures in Sardinia and in older Italian samples (Remedello and
308 Iceman) but the occurrence in modern-day Southern Italians might be explained by different
309 scenarios, not mutually exclusive: 1) population structure among early foraging groups across Italy,
310 reflecting different affinities to CHG; 2) the presence in Italy of different Neolithic contributions,
311 characterised by different proportion of CHG-related ancestry; 3) the combination of a post-
312 Neolithic, prehistoric CHG-enriched contribution with a previous AN-related Neolithic layer; 4) A
313 substantial historical contribution from Southern East Europe across the whole of Southern Italy.
314 No substantial structure has been highlighted so far in pre-Neolithic Italian samples (*8*). An arrival
315 of the CHG-related component in Southern Italy from the Southern part of the Balkan Peninsula is
316 compatible with the identification of genetic corridors linking the two regions (Figure 1E, (*11*)) and
317 the presence of Southern European ancient signatures in Italy (Figure 2). The temporal appearance
318 of CHG signatures in Anatolia and Southern East Europe in the Late Neolithic/Bronze Age suggests
319 its relevance for post-Neolithic contributions (*37*). Additional analyses of aDNA samples from
320 around this time in Italy are expected to clarify what scenario might be best supported.
321 Historical events possibly involving continental groups at the end of Roman Empire and African
322 contributions following the establishment of Arab kingdoms in Europe around 1,000 ya (*20*, *21*,
323 *38–40*) played a role in further shaping the ancestry profiles of the Italian populations.
324 Despite Sardinia was confirmed as being the most closely related population to Early European
325 Neolithic farmers (Figure 2D, I), there is no evidence for a simple genetic continuity between the
326 two groups. Sardinia, and the rest of Italy, experienced in fact historical episodes of gene-flow (*4*)
327 (Fig. 2, Fig. 3, table S3, table S4) that contributed to the further dispersal of ancient ancestries and
328 the introduction of other components, including African ones.

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 (*24, 27, 41*). 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 (*42*). The spatial heterogeneity of Neanderthal legacy within Europe here reported appears as the result of ancient and historical events which 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 (*30, 31*).

The variation in ancestry composition detected across Italy extends to neighbour regions and appears to combine historical contributions and ancient stratification. The differences between Northern and Southern Italian populations are possibly reflecting long-term differential links with Central and Southern Europe respectively, with additional contributions from the African continent for the Southern part of Italy and Sardinia.

The multifaceted admixture profile here sketched provides an interpretative framework for the processes that have shaped Southern European genetic variation. The inclusion of ancient samples spanning diachronic and geographic transects from the Italian peninsula and nearby regions will help in clearing up further questions about the temporal and spatial dynamics of these processes.

Materials and Methods

Analysis of modern samples

Dataset. Two hundred and twenty-four samples are here present for the first time. Of these, 167 Italians and 6 Albanians were specifically selected and sequenced for this project with two versions

(1.2 and 1.3) of the Infinium Omni2.5-8 Illumina beadchip, while 57 additional Italians and Europeans were previously sequenced with Illumina 660W and are presented here for the first time (Supplementary materials, table S1). Two separate world-wide datasets were prepared. The Full Modern Dataset (FMD) included 4,852 samples (1,589 Italians) and 218,725 SNPs genotyped with Illumina arrays; the High Density Dataset (HDD) contained 1,651 samples (524 Italians) and 591,217 SNPs genotyped with the Illumina Omni array (Supplementary materials).

The merging, the removal of ambiguous C/G and A/T and triallelic markers, the exclusion of related individuals and the discarding of SNPs in linkage disequilibrium (LD) were performed using PLINK1.9 (*43*, *44*). Only autosomal markers were considered.

Haplotype analysis (CHROMOPAINTER, CP, and fineSTRUCTURE, fs). Phased haplotypes were generated using SHAPEIT(*45*) and applying the HapMap b37 genetic map.

CP was employed to generate a matrix of recipient individuals "painted" as a combination of donor samples (copying vector). Three runs of CP were done for each dataset generating three different outputs: (i) a matrix of all the individuals "painted" as a combination of all the individuals, for cluster identification and GT analysis; (ii) a matrix of all Italians as a combination of all Italians, for F_{ST} analysis; (iii) a matrix of all the samples as a combination of all the other samples but excluding Italians, for "local" GT analysis.

Clusters were inferred using fineSTRUCTURE (fS). After an initial search based on the "greedy" mode, the dendrogram was processed by visual inspection (*18*, *20*) according to the geographical origin of the samples. The robustness of the cluster was obtained by processing the MCMC pairwise coincidence matrix (Supplementary materials).

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 (Supplementary materials).

Principal Component Analysis (PCA). PCA was performed on CP chunkcount matrix (Supplementary materials) and was generated using the prcomp() function on R software (*46*). Allele frequencies PCA was performed using smartpca implemented in the EIGENSOFT(*47*) after pruning the datasets for LD.

Characterization of the migration landscape (EEMS analysis). Estimated Effective Migration Surfaces analysis (EEMS) (*17*) was performed estimating the average pairwise distances between population using bed2diffs tool and the resulting output was visualised by using the Reems package (*17*).

ADMIXTURE analysis. ADMIXTURE1.3.0 software (*48*) was used performing 10 different runs using a random seed. The results were combined with CLUMPP (*49*) using the largeKGreedy algorithm and random input orders with 10,000 repeats. *Distruct* implemented in CLUMPAK (*50, 51*) was then used to identify the best alignment of CLUMPP results. Results were processed using R statistical software (*46*) .

F_{ST} estimates among clusters. Pairwise F_{ST} estimates among newly generated Italian clusters and among originally generated European clusters (Supplementary materials) were inferred using smartpca software implemented in the EINGESOFT package (*47*). Comparisons between the F_{ST} distributions were performed using a Wilcoxon rank sum test in R programming language environment.

The time and the sources of admixture events (GT analysis and MALDER analysis). Times of haplotype-dense data admixture events were investigated using GLOBETROTTERv2 software. GT was employed using two approaches: complete and non-local (referred as “noItalian”, Supplementary materials), in default modality (*13, 20, 52*). The difference between the two approaches was the inclusion or the exclusion respectively of all the Italian clusters as donors in the CP matrix used as input file. To improve the precision of the admixture signals, “null.ind 1” parameter was set (*52*). Unclear signals were corrected using the default parameters and a total of

100 bootstraps were performed. MALDER uses allele frequencies to dissect the time of admixture signals. The best amplitude was identified and used to calculate a Z-score (Supplementary materials). A Z-score equal or lower than 2 identifies not significantly different amplitude curves (*53*, *54*) (Supplementary materials).

Sources for both GT and MALDER were grouped in different ancestries as indicated in the legend of Fig. 3, fig. S16.

The expression (1950 – (g +1)* 29), where g is the number of generation, was used to convert into years the GT and MALDER results, negative numbers were preceded by BCE (Before Current Era) letters.

Analyses including ancient samples

Dataset. In order to explore the extent to which the European and Italian genetic variation has been shaped by ancient demographic events, we merged modern samples from FMD with 63 ancient samples selected from recent studies (*6*, *7*, *10*, *22*, *37*, *55–57*) (data file S1).

Principal Component Analysis (PCA). We performed two principal components analyses with the EIGENSOFT (*47*) smartpca software and the *"lsqproject"* and *"shrinkmode"* option, projecting the ancient samples on the components inferred from modern European, West Asian and Caucasian individuals and, then, only on modern European clusters. In order to evaluate the potential impact of DNA damage in calling variants from aDNA samples, we repeated the PCA with the 63 ancient samples and modern European, Caucasian and West Asian samples by removing transition polymorphisms and recorded significant correlations for the localisation of ancient samples along PC1 and PC2 ($r > 0.99$, p-value < 0.05).

ADMIXTURE analysis. We projected the ancient samples on the previously inferred ancestral allele frequencies from 10 ADMIXTURE (*48*) runs on modern samples (see "Analysis of modern

samples" section and Supplementary materials). We used CLUMPP(*49*) for merging the resulting matrices and *distruct (51)* for the visualization.

D-STATISTICS. We tested for admixture using the D-statistics as implemented in the qpDstat tool in the software ADMIXTOOLS v4.2 (*58*). We performed the D-statistic analyses evaluating the relationship of Italian cluster with AN, ABA and SBA. In details, we performed the the D-statistics D(Ita1,Ita2,AN/ABA/SBA,Mbuti) where Ita1 and Ita2 are the different clusters composed mainly by italian individuals as inferred by fineStructure.

CHROMOPAINTER (CP)/Non-Negative Least Squares (NNLS) analysis. We used an approach based on the software CP (*12*, *59*) and a slight adaptation of the non-negative least square (NNLS) function (*13, 18, 19*) to estimate the proportions of the genetic contributions from ancient population to our modern clusters. We run CP using the "unlinked" mode (*55*) and the same Ne and θ parameters of the modern dataset and we painted both modern and ancient individuals, using only modern samples as donors (*55, 56*). Then we "inverted" the output of CP by solving an appropriately formulated NNLS problem, producing a painting of the modern clusters in terms of the ancients. We applied this combined approach on different sets of ancient samples (*Ultimate* and various combinations of *Proximate* sources).

The goodness of fit of the NNLS was measured evaluating the residuals of the NNLS analysis. In details, we focused on the Proximate sources, and compared the sum of squared residuals when ABA or SBA were included/excluded as putative sources.

qpAdm analysis. We used the ancestral reconstruction method qpAdm, which harnesses different relationships of populations related to a set of outgroups (**eg.** f4[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..5}, and one set of right/left Outgroups (Supplementary materials).

For each of the tested combinations we used qpWave to evaluate if the set of chosen outgroups is able to I) discriminate the combinations of sources and II) 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 and reported it and the associated standard errors in Supplementary table S3 and table S4. In addition, we performed the same analysis for Iceman, Remedello and Bell Beaker individuals from Sicily and North Italy (table S5).

Archaic contribution

Dataset. We assembled an additional high density dataset by retaining only samples genotyped on the Illumina Infinium Omni2.5-8 BeadChip from our larger modern dataset. In particular, we included seven populations from the 1000 Genomes Project: the five European populations (Northern European from Utah - CEU, England - GBR, Finland - FIN, Spain - IBS, Italy from Tuscany - TSI), one from Asia (Han Chinese - CHB) and one from Africa (Yoruba from Nigeria - YRI). We also retained 466 Italian samples, whose four grandparents were born in the same Italian region. The Italian samples were broadly clustered according to their geographical origin into Northern (ITN), Central (ITC), Southern (ITS) Italians and Sardinians (SAR), while TSI samples from 1000 Genome Project formed a separate cluster (table S10).

From this dataset, we extracted 7,164 Neanderthal SNPs tagging Neanderthal introgressed regions (*24*). In order to select which allele was inherited from Neanderthals, we chose the one from the Altai Neanderthal (*41*) genome when it was homozygous and the minor allele in YRI when it was heterozygous.

Number of Neanderthal alleles in present-day human populations. After pruning variants in linkage disequilibrium, we counted the number of Neanderthal alleles considering all the tag-SNP across all samples. Then, we compared the distribution of Neanderthal allele counts across

475 populations with the two-sample Wilcoxon rank sum test. We repeated the same analyses after
476 removing outlier individuals.

477 **Basal Eurasian ancestry and Neanderthal contribution.** In order to infer the proportion of Basal
478 Eurasian present in European populations (*6*, *7*), we used the f4 ratio implemented in the
479 ADMIXTOOLS package (*58*) in the form f4(Target, Loschbour, Ust_Ishim, Kostenki14)/
480 f4(Mbuti, Loschbour, Ust_Ishim, Kostenki14). We repeated this approach to infer the Neanderthal
481 ancestry, in the form f4 (Mbuti, Chimp Target, Altai)/ f4(Mbuti, Chimp, Dinka, Altai) (fig. S18,
482 fig. S19). We then performed the same analyses by grouping the modern individuals according to
483 the CP/fS inferred clusters ("Analysis of modern samples" section) and retained only clusters with
484 at least 10 samples (Fig. 4)

485 **African ancestry and Neanderthal legacy.** The impact of African contributions in shaping the
486 amount of Neanderthal occurrence was evaluated by exploring how the removal of the clusters
487 showing African gene-flow as detected by GT analysis (Fig. 3) and how individuals belonging to
488 these clusters affected the correlation between Basal Eurasian/Neanderthal estimates and the degree
489 of population differentiation in the amount of Neanderthal alleles, respectively (Supplementary
490 materials; fig. S17).

491 **Comparison of Neanderthal allele frequencies across modern populations.** We computed the
492 allele frequency differences for every SNPs for each of the possible pairs of the eleven populations
493 in our dataset, thus obtaining 55 distributions (Supplementary materials). Then, we selected the
494 NTT SNPs, i.e. the Neanderthal-Tag SNPs in the Top 1% of each distribution (data file S4).

495 **The biological implications of Neanderthal introgression.** Given the list of genes overlapping
496 the Neanderthal introgressed regions harbouring the NTT SNPs and the list of genes directly
497 harbouring the NTT SNPs, we performed different enrichment tests with the online tool EnrichR
498 (*60*, *61*). Particularly, we searched for significant enrichments compared to the human genome
499 using the EnrichR collection of database, e.g. dbGaP (*62*, *63*), Panther 2016 (*64*), HPO (*65*) and

KEGG 2016 (*66–68*) (data file S4). We then investigated known direct associations between the Neanderthal alleles of the NTT SNPs and phenotypes, by looking in the GWAS and PheWAS catalogues (*32*, *33*) and by applying the PheGenI tool (*69*) (Supplementary Data 5). We used the circos representation as in Kanai et al. (*70*), to highlight different sets of NTT SNPs (Figure 4F).

References

1. I. Lazaridis *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. **513**, 409–413 (2014).

2. T. Günther *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 (2015).

3. W. Haak *et al.*, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*. **522**, 207–211 (2015).

4. C. W. K. Chiang *et al.*, Genomic history of the Sardinian population. *Nat. Genet.* **50**, 1426–1434 (2018).

5. M. E. Allentoft *et al.*, Population genomics of Bronze Age Eurasia. *Nature*. **522**, 167–172 (2015).

6. I. Lazaridis *et al.*, Genomic insights into the origin of farming in the ancient Near East. *Nature*. **536**, 419–424 (2016).

7. Q. Fu *et al.*, The genetic history of Ice Age Europe. *Nature*. **534**, 200–205 (2016).

8. E. R. Jones *et al.*, Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).

9. D. W. Anthony, *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World* (Princeton University Press, 2010).

10. I. Lazaridis *et al.*, Genetic origins of the Minoans and Mycenaeans. *Nature*. **548**, 214–218 (2017).

11. P. Paschou *et al.*, Maritime route of colonization of Europe. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9211–9216 (2014).

12. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).

13. S. Leslie *et al.*, The fine-scale genetic structure of the British population. *Nature*. **519**, 309–314 (2015).

14. G. Athanasiadis *et al.*, Nationwide Genomic Study in Denmark Reveals Remarkable Population Homogeneity. *Genetics*. **204**, 711–722 (2016).

15. R. P. Byrne *et al.*, Insular Celtic population structure and genomic footprints of migration. *PLoS Genet.* **14**, e1007152 (2018).

16. C. Bycroft *et al.*, Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula (2018), , doi:10.1101/250191.

17. D. Petkova, J. Novembre, M. Stephens, Visualizing spatial population structure with estimated

effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).

18. F. Montinaro *et al.*, Unravelling the hidden ancestry of American admixed populations. *Nat. Commun.* **6**, 6596 (2015).

19. C. L. Lawson, R. J. Hanson, *Solving Least Squares Problems* (SIAM, 1995).

20. G. B. J. Busby *et al.*, The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Curr. Biol.* **25**, 2518–2526 (2015).

21. G. Hellenthal *et al.*, A genetic atlas of human admixture history. *Science*. **343**, 747–751 (2014).

22. I. Olalde *et al.*, The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*. **555**, 190–196 (2018).

23. B. Vernot, J. M. Akey, Resurrecting surviving Neandertal lineages from modern human genomes. *Science*. **343**, 1017–1021 (2014).

24. C. N. Simonti *et al.*, The phenotypic legacy of admixture between modern humans and Neandertals. *Science*. **351**, 737–741 (2016).

25. K. Prüfer *et al.*, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*. **358**, 655–658 (2017).

26. F .Arcuri *et al.*, . (2016). Influssi balcanici e genesi del Bronzo antico in Italia meridionale: la koinè Cetina e la facies di Palma Campania. Riv. di Sci. Preist**. LXVI**, 77–95 (*2016*).

27. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–722 (2010).

28. B. Vernot *et al.*, Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*. **352**, 235–239 (2016).

29. S. Castellano *et al.*, Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6666–6671 (2014).

30. K. Harris, R. Nielsen, The Genetic Cost of Neanderthal Introgression. *Genetics*. **203**, 881–891 (2016).

31. I. Juric, S. Aeschbacher, G. Coop, The Strength of Selection against Neanderthal Introgression. *PLoS Genet.* **12**, e1006340 (2016).

32. J. C. Denny *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 (2013).

33. J. MacArthur *et al.*, The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

34. M. Dannemann, K. Prüfer, J. Kelso, Functional implications of Neandertal introgression in modern humans. *Genome Biol.* **18**, 61 (2017).

35. C. Bornstein *et al.*, SPATA18, a spermatogenesis-associated gene, is a novel transcriptional target of p53 and p63. *Mol. Cell. Biol.* **31**, 1679–1689 (2011).

36. R. M. Gittelman *et al.*, Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. *Curr. Biol.* **26**, 3375–3382 (2016).

37. I. Mathieson *et al.*, The genomic history of southeastern Europe. *Nature*. **555**, 197–203 (2018).

38. C. Capelli *et al.*, Moors and Saracens in Europe: estimating the medieval North African male legacy

572 in southern Europe. *Eur. J. Hum. Genet.* **17**, 848–852 (2009).

573 39. M. Sazzini *et al.*, Complex interplay between neutral and adaptive evolution shaped differential
574 genomic background and disease susceptibility along the Italian peninsula. *Sci. Rep.* **6**, 32513 (2016).

575 40. S. Sarno *et al.*, Ancient and recent admixture layers in Sicily and Southern Italy trace multiple
576 migration routes along the Mediterranean. *Sci. Rep.* **7**, 1984 (2017).

577 41. K. Prüfer *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*.
578 **505**, 43–49 (2014).

579 42. A. B. Wolf, J. M. Akey, Outstanding questions in the study of archaic hominin admixture. *PLoS*
580 *Genet.* **14**, e1007349 (2018).

581 43. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage
582 analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

583 44. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets.
584 *Gigascience*. **4** (2015), doi:10.1186/s13742-015-0047-8.

585 45. O. Delaneau, J.-F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and
586 population genetic studies. *Nat. Methods*. **10**, 5–6 (2013).

587 46. Team, R Core, R: A language and environment for statistical computing. R Foundation for Statistical
588 Computing, Vienna, Austria. 2016 (2017).

589 47. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190
590 (2006).

591 48. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated
592 individuals. *Genome Res.* **19**, 1655–1664 (2009).

593 49. M. Jakobsson, N. A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing
594 with label switching and multimodality in analysis of population structure. *Bioinformatics*. **23**, 1801–
595 1806 (2007).

596 50. N. M. Kopelman, J. Mayzel, M. Jakobsson, N. A. Rosenberg, I. Mayrose, Clumpak: a program for
597 identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol.*
598 *Resour.* **15**, 1179–1191 (2015).

599 51. N. A. Rosenberg, distruct: a program for the graphical display of population structure. *Mol. Ecol.*
600 *Notes*. **4**, 137–138 (2004).

601 52. G. Hudjashov *et al.*, Complex Patterns of Admixture across the Indonesian Archipelago. *Mol. Biol.*
602 *Evol.* **34**, 2439–2452 (2017).

603 53. F. Montinaro *et al.*, Complex Ancient Genetic Structure and Cultural Transitions in Southern African
604 Populations. *Genetics*. **205**, 303–316 (2017).

605 54. G. B. Busby *et al.*, Admixture into and within sub-Saharan Africa. *Elife*. **5** (2016),
606 doi:10.7554/eLife.15266.

607 55. Z. Hofmanová *et al.*, Early farmers from across Europe directly descended from Neolithic Aegeans.
608 *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6886–6891 (2016).

609 56. F. Broushaki *et al.*, Early Neolithic genomes from the eastern Fertile Crescent. *Science*. **353**, 499–503
610 (2016).

57. I. Mathieson *et al.*, Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. **528**, 499–503 (2015).

58. N. Patterson *et al.*, Ancient admixture in human history. *Genetics*. **192**, 1065–1093 (2012).

59. N. Li, M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*. **165**, 2213–2233 (2003).

60. E. Y. Chen *et al.*, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. **14**, 128 (2013).

61. M. V. Kuleshov *et al.*, Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–7 (2016).

62. M. D. Mailman *et al.*, The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).

63. K. A. Tryka *et al.*, NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* **42**, D975–D979 (2013).

64. H. Mi *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 (2017).

65. S. Köhler *et al.*, The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45**, D865–D876 (2017).

66. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–62 (2016).

67. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2016).

68. M. Kanehisa, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

69. E. M. Ramos *et al.*, Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **22**, 144–147 (2013).

70. M. Kanai *et al.*, Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

71. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

72. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. **319**, 1100–1104 (2008).

73. S. Parolo *et al.*, Characterization of the biological processes shaping the genetic structure of the Italian population. *BMC Genet.* **16**, 132 (2015).

74. A. R. Martin *et al.*, Haplotype sharing provides insights into fine-scale population history and disease in Finland (2017), , doi:10.1101/200113.

75. P. Skoglund *et al.*, Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*. **336**, 466–469 (2012).

76. J. K. Pickrell *et al.*, Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad.*

Sci. U. S. A. **111**, 2632–2637 (2014).

77. T. Günther, M. Jakobsson, Genes mirror migrations and cultures in prehistoric Europe—a population genomic perspective. *Curr. Opin. Genet. Dev.* **41**, 115–123 (2016).

78. R. Nielsen *et al.*, Tracing the peopling of the world through genomics. *Nature*. **541**, 302–310 (2017).

79. G. M. Kılınç *et al.*, The Demographic Development of the First Farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016).

80. M. Lipson *et al.*, Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*. **551**, 368–372 (2017).

81. L. Abi-Rached *et al.*, The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*. **334**, 89–94 (2011).

82. S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012).

83. L. C. Jacobs *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 (2013).

84. S. Sankararaman *et al.*, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*. **507**, 354–357 (2014).

85. H. C. Stanescu *et al.*, Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N. Engl. J. Med.* **364**, 616–626 (2011).

86. P. Sekula *et al.*, Genetic risk variants for membranous nephropathy: extension of and association with other chronic kidney disease aetiologies. *Nephrol. Dial. Transplant.* **32**, 325–332 (2017).

87. C. A. McCarty *et al.*, The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics*. **4** (2011), doi:10.1186/1755-8794-4-13.

Acknowledgments

General: We would like to thank St Hugh's College and the Department of Zoology for facilitating the visits of A.R. and S.A. to the University of Oxford and the PhD programs of the University of Pavia and University of Turin for supporting these visits; the High Performance Computing Facility of the Oxford University and CINECA for the computational resources, the programming assistance and advices given during this project; the SU.VI.MAX for the access to the F_{ST} estimates of their unpublished work (C.D., J.G., P.G.); Tony Capra for sharing the list of Neanderthal introgressed regions in humans; Ryan Daniels and Miguel Gonzales Santos for the computing advices during the early stages of this project; Luca Alessandri for his comments on the archaeological context of the Bronze Age in Italy and surrounding regions; Simonetta Guarrera for

technical support (C.D.G., G.M.); the National Alpini Association (Associazione Nazionale Alpini) for their help in collecting Italian DNA samples at the 86th national assembly in Piacenza in 2013, in particular Bruno Plucani, Giangaspare Basile, Claudio Ferrari and the municipality of Piacenza (A.O., A.A.). Finally, the authors would like to acknowledge all the people that donated their DNA and made this work possible.

Funding: The Leverhulme Trust (F.M., C.C.); the Italian Ministry of Education, University and Research (MIUR): "Progetti Futuro in Ricerca 2012" (RBFR126B8I) (A.O, A.A); the "Dipartimenti di Eccellenza (2018–2022)" (Dept. of Medical Sciences of Turin; G.M.; Dept. of Biology and Biotechnology of Pavia; A.A., A.O., O.S., A.T.); the Italian Institute for genomic Medicine (IIGM) and Compagnia di San Paolo Torino, Italy (G.M.); the European Community, Sixth Framework Program (PROCARDIS: LSHM-CT-2007-037273) (S.B.); the Italian Ministry of Health (Besta CEDIR project: RC 2007/LR6, RC 2008/LR6; RC 2009/LR8; RC 2010/LR8; GR-2011-02347041) (G.B.B.); "Progetti di Ricerca finanziati dall'Università degli Studi di Torino (ex 60%) (2015)" (C.D.G, G.M); ANR-14-CE10-0001 and Région Pays de la Loire (J.G.).

Author Contributions: A.O., A.A., G.M., F.M., and C.C. conceived the idea for the study; A.R., S.A., F.M. and C.C. performed, devised or supervised the analyses; A.A., A.O., F.B. and V.L.P. provided reagents for the genotyping of novel samples; all the authors contributed to this study providing data, computational facilities, or other resources; A.R., S.A., F.M. and C.C. wrote the manuscript with inputs from co-authors.

Competing interests: The authors declare no competing interests.

Data and materials availability: Requests for accessing previously published data should be directed to the corresponding author of the publications where they were originally presented. Enquiries for unpublished data should be directed to Mait Metspalu (Genotype data provided by the Estonian Biocentre), Simon Myers (F_{ST} estimates among clusters in Spain), Christian Dina (F_{ST} estimates among clusters in France). Data genotyped as part of this project and presented here for

709 the first time (135 Italian samples and 6 samples from Albania, genotyped on the Infinium Omni2.5-

710 8 Illumina beadchip) can be downloaded at the following webpage: XXX

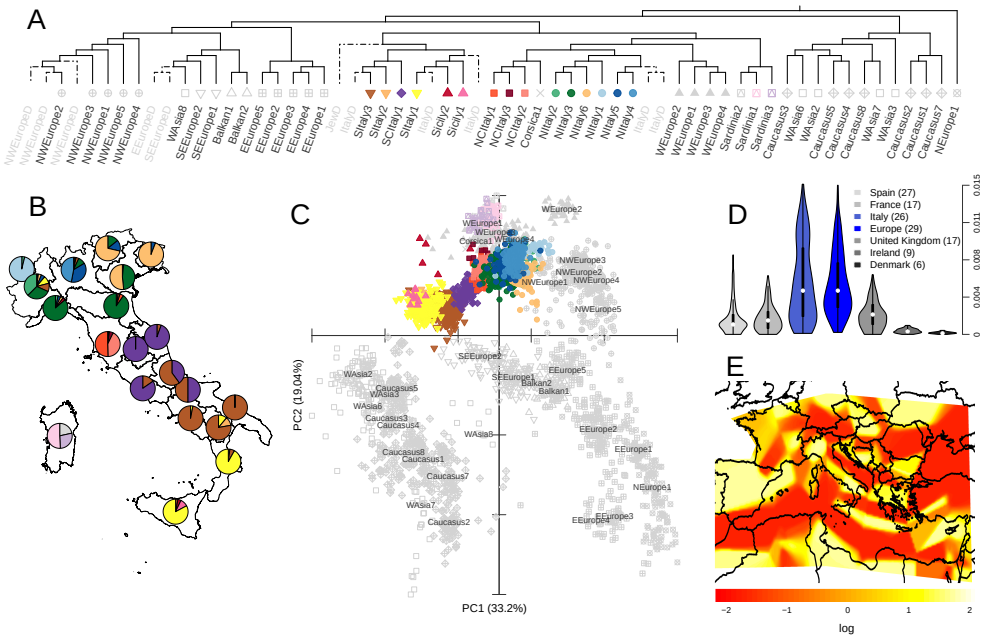


Fig. 1. Genetic structure of the Italian populations.
A) Simplified dendrogram of 3,057 Eurasian samples clustered by the fS algorithm using the CP output (complete dendrogram in fig. S2A); each leaf represents a cluster of individuals with similar copying vectors; clusters with more than five individuals are labelled in black; Italian clusters are colour coded; grey labels ending with the D letter refer to clusters containing less than five individuals or individuals of uncertain origin that have been removed in the following analyses. B) Pie charts summarizing the relative proportions of inferred fS genetic clusters for all the 20 Italian administrative regions (colours as in A). C) PCA based on CP chunkcount matrix (colours as in A); the centroid of the individuals belonging to non-Italian clusters is identified by the label for each cluster. D) Between-clusters F_{st} estimates within European groups; clusters were generated using only individuals belonging to the population analysed (Materials and Methods, Supplementary materials); the number of genetic clusters analysed for each population is reported within brackets; for the comparisons across Europe, the cluster NEurope1 containing almost exclusively Finnish individuals was excluded (F_{st} estimates for Italian and European clusters are in data file S3); F_{st} distributions statistically different from the Italian set are in grey. E) Estimated Effective Migration Surfaces (EEMS) analysis in Southern Europe; colours represent the log10 scale of the effective migration rate, from low (red) to high (yellow).

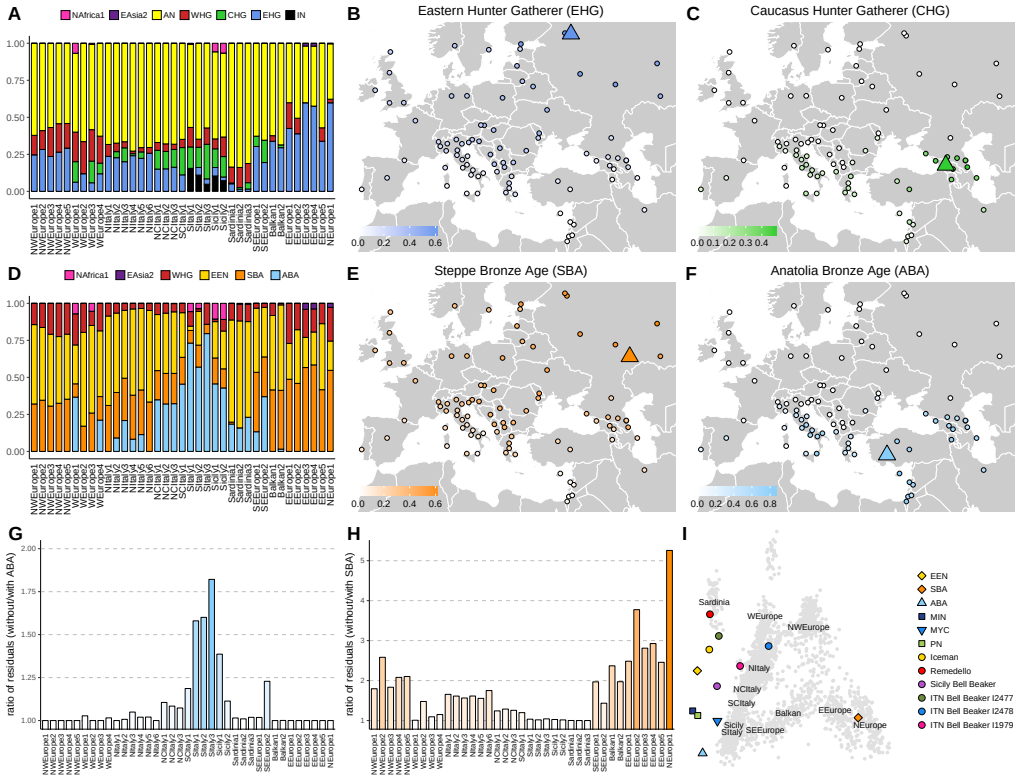


Fig. 2. Ancient ancestries in Western Eurasian modern-day clusters and Italian ancient samples. A, D) CP/NNLS analysis on all Italian and European clusters using as donors different sets of ancient samples and two modern clusters (NAfrica1: North Africa, EAsia2: East Asia) (full results in fig. S8). A) *Ultimate* sources: AN, Anatolian Neolithic (Bar8); WHG, Western Hunter Gatherer (Bichon); CHG, Caucasus Hunter Gatherer (KK1); EHG, Eastern Hunter Gatherer (I0061); IN, Iranian Neolithic (WC1). B) EHG and C) CHG ancestry contributions in Western Eurasia, as inferred in A and fig. S8A (Supplementary materials). D) Same as in A, using *Proximate* sources: WHG, Western Hunter Gatherer (Bichon); EEN, European Early Neolithic (Stuttgart); SBA, Bronze Age from Steppe (I0231); ABA, Bronze Age from Anatolia (I2683). E) SBA and F) ABA ancestry contributions, as inferred in D and fig. S8B. Triangles refer to the location of ancient samples used as sources (see data file S1). G): ratio of the residuals in the NNLS analysis (Materials and Methods, Supplementary materials) for all the Italian and European clusters when ABA was excluded and included in the set of *Proximate* sources; H) as in G), but excluding/including SBA instead of ABA; J) 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.

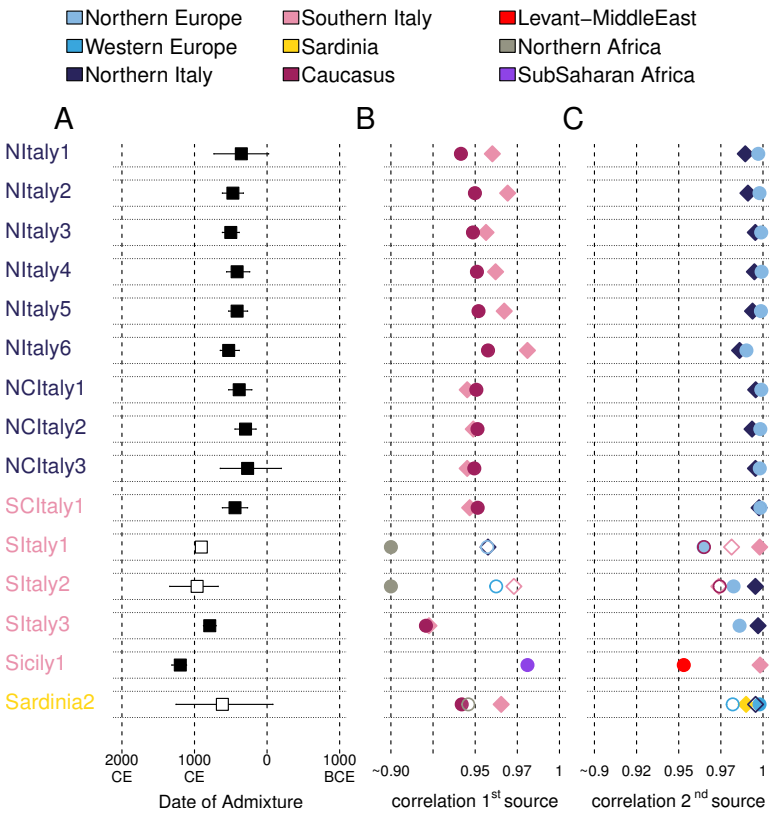


Fig. 3. Admixture events inferred by GLOBETROTTER (GT).
A) Dates of the events inferred in the GT "noItaly" analysis on all the Italian clusters (els as in Fig. 1A and data file S2; full results in fig. S16 and table S7; see Materials and Methods, Supplementary materials); lines encompassed the 95% CI. GT events were distinguished in "one date" (black squares; 1D in table S7) and "one date multiway" (white squares; 1MW). B) Correlation values between copying vectors of 1st source(s) identified by GT and the best proxy in the noItaly analysis (circles) or the best proxy among Italian clusters (diamonds). C) Same as in B, referring to 2nd source(s) copying vectors. Empty symbols refer to additional 1st (B) and 2nd (C) sources detected in multiway events. African best proxies in (B) for clusters SItaly1 and SItaly2 were plotted on the 0.90 boundary for visualisation only, the correlation values being 0.78 and 0.87 respectively. Colours of symbols refer to the ancestry to which proxies were assigned (see Materials and Methods, Supplementary materials).

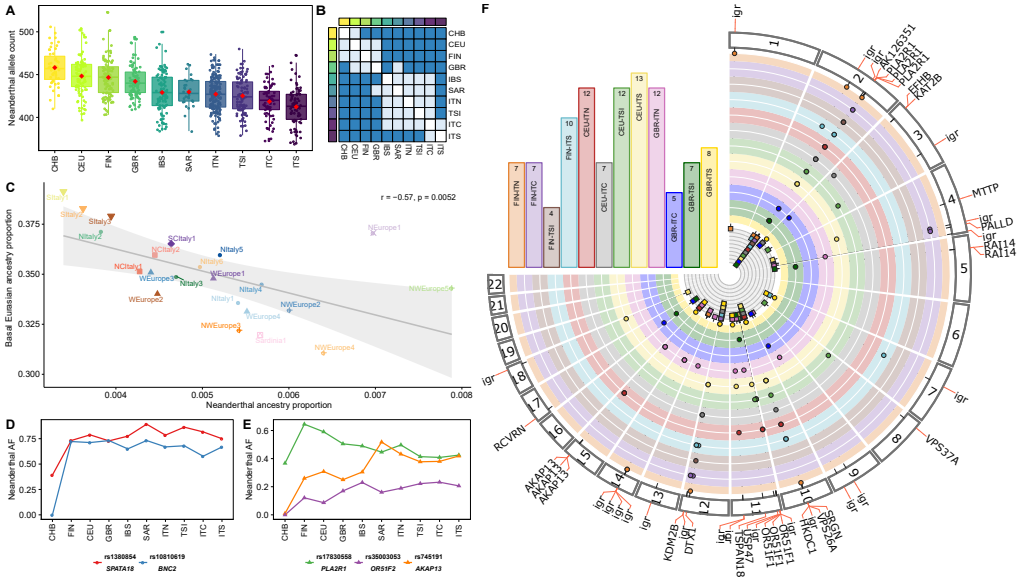


Fig. 4. Neanderthal ancestry distribution in Eurasian populations.
A) Neanderthal allele counts in individuals from Eurasian populations, sorted by median values on 3,969 LD-pruned Neanderthal tag-SNPs. CEU, Utah Residents with Northern and Western European ancestry; GBR, British in England and Scotland; 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 (Materials and Methods, Supplementary materials; 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 (Materials and Methods, Supplementary materials). D, E) Neanderthal allele frequency (AF) for selected SNPs within the indicated genes: D) high frequency alleles in Europe; E) North-South Europe divergent alleles. F) 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.

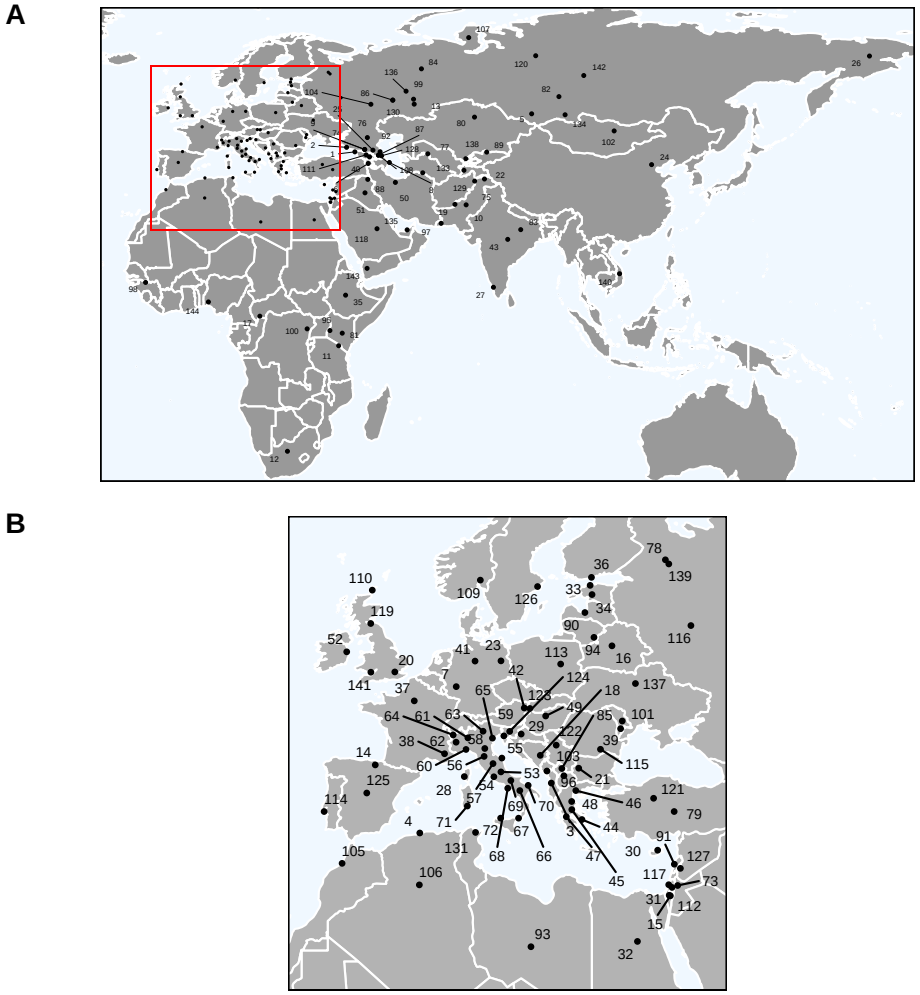


Fig. S1. Geographic location of populations included in FMD and HDD.
A) European, North African and Western Eurasia samples; B) World-wide samples. Numbers as in table S1.

Fig. S2. fineSTRUCTURE dendrogram of all the 4,852 (A, FMD) and 1,641 (B, HDD) samples. 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. At the end, within brackets, the name given to the cluster. Thick lines in black refer to the Italian clusters. The details of cluster assignation are reported in data file S2.

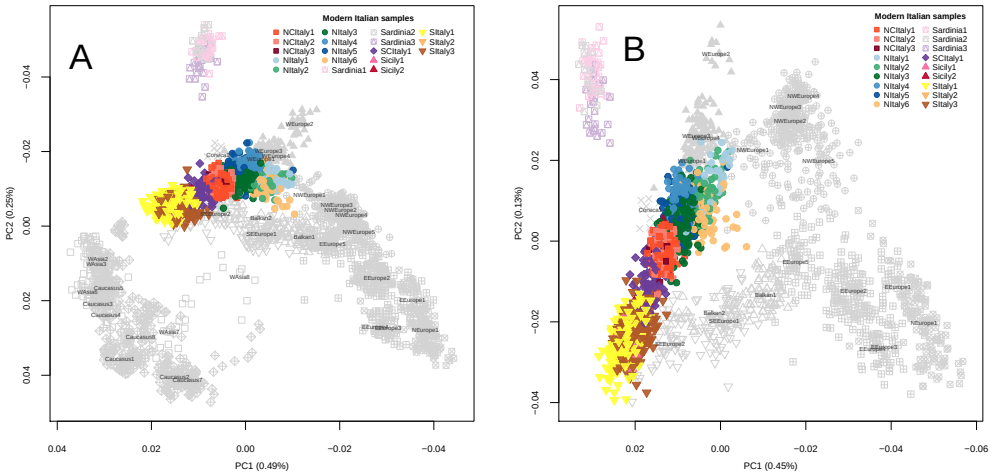


Fig. S3. Allele frequency Principal Components Analysis (PCA) of modern samples (genotype-based).

A) PCA of 3,057 modern samples included in Eurasian CP/fS inferred clusters; all the samples are labelled and coloured as in Fig. 1A. B) PCA of 2,469 modern European samples as displayed from the dendrogram resulting from CP/fS (Fig. 1A).

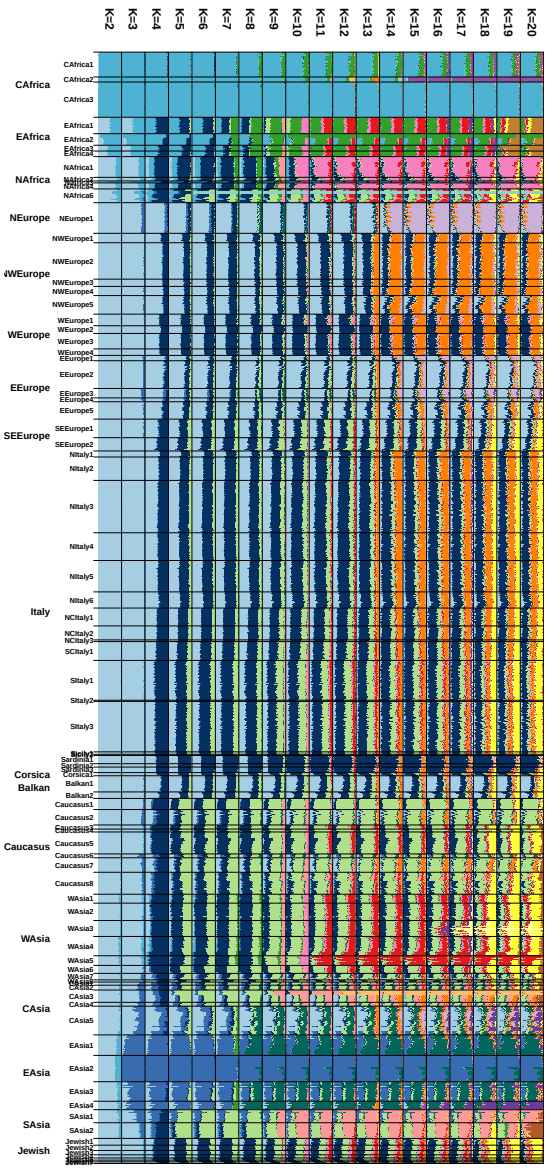


Fig. S4. Individual-level ADMIXTURE analysis of modern samples.
Samples are grouped according to the genetic clusters inferred by the CP/fS pipeline and named as in fig. S2.

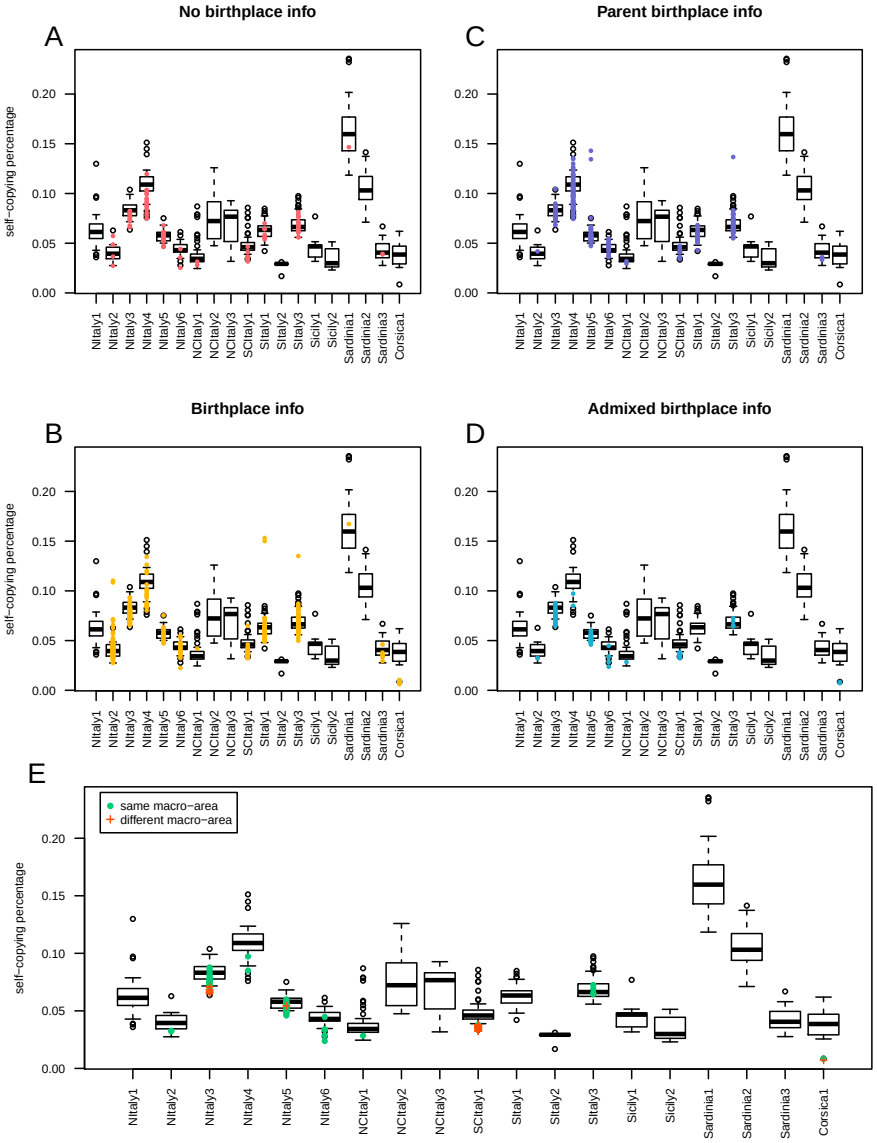


Fig. S5. "Cluster self-copy" analysis.
Box plots refer to the distributions of the self-copying vectors for each cluster for samples with same birthplace region for the four grandparents; coloured points refer to individual samples with other/no information; outliers are indicated as white circles. Coloured points refer to: A) subjects with no information available on their place of birth (red); B) subjects with only their own birthplace information (yellow); C) subjects with parents birthplace information (violet); D) subjects with "mixed" parental ancestry (parents from different regions) (blue); E) same as in D), red crosses identify individuals with parents born in different macro-areas (North and South Italy) indicated as suffix in each Italian population (table S1), while green dots refer to samples with parents born in the same macro-area.

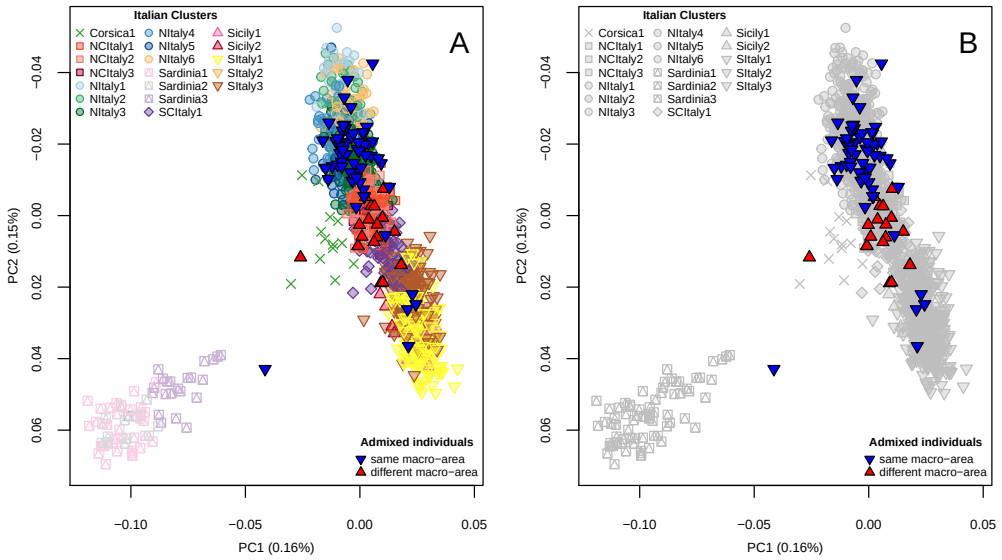


Fig. S6. PCA with Admixed Italian individuals.
Individuals with parents known to be born in two different macro-areas (see Materials and Methods, Supplementary materials - Cluster Self-Copy analysis) are plotted in red together with all the other Italian individuals, these coloured according either to the clusters they belong to (A) or in grey (B). Macro-areas are separated in Northern and Southern, where the central regions of Tuscany and Emilia are considered as part of the Northern macroarea and Latium, Abruzzo, Marche and Sardinia were considered as part of the Southern macro-area.

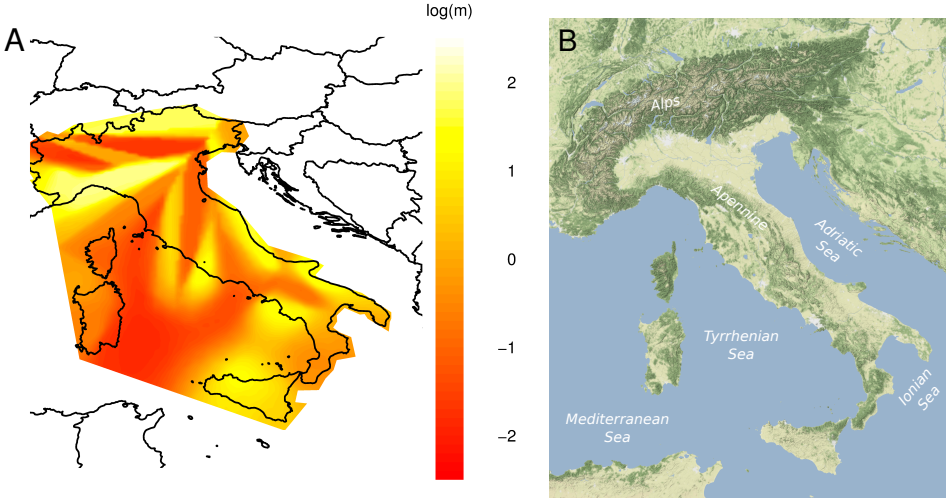


Fig. S7. Results of the EEMS analysis on Italy-only populations.
A) Colours represent the log10 scale of the effective migration rate from low (red) to high (yellow). Samples as reported in table S1. B) Physical map of Italy.

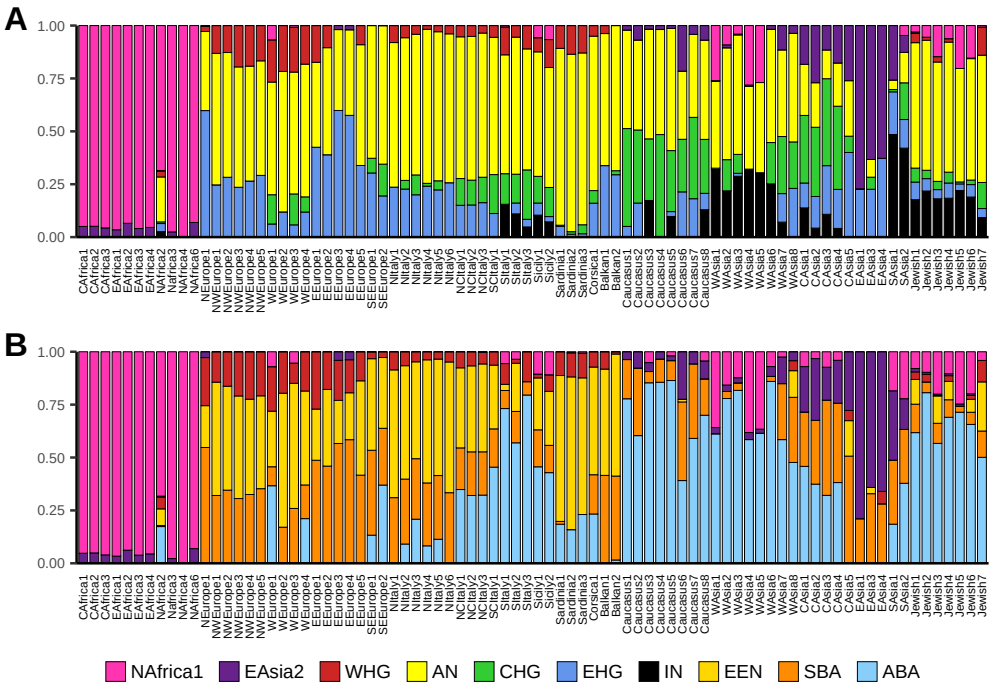


Fig. S8. CP/NNLS results for *Ultimate* and emphProximate sources for all modern clusters.
A) *Ultimate* (A) and *Proximate* (B) sources analysis reporting all modern Eurasian and African clusters and including WHG among the sources (main text; Supplementary Material).

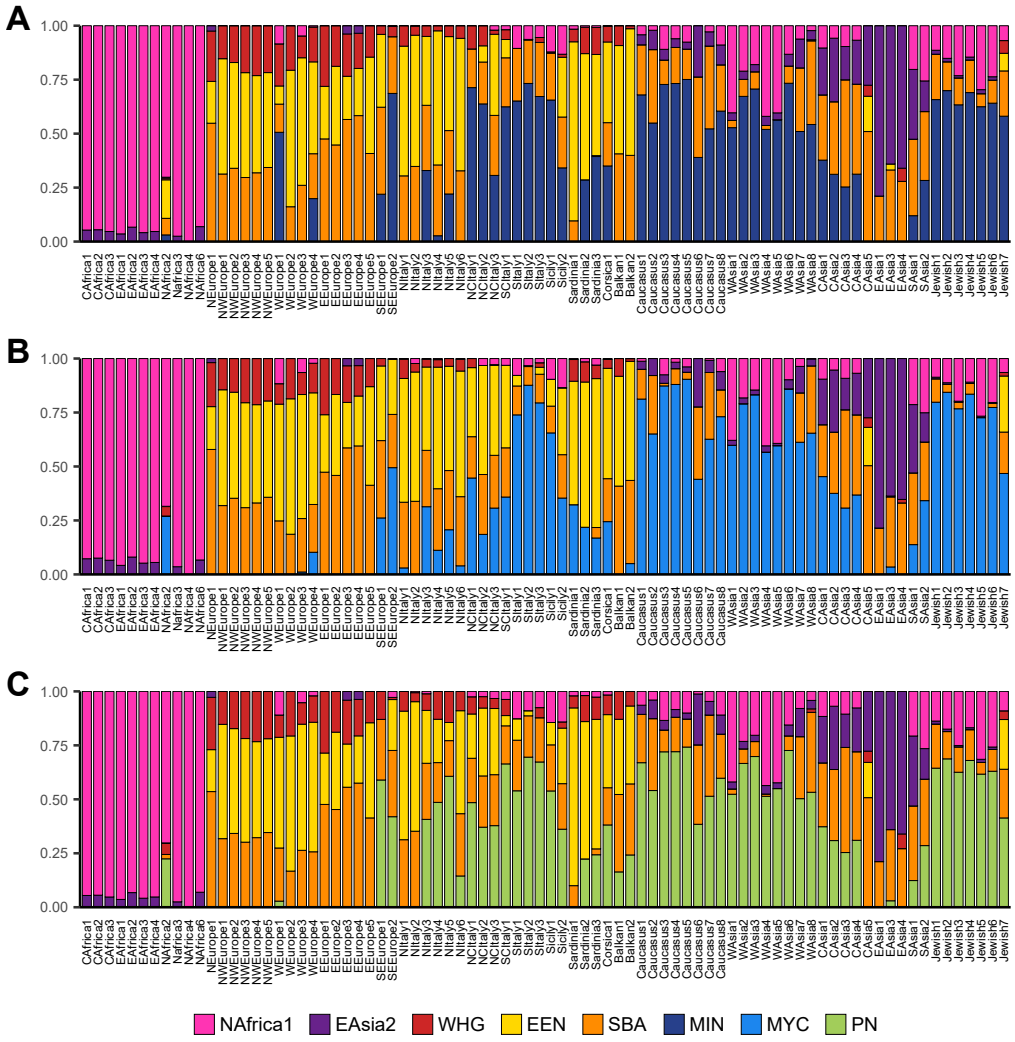


Fig. S9. CP/NNLS results for *emphProximate* sources for all modern clusters using alternative SEE sources.

Proximate sources analysis replacing ABA with alternative SEE sources: A) Minoan, MIN: B) Mycenaean , MYC: C) Peloponnese Neolithic, PN. In all the analyses, WHG was included among the possible sources (Supplementary Material).

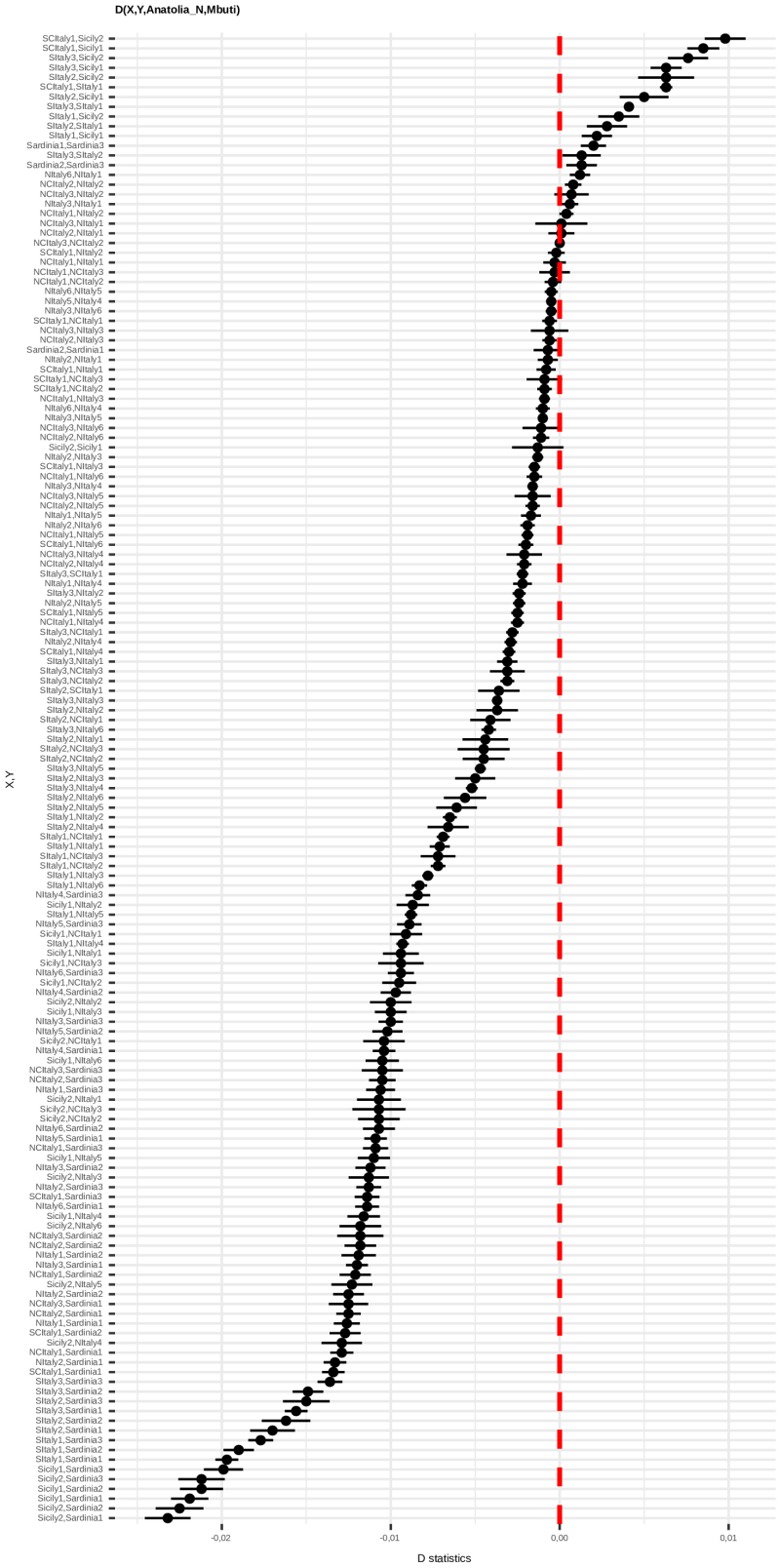


Fig. S10. D statistics in the form D(X,Y, AN,Mbuti) for all the possible pairs of Italian clusters.

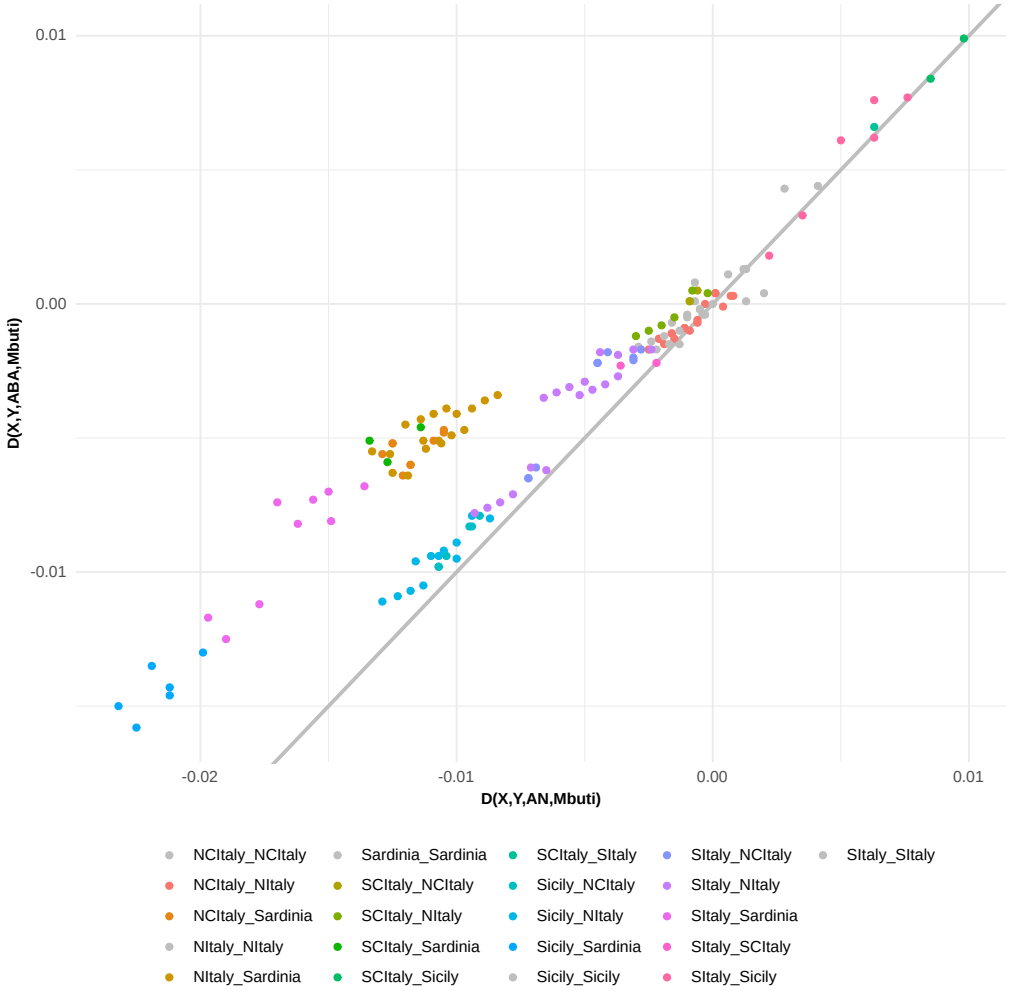


Fig. S11. Comparison of AN and ABA affinity to Italian clusters using D-statistics. Scatter plot of D(Ita1, Ita2, AN, Mbuti) and D(Ita1,Ita2,ABA,Mbuti) for all the Italian clusters. Points for pairs of clusters from the same (grey points) or closely related geographic location fall in proximity of the 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.

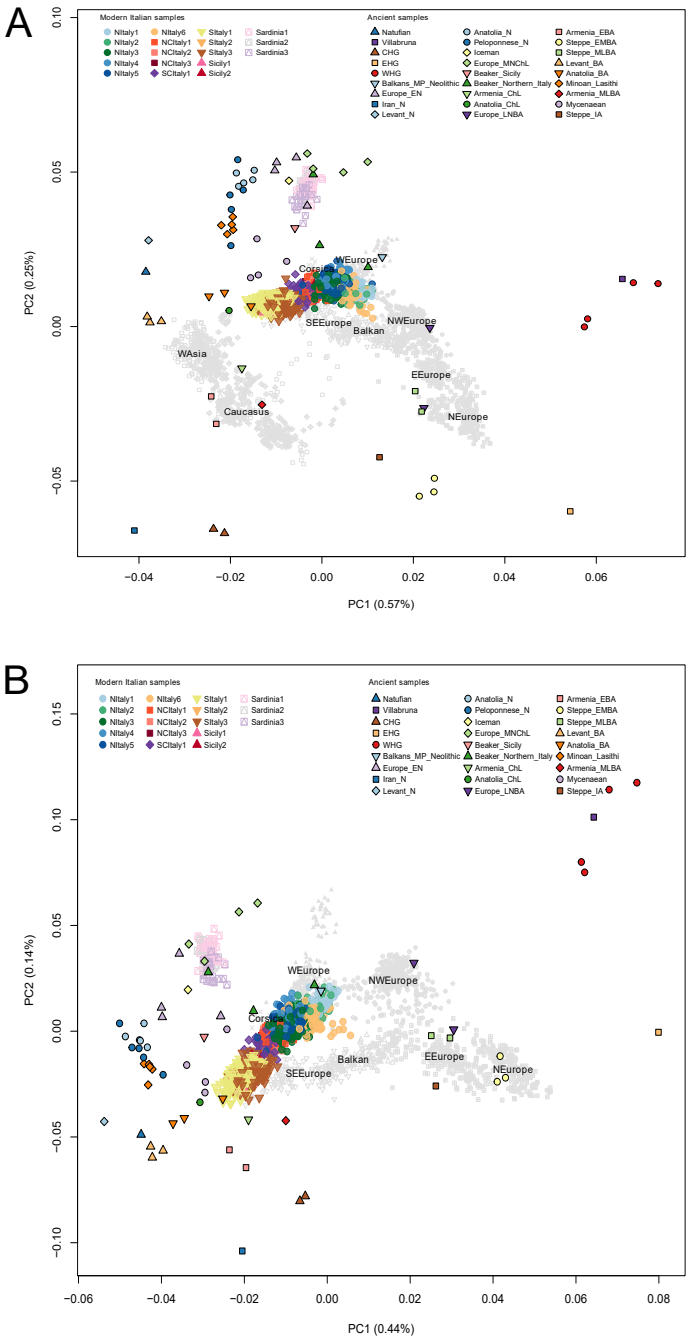


Fig. S12. Principal component analysis projecting 63 ancient individuals onto the components inferred from modern individuals. A) Principal component analysis projecting 63 ancient individuals onto the components inferred from 3,282 modern individuals assigned, through a CP/fS analysis, to European West Asian and Caucasian clusters (data file S2). B) 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 (data file S2). 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.

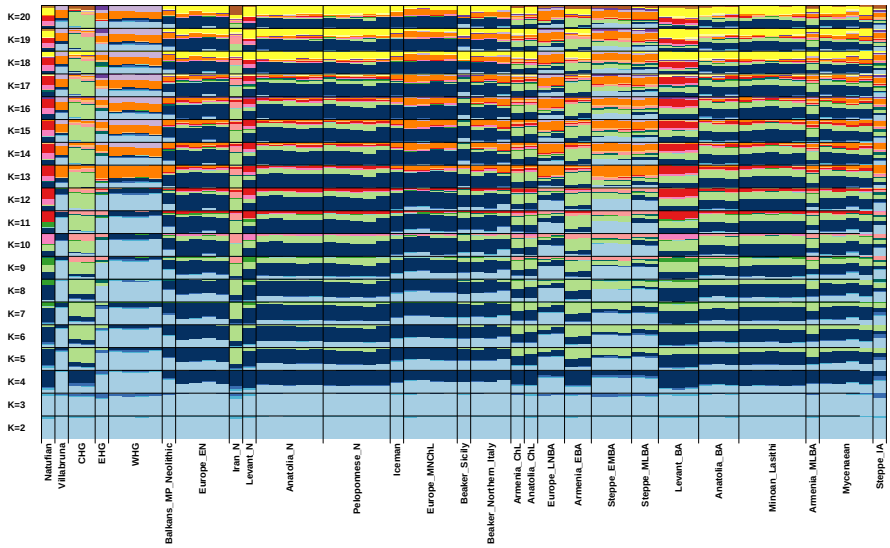


Fig. S13. ADMIXTURE analysis of 63 ancient samples.
Ancestral allele frequencies were inferred from ten different ADMIXTURE runs on 4,606 modern samples and projected onto the ancient samples. Each bar represents an individual grouped into ancient groups (data file S1).

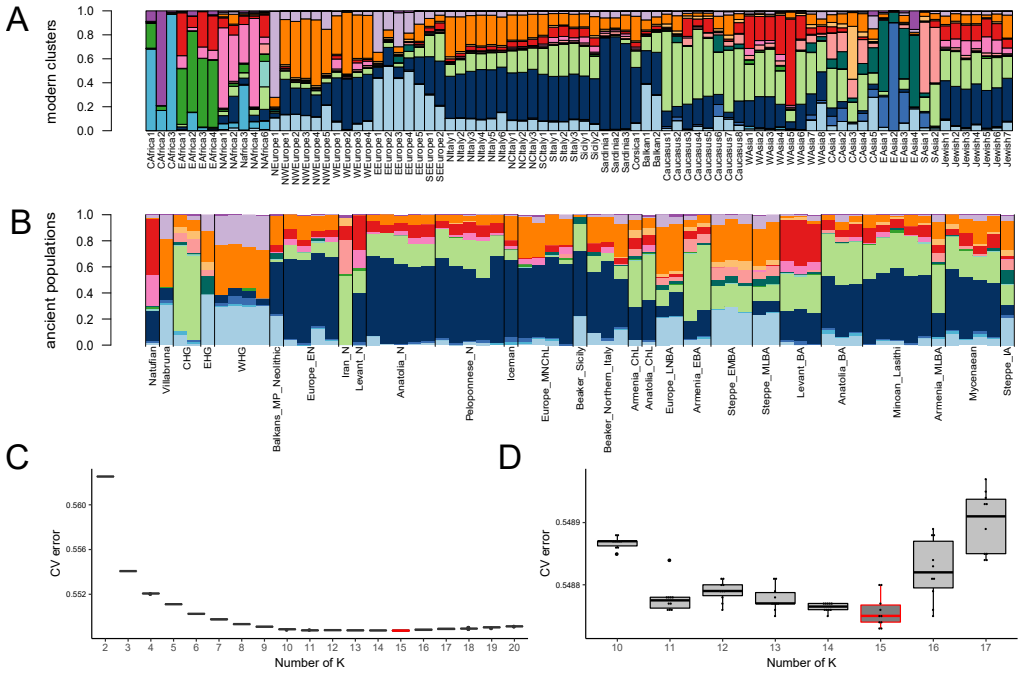


Fig. S14. ADMIXTURE analysis of 63 ancient samples and 4,606 modern samples for K=15. A-B) Results of the ADMIXTURE analysis as in fig. S4 and fig. S13 for K=15 including both modern (A) and ancient samples (B). C) Box plots of the ten CV-errors of each K from 2 to 20. D) Detailed box plots for the ten CV-errors for each K from 10 to 17.

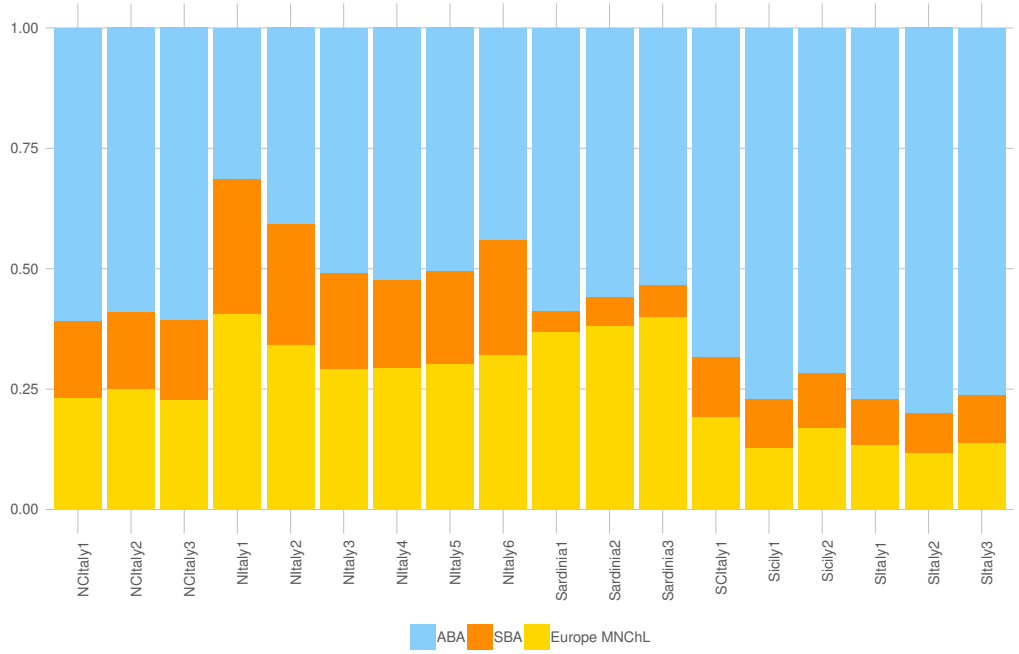


Fig. S15. Mixture proportions on modern Italian clusters inferred by qpAdm as a combination of ABA, SBA and European Middle-Neolithic/Chalcolithic.
For each tested cluster, we have evaluated all the possible combinations of N "left" sources with N={2..5}, and one set of right/left Outgroups (Supplementary materials).

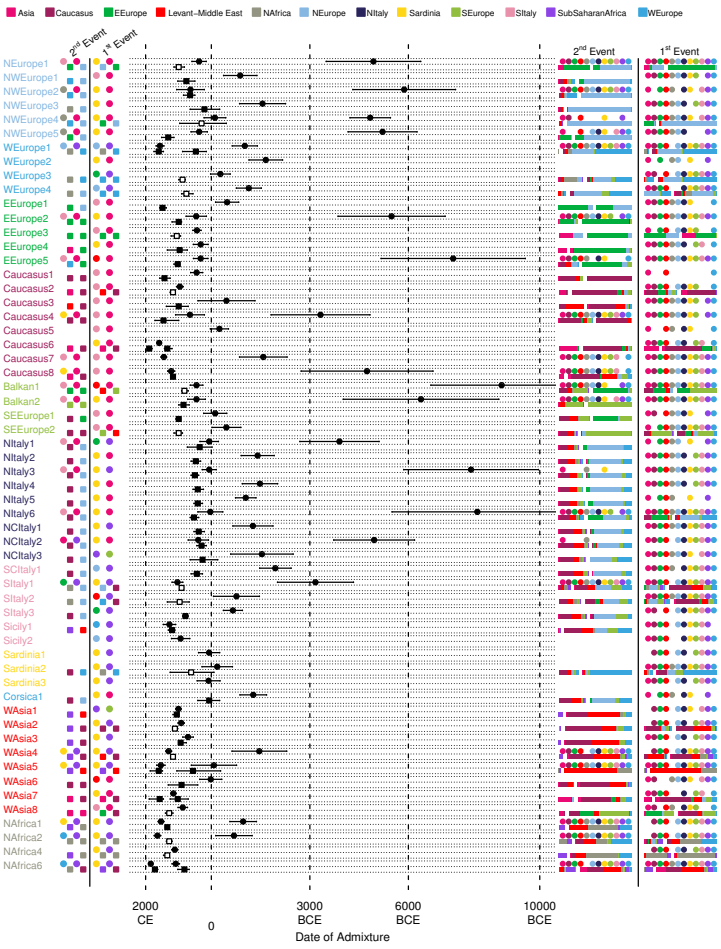


Fig. S16. GT and MALDER analyses for all the Eurasian and North African clusters.
Dates of the events inferred by "noItaly" GT (squares) and MALDER (circles) for clusters as in Fig. 1A and data file S2 are reported in the central part of the plot; lines encompassed the 95% CI for GT and ±1 Standard Error for MALDER. GT events were distinguished in "one date" (black squares; 1D in table S7), "one date multiway" (white squares; 1MW) or "two events" (two black squares; 2D). The best sources are indicated in a staggered way as circles and squares for MALDER and GT, respectively ("1st/2nd event" columns, on the left; four sources are highlighted for 1MW events). Colours refer to the ancestry to which the sources were assigned (see Materials and Methods; Supplementary materials). We additionally included a sub-Saharan African ancestry comprising CAfrica and EAfrica clusters (Fig. S2, data file S2). GT sources for single date events are plotted in the column "2nd event", as overlapping with second events detected by MALDER. The composition of the sources for GT and the geographical regions of the sources in MALDER, for which no significant differences in the amplitude of the fitted curve were found, are reported in the "1st/2nd event" columns on the right. GT sources are divided by a white space; the length of the bars indicates the contribution of each source; for 1MW events, two bar plots are indicated in the "1st/2nd event" columns on the right.

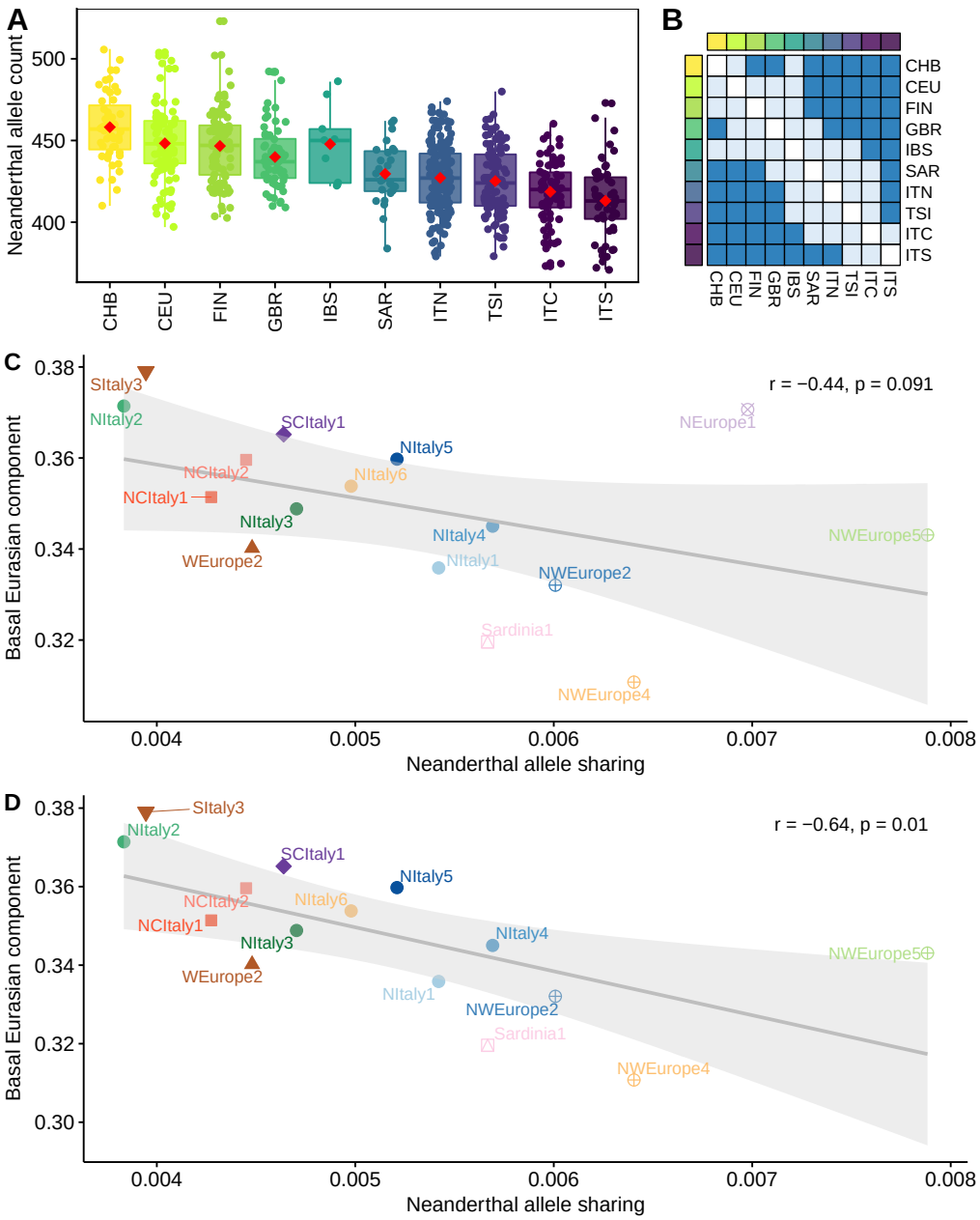


Fig. S17. Exploring the relationship between Neanderthal ancestry and admixture with African sources.

Same as in Fig. 4A, B, C 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, Figure 3 and fig. S16) or the whole set of the clusters listed above (see Supplementary materials). 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 (see Supplementary Materials). Clusters with less than 10 individuals were excluded in C and D.

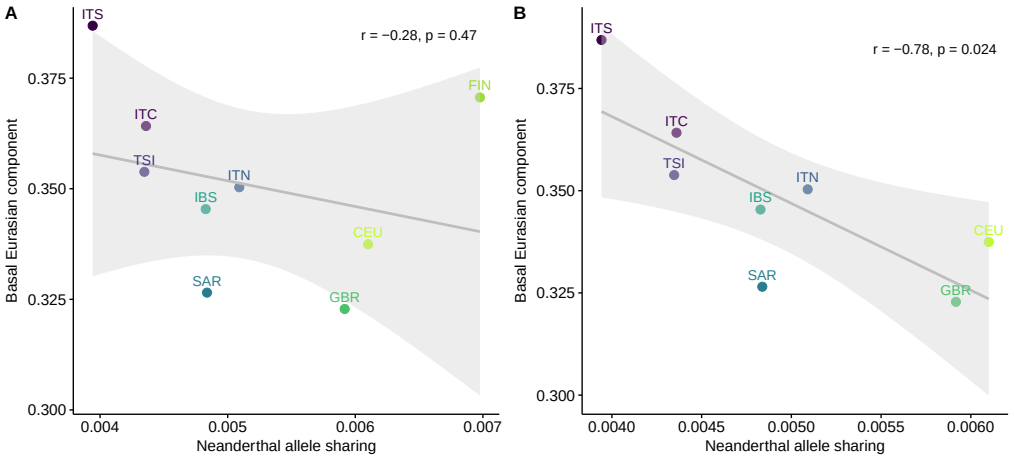


Fig. S18. Correlation between the proportion of Neanderthal allele sharing and the amount of ancestry derived from a Basal Eurasian population in European populations.
A) Correlation considering FIN (Finnish in Finland) population. B) Correlation excluding FIN (Finnish in Finland) population (see Materials and Methods, Supplementary materials).

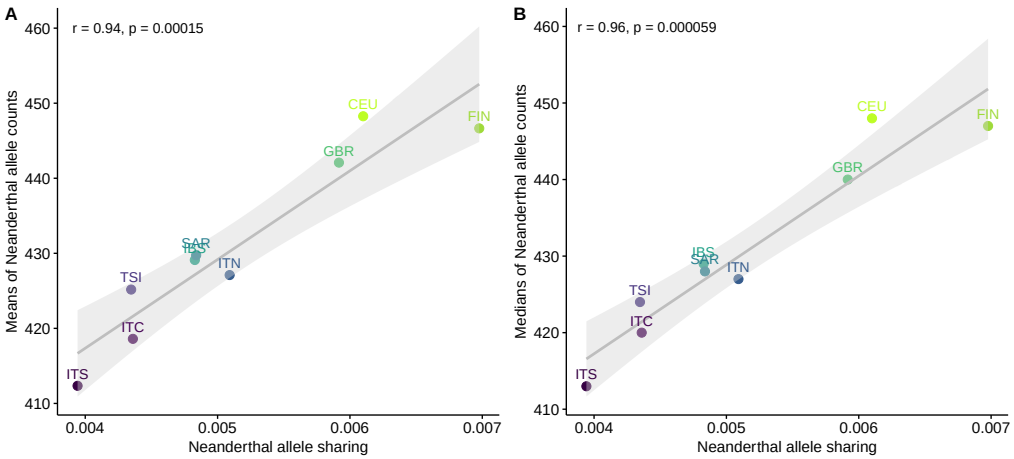


Fig. S19. Correlation between the proportions of Neanderthal allele sharing computed with F4-ratio and the counts per population of Neanderthal alleles in European populations.
A) Correlation between the proportions of Neanderthal allele sharing computed with F4-ratio and the means per population of Neanderthal allele counts. B) Correlation between the proportion of Neanderthal allele sharing computed with F4-ratio and the medians per population of Neanderthal allele counts.

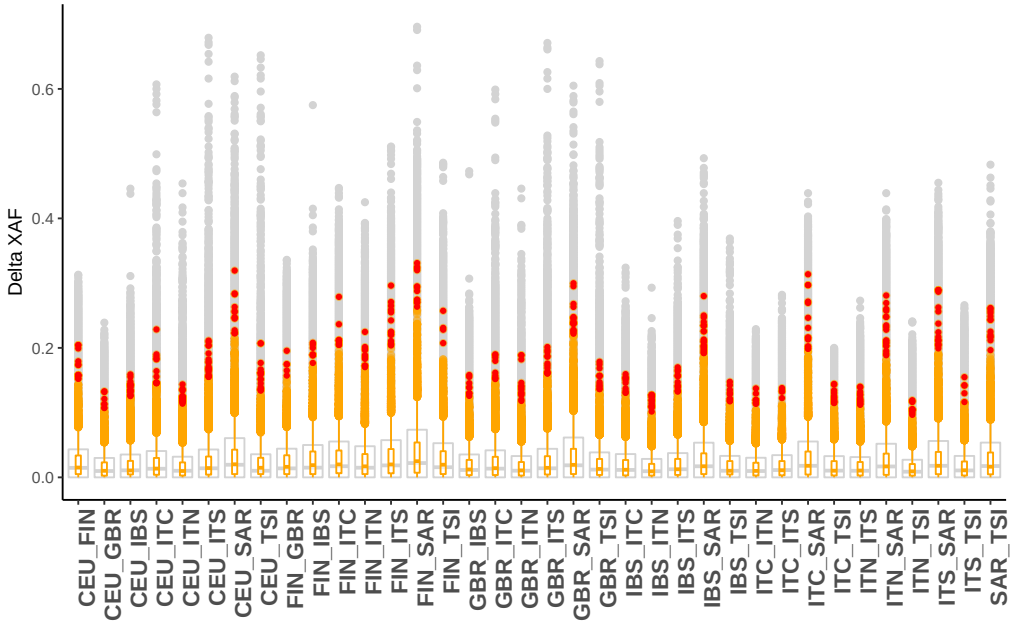


Fig. S20. Absolute allele frequency differences (ΔXAF, where X is the minor allele for each SNP or the Neanderthal allele when considering Neanderthal regions tag-SNPs) for each pair of European populations.

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, as also reported in data file S4.

Glossary

admixture the result of interbreeding between two or more previously isolated populations within a species. 15

allele a variant form of a given gene. 13

ascertainment bias systematic deviations from an expected theoretical result attributable to the sampling processes used to find (ascertain) SNPs and measure their population-specific allele frequencies. 13

effective population size the size that a theoretical population evolving under a Wright–Fisher model would need to be in order to match aspects of the observed genetic data. 19

gene flow the transfer of genetic variation from one population to another. 15

genetic drift random fluctuations in allele frequencies from one generation to the next due to the random sampling of gametes. 13

haplotype a group of genes within an organism that was inherited together from a single parent. 33

Hardy–Weinberg equilibrium a principle that states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. 36

homozygosity a genotype is homozygous if it has identical alleles. Contiguous lengths of homozygous genotypes are called runs of homozygosity. Homozygosity by descent means that identical alleles were descenced from a single source, as may occur in consanguineous mating. 33

linkage disequilibrium the nonindependence, at a population level, of the alleles carried at different positions in the genome. 36

locus a fixed position on a chromosome, like the position of a gene or a genetic marker. 15

single-nucleotide polymorphism a variation in a single nucleotide. For example, at a specific base position in the genome, the C nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by an A. 28

total variation distance a distance measure for probability distributions. Informally, this is the largest possible difference between the probabilities that the two probability distributions can assign to the same event. 39

Acronyms

1000G 1000 Genomes Project. 91

AMH anatomically modern humans. 16

cM centiMorgans. 87

FS fineSTRUCTURE. 36

GWASs genome-wide association studies. 28

HBD homozygosity by descent. 87
HGDP Human Genome Diversity Project. 15
HWE Hardy–Weinberg equilibrium. 91

IBD identity-by-descent. 40

kyr thousand years. 17

LD linkage disequilibrium. 31
LGM Last Glacial Maximum. 19

MAF minor allele frequency. 91
MCMC Markov chain Monte Carlo. 36
mtDNA mitochondrial DNA. 16

PCA principal component analysis. 31
POPRES Population Refence Study. 31

QC quality control. 89

ROH runs of homozygosity. 87, 92

SNP single nucleotide polymorphism. 31

TVD total variation distance. 39

WGS whole genome sequencing. 124
WHG Western European hunters-gatherers. 19

Bibliography

[1] A. F. Agrawal and M. C. Whitlock. Mutation Load: The Fitness of Individuals in Populations Where Deleterious Alleles Are Abundant. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):115–135, Dec. 2012. 126

[2] D. H. Alexander and K. Lange. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12(1):246, 2011. 38

[3] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, Sept. 2009. 37, 38, 93

[4] M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, A.-S. Malaspinas, A. Margaryan, T. Higham, D. Chivall, N. Lynnerup, L. Harvig, J. Baron, P. D. Casa, P. Dąbrowski, P. R. Duffy, A. V. Ebel, A. Epimakhov, K. Frei, M. Furmanek, T. Gralak, A. Gromov, S. Gronkiewicz, G. Grupe, T. Hajdu, R. Jarysz, V. Khartanovich, A. Khokhlov, V. Kiss, J. Kolář, A. Kriiska, I. Lasak, C. Longhi, G. McGlynn, A. Merkevicius, I. Merkyte, M. Metspalu, R. Mkrtchyan, V. Moiseyev, L. Paja, G. Pálfi, D. Pokutta, u. Pospieszny, T. D. Price, L. Saag, M. Sablin, N. Shishlina, V. Smrčka, V. I. Soenov, V. Szeverényi, G. Tóth, S. V. Trifanova, L. Varul, M. Vicze, L. Yepiskoposyan, V. Zhitenev, L. Orlando, T. Sicheritz-Pontén, S. Brunak, R. Nielsen, K. Kristiansen, and E. Willerslev. Population genomics of Bronze Age Eurasia. *Nature*, 522(7555):167–172, June 2015. 19, 20

[5] A. J. Ammerman, editor. *The widening harvest: the Neolithic transition in Europe: looking back, looking forward ; [proceedings of "The Neolithic Transition in Europe: Looking Back, Looking Forward", a conference held in Venice, Italy on 29-31 October 1998].* Number 6 in Colloquia and conference papers / Archaeological Institute of America. Archaeological Institute of America, Boston, Mass, 2003. 17

[6] E. Asouti and D. Q. Fuller. A Contextual Approach to the Emergence of Agriculture in Southwest Asia: Reconstructing Early Neolithic Plant-Food Production. *Current Anthropology*, 54(3):299–345, June 2013. 19

[7] G. Athanasiadis, J. Y. Cheng, B. J. Vilhjalmsson, F. G. Jorgensen, T. D. Als, S. Le Hellard, T. Espeseth, P. F. Sullivan, C. M. Hultman, P. C. Kjaergaard, M. H. Schierup, and T. Mailund. Nationwide Genomic Study in Denmark Reveals Remarkable Population Homogeneity. *Genetics*, 204(2):711–722, Oct. 2016. 27, 32, 123

[8] P. L. Auer and G. Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1):16, 2015. 130, 131

[9] A. Auton, G. Abecasis, D. Altshuler, R. Durbin, G. Abecasis, D. Bentley, A. Chakravarti, A. Clark, P. Donnelly, E. Eichler, P. Flicek, S. Gabriel, R. Gibbs, E. Green, M. Hurles, B. Knoppers, J. Korbel, E. Lander, C. Lee, H. Lehrach, E. Mardis, G. Marth, G. McVean, D. Nickerson, J. Schmidt, S. Sherry, J. Wang, R. Wilson, R. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. Reid, Y. Zhu, J. Wang, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. Lander, D. Altshuler, S. Gabriel, N. Gupta, N. Gharani, L. Toji, N. Gerry, A. Resch, P. Flicek, J. Barker,

L. Clarke, L. Gil, S. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. Albrecht, V. Amstislavskiy, T. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, E. Mardis, R. Wilson, L. Fulton, R. Fulton, S. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. McVean, R. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. Schmidt, C. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C. Campbell, Y. Kong, A. Marcketta, R. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang, L. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. Marth, E. Garrison, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A. Ward, J. Wu, M. Zhang, M. Daly, M. DePristo, R. Handsaker, D. Altshuler, E. Banks, G. Bhatia, G. del Angel, S. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. Lander, S. McCarroll, J. Nemesh, R. Poplin, S. Yoon, J. Lihm, V. Makarov, A. Clark, S. Gottipati, A. Keinan, J. Rodriguez-Flores, J. Korbel, T. Rausch, M. Fritz, A. Stütz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W. McLaren, G. Ritchie, R. Smith, D. Zerbino, X. Zheng-Bradley, P. Sabeti, I. Shlyakhter, S. Schaffner, J. Vitti, D. Cooper, E. Ball, P. Stenson, D. Bentley, B. Barnes, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. Kenny, M. Batzer, M. Konkel, J. Walker, D. MacArthur, M. Lek, R. Sudbrak, V. Amstislavskiy, R. Herwig, E. Mardis, L. Ding, D. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. Frederick Willems, J. Simpson, M. Shriver, J. Rosenfeld, C. Bustamante, S. Montgomery, F. De La Vega, J. Byrnes, A. Carroll, M. DeGorter, P. Lacroute, B. Maples, A. Martin, A. Moreno-Estrada, S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, C. Lee, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F. Hyland, D. Craig, A. Christoforides, N. Homer, T. Izatt, A. Kurdoglu, S. Sinari, K. Squire, S. Sherry, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E. Burchard, R. Hernandez, C. Gignoux, D. Haussler, S. Katzman, W. James Kent, B. Howie, A. Ruiz-Linares, E. Dermitzakis, S. Devine, G. Abecasis, H. Min Kang, J. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. Kate Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, G. McVean, J. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. Xifara, T. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, E. Eichler, B. Browning, S. Browning, F. Hormozdiari, P. Sudmant, E. Khurana, R. Durbin, M. Hurles, C. Tyler-Smith, C. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T. Keane, S. McCarthy, K. Walter, Y. Xue, M. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. Harmanci, M. Jin, D. Lee, J. Liu, X. Jasmine Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. Marth, E. Garrison, D. Kural, W.-P. Lee, A. Ward, J. Wu, M. Zhang, S. McCarroll, R. Handsaker, D. Altshuler, E. Banks, G. del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J. Nemesh, K. Shakir, S. Yoon, J. Lihm, V. Makarov, J. Degenhardt, J. Korbel, M. Fritz, S. Meiers, B. Raeder, T. Rausch, A. Stütz, P. Flicek, F. Paolo Casale, L. Clarke, R. Smith, O. Stegle, X. Zheng-Bradley, D. Bentley, B. Barnes, R. Keira Cheetham, M. Eberle, S. Humphray, S. . A global reference for human genetic variation. *Nature*, 526(7571):68–74, Sept. 2015. 17, 20, 129

[10] M.-C. Babron, M. de Tayrac, D. N. Rutledge, E. Zeggini, and E. Génin. Rare and Low Frequency Variant Stratification in the UK Population: Description and Impact on Association Tests. *PLoS ONE*, 7(10):e46519, Oct. 2012. 131

[11] B. Bachran. The Alans in Gaul. *Traditio*, (23):476–489, 1967. 22

[12] A. A. Behr, K. Z. Liu, G. Liu-Fang, P. Nakka, and S. Ramachandran. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, 32(18):2817–2823, Sept. 2016. 93

[13] S. Benazzi, K. Douka, C. Fornai, C. C. Bauer, O. Kullmer, J. Svoboda, I. Pap, F. Mallegni, P. Bayle, M. Coquerelle, S. Condemi, A. Ronchitelli, K. Harvati, and G. W. Weber. Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature*, 479(7374):525–528, Nov. 2011. 19

[14] A. Bhaskar, Y. R. Wang, and Y. S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, 25(2):268–279, Feb. 2015. 37

[15] G. Bhatia, N. Patterson, B. Pasaniuc, N. Zaitlen, G. Genovese, S. Pollack, S. Mallick, S. Myers, A. Tandon, C. Spencer, C. D. Palmer, A. A. Adeyemo, E. L. Akylbekova, L. A. Cupples, J. Divers, M. Fornage, W. L. Kao, L. Lange, M. Li, S. Musani, J. C. Mychaleckyj, A. Ogunniyi, G. Papanicolaou, C. N. Rotimi, J. I. Rotter, I. Ruczinski, B. Salako, D. S. Siscovick, B. O. Tayo, Q. Yang, S. McCarroll, P. Sabeti, G. Lettre, P. De Jager, J. Hirschhorn, X. Zhu, R. Cooper, D. Reich, J. G. Wilson, and A. L. Price. Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. *The American Journal of Human Genetics*, 89(3):368–381, Sept. 2011. 29

[16] S. Biswas, L. B. Scheinfeldt, and J. M. Akey. Genome-wide Insights into the Patterns and Determinants of Fine-Scale Population Structure in Humans. *The American Journal of Human Genetics*, 84(5):641–650, May 2009. 36

[17] H. Blayo and L. Henry. Données démographiques sur la Bretagne et l'Anjou de 1740 à 1829. *Ann Démographie Hist 1967*, pages 91–171, 1967. 129

[18] R. Bollongino, O. Nehlich, M. P. Richards, J. Orschiedt, M. G. Thomas, C. Sell, Z. Fajkosova, A. Powell, and J. Burger. 2000 Years of Parallel Societies in Stone Age Central Europe. *Science*, 342(6157):479–481, Oct. 2013. 20

[19] R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson. Mapping the Origins and Expansion of the Indo-European Language Family. *Science*, 337(6097):957–960, Aug. 2012. 20

[20] R. Bowden, T. S. MacFie, S. Myers, G. Hellenthal, E. Nerrienet, R. E. Bontrop, C. Freeman, P. Donnelly, and N. I. Mundy. Genomic Tools for Evolution and Conservation in the Chimpanzee: Pan troglodytes ellioti Is a Genetically Distinct Population. *PLoS Genetics*, 8(3):e1002504, Mar. 2012. 27

[21] A. Brisbin, K. Bryc, J. Byrnes, F. Zakharia, L. Omberg, J. Degenhardt, A. Reynolds, H. Ostrer, J. G. Mezey, and C. D. Bustamante. PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human Biology*, 84(4):343–364, Aug. 2012. 37

[22] C. Bromberger and A. Morel, editors. *Limites floues, frontières vives: Des variations culturelles en France et en Europe*. Éditions de la Maison des sciences de l'homme, 2001. 87, 128

[23] S. Browning and B. Browning. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *The American Journal of Human Genetics*, 97(3):404–418, Sept. 2015. 39, 40, 84, 92, 119, 121, 125

[24] S. R. Browning, B. L. Browning, M. L. Daviglus, R. A. Durazo-Arvizu, N. Schneiderman, R. C. Kaplan, and C. C. Laurie. Ancestry-specific recent effective population size in the Americas. *PLOS Genetics*, 14(5):e1007385, May 2018. 121

[25] C. D. Bustamante, F. M. De La Vega, and E. G. Burchard. Genomics for the world. *Nature*, 475(7355):163–165, July 2011. 129

[26] C. Bycroft, C. Fernandez-Rozadilla, C. Ruiz-Ponte, I. Quintela, n. Carracedo, P. Donnelly, and S. Myers. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nature Communications*, 10(1), Dec. 2019. 27, 31, 32, 33, 124

[27] J. Caesar. COMMENTARIUS PRIMUS, Chapter 1, Section 1. In *De bello Gallico*. 21

[28] C. S. Carlson, T. C. Matise, K. E. North, C. A. Haiman, M. D. Fesinmeyer, S. Buyske, F. R. Schumacher, U. Peters, N. Franceschini, M. D. Ritchie, D. J. Duggan, K. L. Spencer, L. Dumitrescu, C. B. Eaton, F. Thomas, A. Young, C. Carty, G. Heiss, L. Le Marchand, D. C. Crawford, L. A. Hindorff, C. L. Kooperberg, and for the PAGE Consortium. Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS Biology*, 11(9):e1001661, Sept. 2013. 29, 31

[29] L. M. Cassidy, R. Martiniano, E. M. Murphy, M. D. Teasdale, J. Mallory, B. Hartwell, and D. G. Bradley. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proceedings of the National Academy of Sciences*, 113(2):368–373, Jan. 2016. 20

[30] L. L. Cavalli-Sforza. Population Structure and Human Evolution. *Proceedings of the Royal Society B: Biological Sciences*, 164(995):362–379, Mar. 1966. 15

[31] L. L. Cavalli-Sforza. Genes, peoples, and languages. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15):7719–7724, July 1997. 16

[32] L. L. Cavalli-Sforza. The Human Genome Diversity Project: past, present and future. *Nature Reviews Genetics*, 6(4):333–340, Apr. 2005. 15

[33] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The history and geography of human genes*. Princeton Univ. Press, Princeton, NJ, abridged paperback ed edition, 1996. 13

[34] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), Dec. 2015. 82, 89

[35] J. Chaurand, editor. *Nouvelle histoire de la langue française*. Editions du Seuil, Paris, 1999. 26

[36] C. W. K. Chiang, J. H. Marcus, C. Sidore, A. Biddanda, H. Al-Asadi, M. Zoledziewska, M. Pitzalis, F. Busonero, A. Maschio, G. Pistis, M. Steri, A. Angius, K. E. Lohmueller, G. R. Abecasis, D. Schlessinger, F. Cucca, and J. Novembre. Genomic history of the Sardinian population. *Nature Genetics*, 50(10):1426–1434, Oct. 2018. 27, 31, 32, 33

[37] COLUMBUS Consortium, L. Fejerman, N. Ahmadiyeh, D. Hu, S. Huntsman, K. B. Beckman, J. L. Caswell, K. Tsung, E. M. John, G. Torres-Mejia, L. Carvajal-Carmona, M. M. Echeverry, A. M. D. Tuazon, C. Ramirez, C. R. Gignoux, C. Eng, E. Gonzalez-Burchard, B. Henderson, L. L. Marchand, C. Kooperberg, L. Hou, I. Agalliu, P. Kraft, S. Lindström, E. J. Perez-Stable, C. A. Haiman, and E. Ziv. Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. *Nature Communications*, 5(1), Dec. 2014. 29

[38] G. Coop, J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R. M. Myers, L. L. Cavalli-Sforza, M. W. Feldman, and J. K. Pritchard. The Role of Geography in Human Adaptation. *PLoS Genetics*, 5(6):e1000500, June 2009. 131

[39] O. Delaneau, J.-F. Zagury, and J. Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1):5–6, Dec. 2012. 82, 91

[40] DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, A. Mahajan, M. J. Go, W. Zhang, J. E. Below, K. J. Gaulton, T. Ferreira, M. Horikoshi, A. D. Johnson, M. C. Y. Ng, I. Prokopenko, D. Saleheen, X. Wang, E. Zeggini, G. R. Abecasis, L. S. Adair, P. Almgren, M. Atalay, T. Aung, D. Baldassarre, B. Balkau, Y. Bao, A. H. Barnett, I. Barroso, A. Basit, L. F. Been, J. Beilby, G. I. Bell, R. Benediktsson, R. N. Bergman, B. O. Boehm, E. Boerwinkle,

L. L. Bonnycastle, N. Burtt, Q. Cai, H. Campbell, J. Carey, S. Cauchi, M. Caulfield, J. C. N. Chan, L.-C. Chang, T.-J. Chang, Y.-C. Chang, G. Charpentier, C.-H. Chen, H. Chen, Y.-T. Chen, K.-S. Chia, M. Chidambaram, P. S. Chines, N. H. Cho, Y. M. Cho, L.-M. Chuang, F. S. Collins, M. C. Cornelis, D. J. Couper, A. T. Crenshaw, R. M. van Dam, J. Danesh, D. Das, U. de Faire, G. Dedoussis, P. Deloukas, A. S. Dimas, C. Dina, A. S. F. Doney, P. J. Donnelly, M. Dorkhan, C. van Duijn, J. Dupuis, S. Edkins, P. Elliott, V. Emilsson, R. Erbel, J. G. Eriksson, J. Escobedo, T. Esko, E. Eury, J. C. Florez, P. Fontanillas, N. G. Forouhi, T. Forsen, C. Fox, R. M. Fraser, T. M. Frayling, P. Froguel, P. Frossard, Y. Gao, K. Gertow, C. Gieger, B. Gigante, H. Grallert, G. B. Grant, L. C. Groop, C. J. Groves, E. Grundberg, C. Guiducci, A. Hamsten, B.-G. Han, K. Hara, N. Hassanali, A. T. Hattersley, C. Hayward, A. K. Hedman, C. Herder, A. Hofman, O. L. Holmen, K. Hovingh, A. B. Hreidarsson, C. Hu, F. B. Hu, J. Hui, S. E. Humphries, S. E. Hunt, D. J. Hunter, K. Hveem, Z. I. Hydrie, H. Ikegami, T. Illig, E. Ingelsson, M. Islam, B. Isomaa, A. U. Jackson, T. Jafar, A. James, W. Jia, K.-H. Jöckel, A. Jonsson, J. B. M. Jowett, T. Kadowaki, H. M. Kang, S. Kanoni, W. H. L. Kao, S. Kathiresan, N. Kato, P. Katulanda, S. M. Keinanen-Kiukaanniemi, A. M. Kelly, H. Khan, K.-T. Khaw, C.-C. Khor, H.-L. Kim, S. Kim, Y. J. Kim, L. Kinnunen, N. Klopp, A. Kong, E. Korpi-Hyövälti, S. Kowlessur, P. Kraft, J. Kravic, M. M. Kristensen, S. Krithika, A. Kumar, J. Kumate, J. Kuusisto, S. H. Kwak, M. Laakso, V. Lagou, T. A. Lakka, C. Langenberg, C. Langford, R. Lawrence, K. Leander, J.-M. Lee, N. R. Lee, M. Li, X. Li, Y. Li, J. Liang, S. Liju, W.-Y. Lim, L. Lind, C. M. Lindgren, E. Lindholm, C.-T. Liu, J. J. Liu, S. Lobbens, J. Long, R. J. F. Loos, W. Lu, J. Luan, V. Lyssenko, R. C. W. Ma, S. Maeda, R. Mägi, S. Männistö, D. R. Matthews, J. B. Meigs, O. Melander, A. Metspalu, J. Meyer, G. Mirza, E. Mihailov, S. Moebus, V. Mohan, K. L. Mohlke, A. D. Morris, T. W. Mühleisen, M. Müller-Nurasyid, B. Musk, J. Nakamura, E. Nakashima, P. Navarro, P.-K. Ng, A. C. Nica, P. M. Nilsson, I. Njølstad, M. M. Nöthen, K. Ohnaka, T. H. Ong, K. R. Owen, C. N. A. Palmer, J. S. Pankow, K. S. Park, M. Parkin, S. Pechlivanis, N. L. Pedersen, L. Peltonen, J. R. B. Perry, A. Peters, J. M. Pinidiyapathirage, C. G. P. Platou, S. Potter, J. F. Price, L. Qi, V. Radha, L. Rallidis, A. Rasheed, W. Rathmann, R. Rauramaa, S. Raychaudhuri, N. W. Rayner, S. D. Rees, E. Rehnberg, S. Ripatti, N. Robertson, M. Roden, E. J. Rossin, I. Rudan, D. Rybin, T. E. Saaristo, V. Salomaa, J. Saltevo, M. Samuel, D. K. Sanghera, J. Saramies, J. Scott, L. J. Scott, R. A. Scott, A. V. Segrè, J. Sehmi, B. Sennblad, N. Shah, S. Shah, A. S. Shera, X. O. Shu, A. R. Shuldiner, G. Sigurðsson, E. Sijbrands, A. Silveira, X. Sim, S. Sivapalaratnam, K. S. Small, W. Y. So, A. Stančáková, K. Stefansson, G. Steinbach, V. Steinthorsdottir, K. Stirrups, R. J. Strawbridge, H. M. Stringham, Q. Sun, C. Suo, A.-C. Syvänen, R. Takayanagi, F. Takeuchi, W. T. Tay, T. M. Teslovich, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, E. Tikkanen, J. Trakalo, E. Tremoli, M. D. Trip, F. J. Tsai, T. Tuomi, J. Tuomilehto, A. G. Uitterlinden, A. Valladares-Salgado, S. Vedantam, F. Veglia, B. F. Voight, C. Wang, N. J. Wareham, R. Wennauer, A. R. Wickremasinghe, T. Wilsgaard, J. F. Wilson, S. Wiltshire, W. Winckler, T. Y. Wong, A. R. Wood, J.-Y. Wu, Y. Wu, K. Yamamoto, T. Yamauchi, M. Yang, L. Yengo, M. Yokota, R. Young, D. Zabaneh, F. Zhang, R. Zhang, W. Zheng, P. Z. Zimmet, D. Altshuler, D. W. Bowden, Y. S. Cho, N. J. Cox, M. Cruz, C. L. Hanis, J. Kooner, J.-Y. Lee, M. Seielstad, Y. Y. Teo, M. Boehnke, E. J. Parra, J. C. Chambers, E. S. Tai, M. I. McCarthy, and A. P. Morris. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234–244, Mar. 2014. 31

[41] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, Aug. 2012. 37

[42] M. Druon, M. Druon, M. Druon, M. Druon, M. Druon, M. Druon, M. Druon, and M. Druon. *Les rois maudits: roman historique*. 2014. 26

[43] C. Dupin. *Forces productives et commerciales de la France...* Forces productives et commerciales de la France. Bachelier, 1827. 128

[44] J. Dupâquier. *Histoire de la population française Coffret 4 volumes : Volume 1, Des origines à la Renaissance. Volume 2, De la Renaissance à 1789. Volume 3, De 1789 à 1914. Volume 4, De 1914 à nos jours*. 125

[45] J. Y. Dutheil, G. Ganapathy, A. Hobolth, T. Mailund, M. K. Uyenoyama, and M. H. Schierup. Ancestral Population Genomics: The Coalescent Hidden Markov Model Approach. *Genetics*, 183(1):259–274, Sept. 2009. 37

[46] M. D. Edge and N. A. Rosenberg. Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 52:32–45, Aug. 2015. 36

[47] L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10):e1003905, Oct. 2013. 37, 125

[48] L. Excoffier and M. Foll. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27(9):1332–1334, May 2011. 37

[49] L. Fejerman, G. K. Chen, C. Eng, S. Huntsman, D. Hu, A. Williams, B. Pasaniuc, E. M. John, M. Via, C. Gignoux, S. Ingles, K. R. Monroe, L. N. Kolonel, G. Torres-Mejía, E. J. Pérez-Stable, E. González Burchard, B. E. Henderson, C. A. Haiman, and E. Ziv. Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. *Human Molecular Genetics*, 21(8):1907–1917, Apr. 2012. 29

[50] J. N. Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, 128(2):415–423, Oct. 2005. 92

[51] Y. Fichou, E. Génin, C. Le Maréchal, M.-P. Audrézet, V. Scotet, and C. Férec. Estimating the age of CFTR mutations predominantly found in Brittany (Western France). *Journal of Cystic Fibrosis*, 7(2):168–173, Mar. 2008. 126

[52] L. Fleuriot. *Les origines de la Bretagne: l'émigration*. Payot, 1980. 87

[53] O. Francois, M. Currat, N. Ray, E. Han, L. Excoffier, and J. Novembre. Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture. *Molecular Biology and Evolution*, 27(6):1257–1268, June 2010. 15

[54] M. L. Freedman, C. A. Haiman, N. Patterson, G. J. McDonald, A. Tandon, A. Waliszewska, K. Penney, R. G. Steen, K. Ardlie, E. M. John, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, S. A. Ingles, D. Altshuler, B. E. Henderson, and D. Reich. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences*, 103(38):14068–14073, Sept. 2006. 29

[55] Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, K. Prüfer, M. Meyer, J. Kelso, D. Reich, and S. Pääbo. An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, 524(7564):216–219, Aug. 2015. 17, 19

[56] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. F. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, M. Meyer, N. Zwyns, D. C. Salazar-García, Y. V. Kuzmin, S. G. Keates, P. A. Kosintsev, D. I. Razhev, M. P. Richards, N. V. Peristov, M. Lachmann, K. Douka, T. F. G. Higham, M. Slatkin, J.-J. Hublin, D. Reich, J. Kelso, T. B. Viola, and S. Pääbo. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7523):445–449, Oct. 2014. 17

[57] Q. Fu, C. Posth, M. Hajdinjak, M. Petr, S. Mallick, D. Fernandes, A. Furtwängler, W. Haak, M. Meyer, A. Mittnik, B. Nickel, A. Peltzer, N. Rohland, V. Slon, S. Talamo, I. Lazaridis, M. Lipson, I. Mathieson, S. Schiffels, P. Skoglund, A. P. Derevianko, N. Drozdov, V. Slavinsky, A. Tsybankov, R. G. Cremonesi, F. Mallegni, B. Gély, E. Vacca, M. R. G. Morales, L. G. Straus, C. Neugebauer-Maresch, M. Teschler-Nicola, S. Constantin, O. T. Moldovan, S. Benazzi, M. Peresani, D. Coppola, M. Lari, S. Ricci, A. Ronchitelli, F. Valentin, C. Thevenet, K. Wehrberger, D. Grigorescu, H. Rougier, I. Crevecoeur, D. Flas, P. Semal, M. A. Mannino, C. Cupillard, H. Bocherens, N. J. Conard, K. Harvati, V. Moiseyev, D. G. Drucker, J. Svoboda, M. P. Richards, D. Caramelli, R. Pinhasi, J. Kelso,

N. Patterson, J. Krause, S. Pääbo, and D. Reich. The genetic history of Ice Age Europe. *Nature*, 534(7606):200–205, June 2016. 19, 20

[58] W. Fu, T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, NHLBI Exome Sequencing Project, and J. M. Akey. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, Nov. 2012. 129

[59] C. Gamba, E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboróczki, I. Kővári, I. Pap, A. Anders, A. Whittle, J. Dani, P. Raczky, T. F. G. Higham, M. Hofreiter, D. G. Bradley, and R. Pinhasi. Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5(1), Dec. 2014. 19

[60] M. Gautier, A. Klassmann, and R. Vitalis. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources*, 17(1):78–90, Jan. 2017. 84

[61] M. Gautier and R. Vitalis. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, 28(8):1176–1177, Apr. 2012. 84

[62] A. Gibbons. Eruption made 536 'the worst year to be alive'. *Science*, 362(6416):733–734, Nov. 2018. 125

[63] E. Gilbert, S. O'Reilly, M. Merrigan, D. McGettigan, A. M. Molloy, L. C. Brody, W. Bodmer, K. Hutnik, S. Ennis, D. J. Lawson, J. F. Wilson, and G. L. Cavalleri. The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland. *Scientific Reports*, 7(1), Dec. 2017. 27, 31, 32, 33, 34, 123

[64] T. Günther and M. Jakobsson. Genes mirror migrations and cultures in prehistoric Europe — a population genomic perspective. *Current Opinion in Genetics & Development*, 41:115–123, Dec. 2016. 19

[65] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, The 1000 Genomes Project, C. D. Bustamante, D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. Wang, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, J. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, E. S. Lander, D. L. Altshuler, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, D. R. Bentley, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, H. Lehrach, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, M. Egholm, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, E. R. Mardis, R. K. Wilson, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, R. M. Durbin, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, R. A. Gibbs, D. Wheeler, M. Bainbridge, D. Challis, A. Sabo, F. Yu, J. Yu, J. Wang, X. Fang, X. Guo, R. Li, Y. Li, R. Luo, S. Tai, H. Wu, H. Zheng, X. Zheng, Y. Zhou, G. Li, J. Wang, H. Yang, G. T. Marth, E. P. Garrison, W. Huang, A. Indap, D. Kural, W.-P. Lee, W. F. Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, D. L. Altshuler, A. D. Ball, E. Banks, T. Bloom, B. L. Browning, K. Cibulskis, T. J. Fennell, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, D. B. Jaffe, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C.

Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, E. Shefler, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, C. D. Bustamante, A. G. Clark, A. Boyko, J. Degenhardt, S. Gravel, R. N. Gutenkunst, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, P. Flicek, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stutz, S. Humphray, M. Bauer, R. K. Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, A. Chakravarti, K. Ye, F. M. De La Vega, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, S. T. Sherry, R. Agarwala, H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, G. A. McVean, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, B. Desany, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. V. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. C. Melton, G. R. Abecasis, Y. Li, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. M. Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. Snyder, X. Zhan, S. Zollner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. C. Sahinalp, P. H. Sudmant, E. R. Mardis, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, D. Dooling, G. Weinstock, J. W. Wallis, M. C. Wendl, Q. Zhang, R. M. Durbin, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, D. M. Carter, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Q. Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, J. Stalker, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, M. Snyder, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, Y. Li, R. Luo, G. T. Marth, E. P. Garrison, D. Kural, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. S. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, July 2011. 17, 28, 129

[66] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Y. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prufer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Hober, B. Hoffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Paabo. A Draft Sequence of the Neandertal Genome. *Science*, 328(5979):710–722, May 2010. 17

[67] R. Grün, C. Stringer, F. McDermott, R. Nathan, N. Porat, S. Robertson, L. Taylor, G. Mortimer, S. Eggins, and M. McCulloch. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *Journal of Human Evolution*, 49(3):316–334, Sept. 2005. 17

[68] I. Gronau, M. J. Hubisz, B. Gulko, C. G. Danko, and A. Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43(10):1031–1034, Oct. 2011. 17, 37

[69] J. Gudmundsson, P. Sulem, D. F. Gudbjartsson, G. Masson, B. A. Agnarsson, K. R. Benediktsdottir, A. Sigurdsson, O. T. Magnusson, S. A. Gudjonsson, D. N. Magnusdottir, H. Johannsdottir, H. T. Helgadottir, S. N. Stacey, A. Jonasdottir, S. B. Olafsdottir, G. Thorleifsson, J. G. Jonasson, L. Tryggvadottir, S. Navarrete, F. Fuertes, B. T. Helfand, Q. Hu, I. E. Csiki, I. N. Mates, V. Jinga, K. K. H. Aben, I. M. van Oort, S. H. Vermeulen, J. L. Donovan, F. C. Hamdy, C.-F. Ng, P. K. F. Chiu, K.-M.

Lau, M. C. Y. Ng, J. R. Gulcher, A. Kong, W. J. Catalona, J. I. Mayordomo, G. V. Einarsson, R. B. Barkardottir, E. Jonsson, D. Mates, D. E. Neal, L. A. Kiemeney, U. Thorsteinsdottir, T. Rafnar, and K. Stefansson. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics*, 44(12):1326–1329, Dec. 2012. 131

[70] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, 5(10):e1000695, Oct. 2009. 37

[71] W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, and D. Reich. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, June 2015. 19, 20

[72] K. Harris and R. Nielsen. Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, 9(6):e1003521, June 2013. 17, 37

[73] K. Hatzikotoulas, A. Gilly, and E. Zeggini. Using population isolates in genetic association studies. *Briefings in Functional Genomics*, 13(5):371–377, Sept. 2014. 131

[74] J. Hawks. *The man who tried to catalog humanity*. 13

[75] S. C. Heath, I. G. Gut, P. Brennan, J. D. McKay, V. Bencko, E. Fabianova, L. Foretova, M. Georges, V. Janout, M. Kabesch, H. E. Krokan, M. B. Elvestad, J. Lissowska, D. Mates, P. Rudnai, F. Skorpen, S. Schreiber, J. M. Soria, A.-C. Syvänen, P. Meneton, S. Herçberg, P. Galan, N. Szeszenia-Dabrowska, D. Zaridze, E. Génin, L. R. Cardon, and M. Lathrop. Investigation of the fine structure of European populations with applications to disease association studies. *European Journal of Human Genetics*, 16(12):1413–1429, Dec. 2008. 31, 33, 93

[76] P. J. Heather, C. f. I. R. o. S. Stress, and C. o. S. i. H. Archaeoethnology, editors. *The Visigoths from the migration period to the seventh century: an ethnographic perspective ; [this volume contains the papers presented at the Fourth Conference on Studies in Historical Archaeoethnology, organized by the Center for Interdisziplinary Research on Social Stress, which was held in San Marino from 5th to 9th September 1996]*. Number 4 in Studies in historical archaeoethnology. Boydell Press [u.a.], Woodbridge, Suffolk, 1999. 22

[77] G. Hellenthal, G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. A Genetic Atlas of Human Admixture History. *Science*, 343(6172):747–751, Feb. 2014. 37

[78] B. M. Henn, S. Gravel, A. Moreno-Estrada, S. Acevedo-Acevedo, and C. D. Bustamante. Fine-scale population structure and the era of next-generation sequencing. *Human Molecular Genetics*, 19(R2):R221–R226, Oct. 2010. 132

[79] S. Hercberg, P. Galan, P. Preziosi, S. Bertrais, L. Mennen, D. Malvy, A.-M. Roussel, A. Favier, and S. Brian?on. The SU.VI.MAX Study: A Randomized, Placebo-Controlled Trial of the Health Effects of Antioxidant Vitamins and Minerals. *Archives of Internal Medicine*, 164(21):2335, Nov. 2004. 82

[80] T. Higham, K. Douka, R. Wood, C. B. Ramsey, F. Brock, L. Basell, M. Camps, A. Arrizabalaga, J. Baena, C. Barroso-Ruíz, C. Bergman, C. Boitard, P. Boscato, M. Caparrós, N. J. Conard, C. Draily, A. Froment, B. Galván, P. Gambassini, A. Garcia-Moreno, S. Grimaldi, P. Haesaerts, B. Holt, M.-J. Iriarte-Chiapusso, A. Jelinek, J. F. Jordá Pardo, J.-M. Maíllo-Fernández, A. Marom, J. Maroto, M. Menéndez, L. Metz, E. Morin, A. Moroni, F. Negrino, E. Panagopoulou, M. Peresani, S. Pirson, M. de la Rasilla, J. Riel-Salvatore, A. Ronchitelli, D. Santamaria, P. Semal, L. Slimak, J. Soler, N. Soler, A. Villaluenga, R. Pinhasi, and R. Jacobi. The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, 512(7514):306–309, Aug. 2014. 17, 19

[81] L. A. Hindorff, V. L. Bonham, L. C. Brody, M. E. C. Ginoza, C. M. Hutter, T. A. Manolio, and E. D. Green. Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, 19(3):175–185, Nov. 2017. 29

[82] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, June 2009. 29

[83] A. Hobolth, O. F. Christensen, T. Mailund, and M. H. Schierup. Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLoS Genetics*, 3(2):e7, 2007. 37

[84] A. Hobolth, J. Y. Dutheil, J. Hawks, M. H. Schierup, and T. Mailund. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Research*, 21(3):349–356, Mar. 2011. 37

[85] Z. Hofmanová, S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, D. Díez-del Molino, L. van Dorp, S. López, A. Kousathanas, V. Link, K. Kirsanow, L. M. Cassidy, R. Martiniano, M. Strobel, A. Scheu, K. Kotsakis, P. Halstead, S. Triantaphyllou, N. Kyparissi-Apostolika, D. Urem-Kotsou, C. Ziota, F. Adaktylou, S. Gopalan, D. M. Bobo, L. Winkelbach, J. Blöcher, M. Unterländer, C. Leuenberger, i. Çilingiroğlu, B. Horejs, F. Gerritsen, S. J. Shennan, D. G. Bradley, M. Currat, K. R. Veeramah, D. Wegmann, M. G. Thomas, C. Papageorgopoulou, and J. Burger. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences*, 113(25):6886–6891, June 2016. 20

[86] P. Hohenberg. Migrations et fluctuations démographiques dans la France rurale, 1836-1901. *Ann Économies Sociétés Civilis*, (29):461–497, 1974. 129

[87] H. Holm, D. F. Gudbjartsson, P. Sulem, G. Masson, H. T. Helgadottir, C. Zanon, O. T. Magnusson, A. Helgason, J. Saemundsdottir, A. Gylfason, H. Stefansdottir, S. Gretarsdottir, S. E. Matthiasson, G. Thorgeirsson, A. Jonasdottir, A. Sigurdsson, H. Stefansson, T. Werge, T. Rafnar, L. A. Kiemeney, B. Parvez, R. Muhammad, D. M. Roden, D. Darbar, G. Thorleifsson, G. B. Walters, A. Kong, U. Thorsteinsdottir, D. O. Arnar, and K. Stefansson. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genetics*, 43(4):316–320, Mar. 2011. 131

[88] K. E. Holsinger and B. S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, 10(9):639–650, Sept. 2009. 36

[89] J.-J. Hublin, A. Ben-Ncer, S. E. Bailey, S. E. Freidline, S. Neubauer, M. M. Skinner, I. Bergmann, A. Le Cabec, S. Benazzi, K. Harvati, and P. Gunz. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, 546(7657):289–292, June 2017. 17

[90] R. Hudson. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990. 39

[91] J. P. Huelsenbeck and P. Andolfatto. Inference of Population Structure Under a Dirichlet Process Model. *Genetics*, 175(4):1787–1802, Feb. 2007. 37

[92] J. Hussenet, P. Contant, and C. vendéen de recherches historiques, editors. *Détruisez la Vendée! regards croisés sur les victimes et destructions de la guerre de Vendée*. Centre vendéen de recherches historiques, La Roche-sur-Yon, 2007. OCLC: ocn173273949. 125

[93] J. E. Jackson. *A User's Guide to Principal Components.* John Wiley & Sons, Hoboken, 2005. 36

[94] P. R. Jansen, K. Watanabe, S. Stringer, N. Skene, J. Bryois, A. R. Hammerschlag, C. A. de Leeuw, J. Benjamins, A. B. Munoz-Manchado, M. Nagel, J. E. Savage, H. Tiemeier, T. White, J. Y. Tung, D. A. Hinds, V. Vacic, P. F. Sullivan, S. van der Sluis, T. J. Polderman, A. B. Smit, J. Hjerling-Leffler, E. J. van Someren, and D. Posthuma. Genome-wide Analysis of Insomnia (N=1,331,010) Identifies Novel Loci and Functional Pathways. Feb. 2018. 28

[95] Y. Jiang, M. P. Epstein, and K. N. Conneely. Assessing the Impact of Population Stratification on Association Studies of Rare Variation. *Human Heredity*, 76(1):28–35, 2013. 131

[96] E. R. Jones, G. Gonzalez-Fortes, S. Connell, V. Siska, A. Eriksson, R. Martiniano, R. L. McLaughlin, M. Gallego Llorente, L. M. Cassidy, C. Gamba, T. Meshveliani, O. Bar-Yosef, W. Müller, A. Belfer-Cohen, Z. Matskevich, N. Jakeli, T. F. G. Higham, M. Currat, D. Lordkipanidze, M. Hofreiter, A. Manica, R. Pinhasi, and D. G. Bradley. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, 6(1), Dec. 2015. 20

[97] T. Jonsson, J. K. Atwal, S. Steinberg, J. Snaedal, P. V. Jonsson, S. Bjornsson, H. Stefansson, P. Sulem, D. Gudbjartsson, J. Maloney, K. Hoyte, A. Gustafson, Y. Liu, Y. Lu, T. Bhangale, R. R. Graham, J. Huttenlocher, G. Bjornsdottir, O. A. Andreassen, E. G. Jönsson, A. Palotie, T. W. Behrens, O. T. Magnusson, A. Kong, U. Thorsteinsdottir, R. J. Watts, and K. Stefansson. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature*, 488(7409):96–99, Aug. 2012. 131

[98] I. Juric, S. Aeschbacher, and G. Coop. The Strength of Selection against Neanderthal Introgression. *PLOS Genetics*, 12(11):e1006340, Nov. 2016. 20

[99] M. Karakachoff, N. Duforet-Frebourg, F. Simonet, S. Le Scouarnec, N. Pellen, S. Lecointe, E. Charpentier, F. Gros, S. Cauchi, P. Froguel, N. Copin, the D.E.S.I.R. Study Group, T. Le Tourneau, V. Probst, H. Le Marec, S. Molinaro, B. Balkau, R. Redon, J.-J. Schott, M. Blum, and C. Dina. Fine-scale human genetic structure in Western France. *Eur J Hum Genet*, 23(6):831–836, June 2015. 27, 31, 33, 35, 84, 93, 123, 125, 126

[100] A. Keinan and A. G. Clark. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science*, 336(6082):740–743, May 2012. 129

[101] S. Kerminen, A. S. Havulinna, G. Hellenthal, A. R. Martin, A.-P. Sarin, M. Perola, A. Palotie, V. Salomaa, M. J. Daly, S. Ripatti, and M. Pirinen. Fine-Scale Genetic Structure in Finland. *G3: Genes\textbarGenomes\textbarGenetics*, 7(10):3459–3468, Oct. 2017. 27, 31, 39, 91, 111, 123, 124

[102] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43, 1982. 39

[103] W. C. Knowler, R. C. Williams, D. J. Pettitt, and A. G. Steinberg. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics*, 43(4):520–526, Oct. 1988. 131

[104] H. G. Koenigsberger. *Medieval Europe 400 - 1500*. Routledge, 1987. 24

[105] M. K. Kuhner. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, 22(6):768–770, Mar. 2006. 37

[106] P. Lanoë. Les noms en « -ac » et la ligne Loth, Nov. 2016. 108

[107] O. Lao, T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, M. Balascakova, J. Bertranpetit, L. A. Bindoff, D. Comas, G. Holmlund, A. Kouvatsi, M. Macek, I. Mollet, W. Parson, J. Palo, R. Ploski, A. Sajantila, A. Tagliabracci, U. Gether, T. Werge, F. Rivadeneira, A. Hofman, A. G. Uitterlinden, C. Gieger, H.-E. Wichmann, A. Rüther, S. Schreiber, C. Becker, P. Nürnberg, M. R. Nelson, M. Krawczak, and M. Kayser. Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18(16):1241–1248, Aug. 2008. 31, 33

[108] D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8(1):e1002453, Jan. 2012. 37, 39, 111, 124

[109] I. Lazaridis. The evolutionary history of human populations in Europe. *Current Opinion in Genetics & Development*, 53:21–27, Dec. 2018. 20, 21

[110] I. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, N. Rohland, S. Pfrengle, A. Furtwängler, A. Peltzer, C. Posth, A. Vasilakis, P. J. P. McGeorge, E. Konsolaki-Yannopoulou, G. Korres, H. Martlew,

M. Michalodimitrakis, M. Özsait, N. Özsait, A. Papathanasiou, M. Richards, S. A. Roodenberg, Y. Tzedakis, R. Arnott, D. M. Fernandes, J. R. Hughey, D. M. Lotakis, P. A. Navas, Y. Maniatis, J. A. Stamatoyannopoulos, K. Stewardson, P. Stockhammer, R. Pinhasi, D. Reich, J. Krause, and G. Stamatoyannopoulos. Genetic origins of the Minoans and Mycenaeans. *Nature*, Aug. 2017. 20

[111] I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak, S. Connell, K. Stewardson, E. Harney, Q. Fu, G. Gonzalez-Fortes, E. R. Jones, S. A. Roodenberg, G. Lengyel, F. Bocquentin, B. Gasparian, J. M. Monge, M. Gregg, V. Eshed, A.-S. Mizrahi, C. Meiklejohn, F. Gerritsen, L. Bejenaru, M. Blüher, A. Campbell, G. Cavalleri, D. Comas, P. Froguel, E. Gilbert, S. M. Kerr, P. Kovacs, J. Krause, D. McGettigan, M. Merrigan, D. A. Merriwether, S. O'Reilly, M. B. Richards, O. Semino, M. Shamoon-Pour, G. Stefanescu, M. Stumvoll, A. Tönjes, A. Torroni, J. F. Wilson, L. Yengo, N. A. Hovhannisyan, N. Patterson, R. Pinhasi, and D. Reich. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, Aug. 2016. 20

[112] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J.-M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J.-M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, and J. Krause. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, Sept. 2014. 18, 19, 20, 21

[113] J.-Y. Le Moing. *Noms de lieux de Bretagne: plus de 1200 noms expliqués*. Noms de lieux. Bonneton, Paris, 2004. OCLC: 232002154. 108

[114] R. Leprohon. *Vie et mort des Bretons sous Louis XIV*. Les bibliophiles de bretagne, 1984. 129

[115] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E. Royrvik, B. Cunliffe, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, D. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, and W. Bodmer. The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314, Mar. 2015. 27, 31, 32, 33, 39, 91, 111, 123, 124

[116] E. Levy-Lahad, R. Catane, S. Eisenberg, B. Kaufman, G. Hornreich, E. Lishinsky, M. Shohat, B. L. Weber, U. Beller, A. Lahad, and D. Halle. Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. *American Journal of Human Genetics*, 60(5):1059–1067, May 1997. 132

[117] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, July 2011. 17, 37

[118] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 319(5866):1100–1104, Feb. 2008. 15, 28, 33

[119] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, Dec. 2003. 39

[120] M. Lipson, A. Szécsényi-Nagy, S. Mallick, A. Pósa, B. Stégmár, V. Keerl, N. Rohland, K. Stewardson, M. Ferry, M. Michel, J. Oppenheimer, N. Broomandkhoshbacht, E. Harney, S. Nordenfelt, B. Llamas, B. Gusztáv Mende, K. Köhler, K. Oross, M. Bondár, T. Marton, A. Osztás, J. Jakucs, T. Paluch, F. Horváth, P. Csengeri, J. Koós, K. Sebők, A. Anders, P. Raczky, J. Regenye, J. P. Barna, S. Fábián, G. Serlegi, Z. Toldi, E. Gyöngyvér Nagy, J. Dani, E. Molnár, G. Pálfi, L. Márk, B. Melegh, Z. Bánfai, L. Domboróczki, J. Fernández-Eraso, J. Antonio Mujika-Alustiza, C. Alonso Fernández, J. Jiménez Echevarría, R. Bollongino, J. Orschiedt, K. Schierhold, H. Meller, A. Cooper, J. Burger, E. Bánffy, K. W. Alt, C. Lalueza-Fox, W. Haak, and D. Reich. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, 551(7680):368–372, Nov. 2017. 20

[121] Q. Liu, D. L. Nicolae, and L. S. Chen. Marbled Inflation From Population Structure in Gene-Based Association Studies With Rare Variants. *Genetic Epidemiology*, 37(3):286–292, Apr. 2013. 131

[122] W. Liu, M. Martinón-Torres, Y.-j. Cai, S. Xing, H.-w. Tong, S.-w. Pei, M. J. Sier, X.-h. Wu, R. L. Edwards, H. Cheng, Y.-y. Li, X.-x. Yang, J. M. B. de Castro, and X.-j. Wu. The earliest unequivocally modern humans in southern China. *Nature*, 526(7575):696–699, Oct. 2015. 17

[123] K. Lohse and L. A. F. Frantz. Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes. *Genetics*, 196(4):1241–1251, Apr. 2014. 37

[124] K. Lohse, R. J. Harrison, and N. H. Barton. A General Method for Calculating Likelihoods Under the Coalescent Process. *Genetics*, 189(3):977–987, Nov. 2011. 37

[125] T. Mailund, J. Y. Dutheil, A. Hobolth, G. Lunter, and M. H. Schierup. Estimating Divergence Time and Ancestral Effective Population Size of Bornean and Sumatran Orangutan Subspecies Using a Coalescent Hidden Markov Model. *PLoS Genetics*, 7(3):e1001319, Mar. 2011. 37

[126] T. Mailund, A. E. Halager, M. Westergaard, J. Y. Dutheil, K. Munch, L. N. Andersen, G. Lunter, K. Prüfer, A. Scally, A. Hobolth, and M. H. Schierup. A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS Genetics*, 8(12):e1003125, Dec. 2012. 37

[127] A.-S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergström, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvåg, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F.-X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, and E. Willerslev. A genomic history of Aboriginal Australia. *Nature*, 538(7624):207–214, Oct. 2016. 17

[128] G. Malécot. *Les mathématiques de l'hérédité*. Barnéoud frères, 1948. 36

[129] E. M. Mann. Little Ice Age, 2002. 126

[130] J. Mann. *The settlement of veterans in the Roman Empire*. PhD Thesis, University of London, 1956. 21

[131] A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo, and I. S. Kohane. Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*, 375(7):655–665, Aug. 2016. 29

[132] J. Marchini, L. R. Cardon, M. S. Phillips, and P. Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, May 2004. 28, 29

[133] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100(4):635–649, Apr. 2017. 29, 30, 31

[134] A. R. Martin, K. J. Karczewski, S. Kerminen, M. I. Kurki, A.-P. Sarin, M. Artomov, J. G. Eriksson, T. Esko, G. Genovese, A. S. Havulinna, J. Kaprio, A. Konradi, L. Korányi, A. Kostareva, M. Männikkö, A. Metspalu, M. Perola, R. B. Prasad, O. Raitakari, O. Rotar, V. Salomaa, L. Groop, A. Palotie, B. M. Neale, S. Ripatti, M. Pirinen, and M. J. Daly. Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. *The American Journal of Human Genetics*, 102(5):760–775, May 2018. 27, 31, 32, 33, 34

[135] R. Martiniano, L. M. Cassidy, R. Ó'Maoldúin, R. McLaughlin, N. M. Silva, L. Manco, D. Fidalgo, T. Pereira, M. J. Coelho, M. Serra, J. Burger, R. Parreira, E. Moran, A. C. Valera, E. Porfirio, R. Boaventura, A. M. Silva, and D. G. Bradley. The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLOS Genetics*, 13(7):e1006852, July 2017. 20

[136] I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, D. Fernandes, M. Ferry, B. Gamarra, G. G. Fortes, W. Haak, E. Harney, E. Jones, D. Keating, B. Krause-Kyora, I. Kucukkalipci, M. Michel, A. Mittnik, K. Nägele, M. Novak, J. Oppenheimer, N. Patterson, S. Pfrengle, K. Sirak, K. Stewardson, S. Vai, S. Alexandrov, K. W. Alt, R. Andreescu, D. Antonović, A. Ash, N. Atanassova, K. Bacvarov, M. B. Gusztáv, H. Bocherens, M. Bolus, A. Boroneanţ, Y. Boyadzhiev, A. Budnik, J. Burmaz, S. Chohadzhiev, N. J. Conard, R. Cottiaux, M. Čuka, C. Cupillard, D. G. Drucker, N. Elenski, M. Francken, B. Galabova, G. Ganetsovski, B. Gély, T. Hajdu, V. Handzhyiska, K. Harvati, T. Higham, S. Iliev, I. Janković, I. Karavanić, D. J. Kennett, D. Komšo, A. Kozak, D. Labuda, M. Lari, C. Lazar, M. Leppek, K. Leshtakov, D. L. Vetro, D. Los, I. Lozanov, M. Malina, F. Martini, K. McSweeney, H. Meller, M. Menđušić, P. Mirea, V. Moiseyev, V. Petrova, T. D. Price, A. Simalcsik, L. Sineo, M. Šlaus, V. Slavchev, P. Stanev, A. Starović, T. Szeniczey, S. Talamo, M. Teschler-Nicola, C. Thevenet, I. Valchev, F. Valentin, S. Vasilyev, F. Veljanovska, S. Venelinova, E. Veselovskaya, B. Viola, C. Virag, J. Zaninović, S. Zäuner, P. W. Stockhammer, G. Catalano, R. Krauß, D. Caramelli, G. Zariņa, B. Gaydarska, M. Lillie, A. G. Nikitin, I. Potekhina, A. Papathanasiou, D. Borić, C. Bonsall, J. Krause, R. Pinhasi, and D. Reich. The genomic history of southeastern Europe. *Nature*, 555(7695):197–203, Feb. 2018. 20

[137] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, K. Sirak, C. Gamba, E. R. Jones, B. Llamas, S. Dryomov, J. Pickrell, J. L. Arsuaga, J. M. B. de Castro, E. Carbonell, F. Gerritsen, A. Khokhlov, P. Kuznetsov, M. Lozano, H. Meller, O. Mochalov, V. Moiseyev, M. A. R. Guerra, J. Roodenberg, J. M. Vergès, J. Krause, A. Cooper, K. W. Alt, D. Brown, D. Anthony, C. Lalueza-Fox, W. Haak, R. Pinhasi, and D. Reich. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, Dec. 2015. 19, 20

[138] I. Mathieson and G. McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–246, Mar. 2012. 28, 129, 131

[139] I. Mathieson and G. McVean. Demography and the Age of Rare Variants. *PLoS Genetics*, 10(8):e1004528, Aug. 2014. 129

[140] S. Matsumura and P. Forster. Generation time and effective population size in Polar Eskimos. *Proceedings of the Royal Society B: Biological Sciences*, 275(1642):1501–1508, July 2008. 92

[141] J. McClellan and M.-C. King. Genetic Heterogeneity in Human Disease. *Cell*, 141(2):210–217, Apr. 2010. 131

[142] G. McVean. A Genealogical Interpretation of Principal Components Analysis. *PLoS Genetics*, 5(10):e1000686, Oct. 2009. 36

[143] D. J. Meltzer. *First peoples in a new world: colonizing ice age America*. University of California Press, 2009. 17

[144] A. Mittnik, C.-C. Wang, S. Pfrengle, M. Daubaras, G. Zariņa, F. Hallgren, R. Allmäe, V. Khartanovich, V. Moiseyev, M. Tõrv, A. Furtwängler, A. Andrades Valtueña, M. Feldman, C. Economou, M. Oinonen, A. Vasks, E. Balanovska, D. Reich, R. Jankauskas, W. Haak, S. Schiffels, and J. Krause. The genetic prehistory of the Baltic Sea region. *Nature Communications*, 9(1), Dec. 2018. 20

[145] T. Möller, H. Anderson, T. Aareleid, T. Hakulinen, H. Storm, L. Tryggvadottir, I. Corazziari, E. Mugno, and EUROPREVAL Working Group. Cancer prevalence in Northern Europe: the EUROPREVAL study. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 14(6):946–957, June 2003. 28

[146] A. E. Mourant, A. C. Kopeć, and K. Domaniewska-Sobczak. *The distribution of the human blood groups, and other polymorphisms*. Oxford monographs on medical genetics. Oxford University Press, London, 2d ed edition, 1976. 14

[147] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zollner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science*, 337(6090):100–104, July 2012. 129

[148] R. Nielsen, J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, and E. Willerslev. Tracing the peopling of the world through genomics. *Nature*, 541:302, Jan. 2017. 16, 17, 18, 19

[149] M. Nordborg. On the Probability of Neanderthal Ancestry. *The American Journal of Human Genetics*, 63(4):1237–1240, Oct. 1998. 17

[150] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, Nov. 2008. 28, 31, 32, 33, 38

[151] J. Novembre and B. M. Peter. Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics & Development*, 41:98–105, Dec. 2016. 27, 36, 38

[152] J. Novembre and S. Ramachandran. Perspectives on Human Population Structure at the Cusp of the Sequencing Era. *Annual Review of Genomics and Human Genetics*, 12(1):245–274, Sept. 2011. 28, 39

[153] J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, May 2008. 15, 36

[154] T. D. O'Connor, W. Fu, J. C. Mychaleckyj, B. Logsdon, P. Auer, C. S. Carlson, S. M. Leal, J. D. Smith, M. J. Rieder, M. J. Bamshad, D. A. Nickerson, and J. M. Akey. Rare Variation Facilitates Inferences of Fine-Scale Population Structure in Humans. *Molecular Biology and Evolution*, 32(3):653–660, Mar. 2015. 129

[155] T. D. O'Connor, A. Kiezun, M. Bamshad, S. S. Rich, J. D. Smith, E. Turner, NHLBIGO Exome Sequencing Project, ESP Population Genetics, Statistical Analysis Working Group, S. M. Leal, and J. M. Akey. Fine-Scale Patterns of Population Stratification Confound Rare Variant Association Tests. *PLoS ONE*, 8(7):e65834, July 2013. 131

[156] T. E. of Encyclopaedia Britannica. Poitou, Sept. 2017. 126

[157] I. Olalde, S. Brace, M. E. Allentoft, I. Armit, K. Kristiansen, T. Booth, N. Rohland, S. Mallick, A. Szécsényi-Nagy, A. Mittnik, E. Altena, M. Lipson, I. Lazaridis, T. K. Harper, N. Patterson,

N. Broomandkhoshbacht, Y. Diekmann, Z. Faltyskova, D. Fernandes, M. Ferry, E. Harney, P. de Knijff, M. Michel, J. Oppenheimer, K. Stewardson, A. Barclay, K. W. Alt, C. Liesau, P. Ríos, C. Blasco, J. V. Miguel, R. M. García, A. A. Fernández, E. Bánffy, M. Bernabò-Brea, D. Billoin, C. Bonsall, L. Bonsall, T. Allen, L. Büster, S. Carver, L. C. Navarro, O. E. Craig, G. T. Cook, B. Cunliffe, A. Denaire, K. E. Dinwiddy, N. Dodwell, M. Ernée, C. Evans, M. Kuchařík, J. F. Farré, C. Fowler, M. Gazenbeek, R. G. Pena, M. Haber-Uriarte, E. Haduch, G. Hey, N. Jowett, T. Knowles, K. Massy, S. Pfrengle, P. Lefranc, O. Lemercier, A. Lefebvre, C. H. Martínez, V. G. Olmo, A. B. Ramírez, J. L. Maurandi, T. Majó, J. I. McKinley, K. McSweeney, B. G. Mende, A. Mod, G. Kulcsár, V. Kiss, A. Czene, R. Patay, A. Endrődi, K. Köhler, T. Hajdu, T. Szeniczey, J. Dani, Z. Bernert, M. Hoole, O. Cheronet, D. Keating, P. Velemínský, M. Dobeš, F. Candilio, F. Brown, R. F. Fernández, A.-M. Herrero-Corral, S. Tusa, E. Carnieri, L. Lentini, A. Valenti, A. Zanini, C. Waddington, G. Delibes, E. Guerra-Doce, B. Neil, M. Brittain, M. Luke, R. Mortimer, J. Desideri, M. Besse, G. Brücken, M. Furmanek, A. Hałuszko, M. Mackiewicz, A. Rapiński, S. Leach, I. Soriano, K. T. Lillios, J. L. Cardoso, M. P. Pearson, P. Włodarczak, T. D. Price, P. Prieto, P.-J. Rey, R. Risch, M. A. Rojo Guerra, A. Schmitt, J. Serralongue, A. M. Silva, V. Smrčka, L. Vergnaud, J. Zilhão, D. Caramelli, T. Higham, M. G. Thomas, D. J. Kennett, H. Fokkens, V. Heyd, A. Sheridan, K.-G. Sjögren, P. W. Stockhammer, J. Krause, R. Pinhasi, W. Haak, I. Barnes, C. Lalueza-Fox, and D. Reich. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*, 555(7695):190–196, Feb. 2018. 20

[158] P. F. Palamara. ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinformatics*, 32(19):3032–3034, Oct. 2016. 92

[159] P. F. Palamara, T. Lencz, A. Darvasi, and I. Pe'er. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *The American Journal of Human Genetics*, 91(5):809–822, Nov. 2012. 37

[160] P. F. Palamara and I. Pe'er. Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):i180–i188, July 2013. 37

[161] B. Pasaniuc, N. Zaitlen, G. Lettre, G. K. Chen, A. Tandon, W. H. L. Kao, I. Ruczinski, M. Fornage, D. S. Siscovick, X. Zhu, E. Larkin, L. A. Lange, L. A. Cupples, Q. Yang, E. L. Akylbekova, S. K. Musani, J. Divers, J. Mychaleckyj, M. Li, G. J. Papanicolaou, R. C. Millikan, C. B. Ambrosone, E. M. John, L. Bernstein, W. Zheng, J. J. Hu, R. G. Ziegler, S. J. Nyante, E. V. Bandera, S. A. Ingles, M. F. Press, S. J. Chanock, S. L. Deming, J. L. Rodriguez-Gil, C. D. Palmer, S. Buxbaum, L. Ekunwe, J. N. Hirschhorn, B. E. Henderson, S. Myers, C. A. Haiman, D. Reich, N. Patterson, J. G. Wilson, and A. L. Price. Enhanced Statistical Tests for GWAS in Admixed Populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genetics*, 7(4):e1001371, Apr. 2011. 29

[162] N. Patterson, A. L. Price, and D. Reich. Population Structure and Eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006. 82, 89

[163] N. Pellen. *Hasard, coïncidence, prédestination... et s'il fallait plutôt regarder du côté de nos aïeux? Analyse démographique et historique des réseaux généalogiques et des structures familiales des patients atteints de mucoviscidose en Bretagne.* PhD Thesis, Université de Versailles-Saint-Quentin-en-Yvelines, 2012. 129

[164] E. Persyn, R. Redon, L. Bellanger, and C. Dina. The impact of a fine-scale population stratification on rare variant association test results. *PLOS ONE*, 13(12):e0207677, Dec. 2018. 131

[165] D. Petkova, J. Novembre, and M. Stephens. Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, 48(1):94–100, Jan. 2016. 38

[166] J. K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Güldemann, B. Kure, S. W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P.-R. Loh, J. Lachance, J. Mountain, C. D. Bustamante, B. Berger, S. A. Tishkoff, B. M. Henn, M. Stoneking, D. Reich, and B. Pakendorf. The genetic prehistory of southern Africa. *Nature Communications*, 3(1), Jan. 2012. 17

[167] J. K. Pickrell and J. K. Pritchard. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8(11):e1002967, Nov. 2012. 37

[168] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, Jan. 2014. 17

[169] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics*, 5(6):e1000519, June 2009. 29, 37

[170] A. L. Price, M. E. Weale, N. Patterson, S. R. Myers, A. C. Need, K. V. Shianna, D. Ge, J. I. Rotter, E. Torres, K. D. Taylor, D. B. Goldstein, and D. Reich. Long-Range LD Can Confound Genome Scans in Admixed Populations. *The American Journal of Human Genetics*, 83(1):132–135, July 2008. 84

[171] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, June 2000. 36, 37

[172] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, Sept. 2007. 82, 84, 89

[173] M. Raghavan, P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T. W. Stafford Jr, L. Orlando, E. Metspalu, M. Karmin, K. Tambets, S. Rootsi, R. Mägi, P. F. Campos, E. Balanovska, O. Balanovsky, E. Khusnutdinova, S. Litvinov, L. P. Osipova, S. A. Fedorova, M. I. Voevoda, M. DeGiorgio, T. Sicheritz-Ponten, S. Brunak, S. Demeshchenko, T. Kivisild, R. Villems, R. Nielsen, M. Jakobsson, and E. Willerslev. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505(7481):87–91, Jan. 2014. 17

[174] A. Raj, M. Stephens, and J. K. Pritchard. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, 197(2):573–589, June 2014. 37

[175] S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, 102(44):15942–15947, Nov. 2005. 17

[176] B. Rannala and Z. Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656, Aug. 2003. 37

[177] M. Rasmussen, Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, M. Bertalan, K. Nielsen, M. T. P. Gilbert, Y. Wang, M. Raghavan, P. F. Campos, H. M. Kamp, A. S. Wilson, A. Gledhill, S. Tridico, M. Bunce, E. D. Lorenzen, J. Binladen, X. Guo, J. Zhao, X. Zhang, H. Zhang, Z. Li, M. Chen, L. Orlando, K. Kristiansen, M. Bak, N. Tommerup, C. Bendixen, T. L. Pierre, B. Grønnow, M. Meldgaard, C. Andreasen, S. A. Fedorova, L. P. Osipova, T. F. G. Higham, C. B. Ramsey, T. v. O. Hansen, F. C. Nielsen, M. H. Crawford, S. Brunak, T. Sicheritz-Pontén, R. Villems, R. Nielsen, A. Krogh, J. Wang, and E. Willerslev. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757–762, Feb. 2010. 17

[178] A. Raveane, S. Aneli, F. Montinaro, G. Athanasiadis, S. Barlera, G. Birolo, G. Boncoraglio, A. M. 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, J.-M. Dugoujon, P. Galan, J. Giemza, T. Kivisild, M. Melhaoui, M. Metspalu, S. Myers, L. M. Pereira, F.-X. Ricaut, F. Brisighelli, I. Cardinali, V. Grugni, H. Lancioni, V. L. Pascali, A. Torroni, O. Semino, G. Matullo, A. Achilli, A. Olivieri, and C. Capelli. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *bioRxiv*, Dec. 2018. 137

[179] D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, and S. Pääbo. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, Dec. 2010. 17

[180] J. Roe. Lactose intolerance around the world, July 2017. 14

[181] N. A. Rosenberg, M. D. Edge, J. K. Pritchard, and M. W. Feldman. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evolution, Medicine, and Public Health*, 2019(1):26–34, Jan. 2019. 31

[182] L. Saag, L. Varul, C. L. Scheib, J. Stenderup, M. E. Allentoft, L. Saag, L. Pagani, M. Reidla, K. Tambets, E. Metspalu, A. Kriiska, E. Willerslev, T. Kivisild, and M. Metspalu. Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. *Current Biology*, 27(14):2185–2193.e6, July 2017. 20

[183] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, Oct. 2002. 84

[184] S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, and D. Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357, Mar. 2014. 17, 20

[185] S. Sankararaman, N. Patterson, H. Li, S. Pääbo, and D. Reich. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics*, 8(10):e1002947, Oct. 2012. 17

[186] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics*, 82(2):290–303, Feb. 2008. 37

[187] S. Sanna, B. Li, A. Mulas, C. Sidore, H. M. Kang, A. U. Jackson, M. G. Piras, G. Usala, G. Maninchedda, A. Sassu, F. Serra, M. A. Palmas, W. H. Wood, I. Njølstad, M. Laakso, K. Hveem, J. Tuomilehto, T. A. Lakka, R. Rauramaa, M. Boehnke, F. Cucca, M. Uda, D. Schlessinger, R. Nagaraja, and G. R. Abecasis. Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS Genetics*, 7(7):e1002198, July 2011. 132

[188] S. Sawyer, G. Renaud, B. Viola, J.-J. Hublin, M.-T. Gansauge, M. V. Shunkov, A. P. Derevianko, K. Prüfer, J. Kelso, and S. Pääbo. Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proceedings of the National Academy of Sciences*, page 201519905, Nov. 2015. 17

[189] A. Scally and R. Durbin. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13(10):745–753, Oct. 2012. 17

[190] S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, Aug. 2014. 17, 37, 92, 126

[191] S. Schiffels, W. Haak, P. Paajanen, B. Llamas, E. Popescu, L. Loe, R. Clarke, A. Lyons, R. Mortimer, D. Sayer, C. Tyler-Smith, A. Cooper, and R. Durbin. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications*, 7(1), Dec. 2016. 20, 129

[192] C. M. Schlebusch, P. Skoglund, P. Sjodin, L. M. Gattepaille, D. Hernandez, F. Jay, S. Li, M. De Jongh, A. Singleton, M. G. B. Blum, H. Soodyall, and M. Jakobsson. Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science*, 338(6105):374–379, Oct. 2012. 17

[193] C. M. Schlebusch and H. Soodyall. Extensive Population Structure in San, Khoe, and Mixed Ancestry Populations from Southern Africa Revealed by 44 Short 5-SNP Haplotypes. *Human Biology*, 84(6):695–724, Dec. 2012. 17

[194] J. G. Schraiber and J. M. Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, Dec. 2015. 36, 37, 38

[195] V. Scotet, D. Gillet, I. Duguépéroux, M.-P. Audrézet, G. Bellis, B. Garnier, M. Roussey, G. Rault, P. Parent, M. De Braekeleer, C. Férec, and R. Mucoviscidose Bretagne et Pays de L. Spatial and temporal distribution of cystic fibrosis and of its mutations in Brittany, France: a retrospective study from 1960. *Human Genetics*, 111(3):247–254, Sept. 2002. 126

[196] A. Seguin-Orlando, T. S. Korneliussen, M. Sikora, A.-S. Malaspinas, A. Manica, I. Moltke, A. Albrechtsen, A. Ko, A. Margaryan, V. Moiseyev, T. Goebel, M. Westaway, D. Lambert, V. Khartanovich, J. D. Wall, P. R. Nigst, R. A. Foley, M. M. Lahr, R. Nielsen, L. Orlando, and E. Willerslev. Genomic structure in Europeans dating back at least 36,200 years. *Science*, 346(6213):1113–1118, Nov. 2014. 17

[197] S. Sheehan, K. Harris, and Y. S. Song. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics*, 194(3):647–662, July 2013. 37

[198] P. Skoglund and M. Jakobsson. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences*, 108(45):18301–18306, Nov. 2011. 17

[199] P. Skoglund, H. Malmstrom, A. Omrak, M. Raghavan, C. Valdiosera, T. Gunther, P. Hall, K. Tambets, J. Parik, K.-G. Sjogren, J. Apel, E. Willerslev, J. Stora, A. Gotherstrom, and M. Jakobsson. Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. *Science*, 344(6185):747–750, May 2014. 20

[200] P. Skoglund, H. Malmstrom, M. Raghavan, J. Stora, P. Hall, E. Willerslev, M. T. P. Gilbert, A. Gotherstrom, and M. Jakobsson. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science*, 336(6080):466–469, Apr. 2012. 19, 20

[201] V. Slon, F. Mafessoni, B. Vernot, C. de Filippo, S. Grote, B. Viola, M. Hajdinjak, S. Peyrégne, S. Nagel, S. Brown, K. Douka, T. Higham, M. B. Kozlikin, M. V. Shunkov, A. P. Derevianko, J. Kelso, M. Meyer, K. Prüfer, and S. Pääbo. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, 561(7721):113–116, Sept. 2018. 17

[202] M. Steinrücken, J. S. Paul, and Y. S. Song. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theoretical Population Biology*, 87:51–61, Aug. 2013. 37

[203] Strabon. De l'Aquitaine a la vallee du Rhone. In *Géographie*, volume Livre Quatre. 21

[204] C. B. Stringer and I. Barnes. Deciphering the Denisovans. *Proceedings of the National Academy of Sciences*, page 201522477, Dec. 2015. 17

[205] P. Sulem, D. F. Gudbjartsson, G. B. Walters, H. T. Helgadottir, A. Helgason, S. A. Gudjonsson, C. Zanon, S. Besenbacher, G. Bjornsdottir, O. T. Magnusson, G. Magnusson, E. Hjartarson, J. Saemundsdottir, A. Gylfason, A. Jonasdottir, H. Holm, A. Karason, T. Rafnar, H. Stefansson, O. A. Andreassen, J. H. Pedersen, A. I. Pack, M. C. H. de Visser, L. A. Kiemeney, A. J. Geirsson, G. I. Eyjolfsson, I. Olafsson, A. Kong, G. Masson, H. Jonsson, U. Thorsteinsdottir, I. Jonsdottir, and K. Stefansson. Identification of low-frequency variants associated with gout and serum uric acid levels. *Nature Genetics*, 43(11):1127–1130, Nov. 2011. 131

[206] Tacitus. *Histories*. 21

[207] H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, May 2005. 37

[208] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad GO, Seattle GO, and on behalf of the NHLBI Exome Sequencing Project. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, 337(6090):64–69, July 2012. 129

[209] J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, Feb. 2017. 126

[210] The Genome of the Netherlands Consortium, L. C. Francioli, A. Menelaou, S. L. Pulit, F. van Dijk, P. F. Palamara, C. C. Elbers, P. B. T. Neerincx, K. Ye, V. Guryev, W. P. Kloosterman, P. Deelen, A. Abdellaoui, E. M. van Leeuwen, M. van Oven, M. Vermaat, M. Li, J. F. J. Laros, L. C. Karssen, A. Kanterakis, N. Amin, J. J. Hottenga, E.-W. Lameijer, M. Kattenberg, M. Dijkstra, H. Byelas, J. van Setten, B. D. C. van Schaik, J. Bot, I. J. Nijman, I. Renkens, T. Marschall, A. Schönhuth, J. Y. Hehir-Kwa, R. E. Handsaker, P. Polak, M. Sohail, D. Vuzman, F. Hormozdiari, D. van Enckevort, H. Mei, V. Koval, M. H. Moed, K. J. van der Velde, F. Rivadeneira, K. Estrada, C. Medina-Gomez, A. Isaacs, S. A. McCarroll, M. Beekman, A. J. M. de Craen, H. E. D. Suchiman, A. Hofman, B. Oostra, A. G. Uitterlinden, G. Willemsen, L. C. Study, M. Platteel, J. H. Veldink, L. H. van den Berg, S. J. Pitts, S. Potluri, P. Sundar, D. R. Cox, S. R. Sunyaev, J. T. d. Dunnen, M. Stoneking, P. de Knijff, M. Kayser, Q. Li, Y. Li, Y. Du, R. Chen, H. Cao, N. Li, S. Cao, J. Wang, J. A. Bovenberg, I. Pe'er, P. E. Slagboom, C. M. van Duijn, D. I. Boomsma, G.-J. B. van Ommen, P. I. W. de Bakker, M. A. Swertz, and C. Wijmenga. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8):818–825, Aug. 2014. 27, 31, 33, 91, 129

[211] The UK10K Consortium. The UK10k project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, Oct. 2015. 28, 129, 130

[212] D. C. Thomas and J. S. Witte. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 11(6):505–512, June 2002. 28

[213] C. Tian, R. M. Plenge, M. Ransom, A. Lee, P. Villoslada, C. Selmi, L. Klareskog, A. E. Pulver, L. Qi, P. K. Gregersen, and M. F. Seldin. Analysis and Application of European Genetic Substructure Using 300 K SNP Information. *PLoS Genetics*, 4(1):e4, 2008. 31

[214] N. Tintle, H. Aschard, I. Hu, N. Nock, H. Wang, and E. Pugh. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17. *Genetic Epidemiology*, 35(S1):S56–S60, 2011. 131

[215] H. K. Tiwari, J. Barnholtz-Sloan, N. Wineinger, M. A. Padilla, L. K. Vaughan, and D. B. Allison. Review and Evaluation of Methods Correcting for Population Stratification with a Focus on Underlying Statistical Principles. *Human Heredity*, 66(2):67–86, 2008. 131

[216] K. R. Veeramah, D. Wegmann, A. Woerner, F. L. Mendez, J. C. Watkins, G. Destro-Bisol, H. Soodyall, L. Louie, and M. F. Hammer. An Early Divergence of KhoeSan Ancestors from Those of Other Modern Humans Is Supported by an ABC-Based Analysis of Autosomal Resequencing Data. *Molecular Biology and Evolution*, 29(2):617–630, Feb. 2012. 17

[217] B. Vernot and J. M. Akey. Complex History of Admixture between Modern Humans and Neandertals. *The American Journal of Human Genetics*, 96(3):448–453, Mar. 2015. 17

[218] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. African populations and the evolution of human mitochondrial DNA. *Science (New York, N.Y.)*, 253(5027):1503–1507, Sept. 1991. 16

[219] B. J. Vilhjálmsson, J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, T. Hayeck, H.-H. Won, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P. M. Visscher, P. Kraft, N. Patterson, A. L. Price, S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K.-H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. DeLisi, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodrguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, J. Grove, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. C. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K.-Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lnnqvist, M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, P. B. Mortensen, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Mller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S.-Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietilinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H.-C. So, C. C. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Sderman, S. Thirumalai, D. Toncheva, P. A. Tooney, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, J. Q. Wu, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St. Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, M. C. O'Donovan, P. Kraft, D. J. Hunter, M. Adank, H. Ahsan, K. Aittomäki, L. Bagli-

etto, S. Berndt, C. Blomquist, F. Canzian, J. Chang-Claude, S. J. Chanock, L. Crisponi, K. Czene, N. Dahmen, I. d. S. Silva, D. Easton, A. H. Eliassen, J. Figueroa, O. Fletcher, M. Garcia-Closas, M. M. Gaudet, L. Gibson, C. A. Haiman, P. Hall, A. Hazra, R. Hein, B. E. Henderson, A. Hofman, J. L. Hopper, A. Irwanto, M. Johansson, R. Kaaks, M. G. Kibriya, P. Lichtner, S. Lindström, J. Liu, E. Lund, E. Makalic, A. M. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97(4):576–592, Oct. 2015. 31

[220] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, 4(3):e72, Mar. 2006. 84

[221] J. D. Wall, M. A. Yang, F. Jay, S. K. Kim, E. Y. Durand, L. S. Stevison, C. Gignoux, A. Woerner, M. F. Hammer, and M. Slatkin. Higher Levels of Neanderthal Ancestry in East Asians than in Europeans. *Genetics*, 194(1):199–209, May 2013. 17

[222] K. Wang, H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. A. Sleiman, C. E. Kim, C. Hou, E. Frackelton, R. Chiavacci, N. Takahashi, T. Sakurai, E. Rappaport, C. M. Lajonchere, J. Munson, A. Estes, O. Korvatska, J. Piven, L. I. Sonnenblick, A. I. Alvarez Retuerto, E. I. Herman, H. Dong, T. Hutman, M. Sigman, S. Ozonoff, A. Klin, T. Owley, J. A. Sweeney, C. W. Brune, R. M. Cantor, R. Bernier, J. R. Gilbert, M. L. Cuccaro, W. M. McMahon, J. Miller, M. W. State, T. H. Wassink, H. Coon, S. E. Levy, R. T. Schultz, J. I. Nurnberger, J. L. Haines, J. S. Sutcliffe, E. H. Cook, N. J. Minshew, J. D. Buxbaum, G. Dawson, S. F. A. Grant, D. H. Geschwind, M. A. Pericak-Vance, G. D. Schellenberg, and H. Hakonarson. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, 459(7246):528–533, May 2009. 131

[223] K. M. Waters, D. O. Stram, M. T. Hassanein, L. Le Marchand, L. R. Wilkens, G. Maskarinec, K. R. Monroe, L. N. Kolonel, D. Altshuler, B. E. Henderson, and C. A. Haiman. Consistent Association of Type 2 Diabetes Risk Variants Found in Europeans in Diverse Racial and Ethnic Groups. *PLoS Genetics*, 6(8):e1001078, Aug. 2010. 29

[224] L. A. Weiss, D. E. Arking, M. J. Daly, A. Chakravarti, D. E. Arking, C. W. Brune, K. West, A. O'Connor, G. Hilton, R. L. Tomlinson, A. B. West, E. H. Cook Jr, A. Chakravarti, L. A. Weiss, T. Green, S.-C. Chang, S. Gabriel, C. Gates, E. M. Hanson, A. Kirby, J. Korn, F. Kuruvilla, S. McCarroll, E. M. Morrow, B. Neale, S. Purcell, R. Sasanfar, C. Sougnez, C. Stevens, D. Altshuler, J. Gusella, S. L. Santangelo, P. Sklar, R. Tanzi, M. J. Daly, R. Anney, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, C. Betancur, S. Bölte, P. F. Bolton, J. Brian, S. E. Bryson, J. D. Buxbaum, I. Cabrito, G. Cai, R. M. Cantor, E. H. Cook Jr, H. Coon, J. Conroy, C. Correia, C. Corsello, E. L. Crawford, M. L. Cuccaro, G. Dawson, M. de Jonge, B. Devlin, E. Duketis, S. Ennis, A. Estes, P. Farrar, E. Fombonne, C. M. Freitag, L. Gallagher, D. H. Geschwind, J. Gilbert, M. Gill, C. Gillberg, J. Goldberg, A. Green, J. Green, S. J. Guter, J. L. Haines, J. F. Hallmayer, V. Hus, S. M. Klauck, O. Korvatska, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, X.-Q. Liu, C. Lord, L. J. Lotspeich, E. Maestrini, T. Magalhaes, W. Mahoney, C. Mantoulan, H. McConachie, C. J. McDougle, W. M. McMahon, C. R. Marshall, J. Miller, N. J. Minshew, A. P. Monaco, J. Munson, J. I. Nurnberger Jr, G. Oliveira, A. Pagnamenta, K. Papanikolaou, J. R. Parr, A. D. Paterson, M. A. Pericak-Vance, A. Pickles, D. Pinto, J. Piven, D. J. Posey, A. Poustka, F. Poustka, R. Regan, J. Reichert, K. Renshaw, W. Roberts, B. Roge, M. L. Rutter, J. Salt, G. D. Schellenberg, S. W. Scherer, V. Sheffield, J. S. Sutcliffe, P. Szatmari, K. Tansey, A. P. Thompson, J. Tsiantis, H. Van Engeland, A. M. Vicente, V. J. Vieland, F. Volkmar, S. Wallace, T. H. Wassink, E. M. Wijsman, K. Wing, K. Wittemeyer, B. L. Yaspan, L. Zwaigenbaum, E. M. Morrow, S.-Y. Yoo, R. Sean Hill, N. M. Mukaddes, S. Balkhy, G. Gascon, S. Al-Saad, A. Hashmi, J. Ware, R. M. Joseph, E. LeClair, J. N. Partlow, B. Barry, C. A. Walsh, D. Pauls, I. Moilanen, H. Ebeling, M.-L. Mattila, S. Kuusikko, K. Jussila, J. Ignatius, R. Sasanfar, A. Tolouei, M. Ghadami, M. Rostami, A. Hosseinipour, M. Valujerdi, S. L. Santangelo, K. Andresen, B. Winkloski, S. Haddad, L. Kunkel, Z. Kohane, T. Tran, S. Won Kong, S. B. O'Neil, E. M. Hanson, R. Hundley, I. Holm, H. Peters, E. Baroni, A. Cangialose,

L. Jackson, L. Albers, R. Becker, C. Bridgemohan, S. Friedman, K. Munir, R. Nazir, J. Palfrey, A. Schonwald, E. Simmons, L. A. Rappaport, J. Gauthier, L. Mottron, R. Joober, E. Fombonne, G. Rouleau, K. Rehnstrom, L. von Wendt, and L. Peltonen. A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, 461(7265):802–808, Oct. 2009. 131

[225] S. Wright. THE GENETICAL STRUCTURE OF POPULATIONS. *Annals of Eugenics*, 15(1):323–354, Jan. 1949. 36

[226] C. Zhang, S.-S. Dong, J.-Y. Xu, W.-M. He, and T.-L. Yang. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, Oct. 2018. 88

[227] E. Ziv and E. G. Burchard. Human population structure and genetic association studies. *Pharmacogenomics*, 4(4):431–441, July 2003. 28

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Joanna GIEMZA

Fine-scale genetic population structure in France

La structure génétique à fine échelle de population en France

Résumé
La structure génétique à fine échelle des populations humaines est intéressante pour deux raisons principales : 1) elle reflète des événements historiques et démographiques, 2) elle informe la recherche sur les études d'association de maladies. Cette thèse a pour objectif de procéder à une analyse approfondie de la structure génétique de la population de France métropolitaine dans un premier temps, en de façon plus détaillée de la population du nord-ouest de la France, et de mettre en lumière les événements historiques, démographiques et culturels qui l'ont façonnés, en tirant parti de trois jeux de données (SU.VI.MAX/3C et PREGO).
Au niveau de la France, nous rapportons la corrélation entre les données génétiques et les lieux de naissance d'individus appartenant à deux cohortes françaises indépendantes (1 414 et 770 individus) et identifions six groupes, concordants entre les jeux de données. La deuxième étude tire parti de la cohorte PREGO, qui comprend 3 234 personnes ayant trois générations d'ascendance liée à des régions spécifiques du nord-ouest de la France. Je révèle une structure à fine échelle à un niveau sans précédent (154 sous-populations).historique de la France et des explications potentielles de la prévalence de différentes maladies dans cette région du nord-ouest. Dans l'ensemble, mes travaux de thèse indiquent des niveaux substantiels de stratification de la population dans une région géographiquement limitée, probablement en raison de différents antécédents démographiques dans la région.

Mots clés
génétique des populations, structure démographique à échelle fine, histoire démographique

Abstract
Fine-scale genetic structure in human populations is interesting for two main reasons: 1), it reflects historical and demographic events, 2) it informs research on disease association studies.
This thesis aims to perform a thorough analysis of the genetic structure of the population from continental France, in particular Northwestern France, and shed light on the historical, demographic and cultural events that have shaped it, by taking advantage of three genome-wide datasets (SU.VI.MAX/3C and PREGO)
At the country level we report the correlation between genetic data and birthplaces of individuals in two independent French cohorts (1,414 and 770 individuals in SU.VI.MAX and 3C, respectively) and identify six clusters, concordant between datasets, and may correspond to ancient political, cultural and geographical borders. The second study takes advantage of the PREGO cohort including 3,234 individuals with three generations of ancestry linked to specific regions of Northwestern France and reveals fine-scale structure at an unprecedented level (154 subpopulations). The resulting genetic clusters and the characterisation of their effective population size and ancestry proportions compared to other European groups provide important and novel insights into the historical peopling of France and potential explanations for different disease prevalence within this northwestern region.
Overall, my thesis work indicate substantial levels of population stratification within a geographically limited region likely caused by different demographic histories across the region.

Key Words
Population genetics, fine-scale population structure, demographic history