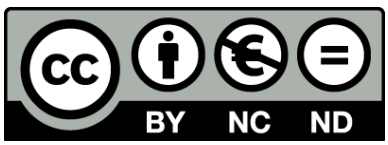
UNIVERSITAT DE BARCELONA

Human population genetics in the Mediterranean region. From single markers to whole-genome sequencing

Genètica de poblacions humanes del Mediterrani. Des de marcadors únics fins la seqüenciació del genoma complet

Miguel Martin Álvarez Álvarez



Human population genetics in the Mediterranean region. From single markers to whole-genome sequencing

Doctoral thesis presented by

Miguel Martín Álvarez Álvarez

in solicitation of the degree of

Doctor of Philosophy awarded by the University of Barcelona

Directed by

Dr. Pedro Moral Castrillo

and

Dr. Georgios Athanasiadis

Doctorate Programme of Biodiversity

Department of Evolutionary Biology, Ecology and Environmental Sciences

Faculty of Biology

Dr. Pedro Moral Castrillo
Director

Dr. Georgios Athanasiadis
Director

Dr. Maria Esther Esteban Torné
Tutor

Miguel Martín Álvarez Álvarez
Doctorate student

Genètica de poblacions humanes del Mediterrani. Des de marcadors únics fins la seqüenciació del genoma complet

Memòria presentada per

Miguel Martín Álvarez Álvarez

per optar al grau de

Doctor per la Universitat de Barcelona

Dirigida per

Dr. Pedro Moral Castrillo

i

Dr. Georgios Athanasiadis

Programa de Doctorat en Biodiversitat

Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals

Facultat de Biologia

Dr. Pedro Moral Castrillo
Director

Dr. Georgios Athanasiadis
Director

Dr. Maria Esther Esteban Tomé
Tutor

Miguel Martín Álvarez Álvarez
Doctorand

A la memoria de mi abuela, Julia

In memory of my grandmother, Julia

ACKNOWLEDGEMENTS

I would like to thank everybody that has accompanied and supported me along these PhD years.

My supervisors, Pedro Moral and Georgios Athanasiadis, who have taught me so much.

All the people in the Moral lab: Esther, Daniela, Robert, Magda, Marc, Elies, it was great to do research with you guys.

Claudia, Jordi, Helena, Laia, Flors, Maria, Bia, Ana, Àlex, Águeda, Nora, Marta, thanks for being there.

My family, Salva, Marti, Nico, Arturo and Santa and, of course, Rosie, I love you all.

ABSTRACT

A crucial part in designing a population genetics study is to determine the type and amount of genetic data required. In the case of analyses that deal with large-scale population processes, or that focus on some specific candidate genes, sparse genetic data with a high degree of variation is an option that allows for larger sample sizes. However, to obtain more refined results, such as estimating the date of an admixture event, or inferring historical changes in population size, the use of array-based or whole-genome sequencing data is required. Here, I present four studies in which the density of genetic markers employed vary in accordance to the depth of the analysis, contributing to the knowledge on the population genetics of Mediterranean populations.

The first survey supports the role of the Mediterranean sea as a historical barrier to gene flow, making use of three single nucleotide polymorphisms (SNP) located within or around *LIN28B*, a gene associated with cancer. One of these markers, rs221639, shows a specially high degree of variation between Mediterranean populations.

The second article analyzes the presence of a sub-Saharan genetic component in four CAD-associated genomic regions in Mediterranean populations. This component is more prevalent in the North African coast, suggesting a more intense sub-Saharan gene flow.

In the third study, the use of array-based data allows to refine previous estimates of the Sephardic component in current Iberian populations. The incorporation of neighbouring populations to the analyses shows a gene flow process from the Iberian Peninsula outwards, reflecting a migration pattern followed by the expelled Sephardic Jews.

Finally, deep-coverage whole-genome sequencing data from the Spanish Eastern Pyrenean population, combined with accurate genealogical information, allows to analyze the demographic history of this human group with a high degree of detail. Namely, the Spanish Eastern Pyreneans appear as a distinct group within the Iberian populations, closely related to the Basques. In addition, this human group presents fine-scale population structure, and it has undergone a historical isolation process involving a reduction of the effective size, whose epidemiological consequence has been a significant depletion of many rare and highly deleterious mutations.

CONTENTS

Table of Contents

GENERAL INTRODUCTION

HUMAN POPULATION GENETICS

The demographic and epidemiological history of human populations from the Mediterranean region has been thoroughly studied historically, although a number of questions remain largely unresolved. In this work, I present four original research articles in which gradually extensive genetic data are used to provide insights into this topic. To give context to these results, the following introduction summarises important topics on human population genetics, as well as on the history of the human populations from the Mediterranean region.

Milestones in human population genetics

The modern concept of anthropology as the study of the nature and origins of the different human populations dates back to the 18th century, although humans have been trying to provide answers to these questions since the first signs of behavioural modernity: first through legends and religion, and later with philosophy, history and natural science. The development of human population genetics – the systematic study of human genetic variation – has proved to be an invaluable complement to history, archaeology, cultural anthropology, linguistics and palaeoanthropology in the task of unravelling human demographic history.

Early milestones

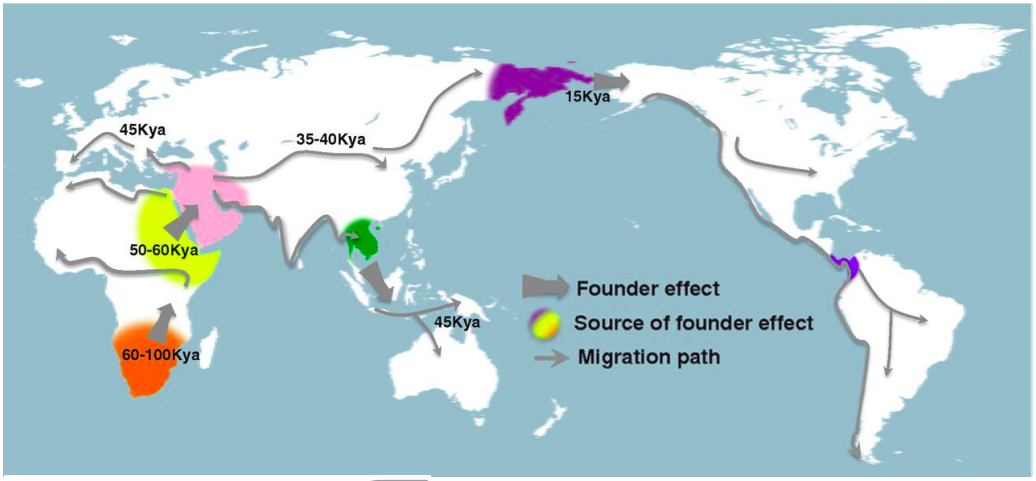
There have been a number of milestones in the field of human population genetics, which are concisely enumerated in Figure 1. The mid-19th century saw the enunciation of the theory of evolution by Charles Darwin and the discovery of the fundamental laws of genetic inheritance by Gregor Mendel. In the early 20th century, the integration of these paradigms by Fisher[1], and the formulation of the Hardy-Weinberg principle[2], settled the ground for the development of

population genetics by Fisher, alongside Haldane and Wright[3]. Later on, the discovery of the nature of DNA[4] and its double-helix structure[5], as well as the invention of DNA sequencing methods[6] and of the polymerase chain reaction (PCR)[7], allowed for the comparison of genetic variation between populations through the study of genetic polymorphisms – DNA loci that present two or more variants. Other contributions worth mentioning are, for example, the neutral theory of molecular evolution[8] and the development of the coalescent theory[9] (see *Coalescent-based demographic analyses*). The latter lead to important discoveries, such as finding that the most recent common ancestor (MRCA) of present-day human mitochondria existed in Africa 200-140 thousand years ago (KYA), thus supporting the African origin of modern humankind[10].

The early study of genetic variation between populations resolved other important human evolutionary debates. For instance, the general migration patterns in the worldwide dispersion of humans were inferred through the study of protein polymorphisms mainly from aboriginal populations[11] (Figure 2). Meanwhile, the debatable concept of race was debunked by recognising a mostly continuous genetic variation across continents with the intra-continental level as bearer of most of this variation (based on the allele distribution of several sets of microsatellites, restriction fragment length polymorphisms (RFLP), and *Alu* insertions, genotyped in populations from Asia, Africa and Europe[12]) (see *Types of genetic markers*).

Figure 1 (next page). Major milestones in the field of human population genetics (taken from Jobling et al., 2013)[13].

1786	Recognition of language families
1856	Discovery of Neanderthal type specimen
1859	Publication of Darwin's "The Origin of Species"
1866	Publication of Mendel's "Experiments in Plant Hybrids"
1871	Publication of Darwin's "The Descent of Man"
1900	Discovery of first genetic polymorphism—ABO blood group (Landsteiner)
1908	Hardy–Weinberg principle formulated
1918	Fisher reconciles Darwin's natural selection and Mendel's mechanism of inheritance
1925	*Australopithecus* fossil described from South Africa
1930–32	Fisher, Haldane & Wright publish the foundations of modern population genetics
1944	DNA shown to be heritable material
1949	Radiocarbon dating introduced
1953	Double-helical structure of DNA described
1956	Human chromosome number described
1957	Hemoglobin amino acid sequences determined
1959	Y chromosome shown to be sex-determining
1966	Genetic code deciphered
1968	Neutral theory of molecular evolution (Kimura)
1969	Internet first successfully tested
1977	Publication of DNA sequencing methods
1978	First human restriction fragment length polymorphisms (RFLPs) described
1978	First human *in vitro* fertilization
1980	First genome (φX174 bacteriophage) sequenced
1981	Human mitochondrial DNA (mtDNA) genome sequenced
1984	DNA fingerprinting (minisatellites) discovered
1984	DNA-DNA hybridization shows human–chimpanzee common ancestry
1985	Invention of polymerase chain reaction (PCR)
1985	First human ancient DNA results published
1985	First Y-chromosomal polymorphism described
1987	Development of laser-induced fluorescent detection of DNA
1987	African origin of human mtDNA identified
1988	Launch of Human Genome Project
1989	Development of capillary electrophoresis for sequencing
1990	First human microsatellites described
1991	Human Genome Diversity Project proposed
1994	Publication of "The History and Geography of Human Genes" (Cavalli-Sforza et al.)
1996	First mammal cloned from adult cell (Dolly)
1997	First Neanderthal mtDNA sequence
1999	First human chromosome sequenced (Chr 22)
2001	Release of draft human genome sequence
2002	Release of draft mouse and *Plasmodium* genome sequences
2002	Human Genome Diversity Project (HGDP) Cell Line Panel released
2004	First maps of copy-number variation published
2005	First-generation human Haplotype Map (HapMap) published
2005	Release of draft chimpanzee genome sequence
2005	First development of next-generation sequencing methods
2006	1 Mb of Neanderthal genomic sequence published
2007	First large-scale genomewide association studies
2007	First personal human genome resequenced (Venter)
2007	Second-generation human Haplotype Map (HapMap) published
2009	Exome capture and sequencing methods published
2010	Denisovan mtDNA and genome sequences published
2010	1000 Genomes Project pilot study published
2012	All great ape genomes now sequenced



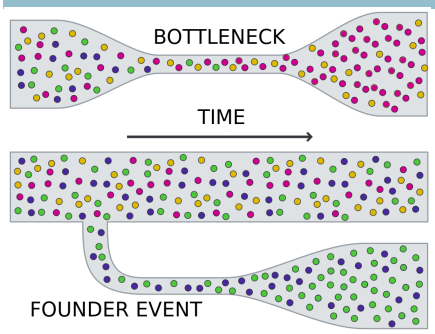


Figure 2. (up) Times and routes for the main migrations followed by modern humans. Founder effect refers to the loss of genetic diversity and drastic changes in allele frequencies that occur to small migrant groups with respect to the larger parent population (adapted from Henn et al., 2012)[14], equivalent to the effect of a population bottleneck (left; adapted from Jobling et al., 2013)[13].

The Human Genome Project

A genome is the complete DNA sequence of an individual, including both coding and non-coding nuclear and mitochondrial DNA. The haploid human genome (i.e. the genetic information provided by one parent) has been found to consist of almost 3 billion nucleotide bases. The first whole-genome sequencing (WGS) of the human genome was carried out as part of the Human Genome Project (HGP), launched in 1990 by the National Institutes of Health (NIH) in the United States, alongside a private initiative by Craig Venter[15]. The researchers involved followed a hierarchical shotgun method[16], based on the sequencing technique mainly used at that time: the Sanger sequencing[6]. In 2001, a general draft of the human genome was published in two separated papers[17,18], claiming to have covered more than 96% of the

euchromatin. In 2004, it was announced that ~99% of the euchromatic genome had been sequenced[19]. The main legacy of the HGP was the creation of a publicly available human reference genome sequence for the assembly of new sequence reads (see *Advances in sequencing techniques*). The reference genome has since been updated by the Genome Reference Consortium (GRC) to include several corrections and the closing of most gaps. Since 2013, the official reference is version GRCh38, in which 12 gaps have been recently closed thanks to the ultralong-read Nanopore sequencing[20].

The introduction of high-throughput technologies in the 2000s has revolutionised the study of human evolution. Before that, the scope of such studies tended to be limited to the analysis of inter-continental genetic differences, but the use of thousands to millions of genetic markers has since facilitated the study of population structure, migration, admixture, natural selection, and even hybridisation with archaic humans on a sub-continental level. For more on these technologies, see G*enotyping arrays* and *Advances in sequencing techniques*.

Genetic databases

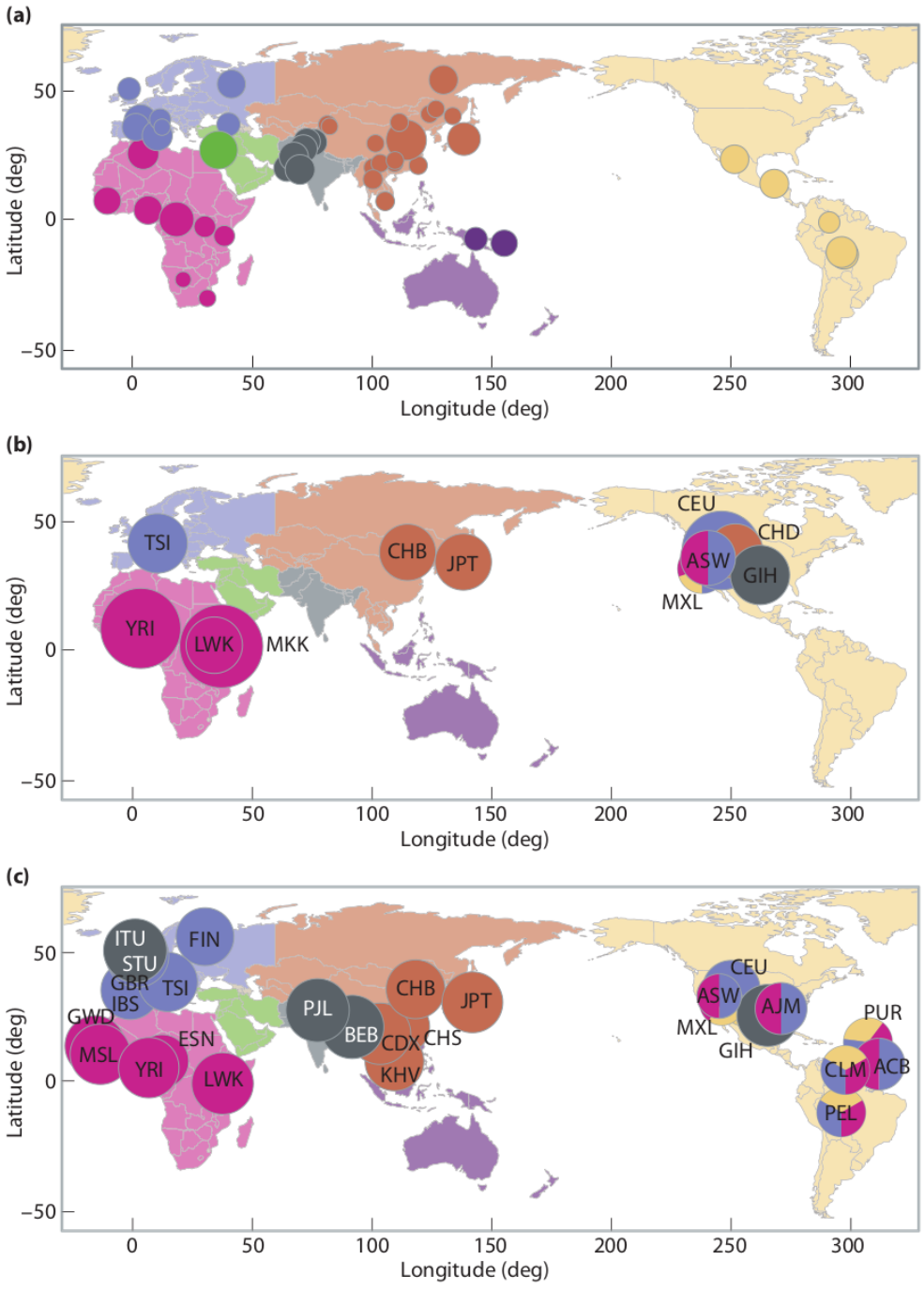
The decrease of sequencing and genotyping costs has been key in the creation of large genetic databases of general human populations. Among the pioneers we could mention the HapMap project[21], which sequenced 11 human groups of African, Asian and European ancestry, cataloguing ~6 millions of new SNPs, as well as the SNP array-based Human Genome Diversity Project (HGDP)[22], which covered 1,043 individuals from 51 world-wide populations. More recently, WGS data have taken the lead, starting with the 1000 Genomes Project (1000G)[23], which includes 2,504 low-coverage genome sequences from 26 populations in representation of all continents except Oceania. This has been followed by the Simons Genome Diversity Project (SGDP)[24] and the Estonian Biocentre Human Genome Diversity Panel (EGDP)[25], encompassing 300 and 483 high-coverage genomes, respectively, each from more than 140 world-wide populations. The two latter examples show a shift towards very deep sequencing of many diverse populations and, even though sample sizes are small, the further reduction in sequencing costs will presumably allow the increase of sample sizes yet again.

These databases have been a source of samples for many demographic studies[26], and have provided control data to epidemiological ones[27]. They can also be used as a reference panel for inferring the haplotypes of unphased data (i.e. "phasing") with programs like Eagle2[28], as well as for the imputation of missing variants in more sparse genetic data[29].

Genome browsers store and administer the information produced by international projects such as HGP, ENCODE or 1000G. They are a useful tool for fast accessing and browsing online demographic and clinical information about a group of polymorphisms or a genomic region. Some examples are the European Bioinformatics Institute (EBI-EMBL) Ensembl, and the University of California-Santa Cruz (UCSC) Genome Browser. Other browsers are provided by the NIH's National Center for Biotechnology Information (NCBI): GenBank, for instance, contains information about entire genes or other sequences, while dbSNP provides information about single SNPs, and OMIM is a database for Mendelian diseases, just to name a few.

There are genetic databases that also include phenotypic information, such as anthropometric measurements or disease status. These databases generally allow data access to research groups upon request. The provided datasets can then be used to increase the statistical power of an association study. On the other hand, there are repositories of data from association studies that allow unrestricted access for scientific purposes, such as Broad Institute's LD-hub[30]. In this case, to prevent tracking participants from their genetic information, raw data are substituted by summary statistics obtained in the original association study (usually effect sizes and p-values). These can be combined with the results of a new study through meta-analysis[31].

Figure 3 (next page). Samples included in three public genetic databases, (a) HGDP, (b) HapMap, and (c) 1000G. Populations are represented by circles of area proportional to sample size. Circle colour indicates the geographical ancestry of the samples. A description of the 1000G population codes is provided in the Appendix (I) (taken from Jobling et al., 2013)[13].


(a)
Latitude (deg)
50
0
−50
0
50
100
150
200
250
300
Longitude (deg)
(b)
Latitude (deg)
TSI
CHB
JPT
CEU
CHD
ASW
GIH
MXL
YRI
LWK
MKK
Longitude (deg)
(c)
Latitude (deg)
ITU
FIN
STU
GBR
TSI
IBS
GWD
MSL
ESN
YRI
LWK
PJL
BEB
CHB
JPT
CDX
CHS
KHV
CEU
ASW
AJM
MXL
GIH
PUR
CLM
ACB
PEL
Longitude (deg)

In the last years, very ambitious large-scale projects that include tens of thousands of both case and control samples from the same population are taking place. One purpose of these studies is to identify virtually all the rare variants present in a given population, which will allow to explain the missing heritability for many complex diseases. A recently finished example is the deCODE project[32], which sequenced 2,636 Icelanders to a median depth of 20× to be used as a reference panel for imputation on another 104,220 chip-typed Icelandic samples. Interestingly, the whole set of samples accounts for a 35.47% of Iceland's native population (293,858 as of 2018, according to the National Statistical Institute of Iceland – https://statice.is/).

Bioinformatics

Finally, it is important to point out the pivotal role of bioinformatics in the recent advances of human population genetics. The analysis of massive amounts of genetic data produced by high-throughput techniques implies high levels of computational complexity, making the adoption of computational approaches necessary in order to conduct the research within a reasonable period of time.

For instance, the alignment of sequence reads to the reference genome sequence is executed by the use of specialised software such as the Burrows-Wheeler Aligner (BWA)[33]. Then, variant calling – determination of the most probable genotype at each locus – is performed using software packages such as SAMtools[34] (Figure 4). In order to efficiently explore and modify large genotype datasets, there are programs like VCFtools[35] and PLINK[36], as well as UNIX-environment commands like *awk* or *grep*.

For statistical analysis of omics data, R is a popular programming language that allows to quickly perform a wide variety of tests and easily customise plots for the visualisation of the results. R further provides a flexible framework in which bioinformaticians develop new packages that can be shared with the research community.

There is also a whole variety of specialised software that implements state-of-the-art methods for genetic analysis, such as ADMIXTURE[37] (see *Applications of human genetic data*). Most of these programs require to be run in a command-line interpreter, or even on a remote computer

cluster when the analyses demand high computing resources. Coding *ad hoc* scripts in high-level programming languages, such as Python and Perl, is customary when specialised software for the task in question does not exist. The different programs and scripts needed for the analysis of a dataset can be efficiently organised in a pipeline using UNIX commands.

Figure 4. Visualisation of the sequence reads mapped to a section of the reference genome sequence using SAMtools. The first line shows the genome coordinates, the second line shows the reference sequence, and the third line shows the consensus sequence inferred from the aligned reads, which are shown along the following lines. Positions containing a dot indicate a match to the reference sequence, while an A, C, G or T indicate the potential presence of an alternate allele in the read. In each line, two different reads are separated by empty positions. In this case, the consensus sequence matches the reference sequence entirely, since the low number of observed alternate alleles in some positions do not conclusively suggest the existence of a polymorphism (taken from Cerami, 2013)[38].

Human genetic variation

Early sequencing techniques allowed population geneticists to gradually shift from the use of anthropometric measures and blood protein groups (which act as proxies of genetic variation) to the direct observation of DNA polymorphisms. Along the next sections, I present i)a summary of some important types of genetic variation, ii)a description of the main high-throughput genotyping techniques currently used, and iii)an introduction to neutral and non-neutral variation.

Types of genetic markers

A single-nucleotide polymorphism (SNP) is a genomic position in which at least two of the possible four nitrogenous bases occur within a population[39]. SNPs are the most commonly used genetic markers nowadays, and the most abundant: out of the 88 million polymorphisms identified in the human genome by the 1000G[23], 84.7 million are SNPs. From these, only ~8 million are considered common – i.e. they have a minor allele frequency (MAF) >5%. Thanks to the development of high-throughput techniques, which have dedicated sections ahead, it is now customary to analyse large numbers of genome-wide distributed SNPs in order to achieve high discriminant power. This is important in population genetics studies, as it allows to find differences between closely related groups. SNPs also have an important role in the study of the heritability of complex diseases, since their predominance makes them suitable tags of risk factors spread across the whole genome.

Structural variants are another group of polymorphisms that collectively affect ~20 million bases in a typical human genome[23]. These include: large indels, consisting in the presence or absence of a nucleotide sequence in relation to the ancestral state; copy-number variants (CNVs), formed by copies of genomic sections in tandem; transposable elements (e.g. *Alu* elements), which are sequences that can change their location within the genome as well as make copies of themselves; nuclear mitochondrial DNA segments (NUMTs); and chromosomal inversions.

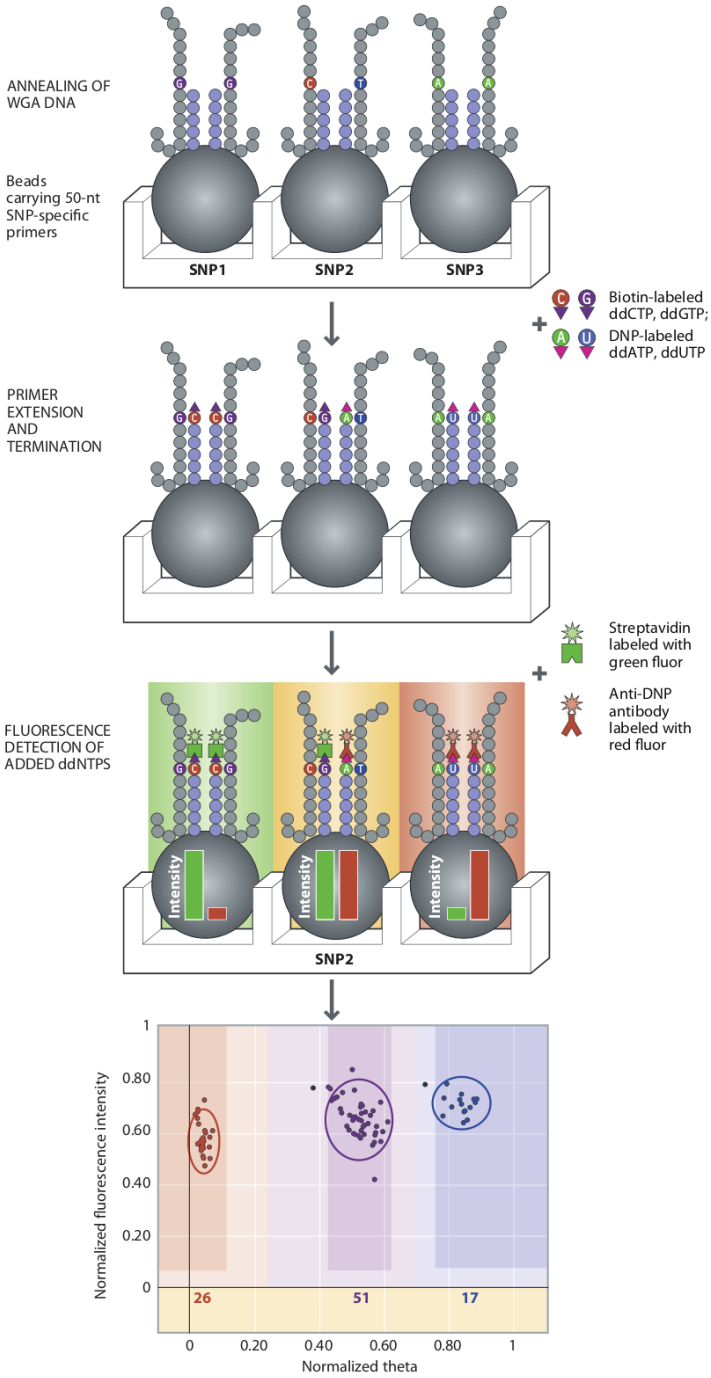
Before the popularisation of high-throughput techniques, population genetics typically employed low numbers of genetic markers, some of which showed high polymorphic levels. An instance of these classic markers are short tandem repeats (STR), also known as microsatellites. A STR consists on a short motif of 1-7 base pairs (bp) repeated in tandem a variable number of times. The abundance of STRs, and their high polymorphism and heterozygosity, are among the reasons that they are extensively used in forensics: a group of just 13 of these loci conforms a unique genetic fingerprint of a person[40].

Finally, the non-recombinant fraction of the Y chromosome and the mitochondrial DNA have also been widely used in population genetics, due to their relatively manageable size, high degree of polymorphism, and the fact that they follow a patrilineal and matrilineal pattern of inheritance, respectively, which can provide information about the sex bias involved in gene flow[41]. Different haplogroups are classified according to variation of STRs and SNPs.

Genotyping arrays

SNP genotyping arrays, also known as genotyping chips, allow the profiling of samples for common variation while saving part of the costs of sequencing a whole genome. A typical array consists in a solid surface that contains, for each SNP of interest, an allele-specific oligonucleotide cluster (Figure 5). Custom arrays are also designed to genotype SNPs for which significant or suggestive association with a specific trait has been proposed. This way, researchers can aim to replicate significant findings and test for association at promising variants[42].

Figure 5 (next page). Example of the working mechanism of a genotyping array[10]. In this case, the oligonucleotides are attached to beads. Upon annealing of the amplified DNA regions (WGA: whole-genome amplification) with their complementary primers, each cluster produces a characteristic fluorescent signal according to the genotype of the sample (taken from Jobling et al., 2013)[13].


ANNEALING OF WGA DNA
Beads carrying 50-nt SNP-specific primers
SNP1
SNP2
SNP3
Biotin-labeled ddCTP, ddGTP;
DNP-labeled ddATP, ddUTP
PRIMER EXTENSION AND TERMINATION
Streptavidin labeled with green fluor
Anti-DNP antibody labeled with red fluor
FLUORESCENCE DETECTION OF ADDED ddNTPS
Intensity
SNP2
Normalized fluorescence intensity
1
0.80
0.60
0.40
0.20
0
26
51
17
0
0.20
0.40
0.60
0.80
1
Normalized theta

Ascertainment bias

The design of genotyping arrays has an intrinsic problem, known as ascertainment bias, operating on several levels. The insufficient coverage with which the target population is originally sequenced causes many variants in each individual to not be captured. Since rare variants have less chances than common ones of being "called" in at least one individual[43], this results in a systematic deviation of the distribution of MAFs in a population, also known as site frequency spectrum (SFS), towards alleles of intermediate frequency[44] (Figure 6). This poses a problem for methods that use the SFS to infer demographic history, as well as for genome-wide association studies (GWAS) due to the important role that rare variants take in the heritability of many traits. Secondly, genotyping arrays are strongly biased towards capturing variation and disease-linked variants in populations of European ancestry, since those are the populations where the SNPs that are covered by the majority of arrays have been ascertained. Therefore, they fail to capture variation specific to populations from other continents[45]. This can lead to mis-estimates of heterozygosity rates[46] and effective population size (N_e)[47], problems in the inference of population structure[48], and cause missing heritability in GWAS.

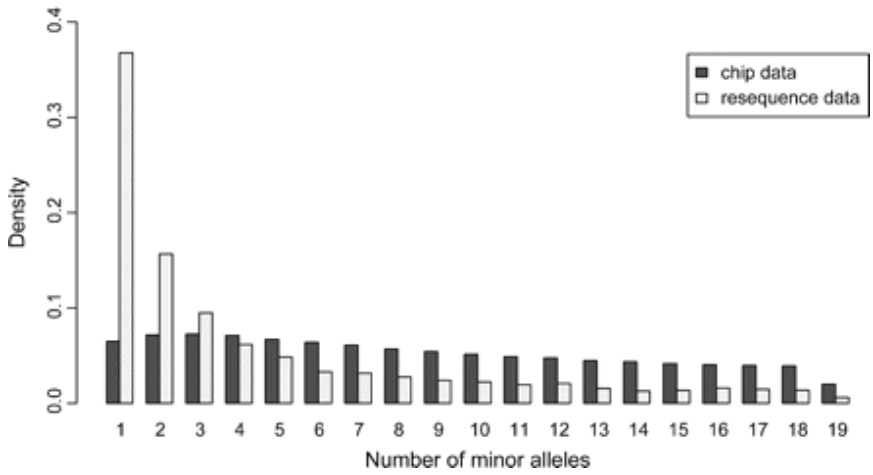


Figure 6. Comparison of the resulting SFS of 19 African Americans employing either chip or WGS data. Chip data has a clear deficiency of rare variants and an excess of variants of intermediate frequencies (taken from Albrechtsen et al., 2010)[49].

Some initiatives have contributed to palliate the ascertainment bias through the design of arrays for genotyping SNP variation that has been carefully ascertained in different world populations, like the Affymetrix Human Origins array[50] and the African Genome Variation Project[51].

Advances in sequencing techniques

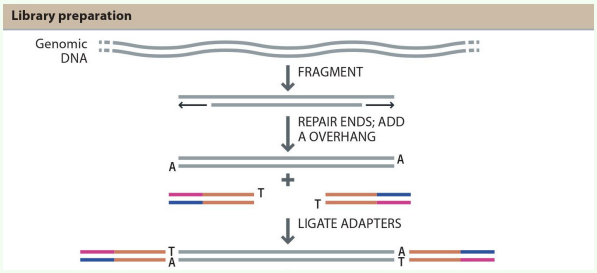
The number of SNPs that can be genotyped with array technologies has increased simultaneously with a reduction of the price per sample. Nevertheless, since technological advances have also lowered sequencing costs, many studies are shifting towards the use of WGS data, as this provides information on population-specific low-frequency variants.

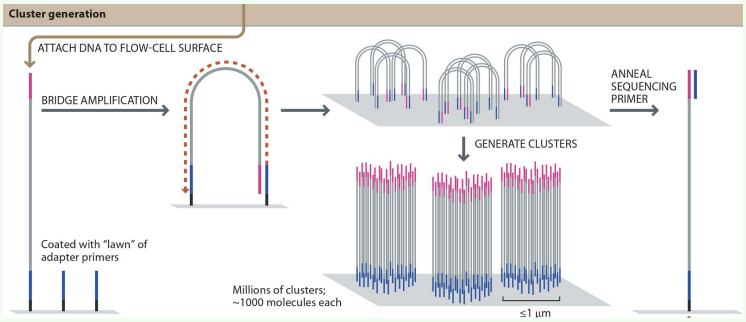
Next-generation sequencing

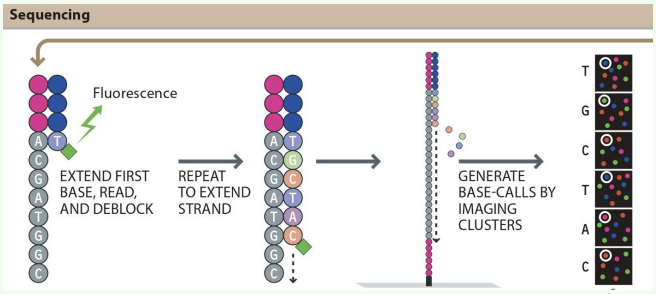
Sequencing methods have advanced substantially since the completion of the HGP. Next-generation sequencing (NGS) technologies are characterized by a PCR amplification phase followed by the sequencing of millions of DNA fragments (reads) in parallel[52]. The basic workflow of the most used NGS method, Illumina dye sequencing[53], is outlined in Box 1. NGS dramatically increased sequencing speed, which also led to a fast decrease of the cost per base[54]. Still, high-coverage NGS remains more expensive than genotyping arrays, which constitutes a problem particularly when large sample sizes are required (see *GWAS*). However, performing ultra-low (0.1-0.5×) or low-coverage (1×) WGS is now cheaper than chip-genotyping. Furthermore, ultra-low and low-coverage WGS are almost as powerful as genotyping arrays in capturing common and low-frequency (1-5%) SNPs[55,56].

On the other hand, NGS has the trade-off of producing smaller reads to work with, which makes their alignment more difficult[57]. They also present higher error rates than previous techniques[58]. However, these problems can be addressed by increasing the mean coverage of the sequenced regions, and the existence of a human reference genome helps to then align the multiple reads in the proper order. In addition, error modelling methods can improve the accuracy of genotype calling, thus reducing the number of false-positive variant calls[59].

Box 1. Illumina dye sequencing. DNA fragments are trimmed until their desired length is reached (200-800 bases), and adapters are added to the ends. The denatured strands are loaded into one of the lanes of a flow cell, which is covered with two types of oligos (adapter primers). The 3' adapters of the strands are hybridized with one type of oligo, the complementary strand is synthesized, and the template is discarded. Then, the synthesized free 3' adapter bends over and hybridizes with the second kind of oligo for another strand synthesis, and the resulting double-strand bridge is denatured. This process is repeated many times in parallel at millions of points in each lane, creating clusters consisting on clones (and complementary strands) of the original fragments. After this, the complementary strands are washed off, and sequencing by synthesis begins. A primer hybridizes with a specific region of the free adapter, and NTPs with a fluorescent dye-labeled base are added to the flow cell. The dye of the first nucleotide that is used for the synthesis is activated upon hybridization, and it blocks any further elongation. After the base is identified thanks to the cluster-specific fluorescence, the dye is cut off and the synthesis continues in the same manner until the desired read length (100-150 bases). The read product is washed off, and the complementary strand is generated and sequenced in the same manner. This double reading of the same fragment in both directions produces paired-end reads, which helps aligning them to the reference genome (adapted from Jobling et al. 2013)[13].







Single-molecule sequencing

New single-molecule sequencing technologies do not require PCR amplification and produce very long reads, which can resolve the alignment of ambiguous regions[20]. Long reads also provide the haplotype phase information – the group of alleles in the genome of an individual that are inherited from the same parent – without having to resort to imputation methods or family trios. Namely, nanopore sequencing[60] can produce reads of up to ~1Mb[20], while PacBio yields read lengths of over 60kb[61]. However, both technologies still produce error rates of up to 15%, which has to be refined before being able to be widely adopted by the research community.

Genetic variation from an evolutionary perspective

The observed patterns of genetic variation in humans are mainly generated by a combination of genetic drift and natural selection that operate differentially on mutually isolated populations. Isolation-by-distance[62,63] is the predominant isolation model in population genetic studies, although geographic and cultural barriers are also very common in human populations. Other contributions to the genetic landscape of a region are migration-related (founder effects, pulse admixture events, recurrent gene flow)[64] or due to changes in population size (bottlenecks and population expansion)[65].

The number of nucleotide differences between a typical genome and the human reference genome ranges between 4.1 and 5.0 million sites[23]. This range reflects world-wide differences, with African and African-admixed populations ranking on the rightmost side of the distribution: this is predicted by the "Out of Africa" (OoA) mechanism of dispersion, since the successive bottleneck processes undergone by non-Africans supposed a drastic reduction in diversity[66] (Figure 7). Although fixation of a mutation is possible, population-pairwise differences for biallelic SNPs usually lie in their MAF and haplotype frequencies.

In the following sections I present the two basic types of genetic variation, namely neutral and non-neutral variation.

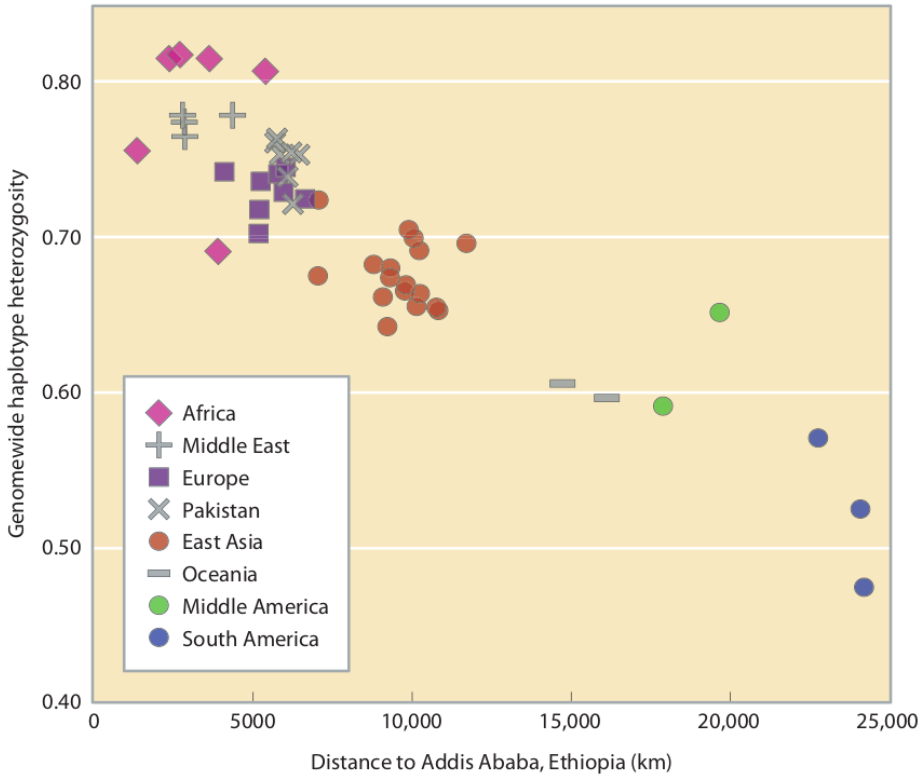


Figure 7. Decline of genetic diversity, calculated from haplotype heterozygosity, as as function of distance from East Africa. Each point represents a sample, shaped and coloured according to their world region ancestry (taken from Jobling et al., 2013)[13].

Neutral variation

The neutral theory of evolution[8] describes a model for the prediction of changes in population frequency of a new (*de novo*) mutation until its eventual loss or fixation, under the assumption that stochasticity, or genetic drift, is the only acting force. This, in turn, is conditioned by population size and allele frequencies. Most genetic variation in humans is considered to evolve under a neutral model, as it either alters the phenotype in a way that does not have a significant impact in the individual's fitness, or it takes place in overall non-functional intergenic regions

and therefore does not intervene in the phenotype. However, the latter assertion is contested by the results from the ENCODE consortium, a follow-up of the HGP, which suggest that up to 80% of the genome might have a function and that, therefore, most genomic regions are conserved to some degree[67]. It is worth noting that the results of this study have faced strong criticism[68,69].

The section *Population genetics* summarises the application of neutral markers for unravelling the demographic history of human groups, including migration events, population bottlenecks and expansions, isolation processes, and population splits.

Non-neutral variation

As a corollary of the above, the genetic component that is subjected to natural selection lies within a reduced fraction of the genome. This mainly consists of mutations whose effects at the protein level have an impact on the individual's fitness. These mutations can be *de novo* or previously neutral variants that, due to an environmental change, become either beneficial or detrimental.

Natural selection

A genetic variant is selected with a strength that is proportional to the magnitude of its effect on fitness, reflected in the selection coefficient *s*, as well as to the dominance model, reflected in the dominance coefficient *h* (Figure 8). Positive *s* values occur when fitness is improved, leading to positive selection. Meanwhile, negative *s* values indicate the opposite: purifying selection, also called background selection. In balancing selection, heterozygotes are selected against either homozygote, which can happen in the case of incomplete dominance ($1 > h < 0$).

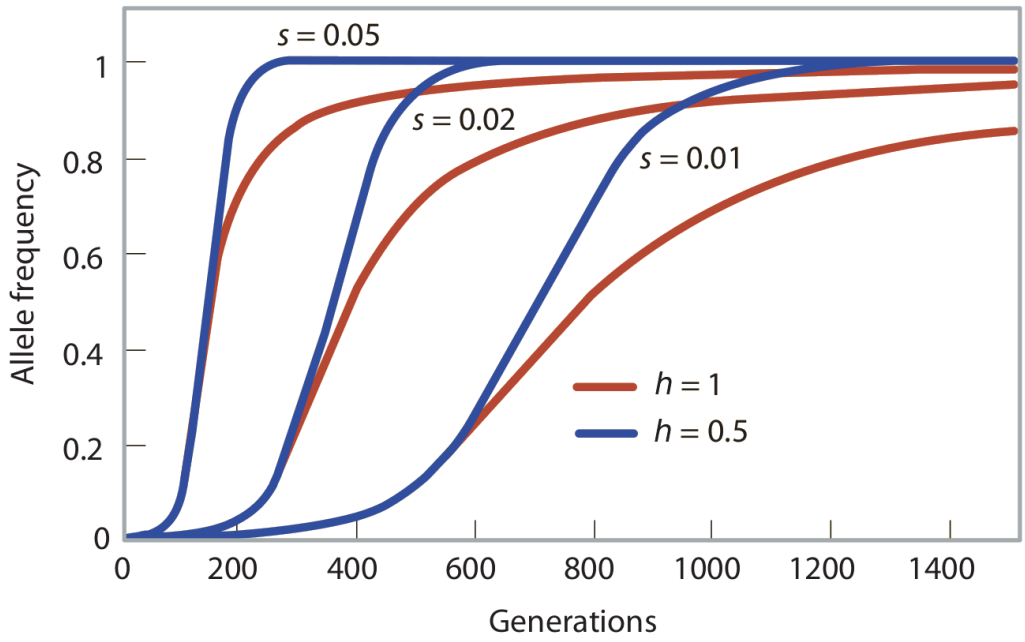


Figure 8. Prediction of changes in frequency of a beneficial allele (y axis) along multiple generations (x axis), depending on s *and* h*. As* s *increases, the time to an eventual fixation shortens. For a given value of* s*, fixation will be reached faster if the model is additive (*h *= 0.5) than if the beneficial allele is fully dominant (*h *= 1), and in this case the mutation might never reach fixation if the selection is not strong enough (*s *= 0.01) (adapted from Jobling et al., 2013)*[13].

Selection tests

Selection can be detected through the characteristic signatures that it leaves on the genomic regions involved. Both background selection and selective sweeps generate long linkage-disequilibrium (LD) blocks (Figure 9) that are present in unexpectedly high frequencies. As a consequence, there is a decrease in nucleotide diversity and an excess of rare variants in the affected genomic region[71]. Significant differences in allele frequencies between nearby populations can also be indicative of selection, although if no convincing mechanism is found, other demographic processes, like admixture or genetic drift, could be the actual cause.

Balancing selection, on the other hand, can be detected by methods such as Tajima's D. This statistic can be used to test whether the observed average number of pairwise nucleotide differences deviates from the expected value, given the number of existing segregating sites: if the statistic is significantly lower than expected, it suggests an excess of rare variants caused by a selective sweep or background selection; if it is significantly higher than expected, it is indicative of an elevated genetic diversity, possibly caused by balancing selection[72].

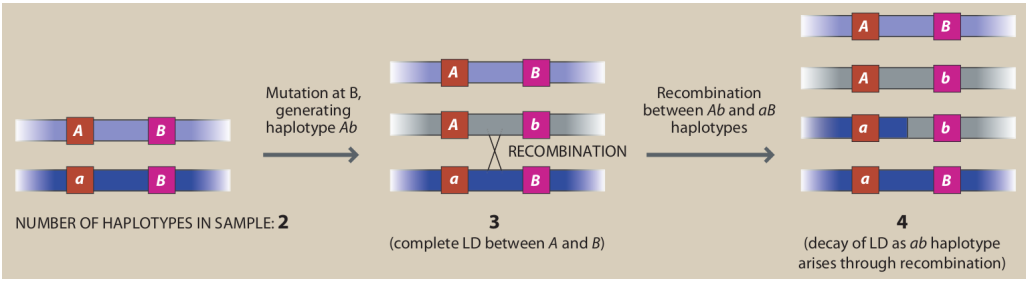


*Figure 9. Linkage disequilibrium (LD). Two biallelic polymorphisms Aa/Bb present at least some degree of LD in a population when the LD coefficient (*D*) – the difference between the observed haplotype frequencies (e.g.* p_{AB} *for haplotype AB) and the values predicted by the allele frequencies (*$p_A \times p_B$*) – is any value other than zero. Several measures derived from* D *exist, such as* D' *or* r^2*. Long haplotypes in LD are produced by a series of phenomena, such as mutation, selective sweeps, background selection, genetic drift, population bottlenecks, inbreeding, and admixture. Due to the chromosome recombinations that take place every generation, LD blocks experiment a decay in length as a function of time, which can be used to infer the date of the demographic events that caused them. Even in the absence of these processes, haplotypes that span less than 100kb usually present LD, since recombination rates (*c*) are lower between variants next to each other. Nonetheless, recombination hot-spots with higher values of* c *do also exist (taken from Jobling et al., 2013)*[13].

Adaptation in humans

Selective pressure drove a great part of human evolution since the split from the chimpanzee lineage, particularly regarding intelligence, speech, bipedalism, and tool use[73]. At the social

level, pair-bonding, biparental care, and modest sexual dimorphism are suggestive of selection for monogamous mating, although higher sexual dimorphism than other monogamous primates also suggests an adaptation for multi-level societies consisting of nested sets of modular family units[74]. Regarding anatomically modern humans, it appears that the majority of selective processes have concerned i)adaptation to different environments (involving, among others, adaptation to cold[75], changes in skin pigmentation as an adaptation to solar radiation[76], and adaptation to hypoxia related to high altitude[77] (Figure 10) or diving[78]); ii)adaptive defense against pathogens[79]; and iii)adaptation to different types of diet (like lactose consumption[80] or diets rich in polyunsaturated fatty acids)[81]. Finally, there has been reported selection of introgressed alleles from archaic humans in some populations for which these were advantageous[82].

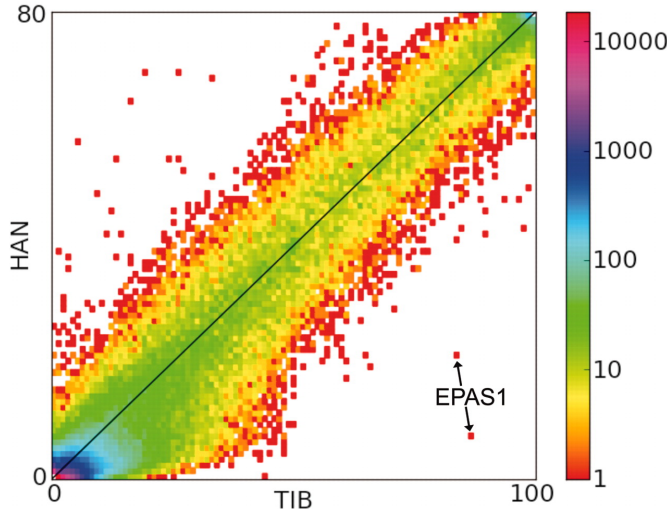


Figure 10. Two-dimensional unfolded SFS of Tibetan (TIB) and Han Chinese populations. The number of SNPs at each frequency coordinate is represented as a colour from the gradient on the right. A pair of intronic SNPs from the EPAS1 *gene show much higher derived allele frequencies in Tibetan vs. Han samples, suggesting adaptive selection to altitude (taken from Yi et al., 2010)*[83].

Applications of human genetic data

Population genetics

Generally speaking, population genetics makes use of neutral mutations for inferring the demographic history of the different human groups. Since neutral loci are not under selective pressure, they can be employed as a molecular clock by assuming a constant mutation rate (μ) and generation time (g). Following this rationale, an accumulation of differences in the form of private neutral alleles is proportional to the time of split of the populations. However, researchers have detected variation in μ for different human groups[84], possibly due to changes in g[85], environment, and population size[86], as well as between sexes due to differences in gametogenesis[87].

Various types of demographic processes produce additional genetic signatures that complement the information provided by the molecular clock. Namely, a great portion of the inter-population neutral variation is originated from differential genetic drift, N_e fluctuations, and gene flow from external groups. Along the following sections, I discuss how the different aspects of population genetic analysis of neutral variation provide valuable information about historical demographic events.

Genetic diversity

Genetic diversity is the degree of genetic variation, or amount of genetic differences, that exists among individuals of a population. The study of genetic diversity has been one of the first to be employed as a source of information about the demographic history of human populations, given that it does not require large amounts of markers nor complex calculations. For example, a simple way to measure genetic diversity is by calculating the nucleotide diversity (π), which corresponds to the average number of polymorphic sites between individual pairs.

Heterozygosity

Another popular measurement of genetic diversity is the heterozygosity (*H*). This can be defined as the probability that a random sample is heterozygous at a given locus, i.e. that two randomly drawn alleles differ. For a biallelic locus, let us define p and q as the population frequencies for the major and minor alleles respectively, so that $q = 1 - p$. The observed H (O[H]) is the frequency of samples that are found to be heterozygous for a given locus, while the expected heterozygosity (E[H]) across loci could be expressed as

$$1 - \frac{1}{m} \sum_{l=1}^{m} \sum_{i=1}^{k} p_i^2,$$

where m represents the number of loci and k the number of individuals in the population, assuming that all biallelic loci are in Hardy-Weinberg equilibrium ($p^2 + 2\,p\,q + q^2$).

Inbreeding coefficient

Genetic diversity is shaped by the history of the population, via diverse processes: mutation, recombination, balancing selection and gene flow increase it; while positive and background selection, as well as genetic drift linked to population bottlenecks and isolation, cause it to decrease. The inbreeding coefficient (*F*) is a way of measuring the lack of genetic diversity. For an individual, *F* can be defined as half the kinship coefficient between their parents (see *Relatedness and genetic distance*). For a population, *F* can be calculated, with relation to *H*, as

$$F = (\mathrm{E}[H] - \mathrm{O}[H]) \div \mathrm{E}[H],$$

where a positive value suggests inbreeding within the population; however, the Wahlund effect – excess of homozygotes due to population structure – could also apply. Another way of measuring the inbreeding is by determining the number and length of the runs of homozygosity (RoH) – overall homozygous genomic regions – within and across individuals.

Population mutation rate (θ)

Finally, the Watterson estimator (θ)[88] is a measurement of genetic diversity that represents the population mutation rate, defined as $\theta = 4\,N_e\,\mu$ for diploid individuals in an ideal population. θ can be estimated as

$$\widehat{\theta}_w = S \div \sum_{i=1}^{n-1} \frac{1}{i},$$

where *S* represents the total number of polymorphic sites and *n* is the number of haploid sequences. The Watterson estimator has many applications in population genetics, such as in coalescent theory (see *Coalescent-based demographic analyses*).

Relatedness and genetic distance

All humans are ultimately related, and the degree of neutral genetic divergence between individuals or populations reflects how many generations back their MRCAs lived. The kinship coefficient, or relatedness, numerically expresses the degree of relationship between a pair of individuals, computed in various ways. For instance, the study of family pedigrees considers the relatedness for each type of relationship to be the expected proportion of Identical by Descent (IBD) genetic material inherited from recent common ancestors (Figure 11), e.g. 1st degree relatives are 50% IBD, 2nd degree relatives are 25% IBD, and so on. However, this method has some limitations, such as the dependence on the availability of a pedigree for the study data, the assumption that the founders of the pedigree are unrelated, and the great variance of IBD values for a given degree of relationship[89]. Instead, it is now possible to estimate relatedness directly from genetic data. There are methods that estimate relatedness from IBD LD block length and density, under the assumption that Identical by State (IBS) alleles, i.e. those that are identical but not necessarily because of sharing a common ancestor, are most probably IBD.

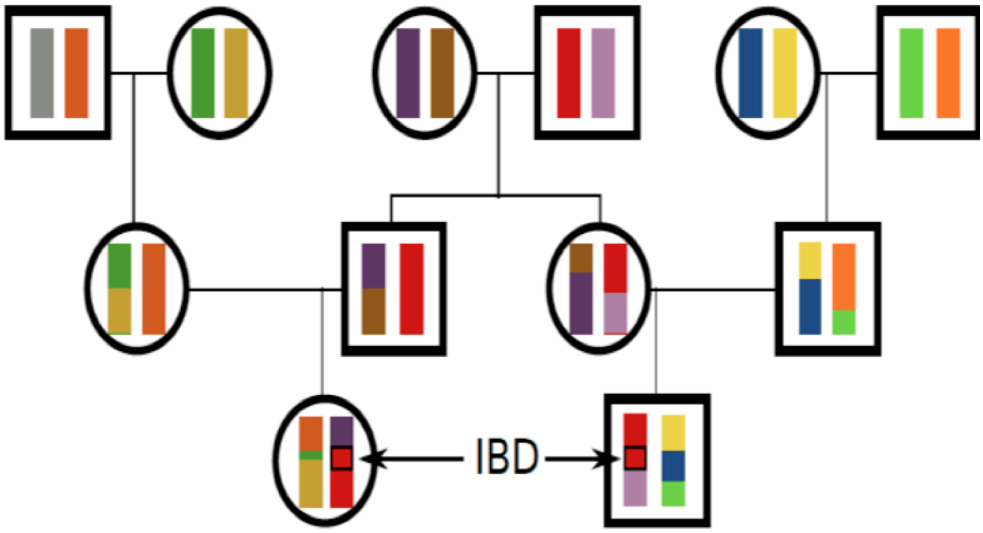


Figure 11. First cousins sharing an IBD genomic region (red colour) inherited from their common grandfather (taken from Coop, 2013)[70].

Pairwise relatedness, be it IBD or IBS, can be summarised in a genetic relationship matrix (GRM). This represents the averaged correlation coefficient between the genotypes of each pair of individuals. For the case of IBS, it can be calculated as

$$\frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)\,(x_{ik}-2p_i)}{2p_i\,(1-p_i)},$$

where p_i represents the MAF of the ith of a total of N SNPs, and x_{ij} and x_{ik} are the genotypes for SNP i (coded 0, 1 or 2) of individual j and k, respectively[90].

Genetic distance

Divergence between pairs of human populations can be assessed by calculating their genetic distances. There is a number of methods based on differences in allele frequencies, such as Nei's standard genetic distance (D)[91], which assumes that mutation and genetic drift are the only generators of population divergence. Thus, the formula

$$D = -\ln \frac{\sum_l \sum_u X_u Y_u}{\sqrt{\left(\sum_u X_u^2\right)\left(\sum_u Y_u^2\right)}}$$

yields the genetic distance between populations X and Y, where X_u represents the u^{th} allele frequency in X at the l^{th} locus. Another measure, Reynolds, Weir, and Cockerham's genetic distance (Θ_w)[92], only considers genetic drift. Pairwise genetic distances are commonly arranged in matrices when more than two populations are studied.

Phylogenetic trees

Phylogenetic trees are another useful way to visually represent the relationships between population groups. They consist in a succession of branches bifurcating from a node, which represents the MRCA of the groups stored in the branches. By including a known outgroup, one can root a tree in a way that reflects the temporal succession of nodes since the split from the outgroup. In the case of cladograms, populations belonging to a clade (i.e. a group of populations originating from a specific node) have a more recent MRCA than the MRCA that they share with a population from another clade (Figure 12).

There are several methods for calculating the composition of nodes and branch lengths of a phylogenetic tree. Neighbour-joining (NJ) uses a clustering algorithm to group samples according to the genetic distances between them[93]. Maximum parsimony methods return the tree that requires fewer evolutionary changes (mutations, nodes) while explaining the observed data. Increased computing power now allows to use Maximum likelihood (ML) and Bayesian approaches to produce more accurate phylogenetic trees.

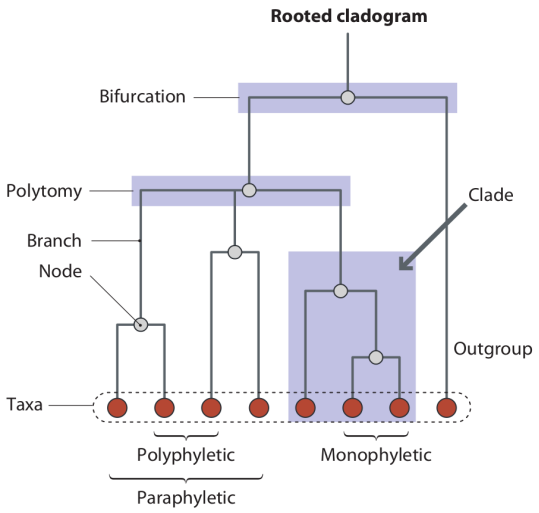


Figure 12. Scheme of a cladogram with standard terminology (taken from Jobling et al., 2013)[13].

Population structure

Population structure or stratification is an important concept in population genetics, and can be defined as the systematic differences in allele frequencies between populations due to geographic or other (e.g. cultural) factors. Geographic population structure can be explained by the isolation-by-distance model or the presence of physical barriers like rivers or mountains. As for stratification due to cultural differences, this can spring from historical inequalities between subgroups of different ancestry, language or religion barriers, among other things.

Population structure has to be taken into consideration when running demographic analyses. For instance, it can emulate the genetic signature of a bottleneck (see *Coalescent-based demographic analyses*), as well as cause the already mentioned Wahlund effect. Besides, it can act as a problematic confounding element in association studies.

AMOVA

Considering that population stratification is a structured form of genetic relatedness, it can be detected using tools like the ones previously mentioned, such as phylogenetic trees. Another way is through the analysis of molecular variance (AMOVA)[94], which makes use of Wright's *F*-statistics for determining the amount of variance (heterozygosity) in a meta-population that can be explained by differences between sub-populations. A typically used *F*-statistic is the fixation index (F_{ST}). It can be defined as

$$F_{ST} = (H_T - H_S) \div H_T,$$

where H_T is the E[*H*] in the meta-population and H_S the E[*H*] in the sub-populations. F_{ST} values range from 0 (no population structure) to 1, which would mean that all the variance observed in the meta-population is due to the sub-populations genotype composition being diametrically opposed (Figure 13). The statistical significance of F_{ST} can be estimated through permutation tests.

AMOVA can be performed for varying levels of complexity, taking as basis several sub-populations or, assuming a hierarchical island model, several groups of sub-populations. F_{ST} can be then further decomposed into F_{CT} and F_{SC}, which represent the part of the fixation index that corresponds to differences between different groups, and between sub-populations of the same group, respectively.

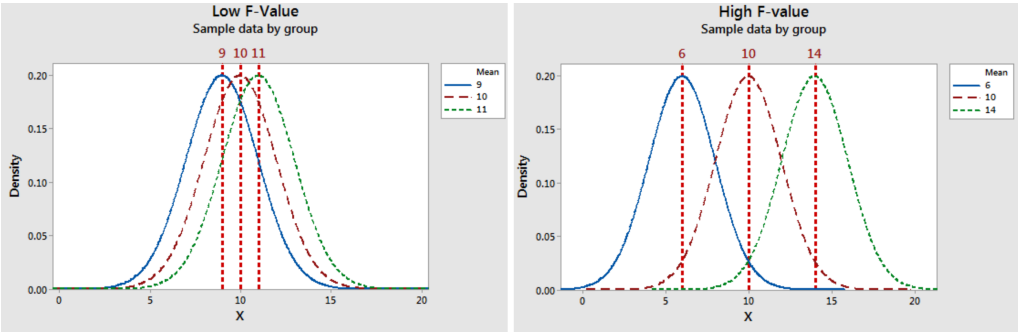


Figure 13. F-statistics can reveal population structure. The low F-*value plot (left) shows a meta-population formed by three sub-populations that present similar mean and variance for the variable X. Therefore, most of the meta-population variability is explained by the variability existent within each sub-population. In contrast, the three sub-populations in the high* F-*value plot (right) present similar variance but large differences in their mean values, thus raising the meta-population variability. A high enough* F-*value suggests population structure through the rejection of the null hypothesis, i.e. that the sub-population means are equal (taken from Coop, 2013)*[70].

Multivariate analyses

A classical approach for a better visualisation of the genetic relationships between samples, and for the detection of population structure, is by use of principal components analysis (PCA), first applied to human gene frequencies by Cavalli-Sforza and Edwards[95]. PCA consists in a reduction of the number of dimensions (e.g. genotypes from polymorphic loci) that define the variance stored in a GRM. This is achieved by the eigendecomposition of the GRM into a matrix of eigenvectors and their corresponding eigenvalues. When represented on a plane, the two eigenvectors that explain a larger fraction of the variance can potentially display the samples clustered according to their geographic relationships. Higher order eigenvectors can capture other more subtle relationships.

Multidimensional Scaling (MDS) is another dimension-reduction method for efficient graphical representation of the population distances contained in a genetic distance matrix.

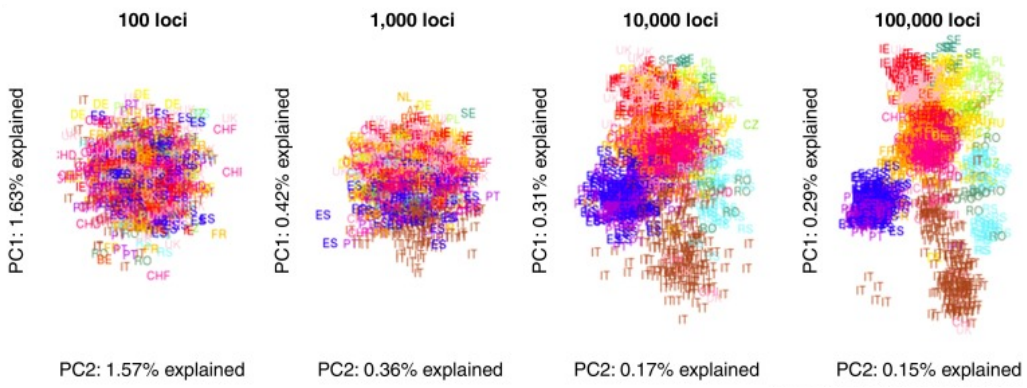
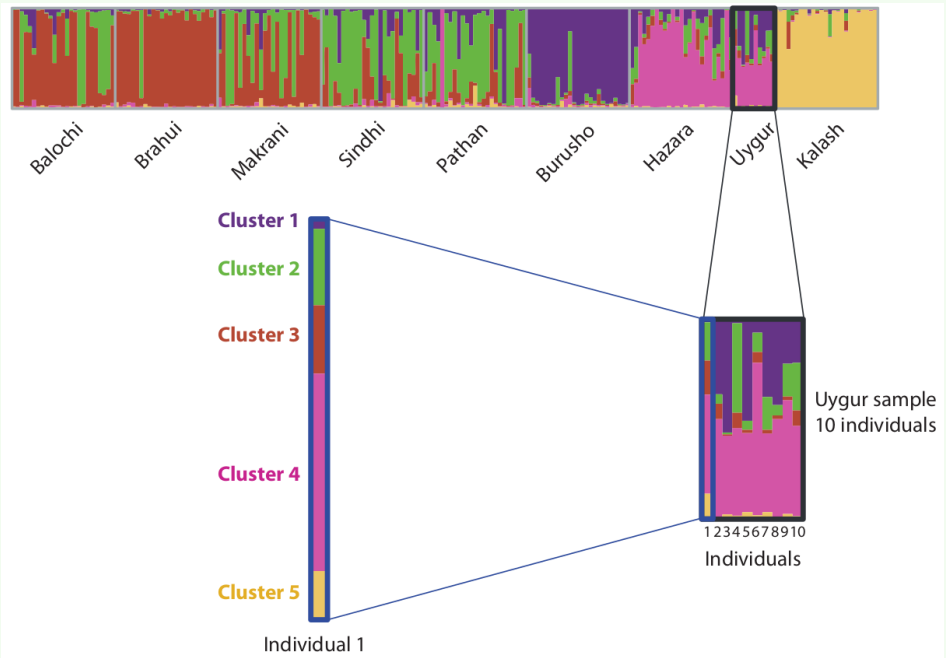


Figure 14. An increase in the number of loci employed in a PCA allows to reveal population structure in an otherwise apparently panmictic population. Each acronym represents the European country of origin of a sample. When using 100,000 loci, the sample arrangement starts to resemble the map of Europe, which indicates a high correlation between genetic and geographic distances in this continent (taken from Novembre et al., 2016)[96].

G*lobal ancestry*

Global ancestry (GA) is the average proportion of an individual's genome originating in a given, presumably ancestral, population. Model-based admixture analyses pre-specify a number *K* of genetic clusters and then estimate the probabilities that an individual belongs to each cluster, which can reveal patterns of population structure and the existence of admixture events[97] (Box 2). STRUCTURE[98] was the first widely-used software to implement model-based admixture analysis. By means of a Markov chain Monte Carlo (MCMC) procedure, STRUCTURE performs a repeated sampling from the prior probability distributions of the parameters of interest (i.e. allele frequencies in each cluster, clusters of origin of each individual, and admixture proportions), with the aim of maximizing the posterior probability of these parameters given the observed genotypes (see *Approximate Bayesian computation* for more on Bayesian inference). ADMIXTURE is another popular program with the same purpose that calculates the ML estimates of the parameters of interest, increasing computation speed without a decrease in accuracy. To determine the most likely value of *K* attending to variation in the data, ADMIXTURE uses a cross-validation approach, while STRUCTURE computes the model evidence for each value of *K*[99].



*Box 2. Example of barplot produced in model-based admixture analysis. The top barplot represents 210 individuals from nine populations from Pakistan, assuming five genetic clusters (*K *= 5) represented by different colours. Each individual is shown as a vertical bar, in which the probability that the individual belongs to a given cluster is proportional to the length of the corresponding coloured segment. For example, the highlighted Uygur sample has a higher probability of being a member of cluster 4 than of any other cluster.*

Kalash, Brahui and Burusho appear as full representatives of three out of the five clusters. The presence of populations whose individuals have more than one significantly long segment (e.g. Uygur) could be attributed to i)a lack of discrimination power, ii)an isolation-by-distance cline[100], or iii)admixture events, in which case the segment lengths can be interpreted as the individual's genomic fractions inherited from each cluster (taken from Jobling et al., 2013)[13].

Local ancestry

Local ancestry (LA) estimation specifies the ancestry of each genomic region within an individual[97]. Original approaches to LA estimation (e.g. LAMP[101]) make use of a hidden

Markov model (HMM), or its extensions, to fit an explicit probabilistic model for the observed variables (alleles) and unobserved variables (ancestry) to the data. Because of this, these programs rely on large reference panels that are good proxies for the true ancestries of the admixed samples[102]. In contrast, RFMix is a recent software that does not require large reference panels and outperforms previous LA estimation methods, by using a machine learning algorithm trained on phased reference panels[102]. While classical analyses usually assume and/or require independence between loci, methods like RFMix take advantage of the more informative haplotype data, which allows to detect population structure even at a very fine scale.

Besides allowing the inference of ancestry proportions, the information about number and lengths of the haplotypes originating in each ancestral population can be used to determine the relative timing of successive admixture events[103,104]. In addition, LA inference has many other applications in clinical and evolutionary genetics, such as mapping genes to disease, mapping sequences of unknown location onto the human reference genome, studying recombination rate variation, and inferring natural selection[97].

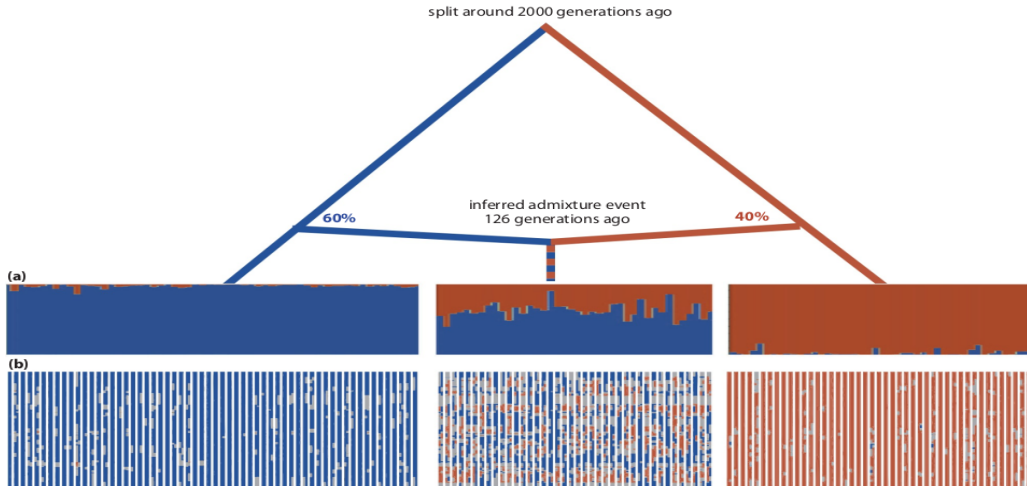


Figure 15. Comparison of (a) GA and (b) LA estimation of admixture proportions. GA estimation produces average ancestry proportions across an individual's genome, while LA estimation specifies how that ancestry is distributed along the genome. Gray colour represents ambiguous ancestry (taken from Jobling et al., 2013)[13].

Chromosome painting

Another novel ancestry inference method that is based on haplotype information is chromosome painting, implemented in Chromopainter[105]. Here, each *recipient* individual is reconstructed with the most likely combination of haplotype segments shared with *donor* individuals. This information can be organised in two types of relationship matrices – one containing the pairwise haplotype "chunk" counts and one containing the total haplotype chunk lengths, respectively. The chunk counts can be employed by the companion software fineSTRUCTURE[105] to infer a phylogenetic tree that represents the relationships between samples. fineSTRUCTURE works by first obtaining the maximum a posteriori probability (MAP) state through MCMC iterations, then improving its posterior probability with additional hill-climbing, and finally merging the populations successively, choosing the highest probability merging at each step. In addition, one can infer the *donor* ancestry proportions in each *recipient* by analysing instead the matrix of total haplotype lengths with GLOBETROTTER[64].

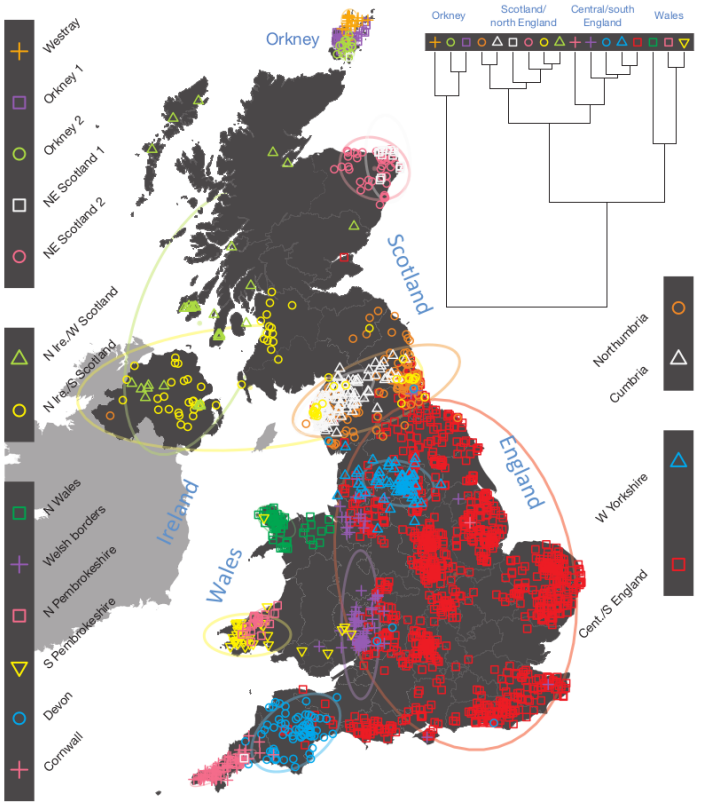


Figure 16. Phylogenetic tree created with fineSTRUCTURE using genome-wide array data from rural British populations, and the geographic location of the samples labelled according to their assigned cluster. Part of this population structure can be explained by historical events, such as migrations (taken from Leslie et al., 2015)[106].

f-statistics

f-statistics comprise a group of tests for evaluating admixture and gene flow by fitting phylogenetic tree models to allele frequency correlations between populations. Using the tree in Figure 17 as an example of a phylogeny, we can define the *F*-values as follows:

$F_2(A, B) = E[(a' - b')^2]$,

$F_3(C; A, B) = E[(c' - a')\ (c' - b')]$,

and

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

where a', b', c' and d' represent the allele frequency for a given biallelic SNP in the population *A*, *B*, *C* and *D*, respectively. We can then obtain the *f*-statistic for each *F*-value by averaging the latter across all markers.

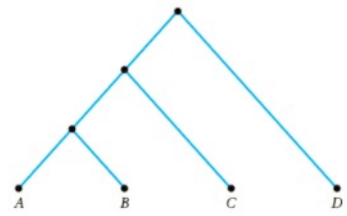


Figure 17. Example of a phylogeny with three populations A, B *and* C, *and an outgroup* D.

Intuitively, $F_2(A, B)$ represents the branch length between *A* and *B* for a randomly drifting neutral allele[107]. Meanwhile, the three-population test uses the f_3 statistic to formally test whether a population *C* is the result of an admixture event between two ancestral populations for which *A* and *B* act as proxies: if *C* is the result of an admixture event between *A* and *B*, f_3 will be significantly negative. However, if *B* has undergone strong genetic drift after the admixture event, due to a bottleneck or founder event, for example, this could effectively mask the admixture event[50]. An example of a software employing this method is TreeMix[107].

D-statistics

D-statistics[108], or *four-population test*, are based on the f_4 statistic, and test whether the gene genealogies of the populations *A*, *B*, *C* and *D* follow a pattern BABA (*A* similar to *C* and *B* similar to *D*) or ABBA (*A* similar to *D* and *B* similar to *C*). For example, assuming that a pair of European (*A*) and Asian (*B*) populations are phylogenetically closer between them than to a sub-Saharan population (*C*), one could be interested in the possibility of genetic introgression from *C* to either *A* (BABA true) or *B* (ABBA true). *D* is an external group used for determining

the ancestral allele, e.g. *Pan troglodytes* (chimpanzee) is a common choice. The *D*-statistic is then calculated as

$$D = \Sigma_i^m(a'_i - b'_i)\,(c'_i - d'_i) \div \Sigma_i^m(a'_i + b'_i - 2\,a'_i\,b'_i)\,(c'_i + d'_i - 2\,c'_i\,d'_i),$$

where *i* represents a locus out of *m* loci. Note that if a' and b' are similar, $D = 0$. If *D* is significantly different from zero, we can accept introgression from *C* to either *A* (*D* positive) or *B* (*D* negative)[109].

Incomplete lineage sorting (ILS) is a special case of BABA/ABBA pattern caused by the population and the genetic trees not matching, resulting in a genetic signature that mimics genetic introgression[50] (Figure 18). However, since shared variation due to drift and ILS follows a random distribution, a significant departure from the expected amount of shared alleles may be indicative of gene flow[109]. Furthermore, in the case of relatively recent admixture events, introgressed haplotypes tend to be longer than those caused by ILS[110].

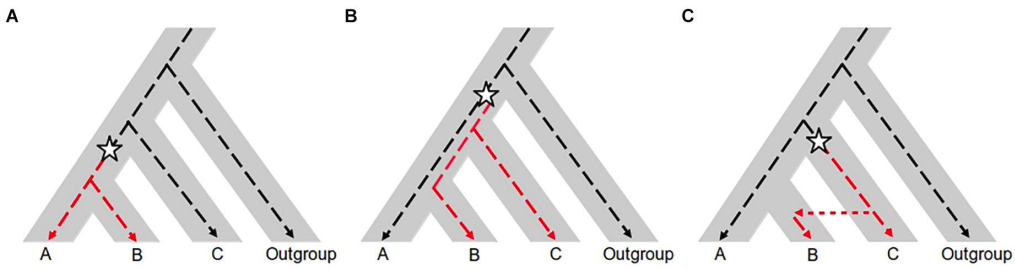


Figure 18. Effect of introgression and ILS in the disparity between population and gene trees. The population tree is represented by the gray area. The dotted line represents the genealogy of a single locus, where a star represents a mutation event. If the D*-statistic calculated for this locus is not significantly different from 0, it means that (A) the population tree matches the gene tree, and no introgression has taken place. If, instead, a significant ABBA pattern is found, i.e. populations B and C share the same private allele, there are two possible demographic scenarios that could have originated it, namely (B) ILS or (C) introgression from C to B (adapted from Burgarella et al., 2019)[109].*

Admixture dating

Once an admixture event has been identified, a number of methods allow to infer the date at which it took place, such as the ALDER software[111]. The principle is that, although LD blocks are present in all human populations, those populations that are not recently admixed do not present LD blocks of more than a few hundred kilobases long[112], but if an admixture event took place relatively few generations ago, this will have generated longer admixture LD (ALD) blocks. However, in each generation, genetic recombination breaks down these blocks, so the more generations from the admixture event, the weaker ALD will exist between distant variants. Therefore, the rate of ALD decay as a function of genetic distance between variants (d) indicates the number of generations ago that the admixture event took place (n). Formally, this can be represented in the formula

$$D_{xy} = \alpha \beta \delta_x \delta_y e^{-nd},$$

where D is the haploid ALD measured as the covariance of alleles at biallelic SNPs x and y, α and β are the proportional genetic contributions from the parent populations, and δ is the difference in allele frequencies between the parent populations. One caveat of this method is that other demographic events, like recent bottlenecks or extended periods of low population size, produce long-range LD as well, which could lead to spurious admixture dates[111].

GLOBETROTTER infers admixture dates in a similar way, using the matrices of haplotype lengths generated by Chromopainter as an input. GLOBETROTTER can also test whether one or more events of admixture took place and date each of them, as well as infer the contributing proportions of the donor populations.

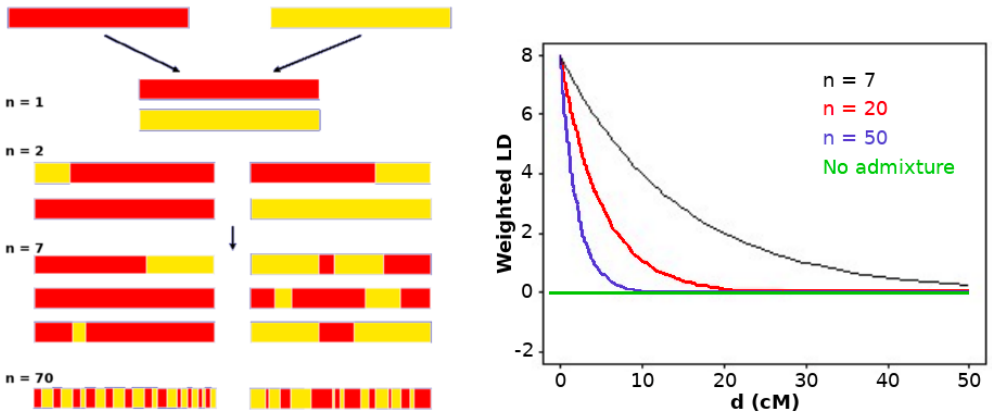


Figure 19. Dating of admixture events based on the decay rates of ALD blocks. The figure at the left depicts how the ALD haplotypes resulting from an admixture event (n = 1) are broken down along generations due to recombination. The plot at the right shows how the weighted ALD decay as a function of d *reflects the number of generations ago that the admixture event took place. cM: centimorgans (adapted from Myers)*[113].

Coalescent-based demographic analyses

The coalescent theory, originally developed by Kingman[9], has numerous applications in population genetics. In brief, it models the probability that the MRCA of *n* homologous haplotypes existed *t* generations ago. In the next sections, I provide an introduction to the basics of the coalescent theory, as well as its application in demographic inference through the sequentially Markovian coalescent methodology and the genetic data simulators.

Coalescent theory

Given $n = 2$ haploid individuals that present the same allele at a given locus, the probability p that this duplicated allele comes from a single chromosome in the previous generation is $p = 1 \div N_e$ and, in the case of two heterozygous diploid individuals, $p = 1 \div (2\ N_e)$. We can then calculate the probability of coalescence in function of time (t, number of generations ago) as $P_c(t) = (1 - p)^{t-1} p$.

If N_e is large enough, the probability mass function approximates to an exponential distribution of the form

$P_c(t) = p\ e^{-t \div (2\ Ne)}$,

with mean

$E[t] = 1 \div p = 2\ N_e$

and variance

$Var[t] = (E[t])^2$.

This means that the coalescence event is expected to happen $t = 2\ N_e$ generations back. If instead we were to calculate the time of coalescence of an allele present in $n > 2$ heterozygous individuals, the probability that at least two of them coalesce in the previous generation is then $p = \binom{n}{2} \div (2\ N_e)$,

and therefore

$E[t] = (2\ N_e) \div \binom{n}{2} = (4\ N_e) \div (n\ (n-1))$.

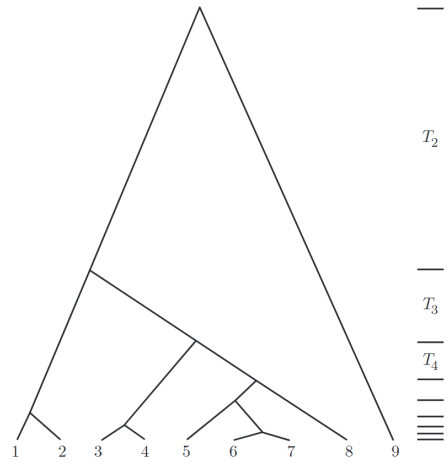


Figure 20. Example of a coalescence tree involving nine samples (taken from Didelot)[114].

From this we can expect that the larger n and the smaller N_e, the lesser generations towards the past until a coalescence event takes place are to be expected. Furthermore, taking as reference a specific coalescence event, we can expect the time to the previous one to take increasingly longer time spans. For example, in Figure 20, $T_9 = (4\ N_e) \div (9\ (9-1)) = N_e \div 72$ generations back, but $T_8 = N_e \div 56$ generations from T_9, and so on, until the final coalescence $T_2 = 2\ N_e$ generations back from T_3. A consequence of this is that there will be more neutral mutations in the deepest branches, since they last more generations. Finally, the total time until all coalescences have happened can be calculated as

$T = \Sigma_i^n (4\ N_e) \div (i\ (i-1)) = (4\ N_e)\ (1 - (1 \div n)) \approx 4\ N_e$

if n is large enough. When mitochondrial DNA or the non-recombinant Y chromosome coalescence times are studied, $T = N_e$ because of, on one hand, being haploid sequences, and on the other, having a pattern of inheritance that involves only half of the population, i.e. either males or females.

Sequentially Markovian coalescent

The sequentially Markovian coalescent (SMC)[115] is a popular seminal model for demographic inference that applies coalescent theory. It takes into account the population recombination rate $\rho = 4\ N_e\ r$, where r is the per generation recombination rate. This Markovian approach to the recombination has the advantage that ρ at a given t only depends on N_e and r, and as a result it is much less computationally expensive than non-Markovian methods[116] in which the probability of recombination at a given t depends on the previous generations[115].

Pairwise SMC (PSMC)[117] infers historical changes in N_e in a population by generating a distribution of the times of coalescence for multiple heterozygous loci of a single phased individual. Intuitively, since N_e is directly related to the number of generations until coalescence, one can infer changes in N_e through time. For example, if many loci coalesce within the same period, this could be indicating a bottleneck at that time. However, there are some caveats to this, like sensitivity to the chosen value of μ, lack of accuracy for times more recent than 10 KYA, and the confounding effect of population structure: if the two alleles at a heterozygous locus come from different demes, for them to coalesce there has to pass enough backward-time to revert the migration event that brought one of the alleles from its original deme, and then more time for the merging of both demes that will allow for the coalescence. Depending on the migration rates, the number of demes, and their respective N_e, this could take a much longer time than for local variants, so the result would be many of the local variants coalescing relatively soon, and a few introgressed variants coalescing much later, effectively emulating a population bottleneck in the meantime[118].

Multiple SMC (MSMC)[119] extends the method to two or more individuals, which, if coming from different populations, allows to infer the time of population split. SMC++[120] employs a different algorithm that makes feasible the use of unphased genomes, as well as the analysis of sample sizes that are orders of magnitude higher than MSMC, which results in better accuracy for the inference of recent events.

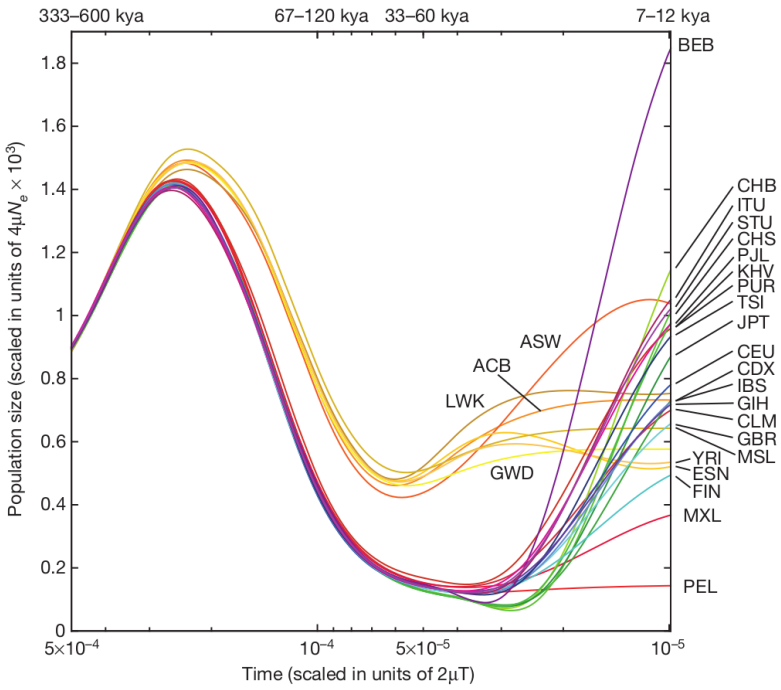


Figure 21. PSMC analysis of 1000G samples, showing N_e *changes (y axis) along time (x axis). The separation of two graphs is interpreted as population divergence. Some points to remark here are the early split of sub-Saharan populations, and the recent general population growth. A description of the population codes is provided in the Appendix (I) (taken from The 1000 Genomes Project Consortium et al., 2015)*[23].

Genetic data simulators

The methods of genetic data analysis introduced so far provide information about demography, such as degree of inbreeding, population structure, or the time of an admixture or split event. However, a caveat of these methods is that, in many occasions, the results could be *a priori* interpreted in several mutually exclusive ways[121]. As an alternative, there are algorithms that generate simulated genetic data based on a specified demographic history, i.e. considering the time and magnitude of events such as N_e changes, population splits, and gene flow. The resulting summary statistics can be then compared with real data, therefore allowing to test the likelihood of different demographic parameters or models.

One approach is the forward-time reconstruction of the population history, which consists in initializing an ancestral population and then simulating the specified demographic changes and their consequences in the genetic pool of the population[122]. An example of a software that makes use of this is simuPOP[123]. However, forward-time simulations have several important limitations, such as the difficulty in accurately determining the ancestral population, and in being computationally expensive since they track the history of all lineages, including those not leading to the samples of interest[124]. For this reason, the most employed methods for simulating the demographic history of a population are based on the coalescent, which is a backward-time approach.

The typical implementation of a coalescent-based simulator starts with a sample with some specified characteristics (sample size, number of considered loci, etc.) whose genotypes are not yet determined. Then, recreating the specified demographic history (Figure 22) backward-time, the algorithm successively merges each independent locus – or haplotype, instead – into the common ancestors, until the MRCA is reached. Genetic information is assigned to the MRCA, and the process runs forward-time. The distribution of genetic variation of the resulting sample is determined by the probability of a random mutation in each generation, which creates new polymorphisms (recurrent mutations and backmutations are not allowed), and by the random distribution of alleles in the offspring.

In its most basic form, the algorithm assumes constant N_e, non-overlapping generations, no gene flow, no recombination, no population structure, no selection, random mating, and the probability of more than one event of coalescence in the same generation is ignored. Nevertheless, modern software has seen an increase in speed and model complexity, adding features such as changes in N_e, variable chromosome recombination rates along sequences, gene conversion, migration, population structure, sampling at multiple time points, and single-locus selection, thus allowing for more realistic scenarios[125].

Overall, genetic data simulators have many applications, since they can simulate data under a wide variety of scenarios, such as haplotypes with pre-specified allele frequencies and LD structure, GWAS samples, and pedigrees[126]. To finish this section, I describe the application of simulated genetic data in approximate Bayesian computation and deep learning.

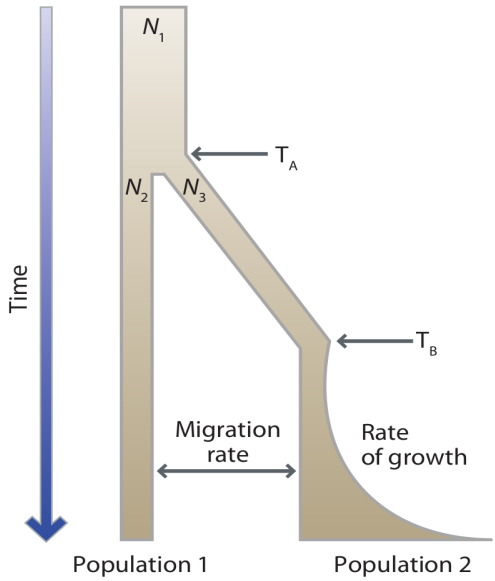


Figure 22. Example of a demographic model. The defining parameters are the number of sampled populations (Population 1 and Population 2), the time of population split (T_A), the effective population sizes of the ancestral population (N_1) and of the child populations at T_A (N_2 and N_3), the time at which the population expansion of Population 2 began (T_B), the rate of growth of this expansion, and the migration rate between Population 1 and Population 2 (taken from Jobling et al., 2013)[13].

Approximate Bayesian computation

Approximate Bayesian computation (ABC) makes use of simulated population data to infer the demographic history that better fits the observed data. The rationale of ABC can be better understood by first introducing Bayes' theorem, which is commonly expressed as

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

where, in the context of demographic studies, θ is a demographic model parameter and D is the observed population data. Applying Bayes' theorem, one can calculate the posterior probability $p(\theta_i \mid D)$, where θ_i is a value of θ sampled from a prior probability distribution with probability $p(\theta_i)$. If one is just interested in which values of θ yield the *relative* maximum posterior probabilities, the constant $p(D)$ can be removed from the equation. However, $p(D \mid \theta)$ still has to be calculated for each sampled θ_i, which can result computationally infeasible in some scenarios[127]. ABC avoids the direct calculation of the likelihood by identifying which values of θ produce simulated data that is closest to D, as the frequency distribution of these values approximates the posterior probability distribution of θ[127]. The basic procedure is detailed in Figure 23.

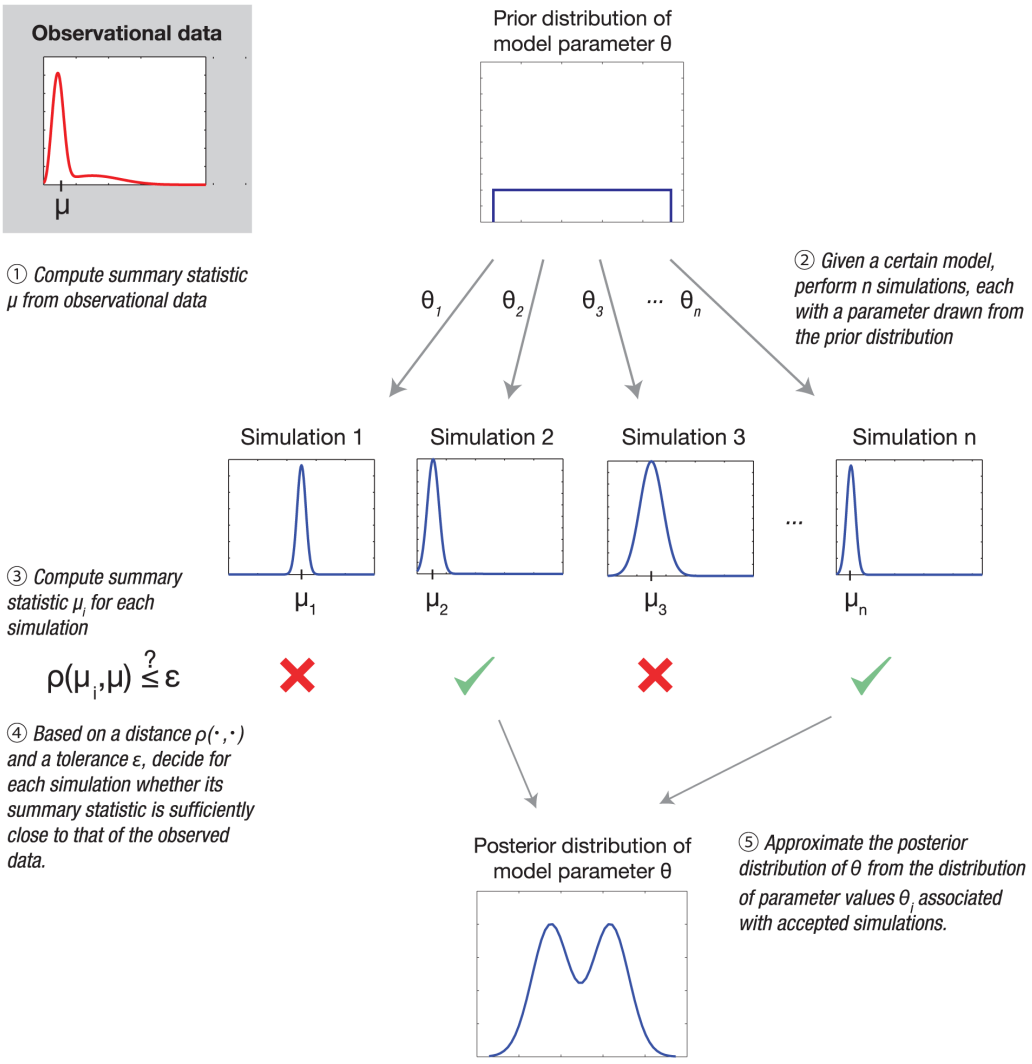
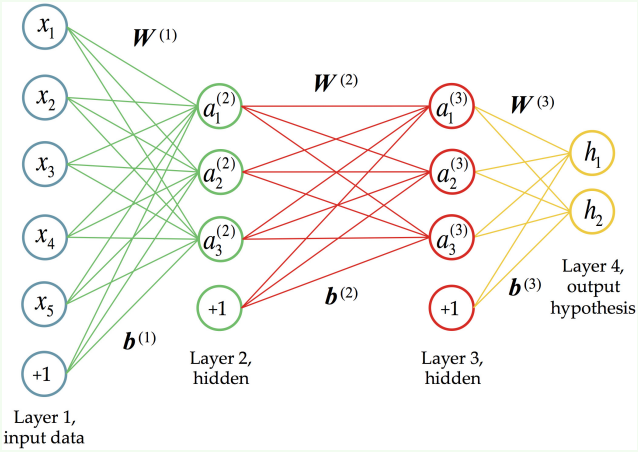


Figure 23. ABC overview. Multiple simulations are run under a specific model, randomly sampling the values for the parameter of interest θ from a prior distribution (e.g. a uniform distribution, when no further assumptions are taken). The generated genetic information is stored in the form of summary statistics (μ_i in the example). Finally, the posterior distribution of θ is approximated with those values whose summary statistics were the closest to that of the observed real data (taken from Sunnåker et al., 2013)[128].

Deep learning

Summary statistics of simulated genetic data, such as the joint multi-population SFS[129], can be used to train artificial neural networks – a machine learning framework – to categorize real genetic data. Usually, artificial neural networks of more than two layers are used, a field known as deep learning (see Box 3 for an explanation on the basic structure of a deep neural network). Currently, deep learning approaches are being incorporated to the study of human demographic history[130,131].



Box 3. Basic architecture of a deep neural network (DNN). A DNN is structured in several layers, where the first layer is the input data, and the inner and output layers are formed by neurons. A neuron is a mathematical function that transforms the weighted input from all the neurons in the previous layer into some value by using an activation function. For example, a sigmoid activation function will output a value within the range {0, 1}.

Formally, the $a_1^{(2)}$ *neuron can be expressed as*

$$\mathrm{sig}\left(\left(\Sigma_{i=1}^{5} w_i^{(1)} x_i \right) + b^{(1)} \right),$$

where sig *is the sigmoid activation function,* w_i *is the weight for the input value* x_i *at neuron* $a_1^{(2)}$*, and* b^1 *is the bias neuron for* $a_1^{(2)}$*, which shifts the activation function to either side. In the figure,* W^i *denotes the matrix that contains the weight for each each connection (coloured lines) between the* i^{th} *layer and the next. The output of a neuron is in turn transferred to the neurons in the next layer, where the same procedure applies, until the output layer produces the prediction* h_i*. This prediction can be, for example, the probability that the input data belongs to the* h_i *category. The error in the prediction is calculated through a loss function, e.g. if it is a supervised DNN, the squared difference between the known real value and the predicted value can be used. By iteratively feeding a DNN the output of multiple simulations, it can update the weights to reduce the error of the predicted parameter values. Once the error is low enough, one can test the accuracy using additional simulations (taken from Sheehan and Song, 2010)*[132].

Clinical genetics

Clinical genetics studies those mutations whose effects contribute to the development of a disease. The first kind of genetic disease studied was monogenic (or Mendelian) disease. A Mendelian disease, like sickle cell anaemia[133], is determined exclusively by a single specific polymorphic locus. Meanwhile, complex phenotypes, which include most common diseases as well as quantitative traits like height and weight, are controlled by multiple genetic markers (thus the term polygenic) interacting with environmental factors[134].

Association study design

In the context of clinical genetics, association studies are designed to detect which genetic polymorphisms are linked to a given disease, by testing for correlation between phenotype and genetic variation[135]. A popular family-based approach is the transmission disequilibrium test (TDT), which consists in testing the over-transmission of each allele of interest across trios of two unaffected heterozygous parents and one affected homozygous child. Population-based studies include the case-control study, which consists in testing for genetic differences between subjects affected by a disease (cases) and unaffected controls, and the prospective cohort study, in which only unaffected subjects are initially sampled; after a period of time, genetic differences between samples that developed the disease and the samples that remain unaffected are tested[136].

Regression analysis

The choice of regression model to test for association between a locus and a disease will depend on the categorical or quantitative nature of the studied phenotype. When the phenotype is quantitative, e.g. blood pressure, a linear regression model is employed. It follows the form

$f(x) = a_0 + \hat{\beta} x + \epsilon$,

where $f(x)$ is the phenotype as a function of genotype, a_0 is the intercept, $\hat{\beta}$ is an estimate of an unbiased regression coefficient β, and ϵ is the statistical noise. When significantly different from zero, the regression coefficient can be used for making predictions of the outcome of an individual carrying a specific genotype.

If, instead, the phenotype is categorical, as in a case-control study, the logistic regression has to be used instead, expressed as

$f(x) = 1 \div (1 + e^{-\hat{\beta} x + \epsilon})$.

The regression coefficient $\hat{\beta}$ can be thought of as the effect size of a genetic variant, i.e. the relation between genotype and phenotype, in the way that a change of $x = 1$ genotype units implies a change of $\hat{\beta}$ phenotype units. In addition, $\hat{\beta}$ corresponds to the correlation coefficient (r) if the variables are standardized. It can be calculated by ordinary least squares (OLS), which consists on finding the values a_0 and $\hat{\beta}$ that minimize the summation of residuals, expressed as

$\Sigma^{n}_{i} (f(x)_i - f(x))^2$,

where i represents a sample from a total of n. In order to test whether $\hat{\beta}$ is significantly different from zero, the Wald test can be applied, which consists in dividing $\hat{\beta}$ by its SE; the resulting statistic belongs to either a t-distribution, if it is a linear regression, or to a Z-distribution, if logistic.

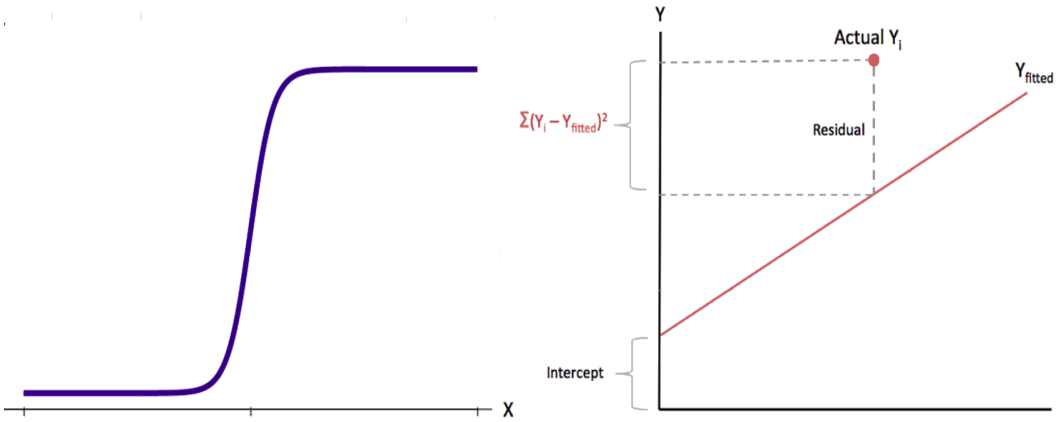


Figure 24. Visual comparison of logistic (left) and linear (right) models. The linear model plot further highlights the intercept and the residual for a data point (adapted from Chávez, 2019, and Suvarna, 2019)[137,138].

Odds-ratio

Another way to express the effect size is in the form of an odds-ratio (OR), which consists in the proportion of the number of samples that reject the null hypothesis vs. those that do not reject it (Table 1). In other terms, the OR indicates how likely it is for a carrier of the studied

allele to present the disease compared to the average population, assuming the carrier is overall average for the remaining causal variants. It ranges from 0 to infinite, where no significant deviation from 1 means no association, values significantly larger than 1 suggest harmful effects of the studied allele, and vice versa.

		Disease status	
		Cases	Controls
Genotype	aa	*a*	*b*
	AA|Aa	*c*	*d*

(1) $OR = (a \times d) \div (b \times c)$

(2) $SE = \sqrt{1 \div a + 1 \div b + 1 \div c + 1 \div d}$

(3) $\ln(OR) = \hat{\beta}$

*Table 1. Example of a 2×2 contingency table including the number of cases (*a *and* c*) and controls (*b *and* d*) for a given recessive disease, which have been genotyped for a SNP with a dominant allele A and a recessive derived allele a. On the right side, formulae for the calculation of the OR (1) and its correspondent standard error (*SE*) (2), as well as the relation between the OR and* $\hat{\beta}$ *(3), are provided.*

GWAS

To determine the genetic variants to be tested for association, one can follow either a hypothesis-driven or a hypothesis-free approach. The former corresponds to a candidate gene study, in which a limited number of genes or genetic markers are hypothesized to have an effect on the phenotype because of their involvement in related biological pathways.

However, this approach is prone to problems when applied to polygenic disease: in this case, genetic risk factors are diffusely distributed across the genome, and variants with a stronger effect are more likely to be located in poorly understood regions of the genome than in candidate genes[139]. For this reason, and given the availability of massive amounts of genetic information from across the genome, a more efficient study of polygenic disease can be carried out through genome-wide association studies (GWAS). A GWAS is a hypothesis-free assay of

the genetic component of a complex trait. GWAS employ large numbers of markers, generally SNPs, distributed genome-wide, which are tested for association with the trait in question. This approach relies on the basis that not only genes, but also functional intergenic regions, can present mutations involved in a complex disease. To compensate for the rise of false-negatives resulting from multiple testing correction, GWAS require large sample sizes to increase statistical power[140].

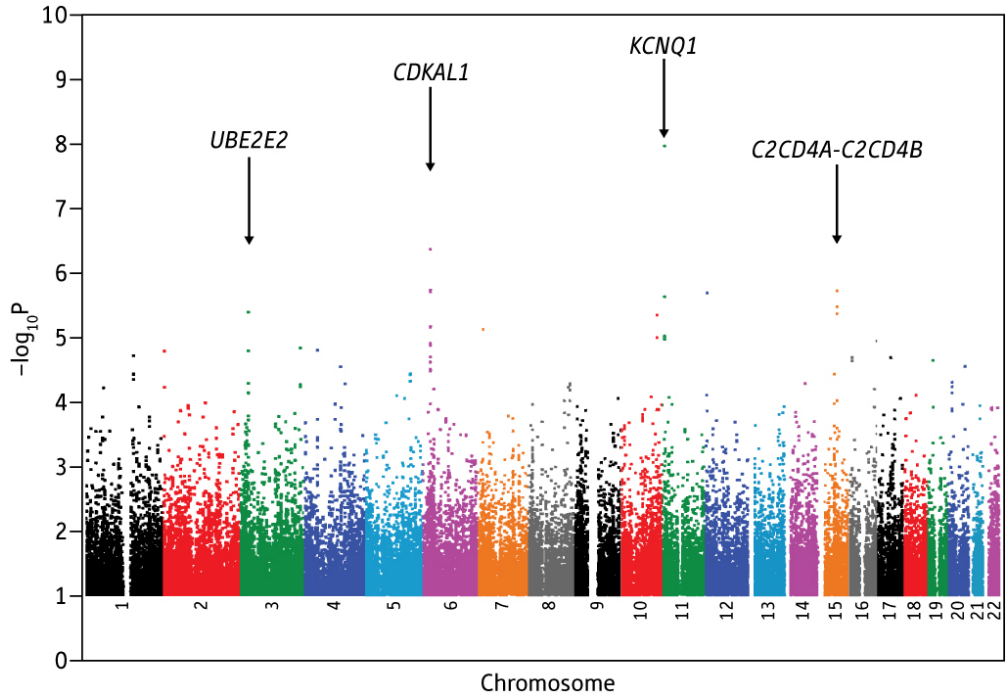


Figure 25. Example of a Manhattan plot, a common way of representing significant variants in GWAS. The x *axis is formed by the studied SNPs arranged by physical location. The* y *axis represents the* $-\log_{10}$ *of the association test p-value, so that strong signals of association are highlighted. In the example, SNPs located in the* KCNQ1, CDKAL1, UBE2E2 *and* C2CD4A-C2CD4B *genes are identified as susceptibility loci for type 2 diabetes (taken from Yamauchi et al., 2010)*[141].

Genetic risk

Genetic risk calculation allows to predict whether an individual will develop a given complex disease, based on their genotype information for variants that have been found associated with the disease. Genetic risk score (GRS), or polygenic score, is a common way to assess the genetic risk of an individual[142]. It can be calculated for a phenotype of interest as

$$\mathrm{GRS} = \Sigma^{m}_{i}\, \hat{\beta}_i\, n_i,$$

where $\hat{\beta}_i$ is the estimated effect size of a given allele at the i^{th} locus of a total of m loci, and n_i is the number of copies of that allele. GRS assumes allele additivity, since complex traits are typically controlled by low-effect additive mutations[65,143].

A related measure, the genetic mutation load, can be defined as the burden of deleterious mutations that produces a decrease in fitness from its maximum absolute value. The average population fitness can be calculated as

$$W \approx \exp(-\Sigma^{m}_{i}\, s_i\, h_i\, \mathrm{Hz}_i + s_i\, \mathrm{Hm}_i),$$

where, for a deleterious allele at the i^{th} locus, s_i is the selection coefficient, h_i the dominance coefficient, Hz_i the number of heterozygotes, and Hm_i the number of homozygotes[66].

HUMAN POPULATIONS IN THE MEDITERRANEAN REGION

The dual nature of the Mediterranean sea, at the same time a barrier and a means of travel, have played a crucial role in the history or this region. While different human peopling patterns took place originally in Southern Europe, North Africa, and the Levant, large-scale population movements gradually forged a common history for these coasts, while always keeping different idiosyncrasies.

Population movements

First humans

To date, the oldest *Homo sapiens* remains in the Mediterranean region are those of Jebel Irhoud (Morocco)[144]. These consist on what has been considered an early stage of anatomically modern human dated as ~315 thousand years old. Meanwhile, a ~210,000 years old fossil found in Greece has been recently identified as another early *H. sapiens*, reflecting a failed dispersal from Africa[145]. However, the oldest findings of *Homo sp.* in Southern Europe are those of *H. heidelbergensis*, with the paradigmatic example of the Sima de los Huesos (Spain), whose individuals have been dated as 430,000 years old[146]. The European *H. heidelbergensis* has been suggested as an ancestral form of *H. neanderthalensis* (Neanderthal)[147], which in turn has been found in a number of sites in Southern Europe and the Levant. Regarding this latter region, a fossil at Zuttiyeh cave (Israel) dated to between 500-200 KYA has been proposed to have affinities to *H. erectus*, *H. neanderthalensis*, and early *H. sapiens*[148].

After the OoA, which some authors consider to have taken place via Egypt[149], modern *H. sapiens* migration waves originated in Southwestern Asia reached the Levant ~48 KYA[150], Southern Europe ~45-43 KYA[151], and North Africa ~45-40 KYA[152].

Neanderthal introgression

Although modern humans coexisted with the Neanderthals in Europe and the Levant, and some admixture pulses have been detected within Europe[153], it is considered that the introgression event that led to the Neanderthal component in Mediterranean populations had already taken place 65-47 KYA in Western Eurasia[154]. Neanderthals are proposed to have undergone general extinction ~40 KYA[155], although the site at Gorham's Cave, Gibraltar, has Neanderthal fossils dated as just 28,000 years old[156]. Some proposed non-exclusive causes of the extinction are i)direct and/or indirect competition with *H. sapiens*[157], ii)pathogens introduced by the latter[158], and iii)failure in adapting to climate change[159].

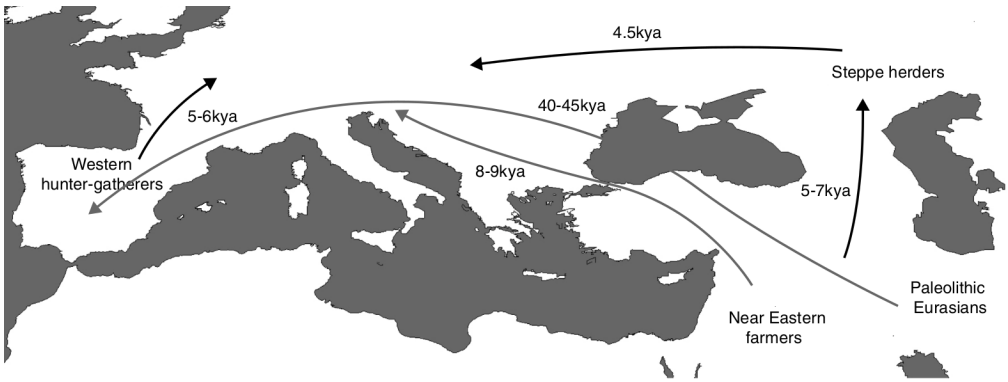


Figure 26. Main prehistoric population movements in Southern Europe inferred from modern DNA (grey arrows) and ancient DNA (black arrows) analysis (adapted from Haber et al., 2016)[160].

Last glacial maximum

The whole human history has taken place within the Quaternary Glaciation, which has consisted of glacial periods 40,000-100,000 years long, intercalated with milder interglacials. Within the Last Glacial Period (115-12 KYA), the Last Glacial Maximum (LGM, 27-19 KYA) was the period of maximum extension of the glacial sheets. This affected European populations, as the glacial sheets reached current Poland, Germany and the British Isles[161], and permafrost extended south of it. Southern France, and the Iberian, Italian, and Balkan peninsulas, acted as refugia where Southern European populations survived this time period[162].

Neolithic

The retreat of the ices that came with the ongoing interglacial period marked the start of the Neolithic in the Fertile Crescent, with the development of agricultural and herding techniques, pottery, polished stone tools, and the adoption of sedentariness. This region saw a population increase and outwards migration waves, bringing with them the new technologies. In this line, gene flow >12 KYA seems to be the main genetic source of nowadays Berber tribes in North Africa[163]. In Europe, humans expanded out of the LGM refugia, and 9-8 KYA the Neolithic migration wave expanded throughout Southern Europe. The resulting Neolithic genetic component is detectable in all current Southern European populations, though a smaller Paleolithic component is present as well in some of them, which suggests a differential degree of population replacement[164]. Numerous megaliths in the Western Mediterranean date from the late Neolithic, although the production of these monuments spanned into the Bronze Age[165].

Bronze Age

There is attested bronze usage in the Levant, Egypt and the Aegean region since ~3200 BPE. This was accompanied by the development of the first Mediterranean civilizations: Ancient Egyptians, Minoans (Crete), Mycenaeans (Southern Greece) and Hittites (Anatolia), and their associated writing systems (Egyptian hieroglyphs, Minoan Linear A, Mycenaean Linear B and

Hittite cuneiform). Meanwhile, at the beginning of the Bronze Age the Western Mediterranean saw the development of the Bell Beaker culture, followed in the Western Iberian Peninsula by the Atlantic Bronze Age culture, and by the expansion of the Urnfield culture from Central Europe to several regions of the Mediterranean[166]. An important genetic component virtually ubiquitous in Southern Europeans is thought to have been originated ~4.5 KYA via a mass migration associated with the Yamnaya and Corded Ware cultures, both based in the steppe north of the Black and Caspian Seas, after the invention of wheeled vehicles[164]. Haak et al. also suggest that the Indo-European languages in Europe could have been introduced by these migrants. The Late Bronze Age collapse (~1200-1150 BPE) marks the abrupt end of the Mediterranean civilizations, probably caused by a combination of factors such as natural disasters, invasions and the introduction of iron metallurgy.

Iron Age

The power vacuum left in the Eastern Mediterranean led to the rise of the Phoenician culture, a group of coastal city-states located within current Lebanon, Syria and Israel territories. Between ~1200-800 BPE the Phoenicians created trade routes and colonies all along the North African coast, parts of the Iberian Peninsula and Asia Minor, and the main Mediterranean islands. They thus transmitted many elements of their culture, one of the most important being their alphabet, which was the base for, among others, the Greek alphabet[167].

Another Mediterranean civilizations of this period were the Tartesians, Iberians, and Aquitanians in the Iberian Peninsula, the Etruscans in modern Tuscany and Lazio and the Ancient Egyptians, Ancient Greeks, Lydians and Phrygians in the Eastern Mediterranean. Celtic tribes related to the Hallstatt culture – which had originated in Central Europe around 1200 BPE – and to La Tène culture – derived from the former – reached the Iberian Peninsula by ~650 BPE and the Balkans and Anatolia by ~300 BPE, respectively[168].

Classical antiquity and Middle Ages

The Iron Age gave way to classical antiquity, an era characterized by the fight over the cultural and political domination of the Mediterranean between Carthage, the Greek states, and Rome, culminating in the control of the whole region by the Roman Empire in the 2nd century PE. Slavery, trade, and political and militar activity promoted people movements between different parts of the Mediterranean and beyond, albeit not at the scale of previous eras[169].

After a long decline, the Roman Empire fell due to the invasion of Germanic and Hunnic tribes in the 4th-5th centuries. During the 7th-8th centuries, the Arabs conquered all the Levant and North Africa, as well as most of the Iberian Peninsula. The Northern Iberian Christian kingdoms expanded southwards during centuries, until the last Muslim ruled territory in the Iberian Peninsula was conquered in the 15th century[170].

Concerning the Eastern Mediterranean, the Byzantine Empire retained most of the territories of the Eastern Roman Empire for a long time. In the 6th-7th centuries, Slavic tribes reached the Balkan Peninsula. In the 11th century, Anatolia, ruled by the Byzantine Empire, was conquered by the Turkic Seljuk Empire, and then by the Mongols in the 13th century. The Ottoman Empire, originated from one Anatolian petty kingdom that achieved control of the region, began the conquest of the Balkans by the 14th century, causing the fall of the Byzantine Empire in the 15th century[171].

Figure 27. Map of the Roman Empire at its greatest extent in 117 PE (taken from Nacu, 2008)[172].

Recent genome-wide studies

Since the introduction of genetics in anthropological studies, there have been numerous attempts to shed light on long-standing questions regarding the history of the Mediterranean region. For example, a reduced number of markers suggests a historical role of the Mediterranean sea as a barrier to gene flow, finding population structure dividing both coasts[173,174]. However, the relatively low discriminant power that classical, uniparental and sparse markers present within such restricted geographic area, has generally supposed a limitation for finer-scale studies. Now, thanks to the growing use of high-throughput genetic data, important discoveries are been made in unravelling the complex demographic history of this region. For instance, genome-wide array data detect significant levels of North African and sub-Saharan ancestry in Southern Europe, together with a westwards Near Eastern ancestry gradient, all of which contributes to a higher genetic diversity in Southern Europe compared to its northern counterpart[175].

Population structure

In this line, an increasing number of genome-wide datasets – mostly array-based – that include Mediterranean populations have been obtained in the last decade, many of which show patterns of population structure within the region. One of them found that Sephardic and Moroccan Jews form a genetic cluster with other Jewish groups like the Ashkenazi, Middle-Eastern and Caucasian Jews, and non-Jewish Cypriots, Druze and Samaritans, but not with other Levantine or historically neighbouring populations[176]. A survey on the peopling history of the European Romani populations identified a single initial founder population from North/Northwestern India ~1,500 ya: after a rapid migration through the Near and Middle East, where moderate gene flow took place, the Romani started to expand across Europe from the Balkans ~900 ya. Further characteristics of these populations are strong population structure, genetic isolation, and differential rates of gene flow with non-Romani Europeans[177]. Another study, focusing this time on Levantine populations, reveals recent population structure based on the closer genetic

affinity to either Europeans and Central Asians or Middle Easterners and Africans[178]. One recent study covering populations from the Iberian Peninsula suggests a north to south stratification that matches the pattern of expansion of the Christian medieval kingdoms[179] (Figure 28), while another finds genetic differentiation between the Ibizan population and other Spaniards, possibly attributable to genetic drift caused by historical isolation[180]. Finally, WGS data of several Sardinian populations[181] confirm elevated levels of an early Neolithic component proposed elsewhere[164], as well as report within-island population structure and male-driven gene flow of Bronze Age steppe ancestry.

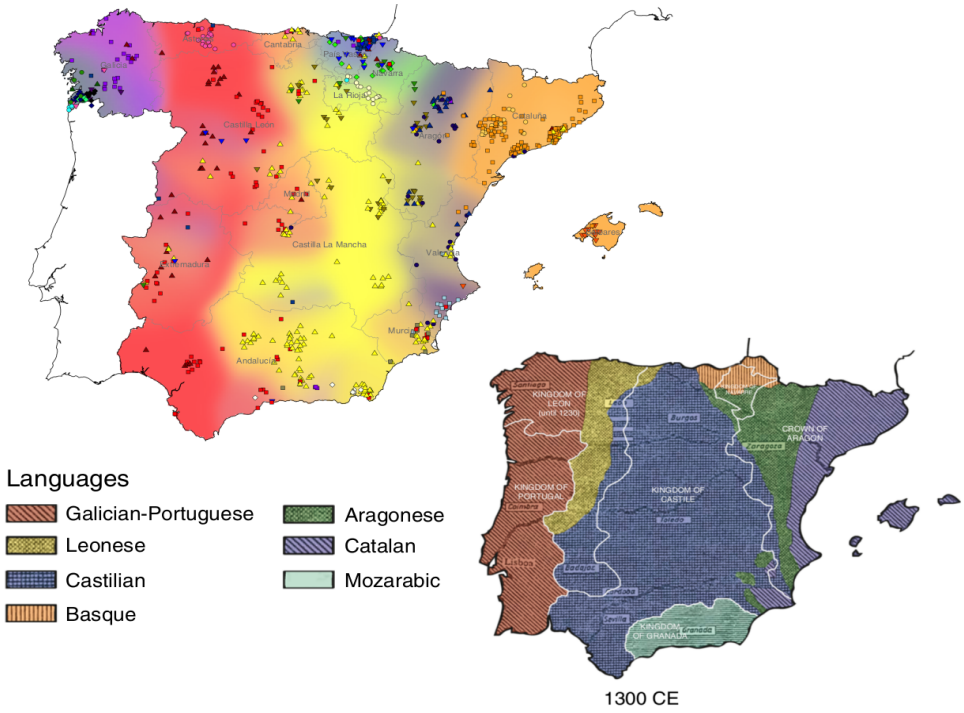


Figure 28. Overlapping of genetic structure (up) and medieval linguistic and political borders (down) in Spain. Array data of 1,413 individuals was analysed with fineSTRUCTURE, and the spatial densities of the 14 inferred clusters are represented as a background colour gradient in the map. Individuals whose four grandparents were born within 80 km of their average birthplace (centroid) are represented by points placed at this centroid. The colour and symbol of each point corresponds to the cluster the individual was assigned to (adapted from Bycroft et al., 2019)[179].

North Africa's complex demographic history

Regarding North African populations, there is a number of recent publications using genome-wide array data that deal with the complex history of this region. For instance, a Neanderthal component has been detected in North African populations, not considered to be due to recent European or Near-Eastern migrations[182]. In this line, two studies estimate that the autochthonous Berber component is a mixture of ancestral inhabitants, related to OoA populations, and a migrational wave coming from the Near East 15 KYA[163,183]. The first study also proposes two pulses of sub-Saharan gene flow that reached Morocco and Egypt 1,200 and 750 years ago, respectively. The second study further shows that Tunisian Berbers appear to have diverged from surrounding populations due to genetic isolation, and also that, due to continuous gene flow from the Middle East, Egyptians are genetically closer to Eurasians than to other North Africans.

Finally, a different study found the current cultural stratification between Berbers and Arabs not to be correlated with genetic structure, with all populations, except two Berber groups, presenting a high degree of genetic heterogeneity[184]. The researchers estimate the Arab component to date from the Arabization of the region in the 7th century. They have also identified sub-Saharan migration waves since the 1st century B.C., with a strong peak in the 17th century, due to slave trade.

Neolithic expansion

Another topic of interest is that of the nature of the Neolithic expansion (Figure 29), which genetic studies have proven to be not just cultural, but that it involved extensive migrations[164]. For example, it has been seen that early European farmers descended from a population related to Northwestern Anatolians[185]. Genomic data from ancient individuals from Anatolia, Levant and the Caucasus confirm these previous results, but they also reveal a culturally-driven Neolithic transition within the Near East that kept the previous hunter-gatherer population structure[186].

Focusing on the Neolithic expansion in North Africa, a recent survey using whole X chromosome sequences estimated that it was synchronous and very similar to that occurred in Europe, both in terms of carrying capacity and speed of expansion[187].

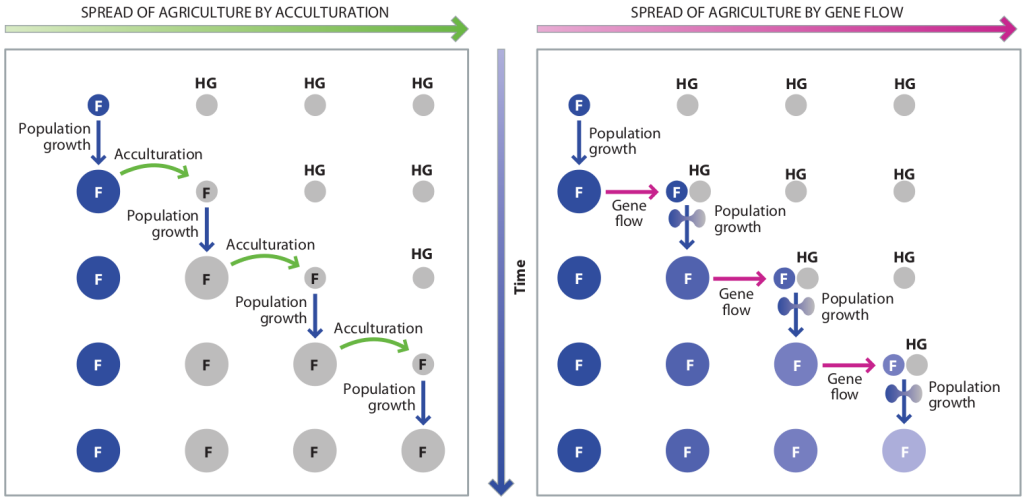


Figure 29. Proposed models for the spread of agriculture. F and HG indicate a farming or hunter-gatherer economy, respectively. Blue shading indicates the degree of ancestry related to the population that developed farming in origin. The acculturation hypothesis (left) suggests that the farming technology was learnt by hunter-gatherers from their neighbours, leading to population growth and the transmission of the knowledge to other hunter-gatherers. The population replacement hypothesis (right) proposes instead that migrant groups from the original farming population were the transmission vector of the technology, by admixing with, or replacing, the original hunter-gatherers (taken from Jobling et al., 2013)[13].

A recent WGS analysis including current and ancient individuals from the Iberian Peninsula, Europe, North Africa and the Middle East[188] sheds some light on this and related topics, mainly in the context of the Iberian Peninsula (Figure 30). Namely, it points out to the existence of high genetic structure between Northwestern and Southeastern Iberian Mesolithic hunter-gatherers, before the Neolithic wave that greatly altered the Iberian genetic pool. It also replicates previous findings that show an increase of Central European-related hunter-gatherer ancestry

after 4000 BPE, the introduction of a 40% Steppe ancestry component by ~2000 BPE, with an almost total replacement of the Y-chromosome haplogroups (suggesting a sex-biased migration), and the description of present-day Basques as an Iron Age population largely unaffected by later gene flow from North Africa and the Central and Eastern Mediterranean regions. Additionally, this study reveals early sporadic contacts between Iberia and North Africa as far as ~2500 BPE. Overall, this study stressing the usefulness of ancient DNA for depicting migration events.

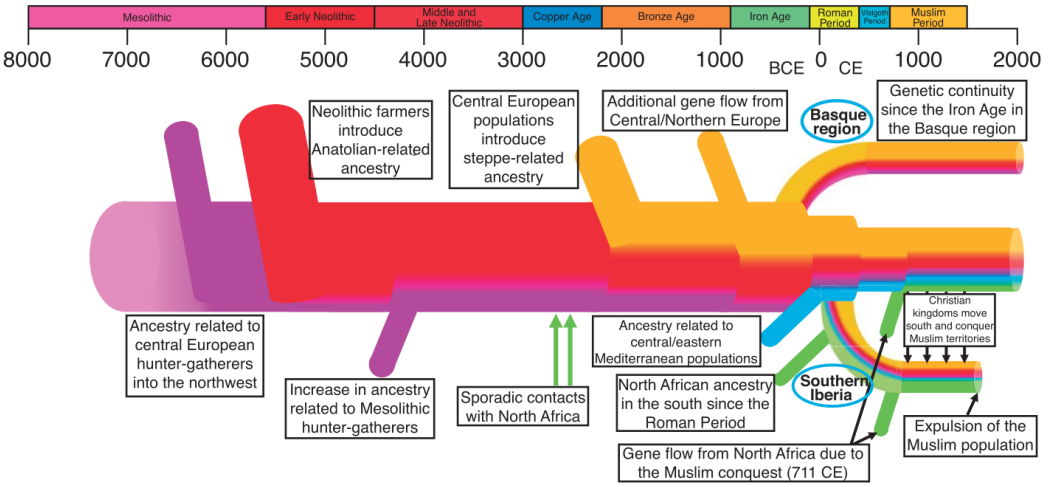


Figure 30. Timeline of the main migration events in the Iberian Peninsula (taken from Olalde et al., 2019)[188].

Population groups studied in this work

The populations studied in this work encompass the whole Mediterranean and adjacent regions, including i)the Iberian Peninsula, ii)South France, iii)Italy, iv)the Balkans, v)Turkey, vi)Israelis of Turkish Sephardic or Iraqi Jewish descent, vii)Jordania, viii)several Berber and Arab Moroccan populations, ix)Tunisia, x)M'zab people from Algeria, xi)Libya, and xii)the islands of Tenerife, Minorca, Sardinia, Sicily and Crete. The genetics of these human groups were

analysed in order to give answers to several demographic and epidemiological questions, as well as to provide context for the study of two populations of particular historical interest: the Sephardic Jews and the Spanish Eastern Pyreneans, whose historical backgrounds are detailed in the following sections.

Sephardic Jews

Judaism is an Abrahamic religion originally practiced in Levant by Israelite tribes from 1200-1000 BPE[189]. Despite its ultimate Near Eastern origin, the Sephardic culture and ethnic group refers to that of the people that historically practiced Judaism in the Iberian Peninsula – having documented presence there since Roman times[190] – as well as their subsequent diaspora. The Sephardic Jews were traditionally dedicated to trade and craftsmanship, and thus mainly lived in urban spaces[191]. They were always a religious minority, alternating periods of general tolerance towards them with eras of persecution.

By the 15th century, approximately 400,000 Jews lived in Spain[192]. Since the 14th century, however, persecution against Jews began to escalate, leading to the Massacre of 1391 and the edict of expulsion in 1492. These events caused the conversion of 200,000-300,000 Sephardic Jews, known since then as Conversos, and the exile of between 50,000 and 80,000[193] to Portugal, Italy, France, Holland, the Western Maghreb, the Ottoman Empire, and Latin American countries[194,195]. The Portuguese edict of 1497 forced Jews to convert to Christianity and stay in Portugal rather than allow their departure[196].

The Spanish and Portuguese Inquisitions heavily persecuted Conversos under the alleged secret practice of Judaism, which incited successive departures until the 18th century. For those who remained, the prohibition to practice Jewish rites, and increasing intermarriage between Conversos and Catholics, caused the loss of the Sephardic identity and culture over time. The Sephardic traditions and language (Ladino, a variety of Old Spanish) were conserved until the 20th century mainly in the exile territories of Western Maghreb, the Balkans, Turkey and Levant. However, the genocide of Balkan Jews perpetrated by Nazi Germany[194], and the posterior migrations to Israel, where Modern Hebrew was the *lingua franca* among the different Jewish groups, further led to the decline of this culture.

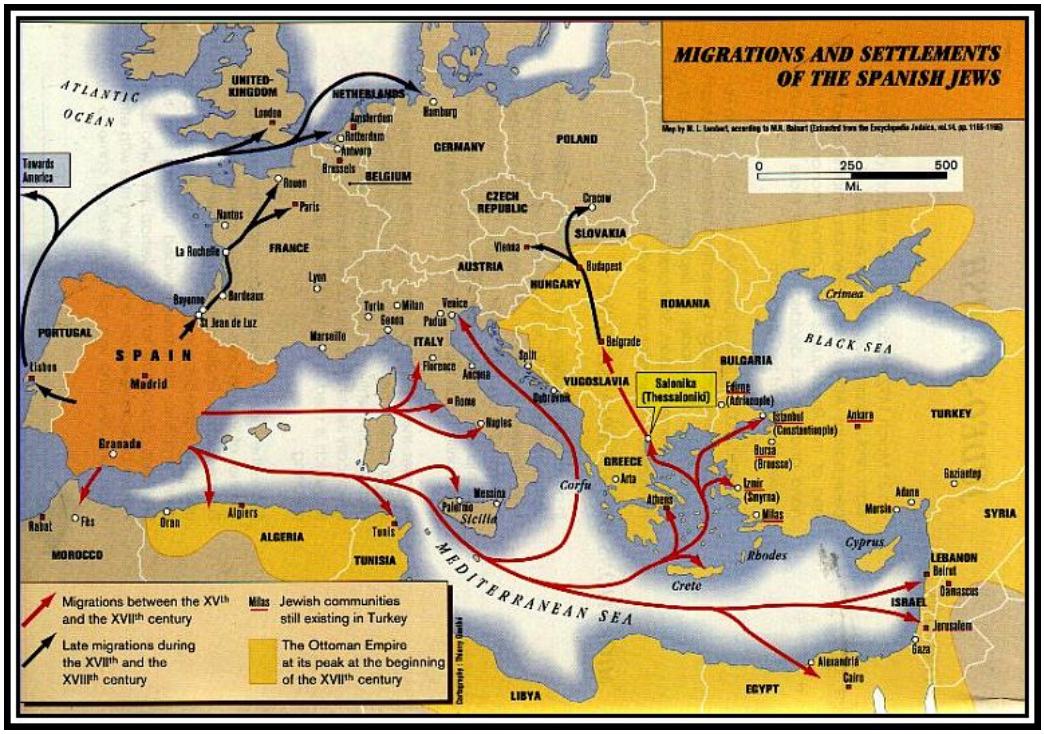


Figure 31. Migrations and settlements of the Sephardic Jews after their expulsion from Spain and, later, Portugal (taken from Encyclopaedia Judaica, 2006)[197].

Spanish Eastern Pyrenees

The Pyrenees have been traditionally classified into Western, Central, and Eastern Pyrenees, as well as into French and Spanish Pyrenees (northern and southern slopes). These divisions have witnessed differential population and cultural processes since the Paleolithic, with the main cultural and genetic gradient following an East-West axis[198]. In this dissertation, the focus will be made on the anthropological Prehistory and History of the Spanish Eastern Pyrenees.

First peoples and the arrival of Neolithic populations

There are sites indicating a continuous human presence in the Spanish Eastern Pyrenees dating back as far as the Lower Paleolithic, with a significant decrease in the Last Glacial Period that reverts with the melting of glaciers in 12000-8000 BPE[199]. Sites dating from 6000 BPE suggest the arrival of Neolithic populations to the area linked to a population growth[200]. For instance, a site in Andorra dating from 5000-4500 BPE contains remains that indicate cultivation of cereals, breeding of cattle, and crafting of Neolithic tools and decorated ceramics[201]. There is also an identified deforestation process taking place in 5000-4000 BPE[202]. In the late Neolithic-Chalcolithic (3300-2200 BPE) there is a peak in the number of megaliths[200], and from 2300 BPE metal elements like gold and copper are introduced in the area[203,204].

Bronze Age migrations

In the early-middle Bronze Age the use of bronze and other metals substituted flint and other types of stone in the production of everyday tools. During this period, there was a high connection between the northern and southern sides of the Eastern Pyrenees, sharing more common cultural features than with the rest of the Iberian Peninsula or Europe. However, the traditional hypothesis of a culturally homogeneous Pyrenees is no longer supported by the data[200].

In the middle Bronze Age (1400-1200 BPE), the apparition of new features on ceramic vases, new types of knifes and axes, and new funeral rituals, all of them present in older strata in the French Pyrenees, suggests a migration of people from the European side of the mountain chain[205]. The result of this process was probably a homogenization of the Spanish Eastern Pyrenean population, or at least closer relationships between some areas[206].

During the late Bronze Age (1100-700 BPE) there is a decline in megalithic sites, and the introduction of people associated with the Urnfield culture[207] indicated by the presence of remains that suggest cremation and depositing of the ashes in buried urns, and new types of ceramic styles, bronze tools, dwelling organisation and subsistence models. It is estimated that the total numbers of newcomers did not surpass the few thousands[208].

Pre-Roman tribes and Roman conquest

Findings suggest an isolation of the Spanish Eastern Pyreneans with respect to the cultural and technological changes introduced since ~625 BPE by the Phoenicians and, later, the Greeks, in the Catalan northern coast, like iron tools and the potter's wheel. There are several pre-Roman Spanish Eastern Pyrenean tribes mentioned by ancient sources (see Figure 32). *Graffiti* found in Cerdanya suggest that they spoke a variety of the Iberian language[209]. The Roman conquest of the Spanish Eastern Pyrenees took place around 39 BPE, but by the beginning of the Middle Ages the romanization process had not been completed[208].

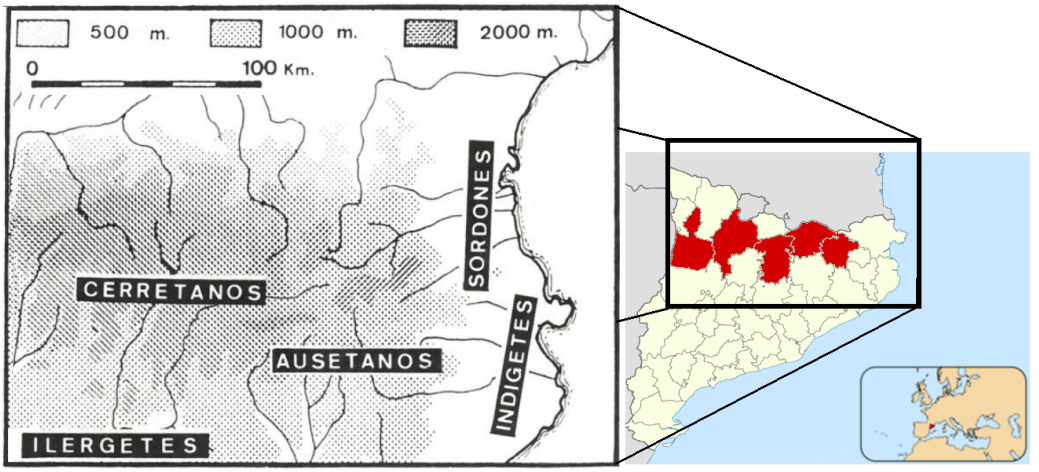


*Figure 32. Main pre-Roman tribes in the Spanish Eastern Pyrenees in the 1st millennium BPE (map at the left). Shading indicates elevation. The map at the right highlights in red the current administrative regions (*comarques*) of the Spanish Eastern Pyrenees included in this work, within the context of Catalonia. In turn, the small box at the bottom right corner highlights Catalonia within a global context (adapted from Ruiz, 1995, and Wikimedia Commons, 2016)[208,210].*

Middle Ages

After the collapse of the Western Roman Empire in the 5th century, the region fell under the govern of the Visigoths. It is estimated that the Roman and Visigoth people only represented a 2.2-4.4% of the Spanish Eastern Pyreneans populati[191]on182. The Islamic conquest of the Spanish Eastern Pyrenees in the 8th century lasted 80 years, before if was taken by the Franks and became part of the Marca Hispanica, a buffer zone of the Carolingian Empire. In the 10th century the Catalan counties gained independence from the West Franks, before their annexation to the Crown of Aragon in the 12th century.

GOALS

GOALS

The field of human population genetics is being transformed by the fast development of genotyping and sequencing techniques. Times and costs reduction have boosted the production of genetic data, which, in combination with sophisticated methods and extensive computer resources, allows for more refined genetic analyses. However, modest dataset sizes could still be useful for obtaining information about large-scale demographic events, while reducing costs. In this work, I present four research articles that make use of increasingly dense genetic data in order to shed light on different topics of the history of human Mediterranean populations.

- Testing for differences in the allele distribution of a cancer-associated gene, potentially caused by the role of the Mediterranean sea as a genetic barrier[211].

- Testing whether historical sub-Saharan gene flow has left detectable signals in genomic regions associated with coronary artery disease[212].

- Measurement of the Sephardic genetic component in current populations from the Iberian Peninsula and detection of signatures of the diasporic process[213].

- Use of deep-coverage WGS data from Spanish Eastern Pyreneans to assess population structure patterns, analyse their demographic history, and evaluate their classification as a genetic isolate as well as the epidemiological implications of this.

SUPERVISORS' REPORT
ON THE IMPACT FACTOR
OF THE PUBLISHED ARTICLES

SUPERVISORS' REPORT ON THE IMPACT FACTOR OF THE PUBLISHED ARTICLES

The doctoral thesis 'Human population genetics in the Mediterranean region. From single markers to whole-genome sequencing' is based on the results of four original research studies carried out by Miguel Martín Álvarez Álvarez. Two of the studies have been published in international peer reviewed journals, and the remaining two studies have been recently submitted for their consideration.

In all four studies, several questions regarding the demographic and biological history of human populations of the Mediterranean region are addressed through the analysis of genetic variation. The quality of the samples obtained and employed, and the state-of-the-art methods that were carried out, have undoubtedly provided insights to the field of the human population genetics of the Mediterranean. The importance of the research conducted is demonstrated by the quality of the journals in which two of the studies have been published.

- The article **Population variation of *LIN28B* in the Mediterranean: Novel markers for microgeographic discrimination** has been published in *The American Journal of Human Biology*, which is the official journal of the Human Biology Association, publishing articles in the interdisciplinary field of Human Biology. It had an impact factor of 1.780 in the year of publication (2016), and it is classified in the Scimago Journal Rank (SJR) in the first quartile of the area 'Anthropology'. The participation of Miguel Martín Álvarez Álvarez in this article consisted of the following tasks:

- Participation in the design of the study together with Dr. Pedro Moral Castrillo
- Genotype determination of the single nucleotide polymorphisms (SNPs)
- Creation of the genotype database and statistical analysis of the data
- Participation in the manuscript drafting together with Dr. Pedro Moral Castrillo

- The article **A survey of sub-Saharan gene flow into the Mediterranean at risk loci for coronary artery disease** has been published in *The European Journal of Human Genetics*, which is the official journal of the European Society of Human Genetics, publishing high-quality papers in the field of human genetics and genomics. It had an impact factor of 3.636 in the year of publication (2017), and it is classified in SJR in the first quartile of the area 'Genetics'. The participation of Miguel Martín Álvarez Álvarez in this article consisted of the following tasks:

- Participation in the design of the study together with Dr. Pedro Moral Castrillo and Dr. Georgios Athanasiadis
- Statistical analysis of the data
- Participation in the manuscript drafting together with Dr. Pedro Moral Castrillo and Dr. Georgios Athanasiadis

- The article **Genetic analysis of Sephardic ancestry in the Iberian Peninsula** has been submitted to *Human Genetics*. The participation of Miguel Martín Álvarez Álvarez in this article consisted of the following tasks:

- Participation in the design of the study together with Dr. Georgios Athanasiadis
- Statistical analysis of the data
- Participation in the manuscript drafting together with Dr. Pedro Moral Castrillo and Dr. Georgios Athanasiadis

- The article **High-coverage sequence data from the Spanish Eastern Pyrenees suggest patterns of population structure and isolation** has been submitted to *The New England Journal of Medicine.* The participation of Miguel Martín Álvarez Álvarez in this article consisted of the following tasks:

- Participation in the design of the study together with Dr. Pedro Moral and Dr. Oscar Lao
- Sample collection together with Dr. Pedro Moral
- Creation of a digital database containing the genealogical information of the samples

- Sample preparation for the shipment to the sequencing facility
- Creation of the genotype database and statistical analysis of the data together with Dr. Pedro Moral, Dr. Oscar Lao and Iago Maceda
- Participation in the manuscript drafting together with Dr. Pedro Moral, Dr. Georgios Athanasiadis, Dr. Oscar Lao and Iago Maceda

In addition, it is important to note that none of the co-authors of this article have used the results of this work in any implicit or explicit way to develop another doctoral thesis. As a consequence, these articles form part of the doctoral thesis of Miguel Martín Álvarez Álvarez exclusively.

Signed by

Dr. Pedro Moral Castrillo
Director
Barcelona, 23 July 2019

Dr. Georgios Athanasiadis
Director
Barcelona, 23 July 2019

PUBLICATIONS

Article I

Álvarez-Álvarez et al., 2016

Population variation of *LIN28B* in the Mediterranean: Novel markers for micro-geographic discrimination

Miguel M. Álvarez-Álvarez, Robert Carreras-Torres, Daniela Zanetti, Esteban Vegas and Pedro Moral

American Journal of Human Biology 2016; 28(6):905-12; doi: 10.1002/ajhb.22887

Resumen en castellano

El objetivo de este estudio fue determinar si las variantes del gen *LIN28B* se encuentran distribuidas de una manera desigual en la región mediterránea. Para ello, se realizó un análisis de las distribuciones alélicas de tres polimorfismos de nucleótido único (SNPs) presentes en dicho gen – concretamente rs7759938, rs314277 y rs221639 – en 24 poblaciones de este ámbito geográfico. Estos SNPs han sido recientemente relacionados con la edad a la menarquia, incremento de la estatura en la pubertad, "body mass index" (BMI) en la pubertad, niveles de exposición a la testosterona en fase prenatal y supervivencia al cáncer.

Un total de 1.197 muestras de ADN fueron genotipadas. Las frecuencias alélicas se emplearon para determinar las relaciones interpoblacionales, usándose poblaciones del Proyecto 1000 Genomas como grupos externos al área mediterránea. También se determinaron las distribuciones genotípicas poblacionales, y se corroboró la presencia de una estructuración poblacional en el Mediterráneo.

Los resultados indican un grado de variación significativo ($F_{ST} = 0.043$, $P < 0.0001$). Las frecuencias alélicas muestras diferencias significativas entre poblaciones. Un análisis jerárquico de la varianza es consistente con una diferenciación principal entre las poblaciones de las costas norte y sur del Mediterráneo. Esta diferencia es especialmente evidente en la inesperada distribución del SNP rs221639, el cual muestra uno de los valores F_{ST} más altos describidos en la región mediterránea hasta la fecha (11.5%, $P < 0.0001$).

La diferenciación poblacional y la estructuración de la variación genética, de acuerdo con estudios previos, indican que los SNPs estudiados son buenas herramientas para el estudio de las poblaciones humanas mediterráneas, incluso a un nivel microgeográfico.

Original Research Article

Population Variation of LIN28B in the Mediterranean: Novel Markers for Microgeographic Discrimination

MIGUEL M. ÁLVAREZ-ÁLVAREZ,[1] ROBERT CARRERAS-TORRES,[1] DANIELA ZANETTI,[1] ESTEBAN VEGAS,[2] AND PEDRO MORAL[1]*
[1]*Anthropology Unit, Department of Animal Biology, Faculty of Biology, University of Barcelona, Barcelona, Spain. Biodiversity Research Institute (IRBio)*
[2]*Department of Statistics, Faculty of Biology, University of Barcelona, Barcelona, Spain*

Objectives: The aim of this study is to determine whether the *LIN28B* gene is differentially distributed in the Mediterranean region through the analysis of the allele distribution of three single nucleotide polymorphisms (SNPs), namely rs7759938, rs314277, and rs221639, in 24 populations. These SNPs have been recently related to the age at menarche, pubertal height growth, peripubertal body mass index, levels of prenatal testosterone exposure, and cancer survival.

Methods: A total of 1,197 DNA samples were genotyped. The allele frequencies were used to determine the relationship between populations, with data from the 1000 Genomes Project being used for external comparisons. The genotype distributions and the population structure between populations and groups of populations were determined.

Results: The population results indicate a significant degree of variation ($F_{ST} = 0.043$, $P < 0.0001$). Allele frequencies show significant differences among populations. A hierarchical variance analysis is consistent with a primary differentiation between populations on the North and South coasts of the Mediterranean. This difference is especially evident in the unexpected distribution of the SNP rs221639, which shows one of the highest F_{ST} (11.5%, $P < 0.0001$) values described in the Mediterranean region thus far.

Conclusion: The population differentiation and the structuring of the genetic variance, in agreement with previous studies, indicate that the SNPs in question are good tools for the study of human populations, even at a microgeographic level. Am. J. Hum. Biol. 00:000–000, 2016.

INTRODUCTION

Advances in molecular methodologies are uncovering a staggering number of variable sites in the human genome, most of them still not systematically analyzed in population studies (The 1000 Genomes Project Consortium, 2012). Therefore, this is a first contribution to the study of human population genetics regarding population variation in three single nucleotide polymorphisms (SNPs) located within or next to the human *LIN28B* gene in a set of 24 populations of the Mediterranean region. The *LIN28B* gene, located on chromosome 6q21 (Fig. 1), encodes a protein that selectively blocks the processing of let-7 miRNA precursor molecules into mature miRNAs (Piskounova et al., 2011; Rybak et al., 2008).

LIN28B reportedly contains about 2,000 SNPs in its sequence (single nucleotide polymorphism database, http://www.ncbi.nlm.nih.gov/SNP/). Several studies have found associations between different *LIN28B* variants and developmental timing traits such as prenatal testosterone exposure (Lawrance-Owen et al., 2013; Medland et al., 2010), age at menarche (Carty et al., 2013; He et al., 2009; Ong et al., 2009; Perry et al., 2009; Sulem et al., 2009), pubertal height growth (Cousminer et al., 2013; Perry et al., 2015; Widen et al., 2010), female adiposity levels at puberty (Johnson et al., 2013; Ong et al., 2011), transition from fetal to adult cells in the thymus (Shiyun et al., 2014), and development and tissue degeneration related to aging (Keane & De Magalhães, 2013). Early menarche has been correlated with a higher risk of subsequent obesity, Type 2 diabetes, cardiovascular disease, and a shorter adult height (Day et al., 2015; Elks et al., 2013; Onland-Moret, 2005). It is also associated with breast cancer (Collaborative Group on Hormonal Factors in Breast Cancer, 2012) and all-cause mortality (Jacobsen et al., 2007). *LIN28B* itself seems to play a role in the proliferation of tumor cell lines (Guo et al., 2006), and it contributes to the development of several types of cancer (Anneleen et al., 2014; Borrego-Diaz et al., 2014; Chong et al., 2014; Hamano et al., 2012; Hetty et al., 2014; Hu et al., 2014; Lu et al., 2012; Pang et al., 2014; Powers et al., 2013; Viswanathan et al., 2009; Wen et al., 2014; Wu et al., 2013; You et al., 2014).

The three markers were selected for study because they were polymorphic (the heterozygosities in all of the 1000 Genomes populations were 0.450 for rs7759938, 0.274 for rs314277, and 0.212 for rs221639 (http://browser.1000genomes.org/index.html) and did not exhibit linkage disequilibrium (LD) (The International HapMap Consortium, 2003). Two of the examined SNPs, rs7759938 and rs314277, have been previously reported in independent surveys to be directly associated with developmental timing and/or adult related phenotypes (Cousminer et al., 2013;

Additional Supporting Information may be found in the online version of this article.

M.M. Álvarez-Álvarez analyzed the data and drafted the manuscript. R. Carreras-Torres performed the DNA extractions. R. Carreras-Torres and D. Zanetti assisted M.M. Álvarez-Álvarez with the data analysis. Vegas designed the resampling method for the distance matrices. Moral designed the study, and directed implementation and data collection.

Abbreviations: AIM, ancestry informative marker; FDR, false discovery rate; HW, Hardy–Weinberg; LD, linkage disequilibrium; MDS, multidimensional scaling; SNPs, single nucleotide polymorphisms.

Contract grant sponsor: Spanish Ministry of Science and Innovation; Contract grant number: CGL2011-27866.

*Correspondence to: Pedro Moral; Anthropology Unit, Animal Biology Department, Faculty of Biology, University of Barcelona, Barcelona 08028, Spain. E-mail: pmoral@ub.edu

Received 24 December 2015; Revision received 11 April 2016; Accepted 6 June 2016

DOI: 10.1002/ajhb.22887
Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com).

Johnson et al., 2013; Lawrance-Owen et al., 2013; Lu et al., 2012; Perry et al., 2015; Ye et al., 2012; Tu et al., 2015).

Comparison of data from the 1000 Genomes Project for *LIN28B* and the two genes that are most related to its primary pathways, processes and functions, namely *LIN28A* and *ZCCHC11* (NCBI AceView: http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html?human), indicated different patterns of interpopulation differences. F_{ST} and F_{SC} values obtained for *LIN28A* and *ZCCHC11* indicated strong differentiation for *ZCCHC11* ($F_{ST} = 0.27$ and $F_{SC} = 0.27$, with $P < 0.0001$) but a smaller differentiation for *LIN28A* ($F_{ST} = 0.07$ and $F_{SC} = 0.06$, with $P < 0.0001$).

Working from these findings, the main objective of this study was to determine whether the three SNP variants of the *LIN28B* gene present a differential distribution in populations of the Mediterranean region. The genetic diversity of Mediterranean populations has been a topic of interest for anthropological scholars for a long time (Simoni et al., 1999). Many previous studies, using different kinds of genetic markers, have reported a North-South differentiation as the most remarkable feature (Athanasiadis et al., 2010; Athanasiadis & Moral, 2013; Bosch et al., 2001; Capelli et al., 2006; Comas et al., 2000; González-Pérez et al., 2010). Others have identified gene flow between the North and South Mediterranean coasts (Botigue et al., 2013). Recent genome wide studies of Mediterranean populations have further examined the influence of natural selection on certain genomic regions (Piras et al., 2012) and demographic gene flow within this region (Botigue et al., 2013). In this context, our study provides new genetic data from markers that have thus far not been analyzed in Mediterranean populations in order to study the genetic differentiation in this geographic area.

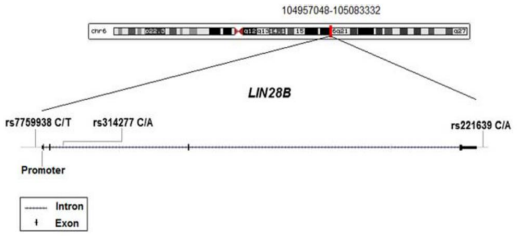


Fig. 1. Location of the three SNPs analyzed in the LIN28B gene, shown in the 5′ to 3′ direction. The rs7759938 is located in the promoter region, the rs314277 in the second intron and the rs221639 in the 3′ region. The unit of the genomic coordinates is the base pair (bp). The image has been modified from the UCSC browser (http://genome.ucsc.edu).

METHODS

Populations

A total of 1,197 DNA samples of individuals from 24 Mediterranean populations were genotyped according to an initial design of around 50 individuals per population. The DNA samples came from the population sample collection in the Unit of Anthropology of the Animal Biology Department at the University of Barcelona. The chosen populations encompass the whole geographic Mediterranean range and include Asturias, Pas Valley, Basque Country; Valencia, Las Alpujarras in Southeast Andalusia; Minorca in the Balearic Islands; Tenerife in the Canary Islands (Spain); Khenifra, Asni, Amizmiz, Sidi Bouhria and Arabs from Doukkala-Abda (Morocco); M'zab (Algeria); North/Central Tunisia and South Tunisia (Tunisia); Sardinia and North/Central Italy (Italy); Bosnia (Bosnia and Herzegovina); Greece and Crete (Greece); Turkey; Northwest Libya, Northeast Libya and South Libya (Libya). The final sample sizes ranged from 22 to 56 individuals as specified in Table 1.

TABLE 1. SNP allele frequencies, P-values of Hardy–Weinberg equilibrium and gene diversities for all populations and loci

	rs7759938				rs314277				rs221639				
Population	*N*	Freq.allele (C)	p.v. H-W	Heterozyn.b.	*N*	Freq.allele (C)	p.v. H-W	Heterozyn.b.	*N*	Freq.allele (C)	p.v. H-W	Heterozyn.b.	H. across loci
Asturias	37	0.162	0.214	0.276	39	0.936	1	0.122	39	0.077	0.187	0.144	0.180
Pas	49	0.214	0.190	0.340	45	0.911	1	0.164	50	0.010	1	0.020	0.175
Basques	51	0.226	0.244	0.353	48	0.885	0.476	0.205	47	0.032	1	0.063	0.207
Valencia	49	0.276	1	0.403	50	0.870	1	0.229	47	0.043	0.064	0.082	0.238
Alpujarras	50	0.260	1	0.389	52	0.875	0.575	0.221	48	0.042	1	0.081	0.230
Minorca	48	0.292	0.290	0.418	48	0.844	1	0.266	48	0.063	0.153	0.118	0.268
Sardinia	47	0.287	0.289	0.414	47	0.840	0.578	0.271	47	0.043	1	0.082	0.256
Italy N&C	47	0.287	0.153	0.414	46	0.946	1	0.104	42	0.012	1	0.024	0.181
Bosnia	53	0.311	1	0.433	54	0.870	0.578	0.228	54	0.028	1	0.055	0.238
Greece	28	0.232	0.121	0.363	29	0.914	0.171	0.160	28	0.018	1	0.036	0.186
Crete	44	0.352	1	0.462	38	0.842	0.563	0.270	34	0.015	1	0.029	0.254
Turkey	47	0.319	0.501	0.439	46	0.837	0.324	0.276	46	0.033	1	0.064	0.260
Canary Is.	54	0.287	0.100	0.413	54	0.824	0.335	0.293	28	0.054	1	0.103	0.270
Amizmiz	37	0.311	1	0.434	41	0.744	1	0.386	35	0.186	0.066	0.307	0.382
Asni	54	0.269	0.489	0.397	53	0.764	0.130	0.364	30	0.183	0.032	0.305	0.362
Doukkala	34	0.397	0.480	0.486	23	0.717	0.626	0.415	22	0.182	0.107	0.304	0.402
Khenifra	51	0.304	0.751	0.427	30	0.833	1	0.283	31	0.065	1	0.123	0.278
Bouhria	37	0.284	0.223	0.412	15	0.800	0.460	0.331	5	-	-	-	0.248
M'zab	40	0.313	1	0.435	40	0.725	0.693	0.404	37	0.270	0.086	0.400	0.440
Tunisia NC	55	0.291	0.754	0.416	53	0.896	0.439	0.188	51	0.186	0.051	0.306	0.303
Tunisia S	53	0.434	1	0.496	53	0.679	0.349	0.440	50	0.250	0.150	0.379	0.438
Libya NW	25	0.220	1	0.350	26	0.827	1	0.292	26	0.154	0.077	0.266	0.320
Libya S	32	0.313	1	0.437	32	0.844	0.563	0.268	32	0.109	0.307	0.198	0.301
Libya NE	56	0.446	0.598	0.499	56	0.732	0.734	0.396	56	0.179	1	0.296	0.397
Average Heterozygosity				0.413				0.274				0.172	0.284

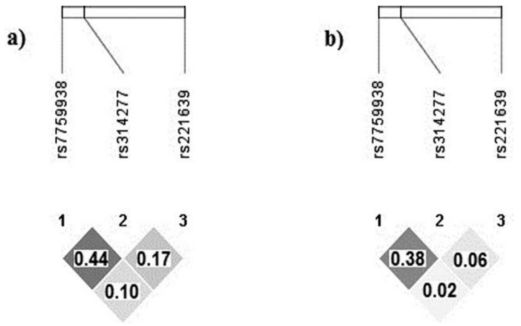


Fig. 2. Linkage disequilibrium (LD) patterns between the three SNPs analyzed, using R-squared measures: (a) LD in the North African populations, (b) LD in the Southern European populations. The darker the intersection boxes, the higher the LD value. All numbers are R-squared values (range from 0 to 1, where 1 means complete predictability of one locus from the other). The plots and LD values were obtained with the Haploview software (Barrett et al., 2005).

In all cases, DNA was extracted from blood obtained from healthy and unrelated subjects of both sexes born in the same geographic region, most of them in rural areas. All of the participants signed an informed consent form before sample donation, and the study was approved by the Ethics Committee of the University of Barcelona.

Genetic markers and laboratory methods

Three SNPs were genotyped in all DNA samples. These included the rs7759938 (chromosomal location 6:105378954, in the promoter region; ancestral allele C, derived allele T), the rs314277 (chromosomal location 6:105407662, in the second intron; ancestral allele C, derived allele A), and the rs221639 (chromosomal location 6:105532843, in the 3′ region; ancestral allele C, derived allele A) (all chromosomal locations are presented in GRCh37 coordinates).

All SNP assays were genotyped on an ABI 7900HT RT-PCR machine using Taqman assays (Applied Biosystems, Warrington, UK) in the Genomic Area of the Parc Cientific of Barcelona. DNA dilutions of 5 ng/μl were prepared in 384-well plates and the TaqMan protocol suggested by Applied Biosystems was followed. The volume of reaction was 5 μl per well, including 1 μl of DNA dilution, 2.5 μl of Genotyping Master Mix for real-time PCR (Applied Biosystems), 0.05 μl of TaqMan fluorescent reporter probes for real-time PCR genotyping (with a stock concentration of 40×), and 1.45 μl of distilled water. As quality controls, blanks for each probe and replications of some individuals were included in all of the runs. The allelic discrimination was performed using the same ABI 7900HT and the SDS 2.3 software through the total amount of fluorescence intensity for both reporters (VIC® and FAM™)for each locus.

Statistical analysis

Standard parameters of human population genetics were computed. Allele frequencies, genetic diversity (Nei, 1978), and Hardy–Weinberg equilibrium were calculated through Genetix (Belkhir et al., 1996) and Arlequin ver3.5 (Excoffier & Lischer, 2010) software. The Bonferroni correction was applied when running the Hardy–Weinberg equilibrium tests. The LD between the three examined loci was measured, sorting the populations by belonging to either North Africa or Southern Europe, employing the Haploview software (Barrett et al., 2005) (Fig. 2). Comparisons of the genotype distributions between pairs of populations were estimated with Genepop software (Raymond & Rousset, 1995), using exact tests for population differentiation. PHASE software (Stephens & Donnelly, 2003; Stephens & Scheet, 2005; Stephens et al., 2001) was employed for the imputation of missing SNP genotypes (3.70% of the total individual alleles) from the haplotype reconstruction applying high stringent criteria (threshold ≥90%).

The relationships between populations were assessed by estimating pairwise genetic distances. Genetic distances using the Reynolds coefficient were estimated with Phylip software (Felsenstein, 1989) and visually represented by multidimensional scaling (MDS) with the R software *stats* package (R Core Team, 2013). In this analysis, the Bouhria population, which lacks information for the SNP rs221639, was excluded. However, to widen the population context, other European (Toscan Italians, Iberian populations, Northwestern Europeans (CEU), British and Finnish) and Sub-Saharan African (Yoruba from Nigeria, Luhya from Kenya, Mende from Sierra Leone, Esan from Nigeria and Gambian) populations from the 1000 Genomes Project phase III were included (http://www.1000genomes.org/).

The structuring of the genetic variation was tested with an Analysis of Molecular Variance (AMOVA) analysis and with the Structure software (Falush et al., 2003, 2007; Pritchard et al., 2000). The AMOVA was performed in Arlequin ver3.5 (Excoffier & Lischer, 2010) to estimate the allele frequency variation among populations (F_{ST}) and its hierarchical distribution between Southern Europe and North Africa groups (F_{CT}) and between populations within groups (F_{SC}). The Structure software was employed assuming an admixture model. The length of the burn-in period was 50,000, and the number of Markov Chain Monte Carlo (MCMC) repeats after burn-in was 100,000 (Hubisz et al., 2009).

Barrier vs. 2.2 (Manni et al., 2004) was employed for checking whether the main barrier to gene flow for the three studied markers was located in the Mediterranean basin or elsewhere in Europe or Africa. This program defines a barrier to gene flow as the geographic area where differences between pairs of populations are the largest. For this purpose, the previously mentioned European and Sub-Saharan African samples from the 1000 Genomes Project were also included in the analysis. A resampling method for the production of an extra number of distance matrices was employed in order to assess the robustness of the barrier to gene flow. This step consisted in the application of a binomial distribution on the original values of allele frequencies and population sizes existing for each SNP, which produced a new table of resampled allele frequencies that was subsequently converted into a Euclidean distance matrix. Two loops were performed, one for 100 matrices and a second for 1,000 matrices. For the resampling, the R software (R Core Team, 2013) was employed.

To find potential signals of selection in the *LIN28B* region, two approaches were followed. In the first method, the loci analyzed were 15 tagged SNPs from the genomic region that surrounds *LIN28B* (namely 6:105404923-105531207) obtained with the Haploview software from HapMap Phase III genotype data. In this set of SNPs, the three studied loci

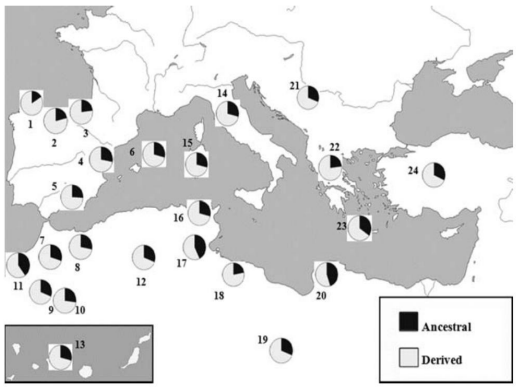


Fig. 3. Allele frequencies of rs7759938 represented by pie charts, where the lighter slice corresponds to the derived allele T and the darker slice corresponds to the ancestral allele C. Populations: 1, Asturias; 2, Pas; 3, Basque Country; 4, Valencia; 5, Las Alpujarras; 6, Minorca; 7, Khenifra; 8, Bouhria; 9, Amizmiz; 10, Asni; 11, Arabs from Doukkala-Abda; 12, M'zab; 13, Canary Islands; 14, North and Central Italy; 15, Sardinia; 16, North and Central Tunisia; 17, South Tunisia; 18, Northwest Libya; 19, South Libya; 20, Northeast Libya; 21, Bosnia; 22, Greece; 23, Crete; 24, Turkey.

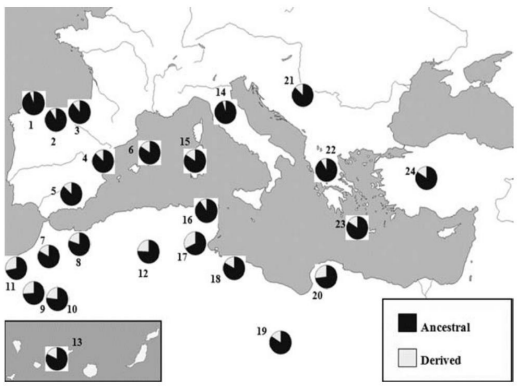


Fig. 4. Allele frequencies of rs314277 represented by pie charts, where the lighter slice corresponds to the derived allele A and the darker slice corresponds to the ancestral allele C.

were included as tagged SNPs by the software. The genotype information for these 15 tagged SNPs was then extracted from the previously mentioned 1000 Genomes European and Sub-Saharan populations. Next, to look for selection signals in these SNPs, we employed Arlequin v3.5 software (Excoffier & Lischer, 2010). The method implemented was based on the probability of observing locus by locus AMOVA statistics as a function of heterozygosity, given a null distribution generated under a hierarchically structured island model of population differentiation (Excoffier et al., 2009). This test detects loci under selection from genome scans that contrast patterns of genetic diversity within and between populations. A total of 20,000 coalescent simulations in 10 groups, with 100 simulated demes per group, were performed. The observed locus-specific measures of population differentiation (F_{ST}) were compared to a null distribution obtained by simulation samples. The

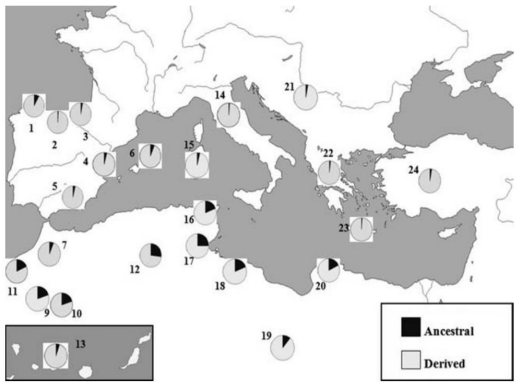


Fig. 5. Allele frequencies of rs221639 represented by pie charts, where the lighter slice corresponds to the derived allele A and the darker slice corresponds to the ancestral allele C.

P-value of each locus was estimated from the joint distribution of heterozygosity and F_{ST} using a kernel density estimation procedure. The null distribution generated was summarized by quantiles of the joint distribution. The 1 and 99% quantiles of the distribution correspond to markers potentially under balancing or directional selection, respectively, at the 1% level, without multiple-test correction and assuming a one-tailed test.

For the second approach, the 1000 Genomes Selection Browser was employed (http://hsb.upf.edu/) (Pybus et al., 2014). Here, we looked for signals of selection in the *LIN28B* region, specifically for the CEU and Yoruba populations, using the F_{ST} rank-scores method, and applying a false discovery rate (FDR) correction with the R software.

RESULTS

Allele frequencies

SNP ancestral-allele frequencies (from http://www.1000genomes.org/), Hardy–Weinberg *P*-values, and gene diversities of the three loci studied in the 24 populations are shown in Table 1. Reliable genotype data were obtained in all populations except the Bouhria population sample from Morocco for which, for unknown technical reasons, only a few individual genotypes were unequivocally identified in the SNP rs221639 and were therefore not considered. In all cases, the observed genotype distribution fit well with Hardy–Weinberg (HW) expectations. Only the Asni population showed a HW *P*-value lower than 0.05 for the rs221639 SNP; it was considered in equilibrium after applying the Bonferroni correction (*P*-value threshold of 0.0007). Low R-squared LD values were obtained (maximum value 0.44) (Fig. 2), which are similar to those reported in public databases.

The population distribution of SNP allele frequencies in the Mediterranean Basin is shown in Figures 3–5. Among the populations analyzed, the frequency of the ancestral allele of rs221639(C) ranged between 0.270 (M'zab) and 0.010 (Pas Valley), for rs314277(C) the highest value corresponded to the continental Italian sample (0.946) and the lowest to South Tunisia (0.679), whereas in the case of rs7759938(C) the highest frequency was found in the Northeast Libya sample (0.446) and the lowest in Asturias (0.162).

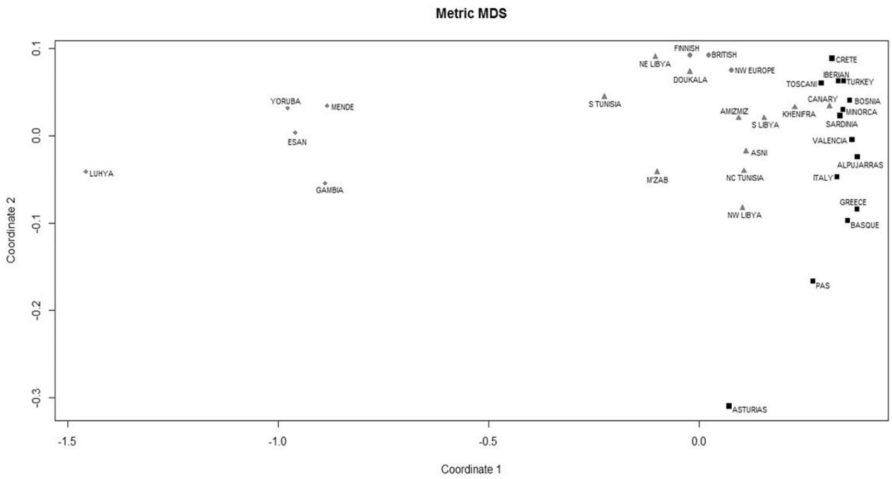


Fig. 6. MDS plot of the Mediterranean populations plus the European and Sub-Saharan African samples from the 1000 Genomes Project, based on the genetic distances obtained from the allele frequencies variation of the three SNPs studied. Diamonds correspond to Sub-Saharan populations, triangles to North African populations, squares to Southern European populations and circles to Northern European populations. The genetic distances separate the populations into two distinct clusters, one containing Sub-Saharan populations and the other containing the remaining populations. Kruskal's stress = 0.10.

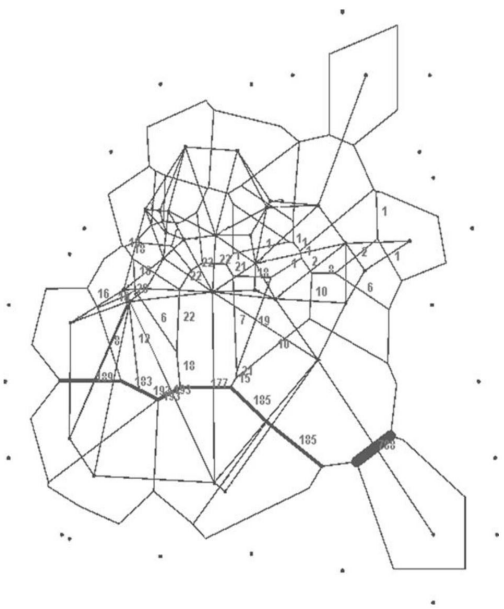


Fig. 7. Barrier plot of the analysis of 1,000 distance matrices. The matrices were obtained through a resampling method explained in *Methods* section, employing the populations from the study plus the five European populations and five Sub-Saharan populations from the 1000 Genomes Project. Each point represents a population, except the external points which are virtual points (the purpose of the virtual points is to obtain a closed Voronoï tessellation enclosing all the populations). The number of matrices in which a barrier to gene flow was the main one is represented next to its corresponding line. For clarity, population names have been removed. The main barrier to gene flow is located between the Sub-Saharan and the remaining populations, with an average statistical robustness of 0.25. The barrier to gene flow is especially strong in its eastern area, namely next to the Luhya population (Kenya).

In general, the geographic allele frequency distribution showed a primary population differentiation between Southern Europe and North Africa, especially for the SNP rs221639 (Figs. 3–5). There was also European–African differentiation for gene diversity with lower values corresponding to Southern Europe when comparing with North African populations (Table 1), especially for the rs221639 locus (ranges: 0.020–0.144 in Southern Europe vs. 0.103–0.400 in North Africa). Comparisons of the SNP genotype distribution by pairs of populations indicated that variation within the North African group was clearly higher than within the Southern European group (22.73% of the tests were significant among North Africans while only 6.06% among Europeans) and that the differences were more important between the Southern European and North African groups (57.14% of the tests were significant) than within each group.

Population relationships and genetic structure

On a worldwide level, the highest Reynolds genetic distances, based on allelic frequencies, corresponded to those between non-Sub-Saharan and Sub-Saharan populations, and the lowest correspond to between-groups of the same region as among Europeans (see Supporting Information). In the Mediterranean region, the average genetic distance was 0.040, with distances between the two coasts being higher (average 0.056) than those within each coast (0.014 in Southern Europe and 0.032 in North Africa) (see Supporting Information).

The genetic distances represented in the MDS plot show in Figure 6 are consistent with the main genetic differentiation being between Sub-Saharan populations and the rest. The remaining populations, Europeans and North Africans, clustered in a global pool reflecting a general similarity of the frequencies of the three loci analyzed. On a finer scale, there was a subclustering of these populations into two groups—North Africans and North Europeans on one side and Southern Europeans on the other.

TABLE 2. F_{ST}, F_{SC} and F_{CT} values for the hierarchical AMOVA test, with P-values for each locus

Locus	F_{ST}	*P*-value	F_{SC}	*P*-value	F_{CT}	*P*-value
rs7759938	0.013	0.006	0.006	0.064	0.006	0.040
rs314277	0.043	<0.0001	0.010	0.013	0.033	<0.0001
rs221639	0.115	<0.0001	0.015	0.055	0.101	<0.0001
All loci	0.043	<0.0001	0.009	0.002	0.034	<0.0001

The main barrier to gene flow depicted by Barrier ver2.2 (Fig. 7), based on the genetic distances obtained from the allele frequencies of the three SNPs when the populations of the study and the European and African samples from 1000 Genomes were joined, traversed the Saharan area. Both performed loops (one for 100 matrices and another one for 1,000 matrices) showed similar levels of average statistical robustness (0.25 aprox.). Figure 7 shows the results of the 1,000 matrices loop.

Regarding genetic structure, a hierarchical analysis of the variance (Table 2) showed that most of the variation occurred within populations (95.7%). Although small, the between-population variation for the three loci frequencies (4.3%) was statistically significant and its most important part corresponded to variation between Southern European vs. North African groups (3.4% variation between groups vs. 0.9% variation within groups). Looking at the loci, the most variable between-populations was the rs221639 ($F_{ST} = 11.5\%$) followed by rs314277 and rs7759938, with F_{ST} values of 4.3 and 1.3%, respectively. In almost all individual cases, the *F*-values were statistically significant except in the F_{SC} of rs314277 and rs7759938. The results of the Structure analysis (Supporting Information) were compatible with $K = 2$ ancestral components with unequal distribution in European vs. North African populations (*t*-test $P = 0.0002$). However, we did not find significant selection signals for the *LIN28B* region.

DISCUSSION

This study provides new population data for three markers on the *LIN28B* gene that have been repeatedly associated with adult phenotypes related to developmental timing in independent studies. Specifically, this is the first systematic population study of the SNPs rs221639, rs314277, and rs7759938, in a set of population groups covering the whole Mediterranean region.

Barrier results suggest that the main barrier to gene flow between African and European populations is located in the Saharan area, rather than in the Mediterranean Basin. Therefore, these results agree with recent studies indicating that the Mediterranean Sea is not a barrier to gene flow (Botigue et al., 2013) (Fig. 7).

From a population point of view, the analysis of the three SNPs reveals significant genetic heterogeneity between populations even at the micro-geographic level of the Mediterranean Basin. The population relationships provided by these markers are consistent with previous population genetics studies in the Mediterranean, which stress the North-to-South differentiation as the main genetic differences in this region (Athanasiadis & Moral, 2013; Athanasiadis et al., 2010; Bosch et al., 2001; Capelli et al., 2006; Comas et al., 2000; González-Pérez et al., 2010). This level of population differentiation is especially evident from the genetic distances (Fig. 6), the Structure analysis (Supporting Information) and the hierarchical AMOVA (Table 2). At first glance, this important genetic differentiation between the two Mediterranean shores could be related to natural selection. However, the negative results obtained from the selection analyses using 1000 Genomes data as well as those from previous studies (Simoni et al., 1999) suggest that natural selection does not explain the genetic differentiation found in the Mediterranean. Regarding the genetic distances between populations (Fig. 6), the presence of a single population as an outlier could be explained simply as an effect of genetic drift, due to the fact that we are using only a few markers.

In general, the population patterns of variation found in this study suggest that this genomic region might be useful for studying genetic variation of potential importance in human population genetics. In this way, it is worth mentioning the unexpected high between-population variation showed by the frequencies of the rs221639. Apart from the clear differentiation between North and South ($F_{CT} \gg F_{SC}$), the variance in allele frequencies in the whole area indicates a surprisingly high value of between-population variation ($F_{ST} = 11.5\%$). This value is higher than other values reported in the same populations (Henn et al., 2012), suggesting that this SNP is a good indicator for population differences in the Mediterranean region.

The strong population variation leads us to think that this SNP, rs221639, could be a good ancestry informative marker (AIM) in the context of the Mediterranean region. In fact, this marker shows an absolute allele frequency difference of 0.124, and an allele frequency variance between Northern and Southern groups of Fct = 10.1%. These values are similar to others identified for AIMs in European populations in other studies (Bauchet et al., 2007). However, further research is needed in order to reach a definitive conclusion about this variant. In this case, the significantly higher frequencies shown by groups from North Africa in comparison with the Southern Europe populations could likely be explained by a higher gene flow from Sub-Saharan African populations, whose rs221639(C) allele presents an average frequency of 0.41 (http://www.1000genomes.org/).

In conclusion, the three SNPs of the *LIN28B* gene examined appear to be useful tools for assessing genetic variation among the human populations of the Mediterranean region. The high between-population variation observed for the rs221639, a polymorphism not studied thus far, suggests that this marker is of potential utility for discriminating between the populations living in the northern and southern Mediterranean coasts. The population genetics results of this study suggest that checking for population variability of these same markers in other geographic areas and studying a higher number of SNPs in this genomic region may be of great interest in future research.

ACKNOWLEDGMENTS

The authors would like to thank all of the collaborators who provided samples for general populations: N. Harich and M. Kandil (Morocco), J.M. Dougoujon (Algeria), F. Cruciani (Italy), C. Calo (Sardinia), A. Kouvatsi (Greece), N. Moschonas (Crete), H. Chaabani (Tunisia and Libya), N. Pojskic (Bosnia), N. Bissar-Tadmouri (Turkey), J.

Santamaría (Valencia, Spain), and G. Pons-Monjo (Minorca, Spain). All volunteers are gratefully acknowledged for their sample donations. The authors declare that they have no conflict of interest.

LITERATURE CITED

Anneleen B, Andrew L, Gert VP, Daniel C, Belamy C, Sara DB, Vandesompele J, De Preter K, Shohet J, Speleman F. 2014. The MYCN/miR-26a-5p/LIN28B regulatory axis controls MYCN-driven LIN28B upregulation in neuroblastoma. ANR Cologne. Cologne, Germany: Uniklinik Köl.

Athanasiadis G, González-Pérez E, Esteban E, Dugoujon J-M, Stoneking M, Moral P. 2010. The Mediterranean Sea as a barrier to gene flow: evidence from variation in and around the F7 and F12 genomic regions. BMC Evol Biol 10:84.

Athanasiadis G, Moral P. 2013. Spatial principal component analysis points at global genetic structure in the Western Mediterranean. J Hum Genet 58:762–765.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD. 2007. Measuring European population stratification with microarray genotype data. Am J Hum Genet 80:948–956.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F. 1996. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire génome, populations, interactions, CNRS UMR 5000:1996–2004.

Borrego-Diaz E, Powers B, Azizov V, Lovell S, Reyes R, Chapman B, Tawfik O, McGregor D, Diaz FJ, Wang X, Veldhuizen PV. 2014. A potential regulatory loop between Lin28B:miR-212 in androgen-independent prostate cancer. Int J Oncol 45:2421–2429.

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J. 2001. High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68:1019–1029.

Botigue LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrerg H, Flores C, Bertranpetit J, Comas D, Bustamante CD. 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proceedings of the National Academy of Sciences 110:11791–11796.

Capelli C, Redhead N, Romano V, Calì F, Lefranc G, Delague V, Megarbane A, Felice AE, Pascali VL, Neophytou PI, Poulli Z, Novelletto A, Malaspina P, Terrenato L, Berebbi A, Fellous M, Thomas MG, Goldstein DB. 2006. Population structure in the Mediterranean basin: a Y chromosome perspective. Ann Hum Genet 70:207–225.

Carty CL, Spencer KL, Setiawan VW, Fernandez-Rhodes L, Malinowski J, Buyske S, Young A, Jorgensen NW, Cheng I, Carlson CS, Brown-Gentry K, Goodloe R, Park A, Parikh NI, Henderson B, Le Marchand L, Wactawski-Wende J, Fornage M, Matise TC, Hindorff LA, Arnold AM, Haiman CA, Franceschini N, Peters U, Crawford DC. 2013. Replication of genetic loci for ages at menarche and menopause in the multi-ethnic population architecture using genomics and epidemiology (PAGE) study. Hum Reprod 28:1695–1706.

Chong C, Lipeng B, Fengqi C, Liu Y, Junling X, Wei W, Qin S, Jian Y, Antao C, Rong X, Yunping L. 2014. Lin28B mediated IKK-β sustains the stemness of breast cancer stem cell via regulating Wnt/TCF4 and miR-34a/LEF1 signaling pathway. Cancer Res 74:3046.

Collaborative Group on Hormonal Factors in Breast Cancer. 2012. Menarche, menopause, and breast cancer risk : individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. Lancet Oncol 13:1141–1151.

Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, Stoneking M, Batzer MA, Bertranpetit J, Sajantila A. 2000. Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. Hum Genet 107: 312–319.

Cousminer DL, Berry DJ, Timpson NJ, Ang W, Thiering E, Byrne E, Taal HR, Huikari V, Bradfield JP, Kerkhof M, Groen-Blokhuis MM, Kreiner-Møller E, Marinelli M, Holst C, Leinonen JT, Perry JRB, Surakka I, Pietiläinen O, Kettunen J, Anttila V, Kaakinen M, Sovio U, Pouta A, Das S, Lagou V, Power C, Prokopenko I, Evans DM, Kemp JP, St Pourcain B, Ring S, Palotie A, Kajantie E, Osmond C, Lehtimäki T, Viikari JS, Kähönen M, Warrington NM, Lye SJ, Palmer LJ, Tiesler CMT, Flexeder C, Montgomery GW, Medland SE, Hofman A, Hakonarson H, Guxens M, Bartels M, Salomaa V, The ReproGen Consortium , Murabito J, Kaprio J, Sørensen TIA, Ballester F, Bisgaard H, Boomsma DI, Koppelman GH, Grant SFA, Jaddoe VWV, Martin NG, Heinrich J, Pennell CE, Raitakari O, Eriksson JG, Smith GD, Hypp—nen E, Järvelin M, McCarthy MI, Ripatti S, WidÖn E, for the Early Growth Genetics (EGG) Consortium. 2013. Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth, pubertal timing, and childhood adiposity. Hum Mol Genet 22:2735–2747.

Day FR, Elks CE, Murray A, Ong KK, Perry JRB. 2015. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women : the UK Biobank study, Scientific reports 5:11208. doi:10.1038/srep11208.

Elks CE, Ong KK, Scott RA, van der Schouw YT, Brand JS, Wark PA, Amiano P, Balkau B, Barricarte A, Boeing H, Fonseca-Nunes A, Franks P, Grioni S, Halkjaer J, Kaaks R, Key T, Khaw K, Mattiello A, Nilsson P, Overvad K, Palli D, Quirœs J, Rinaldi S, Rolandsson O, Romieu I, Sacerdote C, Sanchez M, Spijkerman A, Tjonneland A, Tormo M, Tumino R, Vandera D, Forouhi N, Sharp S, Langenberg C, Riboli E, Wareham N,, The Interact Consortium. 2013. Age at menarche and Type 2 diabetes risk. Diabetes Care 36:3526–3534.

Excoffier L, Hofer T, Foll M. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298.

Excoffier L, Lischer H. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574–578.

Felsenstein J. 1989. PHYLIP-Phylogeny inference package (Version 3.2). Cladistics 5:164–166.

González-Pérez E, Esteban E, Via M, Gayà-Vidal M, Athanasiadis G, Dugoujon JM, Luna F, Mesa MS, Fuster V, Kandil M, Harich N, Bissar-Tadmouri N, Saetta A, Moral P. 2010. Population relationships in the Mediterranean revealed by autosomal genetic data (Alu and Alu/STR compound systems). Am J Phys Anthropol 141:430–439.

Guo Y, Chen Y, Ito H, Watanabe A, Ge X, Kodama T, Aburatani H. 2006. Identification and characterization of lin-28 homolog B (LIN28B) in human hepatocellular carcinoma. Gene 384:51–61.

Hamano R, Miyata H, Yamasaki M, Sugimura K, Tanaka K, Kurokawa Y, Nakajima K, Takiguchi S, Fujiwara Y, Mori M, Doki Y. 2012. High expression of Lin28 is associated with tumour aggressiveness and poor prognosis of patients in oesophagus cancer. Br J Cancer 106:1415–1423.

He C, Kraft P, Chen C, Buring JE, Pare G, Hankinson SE, Chanock SJ, Ridker PM, Hunter DJ, Chasman DI. 2009. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. Nat Genet 41:724–728.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, Bustamante CD, Comas D. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet 8:e1002397.

Hetty H, Tim L, Farzaneh G, Caye A, Hélène C, Christian F, Niemeyer C, Stary J, Bresolin S, Masetti R. 2014. LIN28B defines an aggressive subtype of juvenile myelomonocytic leukemia. Haematologica 99:235–236.

Hu Q, Peng J, Liu W, He X, Cui L, Chen X, Yang M. 2014. Lin28B is a novel prognostic marker in gastric adenocarcinoma. Int J Clin Exp Pathol 7:5083–5092.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9:1322–1332.

Jacobsen BK, Heuch I, Kva G. 2007. Association of low age at menarche with increased all-cause mortality : A 37-year follow-up of 61,319 Norwegian women. Am J Epidemiol 166:1431–1437.

Johnson W, Choh AC, Curran JE, Czerwinski SA, Bellis C, Dyer TD, Blangero J, Towne B, Demerath EW. 2013. Genetic risk for earlier menarche also influences peripubertal body mass index. Am J Phys Anthropol 20:10–20.

Keane M, De Magalhães JP. 2013. MYCN/LIN28B/Let-7/HMGA2 pathway implicated by meta-analysis of GWAS in suppression of post-natal proliferation thereby potentially contributing to aging. Mech Ageing Dev 134: 346–348.

Lawrance-Owen AJ, Bargary G, Bosten JM, Goodbourn PT, Hogg RE, Mollon JD. 2013. Genetic association suggests that SMOC1 mediates between prenatal sex hormones and digit ratio. Hum Genet 132: 415–421.

Lu L, Katsaros D, Mayne ST, Risch H a, Benedetto C, Canuto EM, Yu H. 2012. Functional study of risk loci of stem cell-associated gene lin-28B and associations with disease survival outcomes in epithelial ovarian cancer. Carcinogenesis 33:2119–2125.

Manni F, Guerard E, Heyer E. 2004. Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by using Monmonier's algorithm. Hum Biol 76:173–190.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1092 human genomes. Nature 491:56–65.

Medland SE, Zayats T, Glaser B, Nyholt DR, Gordon SD, Wright MJ, Montgomery GW, Campbell MJ, Henders AK, Timpson NJ, Peltonen L, Wolke D, Ring SM, Deloukas P, Martin NG, Smith GD, Evans DM. 2010. A variant in LIN28B is associated with 2D : 4D finger-length ratio, a putative retrospective biomarker of prenatal testosterone exposure. Am J Hum Genet 86:519–525.

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590.

Ong KK, Elks CE, Li S, Zhao JH, Luan J, Andersen LB, Bingham SA, Brage S, Davey Smith G, Ekelund U, Gillson CJ, Glaser B, Golding J, Hardy R, Khaw K, Kuh D, Luben R, Marcus M, McGeehin MA, Ness AR, Northstone K, Ring SM, Rubin C, Sims MA, Song K, Strachan DP, Vollenweider P, Waeber G, Waterworth DM, Wong A, Deloukas P, Barroso I, Mooser V, Loos RJ, Wareham NJ. 2009. Genetic variation in LIN28B is associated with the timing of puberty. Nat Genet 41:729–733.

Ong KK, Elks CE, Wills AK, Wong A, Wareham NJ, Loos RJF, Kuh D, Hardy R. 2011. Associations between the pubertal timing-related variant in LIN28B and BMI vary across the life course. J Clin Endocrinol Metabol 96:125–129.

Onland-Moret NC. 2005. Age at menarche in relation to adult height: The EPIC study. Am J Epidemiol 162:623–632.

Pang M, Wu G, Hou X, Hou N, Liang L, Jia G, Shuai P, Luo Bin, Wang K, Li G. 2014. LIN28B promotes colon cancer migration and recurrence. PloS one 9:e109169.

Perry J, Day F, Elks C, Sulem P, Thompson D, Ferreira T, He C, Chasman DI, Esko T, Thorleifsson G, Albrecht E, Ang WQ, Corre T, Cousminer DL, Feenstra B, Franceschini N, Ganna A, Johnson AD, Kjellqvist S, Lunetta KL, McMahon G, Nolte IM, Paternoster L, Porcu E, Smith AV, Stolk L, Teumer A, Ternikova N, Tikkanen E, Ulivi S, Wagner EK, Amin N, Bierut LJ, Byrne EM, Hottenga JJ, Koller DL, Mangino M, Pers TH, Yerges-Armstrong LM, Hua Zhao J, Andrulis IL, Anton-Culver H, Atsma F, Bandinelli S, Beckmann MW, Benitez J, Blomqvist C, Bojesen SE, Bolla MK, Bonanni B, Brauch H, Brenner H, Buring JE, Chang-Claude J, Chanock S, Chen J, Chenevix-Trench G, Collée JM, Couch FJ, Couper D, Coviello AD, Cox A, Czene K, D'adamo AP, Davey Smith G, De Vivo I, Demerath EW, Dennis J, Devilee P, Dieffenbach AK, Dunning AM, Eiriksdottir G, Eriksson JG, Fasching PA, Ferrucci L, Flesch-Janys D, Flyger H, Foroud T, Franke L, Garcia ME, García-Closas M, Geller F, de Geus EE, Giles GG, Gudbjartsson DF, Gudnason V, Guénel P, Guo S, Hall P, Hamann U, Haring R, Hartman CA, Heath AC, Hofman A, Hooning MJ, Hopper JL, Hu FB, Hunter DJ, Karasik D, Kiel DP, Knight JA, Kosma VM, Kutalik Z, Lai S, Lambrechts D, Lindblom A, Mägi R, Magnusson PK, Mannermaa A, Martin NG, Masson G, McArdle PF, McArdle WL, Melbye M, Michailidou K, Mihailov E, Milani L, Milne RL, Nevanlinna H, Neven P, Nohr EA, Oldehinkel AJ, Oostra BA, Palotie A, Peacock M, Pedersen NL, Peterlongo P, Peto J, Pharoah PD, Postma DS, Pouta A, Pylkäs K, Radice P, Ring S, Rivadeneira F, Robino A, Rose LM, Rudolph A, Salomaa V, Sanna S, Schlessinger D, Schmidt MK, Southey MC, Sovio U, Stampfer MJ, Stöckl D, Storniolo AM, Timpson NJ, Tyrer J, Visser JA, Vollenweider P, Völzke H, Waeber G, Waldenberger M, Wallaschofski H, Wang Q, Willemsen G, Winqvist R, Wolffenbuttel BH, Wright MJ, Australian Ovarian Cancer Study; GENICA Network; kConFab; LifeLines Cohort Study; InterAct Consortium; Early Growth Genetics (EGG) Consortium, Boomsma DI, Econs MJ, Khaw KT, Loos RJ, McCarthy MI, Montgomery GW, Rice JP, Streeten EA, Thorsteinsdottir U, van Duijn CM, Alizadeh BZ, Bergmann S, Boerwinkle E, Boyd HA, Crisponi L, Gasparini P, Gieger C, Harris TB, Ingelsson E, Jørvelin MR, Kraft P, Lawlor D, Metspalu A, Pennell CE, Ridker PM, Snieder H, S½rensen TI, Spector TD, Strachan DP, Uitterlinden AG, Wareham NJ, Widen E, Zygmunt M, Murray A, Easton DF, Stefansson K, Murabito JM, Ong KK. 2015. Parent-of-origin specific allelic associations among 106 genomic loci for age at menarche. Nature 514:92–97.

Perry JRB, Stolk L, Franceschini N, Lunetta KL, Zhai G, Mcardle PF, Smith AV, Aspelund T, Bandinelli S, Boerwinkle E, Cherkas L, Eiriksdottir G, Estrada K, Ferrucci L, Folsom AR, Garcia M, Gudnason V, Hofman A, Karasik D, Kiel DP, Launer LJ, van Meurs J, Nalls MA, Rivadeneira F, Shuldiner AR, Singleton A, Soranzo N, Tanaka T, Visser JA, Weedon MN, Wilson SG, Zhuang V, Streeten EA, Harris TB, Murray A, Spector TD, Demerath EW, Uitterlinden AG, Murabito JM. 2009. Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. Nat Genet 41:648–650.

Piras IS, De Montis A, Calò CM, Marini M, Atzori M, Corrias L, Sazzini M, Boattini A, Vona G, Contu L. 2012. Genome-wide scan with nearly 700,000 SNPs in two Sardinian sub-populations suggests some regions as candidate targets for positive selection. Eur J Hum Genet 20:1155–1161.

Piskounova E, Polytarchou C, Thornton JE, Lapierre RJ, Pothoulakis C, Hagan JP, Iliopoulos D, Gregory RI. 2011. Lin28A and Lin28B inhibit let-7 MicroRNA biogenesis by distinct mechanisms. Cell 147:1066–1079.

Pritchard J, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pybus M, Olio GMD, Luisi P, Uzkudun M, Pavlidis P, Laayouni H, Bertranpetit J, Carren A, Engelken J. 2014. 1000 Genomes Selection Browser 1.0 : a genome browser dedicated to signatures of natural selection in modern humans. Nucl Acids Res 42:903–909.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raymond M, Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86:248–249.

Rybak A, Fuchs H, Smirnova L, Brandt C, Pohl EE, Nitsch R, Wulczyn FG. 2008. A feedback loop comprising lin-28 and let-7 controls pre- let-7 maturation during neural stem-cell commitment. Nat Cell Biol 10: 987–993.

Powers JT, Sacco A, Reagan MR, Moschetta M, Glavey S, Maiso P, Sahin I, Mishima Y, Aljaway Y, Leleu X, Roccaro AM, Daley GQ, Ghobrial IM. 2013. Lin28B/Let-7 axis regulates multiple myeloma proliferation by enhancing c-Myc and Ras survival pathways. Blood 122:273.

Shiyun X, Zhang W, Manley N. 2014. Premature thymic involution impairs let7 up-regulation in the transition from fetal to adult hematopoietic stem cells in the thymus (HEM4P.236). J Immunol 192:116.

Simoni L, Gueresi P, Pettener D, Barbujani G. 1999. Patterns of gene flow inferred from genetic distances in the Mediterranean region. Hum Biol 71:399–415.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Sulem P, Gudbjartsson DF, Rafnar T, Holm H, Olafsdottir EJ, Olafsdottir GH, Jonsson T, Alexandersen P, Feenstra B, Boyd HA, Aben KK, Verbeek AL, Roeleveld N, Jonasdottir A, Styrkarsdottir U, Steinthorsdottir V, Karason A, Stacey SN, Gudmundsson J, Jakobsdottir M, Thorleifsson G, Hardarson G, Gulcher J, Kong A, Kiemeney LA, Melbye M, Christiansen C, Tryggvadottir L, Thorsteinsdottir U, Stefansson K. 2009. Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche. Nat Genet 41:734–738.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.

Tu W, Wagner EK, Eckert GJ, Yu Z, Hannon T, Pratt JH, He C. 2015. Associations between menarche-related genetic variants and pubertal growth in male and female adolescents. J Adolescent Health 56:66–72.

Viswanathan SR, Powers JT, Einhorn W, Hoshida Y, Ng TL, Toffanin S, O'Sullivan M, Lu J, Phillips LA, Lockhart VL, Shah SP, Tanwar PS, Mermel CH, Beroukhim R, Azam M, Teixeira J, Meyerson M, Hughes TP, Llovet JM, Radich J, Mullighan CG, Golub TR, Sorensen PH, Daley GQ. 2009. Lin28 promotes transformation and is associated with advanced human malignancies. Nat Genet 41:843–848.

Wen J, Liu H, Wang Q, Liu Z, Li Y, Xiong H, Xu T, Li P, Wang LE, Gomez DR, Mohan R, Komaki R, Liao Z, Wei Q. 2014. Genetic variants of the LIN28B gene predict severe radiation pneumonitis in patients with non-small cell lung cancer treated with definitive radiation therapy. Eur J Cancer 50:1706–1716.

Widen E, Ripatti S, Cousminer DL, Surakka I, Lappalainen T, Ja M, Hirschhorn JN, Peltonen L. 2010. Distinct variants at LIN28B influence growth in height from birth to adulthood. Am J Hum Genet 86:773–782.

Wu T, Jia J, Xiong X, He H, Bu L, Zhao Z, Huang C, Zhang W. 2013. Increased expression of Lin28B associates with poor prognosis in patients with oral squamous cell carcinoma. PLoS ONE 8:e83869.

Ye Y, Madison B, Wu X, Rustgi AK. 2012. A LIN28B polymorphism predicts for colon cancer survival. Cancer Biol Ther 13:1390–1395.

You X, Liu F, Zhang T, Lv N, Liu Q, Shan C, Du Y, Kong G, Wang T, Ye L, Zhang X. 2014. Hepatitis B virus X protein upregulates Lin28A/Lin28B through Sp-1/c-Myc to enhance the proliferation of hepatoma cells. Oncogene 33:449–460.

Article II

Álvarez-Álvarez et al., 2017

A survey of sub-Saharan gene flow into the Mediterranean at risk loci for coronary artery disease

Miguel M. Álvarez-Álvarez, Daniela Zanetti, Robert Carreras-Torres, Pedro Moral and Georgios Athanasiadis

European Journal of Human Genetics 2017; 25:472–6; doi:10.1038/ejhg.2016.200

Resumen en castellano

Este estudio evalúa la presencia de señales de flujo génico de origen subsahariano en poblaciones del Mediterráneo, concretamente en cuatro regiones genómicas previamente asociadas a la enfermedad de las arterias coronarias.

Un total de 366 SNPs fueron genotipados en 772 individuos de 17 poblaciones mediterráneas. Además, se incluyeron en los análisis las poblaciones Yoruba (YRI) y Han (CHB) del Proyecto 1000 Genomas. Los análises de estructuración poblacional identifican un notorio componente subsahariano en las muestras Mediterráneas estudiadas. Este componente presenta valores más elevados en las poblaciones del norte de África, las cuales además muestran una mayor proporción de haplotipos subsaharianos. Ello es posiblemente atribuible a unos mayores grados de flujo génico subsahariano en el norte de África que en el sur de Europa. Análisis complementarios empleando *D*-statistics sugieren una posible introgresión de material genético subsahariano en una de las regiones genómicas estudiadas (10q11).

European Journal of Human Genetics (2017) 25, 472–476

www.nature.com/ejhg

ARTICLE

A survey of sub-Saharan gene flow into the Mediterranean at risk loci for coronary artery disease

Miguel M Álvarez-Álvarez[1], Daniela Zanetti[1], Robert Carreras-Torres[1], Pedro Moral[1,3] and Georgios Athanasiadis*[,2,3]

This study tries to find detectable signals of gene flow of Sub-Saharan origin into the Mediterranean in four genomic regions previously associated with coronary artery disease. A total of 366 single-nucleotide polymorphisms were genotyped in 772 individuals from 10 Mediterranean countries. Population structure analyses were performed, in which a noticeable Sub-Saharan component was found in the studied samples. The overall percentage of this Sub-Saharan component presents differences between the two Mediterranean coasts. D-statistics suggest possible Sub-Saharan introgression into one of the studied genomic regions (10q11). We also found differences in linkage disequilibrium patterns between the two Mediterranean coasts, possibly attributable to differential Sub-Saharan admixture. Our results confirm the potentially important role of human demographic history when performing epidemiological studies.
European Journal of Human Genetics (2017) **25,** 472–476; doi:10.1038/ejhg.2016.200; published online 18 January 2017

INTRODUCTION

Cardiovascular diseases are the leading cause of morbidity and mortality in Western societies.[1] According to the World Health Organisation, 17.3 million people died of cardiovascular diseases in 2008, representing 30% of all deaths worldwide (http://www.who.int/cardiovascular_diseases/about_cvd/en/). One of the most common types of cardiovascular disease is coronary artery disease (CAD), which has increased by 44% during the last 20 years in North Africa and the Middle East (http://www.who.int/whr/2013/report/en/), and is manifested as stable and unstable angina, myocardial infarction or sudden cardiac death. In all these heart conditions, genetic and environmental factors interact to determine the clinical phenotype.[2] Recently, several genetic variants have been robustly associated with CAD through genome-wide association studies (GWASs).[3]

Genotype and phenotype variation can present geographic patterns, as is for instance the case of height across Europe.[4] Such patterns can be shaped by selection and/or various demographic phenomena (population expansions, subdivisions, gene flow and/or bottlenecks). As an example, a recent study reported that genetic differences in CAD risk among worldwide populations are due to demographic processes.[5] In some occasions, allele frequencies at specific genomic regions increase through introgression from other populations with which there was admixing. These processes can also affect linkage disequilibrium (LD) patterns across populations. As a result, allele frequencies and LD patterns in introgressed regions for a given population tend to be more similar to distantly related populations than to others otherwise closer. An example of this is the Neanderthal introgression into Eurasian populations.[6]

In this study, we carried out a fine-scale genetic analysis on a broad collection of European populations using variation from four genomic regions putatively or provenly[2] associated with CAD risk. These four regions have shown disparity in the association of specific markers with CAD risk between Southern European and North African samples.[7] To see if this disparity could be partially attributed to introgression, we looked for signals of gene flow of Sub-Saharan origin into the Mediterranean at the four CAD-related genomic regions mentioned above.

MATERIALS AND METHODS

Populations

Our analysis was based on genetic data from 772 individuals of Southern European and North African ancestry. The studied populations encompass 10 Mediterranean countries: Spain (Basque Country, Girona, Valencia and Las Alpujarras); France (Toulouse); Italy (Sardinia and Sicily); Bosnia-Herzegovina; Greece (Crete); Turkey; Morocco (Khenifra and Chouala); Tunisia; Algeria (M'zab); Libya; and Jordan (Bedouin Jordanians and non-Bedouin Jordanians).[7,8] In addition, we considered the Yoruba (YRI) and Han Chinese (CHB) populations from the phase III 1000 Genomes Project in order to have genetic representation of Sub-Saharan Africa and East Asia. The geographic location of the samples is shown in Figure 1.

Genetic data

Samples were genotyped for a total of 366 single-nucleotide polymorphisms (SNPs) distributed along four CAD-associated chromosomal regions: 1p13.3, 1q41, 9p21 and 10q11[2] (Table 1). Genotyping was carried out with the Custom GoldenGate Panel (Illumina Inc., San Diego, CA, USA) genotyping platform. SNPs were previously selected as representatives of common variation in each of the four genomic regions according to the following criteria: (i) coverage of one SNP every ~1.5 Kb; (ii) minor allele frequency (MAF) > 0.05 in European populations; (iii) avoiding markers in strong LD in European populations ($r^2 > 0.8$); and (iv) giving priority to tag SNPs as well as markers previously reported to be associated with CAD.[7] All of the participants signed an informed consent form before sample donation, and the Ethics Committee of the University of Barcelona approved the study. Individual genotypes and SNP

[1]Faculty of Biology, Department of Evolutionary Biology, Ecology and Environmental Sciences, Biodiversity Research Institute, University of Barcelona, Barcelona, Spain; [2]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
[3]These authors contributed equally to this work.
*Correspondence: Dr G Athanasiadis, Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé, Building 1110, Aarhus 8000, Denmark.
Tel: +45 87155569; Fax: +45 87154102; E-mail: athanasiadis@birc.au.dk
Received 21 July 2016; revised 24 November 2016; accepted 14 December 2016; published online 18 January 2017

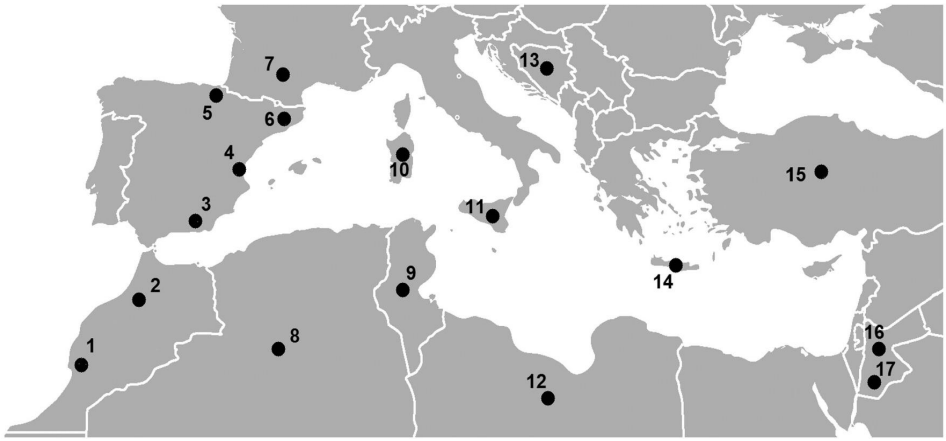


Figure 1 Populations studied. 1, Chouala; 2, Khenifra; 3, Las Alpujarras; 4, Valencia; 5, Basque Country; 6, Girona; 7, Toulouse; 8, M'zab; 9, Tunisia; 10, Sardinia; 11, Sicily; 12, Libya; 13, Bosnia-Herzegovina; 14, Crete; 15, Turkey; 16, General Jordan; 17, Bedouin.

Table 1 Chromosomal regions studied

Chromosomal region	*Number of SNPs*	*Span (kbp)*	*Known genes*
1p13.3	59	150	*CELSR2, PSRC1, MYBPHL, SORT1*
1q41	37	100	*TAF1A, MIA3, AIDA*
9p21	155	300	*CDKN2A, CDKN2B*
10q11	115	200	*CXCL12*

Abbreviation: SNP, single-nucleotide polymorphism.

information is available through the EMBL-EBI ArrayExpress public repository (accession number: E-MTAB-5265).

Statistical analysis

Standard quality control was performed with Plink v1.9.[9] Only one marker was excluded for not fitting Hardy–Weinberg proportions (P-value $<10^{-5}$). Per-individual and per-locus genotype missingness was zero, whereas none of the SNPs was monomorphic. In all subsequent analyses, we used all Mediterranean populations together with YRI and CHB. We first explored population structure in our samples with PCA and ancestry component analysis, using Plink and Admixture v1.3.0,[10] respectively.

We further used *qpdstat* from AdmixTools[11] to calculate D-statistics. We applied *qpdstat* to each of the four genomic regions separately, setting jackknife block size to 0.02–0.06 cm depending on the number of markers in each genomic region. We set block size to a smaller than the default value in order to compensate for the relatively low number of SNPs analysed. Each Mediterranean population was compared with CHB, YRI, and a dummy individual representing a hypothetical outgroup. This dummy individual was homozygous for the ancestral allele for each of the studied SNPs as defined by comparison with six primate species. The Z-scores reported by AdmixTools indicate whether there has been gene flow between two out of four given populations when one of these populations is an outgroup (in our case the dummy individual). In particular, for the topology (((X, CHB); YRI); Outgroup) shown in Figure 2a, where X is a given Mediterranean population, each of the four genomic regions is tested to define whether allele frequencies are more similar between X and CHB (which would denote lack of Sub-Saharan gene flow) or between X and YRI (which could be interpreted as suggestive of Sub-Saharan gene flow). Positive Z-scores indicate more similarity of a Mediterranean population with YRI, whereas negative Z-scores indicate more similarity of

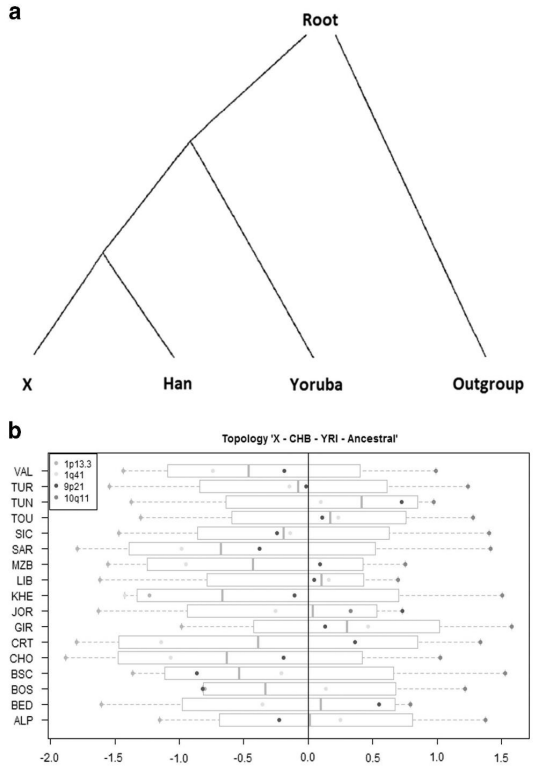


Figure 2 (**a**) Topology assumed for the D-statistics, where X represents a Mediterranean population. (**b**) Boxplots showing the results of the D-statistics. The points represent the Z-score values (abscissas axis) obtained for each genomic region in a given Mediterranean population X (ordinates axis). As the topology assumed is (((X, CHB), YRI), Ancestral), significant Z-score-positive values indicate a gene flow event between YRI and X. The significance threshold is set at ± 2. See legend at Figure 3 for clarification of the population names.

CHB with YRI. We set a threshold of |Z-score| > 2 necessary for a D-statistic to be significant.

In order to check whether there were significant differences in the *Z*-scores between North African and Southern European populations for each genomic region, we ran a Mann–Whitney rank sum test. Furthermore, for the genomic regions for which the Mann–Whitney test was significant ($P < 0.0125$ after Bonferroni correction for multiple testing), populations were further grouped into four geographic categories (Southwest Europe, Southeast Europe, Northwest Africa and Northeast Africa) and a follow-up Kruskal–Wallis test was carried out to see if the structuring pattern could be further explained by an East–West axis.

We also analysed the differences in LD patterns between the Southern European and North African groups using varLD v1.0[12] with 1000 iterations per test. Finally, we obtained phased haplotypes with SHAPEIT[13] for nine LD blocks that we identified with Haploview[14] and explored haplotype sharing between Southern Europe, North Africa and Sub-Saharan groups with Venn diagrams.

RESULTS

Population structure

Figure 3a shows the centroid coordinates for each population along the first two principal components. We can see a North–South separation along PC1, with PC2 defining an East–West gradient. In general, populations were visually separated in four groups (North African, Southern European, Sub-Saharan and East Asian) with the Mediterranean populations naturally showing the closest proximity. The centroids match roughly the geography – with the exception of the Basques who appear as an outlier in the Southern European cluster.

Regarding the Admixture analysis, due to the lack of an optimal cross-validation value for the tested numbers of ancestral components ($K = \{1, 2, \ldots, 30\}$; Supplementary Information 1), we focused our analysis on $K = 3$ ancestral components, roughly matching the three major geographic regions of our study: Mediterranean, East Asia and Sub-Saharan Africa. Figure 3b shows the mixture profiles of the studied populations. We observed a component that was predominant in Sub-Saharan Africa (green), as well as two other components: one that was present in both the Mediterranean and the Asian samples at varying proportions (blue), and one that was present in all samples (red). Due to the small number of SNPs used, these two components were not representative of continent-level geographic regions. The Mediterranean samples show rather homogeneous proportions for all three

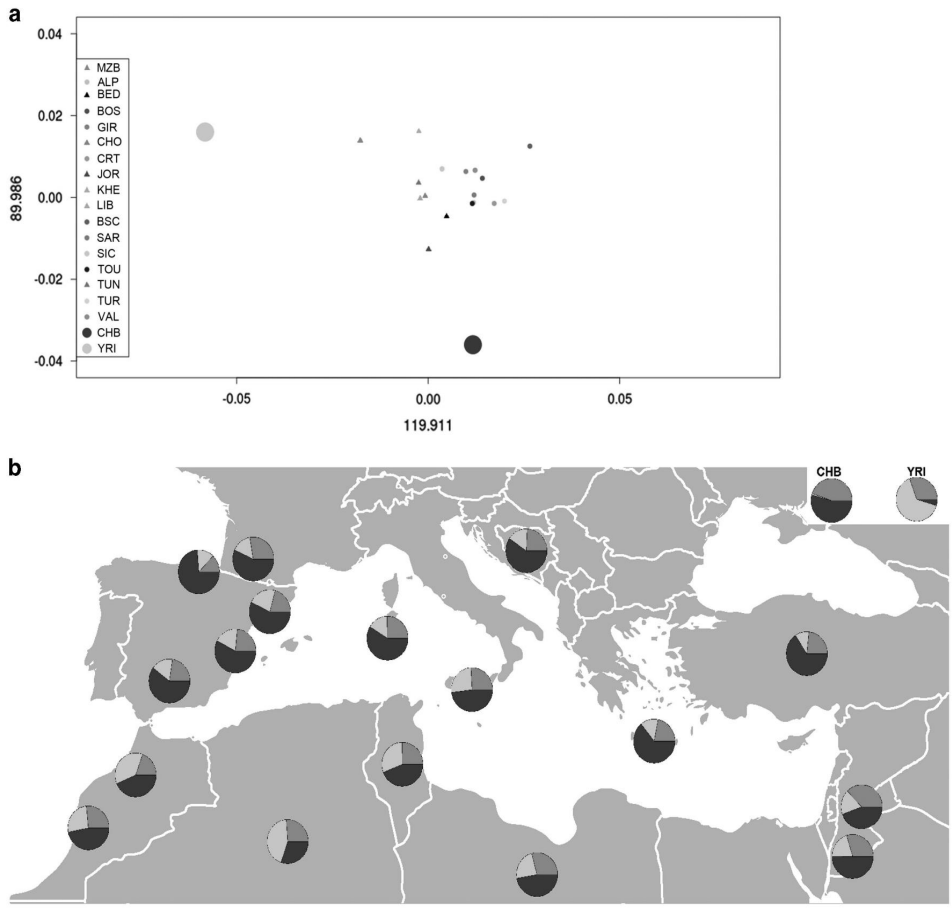


Figure 3 (**a**) PCA based on allele frequencies. Big green and red points are the centroids of YRI and CHB populations, respectively. Triangles represent the centroids of North African populations, whereas small circles represent the centroids of the Southern European populations. (**b**) *Admixture* analysis for $K = 3$ ancestral components represented as pie charts. The pie chart of each population is located on its geographic position. ALP, Las Alpujarras; BED, Bedouin; BOS, Bosnia-Herzegovina; BSC, Basque Country; CHB, Han Chinese from Beijing; CHO, Chouala; CRT, Crete; GIR, Girona; JOR, Jordan; KHE, Khenifra; LIB, Libya; MZB, M'zab;SAR, Sardinia; SIC, Sicily; TOU, Toulouse; TUN, Tunisia; TUR, Turkey; VAL, Valencia; YRI, Yoruba. A full colour version of this figure is available at the *European Journal of Human Genetics* journal online.

Table 2 ***P*-values for the D-statistic comparisons between the two Mediterranean coasts**

	1p13.3	*1q41*	*9p21*	*10q11*
Southern Europe *vs* North Africa	0.16	0.23	0.06	**0.01**
SW Europe *vs* SE Europe *vs* NW Africa *vs* NE Africa				*0.03*
SW Europe *vs* SE Europe				0.07
SW Europe *vs* NW Africa				0.38
SW Europe *vs* NE Africa				*0.02*
SE Europe *vs* NW Africa				0.86
SE Europe *vs* NE Africa				0.06
NW Africa *vs* NE Africa				0.20

Abbreviations: NE, North East; NW, North West; SE, South East; SW, South West. Bold number denotes values below significance (*P*-value <0.0125 for first two tests, *P*-value <0.0083 for the last two tests); numbers in italics denote nominal significance.

components, with slightly higher rates of Sub-Saharan ancestry in the North African populations. For visual comparisons, we provide the Admixture plots for $K=\{2, 3, 4, 5\}$ in Supplementary Information 2.

Signals of Sub-Saharan gene flow

The virtually indistinguishable mixture patterns in all Mediterranean populations motivated the search for signals of differential Sub-Saharan gene flow with more sensitive methods – in particular AdmixTools (Figure 2b). Despite the effort, we did not obtain consistent positive *Z*-scores for either 1q41 or 9p21, and we even found suggestively negative values for 1p13.3, contrasting a possible gene flow from YRI to the Mediterranean samples for these three regions. However, *Z*-scores were consistently positive and close to the significance threshold ($|Z\text{-score}|>2$) for the 10q11 region. Moreover, Mann–Whitney comparisons showed significant differences in *Z*-scores between Southern Europe and North Africa (Table 2), which can be refined in some cases including a West–East differentiation grouping, namely Southwest Europe and Northeast Africa (Mann–Whitney, $P=0.02$).

Differences in LD patterns

All comparisons of LD patterns among the four studied genomic regions were significant ($P<0.01$). Pairwise score matrices are reported in the Supplementary Information 3–6. The highest LD pattern differences occurred at the 9p21 genomic region, followed by those at the 10q11 genomic region. This could be reflecting the differences in the degree of Sub-Saharan gene flow between the two Mediterranean coasts also seen in the structure analyses, as higher levels of admixture between two relatively non-admixed populations result in higher levels of LD in the admixed individuals.[15]

Haplotype sharing

We identified a total of nine LD blocks in the four studied genomic regions, which correspond to a total 2372 unique haplotypes. Comparisons between Southern European, North African and Sub-Saharan African groups showed that there is generally higher haplotype sharing between North Africa and Sub-Saharan Africa than between Southern Europe and Sub-Saharan Africa, providing additional independent evidence for greater Sub-Saharan gene flow towards the former (Figure 4).

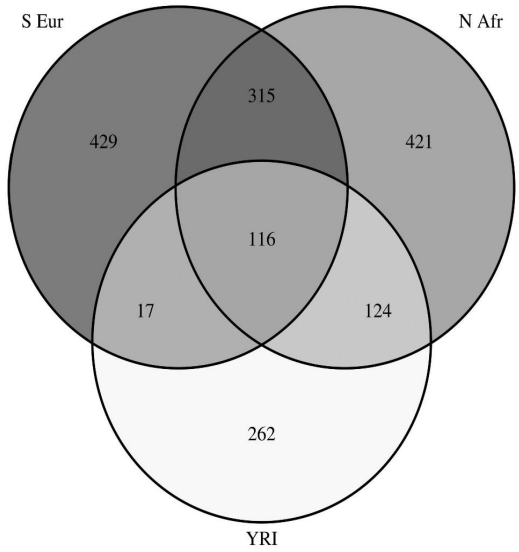


Figure 4 Venn diagram containing the haplotypes from all the LD blocks identified in the four genomic regions studied. The population groups are Southern Europe, North Africa and Sub-Saharan Africa (Yoruba).

DISCUSSION

This study provides insights into how demographic events that are stochastic by nature (such as introgression) have the potential of affecting the differential geographic distribution of variants associated with common diseases. Specifically, through the analysis of genotype data from top CAD risk loci, we suggest that gene flow of Sub-Saharan origin may have played a role in the current geographic distribution of variants associated with CAD.

Our Mediterranean samples presented a considerable proportion of Sub-Saharan admixture, suggesting introgression in at least some of the four genomic regions. The Sub-Saharan component was noticeably more prevalent in North Africa than in Southern Europe, a fact also reflected in the elevated proportion of shared haplotypes of YRI with North Africa than with Southern Europe (Supplementary Information 7). These results are in accord with previous studies suggesting a more intense Sub-Saharan gene flow into North Africa than into Southern Europe due to geographic proximity and/or the potential role of the Mediterranean sea as a genetic barrier.[8,16–21] Among the European populations, the ones that present a higher Sub-Saharan proportion are Sicily, Girona, Valencia and Andalusia (26%, 22%, 19% and 17%, respectively). This observation matches historical of North African influence in these regions during the Middle Ages.[22] Likewise, the LD differences found between the two Mediterranean coasts match the observation of North African populations presenting higher levels of Sub-Saharan admixture.

The results from the D-statistics together with those from Admixture also suggest a potential gene flow of Sub-Saharan origin into the Mediterranean at least for the 10q11 region. This region includes *CXCL12*, a gene that codes for a chemokine ligand linked to cardiovascular disease with protective effects.[23] It is also worth noting that potential signals of balancing selection have been identified at 10q11,[8] implying that natural selection could have maintained this signal of Sub-Saharan gene flow in this genomic region by favouring admixed individuals against cardiovascular disease. Further research is warranted to shed more light on this hypothesis.

It is important to note that the relatively small size of the studied chromosomal regions and the low number of markers pose

a limitation to the robustness of the results obtained, with potential bias also due to fact that the SNP used in the analyses was ascertained primarily in European populations. In addition, the studied loci are not the only ones associated with CAD and therefore an extension to more disease-associated loci and samples would be desirable. The small number of SNPs is probably also the reason behind the lack of an optimal cross-validation value in the Admixture analysis, warranting caution at interpreting the obtained results. Finally, given that none of the *Z*-scores for 10q11 passed the established significance threshold, though very close to it, our results are subject to be false positives. Future work could address efficiently the above issues by analysing a higher number of regions and markers.

Our results build on the notion of a differential gene flow from Sub-Saharan Africa into North Africa that, according to recent studies, could have taken place 750–1200 years ago during the trans-Saharan slave trade.[24] Sub-Saharan introgression into Europe could have been the result of indirect contact of Europeans with North Africans already admixed with Sub-Saharan populations.[25,26] This two-step Sub-Saharan introgression into Europe could be an interesting subject of future research validated through demographic simulations.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported by the Ministerio de Ciencia e Innovación CGL2011-27866 project. We would like to thank all collaborators who provided samples of general populations: N Harich and M Kandil (Morocco), H Chabaani (Tunisia and Libya), M Sadiq (Jordan), JM Dougoujon (France and Algeria), C Calò (Sardinia), N Pojskic (Bosnia), N Moschonas (Crete, Greece), N Bissar-Tadmouri (Turkey), J Santamaría (Valencia, Spain) and F Luna (Las Alpujarras, Spain). All volunteers are gratefully acknowledged for sample donation.

1 Wilson PWF, Agostino RBD, Levy D, Belanger AM, Silbershatz H, Kannel WB: Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**: 1837–1847.
2 Girelli D, Martinelli N, Peyvandi F, Olivieri O: Genetic architecture of coronary artery disease in the genome-wide era: implications for the emerging 'golden dozen' loci. *Semin Thromb Hemost* 2009; **35**: 671–682.
3 Lieb W, Vasan RS: Genetics of coronary artery disease. *Circulation* 2013; **128**: 1131–1138.
4 Robinson MR, Hemani G, Medina-gomez C *et al*: Population genetic differentiation of height and body mass index across Europe. *Nat Genet* 2015; **47**: 1357–1362.
5 Corona E, Dudley JT, Butte AJ: Extreme evolutionary disparities seen in positive selection across seven complex diseases. *PLoS One* 2010; **5**: 1–10.
6 Green RE, Krause J, Briggs AW *et al*: A draft sequence of the Neandertal genome. *Science* 2010; **328**: 710–722.
7 Zanetti D, Via M, Carreras-Torres R *et al*: Analysis of genomic regions associated with coronary artery disease reveals continent-specific single nucleotide polymorphisms in North African populations. *J Epidemiol* 2016; **643**: 1–8.
8 Zanetti D, Carreras-Torres R, Esteban E, Via M, Moral P: Potential signals of natural selection in the top risk loci for coronary artery disease: 9p21 and 10q11. *PLoS One* 2015; **10**: 1–21.
9 Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ: Second-generation PLINK : rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 1–16.
10 Alexander DH, Novembre J, Lange K: Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1–11.
11 Patterson N, Moorjani P, Luo Y *et al*: Ancient admixture in human history. *Genetics* 2012; **192**: 1065–1093.
12 Ong RT, Teo Y: varLD : a program for quantifying variation in linkage disequilibrium patterns between populations. *Bioinformatics* 2010; **26**: 1269–1270.
13 O'Connell J, Gurdasani D, Delaneau O *et al*: A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* 2014; **10**: e1004234.
14 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
15 Shiheng T, Rongmei Z, Jianhua C: A population genetics model of linkage disequilibrium in admixed populations. *Chinese Sci Bull* 2001; **46**: 193–197.
16 Athanasiadis G, González-pérez E, Esteban E, Dugoujon J, Stoneking M, Moral P: The Mediterranean Sea as a barrier to gene flow : evidence from variation in and around the F7 and F12 genomic regions. *BMC Evol Biol* 2010; **10**: 84.
17 Athanasiadis G, Moral P: Spatial principal component analysis points at global genetic structure in the Western Mediterranean. *J Hum Genet* 2013; **58**: 762–765.
18 Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA: High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between Northwestern Africa and the Iberian peninsula. *Am J Hum Genet* 2001; **68**: 1019–1029.
19 Capelli C, Redhead N, Romano V *et al*: Population structure in the Mediterranean Basin : a Y chromosome perspective. *Ann Hum Genet* 2006; **70**: 207–225.
20 Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, Stoneking M: Alu insertion polymorphisms in NW Africa and the Iberian Peninsula : evidence for a strong genetic boundary through the Gibraltar Straits. *Hum Genet* 2000; **107**: 312–319.
21 Gonzalez-Perez E, Esteban E, Via M *et al*: Population relationships in the Mediterranean revealed by autosomal genetic data (Alu and Alu/STR Compound Systems). *Am J Phys Anthropol* 2010; **141**: 430–439.
22 Humphreys RS: *Islamic History: a Framework for Inquiry*. Princeton University Press: Princeton, 1991.
23 Döring Y, Pawig L, Weber C, Noels H: The CXCL12/CXCR4 chemokine ligand/receptor axis in cardiovascular disease. *Front Psychol* 2014; **5**: 1–23.
24 Henn BM, Botigue LR, Gravel S *et al*: Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* 2012; **8**: e1002397.
25 Piancatelli D, Canossi A, Aureli A *et al*: Human leukocyte antigen-A, -B, and -Cw polymorphism in a Berber population from North Morocco using sequence-based typing. *Tissue Antigens* 2004; **63**: 158–172.
26 Coudray C, Olivieri A, Achilli A, Pala M, Melhaoui M, Cherkaoui M: The complex and diversified mitochondrial gene pool of Berber populations. *Ann Hum Genet* 2009; **73**: 196–214.

Supplementary Information accompanies this paper on *European Journal of Human Genetics* website (http://www.nature.com/ejhg)

Article III

Álvarez-Álvarez et al.

Genetic analysis of Sephardic ancestry in the Iberian Peninsula

Miguel M. Álvarez-Álvarez, Neil Risch, Christopher R Gignoux, Scott Huntsman, Elad Ziv, Laura Fejerman, Maria Esther Esteban, Magdalena Gayà-Vidal, Beatriz Sobrino, Francesca Brisighelli, Nourdin Harich, Fulvio Cruciani, Hassen Chaabani, Ángel Carracedo, Pedro Moral, Esteban González Burchard, Marc Via, and Georgios Athanasiadis

https://doi.org/10.1101/325779 (preprint)

Submitted to *Human Genetics*

Resumen en castellano

Los judíos sefardíes o sefarditas son una de las principales divisiones de la etnia judía, con presencia en la Península Ibérica ya en época romana. En este estudio empleamos datos a nivel genómico para investigar el grado de ascendencia genética sefardita en muestras de siete poblaciones de la Península Ibérica y regiones adyacentes, resultante de las conversiones forzosas que comenzaron en España en el siglo XIV y de la expulsión de aquellos que no se convirtieron.

Para llevar a cabo este trabajo, usamos también poblaciones del este del Mediterráneo (sur de Italia, Grecia e Israel) y norteafricanas (Túnez y Marruecos) como representantes actuales de los componentes ancestrales presentes en las poblaciones a estudiar, y realizamos tanto análisis clásicos (PCA, ADMIXTURE) como análisis basados en la información haplotípica (CHROMOPAINTER, fineSTRUCTURE, GLOBETROTTER, RFMix).

Detectamos evidencia de ascendencia genética sefardita en parte de las muestras ibéricas, así como en el norte de Italia y en Túnez. El componente sefardita parece ser de más reciente integración que el componente bereber en el acervo genético de la Península Ibérica. Así mismo, la menor antigüedad del componente sefardita en el norte de Italia que en la Península Ibérica podría reflejar la salida de los judíos expulsados en 1492.

Manuscript

Genetic analysis of Sephardic ancestry in the Iberian Peninsula

Miguel Martín Álvarez-Álvarez,[1,*] Neil Risch,[2,3] Christopher R. Gignoux,[4] Scott Huntsman,[5] Elad Ziv,[2,5] Laura Fejerman,[2,5] Maria Esther Esteban,[1] Magdalena Gayà-Vidal,[6] Beatriz Sobrino,[7] Francesca Brisighelli,[8] Nourdin Harich,[9] Fulvio Cruciani,[10] Hassen Chaabani,[11] Ángel Carracedo,[7,12,13] Pedro Moral,[1] Esteban González Burchard,[2,5,14] Marc Via,[15,16,#] Georgios Athanasiadis[17,*,#]

Affiliations

1. Department of Evolutionary Biology, Ecology and Environmental Sciences and Biodiversity Research Institute (IRBio), Universitat de Barcelona, Barcelona, Barcelona 08028, Spain

2. Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94118 USA.

3. Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA 94118, USA

4. Colorado Center for Personalized Medicine and Department of Biostatistics, University of Colorado, Denver, CO 80204, USA

5. Department of Medicine, University of California San Francisco, San Francisco, CA 94118, USA

6. CIBIO Research Center in Biodiversity and Genetic Resources, University of Porto, Vairão 4485-661, Portugal

7. Fundación Pública Galega de Medicina Xenómica-IDIS SERGAS, University of Santiago de Compostela, Santiago de Compostela, A Coruña 15706, Spain.

8. Section of Legal Medicine, Institute of Public Health, Catholic University of the Sacred Heart, Rome, RM 00168, Italy

9. Equipe des Sciences Anthropogénétiques et Biotechnologies, Faculté des Sciences, Université Chouaib Doukkali, El Jadida, Doukkala Abda 24000, Morocco

10. Dipartimento di Biologia e Biotecnologie "Charles Darwin", Sapienza Università di Roma, Rome, RM 00185, Italy

11. Laboratory of Human Genetics and Anthropology, Faculty of Pharmacy, University of Monastir, Monastir, Monastir 5000, Tunisia

12. Grupo de Medicina Xenómica-CIBERER-University de Santiago de Compostela, Santiago de Compostela, A Coruña 15782, Spain.

13. Center of Excellence in Genomic Medicine, King Abdulaziz University, Jeddah, Jeddah 21589, KSA.

14. Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94118, USA

15. Department of Clinical Psychology and Psychobiology and Institute of Neurosciences, Universitat de Barcelona, Barcelona, Barcelona 08035, Spain.

16. Institut de Recerca Sant Joan de Déu (IRSJD), Esplugues de Llobregat, Barcelona 08950, Spain.

17. Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen, Copenhagen 2100, Denmark

# These authors contributed equally to this work

*Corresponding authors

Abstract

The Sephardim are a major Jewish ethnic division whose origins can be traced back to the Iberian Peninsula. We used genome-wide SNP data to investigate the degree of Sephardic admixture in seven populations from the Iberian Peninsula and surrounding regions in the aftermath of their religious persecution starting in the late 14th century. To this end, we used Eastern Mediterranean (from South Italy, Greece and Israel) and North African (Tunisian and Moroccan) populations as proxies for the major ancestral components found in the target populations and carried out unlinked- and linked-marker analyses on the available genetic data. We report evidence of Sephardic ancestry in some of our Iberian samples, as well as in North Italy and Tunisia. We find the Sephardic admixture to be more recent relative to the Berber admixture following an out-of-Iberia geographic dispersal, suggesting Sephardic gene flow from Spain outwards. We also report some of the challenges in assigning Sephardic ancestry to potentially admixed individuals due to the lack of a clear genetic signature.

Main text

The Sephardic Jews – or Sephardim – are a major Jewish ethnic division whose origins can be traced back to Spain and Portugal. There is documented presence of Jewish populations in the Iberian Peninsula and the Maghreb (North Africa, west of Egypt) since the Roman period[1]. The fate of the Sephardim waxed and waned over the centuries, from periods of tolerance and acceptance to eras of persecution, but the turning point occurred in 1391, when the Jews of Spain started to opt for conversion to Catholicism in increasing numbers as a response to violent attacks and murders. It is estimated that by early 15th century, half of the approximately 400,000 Jews of Spain had already converted to Catholicism (becoming known as Conversos), while only about one-quarter remained as openly practicing Jews[2]. In 1492, the Catholic Monarchs of Spain issued the Alhambra Decree – an edict of expulsion whereby all Jews had to leave Spain by July of the same year.

The total number of Jews who left Spain has been a subject of controversy, although most recent estimates are between 50,000 and 80,000[3]. Part of the expelled Sephardic population settled in the Western Maghreb, whereas many others found shelter further east in the less

hostile Ottoman Empire, which at that time included many Levantine and Balkan countries, as well as Turkey[4]. Sephardic communities were also established in other parts of Europe, such as France and Holland, as well as in Latin American countries[5]. Early on, many Spanish Jews went to Portugal, which at the time was more tolerant. However, in 1497, Portugal followed suit with its own Edict of Expulsion, forcing many Spanish and Portuguese Jews to leave Portugal or convert. It has actually been suggested that Portugal was less tolerant of their Jews to depart, and more were forced to stay in Portugal and live as Catholics[6]. Today, both Spain and Portugal offer citizenship to Jews of documented Sephardic descent.

By 1492, the number of Conversos remaining in Spain may have risen to 200,000 (or even 300,000) out of a total population of approximately 7.5 million and in early 16th century, the proportion of the population of Sephardic origin is estimated to have been approximately 3-4%[3]. While soon after the Spanish and Portuguese Inquisitions Conversos were easily identified, their identity as former Jews was lost over time. Even though intermarriage between Conversos and Catholics was not commonplace at first, this changed over time, as the Jewish identity began to be lost. Therefore, one would expect to see evidence of Sephardic genetic ancestry among present day Iberians on a broad scale, and regionalization of the genetic signal is possible based on where the largest concentrations of Conversos had been living at the time of conversions.

Even though there have been many genetic studies on the origin of worldwide Jewish populations[7–11], the genetic aftermath of the persecution of the Sephardim in the Iberian Peninsula is still poorly understood. A natural question to ask is the degree to which populations in the Iberian Peninsula show evidence of Sephardic genetic ancestry – and to what extent it may be regionalized. This question has been primarily addressed with uniparental markers[12,13] – often with contrasting results. For instance, there is mtDNA-based evidence for Iberian admixture in Turkish Jewish communities[14]. Similarly, another mtDNA study showed that the Sephardim bore more resemblance to Spanish than to Jewish populations, but the opposite trend was observed for Y chromosomes[15,16], a signal possibly reflecting the original Jewish colonization of Spain. On the contrary, the crypto-Jewish Xuetes from Majorca showed high frequency in R0, a mitochondrial lineage typically found in Middle Eastern populations – including the Jews[17]. Recent whole-genome studies have provided valuable insights into the

genetic structure of the Iberian Peninsula[18] (C.B., unpublished data). However, these studies lack at best an explicit reference to Sephardic genetic variation, limiting their scope primarily to North African genetic influences.

In light of the above, this study aims to provide a more detailed picture of the genetic influence of the Sephardim on a number of populations currently living in the Iberian Peninsula. Such influence would have probably been the result of Sephardic admixture into the majority Christian community through the forced conversions that started on a large scale in the late 14th century, continuing through the time of the expulsion and beyond. To examine possible Sephardic admixture into other populations, we extended the geographic scope of our study to two Maghrebi populations from Morocco and Tunisia, as well as two European populations from South France and North Italy.

For this study, we used genotype data from 13 Mediterranean or near-Mediterranean populations of similar sample size (mean N ≈ 40; Table 1). More specifically, we typed 518 individuals (277 males and 241 females) on the Affymetrix 250K Sty array (Affymetrix, Santa Clara, CA, USA). The geographic origin of the samples is shown in Figure 1. The Iraqi and Sephardic Jewish samples were collected in Israel and represent individuals with self-reported Iraqi Jewish or Turkish Sephardic grandparents, respectively. We included Iraqi Jews in our panel in order to account for Middle Eastern Jewish genomic ancestry. All participants provided written informed consent. Our study was conducted according to the Declaration of Helsinki II and was approved by the Bioethics Committees of the University of Barcelona, Spain and the University of California San Francisco, USA.

Country	Geographic region or ethnic group	Acronym	Initial N	N after QC	N after fineSTRUCTURE
Spain	Andalusians (Las Alpujarras)	AND	40	40	40
Spain	Basques (Gipuzkoa)	BAS	37	37	37
Spain	Catalans (Olot)	CAT	45	41	39
Spain	Galicians (Santiago de Compostela)	GAL	34	34	34
Portugal	Portuguese (Porto)	POR	41	39	39
France	South French (Toulouse)	SFR	46	46	44
Italy	North Italians (Liguria)	NIT	31	29	28
Italy	South Italians (Calabria, Puglia, Campania & Basilicata)	SIT	35	35	35
Greece	Cretans (Crete)	CRT	44	44	44
Morocco	Berbers (Khenifra)	BER	40	39	38
Tunisia	Tunisians (Monastir)	TUN	41	39	36
Israel	Iraqi Jews	IRQ	42	38	37
Israel	Sephardic Jews (Turkish origin)	SPH	42	39	39
		Total	**518**	**500**	**490**

Table 1*: Summary of the analysed population samples. Sample sizes are shown before and after quality control (QC), as well as after fineSTRUCTURE clustering.*

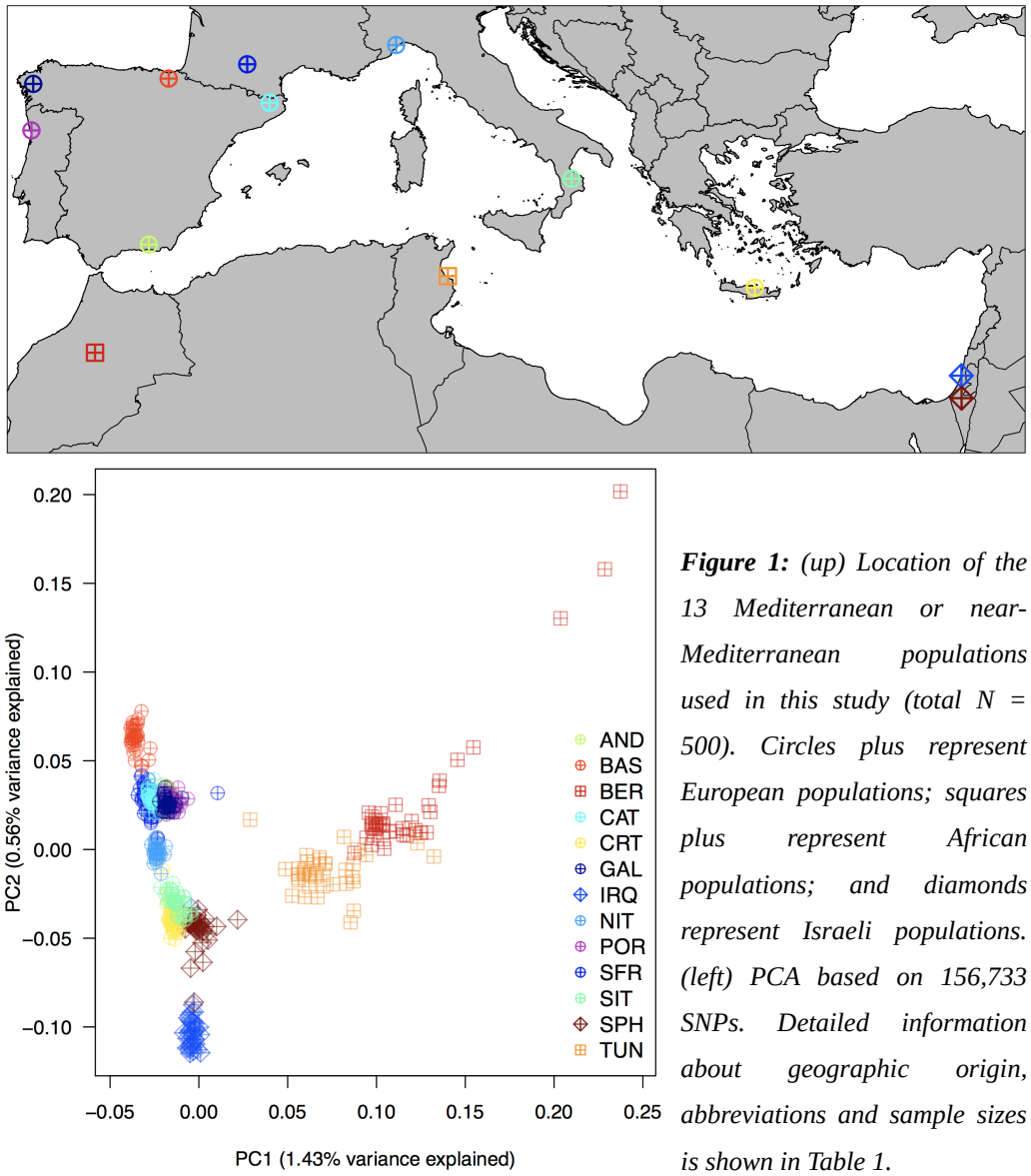


Figure 1: *(up) Location of the 13 Mediterranean or near-Mediterranean populations used in this study (total N = 500). Circles plus represent European populations; squares plus represent African populations; and diamonds represent Israeli populations. (left) PCA based on 156,733 SNPs. Detailed information about geographic origin, abbreviations and sample sizes is shown in Table 1.*

Standard quality control was performed with PLINK v1.9[19] and included the removal of duplicates, close relatives, extreme outliers, as well as individuals with >5% genotype missingness. We subsequently removed variants with >5% genotype missingness and those that did not fit per-population Hardy-Weinberg proportions after Bonferroni correction. After quality

control, a total of 500 samples and 214,338 autosomal single nucleotide polymorphisms (SNPs) were available for downstream analyses.

We performed principal component analysis (PCA) on all 13 populations with PLINK using a set of 156,733 SNPs obtained after linkage disequilibrium (LD) pruning (r^2 threshold = 0.8; window size = 50; step size = 5). The first PC separates North Africans (Berbers and Tunisians) from the rest of the populations, whereas the second PC roughly corresponds to the distribution of the remaining populations along an east-west axis with the Iraqi Jews and the Basques occupying the two extremes (Figure 1). The PCA plot revealed that the Sephardic Jews are genetically most similar to Eastern Mediterranean populations, such as Cretans and South Italians, and to a lesser degree to the Iraqi Jews, in agreement with previous observations[7,8]. The observed genetic affinities remained the same after first removing the North Africans (Figure S1), and then the Eastern Mediterraneans and Jewish populations (Figure S2) from the PCA.

We also ran an unsupervised ADMIXTURE analysis[20] using the same SNP set as in PCA. ADMIXTURE corroborated our PCA findings regarding the genetic relationship between the Jewish populations and the rest of the samples (Figure 2). For K = 3 (i.e. the model with the smallest cross-validation error = 0.457), Iraqi Jews, Basques and Berbers appear as the most differentiated and homogeneous clusters (for comparison see Figure 1). For $K \geq 4$, the Iraqi Jews persist as a largely homogeneous cluster, whereas for K = 7 in particular, the Sephardic Jews acquire their own cluster (in brown), which we can then trace in other populations. This putatively Sephardic component was notably present in Tunisia (27.23%), but also in the Iberian Peninsula with the exception of the Basque Country (Galicia: 4.25%; Portugal: 2.39%; Andalusia: 2.36%; Catalonia: 1.32%), South France (1.56%), North Italy (1.70%) and South Italy (7.50%).

We then used CHROMOPAINTER, fineSTRUCTURE and GLOBETROTTER[21,22] in their default settings to achieve a more detailed picture of population structure in our samples. Because our SNP chip was of an older generation, we used liftOver from the University of California Santa Cruz Genome Browser to update SNP positions from the NCBI36/hg18 (March 2006) to the GRCh37/hg19 build (February 2009). To increase the total number of available SNPs necessary for these analyses, we carried out genotype imputation using the

Michigan Imputation Server[23] (phasing software: Eagle v2.3[24]; imputation software: Minimac v3[23]; reference panel: hrc.r1.1.2016; QC's reference population: mixed population). After filtering for info ≥ 0.995 and removing monomorphic loci and singletons, there were 527,379 phased autosomal SNPs for the painting analyses.

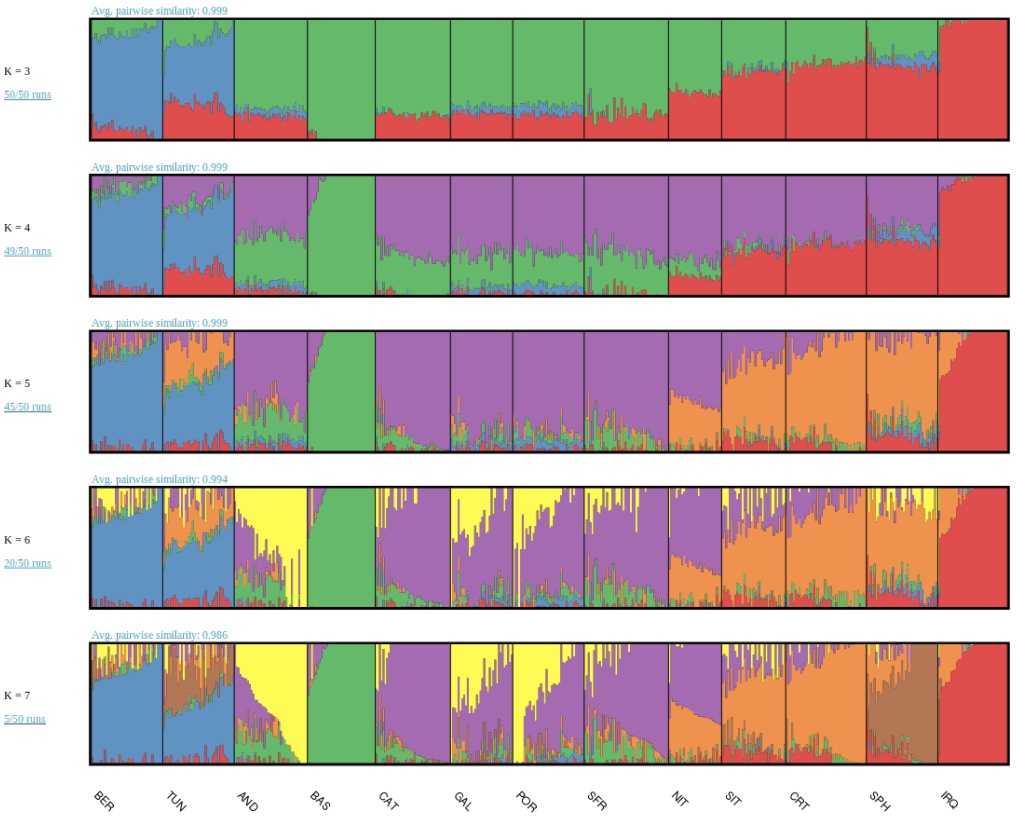


Figure 2: *Ancestral component analysis of the 13 Mediterranean or near-Mediterranean studied populations assuming K = {3, 4, 5, 6, 7} admixing populations. ADMIXTURE was run 50 times for each K and results were plotted with PONG[31]. Bar plots show the consensus per-individual membership to each of the specified ancestral components. Population labels are described in Figure 1.*

We first carried out a painting in which each of the 500 samples was allowed to copy DNA segments from all of the remaining samples. CHROMOPAINTER returns similarity matrices of

shared haplotype counts and shared haplotype lengths. We used the count matrix together with fineSTRUCTURE to hierarchically cluster the 500 individuals into a Bayesian consensus phylogenetic tree. fineSTRUCTURE returned a tree of nine largely homogeneous clusters (Figure S3), which resemble to some extent the results from PCA (Figure 1) and ADMIXTURE (Figure 2). As before, the most differentiated clusters were the ones from North Africa (Tunisians and Berbers), whereas the next node divided the remaining populations into two clusters: Southwest vs. Southeast Europe (including the two Jewish populations). The clustering is consistent with geography with the exception that North Italy was clustered with the Southwestern populations rather than Southeastern. South Italy and Crete clustered together, and the Sephardic Jews appeared more closely related to these two populations than to the Iraqi Jews. To add more resolution, we repeated the painting using only the Iberian (without the Basques) and French populations, in which as above each individual was allowed to copy DNA segments from all other individuals. In this case, fineSTRUCTURE returned a tree with four overly homogeneous clusters, in which the Galician and Portuguese were the only samples that did not split according to their labels (and were therefore treated as one group in the subsequent analyses), while Catalonia, South France and Andalusia formed virtually homogeneous groups (Figure S4).

We used the information from the two fineSTRUCTURE analyses to select recipient and donor groups and carried out two additional paintings in which (i) recipients were allowed to copy DNA segments only from donors (i.e. donor-to-recipient painting) and (ii) each donor was allowed to copy DNA segments only from other donors (i.e. donor-to-donor painting). For this particular analysis, we used CHROMOPAINTER's length matrix together with GLOBETROTTER to estimate South European, Sephardic and North African (i.e. donors) admixture proportions in the Iberian, French, North Italian (i.e. recipients) clusters.

Figure 3A shows the main Mediterranean donor group contributions to six Southwestern European clusters. Among these, the highest Sephardic admixture appeared in Andalusia (12.3%; 95%CI: 11.1-13.5%), followed by Galicia and Portugal (11.3%, 95%CI: 10.6-12.1%), while it was 5.9% in North Italy (95%CI: 5-6.8%). The Sephardic component was notably absent in the Basque Country and South France, as well as in Catalonia (point estimate of ~0.3%, not statistically different from zero).

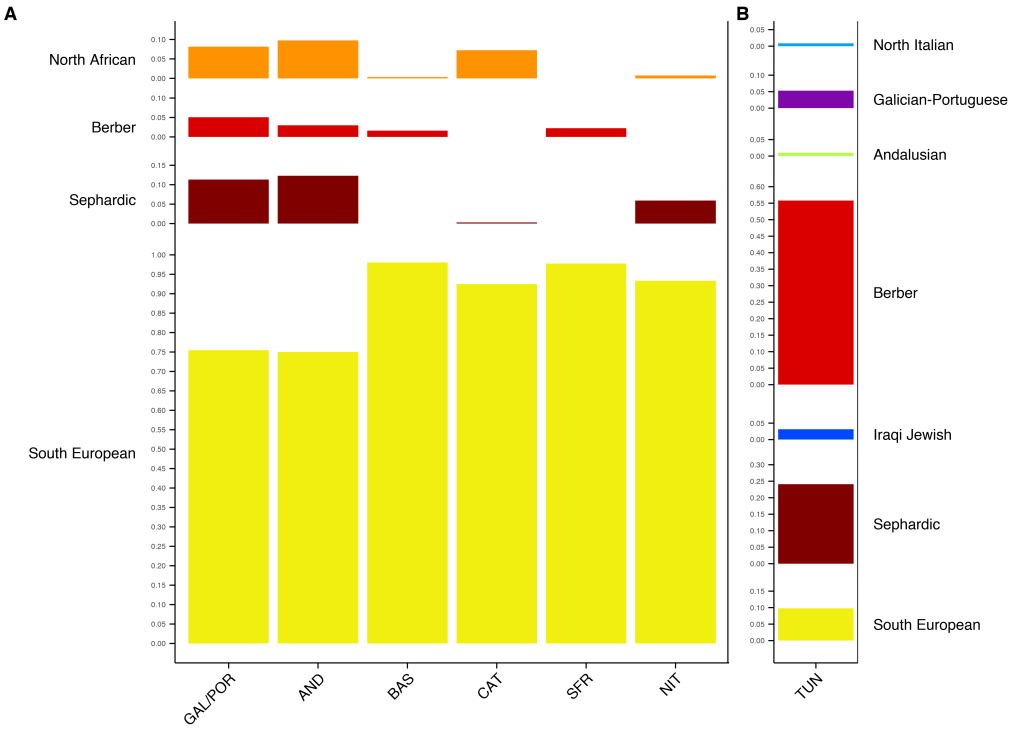


Figure 3: (A) *GLOBETROTTER admixture profiles for six populations from the Iberian Peninsula, France and Italy. Of the five admixture components used in the model (South European, Berber, North African, Sephardic and Iraqi Jewish), only the first four contributed substantially to the recipient populations. The South European component included samples from South Italy and Crete.* ***(B)*** *GLOBETROTTER admixture profile for the Tunisian population using seven admixture components (South European, Sephardic, Iraqi Jewish, Berber, Andalusian, Galician-Portuguese and North Italian). Population legends are described in Figure 1.*

These results reflect the regionalization of Sephardic admixture on the Iberian map and surrounding regions. Our Sephardic admixture estimates in the Iberian Peninsula were overall more conservative than those reported elsewhere for Y chromosome data[25] (their mean value of Sephardic ancestry: 19.8% vs. ours: 6%). The Iraqi Jews were the only donor group that did not contribute to any of the recipient clusters, while the Berber component was present in all of the Iberian populations (with the exception of Catalonia) and South France, matching historical

knowledge about the Moorish presence in the Iberian Peninsula during the Middle Ages[26], as well as results from a recent study in the same region (C.B., unpublished data). In addition, the higher Sephardic ancestry in Galicia and Portugal compared to the rest of the peninsula could be reflecting the ban on the departure of Jews from the kingdom of Portugal combined with greater pressure to convert to Catholicism (thus corroborating historical hypotheses) and/or Sephardic gene flow from Spain to Portugal[27,28].

Historical records show that many Sephardic Jews sought shelter in the Maghreb after they were expelled from Spain and Portugal[1]. In light of this, we performed an additional CHROMOPAINTER and GLOBETROTTER analysis, this time using Tunisians as the recipient population, and all of the remaining clusters as donors, in order to estimate the Sephardic influence in North Africa (Figure 3B). The analysis showed that the main ancestry contributors for Tunisia were the Berbers (55.8%; 95%CI: 55.3-56.3%) followed by the Sephardic Jews (24.1%; 95%CI: 23.2-25%), and to a lesser extent by the remaining donor clusters. Even though the Sephardim were present in Tunisia, it is unlikely that their genetic contribution is as high as reported here by either ADMIXTURE or GLOBETROTTER. The rate of conversion of Sephardic Jews into the local Muslim population was low and intermarriage between Jews and Muslims has been limited throughout their history[29]. Rather, we attribute this high percentage to lacking an appropriate reference population that reflects better the historical background of Tunisia (e.g. a proxy population for the Phoenicians). In Antiquity, the Phoenicians not only had a major capital in Tunisia (Carthage), but also significant outposts in the Iberian Peninsula. As a Levantine population, the Phoenicians were probably genetically very similar to the ancestral Jewish populations of the Mediterranean. Bearing this in mind, we note that our Sephardic estimates for the Iberian Peninsula are potentially also liable to the same issue and should therefore be interpreted as an upper bound.

Finally, in order to provide temporal context for our observations in Southwestern Europe, we inferred local ancestry on phased chromosomes using RFMix[30]. RFMix takes into account LD among markers and identifies ancestry tracts originating in each of the chosen admixing populations. For this analysis, target populations included only samples with considerable Sephardic ancestry, i.e. Andalusia, Galicia-Portugal and North Italy. We inferred local ancestry patterns on target populations as the result of a three-way admixture between Sephardim,

Berbers and Iberians. Sephardic genetic variation was represented by the Sephardic Jews (N = 39); Berber variation by the Moroccan Berbers (N = 38); and Iberian variation by 10 samples from Andalusia and Portugal that showed high membership (≥99%) to the yellow cluster for K = 7 in Figure 2. We ran RFMix three times, assuming time of admixture g = {10, 25, 40} generations ago. The patterns of tract length distribution were qualitatively similar across all three choices of g (differing only in scale), so we chose to focus the rest of our report on g = 25 (~625 years before present).

Figure 4A shows the distribution of migrant tract lengths for the pooled Iberian populations (Andalusia and Galicia/Portugal), whereas Figures S5-7 show the same distribution for each population and each time of admixture g separately. In all cases, Sephardic tracts were predominant over Berber tracts, both in frequency and in length, implying more recent admixture with the Sephardim than with the Berbers.

We then checked whether Sephardic tract length for each of the three populations under study correlated with geography. Figures 4B-C show the relationship (i) between geographic location and median Sephardic tract length, and (ii) between geographic location and variance in Sephardic tract length for the three target populations, whereas Figure S8 shows the same information for g = {10, 40}. Even though the effect is subtle, we detected a geographic pattern, with ancestry tracts looking younger (higher median length and variance) outside Iberia and older (lower median length and variance) inside Iberia. Combined with the relative tract length distribution (Figure 4A), we interpret these observations as suggestive of recent Sephardic gene flow from the Iberian Peninsula towards the East of the Mediterranean. In addition, tract lengths in the two Iberian populations (i.e. Andalusia and Galicia-Portugal) followed very similar distributions (Figures 4B-C), probably reflecting near-contemporary admixing events.

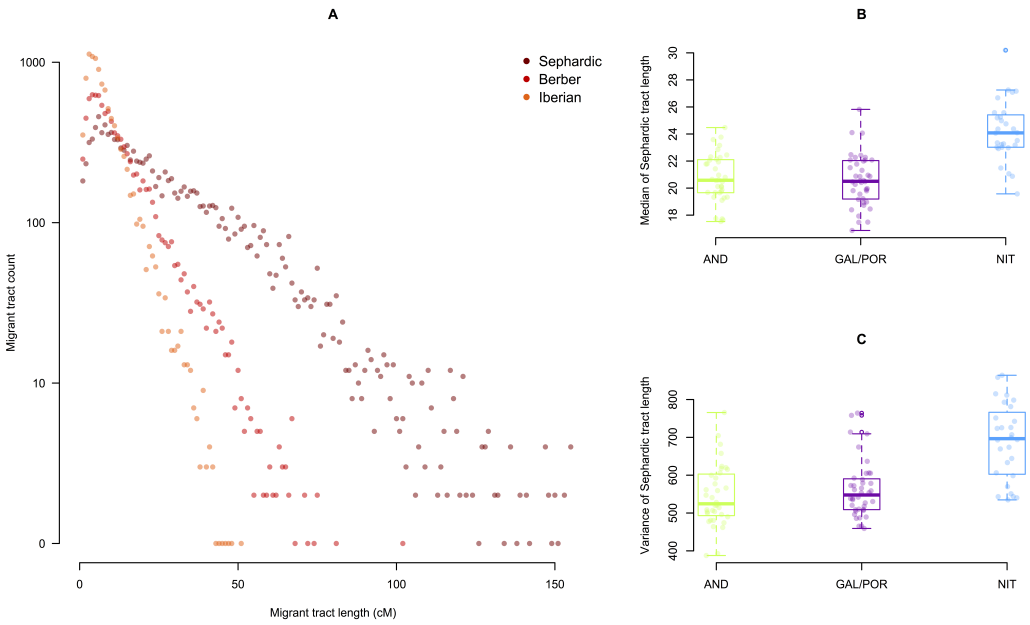


Figure 4: (A) *Distribution of Sephardic, Berber and Iberian migrant tracts (1-cM bins) in a pooled sample made up of Andalusians and Galician-Portuguese assuming time of admixture g = 25 generations ago. Migrant tract length was calculated by subtracting initial from final genomic position (in cM) of each tract as reported by RFMix. Tract length calculations were restricted within chromosome arms, thus omitting centromeres. Berber tracts were notably shorter than Sephardic tracts, suggesting that Berber admixture is older than Sephardic admixture in the Iberian Peninsula.* ***(B)*** *Jitter and box plot of (i) median Sephardic tract length and (ii) variance of Sephardic tract length for Andalusians, Galician-Portuguese and North Italians. Sephardic migrant tracts look younger (i.e. with larger median and variance) outside the Iberian Peninsula. Population legends are described in Table 1.*

Recent whole-genome studies on the structure of worldwide Jewish groups showed that current Sephardic populations are more related to other Jewish and Middle Eastern groups than to non-Jewish Spaniards and Northwest Africans[7–9]. Here, we detected a Sephardic component in the Iberian Peninsula, as well as North Italy and Tunisia. A possible reason for this discrepancy could be the methods used; we observed that PCA did not detect the Sephardic genetic influence in the Iberian Peninsula (Figures 1B and 2), yet haplotype-based methods such as CHROMOPAINTER and RFMix, were more powerful at picking up signatures of Sephardic

admixture (Figures 3 and 4). We therefore recommend haplotype-based methods for the discovery of Sephardic admixture signals in follow-up studies.

An interesting question regarding the era of the Unification of Spain is whether admixture between the Christian, Muslim and Jewish groups was mainly ancient or promoted by the persecution and consecutive conversion. The relative frequency of migrant tracts suggests that Sephardic admixture is notably more recent than Muslim admixture. Given that the presence of Sephardic populations in the Peninsula predates the Arab invasion, this can be interpreted as Sephardic admixture occurring in more recent times, perhaps as intermarriage under the newly acquired Catholic label. An additional hint that corroborates this idea comes from the distribution of Sephardic tracts along Southwestern Europe, as Sephardic gene flow seems to follow an out-of-Iberia pattern, being more recent outside the Iberian Peninsula, although more samples from their potential route to the East are warranted to draw clearer conclusions.

The high percentage of Sephardic ancestry in Tunisia (and probably also Northern Italy, even though the numbers there are quite smaller), as reported both by ADMIXTURE and GLOBETROTTER, does not match historical expectations. We believe that this could be due to the lack of a more appropriate Levantine donor population for the painting of Tunisian chromosomes. This observation hints at the lack of sufficient specificity in the Sephardic genetic signature, probably reflecting a less tumultuous demographic history for the Sephardim compared to e.g. the Ashkenazim. We therefore interpret our admixture estimates as an upper bound.

In this study, we reassessed Sephardic ancestry in the genetic pool of the Iberian Peninsula and parts of Northwest Africa. To do so, we used a sample of Israelis with all four of their grandparents of Turkish Sephardic descent as a proxy for a Sephardic-descended population. In our analyses, Sephardic ancestry was present in many but not all of the studied samples from the Western Mediterranean, possibly reflecting the regionalization of Sephardic admixture and the historical particularities of each region. The observed geographic patterns serve as a starting point to test more specific hypotheses about the course of events around the time of the persecution of the Sephardim in the Iberian Peninsula. Future research should include a broader

geographic scope encompassing not only the Mediterranean but also other parts of the globe (J.C., unpublished data) and, ideally, denser SNP arrays.

Acknowledgements

We would like to thank all participants that provided samples for the study. Genotyping services were provided by the Spanish “Centro Nacional de Genotipado” (CEGEN-ISCIII)” and we also thank the CEGEN coordination team for their support. E.G.B. is supported by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, the Harry Wm. and Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, the National Heart, Lung, and Blood Institute (R01HL117004, R01HL128439, R01HL135156, X01HL134589), the National Institute of Environmental Health Sciences (R01ES015794, R21ES24844), the National Institute on Minority Health and Health Disparities (P60MD006902, R01MD010443, RL5GM118984) and the Tobacco-Related Disease Research Program (24RT-0025).

Declaration of Interests

The authors declare no competing interests.

Literature cited

1. Gerber, J.S. (1994). Jews of Spain: a History of the Sephardic Experience (New York: Simon and Schuster).

2. Pérez, J. (1993). History of a tragedy. The expulsion of Spanish Jews (Barcelona: Crítica).

3. Valdeón-Baruque, J. (2007). El reinado de los Reyes Católicos. Época crucial del antijudaísmo español. In El Antisemitismo En España (Cuenca: Ediciones de la Universidad de Castilla-La Mancha).

4. Mazower, M. (2006). Salonica, City of Ghosts: Christians, Muslims and Jews 1430-1950 (Vintage).

5. Sachar, H.M. (1994). The world of the Sephardim remembered (Alfred A. Knopf Inc.).

6. Soyer, F. (2007). The Persecution of the Jews and Muslims of Portugal. King Manuel I and the End of Religious Tolerance (Leiden: Brill).

7. Haber, M., Gauguier, D., Youhanna, S., Patterson, N., Moorjani, P., Platt, D.E., Matisoo-smith, E., Soria-hernanz, D.F., Wells, R.S., Botigue, L.R., et al. (2013). Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture. PLoS Genet. *9*.

8. Atzmon, G., Hao, L., Pe, I., Velez, C., Pearlman, A., Palamara, P.F., Morrow, B., Friedman, E., Oddoux, C., Burns, E., et al. (2010). Abraham's Children in the Genome Era : Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. Am. J. Hum. Genet. *86*, 850–859.

9. Behar, D.M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., et al. (2010). The genome-wide structure of the Jewish people. Nature *466*, 238–242.

10. Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Ostrer, H., Price, A.L., and Reich, D. (2011). The History of African Gene Flow into Southern Europeans , Levantines , and Jews. PLoS Genet. *7*.

11. Ostrer, H., and Skorecki, K. (2013). The population genetics of the Jewish people. Hum. Genet. *132*, 119–127.

12. Nogueiro, I., Manco, L., Gomes, V., Amorim, A., and Gusmão, L. (2010). Phylogeographic analysis of paternal lineages in NE Portuguese Jewish communities. Am. J. Phys. Anthropol. *141*, 373–381.

13. Nogueiro, I., Teixeira, J., Amorim, A., Gusmão, L., and Alvarez, L. (2015). Echoes from Sepharad: signatures on the maternal gene pool of crypto-Jewish descendants. Eur. J. Hum. Genet. *23*, 693–699.

14. Behar, D.M., Metspalu, E., Kivisilid, T., Rosset, S., Tzur, S., Hadid, Y., Yudkovsky, G., Rosengarten, D., Pereira, L., Amorim, A., et al. (2008). Counting the founders: The matrilineal genetic ancestry of the Jewish Diaspora. PLoS One *3*.

15. Picornell, A., Giménez, P., Castro, J.A., and Ramon, M.M. (2006). Mitochondrial DNA sequence variation in Jewish populations. Int. J. Legal Med. *120*, 271–281.

16. Picornell, A., Jiménez, G., Castro, J., and Ramon, M. (2004). Minimal Y-chromosome haplotypes plus DYS287 in Jewish populations. J Forensic Sci *49*, 410–412.

17. Picornell, A., Gómez-Barbeito, L., Tomàs, C., Castro, J.A., and Ramon, M.M. (2005). Mitochondrial DNA HVRI variation in Balearic populations. Am. J. Phys. Anthropol. *128*, 119–130.

18. Botigue, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., et al. (2013). Gene flow from North Africa contributes to differential human genetic diversity in Southern Europe. Proc. Natl. Acad. Sci. *110*, 11791–11796.

19. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2014). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 1–16.

20. Alexander, D.H., and Novembre, J. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. Genome Res. *19*, 1655–1664.

21. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, 11–17.

22. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A Genetic Atlas of Human Admixture History. Science (80-.). *343*, 747–751.

23. Das, S., Forer, L., Schönherr, S., Sidore, C., and Locke, A.E. (2017). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

24. Loh, P., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, A., Finucane, H.K., Schoenherr, S., Forer, L., Mccarthy, S., Abecasis, G.R., et al. (2017). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.

25. Adams, S.M., Bosch, E., Balaresque, P.L., Ballereau, S.J., Lee, A.C., Arroyo, E., L??pez-Parra, A.M., Aler, M., Grifo, M.S.G., Brion, M., et al. (2008). The Genetic Legacy of Religious Diversity and Intolerance: Paternal Lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am. J. Hum. Genet. *83*, 725–736.

26. Watt, M. (1967). A History of Islamic Spain (Edinburgh).

27. Pignatelli, M. (2000). A comunidade israelita de Lisboa: o passado e o presente na construccão da etnicidade dos judeus de Lisboa (Lisboa: Universidade Técnica de Lisboa, Instituto Superior de Ciências Sociais e Políticas).

28. Martins, J. (2006). Portugal e os judeus: Judaísmo e anti-semitismo no século XX (Lisbon: Vega).

29. Niremberg, D. (2004). Love Between Muslim and Jew in Medieval Spain: a Triangular Affair. In Jews, Muslims, and Christians in and around the Crown of Aragon: Essays in Honour of Professor Elena Lourie, H.J. Hames, ed. (Leiden), pp. 60–76.

30. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix : A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. *93*, 278–288.

31. Behr, A.A., Liu, K.Z., Liu-fang, G., Nakka, P., and Ramachandran, S. (2017). Genetics and population analysis pong : fast analysis and visualisation of latent clusters in population genetic data. Bioinformatics *32*, 2817–2823.

Article IV

Álvarez-Álvarez et al.

High-coverage sequence data from the Spanish Eastern Pyrenees suggest patterns of population structure and isolation

Miguel M Álvarez-Álvarez*, Iago Maceda*, Georgios Athanasiadis, Pedro Moral, and Oscar Lao

**These authors contributed equally to the study*

Submitted to *The New England Journal of Medicine*

Resumen en castellano

Las poblaciones aisladas son infrecuentes en Europa, dada la historia demográfica del continente. No obstante, las poblaciones aisladas genéticamente son importantes desde un punto de vista médico, ya que suelen mostrar patrones singulares de mutaciones deletéreas y una elevada incidencia de ciertas enfermedades monogénicas recesivas raras. Este fenómeno se ha interpretado como una consecuencia de la incapacidad de la selección purificadora de eliminar estas variantes dañinas en poblaciones de tamaño reducido.

En este estudio, se ha analizado la secuencia genómica completa con alta cobertura de 30 individuos de cinco regiones del Pirineo español oriental, el cual, por su situación geográfica, podría ser una población genéticamente aislada.

Los análisis de estructuración poblacional en el contexto de la Península Ibérica sugieren que esta región ha estado aislada históricamente. Además, análisis basados en *deep learning* sugieren una reducción drástica del tamaño efectivo poblacional en tiempos históricos. Por otra parte, análisis *in silico* muestran que las muestras portan en su genoma una carga menor de mutaciones raras altamente deletéreas que las muestras del Proyecto 1000 Genomas IBS (España) y FIN (Finlandia). Ello se ve apoyado en estudios anteriores que sugieren que un aislamiento a largo plazo y un declive en el tamaño poblacional pueden causar la eliminación de una gran parte de las mutaciones dañinas.

Manuscript

High-coverage sequence data from the Spanish Eastern Pyrenees suggest patterns of population structure and isolation

Miguel M Álvarez-Álvarez[1]*, Iago Maceda[2]*, Georgios Athanasiadis[3], Pedro Moral[1], and Oscar Lao[2]

Affiliations:

1 Department of Evolutionary Biology, Ecology and Environmental Sciences. University of Barcelona, Barcelona 08028, Spain

2 CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixach 4, 08028 Barcelona, Spain

3 Institute of Biological Psychiatry (IBP), Mental Health Services, Sct. Hans, Roskilde (Denmark)

** These authors contributed equally to the study*

Abstract

Given the demographic history of Europe, human genetic isolates in the continent are unfrequent. Nevertheless, genetic isolates are important from a medical point of view, because they usually show distinctive patterns of rare deleterious mutations and a higher frequency of rare recessive monogenic diseases. This phenomenon has been interpreted as a consequence of the failure of purifying selection to successfully remove these variants from the population. Nevertheless, it has also been shown in other species that long term genetic isolation and population decline can facilitate the cleaning of damaging recessive mutations. In this work we have analysed the genome of 30 individuals from five different valleys from the Spanish Eastern Pyrenees (SEP) sequenced at 40×, which for their geographical location could be

potentially a genetic isolated population. All our analyses of population structure with the Iberian IBS from the 1000 Genomes Project (1000G) suggest that SEP is a genetic isolate. In silico burden allele analyses suggest that SEP individuals carry in their genome a lower amount of rare damaging mutations compared to IBS and the Finnish FIN population of 1000G. Furthermore, our ABC-DL analyses suggest a strong 10-fold decline in the effective population size of the SEP during the last 100 generations, thus supporting the hypothesis that long term isolation and population decline can explain the reduction in the pool of burden alleles in SEP.

Introduction

From a genetic point of view, genetic isolated populations are characterized by an enrichment of homozygote segments and linkage disequilibrium compared to non-isolated populations[1]. Usually, genetically isolated populations show a set of particular recessive monogenic diseases due to unique mutations not shared with other population[2]. However, it has also been shown in other studies that long term isolation and endogamy can clean the genome from deletereous mutations by increasing homozygosity, leading to a purge of most recessive deleterious mutations[3].

Previous studies have consistently shown that, as a general trend, the genetic diversity of the European population at a continental level correlates with geography and it is not shaped by strong genetic barriers[4,5]. Similar results have been obtained when analysing at a regional level the population structure within some countries[6–8]. However, in other countries more complex patterns have been found, reflecting geographic barriers[9], and historical migration events[10,11]. Overall, these results suggest that the genetic diversity of the European continent has been shaped by a pattern of isolation by distance within the European continent and a large number of migrations[12]. So far, very few European populations have been described as genetic isolates. Genetic isolation within the European continent can be due to cultural factors such as religion or language[13,14] and/or the presence of geographic barriers such as oceans[15–18] or mountains[19].

In this context, the human populations of the Pyrenees appear as a particularly interesting genetic isolate candidate within Europe[20]. These populations are of an almost exclusively rural nature, and they are settled within a mountainous chain that creates a barrier between the

Iberian Peninsula and the rest of Europe, as well as between the Pyrenean valleys themselves. The ancient Pyrenees are considered to have been culturally heterogeneous[21], and several pre-Roman tribes from the Spanish Eastern Pyrenees (SEP) mentioned by ancient sources are already distributed roughly following a West-East axis, such as the Airenosini (Pallars, Alt Urgell), Bergistani (Alt Urgell, Bergueda) and Castellani (Ripolles, Garrotxa). It is estimated that the Roman and Visigoth people only represented a 2.2-4.4% of the SEP population, while the Islamic conquest lasted only 80 years[22]. Current administrative regions are somewhat represented by the medieval counties Pallars Jussa, Urgell (Alt Urgell), Berga (Bergueda; actually a *pagus* belonging to the county of Cerdanya), and Besalu, which included current Ripoll (Ripolles) and Olot (Garrotxa). Furthermore, their rural condition ensures that they have been barely influenced by the internal migration processes of the last century, which have operated mainly on an urban level, and the barrier condition of the Pyrenees suggests that these populations could present some degree of genetic structure. Additionally, these facts combined could also imply some degree of genetic isolation and endogamy, which could lead to a significantly higher genetic risk for complex diseases compared to that of general urban populations[2].

To the best of our knowledge, few genetic studies have included samples from the human populations of SEP. Previous studies used classical markers, such as blood markers, proteins and HLA antigens[23] and immunoglobulin data[20], but they fail in finding any pattern of population structure at all. A Y chromosome polymorphism study detects a subtle degree of structure in the whole Spanish Pyrenees mountain range[24]. The most recent one including a reduced number of samples from the Pyrenees and considering microarray data did not identify any genetic difference with other Iberian samples[18].

In the present study we have characterized the genetic variation of the SEP population from five administrative regions (Pallars, Alt Urgell, Berga, Ripolles and Garrotxa) making use, for the first time, of high-coverage whole-genome sequencing (WGS) data. This allowed the use of powerful haplotype-based methods, revealing genetic differences between close groups. Likewise, it ensured a non-biased capture of the allele frequency spectrum, something necessary in demographic model inference and in the assessment of the relative genetic risk for disease of the population. Another novelty of the study design is that each participant was

selected on condition that the coordinates of the birthplaces of their four grandparents be located within the same administrative region. This requisite, in combination with a mean sample age of 74.14 years, resulted in a sample containing fine-scale geographic information that is effectively unaffected by the demographic changes occurred during the 20th century.

From a methodological perspective, this work provides new information on the possibilities of detecting isolated populations in mainland Europe, and, most importantly, we show that long-term isolation in humans can lead to a decrease in the burden of deleterious mutations.

Material and methods

WGS datasets

We sequenced the genomes of 30 individuals coming from five SEP administrative regions. All samples had all their grandparents born in the same sampled region. This dataset represents the oldest extract of the population, with an average age of 74.14 years and equal proportions of both sexes. The chosen regions correspond to the political separations stablished by the government and are named *comarca (*pl. *comarques*). In total, five *comarques* were sampled (Garrotxa, Ripollés, Berguedà, Alt Urgell and Pallars), with a total of six samples per region. The sampling consisted on blood extraction together with the information of the place of birth from the sampled individual and for the four grandparents. All subjects signed an informed consent and the study had the approval of the Ethics Committee of the University of Barcelona.

Figure 1. (large map) SEP regions sampled in the context of Catalonia, namely a-Pallars; b-Alt Urgell; c-Berguedà; d-Ripollès; e-Garrotxa. (small map) The location of Catalonia within Europe. Adapted from Wikipedia.

Sequencing, SNP calling and data cleaning

Whole-genome sequencing (WGS) was performed at Centre Nacional d'Anàlisi Genòmica of the Centre for Genomic Regulation (CNAG-CRG) at Barcelona, Spain, using standard Illumina Paired-ends sequencing technology with a read-length of 150 bp. The average sequencing coverage of all samples was >40× (36.98x - 43.416x).

We carried out the single nucleotide polymorphism (SNP)-calling using GATK HaplotypeCaller v3.6[25], applying a threshold of at least five Phred >30 reads to call a position, at least 2 reads to call a variant, and the remaining default settings from the GATK manual[26]. We then used BCFtools to extract the biallelic and polymorphic SNPs. After this step, the VCF file was converted to PLINK[27] binary file system. The first step of the quality control consisted in the removal of variants with more than 0% missingness, and samples with more than 10% missingness. Secondly, we checked for kinship between our samples using KING[28], which resulted in the exclusion of one of the individuals from Alt Urgell. The final dataset included a total of 29 individuals and 9,309,056 biallelic polymorphic SNPs.

Publicly available WGS datasets

In order to make comparisons with populations from other datasets, we merged our data with i) West Eurasian samples from the Simons Genome Diversity Project (SGDP)[29], and ii)Spanish (IBS) samples from the 1,000 Genomes Project phase 3 (1000G)[30], and iii)exome data from the IBS and Finnish (FIN) samples from 1000G. SGDP sequenced 279 individuals from 129 different populations located in 7 world regions at a mean coverage of 43.29×. On the other hand, 1000G comprises 2,504 individuals from 26 worldwide populations sequenced at a mean coverage of 7.4×. Upon request to the Spanish National DNA Bank (http://www.bancoadn.org/), we obtained information on the region of birth of the grandparents of the 1000G Spanish samples (IBS), being the same for the four of them in each case.

West Eurasian samples from the SGDP dataset were converted to PLINK binary file system and merged with SEP (hereafter SEP-SGDP). After applying the same quality control described above, SEP-SGDP included 104 individuals and 5,388,964 SNPs. The merging of SEP and IBS datasets is detailed in the section *Coverage downsampling* (below). Finally, the merging of

exome data from SEP, IBS and FIN samples is covered in the section *Analyses - Burden analysis.*

Coverage downsampling

To avoid a potential bias arising from the merging of datasets whose samples were sequenced at different coverage, i.e. SEP and 1000G, we downsampled the SEP sequence reads to match the mean coverage of 1000G (7.4×) using the *DownsampleSam* tool from GATK, with the proper probability and using the strategy named *Chained*. We then applied the same stringent variant calling algorithm described above to both datasets before merging them. 43 IBS individuals were removed due to specially lower calling rates, and after removing markers with calling rates <95%, 5,357,569 markers and 93 individuals remained. This dataset will be referred to, hereafter, as SEP-IBS.

Analyses

Detection of population structure at a macro- and micro-geographic level

We used the SEP-SGDP dataset to generate a matrix of similarity between pairs of individuals using the *1-ibs* algorithm provided by PLINK. A classical multidimensional scaling (MDS) analysis on the matrix was carried out to plot the genetic relationship of SEP with other worldwide samples. We applied the same analysis on the SEP-IBS dataset.

To build further on the genetic situation of SEP on a global context, we ran an analysis on the SEP-SGDP dataset accounting for the haplotype information to increase the discrimination power, using CHROMOPAINTER[31] and fineSTRUCTURE[31] software. This method requires the use of phased genotypes, which were previously obtained with SHAPEIT2[32] using all defaults and a publicly available genetic map based on the 1000G phase 3 sample. Intuitively, CHROMOPAINTER finds the longest possible haplotypes that each individual shares with the remaining samples, and returns similarity matrices of shared haplotype counts and lengths. The haplotype counts matrices are used by fineSTRUCTURE to hierarchically cluster the individuals into a Bayesian phylogenetic tree.

In order to seek differential gene flow processes, the haplotype lengths matrices were analysed with GLOBETROTTER[12] to calculate the ancestral components.

We ran two additional CHROMOPAINTER and fineSTRUCTURE analyses, one with the SEP dataset to look for fine-scale population structure within the Pyrenees, and another with the SEP-IBS dataset to look for differences in the structure patterns between SEP and the Spanish general population.

Identification of genetic barriers

We used the Estimated Effective Migration Surfaces (EEMS)[33] algorithm to obtain an estimate of migration rates between the SEP individuals. The algorithm was run using the similarity matrices produced by CHROMOPAINTER, applying the default parameters and a total of 1,000 demes to conform the surface on which to situate the individuals. EEMS was also applied on SEP-IBS to estimate migration rates between SEP and nearby Spanish samples. In this case, the default parameters and a total of 1,500 demes were specified.

Estimation of the effective population sizes and time of split of SEP populations

To obtain an orientation of the time of split of the sampled SEP populations, we applied two different statistical approaches. First, we used SMC++ 1.15.2[34], which implements a Markovian coalescence algorithm, assuming a generation time of 29 years and a mutation rate of 1.61e-8[35].

Second, we modelled a simple demographic history of SEP populations and used an Aproximate Bayesian Computation (ABC) approach coupled to deep learning (ABC-DL)[36] to estimate the posterior probability distributions of the parameters defining the model. A total of 300,000 simulations were generated using fastsimcoal2[37], each simulating 7,314 genomic regions encompassing ~713 Megabases (Mb) assuming a generation time of 29 years and a mutation rate of 1.61e-8 with a standard deviation of 0.13e-8. 270,000 simulations were used in the training of the artificial neural network (ANN) and the remaining 30,000 in the ABC modelling of the algorithm. In the training step, one sample from each *comarca* was used for generating the observed jSFS, which was repeatedly merged with the normalized simulations

after the addition of random noise to avoid overfitting. A total of 10 independent ANN's featuring 3 neural layers with 100 neurons each were trained using resilient backpropagation for a maximum duration of 2.5 hours, or until an error <0.01 was reached. Then, another set of jSFSs was created using a different sample from each *comarca* and the remaining normalized simulations, so that the trained ANNs predicted the corresponding values for each parameter. This prediction was repeated 100 times with different combinations of individuals to reduce bias. The obtained values were used in a classic ABC approach to obtain the posterior distribution of each parameter. Results are interpreted as a comparison between the prior distribution (defined by the user) and the posterior distribution given by the ABC modelling.

Burden analysis

Runs of homozygosity (RoH) were defined in the individuals from the SEP-IBS dataset using an approach proposed elsewhere[38]. This method estimates the likelihood of a genomic track being homozygote given the population SNP frequencies.

Next, in order to compare the *in silico* predicted burden of deletereous variants in SEP in the context of other Spanish samples and a traditionally considered endogamous population, the Finnish[2], we extracted exome data from the SEP, IBS and FIN datasets. The same variant calling algorithm detailed above was applied. No downsampling of SEP was required in this case, as the coverage values in our data are similar to those of the exome regions in 1000G samples. The merging of SEP, IBS and FIN exome data resulted in 210,529 markers and 235 individuals. We annotated this subset of SNPs using SNPSift[39], and then extracted those predicted to be damaging by Polyphen2[40] (both HVAR and HDIV were used), MutationAssessor[41], and SNPSift. After this filtering, a total of 2,002 SNPs containing a potentially highly deleterious allele remained. We then tested for population-pairwise differences in the burden of deleterious alleles present in each sample.

Results

Genetic variation of SEP in a global context

MDS analysis of SEP-SGDP distributes SEP samples within a Southwest European cluster (Figure 2).

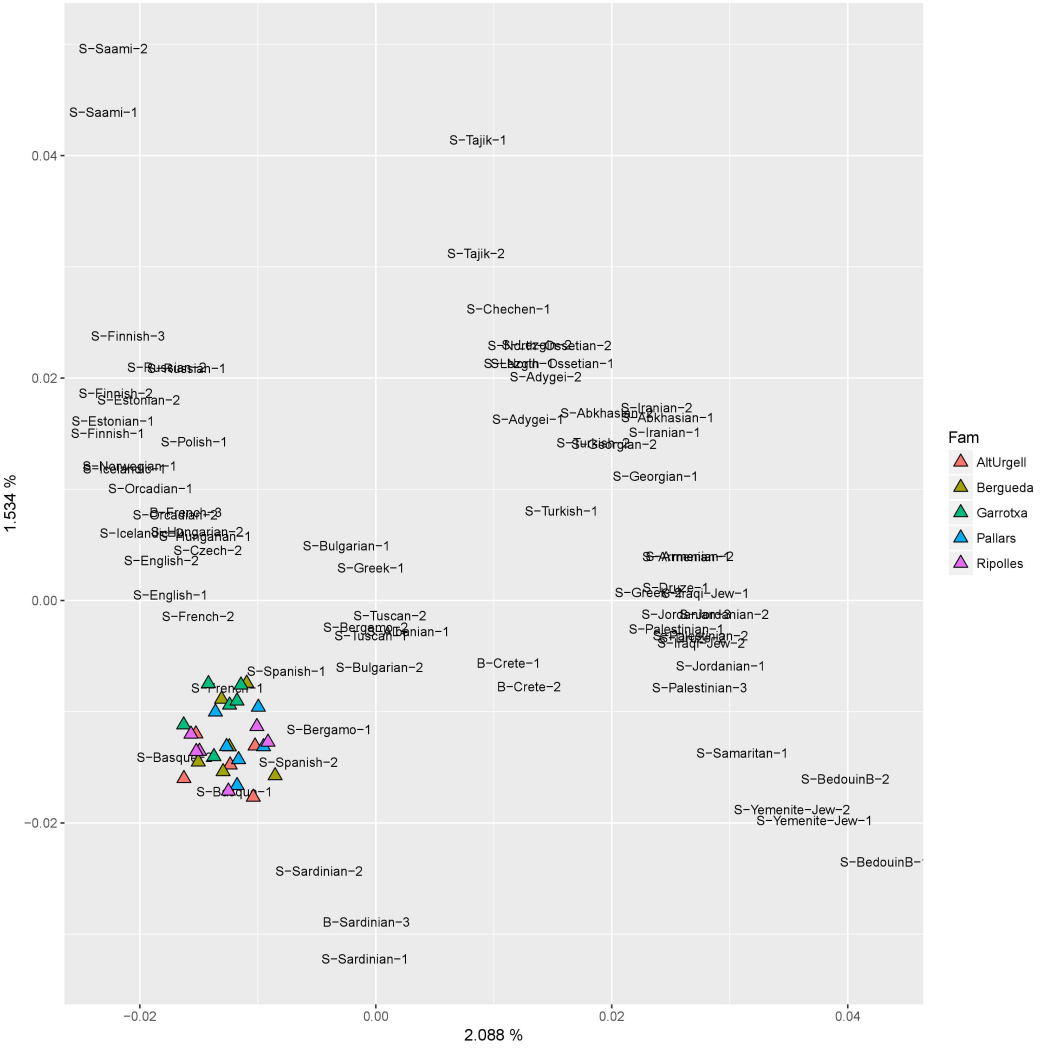


Figure 2. MDS plot showing the relationship of SEP regions with SGDP West Eurasian populations.

The phylogenetic tree generated in the fineSTRUCTURE analysis of SEP-SGDP distributes all SEP samples in a private cluster shared with Basque samples from the French Western Pyrenees, and closely related to other samples from the Iberian peninsula and Sardinia (Figure 3).

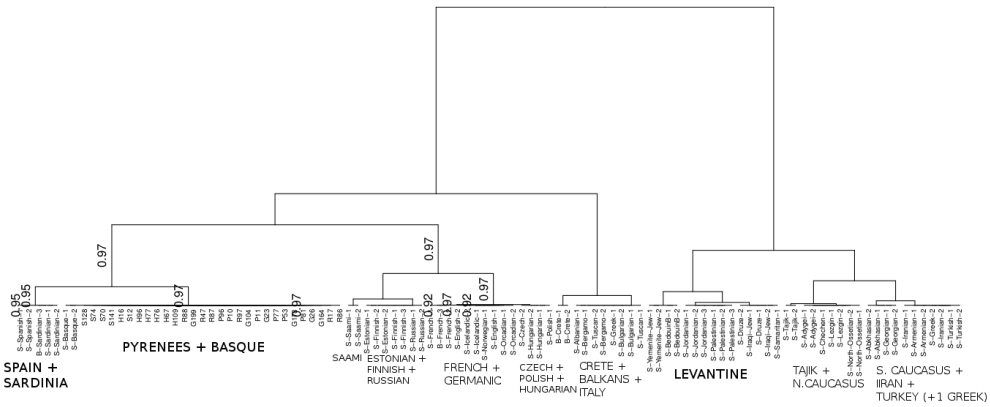


Figure 3. fineSTRUCTURE tree of SEP-SGDP dataset. The general geographic origin of the samples is indicated below each cluster.

The GLOBETROTTER analyses do not support differences between SEP *comarques* regarding their West Eurasian genetic ancestry profiles (Supplementary material 1).

Regarding the fine-STRUCTURE analyses of SEP-IBS, we do not detect population structure in the IBS dataset, while all SEP samples were assigned to a private cluster within the general Spanish population (Supplementary material 2). A fine-STRUCTURE analysis on the original (not downsampled) IBS dataset confirms this lack of population structure (Supplementary material 3).

Fine-scale population structure and demographic history of the Pyrenees

The fine-STRUCTURE analysis of SEP identifies two main geographic clusters that broadly correspond to two set of *comarques*: Garrotxa-Ripolles and Pallars-Alt Urgell-Bergueda (Figure 4).

In order to confirm this genetic barrier by means of a method that takes into account geographic information and includes genetic drift in the model, we ran the EEMS algorithm using SEP data. EEMS identifies a likely migration barrier between the two sets of *comarques* previously detected.

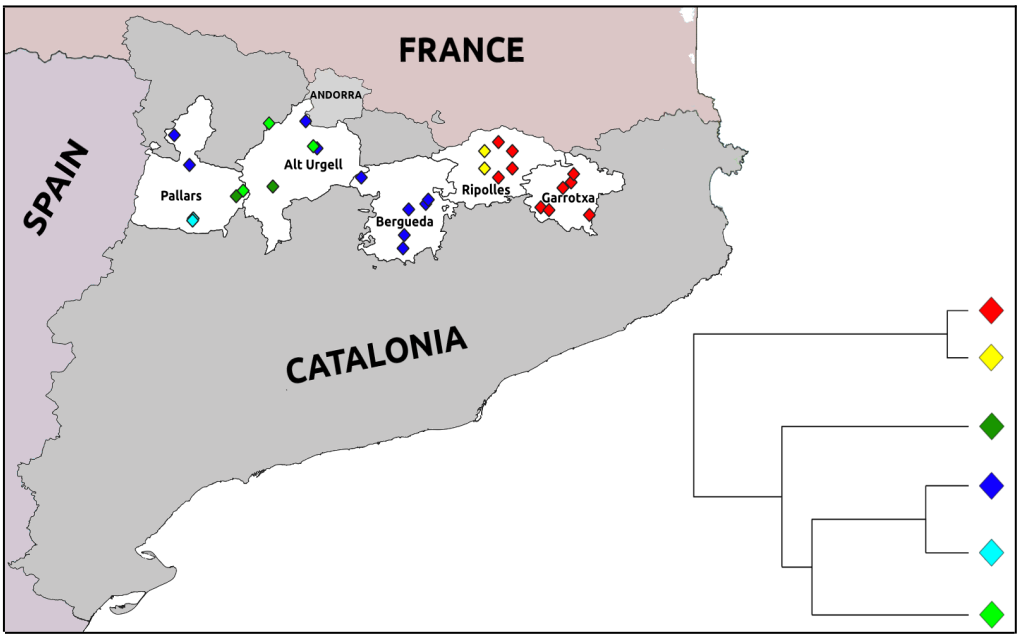


Figure 4. (left) Map of SEP showing the samples on the average coordinates of birth place of their grandparents and painted accordingly to the cluster they belong to. Ripolles samples are artificially dispersed for the sake of clarity since their coordinates overlap. The map was generated using the QGIS software[42]. (right) Simplified fineSTRUCTURE tree of SEP samples, showing six clusters which can be further summarised in two main groups: Garrotxa-Ripollès (red and yellow) and Pallars-Alt Urgell-Bergueda (dark green, dark blue, light blue and light green).

Next, we repeated the EEMS analyses on SEP-IBS, including only the IBS samples that are located geographically close to the SEP samples, namely those from Valencia, Aragon, and Catalonia. In addition to the previous result, EEMS identifies a likely migration barrier between the SEP samples and the IBS samples (see Figure 5). Overall, these analyses support that SEP is a potential genetic isolate compared to other Spanish populations.

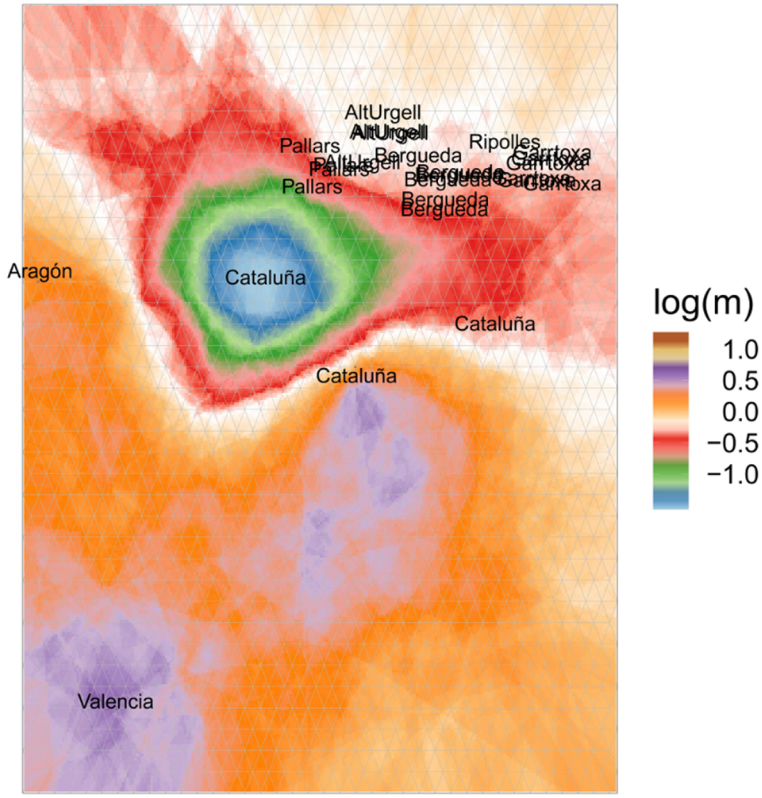


Figure 5. EEMS results show a genetic barrier (colours representing values below zero) between Bergueda and Ripolles. 1500 demes were assumed.

Since genetic isolates tend to show a particular demographic history, we applied different methods to the SEP dataset to describe recent demographic processes. The SMC++ results suggest a split time between the two SEP population sets occurring in ~300PE (Present Era),

95%CI(~1300PE - ~2700BPE). This result was used as a reference for, and refined by, the ABC-DL analyses. ABC-DL results show that the five SEP populations have endured a strong (10-fold) population decline during the last ~100 generations.

Statistics of genetic isolates in the Pyrenees and comparison with IBS

The long-time reduction in effective population size in SEP points to a traditionally endogamous population. To confirm this, we analysed the distribution of RoH in the SEP-IBS dataset, as this is a classical measure for defining genetic isolates. As shown in Figure 6, RoH tend to be longer in the case of the SEP samples than in the general IBS population, thus supporting higher inbreeding in SEP.

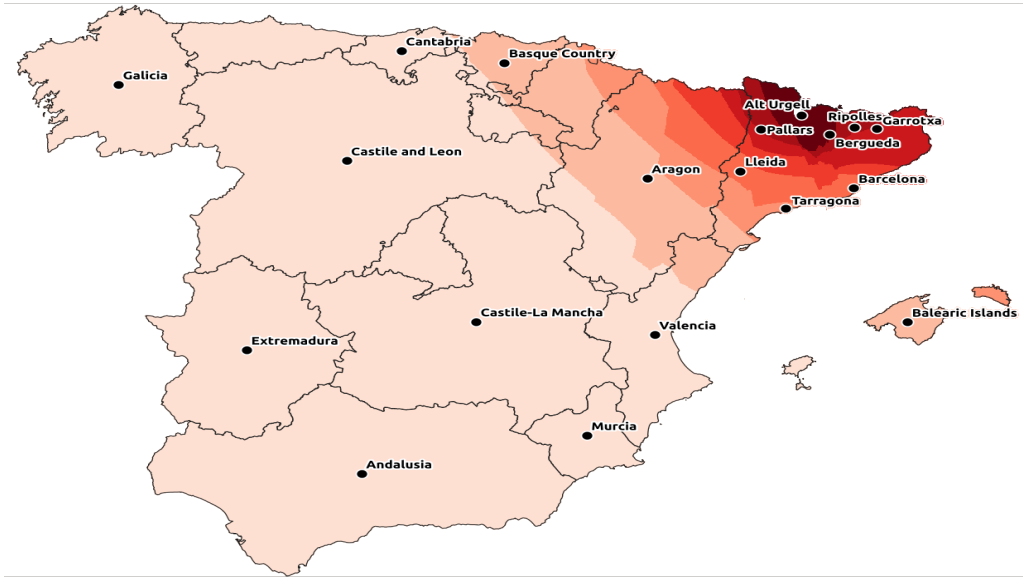


Figure 6. Ordinary kriging with discrete interpolation based on SEP-IBS mean RoH lengths, increasing from white to red. The distribution of mean RoH lengths was normalized using a logarithmic transformation. The respective Comunidad Autónoma, *Province or* comarca *names are shown. The ordinary kriging and maps were generated with the QGIS software.*

Within the SEP dataset, RoHs analyses suggest that the western comarques, namely Pallars and Alt Urgell tend to show an enrichment of RoHs, suggesting that inbreeding could have been more severe in these populations. Further analyses on LD patterns show that Alt Urgell has a smaller LD decay over genomic distance compatible with a higher amount of inbreeding and lower effective population size.

Burden analysis

It has been previously shown that recent human genetic isolates and/or highly inbred populations tend to be enriched for deleterious mutations due to the ineficiency of purifying selection to remove these mutations from the genome[3]. We wondered whether this would apply in the case of SEP, since our previous results suggest that it might be a long-term inbred population. We thus considered 2,002 exomic loci in which the presence of a highly deletereous mutation is predicted. Comparing the distributions of the individual burden of these mutations between SEP, IBS and FIN populations showed no differences regarding the number of variants homozygous for the deletereous mutation. However, in agreement with the categorization of FIN as an isolated and endogamous population, we observed that FIN samples present a higher number of such variants in heterozygosity compared to SEP and IBS samples, and proportional to non-synonymous benign genetic variation in heterozygosity.

In addition, SEP shows an overall depletion of variants in heterozygosity compared to IBS, as well as a lack of association between the number of deleterious and non-deleterious variants per individual. We wondered whether this result could be due to some WGS bias for depletion of rare alleles in SEP when compared to samples from the other datasets, since highly deletereous alleles tend to be rare and both datasets have been generated with different sequencing technologies. Therefore, it could be the case that the lack of association between rare deletereous alleles and benign alleles in SEP compared to the other two populations is a methodological artifact. To account for this putative bias, we randomly matched each SNP containing a highly deleterious allele with a SNP containing a benign mutation in the three populations. We resampled 1,000 times and calculated the Spearman correlation coefficients for each population in the same manner as before. If the observed lack of correlation between the

number of heterozygous variants containing either a deleterious or a benign allele in SEP were due to sequencing bias, we would expect that the association between both variables would increase by this procedure. However, the distribution of correlation coefficients of SEP consists of values significantly non-different from zero, contrary to those of IBS and FIN.

Another concern arises from the fact that, unlike the 1000G dataset, the SEP sample includes only elderly (>70 years old) individuals. Damaging mutations could have prevented reaching such advanced age, and therefore the SEP sample would be biased towards individuals with a low burden of such mutations. However, the highly deletereous condition of these mutations means that they are largely incompatible with life from an early age. This hypothesis is supported by the absence of between-population differences in homozygous rates for this mutations.

Discussion

Genetic isolates are unfrequent in Europe. The European genetic variation has been mostly shaped by large migratory events and by isolation by distance processes, so geographically close individuals tend to be also genetically related and this relationship decays with distance[4,5]. However, genetic isolates are interesting from both a demographic and a medical point of view. In particular, European genetic isolates have been traditionally associated to showh a higher number of rare Mendelian diseases[3] and they can help on identifying genetic variants associated to complex diseases[43].

The potential condition of the Spanish Eastern Pyrenean (SEP) population as a genetic isolate, due to the orography of the Pyrenean mountains, makes it of particular interest. However, so far it has been poorly characterized from a genetic point of view, with the few existing studies using only classical markers[20,23] or microarray data on a small sample size[18]. In this study, we have analysed for the first time the genetic variation of individuals from SEP using next generation sequencing data of high quality.

Our MDS analyses broadly situate the genetic diversity of the SEP samples within the Iberian context (Figure 2), in accordance to the results obtained elsewhere using microarray data[18]. However, when we applied haplotype-based methods, statistically more powerful[31], SEP

samples appear as a distinct group within the Iberian populations, and closely related to the Basque population from the SGDP dataset (Figure 3), as suggested elsewhere[24]. Furthermore, SEP presents fine-scale population structure, with two clearly separated clusters splitting the region into Western (Pallars, Alt Urgell and Bergueda) and Eastern (Ripolles, Garrotxa) subgroups (Figure 4). This is confirmed by the EEMS analyses.

These results contradict previous studies that employ microarray data[18]. This could be due to differences in methodology, since analysing array-based data using methods that do not account for the haplotype information lack in discrimination power. In fact, when we ran fineSTRUCTURE on the SEP dataset using the SNPs common to those used by Biagini et al., we failed to retrieve the observed population structure observed when all the genome was considered (Supplementary material 4). However, using this SNP set in the SEP-IBS dataset still assigns SEP samples to a private cluster within the fineSTRUCTURE tree of IBS samples, albeit in this case the SEP cluster contains also some IBS samples (Supplementary material 5). This could be due to our SEP sample being fully represented by the sample size of 29 individuals, while Biagini et al. incorporates just four samples from this region.

The fine-scale population structure observed in SEP could be caused in part by the orography of the considered geographic region, with long mountainous ranges and river valleys that could have kept both sides of SEP mutually isolated until recent times. Prolonged isolation from neighbouring regions could have reinforced[20] the cultural stratification in reported by classical sources from the 1st millennium BPE. In this line, ABC-DL analyses suggest that the population structure in SEP could have an origin in the 5th century BPE. Moreover, EEMS analyses identify a depletion of migration between SEP and surrounding Spanish populations (Figure 5). Furthermore, the absence of differences between *comarques* with respect to their patterns of Western Eurasian ancestry, as shown in the GLOBETROTTER results, suggests that recent differential gene flow can not explain such stratification (Supplementary material 1). Finally, our ABC-DL analyses suggest that the effective population sizes of the analysed Pyrenean populations have suffered a systematic decline (10-fold) during the past 100 generations. This is in agreement with the observed longer RoH in SEP than in the general Spanish population (Figure 6), which constitutes an indicative of inbreeding.

While similarly low numbers of sites homozygous for rare and highly deletereous mutations suggest the same rates of purifying selection on homozygotes both in IBS and SEP populations, the significantly lower amount of heterozygous sites containing such a mutation in SEP seems to be caused by a general loss of rare variants associated with the pronounced population decline reported here. This agrees with several studies on long-term isolated populations where this results in removal of deleterious alleles[3,44].

In this work, a genetic isolate displaying fine-scale population structure has been identified in mainland Europe, through the analysis of high coverage WGS data accompanied by detailed genealogical information. Further analyses will be required to study if the observed pattern extends to other geographic regions, both at the Spanish and French side, of the Pyrenees.

References

1. Pedersen, C.-E. T. *et al.* The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. *Genetics* **205**, 787–801 (2017).

2. Kääriäinen, H., Muilu, J., Perola, M. & Kristiansson, K. Genetics in an isolated population like Finland: a different basis for genomic medicine? *Journal of Community Genetics* **8**, 319–326 (2017).

3. Simons, Y. B. & Sella, G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current Opinion in Genetics & Development* **41**, 150–158 (2016).

4. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

5. Lao, O. *et al.* Correlation between Genetic and Geographic Structure in Europe. *Current Biology* **18**, 1241–1248 (2008).

6. Humphreys, K. *et al.* The Genetic Structure of the Swedish Population. *PLoS ONE* **6**, e22547 (2011).

7. Lao, O. *et al.* Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history. *Invest Genet* **4**, 9 (2013).

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

9. Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nature Genetics* **50**, 1426–1434 (2018).

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

11. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun* **10**, 551 (2019).

12. Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343**, 747–751 (2014).

13. Mendizabal, I. *et al.* Reconstructing the Population History of European Romani from Genome-wide Data. *Current Biology* **22**, 2342–2349 (2012).

14. Bray, S. M. *et al.* Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proceedings of the National Academy of Sciences* **107**, 16222–16227 (2010).

15. Binzer, S. *et al.* High inbreeding in the Faroe Islands does not appear to constitute a risk factor for multiple sclerosis. *Mult Scler* **21**, 996–1002 (2015).

16. Southam, L. *et al.* Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat Commun* **8**, 15606 (2017).

17. Ebenesersdóttir, S. S. *et al.* Ancient genomes from Iceland reveal the making of a human population. *Science* **360**, 1028–1032 (2018).

18. Biagini, S. A. *et al.* People from Ibiza: an unexpected isolate in the Western Mediterranean. *Eur J Hum Genet* **27**, 941–951 (2019).

19. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet* **21**, 659–665 (2013).

20. Giraldo, M. P. *et al.* Gm and Km alleles in two Spanish Pyrenean populations (Andorra and Pallars Sobira): a review of Gm variation in the Western Mediterranean basin. *Annals of Human Genetics* **65**, 537–548 (2001).

21. Martín, Araceli & Vaquer, Jean. El poblament del Pirineus a l'Holocè, del Mesolític a l'Edat del Bronze. in *Muntanyes i població. El passat dels Pirineus des d'una perpectiva multidisciplinària* (Centre de Trobada de les Cultures Pirinenques, 1995).

22. Riu, M. El poblament dels Pirineus, segles VII-XIV. in *Muntanyes i població. El passat dels Pirineus des d'una perpectiva multidisciplinària* (Centre de Trobada de les Cultures Pirinenques, 1995).

23. Calafell, F. & Bertranpetit, J. Mountains and genes: population history of the Pyrenees. *Human Biology* **66**, 823–42 (1994).

24. López-Parra, A. M. *et al.* In search of the Pre- and Post-Neolithic Genetic Substrates in Iberia: Evidence from Y-Chromosome in Pyrenean Populations. *Annals of Human Genetics* **73**, 42–53 (2009).

25. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analysing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

26. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–498 (2011).

27. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* **4**, 7 (2015).

28. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

29. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

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

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

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

33. Petkova, D., Novembre, J. & Stephens, M. visualising spatial population structure with estimated effective migration surfaces. *Nat Genet* **48**, 94–100 (2016).

34. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**, 303–309 (2017).

35. Lipson, M. *et al.* Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. *PLoS Genet* **11**, e1005550 (2015).

36. Mondal, M., Bertranpetit, J. & Lao, O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nat Commun* **10**, 246 (2019).

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

38. Pemberton, T. J. *et al.* Genomic Patterns of Homozygosity in Worldwide Human Populations. *The American Journal of Human Genetics* **91**, 275–292 (2012).

39. Cingolani, P. *et al*. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Gene.* **3**, (2012).

40. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39**, e118–e118 (2011).

41. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010).

42. QGIS Development Team. *QGIS Geographic Information System*. (Geospatial Foundation Project., 2019).

43. Lim, E. T. *et al.* Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genet* **10**, e1004494 (2014).

44. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**, 220–224 (2014).

OVERALL DISCUSSION OF THE RESULTS

OVERALL DISCUSSION OF THE RESULTS

The human demographic history of the Mediterranean region has long been studied, employing the methodologies available in each era. Nonetheless, there are still numerous unanswered questions that the recent incorporation of massive genetic data analyses is contributing to resolve. In this work, I presented four original research articles that make use of either sparse or dense genetic data for addressing some increasingly complex issues, namely the role of the Mediterranean sea as a historical barrier to gene flow, the impact of sub-Saharan genetic introgression on susceptibility to coronary artery disease (CAD) in Mediterranean populations, the presence of signatures of Sephardic Jewish ancestry in current Iberian populations, and the existence of a population isolate in the Spanish Eastern Pyrenees that shows microgeographical population structure and a depletion of highly deleterious mutations

On the study of the Mediterranean as a historical barrier to gene flow

LIN28B is a gene of clinical relevance due to its reported role in development and cancer. In this study, we assessed the geographic allelic distribution in 24 Mediterranean populations of three SNPs located at, or surrounding, this gene: rs7759938, located in the promoter region; rs314277, in the second intron; and rs221639, in the 3' region. Furthermore, rs7759938 and rs314277 have been related to the age at menarche, pubertal height growth, peripubertal body mass index, levels of prenatal testosterone exposure, and cancer survival. To date, there are no associations found for rs221639.

We discovered significant genetic heterogeneity (F_{CT} = 0.034, p-value < 0.0001) between the North and South coasts of the Mediterranean Basin[211], in consistency with other studies that stress this axis as the main component of genetic differentiation in the region[173,174,214–218]. Natural

selection does not seem to be an explanation for this differentiation, also in agreement with previous studies[219].

The higher MAF of rs221639 in North African groups compared to that of Southern Europe populations could be explained by higher gene flow from sub-Saharan populations, which show a MAF of 0.41 in the 1000G database. This could be also the reason for the higher heterozygosity (*H*) levels observed in North Africa since, on one hand, sub-Saharan populations are characterized by a high *H* compared to other populations and, on the other, the increase in *H* after an admixture event in directly correlated to the proportion of ancestry contributed by the minor group. In addition, the inclusion of sub-Saharan populations from the 1000G in a genetic barrier analysis shows that the main barrier to gene flow between African and European populations is located in the Saharan area, rather than in the Mediterranean Basin.

From a methodological point of view, this work shows that sparse genetic data can provide enough statistical power to detect population structure caused by geographic barriers. Furthermore, *LIN28B* appears to be a useful genomic region for studying genetic variation in Mediterranean populations. In particular, the between-population variation observed for the rs221639 SNP (F_{CT} = 0.101, p-value < 0.0001) is higher than what was previously reported using other loci in the same populations[163], and it lies within the range shown by ancestry informative markers (AIM) in European populations[220], making it a good ancestry informative marker (AIM) candidate in the context of the Mediterranean region.

Finally, it should be pointed out that the use of a low number of markers in population genetics studies can lead to problems such as the presence of outlier populations as a result of genetic drift, and lack of statistical power for detecting more subtle differentiation patterns than the ones observed in this case.

On the sub-Saharan introgression in genomic regions linked to coronary artery disease

Through the comparative study of the variation of 366 SNPs located in four coronary artery disease (CAD)-linked genomic regions (1p13.3, 1q41, 9p21 and 10q11) between Mediterranean populations, in combination with the inclusion of world populations in the analyses (Han Chinese and Yoruba from 1000G), we provide insights into how gene flow of sub-Saharan origin may have played a role in the current geographic distribution of variants associated with CAD in the Mediterranean[212].

The Mediterranean populations studied present a significant sub-Saharan component, which is more prevalent in North Africa than in Southern Europe, in accord with previous studies that suggest a more intense sub-Saharan gene flow into North Africa due to geographic proximity and/or the potential role of the Mediterranean sea as a genetic barrier. Among the European populations, Sicily, Girona, Valencia and Andalusia present a higher sub-Saharan component in the studied genome regions (26%, 22%, 19% and 17%, respectively). This, in agreement with historical North African influence in these regions during the Middle Ages[221], suggests that sub-Saharan introgression in North Africa was secondarily transmitted to European populations through gene flow related to conquest and migration.

D-statistics further suggest potential sub-Saharan introgression into the Mediterranean at the 10q11 region. This region includes *CXCL12*, a gene that codes for a chemokine ligand linked to cardiovascular disease with protective effects[222]. Potential signals of balancing selection have been identified at 10q11[214], implying that natural selection could have maintained this signal of sub-Saharan gene flow in this genomic region by favouring admixed individuals against cardiovascular disease.

The relatively small size of the studied chromosomal regions and the low number of markers, although sufficient to suggest introgression, pose a limitation to the robustness of the results obtained. This could be the reason why none of the *Z*-scores for 10q11 passed the established |

Z-score| > 2 significance threshold, as well as for the lack of an optimal cross-validation value in the ADMIXTURE analysis, warrantying caution at interpretation. In addition, the studied loci are not the only ones associated with CAD, and another potential bias comes from the fact that the SNPs used in the analyses were ascertained primarily in European populations. Future work could address efficiently the above issues by analysing a higher number of CAD regions and markers, preferably ascertained in unbiased sequence data.

On the signatures of Sephardic ancestry in Iberian populations

Using genome-wide array-based data, we detected Sephardic ancestry in some of the studied populations from the Iberian Peninsula and surrounding regions, suggesting a non-homogenous presence of the Sephardic admixture. Namely, a significant Sephardic component is present in Andalusia (12.3%; 95%CI: 11.1-13.5%), Galicia and Portugal (11.3%; 95%CI: 10.6-12.1%), North Italy (5.9%; 95%CI: 5-6.8%), and Tunisia (24.1%; 95%CI: 23.2-25%), but not in the Basque Country, Catalonia or South France. The high Sephardic ancestry in Galicia and Portugal could be reflecting the ban on the departure of Jews from the kingdom of Portugal combined with greater pressure to convert to Catholicism and/or Sephardic gene flow from Spain to Portugal[223,224]. Our mean Sephardic admixture estimate in the Iberian Peninsula is 6%, against the 19.8% reported elsewhere for Y chromosome data[225]. Interestingly, our analyses did not replicate Adams et al. findings in Basque Country and Catalonia, as we did not obtain a significantly non-zero Sephardic component in these populations. Our overall more conservative results can be explained by a potential inflation of ancestry estimates based on uniparental haplogroups, due to genetic drift or haplogroup non-specificity.

On the other hand, the main ancestry contributors for Tunisia were the Berbers (55.8%; 95%CI: 55.3-56.3%) followed by the Sephardic Jews, in agreement with historical records that show that many Sephardic Jews sought shelter in the Maghreb after they were expelled from Spain and Portugal[190]. A Berber component was also present in all the Iberian populations except

Catalonia, matching historical knowledge about the Moorish presence in the Iberian Peninsula during the Middle Ages[170], as well as results from a recent study in the same region[179]. In this line, local ancestry results show that Berber ancestry haplotypes in the Iberian Peninsula samples are notably shorter than Sephardic ancestry haplotypes, further suggesting that Sephardic admixture in the Iberian Peninsula is notably more recent than Muslim admixture. Given that the presence of Sephardic populations in the Peninsula predates the Arab invasion, this can be interpreted as Sephardic admixture occurring in more recent times, perhaps as intermarriage promoted by the persecution and consecutive conversion. This is corroborated by Sephardic gene flow following an out-of-Iberia pattern along Southwestern Europe, although more samples from their potential route to the East are warranted to draw clearer conclusions. Even though historical accounts show that the Sephardim were present in Tunisia, the high genetic contribution that we find in this region is likely to be inflated, since the rate of conversion of Sephardic Jews into the local Muslim population was low and intermarriage between Jews and Muslims has been limited throughout their history[226]. Rather, we attribute this high percentage to a lack of sufficient specificity in the Sephardic genetic signature to resolve North African ancestry in the absence of a more appropriate Levantine reference population that reflects better the historical background of Tunisia (e.g. a proxy population for the Phoenicians). In Antiquity, the Phoenicians not only had a major capital in Tunisia (Carthage), but also significant outposts in the Iberian Peninsula. As a Levantine population, the Phoenicians were probably genetically very similar to the ancestral Jewish populations of the Mediterranean. In addition, another possible limitation in our North African analysis is the lack of a sub-Saharan reference to account for the important gene flow rates suggested in the previous points.

Recent whole-genome studies on the structure of worldwide Jewish groups have shown that current Sephardic populations are more related to other Jewish and Middle Eastern groups than to non-Jewish Spaniards and Northwest Africans[176,178,227]. This implies that Sephardic Jewish haplotypes have a sufficiently specific genetic signature to be detected in the context of the Iberian genetic pool. PCA did not detect the Sephardic genetic influence in the Iberian Peninsula, yet haplotype-based methods were more powerful at picking up these admixture signatures. We therefore believe that haplotype-based methods are more appropriate for the study of Sephardic admixture in follow-up studies. Future research to test more specific

hypotheses about the outcomes of the persecution, conversion, and expulsion of the Sephardim should include a broader geographic scope, encompassing not only the Mediterranean, but also other parts of the globe[228] and, ideally, denser SNP arrays or WGS data.

On the micro-geographic population structure of the Spanish Eastern Pyreneans and their status of genetic isolate

By use of the West Eurasian populations of the SGDP as an external reference, Chromopainter and fineSTRUCTURE analyses show the Spanish Eastern Pyreneans as a distinct genetic group within the Iberian populations, as well as closely related to the Basque population, as suggested elsewhere[229]. A similar analysis, in this case including the Spanish samples from the 1000G (IBS), also shows the Spanish Eastern Pyreneans as a differentiated cluster. Furthermore, employing the genetic relationship matrices (GRM) generated by Chromopainter, the Estimated Effective Migration Surfaces (EEMS) algorithm[230] reveals lower than expected migration rates between the Spanish Eastern Pyrenean samples and nearby samples of the IBS dataset, suggesting isolation.

The isolation hypothesis is further supported by i)the presence in the Spanish Eastern Pyrenean samples of overall RoH lengths within the range of those of the 1000G Finnish samples (FIN), thus longer than in the general Spanish population, which suggests higher inbreeding in the Spanish Eastern Pyreneans than in the former, and ii)a systematic decline (10-fold) in the population size of the Spanish Eastern Pyreneans during the past 100 generations, indicated by the SFS-based ABC-DL analysis.

By applying the haplotype-based methods mentioned above on the Spanish Eastern Pyrenean samples exclusively, we identified two clearly separated clusters that split the region into Western (Pallars, Alt Urgell and Berguedà) and Eastern (Ripollès, Garrotxa) subgroups, with lower than expected migration rates between them. This population differentiation could be

caused in part by the geography of the Spanish Eastern Pyrenees, with long mountainous ranges and river valleys that could have kept both sides mutually isolated until recent times. Prolonged isolation could have reinforced[231] the cultural stratification in this region reported by classical sources (1st millennium BPE). In this line, ABC-DL situates the estimates of the time of population differentiation in SEP in the 5th century BPE. In addition, the GLOBETROTTER ancestry analysis, based on the GRM of the Spanish Eastern Pyrenean samples combined with the SGDP dataset, shows similar admixture patterns for the two Spanish Eastern Pyrenean subgroups, which suggests that no recent differential gene flow seems to have caused such stratification.

A similar number of SNPs homozygous for rare deleterious mutations suggests the same rates of purifying selection on homozygotes both in IBS and the Spanish Eastern Pyrenean populations. On the other hand, the significantly lower number of heterozygous sites containing such mutations in the Spanish Eastern Pyreneans seems to be caused by a general loss of rare variants associated with a pronounced population decline, in agreement with several studies on long-term isolated populations[65,66]. However, these studies report that this is accompanied by an increase in frequency of certain rare recessive mutations, which causes higher incidence of specific rare genetic diseases and raises the overall burden of highly deleterious mutations. In the present study, the latter is not the case, which could be attributed to the artifactual absence of such markers in our dataset.

In conclusion, through the analysis with haplotype and SFS-based methods of high coverage WGS data accompanied by detailed genealogical information, we have been able to identify a new genetic isolate in mainland Europe, which further shows patterns of fine-scale population structure.

CONCLUSIONS

CONCLUSIONS

- Sparse genetic data provide enough power to detect the North-South axis as the main component of genetic differentiation in the Mediterranean, as reported elsewhere.

- The rs221639 polymorphism could be a good ancestry-informative marker candidate in the context of the Mediterranean region.

- The presence of higher heterozygosity levels, as well as a larger sub-Saharan ancestry, in North African populations compared to those observed in Southern Europe, are attributable to higher gene flow from sub-Saharan Africa due to geographic proximity and/or the potential role of the Mediterranean sea as a genetic barrier.

- A sub-Saharan ancestral component is detectable in Mediterranean populations at four sparsely genotyped genomic regions previously associated with coronary artery disease. *D*-statistics further suggest possible sub-Saharan introgression into the genomic region 10q11, that encompasses the chemokine ligand coding gene *CXCL12*, and for which potential signals of balancing selection have been identified.

- Using genome-wide array-based data, we detect a genetic component of Sephardic origin in the Iberian populations of Northern Portugal, Galicia, and Andalusia. These results are a more conservative revision of a previous study based on Y chromosome haplogroups.

- More recent Sephardic than Berber admixture in the Iberian Peninsula suggests that the admixture between Jewish and Christian groups took place fundamentally during, and in posteriority to, the persecution and mass forced conversions.

- Sephardic admixture dates follow an out-of-Iberia dispersal pattern, confirming a Sephardic diaspora route via Southwest Europe.

- Haplotype-based analyses on high-coverage whole-genome sequence data show the Spanish Eastern Pyrenean population as a distinct group within an Iberian context, closely related to the Basques.

- The Spanish Eastern Pyrenees region is genetically divided into two clearly separated Western and Eastern subgroups. This genetic stratification is estimated to have taken place in the 5th century BPE.

- The classification of the Spanish Eastern Pyreneans as a population isolate is supported by the detection of a genetic barrier between the Spanish Eastern Pyrenees and the rest of Spain, a 10-fold decline in the effective population size during the last 100 generations, and the presence of longer runs of homozygosity in the Spanish Eastern Pyrenean population.

- The depletion of heterozygotes at loci containing a deleterious mutation in the Spanish Eastern Pyreneans seems to be in agreement with several studies that show increased purifying selection rates affecting many rare, highly deleterious mutations in populations that have undergone long-term bottlenecks and isolation.

REFERENCES

1. Fisher, R. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433 (1918).
2. Edwards, A. W. F. G. H. Hardy (1908) and Hardy–Weinberg Equilibrium. *Genetics* **179**, 1143–1150 (2008).
3. Crow, J. F. Population genetics history: A personal view. *Ann. Rev. Genet* **21**, 1–22 (1987).
4. Avery, O. T. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine* **79**, 137–57 (1943).
5. Francis, C. & Watson, J. D. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–38 (1953).
6. Sanger, F. & Coulson, A. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441–6 (1975).
7. Bartlett, J. M. S. & Stirling, D. A Short History of the Polymerase Chain Reaction. *Methods in Molecular Biology* **226**, 3–6 (2003).
8. Kimura, M. *The neutral theory of molecular evolution*. (1983).
9. Kingman, J. F. C. The coalescent. *Stochastic Processes and their Applications* **13**, 235–248 (1982).
10. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, (1987).
11. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes*. (Princeton Univ. Press, 1994).
12. Jorde, L. B. & Wooding, S. P. Genetic variation, classification and 'race'. *Nature Genetics* **36**, S28–S33 (2004).
13. Jobling, M. A., Hollox, E., Hurles, M. E., Kivisild, T. & Tyler-Smith, C. *Human Evolutionary Genetics*. (Garland Science, 2014).
14. Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proceedings of the National Academy of Sciences* **109**, 17758–17764 (2012).
15. Shampo, M. A. & Kyle, R. A. J. Craig Venter—The Human Genome Project. *Mayo Clinic Proceedings* **86**, e26–e27 (2011).
16. Chial, H. DNA sequencing technologies key to the Human Genome Project. *Nature Education* **1**, 219 (2008).
17. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
18. Venter, J. C. *et al.* The Sequence of the Human Genome. *THE HUMAN GENOME* **291**, 51 (2001).
19. Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
20. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338–345 (2018).

21. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
22. López Herráez, D. *et al.* Genetic Variation and Recent Positive Selection in Worldwide Human Populations: Evidence from Nearly 1 Million SNPs. *PLoS ONE* **4**, e7888 (2009).
23. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
24. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
25. Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238–242 (2016).
26. Gravel, S. *et al.* Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS Genetics* **9**, e1004023 (2013).
27. Derkach, A. *et al.* Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics* **30**, 2179–2188 (2014).
28. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48**, 1443–1448 (2016).
29. van Leeuwen, E. M. *et al.* Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in *ANGPTL4* determining fasting TG levels. *Journal of Medical Genetics* **53**, 441–449 (2016).
30. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
31. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**, 379–389 (2013).
32. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435–444 (2015).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
35. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
36. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* **4**, 7 (2015).
37. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).
38. Cerami, E. SAMtools: Primer / Tutorial. *eureka* (2013).
39. Gusev, A., Mandoiu, I. I. & Pasaniuc, B. Highly Scalable Genotype Phasing by Entropy Minimization. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **5**, 252–261 (2008).

40. Osborn-Gustavson, A. E., McMahon, T., Josserand, M. & Spamer, B. J. The Utilization of Databases for the Identification of Human Remains. in *New Perspectives in Forensic Human Skeletal Identification* 129–139 (Elsevier, 2018). doi:10.1016/B978-0-12-805429-1.00012-0
41. Battaggia, C. *et al.* Detecting Sex-Biased Gene Flow in African-americans Through the Analysis of Intra- and Inter-Population Variation at Mitochondrial DNA and Y-Chromosome Microsatellites. *Balkan Journal of Medical Genetics* **15**, 7–34 (2012).
42. Voight, B. F. *et al.* The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genetics* **8**, e1002793 (2012).
43. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108**, 11983–11988 (2011).
44. Lachance, J. & Tishkoff, S. A. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it: Prospects & Overviews. *BioEssays* **35**, 780–786 (2013).
45. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
46. Rogers, A. R. & Jorde, L. B. Ascertainment Bias in Estimates of Average Heterozygosity. *Am. J. Hum. Genet.* 9 (1996).
47. Eller, E. Effects of ascertainment bias on recovering human demographic history. *Human Biology* **73**, 411–27 (2001).
48. Guillot, G. & Foll, M. Correcting for ascertainment bias in the inference of population structure. *Bioinformatics* **25**, 552–554 (2009).
49. Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution* **27**, 2534–2547 (2010).
50. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).
51. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
52. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **34**, 666–681 (2018).
53. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry* **55**, 641–658 (2009).
54. Stein, L. D. The case for cloud computing in genome informatics. *Genome Biology* **11**, 207 (2010).
55. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics* **44**, 631–635 (2012).
56. Gilly, A. *et al.* Very low-depth whole-genome sequencing in complex trait association studies. *Bioinformatics* 1–7 (2019). doi:10.1093/bioinformatics/bty1032

57. Salzberg, S. L. & Yorke, J. A. Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320–4321 (2005).
58. Liu, L. *et al.* Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* **2012**, 12 (2014).
59. Wu, S. H., Schwartz, R. S., Winter, D. J., Conrad, D. F. & Cartwright, R. A. Estimating error models for whole genome sequencing using mixtures of Dirichlet-multinomial distributions. *Bioinformatics* **33**, 2322–2329 (2017).
60. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nature Methods* **12**, 351–356 (2015).
61. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* **13**, 278–289 (2015).
62. Ramachandran, S. *et al.* 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**, 15942–15947 (2005).
63. Jay, F., Sjödin, P., Jakobsson, M. & Blum, M. G. B. Anisotropic Isolation by Distance: The Main Orientations of Human Genetic Differentiation. *Molecular Biology and Evolution* **30**, 513–525 (2013).
64. Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343**, 747–751 (2014).
65. Simons, Y. B. & Sella, G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current Opinion in Genetics & Development* **41**, 150–158 (2016).
66. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**, 220–224 (2014).
67. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
68. Graur, D. *et al.* On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and Evolution* **5**, 578–590 (2013).
69. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences* **110**, 5294–5300 (2013).
70. Coop, G. Notes on Population Genetics. *The Coop Lab. Population and Evolutionary Genetics. UC Davis* (2013).
71. Booker, T. R., Jackson, B. C. & Keightley, P. D. Detecting positive selection in the genome. *BMC Biol* **15**, 98 (2017).
72. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet.* **47**, 97–120 (2013).
73. Ko, K. H. Origins of human intelligence: The chain of tool-making and brain evolution. *Anthropological Notebooks* 18 (2016).

74. Wilson, M. L., Miller, C. M. & Crouse, K. N. Humans as a model species for sexual selection research. *Proceedings of the Royal Society B: Biological Sciences* **284**, 20171320 (2017).
75. Daanen, H. A. M. & Van Marken Lichtenbelt, W. D. Human whole body cold adaptation. *Temperature* **3**, 104–118 (2016).
76. Jablonski, N. G. & Chaplin, G. Human Skin Pigmentation as an Adaptation to UV Radiation. in *In the Light of Evolution* **Volume IV: The Human Condition**, (National Academies Press, 2010).
77. Peng, Y. *et al.* Genetic Variations in Tibetan Populations and High-Altitude Adaptation at the Himalayas. *Molecular Biology and Evolution* **28**, 1075–1081 (2011).
78. Ilardo, M. A. *et al.* Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell* **173**, 569-580.e15 (2018).
79. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat Rev Genet* **15**, 379–393 (2014).
80. Luca, F., Perry, G. H. & Di Rienzo, A. Evolutionary Adaptations to Dietary Changes. *Annu. Rev. Nutr.* **30**, 291–314 (2010).
81. Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).
82. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
83. Yi, X. *et al.* Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329**, 75–78 (2010).
84. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc Natl Acad Sci USA* **112**, 3439–3444 (2015).
85. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *eLife* **6**, 17 (2017).
86. Pennisi, E. Human mutation rate a legacy from our past. *Science* **360**, 143–143 (2018).
87. Gao, Z. *et al.* Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc Natl Acad Sci USA* **116**, 9491–9500 (2019).
88. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276 (1975).
89. Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nat Rev Genet* **16**, 33–44 (2015).
90. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).
91. Nei, M. Genetic Distance between Populations. *The American Naturalist* **106**, 283–292 (1972).
92. Reynolds, J., Weir, B. S. & Cockerham, C. C. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* **105**, 767–779 (1983).

93. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* (1987). doi:10.1093/oxfordjournals.molbev.a040454
94. Excoffier, L. & Smouse, P. E. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics* **131**, 479–91 (1992).
95. Arenas, M., François, O., Currat, M., Ray, N. & Excoffier, L. Influence of Admixture and Paleolithic Range Contractions on Current European Diversity Gradients. *Molecular Biology and Evolution* **30**, 57–61 (2013).
96. Novembre, J. & Peter, B. M. Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics & Development* **41**, 98–105 (2016).
97. Padhukasahasram, B. Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics* **5**, (2014).
98. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 15 (2000).
99. Liu, Y. *et al.* Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics* **7**, 1 (2013).
100. Frantz, A. C., Cellina, S., Krier, A., Schley, L. & Burke, T. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology* **46**, 493–505 (2009).
101. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics* **82**, 290–303 (2008).
102. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* **93**, 278–288 (2013).
103. Kidd, J. M. *et al.* Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation. *The American Journal of Human Genetics* **91**, 660–671 (2012).
104. Johnson, N. A. *et al.* Ancestral Components of Admixed Genomes in a Mexican Cohort. *PLoS Genet* **7**, e1002410 (2011).
105. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet* **8**, e1002453 (2012).
106. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
107. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
108. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722 (2010).
109. Burgarella, C. *et al.* Adaptive Introgression: An Untapped Evolutionary Mechanism for Crop Adaptation. *Front. Plant Sci.* **10**, 4 (2019).

110. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat Rev Genet* **16**, 359–371 (2015).
111. Loh, P.-R. *et al.* Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics* **193**, 1233–1254 (2013).
112. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
113. Myers, S. Understanding human admixture, and association mapping in admixed populations.
114. Didelot, X. Short course on statistical population genetics. *Site of the Department of Statistics of the University of Oxford* (http://www.stats.ox.ac.uk/~didelot/popgen/Chapter3.pdf).
115. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1387–1393 (2005).
116. Wiuf, C. & Hein, J. Recombination as a Point Process along Sequences. *Theoretical Population Biology* **55**, 248–259 (1999).
117. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
118. Nielsen, R. & Beaumont, M. A. Statistical inferences in phylogeography. *Molecular Ecology* **18**, 1034–1047 (2009).
119. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* **46**, 919–925 (2014).
120. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**, 303–309 (2017).
121. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun* **9**, 3258 (2018).
122. Peng, B., Amos, C. I. & Kimmel, M. Forward-Time Simulations of Human Populations with Complex Diseases. *PLoS Genetics* **3**, 14 (2007).
123. Peng, B. & Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686–3687 (2005).
124. Carvajal-Rodriguez, A. Simulation of Genomes: A Review. *CG* **9**, 155–159 (2008).
125. Liu, Y., Athanasiadis, G. & Weale, M. E. A survey of genetic simulation software for population and epidemiological studies. *Human Genomics* **3**, 79 (2008).
126. Peng, B. *et al.* Genetic Data Simulators and their Applications: An Overview. *Genet. Epidemiol.* **39**, 2–10 (2015).
127. Busetto, A. G. & Buhmann, J. *Stable Bayesian Parameter Estimation for Biological Dynamical Systems*. (IEEE Computer Society Press, 2009).
128. Sunnåker, M. *et al.* Approximate Bayesian Computation. *PLoS Comput Biol* **9**, e1002803 (2013).

129. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics* **5**, e1000695 (2009).
130. Mondal, M., Bertranpetit, J. & Lao, O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nat Commun* **10**, 246 (2019).
131. Flagel, L., Brandvain, Y. & Schrider, D. R. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution* **36**, 220–238 (2019).
132. Sheehan, S. & Song, Y. S. Deep Learning for Population Genetic Inference. *PLOS Computational Biology* 28 (2016).
133. Serjeant, G. R. One hundred years of sickle cell disease: Review. *British Journal of Haematology* **151**, 425–429 (2010).
134. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences* **107**, 1752–1756 (2010).
135. Lewis, C. M. & Knight, J. Introduction to Genetic Association Studies. *Cold Spring Harbor Protocols* **2012**, pdb.top068163-pdb.top068163 (2012).
136. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nat Rev Genet* **2**, 91–99 (2001).
137. Suvarna, S. Using TensorFlow to conduct simple Linear Regression. *Towards data science* (2019).
138. Chávez, G. Understanding Logistic Regression step by step. *Towards data science* (2019).
139. Duncan, L. E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacol.* (2019). doi:10.1038/s41386-019-0389-5
140. Hong, E. P. & Park, J. W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform* **10**, 117 (2012).
141. Yamauchi, T. *et al.* A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nat Genet* **42**, 864–868 (2010).
142. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet* **9**, e1003348 (2013).
143. Marian, A. J. Molecular genetic studies of complex phenotypes. *Translational Research* **159**, 64–79 (2012).
144. Hublin, J.-J. *et al.* New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* **546**, 289–292 (2017).
145. Harvati, K. *et al.* Apidima Cave fossils provide earliest evidence of Homo sapiens in Eurasia. *Nature* (2019). doi:10.1038/s41586-019-1376-z

146. Arsuaga, J. L. *et al.* Neandertal roots: Cranial and chronological evidence from Sima de los Huesos. *Science* **344**, 1358–1363 (2014).
147. Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504–507 (2016).
148. Freidline, S. E., Gunz, P., Janković, I., Harvati, K. & Hublin, J. J. A comprehensive morphometric analysis of the frontal and zygomatic bone of the Zuttiyeh fossil from Israel. *Journal of Human Evolution* **62**, 225–241 (2012).
149. Pagani, L. *et al.* Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *The American Journal of Human Genetics* **96**, 986–991 (2015).
150. Rebollo, N. R. *et al.* New radiocarbon dating of the transition from the Middle to the Upper Paleolithic in Kebara Cave, Israel. *Journal of Archaeological Science* **38**, 2424–2433 (2011).
151. Benazzi, S. *et al.* Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature* **479**, 525–528 (2011).
152. Olivieri, A. *et al.* The mtDNA Legacy of the Levantine Early Upper Palaeolithic in Africa. *Science* **314**, 1767–1770 (2006).
153. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
154. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics* **8**, e1002947 (2012).
155. Higham, T. *et al.* The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* **512**, 306–309 (2014).
156. Finlayson, C. *et al.* Late survival of Neanderthals at the southernmost extreme of Europe. *Nature* **443**, 850–853 (2006).
157. Banks, W. E. *et al.* Neanderthal Extinction by Competitive Exclusion. *PLoS ONE* **3**, e3972 (2008).
158. Houldcroft, C. J. & Underdown, S. J. Neanderthal genomics suggests a pleistocene time frame for the first epidemiologic transition: NEANDERTHAL INFECTIOUS DISEASE GENETICS. *American Journal of Physical Anthropology* **160**, 379–388 (2016).
159. Staubwasser, M. *et al.* Impact of climate change on the transition of Neanderthals to modern humans in Europe. *Proceedings of the National Academy of Sciences* **115**, 9116–9121 (2018).
160. Haber, M., Mezzavilla, M., Xue, Y. & Tyler-Smith, C. Ancient DNA and the rewriting of human history: be sparing with Occam's razor. *Genome Biol* **17**, 1 (2016).
161. Mangerud, J. Ice-dammed lakes and rerouting of the drainage of northern Eurasia during the Last Glaciation. *Quaternary Science Reviews* **23**, 1313–1332 (2004).
162. Tallavaara, M., Luoto, M., Korhonen, N., Järvinen, H. & Seppä, H. Human population dynamics in Europe over the Last Glacial Maximum. *Proceedings of the National Academy of Sciences* **112**, 8232–8237 (2015).

163. Henn, B. M. *et al.* Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLoS Genetics* **8**, e1002397 (2012).
164. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
165. Schulz Paulsson, B. Radiocarbon dates and Bayesian modeling support maritime diffusion model for megaliths in Europe. *Proc Natl Acad Sci USA* **116**, 3460–3465 (2019).
166. Kristiansen, K. *Europe before history*. (Cambridge University Press, 1998).
167. Harden, D. *The Phoenicians*. (Frederick A. Praeger, 1962).
168. Koch, J. T. *Celtic culture. A historical encyclopedia*. (ABC-CLIO, Inc., 2006).
169. Prowse, T. L. *et al.* Isotopic evidence for age-related immigration to imperial Rome. *Am. J. Phys. Anthropol.* **132**, 510–519 (2007).
170. Watt, M. *A History of Islamic Spain*. (1967).
171. Mango, C. *The Oxford History of Byzantium*. (Oxford University Press, 2004).
172. Nacu, A. The Roman Empire in 117 AD. *Wikimedia Commons* (2008).
173. Athanasiadis, G. *et al.* The Mediterranean Sea as a barrier to gene flow: evidence from variation in and around the F7 and F12 genomic regions. *BMC Evolutionary Biology* **10**, 84 (2010).
174. Athanasiadis, G. & Moral, P. Spatial principal component analysis points at global genetic structure in the Western Mediterranean. *Journal of Human Genetics* **58**, 762–765 (2013).
175. Botigue, L. R. *et al.* Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proceedings of the National Academy of Sciences* **110**, 11791–11796 (2013).
176. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (2010).
177. Mendizabal, I. *et al.* Reconstructing the Population History of European Romani from Genome-wide Data. *Current Biology* **22**, 2342–2349 (2012).
178. Haber, M. *et al.* Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture. *PLoS Genetics* **9**, e1003316 (2013).
179. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun* **10**, 551 (2019).
180. Biagini, S. A. *et al.* People from Ibiza: an unexpected isolate in the Western Mediterranean. *Eur J Hum Genet* **27**, 941–951 (2019).
181. Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nature Genetics* **50**, 1426–1434 (2018).
182. Sánchez-Quinto, F. *et al.* North African Populations Carry the Signature of Admixture with Neandertals. *PLoS ONE* **7**, e47765 (2012).
183. Fadhlaoui-Zid, K. *et al.* Genome-Wide and Paternal Diversity Reveal a Recent Origin of Human Populations in North Africa. *PLoS ONE* **8**, e80293 (2013).

184. Arauna, L. R. *et al.* Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. *Molecular Biology and Evolution* msw218 (2016). doi:10.1093/molbev/msw218
185. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
186. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
187. Pimenta, J., Lopes, A. M., Comas, D., Amorim, A. & Arenas, M. Evaluating the Neolithic Expansion at Both Shores of the Mediterranean Sea. *Molecular Biology and Evolution* **34**, 3232–3242 (2017).
188. Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019).
189. Smith, M. *The Early History of God: Yahweh and Other Deities of Ancient Israel*. (Eerdman's, 2002).
190. Gerber, J. *Jews of Spain: a History of the Sephardic Experience*. (Simon and Schuster, 1994).
191. Riu, M. El poblament dels Pirineus, segles VII-XIV. in *Muntanyes i població. El passat dels Pirineus des d'una perpectiva multidisciplinària* (Centre de Trobada de les Cultures Pirinenques, 1995).
192. Pérez, J. *History of a tragedy. The expulsion of Spanish Jews*. (Crítica, 1993).
193. Valdeón-Baruque, J. El reinado de los Reyes Católicos. Época crucial del antijudaísmo español. in *El antisemitismo en España* (diciones de la Universidad de Castilla-La Mancha, 2007).
194. Mazower, M. *Salonica, City of Ghosts: Christians, Muslims and Jews 1430-1950*. (Vintage, 2006).
195. Sachar, H. M. *The world of the Sephardim remembered*. (Alfred A. Knopf Inc., 1994).
196. Soyer, F. *The Persecution of the Jews and Muslims of Portugal. King Manuel I and the End of Religious Tolerance*. (Brill, 2007).
197. Lambert, M. L. Encyclopaedia Judaica. **14**, 1165–1166
198. Bertranpetit, J., Moral, P. & Calafell, F. Present i passat de les poblacions del Pirineus: Una aproximació des de la genètica. in *Muntanyes i població. El passat dels Pirineus des d'una perpectiva multidisciplinària* (Centre de Trobada de les Cultures Pirinenques, 1995).
199. Fullola, J. M., Garcia-Argüelles, P., Serrat, D. & Bergadà, M. M. El Paleolític i l'Epipaleolític al vessant meridional del Pirineus catalans. Vint anys de recerca a la franja pirinenca sud; interrelacions amb les àrees cicumdants. in (1994).
200. Martín, A. & Vaquer, J. El poblament del Pirineus a l'Holocè, del Mesolític a l'Edat del Bronze. in *Muntanyes i població. El passat dels Pirineus des d'una perpectiva multidisciplinària* (Centre de Trobada de les Cultures Pirinenques, 1995).

201. Guilaine, J. La Balma de la Margineda. Les dossiers. *Histoire et archéologie* **96**, 9–34 (1988).
202. Montserrat, J. M. Evolución glaciar y postglaciar del clima y la vegetación en la vertiente sur del Pirineo: estudio palinológico. *Monografías del Instituto Pirenaico de Ecología* **6**, 147 (1992).
203. Guilaine, J. & Vaquer, J. Les débuts de la métallurgie et les groupes culturels de la fin du Néolithique dans le sud de la France (Languedoc, Causses, Pyrénées). in *Proceedings of the fifth Atlantic Colloquium* 65–79 (1979).
204. Delibes, G., Fernández, M., Martín, A. & Molina, F. El calcolítico en la península ibérica. in *Rassegna dei Archeologia* **7**, 255–82 (1987).
205. Mercadal, O., Aliaga, S. & Bosom, S. Poblament i explotació del territori a la Cerdanya. Assaig de síntesi: del Neolític a l'Edat Mitjana. in *Muntanyes i població. El passat dels Pirineus des d'una perpectiva multidisciplinària* (Centre de Trobada de les Cultures Pirinenques, 1995).
206. Maluquer de Motes, J. El desarrollo de la Primera Edad del Hierro. in (1963).
207. Maya, J. L. Primera Edad del Hierro: los Campos de Urnas. in *Historia de España* **1**, 296–377 (1990).
208. Ruiz, G. El poblamiento del primer milenio A.C. en los Pirineos. in *Muntanyes i població. El passat dels Pirineus des d'una perpectiva multidisciplinària* (Centre de Trobada de les Cultures Pirinenques, 1995).
209. Campmajó, J. & Untermann, J. Les influences ibériques dans la Haute Montagne Catalane: le cas de la Cerdagne. in *Lengua y cultura en la Hispania Prerromana* (1993).
210. Martí888. Catalonia base map 42 comarques. *Wikimedia Commons* (2016).
211. Álvarez-Álvarez, M. M., Carreras-Torres, R., Zanetti, D., Vegas, E. & Moral, P. Population variation of LIN28B in the Mediterranean: Novel markers for microgeographic discrimination: Variation of LIN28B Gene in Human Populations. *Am. J. Hum. Biol.* **28**, 905–912 (2016).
212. Álvarez-Álvarez, M. M., Zanetti, D., Carreras-Torres, R., Moral, P. & Athanasiadis, G. A survey of sub-Saharan gene flow into the Mediterranean at risk loci for coronary artery disease. *Eur J Hum Genet* **25**, 472–476 (2017).
213. Álvarez-Álvarez, M. M. *et al. Genetic analysis of Sephardic ancestry in the Iberian Peninsula*. (Genomics, 2018). doi:10.1101/325779
214. Zanetti, D., Carreras-Torres, R., Esteban, E., Via, M. & Moral, P. Potential Signals of Natural Selection in the Top Risk Loci for Coronary Artery Disease: 9p21 and 10q11. *PLoS ONE* **10**, e0134840 (2015).
215. Bosch, E. *et al.* High-Resolution Analysis of Human Y-Chromosome Variation Shows a Sharp Discontinuity and Limited Gene Flow between Northwestern Africa and the Iberian Peninsula. *The American Journal of Human Genetics* **68**, 1019–1029 (2001).
216. Capelli, C. *et al.* Population Structure in the Mediterranean Basin: A Y Chromosome Perspective. *Ann Human Genet* **70**, 207–225 (2006).

217. Comas, D. *et al.* Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Human Genetics* **107**, 312–319 (2000).
218. González-Pérez, E. *et al.* Population relationships in the Mediterranean revealed by autosomal genetic data (*Alu* and *Alu* /STR compound systems): Population Relationships in the Mediterranean. *Am. J. Phys. Anthropol.* **141**, 430–439 (2010).
219. Simoni, L., Gueresi, P., Pettener, D. & Barbujani, G. Patterns of gene flow inferred from genetic distances in the Mediterranean region. *Human Biology* **71**, 399–415 (1999).
220. Bauchet, M. *et al.* Measuring European Population Stratification with Microarray Genotype Data. *The American Journal of Human Genetics* **80**, 948–956 (2007).
221. Humphreys, R. *Islamic History: a Framework for Inquiry*. (Princeton University Press, 1991).
222. Döring, Y., Pawig, L., Weber, C. & Noels, H. The CXCL12/CXCR4 chemokine ligand/receptor axis in cardiovascular disease. *Front. Physiol.* **5**, (2014).
223. Pignatelli, M. *A comunidade israelita de Lisboa: o passado e o presente na construccão da etnicidade dos judeus de Lisboa*. (Universidade Técnica de Lisboa, Instituto Superior de Ciências Sociais e Políticas, 2000).
224. Martins, J. *Portugal e os judeus: Judaísmo e anti-semitismo no século XX*. (Vega, 2006).
225. Adams, S. M. *et al.* The Genetic Legacy of Religious Diversity and Intolerance: Paternal Lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *The American Journal of Human Genetics* **83**, 725–736 (2008).
226. Niremberg, D. Love Between Muslim and Jew in Medieval Spain: a Triangular Affair. in *Jews, Muslims, and Christians in and around the Crown of Aragon: Essays in Honour of Professor Elena Lourie H.J. Hames* 60–76 (2004).
227. Atzmon, G. *et al.* Abraham's Children in the Genome Era: Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. *The American Journal of Human Genetics* **86**, 850–859 (2010).
228. Chacón-Duque, J.-C. *et al.* Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat Commun* **9**, 5388 (2018).
229. López-Parra, A. M. *et al.* In search of the Pre- and Post-Neolithic Genetic Substrates in Iberia: Evidence from Y-Chromosome in Pyrenean Populations. *Annals of Human Genetics* **73**, 42–53 (2009).
230. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet* **48**, 94–100 (2016).
231. Giraldo, M. P. *et al.* Gm and Km alleles in two Spanish Pyrenean populations (Andorra and Pallars Sobira): a review of Gm variation in the Western Mediterranean basin. *Annals of Human Genetics* **65**, 537–548 (2001).

RESUMEN EN CASTELLANO

INTRODUCCIÓN

Genética de poblaciones humanas

Avances en el estudio de la genética de poblaciones humanas

Durante las primeras décadas del siglo XX, Fisher, Haldane y Wright sentaron las bases de la genética de poblaciones, integrando la teoría de la evolución de Darwin, las leyes fundamentales de la herencia genética de Mendel, y la estadística. Desde entonces, un largo número de hitos ha guiado el estudio temprano de la genética de poblaciones humanas, incluyendo el descubrimiento de la naturaleza y estructura del DNA, el desarrollo de métodos de secuenciación (e.g. Sanger) y amplificación (PCR), y la teoría neutralista de la evolución molecular. La comparación de proteínas polimórficas en diferentes poblaciones por Cavalli-Sforza et al. permitió trazar los patrones de migración generales seguidos por los humanos modernos. Algunos marcadores genéticos típicamente empleados en la comparación de poblaciones son los microsatélites, polimorfismos de restricción, inserciones *Alu*, haplotipos uniparentales (genoma mitocondrial y cromosoma Y), y los polimorfismos de nucleótido único (SNP).

Un SNP es un nucleótido que presenta al menos una base nitrogenada alternativa en parte de la población. Los SNPs son los marcadores genéticos más usados actualmente, principalmente debido a ser los más abundantes en el genoma. Los chips de genotipación permiten evaluar los patrones de variación de SNPs comunes por una fracción del precio de secuenciación. Sin embargo, la falta de SNPs raros en el diseño de estos chips supone un sesgo importante, tanto para el estudio de la demografía como para la identificación de variantes asociadas a enfermedades.

El Proyecto Genoma Humano, llevado a cabo entre 1990 y 2004, generó la primera secuencia genómica de referencia, la cual se actualiza periódicamente. La existencia de un genoma

consenso de referencia, unido a las técnicas de secuenciación masivas introducidas en la primera década de este siglo, permite la genotipación rápida, y relativamente barata, de miles a millones de marcadores genéticos en números elevados de muestras. Esta cantidad de información genética sin precedentes ha facilitado el estudio de procesos demográficos como la migración, la estructuración poblacional, el mestizaje, la selección natural, y la hibridación con humanos arcaicos, incluso a escalas geográficas reducidas. Para procesar y analizar datos genómicos masivos en un tiempo razonable, se requiere el uso de herramientas de computación avanzadas, tales como programas especialisados y superordenadores (clústers).

A su vez, el descenso del coste de secuenciación ha permitido la creación de exhaustivas bases de datos poblacionales, tales como el proyecto HapMap o, más recientemente, el proyecto 1000 genomas, el cual incluye el genoma completo con baja cobertura de 2.504 individuos de varias poblaciones a nivel mundial. Estos datos pueden ser directamente descargados o, en su lugar, la información demográfica y clínica de regiones genómicas específicas puede ser obtenida eficientemente usando buscadores como Ensembl. También existen bases de datos que incluyen información sobre el fenotipo de las muestras, como enfermedades con base genética. Un ejemplo es el UK BioBank, que contiene datos genéticos y fenotípicos de 500.000 muestras del Reino Unido.

Perspectiva evolutiva de la variación genética en humanos

Los patrones de variación genética observados en humanos están generados por la interacción de procesos de deriva genética, migratorios y de selección natural que operan de manera diferente en poblaciones aisladas entre sí, debido a barreras geográficas y/o culturales. El número medio de diferencias nucleotídicas entre un genoma típico y la secuencia de referencia es entre 4,1 y 5 millones. Las poblaciones de ascendencia genética subsahariana son las que más variabilidad presentan, debido a que no han sufrido tantos cuellos de botella como las poblaciones derivadas de la salida de África hace 50-60 mil años.

Análisis demográficos

La mayor parte de la variación genética en humanos es neutra, es decir, consecuencia de procesos estocásticos. La variación neutra tiene numerosas aplicaciones en el estudio de eventos demográficos, tales como migraciones, cambios en el tamaño poblacional, y procesos de aislamiento y diferenciación poblacional. Los métodos empleados se basan en diversos parámetros que recogen información sobre las características de las poblaciones estudiadas, como la diversidad genética, las distancias genéticas interpoblacionales, la estructuración poblacional, los componentes ancestrales de mestizaje, la longitud de haplotipos resultantes de eventos de mestizaje, y los patrones de coalescencia alélica.

Selección natural

Como corolario de lo anterior, la fracción de regiones genómicas cuya alteración causa un impacto en la viabilidad del individuo y que, por tanto, es objeto de selección natural, es muy reducida. Un genotipo puede ser seleccionado positivamente si incrementa las opciones de supervivencia y/o reproducción del individuo, y negativamente en caso contrario. Los humanos anatómicamente modernos han sufrido procesos selectivos en lo relativo a la adaptación a distintos ambientes, patógenos, y dietas.

Estudios de asociación

Los estudios de asociación sirven para identificar variantes genéticas relacionadas con el desarrollo de una enfermedad. La mayor parte de las enfermedades genéticas más frecuentes son complejas, es decir, son el resultado de la interacción de numerosas variantes genéticas con variables ambientales. Un tipo de estudio de asociación muy empleado en la identificación de la base genética de las enfermedades complejas es el estudio de asociación del genoma completo (GWAS). Concretamente, los GWAS destinados al estudio de enfermedades complejas no cuantitativas consisten en la comparación mediante regresión logística de las distribuciones alélicas de miles a millones de SNPs entre individuos afectados por una enfermedad (casos) e individuos sanos (controles). Aunque los GWAS han identificado una gran parte de la variación

genética implicada en enfermedades complejas, una fracción importante de la heredabilidad sigue sin resolverse, debido en parte a falta de poder estadístico y a la escasez de variantes raras en los chips de genotipación.

Poblaciones humanas en la región Mediterránea

Principales movimientos migratorios del Paleolítico a la Edad Media

Los restos de *Homo sapiens* más antiguos en el norte de África son los de Jebel Irhoud (Marruecos), datados en ~315.000 años. Oleadas migratorias procedentes del sudoeste asiático llegaron a la región hace 40-45.000 años. En el sur de Europa, los yacimientos humanos más antiguos son los de *Homo heidelbergensis*, en la Sima de los Huesos, datados en 430.000 años. Los primeros humanos modernos llegaron al Levante hace unos 48.000 años, y a Europa hace unos 43-45.000 años, donde coexistieron con *H. neanderthalensis*. Durante el último máximo glacial, iniciado hace ~27.000 años, las poblaciones humanas europeas quedaron relegadas a las penínsulas Ibérica, Itálica y Balcánica, y al sur de Francia.

La retirada de los hielos hace 12.000 años marcó el inicio del Neolítico en el Creciente Fértil y su expansión hacia el Mediterráneo, en forma de oleadas migratorias que absorbieron o desplazaron a los cazadores-recolectores paleolíticos en diversos grados, dependiendo de la región. Otra importante oleada migratoria se produjo durante la Edad de Bronce, procedente de las estepas al norte de los mares Negro y Caspio, la cual se propone que fue el origen de la familia de lenguas indo-europeas. En esta época se desarrollaron las primeras civilizaciones del Mediterráneo, como la egipcia, las cuales colapsaron hacia el final de la era (~1200-1150 a.e.c.).

La cultura fenicia ocupó el vacío de poder al principio de la Edad del Hierro, desarrollando ciudades-estado en todo el Levante, Anatolia, el norte de África, la Península Ibérica y las principales islas mediterráneas. Desde el ~650 a.e.c., tribus celtas procedentes de centroeuropa se extendieron hasta la Península Ibérica, los Balcanes y Anatolia. El mundo clásico greco-romano dominó la región Mediterránea desde los últimos siglos del I milenio a.e.c. hasta la caída del Imperio Romano de Occidente en el siglo V e.c., marcada por las invasiones de tribus germánicas y hunas.

Durante la Edad Media, la expansión árabe alcanzó todo el Levante y norte de África, así como la mayor parte de la Península Ibérica. Los reinos cristianos del norte de la Península Ibérica fueron ganando terreno durante siglos, hasta la conquista en el siglo XV del último territorio de dominio musulmán de esta región. En cuanto al Mediterráneo oriental, el Imperio Bizantino mantuvo en un inicio la mayor parte de territorios del Imperio Romano Oriental. En los siglos VI y VII, migraciones eslavas alcanzaron la Península Balcánica. En el siglo XI Anatolia fue conquistada por el Imperio Selyúcida, de origen túrquico, y a su vez por el Imperio Mongol en el siglo XIII. En el siglo XIV el reino Otomano logró el control de Anatolia y desde allí extendió un vasto imperio que incluía los Balcanes, parte del norte de África, y todo el Levante.

Estudios genéticos recientes

La incorporación de datos genéticos masivos ha supuesto un importante avance en el estudio de la historia demográfica de las poblaciones humanas de la región Mediterránea. Por ejemplo, estudios basados tanto en poblaciones actuales como en la utilización de ADN antiguo han contribuido a solucionar el debate de la naturaleza de la expansión de la cultura neolítica en el Mediterráneo a favor de la hipótesis de la sustitución (total o parcial) poblacional.

Otros estudios sugieren que el mar Mediterráneo ha sido desde tiempos antiguos una barrera a los movimientos migratorios entre África y Europa. A pesar de esta limitación parcial al flujo génico, se observan niveles significativos de componente genético norteafricano – y, por extensión, subsahariano – en el sur de Europa. El norte de África presenta una mayor diversidad genética, ligada a una historia compleja que incluye varias oleadas migratorias sucesivas procedentes del Oriente Próximo, así como pulsos de flujo génico subsahariano.

También se observan patrones de estructuración poblacional a escala más reducida en diversas poblaciones mediterráneas, como es el caso del Levante, la Península Ibérica y Cerdeña. Los grupos étnicos judío y romaní muestran, además, diferenciación del resto de poblaciones como consecuencia de su aislamiento genético histórico y, en el caso de los romaníes, su origen asiático. Por otra parte, la distribución de grupos étnicos bereber y árabe en el norte de África no parece estar ligada a una estructuración genética.

Poblaciones incluidas en este estudio

Las poblaciones estudiadas en este trabajo abarcan toda la región Mediterránea, incluyendo muestras de i)la Península Ibérica, ii)el sur de Francia, iii)la Península Itálica, iv)los Balcanes, v)Turquía, vi)israelíes de ascendencia judía turca o iraquí, vii)Jordania, viii)Marruecos, ix)Túnez, x)Argelia, xi)Libia y xii)las islas de Tenerife, Menorca, Cerdeña, Sicilia y Creta. A continuación hay un resumen de la historia demográfica de dos poblaciones de especial relevancia en los estudios incluidos en este trabajo: los judíos sefarditas y la población del Pirineo catalán.

Judíos sefarditas

Los sefarditas son los históricos practicantes de la religión judía en la Península Ibérica desde época romana, así como los descendientes de los expulsados. La persecución iniciada en el siglo XIV y continuada por la Inquisición, tras el Edicto de Granada, hasta el siglo XVIII, causó la conversión forzosa de entre 200.000 y 300.000 sefarditas y el exilio de entre 50.000 y 80.000. La cultura sefardita se ha conservado en algunos territorios de exilio, como los Balcanes, Turquía y el Levante, aunque el genocidio nazi y las migraciones a Israel han acelerado el declive de esta cultura.

Pirineo catalán

Hay presencia humana continuada en esta región desde el Paleolítico, con la llegada de grupos neolíticos estimada en torno al VI milenio a.e.c. Además, se ha identificado la introducción de dos oleadas migratorias sucesivas procedentes de centroeuropa en la Edad del Bronce. Las tribus prerromanas mencionadas en fuentes históricas fueron poco influenciadas por los colonizadores fenicios y griegos, y su romanización fue tardía y poco intensa: se estima que el total de colonizadores romanos y, posteriormente, visigodos, sólo representó entre un 2.2% y un 4.4% de la población del Pirineo catalán. Después de 80 años de dominio árabe, la región pasó a formar parte de la Marca Hispánica del Imperio Carolingio hasta el siglo X, y fue finalmente anexionada a la corona de Aragón en el siglo XII.

OBJETIVOS DEL ESTUDIO

- Análisis de diferencias en la distribución alélica de *LIN28B*, un gen asociado a cáncer, entre poblaciones de las costas norte y sur del Mediterráneo.

- Análisis de un componente genético de origen subsahariano en regiones genómicas asociadas a la enfermedad de las arterias coronarias.

- Estimación del componente genético de origen sefardita en poblaciones actuales de la Península Ibérica y análisis del proceso de diáspora.

- Uso de datos de secuenciación de muestras del Pirineo catalán para el análisis de i)la historia demográfica de la región, ii)patrones de estructura poblacional, y iii)el grado de aislamiento genético de la población e implicaciones clínicas de éste.

RESULTADOS Y CONCLUSIONES

- Una cantidad reducida de datos genéticos permite identificar el eje norte-sur como el principal componente de diferenciación genética en el Mediterráneo, algo ya propuesto en otros estudios.

- El polimorfismo rs221639 se presenta como un buen candidato para marcador informativo de ascendencia genética en el contexto de la región mediterránea.

- La presencia de niveles de heterocigosidad más altos, así como de una ascendencia genética subsahariana mayor, en poblaciones norteafricanas comparado con lo observado en el sur de Europa, se puede atribuir a un flujo génico subsahariano más intenso debido a la mayor proximidad geográfica y/o al papel del mar Mediterráneo como barrera genética.

- Se detecta un componente genético subsahariano en cuatro regiones genómicas, asociadas a la enfermedad de las arterias coronarias, genotipadas a baja densidad en poblaciones mediterráneas. Además, el estadístico *D* sugiere una posible introgresión genética subsahariana en la región genómica 10q11, la cual contiene el gen *CXCL12*, codificante de un ligando de quimiocinas, y en la cual se han identificado además señales potenciales de selección estabilizadora.

- Datos a nivel genómico obtenidos con un chip de genotipado permiten detectar un componente genético de origen sefardita en las poblaciones de Portugal, Galicia y Andalucía. Estos resultados revisan a la baja lo observado en un estudio previo basado en el análisis de haplogrupos del cromosoma Y.

- La datación de los componentes genéticos sefardita y bereber en poblaciones de la Península Ibérica indica que el primero es más reciente, lo cual sugiere que el mestizaje entre judíos y cristianos tuvo lugar fundamentalmente durante, y con posterioridad a, la persecución y conversión forzosa de los sefarditas.

- La datación de la introducción del componente genético sefardita en diferentes poblaciones confirma una ruta de salida de los sefarditas a través del sudoeste de Europa.

- Análisis haplotípicos de datos de secuenciación con alta cobertura muestran la población del Pirineo catalán como un grupo diferenciado en el contexto ibérico, y estrechamente emparentado con poblaciones vascas.

- La población del Pirineo catalán está dividida genéticamente en dos grupos, occidental y oriental. Estimamos que esta estratificación tuvo lugar en torno al siglo V a.e.c.

- La clasificación de la población del Pirineo catalán como una población aislada genéticamente se sustenta en la existencia de una barrera genética entre el Pirineo catalán y el resto de España, en un descenso de su tamaño poblacional a una décima parte del original durante las últimas 100 generaciones, y en la presencia de segmentos haplotípicos homocigóticos más largos en la población del Pirineo catalán que en la población general española.

- La reducción en el número de heterocigotos para variantes que contienen una mutación altamente deletérea en la población del Pirineo catalán con respecto a la población general española concuerda con varios estudios que muestran que, en poblaciones que han sufrido una reducción poblacional prolongada, la selección purificadora ha eliminado una gran parte de las mutaciones deletéreas raras. Sin embargo, esto ha ido acompañado de un aumento en la frecuencia de ciertas mutaciones, lo cual produce unas incidencias particularmente altas de

enfermedades genéticas recesivas raras y un aumento de la carga total de mutaciones deletéreas. En nuestro estudio, esto último no se ha observado, lo cual se podría atribuir a una ausencia artificial de estos marcadores en nuestros datos derivada del proceso de secuenciación y/o control de calidad.

APPENDIX

I) Description of 1000G population codes

Population Code	Population Description
CHB	Han Chinese in Beijing, China
JPT	Japanese in Tokyo, Japan
CHS	Southern Han Chinese
CDX	Chinese Dai in Xishuangbanna, China
KHV	Kinh in Ho Chi Minh City, Vietnam
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
TSI	Toscani in Italia
FIN	Finnish in Finland
GBR	British in England and Scotland
IBS	Iberian Population in Spain
YRI	Yoruba in Ibadan, Nigeria
LWK	Luhya in Webuye, Kenya
GWD	Gambian in Western Divisions in the Gambia
MSL	Mende in Sierra Leone
ESN	Esan in Nigeria
ASW	Americans of African Ancestry in SW USA
ACB	African Caribbeans in Barbados
MXL	Mexican Ancestry from Los Angeles USA

PUR	Puerto Ricans from Puerto Rico
CLM	Colombians from Medellin, Colombia
PEL	Peruvians from Lima, Peru
GIH	Gujarati Indian from Houston, Texas
PJL	Punjabi from Lahore, Pakistan
BEB	Bengali from Bangladesh
STU	Sri Lankan Tamil from the UK
ITU	Indian Telugu from the UK

II) Links to the Supplementary Information for the Álvarez-Álvarez et al. 2016 article

Supplementary Information 1.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fajhb.22887&file=ajhb22887-sup-0001-suppinfo1.doc

Supplementary Information 2.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fajhb.22887&file=ajhb22887-sup-0002-suppinfo2.doc

III) Links to the Supplementary Information for the Álvarez-Álvarez et al. 2017 article

Supplementary Information 1.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x1.png

Supplementary Information 2.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x2.png

Supplementary Information 3.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x3.png

Supplementary Information 4.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x4.png

Supplementary Information 5.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x5.png

Supplementary Information 6.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x6.png

Supplementary Information 7.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x7.png

Supplementary Information 8.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5386420/bin/ejhg2016200x8.docx

IV) Links to the Supplementary Information for the article '*Genetic analysis of Sephardic ancestry in the Iberian Peninsula*'

Supplementary Information 1.

PCA based on 156,733 SNPs for the South European and Jewish populations (i.e. without the North Africans).

https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/0/325779-1.pdf

Supplementary Information 2.

PCA based on 156,733 SNPs for the Iberian Peninsula, South France and North Italy.

https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/1/325779-2.pdf

Supplementary Information 3.

fineSTRUCTURE grouping of 500 Mediterranean or near-Mediterranean samples roughly corresponding to well-defined geographic locations.

https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/2/325779-3.jpg

Supplementary Information 4.

fineSTRUCTURE grouping of 198 Iberian (without the Basques) and South French samples roughly corresponding to well-defined geographic locations. Only Galicia and Portugal did not form clear clusters according to their labels.

https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/3/325779-4.jpg

Supplementary Information 5.

Distribution of Sephardic, Berber and Iberian migrant tracts (1 cM bins) in Andalusians, Galician-Portuguese and North Italians assuming time of admixture g = 10 generations ago.

https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/4/325779-5.pdf

Supplementary Information 6.

Distribution of Sephardic, Berber and Iberian migrant tracts (1 cM bins) in Andalusians, Galician-Portuguese and North Italians assuming time of admixture g = 25 generations ago.

https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/5/325779-6.pdf

Supplementary Information 7.

Distribution of Sephardic, Berber and Iberian migrant tracts (1 cM bins) in Andalusians, Galician-Portuguese and North Italians assuming time of admixture g = 40 generations ago.

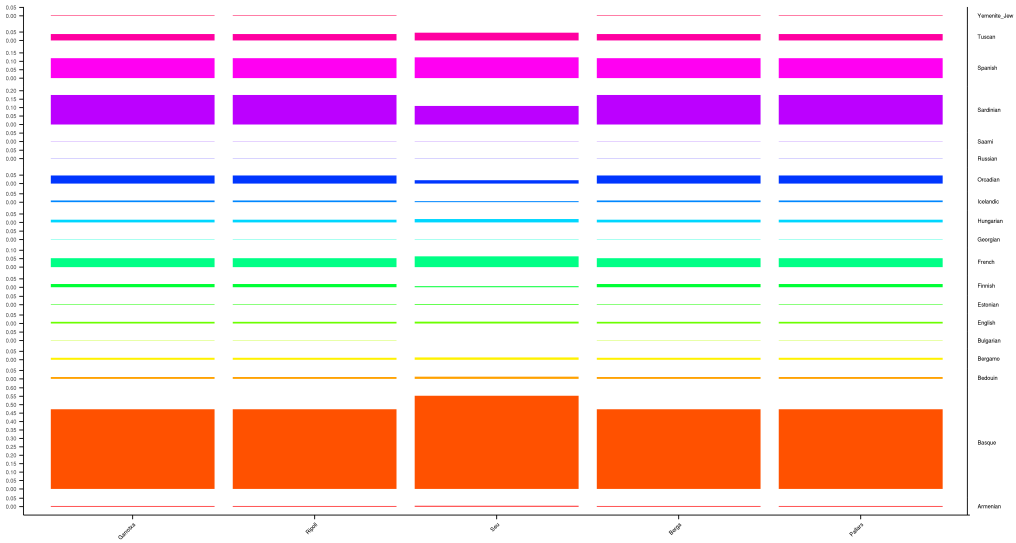
https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/6/325779-7.pdf

Supplementary Information 8.

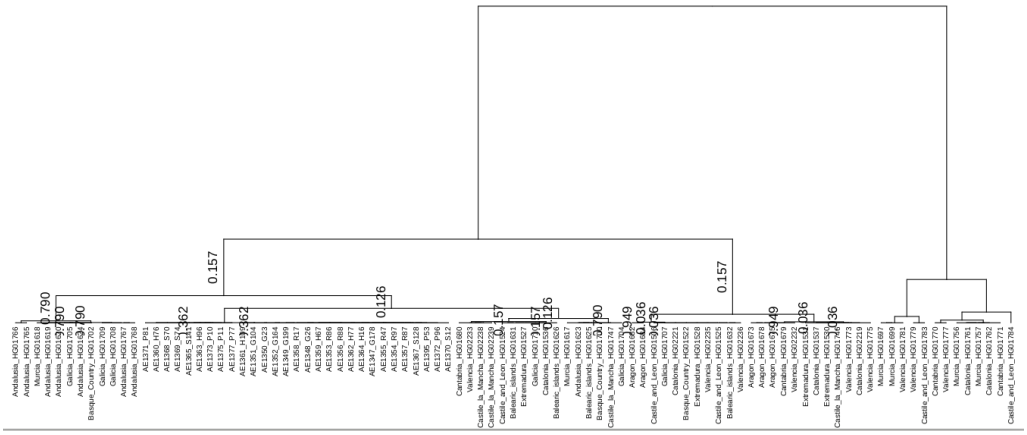
Jitter and box plot of (i) median Sephardic tract length and (ii) variance of Sephardic tract length for Andalusians, Galician-Portuguese and North Italians, assuming time of admixture g = {10, 40} generations ago.

https://www.biorxiv.org/highwire/filestream/99464/field_highwire_adjunct_files/7/325779-8.pdf

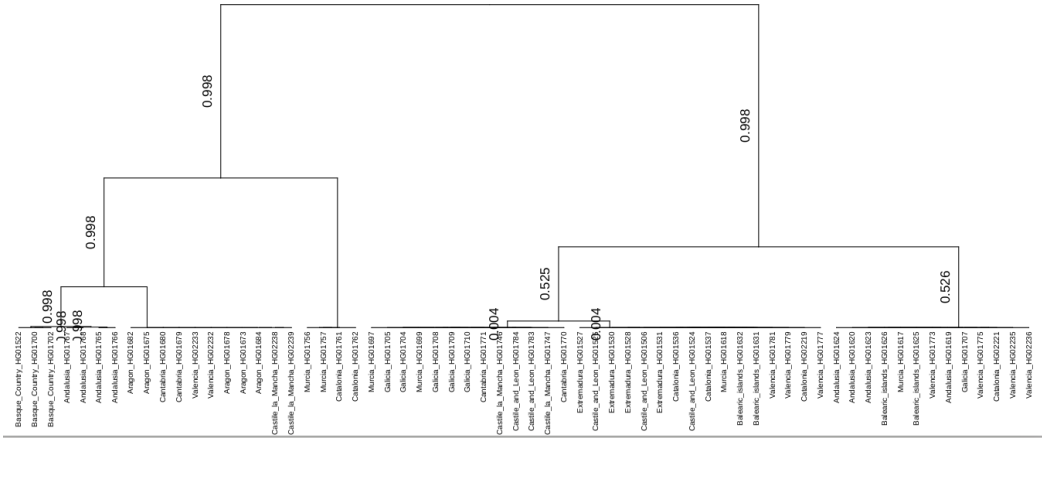
V) Supplementary Information for the article '*High-coverage sequence data from the Spanish Eastern Pyrenees suggest patterns of population structure and isolation*'



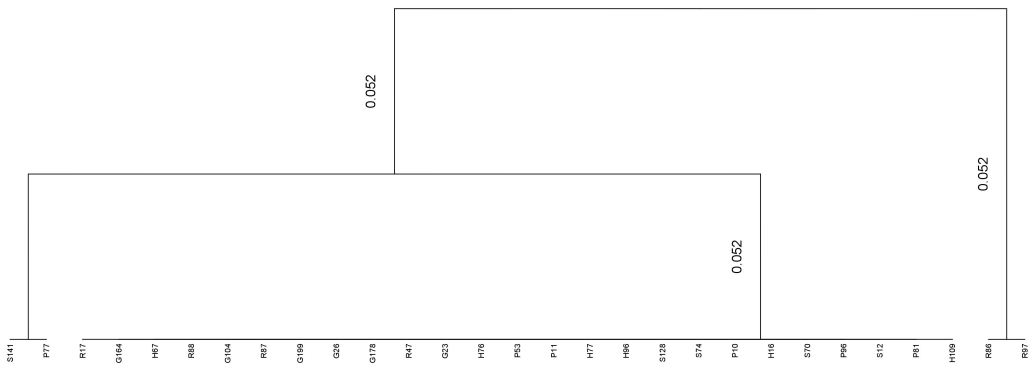
Supplementary material 1. GLOBETROTTER ancestral components plot obtained with the SEP-SGDP dataset. X-axis indicates the recipient populations (the five SEP comarques*), while the y-axis shows the West Eurasian populations from the SGDP database.*



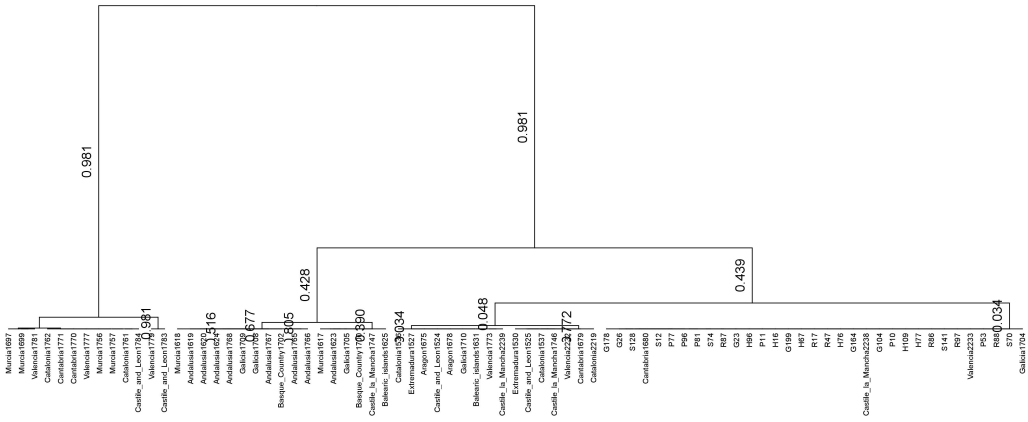
Supplementary material 2. fineSTRUCTURE tree of SEP-IBS, showing lack of structure in IBS samples and SEP samples assigned to a private cluster. SEP sample codes start with "AE", followed by a number and a secondary code indicating the* comarca *of origin; namely, "G*", "R*", "H*", "S*", and "P*", stand for Garrotxa, Ripolles, Bergueda, Alt Urgell, and Pallars, respectively.*



Supplementary material 3. fineSTRUCTURE tree of the original IBS data, confirming lack of structure in IBS samples.



Supplementary material 4. fineSTRUCTURE tree of SEP, using ~247K SNPs common to a recent study that includes array data of Pyrenean samples[18]. As expected, the population structure pattern observed when using the full amount of variation disappears. SEP sample codes start with a letter indicating the comarca of origin, namely "G", "R*", "H*", "S*", and "P*", stand for Garrotxa, Ripolles, Bergueda, Alt Urgell, and Pallars, respectively.*



Supplementary material 5. fineSTRUCTURE tree of SEP-IBS using ~285K SNPs common to a recent study[18]. The private cluster containing the SEP samples persists to some extent. SEP sample codes start with a letter indicating the comarca of origin, namely "G", "R*", "H*", "S*", and "P*", stand for Garrotxa, Ripolles, Bergueda, Alt Urgell, and Pallars, respectively.*