Characterization of the micro-substructure of a rural population from the Pyrenees from a geodesic and technical point of view using NGS data

Quantification of batch effects in whole genome sequence data

Iago Maceda Porto

TESI DOCTORAL UPF / 2020

Thesis supervisor

Dr. Oscar Lao Grueso (Population Genomics Team Leader, Centre Nacional d'Anàlisi Genòmica (CNAG-CRG), Centre de Regulació Genòmica)

Universitat Pompeu Fabra. Departament de Ciències Experimentals i de la Salut. Programa de Biomedicina

“To make a huge success, a scientist must be prepared to get into deep trouble.”

James D. Watson

ACKNOWLEDGEMENTS

I would like to thank the following people for the help and opportunities they gave me during this project:

My supervisor Oscar Lao, for the opportunity, enthusiasm, support, encouragement and patience;

To Olga Dolgova, for the help and talks regarding parts of this work.

To Raúl Tonda and Jordi Camps, for their help on the technical steps of this project.

To Pedro Moral and Miguel Álvarez, the other part of the team that made possible part of this project.

And finally to my parents, Sinda and José Luis, who allowed me to follow in this path with their endless support.

ABSTRACT

In this work we present a new whole genome sequencing dataset with samples gathered from the Spanish Eastern Pyrenees (SEP) with more than 40x coverage. We apply both classical and new methods to unveil their particular demographic histories and we present the use of a newly in-house developed algorithm to detect genetic barriers taking into account the use of geo-statistics. With these analyses we detect fine population substructure for the first time in this region.

We also report the presence of an important batch effect in one of the most important datasets used in genomics: the 1,000 Genomes Project. We find this batch effect when considering very low frequency variants, such as loss of function mutations and the amount of singletons (both ancestral and derived) detected in each sample.

Keywords: population genomics; rural areas; whole-genome sequencing; batch effect; 1,000 Genomes Project

RESUM

En aquest treball presentem un nou dataset de whole genome sequencing amb mostres recollides del Pirineu Oriental espanyol (SEP) amb un coverage superior a 40x. Apliquem mètodes clàssics i nous per descobrir les seves particulars històries demogràfiques i presentem l'ús d'un algorisme desenvolupat recentment en el nostre laboratori per detectar barreres genètiques tenint en compte l'ús de geoestadística. Amb aquestes anàlisis detectem, per primera vegada, una delicada subestructura de poblacions en aquesta regió.

També informem de la presència d'un important batch effect en un dels datasets més importants utilitzats en genòmica: the 1.000 Genomes Project. Trobem aquest batch effect quan considerem variants rares, com per exemple mutacions que comporten pèrdua de funció i la quantitat de singletons (tant ancestrals com derivats) detectats en cada mostra.

Paraules clau: genòmica de poblacions; zones rurals; whole-genome sequencing; batch effect; 1,000 Genomes Project

PREFACE

Whole genome sequencing (WGS) has boosted our current knowledge about the general architecture of the genetic diversity present in human populations. However, as we go deeper into the detection of population substructure, new methods for detecting population substructure are required and infrequent biases associated with the WGS technology become more important.

In this thesis we explore the limits of the detection of fine population substructure and their implications in the context of rural populations from the Pyrenees, and the technical artifacts generated by WGS data due to the different sequencing centres the samples were generated.

In the Introduction I talk about how variation is generated and how the frequency of new variants is modified across generations. I also recapitulate two of the more used technologies to search for variation: microarrays and next-generation sequencing (NGS). To finalize, I briefly describe the history of *Homo sapiens* since the out-of-Africa, recapitulate its demographic history across Europe and the importance of studying rural areas, with a particular emphasis on the Spanish Eastern Pyrenees.

In Material and Methods I describe the different datasets used in this work, with special attention to the SEP dataset. Also, I present a new algorithm to detect genetic barriers between groups of samples taking into account principles of geo-statistics. Furthermore, I also explain the use of old and new techniques to quantify levels of autozigosity in different datasets. To end this part, I show the methodology used to quantify the batch effect in the 1,000 Genomes Project dataset.

In Chapter 1 I present the results of the study of the Spanish Eastern Pyrenees dataset.

In Chapter 2 I present the results of our analysis of a possible batch effect regarding the 1,000 Genomes Project dataset affecting population genetics statistics.

Table of contents

1. INTRODUCTION

1.1 Generation of variation

Genetic variation is generated by a combination of two physical processes: mutation and recombination.

a) Mutation

A mutation is defined as a change that occurs in the DNA and it is the basic source of generating variation in the genome. Mutations usually appear as a result of an error during DNA replication or while repairing DNA damage caused by an external factor. A mutation can involve from a single nucleotide to megabases of nucleotides. The most common type of mutation in our genome is a point mutation (Figure 1), referred to as a single nucleotide variant (SNV) when it is not fixed in the population and single nucleotide polymorphism (SNP) when it reaches a certain frequency (i.e. 1%) in the population.

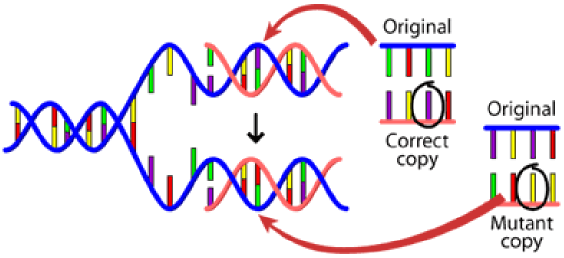


Figure 1: example of mutation. Source: University of California Museum of Paleontology's Understanding Evolution

Mutations present in the germline have the possibility to be transmitted to the next generation. These mutations can represent the somatic variation of the individual, but they can also represent new variation that appears during the gametogenesis process.

Mutations usually appear when DNA is copied during cell division. Although the DNA polymerase used in this step copies DNA with high fidelity (DNA polymerase delta has an error rate of 1 error every 10^4-10^5 incorporated nucleotides *in vitro*, possibly more *in vivo*; (Ganai & Johansson, 2016)), these errors can bypass the error proofing machinery of the cell. These errors tend to be corrected in zones enriched in adenine (A) and thymine (T), as it is easier to maintain the replication fork opened (A-T pairs form only 2 hydrogen bonds, while guanine (G)-cytosine (C) pairs form 3 hydrogen bonds). This phenomenon can be a possible explanation for the higher mutation rate found in GC-rich regions (Bloom et al., 1994; Petruska & Goodman, 1985).

Another important factor that affects the mutation rate is the methylation state of the nucleotide in question. In most mammals the 5' C in a CpG context tends to be methylated, which can undergo spontaneous deamination into T. This mismatch is repaired by low efficiency mechanisms (Schmutte et al., 1995) which results in a higher prevalence of C-to-T transitions in CpG sites (Coulondre et al., 1978; Duncan & Miller, 1980). These same CpG sites are less mutagenic in a high GC content, possibly because deamination of Cs depends on local strand separation (Fryxell & Zuckerkandl, 2000).

The region where the mutation takes place also plays an important role. For example, in CpG islands (regions of the genome with high abundance of C and G) we find a high number of de novo mutations (CpGs account for ~2% of the human genome, but they constitute ~19% of de novo mutations). This is consistent with the hypothesis of spontaneous deamination (where a C becomes a T or a G becomes an A). This process happens, mostly, in regions with medium levels of deamination (Xia et al., 2012). A secondary problem is the low number of polymorphisms that have been found in these regions, which could reflect the effect of gene conversion (process that favours the fixation of strong (G/C) over weak (A/T) alleles) (Capra et al., 2013) and purifying selection to maintain CpGs (Schmidt et al., 2008).

The region on which the mutation takes place also plays a role in the probability of observing a mutation. DNA damage can be sensed by a protein complex that recruits the nucleotide excision repair (NER) machinery to excise the oligonucleotide carrying the damage and then recruiting replication polymerases to copy the strand pair. A special case of NER can occur while transcription, directed towards the transcribed strand of the gene (Pleasance et al., 2010).

Another factor to take into account is mutations occurring in genes responsible for DNA proofreading (NER, for example) that can lead to mutational spectra in populations or even whole species. This is the case regarding an increased TCC to TTC mutation rate found in European populations (Harris & Pritchard, 2017).

Endogenous factors are not the only source of mutation: exogenous effects like ionizing radiation or certain chemicals can also produce mutations. For example, the higher UV part of the light spectrum can produce transitions from C-to-T in dipyrimidine sequences (Pfeifer et al., 2005). Regarding chemical exposure, certain cigarette components have been demonstrated to have effect on germline mutations, for example benzo(a)pyrene can cause G-to-T transversions or DNA adducts, disrupting the normal double helix structure (Yoon et al., 2001; Zenzes, 2000).

Another point of interest when estimating the mutation rate is the mean paternal age at which the individual has been conceived. In multiple pedigree studies (Conrad et al., 2011; Francioli et al., 2015; Kong et al., 2012) it has been proved that paternal age has an effect on the mutation load of the offspring. This effect is exclusive to the age of the father because males produce sperm through all their life, meaning that mutations occurring in the spermatogonial stem cells (the cells which, ultimately, produce spermatozoa) can accumulate through time (Figure 2). In contrast, females are born with most of its oocytes already formed and "frozen" mid maturation, so the probability of generating a new mutation is only depending on the exogenous factors (as duplication of the genetic material has already taken place).

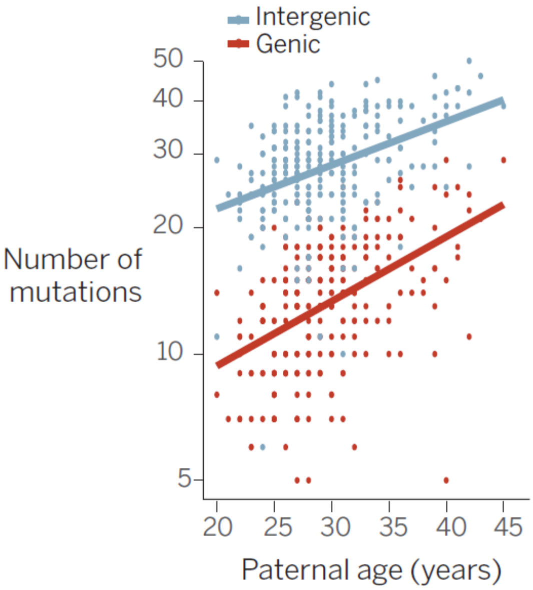
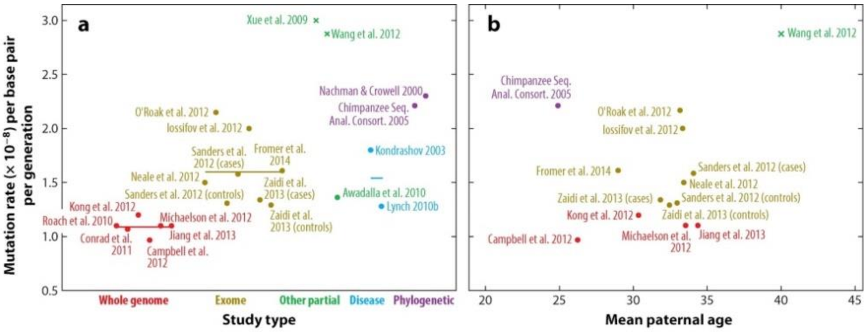


Figure 2: Accumulation of mutations in protein-coding regions at a higher rate in older compared with younger fathers. Source: adapted from Shendure & Akey, 2015

In the literature we can find numerous attempts to estimate the mutation rate in the human lineage. However, they have a high disparity between them, ranging from ~1.0 x 10^{-8} to ~3.0 x 10^{-8} (Figure 3). This heterogeneity in the estimations reflects the fact that mutations in general are rare events and to the genomic regions that have been used to estimate the mutation rate.

For example, we can find that CpG transitions compared with non-CpG transitions are increased 12-13 fold in polymorphism and divergence data; in disease studies this value goes to a 15 fold increase, while in pedigree data this number is close to 18 fold.



Ségurel L, et al. 2014.
Annu. Rev. Genomics Hum. Genet. 15:47–70

Figure 3: Estimates of the human mutation rate per base per generation. Source: Ségurel et al., 2014.

Since the mutation rate defines the tempo at which genetic variation appears, and the genetic variation present in a population depends on the number of chromosomes that reproduce and the mutation rate (see section Demographic factors), the ascertainment of a proper mutation rate is a key point to interpret the parameters of any demographic model.

SNVs that fall within a gene can be functionally classified in four main categories depending on their phenotypic consequences:

- **Non-coding:** these are those SNPs that fall in intergenic regions and, mostly, are not affected by selective forces. They compose the majority of the neutral variation found.

- **<u>Coding:</u>** the SNP occurs in the coding region (or close to one) of a gene. This has different subtypes depending on the effect that the mutation has on the protein. As example, we will see the changes that occur at the amino acid level given different mutations on the codon TTC (Figure 4).

    - <u>Silent:</u> due to the redundancy of the genetic code, this type of mutation does not change the amino acid present in the protein. For example: TTC (Lys) -> TT*<u>T</u>* (Lys)

    - <u>Missense:</u> the mutation causes a change in the amino acid sequence. This change can be conservative or non-conservative depending on the differences of the physicochemical properties between the original and the new amino acid. For example: TTC (Lys; basic polar) -> T*<u>C</u>*C (Arg; basic polar) and TTC (Lys; basic polar) -> T*<u>G</u>*C (Thr: polar), respectively

    - <u>Nonsense/nonstop:</u> this type of variation is the one that produces the most "visible" effect. In the first case the codon changes from an amino acid to a stop codon, producing a shortened, often non-functional, protein. In the second case the stop codon for the protein is changed into an amino acid, elongating the protein and, usually, making it non-functional.

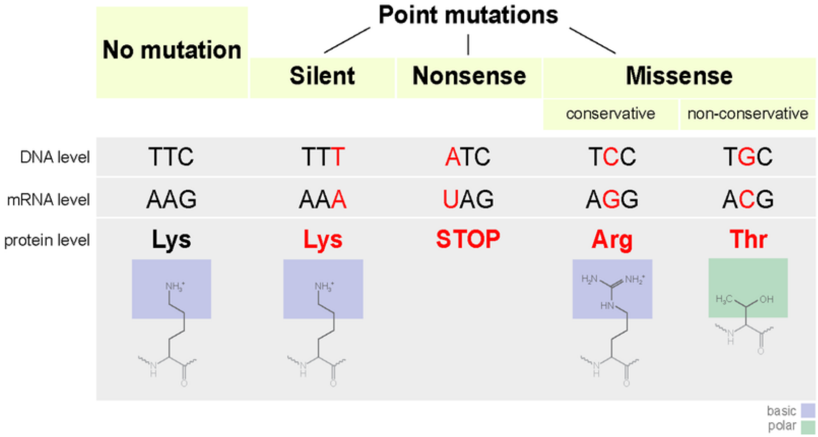


Figure 4: Different types of mutations and its consequences at the protein level. (Source: Jonsta247, n.d.)

Independent of how variation is generated, a mutation can affect the fitness of an individual (its ability to generate descendants) in four main categories:

- Beneficial: variation that increases the fitness of the individual. They are also called advantageous variation.

- Harmful: variation that decreases the fitness of the individual. Also called deleterious variation.

- Neutral: this type of variation does neither increase nor decrease fitness. This variation becomes the basis of the molecular clock.

- Nearly neutral: here we can find variation that is not purely neutral but they are slightly beneficial or slightly harmful.

Therefore, a basic question to address given an observed mutation is to which functional category it belongs, as well as its phenotypic consequences. Usually, predictors focus on the variation present in genes. Several approaches have been developed for predicting the functional effect of a given mutation. Ideally, one should *in vitro* experimentally verify novel variants; however, this approach is often infeasible due to facility limitations. In practice, several algorithms have been developed for predicting the functional effect of a nonsynonymous SNV (nsSNV). These algorithms lie into three different categories:

- Function prediction: refers to scores that predict the likelihood of a given nsSNV causing deleterious functional change of the protein.

- Conservation score: refers to scores that measure the conservativeness of a given nucleotide site across multiple species.

- Ensemble score: refers to scores that combine information of multiple component scores.

The most used ones are the following: SIFT (Vaser et al., 2015), MutationAssessor (Reva et al., 2011) and PolyPhen2 (Adzhubei et al., 2010). These three algorithms correspond to the function prediction category. They work in different ways:

- <u>SIFT:</u> this algorithm is based on a multiple sequence alignment of proteins with similar functions or sequence, and then it calculates a normalized probability of all possible substitutions for the given sequence. If the probability is lower than 0.05 that mutation is marked as deleterious, above that threshold is marked as tolerated.

- <u>MutationAssessor:</u> in this algorithm a score is calculated based on two different alignments. It aligns proteins of the same family (or sub-family) of sequence homologs within the same species and between different species.

- <u>Polyphen2:</u> bases the prediction of the impact on the function and structure of the human protein on features present in the sequence, the structure and the phylogenetic information. This gets fed into a Naive Bayes classifier, trained on two different datasets:

    - <u>HumDiv:</u> all damaging alleles with known effects causing human Mendelian diseases, present in the UniProtKB database, together with differences between human proteins and their closely related mammalian homologs, assumed to be non-damaging.

- HumVar: all human disease-causing mutations from UniProtKB, together with common human non-synonymous SNPs (MAF>1%) without annotated involvement in disease, which were treated as non-damaging.

Depending on the amount of evidence, each classifier ranks a given mutation according to a “risk classification” system (naming of such classification varies depending on the algorithm used). The overlap between the different algorithms is not total, but strong discrepancies between the different classifications have been reported (Dong et al., 2015). This implies that it is difficult to confidently establish from an in silico point of view the functional effect of a genetic variant. Conservative approaches run the different algorithms and establish the effect of a mutation by consensus, only considering it as highly damaging if all the algorithms agree on the damaging status of the mutation. In practice, this means that databases considering the effect of mutations are noisy and reaching useful conclusions is ultimately complex (Narasimhan et al., 2016).

b) Recombination

It is defined as the process that physically interchanges nucleotide sequences between two identical (or near identical) chromosomes. This process starts when two homologous non-sister chromatids align and crossover, usually during prophase I of meiosis, although it can also happen during mitotic division (A. J. Griffiths et al., 1999). In order for the crossover to start, a double-strand break is induced by the Spo11 protein (Keeney et al., 1997). This allows the interchange of non-sister chromatids.

This interchange allows the shuffling of the variation found in the maternal and paternal DNA and allows for new allelic combinations that will inherit the daughter germ cells, breaking the linkage that can be established between SNPs. Recombination events usually happen in localized (1-2 kb) regions of each chromosome, which are called hotspots.

In order for meiosis to proceed correctly, at least one crossover event is needed. If too few crossover events happen, it can lead to aneuploidy or to compromises in genomic integrity (Hassold & Hunt, 2001). Even in some cases, having a mean crossover rate higher than the average slightly increases the number of descendants (Kong et al., 2004).

In some occasions, recombination can be done in regions that are non-homologous in sister chromatids but share a high degree of sequence similarity. In these cases there is a non-homologous crossover, and it can result in one of the two sister chromatids carrying an “extra” copy of a given region (which can contain one or more genes). This process is called gene duplication (Figure 5). These genes then can change in an independent manner.

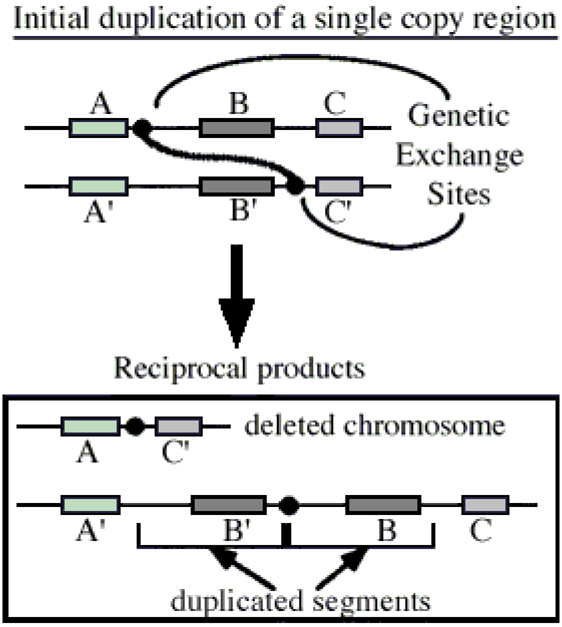


Figure 5: Example of gene duplication.

Source: Silver, 1995, Figure 5.5 (adapted)

Given the importance of recombination, one would think that this process is tightly regulated. This is correct for most species, as they possess mechanisms to ensure that at least one crossover event happens during meiosis, but it is also surprisingly variable between species.

In the following table (Table 1) I present a summary of genetic maps from different species with an average interval between markers studied. In the fifth column there is the measure of centiMorgans (cM) between markers. Morgans are a measurement of how likely a segment of DNA is to recombine from one generation to the next; in this case cM corresponds to a probability of 1% to recombine.

Species	Order	Genetic map length (cM)	Number of markers	Average intermarker interval (cM)
Human	Primates	3615	5136	0.704
Baboon	Primates	2013	352	5.719
Macaque	Primates	2275	326	6.979
Mouse	Rodentia	1361	6336	0.215
Rat	Rodentia	1542	3824	0.403

Table 1. Summary of genetic map distance in different species. Source: adapted from Table 1 on Dumont & Payseur, 2008.

In some cases, these variations can even happen between individuals from the same species. This variation affects all possible scales, from single hotspots to the whole genome. At least in humans, it has been proven a pronounced variation of the female genetic map (Broman et al., 1998; Lenzi et al., 2005).

Recombination in itself can be both good and bad. It can be good for the species as it generates new combinations of pre-existing variants that can lead to a better adaptation. At the same time it can be bad, as it can break favourable combinations of alleles acting on epistasis.

In some cases, crossovers are not carried correctly. One of the chromatids ends up copied in the other chromosome. If the non-sister chromatids contain different alleles in that region (the individual is heterozygous for one or more SNVs), the four derived germ cells end up as homozygous for the "copied" allele. This event is known as gene conversion (GC). In Figure 6 you can see a toy example with the difference between crossover and gene conversion.

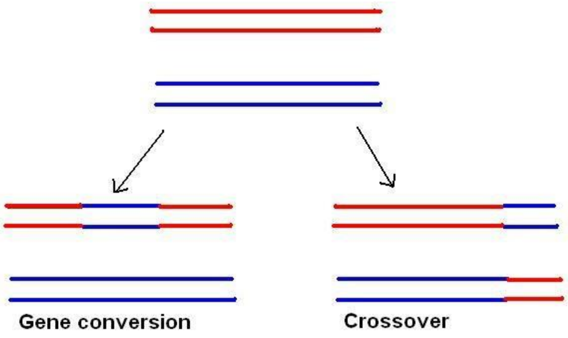


Figure 6: Differences between crossover and gene conversion.

Source: (Häggström, n.d.)

In some cases of gene conversion one of the alleles is more favoured to be the copied allele (or donor allele), in which case is known as biased gene conversion (BGC). In the case of humans there is a large body of evidence suggesting that in cases of GC/AT heterozygotes tend to produce a higher number of GC- than AT-gametes (Duret & Galtier, 2009). Although this can seem an unimportant feature of the molecular machinery, it can explain why local recombination rates tend to be positively correlated with GC content (Fullerton et al., 2001).

Recombination is also an important factor in breaking haplotypes (groups of alleles from different SNVs that tend to be inherited together from a single parent because they are located in the same genomic region). Haplotypes, both in structure and frequency, can vary between different populations. The statistical phenomenon that the frequencies in a population of the different allelic combinations between two loci that are in the same genomic region depart from the expected under the assumption of independence is known as linkage disequilibrium (LD). The patterns of LD for a given genomic distance vary among populations (Figure 7) due to demographic factors such as isolation or recent admixtures and can be used to make predictions about when such events occurred (Patterson et al., 2012).

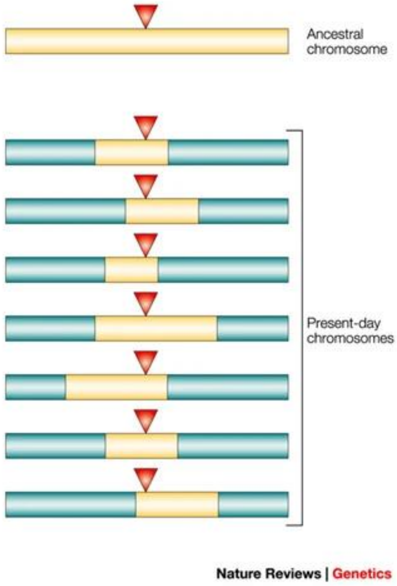


Figure 7: Example of Linkage Disequilibrium. The mutation is indicated by a red triangle; ancestral stretches are shown in yellow; new stretches in blue. Markers that are physically close tend to remain associated. Source: Ardlie et al., 2002.

Because recombination is a much more frequent event than the mutation, sharing a large proportion of haplotypes between individuals is also indicative of relatedness. This feature has been used to develop new algorithms, such as fineSTRUCTURE (Lawson et al., 2012), for identifying fine population substructure, or to predict genomic segments shared between individuals by descendent (identical by descent or IBD). Finally, studying the similarity of two haplotypes in a single individual provides information about the historical inbreeding patterns in her or his pedigree. Long runs of homozygosity (RoHs), indicating that for the two copies of the chromosome the same alleles are observed for a large number of SNVs, is indicative of high amounts of inbreeding. When such analysis is extended to all the sampled individuals from a population, the patterns of RoHs suggest different demographic events (Figure 8).

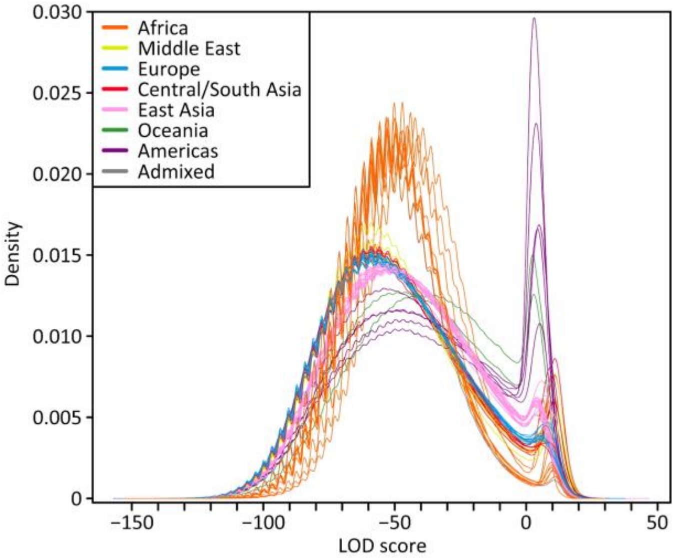


Figure 8: Variation of RoH distribution across different populations. Density function of the LOD scores from all individuals of a given population, coloured by geographic affiliation. Source: Pemberton et al., 2012.

1.2 Demography

Once a mutation has occurred in the germ line of an individual, it has the opportunity to pass to the next generation. The fate of such mutation in the population depends on different factors. If the mutation has no fitness effect in the carriers, then it is said to be neutral, and its change in the population in terms of allele frequency is ruled by demographic factors, ultimately determined by the number of individuals that reproduce at each generation in the population, and the strategy that is used to generate the couples reproducing to the next generation.

The simplest demographic model initially defined by Wright and Fisher (Tran et al., 2013) considers that all the individuals from the same generation mate at random -in panmixia- to produce the next generation and then die. Classical Moran's model (Moran, 1958) assumes that generations can overlap, but individuals mate also at random. Nevertheless, it has been shown that, from a coalescence point of view, the results from Moran's model are extrapolable after scaling to the ones obtained by Wright and Fisher (Kingman, 1982).

Demography has mainly two forces to take into account: genetic drift and migration.

a) Genetic drift

Genetic drift is the change in the frequency of a variant in a population due to random sampling of the organisms that reproduce at each generation. In more layman terms, genetic drift is the evolutionary equivalent of the sampling error from one generation to the next. This process can be easily visualized using a toy example of marbles (Figure 9). Imagine a box of 100 marbles that - by analogy with a SNV- has two possible colour states (green and brown), each at the same frequency (50%). If we sample at random 100 marbles with replacement to be our new box of marbles, the frequency of each colour category is likely to be different from the initial 50%. If this process is repeated over time, we will reach a point where one of the two colours gets fixed in the box and no colour variation is observed.

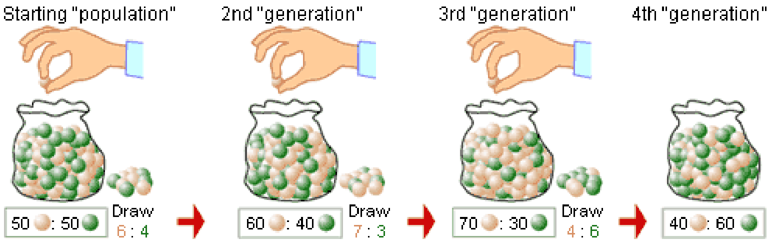


Figure 9: Toy example of how genetic drift works. Source: Source: University of California Museum of Paleontology's Understanding Evolution

From a frequentist point of view, the way how a frequency of a mutation changes over time can be modelled as a Markov Chain, in which its value only depends on the previous generation and the number of chromosomes that reproduce to the next one -called effective population size or Ne (see Figure 10).

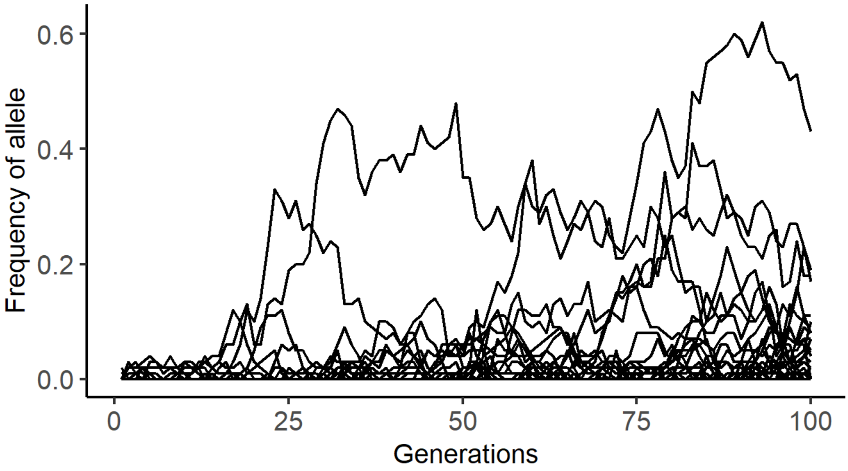


Figure 10: Fate of a new mutation over time in a population of 100 Ne size. Each generation a mutation occurs in the population. New mutations have low chances of surviving to the next generation, and are removed just by chance from the population. In probability proportional to Ne, some mutations start increasing in frequency (Y-axis) over time, and few of them reach fixation.

The change of the frequency of a single allele over time can be approached by means of the diffusion process (Hartl, 1980) and as a Brownian movement for frequent variants (i.e. minimum allele frequency (MAF) in the population > 0.01, for example) (Cavalli-Sforza & Bodmer, 2013).

If we consider the information of which marble we sample each time, in addition to its colour -status-, then the change in frequency of a mutation is the consequence of the fact that some marbles are just by chance more ascertained than others.

In the context of population genetics, this implies that in a finite population some chromosomes reproduce more than others. Using this principle, we can predict that backwards in time there will be, in probability, some chromosomes from one generation that are copies -share a common ancestor between them- from the previous generation. This process of coalescence backward in time of chromosomes towards the previous generation ensures that all copies of a given nucleotide in the current generation are descendants in the past from a single common ancestor (Figure 11). This common ancestor shared by all samples of a population is referred to as most recent common ancestor (MRCA).

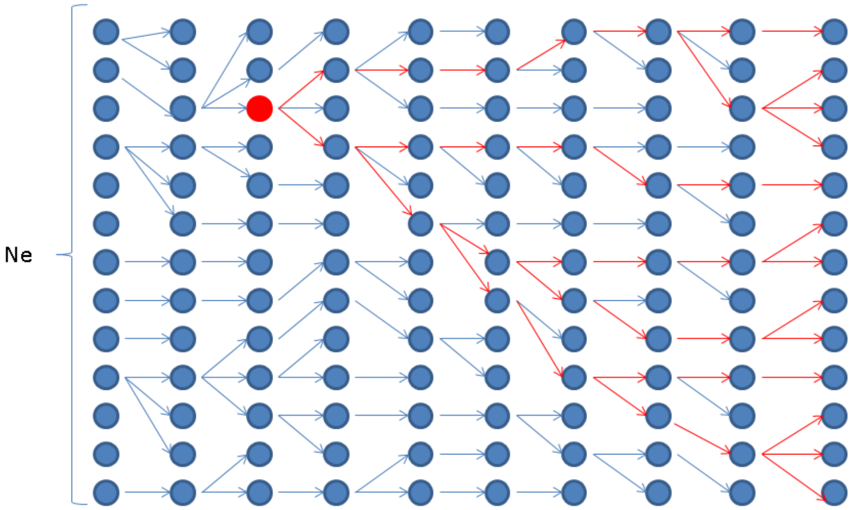


Figure 11. A toy example of the process of coalescence in a Wright-Fisher model. Each circle represents a chromosome, each column a generation, and each arrow how many copies forward in time each chromosome parent produces. Some chromosomes produce by chance more than one copy, and others do not produce offspring. The ultimate consequence of this process is that all the chromosomes from the present generation share a most common ancestor (MRCA, red circle) somewhere in the past.

The rate at which some of the chromosomes from a given generation share a common ancestor, or coalesce, with the previous generation is a function of the number of chromosomes that reproduce at each generation:

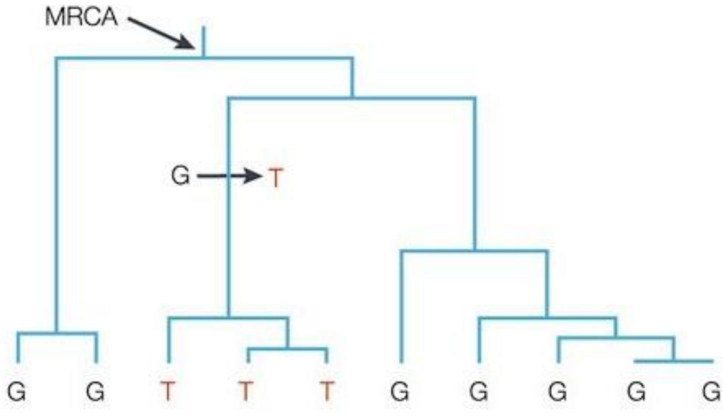
$$E[T_w] = D$$

$$E[T_b] = D\ \left(1 + \frac{D-1}{M\ D}\right)$$

Source: Wakely, 2016.

In the equation shown before T_w corresponds to the time to the MRCA between 2 given sequences in the same subpopulation; T_b corresponds to the time to the MRCA between 2 given sequences in different subpopulation; D corresponds to the number of subpopulations; M corresponds to the movement of lineages between subpopulations.

Because the behaviour of the coalescence process can be mathematically predicted, it is very easy to simulate neutral genomic regions. First, we generate the coalescence tree of the different sampled sequences and then we add mutations to the topology given the length of each branch and the mutation rate of the simulated genomic region (Wakeley 2009; Figure 12).



Nature Reviews | Genetics

Figure 12: Genealogy tree constructed around polymorphism on a given gene. Source: Rosenberg & Nordborg, 2002.

The coalescence process becomes more complex when recombination is taken into account, which is usually approximated by means of the ancestral coalescence graph (R. C. Griffiths, 1991). Several simulators (Hoban et al., 2012), such as *ms* (Hudson, 2002), extend the basic coalescence process to incorporate recombination. Others, such as *fastSimcoal2* (Excoffier et al., 2013) apply shortcuts to the coalescence process to efficiently allow simulating large (i.e. megabases) genomic regions.

Genetic drift has several important effects on evolution:

1. From a frequentist (that is, the study of the frequency of a mutation in a population) point of view, the ultimate fate of a neutral mutation is either being fixed (i.e. reaching a frequency of 1 in the population) or being removed from the population (i.e. frequency = 0). In this sense, genetic drift is a force that reduces the variation present in the population. However, because genetic drift can raise the frequency of a rare mutation towards fixation, at intermediate steps towards the fixation genetic drift increases the variation in the population.

2. It is stronger the smaller the effective population size gets.

3. It can drive speciation processes.

Given the direct dependence between random sampling and effective population size, a process in which genetic drift becomes really important is when a population suffers a bottleneck. This happens when the effective population size contracts to a significantly smaller size in a very short period of time due to a random event (usually environmental). Because of the randomness of the event, the chances of survival of each individual in the population are random and are not improved by any inherent genetic advantage. The bottleneck causes a drastic change of the allele frequencies that is independent from selection.

From a coalescence point of view, a bottleneck process implies multiple coalescent events from the generation at which starts the bottleneck towards the parental generation (Figure 13).

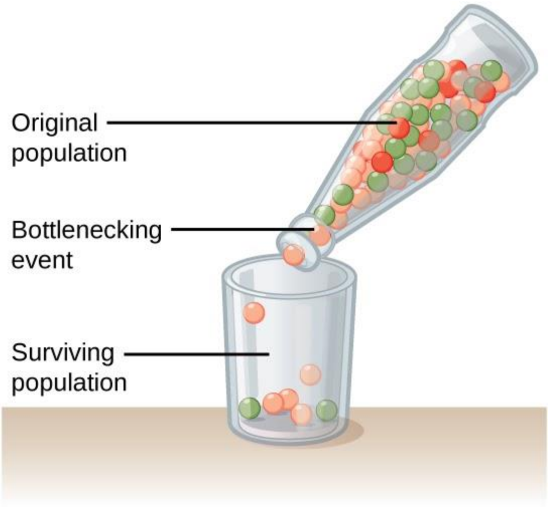


Figure 13: Graphical example of a bottleneck and how it can affect the diversity of generations after the event. Source: Biology at OpenStax, chapter 19, section 2

An example of a bottleneck is the high proportion of individuals with red colour blindness (achromatopsia) in the Pingelap atoll in Micronesia. The bottleneck can be traced to a typhoon in 1775 that left around 20 survivors on the island, one of them being a carrier of this genetic condition. The effect of the bottleneck was first seen in the fourth generation after the event, when around 3% of the population was affected. In the sixth generation around 5% was affected. Nowadays the 10% of the population is affected and another 30% are unaffected carriers (in comparison in the US it has a prevalence of 0.003% are affected) (Hussels & Morton, 1972).

b) Population substructure and Migration

When the probability of ascertaining mate is not uniform, population substructure occurs. The sources of such population substructure are multiple. In species -such as *Homo sapiens*- where the reproduction depends on the physical contact of the mates, and mobility is limited or the species covers a large geographic range compared to the amount of mobility, geography is a strong player in shaping the genetic variation of a population. Mountains (such as the Himalayas, Qiong et al., 2017) or water landmasses (such as between the American continent and the European and African, Luiz et al., 2012) can represent genetic barriers for humans.

In humans, in addition to physical barriers, cultural factors such as the religion or the language among others, can condition the mating preferences of an individual. When the mating preferences are shared between a set of individuals, the population is structured into subpopulations or demes. Between these subpopulations genetic exchange is possible if there exists migrants. From a coalescence point of view, a subdivision of the population implies that the time of the most common recent ancestor must precede, backward in time, the time when the barrier was established (see Figure 14).

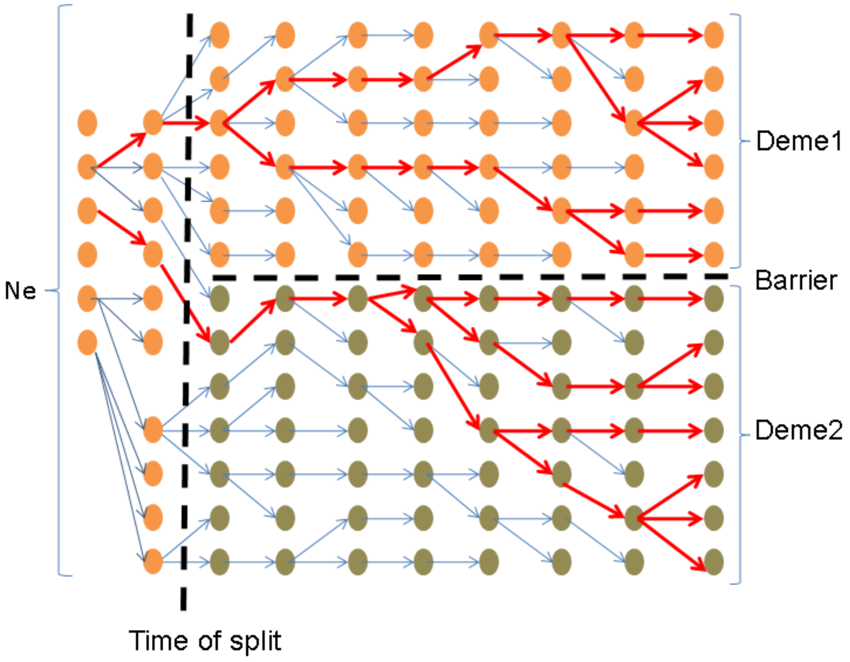


Figure 14. A toy example of the coalescent process under population substructure. Each dot corresponds to a chromosome. Each column corresponds to a generation. The vertical dash line indicates when the population is subdivided in two demes. The horizontal dash line indicates the physical subdivision. Read arrows indicate the ancestors of the current generation of chromosomes. The time of the most common recent ancestor must predate the time of split (not shown in the graph).

Migration, also called gene flow, is the transfer of variants from one sub-population to another, via individuals or gametes. This interchange is very important, as it can prevent two populations from diverging due to genetic drift (theoretically, with only one migrant in ideal populations, Wright, 1969) and, if it is high enough, two populations can have equivalent allele frequencies and be, effectively, a single population.

Migration is also important for the fact that migrants can carry new variation to other populations, even some gene forms that did not exist previously (Figure 15).

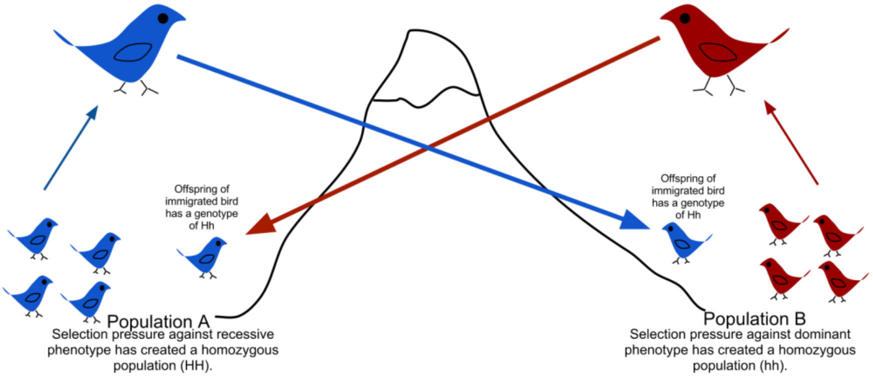

Offspring of immigrated bird has a genotype of Hh
Offspring of immigrated bird has a genotype of Hh
Population A
Selection pressure against recessive phenotype has created a homozygous population (HH).
Population B
Selection pressure against dominant phenotype has created a homozygous population (hh).

Figure 15: Example of migration between two populations. Image source: (Krueger, n.d.).

If migration in a population is too low (or impeded), its inbreeding usually increases. For example, many island populations have low rates of gene flow due to geographical isolation and small population sizes. A usual example of this situation is the state of the Black Footed Rock Wallaby, which has several populations in islands off the coast of Australia. This has led to an increase of the inbreeding in each of these populations (Eldridge et al., 1999).

The amount of genetic differentiation between demes is a consequence of when the population got structured into demes, the amount of genetic drift within each deme (i.e. their particular Ne) and number of migrants that the demes exchange (Weir & Cockerham, 1984).

1.3 Selective factors

Generation of genetic variation, through mutation and recombination, and shaping of the genetic variation by random processes such as gene drift and migration, are not enough to explain the distribution of all the observed allele frequencies. Selection (or natural selection) pushes for some variants to be more frequent, or to disappear from the population, depending on the effect of the variant in the fitness of the carriers. Selection was first described by Charles Darwin in a set of papers, published with Alfred Russel Wallace, in 1858 (Darwin & Wallace, 1858), and elaborated in *On the Origin of the Species* the following year. One has to take into account that selection acts at a phenotype level, which usually is the composition of a number of genetic factors. These forces are the main way to explain adaptive evolution (specialization in certain ecological niches, for example) and, in some cases, are strong enough to result in speciation.

Selective pressures are classified depending on the sign of the selective pressure: positive selection (a particular variant is selected in favour), purifying selection (new variants are selected against) and balancing selection (multiple alleles are actively maintained in a gene pool).

Selective pressures may act in a directional manner: it may be favoured and propagate (positive selection) or disfavoured and eliminated (negative selection) from the population. However, the classification of a particular mutation into one of these categories is not monolithic over space and time. The type of selection can be different depending on the environment the individual is living. For example, in sickle cell disease, the mutated allele is clearly detrimental (as it causes the disease in a homozygous state) but in locations where malaria has become endemic this allele is maintained in heterozygotes as it confers an advantage against malaria. In a meta-analysis done in 2018 (Wastnedge et al., 2018), it was found that the meta-estimate of heterozygotes in Africa (where malaria is endemic inside the tropical region) is around 16,121.91 per 100,000 births, while in Europe (where malaria is not endemic in the larger part of the continent) this meta-estimate falls to 803.57 per 100,000 births.

a) Positive selection

Also called directional selection, is a type of natural selection in which a phenotype is favoured over all others. This effect causes a huge shift in the allele frequencies that are present in the population, favouring those that represent the selected phenotype. The change in allele frequency of the favoured allele is independent of the dominance of the allele. In some cases, even recessive alleles can become fixed in a population.

The identification of the fingerprint of positive selection in the genome has a long tradition in the field of population genetics. Classical methods focus on detecting hard selective sweeps in the genome (Pavlidis & Alachiotis, 2017). Under this model, a new mutation confers a strong fitness advantage to the carriers. As a consequence, the mutation increases rapidly in frequency in the population, reaching fixation in a relatively reduced amount of evolutionary time. Depending on the timing of the selective event, different statistics have been defined for detecting positive selection within a species (Figure 16).

Depending on which is the type of genomic feature that is used for identifying the fingerprint of positive selection, methods can be classified in site frequency spectrum (SFS) based and linkage disequilibrium (LD) based.

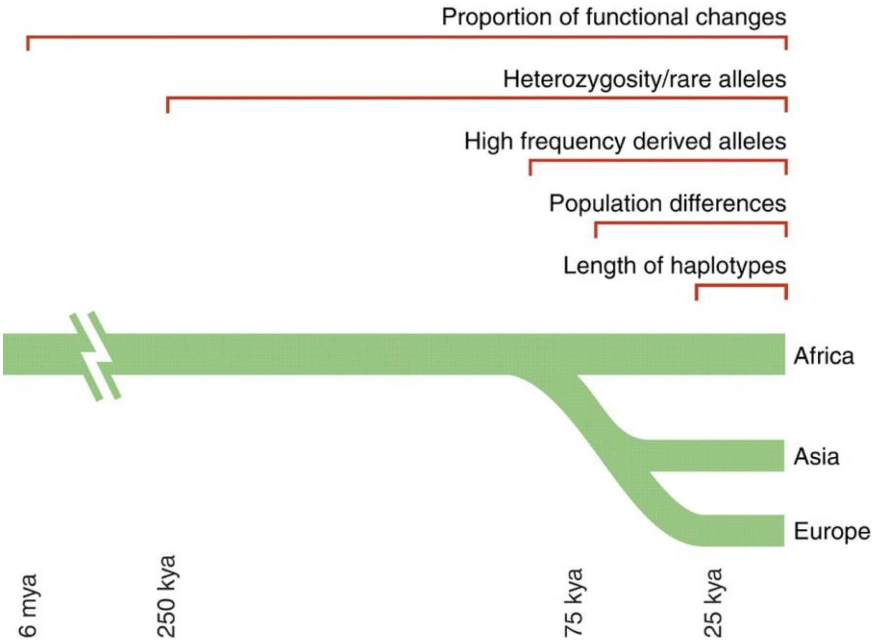

Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

Figure 16: Time scales for the signatures of selection. The five signatures of selection persist over varying time scales, with a rough estimate of how long each is useful for detecting selection in humans. Source: Sabeti et al., 2006.

SFS-based methods

Over a region of the genome, the full SFS is defined as the distribution of the number of SNPs whose derived allele is observed at a particular frequency in the population. Under the null hypothesis of neutrality, and assuming the Wright and Fisher model of neutral demographic evolution, the SFS of a population follows an exponential distribution (Wakely, 2016; Figure 17).

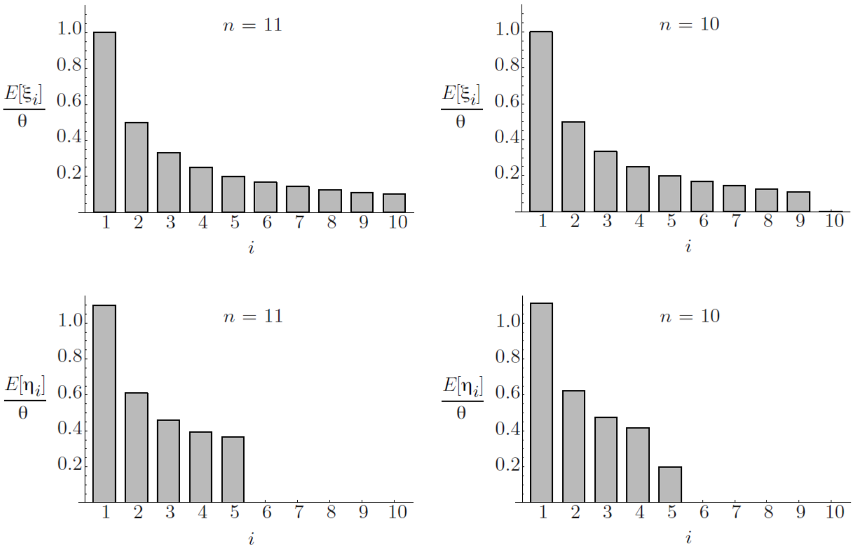


Figure 17: Relative expected numbers of unfolded ($E[\xi_i]$) and folded ($E[\eta_i]$) site frequencies (represent the mutation rate) for 2 different population sizes (n = 10 and n = 11). Source: Wakely, 2016.

That is, under neutrality the derived allele of a large fraction of SNPs is present only in one chromosome in the population (i.e. they are private alleles in the population). Under an event of positive selection from the hard selective sweep model, a rare derived allele, and its surrounding genetic variants, increases rapidly in frequency in the population, creating an excess of SNPs whose derived variants are close to fixation (Figure 18).

Different statistics, such as Tajima's D (Tajima, 1989, 1993) or Fay & Wu's H (Fay & Wu, 2000), take advantage of the shape of the SFS under the neutrality and under the expected bump of derived variants at high frequency in the population under a hard selective sweep.

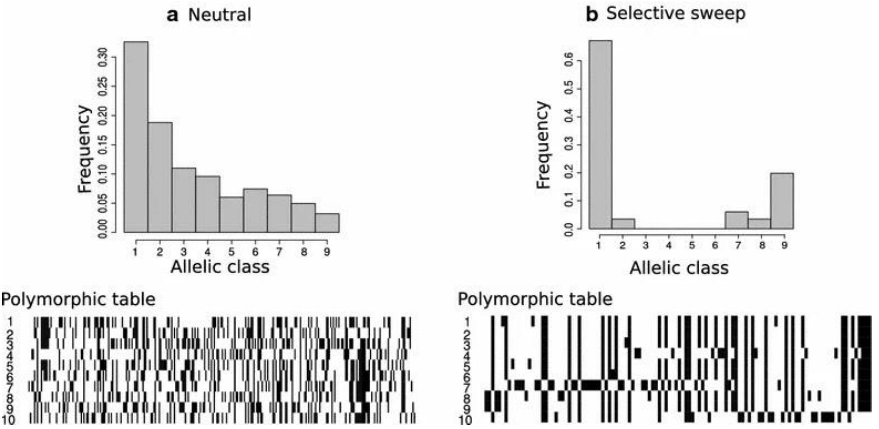


Figure 18: SFS signature of a neutral region (A) and a selective sweep (B). In the polymorphic table *black* squares denote derived alleles, while *white* squares denote ancestral alleles. Source: Pavlidis & Alachiotis, 2017.

Tests based on comparing the allelic frequency among populations (Fst-based and similar statistics) also use the SFS. They use the information of the allelic frequency in populations where it is supposed that the selective event did not take place against a target population where the positive selection is being tested.

In the populations where there was no selective event, the frequency of the derived allele will change over time due to stochastic factors (i.e. genetic drift). In contrast, in the population under selection, the frequency will be driven by positive selection. Thus, we can expect that variants that are highly divergent between genetically similar populations correspond to differential selective pressures.

For example the allele found to be selected in the Bajau people to be better divers is also found in other populations (nearby Saluans), but at a much lower frequency (37.1% in Bajau vs 6.7% in Saluans) (Ilardo et al., 2018).

Linkage disequilibrium methods

LD based methods take advantage of the recombination information around a variant under positive selection. The genomic context where a mutation appears is broken over time by recombination (Figure 19).

Nevertheless, because genetic variants under a hard selective sweep increase rapidly in the population, the speed at which recombination is able to generate new variation is slower than under neutrality. In practice, this means that the haplotypes of genetic variants under a hard selective sweep will tend to keep the genetic background where the mutation initially appeared, and extend this genetic background longer than expected under neutrality.

Several methods have been developed to take advantage of this fact, for example: EHH (Sabeti et al., 2002), iHS (Voight et al., 2006) or EHH-XP (Sabeti et al., 2007).

Hybrid algorithms have been proposed by combining the properties of both types of philosophies, as well as machine learning approaches that extract the information from the different methods.

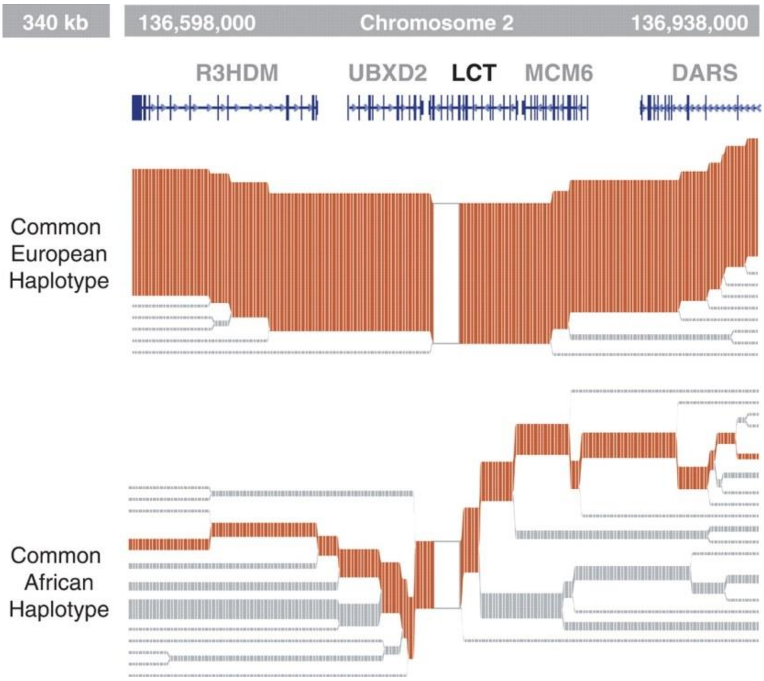


Figure 19: Haplotype around the lactase gene. In the African population it has been broken due to the generations passed since the mutation appeared. Source: Sabeti et al., 2006

b) Positive selection on standing variation and polygenic adaptation

The profusion of methods for identifying hard selective sweeps during the last decade has provided some astonishing results about the recent evolution and local adaptation of particular traits in different human populations (for example, Fan et al., 2016). However, it also pointed out the difficulties of consistently identifying such signatures in the genome when using many tests at the same time (i.e. Pybus et al., 2015), as well as to the surprisingly reduced number of genes that are under hard selective sweeps in the genome.

One explanation for such results is that many of the traits are not monogenic, but are rather complex traits in which many alleles are involved in the expression of a given phenotype.

Another possible explanation is that positive selection in a way that we couldn’t properly detect. Until now we have developed methods to detect hard selective sweeps (a rare allele raise in frequency and lowering variability around it over a few generations), but, perhaps, most of the genome is not subject to a hard sweep. The idea of soft sweep (as opposed to hard sweep) is as follows: selection is not acting over recent rare alleles, but on already present alleles that promote a better adaptation due to changes in the habitat.

An example of a soft sweep is what Hamblin and Di Rienzo (Hamblin & Di Rienzo, 2000) found in Sub-Saharan populations. They found a fixed (or near fixed) null allele for the Duffy blood group that was virtually absent in other populations (individuals from central Italy). But, in four out of the five Sub-Saharan ethnicities studied they did not find a high proportion of rare alleles or a decline in variation in the region studied. At first, they proposed that this panorama can be due to recombination, demography in the form of population structure or other factors.

This type of selection is important as it does not need to create new variation (as classical positive selection) but acts on already present variation that does not have a noticeable effect on fitness.

c) Purifying selection

Also called negative selection (as opposed to positive selection), purifying selection is the selective removal of deleterious alleles. This type of selection is responsible for the stabilization of the genetic background of a species. We can also find a more short term negative selection: most biological structures represent a conditional optimal (as they usually depend on other pieces of the machinery to perform its function). This is further complicated by the fact that it can also depend on the ecological conditions in which the individual lives in.

If negative selection in a locus is strong enough it can lead to removal of spatially linked variation, independently to the effect it has on fitness. The purging, in the long run, decreases the level of variation in the zone around the locus. This effect is also called background selection (Charlesworth et al., 1993). An example of this effect is what happens with housekeeping genes (genes that are required for the maintenance of basic cellular functions). Almost any mutation in any of these genes will be deleterious and it will be taken out, due to the importance of these genes.

We can also consider the opposite case, when purifying selection is too weak. In this case the accumulation of deleterious mutations can lead to the extinction. Sometimes this can be counteracted by back mutations, a type of mutation that restores the original sequence.

We can find an example in humans regarding the IFN-γ in African populations. In the work of Campbell, Smith and Harvey (Campbell et al., 2019) they have found that this gene is under purifying selection, specially 3 variants found in intron 3 from this gene.

d) Balancing selection

In diploid (or polypoid) organisms we can find more subtle combinations of positive and/or negative selection. Balancing selection refers to a type of selective process that maintains multiple alleles of a locus at intermediate frequencies. This can happen in various ways, mainly 2: heterozygote advantage and frequency-dependent selection.

- <u>Heterozygote advantage:</u> in this case the heterozygotes have a higher relative fitness than both the homozygous dominant and homozygous recessive. This happens in environments where the heterozygote is both advantageous and disadvantageous, but the homozygotes are disadvantageous. An example of this is Sickle-cell Anaemia. The homozygous for the normal allele (HgA) is selected against due to malaria while the homozygous for the mutant allele (HgS) is also selected against due to the disease. The heterozygous state confers resistance to malaria. In zones where malaria is endemic, the HgS allele is maintained thanks to the higher fitness of the heterozygotes. In zones, where malaria is not endemic, the allele has decreased fitness compared to the homozygous normal allele.

- Frequency-dependent: this occurs when the fitness of a given phenotype is dependent on its relative frequency to the other phenotypes. This relation can be positive (when the fitness increases the more common an allele is) or negative (when the fitness decreases the more common an allele is). A clear example of the second is the relation between prey and predator. As predators tend to hunt the most common phenotype of the prey, this is selected against, but other less common phenotypes have higher fitness. Then a new cycle starts when one of the once rare alleles becomes the common one.

From a coalescent point of view, each type of selection shows a different fingerprint in the genome (Figure 20)

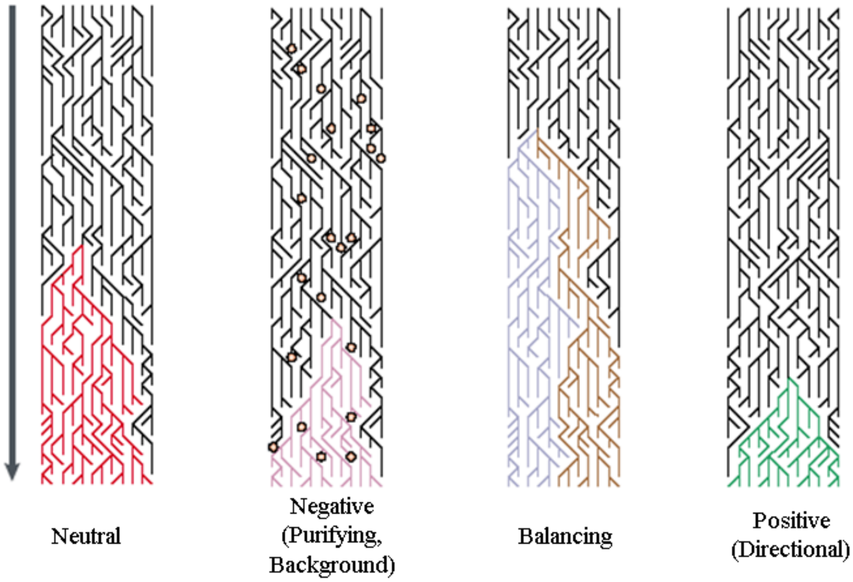


Figure 20. How the shape of the gene genealogy is modified by the different types of evolutionary forces. Source: Bamshad & Wooding, 2003

Given that the different types of selective processes modify the genetic variation in a functional genomic region and the gene genealogy is not independent anymore of the genetic variation (Figure 20), a basic problem when attempting to do demographic modelling is which genomic regions must be considered. The classical view of "*almost everything is neutral in the genome*" is being replaced by an "*almost everything is under background selection*" point of view (Pouyet et al., 2018).

Many of the described selective pressures tend to minimize the genetic variation within a population and to increase the amount of divergence between populations. Therefore, using these regions as neutral ones implies that the demographic estimates will tend to be biased towards demographic processes that produce similar signatures in the genetic variation of the genome. Namely, decreasing the effective population size (so the amounts of genetic variation in a region are lower) and increasing the time since the separation of populations (so the genetic differentiation among populations increases).

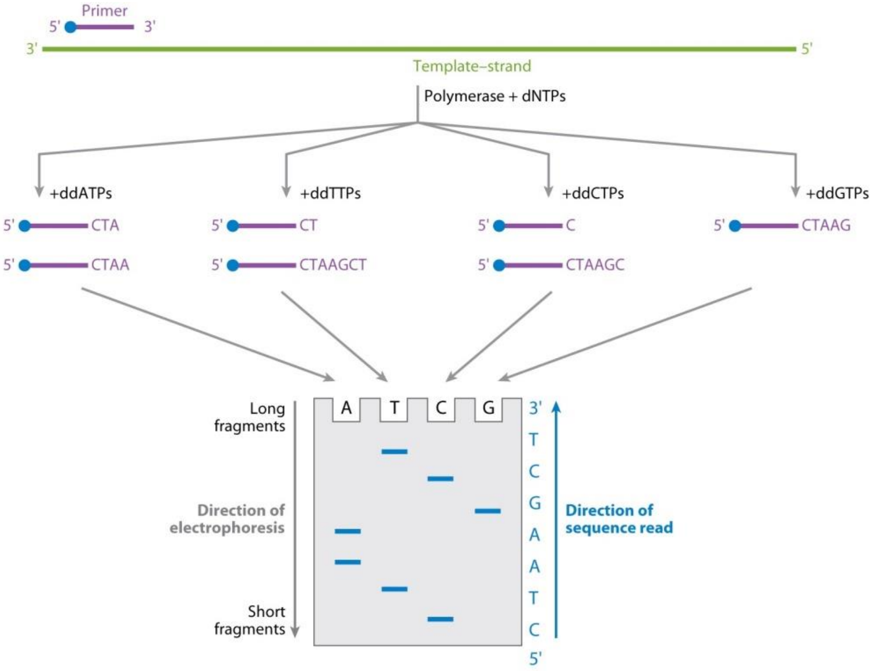
On the other side, the conservative way of considering almost everything under purifying selection produces very limited amounts of data and sparse genomic regions, hindering the inferences (Pouyet et al., 2018). As a compromise, investigators use regions that are putatively not functional (i.e. they are out of genes or CpG islands (Allentoft et al., 2015; Mondal et al., 2019) to do demographic inferences.

1.4 Identification of genetic variation

Until now we have discussed how variation appears and how this variation can change across generations in a given population. Nevertheless, in order to use these changes in evolutionary inferences we have to be able to find and to characterize these changes. This is a very important step, since the way how these changes in DNA are identified can introduce potential biases in the final detected genetic variation and influence our conclusions.

One of the first ways that variation was explored was using blood markers (now called classical markers) such as blood groups or subtypes of proteins found in blood. These markers were easily obtainable, but offered limited information about demographic and evolutionary processes. However, these markers were, by 1978, first used to make geographic and evolutionary inferences in European human populations (Menozzi et al., 1978) and extended to worldwide populations in the monumental "The History and Geography of Human Genes" (Cavalli-Sforza et al., 1996).

The era of sequencing started with the publication in 1977 of the full genome of the bacteriophage φ X174. In 1986 the first semi-automated sequencer was patented and presented, followed in 1987 by the first full-automated sequencer from Applied Biosystems: the ABI 370, which used a novel technique (different fluorescent markings on the terminating nucleotides) that allowed sequencing on a single lane. Before, the method used for sequencing any piece of DNA was Sanger sequencing: four different sequencing reactions (one of each termination nucleotide) that were resolved using a polyacrylamide-urea gel in four different bands and being visualized using autoradiography or UV light (see Figure 21).



Mardis ER. 2013.
Annu. Rev. Anal. Chem. 6:287–303

Figure 21: Overview of how Sanger sequencing works. We have four different PCR reactions, each one containing a mixture between normal nucleotides and different termination nucleotides (ddNTPs). The results of the four reactions are then separated in the polyacrylamide-urea gel and the sequence can be reconstructed. Source: Mardis, 2013.

In 1981 the concept of DNA arrays was developed, although at the time it was a macro array. The array technology is based on the immobilization of DNA or RNA molecules in a solid support. Each spot contains a specific sequence (called probe) that is used to hybridize cDNA extracted from the sample, using high-stringency conditions. Then the probe-target hybridization is detected using a fluorophore or chemiluminescence techniques. The first use of a microarray was by Mark Schena in 1995 (Schena et al., 1995) to measure gene expression in the flowering plant *Arabidopsis thaliana*.

All these different techniques and technologies were the base for the commercial use of sequencing and microarrays at the start of the 2000s. These new methods (called "Next Generation Sequencing, NGS, or Second Generation Sequencing) were characterized by an easy and high scalability, which in turn allowed the sequencing of the entire genome of an individual.

These "easy-to-get" genomes carried a series of problems of their own, requiring bioinformatic tools to process the vast amounts of information that these technologies were getting.

Current high-throughput technologies for defining the genetic variation of a sample can be classified in two main categories: microarrays and next generation sequencing.

a) Microarrays

Microarrays are a technology based on the principle of sequence complementarity (i.e. the more complementarity two sequences have, the stronger the bond is). A sequence of DNA/RNA (called probe) is immobilized on a surface, typically glass or plastic, then the DNA of interest (called target) is added and those targets that are not hybridized are washed away.

The probes are marked (radioactive or fluorescent markings are the most commonly used ones), and then they are detected using the proper methods.

This kind of technology can be used to detect the levels of transcription of certain genes (you can measure the strength of the mark of the label) or to genotype known genetic variation (especially SNPs). This latter use is accomplished using two different probes in the same dot, labelled with different fluorescent colours for example (Figure 22). There were two main technologies: Affymetrix and Illumina.

In the case of Affymetrix we have a collection of 25-mer probes for both alleles fixed on a plastic/glass plate that differentiate between them in the position of the SNP in the sequence. The sample DNA will bind to these probes. However, due to sequence complementarity it will bind stronger in the sequence with the perfect match (the 25-mer with the SNP in the proper place and the proper allele). The one with the brightest signal from the probes corresponds to the correct genotype of the sample. For a simple example check Figure 22 a.

Illumina opted for a different approach. In its technology, the probes were fixed on beads and were longer, 50-mer. In this case the sequence on the bead ended just before the SNP and a step of a single-base extension was performed, adding one of the two possible alleles. These nucleotides were marked with fluorescent colours (red and green in the example). For a simple example check Figure 22 b.

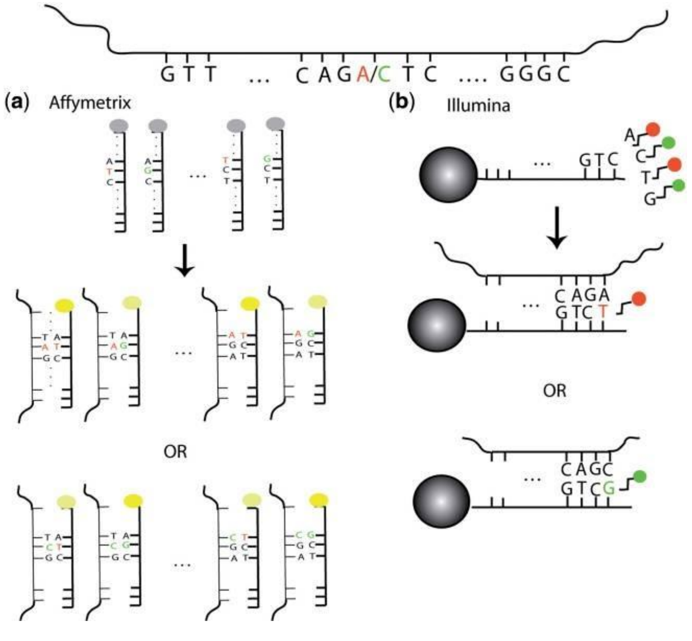


Figure 22: Overview of how the most prominent SNP genotyping arrays work: Affymetrix (a) and Illumina (b). Affymetrix uses a set of different probes that have the same sequence with the exception of the location of the SNP; the probe that has a higher degree of complementarity will have a stronger signal. In the case of Illumina we detect the SNP by incorporating a fluorescently labeled nucleotide using the sample as primer. Source: LaFramboise, 2009.

In both cases a computer algorithm is needed to transform the brightness of the signal (Affymetrix) or the colour in the well (Illumina) into a proper inference of the genotype present on the sample.

In the case of Affymetrix the algorithm produces the inference depending on the relation between the perfect match (PM) probe for allele A or B and the mismatch (MM) probe for allele A or B. The PM is defined as the probe that has perfect complementarity to the target allele and MM is defined as the probe identical to the PM with the exception that the allele in the SNP is altered as to not be complementary to either allele (it can be more than one MM probe per SNP). Depending on the brightness from these probes (PM_A, MM_A, PM_B, MM_B) the algorithm can infer the genotype of the sample: AA, AB or BB.

In the case of Illumina the algorithms had a much easier work: the inference of the genotype was more direct, as we have two different colours for every SNP (green and red in the example). Depending on which colour was more intense in the sample the algorithm determines the genotype of the sample: AA (red), BB (green) or AB (both red and green).

When this technology was first used in humans around 1998 (D. G. Wang, 1998), only 1494 SNPs were genotyped in every sample (Affymetrix HuSNP assay). Advances in the genomic architecture of human populations thanks to projects such as the Human

Genome Project (International Human Genome Sequencing Consortium, 2001) and the HapMap project (The International HapMap Consortium, 2003) in addition to technical advances in Bioinformatics and wet lab have allowed to exponentially increase the number of SNPs that can be interrogated in a single array. In the last version of both Affymetrix and Illumina there has been an effort to detect and genotype copy number probes to interrogate non-SNPs human genetic variation.

Furthermore, in recent times SNP microarrays have become really cheap, allowing the genotyping of hundreds or even thousands of individuals for the same study. This fact can partially solve the batch effect that occurs when comparing datasets generated at different times and labs. In the past, as every study had its own protocols or used different platforms to search for variation, comparison between samples was tricky. Using microarrays, as you are analysing all individuals on that study using the same platform and protocols solves part of the batch effect problem. However, the problem of batch effect, even if it is minimized, it is not completely fixed (J. Luo et al., 2010).

One of the main problems with arrays is the lack of new variation discovered in the samples. As arrays work with fixed DNA molecules, the results we can get from an array are fixed from the start. The decision of what SNVs include is also crucial, as it can be counterproductive: for example, if we choose to include SNVs that are typically found in European populations, we will see that in most other populations these SNVs are genotyped as reference alleles. The reality is that we are losing the real source of variation for these populations (Albrechtsen et al., 2010).

b) Next-generation sequencing

Next-generation sequencing (or NGS) is a series of techniques that operate on the basis of splitting the genome of a sample in individual molecules of different length by digesting it with a restriction enzyme or mechanically breaking it, and then sequencing these individual strands. To be able to capture and sequence the DNA, a first step of ligation with an adapter is needed (owned by each platform). Then, the sequencing is conducted in each molecule by different protocols and the platform is able to detect the incorporation of each individual nucleotide, normally using different fluorescent molecules.

In 2005, the first commercially available NGS instrument was mentioned on a scientific publication (Margulies et al., 2005) and, in the next 2 years, it became obvious that NGS was meant to have a major impact on our ability to explore and answer genome-wide questions with more than 100 related manuscripts in this period (Mardis, 2008).

Initially, three different platforms were available: Roche’s 454, Illumina’s SOLEXA and Applied Biosystems SOLiD.

In the case of Roche’s 454, the molecule is binded (through the adapter) to agarose beads. This population of sample-beads is then isolated into micro-reactors (oil:water micelles that contain the PCR reactants). This is followed by a step of PCR amplification inside the micro-reactors, usually producing a million copies of each DNA molecule. This emulsion is then transferred to a picotiter plate (a fused silica capillary structure) that can hold a single bead in each of its single wells. In these wells takes place the pyrosequencing reaction: the incorporation of single nucleotides is detected thanks to the release of pyrophosphate which starts a chain reaction that ultimately produces light thanks to the firefly enzyme luciferase. This technique has the problem that when the same nucleotide is repeated more than a few times (usually more than 6), the base calling algorithm is not able to properly differentiate the signal and is prone to commit insertion and deletion errors.

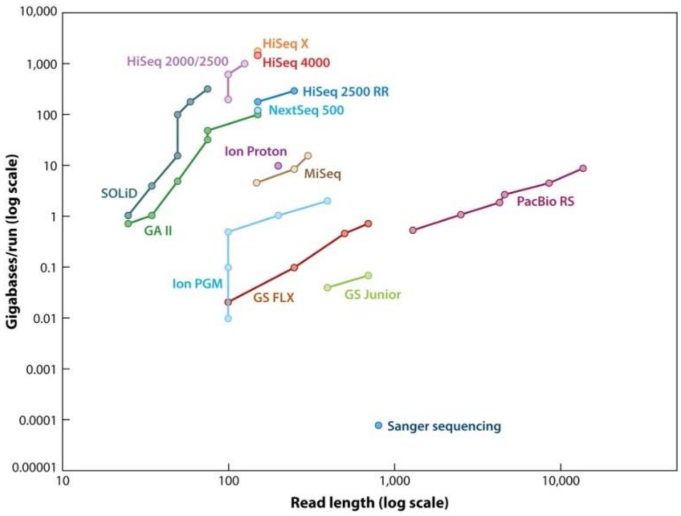
Illumina’s SOLEXA first steps are the same as before. Then, the DNA-adapter molecule is bound (through the adapter) to the inside surface of the flow cell channels. The sample DNA is amplified using a classical PCR reaction. Since the adapter sequence is also copied, the PCR products will stay close to the original molecule, creating a sort of cluster of around a million copies of the same DNA molecule. Then we add the reagents of the sequencing reaction, especially a set of fluorescently labelled nucleotides (each different base is labelled with a different fluorescent colour). When this nucleotide is incorporated, it blocks further nucleotide incorporations and creates a fluorescent event that the platform can detect, because the signal is amplified in each cluster. After this imaging step, the nucleotide is modified so the next nucleotide can be incorporated. This series of steps continues for a specific amount of cycles. After these sequencing steps, a base-calling algorithm is applied to assign the proper base and to establish the quality to each of the imaging steps.

Finally, the third NGS method available was Applied Biosystems SOLiD platform. This platform uses a similar approach as SOLEXA, with the main difference being that it reads two bases at every imaging step and it considers a set of primers for the adapter that moves the frame of the sequencing. After the sequencing steps we still have the need for an algorithm for the base-calling and to assign quality to the bases sequenced.

After the proper sequencing steps in all three platforms end, we obtain a file with the collection of reads for the sample that is being processed. After this an alignment step is needed. This step uses an algorithm to find the proper coordinates in the genome for every read, usually setting aside those that can map to multiple coordinates (as this reads can be useful to determine CNV or indel events). This algorithm assigns a quality score to simplify the fit of the reads for the given coordinates.

A final step is needed to obtain the variation found in the sample: SNP and genotype calling. The SNP-calling algorithm compares the bases in the reads to a reference genome and marks which of those bases are different between said reference and our sample. The genotype-calling algorithm examines the different reads that overlap a given position (as more than one read can map to a given position) and assigns a genotype and a score to the SNP present on the sample based. The genotype depends on the proportion of reads that carry the reference allele or the non-reference allele, while the score depends on the quality score for the alignment of the different reds used to determine the genotype.

In the recent years both the yield (amount of DNA sequenced) and read-length for these platforms has grown, almost at an exponential pace. Also, new technologies are appearing that allow for extremely long reads (up to 10,000 bases long in some cases) (see Figure 23).



Levy SE, Myers RM. 2016.
Annu. Rev. Genom. Hum. Genet. 17:95–115

Figure 23: Comparison between the length of the reads (X axis) and the total DNA (Y axis). Each colour represents a platform and each point represents a different machine of the same platform. Source: Levy & Myers, 2016

A common problem in all these platforms is the possible bias when considering if a base is an SNP. This can happen at multiple levels: the instrument can misread the base that is being incorporated, the base-calling algorithm does not call the proper base, the alignment algorithm matches the read to an incorrect place or the SNP-calling or the genotyping algorithm calls a SNP were it is none (or vice versa). To solve this inconvenience a quality filter is used, filtering out those bases that have a poor quality and are more prone to have a wrong base or genotype assigned (Nielsen et al., 2011). A different approach that can be useful is to examine the distributions of the different alleles across samples and test if said distribution fits within an expected distribution (Muyas et al., 2019).

In these techniques we also have the problem regarding batch effect. In this case it can be more severe than with microarrays, as different techniques can be applied in order to split the genome. This can affect the length of the fragments and, due to the limiting step of the number of amplification cycles, lead to differences in quality and/or mappability (chance of finding a unique location for a read) of the reads. Changes in both pose a problem for the aligner algorithm and, by extension, for the SNP-calling and genotype-calling algorithm (Leek et al., 2010). This can result in major problems if the genotype is associated with an outcome of interest leading to incorrect conclusions.

1.5 What does the observed genetic variation say about the demographic history of human populations?

a) Origin of anatomically modern humans

Already by 1980, a first tree was inferred using mitochondrial DNA (mtDNA) from different populations and the out-of-Africa model was accepted as the best model explaining human evolution. These results represented a milestone for human population genetics, as they unequivocally rejected other competing hypotheses explaining the origin of current anatomically modern humans (AMH), based on what was known from an archaeological point of view. In particular, there were three main classical hypotheses explaining human evolution (Stoneking, 2005):

1. The candelabra hypothesis proposed that AMH would have independently appeared in the different geographic regions as a consequence of the evolution of the first hominins that left Africa around 2 million years ago (mya); local adaptation would have produced the fossils of archaic hominins such as the Neanderthals in Eurasia, that date as far back as 400 thousand years ago (kya) (Higham et al., 2014), and these populations would have produced the current Eurasian populations.

2. The multiregional hypothesis stated that modern features evolved in a fragmented manner across several areas that then were connected through gene flow (Wolpoff & Caspari, 1997).

3. The out of Africa model proposed that AMH originated in Africa and then expanded outwards to the other continents, replacing other archaic populations (Vigilant et al., 1991).

Whereas they were equally likely from an archaeological point of view when these hypotheses were proposed, the expected signature in the DNA of each topology model was different. Both the candelabra hypothesis and the multiregional hypothesis expected higher levels of genetic diversity out of Africa, as well as old (>2 mya) coalescent times out of that continent. In contrast, the Out of Africa hypothesis implied greater diversity within Africa, as well as recent coalescent times in that continent.

The problem with this type of tree, and using mtDNA in general, is that it only reflects the female inheritance (although this is still a controversial issue, see S. Luo et al., 2018), and it behaves as a single marker, making it not representative of the overall genomic pattern and history of humans (Nordborg, 1998).

Further studies using autosomal microsatellite data (Ramachandran et al., 2005; N. A. Rosenberg et al., 2002) observed the predicted patterns under the out of Africa hypothesis, with higher genetic variation in the African continent and a decay of the genetic diversity with space as we moved far away from the African continent. Similar results have been observed in the analysis of haplotypic data (Jakobsson et al., 2008), as well as when using panels of SNPs (J. Z. Li et al., 2008).

The Out of Africa hypothesis became the most accepted hypothesis during the last decade of the last century and the first one of the XXI. However, after the sequencing of the Neanderthal genome (Prüfer et al., 2014) and the Denisovan genome (Meyer et al., 2012), evidence of archaic introgression in AMH were found out of Africa. In particular, all non-African populations contain around 2% of Neanderthal ancestry (Sankararaman et al., 2014). According to patterns of linkage disequilibrium, this admixture happened 50-65 kya (Sankararaman et al., 2012).

This observation, as well as the prediction of other archaic introgressions from currently archaic populations out of Africa (Mondal et al., 2019) but also within Africa (Lorente-Galdos et al., 2019), has forced considering a new model of recent human evolution based on partial assimilation (Smith et al., 2017) and the existence of an African meta-population (Scerri et al., 2019).

Overall, the current picture of archaic populations and interactions with AMH is becoming more and more complex (see Figure 24).

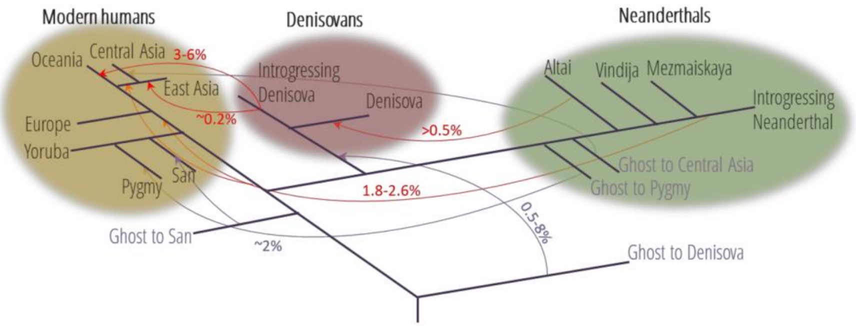
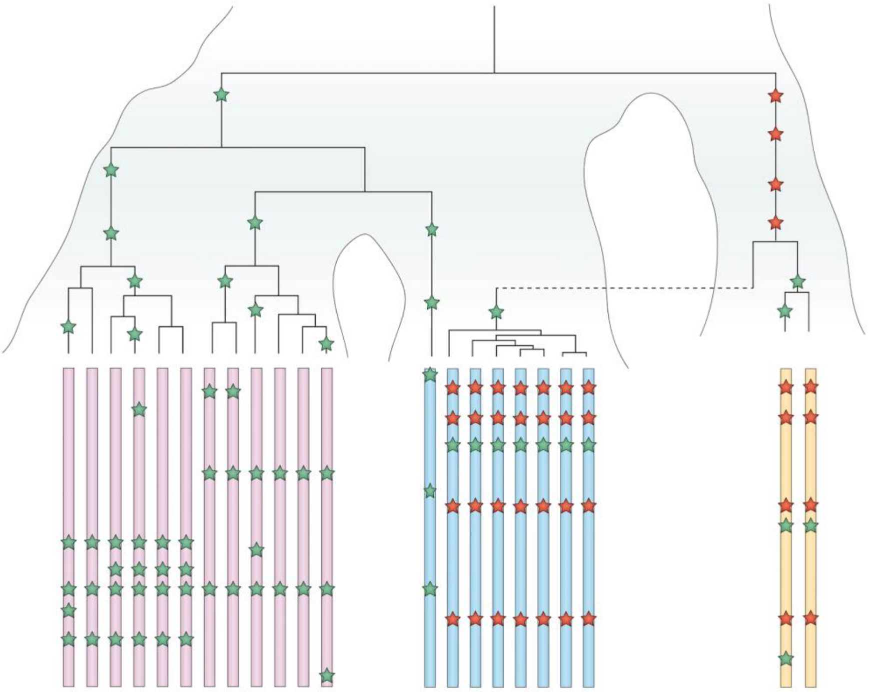


Figure 24: Family tree of the four groups of early humans living in Eurasia 50,000 years ago and the inferred gene flow between the groups due to interbreeding. Source: Dolgova & Lao, 2018.

The evidence of archaic introgression in the human genome has fuelled the development of methods that attempt to identify the regions in our genome that are introgressed from archaic populations using NGS data. The way how these methods usually identify the introgressed fragments is by considering simple demographic models such as the one depicted in Figure 25 (Racimo et al., 2015).

The regions that are found as introgressed in an individual by the different methods tend to be out of functional regions, and particularly in genes expressed in the brain (McCoy et al., 2017), which has been interpreted as evidence of purifying selection in the hybrid individual (Petr et al., 2019; Telis et al., 2020). However, it has been shown that some genomic regions show enrichment for archaic introgression supporting also the role of positive selection in the archaic introgression (Dolgova & Lao, 2018; Kelso & Prüfer, 2014).



Nature Reviews | Genetics

Figure 25: Pink and blue chromosomes represent modern populations. Yellow chromosomes represent archaic populations. Stars represent shared mutations. In the case of red stars, these mutations appeared on the archaic population and were passed to the blue population through an admixture event (dashed line). In this particular case an event made the red star mutation raises to high frequency due to selection. Source: Racimo et al., 2015

The probability of a region to be introgressed is approximated using variation maps built from big population studies such as the HapMap Project (The International HapMap Consortium, 2003) or the 1,000 Genomes Project (1kG) (The 1000 Genomes Project Consortium et al., 2015).

A problem with these gargantuan projects is the fact that samples are divided across different sequencing centres and technologies, which could lead to important batch effect problems. This has already happened in, at least, one of the populations of the 1kG. In the study from Anderson-Trocmé (Anderson-Trocmé et al., 2020) they found a particular mutational signature in the Japanese population from 1kG (JPT), that was not found in a more recent cohort from Japanese individuals (the Nagahama cohort). This fact could mean that the variation maps are biased and the introgression probabilities calculated from them could be skewed.

b) Demographic history of Europe

After the Out of Africa diaspora, human species have gone through multiple regional migrations, independently affecting each continent (Figure 26).

In Europe, AMH arrived around 43 kya (Benazzi et al., 2011). However current European populations are a mixture of these Palaeolithic populations and from others that migrated into Europe in more recent times (Günther & Jakobsson, 2016). Around 11 kya, the Neolithic populations of the Fertile Crescent (actual Middle

East) started to emerge (Asouti & Fuller, 2013) and this type of populations from Central Anatolia expanded to the present Europe (Günther & Jakobsson, 2016). This first wave of migrants absorbed the hunter-gatherers that were already established (Günther et al., 2015). This wave reached the Iberian Peninsula around 7 kya (Günther & Jakobsson, 2016). A second wave of migrants came from the herders of the Yamnaya culture from the Pontic-Caspian steppe, in modern-day Russia, about 4.5 kya (Allentoft et al., 2015).

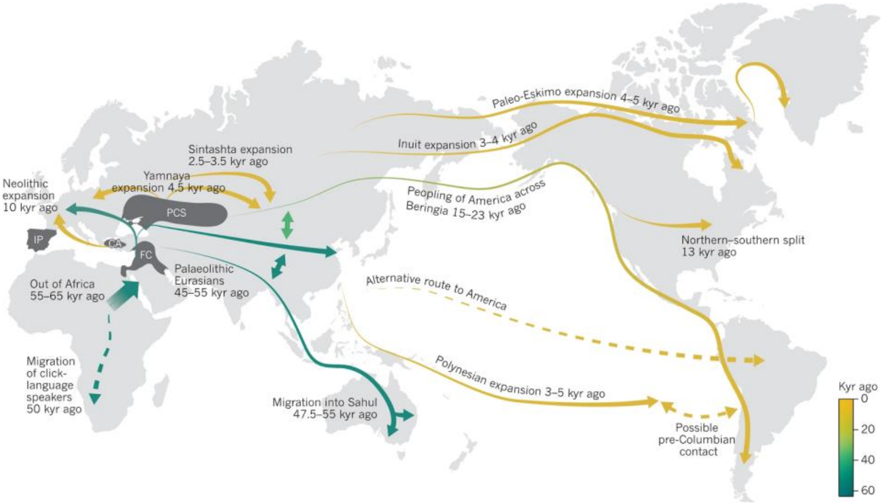

Paleo-Eskimo expansion 4–5 kyr ago
Inuit expansion 3–4 kyr ago
Sintashta expansion 2.5–3.5 kyr ago
Yamnaya expansion 4.5 kyr ago
Neolithic expansion 10 kyr ago
PCS
IP
CA
FC
Palaeolithic Eurasians 45–55 kyr ago
Out of Africa 55–65 kyr ago
Migration of click-language speakers 50 kyr ago
Migration into Sahul 47.5–55 kyr ago
Peopling of America across Beringia 15–23 kyr ago
Northern–southern split 13 kyr ago
Alternative route to America
Polynesian expansion 3–5 kyr ago
Possible pre-Columbian contact
Kyr ago
0
20
40
60

Figure 26: Major human migrations of AMH inferred from genomic data. This map shows a brief summary of the different migrations and estimated times in which they happened. FC = Fertile Crescent; CA = Central Anatolia; IP = Iberian Peninsula; PCS = Pontic-Caspian steppe. Peopling of Asia and America are outside of the scope of the present thesis. Source: Nielsen et al., 2017.

These three different components (early hunter-gatherers, farmers from the Fertile Crescent and herders from the Yamnaya) contribute in different ways to the modern variation found in Europe: for example, the wave of farmers from Central Anatolia has a more prominent ancestry in southern European populations such as the Sardinian people (Cavalli-Sforza et al., 1996).

From a continental scale, the genetic variation of current non-Romani European populations strongly correlates with geography (Lao et al., 2008; Novembre et al., 2008), showing a gradient of decreasing variation with increasing Northern latitude (Auton et al., 2009) (Figure 29).

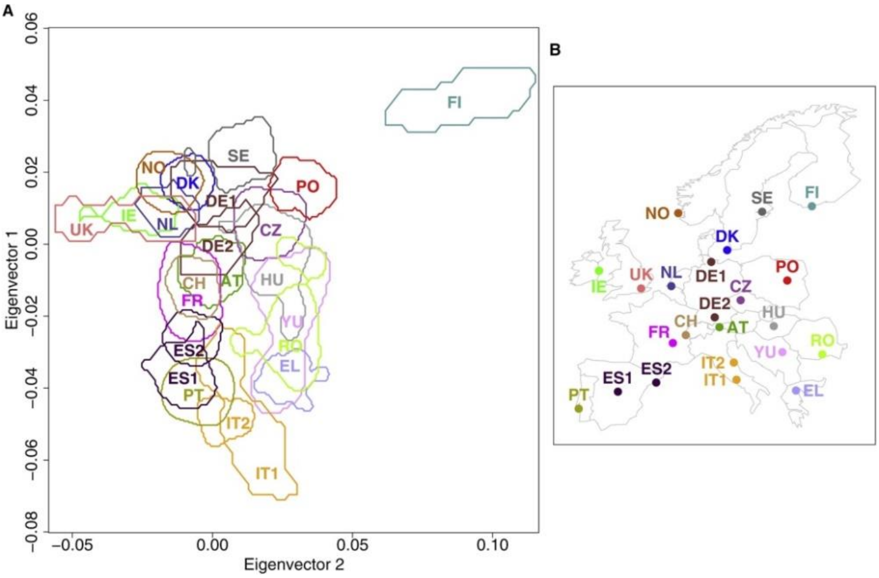


Figure 29: Density plot of the first two dimensions of PCA (A) and geographical distribution of the samples (B). Plot in (A) is the result of applying PCA to the collection of 309,790 SNPs used in the study. Plot in (B) corresponds to the geographical distribution of the country of origin of the samples used in the study. Source: Lao et al., 2008.

c) Regional European demographic history

Although the principal components of the European genetic background were established in these three waves, subsequent processes of gene flow that were limited by geography have shaped its present day landscape.

An example of geography as a limiting factor is the case of the British islands. The British Isles were firstly inhabited by Palaeolithic hunter-gatherers around 11.6 kya (Cunliffe, 2013). The most important migration wave was the Roman colonisation happening in 43 AD that mostly affected the south-east of the British Isles. This has shaped the present day variation in a way that the Romanized zoned can be recuperated from the genomic data. After this period, three main realms were established: Wales, Scotland and England.

The study from Leslie, Winney, Hellenthal et al. (Leslie et al., 2015), using techniques for “painting” the haplotypes to enhance the detection of fine population substructure, found that nowadays the British population can still be clustered into these same groups: Scotland/north England, Central/south England and Wales (see Figure 30 for a map representation).

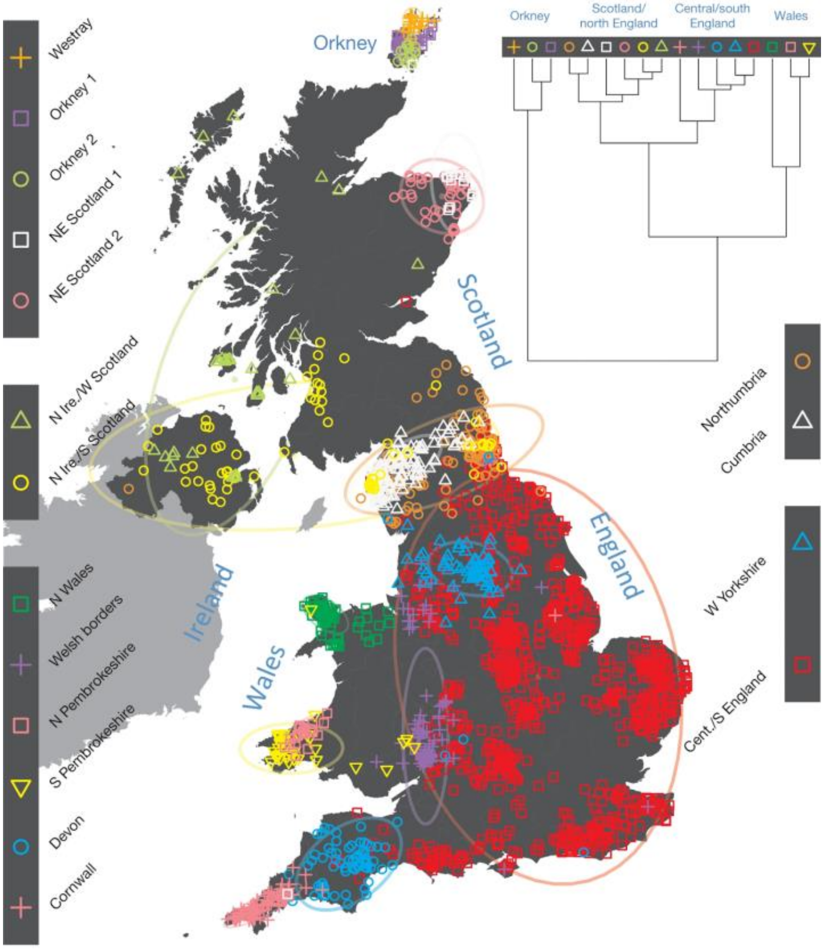


Fig. 30: Map and the hierarchy of the cluster mergings. The map represents the 2,039 samples of the study with colour and shape depending on the cluster they belong to. The tree depicts the order of the hierarchical merging of the clusters. Source: Leslie et al., 2015.

Another effect of geography being a limiting factor in gene flow is the different incidence of genetic diseases across the continent. An example of this is the island of Sardinia, whose inhabitants have frequencies of diabetes type I (Marrosu et al., 2004), multiple sclerosis (Pugliatti et al., 2006) or beta-thalassemia (Cao & Galanello, 2010) that cannot be explained by genotype variation in its causal genes. The case of Sardinia is special given the fact that they constitute a genetic isolate.

Genetic isolates are defined as those populations that have little genetic mixing with other populations. This can lead to enriching some variants and/or deplete others. In these types of populations bottleneck effects, such as wars or pandemics, can have a more profound effect on the genetic pool due to its smaller effective population size (Kääriäinen et al., 2017).

Geography is not the only basis to be genetically isolated in humans. Cultural aspects, such as language or religion, can contribute to the isolation of a population and, given enough time, show the traits associated with genetic isolation.

A clear example of cultural genetic isolation is the case of the Ashkenazi Jews. This population is from medieval Jew populations that were present in Northern France and the Rhineland (banks across the Rhine, Germany) around the 10th century. These medieval populations were founded by migration that started in the Levant (Behar et al., 2004). Given this history, their ancestry is a mix between Levantine and varying degrees of European populations (especially Southern Europe) (J. Xue et al., 2017). This cultural isolation has left a mark in the genetic background of Ashkenazi Jews in the form of particular haplotype groups in the Y chromosome and particular Mendelian diseases incidences, such as the Tay-Sachs disease or Gaucher disease (Ostrer & Skorecki, 2013).

d) Micro-Regional European demographic history: the genetics of rural areas

In a sort of middle ground between isolation due to orography and particular cultural characteristics we find the modern-day rural populations. With the dawn of the Industrial Revolution, the transition from rural to urban communities deeply shaped the demography of different countries and areas within the same country during the last century (Champion, 2012).

In Europe, this transition was due to young adults seeking a better education, work and services, rather than by differences in the fertility rate between rural and urban areas (Brown, 2012). These demographic movements had the effect of depopulating and aging European rural areas (Kulcsár & Curtis, 2012). The genetic effects of such micro-population substructure in European rural areas are still not yet fully understood. It has been shown that rural areas can be more prone to rare diseases due to higher levels of consanguinity (Yali Xue et al., 2017). The latest can be due to a higher isolation of rural populations (i.e. see Nutile et al., 2019) , and/or to the fact that rural populations have shrunk in size during the last generations.

Within this context, the Spanish population appears as a good candidate for analysing the genetic diversity of rural areas. In Spain, by 1900 rural areas still accounted for 68% of the total population (Pinilla & Sáez, 2017) and they have followed the same patterns of depopulation as in other regions of Europe (Silveira et al., 2013). Spanish rural areas tend to have a high number of small municipalities (Vidal & Recano, 1986), which may have experienced isolation for multiple generations (Calderón et al., 2018), and tend to have a higher rate of inbreeding than urban areas independently of the time transect consulted (Fuster & Colantonio, 2003). This situation is the main factor to explain the higher levels of consanguinity found in Spain when compared to other European countries (Mccullough & O'Rourke, 1986).

Multiple sociocultural and socioeconomic factors can also be responsible for this trend (Fuster & Colantonio, 2004). For example, aunt-nephew or uncle-niece marriages or first cousin mating was a practice to maintain and/or expand the family inheritance. However, the main force explaining the higher inbreeding coefficients in Spanish rural areas compared to urban areas is geography. In small and dispersed rural localities the limited amount of suitable local partners conditioned marriage to a point that marrying a distant relative was a likely option; this situation was common even when the population started to increase due to medical advancements but still with restricted mobility (Gamella & Núñez-Negrillo, 2019).

Taking this into context, the Spanish Eastern Pyrenees (SEP) has been suggested as a good representation to understand the particular demographic dynamics of traditional Spanish rural areas (Toledo et al., 2017). The Pyrenees is a mountain chain with a longitude of 430 km West to East oriented that connects the North of the Iberian Peninsula with the rest of Europe. The Pyrenees have a complex orography; mountains can reach more than 3,000 meters high and valleys tend to be narrow and transversal (Martín-Còlliga & Vaquer, 1995). SEP follows a similar pattern as observed in other Spanish rural areas: it reached its maximum recorded population peak around 1860 and has been intensively depopulated since then (Solé et al., 2014) and it is distributed mainly in municipalities of less than 500 inhabitants (in fact only eight of them exceed 2,000 inhabitants).

SEP micro-regions (referred as *comarques*) tend to reflect the medieval counties of Pallars Jussà, Urgell (Alt Urgell), Berga (Berguedà), and Besalú, which included current Ripoll (Ripollès) and Olot (Garrotxa) (Riu-Riu, 1995). Although these counties shared a rural lifestyle each one had different methods of subsistence depending on the geographic location. This caused the Industrial Revolution, and by extension the urban exodus, to affect this regions in different ways. For example, the economy of Berguedà focused on the exploitation of natural resources (in the northern part of the region) and textile (in the southern part), which granted a railway that connected the region to the most industrialized part of Catalonia by 1914 (Serra-Rotés, 2017). This railway was a possible influx of migrants to Berguedà in comparison to the other regions. Despite its close proximity, different recent demographic dynamics in these populations could be expected and it remains unclear to which extent orography could influence its demography.

Studies using classical markers, such as blood markers, proteins and HLA antigens (Calafell & Bertranpetit, 1994) did not detect genetic barriers within the Spanish Pyrenees but a strong West to East gradient that has been explained in terms of ancient demographic events. Others using immunoglobulin data (Giraldo et al., 2001) did not replicate these results but proposed that the observed patterns of diversity are better explained by micro-differentiation. One Y chromosome study detected a subtle degree of substructure in the whole Spanish Pyrenees mountain range (López-Parra et al., 2009).

The most recent study considered autosomal microarray data and a limited number of samples from the Pyrenees (Biagini et al., 2019); it did not identify any genetic difference with other Iberian samples nor detected signals of excess of autozygosity compatible with endogamous practices in the region.

Overall, the discrepancies among these studies suggest that, if the orography of the Pyrenees has shaped the genetic diversity of the rural human populations living within this mountain chain, a much deeper characterization of their genetic variation is required to detect it.

2. OBJECTIVES

The genetic diversity in human populations is not random, but depends on geographic factors reflecting the past demographic history of the species. The limits to detect such population substructure depend on three main factors: the sensitivity of the methods applied to identify population substructure, the type of data that is used and how this data is produced.

The global objective of this thesis is to identify fine population substructure in human populations, using as a case of study WGS from populations from the Catalan Pyrenees, and to better understand the role of batch effects from NGS technologies in the inference of genomic parameters. To achieve this main goal, three different objectives have been conducted:

1. The implementation of a new algorithm that implements geostatistic principles to identify genetic barriers while assuming anisotropic patterns within each geographic group.

2. The population genomics analysis of WGS data at 40X from 30 individuals from the Catalan Pyrenees, covering an area of 140 km.

3. The quantification of the batch effect of the sequencing centre in 1000G in relevant statistics for population genetics such as the loss of function, the amount of archaic introgression, or the proportion of derived alleles.

3 MATERIAL AND METHODS

3.1 Datasets

The outcome of this thesis is based on analysing different genomic datasets, either individually, or after merging them. The main one corresponds to a set of individuals' whole genome sequenced at CNAG from the Southern Eastern Pyrenees (SEP). Publicly available datasets comprised: the Simon Genome Diversity Project (SGDP) (Mallick et al., 2016), the Spanish Exomes (SpExomes) (Dopazo et al., 2016) and the 1,000 Genomes Project phase 3 (1kGp3) (The 1000 Genomes Project Consortium et al., 2015).

a) Spanish Eastern Pyrenees

This dataset is the result of sampling a region of 140 km from the Catalan Pyrenees. The samples come from the *comarques* of Pallars, Alt Urgell, Berga, Ripollès and Garrotxa. All individuals were born in the region they were sampled from, as well as their four grandparents. This dataset also represents the oldest extract of the population, with an average age of 74.14 years and an equal proportion of both sexes. By doing it this way, we are sure that we are capturing variation found in that particular region and we are not introducing a confounding factor in further analyses. A total of six individuals from both sexes were sampled in each region (see Table 2 for details).

Sample name	Region	Age	Sex	Coverage
G178	Garrotxa	86	Male	41,092
G26	Garrotxa	83	Male	38,281
G199	Garrotxa	81	Male	39,56
G23	Garrotxa	87	Female	42,223
G104	Garrotxa	83	Female	40,155
G164	Garrotxa	83	Female	38,175
R86	Ripollès	71	Male	42,485
R97	Ripollès	71	Male	41,885
R47	Ripollès	70	Male	42,16
R88	Ripollès	68	Female	41,512
R87	Ripollès	66	Female	42,306
R17	Ripollès	66	Female	38,256
H67	Berguedà	83	Male	40,679
H76	Berguedà	80	Male	40,844
H109	Berguedà	77	Male	40,891
H77	Berguedà	84	Female	39,465
H96	Berguedà	79	Female	41,743
H16	Berguedà	77	Female	41,432
S141	Alt Urgell	70	Male	39,198
S138	Alt Urgell	64	Male	41,737
S128	Alt Urgell	63	Male	39,319

Sample name	Region	Age	Sex	Coverage
S70	Alt Urgell	69	Female	40,196
S74	Alt Urgell	65	Female	43,123
S12	Alt Urgell	64	Female	39,587
P81	Pallars	81	Male	36,98
P96	Pallars	80	Male	41,821
P10	Pallars	79	Male	38,439
P11	Pallars	77	Female	43,416
P77	Pallars	77	Female	42,563
P53	Pallars	77	Female	42,622

Table 2: Summary of the SEP dataset.

Blood samples were extracted from the individuals and processed to obtain DNA. The samples were sequenced at Centre Nacional de Anàlisis Genòmic of the Centre de Regulació Genòmica (CNAG-CRG) using Illumina® paired ends 150 bp reads in a HiSeq 3000/4000 (Illumina®). The mean coverage per sample was ~40x (see Table 2 for details of the coverage).

The read files were then used for SNP calling, which was performed using the Genome Analysis Toolkit (GATK) HaplotypeCaller v3.6 (McKenna et al., 2010) using the defaults found on the GATK Handbook v3.6 (Depristo et al., 2011). The reference genome used was hs37d5 (Hg19) and all samples were called jointly to obtain a single VCF file.

After this step we decided to continue with data cleaning. In this case we first filtered the VCF to filter out those variants that were not labeled as SNP. After this we filtered out those SNPs that had more than 1 alternative allele. We set a strict threshold of 0% missingness, filtering out those SNPs that did not meet these criteria. As a final step we filtered out those SNPs that do not follow Hardy-Weinberg equilibrium. To have a proper dataset to use in a population genomics framework, we decided to keep only those SNPs that are polymorphic in the SEP populations sampled.

As a last step in data cleaning we decided to run the algorithm KING (Manichaikul et al., 2010) to ascertain if our samples had any family relation of second degree or closer. The results from KING found that individuals S138 and S12, so we proceeded to remove individual S138 from the analysis.

All data cleaning steps were done with a combination of *bcftools* (Heng Li, 2011) and PLINK (Purcell et al., 2007).

The final dataset from SEP has a total of 29 individuals and 9,309,056 biallelic polymorphic SNPs.

To predict loss of function variants (LoF) in SEP first we annotated the variants we found in this work. In order to do so we downloaded the already annotated table of variants from dbNSFP (Liu et al., 2013), a database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome. This table contains prediction scores for several algorithms (37 different algorithms in the last version).

To merge the contents of the table with our VCF file with the variants found in SEP we used the program called SnpSift (Cingolani, Patel, et al., 2012) and to add the functional prediction of said variants we used SnpEff (Cingolani, Platts, et al., 2012).

From this annotated and functionally predicted VCF file we decided to use 3 different algorithms to consider that a variant is a LoF variant: Polyphen2 (Adzhubei et al., 2010), MutationAssessor (Reva et al., 2011) and SIFT (Cingolani, Patel, et al., 2012). If these three algorithms marked a variant as LoF, we accepted that variant into the study. We restricted the analysis to those variants that do not appear in homozygosis because they would reflect redundant or advantageous effects of dispensable genes (Rausell et al., 2020). To account for this bias, the frequency of LoF variants of each individual was estimated using all the variants in the exome that did not report homozygote individuals for the derived allele in the dataset, and that were heterozygote for the individual.

b) Simon Genome Diversity Project (SGDP)

The Simon Genome Diversity Project is an effort to obtain a more detailed picture of human genetic variation. The selection of samples in this dataset pretends to represent most of the genetic, cultural and linguistic human variation. The dataset comprises a total of 278 samples sequenced using standard Illumina to an average coverage of 43x (Mallick et al., 2016). These samples can be assigned to circa 140 populations distributed in seven continental supergroups. In our analyses, we ascertained the samples from the SGDP labeled as "*WestEurasia*" to have a more accurate depiction of the SEP situation inside Europe. Genomes were aligned to hs37d5 (Hg19), and SNP calling was done in a similar way as SEP.

After merging SGDP and SEP and applying data cleaning procedures (the same as SEP minus KING) we have a total of 104 individuals and 5,388,964 SNPs in this dataset.

A problem of this dataset is the number of individuals: one or two per population. This factor makes it less reliable to use in LoF studies as it could easily bias the analysis due to its low sample size.

c) Spanish Exomes (SpExomes)

The SpExomes dataset is a collection of Spanish individuals that were Whole Exome Sequenced (WES) in order to produce controls for the Medical Genome Project. Dopazo et al. sampled 267 individuals from the north (Galicia and Catalonia), centre (Madrid) and south (Andalucia) of Spain. Sequencing was done using SOLiD 5500xl. The dataset was stored in the European Genome-phenome Archive (EGA) under accession number EGAS00001000938 (both VCF and FASTQ files).

Although the SpExomes dataset offers ready-to-use VCF files, when examining these files we realized that they were not processed according to the standards we used in SEP. This produced biases in some of the population statistics that we were using when comparing both datasets. To solve this problem, we downloaded the FASTQ files and conducted mapping and SNP-calling using the same Bioinformatic pipeline as in SEP. We did the same data cleaning (with the exception of KING) as in SEP.

The final dataset comprised a total of 288 individuals and 239,349 SNPs.

d) The 1,000 Genomes Project phase 3 (1kGp3)

This dataset corresponds to the first effort to sequence a large group of individuals from different populations and continents (The 1000 Genomes Project Consortium et al., 2015)This project had the main goal to describe most of the genetic variants with frequencies of at least 1% in the populations studied. It started in 2008 and ended in 2015, publishing the results in three different phases. It finally comprised 2,504 individuals from 26 different populations across five different continental regions.

1kGp3 used a combination of low and high coverage whole genome sequencing, whole exome sequencing, high density microarray genotyping and Complete Genomics (a sequencing company with its own proprietary sequencing and analysis techniques) to generate the sequence data.

This huge effort was done in nine different centres: the Broad Institute, the Baylor College of Medicine (Human Genome Sequencing Center), BGI, Max Planck Institute for Molecular Genetics, Washington University, Wellcome Trust Sanger Institute, Illumina, Affymetrix and Complete Genomics. Each centre did different protocols to produce the sequence data in the form of FASTQ files. These files were then sent to the Sanger Institute to be mapped using *bwa* (H. Li & Durbin, 2009) and proceed through a series of steps for data cleaning. After the mapping the consortia proceeded to do variant calling on the samples. They used a total of ten different SNP-calling tools (not taking into account micro-satellites and structural variants discovery tools), with a final step to create the integrated call set. This integrated call set contained all the variation found in the samples that passed a series of filters.

This strategy was applied to minimize the effect of putative batch effects due to the used Bioinformatic pipeline. However, recent studies (Anderson-Trocmé et al., 2020) point to the fact that these data are not exempt of sequencing errors and other biases. This is particularly relevant because, due to the geographic coverage and the amount of individuals, data from the 1kGp3 are widely used to SNP imputation in Genome Wide Analysis studies (GWAS) or haplotype phasing (Delaneau et al., 2013). Furthermore, this dataset has been studied in depth from an evolutionary point of view (S. R. Browning et al., 2018; Colonna et al., 2014; Huang & Siepel, 2019) and has been used in multiple medical studies (Papadimitriou et al., 2019).

3.2 Detection of population substructure

The identification of (hidden) population substructure in a set of individuals from the same species has been one of the most active fields in population genomics during the last twenty years. Consequently, a large number of algorithms have been proposed for identifying population substructure. Some of them, such as Principal Component Analysis (PCA) or Principal Coordinate Analysis (also called classical multidimensional scaling), are directly inherited from the field of statistical learning (Hastie et al., 2001) for dimensionality reduction. Therefore, they are not population genomic specific.

However, the way how the genomic data is produced allows its use and -in some cases- interpret them from a demographic point of view (Patterson et al., 2006). These methods apply classical matrix decomposition (i.e. the output are two matrices of eigenvectors and eigenvalues) to map the individuals into a set of new orthogonal variables, each explaining a proportion of the variance present in the original data. The main difference is the starting matrix. In PCA, a covariance matrix between SNPs and individuals is generated. In PCoA, a distance matrix between pairs of individuals is estimated and used for the algebraic decomposition.

The ultimate goal is to represent the genetic relationships of the individuals into a lower (usually the first two dimensions) dimensional space, so it is possible to identify sub-groups of individuals more genetically related than others. The latest can be achieved using algorithms such as *mclust* (Scrucca et al., 2016), that assume that the data can be modelled as a mixture of Gaussian distributions. The position of one individual relative to the other individuals can be interpreted in terms of admixture and time since the subpopulations diverged (see Figure 31).

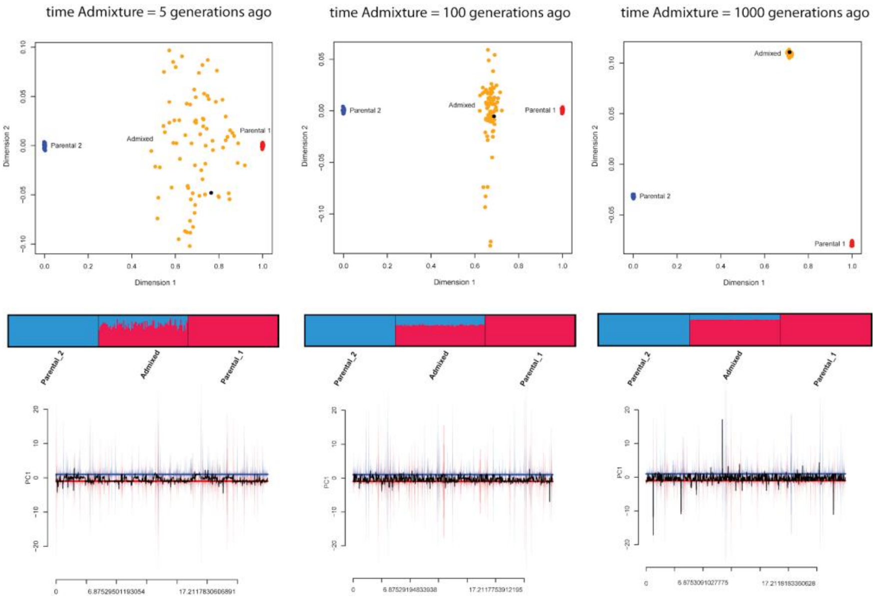


Figure 31. The output from different algorithms for identifying population substructure using simulated data. Two populations (blue and red) diverged for a long time. They create a new population (orange) at time of Admixture. Depending on when it occurred, the position of the individuals in the plot changes. If the admixture occurred a long time ago, the admixed population appears as something completely different from the parental populations. Upper panels report the MDS analysis. Lower panels, an ADMIXTURE (Alexander et al., 2009) analysis. Source: Lao & van Oven, 2015

The application of methods for reducing the dimensionality of the data has shown that genetic maps tend to resemble the physical sampling location of each individual (Lao et al., 2008; Novembre et al., 2008; C. Wang et al., 2010). However, the demographic interpretation of these maps is complex, since multiple demographic events can produce similar plots (Novembre & Stephens, 2008).

The state of art of this approach in population genomics is the fineSTRUCTURE algorithm (Lawson et al., 2012). This algorithm defines a matrix of relationships between individuals taking into account the haplotypic information. In this framework, a donor chromosome is “painted” by the chromosome of other individuals, and the total amount of shared chunks between pairs of individuals is reported. From this matrix of similarity, one can apply different approaches for establishing the genetic relationship between individuals (Lawson et al., 2018).

A second type of algorithms, also unsupervised, models the ancestry of an individual as a mixture of ancestry proportions from different K ancestral populations, all of them equally related, assuming that within each ancestral population each SNP is in HWE. The parameters to estimate are the percentage of admixture of each individual and the frequency in each ancestral population so the likelihood of the model is maximized.

The different algorithms depend on how this optimization problem is solved or the statistical flavour that is applied. For example, STRUCTURE (Pritchard et al., 2000) and fastSTRUCTURE (Raj et al., 2014) take a Bayesian approach to estimate the posterior distributions of the ancestry individual proportions. FRAPPE (Tang et al., 2005) and ADMIXTURE (Alexander et al., 2009) maximize the log likelihood function of the ancestry proportions of each individual given the allele frequencies of each ancestral population. sNMF (Frichot et al., 2014) uses an algebra based approach conditioning the matrix output to estimate the ancestry proportions.

A third type of algorithms include geodesic (spatial) information in the model, either implicitly (unsupervised algorithms such as SPA, Yang et al., 2012) in order to predict the geographic location of an individual, explicitly and implicitly, in which case we use the geographic information for identifying admixture patterns, or explicitly by identifying the migration patterns (i.e. EEMS, Petkova et al., 2015) or genetic barriers (i.e. SAMOVA, Dupanloup et al., 2002, or BARRIER, Manni et al., 2004) in the space. The difference between them is the assumptions that are considered when spatially modelling the genetic variation present in the samples. For example, SAMOVA assumes that the individuals from one geographic group are genetically homogeneous and genetically distinct from other groups (as defined by the AMOVA algorithm, Excoffier et al., 1992).

The objective is to find the geographic groups and assign each population/individual to its geographic group. Such optimization is conducted by means of a natural computing algorithm (Brabazon et al., 2015) called simulated annealing (Kirkpatrick et al., 1983). However, it can be objected that the assumption of geographic homogeneity within each geographic group is unrealistic for many species and, particularly, in the case of humans. As previously described, human populations tend to show isolation by distance patterns.

Therefore, any genetic barrier should take into account that within the geographic groups genetic variation is not going to be distributed homogeneously, but following the patterns of isolation by distance and anisotropy.

The latest is a classical concept in the field of geostatistics, and refers to the Tobler's law (Tobler, 1970): "*everything is related to everything else, but near things are more related than distant things*". Following the same principle, the genetic differentiation between two individuals with regards to the physical distance just due to genetic isolation can be modelled as in Figure 32.

These principles have been previously used for spatially modelling the genetic diversity in human populations (Bradburd et al., 2016) but never used in the context of identifying genetic barriers using genetic matrices between individuals.

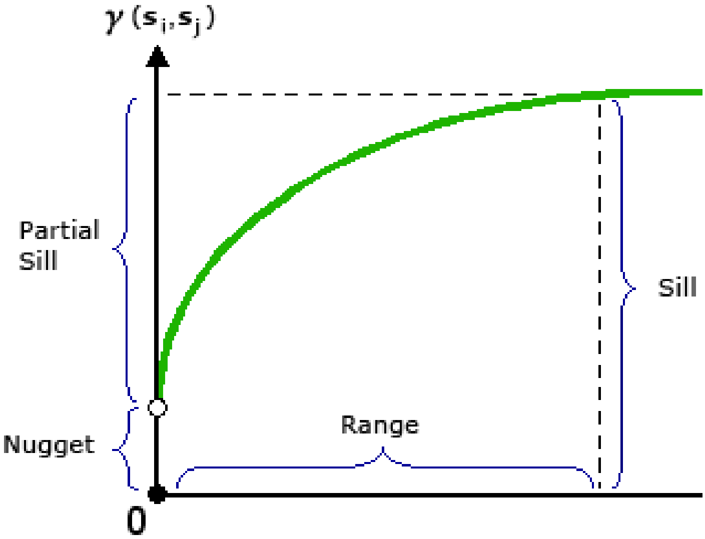


Figure 32. A graphical description of how genetic divergence between two individuals (Y axis) variate with geographic distance (X axis) using a kriging model. The nugget corresponds to the minimum amount of genetic differentiation you can find in a given set of samples that belong to the same population (i.e. the geographic distance between them is 0). Sill is the difference between the maximum amount of genetic distance that can be observed between two samples. The range refers to the geographic distance needed in order for the variable to reach a stable point.

If the genetic differentiation follows a direction in space (a genetic gradient in the data points) we can extend the model to include an angle α of maximum genetic divergence in space.

Given this framework we define genetic distance as follows:

$$(1) \quad GenDist = s * k + n$$

Where s corresponds to the sill, n to the nugget and k to a factor of proportionality defined as:

$$(2)\quad k = \frac{\sqrt{b^2 * (x * \cos(\alpha) + y * \sin(\alpha))^2 + \ a^2 * (x * \sin(\alpha) - y * \cos(\alpha))^2}}{a^2 * b^2}$$

Where α corresponds to the angle that gives the maximum genetic divergence, a and b to the foci on an ellipse and x and y to the position of the sample in the space.

In order to estimate the parameters (a, b and α), we can use multiple regression on distance matrices (Legendre & Legendre, 2012; Lichstein, 2007). In particular we can use nonlinear regression fitting methods, in this case the Nelder and Mead Simplex method implemented at Flanagan's JAVA package (Thomas-Flanagan, 2016).

For a given *K* geographic groups of genetically related points, the identification of the genetic barriers consist on identifying the set of geographically related points that minimizes the goodness of fit of the estimated parameters as the mean sum of square error of each point between the observed genetic distance and the inferred *D'* from the fitted parameters:

$$(3)\ SSE = SquareError_g =$$

$$\sum_{g}^{K} \left\{ \begin{array}{l} \displaystyle\sum_{i=1}^{n_g}\sum_{j=1}^{n_g} \frac{\big(D(i,j) - D'(i,j)\big)^2}{m-1} \quad if\ \frac{n_g(n_g-1)}{2} \geq 3 \\ \displaystyle\sum_{i=1}^{n_g}\sum_{j=1}^{n_g} \frac{(D(i,j) - \overline{D})^2}{m-1} \quad if\ \frac{n_g(n_g-1)}{2} < 3 \end{array} \right\}$$

To optimize the SSE function and to identify the genetic barriers, we define each geographic group by a [x,y] pair of coordinates, and assign each observed point based on its proximity to each geographic group, such as in K-means algorithm. The problem is then to find the geographic coordinates of each group that minimize the SSE. In order to explore the space of possible solutions, we use a genetic algorithm, which has been already applied in optimization problems involving geographic divisions (Sergeeva et al., 2017).

3.3 Methods for quantifying the levels of autozygosity

RoHs are conceptually very easy to define: long genomic segments that for a given individual all the SNPs are homozygotes suggesting recent inbreeding (see Recombination section in the Introduction). However, from an algorithmic point of view, RoHs are challenging, and different algorithms have been proposed for identifying them, each showing a different performance (Howrigan et al., 2011).

The first problem that RoHs face is the definition of SNV. Since the definition of SNPs is ultimately population specific (i.e. a SNP in one population can be fixed in another and therefore not be a SNP), in order to identify recently inbred individuals we need to remove SNPs that are fixed in our population or, more commonly, below a certain minimum allele frequency (MAF) threshold. In addition, RoHs algorithms must be robust against the presence of sequencing errors, which usually introduce heterozygote genotypes thus breaking the RoHs.

In practice, most of the algorithms allow for a certain frequency of heterozygote SNPs before stopping a RoH. Pemberton et al (Pemberton et al., 2012) suggested minimizing these effects by estimating RoHs in terms of the likelihood of observing a particular genotype combination in a genomic region given the allele frequencies of the SNPs in the population from which the individual was sampled. In the present thesis we have used this approach to compute RoHs over all the genome of the SEP samples.

A second problem is that RoHs are dependent on the SNP density in the genome. To estimate the levels of autozygosity in the exomic dataset we used the heterozygosity ratio (HetR) statistic, initially proposed by Guo et al. (Samuels et al., 2016). This statistic is independent of the density of SNVs genotyped. This measure is defined as the ratio of SNPs for which the individual is heterozygote with respect to the number of SNPs for which the individual is homozygote for the non-reference allele. However, we noticed that the HetR depends on the sample size.

Under the assumption of a constant population size, the expected number of SNPs following a particular site frequency i (1 < i < n-1) out of a sample of n chromosomes is determined by (Wakely, 2016):

$$(1) \quad E[\xi_i] = \frac{2Ne\mu}{i} = \frac{\theta}{i}$$

The probability of being heterozygote for a particular SNP with i derived alleles is:

$$(2) \quad P(Het) = \frac{\binom{i}{1}\binom{n-i}{1}}{\binom{n}{2}} = \frac{2i(n-i)}{n(n-1)}$$

The probability of being homozygote derived out of *i* derived alleles is:

$$(3) \quad P(Hom) = \frac{\binom{i}{2}\binom{n-i}{0}}{\binom{n}{2}} = \frac{i(n-i)}{n(n-1)}$$

Using (1), (2) and (3), the expected HetR under the assumption of no inbreeding is defined as:

$$HetR = \frac{\sum_{i=1}^{n-1} \frac{\theta}{i}\frac{2i(n-i)}{n(n-1)}}{\sum_{i=1}^{n-1} \frac{\theta}{i}\frac{i(n-i)}{n(n-1)}} = \frac{2\sum_{i=1}^{n-1}(n-1)}{\sum_{i=1}^{n-1}(i-1)} = \frac{2\left(\sum_{i=1}^{n-1} n - \sum_{i=1}^{n-1} i\right)}{\sum_{i=1}^{n-1} i - \sum_{i=1}^{n-1} 1}$$

$$= \frac{2\left((n-1)n - \left(\frac{n(n+1)}{2} - n\right)\right)}{\frac{n(n+1)}{2} - n - 1 - (n-2)}$$

$$HetR = \frac{2}{n}$$

Therefore, the HetR is not independent of the sample size.

As we have an extremely different sample size between the two sets of populations (maximum of 6 individuals in SEP populations, 259 individuals in the SpExomes) we devised a method to normalize the score. This normalization follows the next steps: we pick a random set of five samples from every population (each subgroup from SEP and SpExomes) and calculate the heterozygosity ratio for each selected sample. We repeat this sampling a total of 5,000 times. The normalized score is the result of dividing the sum of the heterozygosity ratio between all runs and the total number of runs in which that sample has appeared.

3.4 Method for analysing batch effect

In order to quantify the possible batch effect in the 1,000 Genomes Project dataset we downloaded the ready-to-go VCF files from the 1,000 Genomes FTP site. As these files were downloaded on a per chromosome basis we used *bcftools* (Heng Li, 2011) to concatenate them into a single VCF file. Then we proceeded to annotate the single VCF file using the same procedures as with the SEP samples. Information about the sequencing center of every sample was extracted from the supplementary data from the 1,000 Genomes Project 2015 paper (The 1000 Genomes Project Consortium et al., 2015).

To ascertain this possible batch effect we used three different measures: amount of LoF variants per sample, amount of alleles considered to be introgressed by the Sprime algorithm (S. R. Browning et al., 2018), and the amount of derived allele singletons per sample.

Number of LoF variants was calculated using the same procedure as with the SEP samples.

To count the number of introgressed alleles we downloaded the output from the Sprime algorithm used in Browning et al (S. R. Browning et al., 2018) from S. Browning, 2018. Output is divided on a per population and chromosome basis. These files provide the introgressed allele for the SNP according to the algorithm. We used this information to count the number of alleles of each individual of each population (using the proper file to do so). We used the following scoring: if the individual is homozygous for the introgressed allele we add 2 to the score, if the individual is heterozygous we add 1 to the score, if the individual is homozygous for the non-introgressed allele we add 0 to the score. Finally we sum up the score for the individual across all the chromosomes.

To ascertain the number of derived allele singletons present in each sample we used the ancestral allele already present in the VCF file (AA flag in the INFO field), which uses the 6-way EPO alignments available in Ensembl v71 (Flicek et al., 2014) (according to the Supplementary Information of the 1,000 Genome Project). From the general VCF file we extracted those SNPs that were considered singletons using *bcftools* (Heng Li, 2011), selecting those whose allele count was equal to one (AC flag in the INFO field). From these singletons VCF file we checked if the reference allele is the same as the ancestral allele, and assign the derived singleton to the proper sample.

In order to check for the possible effect of the sequencing center over these variables we used the R package *lmer4* (Bates et al., 2015) to construct hierarchical mixed models controlling for the random effects of the continent and population of assignation of each individual. The models were as follows:

$$SeqCenter\ Model\ =\ lmer(log(S) \sim SeqCenter\ +\ (1|Continent/Pop))$$

Where S is the variable of interest. The mixed null model followed the next formula

$$Null\ Model\ =\ lmer(log(S) \sim\ (1|Continent/Pop))$$

To compare both models we used the ANOVA framework implemented in R (*R: The R Project for Statistical Computing*).

4 RESULTS

4.1 Chapter 1

Fine-scale population structure in five rural populations from the Spanish Eastern Pyrenees using high-coverage whole-genome sequence data.

Iago Maceda, Miguel Álvarez-Álvarez, Georgios Athanasiadis, Raúl Tonda, Jordi Camps, Sergi Beltran, Agustí Camps, Jordi Fàbrega, Josefina Felisart, Joan Grané, José Luis Remón, Jordi Serra, Pedro Moral, and Oscar Lao

(Submitted to *European Journal of Human Genetics*)

Fine-scale population structure in five rural populations from the Spanish Eastern Pyrenees using high-coverage whole-genome sequence data

Running title: Genomic structure in the Spanish Eastern Pyrenees

Iago Maceda*[(1)], Miguel Álvarez-Álvarez *[(2)], Georgios Athanasiadis[(3)], Raúl Tonda[(4)], Jordi Camps[(4)], Sergi Beltran[(4)], Agustí Camps[(5)], Jordi Fàbrega[(6)], Josefina Felisart[(7)], Joan Grané[(8)], José Luis Remón[(9)], Jordi Serra[(10)], Pedro Moral*[(2)], and Oscar Lao*+[(1)]

[(1)]Population Genomics, CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, C/ Baldiri i Reixach 4, 08028 Barcelona, Spain.

[(2)]Department of Evolutionary Biology, Ecology and Environmental Sciences, Biodiversity Research Institute, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain.

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

[(4)]Bioinformatics Unit, CNAG-CRG, Centre for Genomic Regulation, C/ Baldiri i Reixach 4, 08028 Barcelona, Spain.

[(5)]Hospital Sant Bernabé, Ctra. de Ribes 47, 08600 Berga, Barcelona, Spain.

[(6)]Fundació Sant Hospital La Seu d'Urgell, Pg. Joan Brudieu 8, La Seu d'Urgell, Lleida, Spain.

[(7)]Hospital d'Olot i Comarcal de la Garrotxa, Avda. Països Catalans 86, 17800 Olot, Girona, Spain.

[8]Hospital de Campdevànol, Ctra. de Gombrèn 20, 17530 Campdevànol, Girona, Spain.

[9]Servei d'atenció primaria Lleida Nord, Gerència Territorial Alt Pirineu i Aran, Institut Català de la Salut, C/ Sant Jordi 13, 25620 Tremp, Lleida, Spain.

[10]Laboratori d'Anàlisis Clíniques, Hospital Comarcal del Pallars, C/ Pau Casals 5, 25500 Tremp, Lleida, Spain.

* These authors contributed equally to this work.

+ Corresponding author (oscar.lao@cnag.crg.eu).

Funding

O.L, I.M, RT, JC and SB acknowledge the support from Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa, CERCA Programme / Generalitat de Catalunya, Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement, Co-financing with funds from the European Regional Development Fund by the Spanish Ministry of Science and Innovation corresponding to the Programa Operativo FEDER Plurirregional de España (POPE) 2014-2020 and by the Secretaria d'Universitats i Recerca, Departament d'Empresa i Coneixement of the Generalitat de Catalunya corresponding to the Programa Operatiu FEDER de Catalunya 2014-2020. O.L. gratefully acknowledges the financial support from Ministerio de Economía y Competitividad (Ministry of Economy and Competitiveness) - RYC-2013-14797, BFU2015-68759-P and

PGC2018-098574-B-I00 and Generalitat de Catalunya (Government of Catalonia) - GRC 2017 SGR 937. I.M. gratefully acknowledges the financial support from the Government of Catalonia | Agència de Gestió d'Ajuts Universitaris i de Recerca (Agency for Management of University and Research Grants) - GRC 2014 SGR 615. M.M.A. was supported by a FPU15/01251 from Ministerio de Ciencia, innovación y Universidades. The samples analyzed in this study are part of a much larger sample whose collection was supported by the project CGL2011-27866 of the Ministerio de Ciencia y Tecnologia, the Fundació Moret i Marguí, and the invaluable collaboration of the Hospital San Bernabé (Berguedà), Fundació Sant Hospital (Alt Urgell), Hospital d'Olot i Comarcal (Garrotxa), Hospital de Campdevànol (Ripollès), and Hospital Comarcal (Pallars).

Abstract

The Spanish Pyrenees is particularly interesting for studying the demographic dynamics of European rural areas given its orography, the main traditional rural condition of its population and the reported higher patterns of consanguinity of the region. Previous genetic studies suggest a genetic continuity of the area in the West to East axis. However, it has been shown that micro-population substructure, compatible with ancient and recent demographic events, can be detected when considering high quality NGS data and using methods specially designed for identifying fine population substructure. In this work we have analyzed the genome of 30 individuals sequenced at 40× from five different valleys in the Spanish Eastern Pyrenees (SEP) covering 140 km. Using haplotype-based tools and spatial analyses we have been able to detect micro-population substructure within SEP not seen in previous studies. Linkage disequilibrium and autozygosity analyses suggest that the SEP populations show diverse demographic histories. In agreement with these results, demographic modelling by means of ABC-DL identify heterogeneity in their effective population sizes despite of their close geographic proximity, and suggest that the population substructure within SEP could have appeared 100 generations ago. Finally, we observed heterogeneity among the populations for the loss of function burden. Overall, these results suggest that each rural population of the Pyrenees could represent a unique entity.

Keywords: fine population substructure, rural population, Spanish Eastern Pyrenees, whole genome sequencing

Introduction

The transition from a rural to an urban world has been mainly triggered by the industrial revolution that started in Europe, which promoted large-scale, differentiated and coordinated activities that were better accomplished by urban communities (1), which caused population movements from rural areas to urban cities (2). From a practical point of view, the division in rural and urban areas has implications for health (3), as well as for generating genetic isolates explaining the predisposition to rare diseases (4).

In Spain, by 1900 rural areas comprised 68% of the total Spanish population (5) and, similarly to other rural European regions, they have been intensively depopulated with massive migratory movements towards industrialized urban areas (6). Spanish rural areas are traditionally characterized by a low demographic density and a high number of small municipalities that may have experienced isolation for generations (7). This situation has been suggested as a main factor for explaining the higher levels of consanguinity of Spain compared to other European countries (8). From a temporal point of view, the levels of consanguinity in urban and particularly rural Spanish areas reached its maximum between the end of the 19th century and 1929 (8).

The main force explaining the higher inbreeding coefficients in Spanish rural areas compared to urban areas is geography (9). In particular, islands and high mountains, as well as altitude within a valley, have been reported as the most effective geographic barriers increasing the levels of inbreeding in Spain (9). In this context, the rural population of the Spanish Eastern Pyrenees (SEP) has been suggested as a particularly interesting system for understanding the demographic dynamics of traditional Spanish rural areas (10). The Pyrenees is a mountain chain of a complex orography with a longitude of 430 km West to East oriented that connects the North of the Iberian Peninsula with the rest of Europe (11). From a demographic point of view, the area reached its maximum-recorded population peak in 1860 and it has been intensively depopulated since then (12).

Studies using classical markers, such as blood markers, proteins and HLA antigens (13) did not detect genetic barriers within the Spanish Pyrenees but a strong West to East gradient that has been explained in terms of ancient demographic events. Others using immunoglobulin data (14) did not replicate these results but proposed that the observed patterns of diversity are better explained by micro-differentiation. One Y chromosome study detected a subtle degree of substructure in the whole Spanish Pyrenees mountain range (15). The most recent study considered autosomal microarray data and it did not find any genetic difference with other Iberian samples nor detected signals of excess of autozygosity compatible with endogamous practices in the region (16). Overall, the discrepancies among these studies suggest that, if the orography of the Pyrenees has shaped the genetic diversity of the rural human populations living within this mountain chain, a much deeper characterization of their genetic variation is required to detect it.

In the present study we have characterized the genetic variation of the SEP rural population from five regions (Pallars (P), Alt Urgell (U), Berga (B), Ripolles (R) and Garrotxa (G) (see Figure 1B) covering around 140 km, making use, for the first time, of high-coverage whole genome sequencing (WGS) data. This allowed us 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 modeling.

Material and methods

Datasets description

In total, five regions corresponding to the political separations established by the government from SEP were sampled (Garrotxa, Ripollés, Berguedà, Alt Urgell and Pallars), covering circa 140 km, with six samples per region. All samples were born and had all their grandparents born in the same sampled region. This dataset represents the oldest extract of the population, with an average age of ~76 years and equal proportions of both sexes (see Supplementary Table 1). Therefore, this sample is effectively unaffected by the demographic changes occurred during the 20th century. All subjects signed an informed consent and the study had the approval of the Ethics Committee of the University of Barcelona. Each individual was whole genome sequenced (WGS) using standard Illumina (San Diego, California, USA) paired-ends sequencing technology with a read-length of 150 bp with an average sequencing coverage of ~40×. SNP-calling used GATK

HaplotypeCaller v3.6 (17), using the default settings according to the GATK Handbook v3.6 (18), with hs37d5 as the reference assembly, using all samples jointly. For an overview on the data cleaning steps, see Supplementary Information. The final dataset contained 29 individuals and 9,309,056 SNPs.

To make comparisons with other European populations, we used samples geographically classified as “West Eurasian” from the Simons Genome Diversity Project (SGDP) (19) (Supplementary Table 1). After the merging with the SEP dataset, we applied the same quality control as conducted with SEP. The SEP-SGDP dataset contains a total of 104 individuals and 5,388,964 SNPs.

Spanish exome data from (20) was accessed from EGA (accession number: EGAS00001000938). Data were downloaded as FASTQ files and mapped. SNPs were called following the same procedure used in the SEP samples. We used the same data cleaning procedures as conducted with SEP. The resulting dataset has a total of 288 individuals and 239,349 SNPs. To make a comparison between our dataset and the one presented in (16), we ascertained the shared SNPs between the two datasets, namely ~231K SNPs.

Analyses

Detection of population substructure of SEP at a macro and microgeographic scale

A classical multidimensional scaling (MDS) analysis on an identical by state (IBS) matrix between pairs of individuals ('1-ibs' function of PLINK (21)) was carried out to summarize the genomic relationships within SEP as well as with SGDP samples. SEP-SGDP data were phased with SHAPEIT2 (22) using all defaults and a publicly available genetic map based on the 1000 Genomes Project phase 3 database (23). Phased data were used in Chromopainter/fineSTRUCTURE (24) to identify fine population substructure. The haplotype lengths matrices from Chromopainter/fineSTRUCTURE were analyzed with GLOBETROTTER (25). Finally, we repeated the haplotype-based analyses with SEP individuals in order to detect fine-scale population substructure in the area.

Identification of genetic barriers and differential migration rates

In order to identify genetic barriers, we developed an algorithm that models the shared co-ancestry matrix from Chromopainter/fineSTRUCTURE in terms of anisotropy within each geographic group (see Supplementary Information). In parallel, we used the Estimated Effective Migration Surfaces (EEMS) algorithm to have an estimate of migration rates between the individuals of SEP dataset, (26). The algorithm was run using the default parameters and a total of 1,000 demes to conform the surface on which to situate the individuals.

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

To estimate the effective population size and the time of split of the SEP populations, we modelled its demographic history using a simple demographic model (see Supplementary Figure 1). We used an ABC approach coupled to deep learning (ABC-DL) (27) to estimate the posterior distribution of the different parameters of the model (see Table 1 for the prior distributions and Supplementary Information for details).

Quantification of Linkage Disequilibrium and levels of autozygosity in SEP and Spanish-exomes samples

The decay of linkage disequilibrium (LD) was estimated for each population with the HR statistic (28). In order to minimize the effects of frequency-dependence in LD measures (29), LD was computed by averaging the HR between pairs of SNVs showing a similar MAFs (|MAFSNVa - MAFSNVb| < 0.05). Runs of homozygosity (RoH) were quantified by means of two different approaches. For WGS data from SEP, we used the RoH as defined in (30). For the exomic data, we used the heterozygosity ratio (HetR) (31). We sampled sets of five samples from every population and calculated the HetR for each selected sample. We repeated this sampling a total of 5,000 times. A normalized estimate (nHetR) was obtained averaging all the replicates for each sample.

Prediction of loss of function (LoF) genetic variants in SEP and Spanish-exomes samples

We used SNPsift (32) to annotate the genomic variants using the table from dbNSFP (33) and SNPEff (34) to add the functional prediction. To classify an SNV as damaging (LoF), we required it to be predicted as such by three different algorithms: PolyPhen2 (35), MutationAssessor (36) and SIFT (32). We restricted our analyses towards LoF variants that do not appear in homozygosis, as these most likely reflect redundant and/or advantageous effects of dispensable human genes (37). To account for this bias, the frequency of LoF SNVs of each individual was estimated using all the variants in the exome that did not report homozygote individuals for the derived allele in the SEP and Spanish-exomes databases, and that were heterozygote for the individual.

Results

Genetic variation of SEP in the European context

In order to describe the genetic relationships of the SEP samples with the European continent, we first performed a classical multidimensional scaling (MDS) analysis. SEP populations cluster with samples from the Iberian Peninsula (Supplementary Figure 2), following the geographic dependence of the genetic diversity observed for whole Europe in other studies (38). Complementary to these analyses, we ran a Chromopainter/fineSTRUCTURE analysis with the SEP-SGDP dataset. In agreement with the previous result, the phylogenetic tree (Supplementary Figure 3) shows all individuals from SEP sharing a private cluster with the Basque samples from the French Western Pyrenees. The GLOBETROTTER analysis (Supplementary Figure 4) did not identify a differential genomic contribution to SEP from any particular European population from the SGDP dataset, thus suggesting that historical migrations did not influenced the population substructure present in the SEP.

Fine-scale population structure in SEP

We wondered if such structure would extend at a micro-geographic level in SEP. We repeated the MDS and Chromopainter/fineSTRUCTURE analyses considering the SEP samples alone. As shown in Figure 1A, the first two dimensions mimic the geographic sampling location (correlation in a symmetric Procrustes rotation = 0.35 (p-value = 0.033 after 99,999 permutations), mainly distributing the SEP samples in the longitudinal axis in the second dimension. Chromopainter/fineSTRUCTURE also identified two main clusters that split the west to east axis (Figure 1B). We wondered if this result could reflect different patterns of spatial anisotropy in our data. We applied an algorithm that describes the genetic relationship between individuals in terms of genetic barriers and different patterns of anisotropy between groups (see Supplementary Information). Our result for two groups (Figure 2A) detected a geographic barrier between Garrotxa-Ripollés and Pallars-Alt Urgell-Berguedà. In order to confirm this identified genetic barrier, we run EEMS in the SEP dataset. As shown in Figure 2B, EEMS identified a migration depletion compared to the rest of the geographic area between the same set of regions previously detected by Chromopainter/fineSTRUCTURE analysis and the detection of genetic barriers. This previously undetected population substructure compared to microarray data (16) could reflect differences in WGS vs array-based data and/or the applied methodology. Using the SNPs that are common with those used in (16), Chromopainter/fineSTRUCTURE failed to detect geographic clusters (see Supplementary Figure 5). Overall, these results suggest the presence of micro-population substructure in SEP that requires WGS to be detected.

Autozygosity, inbreeding and LD in SEP compared to Spanish-exomes

We analyzed the patterns of LD by means of HR, using the genetic variation present in the whole genome of the SEP populations. We observed differences between the SEP populations in the decay of LD heterogeneity (Figure 3A). In particular, the Alt Urgell region showed more LD than others. We wondered whether the observed pattern was specific of SEP, or if it was particular to the Spanish population. We repeated the LD analyses on the exome using the Spanish-exomes dataset. First, we observed that the source of genetic variation (WGS vs exome) did not influence the decay of the HR score in SEP populations (Supplementary Figure 6). When comparing SEP and Spanish-exomes, we observed that all regions have higher HR scores than Spanish-exomes and, again, that Alt Urgell is the one that has the highest HR score among all the populations sampled (Figure 3B).

We wondered whether these results would be in agreement with the reduction of genetic diversity due to a traditionally low demographic density and endogamy. We found that not all the populations showed the same amount of RoHs (Kruskal-Wallis Test P-value = 4.413e-05). Out of the five considered populations, Alt Urgell was the one with the longest RoHs and Bergueda with the shortest ones (Figure 3C). When comparing SEP populations with the SpExomes dataset, Alt Urgell has the lowest nHetR of all the SEP populations and within the nHetR of the SpExomes (Supplementary Figure 7).

Therefore, all the results suggest that the rural populations of the Pyrenees show particular demographic histories compared with the general Spanish population as well as between them, despite of their close geographic proximity.

Demographic history of SEP

We modelled the demography of the SEP populations by means of ABC-DL (see Material and Methods). The ABC-DL analysis suggests that the observed population substructure in SEP generated around the 7th century BPE (15th - 4th BPE CI95%), and that the five SEP populations have endured a population decline during the last ~100 generations (see Supplementary Figure 1 and Table 1). Alt Urgell showed the strongest reduction in effective population size; in contrast, the estimated effective population size of Bergueda was eight times the observed in Alt Urgell. These results agree with the different decay of LD and different RoHs patterns in Alt Urgell compared to the other SEP populations. In particular, a statistically significant negative Pearson correlation is observed between the median of the log(RoHs) by population and the estimated effective population size per population (-0.942, p-value = 0.017).

Loss of Function (LoF) analyses

From these results, we inquired the effect that demographic history of SEP could have on the LoF burden compared to the general population from the Spanish-exomes. Berguedà and Garrotxa showed a significantly lower LoF median than expected if they were sampled from the Spanish-exomes population (Monte Carlo p-value after 100,000 resampling = 0.0022 for Berguedà and 0.0286 for Garrotxa, respectively) (see Figure 4); however, after Bonferroni correction for multiple testing, only Berguedà remained statistically significant.

Discussion

In agreement with previous results based on non-NGS data (16,38), our MDS analyses place the SEP individuals within the South Western context of the genomic diversity within Europe and, particularly, within the Iberian samples of the SGDP dataset. Within SEP, we recover the west to east axis of genetic diversity in the Pyrenees initially reported using classical markers, which has been explained in terms of the original peopling of the area (13). More interestingly, our analyses detected the presence of ultra-fine-geographic differentiation across the five SEP populations when using haplotype-based data and Chromopainter/fineSTRUCTURE. Similar levels of ultra-fine genetic differentiation have been observed in some rural regions of Galicia (39). Nevertheless, in our study we observed that this fine-population substructure can only be detected by using WGS. The two clusters identified by Chromopainter/fineSTRUCTURE show a marked geographical component, as estimated by a spatial model that includes geographic barriers and anisotropy, and independently replicated by EEMS. The orography between the two clusters cannot explain this genetic differentiation, as within the PUB cluster there are greater orographic phenomena than between PUB and RG. Furthermore, the presence in Alt Urgell and Pallars of the genetic component characteristic of Berguedà (in blue in Figure 1B), but not in Ripollés, which is geographically closer to Berguedà, suggests the presence of complex historical demographic processes within SEP on top of the orography.

The absence of differences between regions with respect to their patterns of Western Eurasian ancestry, as shown in the GLOBETROTTER results, suggests that geographic isolation within SEP is likely the cause of the identified substructure. This isolation should have appeared after major migrations into the region, or these had a very limited impact in the genetic makeup of SEP. In particular, it has been claimed that fine genetic variation has been shaped in Spain by linguistic and geopolitical boundaries at the time of Muslim rule in Spain (39). However, the Roman and Visigoth people only represented a 2.2-4.4% of the SEP population, while the Islamic conquest of this region lasted only 80 years (40). In this line, ABC-DL suggests that the population structure observed in SEP originated in the 7th century BPE. Furthermore, the low effective population sizes inferred from genomic data support genetic isolation as the main factor for explaining the geographic structure. However, it is interesting to notice the large heterogeneity in the estimates of the effective population size given the geographic proximity of the SEP populations. These estimations are in agreement with the estimated levels of autozygosity and LD. Therefore, despite the short distances between populations, the particular demographic histories of the different villages played a role in shaping the genomic landscape of the regions. In fact, these counties traditionally shared a rural lifestyle but considered different methods of subsistence depending on their geographic location and period (40). For example, by the end of the XIX century, the economy of Berguedà focused on the exploitation of natural resources (in the upper part) and textile (in the lower part). This type of economy granted a railway to the county connecting it to the most industrialized part of Catalonia by 1914 (41).

Our analyses suggest that the observed individual LoF patterns are consistent with the LoF diversity present in Spain individuals for all SEP, but not for Berguedà, which consistently showed a lower number of LoF than expected in the Spanish population. One possible explanation is the systematic bias in the age of the samples, which could have favored ascertaining individuals with a low number of LoF mutations to reach elderly. Similarly, sequencing errors and batch effects between the SEP and SpEx datasets could bias the amount of reported LoF. However, in both cases we would have expected similar LoF patterns over all the SEP populations. It has been shown that long-term isolation and endogamy can clean the genome from deleterious mutations (42). Nevertheless, this explanation is incompatible with the estimated effective population sizes and the identified levels of micro-population substructure of the region, particularly in Berguedà. Therefore, other factors related to the particular demographic history must be shaping the lower amount of LoF in this population.

In this work we have described the genetic variation of five rural villages from the SEP through the analysis of high coverage WGS data accompanied by detailed genealogical information. Our results suggest that geographically close SEP villages could show particular demographic histories. However, further analyses will be required to study if the observed pattern extends to other geographic regions of the Pyrenees, at both the Spanish and French side.

Acknowledgements

We thank U. M. Marigorta for his helpful comments and suggestions on the manuscript. Thanks also to all the volunteers who decided to participate in the study.

Conflict of interests

All the authors declare no conflict of interest.

References

1. Champion A. Europe’s Rural Demography. In: Kulcsár LJ, Curtis KJ, editors. International Handbook of Rural Demography. 1st ed. Springer Netherlands; 2012. p. 81–93.

2. Brown DL. Migration and Rural Population Change: Comparative Views in More Developed Nations. In: Kulcsár LJ,

Curtis KJ, editors. International Handbook of Rural Demography. 1st ed. Springer Netherlands; 2012. p. 35–48.

3. Sparks J. Rural Health Disparities. In: Kulcsár LJ, Curtis KJ, editors. International Handbook of Rural Demography. 1st ed. Springer Netherlands; 2012. p. 255–71.

4. Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. Nat Commun. 2017 Jun 23;8.

5. Pinilla V, Sáez LA. Rural depopulation in Spain: Genesis of a problem and innovative policies. 2017.

6. Silveira LE Da, Alves D, Painho M, Costa AC, Alcântara A. The evolution of population distribution on the Iberian Peninsula: A transnational approach (1877-2001). Hist Methods. 2013;46(3):157–74.

7. Calderón R, Hernández CL, García-Varela G, Masciarelli D, Cuesta P. Inbreeding in Southeastern Spain. Hum Nat. 2018;29(1):45–64.

8. Mccullough JM, O'Rourke DH. Geographic distribution of consanguinity in europe. Ann Hum Biol. 1986;13(4):359–67.

9. Fuster V, Colantonio SE. Socioeconomic, demographic, and geographic variables affecting the diverse degrees of consanguineous marriages in Spain. Hum Biol. 2004;76(1):1–14.

10. Toledo A, Pámpanas L, García D, Pettener D, González-Martin A. Changes in the genetic structure of a valley in the Pyrenees (Catalonia, Spain). J Biosoc Sci. 2017;49(1):69–82.

11. Martín-Còlliga A, Vaquer J. El poblament dels Pirienus a l'Holocè, del mesolític a l'edat del bronze. In: Vives E, Bertranpetit J, editors. Muntanys i Població El passat dels Pirineus des d'una perspectiva multidisciplinària. 1995. p. 35–73.

12. Solé A, Solana M, Mendizabal E. Integration and international migration in a mountain area The Catalan Pyrenees. Rev géographie Alp. 2014;(102–3):0–13.

13. Calafell F, Bertranpetit J. Mountains and genes: Population history of the Pyrenees. Hum Biol. 1994;66(5):823–42.

14. Giraldo MP, Eesteban E, Aluja MP, Nogués RM, Backés-Duró C, Dugoujon JM, et al. Gm and Km alleles in two Spanish Pyrenean populations (Andorra and Pallars Sobirà): a review of Gm variation in the Western Mediterranean basin. Ann Hum Genet. 2001;65(6):537–48.

15. López-Parra AM, Gusmão L, Tavares L, Baeza C, Amorim A, Mesa MS, et al. In search of the Pre- and Post-Neolithic Genetic Substrates in Iberia: Evidence from Y-Chromosome in Pyrenean Populations. Ann Hum Genet. 2009 Jan 1;73(1):42–53.

16. Biagini SA, Solé-Morata N, Matisoo-Smith E, Zalloua P, Comas D, Calafell F. People from Ibiza: an unexpected isolate in the Western Mediterranean. Eur J Hum Genet. 2019;27:941–51.

17. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

18. Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and

genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May 10;43(5):491–501.

19. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016 Sep 21;538(7624):201–6.

20. Dopazo J, Amadoz A, Bleda M, Garcia-Alonso L, Alemán A, García-García F, et al. 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. Mol Biol Evol. 2016 May 1;33(5):1205–18.

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep 1;81(3):559–75.

22. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013 Jan 1;10(1):5–6.

23. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015 Sep 30;526(7571):68–74.

24. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012 Jan;8(1):1002453.

25. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science (80-). 2014 Feb 14;343(6172):747–51.

26. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet. 2015 Dec 29;48(1):94–100.

27. Mondal M, Bertranpetit J, Lao O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. Nat Commun. 2019 Dec 16;10(1):246.

28. Sabatti C, Risch N. Homozygosity and linkage disequilibrium. Genetics. 2002;160(4):1707–19.

29. VanLiere JM, Rosenberg NA. Mathematical properties of the measure of linkage disequilibrium. Theor Popul Biol. 2008 Aug;74(1):130–7.

30. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet. 2012;91(2):275–92.

31. Samuels DC, Wang J, Ye F, He J, Levinson RT, Sheng Q, et al. Heterozygosity ratio, a robust global genomic measure of autozygosity and its association with height and disease risk. Genetics. 2016 Nov 1;204(3):893–904.

32. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012;3(MAR).

33. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013 Sep 1;34(9):E2393–402.

34. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6(2):80–92.

35. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. 2011 Sep 1;39(17):e118–e118.

36. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

37. Rausell A, Luo Y, Lopez M, Seeleuthner Y, Rapaport F, Favier A, et al. Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. Proc Natl Acad Sci. 2020 Jun 16;117(24):13626–36.

38. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, et al. Correlation between Genetic and Geographic Structure in Europe. Curr Biol. 2008 Aug;18(16):1241–8.

39. Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, Quintela I, Carracedo Á, Donnelly P, et al. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. Nat Commun. 2019;10(1):1–14.

40. Riu-Riu M. El poblament dels Pirineus, segles VII-XIV. In: Vives E, Bertranpetit J, editors. Muntanys i Població El passat dels Pirineus des d'una perspectiva multidisciplinària. 1995. p. 195–220.

41. Serra-Rotés R. Carretera, ferrocarril i industrialització a la comarca del Berguedà (Barcelona) al segle XIX i principis del XX.

In: III Congrés Internacional d'Història dels Pirineus. 2017. p. 487–500.

42. Simons YB, Sella G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Curr Opin Genet Dev. 2016 Dec 1;41:150–8.

43. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 2005 Oct;128(2):415–23.

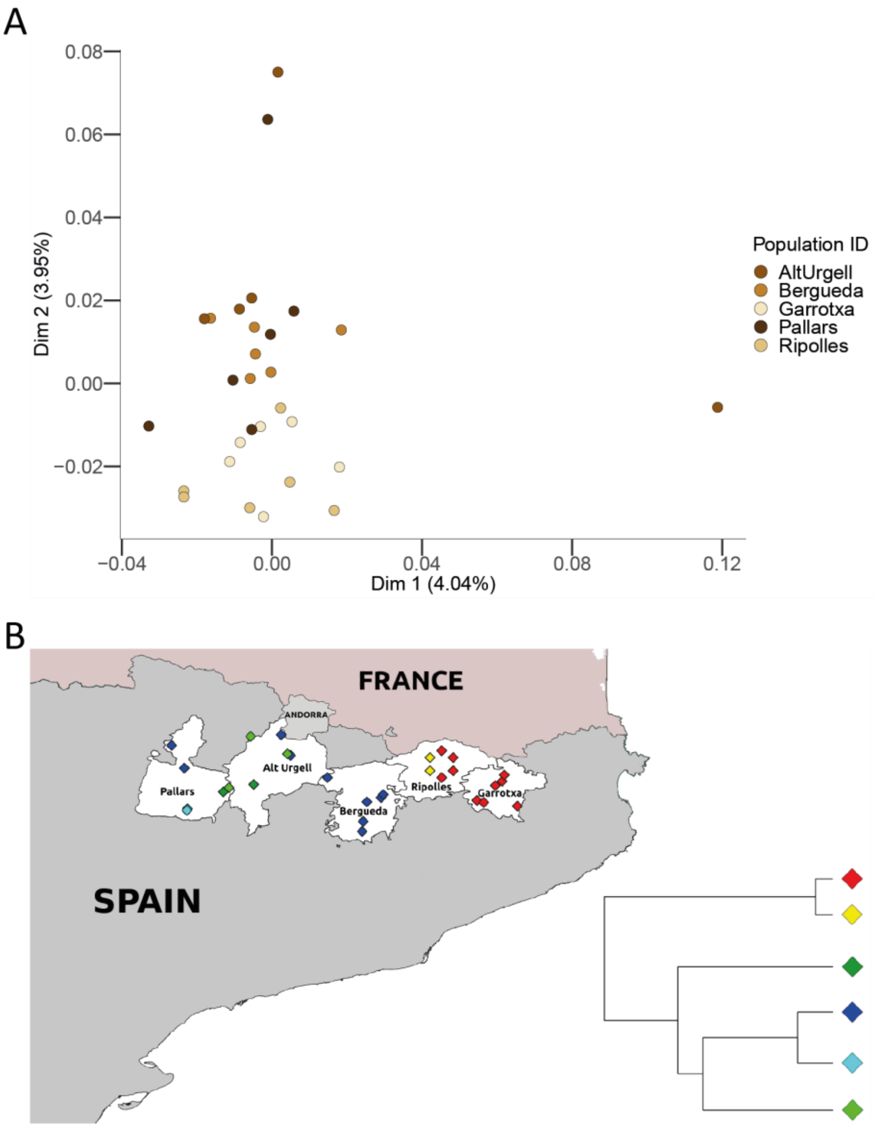


Figure 1. Genetic variation of SEP. A) Classical MDS of the samples from SEP. **B)** Map showing SEP painted accordingly to the cluster they belong to. Ripollès samples are artificially dispersed for the sake of clarity. Simplified fineSTRUCTURE tree of SEP samples, showing six clusters which can be further summarized in two main groups: Garrotxa-Ripollès and Pallars-Alt Urgell-Berguedà.

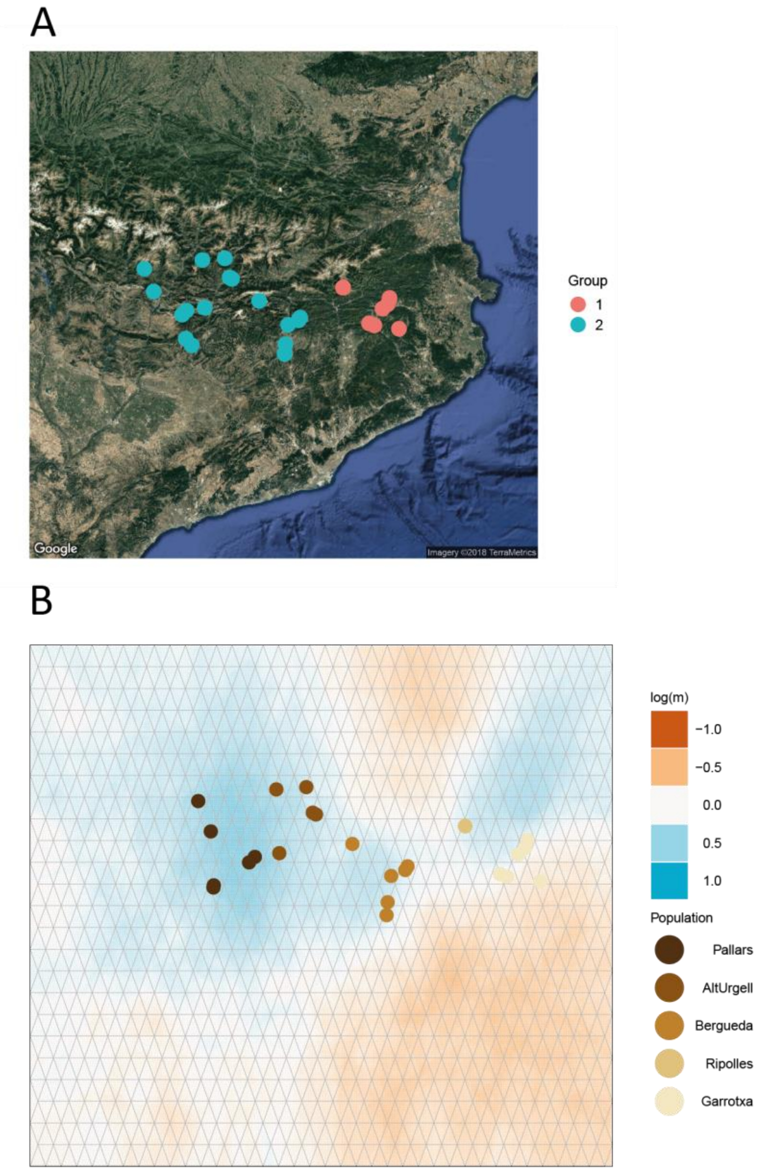


Figure 2. Autozygosity, inbreeding and LD in SEP. A) Genetic barrier between Garrotxa-Ripollés (red dots) and Pallars-Alt Urgell-Berguedà (blue dots) identified by an algorithm that models the genetic variation present in the data in terms of anisotropy and genetic barriers. **B)** EEMS result also state a migration barrier between Garrotxa-Ripollés and Pallars-Alt Urgell-Berguedà.

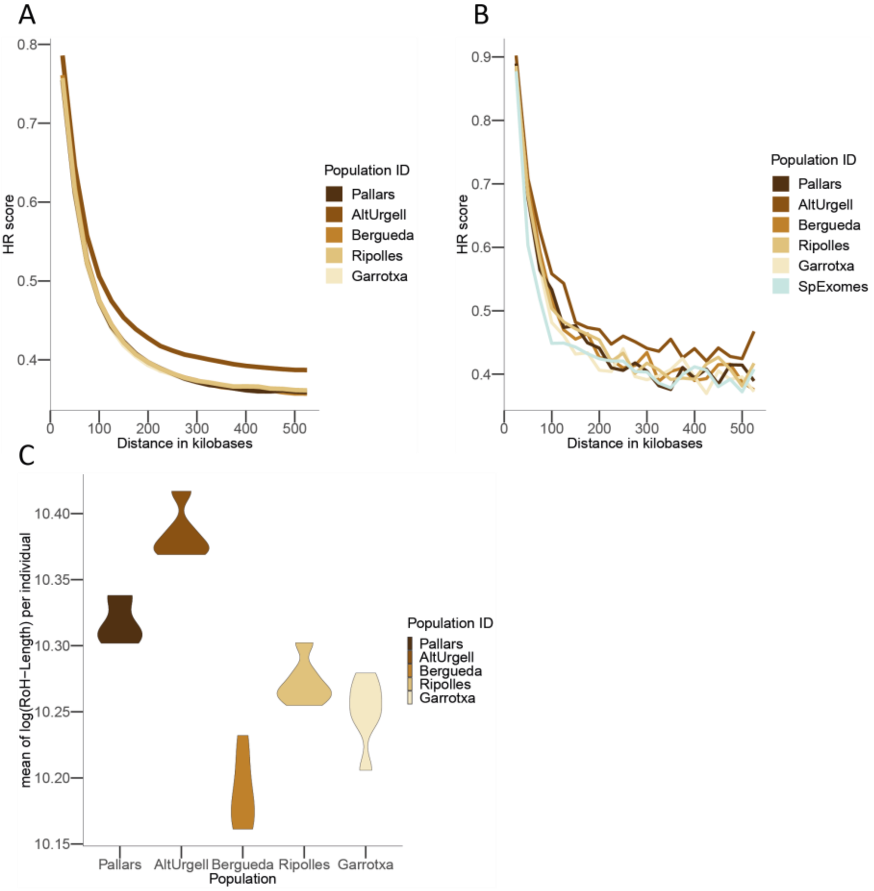


Figure 3. Decay of LD and ROH and HetR of SEP and SpExomes samples. A) LD decay in SEP samples using WGS. **B)** LD decay of SEP samples with the Spanish Exomes dataset using exome sequencing data. **C)** Violin plot of the total amount of homozygous fragments in each SEP individual using WGS.

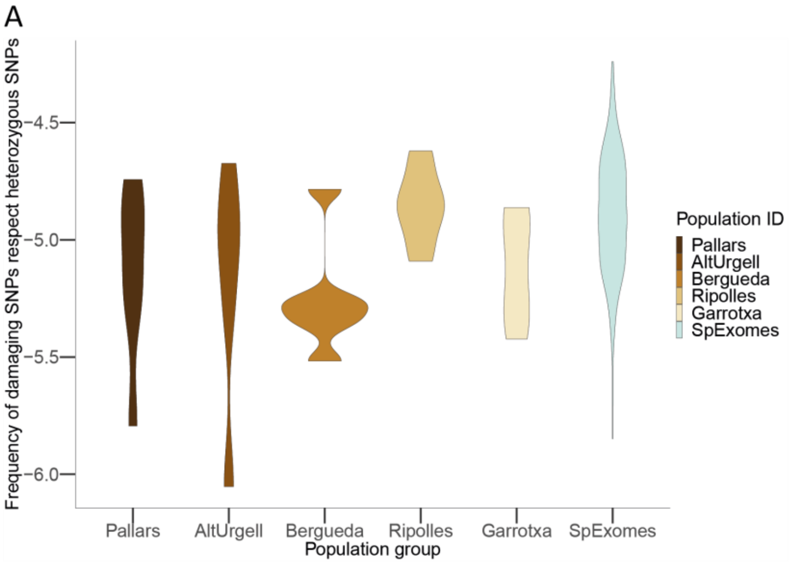


Figure 4. Mutation load analysis. Violin plot of the frequency of Damaging Heterozygote SNPs of each individual compared to all the SNPs in the exome that do not show homozygote derived genotypes, and that are heterozygote in the considered individual.

	Prior		Posterior			
	Minimum	Maximum	CI 2.5%	Median	CI 97.5%	Mean
NeGRUPB	15000.00	40000.00	27992.65	28427.69	28818.08	28421.97
NeG	2000.00	10000.00	2729.9	4107.69	6520.49	4250.97
NeR	2000.00	10000.00	2380.96	3392.74	5040.98	3470.81
NeU	1000.00	10000.00	1005.02	1124.95	1359.1	1142.67
NeB	2000.00	10000.00	5599.93	8568.44	9957.02	8328.27
NeP	2000.00	10000.00	2256.47	3119.74	4663.26	3214.29
tGRUBP	3625.00	2320.00	2319.12	2624.85	3449.17	2697.68

Table 1. 95% Credible interval (2.5% - 97.5%), median and mean for the effective population size of the meta-population (NeGRUBP), Garrotxa (NeG), Ripollés (NeR), Urgell (NeU), Berguedà (NeB) and Pallars (NeP) and the time of the split (tGRUBP, in thousands of years ago, assuming a generation time of 29 years (Fenner, 2005)) estimated using ABC-DL.

Supplementary information

Fine-scale population structure in five rural populations from the Spanish Eastern Pyrenees using high-coverage whole-genome sequence data

Iago Maceda, Miguel Álvarez-Álvarez, Georgios Athanasiadis, Raúl Tonda, Jordi Camps, Sergi Beltran, Agustí Camps, Jordi Fàbrega, Josefina Felisart, Joan Grané, José Luis Remón, Jordi Serra, Pedro Moral, and Oscar Lao

SEP WGS data cleaning

Data cleaning was performed using a strict threshold of 0% missingness and excluding those SNPs that were out of HWE. As a last step for data cleaning, we checked for kinship between our individuals using KING (1), resulting in the exclusion of one individual from Alt Urgell. Only the autosomes were kept at the end of the data cleaning step. The final dataset contained 29 individuals and 9,309,056 biallelic polymorphic SNPs.

Inference of the ancestral allele

In order to define the derived allele needed in the ABC-DL algorithm, we downloaded the best reciprocal alignments between Hg19 and PanTro4 from the UCSC in AXT format. From these files we reconstructed a chimpanzee genome adding "-" (no base) as filler for the gaps in the alignment, using chromosome lengths from Hg19 from the UCSC. An intermediate BED was generated containing the chimpanzee alleles for the SNPs present in the VCF, which was merged with the original VCF. Those SNPs for which the chimpanzee allele was either unknown or not properly aligned were filtered out.

Identification of genetic barriers and anisotropy patterns in the SEP dataset using the shared coancestry matrix

For samples from the same geographic group, one can define the minimum genetic distance expected (nugget, n), as well as a maximum genetic distance no matter how much geographically distant are the two samples (sill, s). If the genetic differentiation follows a direction in space (i.e. there is a genetic gradient in the data points), then the model can be extended to include an angle α of maximum genetic differentiation over space. We define the genetic distance between two points as:

$$(1) \quad GenDist = s * k + n$$

$$(2) \quad k = \frac{\sqrt{b^2 * (x * \cos(\alpha) + y * \sin(\alpha))^2 + \ a^2 * (x * \sin(\alpha) - y * \cos(\alpha))^2}}{a^2 * b^2}$$

In particular, under the Multiple regressions on distance matrices (2) framework, α and β can be estimated using classical nonlinear regression fitting methods. Here we applied the Nelder and Mead Simplex method implemented at Flanagan's JAVA package (3). For a given K geographic groups, the identification of the genetic barriers consists on identifying the set of geographically related points that minimizes the goodness of fit of the estimated parameters. Let be the mean sum of square error of each point between the observed genetic distance and the inferred D' from the fitted parameters:

$$(3)\ SSE =$$

$$SquareError_g =$$

$$\sum_{g}^{K} \left\{ \begin{array}{l} \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} \frac{\left(D(i,j) - D'(i,j)\right)^2}{m-1} \quad if\ \frac{n_g(n_g-1)}{2} \geq 3 \\ \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} \frac{(D(i,j) - \overline{D})^2}{m-1} \quad if\ \frac{n_g(n_g-1)}{2} < 3 \end{array} \right\}$$

In order to optimize the SSE function and to identify the genetic barriers, we define each geographic group by a [x,y] pair of coordinates, and assign each observed point based on its proximity to each geographic group, such as in K-means algorithm. The problem is then to find the geographic coordinates of each group that minimize the SSE. In order to explore the space of possible solutions, we propose using a genetic algorithm. This type of approaches have been already applied in optimization problems involving geographic divisions (4).

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

A total of 300,000 simulations were generated using fastsimcoal2 (5), each simulating 7,314 genomic regions separated by at least 100 kb encompassing ~647 Megabases (Mb), that do not contain CpG islands or genes. The generation time was 29 years (6) and the mutation rate was set to 1.61e-8 with a standard deviation of 0.13e-

8 (7). From the 300,000 simulations, a total of 30,000 simulations were used in the DL training, and the remaining 270,000 in the ABC step. One sample from each comarca was used for generating the jSFS which, by repeatedly being added noise before merging with each normalized simulation, was employed in the artificial neural network (ANN) training (see (8) for details of the implementation). For each parameter of the demographic model, a total of 10 independent ANN's, each featuring four neural layers with 100 neurons, were trained using resilient backpropagation and dropout at 0.1 for a maximum duration of 2.5 hours, or until an error <0.01 was reached. In order to generate a single summary statistic out of all the 10 independent ANN’s, the output from each ANN, namely the prediction of the value of the considered parameter, was averaged over all the ANN’s.

For the ABC step, the remaining individuals not used in the noise injection step were considered. We generated 100 resampled datasets by taking one individual at random from each population. With each dataset we computed the jSFS and used the trained ANN’s to predict the different parameters, which were then used as summary statistics in the ABC approach to infer the posterior distributions of each parameter. For that, we used the abc package (9) with the local linear algorithm (10). The final posterior distribution of each parameter was obtained by combining the posterior distributions out of the 100 sampled datasets.

Supplementary Information References

1. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–73.

2. Legendre P, Legendre L. Numerical Ecology. 3rd Editio. Elsevier Ltd; 2012.

3. Thomas-Flanagan M. Regression Class: Linear and Non-linear Regression. 2016.

4. Sergeeva M, Delahaye D, Mancel C, Vidosavljevic A. Dynamic airspace configuration by genetic algorithm. J Traffic Transp Eng (English Ed. 2017 Jun 1;4(3):300–14.

5. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust Demographic Inference from Genomic and SNP Data. PLoS Genet. 2013 Oct;9(10).

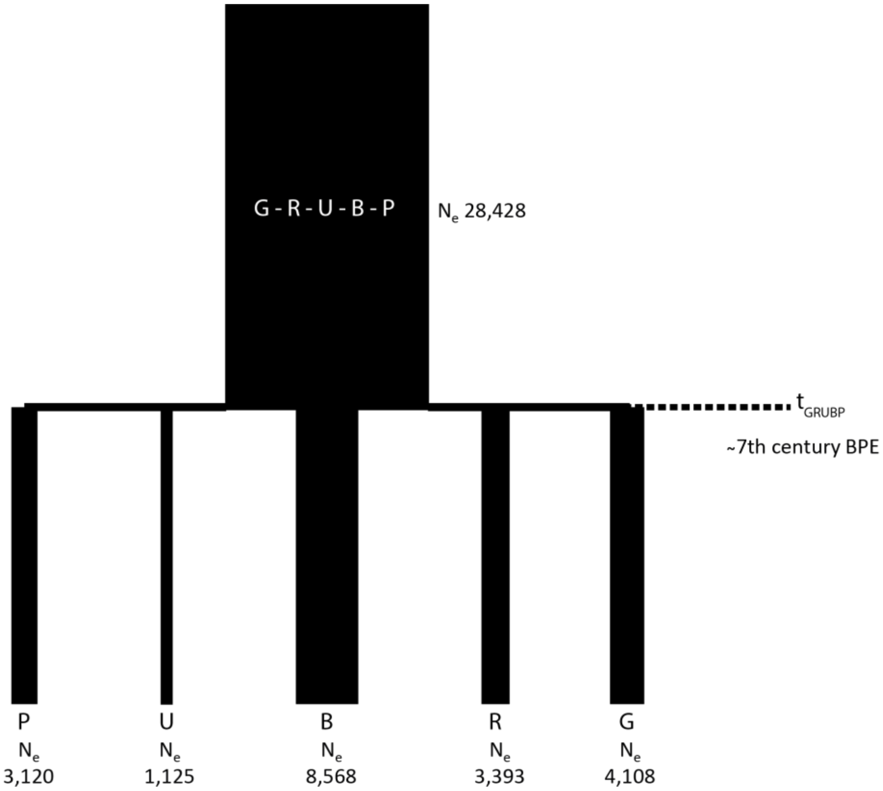
6. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 2005 Oct;128(2):415–23.

7. Lipson M, Loh P-R, Sankararaman S, Patterson N, Berger B, Reich D. Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. Coop G, editor. PLOS Genet. 2015 Nov 12;11(11):e1005550.

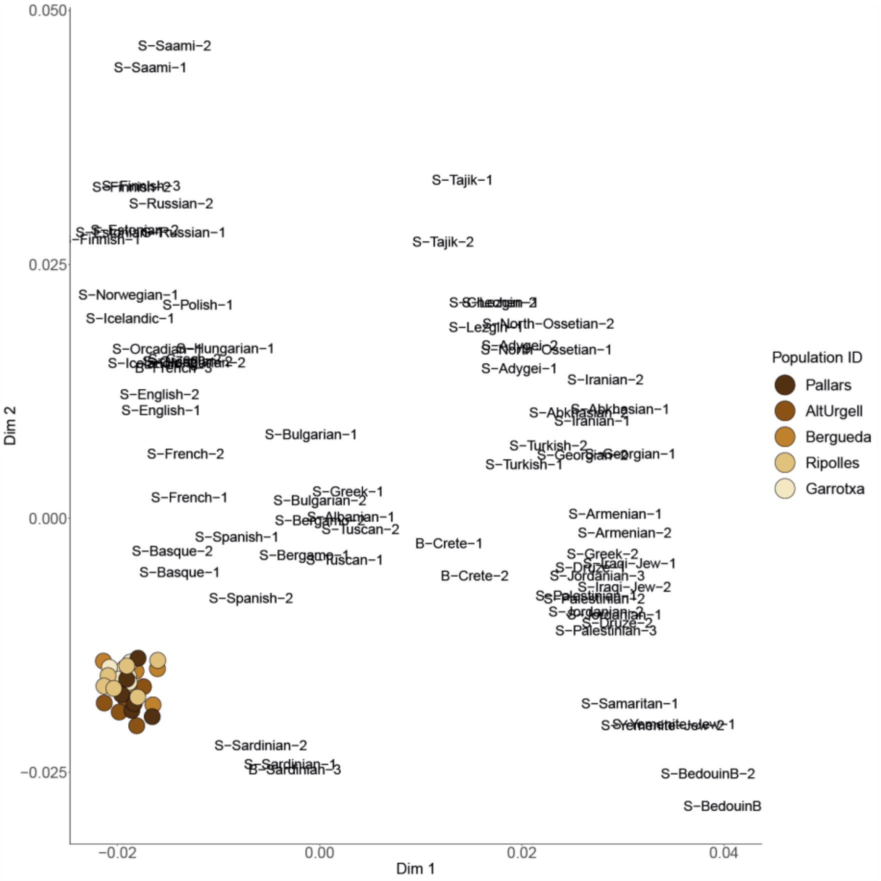
8. Mondal M, Bertranpetit J, Lao O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. Nat Commun. 2019 Dec 1;10(1):1–9.

9. Csilléry K, François O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012 Jun;3(3):475–9.

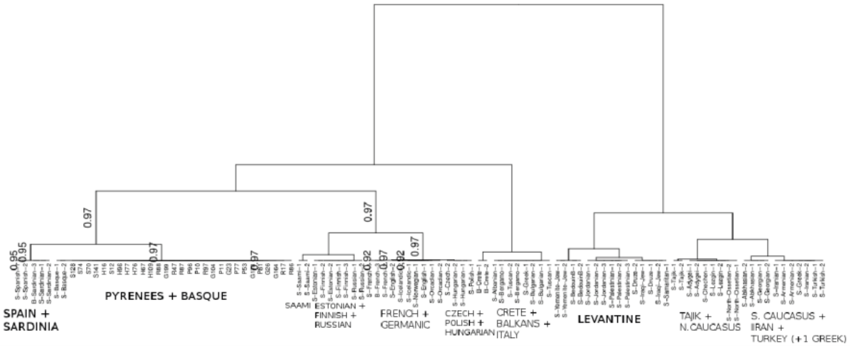
10. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in Population Genetics. Genetics. 2002;162(4):2025–35.



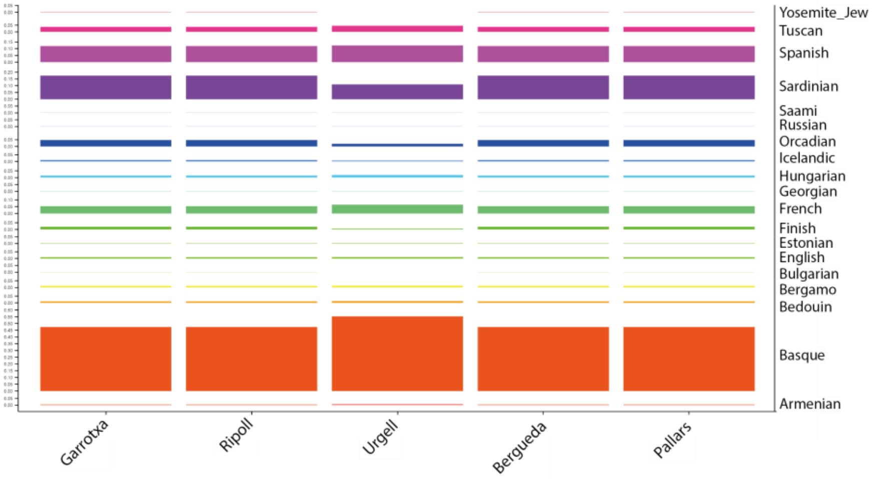
Supplementary Figure 1. Representation of the demography used in the ABC_DL approach. All numbers represent the median of the posterior distribution. (G: Garrotxa; R: Ripollés; U: Alt Urgell; B: Bergadà; P: Pallars; GRUBP: meta-population; tGRUBP: time of split of the meta-population in the present day populations).



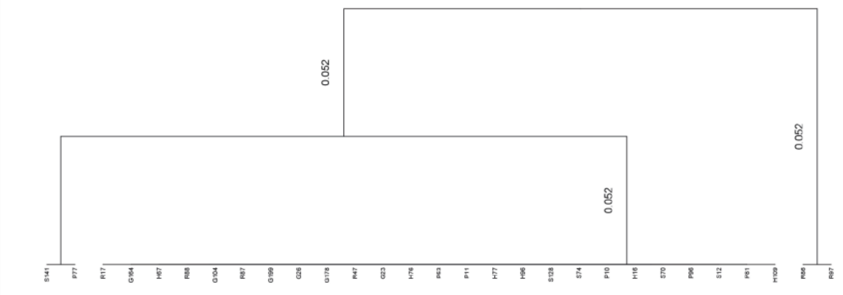
Supplementary Figure 2. Multidimensional scaling using SEP samples (circles, colours denote origin) and SGDP West Eurasian samples.



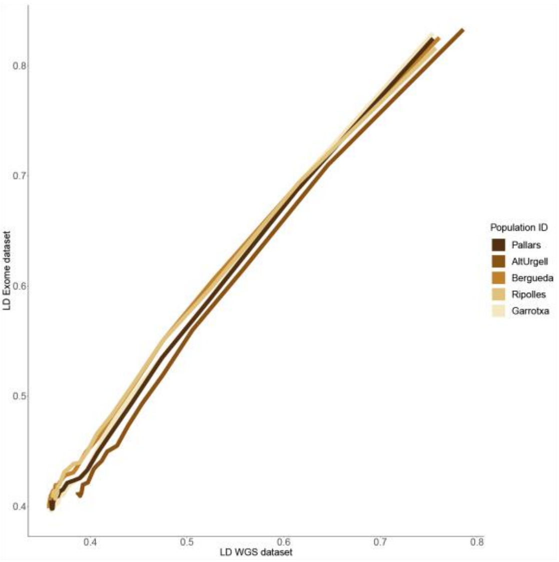
Supplementary Figure 3. fineSTRUCTURE tree showing the relationship of SEP regions with other West-Eurasian populations using data from the SEP-SGDP dataset.



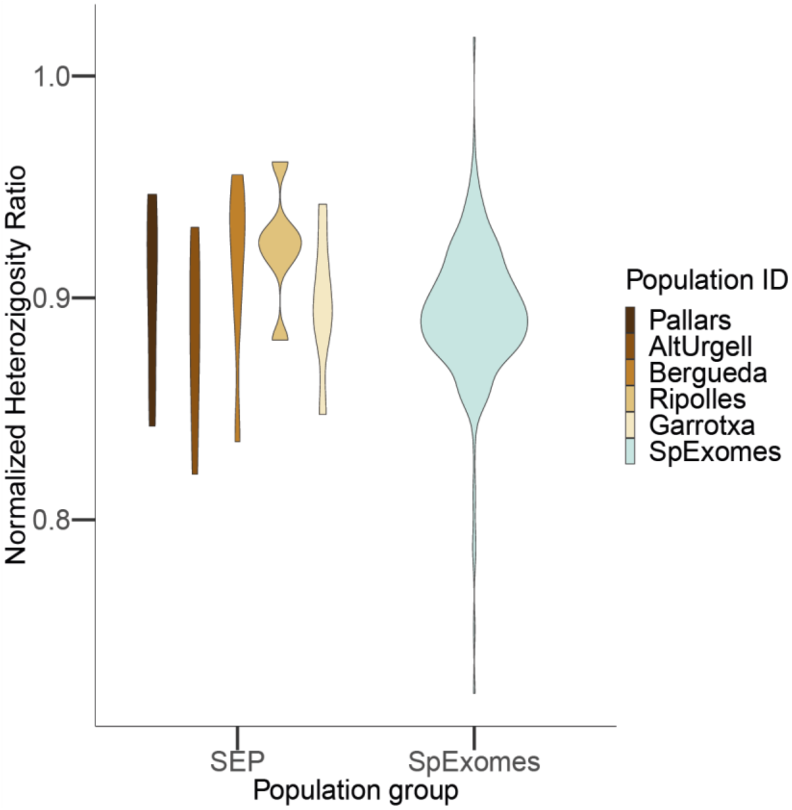
Supplementary Figure 4. GLOBETROTTER ancestral components plot showing the SGDP populations (y-axis) that contributed to the haplotype profiles of the recipient SEP populations (x-axis).



Supplementary Figure 5. fineSTRUCTURE tree including SEP samples and based on 231K SNPs that are common with the Biagini et al. 2019 dataset and present variation in more than one sample. Note that the downsampling causes the patterns of population structure -observed when the full amount of variation is included in the analyses- to disappear. SEP sample codes are explained in the Supplementary table.



Supplementary Figure 6. Comparison between the results of the HR algorithm regarding LD between WGS and Exome datasets in SEP samples. In the x-axis is represented the HR score in the WGS dataset, in the y-axis is represented the HR score in the exome dataset.



Supplementary Figure 7. Normalized Heterozigosity Ratio of SEP and SpExomes samples. SEP samples nHetR fall inside the distribution of the score for SpExomes samples.

Dataset	**Population**	**Sample**	**Sex**	**Age**
SEP	Garrotxa	G178	male	86
SEP	Garrotxa	G26	male	83
SEP	Garrotxa	G199	male	81
SEP	Garrotxa	G23	female	87
SEP	Garrotxa	G104	female	83
SEP	Garrotxa	G164	female	83
SEP	Ripollès	R86	male	71
SEP	Ripollès	R97	male	71
SEP	Ripollès	R47	male	70
SEP	Ripollès	R88	female	68
SEP	Ripollès	R87	female	66
SEP	Ripollès	R17	female	66
SEP	Berguedà	H67	male	83
SEP	Berguedà	H76	male	80
SEP	Berguedà	H109	male	77
SEP	Berguedà	H77	female	84
SEP	Berguedà	H96	female	79
SEP	Berguedà	H16	female	77
SEP	Alt Urgell	S141	male	70
SEP	Alt Urgell	S138	male	64
SEP	Alt Urgell	S128	male	63
SEP	Alt Urgell	S70	female	69
SEP	Alt Urgell	S74	female	65
SEP	Alt Urgell	S12	female	64
SEP	Pallars	P81	male	81
SEP	Pallars	P96	male	80
SEP	Pallars	P10	male	79
SEP	Pallars	P11	female	77
SEP	Pallars	P53	female	77
SEP	Pallars	P77	female	77
SGDP_WestEurasia	Crete	B-Crete-1	-	-
SGDP_WestEurasia	Crete	B-Crete-2	-	-
SGDP_WestEurasia	French	B-French-3	-	-
SGDP_WestEurasia	Sardinian	B-Sardinian-3	-	-
SGDP_WestEurasia	Abkhasian	S-Abkhasian-1	-	-
SGDP_WestEurasia	Abkhasian	S-Abkhasian-2	-	-
SGDP_WestEurasia	Adygei	S-Adygei-1	-	-

Dataset	Population	Sample	Sex	Age
SGDP_WestEurasia	Adygei	S-Adygei-2	-	-
SGDP_WestEurasia	Albanian	S-Albanian-1	-	-
SGDP_WestEurasia	Armenian	S-Armenian-1	-	-
SGDP_WestEurasia	Armenian	S-Armenian-2	-	-
SGDP_WestEurasia	Basque	S-Basque-1	-	-
SGDP_WestEurasia	Basque	S-Basque-2	-	-
SGDP_WestEurasia	BedouinB	S-BedouinB-1	-	-
SGDP_WestEurasia	BedouinB	S-BedouinB-2	-	-
SGDP_WestEurasia	Bergamo	S-Bergamo-1	-	-
SGDP_WestEurasia	Bergamo	S-Bergamo-2	-	-
SGDP_WestEurasia	Bulgarian	S-Bulgarian-1	-	-
SGDP_WestEurasia	Bulgarian	S-Bulgarian-2	-	-
SGDP_WestEurasia	Chechen	S-Chechen-1	-	-
SGDP_WestEurasia	Czech	S-Czech-2	-	-
SGDP_WestEurasia	Druze	S-Druze-1	-	-
SGDP_WestEurasia	Druze	S-Druze-2	-	-
SGDP_WestEurasia	English	S-English-1	-	-
SGDP_WestEurasia	English	S-English-2	-	-
SGDP_WestEurasia	Estonian	S-Estonian-1	-	-
SGDP_WestEurasia	Estonian	S-Estonian-2	-	-
SGDP_WestEurasia	Finnish	S-Finnish-1	-	-
SGDP_WestEurasia	Finnish	S-Finnish-2	-	-
SGDP_WestEurasia	Finnish	S-Finnish-3	-	-
SGDP_WestEurasia	French	S-French-1	-	-
SGDP_WestEurasia	French	S-French-2	-	-
SGDP_WestEurasia	Georgian	S-Georgian-1	-	-
SGDP_WestEurasia	Georgian	S-Georgian-2	-	-
SGDP_WestEurasia	Greek	S-Greek-1	-	-
SGDP_WestEurasia	Greek	S-Greek-2	-	-
SGDP_WestEurasia	Hungarian	S-Hungarian-1	-	-
SGDP_WestEurasia	Hungarian	S-Hungarian-2	-	-
SGDP_WestEurasia	Icelandic	S-Icelandic-1	-	-
SGDP_WestEurasia	Icelandic	S-Icelandic-2	-	-
SGDP_WestEurasia	Iranian	S-Iranian-1	-	-
SGDP_WestEurasia	Iranian	S-Iranian-2	-	-
SGDP_WestEurasia	Iraqi-Jew	S-Iraqi-Jew-1	-	-
SGDP_WestEurasia	Iraqi-Jew	S-Iraqi-Jew-2	-	-

Dataset	Population	Sample	Sex	Age
SGDP_WestEurasia	Jordanian	S-Jordanian-1	-	-
SGDP_WestEurasia	Jordanian	S-Jordanian-2	-	-
SGDP_WestEurasia	Jordanian	S-Jordanian-3	-	-
SGDP_WestEurasia	Lezgin	S-Lezgin-1	-	-
SGDP_WestEurasia	Lezgin	S-Lezgin-2	-	-
SGDP_WestEurasia	North-Ossetian	S-North-Ossetian-1	-	-
SGDP_WestEurasia	North-Ossetian	S-North-Ossetian-2	-	-
SGDP_WestEurasia	Norwegian	S-Norwegian-1	-	-
SGDP_WestEurasia	Orcadian	S-Orcadian-1	-	-
SGDP_WestEurasia	Orcadian	S-Orcadian-2	-	-
SGDP_WestEurasia	Palestinian	S-Palestinian-1	-	-
SGDP_WestEurasia	Palestinian	S-Palestinian-2	-	-
SGDP_WestEurasia	Palestinian	S-Palestinian-3	-	-
SGDP_WestEurasia	Polish	S-Polish-1	-	-
SGDP_WestEurasia	Russian	S-Russian-1	-	-
SGDP_WestEurasia	Russian	S-Russian-2	-	-
SGDP_WestEurasia	Saami	S-Saami-1	-	-
SGDP_WestEurasia	Saami	S-Saami-2	-	-
SGDP_WestEurasia	Samaritan	S-Samaritan-1	-	-
SGDP_WestEurasia	Sardinian	S-Sardinian-1	-	-
SGDP_WestEurasia	Sardinian	S-Sardinian-2	-	-
SGDP_WestEurasia	Spanish	S-Spanish-1	-	-
SGDP_WestEurasia	Spanish	S-Spanish-2	-	-
SGDP_WestEurasia	Tajik	S-Tajik-1	-	-
SGDP_WestEurasia	Tajik	S-Tajik-2	-	-
SGDP_WestEurasia	Turkish	S-Turkish-1	-	-
SGDP_WestEurasia	Turkish	S-Turkish-2	-	-
SGDP_WestEurasia	Tuscan	S-Tuscan-1	-	-
SGDP_WestEurasia	Tuscan	S-Tuscan-2	-	-
SGDP_WestEurasia	Yemenite-Jew	S-Yemenite-Jew-1	-	-
SGDP_WestEurasia	Yemenite-Jew	S-Yemenite-Jew-2	-	-

Supplementary Table 1. List of samples initially considered for the analyses.

4.2 Chapter 2

Analysis of the batch effect due to the sequencing centre in population statistics quantifying rare events in the 1000 Genomes Project

Iago Maceda and Oscar Lao

(Manuscript in preparation)

Analysis of the batch effect due to the sequencing centre in population statistics quantifying rare events in the 1000 Genomes Project

Iago Maceda[1] and Oscar Lao[1+]

[(1)]Population Genomics, CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, C/ Baldiri i Reixach 4, 08028 Barcelona, Spain.
+ Corresponding author (oscar.lao@cnag.crg.eu).

Abstract

The 1,000 Genomes Project (1000G) is one of the most popular whole genome sequencing used in different genomics fields, boosting our knowledge in medical genomics and population genomics, among others. Recent studies have reported the presence of ghost mutation signals in the 1000G. Furthermore, they have shown that these mutations can influence the outcome from follow up studies based on the genetic variation of 1000G, such as SNP imputation in GWAS studies.

In this study we analyze the effect of the sequencing center in the predicted loss of function (LoF) alleles, the amount of singletons and the patterns of archaic introgression. Our results support previous studies showing that the sequencing center is systematically associated with LoF and singletons. Furthermore, we observed that the patterns of archaic introgression were distorted for some populations depending on the sequencing center. When analyzing the frequency of SNPs showing extreme patterns of genotype differentiation among centers for CEU, YRI, CHB and JPT, we observed that the magnitude of the sequencing batch effect was stronger at MAF < 0.2, as well as different profiles between CHB and the other populations. All these results suggest that the data from 1000G must be interpreted with caution when considering statistics considering infrequent events.

Introduction

The 1,000 Genomes Project (1000G) (1) has been one of the cornerstones of population genetics, as it provided the first dataset that considered human worldwide variation. This dataset is normally used as basis for imputation in microarrays (2) or to obtain haplotype phased data (3), in evolutionary studies (4–7), in multiple medical studies (8) or as a basis to identify potential genetic isolates (9).

The 1000G corresponds to the first attempt to characterize the worldwide genetic variation on humans. It was born to provide accurate haplotype information across different human populations. To do so, the project aimed to characterize over 95% of variants in genomic regions that have >1% allele frequency in each of the major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas). The project started with 15 populations in the Pilot Phase, and it had a total of 26 populations by the end of Phase 3, when the project was concluded. In the Pilot Phase the project characterized a total of 1,092 samples (not evenly distributed across the different populations) and, by the end of the Phase 3 it accounted for a total of 2,504 samples (closely to have an even distribution across all populations). The final genomic dataset is heterogeneous in nature, comprising individuals at low coverage and exome sequencing data, produced in nine sequencing centers, using five sequencing technologies and bioinformatic pipelines. An additional problem is that populations were not divided evenly across all sequencing centers. For example, the BGI sequenced 25 individuals, the Broad Institute sequenced 86 individuals and the Washington University sequenced 2 individuals of the GWD population. Also, not all centers did all the types of technologies. For example, the Broad Institute ran low coverage whole genome sequencing and whole exome sequencing. In contrast, Max Planck Institute for Molecular Genetics only conducted low coverage whole genome sequencing. Furthermore, each center followed its own set of protocols to prepare the samples.

Some problems related to the genotypes called in 1000G have been already reported. For example, Anderson-Trocmé et al., 2020 cannot reproduce a particular mutation signature (*AC →*CC) reported in the 1000G JPT population (Japanese in Tokyo, Japan) by (11) using a different cohort from the same population (the Nagahama cohort). Also, (12) evaluated the accuracy of the phasing in the Phase 3 samples, concluding that the 1,000 Genomes Project data is best used to impute common variants (MAF>= 0.01) and has limited utility to impute rare variants. Finally, (13) described sets of SNPs showing patterns of linkage disequilibrium likely due to the presence of sequencing errors, and directly linking them to the sequencing center where the individual was sequenced.

Singletons can be artifactually generated in a variety of ways: the sequencer can misread the base that is being incorporated, the base-calling algorithm does not call the proper base, the alignment algorithm matches the read to an incorrect place or the SNP-calling or the genotyping algorithm calls a SNP were it is none (or vice versa). In order to solve this problems, a series of different solutions have been created: filtering SNPs by means of quality score (an associated measure of uncertainty to the SNP) (14) or either examining allele distributions across individuals and calculating its fit to an expected distribution (15).

In this study we wondered to which extent these batch effects due to the sequencing centre could affect measures of genetic variation that have been previously used in estimating loss of function mutations (4), the patterns of singletons and archaic introgression.

Material and Methods

Dataset

To generate the dataset that we used across this work we downloaded the ready-to-use VCF files from the FTP site of the 1,000 Genomes Project (link). These VCF files are divided by chromosome and we decided to concatenate all the files into a single file using bcftools (16). After this step we selected those variants that correspond to SNV, are biallelic and polymorphic across the whole dataset.

We obtained the sequencing centre information from the spreadsheet available in the 1,000 Genomes Project site (https://www.internationalgenome.org/data/). From this spreadsheet we selected the Exome sequencing centre as our sequencing centre reference for all the samples. This decision was made based on the fact that for the low coverage whole genome sequencing some of the samples seem to be sequenced in two different centres according to the spreadsheet. Also based on this parameter we had to exclude one of the samples from the ACB population (HG02537).

Quantification of LoF variants

To predict loss of function variants (LoF) we annotated using the already annotated table of variants from dbNSFP (17). To merge the contents of the table with the VCF file SnpSift (18) and to add the functional prediction we used SnpEff (19). From this annotated and functionally predicted VCF file we decided to use 3 different algorithms to consider categorize a variant as LoF: Polyphen2 (20), MutationAssessor (21) and SIFT (18). If these three algorithms marked a variant as LoF, we accepted that variant into the study.

We restricted the analysis to those variants that do not appear in homozygosis as they would reflect redundant or advantageous effects of dispensable genes (22). To account for this bias, the frequency of LoF variants of each individual was estimated using all the variants in the exome that did not report homozygote individuals for the derived allele in the dataset, and that were heterozygote for the individual.

Quantification of derived singletons

From the general VCF file we generated we extracted those SNPs that correspond to singletons, using the flag AC (number of alternative alleles for that variant across samples) from the INFO field. From this we used the ancestral allele already present in the dataset (flag AA from the INFO field), which uses the 6-way EPO alignments available in Ensembl v71 (23). We compared the reference allele to the ancestral allele. If they are the same allele, we marked the singleton as derived. We excluded those SNPs that either had no alignment for the ancestral allele (AA=.), that were considered as a lineage-specific insertion (AA=-), or those in which the allele was not present (AA=N).

Quantification of archaic introgressed alleles

To count the number of introgressed alleles, we downloaded the output from the Sprime algorithm used in Browning et al (6) in 1000G samples. The output from Sprime is divided on a per population and chromosome basis. These files provide the introgressed allele for the SNP according to the algorithm. We used this information to count the number of alleles of each individual in each population.

Analyses

In order to quantify the batch effect of the sequencing centre in the studied statistics of human population genetics, we used the R package lmer4 (25) to generate hierarchical mixed models controlling for the random effects of the continent and population of assignation of each individual of the type:

$$lmer(log(S){\sim}\ SeqCeter + (1|Continent/Pop)$$

Where S is the variable of interest. Contrast of hypothesis with the mixed null model:

$$lmer(log(S){\sim}(1|Continent/Pop)$$

were conducted with the anova command of R (26).

We also analysed the variants showing an excess of genetic differentiation due to the sequencing centre in YRI, CEU, JPT and CHB populations. These populations were the first considered in the Phase 1 of the 1000G project (27) and they have been widely used in population genomics. For each population and SNP, we assigned each genotype to each sequencing centre and estimated the Fst between the different sequencing centres. Since it has been shown that the magnitude of Fst is dependent on the MAF (28), we controlled all our analyses by MAF bins of 0.05 by weighting the observed Fst in a given SNP by the maximum Fst that could be obtained given that MAF. Furthermore, we only considered SNPs that in the population the MAF allele was present in 10 or more chromosomes, and selected the same number of individuals by centre. For each MAF category, Fst outliers were defined from the top > 99.9% SNPs with Fst > 0 showing the largest Fst differentiation.

Results & Discussion

The finalization of the 1000G project represented a milestone for the human population genetics community (1). Since then, it has been one of the most widely used human population genetic datasets, medical genetics and genetic epidemiology. In human population genetics, the 1000G project has been widely used for understanding the mutation patterns (4), characterizing the genetic variation of the considered human populations (9), identifying segments of archaic introgression (6), studying signatures of positive selection (7,29) or as a gold standard for comparing the burden of LoF (9). In genetic epidemiology, 1000G is routinely used for data phasing and imputation (3,30) in order to increase the SNP density panels genotyped at microarray. Prevalence of LoF variants in healthy individuals (31), insides in cancer genomics (32), short tandem repeats variation (33), evolution and functional impact of short indels (34) and many of the results from this project, have boosted our understanding of the genomics of the species and about the general patterns of diversity present in human populations. Its usage in combination with other WGS datasets as a reference dataset is also a common practice in the field of medical genomics (9).

However, two different papers (10,12) have recently raised concerns about the presence of batch effects at low frequency variants. Maffessoni et al (13) in particular pointed to the sequencing centre as one of the main factors affecting the presence of rare spurious mutations in 1000G individuals. Given these results, in this study we wondered to which extend the sequencing centre, as reported by the spreadsheet of the 1000G (https://www.internationalgenome.org/data/), could influence statistics of population genomics that quantify phenomena that are infrequent in the human genome. First, we studied the number of LoF alleles in each of the 1000G individuals. Assessing the biological impact of in silico predicted LoF is usually complex (35). Therefore, we adopted a conservative approach for predicting recessive damaging variants that severely affect the function of protein-coding genes. We used variants consistently predicted as highly deleterious by Polyphen2 (20), MutationAssessor (21) and SIFT (18) algorithms. We further restricted our analyses towards LoF variants that do not appear in homozygosis, as these most likely reflect redundant and advantageous effects of dispensable human genes (22). By doing such filtering we created a putative bias among populations (i.e. populations that are more genetically diverse will tend to have more chances to have homozygote pseudo-LoF genotypes and therefore to remove them from the analyses). However, this should not influence the comparisons of centres within each population (i.e. see Figure 1). We run a hierarchical mixed model in which the dependent variable was the log(LoF) per individual, the fixed variable was the sequencing centre (BGI, BI and WUGSC) and the random effects were nested by continent and population. We observed statistically significant differences between a mixed model that includes the sequencing centre as variable (ANOVA p-value = 4.87e-20) (Supplementary Table 1).

Next, we wondered if we would observe such batch effect bias in mutations in coding regions classified by the three LoF predictors as benign (Figure 2). In this case, the hierarchical mixed model also supports the role of the sequencing centre (ANOVA p-value =8.315e-08) (Supplementary Table 2). Taken together, the hierarchical mixed models of the LoF and Benign variables suggest that the sequencing centre plays an important role as a batch effect. The presence of such batch effects relates to what Mafessoni, 2019 reported on the enrichment of mutation artifacts in genes. Moreover, as we are considering LoF, and setting the threshold to absence of homozygotes, any sequencing error occurring in a gene will likely tend to produce a false positive that will be recovered by the three algorithms (31). Therefore, it is not surprising that the statistical significance was bigger in the case of LoF compared to Benign, as well as in the magnitude of the estimated slopes (Supplementary Table 1 and Supplementary Table2).

We wondered whether such bias could also be found in the presence of derived singletons in each individual. Mixed models with the log(derived singletons by individual) and log(singletons by individual) support also a role of the sequencing centre (ANOVA p-value for derived singletons = 9.81e-10; ANOVA p-value for singletons = 1.28e-09). However, in this case not all the sequencing centres equally contribute to the bias (see Supplementary Table 3 and Supplementary Table 4), suggesting some heterogeneity between the centres (see Figure 3 and Figure 4). For example, whereas WUGSC tends to decrease the number of derived singletons observed in the individuals sequenced at that centre, BI increases the number of derived singletons and BGI does not significantly affect this variable.

Given these results, we studied if the batch effect due to the sequencing centre could affect the estimation of variables of population genetics that use the derived alleles to make inferences usually at low frequencies. One of these variables is the identification of chunks of DNA that are enriched for derived alleles, which under certain demographic models are indicative of the presence of archaic introgression (6). Furthermore, it has been shown that the presence of archaic introgression depends on the function of the genomic region, being depleted in genomic regions that contain genes (36,37). We used the archaic regions from S. Browning, 2018 that were identified in the 1000G samples, and computed the number of archaic alleles that are found in each individual. No statistical significant differences are observed between a mixed model using the log(number of introgressed alleles) and the sequencing centre and one without the sequencing centre (ANOVA p-value = 0.88; Supplementary Table 5), thus suggesting that these regions are not enriched by batch effect artifacts due to sequencing centre. However, it is interesting to notice that strong discrepancies were observed in the amount of introgressed alleles for some populations (CHB and CHS; see Figure 5). This is particularly relevant because several studies pointed to the presence of heterogeneous patterns of ghost archaic populations in these populations (6,38).

Next, we studied the properties of genetic markers showing strong deviations between sequencing centres for CEU, CHB, JPT and YRI. Our results show that extreme (>99.99%) SNP outliers per MAF category do not show the same Fst pattern after normalizing it by the maximum Fst value that can be obtained given the observed MAF. For CEU, we observed that the normalized Fst value of the outliers (1878 SNPs) decreased with the MAF (Spearman rho = -0.883; p-value < 2.2e-16). This effect was particularly strong for MAF < 0.2 (Figure 6). A similar pattern is observed in the YRI population (2677 SNPs; Spearman's rho = -0.84; p-value < 2.2e-16; Figure 7) and JPT population (1771 SNPs; Spearman's rho = -0.888; p-value < 2.2e-16; Figure 8). In contrast, for CHB (930 SNPs) we do not observe such strong correlation between MAF and normalized Fst (Figure 9; Spearman's rho = 0.095, p-value = 0.0038). In fact, for this population, an inverted U-shape is observed, with SNPs at MAF = 0.35 showing the strongest normalized Fst values. The fact that JPT and CHB share a common ancestry (i.e. they were usually merged in the first analyses of Phase 1 (27)) suggests that the observed pattern is unlikely to be explained by the different ancestry of CHB. Another possibility is that the way how the CHB samples have been processed is different from the carried in the other populations. In any case, the fact that we observed SNPs showing large Fst after normalizing by MAF supports that the effect that we observed in singletons is also observed at higher frequencies. However, the total amount of SNPs that we have considered is quite reduced due to the stringent filters that we applied, and the reported biases are likely not to strongly affect main conclusions out of the 1000G project. Nevertheless, our results suggest that caution must be taken when using the 1000G data and, particularly, when merging it with another dataset that can have its private batch effects.

Funding

O.L and I.M acknowledge the support from Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa, CERCA Programme / Generalitat de Catalunya, Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement, Co-financing with funds from the European Regional Development Fund by the Spanish Ministry of Science and Innovation corresponding to the Programa Operativo FEDER Plurirregional de España (POPE) 2014-2020 and by the Secretaria d'Universitats i Recerca, Departament d'Empresa i Coneixement of the Generalitat de Catalunya corresponding to the Programa Operatiu FEDER de Catalunya 2014-2020. O.L. gratefully acknowledges the financial support from Ministerio de Economía y Competitividad (Ministry of Economy and Competitiveness) - RYC-2013-14797, BFU2015-68759-P and PGC2018-098574-B-I00 and Generalitat de Catalunya (Government of Catalonia) - GRC 2017 SGR 937. I.M. gratefully acknowledges the financial support from the Government of Catalonia | Agència de Gestió d'Ajuts Universitaris i de Recerca (Agency for Management of University and Research Grants) - GRC 2014 SGR 615.

References

1. The 1000 Genomes Project Consortium, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct 30;526(7571):68–74.

2. Wood AR, Perry JRB, Tanaka T, Hernandez DG, Zheng H-F, Melzer D, et al. Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-frequency Variant - Phenotype Associations Undetected by HapMap Based Imputation. Arking DE, editor. PLoS One. 2013 May 16;8(5):e64343.

3. Delaneau O, Howie B, Cox AJ, Zagury J-F, Marchini J. Haplotype Estimation Using Sequencing Reads. Am J Hum Genet. 2013 Oct;93(4):687–96.

4. Huang Y-F, Siepel A. Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. Genome Res. 2019 Aug;29(8):1310–21.

5. Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, et al. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 2014;15(6):R88.

6. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018 Mar;173(1):53-61.e9.

7. Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, et al. 1000 Genomes Selection Browser 1.0: a

genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. 2014 Jan;42(D1):D903–9.

8. Papadimitriou S, Gazzo A, Versbraegen N, Nachtegael C, Aerts J, Moreau Y, et al. Predicting disease-causing variant combinations. Proc Natl Acad Sci. 2019 May 24;201815601.

9. Nutile T, Ruggiero D, Herzig AF, Tirozzi A, Nappo S, Sorice R, et al. Whole-Exome Sequencing in the Isolated Populations of Cilento from South Italy. Sci Rep. 2019 Dec 11;9(1):4059.

10. Anderson-Trocmé L, Farouni R, Bourgey M, Kamatani Y, Higasa K, Seo J-S, et al. Legacy Data Confound Genomics Studies. Wilson M, editor. Mol Biol Evol. 2020 Jan 1;37(1):2–10.

11. Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. Elife. 2017 Apr 25;6.

12. Belsare S, Levy-Sakin M, Mostovoy Y, Durinck S, Chaudhuri S, Xiao M, et al. Evaluating the quality of the 1000 genomes project data. BMC Genomics. 2019 Dec 16;20(1):620.

13. Mafessoni F. Encounters with archaic hominins. Nat Ecol Evol. 2019 Jan 26;3(1):14–5.

14. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Vol. 12, Nature Reviews Genetics. Nature Publishing Group; 2011. p. 443–51.

15. Muyas F, Bosio M, Puig A, Susak H, Domènech L, Escaramis G, et al. Allele balance bias identifies systematic genotyping errors and false disease associations. Hum Mutat. 2019 Jan 1;40(1):115–26.

16. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov;27(21):2987–93.

17. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013 Sep 1;34(9):E2393–402.

18. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012;3(MAR).

19. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6(2):80–92.

20. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

21. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. 2011 Sep 1;39(17):e118–e118.

22. Rausell A, Luo Y, Lopez M, Seeleuthner Y, Rapaport F, Favier A, et al. Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. Proc Natl Acad Sci. 2020 Jun 16;117(24):13626–36.

23. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res. 2014 Jan;42(D1):D749–55.

24. Browning S. Sprime results for 1000 Genomes non-African populations and SGDP Papuans. 2018.

25. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1).

26. R: The R Project for Statistical Computing [Internet]. [cited 2020 Dec 18]. Available from: https://www.r-project.org/

27. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061–73.

28. Jakobsson M, Edge MD, Rosenberg NA. The Relationship Between F ST and the Frequency of the Most Frequent Allele. Genetics. 2013 Feb;193(2):515–28.

29. Murga-Moreno J, Coronado-Zamora M, Bodelón A, Barbadilla A, Casillas S. PopHumanScan: the online catalog of human genome adaptation. Nucleic Acids Res. 2019 Jan 8;47(D1):D1080–9.

30. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015 Feb 11;518(7538):197–206.

31. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823–8.

32. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative Annotation of Variants from 1092

Humans: Application to Cancer Genomics. Science (80-). 2013 Oct 4;342(6154):1235587.

33. Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov;24(11):1894–904.

34. Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 2013 May 1;23(5):749–61.

35. Narasimhan VM, Xue Y, Tyler-Smith C. Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? Trends Mol Med. 2016;22(4):341–51.

36. McCoy RC, Wakefield J, Akey JM. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell. 2017 Feb;168(5):916-927.e12.

37. Juric I, Aeschbacher S, Coop G. The Strength of Selection against Neanderthal Introgression. Reich D, editor. PLOS Genet. 2016 Nov 8;12(11):e1006340.

38. Mondal M, Bertranpetit J, Lao O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. Nat Commun. 2019 Dec 16;10(1):246.

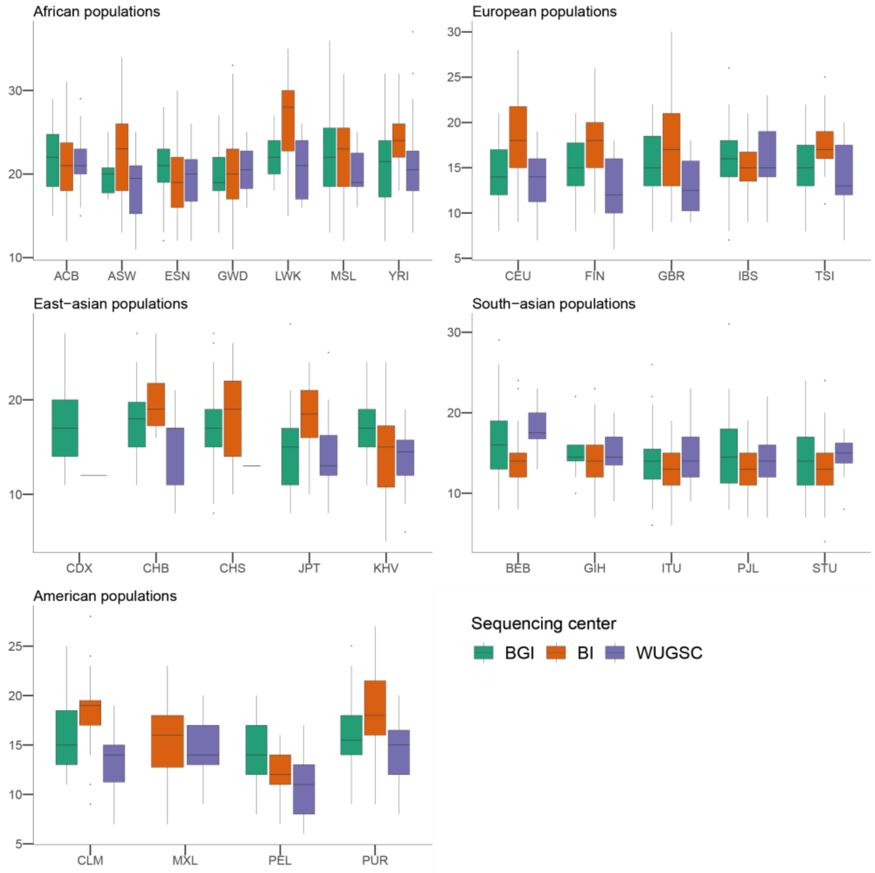


Figure 1 - Number of LoF variants by sequencing centres across continental groups and populations. Each panel corresponds to a continental group. In the x-axis is displayed the name of the population, in the y-axis is displayed the number of LoF snps per individual.

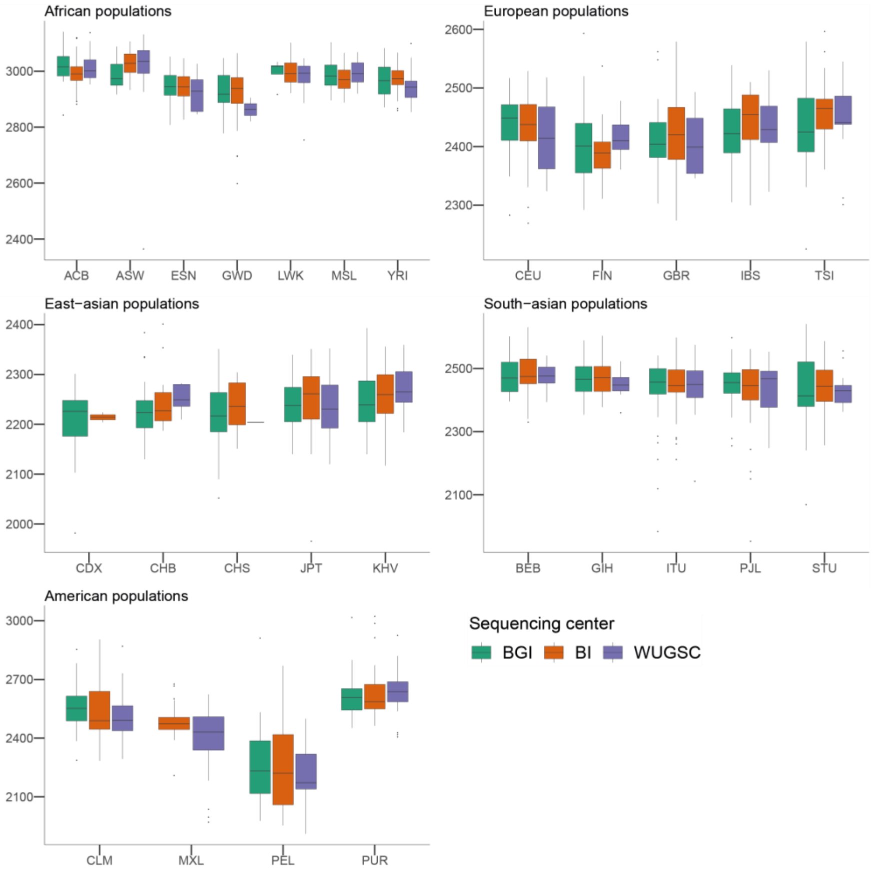


Figure 2 - Number of benign variants by sequencing centre across continental groups and populations. Each panel corresponds to a continental group. In the x-axis is displayed the name of the population, in the y-axis is displayed the number of LoF snps per individual.

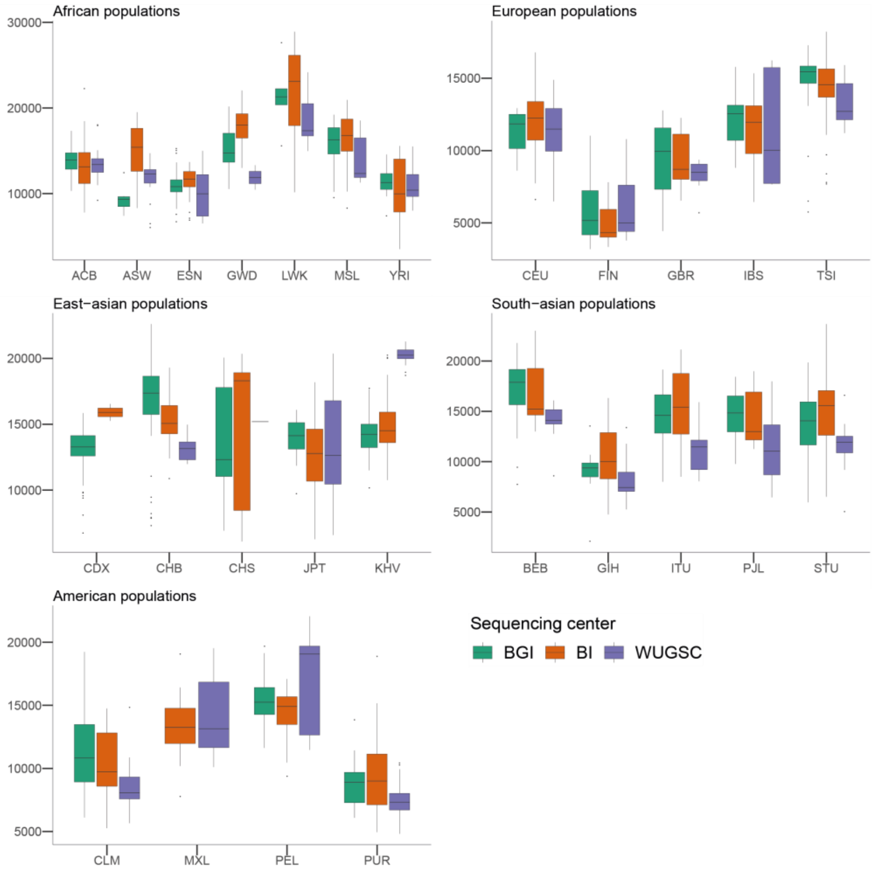


Figure 3 - Number of derived singletons by sequencing centres across continental groups and populations. Each panel corresponds to a continental group. In the x-axis is displayed the name of the population, in the y-axis is displayed the number of derived singletons per individual.

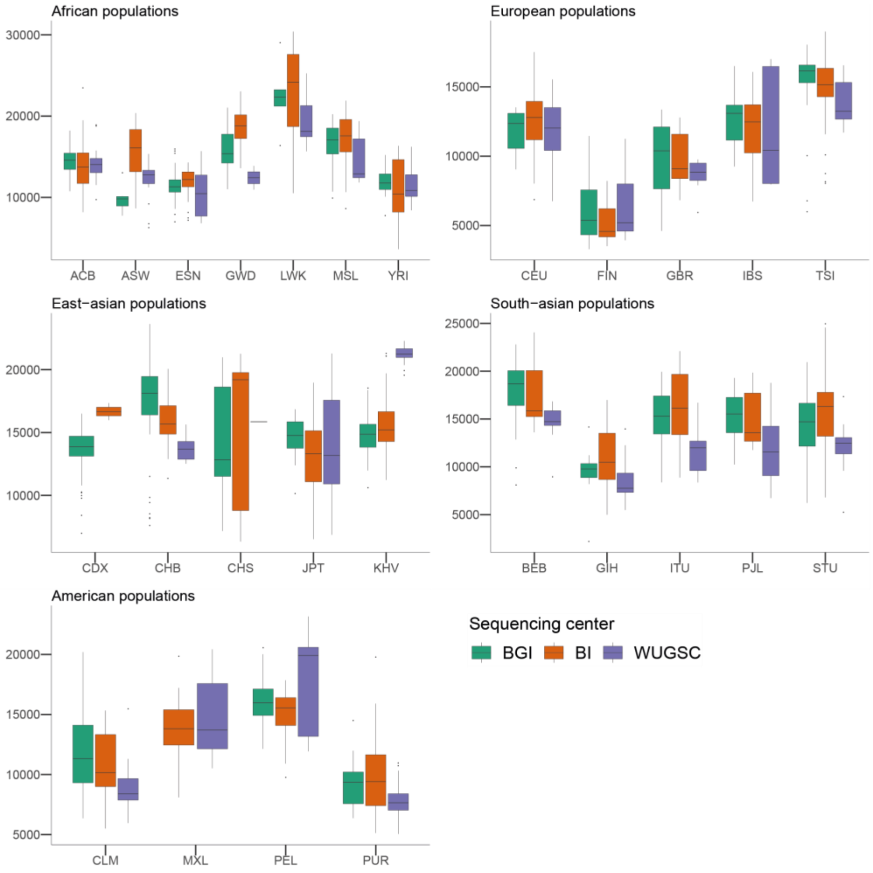


Figure 4 - Number of singletons by sequencing centres across continental groups and populations. Each panel corresponds to a continental group. In the x-axis is displayed the name of the population, in the y-axis is displayed the number of singletons per individual.

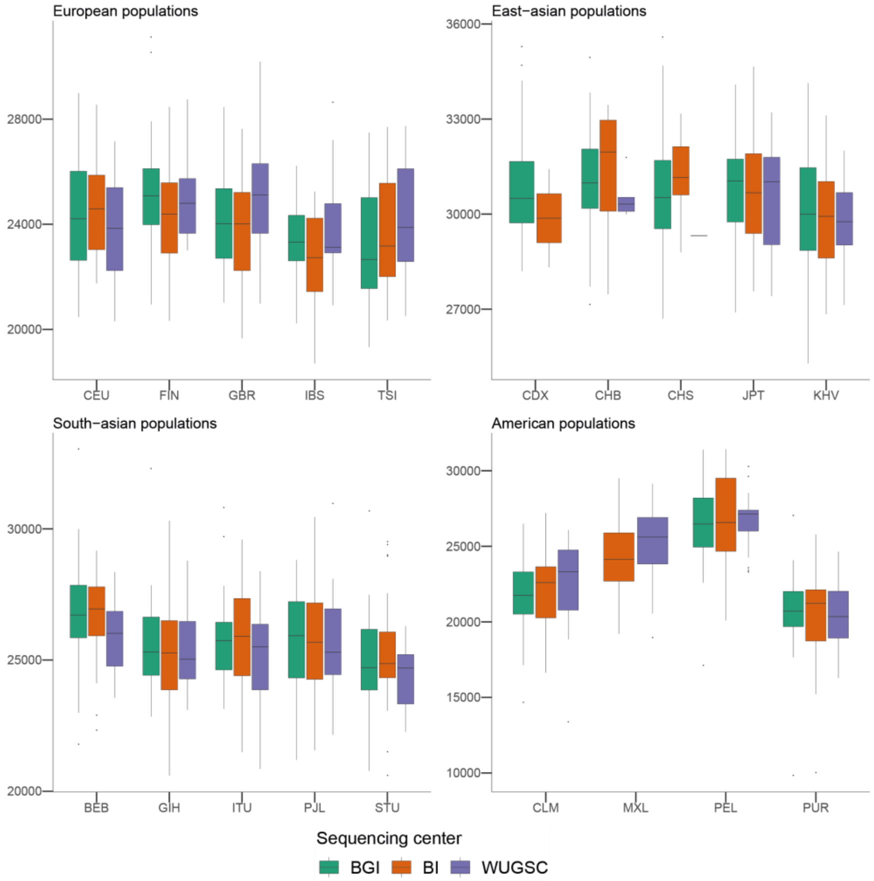


Figure 5 - Number of introgressed alleles by sequencing centre across continental groups and populations. Each panel corresponds to a continental group. In the x-axis is displayed the name of the population, in the y-axis is displayed the number of introgressed alleles as defined by Browning et al. 2018

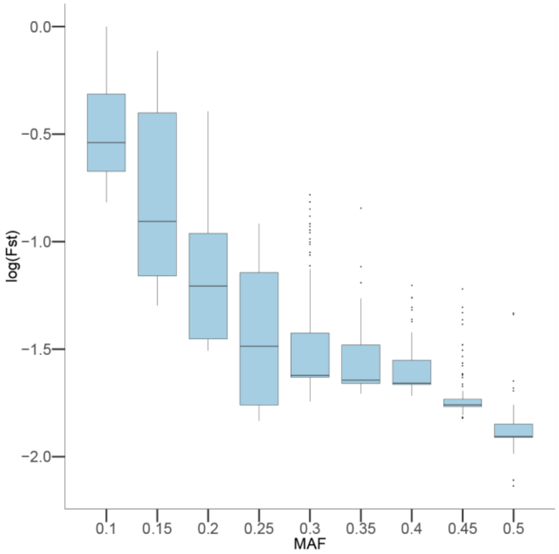


Figure 6 - log(Normalized Fst) values binned by MAF in the CEU population. In the x-axis are the different MAF bins, in the y-axis is logarithm of the normalized Fst value

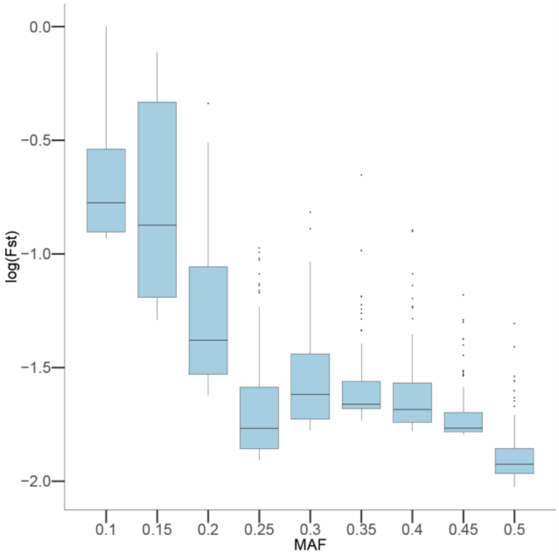


Figure 7 - log(Normalized Fst) values binned by MAF in the YRI population. In the x-axis are the different MAF bins, in the y-axis is logarithm of the normalized Fst value.

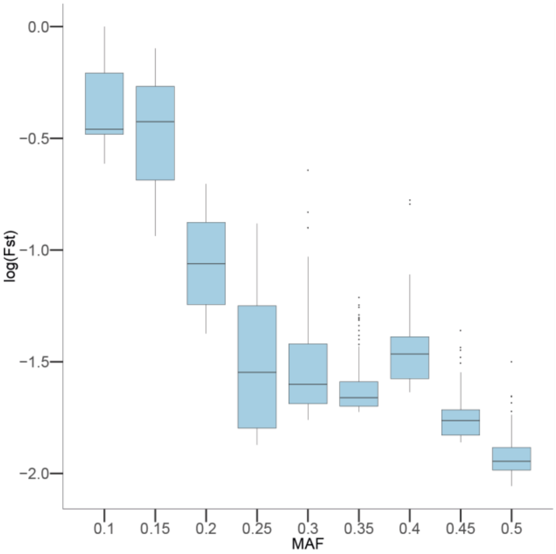


Figure 8 - log(Normalized Fst) values binned by MAF in the JPT population. In the x-axis are the different MAF bins, in the y-axis is logarithm of the normalized Fst value

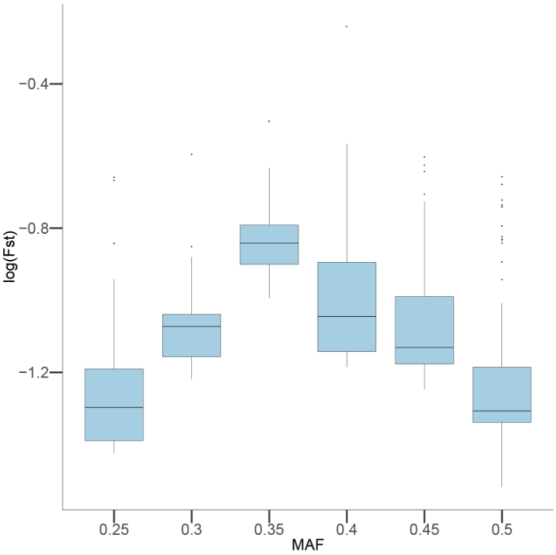


Figure 9 - log(Normalized Fst) values binned by MAF in the CHB population. In the x-axis are the different MAF bins, in the y-axis is logarithm of the normalized Fst value.

Supplementary Material

Supplementary Table 1. Results from the hierarchical mixed model using as dependent variable the number of LoF variants

	Estimate	Std. Error	df	t value
(Intercept)	2.660e+00	7.457e-02	4.274e+00	35.669
SeqCenterBGI	1.260e-01	1.643e-02	2.433e+03	7.668
SeqCenterBI	1.381e-01	1.677e-02	2.469e+03	8.232
SeqCenterWUGSC	4.120e-02	1.891e-02	2.493e+03	2.179

Supplementary Table 2. Results from the hierarchical mixed model using as dependent variable the number of benign variants.

	Estimate	Std. Error	df	t value
(Intercept)	7.812e+00	4.735e-02	4.029e+00	164.959
SeqCenterBGI	1.040e-02	2.146e-03	2.484e+03	4.849
SeqCenterBI	1.277e-02	2.185e-03	2.482e+03	5.845
SeqCenterWUGSC	8.206e-03	2.456e-03	2.480e+03	3.341

Supplementary Table 3. Results from the hierarchical mixed model using as dependent variable the number of derived singletons.

	Estimate	Std. Error	df	t value
(Intercept)	9.41341	0.06802	4.28727	138.388
SeqCenterBGI	0.02382	0.01551	2482.01039	1.536
SeqCenterBI	0.03280	0.01579	2481.35467	2.076
SeqCenterWUGSC	-0.06524	0.01775	2479.05950	-3.674

Supplementary Table 4. Results from the hierarchical mixed model using as dependent variable the number of singletons.

	Estimate	Std. Error	df	t value
(Intercept)	9.45823	0.06829	4.28888	138.508
SeqCenterBGI	0.02295	0.01558	2482.05989	1.473
SeqCenterBI	0.03116	0.01586	2481.39533	1.965
SeqCenterWUGSC	-0.06650	0.01783	2479.08794	-3.730

Supplementary Table 5. Results from the hierarchical mixed model using as dependent variable the log(number of introgressed alleles as defined by S. Browning, 2018.

	Estimate	Std. Error	df	t value
(Intercept)	1.015e+01	6.218e-02	3.033e+00	163.310
SeqCenterBGI	1.179e-03	6.009e-03	1.827e+03	0.196
SeqCenterBI	-2.372e-03	6.340e-03	1.824e+03	-0.374
SeqCenterWUGSC	1.784e-03	7.149e-03	1.825e+03	0.249

5 DISCUSSION

The limits of geographic detection

An almost universal conclusion from studying human genetic variation is that geography matters. Genetic differentiation tends to increase as the geographic distance of the individuals increases (Lao et al., 2008; C. Wang et al., 2012). This differentiation is quite modest compared to other organisms including close relative species such as Chimp (de Manuel et al., 2016). The main reason for this low amount of differentiation must be interpreted in terms of the recent (in evolutionary terms) spread of anatomically modern humans out of Africa (Nielsen et al., 2017). The amount of archaic introgression present in our genome (around 2% up to 5% in some Oceanian populations) is not enough to define strongly genetically stratified populations. However, the observed geographic dependence of genetic variation is due to the fact that most of the demographic processes that dominate humans depend on geography: migrations, isolations and regional selective processes. As a general model, until recently, one of the variables mostly influencing mating in human species was geographic proximity.

This raises some basic questions in human population genetics: which is the minimum unit of geographic differentiation that we can identify and how can we estimate it? It must be stressed that both parts of the question are not independent. Depending on the power we have to estimate population substructure, we will be able to identify putative hidden geographic structure. This power will depend on the assumptions of the model that we are using, and the type of data that we are considering.

During the last 10 years, human population data have been mostly microarray-based. This type of data, as previously explained in the introduction of this thesis (see section 1.4, subsection a), is based on genotyping genetic variants that have been previously defined in human populations. Consequently, this type of data is biased towards genetic variants that tend to be frequent in particular populations, not necessarily the ones we are considering in our particular case (Clark et al., 2005). Furthermore, if they are present in our samples, one can be expected that they are old enough to have spread over different populations, thus limiting the level of geographic resolution that one could observe. Therefore, it is not surprising that previous studies using microarray data and the genotypic information of SNPs in European populations can recover the global geographic structure (Lao et al., 2008; Novembre et al., 2008), but failed to identify more fine population differentiation (for example, see Lao et al., 2013).

Methods that start considering haplotype information, either in terms of IBD (S. R. Browning & Browning, 2012) or chromosome painting (Lawson et al., 2012) overcome this problem by using the higher recombination rate expected in the genome compared to the mutation rate (Bycroft et al., 2019; Byrne et al., 2020). This higher rate of generating genetic variation allows us to find relatively recent events which, according to the previously stated, relate to recent geographic events (Hellenthal et al., 2014). Therefore, deeper population substructure can be identified. However, since the SNP density in the microarray is defined a priori, the fine haplotype architecture is still ignored, and haplotype inference can also be biased. In fact, in our study using samples from SEP (Chapter 1 of results) we observed that fine population substructure cannot be identified when microarray based SNPs are used, even if we apply haplotype based methods. With microarray data we recovered the geographic profile that has been previously described when using classical markers.

In principle, all these putative reported problems are expected to be minimized, and higher population substructure detected, when using NGS data. Previous studies analysing the genetic variation in Sardinia using NGS data (Chiang et al., 2018) reported fine-scale variation in the ancient population ancestry proportions across the island, with the most remote and interior areas of Sardinia to have been the least exposed to contact with outside populations.

In my first work, I show that fine population substructure can be identified in the SEP populations separated by less than 140 km using unsupervised analyses such as fineSTRUCTURE, and that this substructure can be interpreted in geographic terms of near valleys. Such level of fine-population substructure has only been reported in some rural villages of Galicia using fineSTRUCTURE with microarray data (Bycroft et al., 2019), and supports the hypothesis that micro-population substructure due to the same demographic processes that determine the global pattern observed in Europe, mainly isolation by distance, exists.

The second part of the question, how we identify it, points to the need to develop new algorithms that allow the identification of fine population substructure, particularly in a geographic context. This field has been developing for more than 10 years now (Yang et al., 2012), reaching the start of art with software such as EEMS (Petkova et al., 2015), which attempts to estimate the migration surfaces given the expected allele frequencies and the genotypes observed in the geographically sampled individuals.

The identification of genetic barriers normally implies a difference between allele frequencies on both sides of the barrier. In this case, with the SEP populations being geographically close, this could pose a problem. In order to solve this problem, we decided to apply tools from a field that has encountered a similar problem before: geostatistics. We created an algorithm that uses a matrix of distances between individuals to find a genetic barrier between them. The principles of this algorithm can be found in classical works of Sokal and Owen (i.e. the The Bearing Correlogram M. S. Rosenberg, 2010). Using this framework, the genetic distance and the physical distance between two individuals can be related through an angle indicating the maximum amount of differentiation between all the pairs of comparisons. The model can be extended to account for isolation by distance processes using kriging principles (Bradburd et al., 2016). Then, for *K* geographic groups we identify a genetic barrier by identifying the set of sample coordinates that minimizes the goodness of fit of the parameters. We accomplish this by finding the mean sum of square error for each point between the real distance and the inferred new distance. To optimize this function we define geographic groups as a set of coordinates and assign each sample to a geographic group based on its proximity. To find this set of coordinates we use a genetic algorithm (Sergeeva et al., 2017). One of the virtues of this algorithm is that it can use any kind of genetic distance, not only the typical ones such as the Fst or identity-by-descent matrices.

Using this algorithm we found a genetic barrier between the SEP populations that was replicated by fineSTRUCTURE and EEMS.

The importance of studying rural areas for understanding complex global demographic patterns

The importance of rural areas in generating the general genomic landscape has been so far poorly studied in the European continent, despite they represented until recent times the main part of the population until the start of the 20th century (Champion, 2012). The Industrial Revolution shifted the population from rural areas to be concentrated in the cities. This depopulated the rural world and produced a trend of higher inbreeding and consanguinity due to the lack of potential marriage partners.

One of the main reasons for the low number of studies over rural areas is the lack of high quality genomes from samples of rural areas. For example, the Pyrenees, although it constitutes a physical border between the Iberian Peninsula and the rest of Europe, has been poorly studied. Up to our knowledge there has been a study using classical markers (Calafell & Bertranpetit, 1994), a study using immunoglobulin data (Giraldo et al., 2001) one using the Y-chromosome (López-Parra et al., 2009) and the most recent one that used microarrays (Biagini et al., 2019). None of these studies have been able to find the population structure that we have found in this work. One of the problems of these studies is the lack of high coverage whole genome sequenced samples. Also another advantage of using this kind of samples is the ability to ascertain LoF variants and return part of the knowledge we have acquired while working on this study.

In this work we have shown that genetics and demography are interconnected. We observed a trend regarding autozigosity that was already visible using individuals born in the first half of the 20th century. In this dataset we already see the first marks of isolation in terms of higher LD and longer RoHs compared to the general Spanish population. An example of how demography can shape the genetic variation of an area is the case of Berguedà. This region has the shortest RoHs, higher effective population size and lowest LoF diversity from all the studied SEP populations. Given that these patterns are at a microgeographic scale, a possible explanation for the different profile compared to the other studied populations could be the fact that Berguedà has not been as isolated. In particular, the creation of a railway in 1914 (Serra-Rotés, 2017) connected Berguedà with the most industrialized part of Catalonia. This railway could have meant an influx of migrants to this region due to the flourishing economy (based on natural resources and textile manufacturing).

The importance of batch effects in datasets generated at different times and centres

When analysing the demography of SEP, we noticed non-reproducible results when using 1,000 Genomes data (1000G) (The 1000 Genomes Project Consortium et al., 2015). In our particular case, we observed an unusually low amount of LoF variants in the SEP individuals compared to the IBS (Iberian Peninsula samples) and FIN (Finnish samples) from the 1,000G. The latest was used as a prototype of inbred population enriched for rare deleterious mutations (Kääriäinen et al., 2017). We did not find this trend regarding the benign alleles in these populations.

It has been proposed that highly inbred populations for long periods of time can purge the highly deleterious alleles (Y. Xue et al., 2015). Despite what we observed in our ABC-DL modelling that the effective population size of the SEP populations was reduced during the last ~2,700 years, it seemed unlikely to explain the pattern that was observed. Other putative biological bias of the SEP data was the age of the sampled individuals. The SEP sample consisted on old individuals as a part of the study design to avoid recent demographic events masking the inner structure of the rural area. Therefore, one could in principle expect that this population could be pruned of deleterious variants that would have prevented reaching such old age. However, this hypothesis could not explain why we also observed differences in the LoF composition within the SEP populations.

Taking into account that the 1000G data show a large degree of heterogeneity on the way it has been produced, we wondered whether the effects that we were observing were due to sequencing and bioinformatic artifacts. The second point was addressed by using the same bioinformatic pipeline as the one applied to generate our data using whole exome sequenced individuals from IBS and FIN. However, the trend was still present. The first point, which finally provided a plausible explanation, was tested by considering the sequencing centre that produced the samples of the 1000G data. It had been previously reported the presence of sequencing errors in the 1000G, mainly related to the sequencing centre and particularly in genes (Mafessoni, 2019). In the context of studying LoF, it could be expected that any sequencing error occurring at a gene is going to produce a damaging allele, thus enriching the dataset with ghost LoF mutations. Since we were considering micro-population substructure, such bias was large enough to produce the differences we observed between IBS and SEP. We excluded the SNVs that Maffessoni et al had identified as putatively sequencing errors but we still identified the same batch effect in the IBS and FIN samples.

After noticing this fact, we decided to study whether the effect we observed in IBS and FIN could be extended to other 1000G populations and in other statistics used in human population genetics that summarize events related to rare alleles. Although the precision of rare alleles in 1000G has already been questioned (Belsare et al., 2019), their work was more centred regarding the quality of imputing rare alleles.

Our analyses in the amount of singletons, derived singletons and LoF, support that the sequencing centre plays a role in the amount of alleles from these types that are observed in the 1000G individuals over all the populations (see Figures 1, 2, 3 and 4 in Chapter 2).

We also analysed the number of archaic introgressed alleles defined by SPrime (S. R. Browning et al., 2018). This statistic was ascertained because it has been claimed that the fingerprint of archaic introgression is rare in the human genome (Green et al., 2010), mainly due to the deleterious effect of the hybrid archaic-anatomically modern human (Dolgova & Lao, 2018). Since one of the ways to identify such introgressed fragments is based on analyzing tracks of derived alleles compared to reference non-archaic introgressed populations, we wondered whether batch effects due to the sequencing center could affect the estimates obtained in the 1000G individuals by (S. R. Browning et al., 2018). No clear pattern is observed by the sequencing center.

Nevertheless, particular heterogeneous trends were observed in some East Asian populations, such as CHB (see Figure 5 in Chapter 2). This is particularly interesting because this population has been claimed to have different waves of archaic introgressions (S. R. Browning et al., 2018).

We studied the properties of genetic markers showing strong deviations between sequencing centres. In this case we selected the YRI, CEU, JPT and CHB populations. When binning the extreme SNP outliers per MAF category these populations did not show the same pattern Fst pattern (after normalizing by the maximum Fst obtained in each MAF category). While YRI, CEU and JPT showed the same pattern (the normalized Fst value decreases with MAF) (see Figures 6, 7 and 8 in Chapter 2) we observed a different pattern regarding CHB. In the latter instead of a decreasing pattern we observed an inverted U-shape.

The shared ancestry of JPT and CHB s (as they are used together in The 1000 Genomes Project Consortium, 2010) suggests that this pattern is unlikely to be explained by ancestry alone. This particular pattern of CHB is more compatible with the way CHB samples have been processed compared to other populations.

When selecting those SNPs that showed the strongest biases in some populations (YRI, CEU, JPT, CHB) we observed large Fst values, even after normalizing inside each MAF bin. This is and indicative that the effect we observed in singletons is also patent regarding SNPs at higher frequency. A caveat of this analysis is that the number of SNps we used is low due to the filter we applied to obtain these extreme SNPs.

Our results of Chapter 2 show that the scientific community must use 1000G data with caution, especially when merging it with new dataset as the latter can have its private batch effects.

6 CONCLUSIONS

From Chapter 1 we conclude that:

1. Fine population substructure can be identified when using WGS.
2. Particular demographic histories exist even at populations <140 km separated from the Pyrenees.
3. Geography shaped the genetic relationships of the rural areas of the Pyrenees.
4. SEP populations show particular markers associated with genetic isolates but not all of them.
5. Further analysis adding samples from other geographic regions of the Pyrenees are required to study if this pattern extends to the whole Pyrenees.

From Chapter 2 we conclude that:

1. The 1,000 Genomes Project (1000G) is a study used in multitude of projects either as a general population dataset or to perform imputation to complete microarray datasets.
2. The 1000G dataset has been generated with the collaboration of nine different sequencing centers, generating a possible batch effect.
3. The 1000G dataset has quality problems regarding rare variants due to the use of low coverage whole genome sequencing.
4. We have found proof of a batch effect regarding the sequencing centre in most of the cases.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, *27*(11), 2534–2547. https://doi.org/10.1093/molbev/msq148

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P. Della, Dąbrowski, P., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*(7555), 167–172. https://doi.org/10.1038/nature14507

Anderson-Trocmé, L., Farouni, R., Bourgey, M., Kamatani, Y., Higasa, K., Seo, J.-S., Kim, C., Matsuda, F., & Gravel, S. (2020). Legacy Data Confound Genomics Studies. *Molecular Biology and Evolution*, *37*(1), 2–10. https://doi.org/10.1093/molbev/msz201

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, *3*(4), 299–309. https://doi.org/10.1038/nrg777

Asouti, E., & Fuller, D. Q. (2013). A contextual approach to the emergence of agriculture in southwest Asia: Reconstructing early neolithic plant-food production. *Current Anthropology*, *54*(3), 299–345. https://doi.org/10.1086/670679

Auton, A., Bryc, K., Boyko, A. R., Lohmueller, K. E., Novembre, J., Reynolds, A., Indap, A., Wright, M. H., Degenhardt, J. D., Gutenkunst, R. N., King, K. S., Nelson, M. R., & Bustamante, C. D. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Research*, *19*(5), 795–803. https://doi.org/10.1101/gr.088898.108

Bamshad, M., & Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics*, *4*(2), 99–110. https://doi.org/10.1038/nrg999

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Behar, D. M., Garrigan, D., Kaplan, M. E., Mobasher, Z., Rosengarten, D., Karafet, T. M., Quintana-Murci, L., Ostrer, H., Skorecki, K., & Hammer, M. F. (2004). Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. *Human Genetics*, *114*(4), 354–365. https://doi.org/10.1007/s00439-003-1073-7

Belsare, S., Levy-Sakin, M., Mostovoy, Y., Durinck, S., Chaudhuri, S., Xiao, M., Peterson, A. S., Kwok, P.-Y., Seshagiri, S., & Wall, J. D. (2019). Evaluating the quality of the 1000 genomes project data. *BMC*

Genomics, *20*(1), 620. https://doi.org/10.1186/s12864-019-5957-x

Benazzi, S., Douka, K., Fornai, C., Bauer, C. C., Kullmer, O., Svoboda, J., Pap, I., Mallegni, F., Bayle, P., Coquerelle, M., Condemi, S., Ronchitelli, A., Harvati, K., & Weber, G. W. (2011). Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature*, *479*(7374), 525–528. https://doi.org/10.1038/nature10617

Biagini, S. A., Solé-Morata, N., Matisoo-Smith, E., Zalloua, P., Comas, D., & Calafell, F. (2019). People from Ibiza: an unexpected isolate in the Western Mediterranean. *European Journal of Human Genetics*, *27*, 941–951. https://doi.org/10.1038/s41431-019-0361-1

Bloom, L. B., Goodman, M. F., Otto, M. R., Beechem, J. M., Eritja, R., & Reha-Krantz, L. J. (1994). Pre-Steady-State Kinetic Analysis of Sequence-Dependent Nucleotide Excision by the 3′-Exonuclease Activity of Bacteriophage T4 DNA Polymerase. *Biochemistry*, *33*(24), 7576–7586. https://doi.org/10.1021/bi00190a010

Brabazon, A., O'Neill, M., & McGarraghy, S. (2015). *Natural Computing Algorithms*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43631-8

Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2016). A Spatial Framework for Understanding Population Structure and Admixture. *PLoS Genetics*, *12*(1), e1005703. https://doi.org/10.1371/journal.pgen.1005703

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., & Weber, J. L. (1998). Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *American Journal of Human Genetics*, *63*(3), 861–869. https://doi.org/10.1086/302011

Brown, D. L. (2012). Migration and Rural Population Change: Comparative Views in More Developed Nations. In L. J. Kulcsár & K. J. Curtis (Eds.), *International Handbook of Rural Demography* (1st ed., pp. 35–48). Springer Netherlands. https://doi.org/10.1007/978-94-007-1842-5_4

Browning, S. (2018). *Sprime results for 1000 Genomes non-African populations and SGDP Papuans*.

Browning, S. R., & Browning, B. L. (2012). Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*, *46*(1), 617–633. https://doi.org/10.1146/annurev-genet-110711-155534

Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., & Akey, J. M. (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, *173*(1), 53-61.e9. https://doi.org/10.1016/j.cell.2018.02.031

Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C., Quintela, I., Carracedo, Á., Donnelly, P., & Myers, S. (2019). Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-018-08272-w

Byrne, R. P., van Rheenen, W., van den Berg, L. H., Veldink, J. H., & McLaughlin, R. L. (2020). Dutch population structure across space, time and GWAS design. *Nature Communications*, *11*(1), 4556.

https://doi.org/10.1038/s41467-020-18418-4

Calafell, F., & Bertranpetit, J. (1994). Mountains and genes: Population history of the Pyrenees. *Human Biology*, *66*(5), 823–842.

Calderón, R., Hernández, C. L., García-Varela, G., Masciarelli, D., & Cuesta, P. (2018). Inbreeding in Southeastern Spain. *Human Nature*, *29*(1), 45–64. https://doi.org/10.1007/s12110-017-9305-z

Campbell, M. C., Smith, L. T., & Harvey, J. (2019). Population genetic evidence for positive and purifying selection acting at the human IFN-γ locus in Africa. *Genes & Immunity*, *20*(2), 143–157. https://doi.org/10.1038/s41435-018-0016-1

Cao, A., & Galanello, R. (2010). Beta-thalassemia. In *Genetics in Medicine* (Vol. 12, Issue 2, pp. 61–76). Nature Publishing Group. https://doi.org/10.1097/GIM.0b013e3181cd68ed

Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S., & Siepel, A. (2013). A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. *PLoS Genetics*, *9*(8), 1003684. https://doi.org/10.1371/journal.pgen.1003684

Cavalli-Sforza, L. L., & Bodmer, W. F. (2013). *The Genetics of Human Populations Revised ed. Edition*. Dover Publications.

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1996). *The History and Geography of Human Genes*. Princeton University Press.

Champion, A. (2012). Europe's Rural Demography. In L. J. Kulcsár & K. J. Curtis (Eds.), *International Handbook of Rural Demography* (1st ed., pp. 81–93). Springer Netherlands. https://doi.org/10.1007/978-94-

007-1842-5_7

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*(4), 1289–1303.

Chiang, C. W. K., Marcus, J. H., Sidore, C., Biddanda, A., Al-Asadi, H., Zoledziewska, M., Pitzalis, M., Busonero, F., Maschio, A., Pistis, G., Steri, M., Angius, A., Lohmueller, K. E., Abecasis, G. R., Schlessinger, D., Cucca, F., & Novembre, J. (2018). Genomic history of the Sardinian population. *Nature Genetics*, *50*(10), 1426–1434. https://doi.org/10.1038/s41588-018-0215-8

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, *3*(MAR). https://doi.org/10.3389/fgene.2012.00035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, *15*(11), 1496–1502. https://doi.org/10.1101/gr.4107905

Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., Garrison, E., Xue, Y., & Tyler-Smith, C. (2014). Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology*, *15*(6), R88. https://doi.org/10.1186/gb-2014-15-6-r88

Conrad, D. F., Keebler, J. E. M., Depristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., Awadalla, P., & Donald F Conrad, Jonathan E M Keebler, Mark A DePristo, Sarah J Lindsay, Yujun Zhang, Ferran Casals, Youssef Idaghdour, Chris L Hartl, Carlos Torroja, Kiran V Garimella, Martine Zilversmit, Reed Cartwright, Guy A Rouleau, Mark Daly, Eric A Stone, M. E. H. & P. A. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, *43*(7), 712–714. https://doi.org/10.1038/ng.862

Coulondre, C., Miller, J. H., Farabaugh, P. J., & Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, *274*(5673), 775–780. https://doi.org/10.1038/274775a0

Cunliffe, B. (2013). *Britain Begins*. Oxford University Press.

Darwin, C., & Wallace, A. (1858). On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, *3*(9), 45–62. https://doi.org/10.1111/j.1096-3642.1858.tb02500.x

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. https://doi.org/10.1126/science.aag2602

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., & Marchini, J. (2013). Haplotype Estimation Using Sequencing Reads. *The American Journal of Human Genetics*, *93*(4), 687–696. https://doi.org/10.1016/j.ajhg.2013.09.002

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501. https://doi.org/10.1038/ng.806

Dolgova, O., & Lao, O. (2018). Evolutionary and Medical Consequences of Archaic Introgression into Modern Human Genomes. *Genes*, *9*(7), 358. https://doi.org/10.3390/genes9070358

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125–2137. https://doi.org/10.1093/hmg/ddu733

Dopazo, J., Amadoz, A., Bleda, M., Garcia-Alonso, L., Alemán, A., García-García, F., Rodriguez, J. A., Daub, J. T., Muntané, G., Rueda, A., Vela-

Boza, A., López-Domingo, F. J., Florido, J. P., Arce, P., Ruiz-Ferrer, M., Méndez-Vidal, C., Arnold, T. E., Spleiss, O., Alvarez-Tejado, M., … Antiñolo, G. (2016). 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Molecular Biology and Evolution*, *33*(5), 1205–1218. https://doi.org/10.1093/molbev/msw005

Dumont, B. L., & Payseur, B. A. (2008). Evolution of the genomic rate of recombination in mammals. *Evolution*, *62*(2), 276–294. https://doi.org/10.1111/j.1558-5646.2007.00278.x

Duncan, B. K., & Miller, J. H. (1980). Mutagenic deamination of cytosine residues in DNA. *Nature*, *287*(5782), 560–561. https://doi.org/10.1038/287560a0

Dupanloup, I., Schneider, S., & Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology*, *11*(12), 2571–2581. https://doi.org/10.1046/j.1365-294X.2002.01650.x

Duret, L., & Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*, *10*(1), 285–311. https://doi.org/10.1146/annurev-genom-082908-150001

Eldridge, M. D. B., King, J. M., Loupis, A. K., Spencer, P. B. S., Taylor, A. C., Pope, L. C., & Hall, G. P. (1999). Unprecedented Low Levels of Genetic Variation and Inbreeding Depression in an Island Population of the Black-Footed Rock-Wallaby. *Conservation Biology*, *13*(3), 531–541. https://doi.org/10.1046/j.1523-1739.1999.98115.x

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, *9*(10), 1003905. https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics*, *131*(2), 479–491.

Fan, S., Hansen, M. E. B., Lo, Y., & Tishkoff, S. A. (2016). Going global by adapting local: A review of recent human adaptation. In *Science* (Vol. 354, Issue 6308, pp. 54–59). American Association for the Advancement of Science. https://doi.org/10.1126/science.aaf5098

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics*, *155*(3), 1405–1413.

Fenner, J. N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, *128*(2), 415–423. https://doi.org/10.1002/ajpa.20188

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., … Searle, S. M. J. (2014). Ensembl 2014. *Nucleic Acids Research*, *42*(D1), D749–D755. https://doi.org/10.1093/nar/gkt1196

Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Van Duijn, C. M., Swertz, M., Wijmenga, C., Van Ommen, G., Slagboom, P. E., Boomsma, D. I., Ye, K., Guryev, V., Arndt, P. F., Kloosterman, W. P., De Bakker, P. I. W., & Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, *47*(7), 822–826. https://doi.org/10.1038/ng.3292

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Fryxell, K. J., & Zuckerkandl, E. (2000). Cytosine Deamination Plays a Primary Role in the Evolution of Mammalian Isochores. *Molecular Biology and Evolution*, *17*(9), 1371–1383. https://doi.org/10.1093/oxfordjournals.molbev.a026420

Fullerton, S. M., Bernardo Carvalho, A., & Clark, A. G. (2001). Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. *Molecular Biology and Evolution*, *18*(6), 1139–1142. https://doi.org/10.1093/oxfordjournals.molbev.a003886

Fuster, V., & Colantonio, S. E. (2003). Inbreeding coefficients and degree of consanguineous marriages in Spain: A review. *American Journal of Human Biology*, *15*(5), 709–716. https://doi.org/10.1002/ajhb.10198

Fuster, V., & Colantonio, S. E. (2004). Socioeconomic, demographic, and geographic variables affecting the diverse degrees of consanguineous marriages in Spain. *Human Biology*, *76*(1), 1–14.

https://doi.org/10.1353/hub.2004.0021

Gamella, J., & Núñez-Negrillo, A. M. (2019). The Evolution of Consanguineous Marriages in the Archdiocese of Granada, Spain (1900–1979). *Human Biology*, *90*(2).

Ganai, R. A., & Johansson, E. (2016). DNA Replication—A Matter of Fidelity. *Molecular Cell*, *62*(5), 745–755. https://doi.org/10.1016/j.molcel.2016.05.003

Giraldo, M. P., Eesteban, E., Aluja, M. P., Nogués, R. M., Backés-Duró, C., Dugoujon, J. M., & Moral, P. (2001). Gm and Km alleles in two Spanish Pyrenean populations (Andorra and Pallars Sobirà): a review of Gm variation in the Western Mediterranean basin. *Annals of Human Genetics*, *65*(6), 537–548. https://doi.org/10.1017/S0003480001008880

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Paabo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Griffiths, A. J., Gelbart, W. M., Miller, J. H., & Lewontin., R. C. (1999). Chapter 5. Recombination of Genes - Mitotic Crossing-over. In *Modern Genetic Analysis*. W.H.Freeman & Co Ltd.

Griffiths, R. C. (1991). *The Two-Locus Ancestral Graph* (pp. 100–117). https://doi.org/10.1214/lnms/1215459289

Günther, T., & Jakobsson, M. (2016). Genes mirror migrations and cultures in prehistoric Europe — a population genomic perspective. In *Current Opinion in Genetics and Development* (Vol. 41, pp. 115–123). Elsevier Ltd. https://doi.org/10.1016/j.gde.2016.09.004

Günther, T., Valdiosera, C., Malmström, H., Ureña, I., Rodriguez-Varela, R., Sverrisdóttir, Ó. O., Daskalaki, E. A., Skoglund, P., Naidoo, T., Svensson, E. M., De Castro, J. M. B., Carbonell, E., Dunn, M., Storå, J., Iriarte, E., Arsuaga, J. L., Carretero, J. M., Götherström, A., Jakobsson, M., & Willerslev, E. (2015). Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(38), 11917–11922. https://doi.org/10.1073/pnas.1509851112

Häggström, M. (n.d.). *Conversion and crossover*. https://en.wikipedia.org/wiki/File:Conversion_and_crossover.jpg

Hamblin, M. T., & Di Rienzo, A. (2000). Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. *The American Journal of Human Genetics*, *66*(5), 1669–1679. https://doi.org/10.1086/302879

Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *ELife*, *6*. https://doi.org/10.7554/eLife.24284

Hartl, D. L. (1980). *Principles of Population Genetics*. Sinauer Associates.

Hassold, T., & Hunt, P. (2001). To err (meiotically) is human: The genesis of human aneuploidy. In *Nature Reviews Genetics* (Vol. 2, Issue 4, pp. 280–291). Nature Publishing Group. https://doi.org/10.1038/35066065

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518

Higham, T., Douka, K., Wood, R., Ramsey, C. B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., Barroso-Ruíz, C., Bergman, C., Boitard, C., Boscato, P., Caparrós, M., Conard, N. J., Draily, C., Froment, A., Galván, B., Gambassini, P., … Jacobi, R. (2014). The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, *512*(7514), 306–309. https://doi.org/10.1038/nature13621

Hoban, S., Bertorelle, G., & Gaggiotti, O. E. (2012). Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, *13*(2), 110–122. https://doi.org/10.1038/nrg3130

Howrigan, D. P., Simonson, M. A., & Keller, M. C. (2011). Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics*, *12*(1), 460. https://doi.org/10.1186/1471-2164-12-460

Huang, Y.-F., & Siepel, A. (2019). Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *Genome Research*, *29*(8), 1310–1321. https://doi.org/10.1101/gr.245522.118

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Hussels, I. E., & Morton, N. E. (1972). Pingelap and Mokil Atolls: achromatopsia. *American Journal of Human Genetics*, *24*(3), 304–309.

Ilardo, M. A., Moltke, I., Korneliussen, T. S., Cheng, J., Stern, A. J., Racimo, F., de Barros Damgaard, P., Sikora, M., Seguin-Orlando, A., Rasmussen, S., van den Munckhof, I. C. L., ter Horst, R., Joosten, L. A. B., Netea, M. G., Salingkat, S., Nielsen, R., & Willerslev, E. (2018). Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, *173*(3), 569-580.e15. https://doi.org/10.1016/j.cell.2018.03.054

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., Van De Leemput, J., Rafferty, I., … Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, *451*(7181), 998–1003. https://doi.org/10.1038/nature06742

Jonsta247. (n.d.). *Different Types of Mutations*. https://commons.wikimedia.org/wiki/File:Different_Types_of_Mutations.png

Kääriäinen, H., Muilu, J., Perola, M., & Kristiansson, K. (2017). Genetics in an isolated population like Finland: a different basis for genomic medicine? *Journal of Community Genetics*, *8*(4), 319–326. https://doi.org/10.1007/s12687-017-0318-4

Keeney, S., Giroux, C. N., & Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, *88*(3), 375–384. https://doi.org/10.1016/S0092-8674(00)81876-0

Kelso, J., & Prüfer, K. (2014). Ancient humans and the origin of modern humans. *Current Opinion in Genetics & Development*, *29*, 133–138. https://doi.org/10.1016/j.gde.2014.09.004

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, *19*(A), 27–43. https://doi.org/10.2307/3213548

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Kong, A., Barnard, J., Gudbjartsson, D. F., Thorleifsson, G., Jonsdottir, G., Sigurdardottir, S., Richardsson, B., Jonsdottir, J., Thorgeirsson, T., Frigge, M. L., Lamb, N. E., Sherman, S., Gulcher, J. R., & Stefansson, K. (2004). Recombination rate and reproductive success in humans. *Nature Genetics*, *36*(11), 1203–1206. https://doi.org/10.1038/ng1445

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S. W., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., … Stefansson, K. (2012). Rate of de novo mutations and the importance of father-s age to disease risk. *Nature*, *488*(7412), 471–475. https://doi.org/10.1038/nature11396

Krueger, J. (n.d.). *Gene flow final*. https://en.wikipedia.org/wiki/File:Gene_flow_final.png

Kulcsár, L. J., & Curtis, K. J. (2012). Why Does Rural Demography Still Matter? In L. J. Kulcsár & K. J. Curtis (Eds.), *International Handbook of Rural Demography* (1st ed., pp. 1–6). Springer Netherlands. https://doi.org/10.1007/978-94-007-1842-5_1

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, *37*(13), 4181–4193. https://doi.org/10.1093/nar/gkp552

Lao, O., Altena, E., Becker, C., Brauer, S., Kraaijenbrink, T., van Oven, M., Nürnberg, P., de Knijff, P., & Kayser, M. (2013). Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history. *Investigative Genetics*, *4*(1), 9. https://doi.org/10.1186/2041-2223-4-9

Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., … Kayser, M. (2008). Correlation between Genetic and Geographic Structure in Europe.

Current Biology, *18*(16), 1241–1248. https://doi.org/10.1016/j.cub.2008.07.049

Lao, O., & van Oven, M. (2015). Genetic Admixture. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 887–897). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.82054-1

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, *8*(1), 1002453. https://doi.org/10.1371/journal.pgen.1002453

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 3258. https://doi.org/10.1038/s41467-018-05257-7

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733–739. https://doi.org/10.1038/nrg2825

Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (3rd Editio). Elsevier Ltd.

Lenzi, M. L., Smith, J., Snowden, T., Kim, M., Fishel, R., Poulos, B. K., & Cohen, P. E. (2005). Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis I in human oocytes. *American Journal of Human Genetics*, *76*(1), 112–127. https://doi.org/10.1086/427268

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., & Bodmer, W. (2015). The fine-scale genetic structure of the British population. *Nature*, *519*(7543), 309–314. https://doi.org/10.1038/nature14230

Levy, S. E., & Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, *17*(1), 95–115. https://doi.org/10.1146/annurev-genom-083115-022413

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Heng. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, *319*(5866), 1100–1104. https://doi.org/10.1126/science.1153717

Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, *118*, 117–131. https://doi.org/10.1007/s11258-006-9126-3

Liu, X., Jian, X., & Boerwinkle, E. (2013). dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation*, *34*(9), E2393–E2402. https://doi.org/10.1002/humu.22376

López-Parra, A. M., Gusmão, L., Tavares, L., Baeza, C., Amorim, A., Mesa, M. S., Prata, M. J., & Arroyo-Pardo, E. (2009). In search of the Pre- and Post-Neolithic Genetic Substrates in Iberia: Evidence from Y-Chromosome in Pyrenean Populations. *Annals of Human Genetics*, *73*(1), 42–53. https://doi.org/10.1111/j.1469-1809.2008.00478.x

Lorente-Galdos, B., Lao, O., Serra-Vidal, G., Santpere, G., Kuderna, L. F. K. K., Arauna, L. R., Fadhlaoui-Zid, K., Pimenoff, V. N., Soodyall, H., Zalloua, P., Marques-Bonet, T., & Comas, D. (2019). Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. *Genome Biology*, *20*(1), 77. https://doi.org/10.1186/s13059-019-1684-5

Luiz, O. J., Madin, J. S., Robertson, D. R., Rocha, L. A., Wirtz, P., & Floeter, S. R. (2012). Ecological traits influencing range expansion across large oceanic dispersal barriers: insights from tropical Atlantic reef fishes. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1730), 1033–1040. https://doi.org/10.1098/rspb.2011.1525

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H., Zhao, C., Elloumi, F., Shi, W., Thomas, R., Lin, S., Tillinghast, G., Liu, G., Zhou, Y., Herman, D., … Zhang, J. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-

II microarray gene expression data. *Pharmacogenomics Journal*, *10*(4), 278–291. https://doi.org/10.1038/tpj.2010.57

Luo, S., Valencia, C. A., Zhang, J., Lee, N.-C., Slone, J., Gui, B., Wang, X., Li, Z., Dell, S., Brown, J., Chen, S. M., Chien, Y.-H., Hwu, W.-L., Fan, P.-C., Wong, L.-J., Atwal, P. S., & Huang, T. (2018). Biparental Inheritance of Mitochondrial DNA in Humans. *Proceedings of the National Academy of Sciences*, *115*(51), 13039–13044. https://doi.org/10.1073/pnas.1810946115

Mafessoni, F. (2019). Encounters with archaic hominins. *Nature Ecology & Evolution*, *3*(1), 14–15. https://doi.org/10.1038/s41559-018-0729-6

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Manni, F., Guerard`, E., Heyer, E., Guerard, E., & Heyer, E. (2004). Geographic Patterns of (Genetic, Morphologic, Linguistic) Variation: How Barriers Can Be Detected by Using Monmonier’s Algorithm.

Human Biology, *76*(2), 173–190. https://doi.org/10.1353/hub.2004.0034

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. https://doi.org/10.1146/annurev-anchem-062012-092628

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Marrosu, M. G., Motzo, C., Murru, R., Lampis, R., Costa, G., Zavattari, P., Contu, D., Fadda, E., Cocco, E., & Cucca, F. (2004). The co-inheritance of type 1 diabetes and multiple sclerosis in Sardinia cannot be explained by genotype variation in the HLA region alone. *Human Molecular Genetics*, *13*(23), 2919–2924. https://doi.org/10.1093/hmg/ddh319

Martín-Còlliga, A., & Vaquer, J. (1995). El poblament dels Pirienus a l'Holocè, del mesolític a l'edat del bronze. In E. Vives & J. Bertranpetit (Eds.), *Muntanys i Població. El passat dels Pirineus des d'una perspectiva multidisciplinària* (pp. 35–73).

McCoy, R. C., Wakefield, J., & Akey, J. M. (2017). Impacts of Neanderthal-

Introgressed Sequences on the Landscape of Human Gene Expression. *Cell*, *168*(5), 916-927.e12. https://doi.org/10.1016/j.cell.2017.01.038

Mccullough, J. M., & O'Rourke, D. H. (1986). Geographic distribution of consanguinity in europe. *Annals of Human Biology*, *13*(4), 359–367. https://doi.org/10.1080/03014468600008541

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science*, *201*(4358), 786–792. https://doi.org/10.1126/science.356262

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prufer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Paabo, S. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Mondal, M., Bertranpetit, J., & Lao, O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, *10*(1), 246. https://doi.org/10.1038/s41467-018-08089-7

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, *54*(1), 60–71. https://doi.org/10.1017/S0305004100033193

Muyas, F., Bosio, M., Puig, A., Susak, H., Domènech, L., Escaramis, G., Zapata, L., Demidov, G., Estivill, X., Rabionet, R., & Ossowski, S. (2019). Allele balance bias identifies systematic genotyping errors and false disease associations. *Human Mutation*, *40*(1), 115–126. https://doi.org/10.1002/humu.23674

Narasimhan, V. M., Xue, Y., & Tyler-Smith, C. (2016). Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? *Trends in Molecular Medicine*, *22*(4), 341–351. https://doi.org/10.1016/j.molmed.2016.02.006

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302–310. https://doi.org/10.1038/nature21347

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. In *Nature Reviews Genetics* (Vol. 12, Issue 6, pp. 443–451). Nature Publishing Group. https://doi.org/10.1038/nrg2986

Nordborg, M. (1998). On the probability of Neanderthal ancestry [7]. In *American Journal of Human Genetics* (Vol. 63, Issue 4, pp. 1237–1240). University of Chicago Press. https://doi.org/10.1086/302052

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., &

Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, *40*(5), 646–649. https://doi.org/10.1038/ng.139

Nutile, T., Ruggiero, D., Herzig, A. F., Tirozzi, A., Nappo, S., Sorice, R., Marangio, F., Bellenguez, C., Leutenegger, A. L., & Ciullo, M. (2019). Whole-Exome Sequencing in the Isolated Populations of Cilento from South Italy. *Scientific Reports*, *9*(1), 4059. https://doi.org/10.1038/s41598-019-41022-6

Ostrer, H., & Skorecki, K. (2013). The population genetics of the Jewish people. *Human Genetics*, *132*(2), 119–127. https://doi.org/10.1007/s00439-012-1235-6

Papadimitriou, S., Gazzo, A., Versbraegen, N., Nachtegael, C., Aerts, J., Moreau, Y., Van Dooren, S., Nowé, A., Smits, G., & Lenaerts, T. (2019). Predicting disease-causing variant combinations. *Proceedings of the National Academy of Sciences*, 201815601. https://doi.org/10.1073/pnas.1815601116

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient Admixture in Human History. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pavlidis, P., & Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research-Thessaloniki*, *24*(1), 7. https://doi.org/10.1186/s40709-017-0064-0

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics*, *91*(2), 275–292. https://doi.org/10.1016/j.ajhg.2012.06.014

Petkova, D., Novembre, J., & Stephens, M. (2015). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*(1), 94–100. https://doi.org/10.1038/ng.3464

Petr, M., Pääbo, S., Kelso, J., & Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences*, *116*(5), 1639–1644. https://doi.org/10.1073/pnas.1814338116

Petruska, J., & Goodman, M. F. (1985). Influence of Neighboring Bases on DNA Polymerase Insertion and Proofreading Fidelity". *The Journal of Biological Chemistry*, *260*(12), 7553–7539.

Pfeifer, G. P., You, Y. H., & Besaratinia, A. (2005). Mutations induced by ultraviolet light. In *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* (Vol. 571, Issues 1-2 SPEC. ISS., pp. 19–31). Elsevier. https://doi.org/10.1016/j.mrfmmm.2004.06.057

Pinilla, V., & Sáez, L. A. (2017). *Rural depopulation in Spain: Genesis of a problem and innovative policies*.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordõez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., … Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, *463*(7278), 191–196. https://doi.org/10.1038/nature08658

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *ELife*, *7*. https://doi.org/10.7554/eLife.36317

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. https://doi.org/10.1038/nature12886

Pugliatti, M., Rosati, G., Carton, H., Riise, T., Drulovic, J., Vécsei, L., & Milanov, I. (2006). The epidemiology of multiple sclerosis in Europe. In *European Journal of Neurology* (Vol. 13, Issue 7, pp. 700–722). Blackwell Publishing Ltd. https://doi.org/10.1111/j.1468-1331.2006.01342.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., Bertranpetit, J., & Engelken, J. (2015). Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, btv493. https://doi.org/10.1093/bioinformatics/btv493

Qiong, L., Zhang, W., Wang, H., Zeng, L., Birks, H. J. B., & Zhong, Y. (2017). Testing the effect of the Himalayan mountains as a physical barrier to gene flow in Hippophae tibetana Schlect. (Elaeagnaceae). *PLOS ONE*, *12*(5), e0172948. https://doi.org/10.1371/journal.pone.0172948

R: The R Project for Statistical Computing. (n.d.). Retrieved December 18, 2020, from https://www.r-project.org/

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, *16*(6), 359–371. https://doi.org/10.1038/nrg3936

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, *102*(44), 15942–15947. https://doi.org/10.1073/pnas.0507611102

Rausell, A., Luo, Y., Lopez, M., Seeleuthner, Y., Rapaport, F., Favier, A., Stenson, P. D., Cooper, D. N., Patin, E., Casanova, J.-L., Quintana-Murci, L., & Abel, L. (2020). Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. *Proceedings of the National Academy of Sciences*, *117*(24), 13626–13636. https://doi.org/10.1073/pnas.1917993117

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17), e118–e118. https://doi.org/10.1093/nar/gkr407

Riu-Riu, M. (1995). El poblament dels Pirineus, segles VII-XIV. In E. Vives & J. Bertranpetit (Eds.), *Muntanys i Població. El passat dels Pirineus des d'una perspectiva multidisciplinària* (pp. 195–220).

Rosenberg, M. S. (2010). The Bearing Correlogram: A New Method of Analyzing Directional Spatial Autocorrelation. *Geographical Analysis*, *32*(3), 267–278. https://doi.org/10.1111/j.1538-4632.2000.tb00428.x

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*(5), 380–390. https://doi.org/10.1038/nrg795

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic Structure of Human Populations. *Science*, *298*(5602), 2381–2385. https://doi.org/10.1126/science.1078311

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. https://doi.org/10.1038/nature01140

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive natural selection in the human lineage. *Science (New York, N.Y.)*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Samuels, D. C., Wang, J., Ye, F., He, J., Levinson, R. T., Sheng, Q., Zhao, S., Capra, J. A., Shyr, Y., Zheng, W., & Guo, Y. (2016). Heterozygosity ratio, a robust global genomic measure of autozygosity and its association with height and disease risk. *Genetics*, *204*(3), 893–904. https://doi.org/10.1534/genetics.116.189936

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., & Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, *507*(7492), 354–357. https://doi.org/10.1038/nature12961

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., & Reich, D. (2012). The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics*, *8*(10), e1002947. https://doi.org/10.1371/journal.pgen.1002947

Scerri, E. M. L., Chikhi, L., & Thomas, M. G. (2019). Beyond multiregional and simple out-of-Africa models of human evolution. *Nature Ecology & Evolution*, *3*(10), 1370–1372. https://doi.org/10.1038/s41559-019-0992-1

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Schmidt, S., Gerasimova, A., Kondrashov, F. A., Adzuhbei, I. A., Kondrashov, A. S., & Sunyaev, S. (2008). Hypermutable Non-Synonymous Sites Are under Stronger Negative Selection. *PLoS Genetics*, *4*(11), e1000281. https://doi.org/10.1371/journal.pgen.1000281

Schmutte, C., Yang, A. S., Beart, R. W., & Jones, P. A. (1995). Base Excision Repair of U:G Mismatches at a Mutational Hotspot in the p53 Gene Is More Efficient Than Base Excision Repair of T:G Mismatches in Extracts of Human Colon Tumors. *Cancer Research*, *55*(17).

Scrucca, L., Fop, M., Murphy, T., B., & Raftery, Adrian, E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, *8*(1), 289. https://doi.org/10.32614/RJ-2016-021

Ségurel, L., Wyman, M. J., & Przeworski, M. (2014). Determinants of Mutation Rate Variation in the Human Germline. *Annual Review of Genomics and Human Genetics*, *15*(1), 47–70. https://doi.org/10.1146/annurev-genom-031714-125740

Sergeeva, M., Delahaye, D., Mancel, C., & Vidosavljevic, A. (2017). Dynamic airspace configuration by genetic algorithm. *Journal of Traffic and Transportation Engineering (English Edition)*, *4*(3), 300–314. https://doi.org/10.1016/j.jtte.2017.05.002

Serra-Rotés, R. (2017). Carretera, ferrocarril i industrialització a la comarca del Berguedà (Barcelona) al segle XIX i principis del XX. In *III Congrés Internacional d'Història dels Pirineus.* (pp. 487–500).

Shendure, J., & Akey, J. M. (2015). The origins, determinants, and consequences of human mutations. *Science*, *349*(6255), 1478–1483. https://doi.org/10.1126/science.aaa9119

Silveira, L. E. Da, Alves, D., Painho, M., Costa, A. C., & Alcântara, A. (2013). The evolution of population distribution on the Iberian Peninsula: A transnational approach (1877-2001). *Historical Methods*, *46*(3), 157–174. https://doi.org/10.1080/01615440.2013.804787

Silver, L. M. (1995). *Mouse Genetics: Concepts and Applications*. Oxford University Press.

Smith, F. H., Ahern, J. C. M., Janković, I., & Karavanić, I. (2017). The Assimilation Model of modern human origins in light of current genetic and genomic knowledge. *Quaternary International*, *450*, 126–136. https://doi.org/10.1016/j.quaint.2016.06.008

Solé, A., Solana, M., & Mendizabal, E. (2014). Integration and international migration in a mountain area The Catalan Pyrenees. *Revue de Géographie Alpine*, *102–3*, 0–13. https://doi.org/10.4000/rga.2484

Stoneking, M. (2005). DNA and recent human evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *2*(2), 60–73. https://doi.org/10.1002/evan.1360020208

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, *135*(2), 599–607.

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, *28*(4), 289–301. https://doi.org/10.1002/gepi.20064

Telis, N., Aguilar, R., & Harris, K. (2020). Selection against archaic hominin genetic variation in regulatory regions. *Nature Ecology & Evolution*, *4*(11), 1558–1566. https://doi.org/10.1038/s41559-020-01284-0

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

The 1000 Genomes Project Consortium, Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The International HapMap Consortium. (2003). The international HapMap project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

Thomas-Flanagan, M. (2016). *Regression Class: Linear and Non-linear Regression*.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234. https://doi.org/10.2307/143141

Toledo, A., Pámpanas, L., García, D., Pettener, D., & González-Martin, A. (2017). Changes in the genetic structure of a valley in the Pyrenees (Catalonia, Spain). *Journal of Biosocial Science*, *49*(1), 69–82. https://doi.org/10.1017/S0021932016000031

Tran, T. D., Hofrichter, J., & Jost, J. (2013). An introduction to the mathematical structure of the Wright–Fisher model of population genetics. *Theory in Biosciences*, *132*(2), 73–82. https://doi.org/10.1007/s12064-012-0170-3

Vaser, R., Adusumalli, S., Ngak Leng, S., Sikic, M., & Ng, P. C. (2015). SIFT missense predictions for genomes. *Nature Protocols*, *11*.

https://doi.org/10.1038/nprot.2015.123

Vidal, T., & Recano, J. (1986). Rural demography in Spain today. *Espace-Populations-Societes*, *3*, 63–74. https://doi.org/10.3406/espos.1986.1161

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., & Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science*, *253*(5027), 1503–1507. https://doi.org/10.1126/science.1840702

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072

Wakely, J. (2016). *Coalescent Theory: An Introduction*.

Wang, C., Szpiech, Z. A., Degnan, J. H., Jakobsson, M., Pemberton, T. J., Hardy, J. A., Singleton, A. B., & Rosenberg, N. A. (2010). Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis. *Statistical Applications in Genetics and Molecular Biology*, *9*(1). https://doi.org/10.2202/1544-6115.1493

Wang, C., Zöllner, S., & Rosenberg, N. A. (2012). A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS Genetics*, *8*(8), e1002886. https://doi.org/10.1371/journal.pgen.1002886

Wang, D. G. (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science*, *280*(5366), 1077–1082. https://doi.org/10.1126/science.280.5366.1077

Wastnedge, E., Waters, D., Patel, S., Morrison, K., Goh, M. Y., Adeloye, D., & Rudan, I. (2018). The global burden of sickle cell disease in children under five years of age: a systematic review and meta-analysis. *Journal of Global Health*, *8*(2). https://doi.org/10.7189/jogh.08.021103

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Wolpoff, M., & Caspari, R. (1997). *Race and Human Evolution*. Simon & Schuster.

Wright, S. (1969). *Evolution and the Genetics of Populations. Vol. 2, The Theory of Gene Frequencies*. University of Chicago Press.

Xia, J., Han, L., & Zhao, Z. (2012). Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. *BMC Genomics*, *13 Suppl 8*. https://doi.org/10.1186/1471-2164-13-s8-s7

Xue, J., Lencz, T., Darvasi, A., Pe’er, I., & Carmi, S. (2017). The time and place of European admixture in Ashkenazi Jewish history. *PLOS Genetics*, *13*(4), e1006644. https://doi.org/10.1371/journal.pgen.1006644

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N., de Manuel, M., Hernandez-Rodriguez, J., Lobon, I., Siegismund, H. R., Pagani, L., Quail, M. A., Hvilsom, C., Mudakikwa, A., Eichler, E. E., … Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*(6231), 242–245. https://doi.org/10.1126/science.aaa3952

Xue, Yali, Mezzavilla, M., Haber, M., McCarthy, S., Chen, Y., Narasimhan, V., Gilly, A., Ayub, Q., Colonna, V., Southam, L., Finan, C., Massaia, A., Chheda, H., Palta, P., Ritchie, G., Asimit, J., Dedoussis, G., Gasparini, P., Palotie, A., … Zeggini, E. (2017). Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms15927

Yang, W.-Y., Novembre, J., Eskin, E., & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics*, *44*(6), 725–731. https://doi.org/10.1038/ng.2285

Yoon, J.-H., Smith, L. E., Feng, Z., Tang, M.-S., Lee, C.-S., & Pfeifer, G. P. (2001). Methylated CpG Dinucleotides Are the Preferential Targets for G-to-T Transversion Mutations Induced by Benzo[a]pyrene Diol Epoxide in Mammalian Cells: Similarities with the p53 Mutation Spectrum in Smoking-associated Lung Cancers 1. In *Cancer Research* (Vol. 61).

Zenzes, M. T. (2000). Smoking and reproduction: gene damage to human gametes and embryos. *Human Reproduction Update*, *6*(2), 122–131. https://doi.org/10.1093/humupd/6.2.122