Genomic data analyses for population history and population health

Clare Bycroft

St Hildas College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

For submission in Trinity 2017

This thesis is dedicated to my parents, Trevor and Christine Bycroft,

who taught me to be curious.

Abstract

Many of the patterns of genetic variation we observe today have arisen via the complex dynamics of interactions and isolation of historic human populations. In this thesis, we focus on two important features of the genetics of populations that can be used to learn about human history: population structure and admixture. The Iberian peninsula has a complex demographic history, as well as rich linguistic and cultural diversity. However, previous studies using small genomic regions (such as Y-chromosome and mtDNA) as well as genome-wide data have so far detected limited genetic structure in Iberia. Larger datasets and powerful new statistical methods that exploit information in the correlation structure of nearby genetic markers have made it possible to detect and characterise genetic differentiation at fine geographic scales. We performed the largest and most comprehensive study of Spanish population structure to date by analysing genotyping array data for $\sim$1,400 Spanish individuals genotyped at $\sim$700,000 polymorphic loci. We show that at broad scales, the major axis of genetic differentiation in Spain runs from west to east, while there is remarkable genetic similarity in the north-south direction. Our analysis also reveals striking patterns of geographically-localised and subtle population structure within Spain at scales down to tens of kilometres. We developed and applied new approaches to show how this structure has arisen from a complex and regionally-varying mix of genetic isolation and recent gene-flow within and from outside of Iberia. To further explore the genetic impact of historical migrations and invasions of Iberia, we assembled a data set of 2,920 individuals ($\sim$300,000 markers) from Iberia and the surrounding regions of north Africa, Europe, and sub-Saharan Africa. Our admixture analysis

implies that north African-like DNA in Iberia was mainly introduced in the earlier half (860 – 1120 CE) of the period of Muslim rule in Iberia, and we estimate that the closest modern-day equivalents to the initial migrants are located in Western Sahara. We also find that north African-like DNA in Iberia shows striking regional variation, with near-zero contributions in the Basque regions, low amounts ($\sim$3%) in the north east of Iberia, and as high as ($\sim$11%) in Galicia and Portugal.

The UK Biobank project is a large prospective cohort study of $\sim$500,000 individuals from across the United Kingdom, aged between 40-69 at recruitment. A rich variety of phenotypic and health-related information is available on each participant, making the resource unprecedented in its size and scope. Understanding the role that genetics plays in phenotypic variation, and its potential interactions with other factors, provides a critical route to a better understanding of human biology and population health. As such, a key component of the UK Biobank resource has been the collection of genome-wide genetic data ($\sim$805,000 markers) on every participant using purpose-designed genotyping arrays. These data are the focus of the second part of this thesis. In particular, we designed and implemented a quality control (QC) pipeline on behalf of the current and future use of this multi-purpose resource. Genotype data on this scale offers novel opportunities for assessing quality issues, although the wide range of ancestral backgrounds in the cohort also creates particular challenges. We also conducted a set of analyses that reveal properties of the genetic data, including population structure and familial relatedness, that can be important for downstream analyses. We find that cryptic relatedness is common among UK Biobank participants ($\sim$30% have at least one first cousin relative or closer), and a full range of human population structure is present in this cohort: from world-wide ancestral diversity to subtle population structure at sub-national geographic scales. Finally, we performed a genome-wide association scan on a well-studied

and highly polygenic phenotype: standing height. This provided a further test of the effectiveness of our QC, as well as highlighting the potential of the resource to uncover novel regions of association.

Word count: 38,200

Acknowledgements

I wish to first thank my doctoral supervisors, Professors Peter Donnelly[1,2] and Simon Myers[1,2] for their financial and intellectual support over the last four years. More specifically, I thank Peter for giving me the opportunity to come to the other side of the world and study some seriously cool stuff; and Simon, for the many extended discussions and group meetings that gave rise to a lot of the ideas for analysis presented in this thesis.

The analysis of Spanish population structure would not have been possible without the excellent data set that was made available to us through a collaboration with Professor Ángel Carracedo[3], at the University of Santiago de Compostela. I am also indebted to the scientists involved in collecting this data and answering my queries about the data collection process: Dr. Ceres Fernandez–Rozadilla[1,3], Dr. Clara Ruiz-Ponte[3], and Inés Quintela-Garcí[3]. Professor Ángel Carracedo also provided valuable insights into aspects of Spanish history and Galician culture.

The analysis of UK Biobank genetic data was a largely collaborative effort between myself and two other scientists, Dr. Desislava Petkova (Desi) and Dr. Colin Freeman. Desi's involvement in the first two years of the project and the interim data release (May 2015) inspired much of our thinking and analysis of the full cohort of 0.5 million participants. We are also indebted to the hard work and helpful assistance of management staff and scientists at Affymetrix (now Thermo Fisher Scientific): Teresa Webster, Jeanette Schmidt, Amy Ollmann, Jeremy Gollub and Laurent Bellon. UK Biobank staff, especially Samantha Welsh and Alan Young, also provided

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

[2]Department of Statistics, University of Oxford, UK

[3]Galician Public Fundation of Genomic Medicine (FPGMX)-Grupo de Medicina Xenmica-Centro de Investigacin Biomdica en Red de Enfermedades Raras (CIBERer)-Universiy of Santiago de Compostela, Spain

data access and valuable information relating to the phenotype data we used in our analysis. I also wish to thank Professor Jonathan Marchini for his leadership of the project up to and during the full cohort data release in July 2017.

Colleagues in the Donnelly Group and elsewhere have provided me with on-going support of the intellectual and social kind. Specifically, I wish to thank Dr. Liz Batty for reviewing several thesis chapters and keeping the fridge stocked with milk; Dr. Alex Young for assistance with the derivation of equations in Section 4.7.2 and being generally inspiring; Dr. Fabian Wauthier for helpful suggestions about data visualisation, and also being generally inspiring; Dr. Philipp Becker for tolerating my attempts at speaking German; Dr. Chris Gill for reviewing chunks of my thesis; and Kate Distin-Harvey for keeping us all in order, and doing it with panache. Dr. Garrett Hellenthal, Dr. George Busby and Dr. Lucy van Dorp also provided useful advice and discussion around the use of *fineSTRUCTURE* and *GLOBETROTTER*.

The genetic data for hundreds of thousands of individuals has been central to all of this work. I wish to acknowledge the generosity of these individuals in providing samples of DNA, and allowing their data to be used for a wide variety of research.

Finally, writing thesis would not have been possible without the support and understanding of friends and family. There are too many amazing people to list here, but I'd especially like to thank Steinar Halldórsson and Christiane Kowatsch for many cups of tea and getting me out of the office; Ryan Christ and Hilary Martin for many stimulating conversations; Patrick Albers, Kiran Garimella, Winni Kretzschmar, and Robbie Davies for that too, as well as numerous beers; Holly Trochet for keeping me stocked in whisky and a healthy dose of scepticism; Damian and Emma Ryan for many excellent dinners and conversations; Stefan Webb, Gus McFarlane and Na Cai for being superb flatmates; Alasdair Sinclair for understanding;

and Tino Fleischer for getting me through the last mile.

Contents

List of Figures

Chapter 1

Introduction

1.1 Setting the scene

It has long been observed that groups of human beings living in geographically distant places tend to have characteristics that are unique, or found more commonly in one group than another. Skin colour, height, resistance to disease — to name some obvious examples. We now know that many of these characteristics are governed, to varying degrees, by variation in the molecule Deoxyribonucleic acid (DNA). By studying the patterns of variation in DNA carried by modern-day humans and their ancestors we can learn about the underlying processes that lead to variation in heritable human characteristics. The tendency of heritable human characteristics to vary systematically across the globe suggests that DNA itself varies across the globe in a similar way. This turns out to be true [3]. Indeed, we now know, through the study of population genetics, that many of the patterns of genetic variation we observe today have arisen via the complex dynamics of interactions and isolation of historical human populations [4]. In other words, genetic variation provides a window into our past.

In this thesis we use data of variation across the human genome (genomic data) to look broadly in two directions. First, we look backwards by studying population history, focussing on a region, the Iberian peninsula, that has a unique and complex demographic history within Europe. Specifically, we study patterns of population interactions and isolation within Spain at fine-scale geographic resolutions, as well as

the genetic impact of historical migrations and invasions more generally in Iberia. We then look forward by analysing a large-scale collection of genetic data that promises many opportunities, as well as challenges, for research into population health. Specifically, we analyse the genetic data from $\sim$500,000 participants of the UK Biobank project to address one of the first challenges of this multi-purpose research resource: ensuring the quality of data in an experiment that involves DNA samples from hundreds of thousands of participants. We also conduct a series of analyses that provide current and future users with a quantitative description of the quality and content of the genetic data.

In this introductory chapter we discuss the motivations and research aims of the analyses covered in Chapters 2 to 4. In addition, we introduce key concepts that underpin our analyses, and survey the methodological approaches and results in the literature that are relevant to the topics of this thesis. As a preface to the analysis relating to Iberia (Chapters 2 and 3) we also draw from literature in another discipline, history, to introduce what is known about the demographic history of Iberia according to historians. As background to the analysis of UK Biobank genetic data (Chapter 4) we describe the genotyping experiment that generated the data, highlighting aspects that have direct implications for our analyses.

1.2 Studying population history using genomic data

It is well known in the study of the genetics of populations that demographic history plays an important role in shaping the genetic variation inherited by its descendants [4]. In this thesis we focus on two important features of the genetics of populations that can be used to learn about demographic history: population structure and admixture. In studying these features in humans we can learn about the patterns of historical interactions and/or isolation among different groups of people, and ask how this relates to other factors that influence human history, such as geography and culture. Understanding genetic population structure is also relevant for studying the relationship between genetic variation and human disease, which we discuss later in

this chapter (Section 1.3.1). In this section we describe the theoretical frameworks that population geneticists use to study population structure and admixture, and how these relate to population history; we then discuss several methodological approaches used to detect and characterise population structure and admixture in humans.

1.2.1 Theoretical frameworks for studying population structure and admixture using genetic variation data

Genetic structure arises when there exist barriers to mating between some pairs of individuals within a species, for example, due to geographic separation. In the simplest extreme case, a population breaks into two sub-populations (a subset of people migrate elsewhere, for example) and in subsequent generations offspring are only produced between individuals whose ancestors came from the same sub-populations. Complete population splits, or geographic barriers are not in themselves necessary for population structure to arise. All that is required is for pairing to occur in a consistently biased way with respect to the ancestries of the parents. In particular, structure can arise between two sub-populations that still exchange some migrants; as well as if mate-choice is consistently associated with some characteristic other than geographical proximity, such as height, or religious affiliation.

Population structure can be detected and quantified using genetic data by looking at the patterns of genetic variation in a sample[1] of individuals. The basic intuition is that groups of individuals who are descended from the same sub-populations will tend to have more alleles[2] in common than individuals descendent from different sub-populations. There are two main theoretical frameworks that population geneticists use to model how the process of genetic inheritance and the history of a population results in the patterns of allelic variation associated with population

[1] The word 'sample' is somewhat problematic in the interface between statistics and biology. It can refer to a biological component that has been collected from a single individual. In our case this is the DNA, or the genetic data associated with the DNA, from an individual (as opposed to the individual his or herself). We also use 'sample' in the statistical sense to mean a collection of data from a larger group of objects of interest (e.g. a subset of individuals from a larger population). Where it is not obvious from the context we will try to be clear as possible about these distinctions.

[2] The different versions of a genetic variant (e.g. A, G, C, or T in the case of single base variants) that exist in humans are known as alleles.

structure.

The first framework involves thinking forward in time. In this framework there are two processes that cause the descendants of different sub-populations to become genetically divergent – genetic drift and mutation[3]. Genetic drift is the process by which frequency of alleles change over time due to the chance (stochastic) process of genetic inheritance within a finite population[4] [5]. In the context of two separated sub-populations, over time (i.e. generations) genetic drift will act independently, thus the descendants of the two sub-populations will carry allelic variation at different frequencies. If a mutation occurs in one of the sub-populations after the split, then this will exist in one population and not the other. Both of these processes — genetic drift and mutation — will lead to members of the same sub-population sharing more alleles than members in different sub-populations. The predicted dependence of allele frequencies on the demographic history of a population form the basis of 'genetic drift-based' methods for inferring population structure from genetic data.

The second theoretical framework — the coalescent — looks backwards in time, and describes how the ancestral relationships between a sample of individuals within a large population are structured according to the demographic history of their ancestors [6]. The following is an informal description, which will help clarify further discussions in this thesis. Consider a short haplotype[5] within an individual sampled from a population today. The path through the series of ancestors from which the haplotype was directly inherited is a lineage; and a coalescent event is the point at which this lineage has a common ancestor with another lineage (i.e. from a different sampled individual, or different chromosome copy within an individual). Multiple haplotypes form a rooted tree consisting of branches (lineages) that are joined by nodes at each coalescent event. If we now think of the choice of ancestor at each generation as being a stochastic process, then this formulation becomes a *distribution*

[3]For now we will ignore the process of recombination and natural selection.

[4]It can be shown that in a Wright-Fisher population of size N, the heterozygosity of the population (i.e. the chance of the alleles in two randomly-chosen haplotypes being different) is reduced, on average, by a factor of $1/N$ each generation. This means that as time goes on, more individuals in the same population will carry the same allele. The particular allele this it is, is random (if we assume natural selection is not at play) but depends on its initial frequency.

[5]The particular sequence of bases (or more generally, alleles) on a chromosome that an individual inherited from *one* parent is known as a haplotype. This term can be used to refer to a whole chromosome, or a segment of arbitrary length.

over all possible trees joining a set of sampled haplotypes. The statistical properties of the coalescent, such as the expected coalescent times for n lineages, and how this depends on different demographic factors (such as population size, and population structure) are well-understood [7]. Importantly, the coalescent encodes properties of the demographics of a *population* by considering only the ancestral relationships between a *sample* of individuals; and its properties hold independently of the process of mutation (except in the context of natural selection) [8]. Under this framework, in a un-structured (or 'panmictic') population the probability of two randomly sampled lineages coalescing at some time T will be uniform across all lineages. Conversely, in a structured population, the probability of two random lineages coalescing at some time T will depend (in some way) on the two lineages chosen. In the example of two separated sub-populations, pairs of lineages from the same sub-population will coalesce (in expectation) more recently than pairs of lineages from different sub-populations [7]. If we now think of mutations as occurring randomly along lineages, then the haplotypes in different sub-populations will tend to have more allelic differences because their time to coalesce will be longer. So, learning about coalescent times from genetic variation data in a sample can tell us about the demographic history of a population, and this idea forms the basis of 'coalescent-based' statistical inference methods that have been developed to study population structure.

In the discussion above we have so far ignored an important process in genetic inheritance: recombination. This is the process of maternal and paternal chromosomes exchanging DNA in the production of egg and sperm cells. This means that rather than inheriting exact copies of whole chromosomes from their parents, children inherit chromosomes that are a mosaic of their parents' two chromosome copies[6]. Recombination is a relatively rare event, with only about 1–2 recombination event occurring per chromosome per gamete, and positions of recombination occur randomly (but not uniformly) along the genome [9, 10]. In a population setting, this means that the alleles an individual carries in two different places (loci) on the same chromosome cannot be considered as independent draws from an underlying

[6]For simplicity, we refer here to the autosomes only.

distribution of population allele frequencies. Phrased in terms of the coalescent, the lineages through which in an individual inherited genetic material at two different locations on the genome cannot be considered independent draws from the same underlying distribution of coalescent trees. However, over many generations, many recombination events will occur between two positions on the genome, thus breaking down the dependence structure between loci. In terms of the coalescent, this means that the ancestral relationships between lineages can be different for different loci, and these differences occur exactly where recombination has occurred between the loci along one of the lineages. The probability of these events occurring is related to coalescent times between lineages, which depend on properties of population demography. Thus, information in the dependence structure — also referred to as linkage disequilibrium (LD) — between nearby loci is an important tool for studying population history using genetic data.

When people with ancestry from diverged sub-populations have children together, their children (and their descendants) will inherit genetic material that has arisen via a mixture of different population histories. This process is known as 'admixture' and the DNA of their descendants will contain signatures of this process. In terms of the theoretical frameworks described above, alleles in a group of admixed individuals will be mixtures of the allele frequencies in the sub-populations that came together at some point in the past. In terms of the coalescent, the DNA in admixed individuals will be inherited via lineages that trace back through members of different sub-populations. We think of an 'admixture proportion' as the fraction of DNA that an individual inherited from each of the ancestral sub-populations. In a recent admixture scenario (e.g. in the last two or three generations), this fraction may be quite different for different individuals. However, if admixture happened many generations ago, and involved many individuals from the same set of sub-populations having children together, an individual sampled from the descendants of such an event will have admixture fractions that reflect the relative amount of individuals from different sub-populations that came together in the past. We have referred to admixture as being an 'event', which occurs at some point in time. In some cases this may be a

reasonable approximation (such as a single historical migratory event), while in others admixture occurs continuously over a longer period of time.

Another important concept worth describing is 'admixture LD'. This is analogous to LD as described above, but where the correlations among loci along the genome are induced by correlations in the *ancestral population* from which a locus is inherited [11]. At the time of, or soon after an admixture event, descendants will inherit large, contiguous segments of DNA originating from each of the ancestral populations. At successive generations, recombination breaks down these segments into smaller and smaller pieces, thus breaking down the correlations in loci inherited from the same ancestral population. The distribution of the decay in correlation with genetic distance is directly related to the number of generations since the initial admixture [11]. The more generations, the faster the rate of decay, and this idea forms the basis of the more sophisticated (and in general more successful) methods of detecting and dating historical admixture events using multi-locus genetic data.

1.2.2 Survey of methodological approaches for studying population structure and admixture

1.2.2.1 Approaches to studying population structure

Current methods for studying population structure in humans can be broadly categorised into three main approaches. One is to define 'populations' based on external information, such as cultural or geographical labels, and then study the genetic differences or similarities between individuals within those groups, assuming that the ancestors of the individuals within each group have the same underlying population history. Another approach is to look for structure inherent in the patterns of allele sharing among sampled individuals, and relating observed patterns in the data back to other information known about the samples, such as geographic locations. The third approach is to treat population structure as a classification problem, and aim to classify individuals into distinct genetic groups (or mixtures of groups) based on patterns of allele sharing, usually by specifying in advance the number of groups. For

each of these approaches we will focus on what appear to be the most widely-used methods in the literature, and which have been applied in studying the genetic make-up of the Iberian peninsula.

Different methodological approaches are often associated with particular kinds of genetic variation and/or associated technology for measuring it, such as genotype[7] data from genotyping arrays (Appendix A.1 for more details). Prior to the wide-spread use of whole-genome genotyping array data genetic studies of populations were carried out using genetic information from small, but highly variable parts of the genome, such as microsatellites, and which often have many possible allelic states within humans (see Appendix A.1). Methods of analysis of this kind of genetic data usually involve computing various kinds of measures of genetic distance or diversity (e.g. F_{ST}, R_{ST}, Nei & Miller genetic distance), within and across different pre-defined 'populations', from which individuals have been sampled. These measures can then be used to make statements about degrees of genetic differentiation between different populations, or to build trees describing relationships between populations (e.g. neighbour-joining trees [12]), or compared to expected levels (e.g. of allelic diversity) based on different demographic scenarios [13, 14, 15, 16, 17].

One of these measures, the 'fixation index', or F_{ST} has been widely-used as a measure of genetic drift among human populations. It was first proposed independently by Wright [18] and Malécot [19]. It measures the proportion of genetic diversity due to allele frequency differences between populations. F_{ST} can be estimated in a variety of ways, depending on the type of genetic data and sampling scheme used. These are described in a review by Holsinger and Weir [20]. Two points of relevance to this thesis are that estimators of F_{ST} (and related measures such as R_{ST}) treat multiple loci as statistically independent, and sampled individuals need to first be assigned to 'populations', which are assumed to be homogeneous with respect to their demographic history.

Analysing variation in Y-chromosome and mitochondrial DNA (mtDNA) has been a popular way of studying population history. These genomic regions are attractive for

[7] 'Genotype' refers to the *pair* of alleles at a locus that an individual inherited from both parents.

their non-recombining nature, meaning that current-day diversity arose from a single genealogical tree, rather than many trees (as occurs in recombining parts of the genome). It is therefore possible to interpret the topology of this tree (once estimated [21]) as a set of real genealogical relationships, rather than some 'average' genealogy. Analyses of Y-chromosome or mtDNA data usually involve considering 'haplogroups', which are specific haplotypes defined by a unique set of mutations at multiple SNPs, and are thought to have arisen early in human history because they are found in many parts of the world. It is possible to construct a genealogical tree that relates different haplogroups to each other by inferring coalescent times and topology that is consistent with their allelic differences [22]. Tandem repeat loci are also often used to define different Y-chromosome or mtDNA haplotypes and estimate coalescent times based on allelic differences, within and across sub-populations [23, 24]; or more commonly, considering frequencies of Y-chromosome haplotypes in different sub-populations, and relating these back to a previously-estimated geneaology of haplogroups [1, 25, 26, 27, 28]. Because the Y chromosome and mitochondria are inherited exclusively via the paternal and maternal lines, respectively, such data can also reveal aspects of population history that are sex-biased. Again, the majority of studies using this approach rely on pre-defined labels to define discrete 'populations' in advance of the analysis.

Principal components analysis is also a popular way of detecting population structure using genetic data [29, 30, 31, 32, 33, 34]. PCA was first introduced in 1978 [35] as a way of using multi-locus genetic data to analyse the geographic distribution of genetic variation across Europe. Informally, consider a matrix X with S rows and L columns, containing mean-centred frequencies of L genes in S sampled groups. The first principal component (PC) of this matrix is the vector of coefficients (or 'loadings') in a linear combination of the gene frequencies, which has maximum variance when applied to all S groups. The value of the linear combination for each group is known as the PC 'score'. The second PC is the vector of coefficients that maximises the variance of scores after subtracting the first principal component, and successive PCs are defined similarly. It turns out that each principal component defined in this way is an

eigenvector[8] of the matrix $X^T X$, and each PC explains successively smaller amounts of variation in the data. The elements of the matrix X need not be gene frequencies in sampled groups (PCA can be applied to any set of features measured on a set of objects), but PCA was first used in this way to construct 'synthetic maps' of genetic variation [35]. Specifically, Menozzi *et al.* [35] estimated the frequencies of a variety of different genes (forming the columns of X), at a set of regularly-spaced geographic locations across Europe (forming the rows of X). For each PC they drew a map, where the PC scores for each location on the map determined its shade, thus illustrating how the distribution of gene frequencies across Europe could be decomposed into a small number of axes of variation.

There have also been other attempts to combine geographic and genetic information to test, for example, deviations from an isolation by distance model[9] [36, 37], as well as attempts to predict an individual's geographic origin based on their genetic data [29].

Since the development of genotyping array technology that can measure genotypes at hundreds of thousands of genetic loci (see Section A.1), PCA has been commonly-used as a way of studying population structure among a sample of *individuals*. In this case, the matrix X contains the genotypes of individuals, rather than gene frequencies among groups of individuals. PCA does not involve a model of how different demographic histories lead to different patterns in the principal components, such as clines, or clusters of individuals along individual components. However, Patterson *et al.* [38] introduced formal statistical tests for the presence of population structure based on PCA. It has also been shown how PCs relate to underlying genealogical histories [39]. Specifically, McVean showed that PC scores can be predicted from pair-wise coalescent times between individuals in a sample, and F_{ST} between two populations is a simple function of the euclidean distance in the first PC, between individuals across the two populations. However, PCA is limited in several important ways: it is sensitive to uneven sampling (of different sub-populations) as well as the selection of genetic markers[10] (known as

[8] For this reason PCA is sometimes referred to as 'eigen analysis'.

[9] 'Isolation by distance' refers to the notion that population structure can arise because individuals further away from each are less likely to have children together, and so genetic divergence would be observed in a way that is some continuous function of geographic distance.

[10] In this thesis we use the term 'marker' to mean a genetic locus that can be assayed using existing technology.

'ascertainment bias'); and as McVean showed, "any two demographic models that give the same structure of expected coalescence times will also result in the same [PC scores]" [39]. Despite its limitations, PCA remains a popular tool for analysing population structure, probably because of its computational tractability and effective visualisations.

An important development in the study of population structure was a method, '*STRUCTURE*', for clustering individuals into a fixed number of distinct population groups based on their genotypes at many independent loci [40]. Importantly, the method uses an explicit model for the probability of observing the genotypes in the data, given the assignment of individuals to some number of discrete groups. Central to the model is the observation that different sub-populations will have different underlying allele frequencies (due to genetic drift) at a set of genomic loci, and individuals sampled from these populations will carry genotypes which are random draws from the same underlying allele frequency distribution. The parameters in the model — the allele frequencies in each population, and the assignment of individuals to each population — are inferred using standard Bayesian inference methods.

All the above methods assume that different genetic loci (e.g. a gene, or a SNP) are statistically independent. In the setting in which loci are located far apart, this is a reasonable assumption. However, as a consequence, these methods ignore a lot of information contained in the correlation structure among markers along the genome. The *fineSTRUCTURE* method [41] addresses this, and as a result is able to detect subtle population structure, which is likely to have risen in the more recent past and involve small geographic distances (such as within England [42]). Like *STRUCTURE*, the method aims to find clusters of individuals based on their allelic types at many loci. It first measures ancestry sharing between individuals using a model developed by Li and Stephens [43], which reconstructs the haplotypes[11] of each individual as a mosaic of 'nearest neighbour' haplotypes from all the other individuals. We discuss this in more detail in Chapter 2, but it suffices to say here that the model explicitly takes into account correlations in genetic markers along the genome, thus allowing all the information in

[11] In this setting, data that involves genotypes of individuals (such as from genotyping arrays) need to be phased to form haplotypes. That is, the genotypes need to be separated into the maternally- and paternally-inherited alleles. Reliable algorithms exist for doing this [44].

many hundreds of thousands of loci to be utilised.

In *fineSTRUCTURE*, individuals are modelled as samples from discrete population groups, which have the same underlying distribution of ancestry sharing amongst themselves, and with all the other groups. Under this model, assignment of individuals to clusters is carried out using Bayesian inference techniques. *fineSTRUCTURE* also infers a hierarchical tree, which describes relationships between the inferred clusters based on patterns of ancestry sharing between individuals in different clusters. Importantly, the model includes the number of clusters as a parameter to be inferred, so the number of clusters need not be specified in advance. Furthermore, simulations carried out by the authors of *fineSTRUCTURE* [41] showed that by modelling the patterns of linkage across the genome, it out-performs both PCA and *STRUCTURE* in detecting subtle population structure. The method has since been successfully applied in the context of the British Isles. Leslie *et al.* [42] demonstrated the existence of population structure involving finer geographic scales (within the adjacent counties of Devon and Cornwall, for example) than had ever been detected previously.

1.2.2.2 Approaches to studying admixture

Existing methods for studying admixture can be categorised into two broad classes: those that use information contained in 'admixture LD', and those that do not. The latter class of methods have the disadvantage of not being able to infer admixture dates, but have been used to detect the presence of admixture and estimate admixture proportions in studies of admixture in Iberia, so are worth mentioning here.

An example of a method that does not use admixture LD is an extension to the *STRUCTURE* model, which incorporates admixture by allowing individuals to be mixtures of 'ancestral' groups, and the primary parameter of interest is the fraction of an individual's genome that originated from each group [40]. This particular version of the model has proven to be the most popular in the context of human populations, and has since been re-implemented in software called '*ADMIXTURE*', which performs inference much more quickly than the original implementation [45]. This model has

been used extensively to analyse structure and admixture in human populations [3, 46, 47, 2], but has a number of limitations. This is the subject of a recent paper by Falush *et al.* [48]. Specifically, they highlight that it is not straightforward to determine the appropriate number of groups (although attempts have been made to address this [49, 45]), and the authors of *STRUCTURE* caution that the resulting groups do not necessarily represent real historical populations, but rather a partition that best explains the data under the model. They also demonstrate how very different admixture scenarios can lead to the same results (as did another study [50]). In other words, care must be taken in the interpretation of *ADMIXTURE*/*STRUCTURE* results in terms of underlying population history, for example by using other information about the histories of the populations being studied [48]. Another approach commonly-used in the literature, and that does not utilise admixture LD, are so-called f_3 and f_4 statistics. These use allele frequencies in different (pre-defined) population groups to test for departures from simple tree-like phylogenetic relationships between the groups [51].

One method which does utilise admixture LD information, called *'Rolloff'* [52], measures admixture LD by comparing genotypes of individuals in a population of interest, with genotypes from individuals in two reference populations assumed to be un-admixed descendants of the historical admixing populations. It estimates admixture LD by computing the correlation coefficient of pairs of markers at different distances within a population of interest, and weighting them "according to the allele frequency differentiation between two populations that are genetically 'close' to the ancestral mixing populations" [52]. They then use the rate of decay with genetic distance of this of weighted correlation to infer a date of admixture. Two important disadvantages of this method are that it does not allow for modern-day reference populations that might themselves be admixed; and it only allows for a scenario of two source populations, with admixture occuring at a single point in time (i.e. over a few generations) [52].

A method designed to model a wider range of admixture scenarios, and that allows for reference populations that are themselves admixed, is *GLOBETROTTER* [53]. Like

other tests for admixture (e.g. *Rolloff* [52] and ALDER [54]) this method exploits admixture LD. However, *GLOBETROTTER* also tests for different types of admixture scenarios, such as a single date admixture event involving two source populations, or more complex scenarios such as multiple dates of admixture. If identified, this provides strong evidence of admixture into the target group, occurring in the recent past [53]. We describe *GLOBETROTTER* in more detail in Chapter 3, but it is worth noting here that the approach has more power to detect subtle events (such as those that may have impacted Iberia) compared to others [53]. *GLOBETROTTER* also has the advantage of not requiring prior specification of the modern-day populations to use as proxies of the historical admixing populations. Instead, *GLOBETROTTER* infers a mixture of reference populations that best represents the historical source populations.

1.2.3 A brief introduction to the demographic history of Iberia

Before discussing the literature on population structure and admixture in Iberia[12], it is worth describing in broad terms the demographic history of the peninsula as it is understood by historians. We focus on aspects of history that have plausibly had an impact on genetic variation. Specifically, events involving migration into, or within Iberia, and aspects of cultural and linguistic history that imply some degree of isolation among different groups of people in Iberia. We only consider history up until the late 19th Century, as the genetic variation within our samples is likely to have been driven by preceding events (see Section 2.1 for details). We also draw mostly from literature concerning the history of Spain, as this is the primary focus of our main analyses. However, the histories of Spain and Portugal are intimately linked, so we refer to Iberia where appropriate. Historians often talk about Iberian history in terms of broad phases, each of which we will discuss in turn. As a point of reference, we first introduce a map of the peninsula as it is today (Figure 1.1) showing Spain's 17 'Autonomous Communities' [55]. These are modern-day political and administrative regions but, as we shall see, many of them have been shaped by older cultural and

[12]For the purposes of this thesis we define 'Iberia' as the current territories of Spain and Portugal that are located on the Iberian peninsula and the Balearic Islands.

geopolitical processes [56].

Figure 1.1: Terrain map of Iberia showing Spanish Autonomous Communities The boundaries of each Autonomous Community are shown with thin black lines and labelled using their names in Spanish, with English versions in parenthesis if they differ substantially (sourced from the Spanish Statistical Office http://www.ine.es/). The background colours show elevation and major water systems (source from https://maps-for-free.com/). Green and yellow indicates low elevation (∼0–500 meters above sea-level) and higher elevation is shown in brown (∼500–1000m) and white (∼1000–2500m). One Autonomous Community, the Canary Islands, is not shown on this map as we have omitted it from our study. We also omit the 'Autonomous Cities' of Melilla and Ceuta on the north-African coast (see Section 2.2.1 for details).

1.2.3.1 Pre-historic Iberia (until ∼200 BC)

It is thought that prior to Roman influences the Iberian peninsula was occupied by a variety of pre-historic tribes belonging to three broadly different cultures: 'Celts', 'Iberians', 'Basques'. The Celts are believed to have occupied the north and west of the peninsula (some of whom are thought to have arrived around the 6th Century BC [57]), the Iberians in the south Mediterranean and Atlantic coasts, and the Basques in the

western Pyrenees [58, 59]. Archaeologists and historians also talk about a 'Celtiberian' culture that predominated in the central plains of the peninsula, and who are likely to have been a mixture of Celtic and Iberian cultures [58, 57]. From around 1000 BC Iberia began to be colonised by foreign sea-faring groups, Phoenicians, Greeks and Carthaginians, who settled in coastal regions of southern and eastern Iberia. While it seems they did not "encroach on the territory of the native peoples" [58] they brought with them trade and culture (including literacy), which found its way to other parts of the peninsula.

1.2.3.2 Roman Iberia ($\sim$200 – 400 BC)

In the 2nd and 1st Centuries BC a series of wars known as the Punic wars broke out, during which Romans and Carthaginians jostled for power across the Mediterranean and north African coast. In the second Punic war (218 – 201 BC) Roman forces defeated the Carthaginians in Iberia, and by 206 BC were in control of the southern and eastern coast of the peninsula [58]. It took 200 years for the Romans to take control of the rest of the peninsula, by which time the Roman Empire governed a territory covering the regions of modern-day France, Italy, Greece, much of Turkey, and parts of the north-African coast. The subsequent 400 years of Roman presence in the Iberian peninsula had a significant impact on the cultural and economic life in the region. This influence is summed up by A. T. Fear [58]: "Rome had impressed herself physically on the landscape, providing a framework which allowed transport, manufacture, and agriculture to flourish." Furthermore, the "Latin language provided the base from which the region's romance languages would evolve." Christianity was introduced and was flourishing in Iberia by the 4th Century. Jewish communities had also begun to settle in Iberia during Roman rule, probably in the 1st Century CE, although "we know remarkably little about the communities they established" [57].

1.2.3.3 Visigothic Iberia (409 – 711 CE)

The collapse of the Roman Empire during the 5th Century CE introduced a power vacuum in Iberia, which was taken advantage of by a series of Germanic tribes: Alans, Sueves[13], Vandals, and Visigoths. These groups were part of a more general, and significant, period of migration within Europe, which involved groups of people moving from central, eastern and northern Europe into other regions (including Iberia and the British Isles) where the Romans had lost control. In the particular case of Iberia two of the groups, the Sueves and Visigoths, ultimately commanded a lasting presence in the peninsula. Although thought to be a "small minority of the population" [58], Iberia was ruled by leaders associated with these groups for around 300 years. Sueves settled and ruled in a region known as 'Gallaecia' in north-west of Spain (approximately in Galicia today, see Figure 1.2) until 585 CE, and the Visigoths ruled the rest of Iberia until the early 8th Century. The ethnic origins of the original invading tribes are somewhat elusive (and debated). According to historian Roger Collins the "common identity" of the Visigoths is likely to have originated in the Balkans [58]. The origin of the Sueves seems to be less clear, but probably originated from an area along the Danube river that now runs through modern Slovakia [60]. However, as Collins explains, "it seems more likely that such confederacies were continually forming and reforming themselves, drawing in new elements of population, while losing others, and re-creating their sense of ethnic identity around a series of origin myths..." Furthermore, "at every stage intermarriage with local populations contributed further to the heterogeneous nature of the ethnic mix."

[13] Sueves are also referred to as 'Seuvis' or 'Suebi' in the literature. We will follow [58] and [57] and use Sueves.

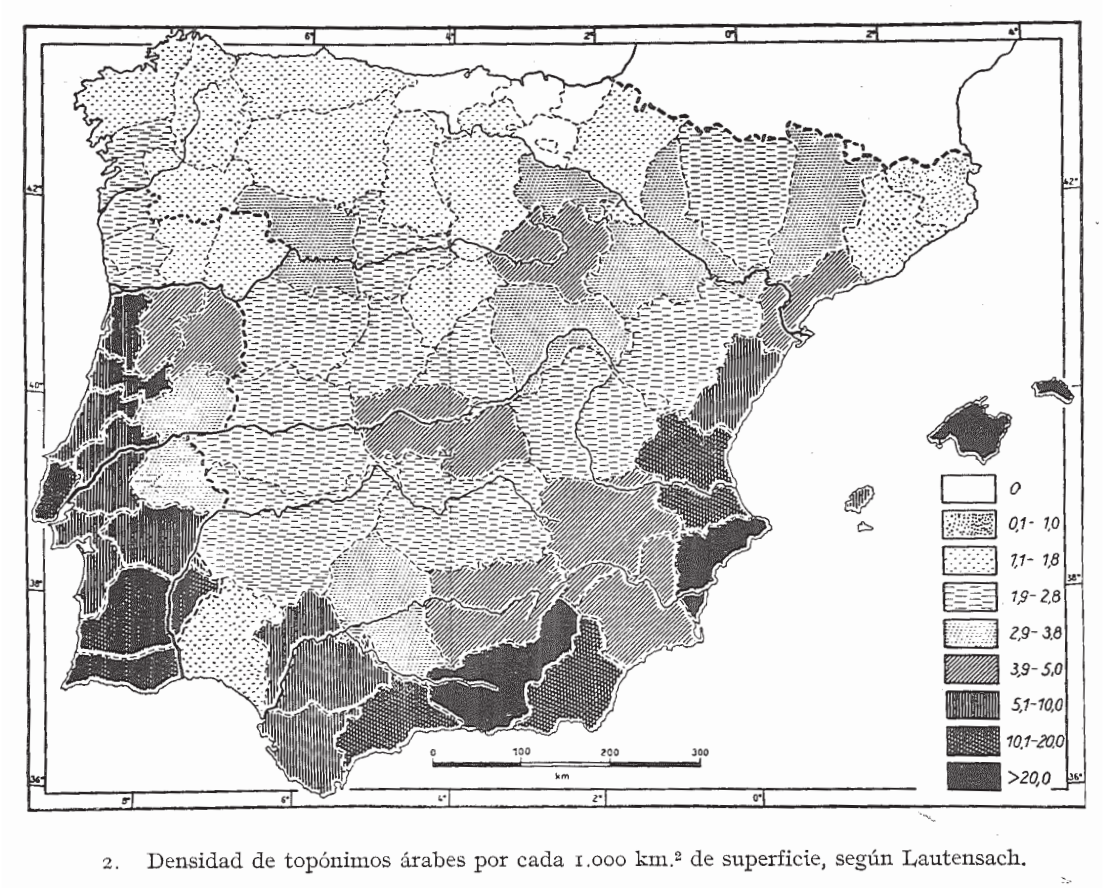
Figure 1.2: Regions of Visigothic Iberia. Different regions of Roman Iberia (e.g. Baetica) were divided up amongst the Sueves and Visigoths in the 5th Century CE. Gallaecia was controlled by Sueves until the 585 CE, when it was annexed by the Visigoths. This map is reproduced from p.g. 40 of [58].

1.2.3.4 Muslim Iberia and the *Reconquista* (711 – 1492 CE)

In the early 8th Century CE, much of north Africa was under the control of the Muslim Umayyad Caliphate. In 711 the Arab-led Caliphate expanded into the Iberian peninsula with an army of largely Berber (indigenous north African) origin, who defeated the ruling Visigoths in less than 10 years [58, 57]. The number of combatants that arrived during the initial invasion events (711 and 712 CE) is well-known: around 30,000 [61] (a relatively small amount compared to the several million indigenous inhabitants [62]). It is also known that an initial migratory wave occurred in the 8th Century, after the Umayyad caliphate took over almost the entire peninsula [61]. Among both the civilian and military newcomers, the majority are thought to have been Berbers from north Morocco, and they settled in many parts of the peninsula [61].

For the next 500 years a variety of Muslim leaders ruled over much of Iberia, except for the mountainous northern regions, which remained the territory of Christian rulers. There is an extensive literature on the significant cultural [62, 63, 64] and linguistic [65, 66, 67] impact of Muslim presence in Iberia. For our purposes, perhaps the most pertinent aspect is the geographical spread of such influences. One indicator of the

regional impact of Muslim presence in Spain is the distribution of Arabic place names in Iberia, as shown in Figure 1.3. Highest densities are seen in Portugal, and regions along the south and east coasts, as well as Balearic Islands. Lowest densities occur in the far north, especially in the regions now known as Galicia, Asturias and País Vasco (Basque Country).



2. Densidad de topónimos árabes por cada 1.000 km.² de superficie, según Lautensach.

Figure 1.3: Density of Arabic place names in Iberia. The caption reads: 'Density of arabic place names for every 1,000 Km^2, according to Lautensach.' Each shaded region on that map is a 20th-century province. This map is reproduced from page 578 of [68], Volume I.

It was from the small, Christian-held territories in the north that the efforts to defeat the southern rulers, collectively known as the *Reconquista*, began. Beginning as early as ∼720, with a battle in the Asturias region, Christian-controlled territory moved gradually southwards. By 1248, when Seville was conquered by the Catholic King Ferdinand III of Castile, almost all of Iberia was again under Christian rule. A region in southern Spain encompassing the city of Granada (Figure 1.4) remained under Muslim rule until the conquest of Granada in 1492, jointly led by Ferdinand II of

Aragón and Isabella I of Castile.

Figure 1.4: State of the *Reconquista* by 1300 CE. Dates next to labelled towns indicate when they were captured by Christian forces. This is a scanned reproduction of Map 7. in [66].

There is little doubt that over the the course of the Middle Ages Iberia experienced significant cultural and demographic flux. In the words of historian Richard Fletcher [58] the peninsula "entered the Middle Ages as a land in which three religions and cultures coexisted and overlapped. Settlements of Arab and Berber immigrants... intermarriage with the indigenous Christian and Jewish populations, shifts of religious allegiance between Judaism, Christianity, and Islam, all combined to produce a thorough cultural mix." Furthermore, it was a time of internal migration and ethnic mixture. Fletcher again: "In northern Spain the Basque and Cantabrian peoples, who had been less affected by Romano-Visigothic Christian culture than their neighbours to the south, were spilling out from their mountains on the plains."

Over the course of the *Reconquista* the linguistic landscape of Iberia appears to have changed substantially. Philologist Ralph Penny writes that "...features which belonged to specific segments of the northern dialect continuum were carried south into areas where they were previously unknown, and where they entered into competition with

features used by the surviving Romance (i.e., Mozarabic) speakers of those areas. The southward movement of population was constant throughout the period of the the Reconquest... The linguistic results of this process, taken together with the gradual hardening of political frontiers in the Peninsula, were the creation of three vertical dialect continua, one in the west (Portugal), one in the centre (Old Castile, New Castile, Extremadura, Andalusia and Murcia, increasingly also including southern Aragón), the third in the east." [65] This effect is illustrated in the work of another philologist, Kurt Baldinger [69], which we reproduce in Figure 1.5.

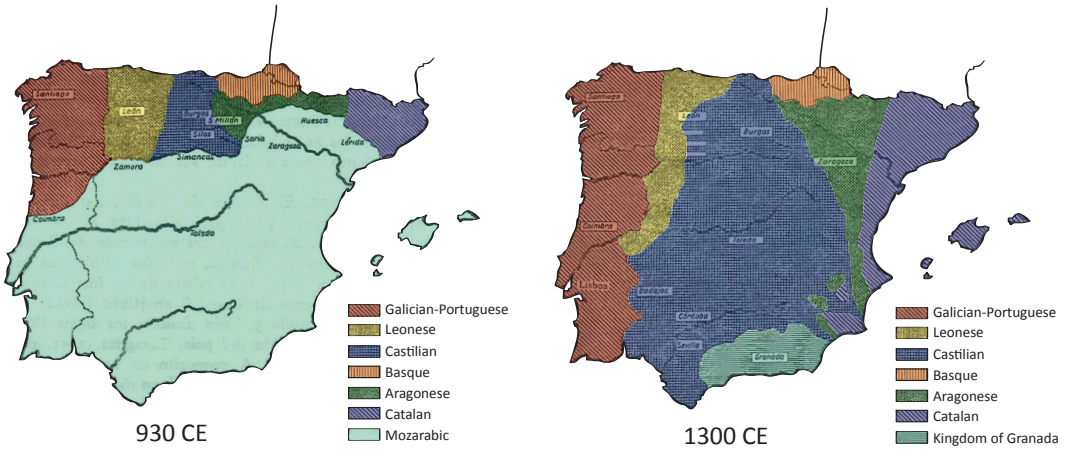


Figure 1.5: Linguistic groups in Iberia near the beginning and end of the *Reconquista*. These maps are adapted from maps in [69] (pages 48 and 54). Colours of the linguistic groups have been added to aid visualization.

1.2.3.5 Imperial Spain and Portugal (∼1500 – 1800 CE)

Political power was first centralised in Spain in 1479 CE when the monarchs of the two largest kingdoms, Ferdinand II of Aragón and Isabella I of Castile, united in marriage. By this time the north-west Kingdom of Leon had become part of the Kingdom of Castile, and Portugal was already a separate kingdom lead by King Alphonso V. The end of the 15th Century marks the beginning of a period of exploration and colonisation carried out by monarchs in both Spain and Portugal. Spanish expeditions occupied the Canary Islands in 1483; and in 1492, under the sponsorship of Ferdinand and Isabella, the maritime explorer, Christopher Columbus,

set sail westwards from Iberia. Spain and Portugal's conquest of the the New World (south America, specifically) began in earnest in the first half of the 16th Century, and by the late 18th Century the Spanish Empire covered large swathes of the central and south American continents [57], and the Portuguese had colonised Brazil. Spanish imperial ambitions during this time also involved other parts of Europe and north Africa. In the mid-16th Century Spanish-claimed territories included what are now the Netherlands, Luxembourg, southern Italy and Sardinia, as well as Melilla and various other townships along the north African coast [57].

Instances of religious and cultural intolerance were certainly not confined to this time, but the 15th and 16th Centuries involved especially large-scale, state-sponsored ethnic and religious persecution. Ferdinand and Isabella initiated the Spanish Inquisition in 1478, and in 1492 — the same year as Columbus' first westward expedition — declared the 'Edict of the Expulsion of the Jews'. According to historian Simon Barton [57] this resulted in around 50,000 Jews from Castile and Aragón leaving for north Africa, Italy and the eastern Mediterranean, while those that stayed in Iberia converted to Christianity. The descendants of Spain's Muslim population ('Moriscos') were also forced to convert to Christianity, and as Barton notes, "subsequent monarchs imitated [Ferdinand and Isabella] with excessive zeal." Hundreds of thousands of Moriscos were systematically expelled by the crowns of Castile and Aragon, with varying degrees of success, and the exact numbers are a debated topic in scholarship covering this period. Barton again: "According to one estimate, some 117,000 Moriscos were expelled from Valencia in 1609, to be followed by a further 150,000 from Aragon and Castile... The majority of emigrants found asylum in North Africa; some travelled, travelled to France, and from there to Italy, Salonika and Istanbul."

1.2.3.6 Population growth and emigration ($\sim$1700 – 1900 CE)

The expanding interests of other European nations, such as Britain and France, checked the domination of the Spanish Empire both in Europe and the Americas. France invaded Spain in 1809, causing the temporary collapse of the Spanish

monarchy, and by the mid-1800s the majority of Spain's American colonies (Mexico, Peru, Bolivia) had declared independence from Spain [57]. In Iberia, the 18th and 19th centuries are marked by demographic change associated with population growth and emigration. It is estimated that the population of Spain grew from 7.5 million people in 1712 to 10.5 million in 1786 [58], and reached 20 million by 1910 [70]. Such expansion was common in most regions of Western Europe. In Iberia, this was also accompanied by episodes of large-scale emigration and epidemic disease. Cholera killed over 500,000 people in various outbreaks in the mid-1800s, and in the late 1800s over 1.5 million Spaniards emigrated to Latin America, driven by rural deprivation, particularly in Galicia and Andalusia [57].

1.2.4 Previous studies of population structure within Spain

As we have seen, Spain has a rich demographic history as well as linguistic and cultural diversity, which suggests genetic population structure is likely to exist. Furthermore, it is likely to have arisen through a complex mixture of migration and isolation driven by a mix of cultural, political or geographic forces. As such, Spain (and more broadly Iberia) has already been the focus of a number of genetic studies.

A variety of other studies have focused on a particular region of Spain — usually an Autonomous Community (Figure 1.1) — and investigated how that region differs genetically[14] from the rest of Spain and/or Europe. Several studies focused on Galicia, using Y-chromosome [26], and other genetic markers [13, 14]. They found no evidence for sub-structure within Galicia, but some evidence of differentiation between Galicia and País Vasco. Another study, focusing only on north-west Spain, reported significant differences in Y-chromosome haplogroup frequencies between the Pas Valley in Cantabria, and Galicia [26].

In the study of Spanish population genetics there has been a sustained interest in the origins of the Basques and characterising their position in the European genetic landscape[15]. This interest is largely due to their cultural isolation despite a number of

[14] 'Genetic distinctiveness' has been measured and statistically tested in a myriad of ways. See Section 1.2.2.1.

[15] Other regions within Spain (Galicia [14, 13, 28], Andalucía [16], and Valencia [15]) have also been the focus of genetic studies, although they have not enjoyed the same general scholarly interest as the Basque region.

major invasions into Iberia over the last 2,000 years, and their remarkable linguistic uniqueness (Basque is the only non-Indo-European language spoken in Western Europe [71]). The region generally referred to as 'Basque' in the literature covers an area that includes the current-day Spanish Autonomous Communities of País Vasco (Spanish for Basque Country) and Navarra, as well as three provinces in the south-west of France which lie contiguous with País Vasco and Navarra. Current-day native speakers of the Basque language are concentrated in the northern parts of this region (see Figure A.1).

It is generally acknowledged in the literature that the Basques are genetically distinct among Spanish groups, as argued by a number of studies using classical markers, as well as Y-chromosome and mtDNA data [33, 14, 72, 1, 30, 24, 37]. However, there is little consensus as to what 'genetically distinct' means. Some studies also claim to have detected structure *within* the Basque region using Y-chromosome and mtDNA data [73, 74] as well as *Alu*-insertions [75]. The most comprehensive study of Basque sub-structure ($\sim$900 samples from Basque-speaking and surrounding regions in Spain and France) [73] argues that genetic structure within the Basque region is best explained by divisions based on pre-Roman tribal boundaries, rather than current-day dialectical boundaries. All these studies pool samples into discrete groups (based on some non-genetic factor such as geographic region or ethnic affiliation) and make inferences based on statistical summaries of these groups.

Only four studies of population structure in Iberia (to our knowledge) sampled comprehensively across Spain [36, 37, 25, 1]. The two earlier studies (published in 1991 [36] and 1994 [37]) used classical genetic markers (see Section A.1) to build 'synthetic maps' using principal components of gene frequencies (e.g. Figure 1.6), as described in Section 1.2.2.1. Both found evidence for genetic structure that differentiated people in the Basque region from the rest of Spain, an observation that has since been made many times using other types of data and methods [1, 30, 24, 72].

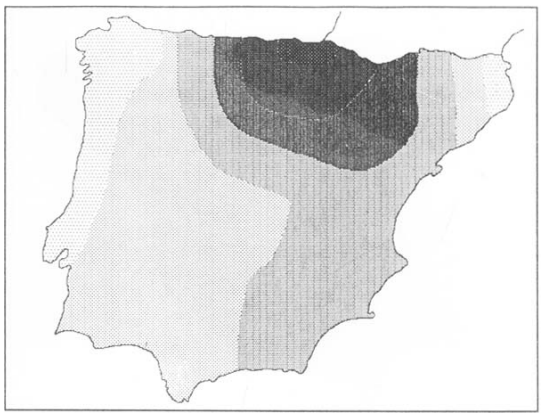


Fig. 1. First principal component of gene frequencies in the Iberian Peninsula from 54 alleles (34 of them independent) of 20 human loci. The percentage of variation explained by this factor is 27·1 %.

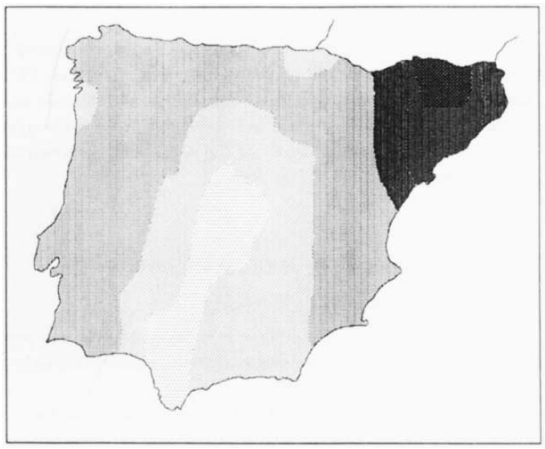


Fig. 2. Second principal component of gene frequencies in the Iberian Peninsula. The percentage of variation explained by this factor is 14·5 % (and 41·6 % accumulated).

Figure 1.6: Synthetic maps of gene frequencies in Iberia by Bertranpetit and Cavalli-Sforza (1991). Each map depicts the PC scores for PC 1 (a) and PC 2 (b) in a PCA of regional gene frequencies (see Section 1.2.2.1). The legend for each map notes the fraction of variation in gene frequencies explained by each PC. These figures are reproduced from [36].

The more recent studies (published in 2004 [25] and 2008 [1]) both used Y-chromosome data and looked for regional differentiation by comparing frequencies of Y-chromosome haplogroups in Spain's Autonomous Communities. The earlier of these studies reported "limited heterogeneity" in Y-chromosome haplogoup frequencies [25][16], and Figure 1.7 shows the estimates of regional haplogroup frequencies in the more recent study [1]. Adams *et al.* found statistically significant differences between samples from the Basque regions and all other sampled regions

[16]Specifically, they used a method ('*SAMOVA*' [76]) that aims to find a partition of the sampled regions that maximizes the proportion of total genetic variance due to differences between the groups in the partition. They found that a partition of their sampled regions into three groups — Basque Country, Catalonia and the rest of Iberians — explained 2.5% of the variance in Y-chromosome haplogroup frequencies ($p < 0.05$)

in Iberia using F_{ST}- and R_{ST}-based tests [1].

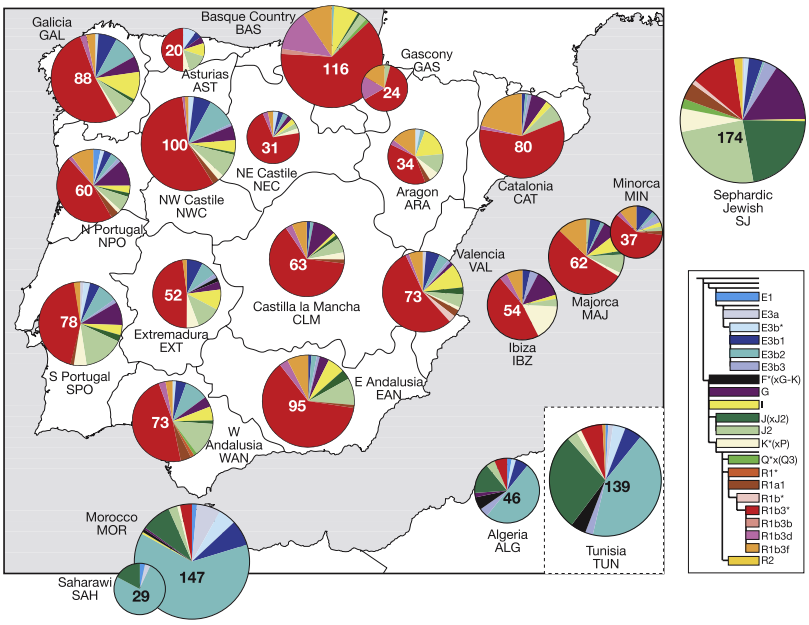


Figure 2. Haplogroup Distributions in Iberian, North African, and Sephardic Jewish Populations
Haplogroup profiles of samples from the Iberian Peninsula and the Balearic Islands, published North African samples,[34,47] and a Sephardic Jewish sample. Sectors in pie charts are colored according to haplogroup in the schematic tree to the right, and sector areas are proportional to haplogroup frequency. Sample names, abbreviations, and sizes (within pie charts) are indicated. Subhaplogroups of R1b3 were not typed in the Sephardic Jewish sample.

Figure 1.7: Frequencies of Y-chromosome 'haplogroups' in Iberia from Adams *et al.* [1]. This study used Y-chromosome data for 1,140 males from the Iberian peninsula and Balearic Islands. We also include the caption from the article for further explanation.

So far only two studies [34, 30] used genome-wide genotyping array data and sampled from more than one region in Spain. The first [34], which used ∼100,000 markers and ∼800 samples, reported two 'clines' of variation in a PCA: one that partially separated individuals recruited in Cataluña from those recruited in central and southern Spain, and another cline partially separating individuals recruited from the south coast (see Figure 1.8a). Despite having a relatively large sample size and using genome-wide genotyping array data, this study is far from a comprehensive description of population structure across Spain due to its restricted sampling scheme. Significantly, there were no individuals from the Basque region or Galicia. The other study was a genome-wide association study (GWAS) for colorectal cancer [30]. It involved cases and controls from all across Spain, and we used the data from this collection in our study (see Section 2.2.1). While the primary aim of the original study was not to investigate population structure, the authors used PCA as a way to

account for population structure in their association analysis. The first two PCs broadly distinguished three subgroups: samples from País Vasco, Galicia, and all others from other parts of Spain (see Figure 1.8b).

(a)

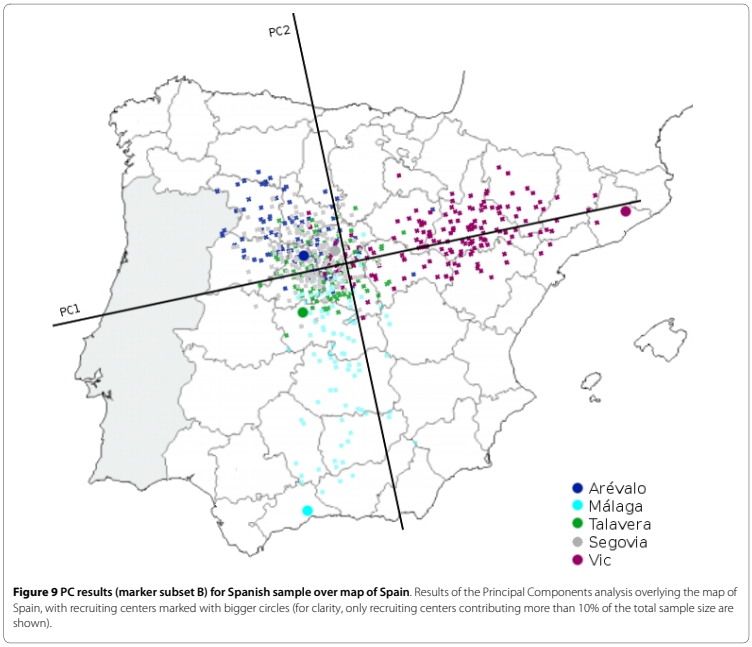


Figure 9 PC results (marker subset B) for Spanish sample over map of Spain. Results of the Principal Components analysis overlying the map of Spain, with recruiting centers marked with bigger circles (for clarity, only recruiting centers contributing more than 10% of the total sample size are shown).

(b)

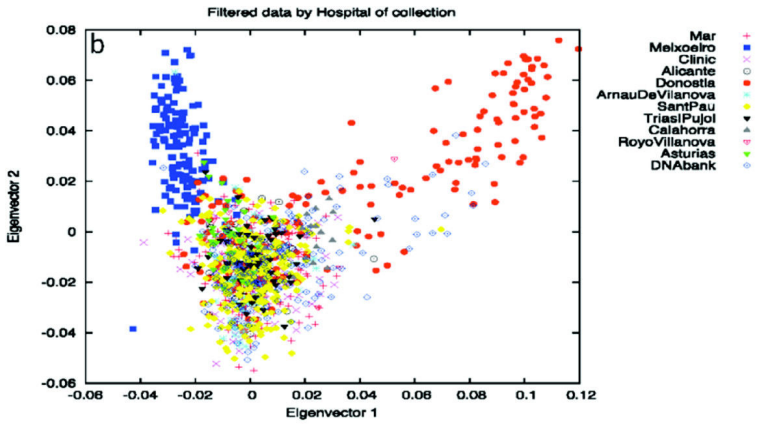


Figure 1.8: Results of PCA from two previous studies that used genome-wide genotyping array data. (a) Figure from Gayan *et al.* [34] showing samples projected onto PCs one and two, coloured according to their centres of recruitment. **(b)** Figure from Fernandez-Rozadilla *et al.* [30] showing samples projected onto PCs one and two and coloured according to the hospital from which the individuals were recruited. The blue and orange points indicate hospitals located in Galicia and País Vasco, respectively.

1.2.5 The aims of the analyses in Chapter 2

Apart from indications of genetic differentiation involving some northern regions of Spain (Cataluña, Galicia, and País Vasco), a comprehensive picture of population structure in Spain — including subtle structure at fine geographic scales — has so far remained elusive. In this thesis we sought to identify and characterise fine-scale population structure within Spain by addressing the main limitations of previous studies. That is, we use a collection of data that covers all of Spain, and which is larger, and contains more detailed geographic information than in any previous population structure studies. Furthermore, we use genotype data of many thousands of markers genome-wide, and apply a method that exploits information in the correlation structure (LD) between markers along the genome [42]. In addition, we sought to explain how the population structure observed in Spain may have been driven by specific events in the demographic history of Iberia.

1.2.6 Previous studies of admixture within Spain or Iberia

Previous studies of admixture in Spain or Iberia have focussed primarily on detecting admixture involving combinations of European- and African-like source populations. This interest has largely been motivated (as it does us) by the known historical link between north Africa and Iberia. Namely, the Islamic invasion of 711 CE, and the subsequent centuries of Muslim presence in Iberia (8th to $\sim$15th Century). The cultural and linguistic impact of Muslim rule in Iberia is well-documented, but the historical record is limited in its ability to inform about the extent, timing and geographic spread of genetic mixing between immigrants and indigenous Iberians over several centuries after the initial invasion [63].

Previous studies using both Y-chromosome and autosomal DNA have reported signals of admixture from sub-Saharan Africa and/or north Africa into Iberia at some point in the past [77, 52, 53, 2]. However, point estimates of the timing of an admixture event involving a north African and European group (or groups) in Iberia vary greatly, from as long as 74 generations ago [77] to 23 generations ago [53] (both of these studies used

an admixture LD-based method), and one study even reported a lower bound as recent as 6 generations ago [2]. Previous estimates of proportions of African-like DNA in Iberia using autosomal data also vary, ranging from 1% [78] (using *GLOBETROTTER*) to as high as 14% [2]. For the latter study, estimates vary considerably depending on the choice of the number of 'source' populations in the *ADMIXTURE* analysis (10 – 25% in Galicia, for example).

The differences in reported ranges within, and across studies may also reflect differences in ancestral make-up in different parts of Iberia. Some regional variation of broadly *north* African-like DNA within Ibera is suggested by two studies [1, 2] (see Figure 1.9 and Figure 1.10). Both studies report a pattern that can be described as highest in the western regions of Iberia (as high as 22% in North-west Castile [1]) and lower in the north-east and Basque regions (as low as 0% in Gascony [1]). Studies that report regional variation do not attempt to, or do not precisely estimate, admixture timing. Nor do they localise the source of admixture to specific regions of north Africa (or the middle-east).

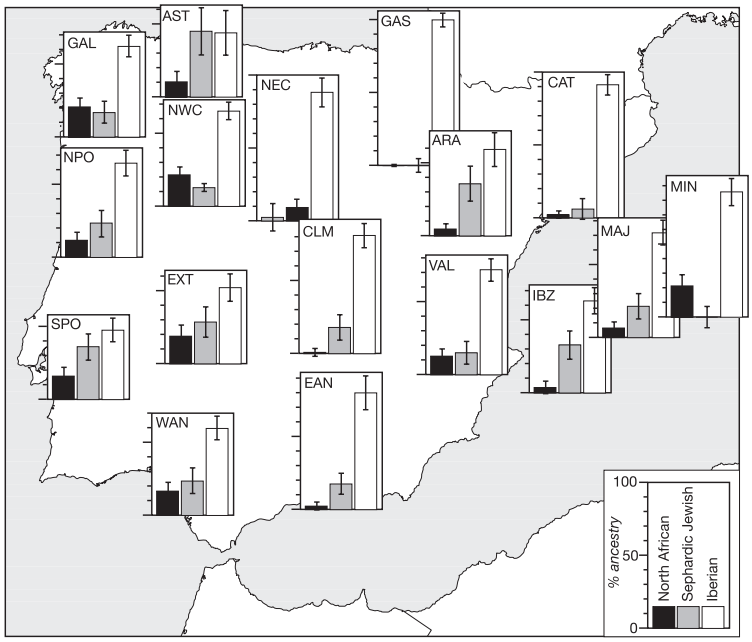


Figure 4. Iberian, North African, and Sephardic Jewish Admixture Proportions among Iberian Peninsula Samples
Mean North African, Sephardic Jewish, and Iberian admixture proportions among Iberian samples, based on the mY estimator and on Moroccan, Sephardic Jewish, and Basque parental populations, are represented on a map as shaded bars on bar charts. Error bars indicate standard deviations, and three-letter codes indicate populations, as given in Figure 1.

Figure 1.9: Estimates of admixture proportions across Iberia using Y-chromosome data from Adams *et al.* [1]. These proportions of ancestry from a set of potential 'parental populations' were estimated using the 'mY' estimator [79]. The method exploits information in the differences in the allele frequencies of the assumed parental populations (based on modern-day references), as well as the amount of differentiation between different, mult-locus *alleles* (in this case between different Y-chromosome haplogroups). It does not use admixture LD. The parental populations for Iberia were defined in this study as Basques, Moroccans and Sephardic Jews. The three-letter codes refer to the same regional groups as shown in Figure 1.7, also from the same article.

Other studies have estimated sub-Saharan African contributions to Iberia, and these generally report lower proportions (0.7% [77] to 3.2% [52]) and an older date (55 [52] to 66 [77] generations ago) than those that involve north African-like mixing groups only. Two studies also claim to observe signals of gene-flow into Iberia involving a middle-eastern source group. Specificially, a component of Qatari-like DNA based on results from *ADMIXTURE* (Figure 1.10) [2], and a source group from the Levant (using *GLOBETROTTER*) [78].

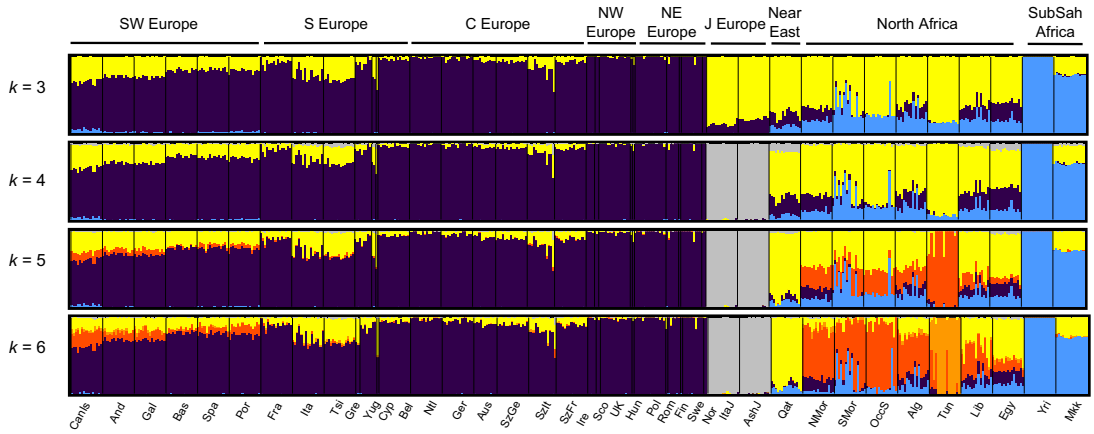


Fig. 1. Allele-based estimates of ancestry in Europe and for European Jews, the Near East, North Africa, and Sub-Saharan Africa. Unsupervised ADMIXTURE results for k = 3–6. Cross-validation indicated k = 4 as the best fit, but higher density datasets (25) and higher values of k continue to identify population-specific ancestries (*SI Appendix,* Fig. S2); we therefore conservatively focused on k = 3:6 ancestral populations.

Figure 1.10: *ADMIXTURE* analysis using genome-wide genotype array data from Botigue *et al.* [2]. Each vertical bar for a given number of ancestral groups (K) represents a sampled individual, and colours indicate the fraction of their ancestry estimated to come from each ancestral group, which does not necessarily represent a real historical population (as we discussed in Section 1.2.2.2). Individuals are grouped according to predefined labels (mostly geographic). For our purposes the groups of interest are those from Iberia (far left) and Africa (far right). The full labels for these groups are [2]: CanIs=Canary Islands; And=Andalucía; Gal=Galicia; Bas=Basque; Spa=Spain General; Por=Portugal; Qat=Qatari; NMor=Morocco North; SMor=Morocco South; OccS=Occidental Sahara; Alg=Algeria; Tun=Tunisia; Lib=Libya; Egy=Egypt; Yri=Nigeria Yoruba; Mkk=Kenya Maasai.

1.2.7 The aims of the analyses in Chapter 3

Previously-reported estimates of the timing and extent of admixture in the population history of Iberia are likely to vary depending on the reference populations assumed to represent the ancestral mixing groups (e.g. from Morocco [1] or Western Sahara [2]), as well as heterogeneity in the ancestral make-up of the modern-day Iberian samples used in the analysis. We sought to clarify the timing of African-like and potentially non-African genetic admixture in the Iberian peninsula; characterise the likely genetic make-up of the source populations(s); as well as map regional variation in external genetic contributions at fine geographic scales. Importantly, we use methods that take into account heterogeneity in the population of interest (Iberia) and avoid strong assumptions about the genetic make-up of the historical admixing groups.

1.3 Studying population health using genomic data

1.3.1 The GWAS era and genomic data for multi-purpose research use

Understanding the role that genetics plays in phenotypic variation[17], and its potential interactions with other factors, provides a critical route to a better understanding of human biology and population health. It is anticipated that this will lead to more successful drug development [80], and potentially to more efficient and personalised treatments and to better diagnoses [81]. One of the primary routes for discovery in this area has been genome-wide association studies (GWAS), which involve testing for statistical association between a phenotype of interest (e.g. a disease) and allelic types at hundreds of thousands of loci across the whole genome (mostly SNPs and indels). This kind of analysis has been enabled through the development and optimisation of high-throughput genotyping array technology (see Section A.1), allowing individual studies to collect genome-wide genetic data on thousands, and in some cases tens of thousands, of individuals. The development of methods for imputing genotypes by exploiting the correlations between nearby loci has also enabled millions more variants to be tested in any given study [82]. In the last 10 years hundreds of different clinical phenotypes and human traits have been studied in this manner, and the reported associations systematically catalogued [83].

The significant impact of the GWAS era on our understanding of the role of human genetics in health and disease has been summurised in a recent review [81]. The authors point to two important considerations for the future of GWAS-based discovery that are of relevance to this thesis: the need for much larger sample sizes (hundreds of thousands or millions of individuals) in order to explore the contribution of rare genetic variation to heritable traits; and the need to collect more comprehensive data on non-genetic factors to explore the complex interactions between our environment, our lifestyle, and our genetics (including ancestry), and how this contributes to different human health outcomes. These considerations have prompted the emergence of very large-scale projects, in which genetic data on hundreds of

[17] We use the term 'phenotype' to mean any measurable human trait (e.g. height, or having a particular disease) that is not a genetic variant.

thousands of participants are being collected, along with extensive phenotypic information. Notable examples of such collections are deCODE Genetics in Iceland (150K individuals with genotyping array data) [84]; the China Kadoorie Biobank (500K individuals, with genetic data for 100K) [85], and the Million Veterans Program in the United States (currently 600K participants, with genetic data for 200K) [86] and the Genetic Epidemiology Research on Adult Health and Ageing study (GERA) (100K individuals with genotyping array data) [87]. Other, commercially-driven collections of similar scales also exist, such as 23&Me [88].

These collections are designed to be multi-purpose. That is, intended to be used for a wide variety of research objectives, from conventional phenotype-genotype association discovery (i.e. GWAS), to the development of better methodology for analysing such data [89]. While these ambitious collections promise many new opportunities for discovery, they also come with several new challenges. The scale of the genotyping experiments means that the necessary laboratory processes can take years to complete, multiplying the chances of experimental confounding factors (e.g. changes in chemical reagents, or personnel handling the samples). Once the genomic data is generated, there is the bioinformatic challenge of storing, handling, and analysing the high volumes of data (typically tens of Terabytes for genotyping array data). The multi-purpose function of these data poses the challenge of how to present it to researchers so that it meets (or can potentially meet) the standard of data quality required for different kinds of analyses. Finally, in the context of very large cohorts and the analysis of rare variation [90], the empirical magnitude and impact of familiar confounding factors — such as population structure [91, 92] and familial relatedness [93] — remains largely unexplored.

1.3.2 The UK Biobank project

The UK Biobank project is a large prospective cohort study of $\sim$500,000 individuals from across the United Kingdom, aged between 40-69 at recruitment [94]. A rich variety of phenotypic and health-related information is available on each participant, making the resource unprecedented in its size and scope. The data contains

self-reported information, including basic demographics, diet, and exercise habits; extensive physical and cognitive measurements; with other sources of health-related information such as medical records and cancer registers being integrated and followed up over the course of the participants' lives [95]. The baseline information has, and will be, extended in a number of ways [96]. For example, many blood and urine biomarkers are being measured; and medical imaging of brain [97], heart, bones, carotid arteries and abdominal fat is being carried out on a large subset ($\sim$100,000) of participants [98].

A key component of the UK Biobank resource has been the collection of genome-wide genetic data on every participant using purpose-designed genotyping arrays [99]. An interim release of genotype data on $\sim$150,000 UK Biobank participants (May 2015) [100] has already facilitated numerous studies [101]. These exploit the UK Biobank's substantial sample size, extensive phenotype information, and genome-wide genetic information to study the often subtle and complex effects of genetics on human traits and disease, and its potential interactions with other factors [102, 103, 104, 105, 106, 107].

The genetic data for 488,377 UK Biobank participants genotyped at 805,426 unique markers (SNPs and indels) was made available to approved researchers in July 2017. The genotyping was carried out by Affymetrix Research Services Laboratory, who produced the primary data files: genotype calls and intensity data (see Section A.1). In the next section we cover details of the array design and the laboratory and bioinformatic processes that were carried out before the data were passed on to us for further quality control and analysis.

1.3.3 The UK Biobank genotyping experiment

1.3.3.1 Two novel genotyping arrays

The genotypes of $\sim$490,000 UK Biobank participants[18] were assayed using two very similar genotyping arrays. A subset of $\sim$50,000 participants included in the UK

[18]Precise numbers of samples and markers at different stages of the experiment are shown in Table A.5

Biobank Lung Exome Variant Evaluation (UK BiLEVE) study[19] were genotyped using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix (807,411 markers), which is described elsewhere [107]. Following this, the other $\sim$440,000 participants were genotyped using the closely-related Applied Biosystems UK Biobank Axiom Array (825,927 markers). Both arrays were designed specifically for the UK Biobank genotyping project and share 95% of marker content [99]. The marker content of the UK Biobank Axiom array was chosen to capture genome-wide genetic variation (single nucleotide polymorphism (SNPs) and short insertions and deletions (indels)), and is summarised in Figure 1.11. Many markers were included because of known associations with, or possible roles in, phenotypic variation, particularly disease. A notable example is the inclusion of two variants, rs429358 and rs7412, which define the isoforms of the apolipoprotein E (APoE) gene known to be associated with risk of Alzheimers disease [99] and other conditions. Neither marker is easy to type using array technologies; as a consequence of this they have not always been assayed on earlier arrays. The array also includes coding variants across a range of minor allele frequencies (MAFs), including rare markers ($<$1% MAF); and markers that provide good genome-wide coverage for imputation in European populations in the common ($>$5%) and low frequency (1-5%) MAF ranges.

[19]The UK BiLEVE project, for which the UK BiLEVE array was designed, aims to study the genetics of lung health and disease, and so these individuals were selected based on lung function and smoking behaviour from participants with self-declared European ancestry [107]. Otherwise, the UK BiLEVE cohort and the rest of UK Biobank differ only in small details of the DNA processing stage (e.g. UK BiLEVE samples were manually transferred from storage to plates for DNA extraction [108]).

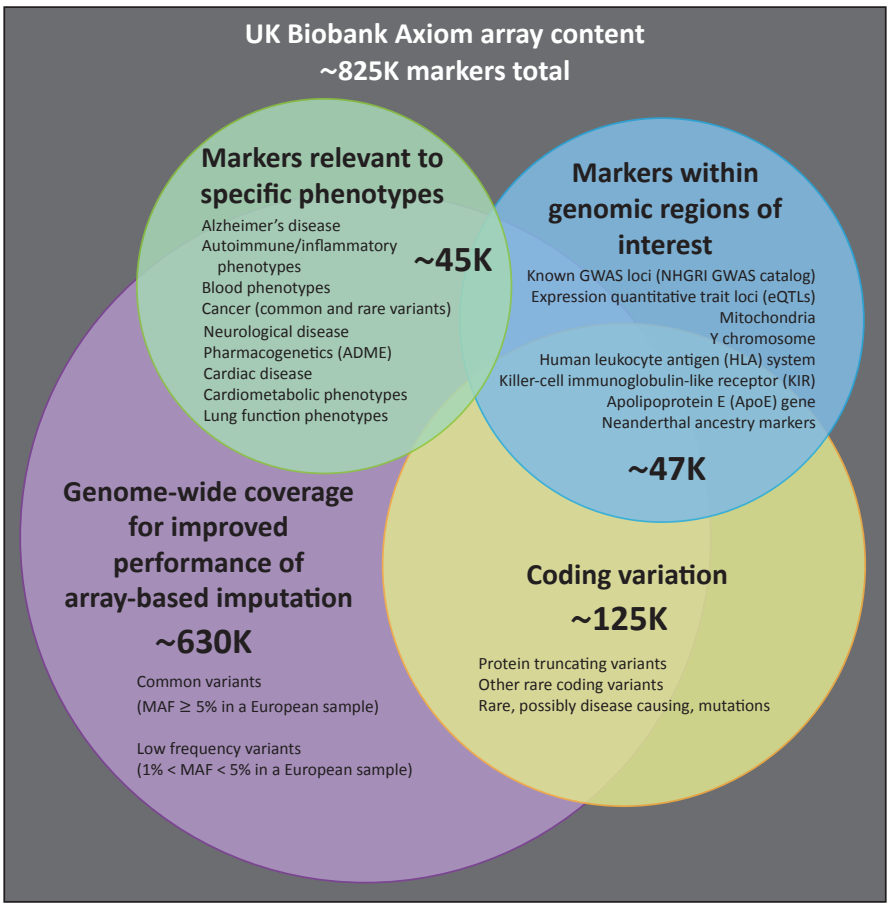


Figure 1.11: Summary of UK Biobank genotyping array content. This is a schematic representation of the different categories of content on the UK Biobank Axiom array. Numbers indicate the approximate count of markers within each category, ignoring any overlap. A more detailed description of the array content is available in [99].

1.3.3.2 DNA extraction and genotype calling

UK Biobank staff collected blood samples from UK Biobank participants on their visit to a UK Biobank assessment centre and the samples are stored at the UK Biobank facility in Stockport, UK [109]. Over a period of 18 months (Nov. 2013 – Apr. 2015) samples were retrieved, DNA was extracted, and 96-well plates of 94 50µl aliquots were shipped to Affymetrix Research Services Laboratory for genotyping. Special attention was paid in the automated sample retrieval process at UK Biobank to ensure that experimental units such as plates or timing of extraction did not correlate systematically with baseline phenotypes such as age, sex, and ethnic background, or the time and location of sample collection. Full details of the UK Biobank sample retrieval and DNA extraction process are described in [108, 110].

On receipt of DNA samples, Affymetrix processed samples on the GeneTitan Multi-Channel (MC) Instrument in 96-well plates containing 94 UK Biobank samples and two control samples from the 1000 Genomes Project [111], which were assigned to the same well positions on each plate: HG00097 to well A12 and HG00264 to well E12. Genotypes were then called from the array intensity data batches, which consist of multiple plates. Across the entire cohort, there were 106 batches of $\sim$4,700 UK Biobank samples each ($\sim$4,800 including the controls). Eleven batches contain individuals typed on the UK BiLEVE Axiom array, and the other 95 batches contain individuals typed on the UK Biobank Axiom array (Table A.4).

Affymetrix assays genetic markers using probe sets: a set of probes targeting a particular marker. The fluorescence intensity of two alleles is measured and used to infer an individual's genotype at the marker. Individuals with the same genotype at any given marker will cluster together in a two-dimensional intensity space (one dimension for each targeted allele). The technical details of Affymetrix's laboratory process are described in [112], and details of the genotyping calling routine specific to the UK Biobank project are available in [113]. Briefly, genotype calling involves inferring properties of these clusters within each batch and assigning each sample a genotype (or leaving the call missing) based on its position in intensity space. Figure 1.12 shows the intensities and genotype calls for an example marker.

Routine quality checks were carried out during the process of sample retrieval, DNA extraction [108], and genotype calling [112]. Any sample that did not pass these checks was excluded from the resulting genotype calls.

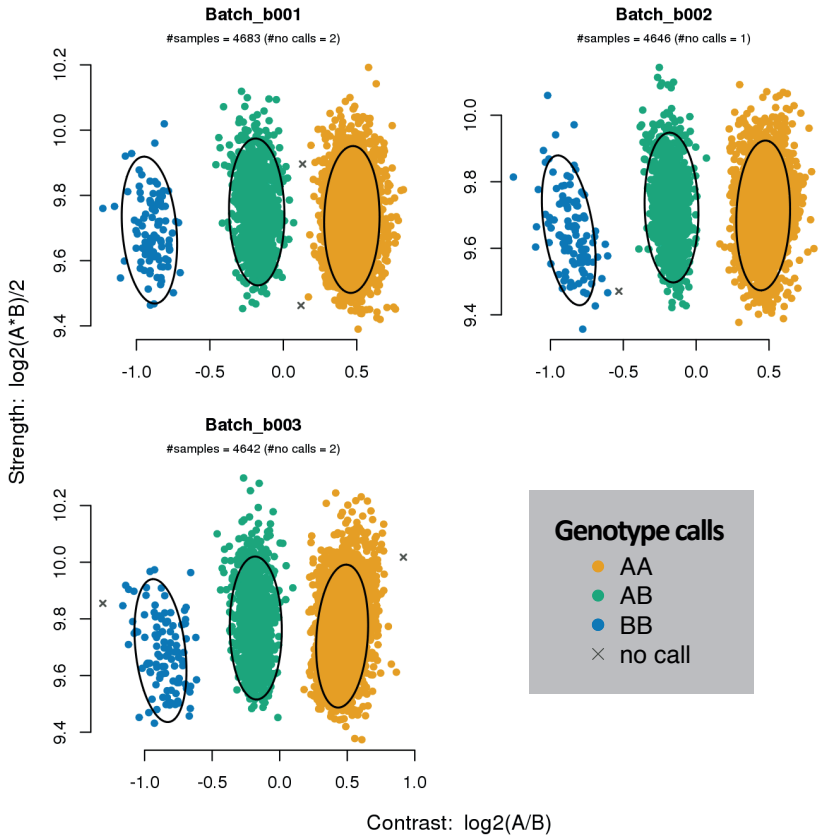


Figure 1.12: Example of intensity data and genotype calls for a marker in three batches. Each point represents one sample and is coloured according to its inferred genotype at this marker. The x and y axes are transformations of the intensities for probes targeting allele 'A' and allele 'B'. The ellipses indicate the location and shape of the posterior probability distribution (2-dimensional multivariate Normal) for each genotype cluster, such that 85% of the probability density falls inside the ellipse. More details of Affymetrix genotype calling for the UK Biobank is available in [113].

1.3.3.3 Filtering by Affymetrix

The purpose-designed UK Biobank Axiom array attempts to assay a large number of markers (SNPs or indels) that have not been previously genotyped using Affymetrix technology. In order to maximize the chances of such markers being successfully assayed, some were typed using more than one probe set. For each of these markers Affymetrix selected a single probe set that had performed best across all batches, and provided only the genotype calls made using that probe set.

Affymetrix also applied a series of checks, largely based on properties of the intensity signal, to determine whether the genotyping assay for a given marker was successful,

either within a single batch, or across all samples. If a marker did not meet Affymetrix's success criteria in a given batch, it was set to missing for all samples in that batch. If a marker did not meet the success criteria across all or many batches, it was excluded from the data delivery altogether. In addition, some markers assayed on the array were known, or suspected to have, more than two different alleles in humans. Such multi-allelic markers require special treatment in array design and genotype calling, and these were also excluded from the data delivery. More information about the Affymetrix calling algorithms and filtering protocols is available in [113, 112].

This filtering resulted in a set of genotype calls for 489,212 samples at 812,428 unique markers (bi-allelic SNPs and indels) from both arrays, which we used to conduct further quality control and analysis.

1.3.4 The aims of the analyses in Chapter 4

As outlined above (Section 1.3.3.3), QC procedures are included in Affymetrix's laboratory and bioinformatic processes. However, these tend to be generic in nature so that they can be applied to any study using the same, commercially-available technology. It was therefore necessary to carry out additional QC that takes into account aspects of the data specific to the UK Biobank genotyping project, such as the ancestral backgrounds of the participants, and the likely uses of the data by the research community (e.g. GWAS). Such additional QC is routinely advised by Affymetrix [114], and in the case of the UK Biobank project, carrying it out centrally saves different members of the research community from repeating the same analyses. Applying quality control was also a prerequisite to adding millions more markers to the UK Biobank genetic data set via imputation[20].

To this end, the author of this thesis, along with two colleagues, Dr. Colin Freeman (C.F.) and Dr. Desislava Petkova (D.P.), designed and implemented a QC pipeline[21] that addresses challenges specific to the experimental design, scale, and diversity

[20] The imputation was carried out by Professor Jonathan Marchini and the details are discussed elsewhere [115].

[21] All three of us were involved in the design and implementation of the pipeline as applied to the interim release data for the full cohort of ∼150,000 UK Biobank samples. C.F and the author of this thesis adapted and augmented the original pipeline, which we then applied to the data for ∼500,000 UK Biobank samples. The methods and results for the latter version is what is discussed in this thesis.

of this dataset. In Chapter 4 we discuss the elements of the pipeline that primarily designed and implemented by the author of this thesis: the sample-based QC. It will be necessary to discuss other elements of the pipeline within this chapter, and wherever possible we will be clear about which parts were designed and/or implemented by D.P. or C.F.

In addition to the QC, we conducted a set of analyses that reveal properties of the genetic data – such as population structure and relatedness – that can be important for downstream uses. Finally, after applying the QC pipeline and other analyses, we performed genome-wide association scans on a well-studied, and highly polygenic phenotype: standing height. These provided a further test of the effectiveness of the QC, as well as highlighting the potential of the resource to uncover novel regions of association.

Chapter 2

Fine-scale genetic structure in the Spanish population

2.1 Chapter overview

In this chapter we study fine-scale population structure in the Iberian peninsula, focussing on Spain. The primary data set for our analyses is high density genotyping array data for a cohort of 1,548 Spanish individuals, sampled from geographically diverse locations across Spain. These data were originally collected for a genome-wide association study (GWAS) [30], and were made available to us through a collaboration with the University of Santiago de Compostela in Spain[1]. A set of 1,413 samples passed our QC, and to these data we applied a powerful, haplotype-based method, *fineSTRUCTURE* [41], which utilizes phased haplotype data to cluster individuals into groups with similar patterns of shared ancestry. As noted in our introduction chapter (Section 1.2.2.1), the method has been used successfully to detect population structure at sub-national geographic scales in the British Isles [42]. The method identifies a set of discrete population groups, without prior specification of the number of groups, or their geographic origin. For a subset of individuals in the Spanish cohort, fine-scale geographic information about

[1] We acknowledge the central role of Professor Ángel Carracedo in this collaboration. His full affiliations are: Galician Public Fundation of Genomic Medicine (FPGMX)-Grupo de Medicina Xenmica-Centro de Investigacin Biomdica en Red de Enfermedades Raras (CIBERer)-Universiy of Santiago de Compostela, Spain

grandparental birthplaces was available, and all four of their grandparents were born less than 80Km from the centroid of their birthplaces. We visualized the *fineSTRUCTURE* results by plotting these individuals (726) as a point on a map of Spain, located at the centroid of their grandparents' birthplaces and labelled according to their cluster assignment. Their grandparents were likely to have been born in the late 1800s (median birth-year of the cohort is 1941), so the spatial distribution of genetic diversity described here would reflect that of Spain around that time.

Within the *fineSTRUCTURE* framework, shared ancestry is measured as the total amount of the genome (in centiMorgans (cM)) for which individual i shares a common ancestor with individual j, more recently than all the other individuals in the sample. This is estimated for each pair of individuals i and j, defining a square matrix referred to as the 'coancestry' matrix. This matrix is then used to cluster individuals into groups with similar patterns of coancestry, i.e. similar rows and columns in the matrix. The method also builds a hierarchical tree, by successively joining pairs of clusters with the most similar patterns of coancestry. This approach is discussed in more detail in Section 2.3.

To our knowledge, this is the largest and most comprehensive study of Spanish population structure to date, and we demonstrate the existence of extensive fine-scale structure across Spain. Examining different layers of the hierarchical tree, along with properties of the coancestry matrix allows for quantitative analysis of the relationships among the inferred clusters. By combining this with the geographic locations of the samples, we are able to form a rich description of the patterns of genetic differentiation in Spain, and relate these back to historical demographic events within the region.

2.2 Data and quality control

2.2.1 Spanish cohort data

For the analysis discussed in this chapter we used genotype data that was originally collected and typed for a colorectal cancer GWAS [30]. Biological samples were sourced from a variety of hospitals across Spain as well as the Spanish National DNA Bank. All samples were assayed together by Affymetrix[2] in the same facility. Full details of sample collection and genotype calling are published elsewhere [30]. We used both cases and controls, which totalled 1,548 individuals prior to removing samples for quality control reasons. All 17 of Spain's Autonomous Communities were represented in this dataset (Figure 2.1), but the Canary Islands and the 'Autonomous Cities' of Melilla and Ceuta are excluded from analyses involving geographic labels due to limited sampling in these areas (4 samples).

Geographic information was available for the sampled individuals, along with their age at the time of data collection, sex, genotyping plate (controls only) and batch used in genotype calling. The geographic information includes region of origin (Autonomous Community) for all individuals, and for 953 individuals (65% of the total sample) the birthplace (municipality) of all four grandparents. Birth dates of grandparents were not available, but the median year of birth for the sampled individuals is 1941 (sd 12 years), so their grandparents were likely to have been born in the early 1900s. We assigned each individual to a geographic coordinate by matching the text (e.g. 'Barcelona') to a municipal region as defined by the Spanish Statistical Office and coding them to the geographic centre of the matching region. Some locations were not themselves a municipality, so we coded these individuals to the centre of the nearest municipality, identified by using Google Maps. In order to maximise power to detect structure in the *fineSTRUCTURE* analysis we included the individuals for whom the exact birthplaces of their grandparents was not known, but these are not used in analysing the spatial distribution of the inferred genetic structure. For that, we used a subset of individuals (726) for whom all four of their grandparents were born less than 80 Km from the

[2]Affymetrix is now called Thermo Fisher Scientific. We will refer to it as Affymetrix throughout, as this was its commercial name at the time the data used in this thesis was generated.

centroid of their birthplaces (following [42]). The locations of these samples are shown in Figure 2.1. Note that this represents a substantial fraction (76%) of all Spanish individuals for whom there was fine-scale geographic information, despite the fact that localisation of grandparental birthplace was not a requirement of the original study.

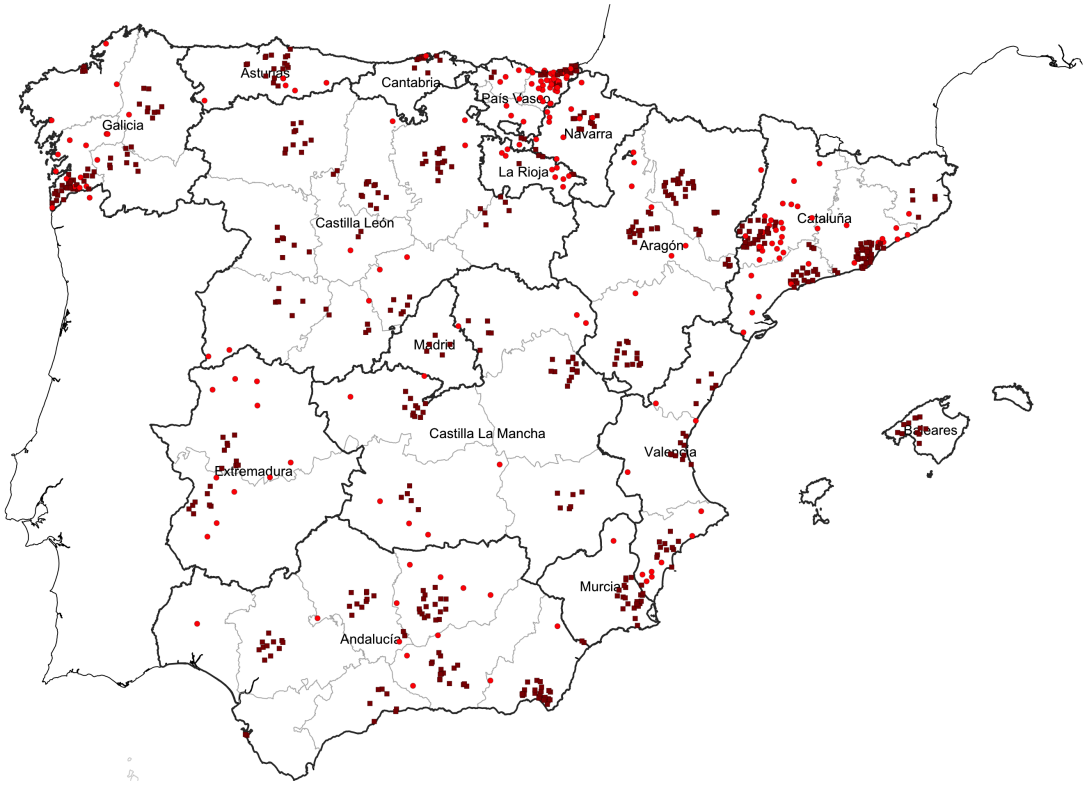


Figure 2.1: Geographic locations of individuals in Spanish cohort. Each point represents one of 726 individuals included in the map-based visualisations, with locations based on the birthplaces of each their grandparents (see 2.2.1). Dark red squares indicate individuals whose locations have been jittered slightly to aid visualisation; red circles indicate precise locations. The boundaries of Spain's Autonomous Communities and provinces (unlabelled) are also shown. Only 4 samples in total were located in the Spanish territories of Canary Islands, Melilla and Ceuta (not shown on this map). These regions are not strictly within the Iberian Peninsula and were excluded from the geographic-based analyses.

The samples were genotyped on the Affymetrix 6.0 array (∼900,000 markers) and we used the output of the genotype calling algorithm Birdseed [116] as the starting point in this analysis. This data contains genotype calls for every marker on the array, along with a measure of certainty for each call. The genotype calls from Birdseed were coded using a mixture of forward and reverse strands, so for compatibility with other data sets used in later analyses, we first flipped all SNPs to the positive strand and converted SNP positions to genome build hg19 coordinates. This could be done with certainty by

using the appropriate Affymetrix SNP annotation file for the Affymetrix 6.0 chip [117] (Release 34), and using the software Birdsuite [116].

2.2.2 Quality control and phasing

We use the quality control procedures applied to the Wellcome Trust Case Control Consortium 2 datasets [118] as a guide to quality control for array-based genotype data, with some adjustments to suit particular properties of this dataset. We first set to missing all genotype calls with a 'confidence'[3] value greater than 0.1, and then excluded SNPs that had one or more of the following properties:

- p-value for departures from Hardy-Weinberg equilibrium $< 10^{-20}$.
- Minor allele frequency < 0.01.
- Missing rate > 0.02.
- On the Mitochondrial genome, Y, or X chromosome.
- Its rsID did not map to genome assembly build hg19.
- Did not match the strand of the reference panel used in phasing.

We also excluded samples with one or more of the following properties, where all metrics were calculated after excluding SNPs as described above.

- Genome-wide missing rate > 0.009.
- At least one grandparent born outside Spain (if information available).
- Outliers (7 samples) in the first two principal components (see Figure A.2).
- Kinship coefficient > 0.1 (excluded one from each related pair), computed using the software *KING* [119].
- Showed signs of poor quality genotyping after running *fineSTRUCTURE* (discussed in Section 2.3.3).

[3]The genotype calling 'confidence' metric refers to the probability that the true genotype is *not* the one called. Thus (somewhat confusingly) values close to zero indicate good quality.

After applying the above quality control filters (using *qctool* v1.4 [120]) there remained 1,413 samples and 693,092 SNPs for further analysis.

2.2.3 Phasing quality

We phased the quality-filtered genotype data using SHAPEIT (v2) [44] with a reference panel and genetic map from 1000 Genomes Project Phase I [111]. Related samples were included in the phasing step, but excluded from the *fineSTRUCTURE* analysis. There are no trios in this data with which to directly check the quality of the phasing. However, using a comparable data set based on the same Affymetrix array used here, the authors of SHAPEIT (v2) estimated the average sequence length between two consecutive phase-swith errors to be 1.37 Mb in a sample of 500 individuals [44]. This is much longer than the average length of 'copied chunks' (see Section 2.3.1) observed in our (and others [42]) *fineSTRUCTURE* analysis, which is around 0.5 cM. Furthermore, in the application of *fineSTRUCTURE* some phase-switch errors would be tolerated because it uses a genome-wide measure of ancestry sharing, which is aggregated across an individual's two haplotypes. This is potentially an issue in studying very recent admixture (1–3 generations), where contiguous chunks inherited from different ancestral groups would be expected to be Mbs long, but is beyond the aims of this chapter.

2.3 Details of *fineSTRUCTURE* analysis and data visualisation

2.3.1 Measuring haplotype sharing among individuals (coancestry)

We inferred clusters of individuals based on genetic data only by applying the *fineSTRUCTURE* method [41]. The method uses a model-based approach to cluster individuals with similar patterns of shared ancestry, as measured by the 'coancestry matrix', which we describe here. *fineSTRUCTURE* first applies a hidden Markov model (HMM) along the genome, which is constructed under the Li and Stephens

copying model [43]. Consider a target haplotype i and a set of haplotypes from a sample of other individuals. The hidden state at each locus (SNP) is the haplotype j with which haplotype i shares a common ancestor which is the most recent among all other haplotypes (at that locus) [41]. This is referred to as haplotype i 'copying' from haplotype j. In the context of *fineSTRUCTURE* this process is known as 'chromosome painting', and is implemented in the software *CHROMOPAINTER*. A haplotype of interest is referred to as a 'recipient', and all other haplotypes in the sample are known as 'donors'.

Genome-wide summaries of the HMM yield measures of haplotype sharing between one individual and all the others, and can be computed efficiently using standard algorithmic apparatus of HMMs [41]. *CHROMOPAINTER* implements two main measures of haplotype sharing, which are expectations over all possible paths through the set of hidden states. The first is the expected number of chunks (i.e. contiguous sets of SNPs) of the genome for which an individual i copies from individual j; the second is the expected total amount of genome (measured in cM), for which an individual i copies from individual j. In practice, for diploid individuals each of their (phased) haplotypes is treated separately in the HMM, but only haplotypes from other individuals can be donors. For each recipient individual the genome-wide measures of haplotype sharing are summed across their two haplotypes. Thus, each of these genome-wide measures forms a matrix of size NxM, where N is the number of recipient individuals and M is the number of donor individuals. The sets of recipient and donor individuals can (but need not be) the same set of individuals, but any element of the matrix where the donor and recipient individual is the same will always take the value 0 because individuals cannot be donors to themselves. Another property worth noting is that coancestry matrices are not necessarily symmetric. That is, the coancestry value corresponding to individual a copying from individual b is not necessarily the same as the value corresponding to individual b copying from individual a.

2.3.2 Clustering Spanish cohort based on coancestry patterns

To apply *fineSTRUCTURE* to the Spanish cohort data (and to other data sets e.g. in Chapter 3), we first computed a coancestry matrix allowing all samples to be both donors and recipients. Using this square coancestry matrix, *fineSTRUCTURE* then applies a Markov chain Monte Carlo (MCMC) procedure to find a high posterior probability partition of individuals into a set of clusters. The number of clusters is not specified in advance, but rather estimated under the *fineSTRUCTURE* probability model. Having found a set of clusters, *fineSTRUCTURE* then infers a hierarchical tree by successively merging pairs of clusters whose merging gives the smallest decrease to the posterior probability (of the partition) among all possible pairwise merges [41]. This shouldn't be interpreted as a phylogeny describing population divergences, but rather describes a hierarchy of similarity in coancestry patterns between the different clusters. We applied *fineSTRUCTURE* using the procedure recommended by the authors, except in one aspect: we used the total amount of genome (in cM) for which an individual i copies from an individual j as a measure of coancestry instead of the default, number of chunks (see Section 2.3.3 for rationale).

We ran several *fineSTRUCTURE* analyses using phased haplotypes from the following sets of individuals:

(A) Spanish individuals

(B) Spanish and Portuguese individuals

(C) Non-Spanish individuals (only discussed in Chapter 3)

In all cases we used *CHROMOPAINTER* software (V2) [41] to estimate the coancestry matrix as described above. We ran *CHROMOPAINTER* using the same genetic map as we used in phasing. Two parameters are required for the HMM: the initial genome-wide average 'switch' rate (n) and global 'emission' rate (M). For analysis (A) we estimated these using 10 iterations of *CHROMOPAINTER*'s Expectation-Maximization (E-M) algorithm, across 10 individuals chosen from a variety of regions across Spain, but allowing all other Spanish individuals to be

donors. We then averaged the results over all chromosomes and the 10 individuals using the auxiliary program ChromoCombine. For each of analyses (B) and (C) we re-estimated these parameters in a similar manner, using a subset of samples from a variety of regions in each case. This procedure mimics that used by a previous successful application of *fineSTRUCTURE* [42]. All other parameters were set to the software defaults.

In all cases, we computed the c-factor parameter required for *fineSTRUCTURE*'s MCMC algorithm as described in Section 2.3.4. Also, for all cases we used 500,000 burn-in iterations and 1,000,000 subsequent iterations, and stored the results from every 10,000th iteration. The MCMC was followed by 100,000 hill-climbing moves before applying *fineSTRUCTURE*'s tree-building algorithm. In analysis (A) we also conducted an iterative procedure after the MCMC and hill-climbing iterations to further refine the final set of clusters at the bottom of the tree. This procedure is described formally in [42]. Informally, it uses an iterative procedure to reassign individuals to a new set of clusters, such that individuals that are often co-clustered across multiple MCMC samples share a cluster assignment. We used the last 50 stored iterations of the MCMC, as these showed no signs of increasing posterior probability (see Figure A.3). We also checked that the MCMC samples were independent of the algorithm's initial position by visually comparing the results of two independent runs starting from different random seeds. Good correspondence in the pairwise coincidence matrices of the two runs indicates convergence of the MCMC samples to the posterior distribution [41] (see Figure A.4). We used the first of these two runs in the main analysis.

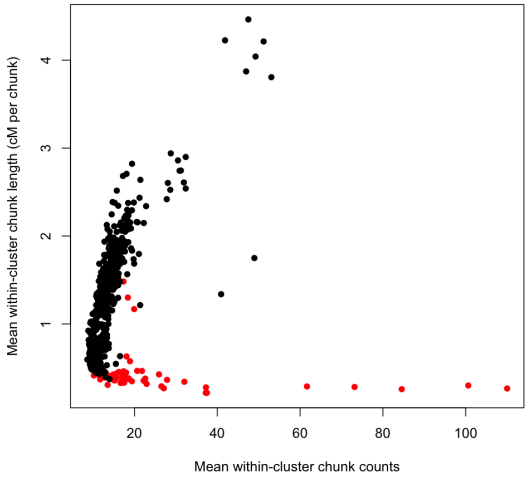
2.3.3 Rationale for using total amount of genome as coancestry measure

Recall that *CHROMOPAINTER* computes two summaries of haplotype sharing between individuals in a sample: the number of chunks of the genome for which an individual i shares its most recent common ancestor with individual j; and the total amount of genome (measured in cM), for which an individual i shares its most recent

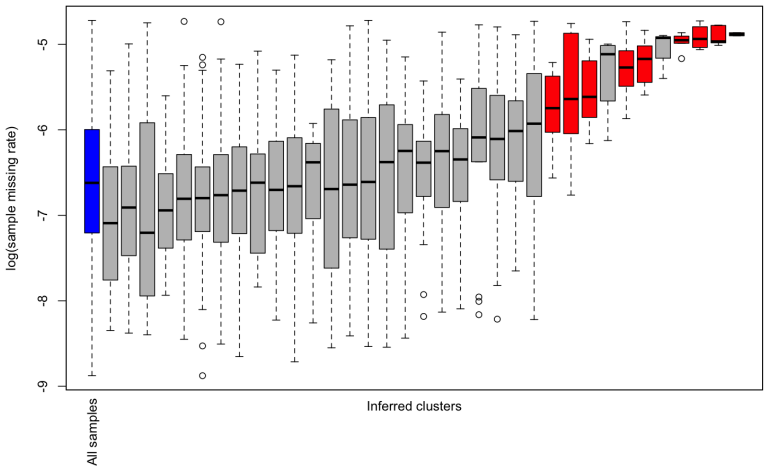
common ancestor with individual j. An observation made in [41] is that individuals with relatively high levels of recent common ancestry will tend to share more chunks (relative to other individuals in the sample), but these chunks will also tend to be longer, because there has been less time for recombination to break up the shared chunks. Therefore, if clusters inferred under the *fineSTRUCTURE* model are capturing groups of individuals that are genuinely more closely-related, we would expect a positive correlation between the number of chunks, and the average length of the chunks shared between individuals within clusters. We exploited this property to identify spurious cluster assignments that were likely a result of poor data quality. Intuitively, more genotyping errors would increase the number of 'switches' occurring in the HMM used to estimate the coancestry matrix, thus inflating the number of chunks copied. The total amount of genome coancestry measure should be less sensitive to this property because it is a sum over all chunks copied from the same donor (regardless of spurious switches in the HMM), so would still capture signals of real ancestry sharing despite the noise due to genotyping errors.

Figures 2.2a and 2.2b show results from a *fineSTRUCTURE* analysis using chunk counts as the coancestry measure. Specifically, there is a set of clusters with individuals who share high numbers of chunks relative to other clusters, but share unexpectedly short chunks on average. These same clusters also have significantly higher sample missing rates than other clusters (Figure 2.2b), a symptom of poor genotype quality. We found that this effect was reduced when using the total amount of genome as the coancestry measure (Figures 2.2c and 2.2d). In that case, the extent of chunk count inflation was much reduced, with only 19 individuals that showed evidence of spurious cluster assignment, compared to 89 when using chunk counts (with exactly the same algorithm parameters, except a different c-factor).

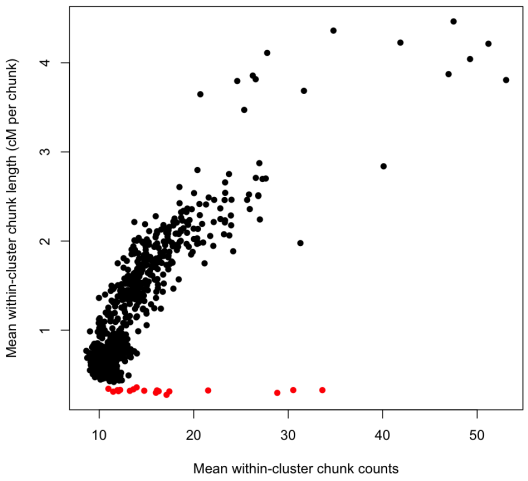
(a) Using number of chunks copied



(b) Using number of chunks copied



(c) Using total amount of genome copied



(d) Using total amount of genome copied

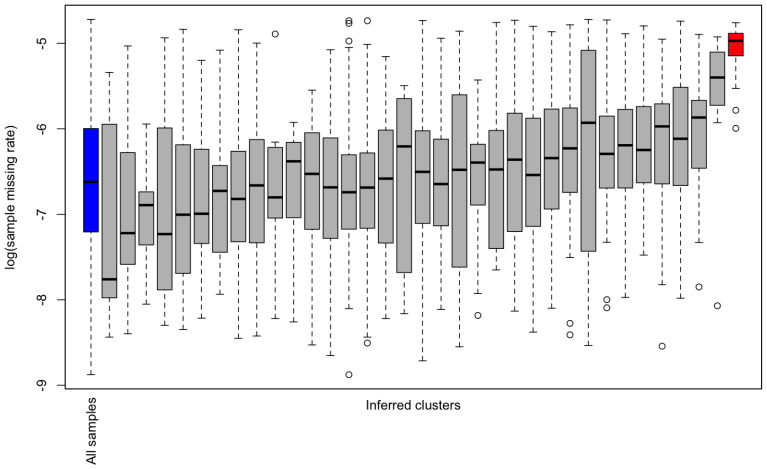


Figure 2.2: Effect of using total lengths verses chunk counts as coancestry measure in *fineSTRUCTURE* algorithm. We ran *fineSTRUCTURE* for the Spanish cohort using two different coancestry measures, and compared their robustness to genotype quality (as discussed Section 2.3.3). Plots **(a)** and **(b)** show results from a *fineSTRUCTURE* analysis using chunk counts as the coancestry measure. Plots **(c)** and **(d)** show results from a *fineSTRUCTURE* analysis using total amount of genome copied as the coancestry measure. For both runs we show these metrics for the level of the hierarchical tree with 35 clusters. The left-hand plots (a) and (c) show, for each individual, the mean chunk lengths (i.e. the average length of copied chunks), and mean number of chunks copied from other individuals inferred to be part of the same cluster. We expect these two measures to be positively correlated under conditions of real population structure [41]. The right-hand plots (b) and (d) show the distribution of genotype missing rates (on the log scale) for the samples in each of the inferred clusters. In all plots, the clusters with significantly higher missing rates from the overall cohort are shown in red ($p < 0.001$, one-sided t-test on log-transformed sample missing rates).

2.3.4 Estimating the *fineSTRUCTURE* c-factor

Recall that *fineSTRUCTURE* applies a MCMC procedure to find a high probability partition of individuals into any number of clusters. The posterior probability of a given partition is inferred using the following multinomial likelihood function (other details such as priors are in [41]):

$$F(x|p,q) = \prod_{i=1,j=1}^{N} \left(\frac{P_{q_i q_j}}{\hat{n}_{q_j}} \right)^{x_{ij}/c} \tag{2.1}$$

where x_{ij} is the measured coancestry between individual i and j; q_i and q_j denote the clusters that individual i and j are assigned to; $\hat{n}_{q_j}$ is the number of samples assigned to cluster q_j; and $P_{a,b}$ is a cluster-level coancestry matrix. That is, the proportion of coancestry donated to any individual in cluster a from any individual in cluster b.

This multinomial model implies that each unit of coancestry (either number of chunks, or cM) that contributes to the coancestry matrix is an independent draw from a set of possible outcomes, where an outcome is a particular donor haplotype. In practice, the independence assumption does not always hold, and so the coancestry matrix should be scaled by a factor of c to account for this. The authors of *fineSTRUCTURE* recommend estimating this value (and show that it is a good approximation to that predicted by theory [41]) by computing the ratio of the empirical variance of the coancestry values, to the theoretical variance assuming the data is generated from the multinomial distribution. The *fineSTRUCTURE* software automatically calculates c, but because we used a different measure of coancestry we also estimated it differently.

Specifically, we measured the empirical variance of the coancestry values in a similar fashion to that used by *fineSTRUCTURE* software by breaking up the genome into segments of equal length, and which are long enough to be approximately independent. Since we used total genome length as the coancestry measure we used segments of a length defined in cM rather than number of chunks. We used 40 cM as there would be sufficient recombination (by definition) between SNPs this distance apart such that linkage disequilibrium across segments would be minimal; and given

that the average length of chunks in the Spanish cohort tended to be about 0.5 cM[4], a segment size of 40 cM corresponds approximately to the size (100 chunks per segment) noted by the authors as working well in real human data sets. This distance is also small enough such that all chromosomes have at least one whole segment. We calculated coancestry values independently for each segment and then estimated the empirical and theoretical variance (assuming the multinomial model) as described in the Supplementary Material (page 30) of [41], with the simplification that the number of segments is the same for every individual. This is because all individuals have the same genome length ($\sim$3,600 cM per haplotype), whereas the total number of shared chunks can vary across individuals. In the case of analysis (A), the c-factor value estimated for total genome length was 1.17, compared to 0.39 for the number of chunks.

2.3.5 Map-based visualisation of clusters

Data visualisation is crucial for exploring relationships between genetic variation and the geographic dispersion of individuals. Having first derived geolocations for individuals, we used a spatial smoothing approach to visually represent this information together with the discrete assignment of individuals to clusters by *fineSTRUCTURE*. In figures showing a map of Spain each individual is represented by a point placed at the average coordinate (centroid) of their grandparents' birthplaces (coordinates were derived as described in Section 2.2.1). The points are coloured according to their assigned genetic cluster at the specified level of the hierarchical tree inferred by *fineSTRUCTURE*. Where many individuals have the same coordinate, such as in Barcelona, points have been randomly shifted (by no more than 24 Km) to aid visualisation, but the exact coordinates were used in determining the background colours (as described below). The map boundaries for Spain were downloaded as polygon files in longitude/latitude coodordinates from the Spanish official statistics website [55]. We transformed the polygons into the Lambert Azimuthal Equal Area (LAEA) projection centred on Madrid (40.5'N, 3.7'W) before

[4]This is very comparable the study of the British Isles [42], which also applied *fineSTRUCTRE* to genotype array data. There, the median chunk length was 0.51 cM with inter-quartile range 0.440.63 cM.

plotting. All maps were drawn using the 'spplot' function contained in the 'sp' package in *R* [121].

We coloured the backgrounds of the maps in Figures 2.3, 2.4, and 2.12, and others in Chapter 3 using the following spatial smoothing procedure:

1. The map is divided up into a fine, regular grid of 3 Km-wide squares, after projecting longitude/latitude coordinates into the above-mentioned coordinate system.
2. Each individual i shown on the map contributes an amount h_{si} to each square s on the grid, where h_{si} is computed by evaluating a normal density (mean=0, sd=3.5) at the Euclidean distance d_{si} between the centre of square s and the coordinate of individual i (in units of 10 Km). This is equivalent, up to a constant scaling factor, to evaluating a symmetric, bivariate Gaussian density with mean (0,0), and equal variance in both directions.
3. Weights are reduced by a factor of one quarter for individuals whose location falls further than 200 Km from the nearest individual in its assigned cluster. There were between 1 and 30 such individuals (depending on the level of the tree), and this prevents them from contributing disproportionate amounts of colour to regions of low sampling density.
4. The total contribution H_{ks} of cluster k to square s is the sum of the weights of individuals assigned to that cluster.
5. Multiple colours (one for each cluster) are applied to each square, with transparency according to the relative contribution of each cluster k.

Parameters for the above procedure, such as grid cell size and variance of the smoothing kernel have been chosen such that the visualization emphasises the geographical localities of genetic clusters without suggesting undue certainty about cluster boundaries.

We use a similar spatial smoothing procedure to visualise the geographic distribution of continuous variables measured for each individual (e.g. cluster assignment certainty

in Figure 2.7). Specifically, for each grid-point s we compute a weighted average (over all individuals) of the continuous variable, where the weights w_{si} are exactly those in Equation 3.4, as discussed in Chapter 3 Section 3.5.2.

2.4 The fine-scale genetic structure of Spain

2.4.1 Overall genetic structure of Spain

We inferred 145 clusters within the Spanish cohort using *fineSTRUCTURE* (analysis (A)) along with a hierarchical tree describing the relationships between the clusters. The maps in Figure 2.3 visualise the first 14 clusters, moving downwards from the top of the hierarchical tree. The coarsest level of genetic differentiation (i.e. two clusters) separates samples located in a small region in south-west Galicia from samples in the rest of Spain. The next level separates samples located primarily in the Basque regions in the north (País Vasco and Navarra) from the rest of Spain, although not exclusively — a point we will address later. Next, almost all the other samples located in Galicia are separated from the rest of Spain, followed by a further split within the cluster in south-west Galicia. The next split (red and yellow) isolates the central region of Spain from two flanking regions to the west and east. Curiously, this implies that the two geographically distant regions (in red) are more genetically similar to each other than they are to the central region (in yellow). We address this point later by considering properties of the coancestry matrix. Next, the cluster in the Basque regions splits into two, which broadly separates samples located in regions surrounding País Vasco, such as Navarra, from those located within País Vasco (again, with some exceptions). Next, the cluster (in red) containing samples in the west and far east of Spain splits cleanly into two (red and orange), with the eastern split corresponding closely to the Autonomous Communities of Cataluña and the Balearic Islands. Finally, further splits occur within the clade located in south-west Galicia, revealing remarkable sub-structure in Galicia, which we discuss in more detail later.

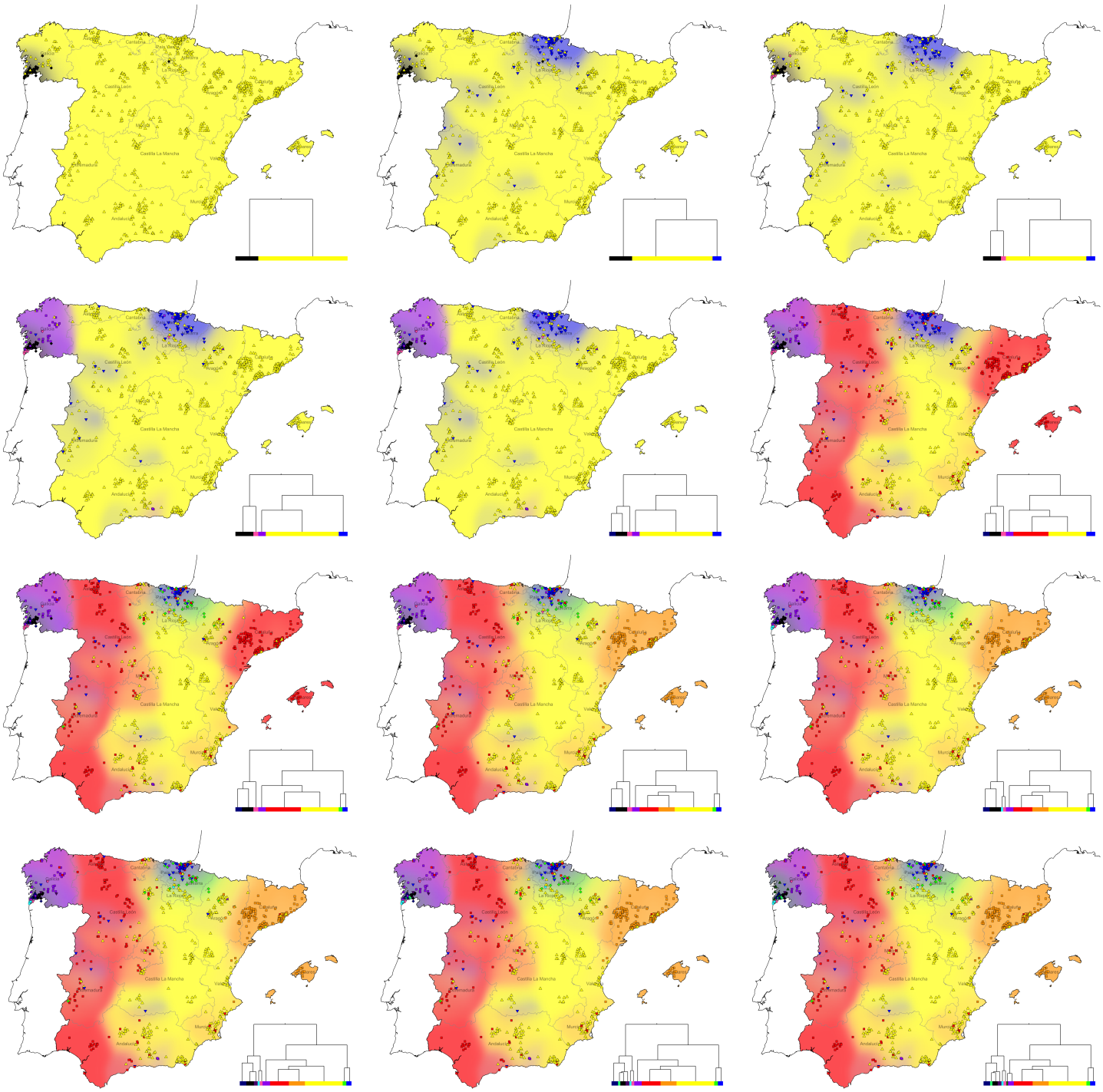


Figure 2.3: Results of *fineSTRUCTURE* clusters at different levels of the hierarchical tree down to thirteen clusters. Maps are shown in order of the tree topology (read left to right, top to bottom). At each successive level of the tree a binary split occurs, and the smaller of the two resulting clusters is assigned a new colour. Each map shows individuals represented by points placed at (or close to) the centroid of their grandparents' birthplaces (as in Figure 2.4). The background of each map is coloured according to the relative contributions of each cluster to each part of the map surface, at the level of the tree shown (see Section 2.3.5). The colour and symbol of each point corresponds to the cluster the individual was assigned to, and the trees show the hierarchical relationship between the clusters.

To aid visualisation of the lower levels of the hierarchical the tree, and because cluster assignments at higher levels of the hierarchical tree are typically more certain than lower-level clusters (as we discuss later in Section 2.4.2), we chose two different levels of the hierarchical tree to examine in more detail. There is no 'right' level of the tree to pick, but we chose them based on the sizes of the clusters and available geographic information (see Section A.2 for details). Figure 2.4 shows *fineSTRUCTURE*-inferred clustering of the Spanish individuals into 14 higher level (indicated by background colour) and 27 lower-level clusters (indicated by colours and shapes of points).

Many clusters closely align with geographic regions within Spain, and together reveal rich fine-scale population structure. Strikingly, at a broad scale the major axis of genetic differentiation in Spain runs from west to east, while conversely there is remarkable genetic similarity in the north-south direction (background colours in Figure 2.4). At the 14-cluster level and along the east-west direction, many cluster boundaries correspond closely to the east-west boundaries of Spain's Autonomous Communities, especially in the north of Spain: Galicia (purple), Asturias (red), País Vasco (blue), Cantabria (yellow), Aragón and Valencia (dark-blue), and Cataluña (orange). In contrast, in the north-south direction several clusters (yellow, red, dark-blue) span almost the entire peninsula, crossing the borders of several Autonomous Communities. This suggests that fine-scale population structure relates only partially to regions of Spain defined by modern political boundaries. Indeed the linguistic frontiers present in the Iberian peninsula around 1300 CE bear a remarkable resemblance to the pattern of genetic clusters we observe in modern Spain (compare Figure 2.4 with Figure 2.5d).

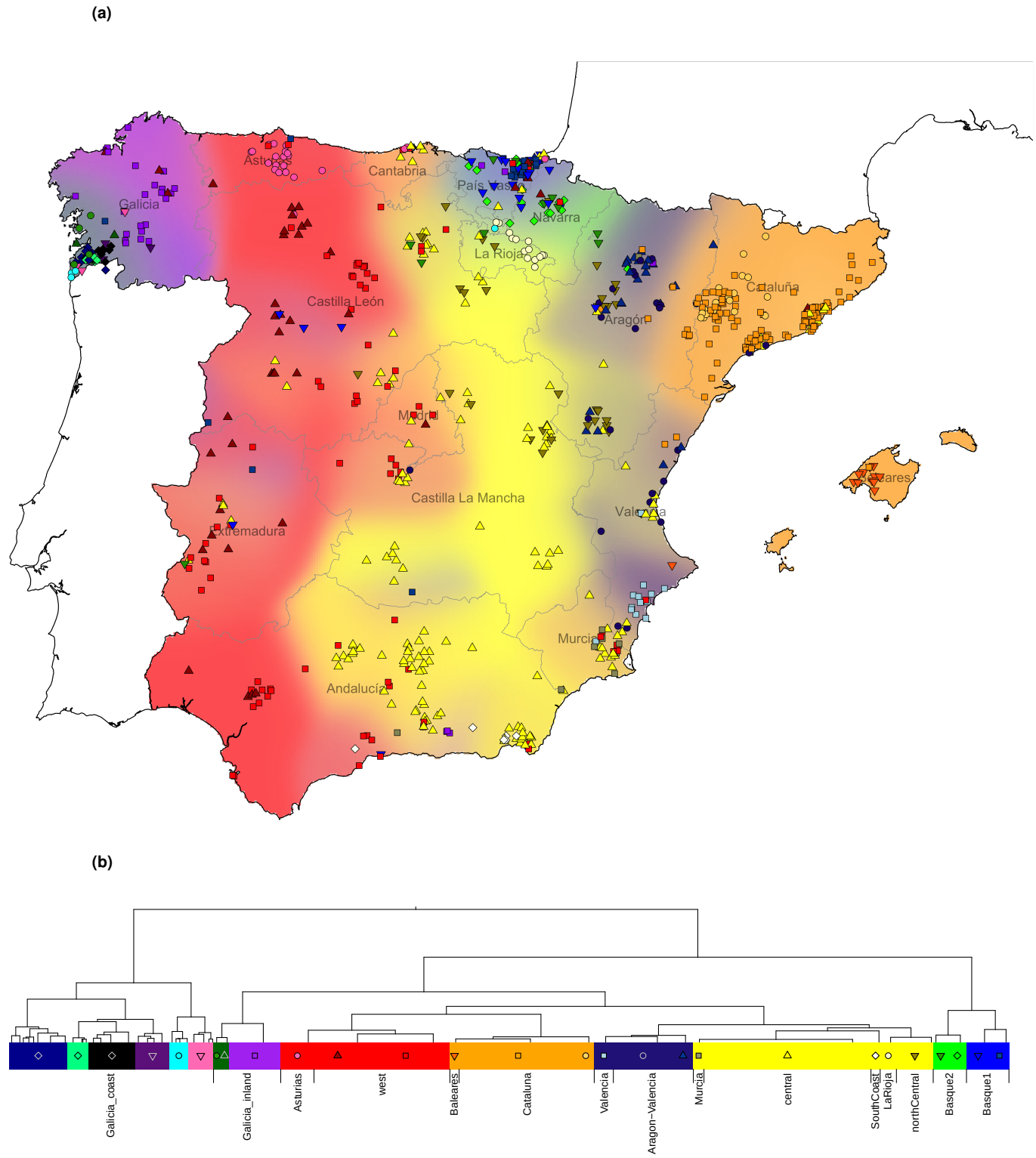


Figure 2.4: Spanish individuals grouped into clusters using genetic data only. (a) Each individual is represented by a point placed at (or close to) the centroid of their grandparents' birthplaces. On this map we only show the individuals for whom all four grandparents were born within 80km of their average birthplace. The background is coloured according to the relative densities of each cluster at the level of the tree where there are 14 clusters (see 2.3.5). The colour and symbol of each point corresponds to the cluster the individual was assigned to at a lower level of the tree, as shown in **(b)**. The labels and boundaries of Spain's Autonomous Communities are also shown. **(b)** Binary tree showing the inferred hierarchical relationships between clusters. The colours and points correspond to each cluster as shown on the map, and the length of the coloured rectangles is proportional to the number of individuals assigned to that cluster. We have labelled each cluster (or group of clusters) based on the general geographical location of the clusters, but no geographic labels were used in the inference of the clusters.

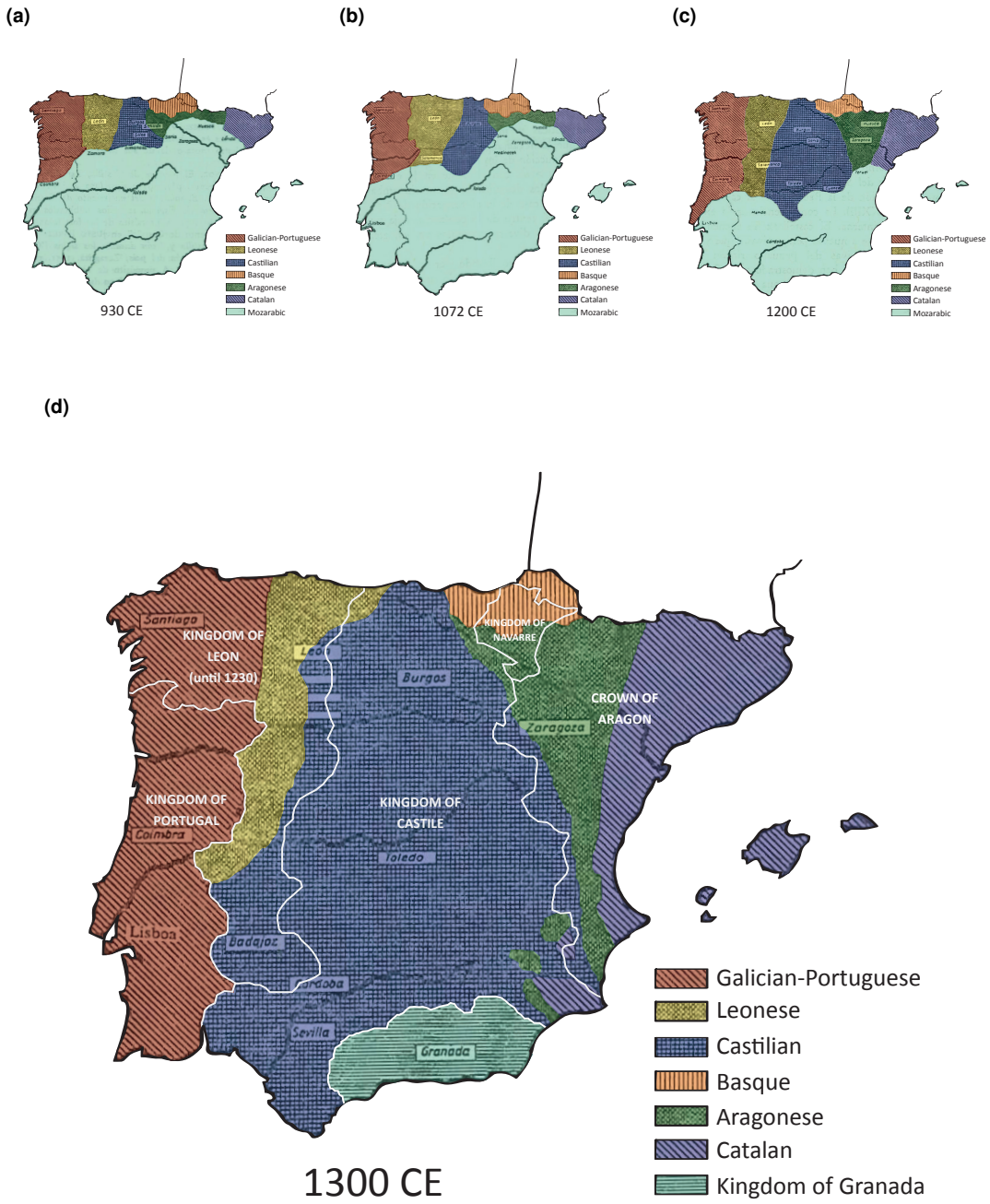


Figure 2.5: Changes in linguistic and political boundaries in Spain 930-1300 CE. The first three maps show changes in linguistic regions only, and the last map includes political boundaries. Different linguistic areas are shown with the colours and shading, and polilitical boundaries with white borders. These maps are adapted from maps in [69] (pgs. 48, 50, 52, 54, 57). That is, the colours and labels of the Christian kingdoms have been added to aid visualization. In 1230 the Kingdoms of Leon and Castile were united under Ferdinand III, king of Castile-Leon [57].

The lower level of the tree (colours and symbols in Figure 2.4) reveals greater geographical localisation. For example, this level identifies a cluster located almost exclusively in the region of Asturias (pink circles), and a cluster located almost exclusively in the Balearic Islands (red triangles), among others. There are two notable exceptions to the trend of more geographically localized clusters in the lower branches of the tree. Within both the yellow and red clusters in the centre of Spain (at level 14), even at the finest level of the tree (145 clusters) there still exists a large sub-group that spans the entire peninsula, north to south (Figure 2.6).

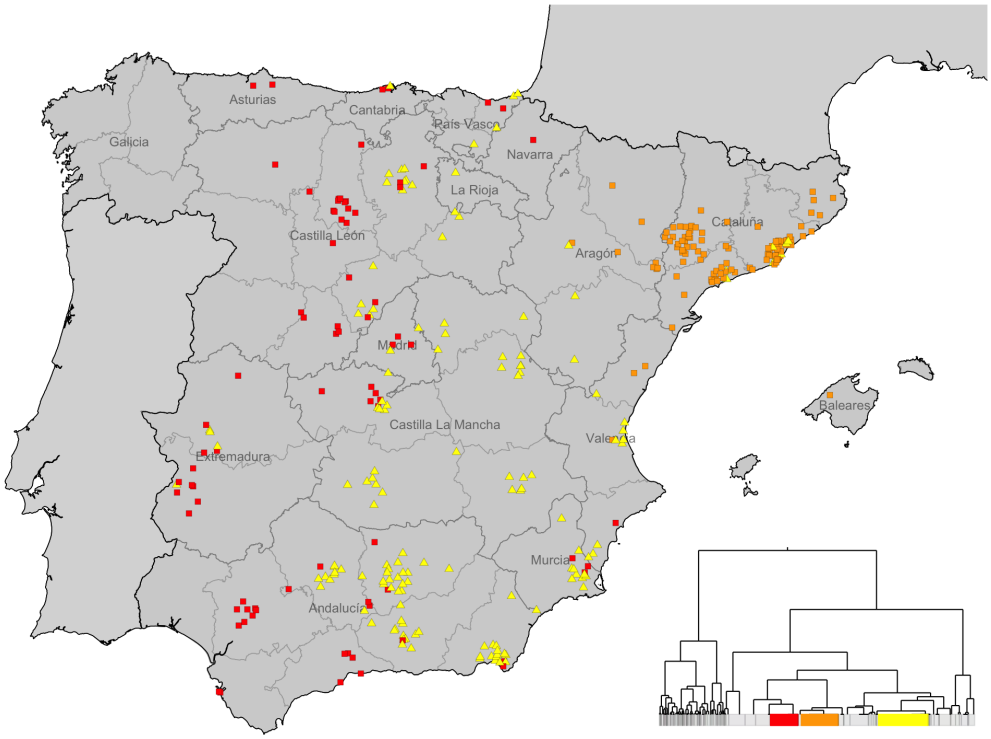


Figure 2.6: Large clusters at the bottom of the inferred hierarchical tree. The map shows geographic spread of the three large clusters that remain at the bottom of the tree inferred in the Spain-only *fineSTRUCTURE* analysis. Each cluster contains more than 100 individuals out of the full set of 1,413. The accompanying tree highlights the three clusters within the full tree structure.

2.4.2 The statistical certainty of cluster assignments

For analysis (A) we measured uncertainty in the assignment of individuals to clusters by using a procedure described formally in [42], which uses the information from multiple samples of the *fineSTRUCTURE* MCMC. Informally, the procedure measures the overlap between a cluster k, and individual i's assigned cluster in each of the

MCMC samples within a single *fineSTRUCTURE* run. This can take values between 0 and 1, and sums to 1 across all clusters for a given individual. It provides a measure of certainty about the assignment of individual i to each cluster k in the final set of clusters. The 'cluster assignment certainty' for an individual is the value corresponding to the final cluster assignment, and will be close to 1 if they are assigned to a cluster with largely the same set of individuals in each MCMC sample. This measure can be obtained for different layers of the hierarchical tree by summing the values for the clusters that merge at each layer.

For *fineSTRUCTURE* analysis (A), cluster assignments at higher levels of the hierarchical tree are typically more certain than lower-level clusters. At the broader level of the tree (14 background colours in Figure 2.4) 94% of individuals have a cluster assignment certainty greater than 0.7; and at the finer level (points in Figure 2.4), this level of certainty is reached by 89% of individuals (see Figure 2.7). Furthermore, the clusters with the highest certainty tend to be those with greater geographic localisation (e.g. 'LaRioja', 'Baleares').

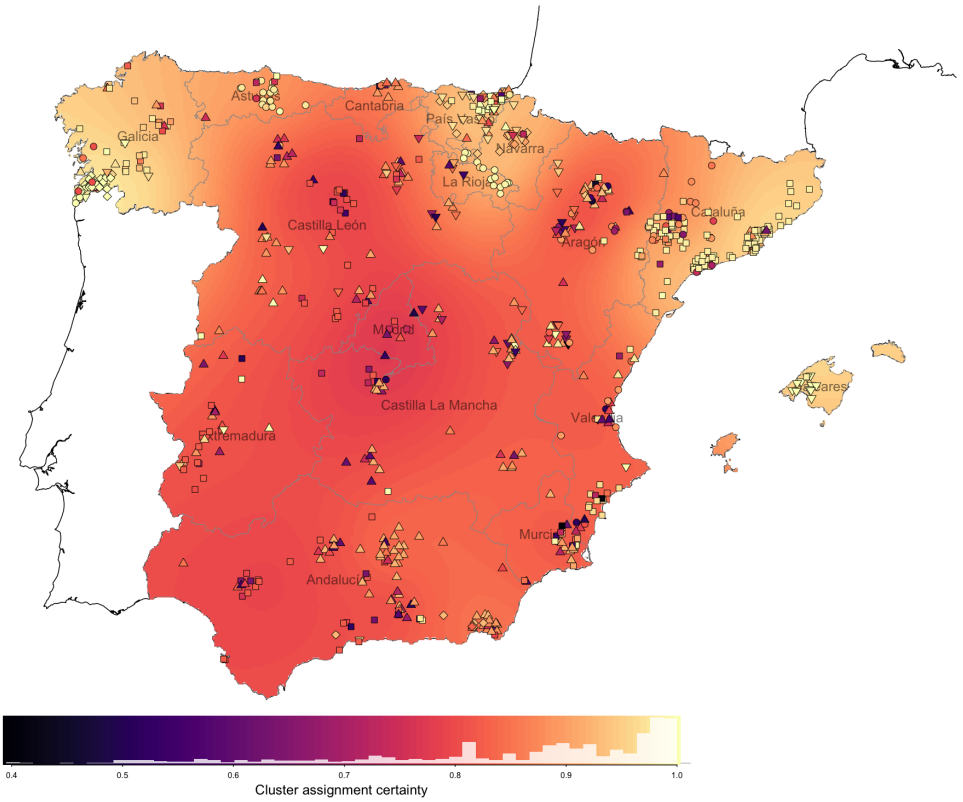


Figure 2.7: Cluster assignment certainty for analysis of fine-scale structure. For the finer level of the tree shown as points in Figure 2.4 (27 clusters), we computed a measure of cluster assignment certainty as described in [42]. It measures the co-clustering of individuals over multiple MCMC samples, and can take values between 0 and 1, where 1 indicates high certainty (see 2.4.2). Points have been coloured according to this certainty measure computed for each individual, and the symbols match those shown in Figure 2.4 to distinguish between different clusters. The histogram shows the distribution of the certainty measure for the individuals shown on the map. The background colour has been determined by applying a spatial smoothing algorithm to the same data (see Section 2.3.5).

2.4.3 Ultra-fine-scale structure in Galicia and elsewhere

Although fine-scale structure is seen generally, by far the strongest substructure exists within a single province in south-west Galicia (Pontevedra) (Figure 2.8a). This small province contains almost half of the inferred clusters in all of Spain, and 'ultra-fine' structure is seen across scales of less than 10 kilometres in some cases. For example, one cluster (black diamonds) stretches down a river valley, and another cluster (blue diamonds) is located on the other side of the mountain range to the north.

Sub-structure is also evident in regions associated with the Basque region, with four different clusters inferred within the Autonomous Communities of País Vasco and

Navarra at the lower level of the tree (Figure 2.8b). The clade shown in Figure 2.8b makes up the majority of all individuals located in the region of País Vasco, and a majority of this clade is located in this region. At the lower level of the tree this clade splits into four clusters, one of which is exclusive to a single province, Guipuzcoa (blue squares). Interestingly, there are also several individuals inferred to be part of the Basque-centred clusters, but whose grandparents were born in distant regions (e.g. blue triangles within Castilla León, Figure 2.4); this is less common for other highly-localised clusters. Another example of remarkably localised structure is the cluster labelled 'LaRioja' in Figure 2.4b. Individuals within this cluster are located exclusively in a $\sim$50km-wide region just south of País Vasco and Navarra. The cluster lies along a valley corresponding to a section of the Ebro river, which also marks the border between the Autonomous Communities of La Rioja and Navarra (Figure 2.8c).

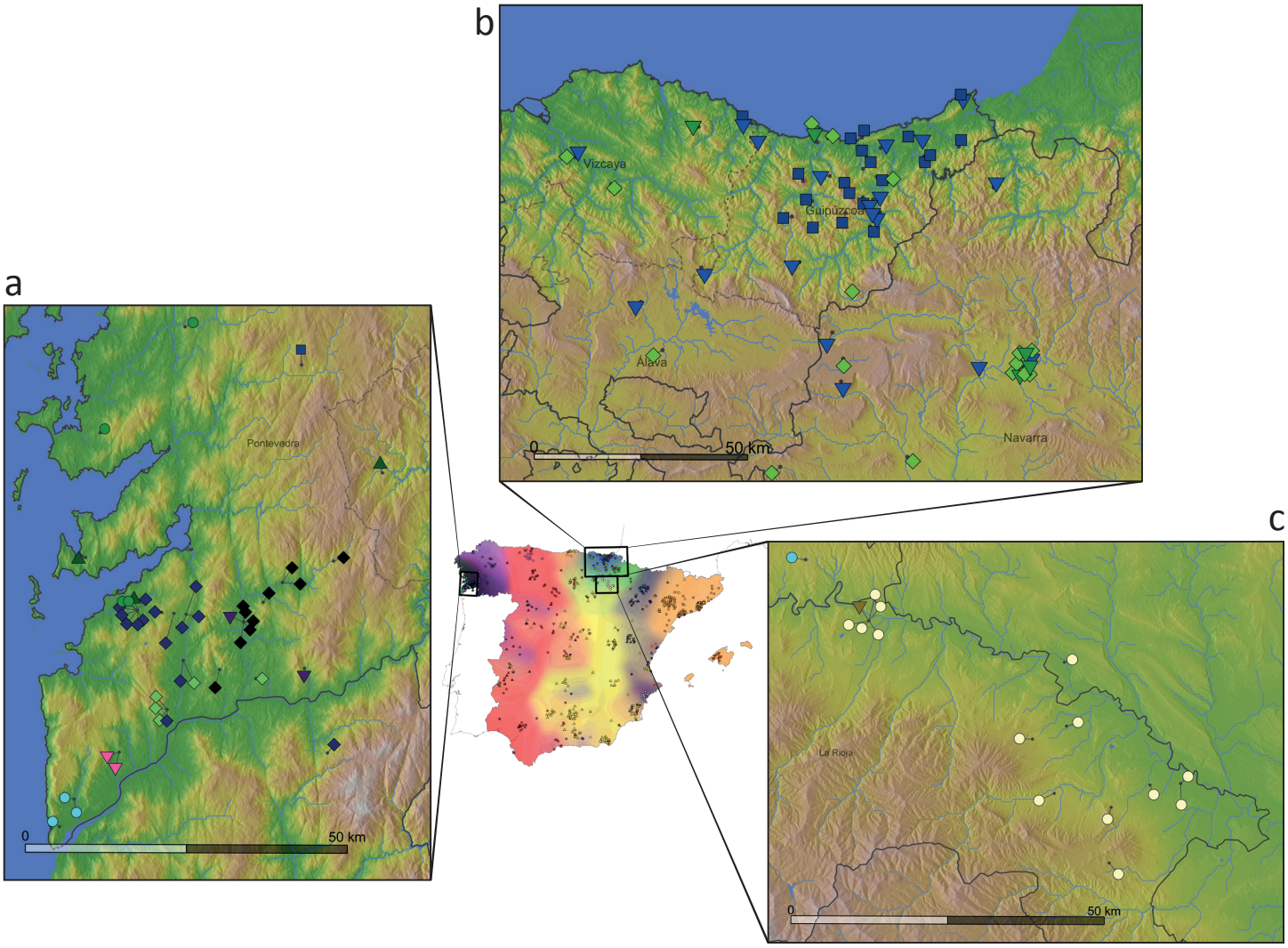


Figure 2.8: Ultra-fine-scale genetic structure within Spain. Points representing individuals are placed on each of the magnified maps with symbols and colours as in Figure 2.4, and with short dark lines pointing to their precise locations (the average birthplace of their grandparents). The three magnified maps show local elevation, rivers, and water bodies, as well as borders of Autonomous Communities (solid black lines) and provinces (dashed lines and text). The scale bars show distances that are accurate at the place they are positioned on each map, and at this scale will well-approximate distances elsewhere on the map (see Section 2.3.5). **(a)** Locations of individuals within the genetic clusters centred in Galicia. Note that we show this region at a higher level of the tree (14) as the lower level yields clusters with fewer than 3 individuals with sufficient geographic information for placing on the map. **(b)** Individuals within the clusters centred in the Basque-speaking regions of País Vasco (Basque Country) and Navarra. For visual clarity we only show the individuals that are within the clade coloured blue and green in Figure 2.4. **(c)** Locations of individuals in a cluster labelled 'La Rioja' in Figure 2.4b.

2.4.4 The effect of sampling density on fine-scale structure in Galicia

Many of the inferred clusters are located in the south-west region of Galicia, but relatively few clusters in the rest of Galicia. This difference can be seen by comparing the two clades labelled Galicia_inland and Galicia_coast in Figure 2.13. However, this region of Galicia was sampled more densely than other regions, which would provide greater power to detect structure. In order to test whether sampling density was a strong factor in the excess structure detected within the south-west region of Galicia, we conducted a complementary *fineSTRUCTURE* analysis after sub-sampling individuals from Galicia. Specifically, we used a subset of individuals such that the number of individuals in the set of clusters located in south-west Galicia (Galicia_coast) was the same as the number of individuals in the clusters located in inland Galicia ($\sim$100). We kept all the individuals shown on the maps, and then randomly sampled (evenly) from each of the six clusters within Galicia_inland to make up the remainder. We then re-computed the coancestry matrix, but with just 1,220 individuals, and ran *fineSTRUCTURE* in the manner exactly as described above for analysis (A), with one exception. In order to focus the analysis on Galicia only we used the 'continental force file' option (-F), where the forced population groups were each of the lower-level clusters not part of the clades labelled Galicia_coast or Galicia_inland (as in Figure 2.13). This option restricts *fineSTRUCTURE* to find clusters within the set of individuals *not* in the 'forced' population groups (i.e. the Galician clusters). Other individuals are allowed to be donors to the individuals of interest, but do not contribute to the parameter inference or the tree-building step (p.g. 11 of the *fineSTRUCTURE* software manual). This allowed for the possibility that structure within Galicia might be driven by different relationships with groups outside Galicia, while at the same time focussing the analysis only on the behaviour of the Galician clusters. Results of this analysis are shown in Figure 2.9. There is good correspondence with the cluster assignments between the two analyses at the level of the tree that is shown in Figures 2.4-2.13, confirming that the excess of fine-scale structure in south-west Galicia is still detectable, even under a more even sampling scheme.

(a)

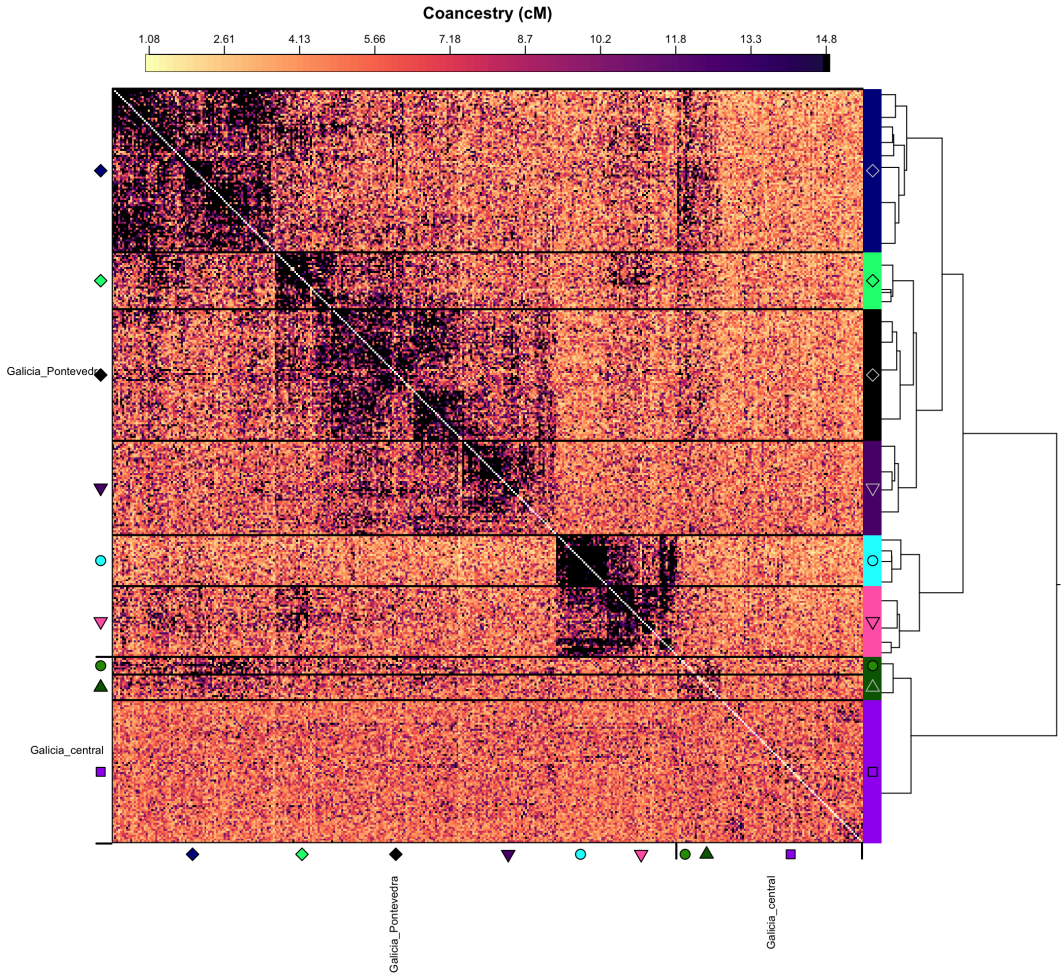


Figure 2.9: Effect of sub-sampling on fine-scale structure in Galicia (continued on next page). We tested the effect of high density sampling in the region of south-west Galicia by conducting a *fineSTRUCTURE* analysis on a subset of individuals such that the number of individuals in the set of clusters located in south-west Galicia, labelled in Figure 2.9a as Galicia_coast, was the same as the number of individuals in the clusters located in inland Galicia (see 2.4.4). **(a)** Section of the coancestry matrix (and the pruned tree) shown in Figure 2.13 that involves only the clusters located primarily in Galicia.

(b)

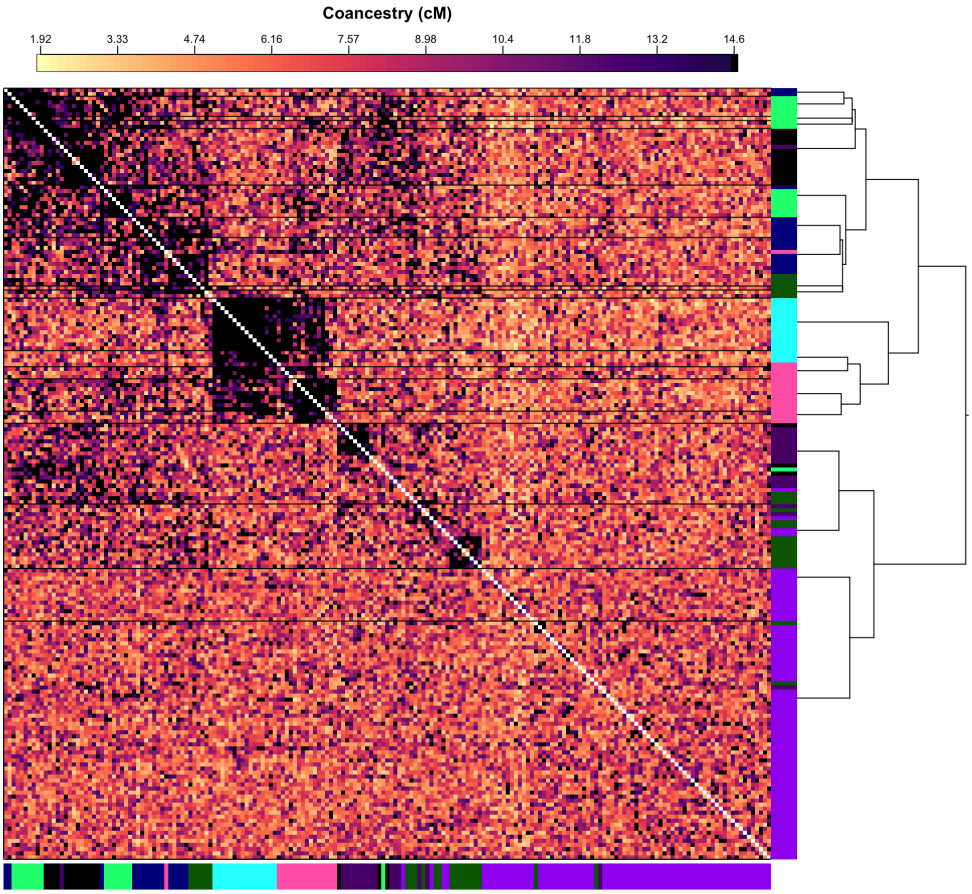


Figure 2.9: Continued from previous page. (b) Coancestry matrix and *fineSTRUCTURE* tree inferred after sub-sampling as described in 2.4.4. The colours in the axes indicate which of the clusters each individual belongs to in the original analysis shown in **(a)**.

2.5 Comparison with principal components and F_{ST}

As a point of methodological comparison, we also used two common approaches to detecting population structure using genotype data but which do not model LD (i.e. use a set of markers assumed to be independent). Namely, PCA and F_{ST} (as discussed in Section 1.2.2.1).

F_{ST}, or the proportion of genetic diversity explained by allele frequency differences between populations, can be used as a way of testing for population structure and

was commonly used in early studies of human population stratification [20, 122]. For the Spanish data set we computed pairwise F_{ST} between different sub-groups using *EIGENSOFT* software (version 5.0.1)[5] [38]. Since F_{ST} is measured between groups of individuals, we conducted two analyses: one using Autonomous Community as group labels, and the other using clusters inferred by *fineSTRUCTURE* as group labels.

We computed principal components of the genotype calls using the software *shellfish* [123]. Both analyses require a set of independent SNPs, so we used a set of 143,599 LD-pruned SNPs ($r^2 < 0.2$), by applying the '--indep-pairwise r2' command in *PLINK* [124]; we also excluded regions of long-range LD derived from [125] (Table A.1).

In the F_{ST} analysis using Autonomous Community as group labels the strongest differentiation (but still weak, at 0.002) is between the Basque-speaking regions and all other regions, and between Galicia and other regions (Figure 2.10b). The principal components analysis revealed a similar pattern, drawing out the same regions (Figure 2.10a). The range of pairwise F_{ST} values is higher when grouping individuals by the clusters inferred by *fineSTRUCTURE* (0 – 0.008) (Figure 2.11b). This highlights *fineSTRUCTURE*'s ability to find clusters of individuals who share genetic drift, and reveals two highly drifted clusters within Galicia and the Basque region.

[5]Note this implementation uses the 'Hudson' estimator [122] and combines the estimator across many loci, by computing a 'ratio of averages'. That is, computing the average of the denominator and numerator across the loci before taking the ratio.

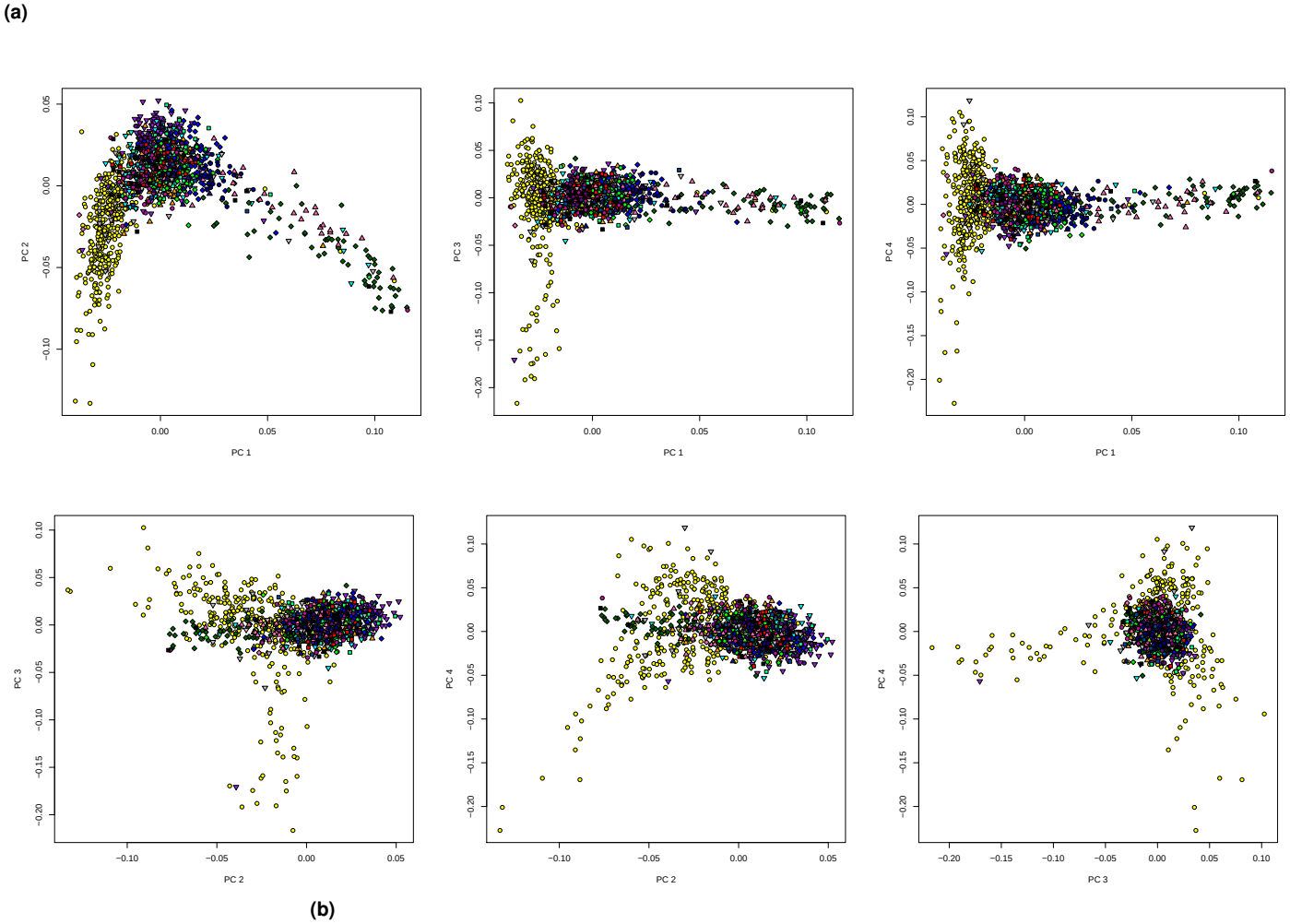


Figure 2.10: F_{ST} and PCA in relation to Autonomous Communities in the Spanish cohort. We computed F_{ST} and PCs using a set of LD-pruned SNPs and show these with respect to Autonomous Community of individuals (based on their grandparents' birthplaces). **(a)** PCs 1-4 for of all Spanish individuals with colours and symbols of points showing Autonomous Community. Refer to Figure 2.10b for labels. **(b)** Pairwise F_{ST} between all Autonomous Communities.

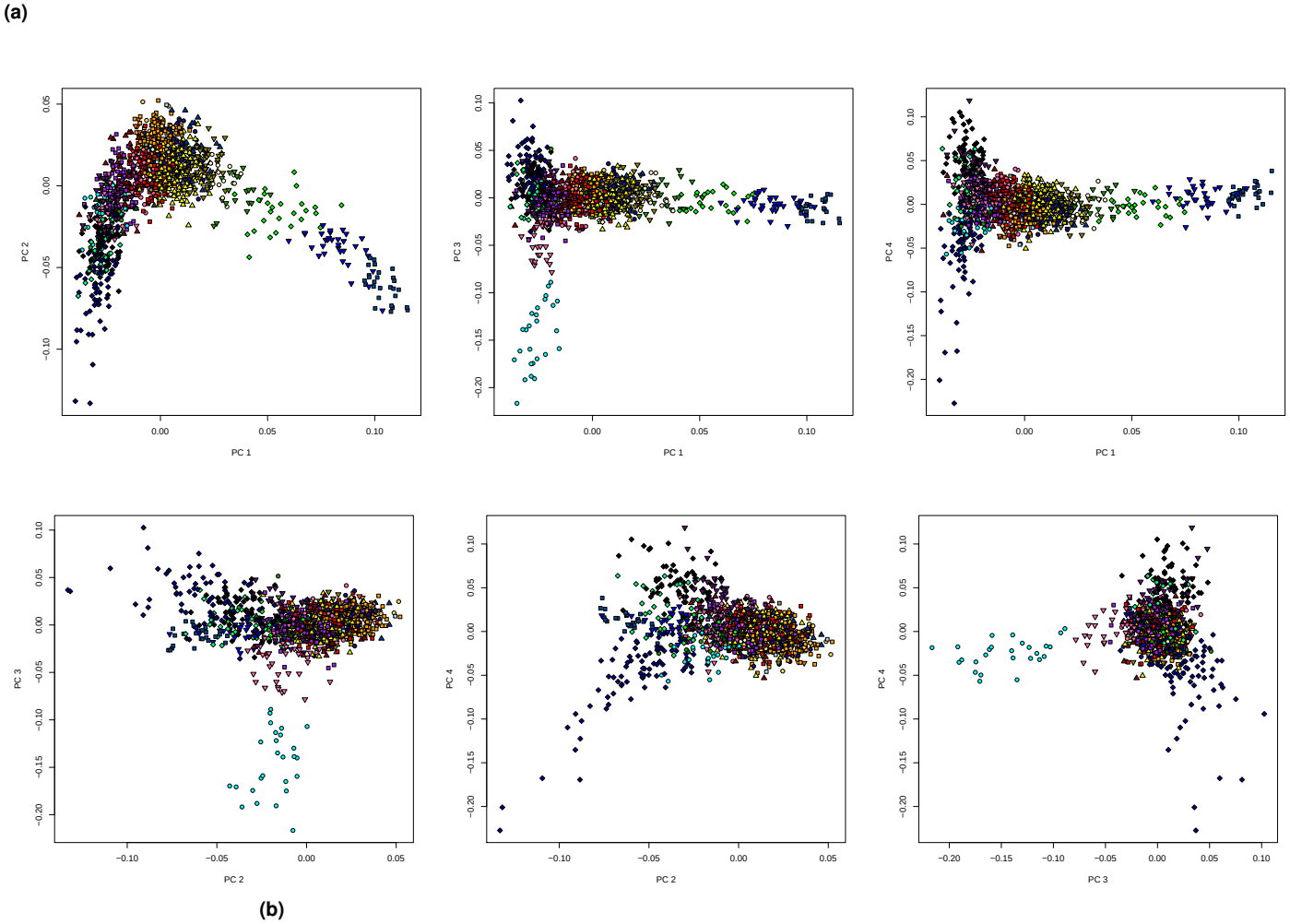


Figure 2.11: F_{ST} and PCA in relation to *fineSTRUCTURE* clusters of Spanish cohort. We computed F_{ST} and PCs using a set of LD-pruned SNPs and show these with respect to clusters inferred by *fineSTRUCTURE* (as represented as points in Figure 2.4). **(a)** The first 4 principal components computed using genotypes only, but coloured according to *fineSTRUCTURE* clusters. **(b)** F_{ST} between each pair of clusters inferred by *fineSTRUCTURE*. The clusters are ordered and labelled as in Figure 2.13, which is informed by the inferred hierarchical tree.

2.6 Relationship with Portugal

To explore the genetic relationship between Spain and the other Iberian nation, Portugal, we performed a *fineSTRUCTURE* analysis of data for the Spanish cohort together with a set of 117 Portuguese individuals from another data source (methods detailed in Section 3.2.1), using $\sim$350,000 SNPs shared across the two data sets (analysis (B)). Fewer SNPs were used in this analysis, so some very fine-scale structure could not be detected, but the broad-scale structure is consistent with the analysis using Spanish individuals only (Figure 2.12). All the Portuguese individuals were assigned to a single cluster also containing Spanish individuals located either in Galicia (97% of the cluster shown in purple squares in Figure 2.4), or less frequently elsewhere along the Spain-Portugal border (dark red triangles in Figure 2.4). This strong similarity between Portugal and Galicia is consistent both with their geographic proximity and linguistic similarities [126]. It also extends the general pattern we observe elsewhere in Spain, of genetic similarity in the north-south direction and genetic divergence in the east-west direction.

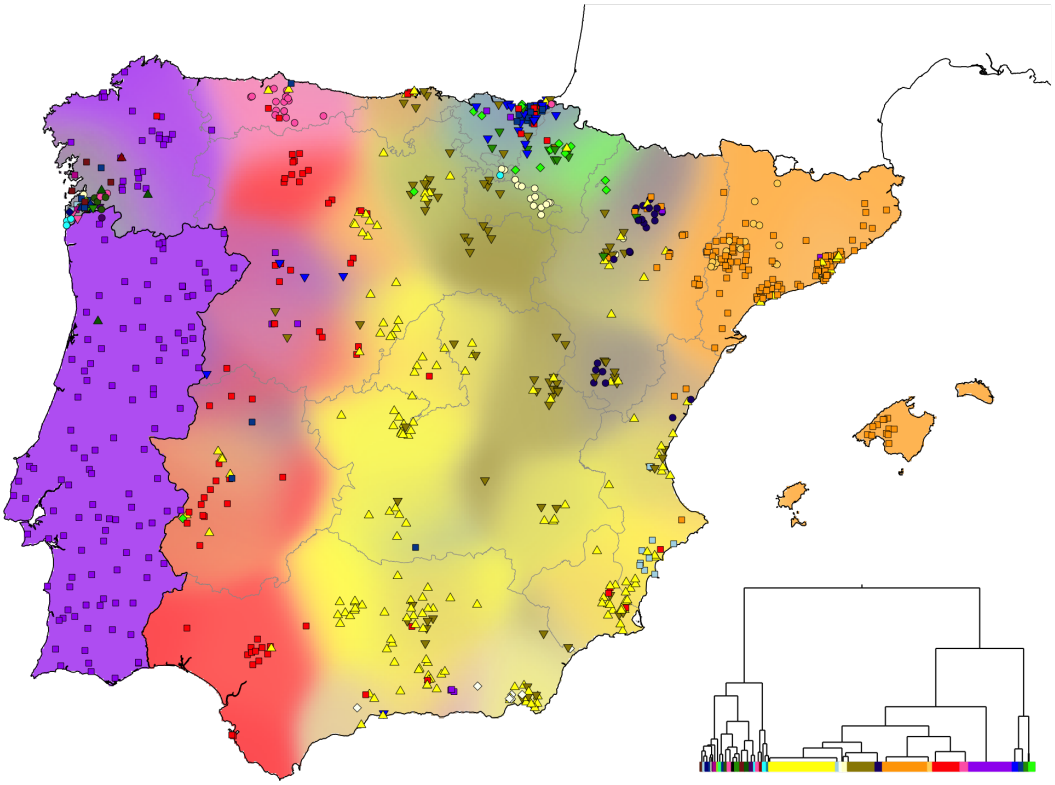


Figure 2.12: *fineSTRUCTURE* analysis including data from Portuguese individuals. Map and tree showing the final set of clusters inferred in *fineSTRUCTURE* analysis (B) that included data from Portuguese individuals but using a smaller set of SNPs (see Section 2.6). Positions of points and background colours are as described in Figure 2.4, with the exception of Portugal. There was no fine-scale geographic information available for the Portuguese individuals (sourced from POPRES [127]) so they are randomly assigned a position within the borders of Portugal. The background colour within Portugal is determined by assuming all individuals contribute the same weight to each grid-point.

2.7 Signals of drift and admixture in the coancestry matrix

The distinct groups within Spain likely reflect regional genetic isolation, but have also likely been further shaped by subsequent migration events within Spain, and which may not be fully represented by a hierarchical tree. We therefore examined properties the full coancestry matrix used to cluster the 1,413 individuals (Figure 2.13). Recall that coancestry (as we have used it) measures the amount of genome (in cM) for which an individual i shares its most recent common ancestor with another individual j, out of all the individuals in the sample. Properties of this matrix are informative of patterns of drift and admixture within and across clusters inferred by *fineSTRUCTURE*.

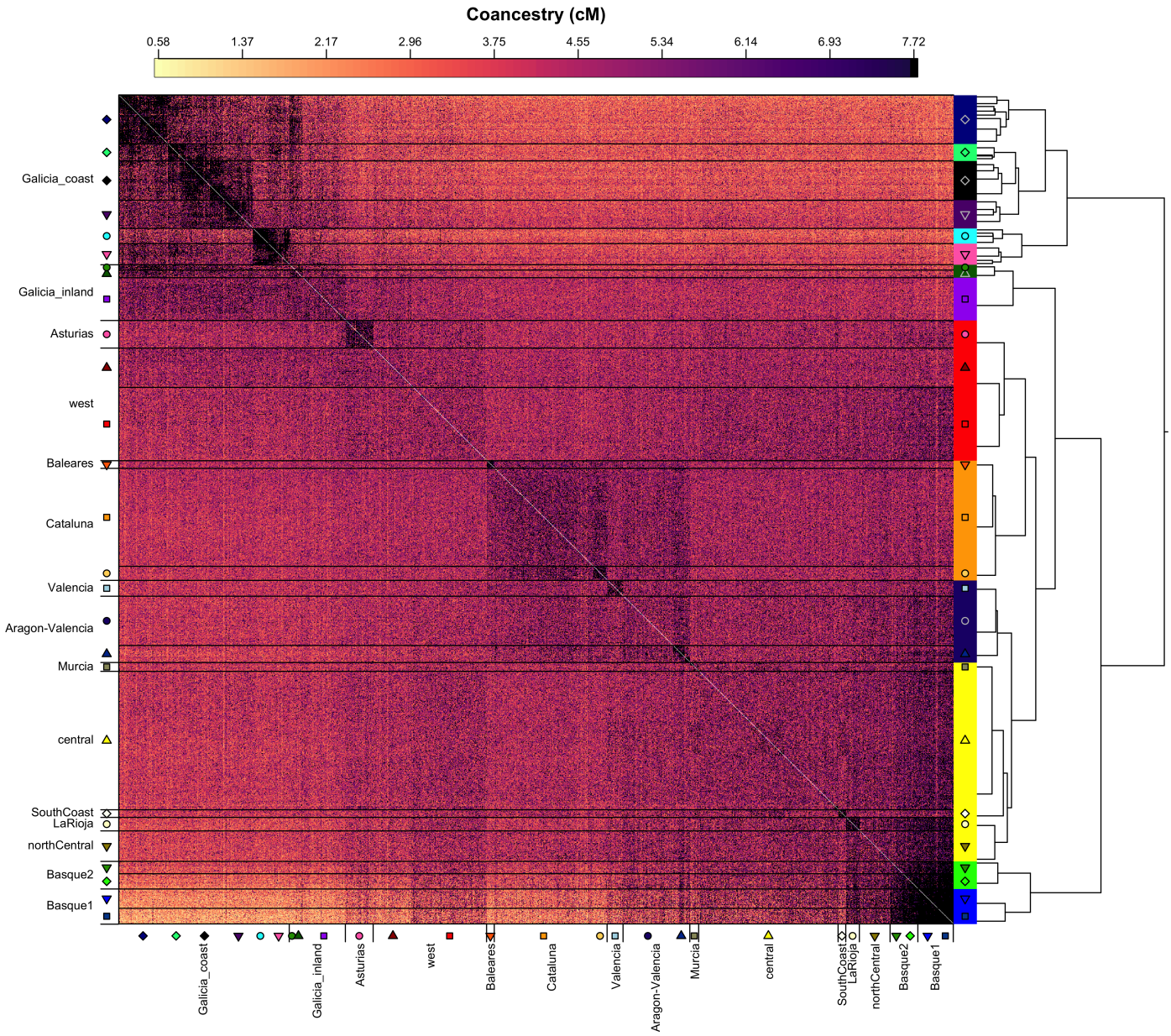


Figure 2.13: Estimates of shared ancestry between individuals in the Spanish cohort. Each individual is represented as a row, where each element is the coancestry (in cM) shared with each of the other individuals (see 2.3 for the definition of coancestry). In order to visualise the bulk of the variation, coancestry values equal to or above the 90th percentile (7.7 cM) are coloured black. Horizontal black lines demarcate the clusters at the lower level of the tree, which are shown as points on Figure 2.4. Colours at the base of the tree correspond to the same clusters as shown in the background of Figure 2.4. The labels correspond to the general geographical location of the clusters, and were not used in the inference of the clusters.

Specifically, excess coancestry between individuals in the same cluster (within-cluster coancestry) is a natural measure of genetic drift of that cluster relative to all the other clusters [41]. In general, individuals are often observed to have the highest levels of coancestry with other individuals in their assigned cluster. This is not a constraint of the *fineSTRUCTURE* model; rather it is because if two individuals have similar patterns of shared ancestry, they are naturally also likely have more recent shared ancestry between them. However, it is possible for this not to be the case, and this is informative of admixture. We reason as follows. Different demographic scenarios could lead to high coancestry between individuals in *different* clusters: either a shared genetic history that is more recent than with the other clusters, followed by a period of isolation (as illustrated in Figure 2.14a); or independent histories followed by a pulse, or period, of admixture, as illustrated in Figure 2.14b.

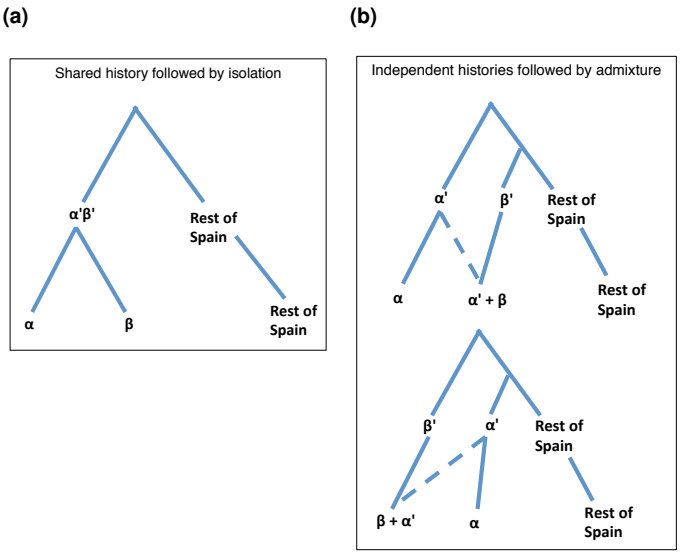


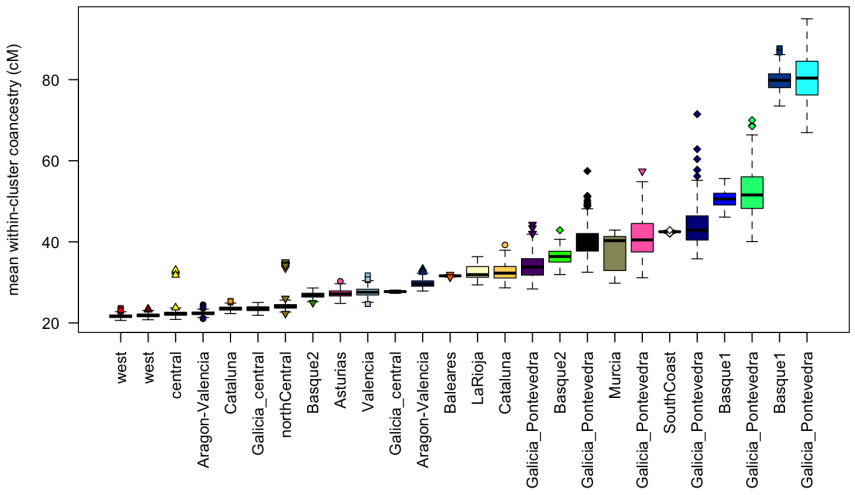
Figure 2.14: Demographic scenarios leading to high coancestry between different clusters. The symbols α and β represent groups of individuals today, and α' and β' represent their ancestral populations. Solid lines represent drift over time, and dashed lines represent admixture from one group into another.

In any pure shared history scenario individuals in β must always have more coancestry — on average — with other individuals in β than they do with individuals in α. This is because at all times in the past, the rate of coalescence between ancestors of two members of population β is at least as high as that between the first population β member and a population α member. So, for a member of population β, their most recent coalescent events are more likely to occur with other individuals in β, thus

elevating their observed coancestry with other members of β. It follows that if individuals in β have more coancestry with individuals in α than they do with individuals in their own cluster, then the shared history scenario cannot be true, so the admixture scenario must have occurred. This property is sufficient, but not necessary for indicating admixture. In other words, it does not necessarily hold that if the admixture scenario had occurred, then β would have more coancestry with α than with itself. This is because the resulting levels of coancestry shared across clusters would depend on the admixture proportions, the relative levels of drift in the ancestral source groups, and the amount of drift since the admixture event.

We looked for signals of admixture in the case of Spain by testing whether a cluster (inferred by *fineSTRUCTURE*) has a within-cluster coancestry that is, on average, smaller than its coancestry with another cluster. We used the 26 clusters (indicated with symbols on the axes in Figure 2.13), which contained at least 13 individuals. To avoid potential bias due to uneven sizes of the clusters, we estimated within-cluster coancestry levels by randomly sub-sampling (without replacement, as coancestry is only defined between two different individuals) each of the 26 clusters such that there were of equal size (13). We re-computed the coancestry matrix using *CHROMOPAINTER*, and the same set of parameters as in *fineSTRUCTURE* analysis (A), but using this smaller subset. We repeated this 200 times and used these resamples to compare coancestry levels across clusters. For each resample and each cluster we computed the mean of the coancestry values within that cluster (excluding zeros on the diagonals), and with each of the other clusters. We then computed a *p*-value using the number of resamples (S) for which the mean within-cluster coancestry is smaller than the mean coancestry with each of the other clusters. That is: $p = \frac{(1+S)}{201}$ [128]. Results are shown in Figure 2.15b. Mean within-cluster coancestry across the resamples is shown in Figure 2.15a.

(a)



(b)

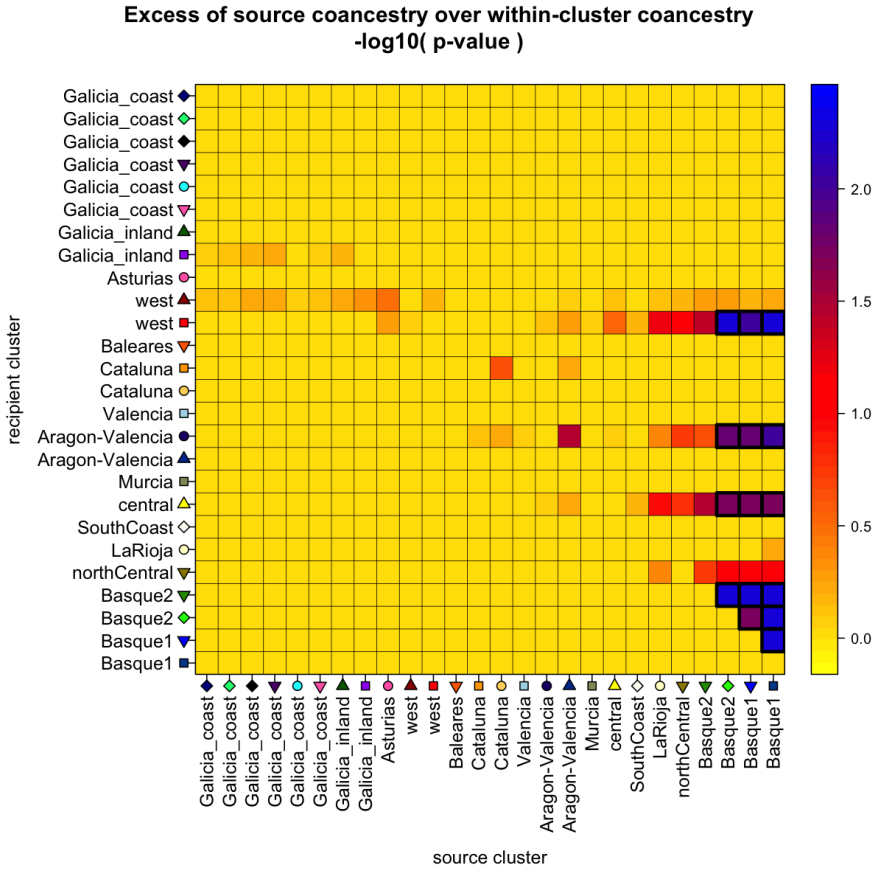


Figure 2.15: Mean within-cluster coancestry and excess coancestry with other clusters across 200 bootstrap resamples. **(a)** For each cluster the distribution of the mean within-cluster coancestry for each bootstrap resample is shown. Clusters are ordered by their median value, and coloured/labelled according to those shown in Figure 2.13. One cluster (part of the clade labelled Galicia_inland) was excluded from this analysis as it only contains 9 individuals. **(b)** Each row of this matrix is a cluster inferred in the *fineSTRUCTURE* analysis as labelled and demarcated in Figure 2.13. For each recipient cluster (rows) we tested whether the mean coancestry among individuals within the recipient cluster is smaller than their mean coancestry with each of the other clusters (columns). p-values are based on 200 bootstrap resamples with even sample sizes across each recipient cluster (see Section 2.7). Dark borders indicate source-recipient pairs with a p-value < 0.02.

We find diagonal entry 'drift' signals for most identified clusters, with the strongest corresponding to highly geographically localized clusters, including the Asturias-centred, Basque-centred, and Galician-centred clusters (Figure 2.15a). These clusters also tend to have higher cluster assignment certainty (Figure 2.7). These clusters have therefore either had small recent effective population sizes, or have remained isolated for long time periods. Contrastingly, the cluster labelled 'central' in Figure 2.13 shows no clear drift signal. In fact, people in this group share — on average — more recent ancestry with the members of Basque-centred clusters (blue squares and triangles) than they do with other individuals in their own cluster ($p < 0.02$, shown in Figure 2.15b). Although perhaps counter-intuitive, our theoretical arguments (Section 2.7) predict that this signal can occur only if admixture from a highly drifted group into another population takes place. Thus, this signal indicates historical admixture into the 'central' cluster from a group related to the Basque populations. We further examine the case of Basque-like admixture more directly in Chapter 3. We see other signals ($p < 0.02$) of admixture within Spain, shown in Figure 2.15b. The absence of such a signal does not preclude admixture as a possibility, so this likely represents only a subset of recent movements within Spain.

We should note that this analysis would not necessarily distinguish 1st generation admixed individuals from those with admixture from further back in time. However, although not guaranteed, groups of individuals with very different admixture timing would likely form different clusters under the *fineSTRUCTURE* model due to the effects of drift (or other mixing) since the admixture event. We assess the timing of Basque-like ancestry using an admixture LD-based method in Chapter 3.

2.8 Discussion

At a broad scale, the major axis of genetic differentiation across Iberia clearly runs from west to east, implying historical barriers to movements of people along this axis. Conversely, there is remarkable genetic similarity in the north-south direction, which breaks down only partially at finer levels of the inferred hierarchical tree. To understand

these patterns, it is key to take into account the historical record. In the case of Spain, many of the patterns we see are most straightforward to interpret in the light of historical information regarding the period of Muslim rule in Iberia, approximately between the start of the 8th, and the latter half of the 13th Century CE.

There is direct historical evidence of migration of peoples from the northern Christian kingdoms into newly conquered regions in the south during the *Reconquista* (see Section 1.2.3.4) and such migration could explain the patterns of genetic differentiation we observe in Spain. Firstly, the genetic structure at broad scales (14 clusters) appears remarkably similar to linguistic and geopolitical boundaries present around the end of the period of Muslim rule in Spain (compare Figures 2.4 and 2.5d). Secondly, the west to east boundaries of the clusters in the north correspond closely to the regions of broad linguistic differences in the Christian-ruled north, which date back to at least the first 200 years of Muslim rule (see Figure 2.5a), and which expanded southwards over 400 years during the *Reconquista*. This phenomenon is well-documented in literature on Spanish linguistic history [69, 65], and is illustrated in 2.5. The generally stronger structure in the northern regions – evidenced by higher coancestry and greater geographic localisation – suggests longer-term isolation there than in southern and central Spain. Thirdly, patterns in the coancestry matrix indicate there has been Basque-like admixture, predominantly in a southerly direction. It is unclear to what extent migrants from the north replaced the existing population, but we should note that complete population replacement is not necessary for it to have a substantial effect on the genetic structure, because any increase in movement between north and south would gradually erode genetic differences along this direction.

We have also detected population structure at ultra-fine scales, especially in the northern regions of Galicia and País Vasco. In Galicia, ultra-fine-scale structure is limited to the province of Pontevedra, with some clusters having geographic ranges of $<$10km. Individuals within these regions have much higher levels of haplotype sharing (coancestry) with other individuals within their clusters than other parts of Spain (Figure 2.13). The demographic history of Galicia goes some way to explaining this

remarkable substructure. Galicia maintained its own cultural identity and language as distinct from the rest of Spain [60], and historically, people in Galicia have tended to inhabit small villages rather than towns [129]. Even today, Galicia remains predominantly rural, with 64% of the Galician population residing in rural areas[6], compared to 44% outside Galicia; and 57% of Galician residents born in Spain before 1961[7] still reside in the same municipality as they were born, compared to 43% in the rest of Spain (Population and Housing Censuses 2011) [130]. Individuals within the clusters located in Galicia do not have unusually low heterozygosity (see Figure A.5), which would indicate a prevalence of parental cosanguinuity. However, studies of marital behaviour over the early-to-mid 20th Century within Galicia have shown that there has been a higher proportion of marriages between uncles/nieces or aunts/nephews and first cousins than in other parts of Spain [131, 132], suggesting that a highly localised marital behaviour is more common in Galicia than elsewhere. Galicia has been the subject of a number of genetic studies, all using small numbers of genetic markers (mtDNA, Y-chromosome, Alu-insertions). While some found significant differences in allele or haplotype frequencies between Galicia and other parts of Spain, such as the Basque region [14], none have been able to detect strong population structure within Galicia [14, 13, 26, 28]. Our analysis indicates that there is significant genetic sub-structure within Galicia, especially in the south-west, and which is probably a result of isolation from the rest of Spain, as well as localised societal organisation within Galicia.

Consistent with previous studies [33, 1, 37], we have observed strong genetic differentiation between individuals in the Basque region and the rest of Spain. However, we have also detected genetic heterogeneity within the region, with different groups distinguished by the extent to which they share ancestry with individuals from outside the Basque region (see Figure 2.13), a conclusion consistent with other studies of the region [133, 134].

[6]This is based on the total resident population minus those that live in urban municipalities or 'cities' in 2016 (including the 'Greater cities' of Barcelona and Bilbao) as defined by the Spanish official statistics agency. See http://www.ine.es/en/prensa/ua_2017_en.pdf for details.

[7]Most (90%) of the Spanish cohort were over 50 years old at the time of data collection (2000-2008), so would have been born around 1960 or earlier.

Chapter 3

The genetic impact of historical migrations and invasions in the Iberian peninsula

3.1 Chapter overview

In Chapter 2 we explored genetic relationships amongst people within Iberia, and it is natural to then explore its relationship with regions surrounding Iberia. As noted earlier (Section 1.2.3), the last 2,000 years of history in Iberia involved several significant changes in political leadership, as well as immigration, and cultural and linguistic transitions. It is plausible that gene-flow from outside Iberia accompanied such changes. Through the analysis in this chapter we sought to understand the extent to which migrations from outside Iberia have influenced the modern-day genetic make-up of the peninsula, and how this might shed light on the demographic forces underlying the population structure we observe within modern-day Iberia.

To place the genetic structure of Spain in the context of surrounding populations (including Portugal) we combined the Spanish data (described in Section 2.2) with genotype data of individuals from Europe [127], north Africa [31] and sub-Saharan Africa [135]. This formed (after QC) a merged data set of genotypes at 300,895 SNPs

for for 2,920 individuals, encompassing all the regions proximal to Iberia.

Using this data set we first explored the population structure of non-Spanish populations themselves. The main aim of that analysis was to identify a set of (generally) geographically localised and genetically distinguishable potential 'donor groups' with which to characterise Iberia (an approach taken in previous work [42, 78]). While population structure within Europe or Africa is not the focus of this study (and has been discussed elsewhere [42],[29],[2]), it is worth examining here for two reasons. It is relevant for interpreting the ancestral relationships between Iberia and non-Iberian groups that we observe in further analyses; and revealing any elements of structure in Europe and north Africa which have not previously been reported in the literature.

To investigate the relationship between these groups and present-day Iberians (combining the Spanish and Portuguese samples), we developed and applied an approach to infer a new set of Iberian clusters based only on their pattern of coancestry with the donor groups, not each other. In contrast to the original 145 clusters (Chapter 2), this approach does not distinguish groups that purely reflect events occurring within Iberia. In the extreme case that all Iberian clusters solely reflect events involving only Iberian ancestors, all individuals should – in this analysis – be indistinguishable. Instead, this analysis is well powered to split groups defined by differing relationships with the donor groups, which would imply differing impacts of migration events, or different events, at some point in the past.

Having identified a set of Iberian clusters in this way, we examined several properties of these, including quantifying the genetic contributions of different external groups; elucidating the timing and character of any historical admixture events; as well as quantifying the geographical distribution (across the peninsula) of genetic contributions, especially of north African-like and Basque-like DNA.

3.2 Data and quality control

3.2.1 Building a data set with multiple data sources

For this analysis, which involves individuals from outside of Spain, we combined four sources of genotype data: European samples from POPRES [127], north African samples from [31], sub-Saharan African samples from Hapmap Phase 3 [135], and the Spanish samples described in Section 2.2. The POPRES and north African samples were typed on the Affymetrix 500K array, which overlaps substantially with the Affymetrix 6.0 array. We first switched all genotypes to the positive strand, shifted to genome build hg19 coordinates and excluded SNPs not in the intersection of the two arrays. We merged all four data sources using the *PLINK* (v1.7) 'merge' function [124] to form a dataset of 359,247 SNPs and 6,954 samples prior to QC and phasing.

Combining data from multiple sources presents unique quality control issues, such as strand misalignments, or subtle biases due to the genotyping taking place in different laboratories and using different arrays. These considerations must be balanced with the need to retain enough SNPs for haplotype-based analyses. As such, we looked for large differences in MAF between each data source and the Spanish cohort, which might indicate strand misalignments. Rare variants ($<$1%) are generally more difficult to genotype than common variants so are often excluded from population genetic analyses. Treating MAF as a proxy for genotype quality within a given study, we excluded SNPs based on MAF computed for each data source separately. Note that we did not use exactly the same set of thresholds as for the Spain-only analysis because that would have resulted in fewer than 200,000 SNPs, which would likely be prohibitively low for the subsequent haploype-based analysis. Specifically, we excluded SNPs using the following criteria:

- MAF < 0.01 within at least one data source.
- *p*-value for departures from Hardy-Weinberg equilibrium $< 10^{-20}$ (computed in Spanish cohort only).
- Overall sample missing rate > 0.05.

- MAF in at least one data source differs from MAF in the Spanish cohort by more than 0.2,0.4, and 0.6, for POPRES, north Africa, and HapMap3 data sources. respectively (values were based on visual examination of the distribution of MAF differences for the different groups).
- Mismatch with the strand of the reference panel used in phasing.
- SNP did not map to genome assembly build hg19.

We also excluded the Spanish samples that were excluded from the previous analysis (Section 2.2.2), along with an additional 221 samples with a call rate (on the filtered SNPs) less than 95% in at least one chromosome, as this is a requirement for phasing. The merged dataset of 6,617 individuals was phased together using the same reference panel as the Spanish-only dataset. After phasing the data we excluded samples using the following criteria:

- A child in a reported trio, or one sample of a related pair with kinship $>$ 0.1, computed using *KING* [119] (excluded the sample with the lowest call rate).
- POPRES individuals labelled as non-European or of mixed ancestry.
- POPRES individuals with non-European ancestry based on principal components analysis.
- Spanish individuals in POPRES (field 'GROUPING_PCA_LABEL1' = 'Spain').

For POPRES individuals we also required that all four grandparents were born in the same country, except for individuals from Ireland and the United Kingdom (UK) as there was no grandparental birthplace information available for them. In POPRES there are many more individuals from UK and Switzerland than from other parts of Europe ($\sim$1,000 from Switzerland; $\sim$400 from the UK). To ensure more even sampling across Europe we picked 90 samples from each of the sets of samples labelled 'Swiss-French' and 'Switzerland' based on the field 'GROUPING_PCA_LABEL1' provided in the POPRES auxiliary data.

This left 2,920 samples and 300,895 SNPs in the merged dataset which we used in the main analyses.

3.3 Defining 'donor' groups

3.3.1 Defining donor groups using *fineSTRUCTURE*

We applied *fineSTRUCTURE* as described in Section 2.3 to define a set of 'donor groups' using the combined European, north African, and sub-Saharan African individuals. We also used these same analyses to investigate population structure outside of Iberia, which we discuss later in this chapter (Section 3.7). Specifically, we conducted three rounds of *fineSTRUCTURE* analyses, each using the following sets of individuals:

(CI) All individuals combined (excluding Spanish but including Portuguese)

(CII) Individuals from north Africa only

(CIII) Individuals from Europe only

In analysis (CI) *fineSTRUCTURE* cleanly split the 3 main groups corresponding to Europe, north Africa, and sub-Saharan Africa, as well as inferring finer sub-structure. However, in order to maximise the power to detect finer scale structure [41], we obtained *fineSTRUCTURE* results for the north African and European groups independently. That is, using coancestry matrices that only allow copying within each set of individuals in (CII) and (CIII), respectively. We then defined 29 donor groups based on the clusters and hierarchical trees inferred by *fineSTRUCTURE* in these three analyses. We considered the following factors in defining donor groups. Ideally, each donor group would contain about the same number of samples, and not be too small. Donor groups should also be relatively homogeneous with respect to their shared ancestry with the population of interest (in this case Iberia), although could be heterogeneous within themselves. We therefore prioritised donor group size over capturing finer scale structure that might exist within donor groups themselves.

For the European samples we used results from analysis (CIII). Some clades of the tree contained clusters with very small numbers of individuals (especially the clade involving mostly Swiss individuals). To avoid using small clusters as donor groups, we

chose a level of the hierarchical tree such that all further splits were smaller than 5, and then excluded samples from any group with a size less than or equal to 15. There were 50 such samples (these are not shown on the figures).

For the north African samples we used analysis CII. Simply removing smaller clusters would have resulted in excluding a significant proportion of samples, so we employed a different strategy. Starting with the smallest cluster, we merged this with the cluster nearest it in the tree, and repeated this process until all resulting clusters were at least size 15.

For the sub-Saharan African samples we used analysis CI. These individuals formed their own clade of the hierarchical tree, and *fineSTRUCTURE* identified the three main population groups: Maasai in Kinyawa, Kenya (MKK); Luhya in Webuye, Kenya (LWK); and Yoruba in Ibadan, Nigeria (YRI). Some sub-structure within these groups was also identified, especially among the MKK individuals (we discuss this later). However, since the coancestry sharing between these smaller clusters and north African and European samples is relatively homogeneous, we merged all the clusters within the MKK and LWK sub-clades, resulting in four sub-Saharan African donor groups.

This procedure resulted in a total of 29 donor groups with median size 30, and minimum size 16. The locations of origin for the individuals are shown in Figure 3.1. Labels of the inferred groups are based on the sampling locations of most of the individuals in a given group. In some cases the majority of individuals were split across two locations, and this is indicated by a multi-region label (e.g. Germany-Belgium).

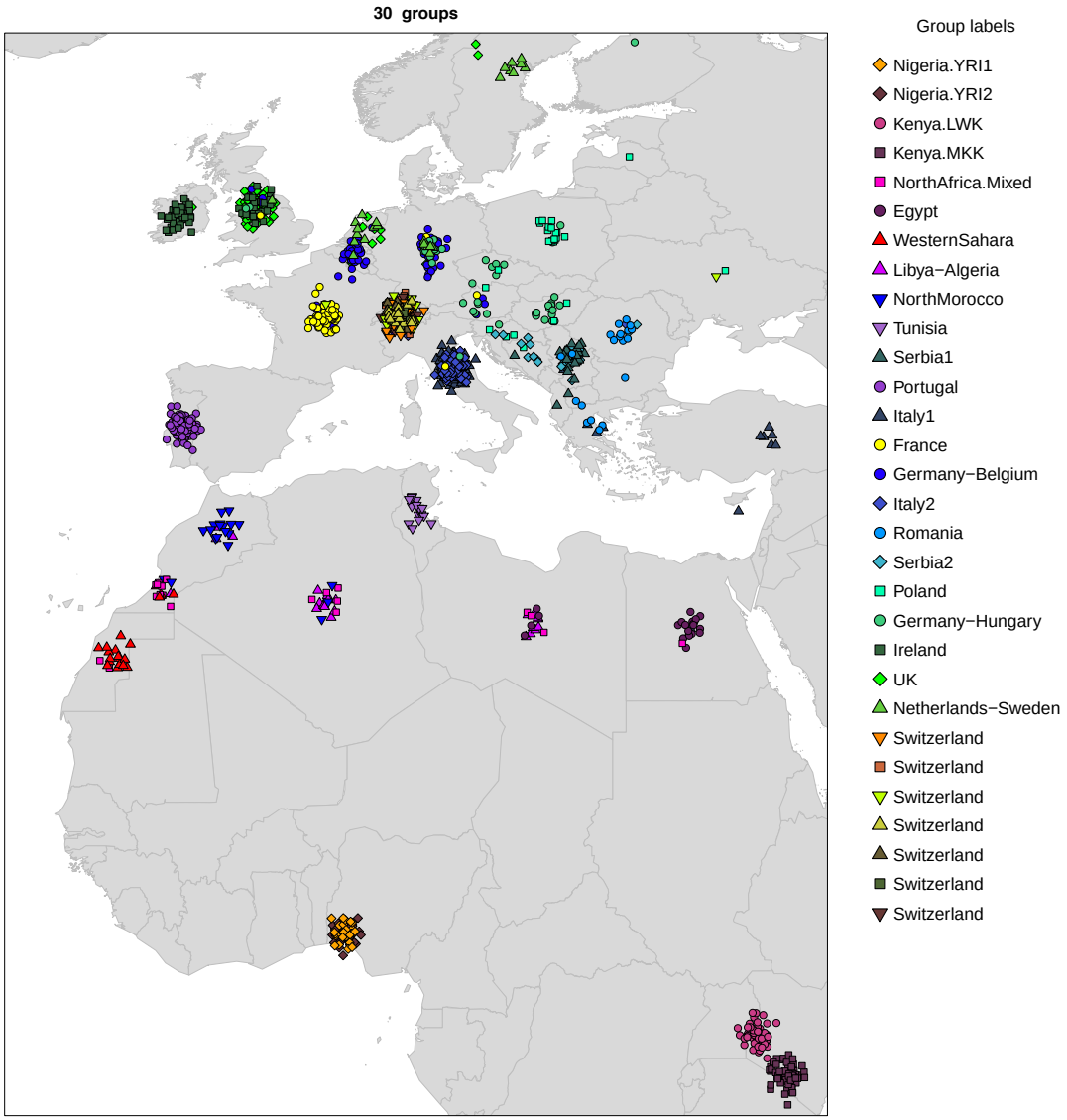


Figure 3.1: Locations of individuals within non-Spanish groups inferred using genetic data only. Each point represents an individual, shown at their country-level location of origin. Colours indicate the different groups inferred using *fineSTRUCTURE* (see Section 3.3.1). Individuals from the same location (country) have been randomly jittered for visual clarity. All groups shown here, except 'Portugal', were used as 'donor groups' in the analyses of Iberia (Section 3.3.2). Genetic relationships among these groups are shown in Figures 3.15 to 3.16b, and Figure 3.14.

3.3.2 Treatment of Portugal

One cluster in the *fineSTRUCTURE* analysis (CIII) overlaps significantly (98%) with the individuals with grandparental origins in Portugal as reported by the data source (POPRES). For the purposes of the analyses in this chapter (and *fineSTRUCTURE* analysis (B) as discussed in Section 2.3), this group of 117 individuals is referred to as 'Portugal' or 'Portuguese individuals' (e.g. in Figure 3.15). The strong genetic

similarity between individuals from Portugal and Spanish individuals (especially those located in Galicia) means they are likely to share a similar admixture history, and including Portugal as a *donor* group would mask the signal from those shared events. We therefore excluded them from the set of donor groups and instead treated them in the same way as the Spanish individuals. This is analogous to the rationale for excluding Ireland as a donor group in the British Isles study [42].

3.4 Clustering Iberian samples on the basis of haplotype sharing with external groups

Having defined a set of non-Iberian donor groups based on genetic data only, we set out to infer clusters of Iberian individuals with distinct patterns of haplotype sharing with these external groups. To do this we used the *fineSTRUCTURE* clustering algorithm exactly as described in Section 2.3 but with a modified version of the input coancestry matrix. Specifically, we compute the coancestry (using *CHROMOPAINTER*) between each Iberian individual and each of the non-Iberian individuals, as described above, but only allowing Iberian individuals to copy from non-Iberian individuals. This results in a rectangular matrix, X, of size N x M, where N is the number of Iberian individuals and M the number of non-Iberian individuals. We then constructed a square matrix C, such that,

$$C = \begin{pmatrix} 0 & X \\ 0 & Y \end{pmatrix} \tag{3.1}$$

and matrix Y contains zeros, except for each of the (block) diagonal entries corresponding to pairs of individuals within the same donor population k. These entries each take value the g_k, which is determined such that the mean of all the entries corresponding to donor population k are the same for the sub-matrix X as the sub-matrix Y. The zeros in the matrix C have the effect of not allowing any copying from or to the Iberian individuals to contribute to the *fineSTRUCTURE* likelihood. we then run *fineSTRUCTURE* algorithm using the 'force file' option (-F), where each 'continental' group is a donor group, thus only allowing splits and merges to take place

among Iberian individuals. We used a c-factor of 0.0579, which was computed in the manner described in Section 2.3.4, but using segments of DNA from a *CHROMOPAINTER* run where we only allowed Iberian individuals to copy from non-Iberian individuals.

By applying this approach we inferred six distinct clusters within Iberia (Figure 3.2), many fewer than the previous approach (Chapter 2) using only Spanish samples. The relatively small number of inferred clusters indicates that much of the fine-scale structure seen in the within-Spain analysis is a result of genetic effects internal to Spain. More specifically, in this analysis there is no sign of sub-structure in the region of Galicia, indicating that the extensive structure seen in the within-Spain analysis is due to local drift effects. Conversely, the cluster located in the Basque region (blue triangles) is almost entirely contained (95%) within the Basque-centred cluster in the within-Spain analysis (blue triangles and squares labelled 'Basque1' in Figure 2.4), showing that the Basque-like ancestry is differentiated due both to drift from the rest of Spain, but also different amounts of ancestry sharing with groups outside of Spain. The new clusters still show geographical localization, predominantly in the east-west direction rather than the north-south direction. However, the greater geographical overlap between groups suggests a somewhat continuous east-west gradient of shared ancestry with non-Iberian populations. The cluster assignment certainty (Figure 3.3) is higher for individuals in the dark-blue and pink clusters located in the north east of Iberia, suggesting that these two groups have sharper geographic boundaries than the others.

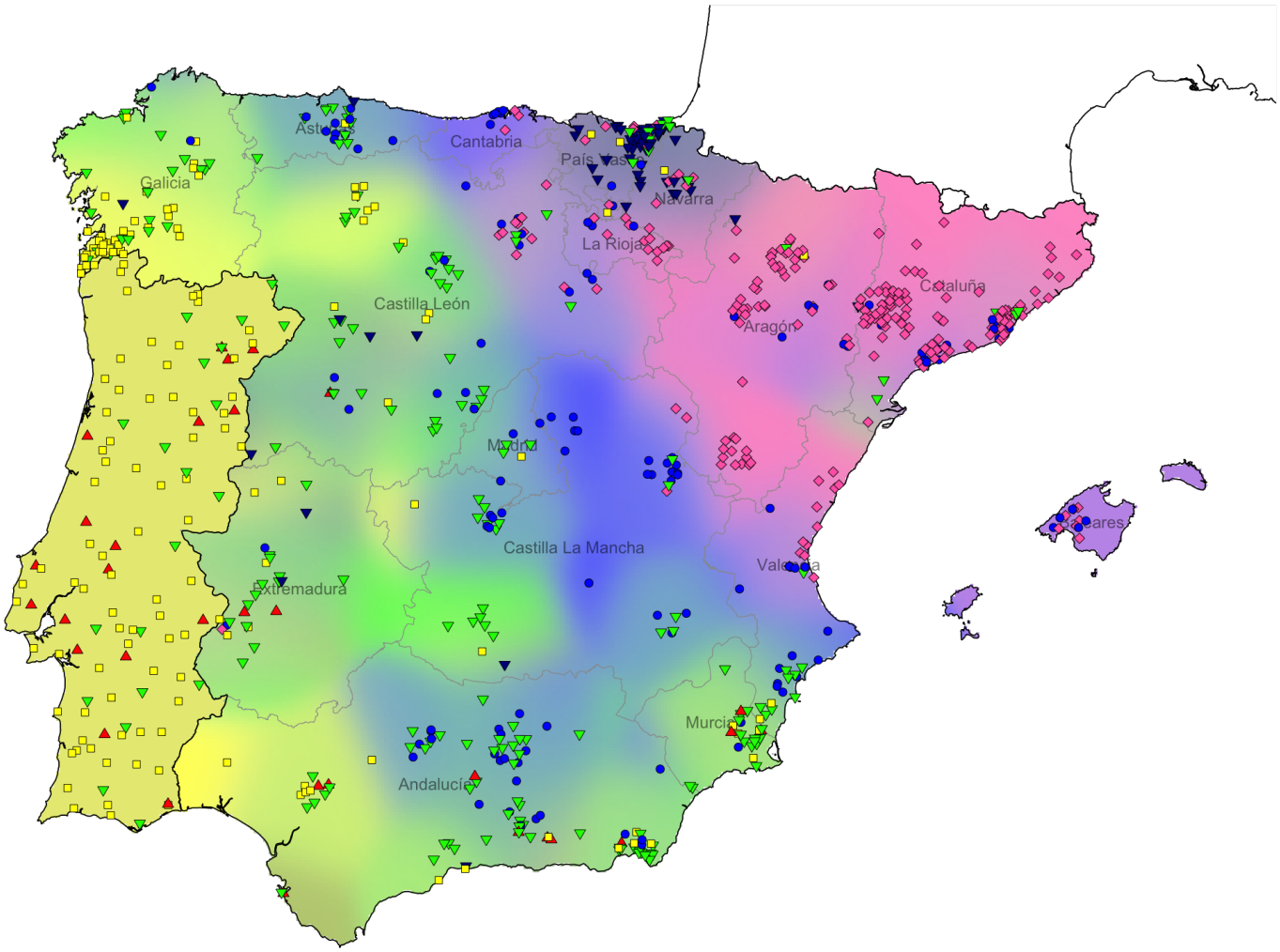


Figure 3.2: Iberian individuals grouped into 6 clusters based on haplotype sharing with external populations. The map shows the geographic distribution of the clusters inferred using the approach described in Section 3.4. Background colours and the positions of points on the map are determined as described in Section 2.3.5 (and similar to Figure 2.4). No fine scale geographic information was available for individuals of Portuguese origin, so we placed them randomly within the boundaries of Portugal, and show a single background colour.

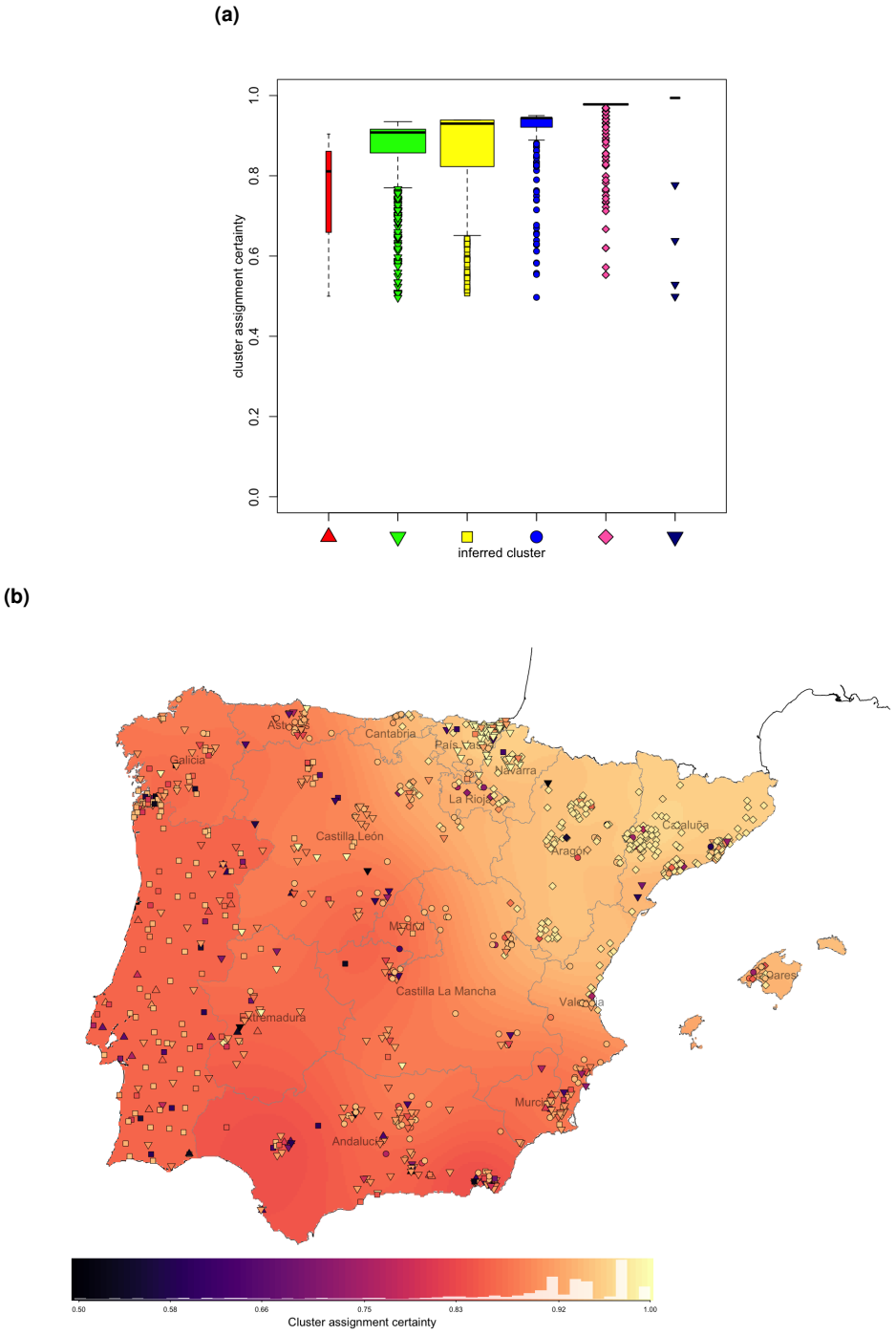


Figure 3.3: Cluster assignment certainty for inference based on haplotype sharing with external populations. For the six Iberian clusters shown in Figure 3.2 we computed a measure of cluster assignment certainty as described in [42]. It measures the co-clustering of individuals over multiple MCMC samples, and can take values between 0 and 1, where 1 indicates high certainty (see 2.4.2). **(a)** The variation in cluster assignment certainty of samples within each cluster. Colours and symbols match those in Figure 3.2, and widths of the box-plots are proportional to the size of each cluster. **(b)** The same data as shown in Figure 3.3a, but showing the geographic spread of cluster uncertainty for the subset of individuals with fine-scale geographic data. The symbols of each point match those shown in Figure 3.2, but are coloured according to their cluster assignment certainty value. The histogram shows the distribution of the certainty measure for the individuals shown on the map. The background colour has been determined by applying a spatial smoothing algorithm to the same data (see Section 2.3.5).

3.5 Iberia as mixtures of external groups

3.5.1 Computing ancestry profiles

In order to characterise genetic contributions to Iberia from external groups, we used an approach previously shown to be informative in the context of the British Isles [42]. This approach accounts for the stochasticity in ancestral relationships along the genome, and uses mixtures of modern-day (sampled) populations to approximate the unknown ancestral groups that actually contributed to modern-day Iberian individuals. Briefly, the approach models a target group (e.g. a cluster from Iberia) as a linear mixture of donor groups, which best explains their haplotype sharing (coancestry) with all the donor groups, while taking into account the haplotype sharing amongst the donor groups themselves. The coefficients of the linear mixture (many if which could be zero) is known as an 'ancestry profile'. Applying this separately to each of the Iberian clusters allows us to compare external contributions to different parts of Iberia.

We estimated ancestry profiles for each of the Iberian clusters based on the procedure used in [42]. Specifically, we use *CHROMOPAINTER* to compute a coancestry vector for each Iberian individual, where we only allow them to copy from haplotypes in the donor groups (as with matrix X above). We sum the elements of these vectors corresponding to individuals from each donor group, thus forming a vector for each Iberian individual where each element corresponds to a donor group. We then average over individuals in each Iberian cluster to form a single coancestry vector for each cluster and normalise so these vectors sum to one. Equivalent vectors are also computed for each of the donor groups, but allowing them to copy from other individuals within their own group. We computed all these vectors using a leave-one-out procedure. That is, since individuals cannot copy from themselves we excluded one individual (sampled randomly) from each of the donor groups, except the group being painted, so that the number of possible donors from each group is always the same.

Using these cluster- and group-averaged copying vectors we find a smaller set of donor groups which together (as a linear mixture) can best account for the raw coancestry

vector of each Iberian cluster. That is, we solve (for $\overline{x}_i$) the non-negative linear least squares problem,

$$\overline{y}_i = \overline{x}_i D + e \quad \text{such that} \quad \overline{x}_i \geq 0 \tag{3.2}$$

where $\overline{y}_i$ is group-averaged coancestry vector for Iberian cluster i, and D is the matrix of group-averaged coancestry vectors of each donor group. The vector of coefficients, $\overline{x}_i$, is the ancestry profile for cluster i, and its elements sum to 1. Results for six Iberian clusters are shown in Figure 3.4.

We also did a complementary analysis where we treated each *donor group* in turn as $\overline{y}_i$ (the results are discussed later in Section 3.7). To do this we first excluded the elements of $\overline{y}_i$ and D that correspond to coancestry within its own group, and re-normalised the coancestry vectors to sum to 1.

We measured uncertainty in the ancestry profiles by re-estimating $\overline{x}_i$ using the coancestry vectors for a set of pseudo individuals. Each pseudo individual is formed by randomly selecting an individual in cluster i for each chromosome, and summing the observed chromosome-level coancestry vectors across all chromosomes. We then compute 1000 such re-estimations and report the range of the inner 95% of the resulting bootstrap distribution.

3.5.2 Computing spatially smoothed ancestry profiles

The availability of fine-scale geographic information for many of the Spanish individuals allowed us to estimate the spatial distribution of shared ancestry. Instead of averaging coancestry over individuals within a cluster, we average across geographic space using the following kernel smoothing method.

Given a set of N coancestry vectors $\overline{y}_i$ (one for each individual i), we computed a new coancestry vector $\overline{y}_s$ for each grid-point s in a fine grid across Spain such that,

$$\overline{y}_s = \frac{\sum_{i=1}^{N} w_{si} \overline{y}_i}{\sum_{i=1}^{N} w_{si}} \tag{3.3}$$

where w_{si} the two-dimensional, zero-centred symmetric Gaussian function evaluated at d_{si}, the Euclidean distance between the centre of grid-point s and the coordinate of individual i (in units of 10 Km). That is,

$$w_{si} = e^{-\frac{d_{si}^2}{2\sigma_s^2}} \ . \tag{3.4}$$

We allow the parameter σ_s^2 to vary to reflect the amount of real data available around each grid-point. Specifically, we set each σ_s^2 such that the sum of the weights w_{si} is the same for all grid-points. That is, for each grid-point, s we solve Equation 3.5 for σ_s^2, where k is set to the maximum value (over all grid-points) of the left-hand-side when using $\sigma_s^2 = 3.5$. In practice we solve this equation numerically for each grid-point using an implementation of Nelder & Mead's simplex algorithm in *R* ('optim' function) [136].

$$\sum_{i=1}^{N} e^{-\frac{d_{si}^2}{2\sigma_s^2}} = k \ . \tag{3.5}$$

This has the effect that all grid-points have $\sigma_s^2 \geq 3.5$ and grid-points with a lower density of nearby data points have weights w_{si} spread more thinly over larger distances. This ensures, for example, that a single isolated data point will not contribute heavily to the coancestry vectors of nearby grid-points.

We then compute ancestry profiles for each of the grid-points in the same way as we did for the clusters (described above), but setting $\overline{y}_i = \overline{y}_s$. We visualize the results by colouring each grid point according to the value of its coefficient for a single donor population of interest (e.g. NorthMorocco). We set w_{si} to 1 for any grid-point s and individual i located within the borders of Portugal because we have no fine-scale geographic information for these individuals. This means in Portugal there is just one colour corresponding to the mean coancestry vector across individuals located in the region.

3.5.3 Ancestry profiles for Iberia

Ancestry profiles for each of the six Iberian clusters are shown in (Figure 3.4). Only seven of the 29 donor groups (six with more than 1% contribution to any group) show any meaningful contribution, and are primarily in western and southern Europe, and north-west Africa. In all six Iberian clusters the largest contribution comes from France, with smaller inferred contributions that relate to present-day Italian and Irish samples. Because these contributions are present and dominate overall ancestry throughout Spain, they might represent ancient ancestry components, rather than recent migration. The remaining contributions come predominantly from North Morocco and Western Sahara, and show strong (and significant) regional differences, varying continuously across Spain. Of particular note is the North Moroccan component. In the cluster-based ancestry profiles it steadily declines from 11% in the far west to as low as 0% in the Basque region. The geographic pattern of contributions from NorthMorocco, as measured by the spatially smoothed ancestry profiles (Section 3.5.1), shows a striking west-east pattern (Figure 3.5). This is similar to the pattern of WesternSahara contributions. In general, the pattern of contributions from north African donor groups is opposite to the pattern of French contributions.

In contrast to African ancestry components, French-like ancestry dominates in all regions (minimum estimate 63%, Table A.2), and shows less marked regional variation, although is highest in regions most proximal to France (Cataluña and the Basque regions). Following an argument used in [42], ancestral components consistently present over large geographic areas need to have arrived long enough in the past to have spread. This, along with France's geographic proximity to Spain, suggests that French-like ancestry is most plausibly a result of long-term (potentially prehistoric), and probably on-going gene-flow between France and Spain.

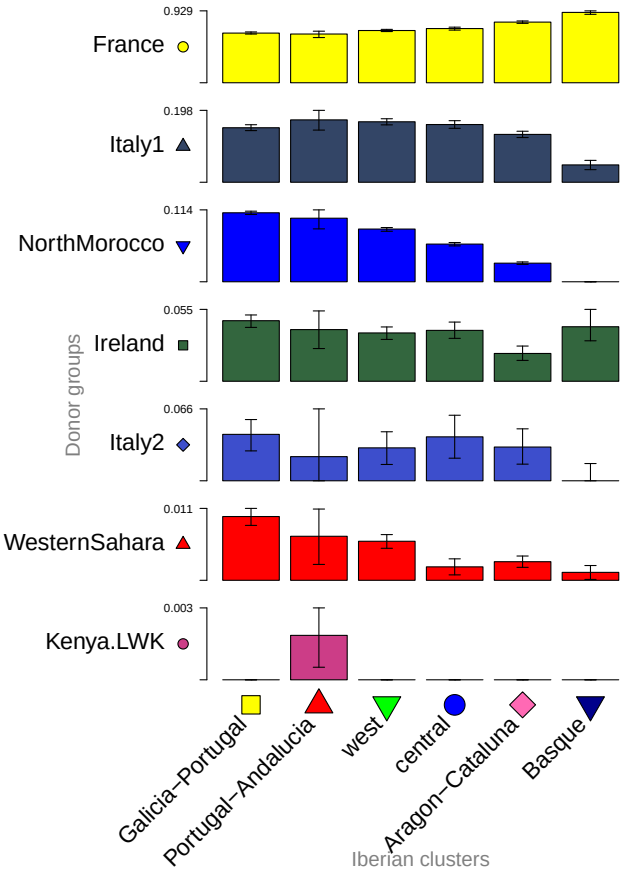


Figure 3.4: Ancestry profiles of Iberian clusters. Each column shows the ancestry profile of each of the inferred clusters shown in Figure 3.2. The heights of the bars show the proportion of each cluster's ancestry which is best represented by that of the labelled non-Iberian group. Details of the inference are described in Section 3.5.1. Note that each row has a different y-axis range for visibility of the smaller components. Error bars show the range of the inner 95% of 1,000 bootstrap re-samples (see Section 3.5.1), and donor groups are only shown if at least one cluster has a range not including zero and a point estimate greater than 0.001. The exact values shown in this plot are tabulated in Table A.2.

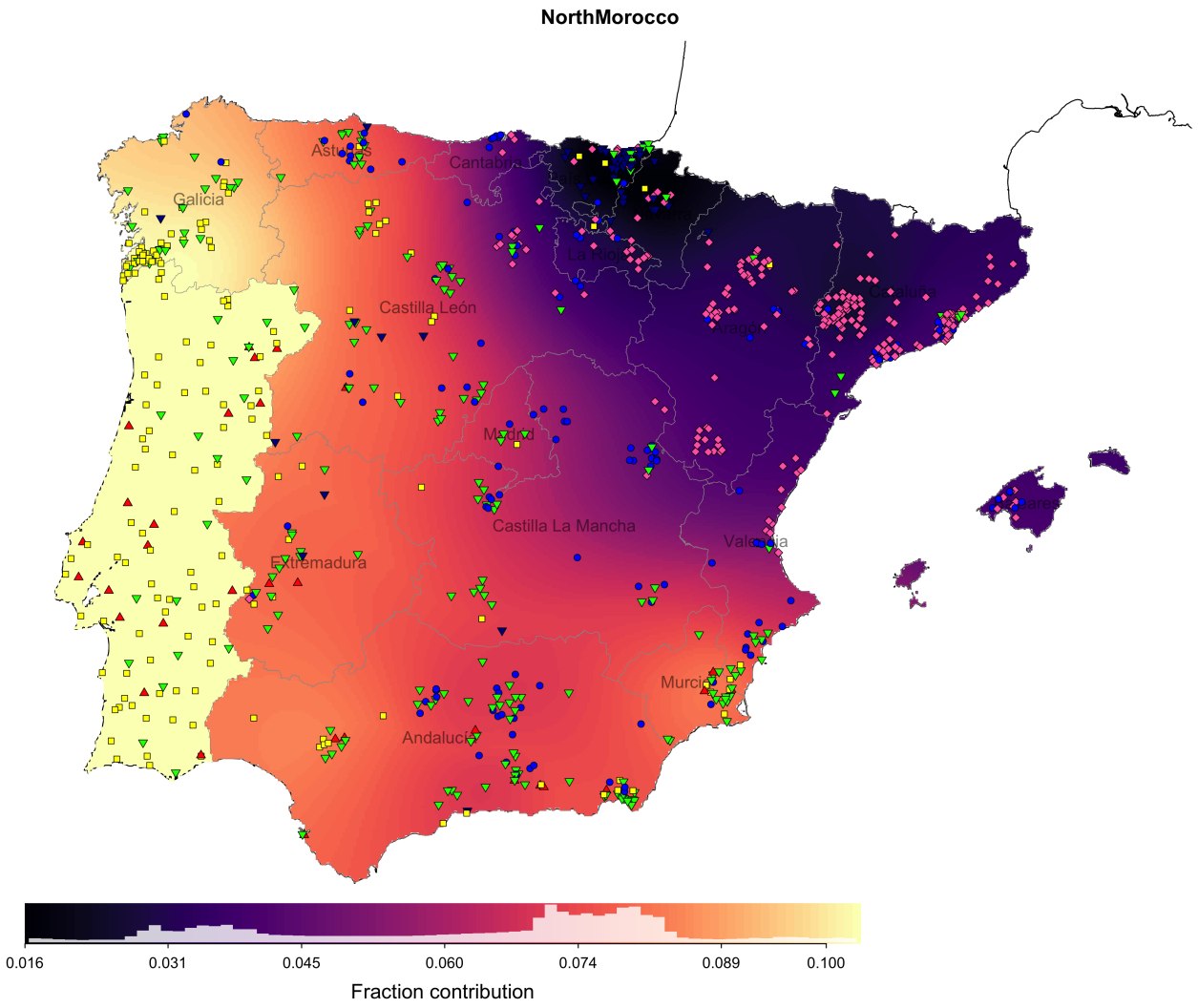


Figure 3.5: North Moroccan component of spatially-smoothed ancestry profiles. We computed spatially smoothed ancestry profiles for each point on a spatial grid across Spain (see Section 3.5.2). This map shows the fraction contributed from the donor group 'NorthMorocco'. Colours and symbols of the points are exactly as in Figure 3.2. The histogram on the scale bars shows the distribution of values across grid-points on the map. Maps of contributions from other donor groups are shown in Figure A.7.

Interestingly, regional variation is less marked for the contributions from the Irish and Italian populations, perhaps implying these relate to older ancestral populations to that of the other groups. A very small (0.2%) but statistically non-zero contribution from sub-Saharan Africa is present for just one cluster (red triangles), which contains individuals largely from Portugal and regions of southern Iberia, such as Andalucía, and Murcia (see Figure 3.2). This group otherwise has a similar profile to the west-most cluster (yellow squares), indicating that this cluster is differentiated only by an excess of haplotype sharing with individuals from sub-Saharan Africa.

Notably, the Basque-centred cluster has a markedly different profile from the rest. Firstly, it has much lower, or zero contributions from donor groups that contribute to all other clusters: Italy1, NorthMorocco, and WesternSahara, and a very large contribution (88-93%) from France. Additionally, the model fit for this cluster is strikingly less good than that for the other clusters (see Figure 3.6), indicating that Basque-like DNA is less well captured by the mixture of donor groups in this data set.

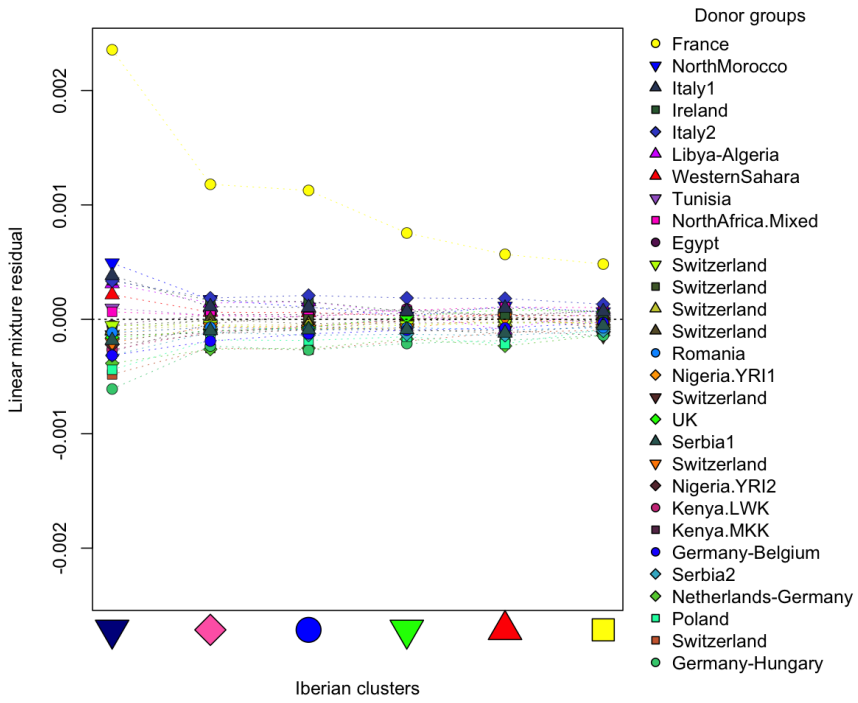


Figure 3.6: Ancestry profiles residuals for Iberian clusters. Ancestry profiles for each of the Iberian clusters are coefficients of a model that fits each clusters coancestry vector as a linear mixture of the coancestry vectors of each donor group (see Section 3.5.1). Residuals for each component of this fit (corresponding to a donor group) are shown here. Each point represents the residual for a donor group indicated by a colour/symbol. Positive values on the y-axis indicate that the observed coancestry component is larger than the fitted component. Locations of each donor group and Iberian cluster are shown in Figures 3.1 and 3.2, respectively.

3.6 The character and timing of admixture events in Iberia

3.6.1 Estimating admixture dates and source populations

Although the mixture decompositions provide insights into the relationships between DNA in Spain and neighbouring regions, they are not able to distinguish between

different historical scenarios involving sharing of DNA, such as stable migration between different regions, as opposed to particular pulses of migration and subsequent admixture [50]. To examine the evidence for, and characterise such admixture events if they are detected, we applied the *GLOBETROTTER* method [53] to each of the six Iberian clusters.

The underlying theory and implementation of the *GLOBETROTTER* model is detailed in [53], but it is worth discussing some details of the approach in order to clarify discussions later in this chapter. Recall (from Section 1.2.1) that admixture LD, i.e. the rate of decay of correlation with genomic distance between genetic loci with ancestry from two (or more) historical source populations, is informative of admixture date(s) and proportions. *GLOBETROTTER* measures this correlation empirically by constructing 'coancestry curves' based on 'sample paintings' from the *CHROMOPAINTER* algorithm. Recall that for a given target haplotype, the *CHROMOPAINTER* algorithm estimates at each SNP (via an HMM) the donor haplotype with which the target haplotype has the shortest coalescence time (Section 2.3). Under the HMM framework, it is possible to sample from the probability distribution of paths through the hidden states. That is, to represent a target haplotype as a mosaic of contiguous SNPs (chunks) of donor haplotypes. This is referred to as a 'sample painting'. Using multiple sample paintings of haplotypes within a population of interest (referred to as a 'target population'), *GLOBETROTTER* constructs coancestry curves for pairs of modern-day donor groups used in the painting. Specifically, *GLOBETROTTER* measures how frequently a pair of chunks g cM apart copy from haplotypes in two donor groups, relative to the genome-wide expected frequency of their co-occurrence (i.e. if there was no correlation between chunks from the two donor groups).

One challenge of admixture inference using modern-day genetic data is that modern-day populations do not necessarily well-represent the ancestral populations that came together at some point in the past. To address this, *GLOBETROTTER* models each of the historical admixing populations (for example two populations) as a mixture of the haplotypes found in modern-day populations. It is natural to then model

the target population itself as a mixture of the same modern-day groups. It remains for *GLOBETROTTER* to partition these groups into each of the admixing populations (referred to as 'sides' of the admixture event) and infer the timing of the event(s).

In practice, *GLOBETROTTER* does this by first reducing the set of donor groups involved in the sample paintings to a smaller subset of 'surrogate groups' by representing the target population as a linear mixture of donor groups with non-negative coefficients. This is based on genome-wide coancestry proportions, and is similar to the way ancestry profiles are constructed (see Section 3.5.1). Then, each chunk in the sample paintings is replaced with a vector of weights, where each element is the probability that a chunk has ancestry from surrogate group m, given the donor group the chunk copies from in the raw painting. Note that there are typically many fewer surrogates chosen than potential donor groups. Coancestry curves for each pair of *surrogate* groups are then constructed as a weighted average (using the weights for the two surrogate groups) of their co-occurrence in pairs of chunks g cM apart. In the context of an admixture scenario involving two historical source populations (2-way admixture), theory predicts (Supplementary Material S3.5 of [53]) that downward-sloping coancestry curves (nearby chunks are positively correlated) will occur for pairs of surrogate groups on the same side of the admixture event; upward-sloping curves (nearby chunks are negatively correlated) will occur for pairs of surrogate groups on the opposite side of the admixture event. Examples of such curves are shown later in Figure 3.8. *GLOBETROTTER* proceeds in an iterative way, by recomputing the coancestry curves after excluding surrogate groups with coancestry curves that do not indicate admixture.

By default, all the donor groups can be potential surrogate groups. However, it is possible to restrict the *GLOBETROTTER* model to only consider a subset of donor groups as 'surrogates'. This can be useful in the context of complex admixture histories because it can allow us to focus the inference on one, or a smaller set of historical events.

3.6.2 Details of *GLOBETROTTER* analyses of Iberia

We conducted two analyses using *GLOBETROTTER*. The first (gtA) was designed to detect admixture event(s) in the history of Iberia that might involve any combination of non-Iberian source populations, without any prior assumptions on the nature of the event. The second analysis (gtB) was designed to detect only admixture event(s) involving a Basque-like source population, i.e. based on a prior hypothesis. In each case we defined a set of target groups within which to look for an admixture event; a set of donor groups, which we allow to be donors in the initial 'painting'; and a set of surrogate populations, which we allowed *GLOBETROTTER* to consider as components of any admixture event. These details are summarised in Table 3.1. Labels of target and surrogate groups refer to those shown in Figure 2.13 (Spanish clusters) and Figure A.6 (non-Iberian groups). In order to speed up computation time, in both analyses we restricted the number of individuals within a target group to 100 by randomly sampling groups with more than 100 individuals. See Table A.3 for resulting sample sizes in each target group. We otherwise ran *GLOBETROTTER* as recommended by the authors (*GLOBETROTTER* Instruction Manual [53]). In all figures and tables relating to *GLOBETROTTER* results we report the results using the 'null.ind: 1' procedure. That is, the coancestry curves are normalised by 'across-individual' coancestry curves, which are constructed exactly as described above, but where only pairs of chunks in the haplotypes of *different individuals* within the target population are compared. In practice we found the results were similar to the 'null.ind: 0' versions, but are less likely to be influenced by any bottle-neck effects since an admixture event [53].

Analysis	Target groups	Donor groups	Allowed surrogate groups
gtA	6 Iberian clusters, inferred on the basis of haplotype sharing with non-Iberians only. These are shown in Figure 3.2.	29 donor groups (see Section 3.3.1, Figure 3.1).	29 donor groups (see Section 3.3.1, Figure 3.1).
gtB	Portuguese cluster (see Section 3.3.1) 20 Spanish clusters, excluding 60 individuals in the clade labelled 'Basque1' (see Chapter 2, Figure 2.13). Note, individuals within the clade labelled 'Galicia_coast' were combined for this analysis.	29 donor groups (see Section 3.3.1, Figure 3.1). Basque1 (see Chapter 2, Figure 2.13)	Basque1 (see Chapter 2, Figure 2.13) Germany-Belgium_1 Germany-Hungary_7 Netherlands-Sweden_27 Poland_19 Romania_23 Serbia_3 Serbia_9 Switzerland_11 Switzerland_14 Switzerland_16 Switzerland_20 Switzerland_21 Switzerland_25 Switzerland_4 UK_2

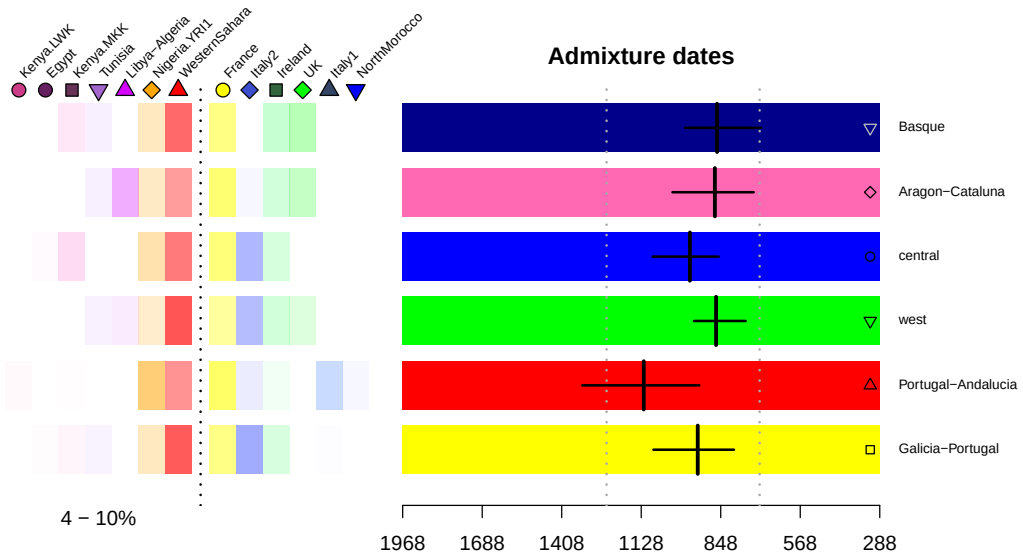
Table 3.1: Details of two *GLOBETROTTER* analyses as discussed in Section 3.6.2.

3.6.3 Admixture involving European-like and north African-like ancestral populations

In analysis (gtA) *GLOBETROTTER* found evidence ($p < 0.01$) of admixture impacting all six clusters, including the Basque cluster (Figure 3.7, Table A.3a). An admixture model of one date and two source groups fit well in all cases (discussed later in Section 3.6.5), and the inferred times for this admixture range from 30-40 generations ago, corresponding to 860 – 1120 CE (assuming a 28-year generation time [53]). All dates fall within the period of Muslim rule within Spain, and all inferred confidence intervals overlap. Futhermore, the inferred event for each of the six clusters has a very similar composition (Figure 3.7). The unsampled source populations — as in the mixture analysis — are modelled as mixtures of sampled (modern-day) groups, based on long-range correlation patterns observed along the genome. In this mixture representation, the major source was inferred to contain almost exclusively European

populations. In contrast, the minor source was inferred to contribute 4-10% of DNA, and is made up of African (mainly north African) groups, with the largest contribution being from the WesternSahara group in all but one case (red triangles), for which western sub-Saharan DNA (YRI) contributes the most. This cluster (Portugal-Andalucia) is the same cluster in which the ancestry profiles (Figure 3.4) show a small sub-Saharan African contribution. This group also showed modest evidence of admixture happening at more than one time (Section 3.6.5; Figure 3.13). Under a two-date admixture model, the admixture event involving primarily sub-Saharan-African-like and European-like source groups was inferred to be more recent, at 1647 CE (1300 – 1774) (see Table A.3a; Figure A.8). Furthermore, in general for Iberia there is a strong linear correlation between north African and sub-Saharan haplotype sharing (coancestry), but for many individuals in this group there is a larger sub-Saharan African component than would be expected given their north African component (Figure 3.9a). This suggests that for this cluster, a small component of sub-Saharan African DNA arrived more recently, and independently from the north African component.

(a)



(b)

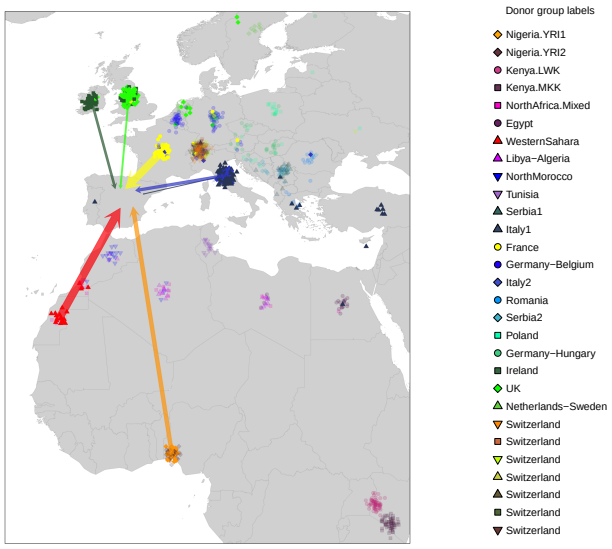


Figure 3.7: Characterisation of a north African-like admixture event in Iberia. Admixture dates and mix of modern-day surrogate groups that best characterises the single-date, two-way admixture events, as inferred using *GLOBETROTTER*. For all Iberian target groups this event involves a north-African-like side and a European-like side. **(a)** Donor groups inferred to best represent the two sides of the admixture event are shown on the left of the panel (separated by a dashed line) along with the range of inferred admixture proportions of the smaller side. Estimated dates, and 95% bootstrap intervals are shown on the right, for each target group. Target groups are the set of Iberian clusters inferred based on haplotype-sharing with non-Iberian groups (see Section 3.4), and symbols match those shown in Figure 3.2. The dates shown assume a 28-year generation time, and a 'now' date of 1940 (the approximate average birth-year of this cohort). The white vertical dashed lines show the time of the initial Muslim invasion (711 CE) and the Siege of Seville (1248 CE), between which around half (or more) of Iberia was under Muslim rule. **(b)** Locations of donor groups, highlighting the set of groups inferred to represent the historical mixing populations (Section 3.6.1). The width of the arrows is proportional to the average inferred contribution across all six target groups, and only those contributing more than 1% to the whole mixture are shown with arrows. Points represent the locations of individuals in each donor group, and those donor groups inferred to contribute zero or less than 1% to the mixture are shown as semi-transparent.

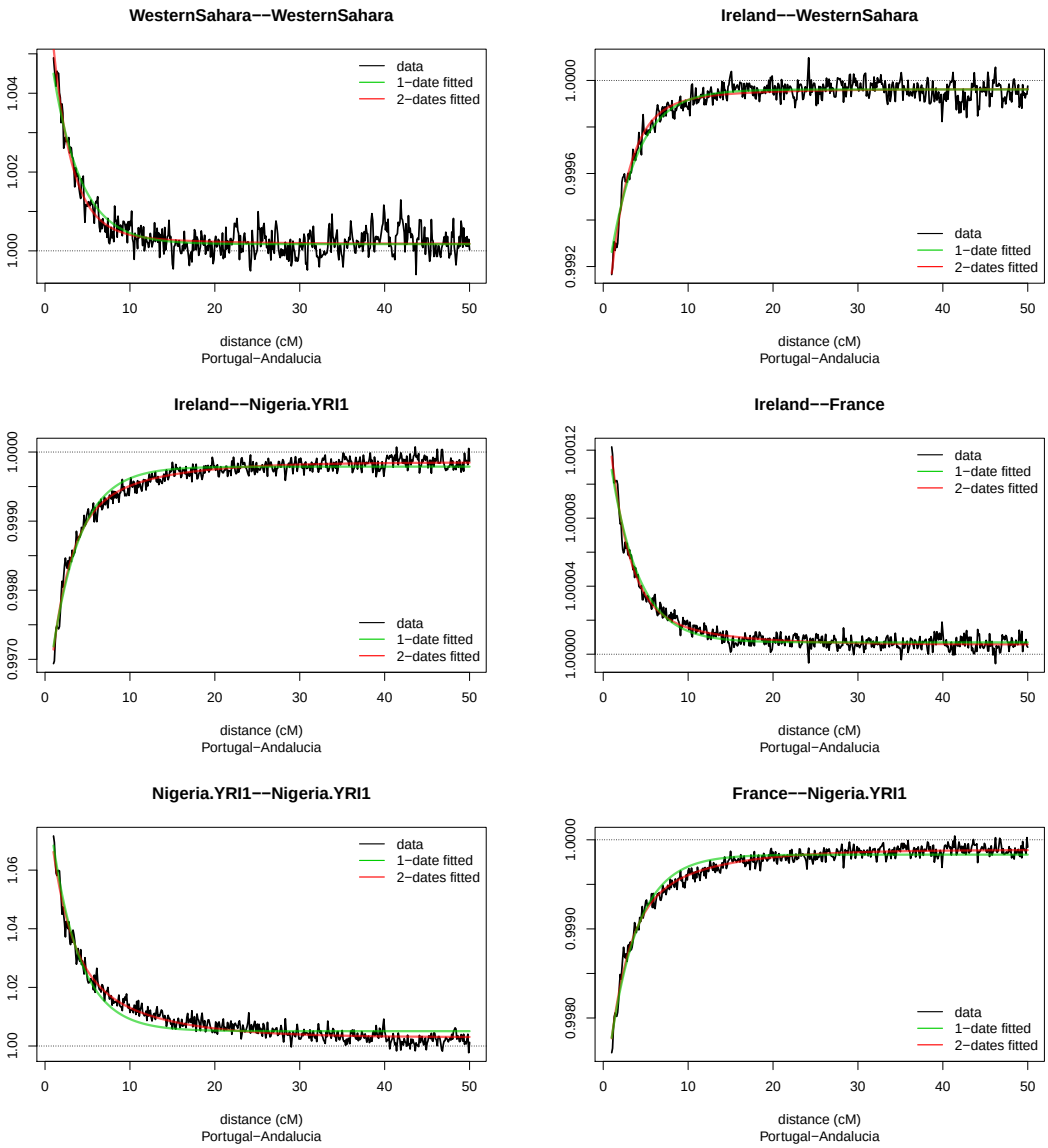


Figure 3.8: Example coancestry curves for target group 'Portugal-Andalucia'. These curves are for an example Iberian cluster ('Portugal-Andalucia'), shown in red triangles in Figure 3.2. Each curve shows the decay of correlation with genomic distance, between segments of DNA with ancestry from the two surrogate groups shown in the titles. The surrogate groups are a subset of the donor groups defined using *fineSTRUCTURE* (see Figure 3.1). Curves with a negative slope indicate pairs of surrogate groups on the same 'side' of the admixture event (e.g. Ireland–France); curves with an upward slope indicate pairs of surrogate groups on the opposite 'side' of the admixture event (e.g. Ireland–WesternSahara). Green lines show the fitted curves for a one-date admixture model, and red lines show the fitted curves for a two-date admixture model. The better fit to the two-date model for curves involving the sub-Saharan African group 'Nigeria.YRI1' is visually evident. See Table A.3a for model fit statistics for all Iberian target groups, and Figure 3.13 for further evidence of a possible two-date admixture event in some target groups.

Encouragingly, these results are in close agreement with the mixture analysis, and imply that the north African-like DNA found in all present-day Iberian groups examined is likely to have largely originated from admixture occurring during the first half of the Muslim rule in Iberia. Notably, the *GLOBETROTTER* analysis favours Western Sahara, rather than the nearby present-day North Moroccan group, as the closest surrogate for the group contributing north African DNA into Spain. This difference, though subtle, might be explained if modern-day north Moroccan haplotypes are more similar to present-day Spanish individuals than the admixing group were. Indeed, work by others [31] found that modern-day north Moroccan haplotypes carry a component of European ancestry that most likely arrived subsequent to the detected admixture event, and this is consistent with the presence of non-zero components from southern Europe (Italy and France) in the ancestry profile for the north Moroccan group itself (Figure 3.14).

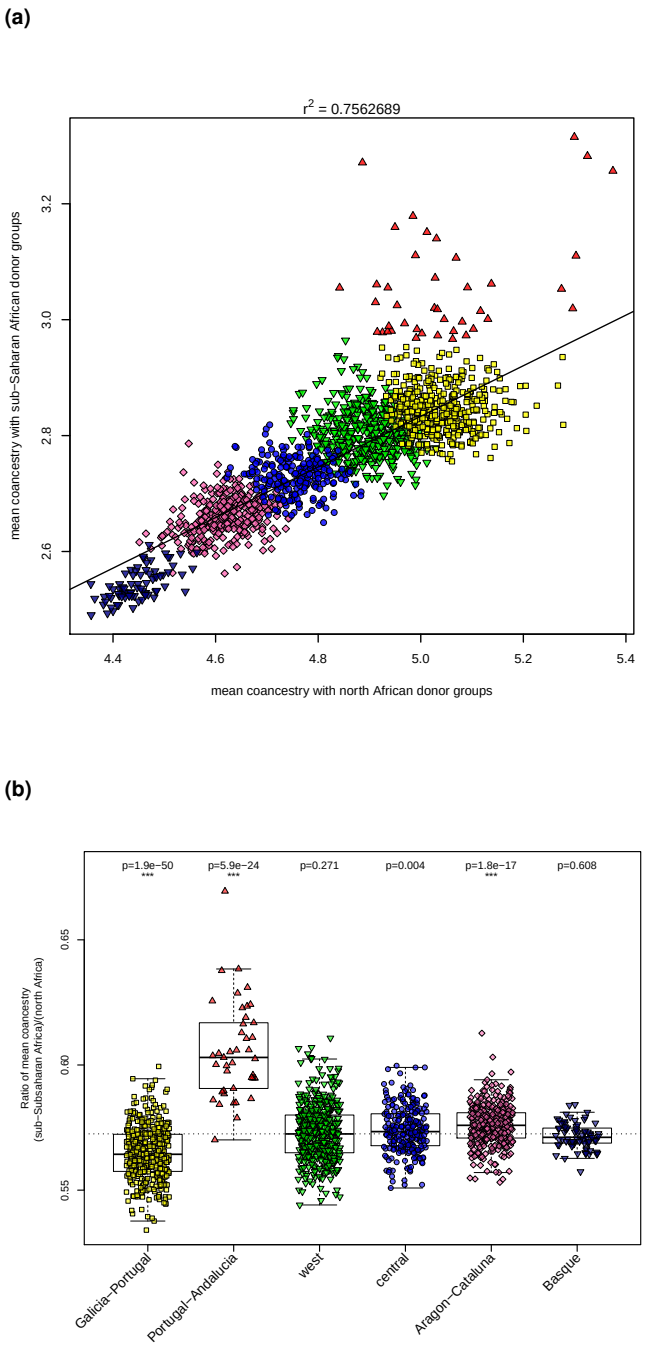


Figure 3.9: Relationship between Iberians' haplotype sharing with north African and sub-Saharan African individuals. (a) Each point represents an Iberian individual, with colours and symbols corresponding to the genetic clusters inferred based on the pattern of haplotype sharing (coancestry) with non-Iberian individuals (see Section 3.4; Figure 3.2). The x and y-axes show the mean coancestry with north African and sub-Saharan African individuals, respectively. For this purpose, sub-Saharan African individuals are made up of the donor groups Kenya.LWK, Kenya.MKK, Nigeria.YRI1 and Nigeria.YRI2; north African individuals are made up of the donor groups NorthAfrica.Mixed, WesternSahara, NorthMorocco, Tunisia, Libya-Algeria, and Egypt. See Figure 3.1 for geographic locations of donor groups, and Section 3.3.1 for details of how donor groups were determined. **(b)** Ratio of sub-Saharan African to north African coancestry. That is, the ratio of the two axes (y/x) in Figure 3.9a. *p*-values are for a two-sided Wilcoxon rank sum test between the given cluster, and all other clusters combined. The dashed line shows the mean value of all individuals combined.

3.6.4 Admixture involving a Basque-like ancestral population

Our earlier results (Chapter 2) imply movement of Basque-like DNA to other parts of Spain. Here we quantify and date this effect, while accounting for other external influences such as those described above. First, we examined the geographic spread of Basque-like ancestry by computing spatially smoothed ancestry profiles as before (Section 3.5.2) but including 60 Basque individuals as a donor group (defined as the clade labelled 'Basque1' in Figure 2.4b). Contributions of Basque-like DNA (Figure 3.10a) are highest in places immediately surrounding the Basque-speaking regions themselves, and also contributions are higher southwards than to the east and west. For example Basque-like DNA levels are higher in Andalucía than in the more proximal regions of Asturias and Cataluña. Once again, this is consistent with more movement of people within Spain along the north-south axis than in the east-west direction.

In order to infer the timing of Basque-like admixture in Iberia, we conducted a second *GLOBETROTTER* analysis (gtB), now introducing the same 60 Basque individuals as a potential surrogate group, and disallowing groups from north Africa and other selected countries as surrogates to avoid confusion with the north African admixture signal (Section 3.6.1). Given we was testing for admixture amongst groups within Iberia, we analysed each of the clusters from the within-Spain analysis, plus Portugal as separate target groups (see Section 3.6.1). We detected an admixture event in all clusters, with one side involving a Basque-like ancestral group, and the other side similar to other European groups (Table A.3b) with proportions generally following the pattern expected from the ancestry profiles (Figure 3.10b). Inferred dates range from 16 to 28 generations ago, corresponding to dates 1190 - 1514 CE and bootstrap confidence intervals all overlap. A signal of Basque-like admixture in non-Basque Spain was reported in a previous study, but no date was precisely estimated [78]. Our result implies that migration of Basque-like peoples to other parts of Spain has taken place, and occurred at a more recent time than the north Africa-related event.

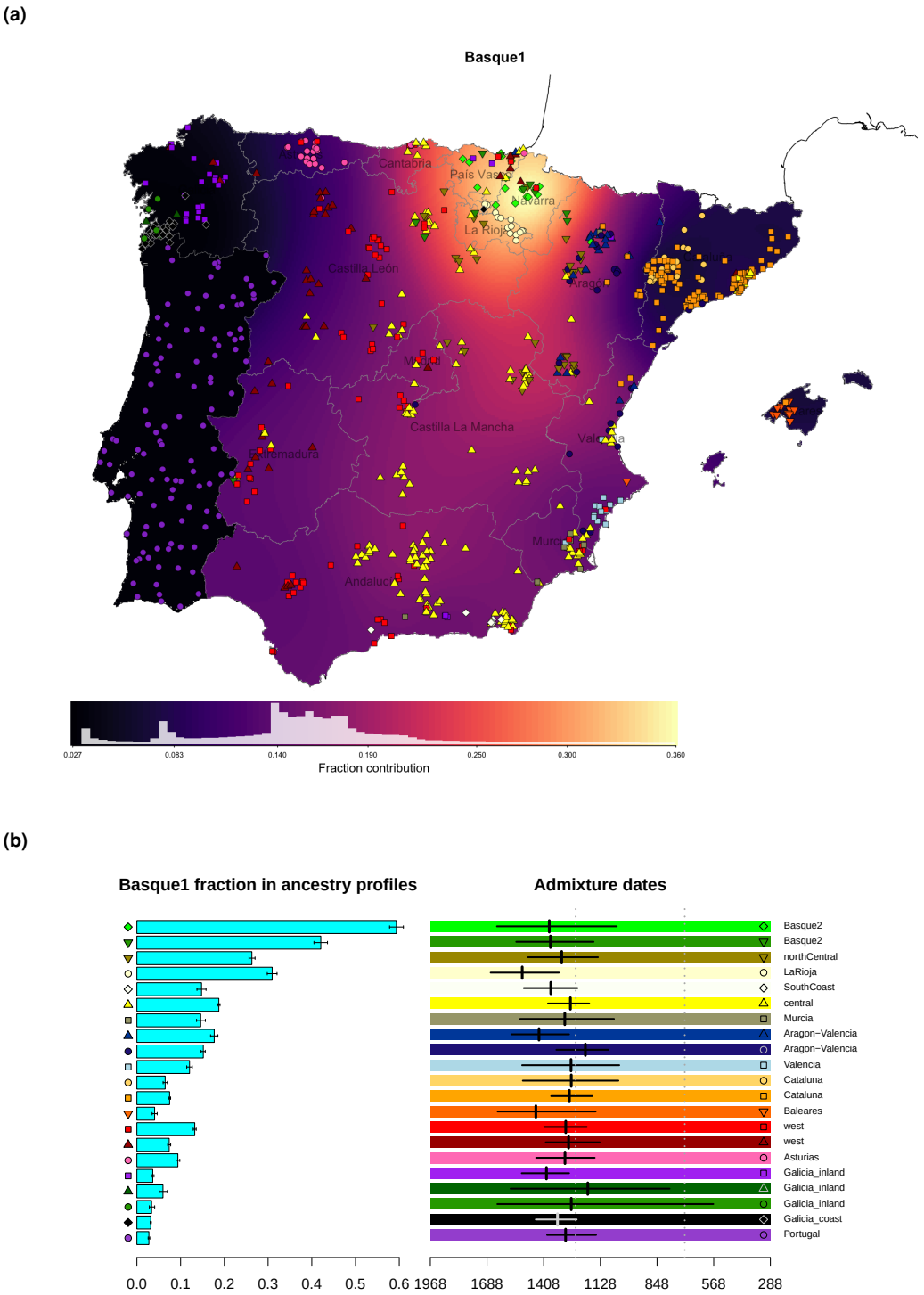


Figure 3.10: Geographic spread and timing of Basque-like genetic contributions to Iberia. (a) Ancestry profiles (which include a Basque-like donor group) have been computed separately at each point on a spatial grid across Spain (see Section 3.5.2), and the map shows the fraction contributed from the Basque donor group (exactly those individuals labelled 'Basque1' in Figures 2.4 and 2.13). The histogram on the scale bars show the distribution of values across grid-points on the map. The points are exactly as shown in Figure 2.4, but without the cluster that was used as the donor group itself. **(b)** Fraction contributions from the Basque-like donor group in ancestry profiles, and Basque-like admixture dates (*GLOBETROTTER*), for each cluster inferred in the Spain-only analysis (as shown in Figure 2.13), plus Portugal. The clade labelled 'Galicia_coast' (Figure 2.4) was combined into one group for this analysis. The admixture dates are for a two-way admixture event involving a Basque-like side and a European-like side, and shown with 95% bootstrap intervals (see Section 3.6.1). The dates shown assume a 28-year generation time, and a 'now' date of 1940 (the approximate average birth-year of this cohort). The white vertical dashed lines show the time of the initial Muslim invasion (711 CE) and the Siege of Seville (1248 CE), exactly as in Figure 3.7a.

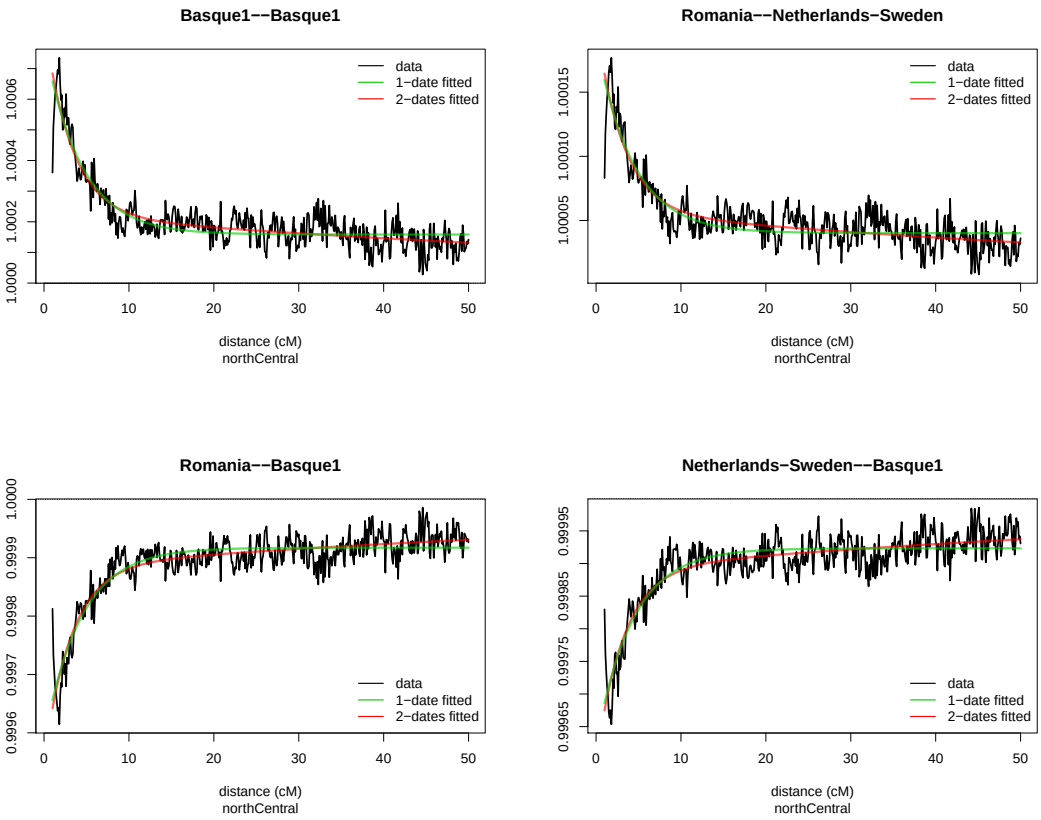


Figure 3.11: Coancestry curves for Basque-like admixture - 'northCentral' cluster. These curves are for an example Spanish cluster ('northCentral') which is centred just south of the Basque-speaking region. Each curve shows the decay of correlation with genomic distance, between segments of DNA with ancestry from the two surrogate groups shown in the titles. The fitted curves for one-date and two-date admixture models are also shown in green and red lines, respectively.

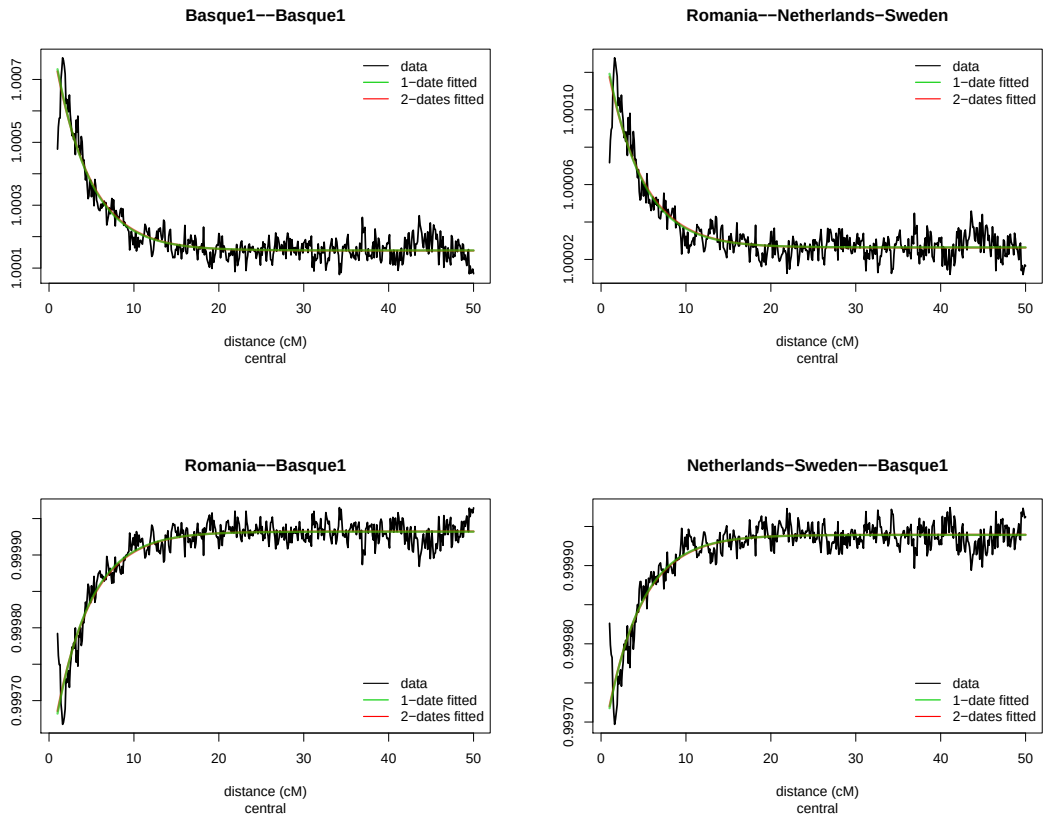


Figure 3.12: Coancestry curves for Basque-like admixture - 'central' cluster. These curves are for an example Spanish cluster ('central'), shown with yellow triangles in Figure 3.10a. Each curve shows the decay of correlation with genomic distance, between segments of DNA with ancestry from the two surrogate groups shown in the titles. The fitted curves for one-date and two-date admixture models are also shown in green and red lines, respectively.

3.6.5 Evaluating the statistical support for inferred admixture events

As recommended by the authors of *GLOBETROTTER* [53] we first tested for evidence of any admixture within each target population by running 100 dating bootstraps using the 'null' procedure under both a one-date and two-date admixture scenario. If D is the number of one-date bootstraps where the inferred date is either greater than 400 or less than 1, then the *p*-value for any admixture is $D/101$. A *p*-value less than 0.01 indicates evidence of admixture. In all target groups for both analyses, all bootstrap date estimates were within this range, implying evidence of an admixture event, or events, for every target group (Table A.3).

After identifying the presence of admixture, we next evaluated evidence for more

complex admixture events (e.g. multiple dates or more than two source populations). *GLOBETROTTER* automatically tests for these using a series of criteria based on how well the coancestry curves fit the models for different types of admixture scenarios [53]. Using *GLOBETROTTER*'s automated criteria, for all target populations there was only evidence for one-date or two-date admixture events. However, given the potentially complex nature of admixture in Iberia, we departed slightly from *GLOBETROTTER*'s automatic criteria when evaluating the evidence for one-date verses two-date admixture events. Specifically, evidence of a two-date admixture scenario can be assessed by measuring how much additional variance is explained by fitting the coancestry curves to a two-date model compared to a one-date model. That is,

$$M = \frac{R^2_{2.date} - R^2_{1.date}}{1 - R^2_{1.date}} \tag{3.6}$$

where $R^2_{1.date}$ and $R^2_{2.date}$ is the goodness-of-fit (coefficient of determination) for a coancestry curve fitted under one-date and multiple-date admixture model, respectively. Under *GLOBETROTTER*'s automatic criteria, if the maximum value of M across all inferred coancestry curves ('maxScore.2events' in Table A.3) is greater than 0.35 this is considered evidence for multiple-date admixture (a value determined by simulations [53]). In our analysis we considered M separately for each pair of surrogate groups inferred to be part of the modern-day mixture representing the admixing sources.

In the case of analysis (gtB), there was no evidence that a two-date admixture model fitted better than a one-date model (the maximum M for any pair of surrogate populations was 0.1). In the case of analysis (gtA), the coancestry curves fit a one-date admixture model very well (across all target groups and coancestry curves the $R^2_{1.date}$ has mean 0.92 and standard deviation 0.1). However, there was some evidence that a two-date admixture model fit better than a one-date model in 4 out of 6 of the target groups, with the strongest evidence for the group 'Portugal-Andalucia' (Figure 3.13). In these cases, *only* the coancestry curves involving a sub-Saharan African surrogate group fit better to a two-date admixture event (see Figure 3.13). The improved fit for the curves involving the sub-Saharan African surrogate group

'Nigeria.YRI1' is visually apparent in Figure 3.8. We therefore consider the one-date admixture event to be a better fit overall, but that there is some evidence for a second event involving sub-Saharan African-like DNA mixing with European-like DNA. In the target groups where there is evidence of this, *GLOBETROTTER* infers dates in the range 1370 - 1700 CE (Table A.3a; Figure A.8).

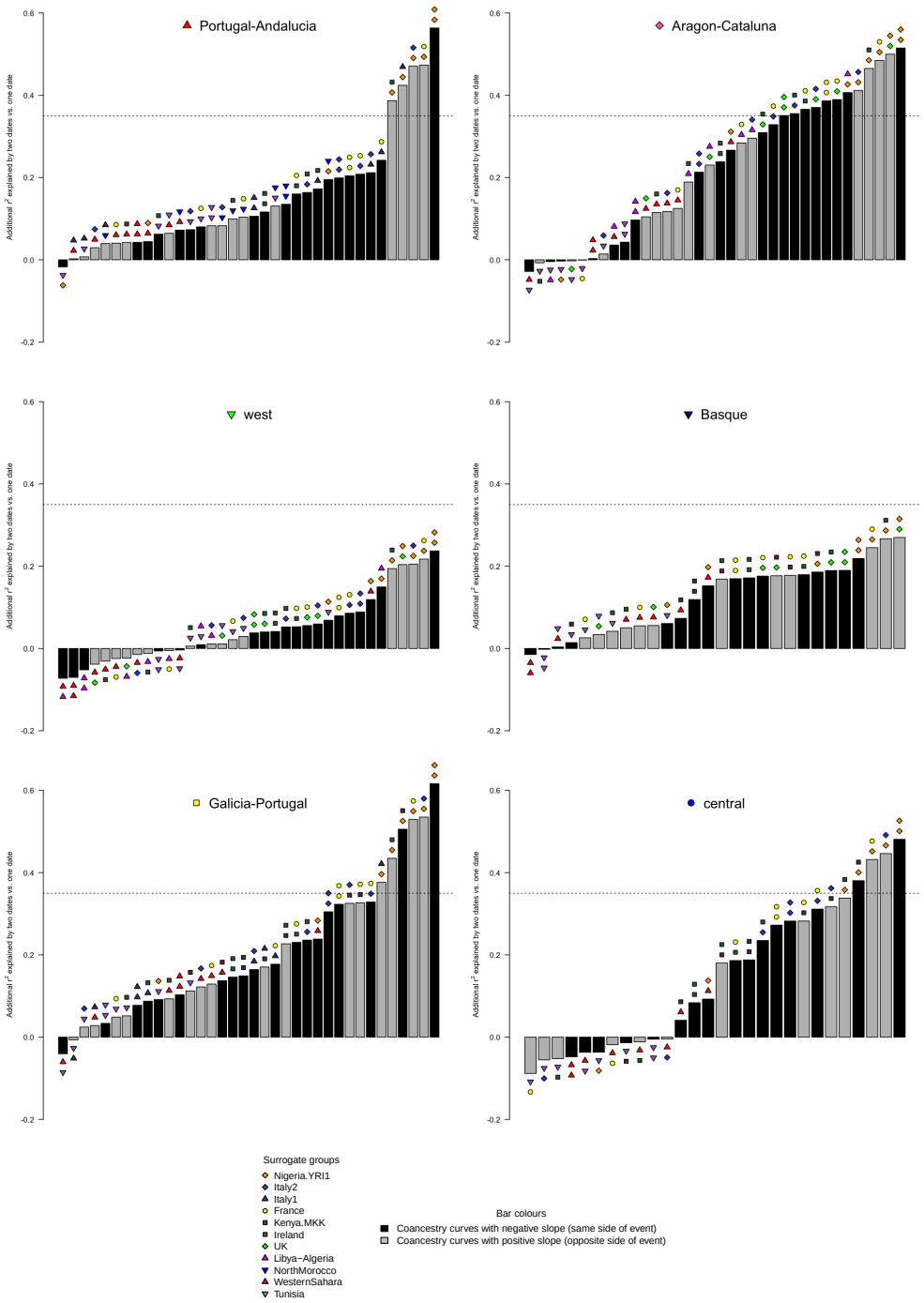


Figure 3.13: *GLOBETROTTER* model fit statistics for one-date verses two-date admixture events All figures refer to results for *GLOBETROTTER* analysis gtA (see Table A.3a). Barplots for each target Iberian group show the fraction of additional R^2 explained by a two-date admixture model compared to a one-date model (see Section 3.6.5). Negative values can occur when the R^2 for a two-date model is lower than for a one-date model, and the dotted line (0.35) is the value above which there is evidence for a two-date admixture event, as recommended by the authors of *GLOBETROTTER*. Pairs of surrogate groups are indicated by colors/symbols above the bars; the color of the bars indicates which pairs have a coancestry curve with a negative (black) or positive (grey) slope, which indicate pairs on the same and opposite side of an admixture event, respectively. Figure 3.8 shows the coancestry curves for the target group 'Portugal-Andalucia' that involve a sub-Saharan African-like surrogate group (YRI), and the fits for two-dates and one-date admixture models.

3.7 Population structure in non-Iberian populations

Here we describe specific aspects of population structure within Europe, north Africa, and sub-Saharan Africa, based on the *fineSTRUCTURE* analyses of non-Iberian groups as described in Section 3.3.1. Most inferred clusters correspond strongly to one (or at most 3) location(s) of origin, as illustrated in Figure 3.1. The following discussion refers to the coancestry matrices for Europe, north Africa, and sub-Saharan Africa shown in Figures 3.15, 3.16 and 3.17 respectively. We also draw insights into the relationships between the different groups from their ancestry profiles shown in Figure 3.14, which have been computed using the procedure described in Section 3.5.1. That is, each group (columns in Figure 3.14) is represented as a linear mixture of all the other groups in the sample (rows in Figure 3.14), and takes into account ancestry sharing between the other groups themselves.

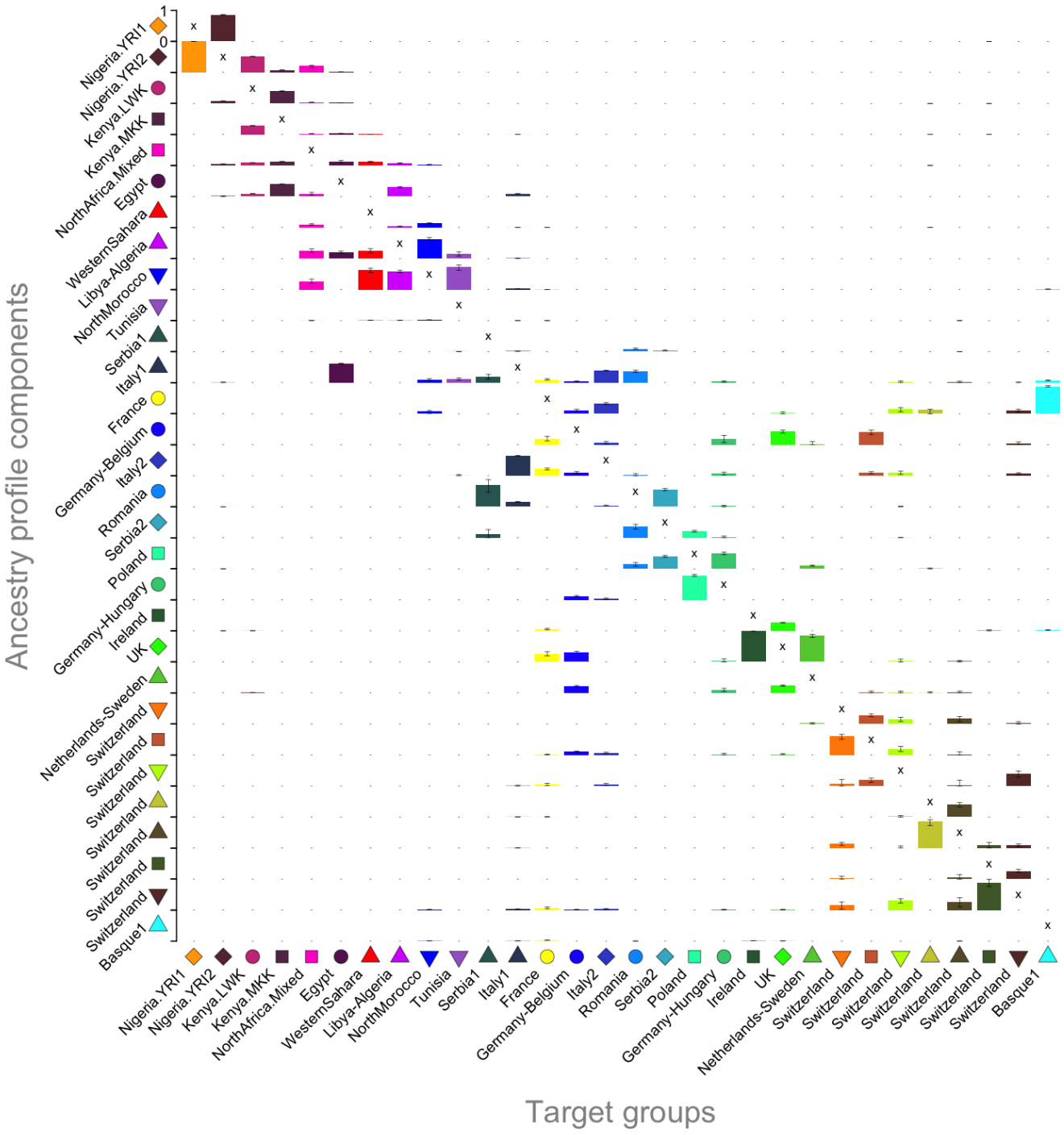


Figure 3.14: Ancestry profiles of non-Iberian groups and Basque cluster. Each column shows the estimated ancestry profile for each of the non-Iberian donor groups plus a Basque cluster (labelled 'Basque1' in Figure 2.13). Ancestry profiles were computed as described in Section 3.5.1, and groups are ordered based on the *fineSTRUCTURE* trees (as in Figure A.6). The heights of the bars within each column sum to one; donor groups are labelled based on the locations of most individuals in each group (see Figure 3.1). Error bars show the range of the inner 95% of 1000 bootstrap re-samples (see Section 3.5.1), and are only shown if the point estimate was greater than zero. Within-group copying was not allowed under the model, indicated by an 'x' on the the diagonal entries.

3.7.1 Structure within Europe

Regions of relatively strong drift are apparent in parts of Europe, as indicated by relatively high levels of within-group coancestry (dark squares in Figure 3.15). Specifically, they are the groups with samples mostly from Serbia, Romania, and Poland, but also Ireland and Portugal. Although the number of samples from Switzerland is similar to the number in UK and Ireland, many more clusters were inferred in Switzerland, indicating a more complex pattern of drift and admixture in the region. This pattern might be expected, given Switzerland's mountainous terrain and multiple official languages (French, German, Italian). Three of the seven Swiss groups (labelled in Figure 3.14 with a red square, a green inverted triangle, and a brown inverted triangle) have non-zero components in their ancestry profiles (i.e. the bootstrap intervals do not contain zero). All three have a small component of the group Italy2, but two of the groups (red square and green inverted triangle) have larger components from the groups Germany-Belgium and France, respectively. This corresponds to the primary language of the majority of individuals in each group: German and French, respectively (data not shown).

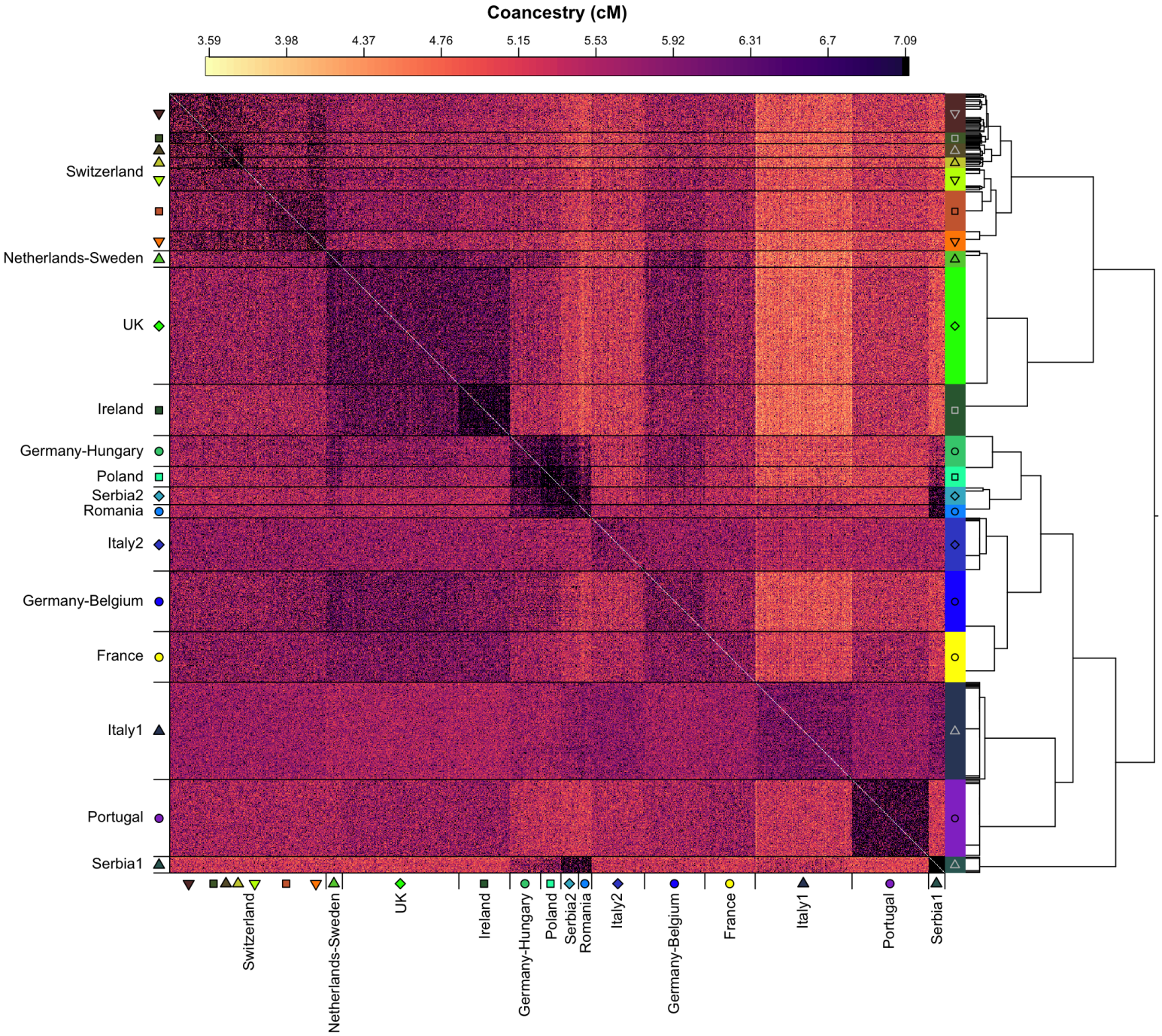


Figure 3.15: Coancestry matrix and inferred *fineSTRUCTURE* clusters for European donor groups. Each row of the matrix is the coancestry vector for a European individual (genotype data sourced from [127]), and ordered according to the results of *fineSTRUCTURE* analysis (CIII) (Section 3.3.1). The groups of individuals indicated by the points and coloured rectangles on the axes are those that we used as European 'donor groups' in various analyses of Iberia, except for the cluster labelled 'Portugal', which was treated the same way as the Spanish cohort (see Section 3.3.2). Some clades of the inferred tree were joined to ensure no group was too small (see Section 3.3.1). Labels were determined based on the locations of the majority of individuals in a given group (see Figure 3.1). Where a cluster was split more evenly across two regions, a double-barrel name is used. In order to visualise the bulk of the variation, coancestry values equal to or above the 90th percentile (7.09 cM) are coloured black.

While most of the inferred groups in Europe contain a majority of individuals from the same country, individuals from Germany do not form their own group (see Figure 3.1). Most German samples cluster with the samples from Belgium, but others cluster either with samples from the east of Germany (Hungary and Austria) or with samples from the west or north of Germany (Netherlands, Sweden). The differences among these groups can be seen most clearly in the ancestry profiles (Figure 3.14). The profile of the Netherlands-Sweden group is dominated by a large component of 'UK', and the ancestry profile for the Germany-Hungary group instead has a significant component from 'Poland', and very little from the 'UK' group. The Germany-Belgium group differs from the other two in having a component from France, and a small (but non-zero) component from the Swiss group (labelled with red squares) mentioned above, with a majority of German speakers. This structure points to a complex mix of population groups within Germany, a country made up of a patchwork of states ('Länder') that have historically had unique cultural and political identities, as well as differing relationships with other regions of Europe [137].

Within the British Isles, the relatively drifted group, 'Ireland' (made up mostly of Irish individuals) differs from the UK group only in the amount of coancestry it shares among its own members, and that it shares less coancestry with two of the continental Europe groups with German individuals, Germany-Belgium and Netherlands-Sweden (Figure 3.15). This is borne out in the ancestry profiles (Figure 3.14), with the Ireland group being almost entirely comprised of the UK group, whereas the profile for the UK group contains a much smaller component from Ireland, with the rest mostly made up of a mixture of the Germany-Belguim and Netherlands-Sweden groups. This may reflect different impacts of migrations into different parts of the British Isles from continental Europe, such as the Angles and Saxons (450 - 500 CE) [42].

Individuals from Italy are almost all contained within two inferred groups. Although the largest components in the ancestry profiles for these groups involve each other, they have quite distinct features. The group Italy1 — which also includes some individuals from Turkey and Greece — has components in its ancestry profile from Romania and north Africa (especially Egypt), and none from France. The other group (Italy2) —

which also includes some individuals from France and Switzerland — only has European components in its ancestry profile, chiefly France. These two groups may broadly reflect genetic differences between north and south Italy [138], although there is not sufficient geographic information within the POPRES resource to directly confirm this.

3.7.2 Structure within north Africa

The pattern of ancestry sharing among north African groups, as shown in the coancestry matrix Figure 3.16a, reveals a more complex structure than the straight-forward west-to-east decline of Western Saharan (or Maghrebi) ancestry as noted by Henn *et al.* [31], which was based on an ADMIXTURE analysis of the same data. The strongest signal of drift among the north African samples appears in the group Tunisia (comprised entirely of samples from Tunisia), with mean within-group coancestry 239 cM as shown in Figure 3.16b. This reflects the observation of long IBD segments shared between Tunisian samples also reported by Henn *et al.*. [31] but other patterns are revealed by the *fineSTRUCTURE* approach. Specifically, the group WesternSahara also has elevated within-group coancestry (mean 97 cM); individuals from south Morocco are spread amongst several genetic groups, depending on their level of coancestry with the group WesternSahara; and in contrast to the groups Tunisia and WesternSahara, the groups NorthMorocco and Libya-Algeria share similar levels of coancestry with each other as they do within their own group, suggesting there has been relatively unobstructed gene-flow across a region that spans over 3,000 kilometres west to east.

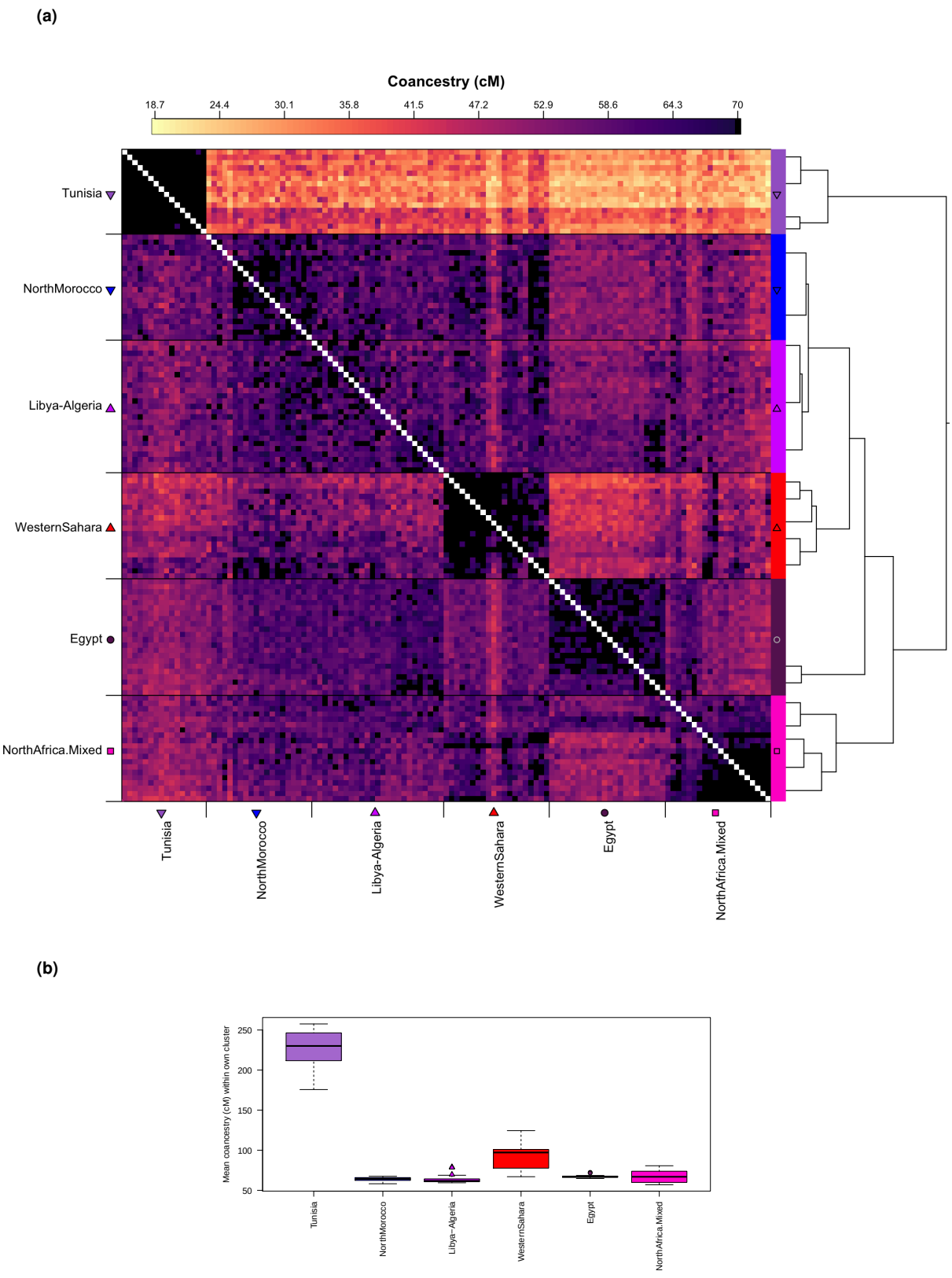


Figure 3.16: Coancestry matrix and inferred *fineSTRUCTURE* clusters for north African donor groups. (a) Each row of the matrix is the coancestry vector for a north African individual (genotype data sourced from [31]), and ordered according to the results of *fineSTRUCTURE* analysis (CII) (see Section 3.3.1). The groups of individuals indicated by the points and coloured rectangles on the axes are exactly those that we used as north African 'donor groups' in various analyses of Iberia. Some clades of the inferred tree were joined to ensure no group was too small (see Section 3.3.1). Labels were determined based on the locations of the majority of individuals in a given group (see Figure 3.1). Where a cluster was split more evenly across two regions, a double-barrel name is used. In order to visualise the bulk of the variation, coancestry values equal to or above the 90th percentile (70 cM) are coloured black. **(b)** The mean coancestry (cM) with other individuals assigned to the same north African donor group, as demarcated in Figure 3.16a.

The relationship between the the North African and sub-Saharan African groups is also worth highlighting. Three of the north African groups have one or more components in their ancestry profile (Figure 3.14) from sub-Saharan Africa: WesternSahara, Egypt and NorthAfrica.Mixed. This is especially pronounced with the group NorthAfrica.Mixed, with 27% of its ancestry profile being made up of sub-Saharan African groups (compared with 6% for Egypt and 0.4% for WesternSahara). The group NorthAfrica.Mixed is made up of individuals from Morocco, Algeria and Libya. This suggests that their genetic similarity could partly be explained by having shared sub-Saharan African ancestry, which is best represented (in this data set) by the Nigerian group Nigeria.YRI2.

3.7.3 Structure among sub-Saharan African groups

In order to account for any sub-Saharan African ancestry in the Iberian genomes, we included three populations from the Hapmap3 data source in the set of donor groups. Namely, LWK, MKK, and YRI. Given that these groups originate from geographically distant regions (shown in Figure 3.1), it is unsurprising that *fineSTRUCTURE* was able to cleanly distinguish between the three groups (Figure 3.17). In addition, this analysis reveals further sub-structure, especially amongst the Maasai samples (Kenya.MKK). Two distinct groups were detected among the YRI individuals, and to our knowledge, this heterogeneity has not been reported before in studies using data for the same set of individuals [135, 139, 140]. Specifically, the group Nigeria.YRI1 has more within-group coancestry, and individuals in the group Nigeria.YRI2 share more coancestry with individuals from the other group than with themselves, indicating gene-flow from the more drifted group (Nigeria.YRI1) into the other (Nigeria.YRI2), analogous to the pattern associated with Basque-like groups discussed in Chapter 2. Another pattern of note is that the Kenyan group LWK (Kenya.LWK) shares more coancestry with the YRI individuals (especially Nigeria.YRI1) than with the other Kenyan group, MKK, despite being geographically much closer. This is consistent with an observation made in a more comprehensive study of sub-Saharan African population structure and admixture [140], that Luhya likely experienced greater levels of gene-flow from western Africa

during the ‘Bantu expansion’ compared to the Maasai (Figure 6A in [140]).

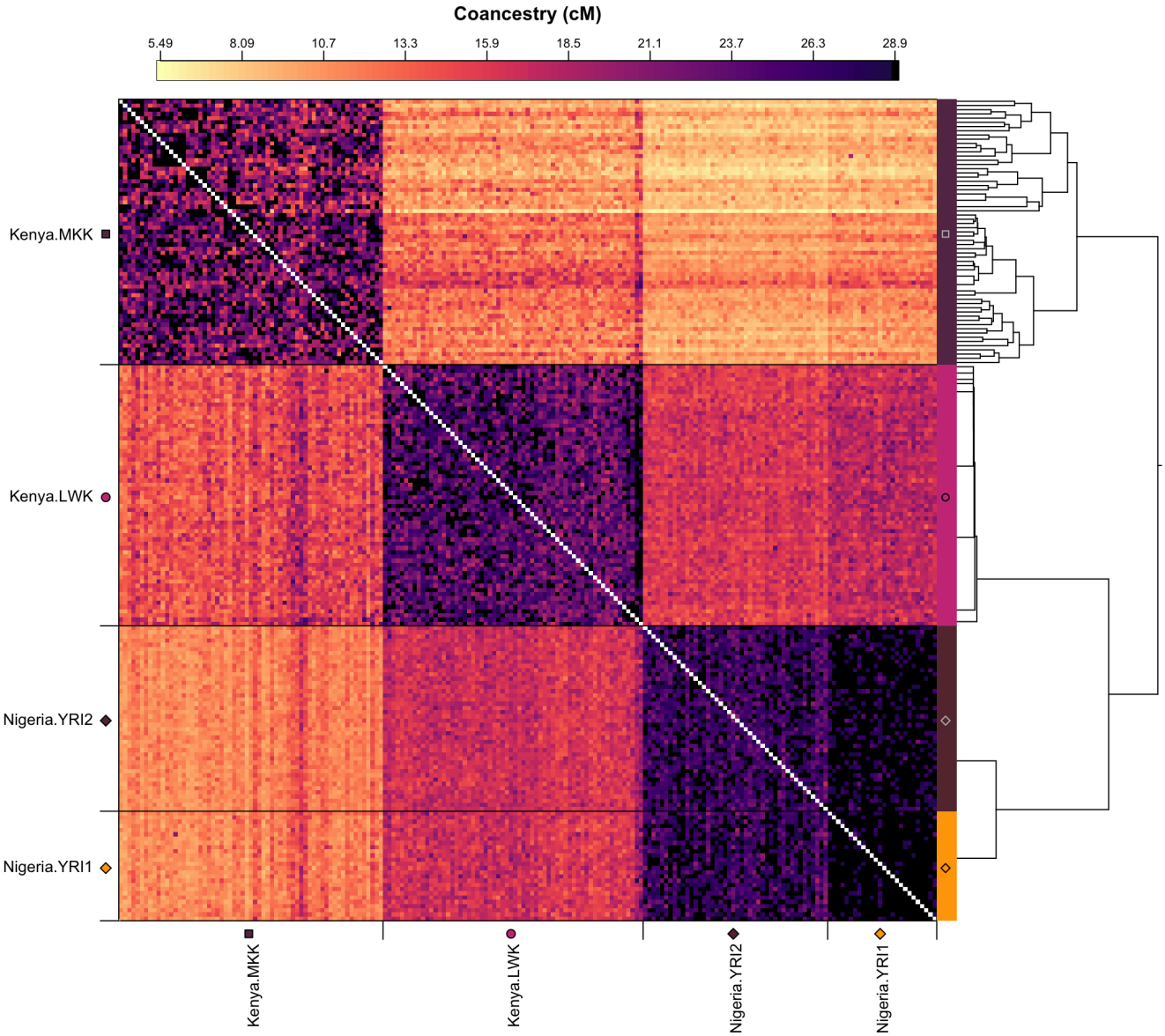


Figure 3.17: Coancestry matrix and inferred *fineSTRUCTURE* clusters for sub-Saharan African donor groups. Each row of the matrix is the coancestry vector for a sub-Saharan African individual (genotype data sourced from [135]), and ordered according to the results of *fineSTRUCTURE* analysis (CI) (see Section 3.3.1). The tree on the right is the specific clade of the inferred hierarchical tree (analysis (CI) involved other samples) which involves only the sub-Saharan African individuals. The groups of individuals indicated by the points and coloured rectangles on the axes are exactly those that we used as sub-Saharan African ‘donor groups’ in various analyses of Iberia. Some clades of the inferred tree were joined to ensure no group was too small (see Section 3.3.1). Labels were determined based on the locations of the majority of individuals in a given group (see Figure 3.1). Where a group was split more evenly across two regions, a double-barrel name is used. In order to visualise the bulk of the variation, coancestry values equal to or above the 90th percentile (28.9 cM) are coloured black.

3.8 Discussion

By applying the *fineSTRUCTURE* method to a combined set of European and African samples, we have uncovered a complex pattern of population structure within the continents, as well as evidence of cross-continental gene-flow. It is clear that different regions have experienced quite different amounts of genetic drift (e.g. compare the group 'Serbia1' or 'Tunisia' to the 'UK'). While in many cases the country-level labels based on the origin of individuals tracks strongly with their assigned genetic clusters, there are several instances where sub-structure can be found within a country (most notably in Italy, Serbia, and Switzerland). The converse is also true, where individuals across several countries are genetically very similar (e.g. across north Africa; Germany and Belgium). This is not too surprising, given that many of the current-day national boundaries have been subject to change, even over the last 100 years (for example in central Europe [137]), and are therefore unlikely to align precisely with the demographic forces that have shaped the population structure we see today.

The analysis of Iberia in relation to external regions is informative of what forces have likely driven the structure seen in within-Spain analysis. Specifically, as discussed in Chapter 2, we detected population structure at ultra-fine scales in Galicia, with some clusters having geographic ranges of less than 10 Km. Contrastingly, when we only consider haplotype sharing between Galicia and non-Iberian groups this structure disappears, and individuals from Pontevedra are indistinguishable from those from Portugal and other parts of western Spain (Figure 3.2). Therefore, the very strong population stratification observed in Galicia can most easily be explained by very recent isolation. That is, a uniquely (within Spain) low level of immigration to Galicia over the last 500-600 years [141, 142], as well as localised societal organisation (as discussed in Section 2.8).

Surprisingly, the similarity between individuals in Galicia and Portugal includes sharing relatively high levels of north African-like DNA, despite the fact that the northern part of Portugal was only briefly under Muslim rule. Berber settlements north of the Douro river were abandoned by 741, and the region of Galicia as it is defined

today (north of the Miño river), was never under Muslim rule [143]. A similar observation has been made using data on Y-chromosome haplotypes [1, 27]. Our *GLOBETROTTER* analysis places the timing of the arrival of this ancestry within the period 760 – 1090 CE (Table A.3a). Given that much of Portugal did experience long-term Muslim rule, and now has a higher density of arabic place names than in other parts of Spain (Figure 1.3), this likely reflects – at least in part – historical migration from modern-day Portugal into Galicia. Despite a generally low immigration rate into Galicia, the region currently has the highest amount of recent immigration from Portugal among all the regions in Spain, and the Portuguese represent the largest non-Galician contingent in Galicia by far. This is especially true for the Galician population sampled in this study, most of which were from Pontevedra, which shares a border with Portugal [144].

As discussed in Chapter 2, we also observe strong genetic differentiation between individuals in the Basque region and the rest of Spain, as well as sub-structure within the region. However, unlike the case of Galicia, individuals in the Basque region also show a distinct pattern of haplotype sharing with non-Iberian groups, indicating that the genetic distinctiveness of Basque-like DNA is a result of isolation older than that in Galicia, and probably pre-dates major migration events into Iberia, such as the Muslim invasion of 711. However, the small amounts of north African-like DNA, and clear evidence of the same admixture event within the DNA of individuals from the Basque region as elsewhere in Spain, indicates that the region has either experienced small amounts of on-going gene-flow from other Spanish groups, or directly from north Africa.

Gene-flow from north Africa has clearly made its mark on Iberia in a way that involves striking regional variation. Our results also imply that north African-like DNA in Iberia was mainly, and perhaps almost exclusively, introduced within the earlier half of the period of Muslim rule (Figure 3.7a). The genetic legacy of migration from north African, as represented by independent contributions from the NorthMorocco group (Figure 3.5), ranges from near (but above) 0% to 10% and varies strongly, but relatively smoothly, across Iberia. Highest levels are in the west, especially within the borders of Galicia and Portugal, and lowest in the east, especially the Basque region,

Cataluña and Beleares (Balearic Islands). This overall pattern is consistent with previous studies [1, 2], but reveals variation with greater spatial resolution. Interestingly, we do not see a strong north-south gradient in north African ancestry, which one might expect given the north-south trajectory of the *Reconquista*, the relative proximity of the southern regions of Iberia to north Africa, and the variation in density of arabic place names in Iberia (Figure 1.3). It is therefore plausible that the pattern we see might be influenced, instead, by later internal migratory patterns. For example, from south to north into Galicia, which would be consistent with the results of the *fineSTRUCTURE* analysis in Chapter 2 involving Portugal (see Figure 2.12). Alternatively, it might be that these patterns instead reflect regional differences in patterns of settlement and integration with local peoples, of north African immigrants themselves. Large-scale expulsion of Muslim populations, post-Reconquista, was common in Iberia, especially in towns and cities [58, 57], and may also have influenced the regional variation in north African-like DNA. By leveraging fine-scale geographic information we have isolated an area (apart from the Basque region) with a distinctly low genetic contribution from north Africa. Interestingly, this covers a similar area to the region known as the Crown of Aragon in the 14th Century (see Figure 1.4).

Chapter 4

Population structure, relatedness, and quality control of genotype data for 0.5 million UK Biobank participants

4.1 Chapter overview

The UK Biobank genetic data that was made available to approved researchers in July 2017 contained genotype calls for 488,377 UK Biobank participants at 805,426 unique markers (SNPs and indels), as well as a number of auxiliary data files containing further information of use in many kinds of downstream analyses[1]. In this chapter we cover the two broad streams of work we carried out to produce the suite of data files released to researchers. Namely, designing and applying a quality control pipeline to the genotype calls provided to us by Affymetrix; quantifying the quality of the resulting data set; and detecting population structure and cryptic relatedness among the UK Biobank participants.

[1] A full list of the data files associated with UK Biobank genetic data can be found here: http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100314

4.1.1 Quality control in a large-scale, ethnically diverse cohort

Our QC pipeline was designed specifically to accommodate the large-scale dataset of ethnically diverse participants, genotyped in many batches (106), using two slightly different novel arrays, and which will be used by many researchers to tackle a wide variety of research questions. Our aim was not to 'fix' the data. Rather, we either set to missing, excluded, or flagged parts of the data set for which there was evidence that the genotyping had performed poorly. This could be a result of something that had gone wrong in the laboratory processing (e.g. contamination or sample mishandling) or because the genotype calling algorithm had misinterpreted signals in the intensity data.

As part of the baseline UK Biobank questionnaire, participants reported their ethnic background by selecting from a fixed set of categories [145]. While the majority (94%) of individuals report their ethnic background as within the broad-level group 'White', there are $\sim$22,000 individuals with a self-reported ethnic background originating outside Europe (Table 4.1). This ethnic diversity implies genetic diversity, which we observe directly in the genotypes as allele frequency differences, and has implications for QC. Some commonly used QC tests are ineffective in the context of strong population structure if applied without taking this into account. For example, extreme values in the fraction of non-missing markers that are called heterozygous (heterozygosity) and/or missing rates can be indicators of poor sample quality due to, for example, DNA contamination [118]. The heterozygosity of a sample can also be sensitive to natural phenomena, including population structure, recent admixture and parental consanguinity. We took extra measures to avoid misclassifying good quality samples because of these effects.

Another challenge related to the size of the cohort, is that rare genetic phenomena are likely to be observed frequently. Rare SNPs or indels are generally more difficult to genotype accurately using array-based technology, but they are also of special interest to researchers, most notably to study rare disease [146] and estimating heritability [147]. Furthermore, rare, large-scale structural variants could cause problems with genotype calling if they do not align with the model assumed by the calling algorithm.

For example, a sample with a chromosomal trisomy may have an excess of missing calls, or miscalled genotypes because the markers on that chromosome are triploid rather than diploid. In this case it would be incorrect to assume that the sample was poorly-performing because of experimental error and to advise researchers to ignore the data. The most common forms of aneuploidy with a life expectancy beyond 40 years involve the sex chromosomes or chromosome 21 (Down's syndrome) [148][2]. We looked for individuals with likely sex chromosome aneuploidy and provide a list of these to researchers. For rare SNPs and indels, we broadly assessed the quality of the genotype calling by comparing the allele frequencies observed in the UK Biobank with those in an independent (and large) cohort of sequenced individuals, the Exome Aggregation Consortium (ExAC).

To address the fact that this data would be used to tackle many different types of research questions, we only removed data where there was strong evidence of poor quality data, and for sample-based QC we did not remove any samples from the data, but instead provided relevant information in the form of an indicator variable, or the numerical values of a QC metric itself.

Finally, as a check of the quality of the data and a demonstration of its utility, we conducted GWAS on the phenotype, standing height. We chose this phenotype because it has many previously-known regions of association and has been studied using large sample sizes ($>$ 100,000), thus providing opportunities for comparison. We conducted the GWAS using both the directly genotyped data ($\sim$800,000 markers) and the imputed data set, which allowed us to test $\sim$96 million more markers.

[2]Babies with three copies of chromosomes 13 ('Patau syndrome') or 18 ('Edwards syndrome') are known to survive to birth, but are much rarer than Down's. Down's has been estimated to occur in about 1.3 of every 1,000 live births, whereas three copies of 13 or 18 occurs in about 2 of every 10,000 live births [148].

Ethnic group	Self-reported ethnic background	Percentage in UK Biobank genotype data
White		**94.23**
	British	88.26
	Any other white background	3.24
	Irish	2.61
	White	0.11
Asian or Asian British		**1.94**
	Indian	1.17
	Pakistani	0.36
	Any other Asian background	0.36
	Bangladeshi	0.05
	Asian or Asian British	0.01
Black or Black British		**1.57**
	Caribbean	0.88
	African	0.66
	Any other Black background	0.02
	Black or Black British	0.01
Chinese		**0.31**
Mixed		**0.58**
	Any other mixed background	0.2
	White and Asian	0.16
	White and Black Caribbean	0.12
	White and Black African	0.08
	Mixed	0.01
Other/Unknown		**1.38**
	Other ethnic group	0.89
	Not stated	0.48

Table 4.1: Proportions of self-reported ethnic groups among 488,377 genotyped UK Biobank participants. Categories of self-reported ethnic background (UK Biobank data field 21000) and broader-level ethnic groups are shown here to reflect the two-layer branching structure of the ethnic background section in the UK Biobank touchscreen questionnaire [145]. Participants first picked one of the broader-level ethnic groups (e.g. White), and were then prompted to select one of the categories within that group (e.g. Irish). The broader-level groups are also shown here as an ethnic background category (e.g. 'White' in column two) because a small proportion of participants only responded to the first question. In this table we also combine the category 'Other ethnic group' with an aggregated non-response category 'Not stated', which includes all participants who did not know their ethnic group, or stated that they preferred not answer, or did not answer the first question.

4.1.2 Characterising population structure and cryptic relatedness among UK Biobank participants

The diversity in ancestral origins of UK Biobank participants is evident from the self-reported ethnic background (Table 4.1) and country of birth information. The

genotype data provides a unique opportunity to study their ancestral origins in a quantitative manner. Accounting for the ancestral background of participants is an essential component of analysis of the UK Biobank resource, both for epidemiological studies [149], and genetic analyses, such as GWAS [92, 91]. We used principal components analysis (PCA) to measure population structure within the UK Biobank cohort. PCA is widely used as a method for assessing and potentially controlling for population structure in GWAS [92, 150], and it was also crucial to many elements of our QC pipeline. We also looked for evidence of population structure amongst participants of white British ancestral backgrounds, and whether (and how) this is related to the geography of the UK.

Close relationships (e.g. siblings) among UK Biobank participants were not recorded during the collection of other phenotypic information. Indeed, many participants may not be aware that a close relative (such as an aunt, or sibling) is also part of the cohort. This information can be important for epidemiological analyses [151], as well as in GWAS [93], and the genetic data provides an opportunity to discover and characterise familial relatedness within the cohort. This analysis, combined with phenotype information, is also useful for identifying samples that are experimental duplicates rather than genuine twins.

In detecting both population structure and cryptic relatedness, it was necessary to use methods and software that could handle the scale and diversity of this data set. For cryptic relatedness we also conducted complementary analyses to assess the robustness of the results, as well as estimating how many related individuals we should expect to see in a sample of this size in the UK.

4.2 Marker-based quality control

4.2.1 Tests for poor quality markers

We[3] identified poor quality markers using statistical tests designed primarily to check for consistency across experimental factors. Specifically, we tested for batch effects, plate effects, departures from Hardy-Weinberg equilibrium (HWE)[4], sex effects, array effects, and discordance across control replicates. The full details or each test are described in the Supplementary Material of [115]. Briefly, for all tests except departures from HWE and discordance across controls, a Fisher's exact test was applied to the 2x3 (or 2x2 in haploid cases) tables of genotype counts, where each row is the counts for an experimental unit. For example, in the batch effect test, one row is the counts of genotypes in the batch being tested, and the other row is the counts of genotypes in all other batches (genotyped using the same array) combined. We tested for departures from HWE using an exact test described in [152] and implemented by the authors of *plink*. In order to attenuate population structure effects we applied all tests using a subset of 463,844 individuals with estimated European ancestry. We identified these individuals from the genotype data prior to conducting any QC by projecting all the UK Biobank samples on to the two major principal components of four 1000 Genomes populations (CEU, YRI, CHB and JPT) [153]. We then selected samples with principal component (PC) scores falling in the neighbourhood of the CEU cluster.

[3]The design and implementation of this part of the pipeline was carried by D.P. and C.F., but the summaries shown in Section 4.5 were carried out by the author of this thesis.

[4]Hardy-Weinberg equilibrium (HWE) refers to the relationship between the frequency of alleles and the frequency of genotypes that is expected to occur in a large, panmictic population and in the context of no selection. It was first described (independently) by Godfrey Harold Hardy and Wilhelm Weinber in 1908 [152]. Consider a diploid organism (e.g. humans) and a SNP with two possible alleles (bi-allelic). The genotype of an individual (AA, Aa, or aa) is equivalent to a binomial random variable with two trials, and where the number of 'successes' is the number of A alleles (without loss of generality) in a given genotype (AA = two successes), and the probability of 'success' is equal to the frequency p of that allele. The expected fractions of each genotype are therefore: $Pr(Aa) = 2p(1-p)$, $Pr(AA) = p^2$, and $Pr(aa) = (1-p)^2$.

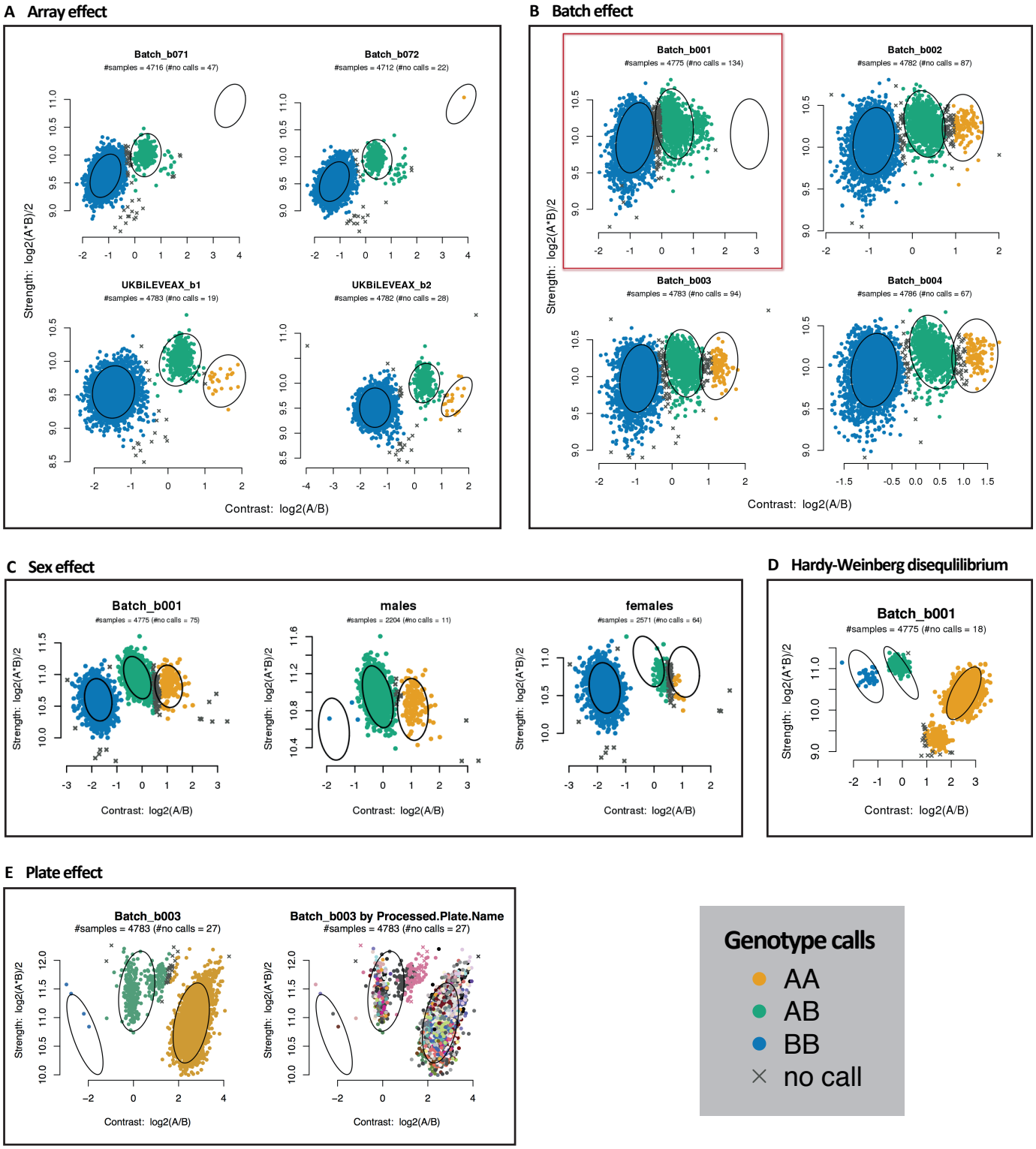


Figure 4.1: Examples of markers failing quality control tests. Each sub-figure shows an example of a marker exhibiting the properties that our QC tests were designed to identify (p-value $< 10^{-12}$). See Section 4.2.1 for details. Each plot shows the samples within the stated batch coloured according to their inferred genotype, as in Figure 1.12. The limits of the axes vary depending on the range of intensities observed in each batch. **(A)** The top two plots show batches typed on the UK Biobank Axiom array, and the bottom two show batches typed on the UK BiLEVE Axiom array. The third cluster (orange; minor homozygotes) has been called as homozygote (green) in the UK Biobank Axiom array batches, likely due to the presence of the outlier in Batch_b072. **(B)** Genotype calls in the highlighted batch (Batch_b001) contain no minor homozygotes (orange), unlike the other three batches shown. **(C)** One batch is shown here, but also with males and females plotted separately. There are only two clusters for each of males and females, but they are shifted relative to each other so form what appears to be three clusters when combined. This is an autosomal marker, so males and females are genotyped together. **(D)** The presence of a fourth cluster suggest that this marker involves variation more complex than a bi-allelic marker. The samples in the fourth cluster that were called as major homozygotes causes the genotype counts to violate HWE. **(E)** This batch for this marker contains two plates (shown as dark brown and pink dots in the right-hand plot) that are systematically shifted in intensity space.

We considered a test to be 'failed' if it yielded a *p*-value smaller than 10^{-12} (see Section 4.2.2 for rationale), and for the concordance test if a marker had a discordance rate[5] greater than or equal to 5% in either of the controls. Four of the tests (batch effect, plate effect, departures from HWE, sex effect) were applied to each marker in each batch separately. For markers that failed at least one test in a given batch, we set the genotype calls in that batch to missing. If a marker failed any one of the tests in every batch, or failed at least one of the other two tests (array effect or concordance across controls), we excluded the marker from the data altogether.

In addition to the six tests, some of the genotype calls for a small number of samples (387) were likely compromised due to a rare artefact involving the digital misalignment of features on the array during image-capture. The intensities for these samples at a subset of SNPs were systematically shifted in a way that imitated a real genotype cluster, so were not highlighted by other QC tests but were identified in a principal components analysis (data not shown). We set the genotype calls to missing only for these samples at a set of 34,921 markers that Affymetrix identified as likely affected (0.0037% of all genotype calls).

4.2.2 Choice of *p*-value for hypothesis-based tests

The batch effect, plate effect, HWE, sex effect and array effect tests are hypothesis tests with an associated *p*-value. Any marker or marker/batch combination with a *p*-value smaller than a fixed threshold we considered as failing the test. Our threshold of 10^{-12} was chosen so that we only set a marker/batch to missing if there is very strong evidence for deviation from the null hypothesis of any of these tests.

There are 5 kinds of hypotheses, 106 batches, $\sim$50 plates per batch and $\sim$800,000 markers, making a total of around 4.6×10^9 tests (accounting for plate-level and batch-level tests). A *p*-value of 10^{-12} can therefore be thought of as equivalent to a family-wise error rate of at most 0.005. Many tests will be positively correlated, especially across batches for the same marker, so this is likely to be an upper bound on the probability

[5]Each of the control samples were genotyped $\sim$5,500 times, and the discordance rate for a marker is the fraction of genotype calls not equal to the most frequent call for that marker.

that we observe an extreme test statistic just by chance.

4.2.3 Results of marker-based quality control

The amount of data affected by applying these tests is summarised in Table 4.2. The most common effect was differences in allele frequencies among the samples typed on the two different genotyping arrays (see Figure 4.1A for an example). While the two arrays have a large number of markers in common (see Table A.5), some aspects of the design of the physical genotyping array differed. These differences can have subtle effects on the distribution of observed intensities for a marker, and in some cases this likely influenced the behaviour of the genotype calling algorithm across the two arrays, for the same marker.

Test	Average number of SNPs failed per batch (sd)	Fraction of all genotype calls affected
Affymetrix cluster QC	1109 (699)	0.00140
1. Batch effect	197 (86)	0.000249
2. Plate effect	284 (266)	0.000358
3. Departure from Hardy-Weinburg equilibrium	572 (77)	0.000723
4. Sex effect	45 (5)	0.0000569
5. Array effect*	5417	0.00683
6. Discordance across controls**	622 (ukbl), 632 (ukbb)	0.000796
Total	7704 (721)	0.00971

Table 4.2: Failure rates for six marker-based quality tests and Affymetrix filters. For all numbered tests a marker (or marker within a batch) was set to missing if the test yielded a p-value $< 10^{-12}$, except in the case of Test 6, for which a marker was set to missing if the test yielded $<$ 95% concordance. See Section 4.2.1 for a summary description of the tests. The total is not equal to the sum of all tests because it is possible for a marker to fail more than one test. Since the two arrays contain slightly different sets of markers, the total number of genotype calls used to compute the fractions is, N_{ukbb} L_{ukbb} + N_{ukbl} L_{ukbl}, where N and L refer to the numbers of markers and samples typed on the UK Biobank Axiom array (ukbb) and samples typed on the UK BiLEVE Axiom array (ukbl) within the Affymetrix data delivery (see Table A.5). *The array effect test was applied across all batches and only for markers present on both arrays, so we simply report the total number of markers that failed this test. **The discordance test was applied across all batches, but not all markers are present on both arrays. The first value is the number of unique markers on the UK BiLEVE Axiom array that failed this test, and the second is for markers on the UK Biobank Axiom array.

It is also worth noting that the participants who were typed on the UK BiLEVE Axiom array were selected from the whole cohort based on phenotypes involved in lung function and smoking behaviour, among those with self-declared European ancestry [107]. Consequently, it is possible to observe what look like array effects as a result of genuine genotypic differences between the two sets of participants; for example, at loci associated with smoking behavior or lung function. While most of the markers with evidence of an array effect are scattered across the genome, we found one set of markers with low *p*-values clustered within and around the gene CHRNA3, a locus known to be associated with smoking behaviour [107] (see Figure 4.2). Since this signal is likely to reflect genuine phenotype-genotype associations and not an experimental artefact, we did not exclude any marker in this region on the basis of the array effect test. Specifically, chromosome 15, positions 78.679.0 Mega-bases (Mb).

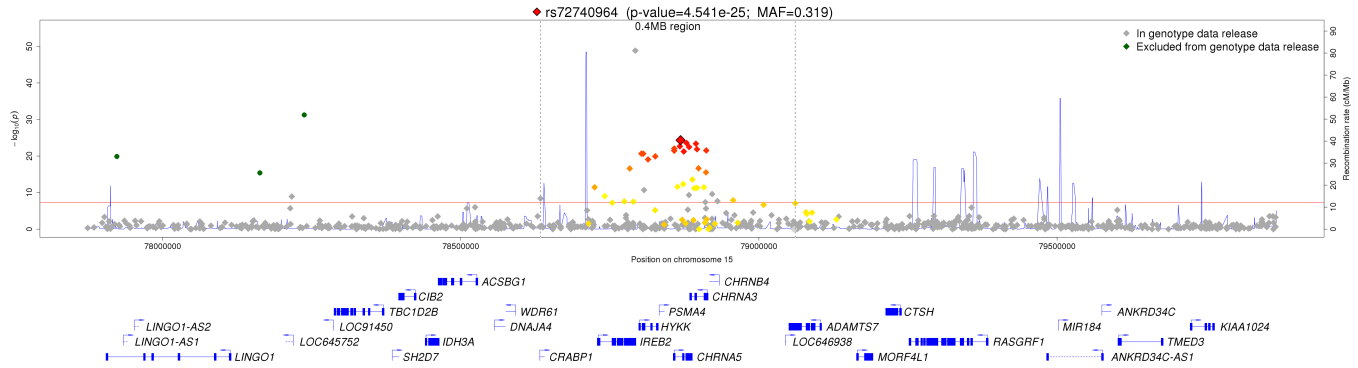


Figure 4.2: Array effect test near smoking-associated region. The array effect test compares genotype counts across the two arrays using a Fisher's exact test (see Section 4.2.1). Each marker is shown as a diamond, and coloured according to it's correlation (r^2) with the marker highlighted with a black border (red is close to 1, gray is < 0.1). The gray marker with the smaller *p*-value $< 10^{-12}$ has $\sim$20% missing data (and exists on both arrays), so is likely to be a generally poorly-performing marker. We did not exclude any of the markers between the dashed vertical lines, even though they failed the array effect test (*p*-value $< 10^{-12}$).

4.3 Overview of sample-based quality control and analysis pipeline

The pipeline for sample-based QC and analysis was designed to identify samples with poor quality genotype calls, find related individuals, and provide a quantitative

description of ancestral diversity of the cohort based on information in the genetic data. We used a set of high quality SNPs that were typed on both arrays to ensure that metrics computed across many markers reflect properties associated with each sample, and will be less affected by noise associated with lower quality markers. Specifically, we selected 621,642 markers (605,876 autosomal; 15,766 on X and Y chromosomes), which fulfil the following criteria:

- In both of the two arrays.
- Is a SNP (not an indel).
- Passed QC in all 106 batches (see Section 4.2.1).
- MAF among all UK Biobank samples $>$ 0.0001.
- Is not in the list of SNPs affected by the 'image' artefact[6].

All analyses described in the rest of this chapter used these SNPs only, or a subset of these where stated. The X and Y chromosome markers were only used for the sex chromosome-specific analysis (Section 4.4.2).

Several of the analyses we conducted were dependent on each other. For example, adjusting the heterozygosity metric to account for population structure first requires computation of principal components. Figure 4.3 shows all the key interdependences within the pipeline, and the relevant sections in this chapter.

[6]This artefact involved a subset of SNPs and a very small number of samples ($\sim$300), so it only affects a very small proportion of all the data (details in Section 4.2.1). We excluded these SNPs in the sample-based QC and analysis as a precaution only.

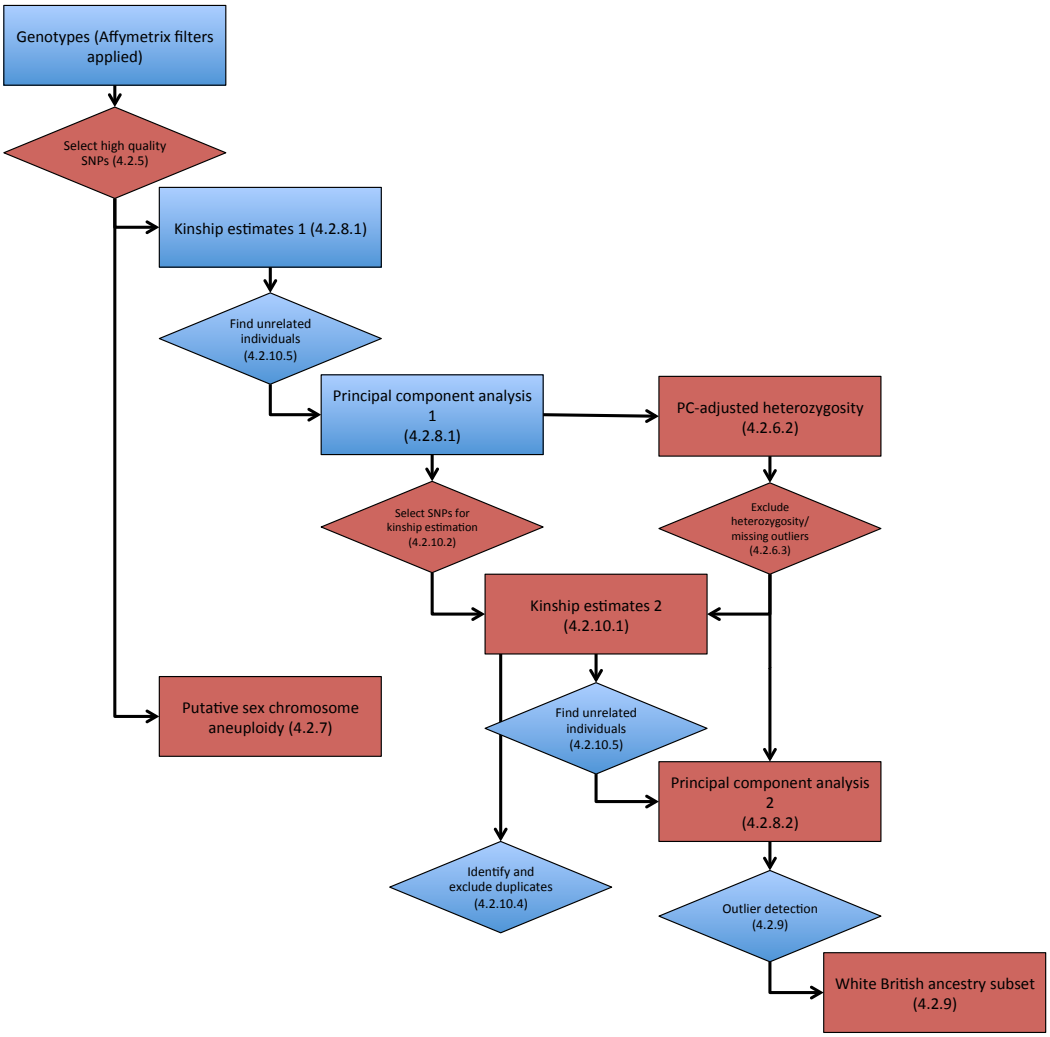


Figure 4.3: Overview of pipeline for sample-based analyses and quality control. Rectangles represent data or computed variables; diamonds indicate key processes that link the different analyses. For example, the computation of 'Principal component analysis 1' requires 'Kinship estimates 1' in order to exclude related individuals from the PC computation. Numbers denote the relevant section in this chapter. Data that is associated with the elements shown in red are made available to researchers.

4.4 Sample-based quality control

4.4.1 Detecting outliers in heterozygosity and missing rates

We identified poor quality samples using the metrics of missing rate and heterozygosity computed using the set of high quality autosomal markers (defined in Section 4.3). Extreme values in one or both of these metrics can be indicators of poor sample quality due to, for example, DNA contamination [118]. However, the

heterozygosity of a sample can also be sensitive to population structure, recent admixture and parental consanguinity. All these effects can be seen in the UK Biobank genotypes. Here we describe the approach we took to identify samples that have unusually high heterozygosity or missing rates while accounting for these effects.

4.4.1.1 Raw heterozygosity and missing rates

Using a set of 605,876 high quality autosomal SNPs (see Section 4.3) we computed raw heterozygosity (h) for each sample. That is, the proportion of non-missing genotypes that are heterozygous:

$$h = \frac{(N_{nm} - N_{hom})}{N_{nm}} \tag{4.1}$$

where N_{nm} is the number of non-missing genotypes, and N_{hom} the number of homozygous genotypes, both computed using the '--het' command in *plink*.

We computed missing rates using the '--miss' command in *plink*, and using the same set of SNPs as for heterozygosity.

4.4.1.2 Adjusting heterozygosity for population structure

The proportion of a sample's non-missing genotypes that are heterozygous (heterozygosity rate) is sensitive to population structure because allele frequency distributions (and thus expected heterozygosity) can differ between populations. This effect can be especially marked in array-based genotype data, where markers on the array are often chosen based on allele frequencies in a particular population (e.g. European). How this effects heterozygosity in the UK Biobank is clear in Figure 4.4a. We[7] controlled for this by fitting the following linear regression model.

Let h denote the heterozygosity and let x be a set of features correlated with ancestry. We used the projections onto the six major UK Biobank principal components (first

[7] This adjustment was applied by the author of this thesis, using *R* scripts originally developed by D.P. for the interim release data but which were adapted to include more PCs in the regression.

round, see Section 4.6.1.1) to characterise ancestry, writing $x = (x_1, x_2, x_3, x_4, x_5, x_6)$ for these six principal component values. Consider the following model for heterozygosity under population structure:

$$h(x) = h_0 + \beta(x) \tag{4.2}$$

where $h(x)$ is the raw heterozygosity, which depends on the features x, h_0 is the ancestry-adjusted heterozygosity, and $\beta(x)$ is a bias term due to population structure. The quadratic form for $\beta(x)$ includes all linear and quadratic terms x_i and x_i^2 as well as all cross terms $x_i x_j$. More specifically, the bias was assumed to have the following functional form:

$$\beta(x) = \sum_{i=1}^{6} \beta_i x_i \; + \sum_{i=1}^{6} \beta_{i^2} x_i^2 \; + \sum_{i,j=1, i \neq j}^{6} \beta_{ij} x_i x_j \; . \tag{4.3}$$

We estimated h_0 with ordinary least squares, and the fitted value $\hat{h}_0$ is the PC-corrected heterozygosity. We plot this on the y-axis in Figure 4.4b with all ethnic background categories combined, and in Figure A.9 with each ethnic background category separately. Both the PC-corrected and raw heterozygosity are provided to researchers.

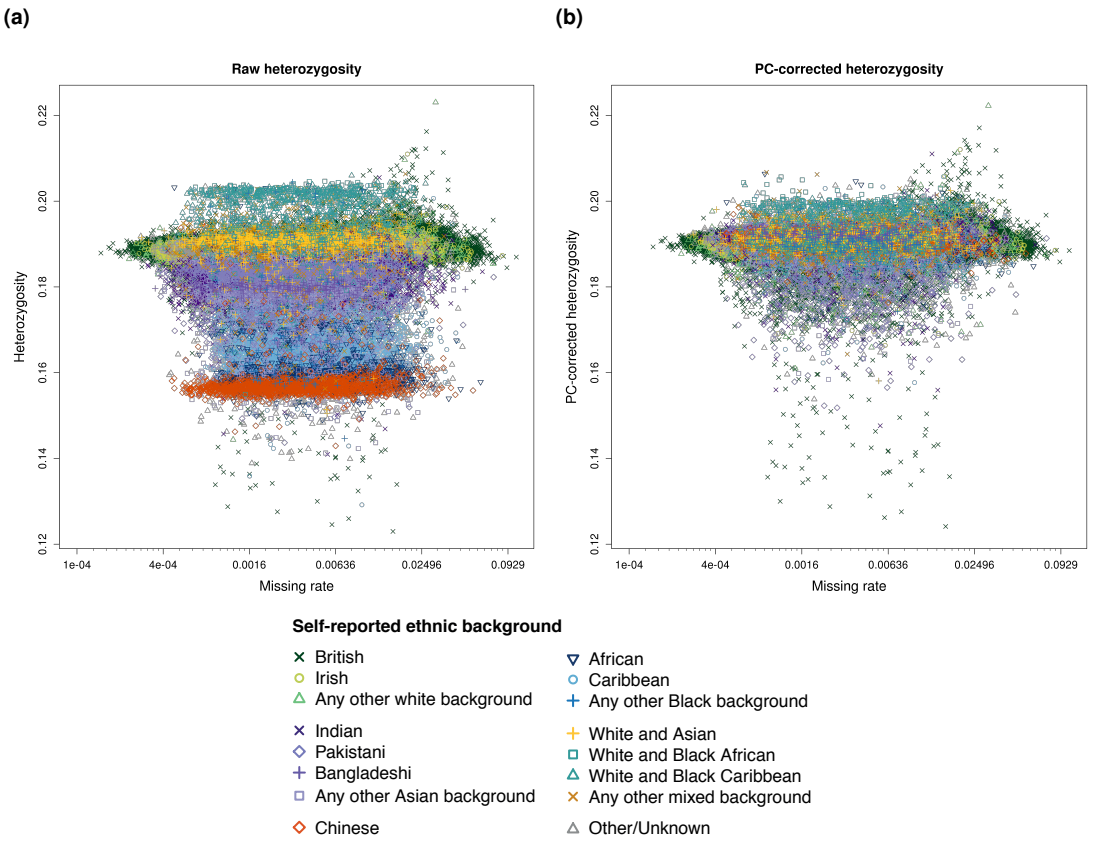


Figure 4.4: The effect of population structure on heterozygosity. Heterozygosity for each sample before **(a)** and after **(b)** correcting for ancestral background using PCs (details in Section 4.4.1.2). The symbols (shapes and colours) indicate the self-reported ethnic background of each participant, as annotated in the legend.

4.4.1.3 The effects of recent admixture and parental consanguinity on heterozygosity

Some samples can have extreme heterozygosity, even after accounting for population structure using PCs. Individuals whose parents are closely related (parental consanguinity) tend to have lower heterozygosity. This is because the expected heterozygosity of an individual is a function of population allele frequencies (assuming HWE) and the relatedness of their parents. Specifically, if f is the kinship coefficient of their parents (the probability of two alleles sampled at random from the two parents are identical by descent (IBD)), and p is the allele frequency at a bi-allelic locus, then the expected heterozygosity $E(h)$ of the child is:

$$E(h) = 2p(1-p)(1-f)$$

So, for example, a child of cousin parents will have $E(h) = 2p(1-p)(1-\frac{1}{16})$. Or in

other words, an expected heterozygosity a factor of $15/16$ smaller than the expected heterozygosity of an individual with unrelated parents in the same population (assuming that background IBD-sharing is negligible).

Conversely, individuals with 'mixed' ethnicity (i.e. recent admixture) tend to have higher heterozygosity, which is not captured by the PCs. This can be seen in Figure 4.5, especially within the categories 'White and Black African' and 'White and Black Caribbean'.

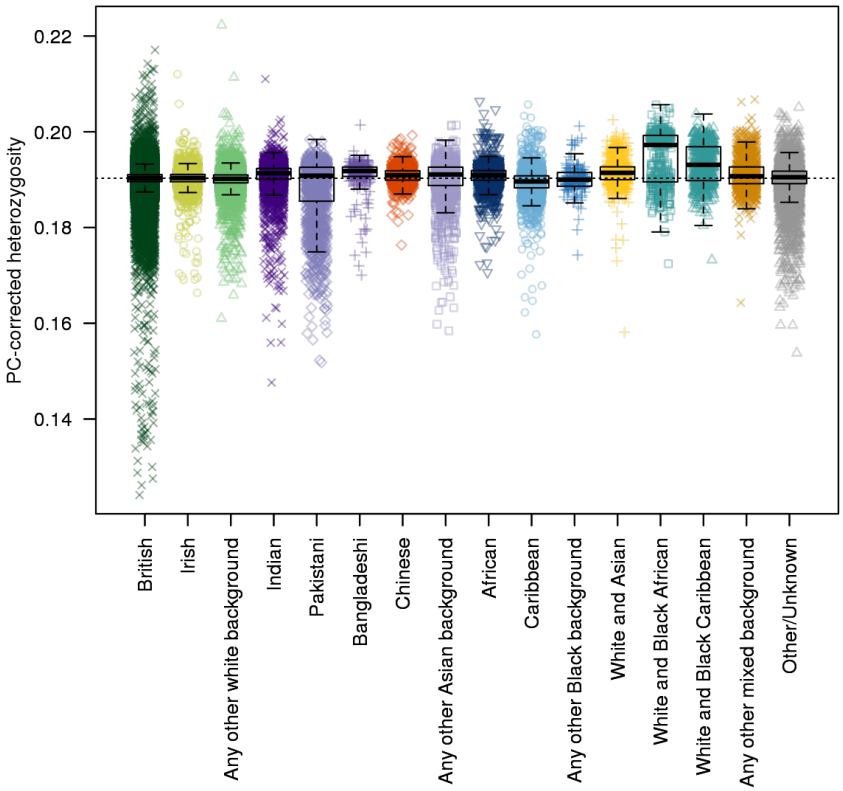


Figure 4.5: PC-corrected heterozygosity for each ethnic background category in UK Biobank. The dashed horizontal line shows the mean heterozygosity across all samples.

To see why this is the case, recall that the PC score for a sample is a linear sum of their genotypes. Therefore, for a given PC with eigenvector (i.e. 'SNP-loads'), the expected PC score is $\sum_{i}^{L} v_i E(X_i)$, where v_i is the SNP-load corresponding to marker i, and $E(X_i)$ is the expected genotype for marker i, treating the three genotypes as 0, 1, or 2. Now, if we consider individuals with ancestry from two different populations, then their expected genotype $E(X_i)$ depends only on the total fraction of genome they

inherited from each population. However, heterozygosity depends both on the fraction of genome and on how the ancestry is distributed across the chromosomal copies.

If we denote the expected genotype of an individual at marker i, conditional on the two chromosomes being inherited from populations j and k as $E(X_i|j,k)$, and denote the allele frequency for marker i within each ancestral population as p_k and p_j, then

$$E(X_i|1,1) = 2p_1$$
$$E(X_i|2,2) = 2p_2$$
$$E(X_i|1,2) = p1 + p2 \ .$$

These follow easily from the mean of the binomial distribution (when the ancestral population is the same on both chromosomes) and the poisson binomial distribution (when the ancestral populations are different). If f_{jk} is the fraction of loci where the chromosomes are inherited from populations j and k, then the total fraction α of the genome inherited from population 1 is $\frac{f_{12}}{2} + f_{11}$. Summing over all three combinations of ancestries we have:

$$\begin{aligned} E(X_i) &= f_{11}E(X_i|1,1) + f_{22}E(X_i|2,2) + f_{12}E(X_i|1,2) \\ &= [\alpha - \frac{f_{12}}{2}]E(X_i|1,1) + [1 - \alpha - \frac{f_{12}}{2}]E(X_i|2,2) + f_{12}E(X_i|1,2) \\ &= [\alpha - \frac{f_{12}}{2}]2p_1 + [1 - \alpha - \frac{f_{12}}{2}]2p_2 + f_{12}(p_1 + p_2) \\ &= \alpha 2p_1 + (1-\alpha)2p_2 \ . \end{aligned}$$

However, the expected heterozygosity $E(H_i)$ depends both on α and f_{12} (equivalently f_{11} or f_{22} since all f_{jk} must sum to 1). The expected heterozygosity conditional on the combination of ancestries is

$$E(H_i|1,1) = 2p_1(1-p_1)$$
$$E(H_i|2,2) = 2p_2(1-p_2)$$
$$E(H_i|1,2) = p_1(1-p_2) + p_2(1-p_1) \ .$$

Now consider $E_x(H_i)$ and $E_y(H_i)$ with the same α but different values, x and y, for f_{12}.

Then,

$$
\begin{aligned}
E_x(H_i) - E_y(H_i) &= (x-y)[E(H_i|1,2) - \frac{E(H_i|1,1)}{2} - \frac{E(H_i|2,2)}{2}] \\
&= [x-y][p_1(1-p_2) + p_2(1-p_1) - (p_1(1-p_1) + p_2(1-p_2)] \\
&= [x-y][(p_1-p_2)^2] \\
&= 0 \iff p_1 = p_2 \\
&\geq 0 \quad \text{if } x > y \\
&\leq 0 \quad \text{if } x < y \,. \qquad (4.4)
\end{aligned}
$$

This means, for example, that the heterozygosity of an individual whose parents are from two different populations (so $f_{12} = 1$) is larger, on average, than another individual with the same ancestry proportions (0.5) but with the ancestry inherited via both parents, such that $f_{11} > 0$ (and so $f_{12} < 1$). Since their expected PC-score will be the same, the PC-corrected heterozygosity will not account for this difference.

4.4.1.4 Details of detecting outliers in heterozygosity and missing rates

We aimed to flag samples whose extreme heterozygosity is not explained by ancestral background (i.e. marker ascertainment bias), recent mixed ancestry or parental consanguinity. We proceeded as follows, with missing rates and PC-corrected heterozygosity computed as described in Section 4.4.1.1.

We first considered individuals within the four largest ethnic background categories ('British', 'Any other white background', 'Irish', 'Indian'). To this combined set we applied a Bayesian outlier detection method, *aberrant* [154], to the two dimensions of logit-transformed missing rate and PC-adjusted heterozygosity. The method classifies samples into two groups, 'inliers' and 'outliers', which are modelled as being drawn from two different 2-dimensional Gaussian distributions. Typically, the 'outliers' have a larger variance and potentially a different mean. We used the logit transformation of missing rate because its distribution is approximately normal under this transformation. *aberrant* requires a user-defined parameter, Lambda, which tunes the

ratio of the variance of the 'outlier' and 'inlier' groups. We used Lambda=120 so that we only identified extreme values as outliers. In this way we identified 744 outliers and with PC-adjusted heterozygosity above the mean (0.1903). To avoid missclassifying samples with mixed ethnicity, we inspected plots of missing rate and PC-adjusted heterozygosity separately for each of the other ethnic background categories, looking for individuals with unusually high heterozygosity within their category (Figure A.9). This resulted in zero further outliers. We also flagged any sample with a missing rate > 0.05. In total we identified 968 samples with unusually high heterozygosity and/or missing rates. These samples are shown in red in Figure 4.6a.

As previously mentioned, low heterozygosity is expected as a consequence of relatedness between an individual's parents. This would also result in long runs of unbroken homozygous genotypes within the individual's genome. Closely-related individuals will share long contiguous regions of their chromosomes IBD, and so a child will be homozygous whenever they inherit these chunks from both parents. We used this observation to confirm that individuals with unusually low heterozygosity were not subject to poor quality genotyping by checking the expected (negative proportional) relationship between heterozygosity and long runs of homozygosity (LROH). We used *plink* to detect LROH, using the '--homozyg-kb' command with a homozygous run required to span at least 1000 Kb. The negative relationship between the heterozygosity and the total length of all runs of homozygosity is clear in Figure 4.6b.

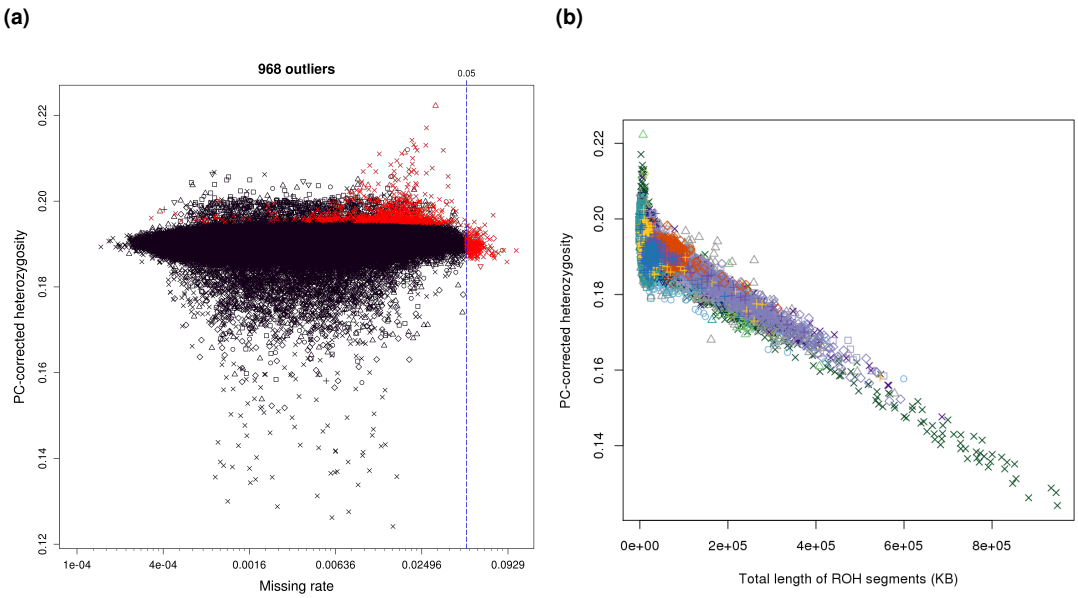


Figure 4.6: Detecting outliers in heterozygosity and missing rate. (a) The set of samples we flagged as outliers (in red), and all other samples (in black), with shapes the same as in (a) and (b). The vertical line shows the threshold we used to call samples as outliers on missing rate. In plots (a)-(c) missing rate data is transformed to the logit scale, but with the axis annotated with the original values. **(b)** Observed relationship between long runs of homozygosity and heterozygosity. The symbols (shapes and colours) show self-reported ethnic background as annotated in the legend.

4.4.2 Putative sex chromosome aneuploidy

We conducted quality control specific to the sex chromosomes using a set of 15,766 high quality markers on the X and Y chromosomes. Affymetrix infers the sex of each individual based on the relative intensity of markers on the Y and X chromosomes [114]. Sex is also reported by participants, and mismatches between these sources ('sex mismatch') can be used as a way to detect sample mishandling or other kinds of clerical error. However, in a data set of this size, some such mismatches would be expected due to transgender individuals, or instances of real (but rare) genetic variation, such as sex-chromosome aneuploidies [155]. Affymetrix genotype calling on the X and Y chromosomes allows only haploid or diploid genotype calls, depending on the inferred sex [114]. Therefore, cases of full or mosaic sex chromosome aneuploidies may result in compromised genotype calls on all, or parts of, the sex chromosomes (but not affect the autosomes). For example, individuals with karyotype XXY will likely have poorer quality genotype calls on the pseudo-autosomal region

(PAR) of the X chromosome, as they are effectively triploid in this region.

We used information about the relative intensities of chromosomes X and Y to identify individuals with sex chromosome karyotypes putatively different from XY or XX. Specifically, 'Log 2 ratios' (L2R) are computed (by Affymetrix) for each sample at each marker, and are one of the primary measures used in copy number detection algorithms [156]. For a given marker and sample, the L2R is the sum of the A and B allele intensities, normalized by the median intensity of that marker in individuals assumed to represent the normal copy number state. To find the median intensity, only individuals in the same genotyping batch were used; and for markers on the X and Y chromosomes only (inferred) females and males, respectively, were used [156].

We first extracted the L2R for a set of high quality SNPs on the X and Y chromosomes (see Section 4.3), then for each individual we computed the mean L2R across each of the sex-specific region of the X chromosome (*L2Rx*) and the Y chromosome (*L2Ry*). Individuals with full chromosome aneuploidies will be expected to have intensity values that are, on average, larger (or smaller) than the most common chromosome copy number state. L2R values tend not to relate directly to the relative copy number because of background intensity associated with the assay. For example, the mean *L2Rx* for males is around -0.44, i.e. an intensity of 0.7 relative to females with copy number 2, rather than 0.5. After visual examination of a scatter plot of these metrics (Figure 4.7) we defined a set of criteria with which to classify the likely karyotype of each individual, as noted in Table 4.3.

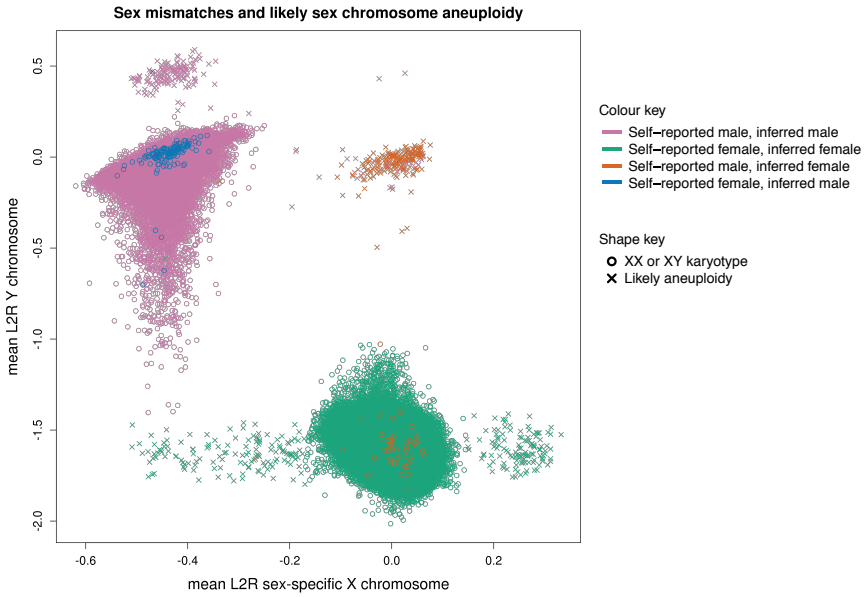


Figure 4.7: Putative sex chromosome aneuploidy. Mean Log2 ratios (L2R) on X and Y chromosomes for each sample, indicating likely sex chromosome aneuploidy. Samples which are most likely XX or XY are depicted with a circle. Other samples, which represent possible instances of sex chromosome aneuploidy, are indicated by a cross. There are 652 such samples in total; see Table 4.3 for counts. The colours of each symbol relate to different combinations of self-reported sex, and sex inferred by Affymetrix (from the genetic data), as indicated by the key. For almost all samples (99.9%) the self-reported and inferred sex are the same, but for a small number of samples (378) they do not match (see Section 4.4.2 for discussion).

Criteria	Putative sex chromosome karyotype	Sex match	Sex mismatch	Total	Rate per 10,000
Inferred female and $L2Rx < -0.17$	X0 (complete, or mosaic)	148	2	150	3.1
Inferred female and $L2Rx > 0.145$	XXX	123	0	123	2.5
$-1 \geq L2Ry < 0.23$ and $L2Rx > -0.2$	XXY	47	178	225	4.6
L2Ry ≥ 0.23	XXYY or YY	153	1	154	3.2
Not any of the above	XX or XY	487528	197	487725	9987
	Total	487999	378	488377	10000

Table 4.3: Criteria used for identifying instances of putative sex-chromosome aneuploidy and associated counts of samples in UK Biobank. 'Sex mismatches' indicate UK Biobank participants for whom their self-reported sex does not match sex as inferred from the genetic data (by Affymetrix). Figure 4.7 shows the distribution of *L2Rx* and *L2Ry* used to determine the criteria for different karyotypes.

Using the above criteria we identified a set of 652 (0.134%) individuals with sex chromosome karyotypes putatively different from XY or XX (Table 4.3). The 225 samples with a possible XXY karyotype are heavily enriched for sex mismatches

(79%). For all but one of the XXY karyotype cases the submitted sex was male, which would be consistent with the occurrence of a Y chromosome, which contains the sex-determining region. There is also a strikingly long tail in Y-chromosome intensities (as measured by mean L2R on the Y chromosome) in males. This is likely to be an indicator of mosaic loss of the Y chromosome, which is thought to be a somatic mutation related to age, as well as smoking [157]. Indeed, L2R*y* is significantly associated with age in UK Biobank, as can be seen in Figure 4.8 (smoking phenotypes were not made available to us for QC purposes).

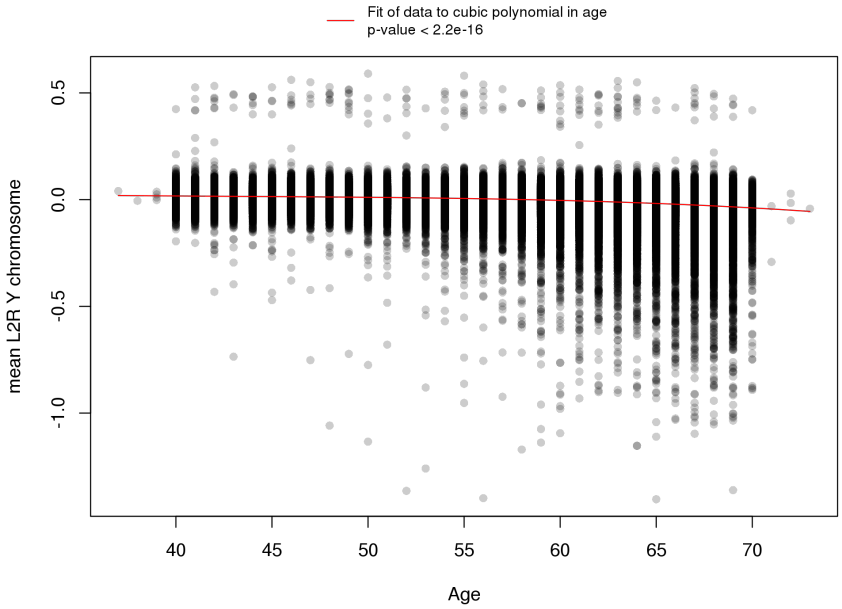


Figure 4.8: Y chromosome intensity (mean L2R) associated with age in UK Biobank. The red line shows the fitted values for the regression model L2Ry $\sim$ age + age^2 + age^3 (p-value $<$ 2.2x10^{-16}).

4.5 Summary of UK Biobank genotype data quality

4.5.1 Overall missing data and duplicate concordance rates

The application of our QC pipeline resulted in the released data set of 488,377 samples and 805,426 markers from both arrays. The proportion of all the genotype calls made by Affymetrix that were set to missing as a result of quality control is 0.0097 (Table 4.2). The proportion of genotyped samples that we identified as poor

quality is 0.002 (968/488377). We did not remove these samples from the data, but rather provided the information as part of the data release. However, we did exclude a small number of samples (835 in total) that we identified as sample duplicates, as opposed to identical twins (see Section 4.7.4) or were likely involved in sample mishandling in the laboratory (∼10), as well as participants who asked to be withdrawn from the project prior to the data release in July 2017 (33). The overall missing rates for samples and marker in the released data are shown in Figure 4.12b, which also illustrate the peculiar effect of two slightly different arrays.

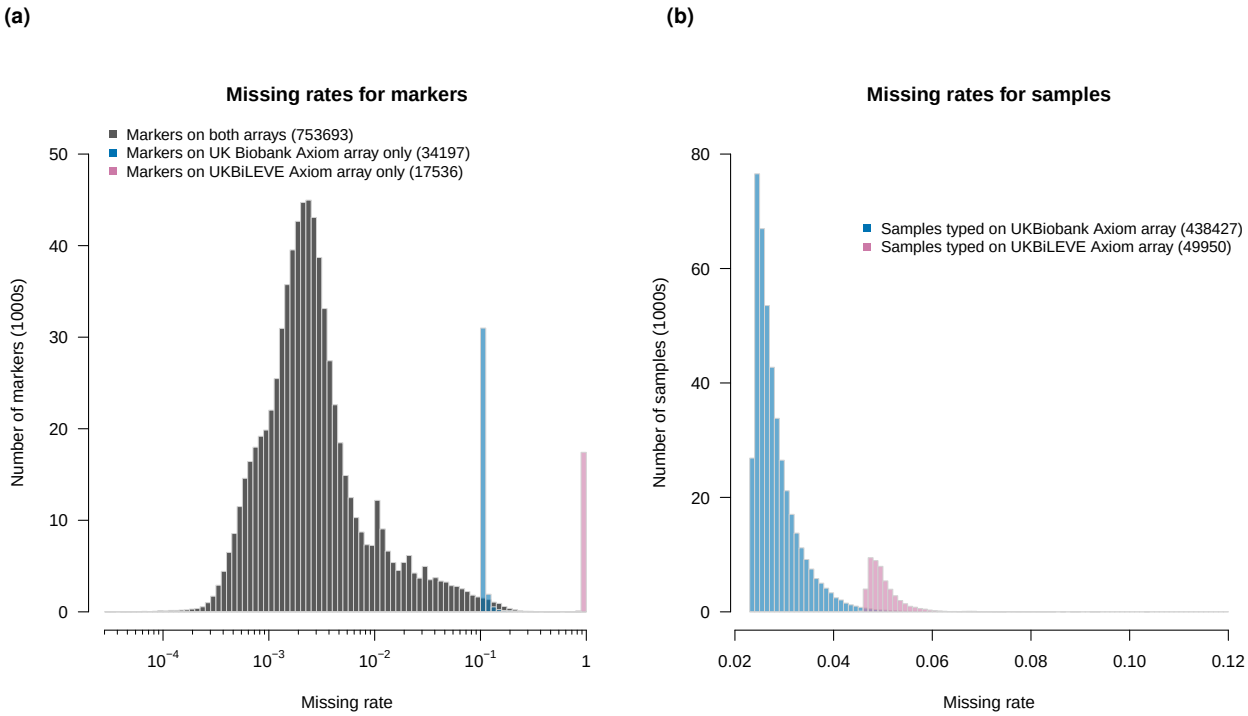


Figure 4.9: Missing rates for markers and samples after applying QC. (a) The distribution of missing rates for markers. Three histograms are overlaid, each showing a different, mutually exclusive, subset of markers, which are indicated by the three colours. Markers from only one array exhibit more missing data because only a subset of samples were typed on each array (10% on UK BiLEVE Axiom array and 90% on the UK Biobank Axiom array). **(b)** The distribution of missing rates for samples. Two histograms are overlaid, each showing a mutually exclusive subset of samples. All samples have some missing data because not all markers were included on both arrays (∼2% are exclusive to the UK BiLEVE Axiom array and ∼4% exclusive to the UK Biobank Axiom array). Additional missing data is also introduced from the batch-based marker QC.

A set of 588 pairs of experimental duplicates ('Blind Spike Duplicates') were genotyped as if they were DNA samples from different individuals (see Section 4.7.4). Concordance rates in the genotype calls for these samples can therefore provide a

natural measure of experimental error. We calculated genotype call concordance rates for all the pairs by computing the fraction of markers with the same genotype call in both samples, excluding any markers missing in one or both samples. On average, 99.87% of a pair's genotype calls are identical and the lowest rate is still 99.39% (Figure 4.10).

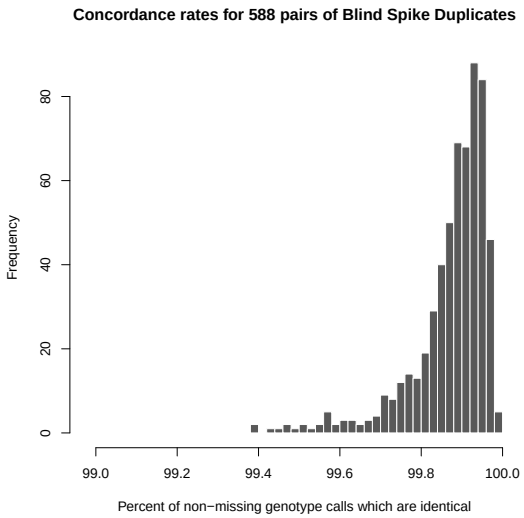


Figure 4.10: Concordance rates for Blind Spike Duplicates. For each Blind Spike Duplicate pair (two genotyped samples from the same individual) we calculated the fraction of markers with the same genotype call in both samples, excluding any markers missing in one or both samples.

4.5.2 Comparison of UK Biobank with ExAC

We compared allele frequencies in the UK Biobank with those estimated from sequencing data sourced from the Exome Aggregation Consortium (ExAC) database[8] [158]. For the ExAC data we used allele counts for the non-Finnish European population group (33,370 samples). For the UK Biobank we used the set of 463,844 European-ancestry samples defined using 1000 Genomes PCs (see Section 4.2.1), and who did not report Finland as their birthplace (there are $\sim$160 Finland-born participants in the UK Biobank cohort). We compared a set of 91,298 markers, which are those in the released genotype dataset, on both genotyping arrays, and have more than 90% call rate in both UK Biobank and ExAC. We merged data for the two studies by requiring markers to match on chromosome, position, and the reference

[8] We downloaded VCF files from ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/ExAC.r1.sites.vep.vcf* .

and alternative alleles (all markers in the UK Biobank released data are bi-allelic). We report the frequency of the allele that is minor in the UK Biobank, so some markers could have allele frequency > 0.5 in ExAC. The results are shown in Figure 4.11.

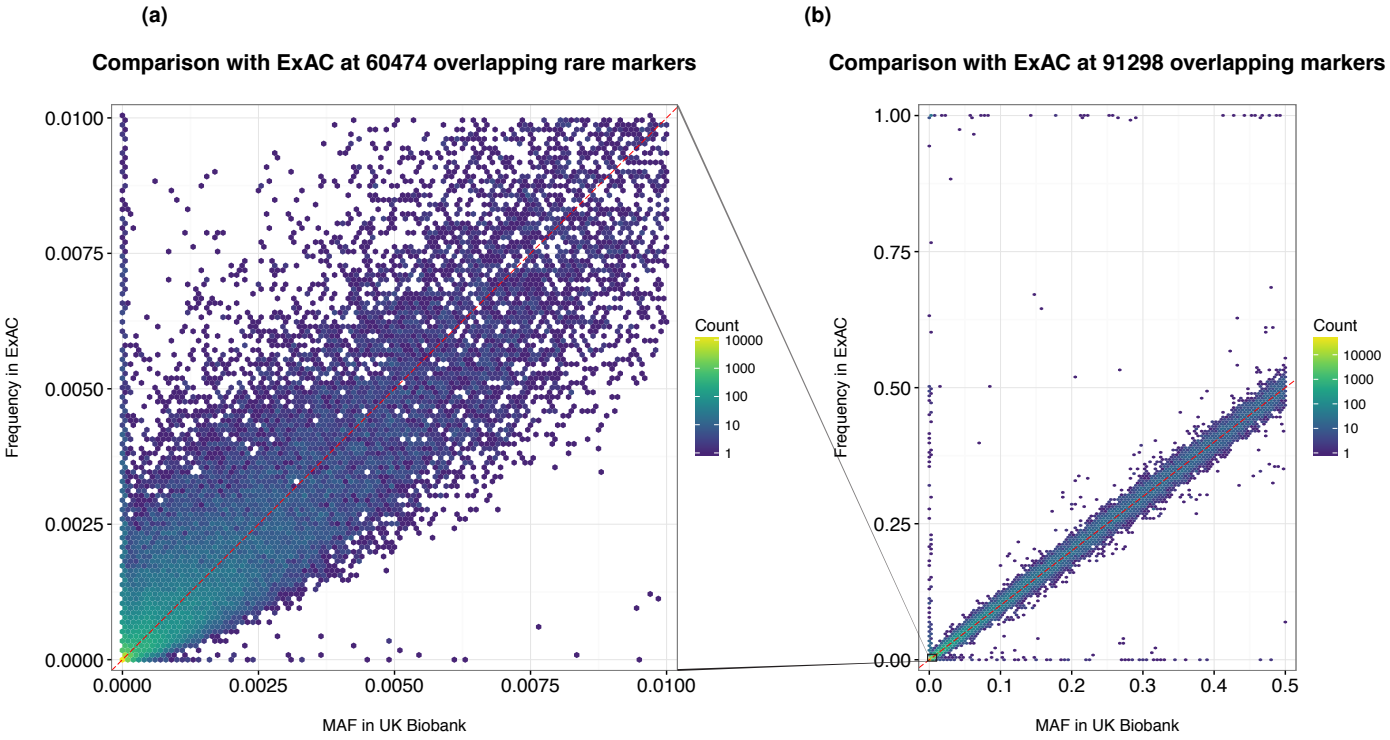


Figure 4.11: Comparison of allele frequencies between UK Biobank and ExAC. We used 33,370 European ancestry samples from ExAC and 463,844 European ancestry samples from UK Biobank, and only markers that have a call rate greater than 0.9 in both studies. In both plots each hexagonal bin is coloured according to the number of markers falling in that bin, as indicated by the keys (note the $\log_{10}$ scale). The dashed red lines show x=y. **(a)** Comparison of rare markers in the subset (both axes < 0.01). That is, only showing markers in the bottom-left corner of (b). **(b)** Comparison of all the markers in the subset. The set of markers with very different allele frequencies seen on the top, bottom, and left-hand sides of the plot comprise $\sim$300 markers. This is $\sim$0.3% of all markers in the comparison, or $\sim$0.5% of all markers with MAF > 0.001 in at least one study.

We do not expect allele frequencies in the two studies to match exactly due to subtle differences in the ancestral backgrounds of the individuals in each study, as well as differences in the sensitivity and specificity of the two technologies (exome sequencing and genotyping arrays). Despite this, overall the allele frequencies are encouragingly similar ($r^2 = 0.93$). There are also a small number of $\sim$300 markers ($\sim$0.3%) that have very different allele frequencies. Namely, 179 markers for which the frequency is $>$ 0.001 in ExAC (usually corresponding to more than 60 copies of the minor allele) but zero in the UK Biobank; and 35 markers where the reverse is the case. A further 73

markers have frequency > 0.75 in ExAC, indicating a mis-annotation of the alternative allele in either the UK Biobank arrays or ExAC. From visual inspection of intensity plots (such as those shown in Figure 1.12) for a subset of these markers we concluded the following. In the cases where MAF is zero in UK Biobank there is no evidence of a heterozygous cluster, possibly because the probes for one or both of the alleles are not working. In the cases where the frequency is zero (or close to 1) in ExAC, most appear to be genotyped well in UK Biobank. However, many of these markers are multi-allelic (in ExAC) or indels, which would be consistent with either annotation error on the UK Biobank arrays or in ExAC, or mapping errors in the sequence data in regions of more complex variation.

4.5.3 Performance of rare markers

Over 110,000 rare markers (MAF < 0.01) were included on the two arrays used for the UK Biobank cohort [99] (see also Figure 4.12a). Variants occurring at very low frequencies present a particular challenge for genotype calling using array technology, especially in cases where only a small number of samples within a genotyping batch are expected to have a copy of the minor allele (almost always with a heterozygous genotype). In these cases it is sometimes difficult in the genotype intensity data to distinguish a sample that genuinely has the minor allele, from one whose intensities are in the tails of the distribution of those in the major homozygote cluster. Examples of two different markers with MAF < 0.001 in UK Biobank are shown in Figure 4.13B and Figure 4.13C. One marker is performing well (4.13B), but for the other marker (4.13C) the heterozygous samples are more difficult to identify. In contrast, Figure 4.13A shows a common SNP (MAF=0.077), which has three well-separated clusters corresponding to the three genotypes. In the UK Biobank, a larger fraction of rare markers fail quality control tests compared to low frequency and common markers, but 84% still pass in all batches (Figure 4.12b). The MAF of rare markers are generally similar to those in ExAC, but very rare markers tend to have a lower MAF in the UK Biobank (Figure 4.11a).

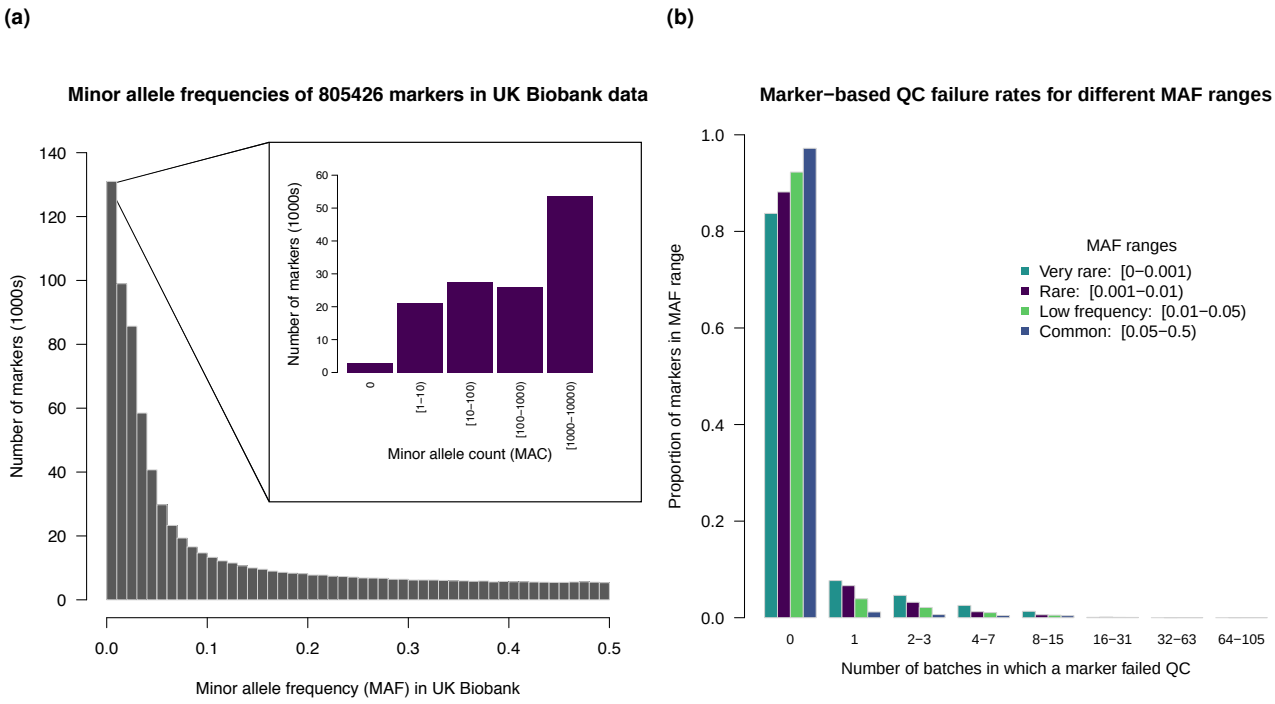


Figure 4.12: Minor allele frequency distribution and QC test failure rates by MAF. (a) Minor allele frequency (MAF) distribution. All samples were used in the calculation of MAF. The inset shows the number of markers within different minor allele count bins for rare markers only (MAF < 0.01). **(b)** The distribution of the number of batch-level QC tests that a marker fails (see Table 4.2; Section 4.2.1). For each of four different MAF ranges (indicated by colours) we show the fraction of markers that fail the specified number of batches. For example, just over 92% of the 'low frequency' markers ($0.01 \leq$ MAF < 0.05) do not fail quality control in any batches. Any marker that failed all 106 batches is excluded from the data release, so such markers are not included here (see Table 4.2).

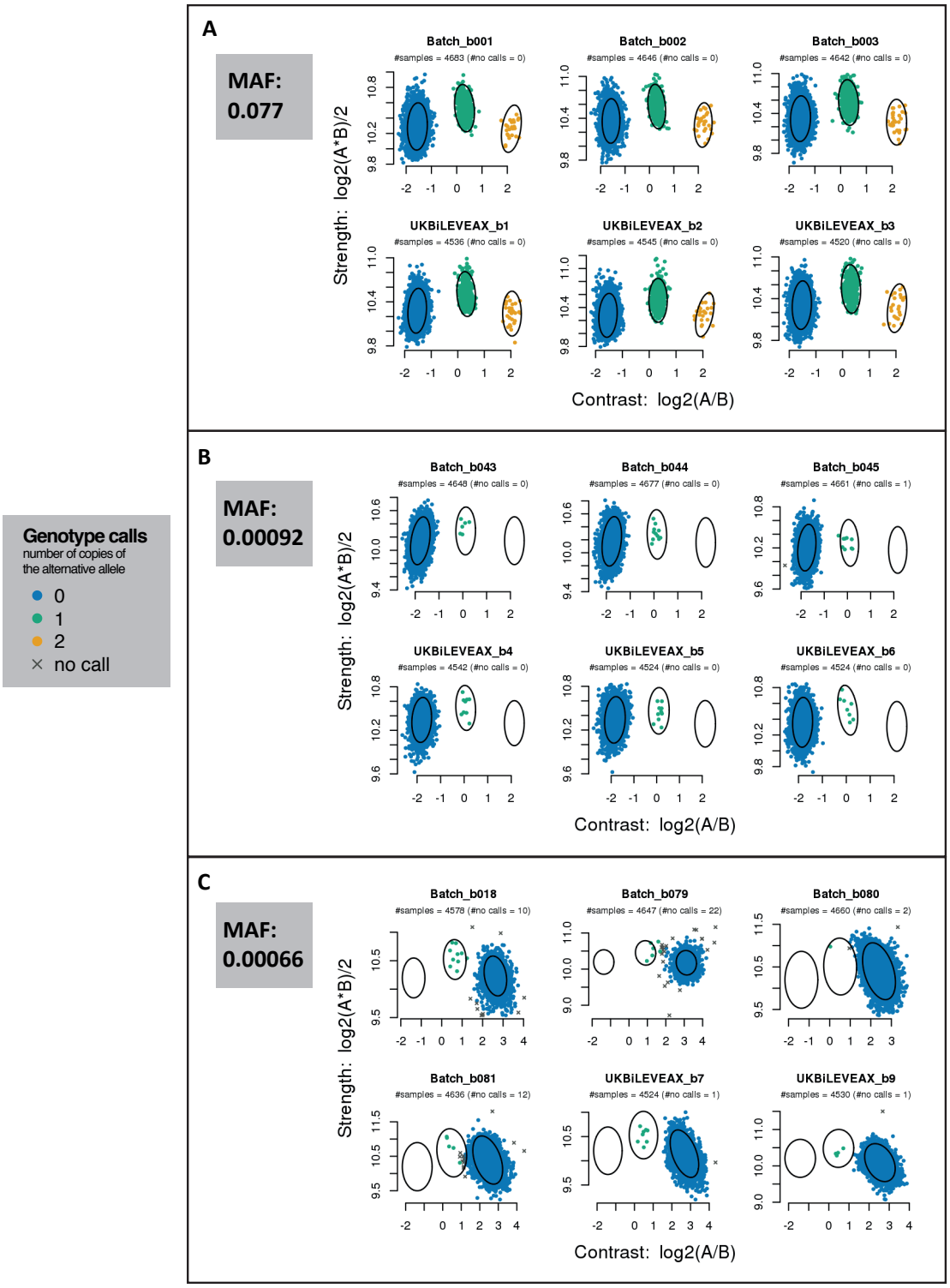


Figure 4.13: Examples of intensity data and genotype calls for markers of different allele frequencies. Each of the three sub-figures shows intensity data for a single marker within six different batches. Batches labelled with the prefix 'UKBiLEVEAX' contain only samples typed using the UK BiLEVE Axiom array, and those with the prefix 'Batch' contain only samples typed using the UK Biobank Axiom array. Each point represents one sample and is coloured according to the inferred genotype at the marker. The x and y axes are transformations of the intensities for probe sets targeting each of the alleles 'A' and 'B' (see Section 1.3.3.2 for definition of probe set). The ellipses indicate the location and shape of the posterior probability distribution (2-dimensional multivariate Normal) of the transformed intensities for the three genotypes in the stated batch. That is, each ellipse is drawn such that it contains 85% of the probability density. See [114] for more details of Affymetrix genotype calling. The minor allele frequency of each of the markers is computed using all samples in the released UK Biobank genotype data. **(A)** A marker with a MAF of 0.077 with well-separated genotype clusters. **(B)** Intensities for a marker with a MAF of 0.00092 with well-separated genotype clusters. As would be expected under HWE, there are no instances of samples with the minor homozygote genotype. **(C)** Intensities for a marker with a MAF of 0.00066, and where the heterozygote cluster is not well separated from the large major homozygote cluster in some batches, making it more difficult to confidently call the heterozygous genotypes.

4.6 Population structure among UK Biobank participants

4.6.1 Detecting population structure using PCA

We computed principal components (PCs) using an algorithm (*fastPCA* [159]), which performs well on datasets with hundreds of thousands of samples by approximating only the top K PCs that explain the most variation, where K is specified in advance. For principal components to reflect population structure (as opposed to technical artefacts or recent relatedness), they should ideally be computed using a subset of high quality, unrelated samples [38]. However, the metrics used to find related samples and poorer quality samples themselves require information about population structure. We therefore conducted an initial round of PCA, computing just the top 8 PCs (although we only ended up using the first 6), using a set of unrelated samples based on an initial round of kinship estimation. We used the results of this analysis to compute PC-adjusted heterozygosity (Section 4.4.1.2) as well as refine the relatedness inference (Section 4.7.1). Having then identified a set of high quality, unrelated samples, we conducted a second round of PCA, computing the first 40 PCs. Results of the second round are made available to researchers and visualised in figures showing PCA results (e.g. Figure 4.14) and discussed in the results section of this chapter.

4.6.1.1 Details of PCA round 1

We first estimated kinship coefficients between all samples using the software *KING* (v1) [119], with the command '--related -degree 3', and used these results to find a set of unrelated individuals. Note this is separate from the final relatedness inference described in Section 4.7.1. Next we excluded samples with the following properties:

- Missing rate on autosomes $>$ 0.02.
- Not in a set of unrelated individuals (see Section 4.7.5).
- Mismatch between inferred sex and self-reported sex.

We also excluded SNPs with the following properties:

- Missing rate $>$ 0.015.
- MAF $<$ 0.01.
- In regions of long-range linkage disequilibrium (LD) e.g. inversions. The boundaries we used are in Table A.1.

We then pruned the remaining 535,045 SNPs to a set of independent markers such that pairwise $r^2 < 0.1$. We used *plink*'s '--indep-pairwise' function with windows of 1000 Kb and a step-size of 80 markers.

We applied these filters to the genotype data using the appropriate commands in *plink*, in the order described. This resulted in a set of 147,551 SNPs and 406,247 samples with which to compute PCs. We computed the top 8 PCs using *fastPCA* with options 'numoutevec' = 8, and otherwise used the program defaults. We computed SNP-loads for each PC by carrying out the appropriate matrix multiplications based on mean-centred and variance-scaled genotypes, and the PC scores computed by *fastPCA*. Finally, we projected all samples onto the PCs using the SNP-loads. For the SNP-loads and projection steps we used scripts developed by D.P.

4.6.1.2 Details of PCA for release

We filtered the genotype data using the same criteria as above, but with additional sample exclusion criteria based on a second round of familial relatedness inference using a specially filtered set of SNPs (see Section 4.7.1), and having identified a small number of lower quality samples (see Section 4.4.1). Specifically, we excluded samples with the following properties:

- Missing rate on autosomes $>$ 0.02.
- Not in a set of unrelated individuals (see Sections 4.7.1 and 4.7.5).
- In the list of outliers based on heterozygosity and missing rates (see Section 4.4.1).

- Mismatch between inferred sex and self-reported sex.

We applied these filters to the genotype data using the appropriate commands in *plink*, in the order described. This resulted in a set of 147,606 SNPs and 407,599 samples with which to compute PCs. We then computed PCs for all samples exactly as described for PCA round 1, except when running *fastPCA* we computed the top 40 PCs with options 'numoutevec' = 40; 'fastdim' = 50; 'fastiter' = 40.

4.6.2 Broad-scale population structure

Figure 4.14 shows results for the first 6 PCs plotted in consecutive pairs, and Figure A.10 shows all pairs of the first 6 PCs. The eigenvalues (Figure A.12) indicate that these PCs capture 93% of the variance explained by the first 40 PCs combined. Results for further PCs are shown in Figure 4.15. As expected, individuals with similar PC scores have similar self-reported ethnic backgrounds. For example, the first two PCs separate out individuals with sub-Saharan African ancestry, European, and east-Asian ancestry. Individuals who self-report as mixed ethnicity tend to fall on a continuum between their constituent groups (e.g. the White and Asian category). Figure 4.15 shows the relationship between each of the 40 PCs, and country of birth as reported by participants. The higher PCs capture population structure at sub-continental geographic scales. For example, high scores in PC 6 are associated with individuals born in Central and South American countries such as Peru, Colombia, and Chile.

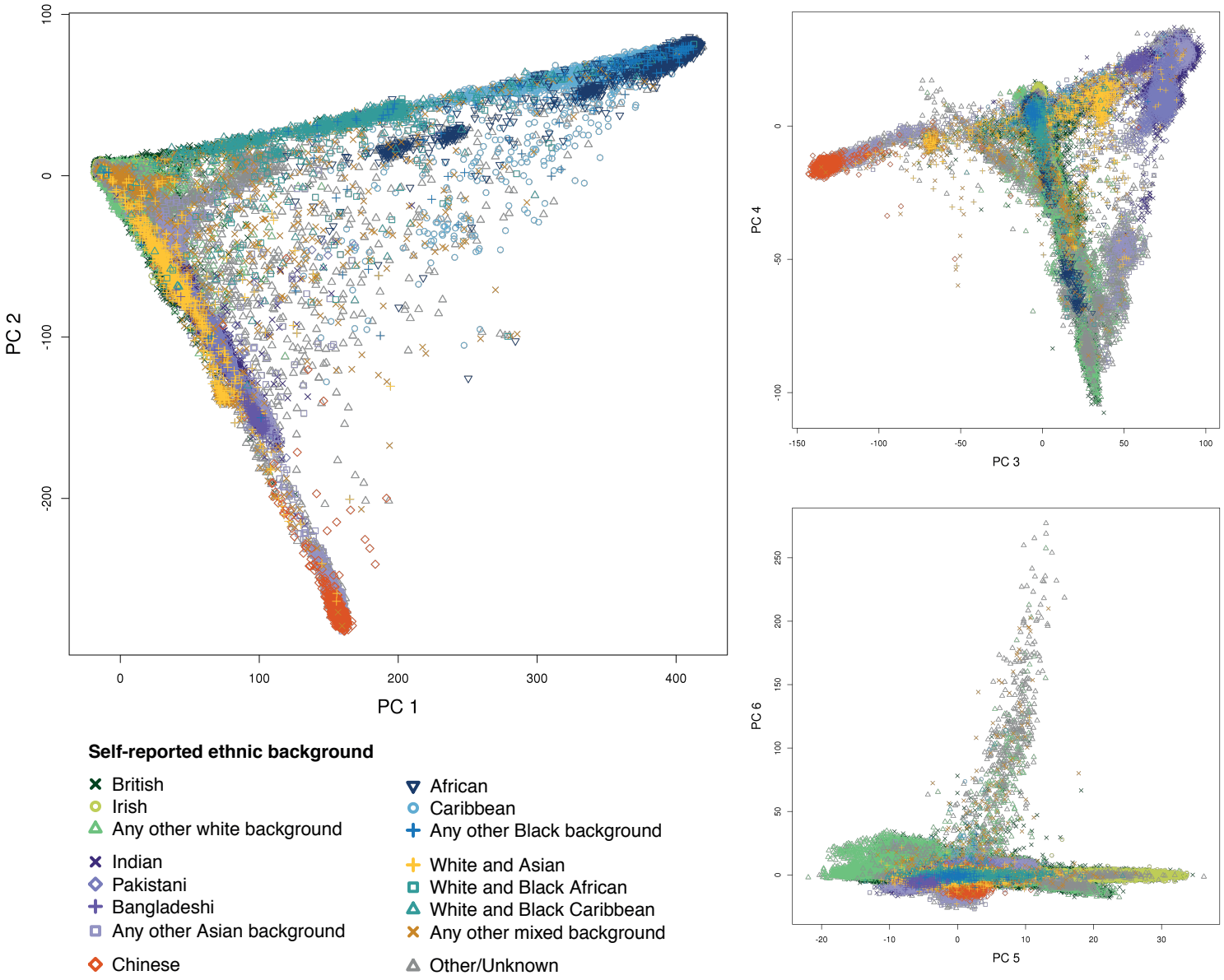


Figure 4.14: Ancestral diversity in the UK Biobank cohort. Plots of consecutive pairs of the first six principal components in a PCA of genotype data for UK Biobank participants (see [ref. chapter section]). Each point represents an individual and is placed according to their principal component scores (using genetic data only), with shapes and colours indicating their self-reported ethnic background as shown in the legend. See Table 4.1 for the proportions of participants in each category. Plots of all pairs of these 6 PCs are shown in Figure A.10, and results of further PCs are represented in Figures 4.15.

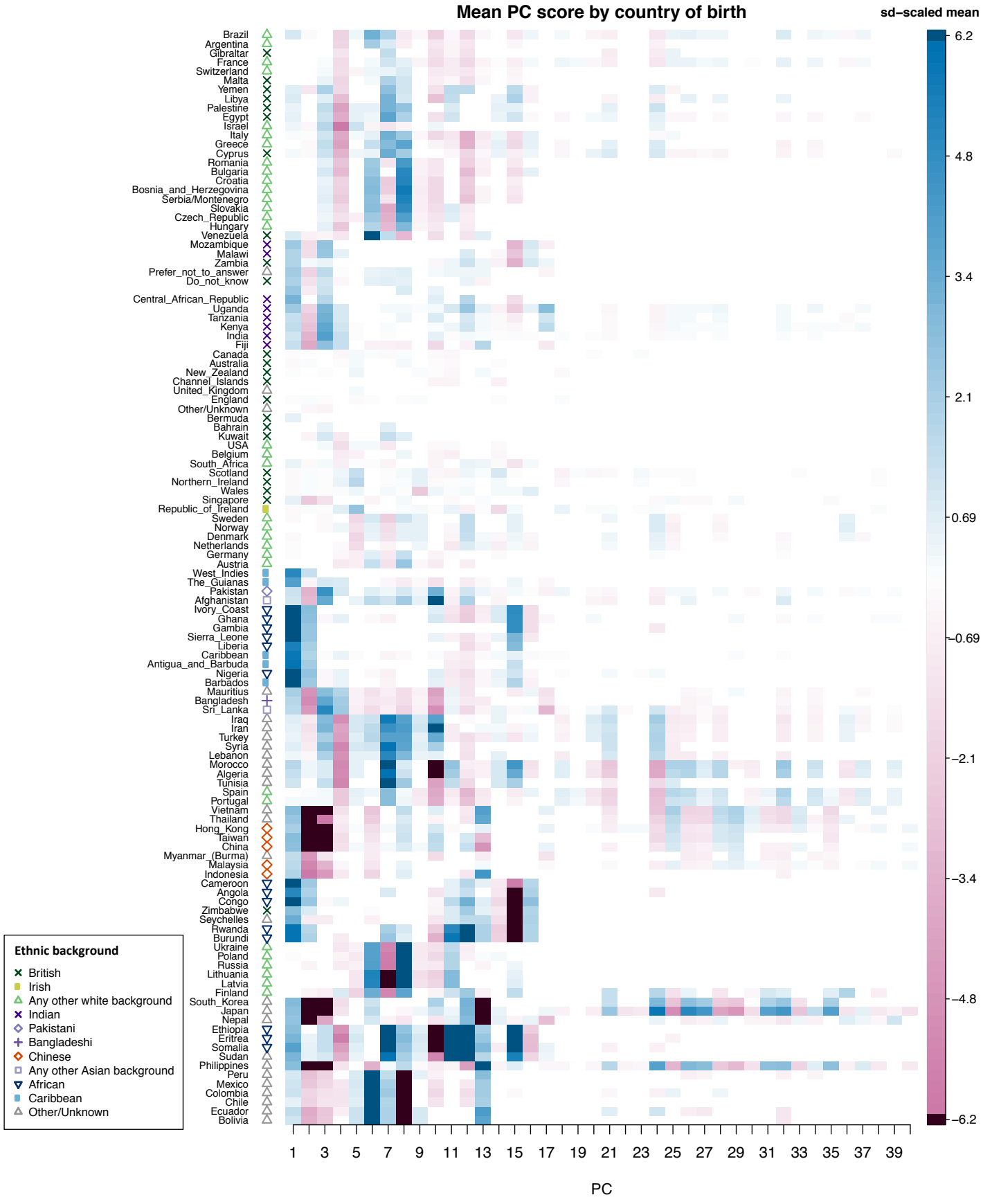


Figure 4.15: Mean principal component scores for each self-reported country of birth. Each column shows one PC and each element is the mean PC score for individuals born the labelled country, scaled by the standard deviation of the scores for that PC. Elements in each column are only coloured if the country has a non-zero coefficient ($p < 10^{-5}$) in a linear model with country of birth as predictor and PC scores as outcome. Countries (rows) have been ordered using hierarchical clustering ('hclust' function in *R*). The symbols next to each country label indicate the most common ethnic background category among the participants born in that country. For example, the most common self-reported ethnic background of participants born in Sri Lanka is 'Any other Asian background'.

4.6.3 Defining a 'white British ancestry' subset

Researchers may want to only analyse a set of individuals with relatively homogeneous ancestry to reduce the risk of confounding due to differences in ancestral background. Although the UK Biobank cohort is ethnically diverse, such analysis is feasible without compromising too much in sample size because a majority (88.26%) of participants in the UK Biobank cohort report their ethnic background as 'British', within the broader-level group 'White'. The PCA revealed population structure even within this category (Figure 4.16), so we used a combination of self-reported ethnic background and genetic information to identify a subset of individuals who self-report as 'British' as well as having very similar ancestral backgrounds based on results of the PCA.

We first selected 431,059 (88.26%) individuals who report their ethnic background as 'British', within the broader-level group 'White' (see Table 4.1). We used the outlier detection algorithm, *aberrant* [154], to isolate the largest cluster of samples from the rest, using PCs 1 – 6 (see Section 4.4.1.4 for a description of *aberrant*). We set the parameter Lambda to 40, which we chose to balance the number of samples excluded with their closeness in PC space. *aberrant* works only in two dimensions, so we applied it separately to pairs of PCs: 1&2; 3&4; 5&6 to find three sets of tightly clustered samples. We then took the intersection of all three sets, and defined this set of 409,728 individuals as the 'white British ancestry subset' (see Figure 4.16).

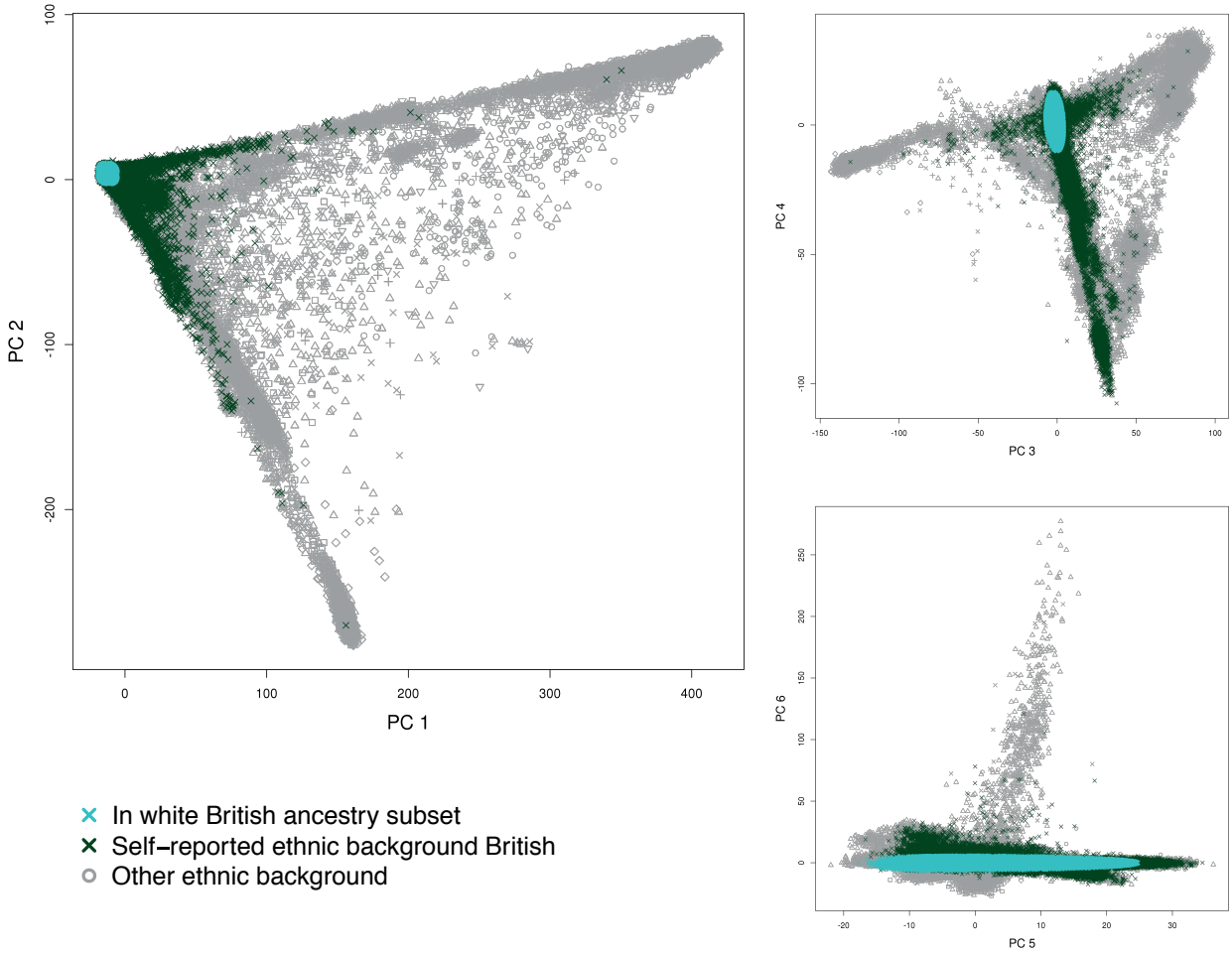


Figure 4.16: Selection of white British ancestry subset using PCA. Each plot shows the principal component scores for all UK Biobank samples, which we used to select the white British ancestry subset (see [ref. wb section]). Non-grey points indicate participants who have self-reported ethnic background 'British' (within the broader-level group 'White', see Table 4.1), and participants with other ethnic backgrounds are coloured grey, but with the same set of symbols as shown in Figure 4.14. Blue crosses show participants within the white British ancestry subset.

4.6.4 Population structure among participants of white British ancestry

Fine-scale population structure is known to exist within the UK [42] but methods for detecting such subtle structure [41] available at the time of analysis are not feasible to apply at the scale of the UK Biobank. The white British ancestry subset may therefore still contain subtle structure present at finer scales. To test this, we computed 40 PCs using a set of 340,040 unrelated samples within this subset, and then projected the rest of the samples in the white British ancestry subset onto these PCs (otherwise we proceeded exactly as described in Section 4.6.1.2). The first 6 PCs plotted by country

of birth in Figure 4.17, indicate that a) population structure exists within this subset, and b) that this reflects, to some degree, geographic regions (e.g. Wales). For a subset of these individuals, geographic coordinates (eastings and northings) of their birth place is available in the UK Biobank phenotype data. This provides an opportunity to examine the relationship between population structure and geography in the UK in more detail.

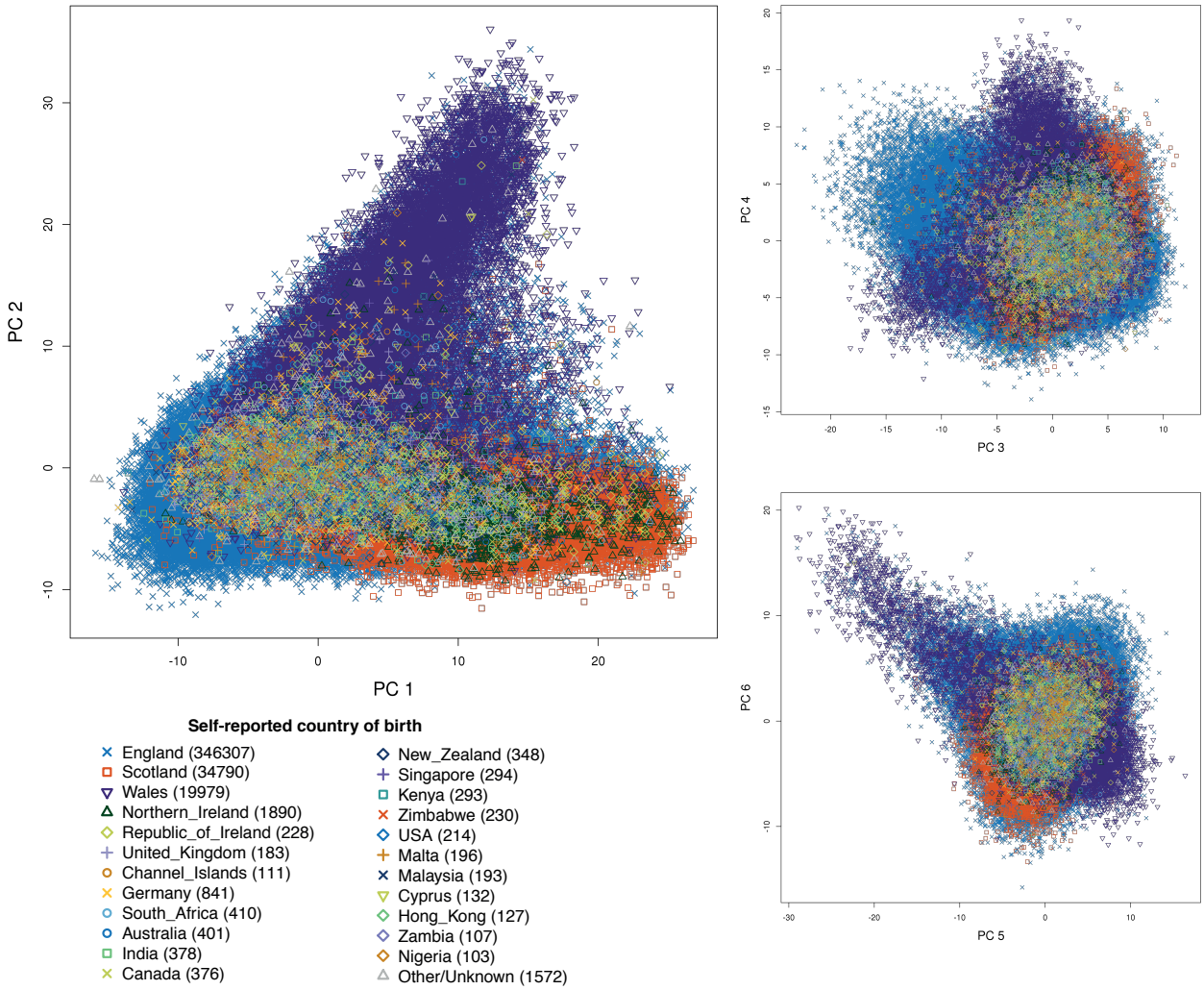


Figure 4.17: The first 6 PCs for the white British ancestry subset. Plots of consecutive pairs of the first six principal components in a PCA of genotype data for UK Biobank participants in the white British ancestry subset (as defined in Section 4.6.3). Each point represents an individual and is placed according to their principal component scores, with shapes and colours indicating their country of birth as shown in the legend.

The birth place information has been derived from the question "What is the town or district you first lived in when you were born?" [160]. All participants were asked their country of birth, but only individuals born in England, Wales, or Scotland (i.e. not

Northern Ireland or the Republic of Ireland) had the option within the UK Biobank verbal interview to state their exact place of birth. We also excluded 281 individuals whose birth-place coordinate was in the sea to the west of the Isles of Scilly[9]. This left a total of 394,603 individuals for spatial analysis.

To test formally whether PC scores correlate with geographic location, we used the global Moran's I statistic, I ([161]), which measures the correlation of some quantity of interest between nearby points (spatial autocorrelation). In general,

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \hat{x})(x_j - \hat{x}))}{\sum_i (x_i - \hat{x})^2} \tag{4.5}$$

where x_i and x_j are the values of the quantity of interest (e.g. PC scores) measured at two points in space, i and j; $\hat{x}$ is the mean value over all points; N is the total number of points; w_{ij} is some measure of closeness between points i and j; and W is the sum of all w_{ij}. By applying this statistic to PC scores in geographic space, we measure the correlation of PC scores among individuals near each other in space, where 'nearness' is defined by w_{ij}.

To allow efficient computation of this statistic for $\sim$400,000 geocoded samples, we first divided the UK into a grid of 10x10 Km squares and computed the mean PC score for individuals born within each square. We then treated each *grid-square* as a 'point' for the purposes of computing I. We constructed w_{ij} such that w_{ij} is 1 if the grid-square j is within the nearest k grid-squares of i, and 0 otherwise (based on the distance between the centres of the grid-squares). We assessed deviations of I from a null hypothesis of no correlation by re-computing the statistic after randomly permuting the PC scores among the individuals. We did this 200 times, and report the fraction of times that these are greater than or equal to the statistic using the un-permuted data. The results for different values of k are shown in Figure 4.18.

[9]These were exactly all the participants within the white British ancestry subset who were interviewed at the UK Biobank assessment centre in Stockport during the UK Biobank *pilot* study [94], and who were born within England, Scotland or Wales. It is therefore likely that the geocoding for their birth place is unreliable.

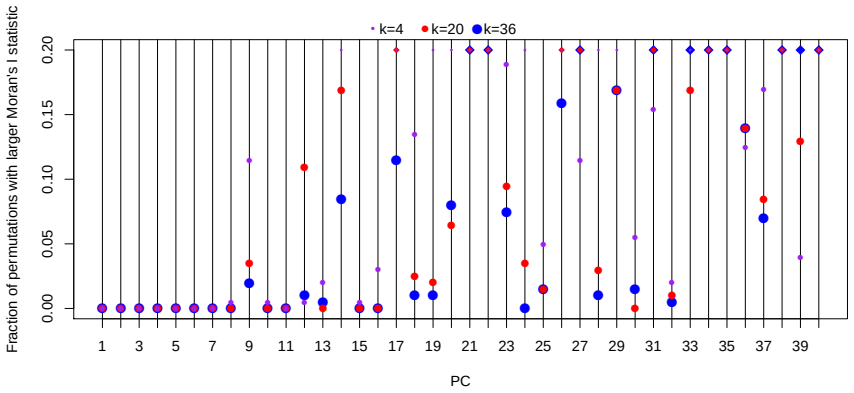


Figure 4.18: Evidence for spatial autocorrelation in PCs using different numbers of nearest neighbours. The y-axis shows the fraction of times that the Moran's I statistic under random permutation is greater than or equal to the statistic using the un-permuted data. Each colour shows different values of k, the number of nearest neighbours using in computing I. For the three values there are always exactly k nearest neighbour grid-squares (i.e. there are no ties).

The results for PC1-6 are shown in Figure 4.19, with the Moran statistics using k=20, which is equivalent to all grid-squares with centres within a 25 Km of the centre of grid-square i being among its nearest neighbours. Note that the average PC scores for individuals born in Ireland are shown in the map-based visualisations (as 2,118 of them are within the white British ancestry subset), but we excluded them from the spatial autocorrelation analysis.

PC 1 broadly separates out individuals born in Ireland, Scotland, and Wales from those in England, with the highest PC scores among individuals born in the Hebrides. PC 2 clearly separates individuals born in Wales from those elsewhere; and PCs 3 to 6 highlight structure present within Wales, and a region of England to the north-east of Wales containing the cities of Liverpool and Manchester. PCs 1-6 all showed a Moran's I statistics deviating from the null (although this is already visually clear). However, of the other 34 PCs, six showed evidence of spatial autocorrelation (PCs 7, 8, 10, 11, 15, and 16). These are shown in the top two rows of Figure 4.20.

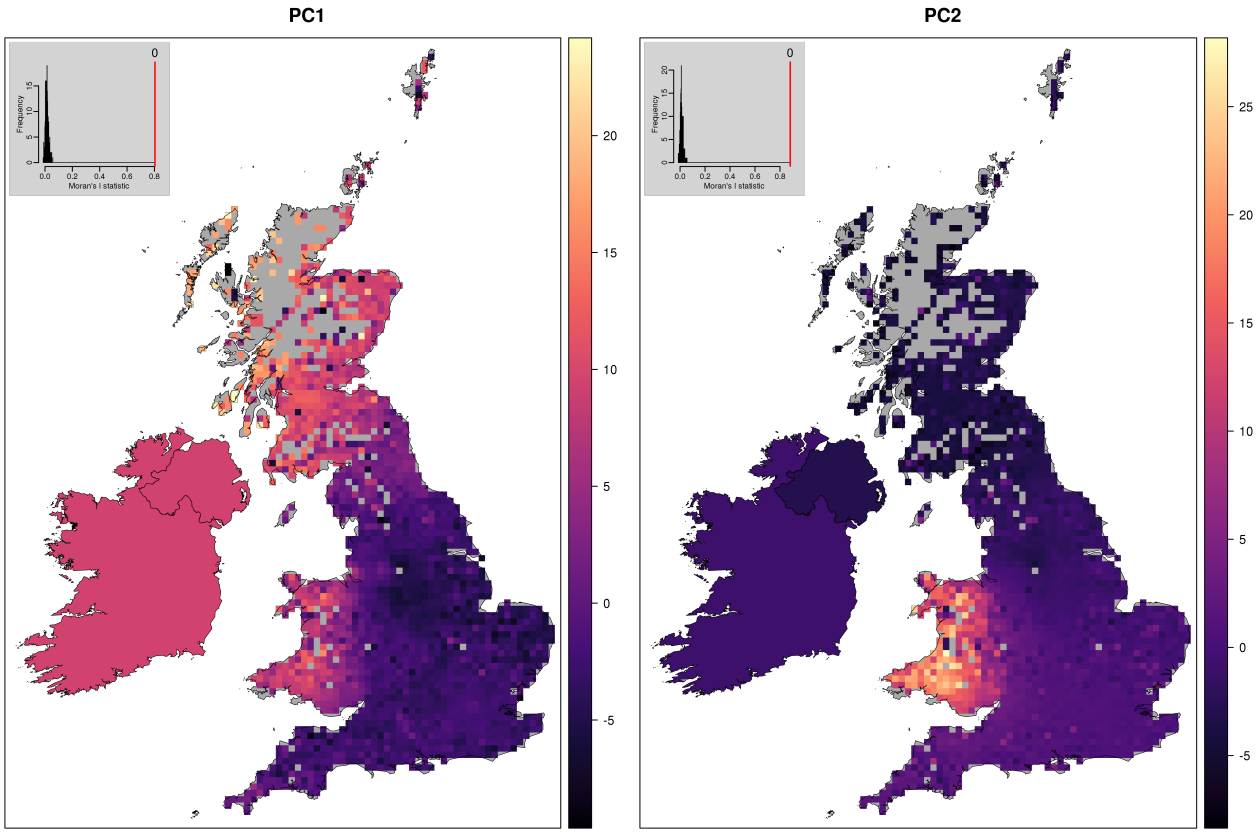


Figure 4.19: Relationship between principal component scores and place of birth for 395,231 UK Biobank participants. Each grid-square is a 10x10Km region and is coloured according to the mean PC score for UK Biobank participants born with that region. All participants were asked their country of birth, but individuals born in Northern Ireland or the Republic of Ireland did not have the option within the UK Biobank verbal interview to state their exact place of birth. There is therefore just one colour within the boundaries of Northern Ireland and the Republic of Ireland, which corresponds to the mean PC score of the individuals born in each region. The histograms in each plot show the Moran's I statistic for spatial autocorrelation computed after randomly permuting the PC scores among the individuals (excluding the Irish-born individuals). Values of Moran's I greater than zero indicate that nearby grid-squares have PC scores more similar than two random grid-squares. The red vertical line shows the test statistic for the real data, and the number above is the fraction of the 'null' scores that are greater than or equal to that value.

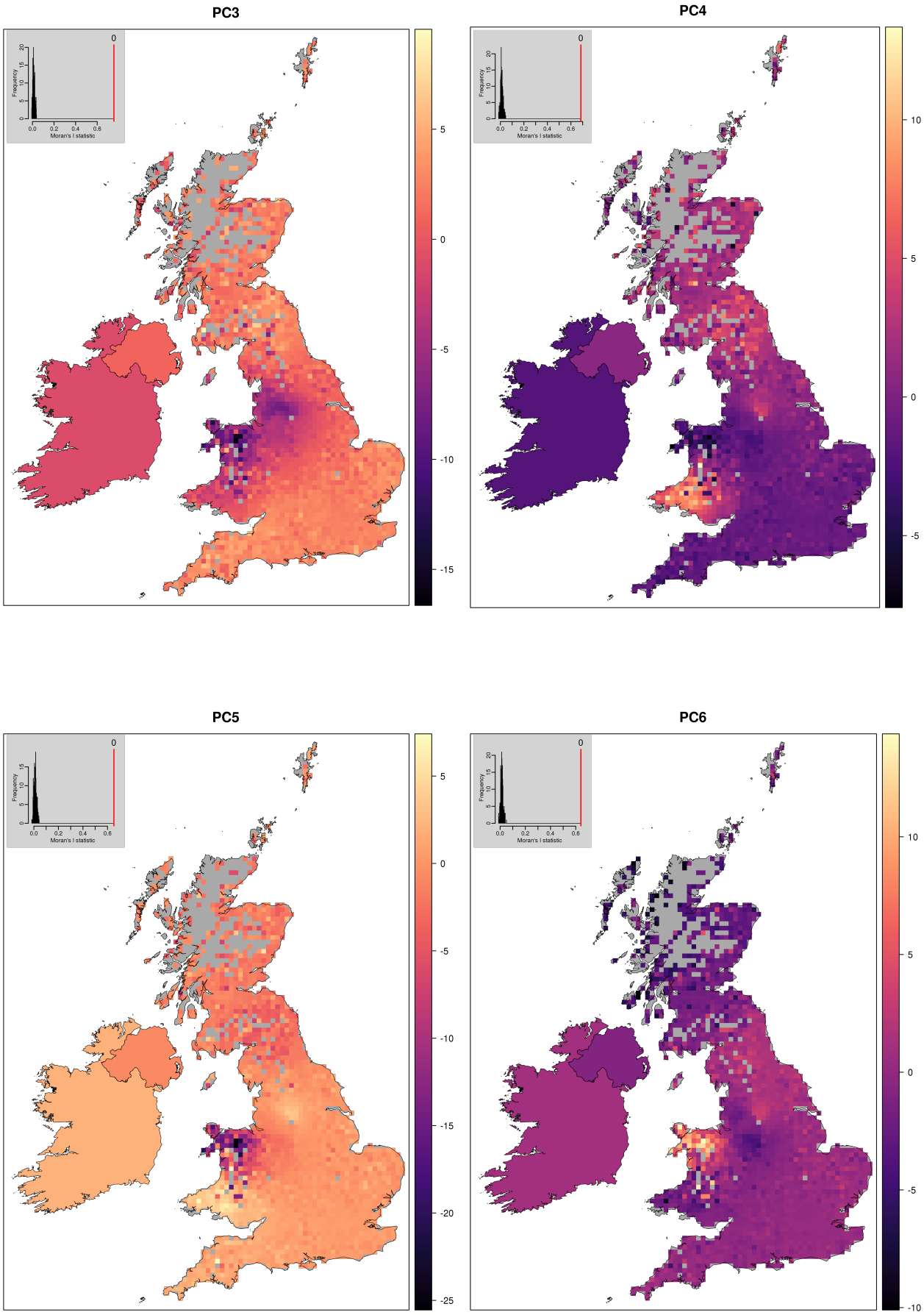


Figure 4.19: Relationship between principal component scores and place of birth (continued from previous page).

Figure 4.20: Relationship between principal component scores and place of birth – later PCs. Plots were generated in the same way as for Figure 4.19. The top six plots show PCs with significant Moran's I statistics using at least two different numbers of nearest neighbours (k) (see Figure 4.18). For comparison, the bottom three plots are examples of PCs with no evidence of spatial autocorrelation.

The SNP-loads of each principal component indicate whether a PC captures genome-wide variation, or variation in particular regions of the genome. In this case, the first 6 PCs capture genome-wide variation (Figure 4.21), as the loads are distributed evenly across the genome. However, this is not the case for further PCs, which appear to be capturing variation within local genomic regions. These later PCs tend to correlate only weakly, or not at all, with geography, and are probably highlighting regions of long-range LD, not otherwise accounted for by LD-pruning.

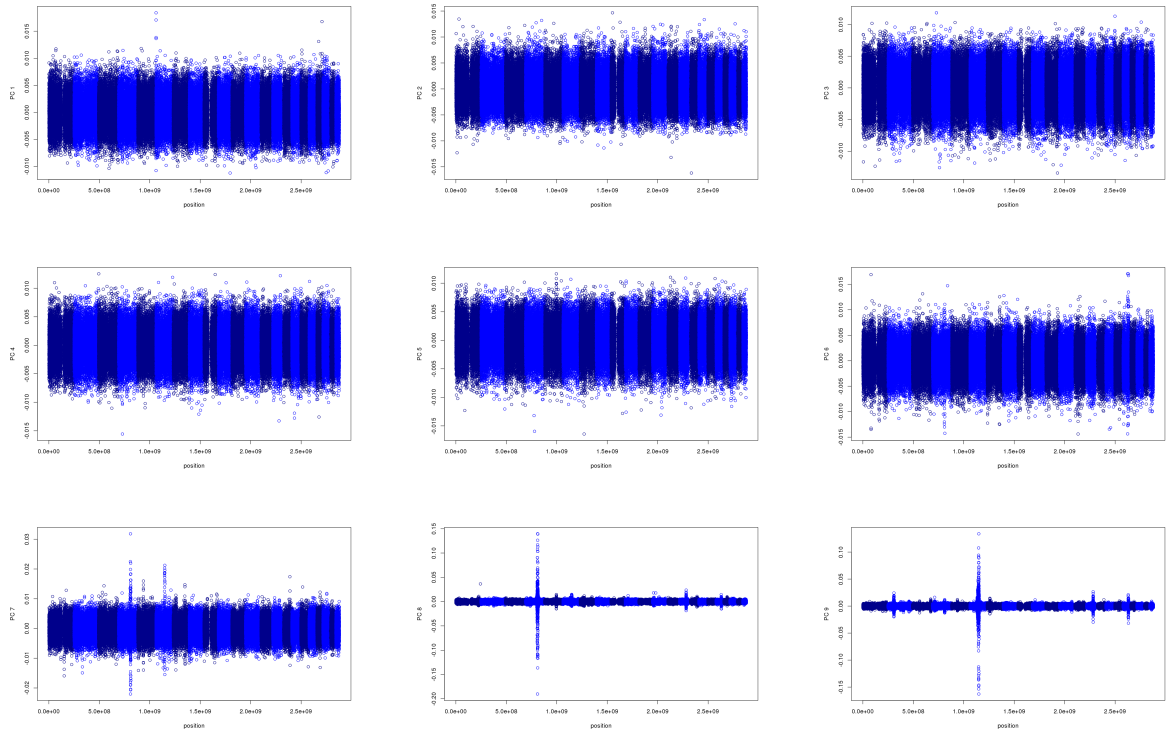


Figure 4.21: SNP-loads for PCA on white British ancestry subset. Each plot shows a PC, and the SNP-load for all the markers used in the PCA. Different chromosomes are indicated by changes in the colour of the points. SNP-loads for PCs 10 onwards show similar patterns as can be seen for PCs 8 and 9.

4.7 Cryptic relatedness in the UK Biobank

4.7.1 Inference of familial relatedness in the presence of population structure and admixture

We identified related individuals by estimating kinship coefficients for all pairs of samples, and recorded coefficients for pairs of relatives who were inferred to be 3rd degree or closer. The kinship coefficient is the probability that two alleles sampled randomly from two individuals are identical by descent (IBD). Here, 'by descent' is defined as DNA which both individuals inherited directly from the same common ancestor, and for the purposes of finding close relatives, we only consider their most recent common ancestors (e.g. grandparents, for cousins). Under this definition, expected kinship coefficients decrease by a multiple of 1/2 for each degree of relatedness. For example, parent-offspring pairs (1st degree relatives) have an expected kinship coefficient of 1/4, and grandparent-grandchild pairs (2nd degree relatives) have an expected kinship coefficient of 1/8. Variation around the expected values is a result of the stochastic nature of genetic inheritance or other effects such as parental consanguinity. Kinship coefficient estimation in a large and diverse cohort presents unique challenges. Specifically: diverse ancestral backgrounds, recent admixture, and computational scalability. We used an estimator implemented in the software, *KING* [119], as it is robust to population structure (i.e. does not rely on accurate estimates of population allele frequencies) and it is implemented in an algorithm efficient enough to consider all $\sim$1.20x10^{11} pairs in a practicable amount of time.

4.7.1.1 Accounting for the effects of recent admixture

The authors of *KING* derived the following equation for the kinship coefficient ϕ_{ij} between two individuals i and j:

$$\phi_{ij} = \frac{1}{2} - \frac{E[(X^{(i)} - X^{(j)})^2]}{4E[2P(1-P)]} \tag{4.6}$$

where $X^{(i)}$ is a random variable representing the number of reference alleles that individual i carries at a SNP, and P is a random variable, representing the allele frequency at a SNP that is randomly picked from the genotyped SNPs of an individual. The full derivation is in the Supplementary Material of [119].

The kinship coefficient estimator, $\hat{\phi}_{ij}$, used by *KING* can be calculated efficiently from the observed genotypes of individual i and j:

$$\hat{\phi}_{ij} = \frac{1}{2} - \frac{\sum_m (X_m^{(i)} - X_m^{(j)})^2}{N_{Aa}^{(i)} + N_{Aa}^{(j)}} \tag{4.7}$$

where $X_m^{(i)}$ is the observed number of reference alleles carried by individual i at SNP m, and $N_{Aa}^{(i)}$ is the observed number of heterozygous sites for individual i. Importantly, this estimator is derived under the assumption that HWE holds among markers with the same underlying allele frequencies within an individual. That is, if I_{Aa} is an indicator variable for a heterozygous genotype, then $E(2P(1-P)) = E(Pr(Aa|P)) = E(I_{Aa})$, which can then be estimated from the observed fraction of heterozygous sites.

As noted by the authors of *KING*, we found that recent admixture (e.g. 'Mixed' ancestral backgrounds) tended to inflate the estimate of the kinship coefficient. An individual with recent admixture will inherit alleles which have two (or potentially more) different population-level frequencies, such that the probability of them carrying a heterozygous site no longer follows HWE. In the extreme case where an individual's parents have different ancestral backgrounds, then the probability of the child carrying a heterozygous genotype is $Pr(Aa|p_1, p_2) = p_1(1-p_2) + p_2(1-p_1)$, where p_1 is the allele frequency in the mother's population, and p_2 is the frequency in the father's population. As noted earlier (Equation 4.4) this departure from HWE always results in an excess of genome-wide heterozygosity (unless $p_1 = p_2$), so $\hat{\phi}_{ij}$ will tend to overestimate the true kinship coefficient.

We alleviated this effect by using a subset of markers that are only weakly informative of ancestral background, based on the SNP-loads in the PCA (details are in Section 4.7.1.2). This means in the above case $p_1 \approx p_2$; and more generally, the

assumption of HWE is more likely to hold, even in the context of recent admixture. The effect of this filtering in UK Biobank data is illustrated in Figure 4.22 and Figure 4.23.

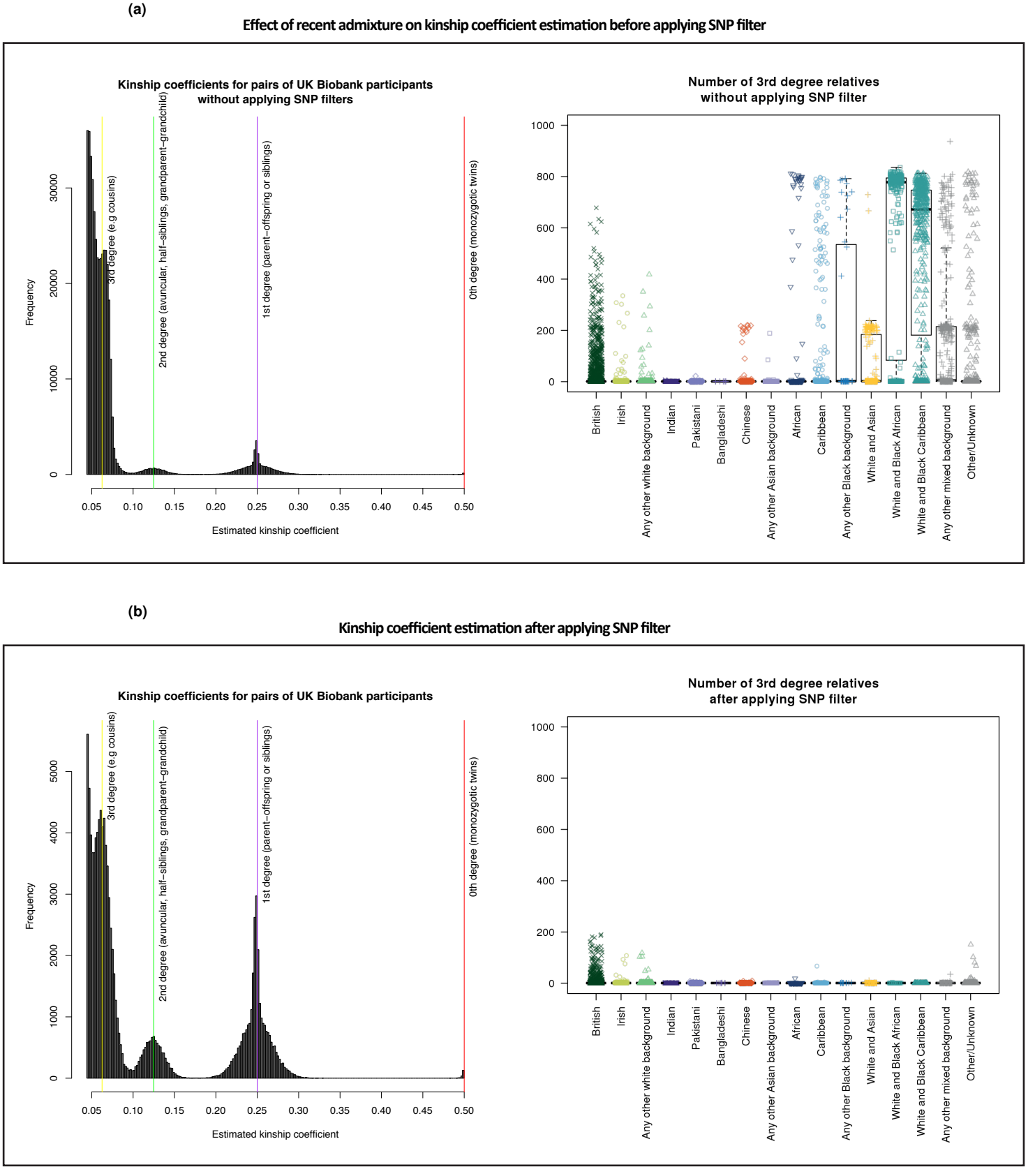


Figure 4.22: The effect of PC-based SNP filtering on kinship coefficient estimation. Both sub-figures show a histogram (on the left) of kinship coefficients estimated by *KING* using a 3rd degree cut-off, with vertical lines placed at the expected coefficients for different degrees of relatedness. The box plots on the right show the distribution of the number of 3rd degree relatives inferred for each sample (excluding zero) within each ethnic background category. **(a)** Plots based on kinship coefficient estimates using a set of autosomal SNPs selected for genotyping quality (see Section 4.3). The excess of 3rd degree pairs is evident in the histogram, and the boxplot shows how the ethnic groups involving mixed ancestry are disproportionately affected. The accumulation of points around ∼800 and ∼200 relatives occurs because there are two sets of samples in which almost all pairs appear 'related' to each other. Namely, those with ethnic backgrounds originating in Africa+Europe and Asia+Europe. **(b)** Plots based on kinship coefficient estimates after excluding SNPs informative of ancestry based on PCA (see Section 4.7.1). The reduction in an excess of 3rd degree relatives among mixed ancestry ethnic groups is clear from the box plot.

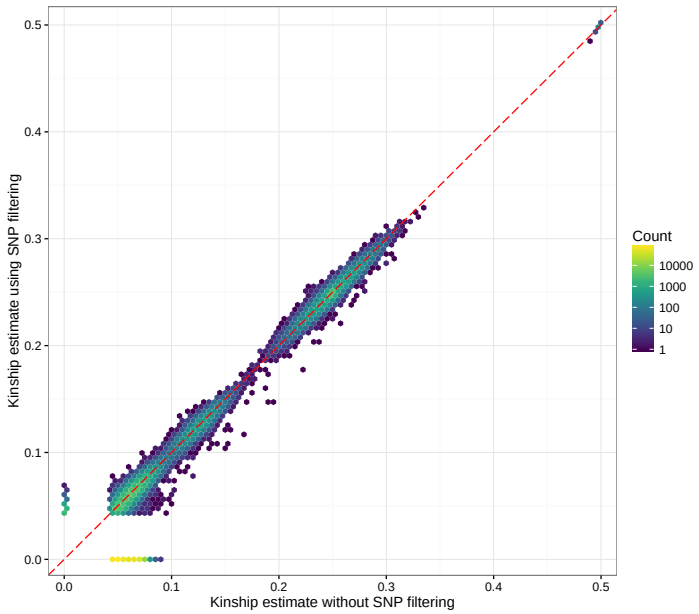


Figure 4.23: Kinship coefficient estimates before and after filtering SNPs. On each axis are the kinship coefficients of pairs of samples inferred to be 3rd degree relatives or closer in two analyses using *KING*. Colours indicate the number of pairs that fall within the range of each hexagonal bin. Most of the pairs that changed relationship class as a result of the SNP filtering were those that shifted from '3rd degree' to 'unrelated' (yellow/green hexagons in the bottom left).

4.7.1.2 Details of relatedness inference

We first selected a set of SNPs that are only weakly informative of ancestry to minimise inflation of the kinship estimates due to recent admixture (as justified above). Using results of the PCA round 1 (see Section 4.6.1.1) we selected SNPs that only contribute very small 'loads' to PCs 1-3. That is, where t_k is the value of SNP-load for PC k, we only used SNPs with $t_k < 0.003$ for all k in 1,2,3. The threshold was chosen to balance the number of SNPs included (too few would lead to noisy kinship estimates), and how informative of ancestry the SNPs are (too large a threshold would lead to inflation in the presence of recent admixture). This resulted in a set of 93,511 SNPs to use for the final kinship inference. We also excluded individuals in the list of outliers in heterozygosity and missing rates (see Section 4.4.1).

With the genotypes filtered as described above, we computed kinship coefficients for all pairs of individuals using *KING* and recorded the pairs of degree 3 or closer (kinship coefficient $\geq \frac{1}{2^{(9/2)}}$) [119]. In practice, we parallelised this computation by combining

data into pairs of batches ('--merge' command in *plink*) and running *KING* with the options '--related –degree 3' on all pairs of batches. We then merged the results into one pairwise kinship table.

A small number of individuals (9) appeared to be related (3rd degree) to a very large number ($>$ 200) of individuals. In some cases this was in the order of 1000s, and their 'relatives' were usually not themselves related to one another. By considering family trees, it is only possible for an individual to have a maximum of four 3rd degree relatives who are not themselves related. These individuals also had slightly elevated heterozygosity and missing rates, but not extreme enough to flag as poor quality (as defined in Section 4.4.1). We concluded that the excess related pairs are likely to be false positives, and being driven by a small number of individuals, so we excluded them from the kinship table. These 9 individuals, along with the pre-filtered samples, comprise a set of 977 samples that are effectively excluded from the kinship inference. So for this small fraction (0.2%) of the cohort we therefore cannot confirm the presence or absence of any of their relatives in the cohort.

We also called the relationship class of each pair (see Table 4.4 and Figure 4.27). That is, we assigned each related pair to one of twins, parent-offspring, siblings, 2nd degree or 3rd degree relatives using the kinship coefficient boundaries recommended by the authors of *KING* (See Table 1 in [119]). We used the fraction of markers with zero alleles identical by state (IBS0) only to distinguish parent-child from sibling pairs, who have the same expected kinship coefficient. Specifically, we called any pair with IBS0 $<$ 0.0012 as parent-offspring, based on visual inspection of the kinship estimates for all 1st degree relatives. A total of 147,731 UK Biobank participants (30.3%) are inferred to be related (3rd degree or closer) to at least one other person in the cohort, and form a total of 107,162 related pairs (Table 4.4).

	Monozygotic twins	Parent-offspring	Full siblings	2nd degree	3rd degree	Total
Number of pairs	179	6,276	22,666	11,113	66,928	107,162

Table 4.4: Summary of related pairs (3rd degree or closer) for the full UK Biobank cohort. Counts are derived from the kinship coefficients as recommended by the authors of *KING* [119] (see Section 4.7.1). Note that parent-offspring and full sibling pairs have the same expected kinship coefficient (0.25) but can be easily distinguished by their IBS0 fraction. The count of monozygotic twins is after excluding samples identified as duplicates (see Section 4.7.4).

4.7.1.3 Comparison with inference using a different estimator

We validated the kinship estimates by applying a different kinship coefficient estimator based on allele frequencies, implemented in *plink*'s '--genome' command [124]. We used the same set of LD-pruned SNPs as with the *KING* analysis, and the option '--min' to apply the same cut-off for 3rd degree (2 x $\frac{1}{2^{(9/2)}}$ = 0.08838835). The multiple of 2 accounts for the fact that *plink* actually estimates the IBD-sharing fraction which is, by definition, twice the kinship coefficient [119]. In order to avoid population structure effects we restricted the analysis to pairs of related individuals (according to the *KING* analysis detailed above) where both are in the white British ancestry subset (see Section 4.6.3). These account for 85% of all the inferred pairs.

Of all the pairs of relatives in the subset 99.9% were also inferred as 3rd degree or closer using *plink*. The small fraction of unconfirmed pairs all had a kinship coefficient (according to the *KING* analysis) smaller than 0.0486, which is close to the recommended cut-off between 3rd and 4th degree. Furthermore, all twins, parent-offspring and sibling pairs were inferred to have the same degree of relatedness in the *plink* analysis. There was some discrepancy between the assignment of 2nd and 3rd degree relatives. A number (5%) of 3rd degree pairs from *KING* were called as 2nd degree pairs in the *plink* analysis, although *plink* also inferred a much larger number ($\sim 10^7$) of 3rd degree pairs, which is unrealistic for this data set.

4.7.2 Estimating the theoretical expectation of the number of related pairs in the cohort

According to the above analysis, around 30% of UK Biobank participants have at least one relative in the cohort. In order to verify that the surprisingly large numbers of related individuals was within reasonable expectations (or if not, by how much), we estimated the theoretical expected number of sibling or cousin pairs in a simple random sample of the population eligible for UK Biobank. To do this we[10] derived the following equations with parameter definitions shown in Table 4.6.

Expected number of cousin pairs in sample of size n
= Pr(Sampling a cousin pair)x(Number of pairs in sample)

$$\begin{aligned} &= \frac{2\mu_1(\hat{\mu}_0 - 1)}{N-1}\frac{n(n-1)}{2} \\ &= \frac{\mu_1(\hat{\mu}_0 - 1)n(n-1)}{N-1} \end{aligned} \tag{4.8}$$

Expected number of sibling pairs in sample of size n
= Pr(Sampling a sibling pair)x(Number of pairs in sample)

$$= \frac{(\hat{\mu}_1 - 1)}{N-1}\frac{n(n-1)}{2} \tag{4.9}$$

In applying these equations to the UK Biobank sample, we assume the following:

- Most eligible individuals are descendants of UK citizens so we can apply historical UK fertility rates.
- The majority of 3rd degree pairs in the cohort are first cousins (as opposed to great-uncle/aunts; or connections involving half-siblings). This is likely to be true given that the age-range only spans about one human generation, and assuming half-siblings are relatively rare.
- The sample size (n) is small compared to the population size (N) so that sampling with replacement can be assumed.

[10] We acknowledge the assistance of Dr. Alex Young in these derivations.

Several factors make estimating these values challenging. Fertility rates of mothers that were having children during the time of the birth years of this cohort (1938 – 1968) changed dramatically (see Figure 4.24). Therefore, the mean family size is likely to depend on the birth-year of the mothers of individuals in the cohort. Secondly, the age-distribution of women bearing children also changed over this time, affecting the likely birth-year of the mothers. Instead of modeling these factors directly, we simply computed a maximum and minimum expected value, based on the maximum and minimum observed fertility rates for UK mothers having children who would be eligible for the UK Biobank cohort. Furthermore, the UK Biobank cohort is not a simple random sample of the eligible population. In particular, people aged 60-70 are over-represented and people aged 40-44 are under-represented (see Figure 4.24; [162]). Given that the estimates depend on the sampling fraction (n/N), and this is different for different age-groups in the cohort, we computed the estimate separately for 5-year age-groups and summed the results. Therefore, any discrepancies in the estimates and the observed values would be a result of sampling bias *independent* of the age-related bias.

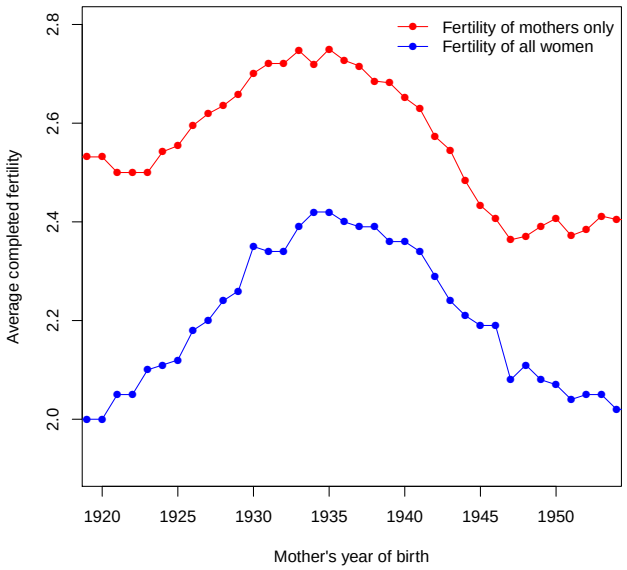


Figure 4.24: Cohort fertility of mothers of the UK Biobank eligible population. The data was sourced from Office of National Statistics, UK. Fertility statistics are reported in cohorts of women born in a particular year. The completed fertility is the number of children born per woman, within a cohort. To calculate the completed fertility for mothers only, we divided the completed fertility for all women in a cohort, by the fraction of women in the cohort who had at least one child by the end of their fertile years. 1920 is the earliest birth year for which these statistics are available, and mothers born after 1953 would likely be too young (13 or younger) to have given birth to someone eligible for the UK Biobank.

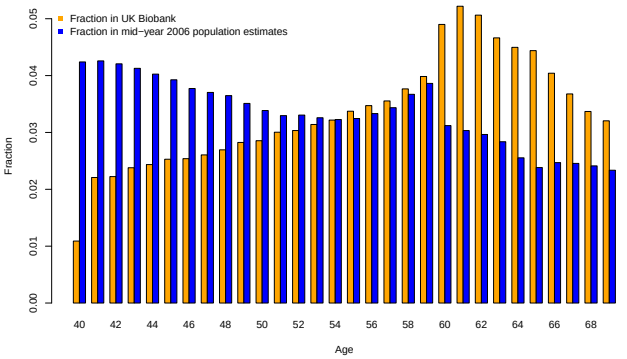


Figure 4.25: Bias in the representation of eligible population in UK Biobank cohort for age. We sourced mid-year 2006 population estimates from the Office of National Statistics (ONS).

Table 4.5 shows the expected and observed numbers of pairs in the UK Biobank, using the parameters shown in Table 4.6. The number of sibling pairs (22,666) is about twice as many as would theoretically be expected (under the conditions stated

about), and between 1.2 and 2 times as many first cousin pairs. We can therefore conclude that the large number of related participants is not driven solely by an excess of 3rd degree relatives (which are more likely to involve some false positives).

	Expected number of pairs (range)	Observed number of pairs
1st degree sibling pairs	9,530 - 11,110	22,667
3rd degree pairs (cousins*)	36,390 - 53,790	66,935

Table 4.5: Expected and observed numbers of pairs of related individuals in the UK Biobank cohort. Expected ranges are based on using the minimum (and maximum) parameters for family sizes given in Table 4.6, and computed using the equations shown in Section 4.7.2. *We estimated the expected number of cousin pairs (column 2), but the observed number (column 3) shows all observed 3rd degree pairs. We make an assumption here that almost all of the 3rd degree pairs in the UK Biobank will be cousins, rather than connections spanning 3 generations (e.g. a great aunt/uncle) or involving half-siblings.

Parameter	Description	Range(s)	Source
N	Total population size eligible for the UK Biobank sample.	Total: 21,734,300	2006 mid-year population estimate for people aged 40-69 (Office of National Statistics, UK)
n	Size of UK Biobank sample (and successfully genotyped).	Total: 485,029	UK Biobank
μ_1	Average completed family size in sampled generation (includes childless mothers). Counts expected number of children per aunt/uncle.	[2.00, 2.42]	Completed cohort fertility for women born between 1920 and 1953 (Office of National Statistics, UK).
$\hat{\mu_1}$	Average completed family size in sampled generation (excludes childless mothers). Counts expected number of siblings.	[2.36, 2.75]	Completed cohort fertility for women born between 1920 and 1953 and have given birth to at least one child (Office of National Statistics, UK).
$\hat{\mu_0}$	Average completed family size in previous generation (excludes childless mothers). Counts expected number of aunts/uncles.	[2.36, 2.75]	No direct data is available on fertility rates of women born before 1920 so assume the range is similar to $\hat{\mu_1}$.

Table 4.6: Parameters for estimating number of expected sibling or cousin pairs in UK Biobank cohort. Note that we used different values for N and n for each 5-year age-group between 40 and 69. For brevity, just the total is shown in this table.

4.7.3 Trios and family groups

Pairs of related individuals within the UK Biobank cohort form networks of related individuals, or 'family' groups. In most cases these are of size two, but there are many groups of size 3 or larger in the cohort, even when restricting to 2nd degree or closer relative pairs (Figure 4.26). By considering the relationship types and the age and sex of individuals within each family group, we identified 1,066 sets of trios (two parents and an offspring), which comprises 1,029 unique sets of parents and 37 quartets (two parents and two children). There are no instances of 3-generation nuclear families (grandparent-parent-offspring), which is not surprising, given that the age-range of the cohort spans only 30 years.

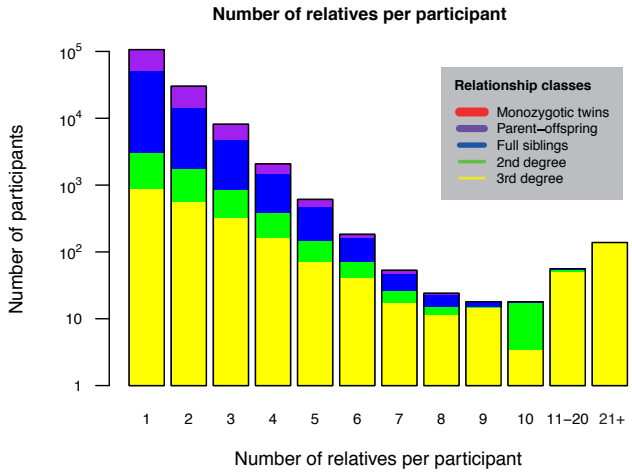


Figure 4.26: Distribution of the exact number of relatives that participants have in the UK Biobank cohort. The height of each bar shows the count of participants who have exactly the stated number of relatives. Note the logarithmic scale. The colours indicate the proportions of each relatedness class for individuals counted within a bar. For example, for each individual that has three relatives in the cohort (3rd bar), we count how many of their relatives are in each relatedness class (i.e. a full sibling, parent-offspring etc.). We then sum these counts over all individuals with three relatives and colour the bar according to the proportions of each relatedness class. In this group $\sim$20% of their relatives are full siblings and $\sim$64% are 3rd degree relatives. There are also 18 participants with exactly 10 relatives. The unusually large fraction of 2nd degree relationships for this group is a result of the set of eleven individuals who are all 2nd degree relatives of each other, as shown in the centre of Figure 4.27.

There are 172 family groups with 5 or more individuals that are related to 2nd degree or closer, and Figure 4.27 illustrates examples of large nuclear families in the cohort. One of these is a group of eleven individuals where every pair is a 2nd degree relationship. The simplest explanation for such a family group is that all of the individuals are half siblings, with one shared parent who is not in the cohort.

Alternatively, one individual in the group could be a sibling or a parent of the shared parent. The pattern of haplotype sharing between the individuals would distinguish these cases but we have not undertaken this analysis. We did, however, confirm that the shared parent must be their father, because the individuals do not all carry the same Mitochondrial alleles, and all the males in the group have the same alleles on their Y chromosome (data not shown).

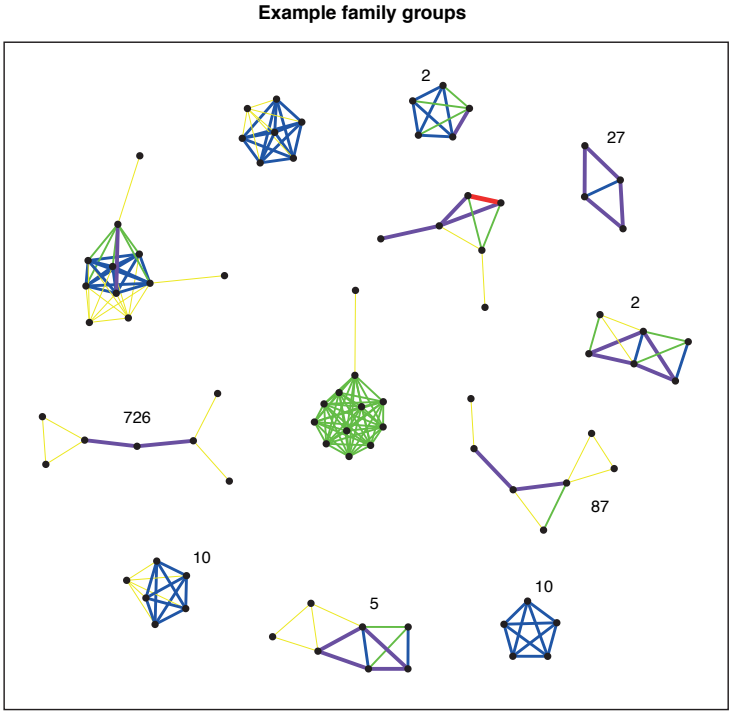


Figure 4.27: Examples of family groups within the UK Biobank cohort. Points indicate participants, and lines between points indicate familial relatedness (3rd degree and closer) as inferred from the genetic data (see Section 4.7.1). The colour and thickness of the lines indicate different relative classes, as shown in the key. An integer next to a network indicates the total number of family networks in the cohort with the same configuration, ignoring 3rd degree pairs. No integer means there is only the one shown. For example, there are 10 networks that comprise exactly 5 full siblings (two examples, which differ with respect to a 3rd degree relative, are shown on this plot); and there is only one network that comprises 6 full siblings (plus one 3rd degree relative who is related to all siblings).

4.7.4 Distinguishing identical twins from duplicated samples

Without considering phenotype information, a pair of duplicated samples in the genotype data will be indistinguishable from genuine identical twins because they will all have kinship coefficient 0.5. To resolve this, UK Biobank staff reviewed phenotype

details of a list of 894 candidate pairs of samples (all those we inferred to have a kinship coefficient close to 0.5). Where evidence was found that the participants may be twins, triplets, or part of a multiple birth, the pair was marked as twins (188 pairs). Any remaining pairs were marked as either 'Blind Spike Duplicates' (588 pairs) or 'unintended' duplicates (118 pairs). Blind Spike Duplicates are those where an extra aliquot from the same participant was deliberately included in the genotyping experiment as a validation tool [108]. Unintended duplicates were pairs of samples that had not been included as Blind Spike Duplicates, and were associated with phenotype information from different participants who were not known to be identical twins.

A total of 1,364 samples were identified as duplicates, some duplicated more than once. We[11] excluded 793 of these from the released genotype data. That is, we excluded all samples within the unintended duplicate pairs because for these samples the correct link between the genotype data and phenotype information cannot be guaranteed; and we kept a single sample from each of the Blind Spike Duplicates (the one with the highest call rate).

4.7.5 Finding a maximal set of unrelated individuals

Given the large proportion of related participants in the UK Biobank, for analyses that require a set of unrelated individuals (e.g. PCA) it is useful to exclude related samples in such a way that the remaining unrelated subset is as large as possible. We used the following procedure to find such a maximal set of unrelated individuals (for which there are many possible solutions) among a set of samples of interest (e.g. after QC filtering prior to PCA). We first pruned the full pairwise kinship table so that it only contained individuals in the set of interest. Using the *i-graph* (v1.0.1) package [163] in *R* we then converted the table into a graph object, where each vertex is an individual, and edges exist between pairs of related individuals. We next identified all the 'family' groups (i.e. a network of nodes joined by edges) using the 'clusters' function in the same *R* package. For each of these groups we found the largest subset of individuals

[11] This exclusion procedure was carried out by C.F.

(vertices) such that there is no relatedness (edges) between them. In the case of trios, for example, the child would be excluded, leaving the two unrelated parents. To do this we used an algorithm implemented in the 'largest_ivs' function in *i-graph*. When there was a choice of solutions (e.g. within a set of 3 siblings), we chose one solution at random.

4.8 GWAS for standing height

4.8.1 Details of GWAS analysis

As a final demonstration of the quality of the genotyped and imputed data we conducted a genome-wide association scan for a well-studied [164], and highly polygenic human trait: standing height. We conducted tests using the directly genotype and imputed data files separately, in the form that they are made available to researchers, but with a subset of samples. Specifically, we only included samples with all of the following properties:

- Imputation was carried out on them (see Section 2.3 of main text).
- In the white British ancestry subset (see Section 4.6.3).
- Inferred sex matches self-reported sex (see Section 4.4.2).

From this group we selected a set of 344,397 unrelated individuals (using the procedure described in Section 4.7.5). For standing height, a further 1,076 individuals were excluded due to missing values for the phenotype, leaving a total of 343,321 for association testing. We used the software *BOLT-LMM* (v2.2) [165] to look for evidence of statistical association between each marker and standing height. We report association statistics based on a linear regression model with the following covariates.

- Array (UK BiLEVE Axiom Array or UK Biobank Axiom Array).
- Sex (inferred by Affymetrix).
- Age when attended UK Biobank assessment centre.

- Principal components[12] 1-20.

Previously published, large cohort studies for this trait also provide a suitable independent set for comparison with UK Biobank results. Specifically, we compared the UK Biobank results to the largest published GWAS for human height that does not use UK Biobank data: a meta analysis involving a total of 253,288 individuals of European ancestry carried out by the Genetic Investigation of Anthropometric Traits (GIANT) Consortium [166] in 2014.

4.8.2 GWAS results and comparison with GIANT

The p-values for associations with height on one chromosome are shown in Figure 4.28, along with p-values for GIANT's meta analysis. Reassuringly, the pattern of associations is very similar in both the UK Biobank and GIANT results. The gain in power in the UK Biobank cohort is clear, with many loci reaching genome-wide significance (p-value $<$ 5x10^{-8}) in the UK Biobank, which do not in the GIANT study (Figure 4.29). The purpose of this analysis was not to discover novel associations for height, rather to indicate the potential of the resource to uncover such findings. As an illustration, Figure 4.30 shows a single associated region on chromosome 2, which does not reach genome-wide significance in GIANT. Correlations (r^2) between markers in this region show a pattern that is as expected in the context of linkage disequilibrium (LD), and the local recombination rates. The stripe-like pattern of the association statistics is indicative of multiple mutations occurring on similar branches of the genealogical tree underlying the data, which are likely linked to varying degrees with the causal marker(s). The correlation (r^2) between the most associated marker and all other markers in the region drops off sharply around the small peak in recombination (Hapmap recombination map [111]) to the right of the most significantly associated marker. Interestingly, this marker was imputed from the genotypes, which points to the success of the imputation in this study, and in general, to the value of imputing millions more markers.

[12]The PCs were computed as described in Section 4.6.1.2, but using only individuals within the white British ancestry subset.

Human height is a highly polygenic trait [164] and this provided an opportunity to examine many such regions of association. Other regions that we visually examined showed similar patterns, and all regions containing the variants reported to be associated with height in the NHGRI-EBI GWAS catalogue [164] (as of Feb 2017, excluding results based on UK Biobank data) were also genome-wide significant, or close to, in the UK Biobank data.

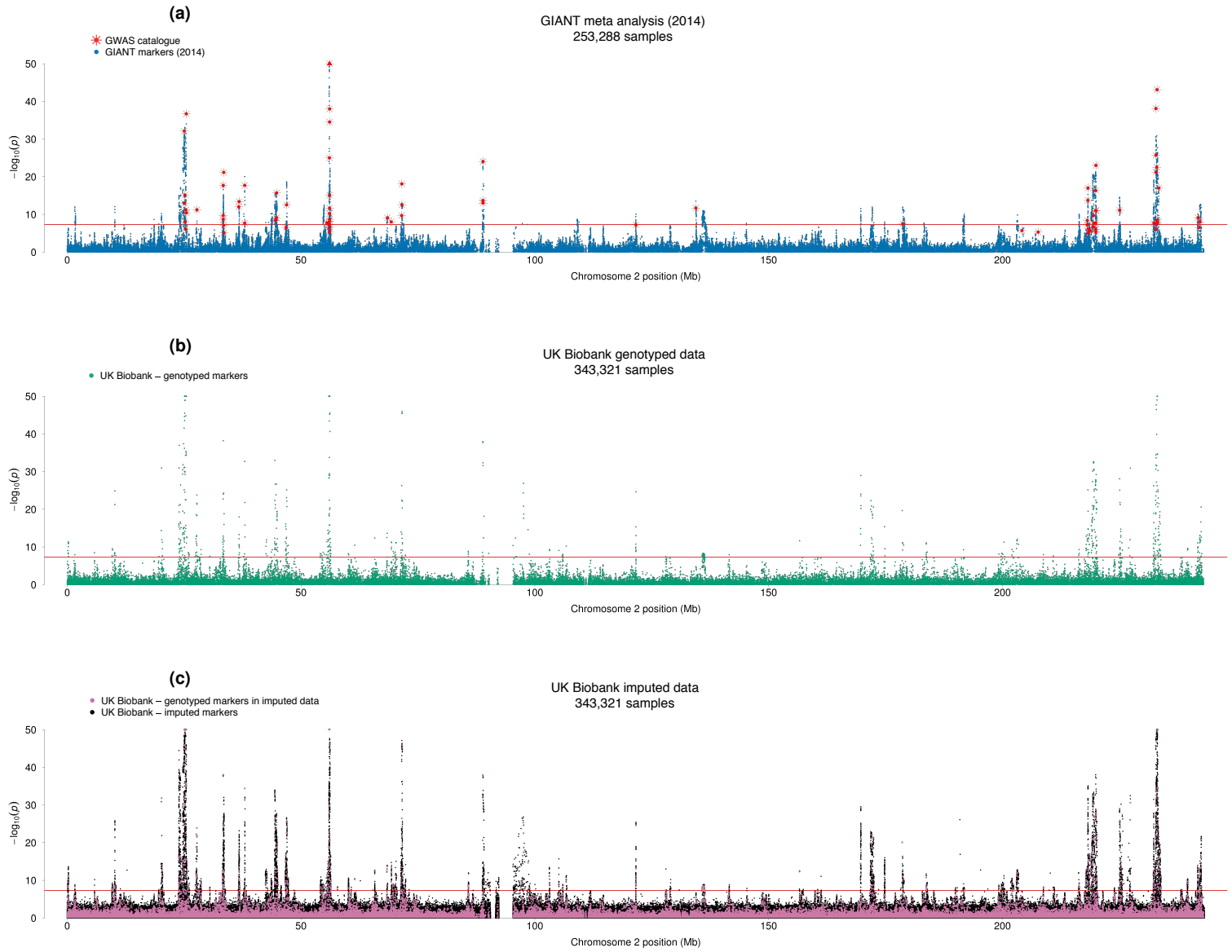


Figure 4.28: Association statistics for human height. Results (*p*-values) of association tests between human height and genotypes using three different sets of data for chromosome 2. *p*-values are shown on the $\log_{10}$ scale and capped at 50 for visual clarity. Markers with $\log_{10}(p) > 50$ are plotted at 50 on the y-axis and shown as triangles rather than dots. **(a)** Results for meta-analysis by GIANT (2014), with NHGRI-EBI GWAS catalogue markers [164] superimposed in red (plotted at the reported p-values). **(b)** Association statistics for UK Biobank markers in the genotype data. **(c)** Association statistics for UK Biobank markers in the imputed data. Points coloured pink indicate genotyped markers that were used in pre-phasing and imputation. This means that most of the data at each of these markers comes from the genotyping assay. Black points (the vast majority, $\sim$8M) indicate fully imputed markers.

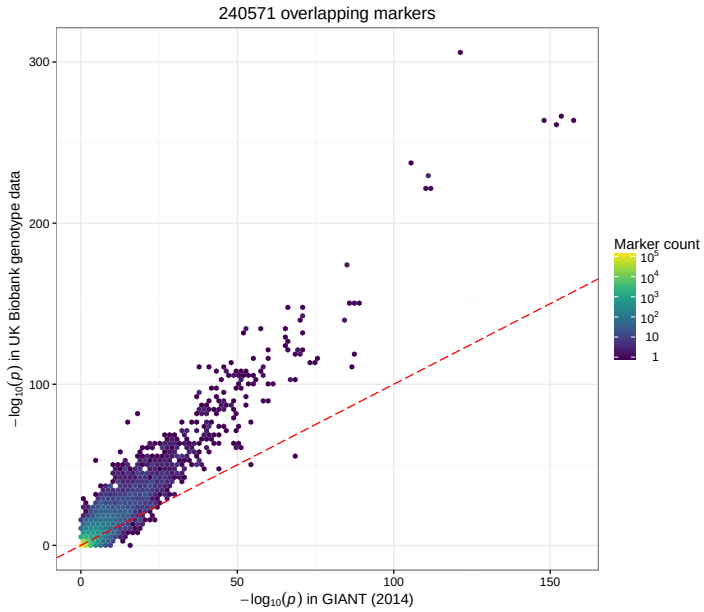


Figure 4.29: Comparison of *p*-values for UK Biobank and GIANT in standing height GWAS. Each point is a marker on chromosome 2 that was included in both UK Biobank genotype data and GIANT (2014). We identified markers common to both studies by matching on chromosome, position, and the two alleles. The red line shows x=y.

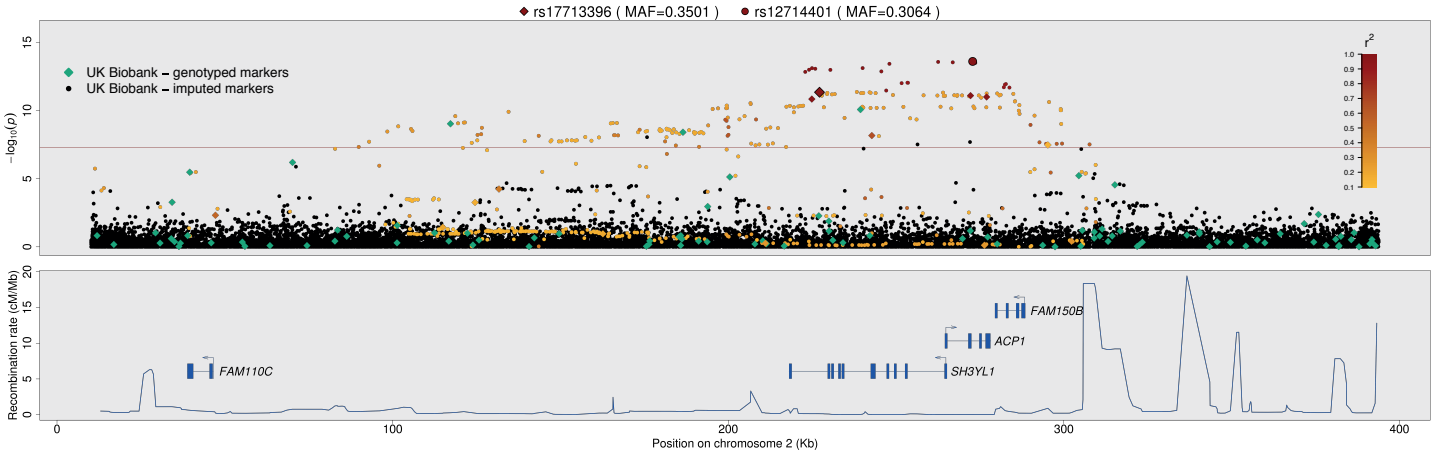


Figure 4.30: Results of standing height GWAS focussing on a ∼3 Mega-base region at the terminal end of the p-arm. Genotyped markers (i.e. markers in Figure 4.28b) are shown as diamonds, and imputed markers (i.e. only markers coloured black in Figure 4.28c) as circles. The two markers with the smallest *p*-value for each of the genotyped data and imputed data are enlarged and highlighted with black outlines, and other UK Biobank markers are coloured according to their correlation (r^2) with one of these two. That is, genotyped markers with the leading genotyped marker (rs17713396), and imputed markers with the leading imputed marker (rs12714401). Markers with r^2 less than 0.1 are shown as black or green.

4.9 Discussion

4.9.1 Quality control of UK Biobank genotype data

Quality control for the UK Biobank genotype data presented several challenges over and above the computational challenge of its size. Aspects of the experimental design, especially the inclusion of two different arrays, and the fact that the genotype calling was carried out separately in many batches, posed potential experimental confounding factors. However, the amount of data impacted by such effects is relatively small (see Table 4.2), which points to the success of the tightly-controlled laboratory protocols and automated sample extraction procedure (Section 1.3.3) of the UK Biobank genotyping experiment. This is further highlighted by the very small fraction (0.002%) of samples that we identified as poor quality (Section 4.5.1), and the high replication rates for deliberately duplicated samples (Figure 4.10). The MAF of markers in UK Biobank data are reassuringly consistent with an external source, ExAC (Figure 4.11). This is less true for very rare markers (MAF < 0.001), which is not surprising given that they are generally more prone to error using genotype array technology.

One further challenge was that we aimed to apply quality control to a data set that will be used by a research community with a diverse range of research questions. Most QC metrics require a threshold beyond which to consider a marker 'not reliable'. As such, we used thresholds such that only strongly deviating markers would fail QC tests (Section 4.2.2), therefore allowing researchers to further refine the QC in whichever way is most appropriate for their study requirements.

The number of individuals with putative sex chromosome aneuploidies is much larger than the number reported in the UK Biobank phenotype data. There are approximately $\sim$130 cases of some form of sex chromosome aneuploidy reported in the hospital admissions diagnosis data (data fields 41202 – 41205), the most common being Klinefelters (XXY) and Turner's (0X) syndromes. The fact that there are around 5 times as many in the genetic data (see Table 4.3) suggests that the phenotypic consequences of these aneuploidies are often not severe enough to require hospitalisation and diagnosis. Even if a participant is aware that they carry a

sex chromosome aneuploidy, this information is not available in the UK Biobank self-reported data, so the genetic data itself is the best source of this information. However, we should note that the karyotype calls we made have not been independently validated, for example using laboratory-based karyotyping. Rather it would be a useful starting point for more detailed analysis of sex chromosome aneuploides, or as a QC tool. For instance, this set of individuals was excluded from phasing and imputation on the X chromosome only, as assumptions about diploidy/haploidy in this chromosome may not hold for these individuals. The list could also be used to differentiate between sex-mismatches likely due to unusual chromosomal configurations, as opposed to other reasons such as clerical error.

Finally, the test GWAS on human height (Section 4.8.2) provides further evidence for the quality of the resulting set of genotype calls, as well as highlighting the potential of this data set to uncover novel associations.

4.9.2 Extensive cryptic relatedness among UK Biobank participants

We observe a surprisingly large number of related individuals among the UK Biobank participants, even after addressing a number of potential methodological artefacts (such as the effects of recent admixture). The large number of related pairs could be explained by sampling bias due to, for example, an individual being more likely to agree to participate because a family member was also involved. Furthermore if, as seems likely, related individuals cluster geographically, rather than being randomly located across the United Kingdom (UK), the assessment-centre based recruitment strategy employed by UK Biobank [94] will naturally tend to oversample related individuals, relative to random sampling of the UK population.

The group of 11 2nd degree relatives (see Figure 4.27) is worth further discussion. The birthplaces and age-range (20 years) of these individuals (data not shown) suggests that they are unlikely to have shared a familial upbringing. The UK Biobank cohort represents about 2.2% of the 2006 resident population aged 40-69. If we assume the cohort is a simple random sample (or that any bias is independent of

whether or not someone shares this same father), we can estimate the likely number of individuals who share the same father as the eleven sampled in the UK Biobank cohort: 491 (201-780)[13], or 446 (170-722) if only 10 of the individuals are half siblings. A possible explanation for this unlikely reproductive excess is that the father was involved in a sperm donation program operating in the middle of the 20th Century. There is indeed such a case, which was reported widely in mainstream media in 2012, after two men successfully determined who their biological father was [167, 168, 169]. They estimated that their father, whose wife ran a controversial (at the time) fertility clinic in London, fathered another 300 to 600 children.

4.9.3 Ancestral diversity in the UK Biobank

Population structure among all UK Biobank participants reflects their world-wide ancestral backgrounds indicated by the self-reported ethnic background data (Figure 4.14). The PCs, however, provide researchers with a quantitative measure of ancestry that reveals structure among the relatively coarse categories (including 'Other/Unknown') of the ethnic background variable.

Structure at fine geographic scales is detectable using PCA, which, while not the most powerful method, has the advantage of scalability. The results of the PCA on the white British ancestry subset (Figure 4.19) are largely consistent with (although less sensitive than) the fineSTRUCTURE results for a study of the People of the British Isles (POBI) [42] (see Figure 4.31). The first PC (if we ignore Wales for a moment) mimics the split involving the two clades, 'Scotland/north England' and 'Central/south England', although the highest scores for PC 1 involve participants born in the regions of South and West Scotland, and northern Ireland, which broadly match the regions of the two clusters (green triangles and yellow circles in Figure 4.31) that appear further down the tree. There are also two branches of the *fineSTRUCTURE* hierarchical tree that separate north and south Wales (as with PCs 3 and 4); and the clade labelled 'Central/south England' involves a cluster located approximately in the region highlighted by PC3 (along with north Wales). A notable difference between these two

[13] These are 95% confidence intervals based on the standard error of the sample mean.

analyses is that the first split in *fineSTRUCTURE* hierachical tree separates individuals in Orkney from everyone else; whereas the Orkney-born samples in the UK Biobank are not differentiated in the PCs. This is probably due to smaller sample size in the region in UK Biobank, and/or the stricter sampling strategy employed by the POBI study, which required all four grandparents to have been born nearby [42]. As expected when using a more powerful method, the *fineSTRUCTURE* analysis detected other elements of structure, such as in south-west England, which PCA fails to detect.

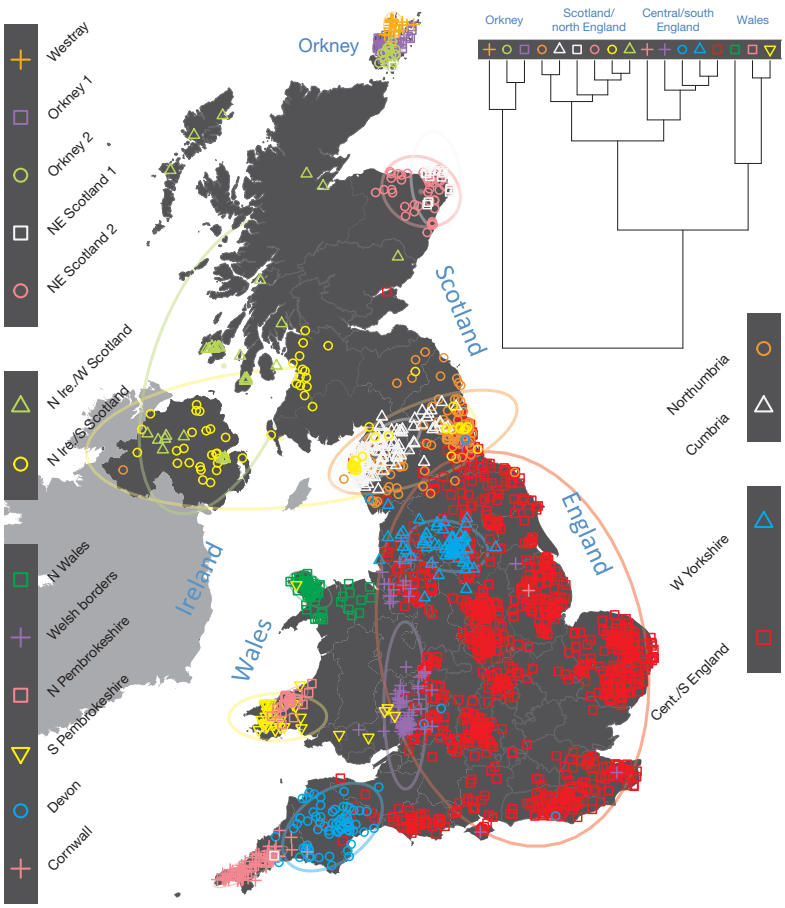


Figure 1 | Clustering of the 2,039 UK individuals into 17 clusters based only on genetic data. For each individual, the coloured symbol representing the genetic cluster to which the individual is assigned is plotted at the centroid of their grandparents' birthplaces. Cluster names are in side-bars and ellipses give an informal sense of the range of each cluster (see Methods). No relationship between clusters is implied by the colours/symbols. The tree (top right) depicts the order of the hierarchical merging of clusters (see Methods for the interpretation of branch lengths). Contains OS data © Crown copyright and database right 2012. © EuroGeographics for some administrative boundaries.

310 | NATURE | VOL 519 | 19 MARCH 2015

Figure 4.31: Fine-scale population structure among the people of the British Isles. This figure is reproduced from [42], and shows the results of a *fineSTRUCTURE* analysis of 2,039 British samples.

Chapter 5

Conclusion

We conclude this thesis by summarising our main findings and how they contribute to the wider literature. We also discuss some limitations of our analyses, and suggest some further lines of enquiry that are motivated by our results.

In Chapter 2 we set out to build a more comprehensive picture of population structure within Spain than what is found in the existing literature. By applying a powerful, haplotype-based method to a large and geographically dispersed data set we have been able to detect extensive population structure within Spain, much of which has not been previously characterised. We have also developed and applied new approaches to show how the structure has arisen from a complex and regionally-varying combination of genetic isolation and recent gene-flow within and from outside of Spain.

Overall, the pattern of genetic differentiation we observe in Spain reflects the linguistic and geopolitical patterns present around the twighlight of Muslim rule in Spain ($\sim$1300 CE). This suggests that the period has had a significant and long-term impact on the genetic structure observed in modern Spain, over 500 years later. We show, using the example of the Basques, how internal population movements from the north to the south during the *Reconquista* were probably the driving force of this structure. Other periods, such as the Roman and Visigothic eras appear to have been less far-reaching, although it was probably during these earlier times that the structure we see in northern

Spain already existed, if we consider the linguistic scholarship (Figure 1.5). In the case of the UK, similar geopolitical correspondence was seen, but to a different period in the past (around 800–950 CE) [42], suggesting that fine-scale structure evolves at different rates in different places, and that observed patterns may often reflect those at the ends of particular periods of significant upheaval, such as the end of the *Reconquista* in Spain, and the end of the Anglo-Saxon and Danish Viking invasions in the UK.

Two Iberian regions that the existing literature already pointed to as being genetically differentiated from the rest of Spain also stand out in our analysis. Namely, Galicia and País Vasco. We were also able to show that population structure exists *within* these regions, and to our knowledge, these results represent the finest scales over which such structure has yet been observed in humans. Our samples from País Vasco and Galicia also have very different genetic relationships with the rest of Europe and Portugal. We do not attempt to estimate 'divergence' times, but we interpret the results of our clustering and mixture-based analyses (Figure 3.2 and Figure 3.4) to mean that the excess of structure in Galicia has likely arisen recently (probably post-dating the Muslim conquest), and that the genetic isolation of Basque ancestors is likely to be much older phenomenon than in other regions (at least older than the Muslim conquest).

Our analysis also indicates that the genetic impact of the Muslim invasion of Iberia is likely to be the most significant of all the instances of migration into the peninsula in the last 2000 years (see Section 1.2.3), at least as far as our methods are able to detect. Taken together with what is known about the history of Muslim rule and settlement in Iberia (see Figure 1.3 and Figure 1.4), our results suggest that population movements subsequent to the conquest (e.g. from Portugal into Galicia) are likely to have been a significant factor in the regional variation of north-African-like DNA by the end of the 19th Century (Figure 3.5). It is also tempting to speculate that the region of surprisingly low north Moroccan ancestry that corresponds closely to the 14 and 15th-Century Crown of Aragon, reflects the efforts of the reigning monarchs to expel the Moriscos from this region. Both of these conclusions are consistent with those made by the most recent study of Iberia using Y-chromosome data [1].

According to our analysis using *GLOBETROTTER*, the genetic make-up of the north African-like migrants in the early half of Muslim rule is best represented as a mixture of modern-day Western Saharan and a smaller amount of sub-Saharan African DNA. However, the most prominent non-European group contributing to Iberia is north Morocco. We interpret this to mean that more recent gene-flow out of Iberia and into Morocco masks the original admixture event. What this means for the interpretation of the north Moroccan component of our Iberian ancestry profiles is not so clear, because we did not allow any Spanish individuals to be donors in that analysis. Any component of Spanish-like ancestry within the Moroccans would instead have to be accounted for by shared ancestry with the nearest proxy, such as French. Further analysis would be required to determine how much of the regional variation in ancestry shared with north Moroccans reflected mostly inward (or internal) rather than outward gene-flow.

We conducted our main analysis of Spanish population structure (Chapter 2) using data originally collected for a GWAS. These sources do not have the advantageous properties of a careful sampling scheme such as that used in the recent study of the fine scale structure of the British population [42], or fine-scale geographical information for all individuals in the sample. The fact that structure is identified and reflects geographical origins (where available) demonstrates that – at least in Spain – local population structure exists in the population as a whole, and general samples from the Spanish population, can be used to identify such structure.

Another aspect worth mentioning is our omission of middle-eastern and Jewish groups to the combined data set we analysed in Chapter 3. Two factors were at play here: the existence of publicly-available[1] data based on a genotyping array that has a sufficiently large overlap with the array used for the main Spanish cohort; and the type of self-reported ethnic background information available within such data sets.[2] Future studies of this kind, i.e. that make use of data from multiple sources, could potentially use

[1] By 'publicly-available' we mean readily accessible for research use.

[2] In the case of Jewish groups, individuals with Jewish ancestry are not readily identifiable from POPRES auxiliary data. Other studies have used genotype data for self-identified Jewish individuals based on the same (or sufficiently similar) genotyping array (see in Figure 1.10). However, to our knowledge, these data had not been made available for use by other researchers at the time of our analysis. Some genotype data for Qatari individuals was available, but based on a much smaller genotyping array ($\sim$250,000 markers) [2], so we omitted these from our analysis as it would have reduced the number of markers we could use by over a half.

imputation to help avoid this problem.

Nevertheless, if gene-flow directly from the middle-east had had a significant impact on genetic variation in Iberia, we would likely have observed ancestry-sharing with the next best proxy in our data: most likely Egypt, based on results from [2] (Figure 1.10). However, Egypt does not appear as a component any of our mixture-based analyses of Iberia, and components of north Moroccan and Western Saharan ancestry are the most dominant of non-European donor groups. This is either because gene-flow directly from the middle-east was too small to detect, or that middle-eastern-like ancestry in Iberia is entirely explained by shared ancestry with one of the Italian groups in our analysis (Italy1), which does have a significant component of Egyptian-like ancestry (see Figure 3.14).

DNA extracted from ancient human remains is increasingly becoming a source of insights into ancient migratory events, or into the ancestral origins of modern-day populations [77, 170, 171, 172]. The use of ancient DNA was beyond the scope of this thesis, but in future research such data may prove fruitful for better understanding Iberian population history. For example, DNA of ancient samples from north Africa dating to before the Muslim invasion would be a more direct source of information about the genetic make-up of the largely Berber migrants. In our analysis of population structure using modern-day samples the patterns we observe are dominated by major demographic events of the last two millennia. Analysing ancient samples from Iberia, such as was recently done for Portugal [172], might reveal genetic structure that may have existed prior to these major events, and which (if any) of the patterns we observe in modern-day Iberia are a result of pre-historic structure.

Our results for Spain indicate that fine-scale genetic structure is dependent on particular, local events during the recent past. More generally, it is becoming clear that fine-scale structure in human populations is highly prevalent, but also highly variable in strength and geographic scale (compare Switzerland to France in our analysis Figure 3.14, or compare different regions of the British Isles in Figure 4.31 [42]). Investigating these patterns illuminates details of our genetic history that would be difficult, or impossible, to infer otherwise. As more large collections of high-quality

whole-genome sequencing data emerge, we anticipate that even finer-scale differences will be revealed due to the increase in precision possible from directly typing rare mutations.

This brings us to the other focus of this thesis, the UK Biobank genetic data. Our results indicate that through rigorous experimental protocols applied to a large cohort and careful quality filtering, that genotype array technology can be used to successfully assay rare markers with MAF around 0.001 to 0.01 (see Section 4.5.3). Many very rare markers (< 0.001) are also likely to be well-assayed although results involving these markers should be examined carefully (e.g. Figure 4.13). Using this data we have also shown that population structure is present within the collection, even amongst a subset of individuals chosen for their relative ancestral homogeneity ('White British'). Furthermore, structure can be seen over small geographic scales (e.g. within Wales), even when using an analysis method (PCA) less sensitive than other methods, such as fineSTRUCTURE (Figure 4.19). In 2012 Mathieson and McVean [90] predicted that rare and common variants in association studies would be differentially confounded by population structure. They argued that rare variants are more likely to be associated with population structure that has arisen in the recent past, and over small geographic distances. Phenotypes that are also geographically stratified, but not genetic, may falsely appear associated with variants that are similarly geographically-localised. It is not yet clear how effective existing methods for correcting for population structure would be in the context of rare variation and fine-scale structure, or which kinds of phenotypes would be more susceptible to this problem. The UK Biobank data is an ideal source of data for research into this question, with its detailed geographic information and many other variables that are likely to exhibit some kind of systematic geographic variability (e.g. obesity or respiratory disease [173]). Haplotype-based methods for characterising subtle population structure are likely to be the way forward, provided they can be developed to be tractable for data sets of hundreds of thousands of individuals.

We found that familial relatedness is common among UK Biobank participants, with $\sim$30% having at least one first cousin relative or closer in the cohort. Furthermore, we

showed that familial relatedness is more common than would be expected if the UK Biobank sample was truly random, and it would be interesting to know whether this bias is a general phenomenon in large-scale, prospective cohort studies. The method we used to detect close relatives does not use haplotype information, so is limited in its ability to detect IBD-sharing that might occur between pairs of individuals more distantly related. Any spurious third-degree relatives could be isolated by testing for long contiguous chunks of DNA shared IBD; and more sophisticated haplotype-based methods, such as those used in 'long-range phasing', could be used to detect more distantly-related pairs [174].

The presence of so many close relatives is likely to be a common element of collections of this size sampled from a single country. There is also likely to be a wide range of degrees of kinship (e.g. second cousins) among UK Biobank participants, which we have not attempted to characterise (see above). GWAS methods that are designed to account for close relatedness [175] (or indeed leverage it [176]) will be important and valuable in this setting. Furthermore, the presence of over 1,000 trios, and other larger family groups in the UK Biobank presents an opportunity to study processes like recombination [177], or to test methods for phasing [178]. Finally, our GWAS for human height provides further evidence of the effectiveness of our QC, as well as highlighting the potential of the resource to uncover novel regions of association. We anticipate that the UK Biobank genetic data, and similar collections, will play a crucial role in the future of genomic research for population health and personalised medicine. The analyses in this thesis provide current and future users of the UK Biobank with a head-start in realising the opportunities for discovery that such data can offer.

Bibliography

[1] Adams, S.M., Bosch, E., Balaresque, P.L., Ballereau, S.J., Lee, A.C., Arroyo, E., *et al.* The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *Am J Hum Genet*, 2008. volume 83(6):725. ISSN 1537-6605 (Electronic) 0002-9297 (Linking). doi:10.1016/j.ajhg.2008.11.007. URL http://www.ncbi.nlm.nih.gov/pubmed/19061982.

[2] Botigue, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., *et al.* Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci U S A*, 2013. volume 110(29):11791. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi:10.1073/pnas.1306223110. URL http://www.ncbi.nlm.nih.gov/pubmed/23733930.

[3] Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., *et al.* Genetic structure of human populations. *Science*, 2002. volume 298(5602):2381. ISSN 1095-9203 (Electronic); 0036-8075 (Linking). doi:10.1126/science.1078311.

[4] Dermitzakis, E.T. Population Genetic Principles and Human Populations, pages 487–506. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-37654-5, 2010. doi:10.1007/978-3-540-37654-5_18. URL https://doi.org/10.1007/978-3-540-37654-5_18.

[5] Wakeley, J. Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, 2009. ISBN 0-9747077-5-9.

[6] Kingman, J.F.C. On the Genealogy of Large Populations. 1982. volume 19:27. doi:10.2307/3213548. URL http://www.jstor.org/stable/3213548.

[7] Hudson, R.R. Gene genealogies and the coalescent process. ID - 19910191040. *Oxford Surveys in Evolutionary Biology*, 1990. volume 7:1.

[8] Donnelly, P. and Leslie, S. The coalescent and its descendants. *ArXiv e-prints*, 2010.

[9] Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Richardsson, B., *et al.* A high-resolution recombination map of the human genome. *Nature Genetics*, 2002. volume 31(3):241. ISSN 1061-4036. doi:10.1038/ng917.

[10] McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. The fine-scale structure of recombination rate variation in the human genome.

Science, 2004. volume 304(5670):581. ISSN 1095-9203 (Electronic); 0036-8075 (Linking). doi:10.1126/science.1092500.

[11] Falush, D., Stephens, M., and Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 2003. volume 164(4):1567. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462648/.

[12] Saitou, Naruya; Nei, M. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, 1987.

[13] Varela, T.A., Farina, J., Dieguez, L.P., and Lodeiro, R. Gene flow and genetic structure in the Galician population (NW Spain) according to Alu insertions. *BMC Genet*, 2008. volume 9:79. ISSN 1471-2156 (Electronic) 1471-2156 (Linking). doi:10.1186/1471-2156-9-79. URL http://www.ncbi.nlm.nih.gov/pubmed/19055739.

[14] Calderon, R., Lodeiro, R., Varela, T.A., Farina, J., Ambrosio, B., Guitard, E., *et al.* GM and KM immunoglobulin allotypes in the Galician population: new insights into the peopling of the Iberian Peninsula. *BMC Genet*, 2007. volume 8:37. ISSN 1471-2156 (Electronic) 1471-2156 (Linking). doi:10.1186/1471-2156-8-37. URL http://www.ncbi.nlm.nih.gov/pubmed/17597520.

[15] Garcia-Obregon, S., Alfonso-Sanchez, M.A., Perez-Miranda, A.M., Vidales, C., Arroyo, D., and Pena, J.A. Genetic position of Valencia (Spain) in the Mediterranean basin according to Alu insertions. *Am J Hum Biol*, 2006. volume 18(2):187. ISSN 1042-0533 (Print) 1042-0533 (Linking). doi:10.1002/ajhb.20487. URL http://www.ncbi.nlm.nih.gov/pubmed/16493641.

[16] Calderón, R., Ambrosio, B., Guitard, E., González Martín, A., Aresti, U., and Dugoujon, J.M. Genetic Position of Andalusians from Huelva in Relation to Other European and North African Populations: A Study Based on GM and KM Allotypes. *Human Biology*, 2006. volume 78(6):663. ISSN 1534-6617. doi:10.1353/hub.2007.0008.

[17] Comas, D., Calafell, F., Benchemsi, N., Helal, A., Lefranc, G., Stoneking, M., *et al.* Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Human Genetics*, 2000. volume 107(4):312. ISSN 0340-6717 1432-1203. doi:10.1007/s004390000370.

[18] Wright, S. The genetical structure of populations. *Ann Eugen*, 1951. volume 15(4):323.

[19] Malécot, G. Les Mathématiques de l'Hérédié. Masson, Paris, 1948.

[20] Holsinger, K.E. and Weir, B.S. Genetics in geographically structured populations: defining, estimating and interpreting Fst. *Nat Rev Genet*, 2009. volume 10(9):639. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi:10.1038/nrg2611. URL http://www.ncbi.nlm.nih.gov/pubmed/19687804.

[21] Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J., and Feldman, M.W. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence

data. *Proceedings of the National Academy of Sciences of the United States of America*, 2000. volume 97(13):7360. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC16550/.

[22] Calafell, F. and Larmuseau, M.H.D. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Hum Genet*, 2017. volume 136(5):559. ISSN 1432-1203 (Electronic); 0340-6717 (Linking). doi:10.1007/s00439-016-1740-0.

[23] Capelli, C., Onofri, V., Brisighelli, F., Boschi, I., Scarnicci, F., Masullo, M., *et al.* Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe. *Eur J Hum Genet*, 2009. volume 17(6):848. ISSN 1476-5438 (Electronic) 1018-4813 (Linking). doi:10.1038/ejhg.2008.258. URL http://www.ncbi.nlm.nih.gov/pubmed/19156170.

[24] Calderon, R., Carrion, M., Perez-Miranda, A., Pena, J.A., Dugoujon, J.M., and Crouau-Roy, B. Allele Variation of DYS19 and Y- Alu Insertion (YAP) Polymorphisms in Basques: An Insight into the Peopling of Europe and the Mediterranean Region. *Human Biology*, 2003. volume 75(1):117. ISSN 1534-6617. doi:10.1353/hub.2003.0018.

[25] Flores, C., Maca-Meyer, N., Gonzalez, A.M., Oefner, P.J., Shen, P., Perez, J.A., *et al.* Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. *Eur J Hum Genet*, 2004. volume 12(10):855. ISSN 1018-4813 (Print) 1018-4813 (Linking). doi:10.1038/sj.ejhg.5201225. URL http://www.ncbi.nlm.nih.gov/pubmed/15280900.

[26] Brion, M., Quintans, B., Zarrabeitia, M., Gonzalez-Neira, A., Salas, A., Lareu, V., *et al.* Micro-geographical differentiation in Northern Iberia revealed by Y-chromosomal DNA analysis. *Gene*, 2004. volume 329:17. ISSN 0378-1119 (Print) 0378-1119 (Linking). doi:10.1016/j.gene.2003.12.035. URL http://www.ncbi.nlm.nih.gov/pubmed/15033525.

[27] Brion, M., Salas, A., Gonzalez-Neira, A., Lareu, M.V., and Carracedo, A. Insights into Iberian population origins through the construction of highly informative Y-chromosome haplotypes using biallelic markers, STRs, and the MSY1 minisatellite. *Am J Phys Anthropol*, 2003. volume 122(2):147. ISSN 0002-9483 (Print) 0002-9483 (Linking). doi:10.1002/ajpa.10231. URL http://www.ncbi.nlm.nih.gov/pubmed/12949835.

[28] Salas, A., Comas, D., Victoria Lareu, M., Bertranpetit, J., and Carracedo, A. mtDNA analysis of the Galician population: a genetic edge of European variation. *European Journal of Human Genetics*, 1998.

[29] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., *et al.* Genes mirror geography within Europe. *Nature*, 2008. volume 456(7218):98. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature07331. URL http://www.ncbi.nlm.nih.gov/pubmed/18758442.

[30] Fernandez-Rozadilla, C., Cazier, J.B., Tomlinson, I.P., Carvajal-Carmona, L.G., Palles, C., Lamas, M.J., *et al.* A colorectal cancer genome-wide association

study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics*, 2013. volume 14:55. ISSN 1471-2164 (Electronic) 1471-2164 (Linking). doi:10.1186/1471-2164-14-55. URL http://www.ncbi.nlm.nih.gov/pubmed/23350875.

[31] Henn, B.M., Botigue, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet*, 2012. volume 8(1):e1002397. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1002397. URL http://www.ncbi.nlm.nih.gov/pubmed/22253600.

[32] Pino-Yanes, M., Corrales, A., Basaldua, S., Hernandez, A., Guerra, L., Villar, J., *et al.* North African influences and potential bias in case-control association studies in the Spanish population. *PLoS One*, 2011. volume 6(3):e18389. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi:10.1371/journal.pone.0018389. URL http://www.ncbi.nlm.nih.gov/pubmed/21479138.

[33] Rodriguez-Ezpeleta, N., Alvarez-Busto, J., Imaz, L., Regueiro, M., Azcarate, M.N., Bilbao, R., *et al.* High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations. *Hum Genet*, 2010. volume 128(1):113. ISSN 1432-1203 (Electronic) 0340-6717 (Linking). doi:10.1007/s00439-010-0833-4. URL http://www.ncbi.nlm.nih.gov/pubmed/20443121.

[34] Gayan, J., Galan, J.J., Gonzalez-Perez, A., Saez, M.E., Martinez-Larrad, M.T., Zabena, C., *et al.* Genetic structure of the Spanish population. *BMC Genomics*, 2010. volume 11:326. ISSN 1471-2164 (Electronic) 1471-2164 (Linking). doi:10.1186/1471-2164-11-326. URL http://www.ncbi.nlm.nih.gov/pubmed/20500880.

[35] P Menozzi, A.P. and Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science*, 1978. volume 201(786-792).

[36] Bertranpetit, J.; Cavalli-Sforza, L.L. A genetic reconstruction of the history of the population of the Iberian Peninsula. *Annals of Human Genetics*, 1991. volume 55:51.

[37] Bertranpetit, J.; Cavalli-Sforza, L.L. Principal Component Analysis of Gene Frequencies and the Origin of Basques. *American Journal of physical anthropology*, 1994. volume 93.

[38] Patterson, N., Price, A.L., and Reich, D. Population Structure and Eigenanalysis. *PLOS Genetics*, 2006. volume 2(12):e190. URL https://doi.org/10.1371/journal.pgen.0020190.

[39] McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet*, 2009. volume 5(10):e1000686. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1000686. URL http://www.ncbi.nlm.nih.gov/pubmed/19834557.

[40] Pritchard, J.K., Stephens, M., and Donnelly, P. Inference of population structure

using multilocus genotype data. *Genetics*, 2000. volume 155(2):945. ISSN 0016-6731 (Print); 0016-6731 (Linking).

[41] Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet*, 2012. volume 8(1):e1002453. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1002453. URL http://www.ncbi.nlm.nih.gov/pubmed/22291602.

[42] Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., *et al.* The fine-scale genetic structure of the British population. *Nature*, 2015. volume 519(7543):309. URL http://dx.doi.org/10.1038/nature14230.

[43] Li, N. and Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 2003. volume 165(4):2213. ISSN 0016-6731 (Print); 0016-6731 (Linking).

[44] Delaneau, O., Marchini, J., and Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods*, 2012. volume 9(2):179. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/nmeth.1785. URL http://www.ncbi.nlm.nih.gov/pubmed/22138821.

[45] Alexander, D.H., Novembre, J., and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 2009. volume 19(9):1655. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi:10.1101/gr.094052.109. URL http://www.ncbi.nlm.nih.gov/pubmed/19648217.

[46] Hunter-Zinck, H., Musharoff, S., Salit, J., Al-Ali, K.A., Chouchane, L., Gohar, A., *et al.* Population genetic structure of the people of Qatar. *Am J Hum Genet*, 2010. volume 87(1):17. ISSN 1537-6605 (Electronic) 0002-9297 (Linking). doi:10.1016/j.ajhg.2010.05.018. URL http://www.ncbi.nlm.nih.gov/pubmed/20579625.

[47] Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet*, 2013. volume 9(11):e1003925. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1003925. URL http://www.ncbi.nlm.nih.gov/pubmed/24244192.

[48] Falush, D., van Dorp, L., and Lawson, D. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. *bioRxiv*, 2016. URL http://biorxiv.org/content/early/2016/07/28/066431.abstract.

[49] Evanno, G., Regnaut, S., and Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, 2005. volume 14(8):2611. ISSN 0962-1083 (Print); 0962-1083 (Linking). doi:10.1111/j.1365-294X.2005.02553.x.

[50] van Dorp, L., Balding, D., Myers, S., Pagani, L., Tyler-Smith, C., Bekele, E., *et al.* Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLoS*

Genet, 2015. volume 11(8):e1005397 EP . URL http://dx.doi.org/10.1371%2Fjournal.pgen.1005397.

[51] Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., *et al.* Ancient Admixture in Human History. *Genetics*, 2012. volume 192(3):1065. doi:10.1534/genetics.112.145037. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522152/.

[52] Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., *et al.* The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet*, 2011. volume 7(4):e1001373. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1001373. URL http://www.ncbi.nlm.nih.gov/pubmed/21533020.

[53] Hellenthal, G., Busby, G.B., Band, G., Wilson, J.F., Capelli, C., Falush, D., *et al.* A genetic atlas of human admixture history. *Science*, 2014. volume 343(6172):747. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi:10.1126/science.1243518. URL http://www.ncbi.nlm.nih.gov/pubmed/24531965.

[54] Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 2013. volume 193(4):1233. ISSN 1943-2631 (Electronic); 0016-6731 (Linking). doi:10.1534/genetics.112.147330.

[55] Instituto Nacional de Estadística. Electronic maps of modern-day Autonomous Communities, 2013. URL http://www.ine.es/ss/Satellite?L=0&c=Page&cid=1254735116596&p=1254735116596&pagename=ProductosYServicios%2FPYSLayout.

[56] Droysen, G. and Andree, R. Professor G. Droysens Allgemeiner historischer Handatlas : in sechsundneunzig Karten mit erläuterndem Text. Velhagen & Klasing, Bielefeld und Leipzig, 1886.

[57] Barton, S. A History of Spain. Palgrave Macmillan, 2009.

[58] Carr, R. Spain: a history. Oxford University Press, 2000.

[59] Pro Ruiz, J. and Rivero Rodríguez, M. Breve Atlas de Historia de España. Alianza Editorial, Madrid, 1999.

[60] D'Emilio, J. Culture and Society in Medieval Galicia : A Cultural Crossroads at the Edge of Europe. BRILL, Leiden, 2015. ISBN 9789004288607. URL http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=2110721.

[61] Ṭāhā, Abd al-Wāḥid Dhannūn. The Muslim conquest and settlement of North Africa and Spain. Routledge, London, 1989. ISBN 9780415004749.

[62] Glick, T.F. Islamic and Christian Spain in the early middle ages. Brill, Leiden, 2005. ISBN 9789004147713.

[63] Boone, J.L. and Benco, N.L. Islamic Settlement in North Africa and the Iberian Peninsula. 1999. volume 28:51. URL http://www.jstor.org/stable/223388.

[64] Safran, J.M. Defining boundaries in al-Andalus : Muslims, Christians, and Jews in Islamic Iberia. Cornell University Press, Ithaca, 2016. ISBN 9780801468018.

[65] Penny, R. Variation and Change in Spanish. Cambridge University Press, Cambridge, UNKNOWN, 2000. ISBN 9780511156175. URL http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=201369.

[66] Pharies, D.A. A brief history of the Spanish language. Chicago, second edition. edition, 2015. ISBN 9780226133942.

[67] Trend, J.B. The language and history of Spain. Hutchinson, London, 1953.

[68] Alvar, M. Enciclopedia lingüística hispánica. Consejo Superior de Investigaciones Científicas, Madrid, 1960.

[69] Baldinger, K. La formación de los dominios lingüísticos en la península ibérica. Gredos, Madrid, 1963.

[70] Ministerio de Instrucción Pública Y Bellas Artes. Anuario Estadístico de España: Año 1912, 1912. URL http://www.ine.es/inebaseweb/treeNavigation.do?tn=29307.

[71] Ruhlen, M. A guide to the world's languages. London : Edward Arnold, 1991.

[72] Corte-Real, H.B., Macaulay, V.A., Richards, M.B., Hariti, G., Issad, M.S., Cambon-Thomsen, A., *et al.* Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet*, 1996. volume 60(Pt 4):331. ISSN 0003-4800 (Print); 0003-4800 (Linking).

[73] Martínez-Cruz, B., Harmant, C., Platt, D.E., Haak, W., Manry, J., Ramos-Luis, E., *et al.* Evidence of Pre-Roman Tribal Genetic Structure in Basques from Uniparentally Inherited Markers. *Molecular Biology and Evolution*, 2012. volume 29(9):2211. URL http://mbe.oxfordjournals.org/content/29/9/2211.

[74] Aguirre, A., Vicario, A., Mazón, L.I., Estomba, A., Martínez de Pancorbo, M., Arrieta Picó, V., *et al.* Are the Basques a single and a unique population? *American Journal of Human Genetics*, 1991. volume 49(2):450. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1683283/.

[75] de Pancorbo, M., López-Martínez, M., Martínez-Bouzas, C., Castro, A., Fernández-Fernández, I., Antúnez de Mayolo, G., *et al.* The Basques according to polymorphic Alu insertions. *Human Genetics*, 2001. volume 109(2):224. ISSN 0340-6717. doi:10.1007/s004390100544. URL http://dx.doi.org/10.1007/s004390100544.

[76] Dupanloup, I., Schneider, S., and Excoffier, L. A simulated annealing approach to define the genetic structure of populations. *Mol Ecol*, 2002. volume 11(12):2571. ISSN 0962-1083 (Print); 0962-1083 (Linking).

[77] Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 2014. volume 513(7518):409. URL http://dx.doi.org/10.1038/nature13673.

[78] Busby, G.B.J., Hellenthal, G., Montinaro, F., Tofanelli, S., Bulayeva, K., Rudan, I., *et al.* The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Current Biology*, 2015. volume 25(19):2518. doi: http://dx.doi.org/10.1016/j.cub.2015.08.007. URL http://www.sciencedirect.com/science/article/pii/S0960982215009495.

[79] Dupanloup, I. and Bertorelle, G. Inferring Admixture Proportions from Molecular Data: Extension to Any Number of Parental Populations. *Molecular Biology and Evolution*, 2001. volume 18(4):672. URL http://mbe.oxfordjournals.org/content/18/4/672.abstracthttp://mbe.oxfordjournals.org/content/18/4/672.full.pdf.

[80] Plenge, R.M., Scolnick, E.M., and Altshuler, D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov*, 2013. volume 12(8):581. URL http://dx.doi.org/10.1038/nrd4051.

[81] Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*, 2017. volume 101(1):5. ISSN 1537-6605 (Electronic); 0002-9297 (Linking). doi:10.1016/j.ajhg.2017.06.005.

[82] Marchini, J. and Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 2010. volume 11(7):499. ISSN 1471-0064 (Electronic); 1471-0056 (Linking). doi:10.1038/nrg2796.

[83] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 2017. volume 45(Database issue):D896. doi: 10.1093/nar/gkw1133. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210590/.

[84] Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet*, 2015. volume 47(5):435. URL http://dx.doi.org/10.1038/ng.3247.

[85] Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol*, 2011. volume 40(6):1652. ISSN 1464-3685 (Electronic); 0300-5771 (Linking). doi:10.1093/ije/dyr120.

[86] Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, 2016. volume 70(Supplement C):214. doi:https://doi.org/10.1016/j.jclinepi.2015.09.016. URL http://www.sciencedirect.com/science/article/pii/S0895435615004448.

[87] Kvale, M.N., Hesselson, S., Hoffmann, T.J., Cao, Y., Chan, D., Connell, S., *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics*, 2015. volume 200(4):1051. ISSN 1943-2631 (Electronic); 0016-6731 (Linking). doi:10.1534/genetics.115.178905.

[88] Tung, J.Y., Do, C.B., Hinds, D.A., Kiefer, A.K., Macpherson, J.M., Chowdry, A.B., *et al.* Efficient Replication of over 180 Genetic Associations with Self-Reported Medical Data. *PLOS ONE*, 2011. volume 6(8):e23473. URL https://doi.org/10.1371/journal.pone.0023473.

[89] UK Biobank Approved Research, 2017. URL http://www.ukbiobank.ac.uk/approved-research/.

[90] Mathieson, I. and McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*, 2012. volume 44(3):243. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi:10.1038/ng.1074. URL http://www.ncbi.nlm.nih.gov/pubmed/22306651.

[91] Marchini, J., Cardon, L.R., Phillips, M.S., and Donnelly, P. The effects of human population structure on large genetic association studies. *Nat Genet*, 2004. volume 36(5):512. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi:10.1038/ng1337. URL http://www.ncbi.nlm.nih.gov/pubmed/15052271.

[92] Balding, D.J. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 2006. volume 7(10):781. ISSN 1471-0056 (Print) 1471-0056 (Linking). doi:10.1038/nrg1916. URL http://www.ncbi.nlm.nih.gov/pubmed/16983374.

[93] Voight, B.F. and Pritchard, J.K. Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLOS Genetics*, 2005. volume 1(3):e32. URL https://doi.org/10.1371/journal.pgen.0010032.

[94] UK Biobank. UK Biobank: Protocol for a large-scale prospective epidemiological resource. *Technical report*, UK Biobank, UK Biobank Coordinating Centre, 2007.

[95] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, 2015. volume 12. doi:10.1371/journal.pmed.1001779. URL http://dx.doi.org/10.1371/journal.pmed.1001779.

[96] Littlejohns, T.J., Sudlow, C., Allen, N.E., and Collins, R. UK Biobank: opportunities for cardiovascular research. *Eur Heart J*, 2017. ISSN 1522-9645 (Electronic); 0195-668X (Linking). doi:10.1093/eurheartj/ehx254.

[97] Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*, 2016. volume 19(11):1523. URL http://dx.doi.org/10.1038/nn.4393.

[98] UK Biobank Imaging Study. URL http://imaging.ukbiobank.ac.uk/.

[99] UK Biobank and Affymetrix. UK Biobank Axiom®Array - Content Summary. URL http://tools.thermofisher.com/content/sfs/brochures/uk_axiom_biobank_contentsummary_brochure.pdf.

[100] UK Biobank. Genotyping and quality control of UK Biobank, a large-scale,

extensivelt phenotyped prospective resource, 2015. URL http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf.

[101] 2017. URL http://www.ukbiobank.ac.uk/published-papers/.

[102] Young, A.I., Wauthier, F., and Donnelly, P. Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nature Communications*, 2016. volume 7:12724 EP . URL http://dx.doi.org/10.1038/ncomms12724.

[103] Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 2016. volume 167(5):1415. doi:http://dx.doi.org/10.1016/j.cell.2016.10.042. URL http://www.sciencedirect.com/science/article/pii/S0092867416314635.

[104] Warren, H.R., Evangelou, E., Cabrera, C.P., Gao, H., Ren, M., Mifsud, B., *et al.* Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat Genet*, 2017. volume 49(3):403. URL http://dx.doi.org/10.1038/ng.3768.

[105] Wain, L.V., Shrine, N., Artigas, M.S., Erzurumluoglu, A.M., Noyvert, B., Bossini-Castillo, L., *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet*, 2017. volume 49(3):416. URL http://dx.doi.org/10.1038/ng.3787.

[106] Lane, J.M., Liang, J., Vlasac, I., Anderson, S.G., Bechtold, D.A., Bowden, J., *et al.* Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat Genet*, 2017. volume 49(2):274. URL http://dx.doi.org/10.1038/ng.3749.

[107] Wain, L.V., Shrine, N., Miller, S., Jackson, V.E., Ntalla, I., Artigas, M.S., *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK biobank. *Lancet respir med*, 2015. volume 3. doi:10.1016/S2213-2600(15)00283-0. URL http://dx.doi.org/10.1016/S2213-2600(15)00283-0.

[108] Welsh, S. Genotyping of 500,000 participants: Description of sample processing workflow and preparation of DNA for genotyping, 2017. URL http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=807.

[109] Elliott, P. and Peakman, T.C. The UK biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int j epidemiol*, 2008. volume 37. doi:10.1093/ije/dym276. URL http://dx.doi.org/10.1093/ije/dym276.

[110] Welsh, S., Peakman, T., Sheard, S., and Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics*, 2017. volume 18(1):26. doi:10.1186/s12864-016-3391-x. URL http://dx.doi.org/10.1186/s12864-016-3391-x.

[111] The 1000 Genomes Project Consortium. A map of human genome variation

from population-scale sequencing. *Nature*, 2010. volume 467(7319):1061. URL http://dx.doi.org/10.1038/nature09534.

[112] Affymetrix. UKB_WCSGAX: UK Biobank 500K Samples Processing by the Affymetrix Research Services Laboratory. *Report*. URL http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=590.

[113] Affymetrix. UKB_WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix Research Services Laboratory. *Report*. URL http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=368.

[114] Affymetrix. Axiom®Genotyping Solution Data Analysis Guide. URL http://tools.thermofisher.com/content/sfs/manuals/axiom_genotyping_solution_analysis_guide.pdf.

[115] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., *et al.* Genome-wide genetic data on ˜500,000 UK Biobank participants. *bioRxiv*, 2017. URL http://biorxiv.org/content/early/2017/07/20/166298.abstract.

[116] Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*, 2008. volume 40(10):1253. URL http://dx.doi.org/10.1038/ng.237.

[117] Affymetrix. Affymetrix 6.0 annotation. URL http://www.affymetrix.com/support/technical/byproduct.affx?product=genomewidesnp_6.

[118] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium, Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C., *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 2011. volume 476(7359):214. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:10.1038/nature10251. URL http://www.ncbi.nlm.nih.gov/pubmed/21833088.

[119] Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 2010. volume 26(22):2867. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi:10.1093/bioinformatics/btq559. URL http://www.ncbi.nlm.nih.gov/pubmed/20926424.

[120] Band, G. and Marchini, J. qctool. URL http://www.well.ox.ac.uk/~gav/qctool/#overview.

[121] Pebesma, E.J. and Bivand, R.S. Classes and methods for spatial data in R. *R News*, 2005. volume 5(2):9. URL http://CRAN.R-project.org/doc/Rnews/.

[122] Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. Estimating and interpreting Fst: The impact of rare variants. *Genome Research*, 2013. volume 23(9):1514. doi:10.1101/gr.154831.113. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3759727/.

[123] Davison, D. Shellfish: Parallel PCA and data processing for genome-

wide SNP data. URL http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php.

[124] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., *et al.* PLINK: a toolset for whole genome association and population-based linkage analyses. *American Journal of Human Genetics*, 2007. volume 81.

[125] Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., *et al.* Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet*, 2008. volume 83(1):132. ISSN 1537-6605 (Electronic) 0002-9297 (Linking). doi:10.1016/j.ajhg.2008.06.005. URL http://www.ncbi.nlm.nih.gov/pubmed/18606306.

[126] Mackenzie, D. Encyclopedia of the Languages of Europe, chapter Galician. Blackwell Publishing, 2000.

[127] Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., *et al.* The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. *American Journal of Human Genetics*, 2008. volume 83(3):347. doi:10.1016/j.ajhg.2008.08.005. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556436/.

[128] Davison, A.C. and Hinkley, D.V. Bootstrap methods and their application. Cambridge University Press, 1997.

[129] Sánchez Pardo, J.C. Territorio y poblamiento en Galicia entre la antigüedad y la plena Edad Media : tesis doctoral. Ph.D. thesis, Universidade de Santiago de Compostela, 2008.

[130] Instituto Nacional de Estadística. Population and Housing Censuses 2011. URL http://www.ine.es/censos2011/tablas/Inicio.do.

[131] Varela, T.A., Ainsua, R.L., and Farina, J. Evolution of consanguinity in the Bishopric of Lugo (Spain) from 1900 to 1979. *Ann Hum Biol*, 2001. volume 28(5):575. ISSN 0301-4460 (Print); 0301-4460 (Linking).

[132] Varela, T.A., Aínsua, R.L., and Fariña, J. Consanguinity in the Bishopric of Ourense (Galicia, Spain) from 1900 to 1979. *Annals of Human Biology*, 2003. volume 30(4):419. doi:10.1080/0301446031000103301. URL http://dx.doi.org/10.1080/0301446031000103301.

[133] Perez-Miranda, A.M., Alfonso-Sanchez, M.A., Pena, J.A., and Calderon, R. HLA-DQA1 polymorphism in autochthonous Basques from Navarre (Spain): genetic position within European and Mediterranean scopes. *Tissue Antigens*, 2003. volume 61(6):465. ISSN 0001-2815 (Print); 0001-2815 (Linking).

[134] Calderón, R., Perez-Miranda, A., Peña, J., Vidales, C., Aresti, U., and Dugoujon, J. The genetic position of the autochthonous subpopulation of Northern Navarre (Spain) in relation to other basque subpopulations. A study based on GM and KM immunoglobulin allotypes. *Human biology*, 2000. volume 72(4):619. URL http://europepmc.org/abstract/MED/11048790.

[135] The International HapMap 3 Consortium. Integrating common and rare genetic

variation in diverse human populations. *Nature*, 2010. volume 467(7311):52. doi:10.1038/nature09298. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3173859/.

[136] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[137] MacGregor, N. Germany : memories of a nation. Penguin Books, UK, 2016. ISBN 9780141979786.

[138] Sazzini, M., Gnecchi Ruscone, G.A., Giuliani, C., Sarno, S., Quagliariello, A., De Fanti, S., *et al.* Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula. *Scientific Reports*, 2016. volume 6:32513. doi:10.1038/srep32513. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5007512/.

[139] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 2015. volume 526(7571):68. URL http://dx.doi.org/10.1038/nature15393.

[140] Busby, G.B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V.D., *et al.* Admixture into and within sub-Saharan Africa. *eLife*, 2016. volume 5:e15266. ISSN 2050-084X. doi:10.7554/eLife.15266. URL https://dx.doi.org/10.7554/eLife.15266.

[141] Guimerá, A. La emigración española a ultramar, 1492-1914. Fundación Española de Historia Moderna, 1991. URL http://digital.csic.es/handle/10261/79316.

[142] Meakin, A.M. Galicia: The Swizerland of Spain. https://archive.org/details/galiciaswitzerla00meakuoft, 1909.

[143] Pallares, M.C; Portela, E. Galicia en la epoca medieval. In Galicia historia (edited by M.d.M. Rodríguez Iglesias Francísco; Pérez Negreira), volume 2, pages 59–61. Hércules de Ediciones, 1998.

[144] Hernández-Borges, J. Inmigración extranjera y estructura demográfica en Galicia. *Xeográfica, Revista de Xeografía, Territorio e Medio Ambiente*, 2007. volume 7:137.

[145] UK Biobank. Touchscreen questionnaire ordering, validation and dependencies. URL http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=113241.

[146] Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet*, 2013. volume 14(10):681. URL http://dx.doi.org/10.1038/nrg3555.

[147] Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., and Balding, D.J. Reevaluation of SNP heritability in complex human traits. *Nat Genet*, 2017. volume 49(7):986. ISSN 1546-1718 (Electronic); 1061-4036 (Linking). doi:10.1038/ng.3865.

[148] Griffiths, A.J.F. An introduction to genetic analysis. Freeman, New York, 2000. ISBN 9780716735205; 9780716737711; 9780716735274.

[149] Mathur, R., Grundy, E., and Smeeth, L. Availability and use of UK based ethnicity data for health research. *Technical report*, National Centre for Research Methods Working Paper, http://eprints.ncrm.ac.uk/3040/1/Mathur-_Availability_and_use_of_UK_based_ethnicity_data_for_health_res_1.pdf, 2013.

[150] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 2006. volume 38(8):904. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi:10.1038/ng1847. URL http://www.ncbi.nlm.nih.gov/pubmed/16862161.

[151] Shibata, K., Hozawa, A., Tamiya, G., Ueki, M., Nakamura, T., Narimatsu, H., *et al.* The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies. *Molecular Genetics & Genomic Medicine*, 2013. volume 1(1):45. doi:10.1002/mgg3.4. URL http://dx.doi.org/10.1002/mgg3.4.

[152] Wigginton, J.E., Cutler, D.J., and Abecasis, G.R. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *American Journal of Human Genetics*, 2005. volume 76(5):887. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1199378/.

[153] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 2012. volume 491(7422):56. URL http://dx.doi.org/10.1038/nature11632.

[154] Bellenguez, C., Strange, A., Freeman, C., Wellcome Trust Case Control, C., Donnelly, P., and Spencer, C.C. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics*, 2012. volume 28(1):134. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi:10.1093/bioinformatics/btr599. URL http://www.ncbi.nlm.nih.gov/pubmed/22057162.

[155] Nielsen, J. and Wohlert, M. Chromosome abnormalities found among 34910 newborn children: results from a 13-year incidence study in Århus, Denmark. *Human Genetics*, 1991. volume 87(1):81. doi:10.1007/BF01213097. URL http://dx.doi.org/10.1007/BF01213097.

[156] Affymetrix. Affymetrix® Axiom CNV Summary Tool. Affymetrix, 2013.

[157] Dumanski, J.P., Rasi, C., Lönn, M., Davies, H., Ingelsson, M., Giedraitis, V., *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science*, 2015. volume 347(6217):81. URL http://science.sciencemag.org/content/347/6217/81.abstract.

[158] Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 2016. volume 536(7616):285. URL http://dx.doi.org/10.1038/nature19057.

[159] Galinsky, K.J., Bhatia, G., Loh, P.R., Georgiev, S., Mukherjee, S., Patterson,

N.J., *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ¡em¿ADH1B¡/em¿ in Europe and East Asia. *The American Journal of Human Genetics*, 2016. volume 98(3):456. doi:10.1016/j.ajhg.2015.12.022. URL http://dx.doi.org/10.1016/j.ajhg.2015.12.022.

[160] UK Biobank. Verbal Interview stage. URL http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=100235.

[161] Cliff, A.D. and Ord, J.K. Spatial processes - models & applications. Pion, London, 1981. ISBN 0850860814.

[162] Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with the General Population. *Am J Epidemiol*, 2017. ISSN 1476-6256 (Electronic); 0002-9262 (Linking). doi:10.1093/aje/kwx246.

[163] Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal*, 2006. volume Complex Systems:1695. URL http://igraph.org.

[164] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 2014. volume 42(D1):D1001. URL http://dx.doi.org/10.1093/nar/gkt1229.

[165] Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*, 2015. volume 47(3):284. URL http://dx.doi.org/10.1038/ng.3190.

[166] Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, 2014. volume 46(11):1173. URL http://dx.doi.org/10.1038/ng.3097.

[167] Nelsen, S.C. Bertold Wiesner, British Man 'Fathered Between 600 – 1,000 Children At Own Clinic', 2012. URL http://www.huffingtonpost.co.uk/2012/04/09/bertold-wiesner-british-man-fathered-between-600-1000-children-at-own-clinic_n_1411758.html.

[168] Smith, R. British man 'fathered 600 children' at own fertility clinic, 2012. URL http://www.telegraph.co.uk/news/9193014/British-man-fathered-600-children-at-own-fertility-clinic.html.

[169] Kreider, R. Did Sperm Bank Founder Father 600 Children?, 2012. URL http://abcnews.go.com/Blotter/sperm-bank-founder-father-600-children/story?id=16104054.

[170] Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 2015. volume advance online publication:. URL http://dx.doi.org/10.1038/nature14317.

[171] Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., *et al.* Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat Commun*, 2016. volume 7. URL http://dx.doi.org/10.1038/ncomms10408.

[172] Martiniano, R., Cassidy, L.M., Ó'Maoldúin, R., McLaughlin, R., Silva, N.M., Manco, L., *et al.* The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLOS Genetics*, 2017. volume 13(7):e1006852. URL https://doi.org/10.1371/journal.pgen.1006852.

[173] Ellis, A. and Fry, R. Regional health inequalities in England. *Office for National Statistics, UK*, 2010.

[174] Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, 2008. volume 40:1068 EP . URL http://dx.doi.org/10.1038/ng.216.

[175] Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 2014. volume 46:100 EP . URL http://dx.doi.org/10.1038/ng.2876.

[176] Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 2012. volume 488:471 EP . URL http://dx.doi.org/10.1038/nature11396.

[177] O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, 2014. volume 10(4):e1004234. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi:10.1371/journal.pgen.1004234. URL http://www.ncbi.nlm.nih.gov/pubmed/24743097.

[178] O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet*, 2016. volume 48(7):817. URL http://dx.doi.org/10.1038/ng.3583.

[179] Antonarakis, S.E. Human Genome Sequence and Variation, pages 31–53. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-37654-5, 2010. doi:10.1007/978-3-540-37654-5_3. URL https://doi.org/10.1007/978-3-540-37654-5_3.

[180] Feuk, L., Carson, A.R., and Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet*, 2006. volume 7(2):85. URL http://dx.doi.org/10.1038/nrg1767.

[181] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 2001. volume 409(6822):860. URL http://dx.doi.org/10.1038/35057062.

Appendix A

A.1 Types of genetic variation and techonologies for measuring them

Genetic variation (or polymorphism) in the human genome occurs in multiple forms. The most common form of variation is single nucleotide polymorphisms (SNPs), which occur (on average) about once every 1,000 base-pairs [179]. These are positions in the genome for which a single nucleotide base can differ between individuals, or between the two inherited copies within the same individual. Another type of variation is small regions (typically less than 1,000 bases long) which are found within some individuals' chromosomes and not others. These are known as short insertions and deletions (indels), and are common, but less common than SNPs. There are also other types of variants including larger genomic alterations (structural variants), such as inversions and copy-number variants [180].

Prior to the first sequence of the human genome [181] and the development of high-throughput technologies for genotyping genome wide, polymorphism in human DNA was measured by targeting specific locations in the genome using two main experimental approaches. One approach uses the fact that DNA polymorphisms can result in variable biological products that can be easily typed (e.g. the ABO blood types). The genotype an individual carries can be determined by an assay that targets the biological product, rather than the DNA itself. These are known as 'classical' genetic markers, and examples include markers of human leucocyte antigen (HLA) alleles and various types of blood groups. The other approach is to target polymorphism in DNA using restriction enzymes, which cut DNA wherever there is a specific, short sequence of bases (for example, in *Alu*-insertions). If an individual carries a mutation within one of these sequences, or doesn't carry the sequence, the enzyme can not cut there and so the resulting DNA fragments will be of different lengths to individuals that do not carry the mutation. Polymorphism in regions of repeating DNA sequences, known as 'tandem repeat loci' (or 'microsatellites'), can also be typed in this way.

Newer technology has since been developed to measure genetic variation at

hundreds of thousands of different loci at once. Data generated from this type of technology we refer to as genomic data, as it typically involves markers from right across the human genome. Oligonucleotide genotyping arrays (genotyping arrays) target genetic variation at pre-specified polymorphic loci (usually SNPs and indels) using DNA microarrays. Briefly, a set of oligonucleotide probes are placed in a regular grid on a microarray (or 'chip'). Each oligonucleotide probe is a stretch of single-stranded DNA that is an exact complement to one particular allele at a polymorphic site of interest, as well as some of the non-polymorphic flanking region. Fragments of DNA from a sample are washed over the microarray and the fragments containing a particular genetic locus will attach (or hybridise) more readily to the probes that match the particular allele present on the sequence. The hybridised DNA molecules are chemically tagged with a fluorescent marker and the amount of hybridisation is measured by the intensity of the fluorescence. The relative intensities associated with different alleles is then used to call the genotypes DNA samples bioinformatically. This genotyping technology is commercially available, and many different types of arrays have been designed to capture specific types of human variation.

Both classical genetic markers and genotyping arrays require pre-existing knowledge of human polymorphism, and only target allelic variation at specific sites on the genome. Whole-genome sequencing (WGS) technology aims to 'read' the whole sequence in a sample of DNA, resulting in an almost complete sequence of bases. This technology is also commercially available, but at a greater cost per sample than genotyping arrays. We mention it here for completeness, but the data used in the literature we survey is largely based on classical genetic markers or genotyping arrays; and for our main analyses we use genomic data from genotyping arrays designed by the company, Affymetrix[1].

[1] Affymetrix is now called Thermo Fisher Scientific. We will refer to it as Affymetrix throughout, as this was its commercial name at the time the data used in this thesis was generated.

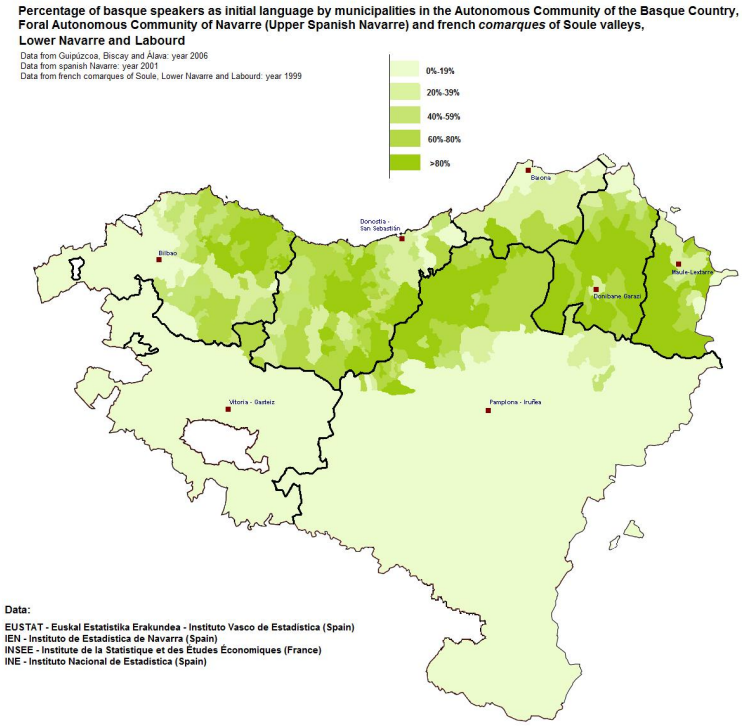


Figure A.1: Density of Basque language speakers in País Vasco (Basque Country). This map was compiled from official statistics data by Andrew Champs.

A.2 Selection of levels of the hierarchical tree in *fineSTRUCTURE* analysis (A)

In the *fineSTRUCTURE* analysis involving the Spanish cohort 145 clusters were inferred. To examine broad-scale structure we chose a level (14 clusters) such that all clusters were larger than size 20. Recall that moving up the tree, at each level a pair of clusters is merged. To examine finer-scale structure we traversed up from the bottom of the tree until the smaller of the two merging clusters contained at least 15 individuals. This occurs at the level of 49 clusters, and has the property that splits below this level only ever involve small numbers of individuals splitting from a similar cluster. However, at this level over half (28) of the clusters are within the clade involving individuals mostly from south-west Galicia, and for many of these clusters fine-scale geographic information was only available for one or two individuals, and/or the cluster contained fewer than 15 individuals. Therefore, in Figures 2.4 and 2.8a we only show the clusters at the higher level of the tree (level 14) for that clade.

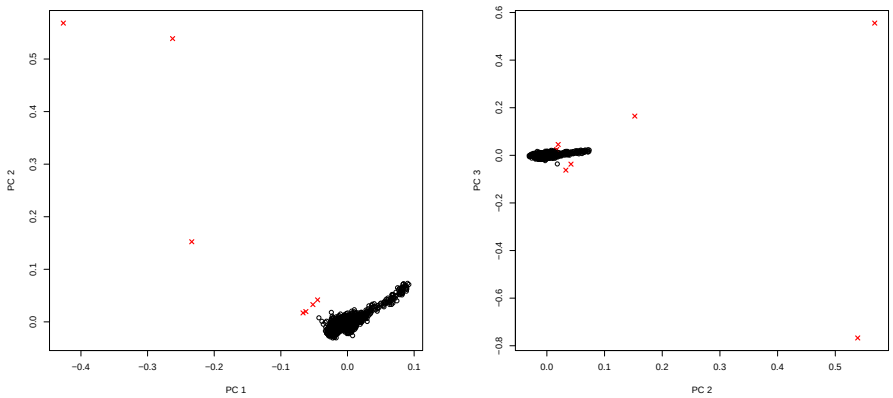


Figure A.2: PCA using genotype data showing outliers excluded in the main analyses. Excluded samples are indicated with red crosses.

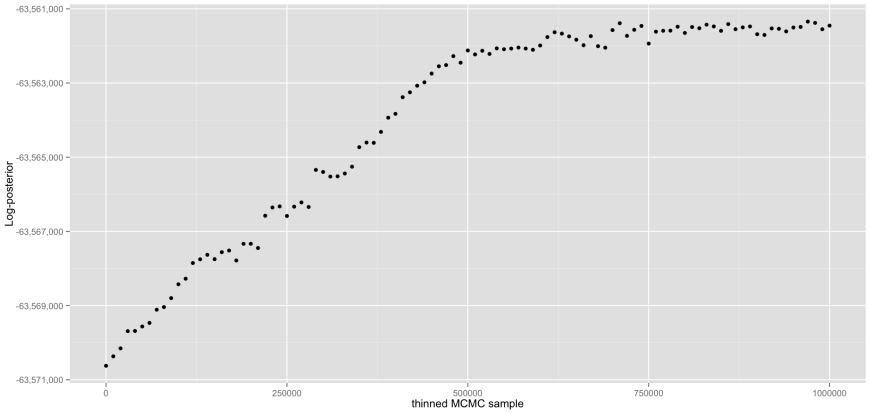


Figure A.3: Posterior probability (log) for each MCMC sample in *fineSTRUCTURE* analysis (A). We ran *fineSTRUCTURE* such that it stored a set of cluster assignments every 10,000 iterations of the MCMC after 500,000 burn-in iterations. This plot shows the log posterior probability of the cluster assignments at each MCMC sample. Because of the clear trend in increasing posterior probability in the first 50 MCMC samples, we only used the last 50 MCMC samples to refine the clusters assignments before the tree-building step of the *fineSTRUCTURE* algorithm (see Section 2.3).

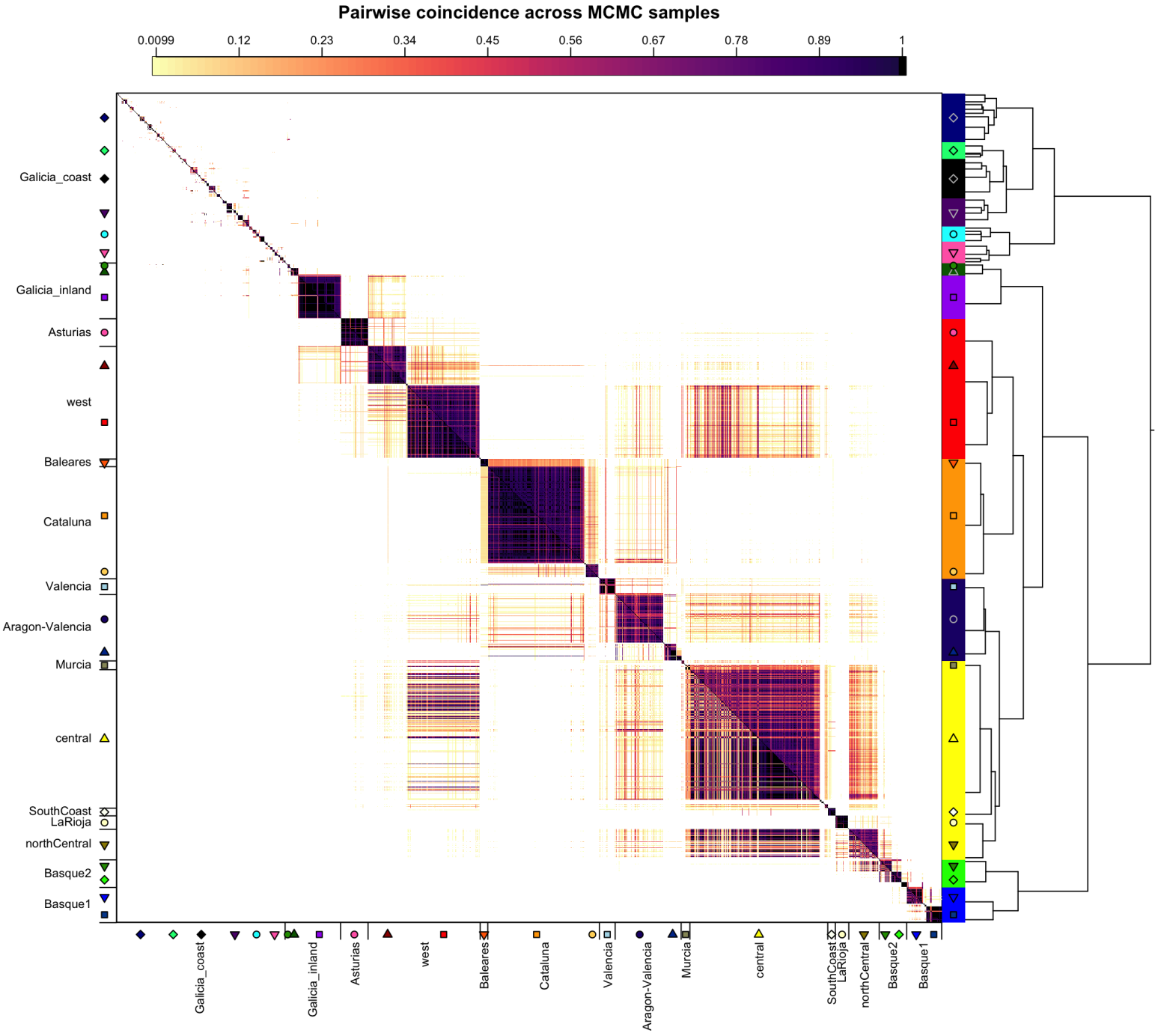


Figure A.4: Pairwise coincidence of cluster assignments for two independent *fineSTRUCTURE* runs. Each element of this matrix shows the fraction of times that a pair of individuals (row and column) are assigned to the same cluster across the set of MCMC samples used to construct the final set of clusters (see 2.3). White (0) indicates no coassignment across all MCMC samples, and black (1) indicates perfect coassignment across all MCMC samples. The upper triangle shows results for analysis (A), as discussed in this chapter, and the lower triangle shows results for an independent run of *fineSTRUCTURE* using exactly the same input data and parameters, but with different a random seed. The sample ordering and tree are from *fineSTRUCTURE* analysis (A), as shown in Figure 2.13. The clear similarity between the two runs indicates convergence of the MCMC samples to the posterior distribution.

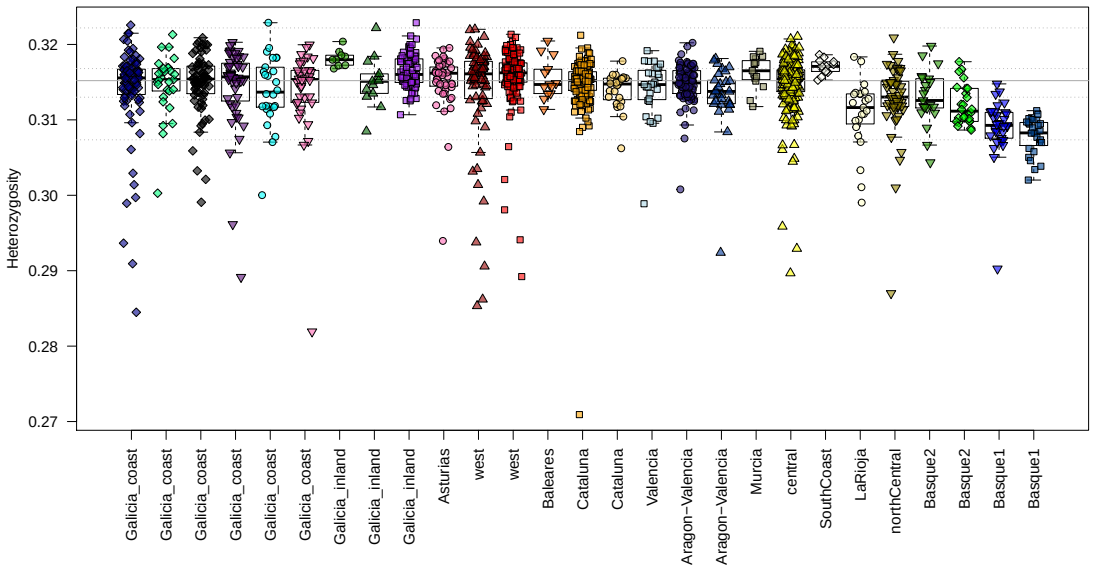


Figure A.5: Heterozygosity of Spanish individuals by their inferred Spanish cluster. Clusters are exactly those shown for the lower level of the tree in Figure 2.4. The horizontal gray lines across all the clusters show the distribution of heterozygosity for the whole cohort (median, quartiles, and the most extreme data point no more that 1.5 times the interquartile range).

Chromosome	Start position	End position
1	48000000	52000000
2	86000000	100500000
2	134500000	138000000
2	183000000	190000000
3	47500000	50000000
3	83500000	87000000
3	89000000	97500000
5	44500000	50500000
5	98000000	100500000
5	129000000	132000000
5	135500000	138500000
6	57000000	64000000
6	140000000	142500000
6	25500000	33500000
7	55000000	66000000
8	8000000	12000000
8	4300000	50000000
8	112000000	115000000
10	37000000	43000000
11	46000000	57000000
11	87500000	90500000
12	33000000	40000000
12	109500000	112000000
20	32000000	34500000

Table A.1: Regions of long-range LD excluded from principal components analyses. Regions were derived from [125] by D.P.

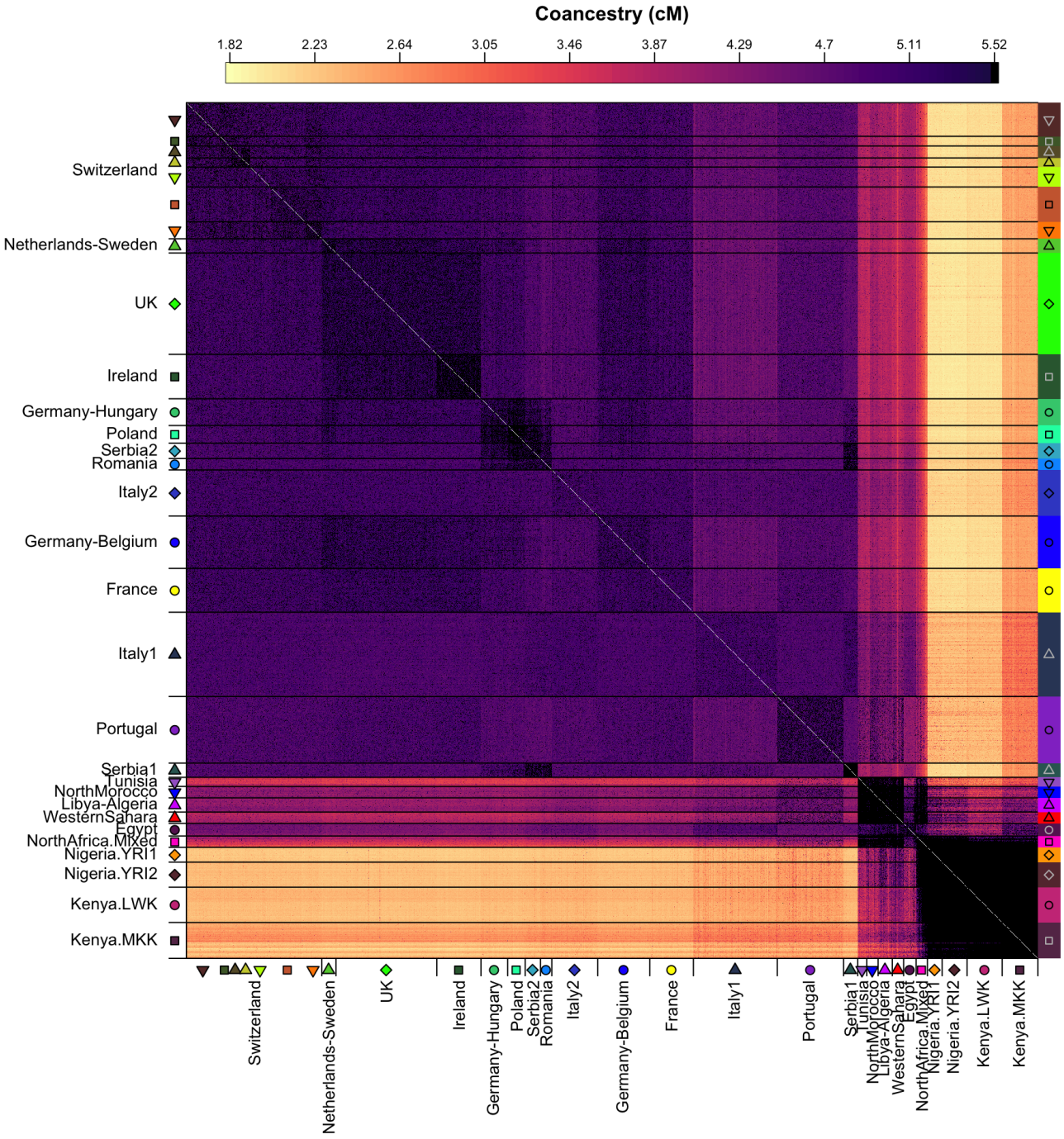


Figure A.6: Coancestry matrix among all non-Spanish individuals. Each row of the matrix is the coancestry vector for a non-Spanish individual, in a *CHROMOPAINTER* painting allowing all other individuals to be donors. That is, analysis (CI) as discussed in Section 3.3.1. However, the rows and columns are ordered according to the *fineSTRUCTURE* results of three separate analyses which we used to infer the donor groups (see Section 3.3.1). Labels were determined based on the locations of the majority of individuals in a given cluster, as shown in Figure 3.1. Where a cluster was split more evenly across two countries, a double-barrel name is used. In order to visualise the bulk of the variation, coancestry values equal to or above the 90th percentile (5.52 cM) are coloured black.

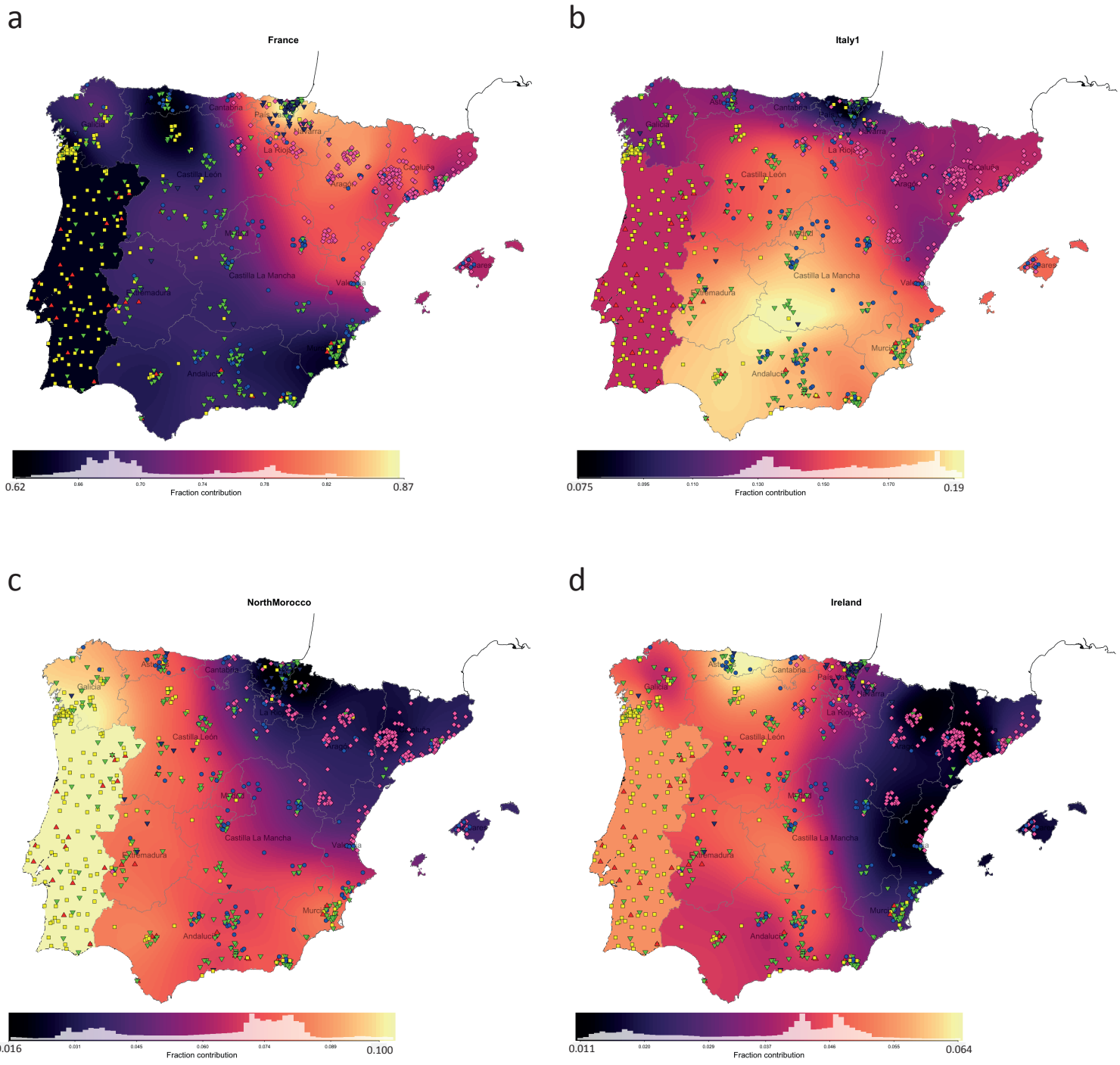


Figure A.7: Components of spatially-smoothed ancestry profiles for main genetic contributors to Iberia (continued on next page). A spatially-smoothed ancestry profile has been computed for each point on a spatial grid across Spain (see Section 3.5.2). Each map shows the fraction contributions from the stated donor group (e.g. 'NorthMorocco'). These seven groups (Figures a-g) are exactly the same contributors to the cluster-based ancestry profiles shown in Figure 3.4). Plots for the other three contributors are shown on the next page. Note the colour scale changes because the maximum contribution differs across donor groups. White areas indicate where the donor group contributed zero to the ancestry profile. Colours and symbols of the points indicate the same Iberian clusters as shown in Figure 3.2. The histogram on the scale bars shows the distribution of values across grid-points on the map (excluding zero).

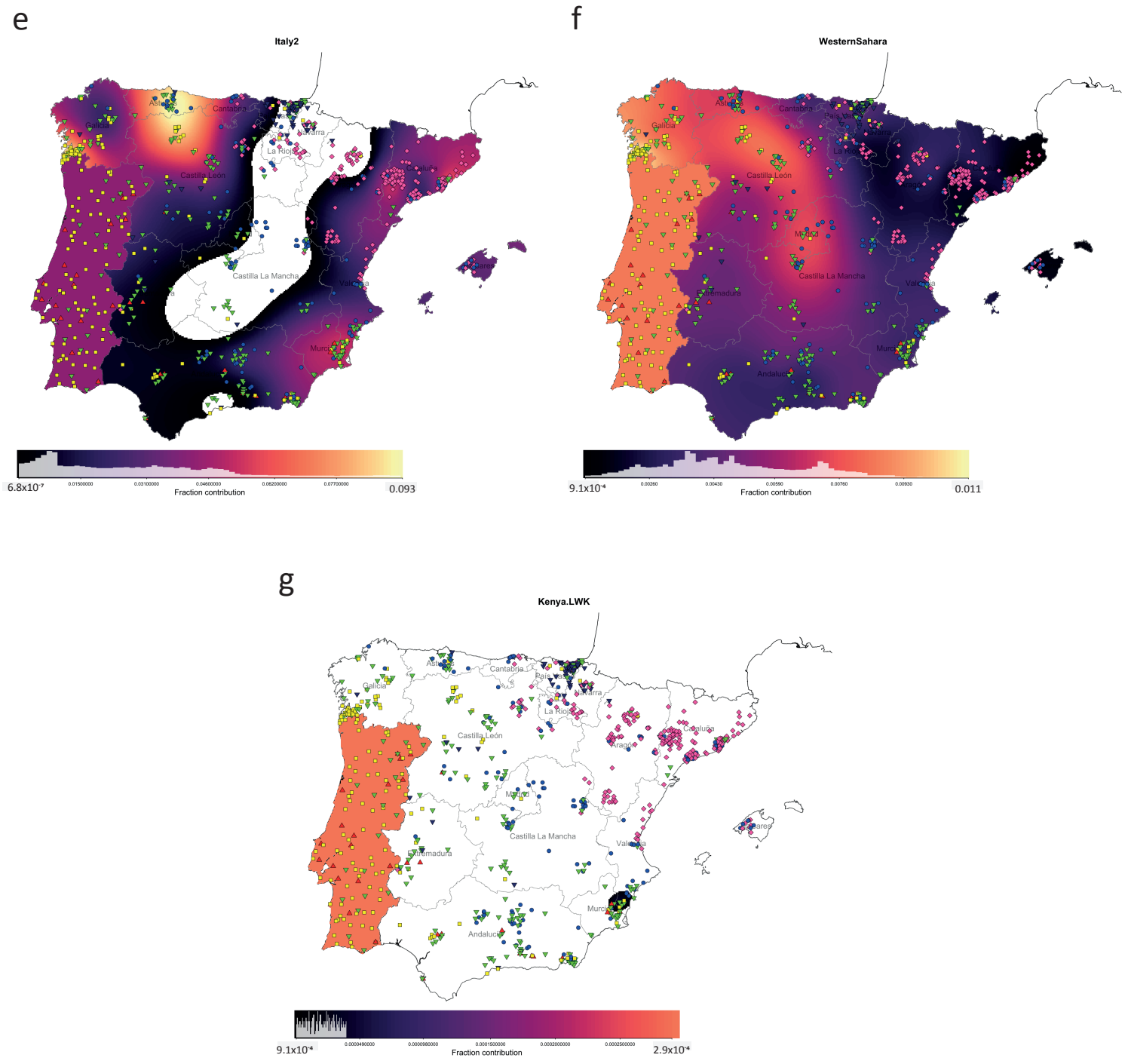


Figure A.7: continued from previous page.

	■ Galicia-Portugal	▲ Portugal-Andalucia	▼ west	● central	◆ Aragon-Cataluna	▼ Basque
France	0.6416 (0.6291 - 0.6566)	0.6296 (0.5838 - 0.6661)	0.6749 (0.6615 - 0.6895)	0.7001 (0.68 - 0.718)	0.7835 (0.7673 - 0.7984)	0.9089 (0.8838 - 0.9289)
Italy1	0.1501 (0.1422 - 0.1584)	0.1717 (0.1436 - 0.1979)	0.1665 (0.158 - 0.1748)	0.1589 (0.1485 - 0.1695)	0.1317 (0.1231 - 0.1405)	0.0477 (0.0347 - 0.0602)
NorthMorocco	0.1092 (0.1065 - 0.1117)	0.1006 (0.0837 - 0.1138)	0.0834 (0.0799 - 0.0858)	0.0596 (0.0568 - 0.0621)	0.0294 (0.0274 - 0.0314)	0 (0 - 0)
Ireland	0.0467 (0.0415 - 0.0511)	0.0398 (0.0252 - 0.0542)	0.0372 (0.0322 - 0.0419)	0.0392 (0.0331 - 0.0456)	0.0214 (0.0161 - 0.0271)	0.042 (0.0312 - 0.0554)
Italy2	0.0423 (0.0273 - 0.0558)	0.022 (0 - 0.0656)	0.03 (0.0149 - 0.0447)	0.0401 (0.0206 - 0.0598)	0.0308 (0.0152 - 0.0474)	0 (0 - 0.0157)
WesternSahara	0.0099 (0.0085 - 0.0112)	0.0068 (0.0025 - 0.0111)	0.0061 (0.005 - 0.0071)	0.0021 (0.0008 - 0.0033)	0.0029 (0.002 - 0.0038)	0.0012 (0.0001 - 0.0023)
Kenya.LWK	0 (0 - 0)	0.0018 (0.0005 - 0.0029)	0 (0 - 0)	0 (0 - 0)	0 (0 - 0)	0 (0 - 0)

Table A.2: Numeric values for ancestry profiles of Iberian clusters as visualised in Figure 3.4. See caption of Figure 3.4 for details.

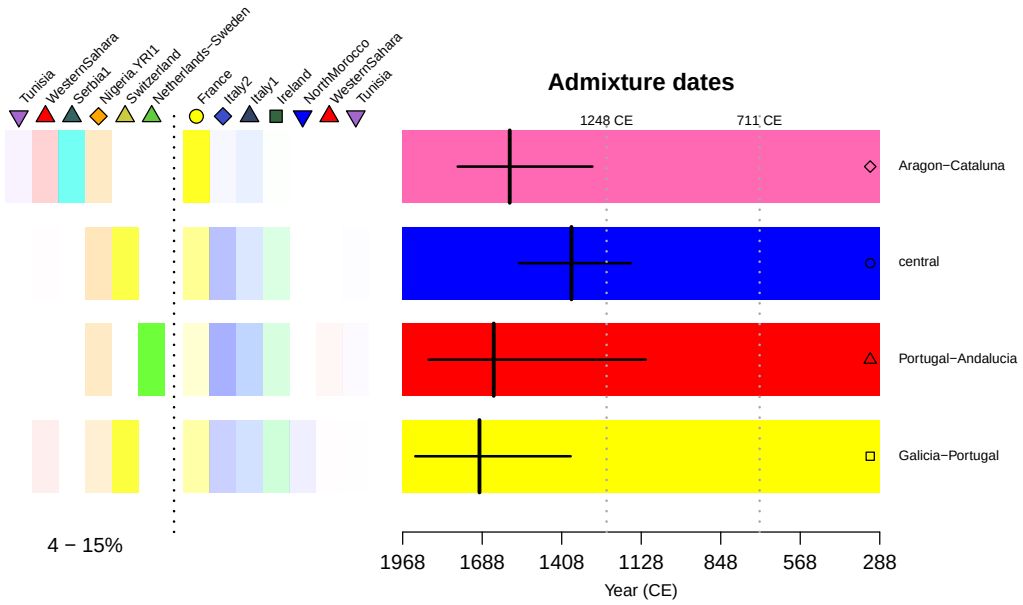


Figure A.8: *GLOBETROTTER* results for analysis (gtA) under a two-date admixture model. This figure is the same as in Figure 3.7a, except showing results for the more recent event under a two-date admixture model. This event involved mostly sub-Saharan-African donor groups on the non-European admixing side (as opposed to mainly north African groups). Only the target groups that showed some evidence of a two-date admixture event are shown. See Figure 3.13, and Section 3.6.5 for a fuller discussion.

(a)

Target population	Sample size	One date (generations)	One date (year)	Two dates date 1 (generations)	Two dates date 1 (year)	Two dates date 2 (generations)	Two dates date 2 (year)	fit.quality.1event	maxR2fit.1date	maxScore.2events
▼ Basque	100	40 (36-45)	861 (706-973)	N/A	N/A	N/A	N/A	0.992	0.978	0.186
◆ Aragon-Cataluna	100	39 (34-44)	869 (732-1018)	13 (3-35)	1591 (984-1873)	57 (45-103)	365 (922BC-714)	0.993	0.976	0.515
● central	100	36 (31-40)	957 (855-1088)	21 (15-29)	1374 (1164-1558)	67 (48-88)	86 (502BC-612)	0.996	0.978	0.481
▼ west	100	39 (37-43)	864 (760-942)	N/A	N/A	N/A	N/A	0.990	0.980	0.237
▲ Portugal-Andalucia	41	30 (23-37)	1119 (923-1336)	11 (7-24)	1647 (1300-1774)	45 (50-74)	696 (104BC-567)	0.996	0.962	0.564
□ Galicia-Portugal	100	37 (32-42)	929 (802-1085)	10 (3-35)	1698 (984-1873)	49 (45-103)	589 (922BC-714)	0.989	0.978	0.616

(b)

Target population	Sample size	One date (generations)	One date (year)	Two dates date 1 (generations)	Two dates date 1 (year)	Two dates date 2 (generations)	Two dates date 2 (year)	fit.quality.1event	maxR2fit.1date	maxScore.2events
◆ Basque2	26	21 (12-33)	1380 (1047-1639)	N/A	N/A	N/A	N/A	1.000	0.520	0.019
▼ Basque2	21	21 (15-29)	1375 (1162-1544)	N/A	N/A	N/A	N/A	1.000	0.597	0.030
▼ northCentral	52	23 (17-30)	1319 (1140-1486)	N/A	N/A	N/A	N/A	1.000	0.811	0.102
○ LaRioja	23	16 (11-23)	1514 (1332-1670)	N/A	N/A	N/A	N/A	1.000	0.677	0.013
◇ SouthCoast	13	21 (16-26)	1373 (1241-1507)	N/A	N/A	N/A	N/A	1.000	0.467	0.010
△ central	100	25 (21-28)	1276 (1183-1387)	N/A	N/A	N/A	N/A	1.000	0.894	0.003
■ Murcia	15	24 (16-32)	1304 (1061-1525)	N/A	N/A	N/A	N/A	1.000	0.468	0.008
▲ Aragon-Valencia	29	19 (14-24)	1431 (1283-1568)	N/A	N/A	N/A	N/A	1.000	0.691	0.013
● Aragon-Valencia	100	27 (22-31)	1202 (1087-1345)	N/A	N/A	N/A	N/A	1.000	0.838	0.037
□ Valencia	27	25 (16-33)	1273 (1035-1515)	N/A	N/A	N/A	N/A	1.000	0.593	0.012
○ Cataluna	24	25 (16-33)	1272 (1039-1511)	N/A	N/A	N/A	N/A	1.000	0.602	0.003
■ Cataluna	100	25 (21-29)	1281 (1167-1371)	N/A	N/A	N/A	N/A	1.000	0.889	0.016
▼ Baleares	13	19 (12-29)	1447 (1151-1636)	N/A	N/A	N/A	N/A	0.999	0.385	0.014
■ west	100	24 (20-28)	1299 (1195-1406)	N/A	N/A	N/A	N/A	1.000	0.863	0.012
▲ west	100	24 (20-30)	1285 (1132-1398)	N/A	N/A	N/A	N/A	1.000	0.834	0.036
○ Asturias	47	24 (19-29)	1303 (1156-1445)	N/A	N/A	N/A	N/A	1.000	0.763	0.008
■ Galicia_inland	100	20 (16-24)	1394 (1282-1517)	N/A	N/A	N/A	N/A	1.000	0.842	0.017
▲ Galicia_inland	13	28 (14-42)	1190 (786-1572)	N/A	N/A	N/A	N/A	1.000	0.275	0.027
● Galicia_inland	9	25 (12-50)	1272 (570-1637)	N/A	N/A	N/A	N/A	1.000	0.310	0.012
◆ Galicia_coast	100	22 (19-26)	1340 (1245-1447)	N/A	N/A	N/A	N/A	1.000	0.892	0.010
● Portugal	100	24 (21-29)	1300 (1149-1391)	N/A	N/A	N/A	N/A	1.000	0.849	0.038

Table A.3: Results of two GLOBETROTTER analyses. (a) Results for analysis gtA. **(b)** Results for analysis gtB. Refer to Section 3.6.1 and Table 3.1 for the details of each analysis. Unless otherwise stated, date estimates are reported in years CE and based on a generation time of 28 years. Date estimates for a two-date event are only shown for target populations where 'maxScore.2events' $>$ 0.35 (see Section 3.6.5 for details).

Batch	Batch size	Number of processed plates	Median plate size
UKBiLEVEAX_b1	4536	52	90.0
UKBiLEVEAX_b2	4545	53	90.0
UKBiLEVEAX_b3	4520	56	89.0
UKBiLEVEAX_b4	4542	52	89.0
UKBiLEVEAX_b5	4524	58	88.0
UKBiLEVEAX_b6	4524	60	89.0
UKBiLEVEAX_b7	4524	60	88.5
UKBiLEVEAX_b8	4551	53	90.0
UKBiLEVEAX_b9	4530	54	89.0
UKBiLEVEAX_b10	4559	58	88.0
UKBiLEVEAX_b11	4595	71	86.0
Batch_b001	4683	52	92.0
Batch_b002	4646	74	68.5
Batch_b003	4642	83	59.0
Batch_b004	4642	91	53.0
Batch_b005	4655	84	59.5
Batch_b006	4675	64	78.5
Batch_b007	4670	74	67.0
Batch_b008	4743	186	19.0
Batch_b009	4679	73	82.0
Batch_b010	4691	85	58.0
Batch_b011	4692	96	63.5
Batch_b012	4686	83	86.0
Batch_b013	4679	56	84.0
Batch_b014	4689	201	10.0
Batch_b015	4677	465	4.0
Batch_b016	4552	171	7.0
Batch_b017	4526	132	8.0
Batch_b018	4578	108	32.5
Batch_b019	4573	129	11.0
Batch_b020	4611	86	68.5
Batch_b021	4548	134	9.0
Batch_b022	4695	79	84.0
Batch_b023	4660	97	53.0
Batch_b024	4650	112	19.5
Batch_b025	4664	88	72.5
Batch_b026	4662	91	64.0
Batch_b027	4652	50	93.5
Batch_b028	4659	74	89.0
Batch_b029	4658	51	93.0
Batch_b030	4650	58	92.5
Batch_b031	4659	54	92.0
Batch_b032	4690	72	88.0
Batch_b033	4667	50	93.5
Batch_b034	4628	50	93.0
Batch_b035	4631	50	93.0
Batch_b036	4658	50	93.5
Batch_b037	4651	50	94.0
Batch_b038	4638	79	91.0
Batch_b039	4602	74	91.5
Batch_b040	4665	50	94.0

Continued on next page

Batch	Batch size	Number of processed plates	Median plate size
Batch_b041	4622	61	80.0
Batch_b042	4643	50	93.0
Batch_b043	4648	62	89.5
Batch_b044	4677	51	93.0
Batch_b045	4661	53	93.0
Batch_b046	4656	70	90.0
Batch_b047	4642	84	74.0
Batch_b048	4643	91	70.0
Batch_b049	4635	64	90.0
Batch_b050	4633	59	91.0
Batch_b051	4631	64	92.0
Batch_b052	4586	136	7.5
Batch_b053	4613	100	40.5
Batch_b054	4608	80	87.5
Batch_b055	4626	70	92.0
Batch_b056	4615	69	92.0
Batch_b057	4617	54	93.0
Batch_b058	4652	58	93.0
Batch_b059	4652	52	93.0
Batch_b060	4610	55	92.0
Batch_b061	4616	57	92.0
Batch_b062	4619	67	91.0
Batch_b063	4625	63	91.0
Batch_b064	4627	84	86.0
Batch_b065	4646	55	93.0
Batch_b066	4630	55	92.0
Batch_b067	4622	59	92.0
Batch_b068	4641	60	90.0
Batch_b069	4634	64	91.0
Batch_b070	4657	56	92.5
Batch_b071	4610	52	92.0
Batch_b072	4618	56	92.0
Batch_b073	4640	64	92.0
Batch_b074	4641	54	91.0
Batch_b075	4644	59	92.0
Batch_b076	4632	59	90.0
Batch_b077	4643	53	91.0
Batch_b078	4638	58	90.0
Batch_b079	4647	61	91.0
Batch_b080	4660	60	86.0
Batch_b081	4636	64	85.0
Batch_b082	4647	64	84.0
Batch_b083	4629	68	84.0
Batch_b084	4664	65	82.0
Batch_b085	4649	66	84.5
Batch_b086	4651	69	87.0
Batch_b087	4660	61	88.0
Batch_b088	4664	69	88.0
Batch_b089	4647	71	88.0
Batch_b090	4658	60	90.0
Batch_b091	4626	66	86.5
Batch_b092	4663	58	92.0

Continued on next page

Batch	Batch size	Number of processed plates	Median plate size
Batch_b093	4626	70	87.5
Batch_b094	2203	59	10.0
Batch_b095	4468	258	1.0
All batches	488377	5625	84.0

Table A.4: Number of UK Biobank participants genotyped within each batch. Intensities for each marker were measured on 96-well plates in groups of 94 UK Biobank samples and two control samples. The intensity data for multiple plates were combined to form batches of $\sim$4,700 UK Biobank samples, and genotypes were called in silico within each batch. In some cases samples from the same plate were genotyped in different batches, so the total number of unique plates is smaller than the sum of column 3, and the median plate size within each batch is often less than 94. Batches labelled with the prefix 'UKBiLEVEAX' contain only samples typed using the UK BiLEVE Axiom array, and those with the prefix 'Batch' contain only samples typed using the UK Biobank Axiom array. See [112] for full details of the sample handling process.

		UK BiLEVE Array only	UK Biobank Array only	Both arrays	Total
Included in experiment	Number of samples sent to Affymetrix (including duplicates)	50561	443568	0	494078
Included in data delivery from Affymetrix	Number of markers	18019	34313	760096	812428
	Number of samples (including duplicates)	50520	438692	0	489212
Included in released data	Number of markers	17536	34197	753693	805426
	Number of unique samples	49950	438427	0	488377

Table A.5: The number of markers and samples by genotyping array at main stages of the UK Biobank genotyping experiment. 'Data delivery from Affymetrix' refers to the data produced by Affymetrix after applying their filtering (see Section 1.3.3.3). 'Released data' refers to the genotype data made available to researchers after applying QC as described in Chapter 4.

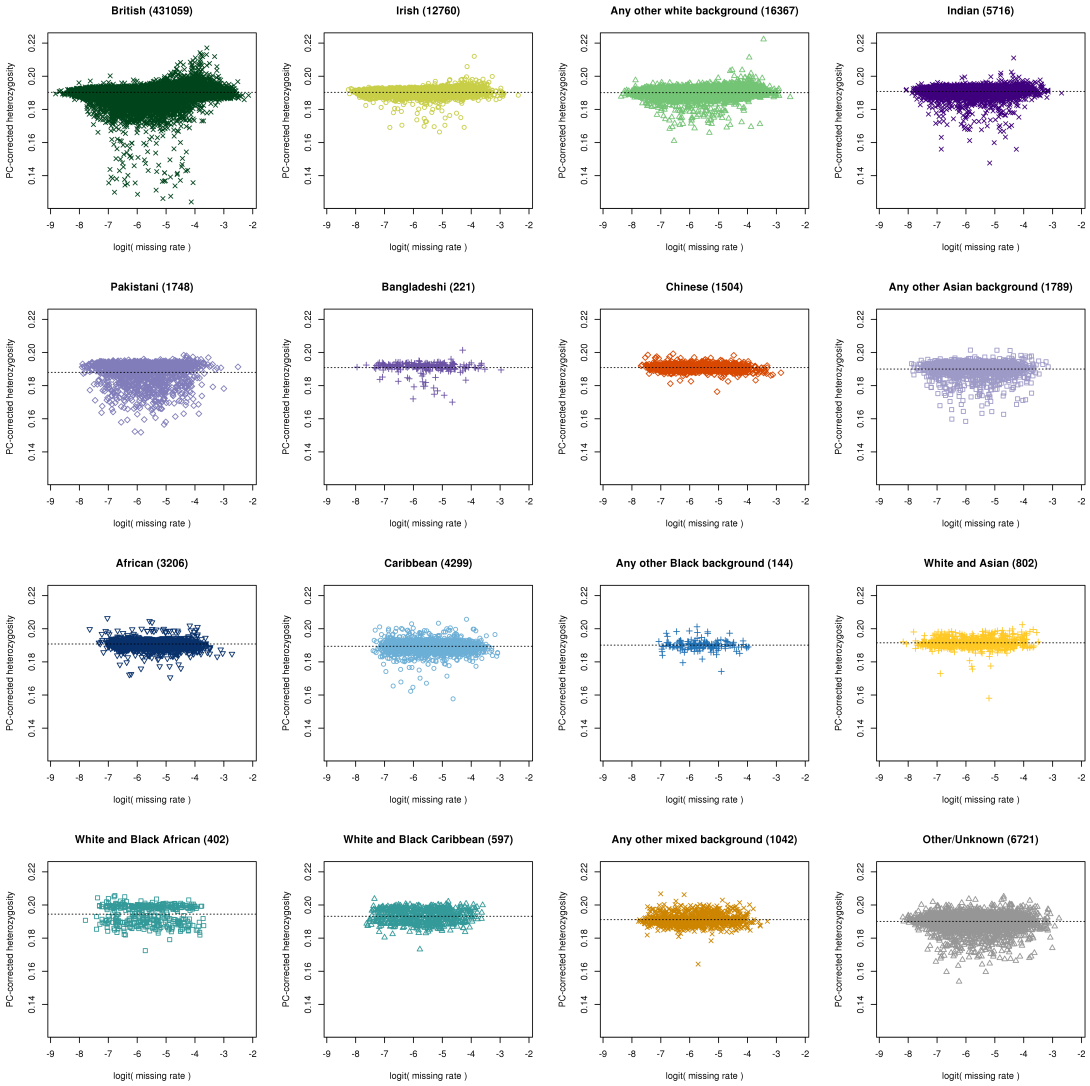


Figure A.9: PC-corrected heterozygosity and missing rates for different ethnic background categories. Horizontal lines show the mean value within each group. Groups with mixed ancestry (e.g. White and Black African) tend to have higher heterozygosity even after correcting for PCs. We therefore only included the largest ethnic background categories (shown in the top four plots) in the automated outlier detection process and for the other ethnic background categories we visually inspected these plots (see Section 4.4.1).

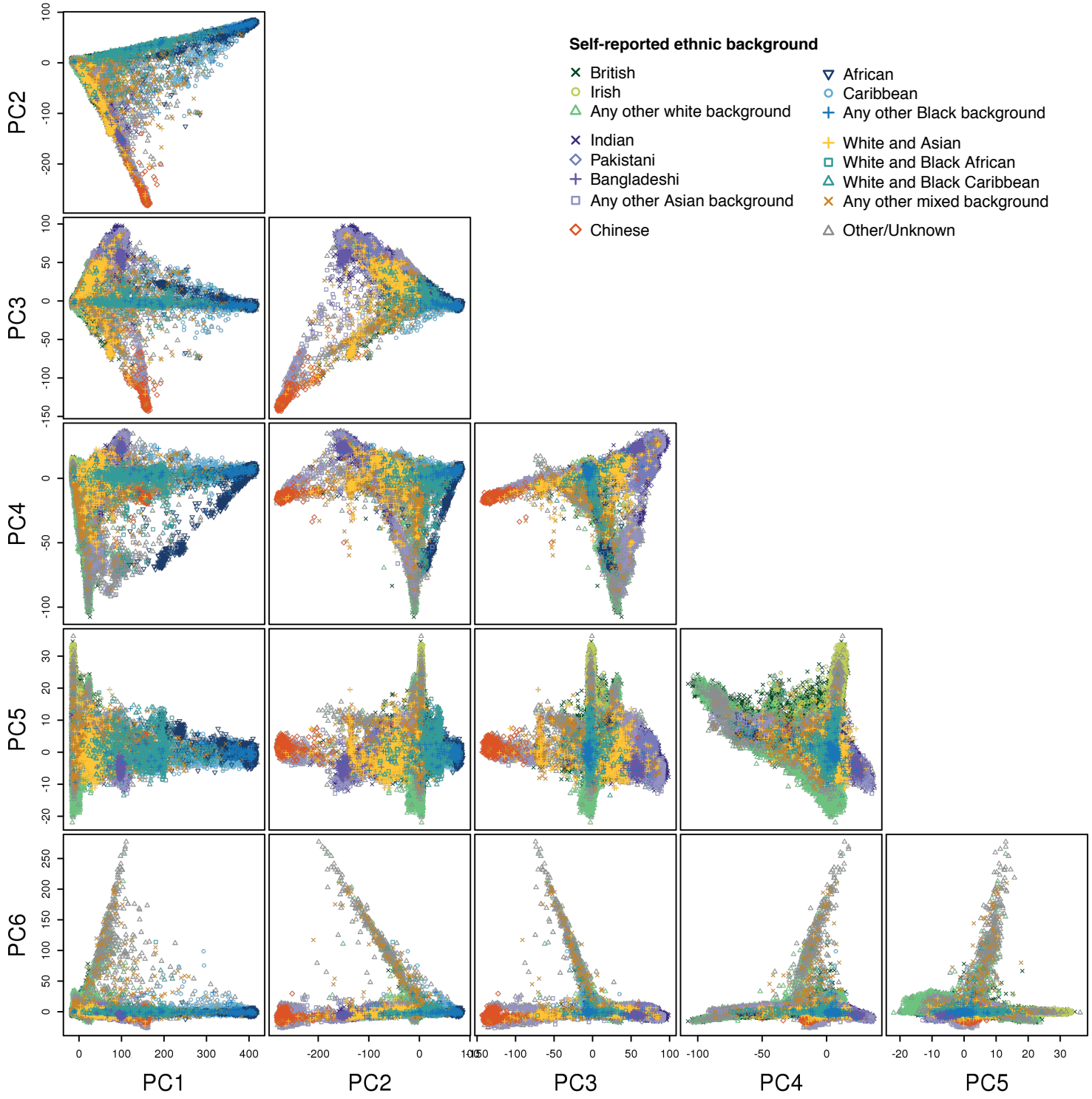
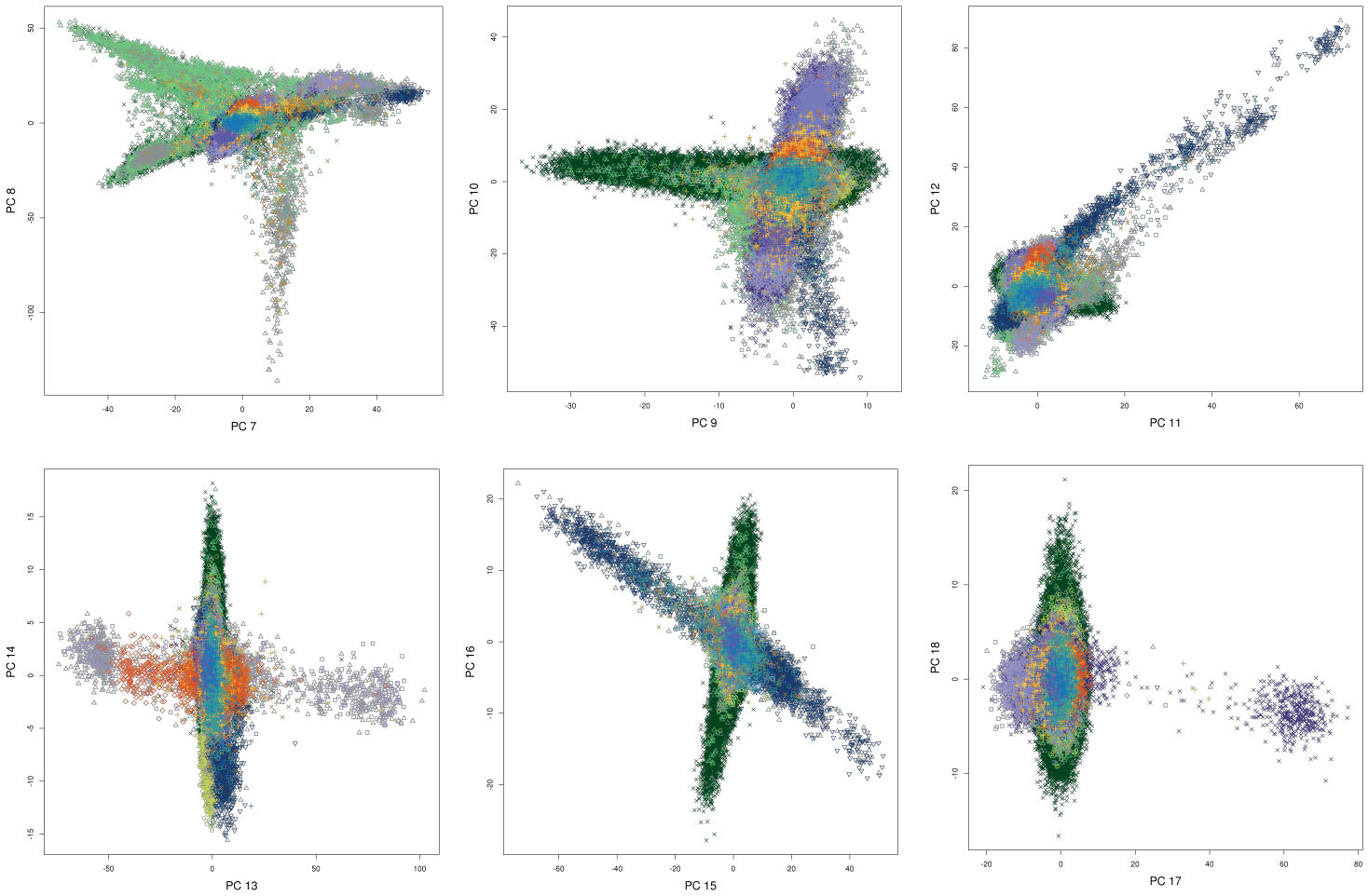


Figure A.10: All pairs of the first 6 principal components in PCA on UK Biobank genotype data. Each plot shows PC scores for UK Biobank samples for pairs of successive principal components. Each point represents a UK Biobank participant and is coloured according to their self-report ethnic background as defined in the key. Results for other PCs are visualised in Figures A.11 and 4.15.



Self-reported ethnic background

- British
- Irish
- Any other white background
- Indian
- Pakistani
- Bangladeshi
- Chinese
- Any other Asian background
- African
- Caribbean
- Any other Black background
- White and Asian
- White and Black African
- White and Black Caribbean
- Any other mixed background
- Other/Unknown

Figure A.11: Principal components 7-18 for UK Biobank genotype data. Each plot shows PC scores for UK Biobank samples for pairs of successive principal components from 7-18. Each point represents a UK Biobank participant and is coloured according to their self-report ethnic background as defined in the key. This figure shows results of the PCA for release (see ref. PCA methods section). Results for the first 6 PCs are visualised in Figure A.10; results for all 40 PCs are visualised in Figure 4.15.

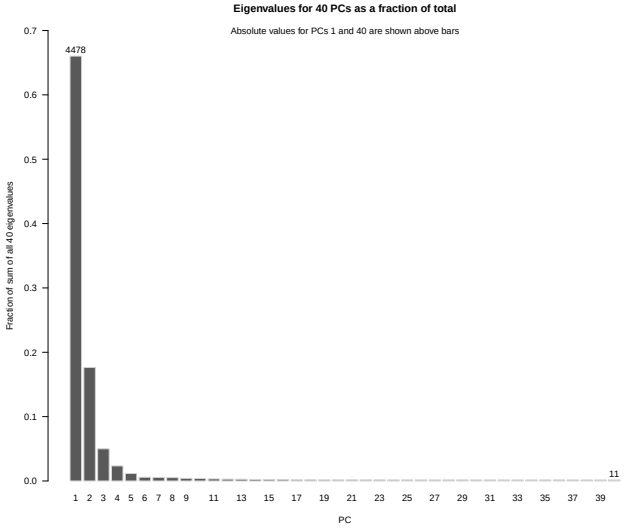


Figure A.12: Eigenvalues for 40 PCs in UK Biobank. Each bar shows the eigenvalue for each PC, as a fraction of the sum of all 40 eigenvalues. The larger the eigenvalue, the more variance is explained by the PC. The fraction of *all* the variance explained by each PC requires the sum of the eigenvalues of all PCs (ix.e. as many PCs as the number of individuals used in the computation). The trace of any square matrix is the sum of its eigenvalues, so this could be extracted from the matrix $X'X$, where X is the matrix of L genotypes for N individuals, mean-centred and variance-scaled [159]. The main computational efficiency that *fastPCA* uses is to avoid explicitly calculating this matrix, so it would require further analysis to calculate the sum of all eigenvalues. We have not done this analysis.

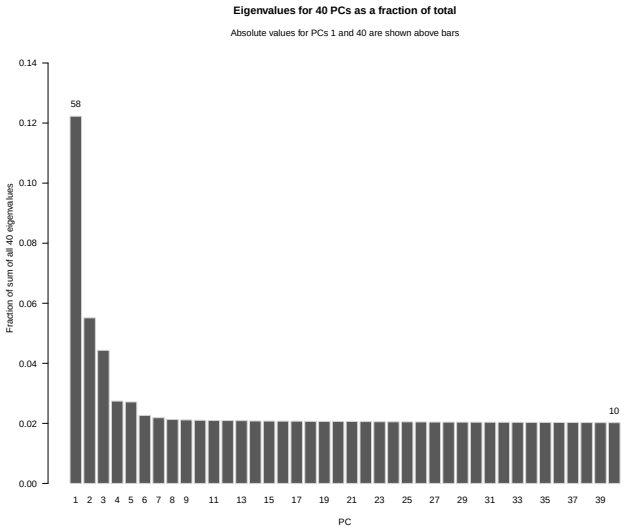


Figure A.13: Eigenvalues for 40 PCs in the white British ancestry subset. Each bar shows the eigenvalue for each PC, as a fraction of the sum of all 40 eigenvalues, exactly as shown in Figure A.12, but for PCs computed only using the white British ancestry subset.

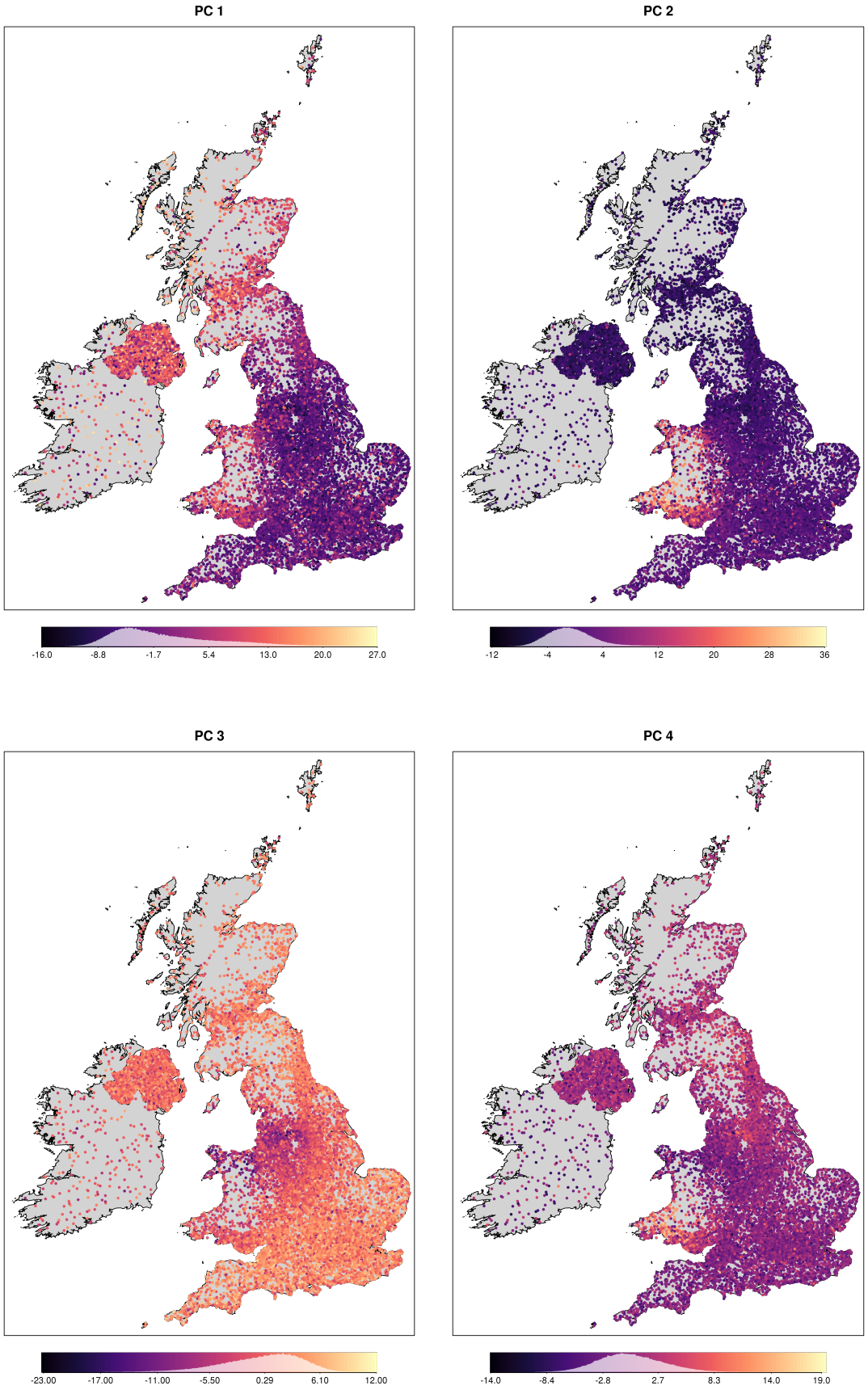


Figure A.14: Relationship between principal component scores (PCs 1-4) and place of birth for 395,231 UK Biobank participants. Each point represents a UK Biobank participant within the white British ancestry subset (see Section 4.6.3), and coloured according to their PC scores for each of the first 6 PCs. The histograms overlaid on a colour scale show the distribution of the scores for each PC.