Fast and efficient statistical methods for detecting genetic admixture events and its applications in large-scale data cohorts

PONGSAKORN WANGKUMHANG

A Thesis submitted for the degree of

Doctor of Philosophy

of

University College London.

UCL Genetics Institute, Department of Genetics, Evolution and Environment
University College London

March 26, 2020

I, Pongsakorn Wangkumhang, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Present-day cohorts of genome-wide DNA provide a powerful means of elucidating admixture events where different human groups intermixed, providing new insights into human history and population movements. The method GLOBETROTTER (Hellenthal et al., 2014) shows increased precision over other available techniques for characterising admixture due to modelling haplotype information, i.e. associations among tightly linked Single Nucleotide Polymorphisms (SNPs). However, because of its computational demands, GLOBETROTTER can only handle relatively small sample sizes of tens to hundreds of admixed individuals.

In this thesis, I present a new statistical method, fastGLOBETROTTER, that both reduces computational time and increases accuracy relative to GLOBETROTTER. In particular, fastGLOBETROTTER more efficiently models admixture linkage disequilibrium by sampling sets of genomic regions within individuals that are the most informative for admixture events. Additionally, I have developed an algorithm for allocating memory more efficiently to enable a factor of up to 20 fold improvement in computation time relative to GLOBETROTTER. Therefore, this technique can cope with the rapidly emerging large-scale cohorts of genetically homogeneous populations sampled from small geographic regions, e.g. within a country (China Kadoorie Biobank, UK Biobank), to provide more precise estimates of admixture dates. Via simulations, I use fastGLOBETROTTER to demonstrate the sample sizes required to characterize admixture between groups with high levels of genetic similarity, and the time depths for which these approaches can reliably detect such past intermixing.

I also apply fastGLOBETROTTER to over 6000 European individuals, using over 2500 individuals as ancestry surrogates, revealing new insights into admixture across Western Europe. These include admixture events dated to ∼500-600CE from sources carrying DNA related to present-day West Asian and North African populations found in individuals within France, Belgium and parts of Germany. I also report admixture from East-Asian/Siberian-like sources in individuals within Finland, Norway and Sweden at different times starting ∼1900 years ago.

Impact Statement

Identifying and describing past events where different populations intermixed (i.e. "admixture") is essential for understanding the processes leading to the present-day genetic diversity of humans and other organisms. While computational methods have been developed previously to infer admixture events, the currently most-powerful techniques to do so are unable to cope with the increasingly large-scale DNA datasets that are emerging.

In this thesis, I propose a new method that can infer the dates and components of admixture events more accurately and efficiently than currently available models. I develop a distributable software, fastGLOBETROTTER, that can handle large-scale data, e.g. decreasing computational time by a factor 20 relative to the most powerful existing approach while increasing precision. I provide a step-by-step tutorial to enable users from non-computational backgrounds to apply this software to their data, which I have already distributed to researchers in the field for on-going feedback. I also apply fastGLOBETROTTER to genome-wide genetic variation data from over 6000 European samples, unearthing new insights into the admixture history of parts of Europe.

This software will accelerate academic studies in several areas beyond population genetics. Inferring admixture can evaluate the genetic impact of well-established historical events (e.g. mass migrations, empires and armies) and unearth details about events that were largely unknown or hotly debated among researchers in other disciplines (anthropology, archaeology, linguistics). Understanding admixture is also essential for the design and analyses of studies to identify genetic loci associated with (historical or on-going) disease susceptibility, and how these loci have been distributed among human groups via historical intermixing. There are also several applications beyond academics. This includes forensics applications, through improved ability to identify the genetic make-up of individuals whose DNA has been found at crime scenes, particularly as some countries move towards storing information at $>$100K genome-wide genetic markers per individual in their crime databases. Admixture inference is also vital for the livestock industry, through understanding mixing of breeds, as well as in conservation efforts by identifying degrees of intermixing between endangered species and other animals. Finally, this work will also impact the views of a widespread audience of lay people interested in the history of humans and other organisms and/or their own genetic origins, e.g. as highlighted by the enormous popularity of ancestry testing companies.

Acknowledgements

The completion of my four-year PhD would never have been possible without support from many people whose names may not all be mentioned here, I truely appreciate all of you. However, I would like to thank the following people particularly those who have helped me carry out my research.

First I would like to express my sincere gratitude to my advisor, Garrett Hellenthal, who allowed me to join his research group. I will always be thankful for his invaluable guidance and advice to me as well as his patience to all our interactions. I also thank my fellow colleagues in Hellenthal group and our collaborations, particularly, Lucy van Dorp, Saioa López and Juan Camilo Chacón-Duque, their suggestions and feedback really helped improve my research. I am also grateful to Matthew Greenfield, a Master's student in our group, for sharing some of his results which I included in this thesis. I was lucky enough to be surrounded by great people at UCL Genetics Institute.

I would like to thank my thesis committee, Prof. Richard Mott from University College London and and Dr. Russell McLaughlin who traveled from the University of Dublin, Ireland, for their insightful discussion, comments and encouragement, and also for their tough questions which challenged me to widen my research perspectives.

I gratefully acknowledge the financial support received during my PhD studies from the Royal Thai Government Scholarship. My deep appreciation goes out to Dr. Sissades Tongsima, National Biobank of Thailand, National Science and Technology Development Agency, for his encouragement and initial support to undertake my PhD.

Many thanks to Aphiwat Luangsomboon and Nattika Nimmano who were always helpful with their supports throughout my dissertation. I thank all of my friends I have met here who opened their homes to me during my time in London. Thank you, Jomar, Jestoni, Cameron and all my badminton family and friends for always being so helpful in many ways.

Lastly, I would like to thank the Wangkumhang family: my parents, my brother, my sister and my two beautiful nieces, Nano and Nadia, for supporting me spiritually throughout my PhD journey and keeping me going. This is for all of you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent findings show that anatomically modern humans are believed to have arisen 315,000 years ago in Africa [1,2]. Following this, the spread of *Homo sapiens* out of Africa to the rest of the world first began around 270,000 years ago [3] with later migrations around 130,000-60,000 years ago [4–7]. Migrating people reached different continents at different times, and due to natural barriers, e.g. geography and climates, they separated, settled and progressed independent of one another. However, throughout the subsequent millennia, migrations, wars, slavery, colonization and other factors influenced intermingling between different populations that had previously been isolated from one another. Such interactions occasionally resulted in the exchange of cultural material and – potentially – intermixing among the different groups. Tracing human population history has been done by incorporating different types of evidence from e.g. anthropological, archaeological, linguistic, cultural and other historical materials. Recent advances in DNA sequencing and genotyping technology, combined with powerful new computational and statistical methods, has now made it possible to use DNA as an additional tool for inferring the history of human populations.

The human genome is made up of 23 pairs of chromosomes, with 22 pairs of autosomal chromosomes and one pair of sex chromosomes that differ between males and females (i.e. the X and Y chromosomes). Along the string of all 46 chromosomes, there are over 6 billion DNA base-pairs in total. This basepair sequence varies among individuals due to processes such as mutation, which results in a change in basepair sequence, and recombination briefly described below. The degree to which genetic variation, or genetic diversity, varies among individuals sampled from one or more populations can be influenced by an array of factors, such as changes in population size, the exchange of genetic material among groups, and natural selection whereby some genetic variants confer an advantage in a particular environment [8, 9]. Population geneticists pay most attention to levels and patterns of genetic diversity that differ between populations, which can lead to a better understanding of human population history.

When two humans reproduce, their offspring inherits one chromosome from each parent within each of the 23 pairs, i.e. a paternal copy and maternal copy from the father and mother, respectively. For each autosome and the maternally-inherited X-chromosome, the chromosome that the offspring inherits from each parent is often a mixture of the DNA that parent inherited from their own parents. This is due to a process known as recombination that – within each chromosome pair – shuffles the DNA between each parent's two chromosomes during meiosis. As a result of this recombination process, genetic regions located far apart along a chromosome can come from different ancestral sources (i.e. different grandparents) within each offspring. At a population level, recombination ensures that genetic variants located far apart along a chromosome can be uncorrelated among unrelated individuals sampled from a population. Alternatively, genetic variants located nearby can be correlated, because relatively little historical recombination has occurred between them. The correlation among nearby SNPs or other genetic markers is referred to as "linkage disequilibrium" (LD), and leads to nearby variants in the genome being "linked" and carrying correlated (or non-independent) information.

In order to learn about population demography, many initial studies analysed mitochondrial DNA (mtDNA) and Y chromosome data (specifically the non-recombining part of the Y-chromosome, or NRY), which was relatively easier to acquire and analyse [10, 11]. However, mtDNA and NRY provide information from only a single linked locus, severely limiting the power of what they can reveal about population history [12]. Subsequent decreases in the cost of capturing whole genome autosomal data, either via sequencing or by targeting specific genetic markers, has made autosomal data increasingly popular for exploring human history. In contrast to mtDNA and NRY, each autosome undergoes recombination and hence can be comprised of many thousands of independent loci inherited from different ancestors, potentially increasing the precision of inference. Of particular prominence is the analysis of biallelic Single Nucleotide Polymorphisms (SNPs), which are the most frequent type of genetic variation, located across the 22 autosomes.

This thesis will focus on using genome-wide autosomal SNP data to shed light on a particular aspect of population history inference: identifying and characterizing events where different groups of individuals intermixed, which is known as genetic admixture.

1.1 Genetic Admixture

Genetic admixture occurs when two or more genetically distinct populations intermix and form an admixed individual. In the autosomes and X-chromosome, recombination shuffles the chromosome segments of admixed individuals over the subsequent generations, such that the chromosomes of the descendants of the admixture event are comprised of a mosaic of DNA inherited from the ancestral sources [13, 14].

One particular model of admixture is known as the "pulse model" [15–17]. An example of the simplest type of a pulse model of admixture is provided in Figure 1.1, where there is an admixture between two groups A (red) and B (yellow), λ generations prior to sampling. Under this model, individuals from the admixed population randomly mate for these λ generations. Following this, and assuming no crossover interference, within the admixed individuals the boundaries between segments of DNA inherited intact from one of the original admixing sources follow a Poisson process with rate λ per Morgan, as it describes the occurrences of a discrete event (crossover in this case) that appear to happen at a certain rate (λ per Morgan). This model enables means of leveraging genetic patterns in the admixed

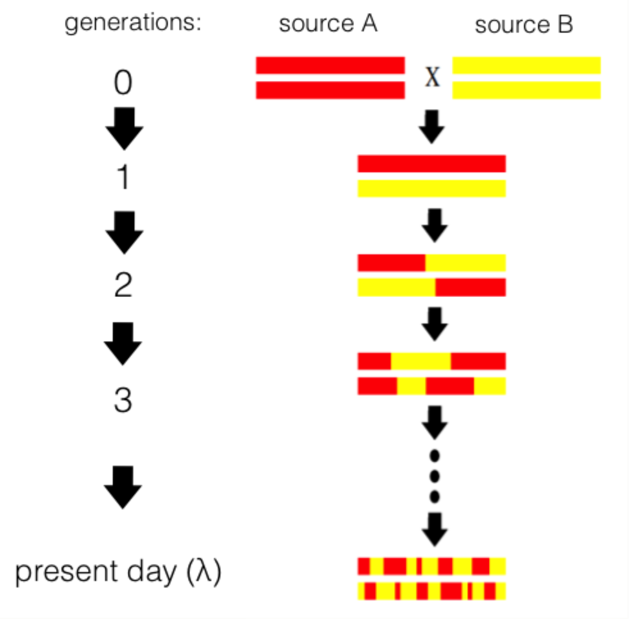

Figure 1.1: Schematic representation of an admixture event when sources *A* (red) and *B* (yellow) mix λ generations ago. The contiguous DNA segments or "chunks" inherited from each ancestry become shorter as the number of generations increases, due to recombination.

individuals to infer admixture times and proportions [16, 18–21].

As an example, Hellenthal et al., 2014 [21] simulated admixture between two sources under the pulse model by following the procedure outlined in Price et al., 2015 [22]. In particular, among these simulations are two sets of 20 simulated individuals each that descend from a single admixture event between two sources occurring 30 generations ago. The sources are African and European in the first set, contributing 80% and 20% of the DNA, respectively, while the sources are Central-South-Asian and European in the second set, each contributing 50% of the DNA. These simulations used DNA from 21 present-day Yoruban individuals from Nigeria, 21 present-day Brahui individuals from Pakistan and 28 present-day individuals from France as the African, Central-South-Asian and European admixing sources, respectively, in order to generate the simulated admixed individuals. Here we assume these admixing populations were not sampled, a likely reflection of reality for at least older admixture. Instead individuals from other populations are used as surrogates for each of these continental admixing sources; in particular here we use the genomes of 22 Mandenka from Africa, 21 Balochi from Central South Asia, and 23

British/Irish from Europe to represent the Yoruba, Brahui and French, respectively. Throughout this chapter, as an illustration, I apply various published methods to these simulated examples.

1.2 Signatures of admixture

Several techniques have been developed to detect population structure, i.e. to quantify the level of heterogeneity among sampled individuals. This genetic structure can be affected by demographic processes such as migration, drift, and admixture between populations [23]. Hence patterns of population structure potentially can shed light on these historical processes, including admixture. Two of the most widely-used approaches for describing population structure are model-free methods, such as Principal Components Analysis (PCA), and model-based clustering algorithms.

1.2.1 Principal components analysis

PCA is a vector decomposition method that projects the genetic variation data of individuals into lower dimensions using orthogonal vectors or principal components that each attempt to capture a large degree of the variation or information of the whole data while describing each individual using a single value. The first principal component explains more overall variation in the data than the second, and so on. When applying PCA to genotype data, individuals are projected and graphically visualized, typically on the first few principal components that explain the largest variation in the genotype data out of all components [24–26]. This projection is sometimes used as evidence for evolutionary processes including migration, geographical isolation, and admixture between populations [27]. For example, Figure 1.2 plots the first two principal components from a popular program EIGENSTRAT [28] applied to our simulated data and the surrogate individuals. Note that surrogate individuals from the three continental groups fall into different corners of the plot, while each set of admixed individuals falls between the two continental groups comprising their ancestry in a manner indicative of admixture and with placement roughly consistent with admixture proportions.

Despite of the appealing intuition and visualization of PCA, the interpretation of the projection is not always reliable as PC projections have been shown to depend on sample size, SNP ascertainment, and and features of demography besides admixture [27, 29].

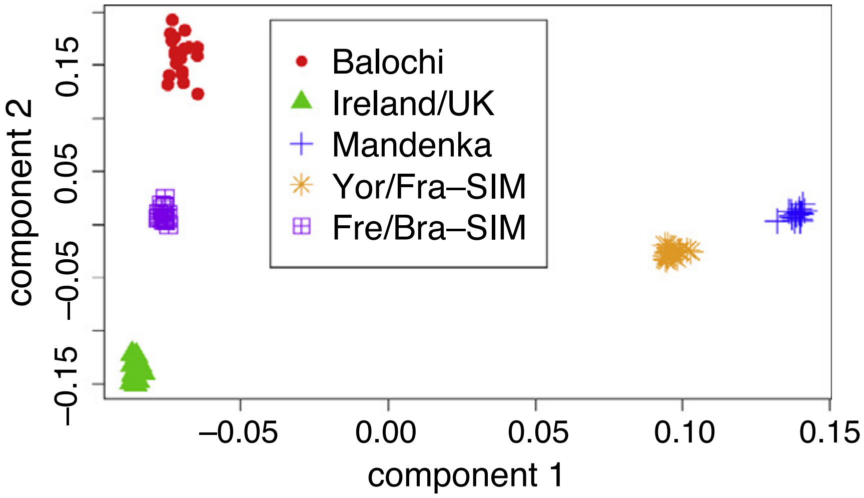

Figure 1.2: EIGENSTRAT on the simulated example and surrogates. Each dot is an individual colored by population and projected on principal components 1 (x-axis) and 2 (y-axis).

1.2.2 Model-based clustering methods

The aim of clustering methods is to classify individuals into a number of groups based on their genetic variation patterns. The most widely-used methods analyse unlinked SNPs and use probabilistic models to infer population substructure. These programs include the widely used software STRUCTURE [30], ADMIXTURE [31], FRAPPE [32] and related techniques (fastSTRUCTURE [33], teraSTRUCTURE [34]). STRUCTURE classifies the genetic variation data of individuals into K clusters, allowing individuals to be assigned to multiple clusters – a pattern that may indicate admixture – using Markov Chain Monte Carlo (MCMC). Later, Falush et al. [15] extended the STRUCTURE model to accommodate linked markers with the aim to capture the mosaic pattern of DNA that descends from the admixing ancestors (e.g. Fig 1.1). Under the pulse model, the breakpoints between successive DNA segments inherited from a single ancestor from the time of admixture occur as a Poisson process with rate r (in generations) per unit of genetic distance. Therefore, Falush et al., 2003, proposed a Markov Chain using the following linkage model:

$$Pr(Z_{i(1)j} = k|r,Q) = q_{ik},$$

and

$$Pr(Z_{i(l+1)j} = k'|Z_{ilj} = k,r,Q) \begin{cases} exp^{-g_l r} + (1 - exp^{-g_l r})q_{ik} & \text{if } k = k' \\ (1 - exp^{-g_l r})q_{ik'}, & \text{otherwise} \end{cases} \tag{1.1}$$

where Z_{ilj} is the cluster from which individual i derives its j^{th} allele at locus l from L total loci. q_{ik} can be viewed as the proportion of DNA for which individual i is most related to that of individuals in cluster k, with Q containing the set of all such proportions across all individuals and clusters. As mentioned above, the rate r can be thought of as the number of generations ago in which the source populations admixed (though the authors caution this may not provide a good estimate of this parameter in practice), and g_l is the genetic distance (in morgans) between loci l and $l+1$. In Equation 1.1, the top part is the probability where there is no switch in cluster membership between l and $l+1$; it is the combination of probability of no recombination $exp^{-g_l r}$ and the probability of at least one recombination $1 - exp^{-g_l r}$ times the probability of switching back to k or q_{ik}. While the bottom is the probability that there is at least one switch between l and $l+1$, resulting in a switch from cluster k to cluster k'.

Normally users select the number of clusters K (though K can also be estimated, [35, 36], and many of these programs then determine the proportion of each individual's DNA derived from K inferred clusters. Therefore, it can be tempting to interpret the K inferred clusters as K ancestral source groups that potentially intermixed, though individuals falling into multiple clusters may not always reflect admixture. Figure 1.3 provides results from ADMIXTURE applied to the simulated and surrogate data using K = 2-5. Here ADMIXTURE accurately describes sources and proportions of admixture in the two simulated groups at K = 3, but does not always characterize this admixture well at other values of K, with cross-validation choosing $K = 2$ as the best-fitting value. Therefore, as in PCA, multiple different demographic histories can lead to similar clustering [30, 37].

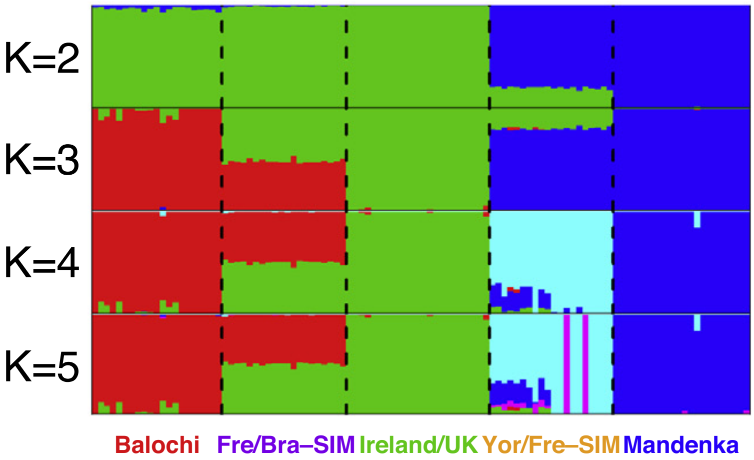

Figure 1.3: ADMIXTURE analysis on the simulated example and surrogates for cluster numbers $K = 2$–5 (columns = individuals, colors = clusters).

1.3 Identifying admixture

1.3.1 f statistics for admixture testing

f statistics estimate the change in allele frequencies or genetic drift between populations under the assumption that populations are related by a pre-defined topology. The ADMIXTOOLS package developed by Patterson et al. [17] and previously described in Reich et al. [38] offers different tests for admixture. This includes the three-population test (f_3), which tests whether there is admixture in populations C from sources related to populations A and B by calculating $(P_C - P_A)(P_C - P_B)$ across all SNPs where P_A, P_B, and P_C are the allele frequencies at any locus in population A, B, and C, respectively. It is based on the fact that if population C descends from an admixture event between two ancestors A and B, the allele frequency of C, or P_C, at SNP loci should fall between the frequencies of P_A and P_B. If this is significantly negative, then population C descends from an admixture event between two sources related to A and B. Significance is assessed using jack-knife re-sampling, based on dividing the genome into independent regions or chromosomes. Applying AD-

MIXTOOLS to calculate f_3 statistics for the two simulated populations using the given surrogates, strong evidence of admixture in both (Z-scores < 30) is found. However, the authors note that the f_3 statistics may no longer be negative if C has experienced a high degree of drift, e.g. due to a bottleneck, following admixture, as this can cause P_C to no longer often fall between P_A and P_B [17].

The four-population test, or f_4, aims to test whether $(A;B)$ and $(C;D)$ form clades by estimating $(P_A - P_B)(P_C - P_D)$ across all SNP loci. This statistic can be viewed as the correlation between the difference of allele frequencies between A,B and C,D. If the assumed clade is true, then there should be no correlation between $P_A - P_B$ and $P_C - P_D$, so that f_4 has expectation zero. In contrast, if f_4 is positive or negative, it indicates there is a correlation between (A,B) and (C,D) that indicates possible admixture. Specifically, if f_4 is significantly negative, it implies admixture between sources related to C and B and/or between sources related to D and A. If it is positive, it implies admixture between sources related to A and C and/or sources related to B and D. In practice, users usually set population A as an outgroup, i.e. assuming no admixture from A to either C and D, and test if f_4 is negative to suggest admixture between C and B or positive to suggest admixture between D and B. Moreover, ratios of $f4$ statistics can also be used to infer admixture proportions from each source, given the phylogeny among populations is known [17, 38].

1.3.2 Tree-based inference

Tree-based inference methods model the admixture among ancestral populations and their descendant populations as a tree where inner nodes represent populations, and edges reflect the genetic drift showing how far descendent populations have drifted from their ancestral populations. For instance, qpGraph included in ADMIXTOOL [17, 38] scores how well observed f statistics fit those predicted by a user-specified bifurcating tree relating the n sampled populations, with this tree potentially containing multiple migration events. The authors caution that the number of populations included should be small in practice.

MixMapper [39] extends qpGraph to a larger number of populations, by first inferring which populations are unadmixed using f_3 statistics, then building a bifurcating tree using these putatively unadmixed populations, and finally adding admixed populations onto this tree.

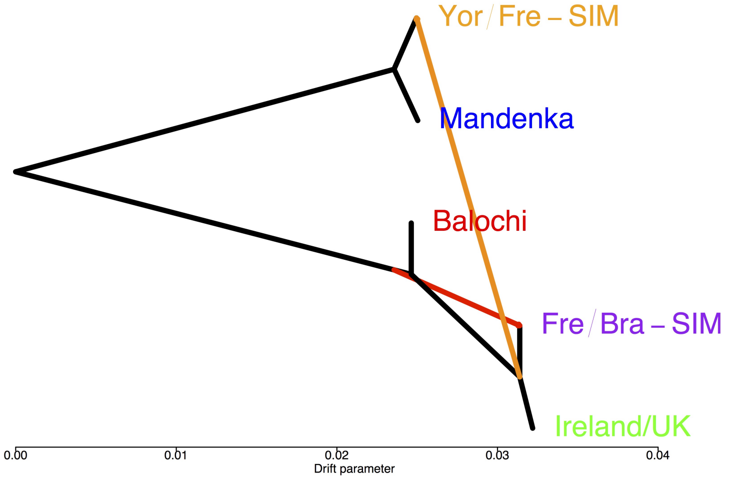

Figure 1.4: TREEMIX result (black lines = topology; red/orange lines = migration edges) inferred from individuals simulated as mixtures of 50% French + 50% Brahui ("Fre/Bra-SIM") or 80% Yoruba + 20% French ("Yor/Fre-SIM"). Balochi, Ireland/UK, and Mandenka are used as surrogates for Brahui, French, Yoruba, respectively.

TreeMix [40] is a related model that builds bifurcating trees relating a large number of groups and then adds links between branches to indicate admixture. TREEMIX uses a multivariate normal distribution to relate observed allele frequencies among populations, incorporating work by [41, 42], with the mean and covariance of this distribution defined by the (unknown) tree branch lengths (measuring drift) and admixture links or "migration edges" among populations. Figure 1.4 shows the inferred tree topology and admixture events when applying TREEMIX to the simulated data and surrogate individuals, correctly characterising the admixture in both cases.

The advantage of these methods is the intuitive way that they visualize the relationship between populations and migration events. However, a major limitation is the computational burden when analysing large number of populations, with the complexity of genetic relationships greatly limiting the search space of possible tree topologies.

1.4 Dating admixture

1.4.1 Local Ancestry Inference

The offspring of two parents from two isolated ancestral populations carries exactly one copy of a chromosome from each of the ancestries (Figure 1.1). These admixed individuals pass chromosomes along to their offspring that can contain a mixture of ancestries, due to recombination. Generations later, the genomes of the descendants of these admixed individuals will consist of contiguous "blocks" of ancestry inherited from each admixing population. Local inference along the genome of the ancestral sources that each individual inherited DNA from has been used to infer selection and map disease–associated genes in admixed populations [43, 44]. Several statistical models have been use to infer the ancestral source at each position in the genome.

Price et al. proposed the method HAPMIX [22] that takes haplotype information to infer individual-level information about ancestry, based on the idea that the haplotype of an admixed individual or offspring can be represented as a mosaic from the parental populations. To represent this idea, it requires phased genotype of surrogate individuals representative of each ancestral source as input. HAPMIX then identifies the surrogate population that best matches the haplotype segments within each admixed individual. HAPMIX offers accurate local ancestry; however, Pugach et al., 2011 [14] has shown from simulations that the HAPMIX date estimator might not be accurate as it underestimates some parameters, e.g. number of recombination, in some cases. Also, one limitation of this method is that it can only analyse a single admixture with two reference populations at a time, making it not flexible for admixture events consisting of multiple sources or where sources are unknown.

Maples et al. developed a computationally efficient method, RFMIX [45], for inferring local ancestry. RFMIX partitions each chromosome of an admixed individual into windows and uses a random forest to infer the ancestry in each window using reference phased chromosomes that are also partitioned into windows of the same size (Figure 1.6). Within a window, a random forest is trained to distinguish

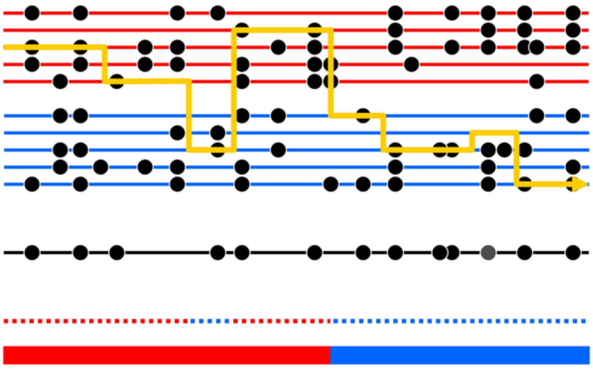

Figure 1.5: Schematic of sampling offspring (black line) from representative parental haplotypes (red and blue lines) taken from Price et al., 2009 [22]. Imperfect copying mimicking the mutation is show at the bottom dotted line when compared to the true admixed haplotype (bottom bar).

ancestry by using the reference panels; the votes for each ancestry are then performed to represent the admixed chromosome. These votes are summed to generate the posterior ancestry probabilities within each window. Finally, the posterior ancestry probabilities are used to determine ancestry across all windows. Figure 1.7b,e shows results from applying RFMix to the simulated data, which is dating admixture based on continuous ancestry tracts. Given the admixture proportion α in the admixed genome at λ generations, the density function $f(d)$ of the ancestry tracts length from the source population is:

$$f(d) = (1-\alpha)\lambda exp^{-(1-\alpha)\lambda d}$$

where d is the is a variable representing the length in cM [16, 46]. There is much better agreement between the expected exponential decay generated using the above equation with λ and α from the truth (dashed red lines) and the observed tract length distribution inferred by RFMix for the Yoruba-French simulation relative to the French-Brahui simulation, because the two admixing sources are considerably more genetically similar in the latter. This highlights an issue with many of these approaches – they typically only perform well when the intermixing source groups have a relatively high level of genetic differentiation.

Churchhouse and Marchini developed the method MULTIMIX [47] to infer the

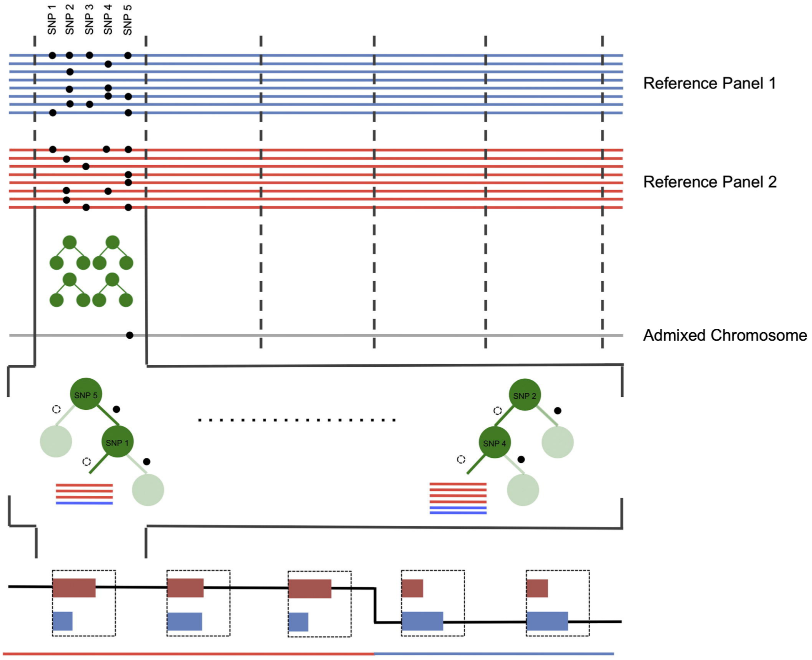

Figure 1.6: Illustration of RFMIX method taken from [45]. A chromosome is parted into multiple windows (dashed vertical lines). Within each window, a random forest (green) is trained to distinguish ancestry by using the reference panels (blue and red) with allelic types (black and white circle on the trees), after which a vote is made for the most probable ancestry for local admixture inference (bottom bar).

local ancestry in both phased and unphased admixed individuals. Unlike HAPMIX, it is based on a multivariate model that makes it applicable to infer multiple ancestral populations, and it does not require phased ancestral haplotypes. The model is a hidden Markov model that accounts for background LD – LD that does not relate to admixture [15,48]. Like RFMix, MULTIMIX describes sources of ancestry within windows, with a transition matrix of the HMM that models the switch of ancestry between adjacent windows. The MULTIMIX transition probability between consecutive window $W_{(l)j}$ and $W_{(l+1)j}$ is denoted as:

$$Pr(W_{(1)j} = a|r,q) = q_{aj}$$

and

$$Pr(W_{(l+1)j} = a | W_{lj} = b, r, q) \begin{cases} exp^{-d_l r} + (1 - exp^{-d_l r}) q_{aj} & \text{if } a = b \\ (1 - exp^{-d_l r}) q_{aj}, & \text{otherwise} \end{cases} \tag{1.2}$$

where q_{aj} is the ancestry proportion of individuals j contributing from ancestral source a, and d_l is the genetic distance in Morgans between the midpoints of the two windows l and $l+1$. r is the rate parameter related to number of generations since an admixture event between the source populations, under a pulse model assuming random mating. The above equation relates to the linkage STRUCTURE model [15], in that they rely on modelling mosaic patterns that descend from multiple ancestors along the chromosome. However, instead of using genetic distance from SNP loci, MULTIMIX uses the distance between windows or blocks that are individually assigned ancestry.

1.4.2 Techniques using linkage disequilibrium decay patterns

Recall Figure 1.1, considering two ancestries A and B that mixed at λ generations ago. The genetic distance g between any two SNPs inherited from A and B can be modelled under the Poisson process as an exponential decay curve with decay rate λ. Based on this property, many tools have been proposed to identify and date admixture. For example, Moorjani et al., 2011 [20] proposed ROLLOFF, which is a method for dating the most recent admixture event by fitting the decay in LD versus genetic distance among segments inherited from two admixing sources A and B. Taking as input SNP data and two pre-specified surrogates reflecting the two ancestral populations, ROLLOFF constructs an LD curve $R(g)$ showing the correlation between pairs of SNPs, with a weighting function that puts more weight on SNPs where the difference in frequency between the surrogates to A and B is large. They fit the model $R(g) = R_0 exp^{-\lambda g} + \varepsilon$ to estimate the number of generations λ simultaneously with the curve amplitude, R_0, and the residuals from regression, ε. Figure 1.7c,f shows results from applying ROLLOFF to the simulated data. Like RFMix, there is much better agreement between the expected exponential decay (dashed red lines) and the LD curve generated by ROLLOFF for the Yoruba-French simulation relative to the French-Brahui simulation. Loh et al. proposed a method, ALDER [19], for inferring the admixture date based on the same idea of ROLLOFF, but with several significant improvements. Like ROLLOFF, ALDER uses the decay curve of weighted LD to estimate an admixture date. However, ALDER offers a new weight function that is more accurate, and uses the amplitude of the weighted LD curve to infer the admixture proportion. Moreover, ALDER adopts a fast Fourier Transform (FFT) to speed up computation of the weighted LD statistic.

1.4.3 Techniques using linkage disequilibrium decay patterns among haplotypes

These methods use haplotypes to infer admixture events rather than just independent SNPs, which is considered to be more informative for capturing the ancestral signal from large-scale genotyping or sequencing data. To do this, Hellenthal et al. proposed the software GLOBETROTTER [21] that incorporates haplotype information when modelling admixture LD decay to infer the admixture date. This approach first constructs the genomes of each putatively admixed individual as a mosaic of that of a set of reference individuals, based on identifying matching haplotype patterns. GLOBETROTTER then measures the decay of LD among segments matched to different reference populations, relating the decay rate to the time of admixture. GLOBETROTTER also provides the inference of the proportion contributed from the different ancestral sources. Unlike ROLLOFF and ALDER that identify a single best surrogate for each source by finding the best model fit out of pairings of available surrogates, GLOBETROTTER infers the genetic make-up of each source as a mixture of DNA from all surrogate groups (i.e. without requiring one pre-specified surrogate per source). In cases where evidence of admixture is found, GLOBETROTTER further examines whether the data fits a single rate of decay, indicative of a single date of admixture, or a multiple rates of decay suggesting multiple dates of admixture. Figure 1.7a summarizes the intuition of using different type of information among 3 approaches i.e. RFMix only observes the decay of the tract-length distribution of DNA contributed from an ancestry; ROLLOFF measures the LD decay between any pair of SNPs with different allele frequencies in the two ancestries, while GLOBETROTTER observes the LD decay from any pair of DNA segments contributed from the two ancestries along the chromosomes. Figure 1.7e-g indicates that GLOBETROTTER is more accurate than RFMix and ROLLOFF in the Brahui-French simulation, a more challenging problem relative to the Yoruba-French simulation where the inferences are similar between the three approaches.

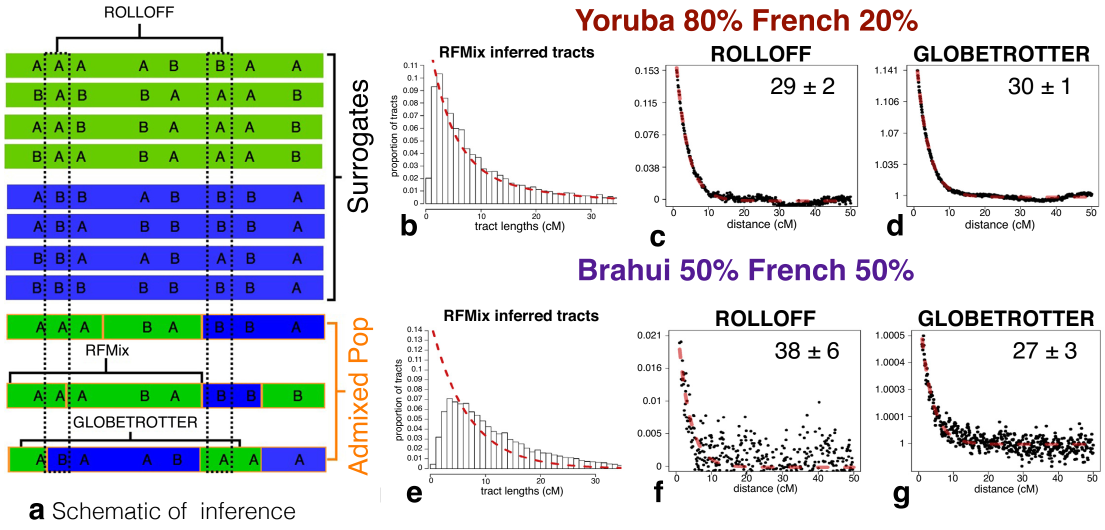

Figure 1.7: RFMix, ROLLOFF, and GLOBETROTTER applied to date admixture in the simulations, with each approach intuitively explained in (a) as measuring (b,e) the distribution of tract lengths matched to the same surrogate group, or the decay of LD (black dots) versus genetic distance between (c,f) SNPs with different allele frequencies in the two surrogates or (d,g) haplotype blocks matched to a surrogate (Ireland). Dashed red lines give the expected decay for the true proportions and true admixture date, and numbers in the top right of (c,d,f,g) give inferred dates and standard errors.

Salter-Townshend and Myers [49] developed a fine-scale local ancestry inference method called MOSAIC. Like HapMix, this method uses a HMM to quantify the local ancestry along the genome. However, it can handle multiple ancestral sources without prior knowledge of admixing groups, making MOSAIC a more flexible method. Similar to GLOBETROTTER, MOSAIC generates LD-decay or co-ancestry curves and infers a best-fitting admixture date. Moreover, it provides a re-phasing step which aims to cope with phasing errors and offers ancestry informed phasing of the target haplotype. The schematic is represented in Figure 1.8. Despite these helpful features, MOSAIC can currently only model one date of admixture involving 2 or more sources; it cannot detect the multiple-date events. The inference accuracy is also not significantly improved when compared to GLOBETROTTER.

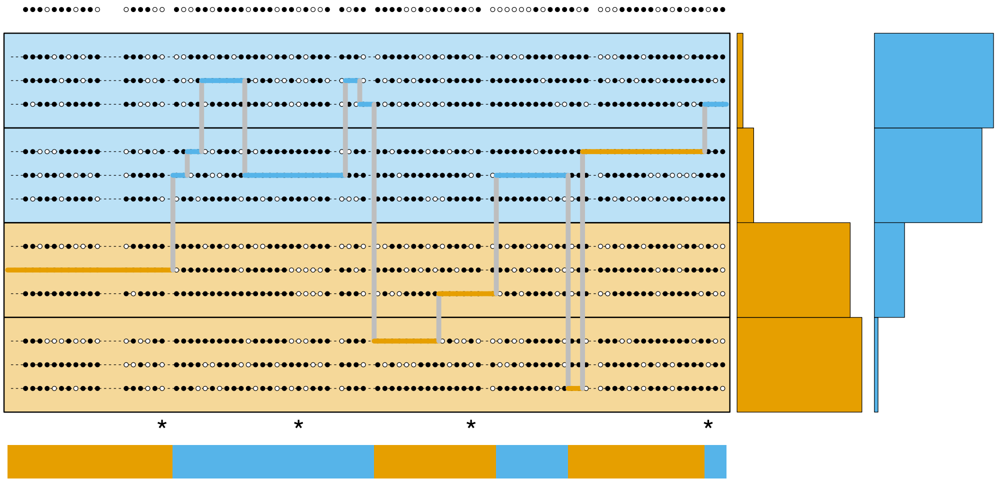

Figure 1.8: Schematic of MOSAIC taken from [49]. The top row is an admixed haplotype. The haplotypes of reference panels are below and divided into 2 groups of ancestors (blue and orange). The local ancestry inference of the admixed haplotype is located at the bottom, indicating the ancestry content from blue and orange source. The sidebars represent the conditional (copying) probability of choosing reference panels given the local ancestry.

1.5 Thesis Aims

Amongst several well-established approaches to studying the genetic history of worldwide populations, GLOBETROTTER has been proven to be the most comprehensive and powerful. However, it is considered to be computational costly for a large-scale genetic data analysis. Therefore, the aim of this thesis is to:

1. develop a new method that increases the speed of inferring admixture events without losing accuracy

2. apply this new method to large-scale data including 6,000 of European samples with $\sim$500,000 SNPs

In Chapters 2, I will describe, in detail, the fundamental intuition behind the method GLOBETROTTER on which my approach is mostly based. In general, the method includes haplotype phasing of SNP data, chromosome painting, performing mixture modelling, generating LD-decay curves, and fitting the curves for admixture events.

With a good understanding of the existing GLOBETROTTER model described in Chapter 2, Chapter 3 presents my new method fastGLOBETROTTER that improves the speed of inferring admixture dates without losing accuracy. I demonstrate that the use of sampling distributions can solve the complexity burden of GLOBETROTTER. Furthermore, there are also other features that can efficiently reduce computational time such as “code optimization” which enables some parts of the algorithm to be optimized in the more efficient way. Moreover, fastGLOBETROTTER speeds up the calculation by combining multiple donors that share a similar genetic background.

To evaluate the performance of fastGLOBETROTTER, the computational time and inference accuracy in a variety of simulated datasets are tested and compared between fastGLOBETROTTER and GLOBETROTTER, as described in Chapter 4. This chapter shows that fastGLOBETROTTER not only significantly improves computational time, but also provides more accuracy in some cases, interestingly. This chapter includes additional simulations designed to assess how to interpret results reported in Chapter 5.

Chapter 5 presents the application of fastGLOBETROTTER to a large-scale dataset consisting of more than 6,000 European samples from Italy, Spain, France, Germany, Belgium, Poland, Denmark, Sweden, Finland and Norway over 2,500 individuals as ancestry reference. The chapter reports new findings of admixture events across Western Europe. These include admixture dated to 500-600CE from sources carrying DNA related to present-day west Asian and African populations found in individuals within Belgium, France and parts of Germany. In Scandinavia, we also detect admixture from sources carrying DNA related to east Asian/Siberian in individuals within Finland, Norway and Sweden at different times starting around 100 CE.

The final Chapter 6 provides a detailed description about how to apply fastGLOBETROTTER as a distributed software. It presents two topics: the software instruction and the software tutorial. The software instruction aims to give readers basic information about fastGLOBETROTTER, e.g. the method's intuition, parameters, input and output file formats. The tutorials present a step-by-step guidance on how to perform inference using a real example. The steps progress through the entire analysis, from description of the data, preparing files for fastGLOBETROTTER, haplotype phasing, estimating parameters, chromosome painting, inferring admixture events and finally ends with interpreting the result.

Chapter 2

Fundamentals of admixture inference using GLOBETROTTER

2.1 Fundamental of Admixture Inference using GLOBETROTTER

GLOBETROTTER is a haplotype-based method that models the segments of DNA inherited from multiple admixing sources to identify and date admixture events. It relies on characterizing sampled admixed individuals that contain a mixture of ancestries that are related to different sampled reference groups (or “surrogates”). For two genetic segments (or “chunks”) separated by a particular genetic distance, GLOBETROTTER infers the probability that one segment shares most recent ancestry with surrogate X while the other shares most recent ancestry with surrogate Y. These probabilities are calculated according to different types of admixture scenarios, including a simple admixture consisting of two groups intermixing at a single time and more complex admixture consisting of admixture involving three or more groups at one or multiple times.

In particular GLOBETROTTER initially uses chromosome painting (Section 2.2.2) to attempt to identify which segments in a target population are inherited from each admixing source. While accurate identification of these segments would enable both determining the proportion of admixture from each source and the date of admixture (i.e. using the distribution of segment lengths matching to each source

– see Section 1.4.1), in practice accurate painting is very challenging in most cases of human admixture. For example, Figure 1.7b,e gives the truth and RFMix inferred paintings for the simulated cases of Chapter 1. Note that for the French-Brahui simulation, results are very noisy, which in turn leads to a segment length distribution that is a poor reflection of the true admixture date (Figure 1.7e). Therefore, GLOBETROTTER (and related techniques ROLLOFF, ALDER, MOSAIC) implements a more robust approach for inferring dates and proportions that does not rely on perfect (or near-perfect) reconstruction of the sources inherited at each segment of each target haploid.

Let consider a single admixture model (Figure 1.1) where an admixed population descends from the mixture of two source groups A and B. The genetic contributions from A and B to the admixed population are $P_A = \alpha$ and $P_B = 1-\alpha$, respectively. Assuming that that the crossovers between any 2 loci occur at random (i.e. no crossover interference), they can be described using a Poisson process [15] with rate related to number of generation λ. Under this model, the probability of no recombination between any two loci separated by a distance g Morgans since the date of admixture λ is $exp^{-g\lambda}$. Therefore, the probability of having 2 loci separated by g along a chromosome within a haploid genome of an admixed individual where one locus derives from A and another from B and $A \neq B$ is:

$$P_{AB}(g) = P_A P_{(recom|copyA,Batendpoints)} P_B + P_A P_{(norecom|copyA,Batendpoints)} P_B = P_{BA}(g)$$

$$P_{AB}(g) = \alpha(1-exp^{-g\lambda})(1-\alpha) + \alpha(0)(1-\alpha)$$

where $P_{(recom|copyA,Batendpoints)}$ refers to the probability of having at least one recombination given one endpoint copies from A and another from B. We can simplify these terms as

$$P_{AB}(g) = \alpha(1-\alpha) - \alpha(1-\alpha)exp^{-g\lambda} \tag{2.1}$$

In the case of A=B:

$$P_{AA}(g) = (\alpha)(1-exp^{-g\lambda})(\alpha) + \alpha(exp^{-g\lambda})(1)$$

$$P_{AA}(g) = \alpha^2 + \alpha(1-\alpha)exp^{-g\lambda} \tag{2.2}$$

and

$$P_{BB}(g) = (1-\alpha)(1-exp^{-g\lambda})(1-\alpha) + (1-\alpha)(exp^{-g\lambda})(1)$$

$$P_{BB}(g) = (1-\alpha)^2 + \alpha(1-\alpha)exp^{-g\lambda}. \tag{2.3}$$

Examples for P_{AA}, P_{AB}, P_{BB} are given in Figure 2.1 for α= 0.5, λ=100. Critically, P_{AB} increases exponentially with g at rate equal to λ, while each of P_{AA} and P_{BB} decrease exponentially with the same rate. GLOBETROTTER uses these properties to both date the admixture event and to identify which surrogate groups best reflect each admixing source, as briefly outlined below.

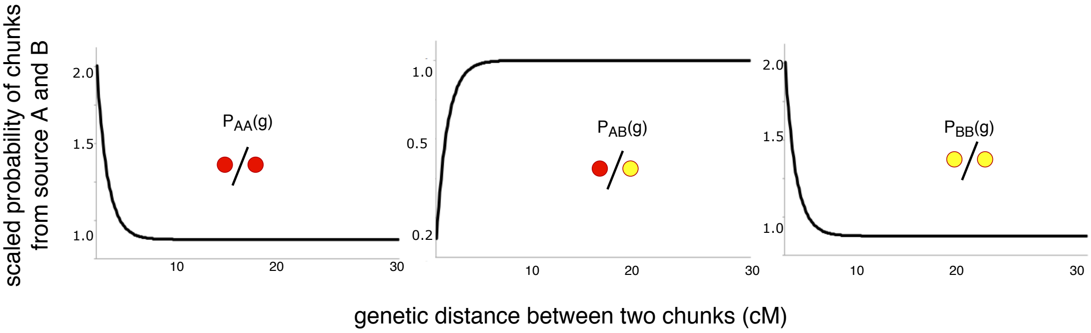

Figure 2.1: Example of P_{AA}, P_{AB}, P_{BB} curves generated from α= 0.5, λ=100. The x-axis is the genetic distance g between any pair of DNA segments where one segment is contributed from source A and the other from source B. The y-axis is the probability calculated from Equations 2.2, 2.1 and 2.3, scaled by genome-wide expectation of each event (i.e. $P_A P_A$, $P_A P_B$ and $P_B P_B$, respectively) as g varies.

In practice we do not observe the true admixing sources A and B but instead use sampled populations to act as surrogates for these sources.

Let Q_{ml} be the probability that segment l within the genome of an admixed individual (resulting from admixture between sources A and B) is most recently related ancestrally to surrogate m, and let $E(Q_{ml}Q_{nr}; g)$ be the expected product of probabilities that two segments at locations l and r separated by distance g share most recent ancestry with surrogates m and n, respectively. Hellenthal et al., 2014 show that:

$$E(Q_{ml}Q_{nr};g) = Q_mQ_n + \alpha(1-\alpha)[Q_m^B - Q_m^A][Q_n^B - Q_n^A]\exp^{-g\lambda}, \tag{2.4}$$

where $Q_m^A = E[Q_{ml} \mid \text{true ancestry at } l \text{ is } A]$ and $Q_m = \sum_{i\in\{A,B\}} Q_m^i \Pr(\text{true ancestry is } i)$. They then also define:

$$\psi(Q_mQ_n;g) = \frac{E(Q_{ml}Q_{nr};g)}{Q_mQ_n} = 1 + \delta_{mn}exp^{-g\lambda} \tag{2.5}$$

which is used to date admixture and help infer the genetic make-up of the admixing sources.

2.2 Steps in GLOBETROTTER

Here I describe the main inference steps of GLOBETROTTER, as detailed in Hellenthal 2014. The aim is to determine whether a target population shows evidence of admixture, and – if so – to identify the sources of this admixture and to date when the admixture event(s) occurred. Throughout this section, I use the term "target" to refer to the putatively admixed population, and "surrogate" to refer to other sampled populations that can be used to describe the sources of admixture in the target population. These terms differ from the concepts of "recipient" and "donor" outlined below, which refer only to the painting process described in Section 2.2.2. Briefly, in this sub-section each individual from the target and surrogate populations are considered "recipients" when their genetic variation patterns are compared to a set of "donor" individuals. In practice, typically the same set of sampled populations are used as surrogates and donors, and for some parts of the analysis the target individuals often are included among the donors as well.

2.2.1 Haplotype phasing

Haplotype phasing is a process to convert the unphased genotype data of each individual to haploid genomes that were each inherited from a different parental source. Using these phased genomes, we can make use of the haplotype information of physically close linked alleles on the same chromosome to increase power to characterize admixture events relative to methods that use ignore haplotype information [21, 36]. Phasing methods infer the most likely haplotype configurations in unphased genotype data, sometimes by using a reference set of individuals whose phase is well-estimated (e.g. from using trios). Several popular methods exist to do so, some of which do not require reference individuals, including fastPHASE [50], BEAGLE [51], IMPUTE [52] and SHAPEIT [53]. Among these tools, SHAPEIT is the fastest, with linear complexity in terms of number of individuals while others are quadratic, while maintaining accuracy, and is hence more readily applicable to large-scale genome-wide data resources such as those considered in this thesis.

2.2.2 Incorporate linkage disequilibrium and haplotype information - "chromosome painting" approach

Lawson et al [36] introduced the software CHROMOPAINTER that aims to find haplotype patterns shared among individuals, which is indicative of recent shared ancestry. CHROMOPAINTER is based on a HMM that was proposed by Li and Stephens [54] originally to estimate recombination rates but which was re-purposed to explore ancestry. CHROMOPAINTER compares the phased SNP data of a recipient (or target) individual to that of a set of donor (or reference) individuals. In particular for each segment along each haploid genome of a recipient individual, CHROMOPAINTER identifies the donor haploid with the most closely matching SNP data. In this manner, CHROMOPAINTER constructs each recipient haploid as a collection of painted segments (chunks) from the donor haploids, where a chunk is a continuous segment of DNA – typically containing multiple SNPs – that matches to (is painted by) the same donor haploid. This painting process is stochastic, so that different CHROMOPAINTER runs can generate different "painting samples". The Forward-Backward algorithm can be used to find the total expected amount of genome-wide DNA that each target individual matches to each donor. Figure 2.2 illustrates the concept of constructing painting samples taken from Li and Stephens [54].

More formally, let $h_r = \{h_{r1}, ..., h_{rL}\}$ be the phased haploid data of a recipient individual across L SNPs, where h_{rl} is the allele type carried by this recipient hap-

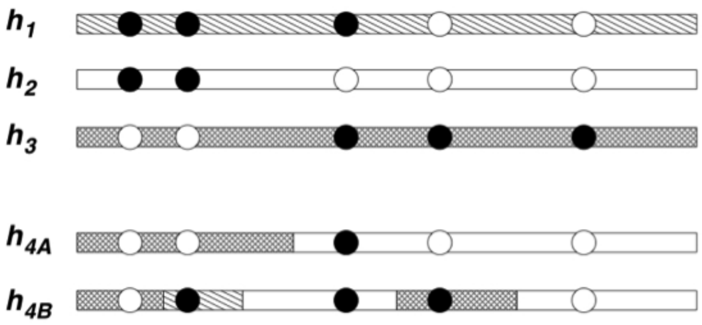

Figure 2.2: Painting samples taken from Li and Stephens [54]. h_{4A} and h_{4B} are recipients while h_1, h_2, h_3, are donors. White and black circles represent two alleles. The shading (or painting) in recipients refers to which haplotype that is copied from donors at each position.

loid r at SNP l. CHROMOPAINTER constructs each such recipient haploid as a mosaic of j phased donors with haploid data $h_1,...,h_j$. Let $\vec{q} = \{q_1,...,q_j\}$ be the vector of copying probabilities, where q_j is the probability that haploid h_r copies from donor j at any particular locus. Let$\vec{\rho} = \rho_1,...,\rho_{L-1}$ where $\rho_l = N_e d_l$ is the population-scaled genetic distance between locus l and locus $l+1$ in the genome, where N_e is the effective population size and d_l is the genetic distance in Morgans between loci l and $l+1$. A HMM is built incorporating the hidden state vector $\bar{Y} = Y_1,...,Y_L$, where Y_l is the donor haploid j that r copies from at locus l. The transition probability from Y_l to Y_{l+1} follows a Poisson Process with rate ρ_l, so that:

$$Pr(Y_1 = y_l) = q_{y_l}$$

and

$$Pr(Y_{l+1} = y_{l+1}|Y_l = y_l) = \begin{cases} \exp^{-\rho_l} + (1 - \exp^{-\rho_l}) q_{y_{l+1}} & \text{if } y_{l+1} = y_l \\ (1 - \exp^{-\rho_l}) q_{y_{l+1}}, & \text{otherwise} \end{cases} \tag{2.6}$$

Let θ be a per locus mutation rate parameter, which allows for mismatches between the allele carried by h_r and the allele carried by the copied donor h_{y_l} at a given SNP l. The emission probability of the HMM is given by

$$Pr(h_{rl} = a|Y_l = y_l) = \begin{cases} 1 - \theta & \text{if } h_{(y_l)l} = a \\ \theta & \text{otherwise} \end{cases}$$

In practice we fix d_l to be the inferred genetic distance using the recombination map combined across all populations for HapMap. We fix q_j to $1/J$ for $j \in [1,...,J]$. Then Lawson et al., 2012 [36] describe how to infer θ and N_e using an Expectation-Maximisation algorithm, with default (or starting) values of Watterson's estimate and 400000/J, respectively.

We can use this HMM to estimate the expected total genome-wide amount (in Morgans) of chunks that haploid r copies from each donor, termed a "copy vector", $\vec{f_r} = f_{r1},...,f_{rj}$ where f_{rj} is the expected genome-wide amount that a recipient individual r copies from donor individual j. Specifically for fixed values of θ, ρ_l for $l \in [1,...,L-1]$ and q_j for $j \in [1,...,J]$:

$$\begin{aligned} f_{ij} = \tfrac{1}{Pr(D)} \textstyle\sum_{l=1}^{L-1} g_l[\alpha_{jl}\beta_{j(l+1)}(exp^{-\rho_l} + (1-exp^{-\rho_l})q_j)Pr_h \\ +(1/2)[\alpha_{jl}\beta_{jl} + \alpha_{j(l+1)}\beta_{j(l+1)} - 2\alpha_{jl}\beta_{j(l+1)}((exp^{-\rho_l} + (1-exp^{-\rho_l})q_j)Pr_h)]]. \end{aligned} \tag{2.7}$$

Here $Pr_h \equiv Pr(h_{r(l+1)}|Y_{l+1} = h_j)$ - the emission probability generated from donor j. $Pr(D) \equiv Pr(h_r|h_1,...,h_j;\vec{\rho};\vec{q};\theta)$ is calculated from a summation is performed over all permutations of the copying process, i.e. all possible y, which can be done using the forward algorithm [55]. Following [55], we define the the forward probabilities as;

$$\alpha_{jl} = Pr(h_{r1},...,h_{rl},Y_l = h_j)$$

and backward probabilities of the HMM as;

$$\beta_{jl} = Pr(h_{r(1+1)},...,h_{rL}|Y_l = h_j)$$

We calculate α_{jl} for $j = 1,...,J$ in the following manner

$$\alpha_{jl} = \begin{cases} Pr(h_{r1} \mid Y_1 = h_j)q_j \text{ for } L = 1 \\ Pr(h_{rl} \mid Y_l = h_j)(\sum_{i=1}^{J} \alpha_{i(l-1)}q_j(1-exp^{-\rho_l}) + exp^{-\rho_l}\alpha_{j(l-1)}) \text{ for } l = 2,...,L. \end{cases}$$

and we calculate β_{jl} for $j = 1,...,J$ in the following manner

$$\beta_{jl} = \begin{cases} \beta_{jL} = 1.0 \\ [\sum_{i=1}^{J} \beta_{i(l+1)} q_i Pr(h_{r(l+1)} \mid Y_{l+1} = i)](1 - exp^{-\rho_l}) + exp^{-\rho_l} Pr(h_{r(l+1)} \mid Y_{l+1} = h_j) \beta_{j(l+1)} \\ \text{for } l = 1,...,L-1 \end{cases}$$

To generate painting samples of $\vec{Y}$ conditional on $(h_r, h_1, ..., h_j; \vec{\rho}; \vec{q}; \theta)$, we perform the following steps;

1. Sample Y_L according to $Pr(Y_L = h_j \mid h_1,...,h_j,h_r) \propto \alpha_{kL}$.

2. For $l = L-1,...,1$, sample Y_l according to: $Pr(Y_l = h_k | h_1,...,h_j,h_r,Y_{l+1},...,Y_L) \propto [\sum_{i=1}^{J} \alpha_{il}](1 - exp^{-\rho_l}) q_{Y_{l+1}} + \alpha_{Y_{(l+1)}l} exp^{-\rho_l}$.

In practice, For each individual in this analysis, we generate 10 such "painting samples" of for each haploid for use in generating coancestry curves (Section 2.2.4). Figure 2.3a shows an example of multiple painting samples of a target as a mosaic of donor haplotypes.

2.2.3 Using copy vector to model groups as mixtures of sampled populations

For any haploid, individual or group, we can describe its genome-wide copy-vector as a mixture of those from a set of surrogate groups. For example, GLOBETROTTER models the copy vector of the admixed group as a linear combination of the copy vectors of the sampled surrogate groups (Figure 2.3b). Recall The right-hand side of the Equation 2.7 gives the genome-wide amount of DNA that individual target i matches to individual donor j. To get the proportion $\hat{f}_{ij}$, we divide this by the total amount that target i matches to all donors $j \in [1,...,J]$ and $\sum_{j=1}^{J} \hat{f}_{ij} = 1$. Specifically in population level, GLOBETROTTER forms the population-averaged copy vector $\hat{f} = \hat{f}_1, ..., \hat{f}_J$ from

$$\hat{f}_j = \sum_{i=1}^{n_i} \hat{f}_{ij} / n_i \qquad (2.8)$$

where n_i is the number of samples in recipient population i. Let $\hat{f}_r$ be a copy vector of the admixed group or target group, we can describe $\hat{f}_r$ as a linear mixture of the population-averaged copy vectors ($\hat{f}_j$) of donor population $j = 1,...,J$

$$\hat{f}_r = \sum_{j=1}^{J} \beta_j \hat{f}_j + \varepsilon, \tag{2.9}$$

where ε is a vector of error terms. The mixture coefficients (β_js) are estimated using non-negative least squares regression (nnls) in R, which assumes $\beta_j \geq 0$ for $j \in [1,...,J]$ and $\sum_{j=1}^{J} \beta_j = 1$. Informally, the above regression identifies which surrogate populations have average genome-wide painting patterns that best match the painting patterns observed in the target population, and at what relative proportions. This mixture model accounts for issues such as unequal sample sizes among donor groups when identifying the surrogates that share recent ancestry with the target. Typically only a subset of the K surrogate populations have $\beta_k > 0.005$, reducing the number out of K total surrogate populations that are subsequently used in analysis. Using these β_ks, the matrix $K \times J$ matrix W, where the mth row and ith column of W is $W(m,i) \equiv \Pr(\text{ancestry shared most recently with surrogate } m \mid \text{copy donor } i)$ calculated as:

$$W(m,i) = \frac{\beta_m \hat{f}_{mi}}{\sum_{k=1}^{K} \beta_k \hat{f}_{ki}}. \tag{2.10}$$

Recall that $\hat{f}_{mi}$ refers to the estimated probability that surrogate group m copies from donor i. The matrix $W(m,i)$ is used as a reweighting matrix to determine which surrogate is best reflected by the particular donors copied.

2.2.4 Generating coancestry curves

Coancestry curves are generated empirically from the painting samples $\vec{Y}$ for the two haploids of each target individual in Section 2.2.2. To build the curves, GLOBETROTTER initially tabulates the counts among these painting samples whereby any two chunks (weighted by their lengths) separated by a distance gcM has one chunk painted by donor group A and the other by donor group B. In practice GLOBETROTTER places g into bins of user-specified size (with a default of 0.1cM). For each target individual, typically 10 painting samples are generated for each of the individual's two haploids, giving 20 samples in total. Separately for each target individual, GLOBETROTTER compares every pair of chunks within and between these 20 painting samples, considering chunks on different haploids to cope with potential phasing errors.

In this manner, GLOBETROTTER constructs a "raw" co-ancestry curve $\phi_r(i, j; g)$ for a target individual r and a donor pair i and j for each distance g as:

$$\phi_r(i, j; g) = \sum_{a=1}^{20} \sum_{b=a}^{20} \sum_{\chi_{a,b;g}} \omega_l \omega_h, \tag{2.11}$$

where ω_l and ω_h are the chunk-lengths of chunks l and h. Here chunks l and h, separated by gcM, are painted by donors i and j, respectively, and found on painting samples a and b, respectively, with $\chi_{a,b;g}$ the set of all chunk pairs meeting these criteria. Note that longer chunks contribute more to $\phi_r(i, j; g)$, though ω_l is capped at a size of 1cM. Figure 2.3c illustrates this step, whereby each node is a chunk and its color represents the donor painting that chunk, with edges the genetic distance between chunks.

As stated in Hellethal 2014 [21], these counts are then multiplied by the weights from (Equation 2.10) and summed over all donor pairs to find $\phi_r(m,n;g)$, the counts of segment pairs separated by distance g that are most recently related to surrogates m,n:

$$\phi_r^*(Q_{ml}, Q_{nr}; g) = \sum_{i=1}^{J} \sum_{j=1}^{J} W(m,i)W(n,j)\phi_r(i,j;g) \tag{2.12}$$

Equation (2.12) is used to find these counts for all pairs of surrogates $(m,n) \in [1,..,K]$. Note that Equation 2.12 represents our empirical estimate of $E(Q_{ml}Q_{nr}; g)$ defined in Equation 2.4.

Next the marginal counts for sharing ancestry with surrogates m and n is calculated as:

$$\tilde{\phi}_r^*(Q_{ml},Q_{nr};g) = [\sum_{h=1}^{K} \phi_r^*(Q_{ml},Q_{hr};g)][\sum_{h=1}^{K} \phi_r^*(Q_{hl},Q_{nr};g)]/[\sum_{h,p}^{K} \phi_r^*(Q_{hl},Q_{pr};g)]. \tag{2.13}$$

For each distance g, this gives an additional $K \times K$ matrix. If the populations m and n do not indicate evidence of admixture in this recipient individual, then Equation 2.13 should be approximately equivalent to Equation 2.12 at every distance g for all m, n in K.

After a step that symmetrizes the two matrices defined by Equations 2.12 and 2.13, the authors then define $\widehat{\psi}(m,n;g)$ to be the average (across target individuals r) of Equation 2.12 divided by Equation 2.13. $\widehat{\psi}(m,n;g)$ is referred to as the "coancestry curve" of the target population for surrogate pair (m,n) – .e.g. the black lines in Figure 2.3d.

2.2.5 Fitting the coancestry curve to estimate the admixture date

The value $\widehat{\psi}(m,n;g)$ is an estimate of $\psi(Q_m Q_n;g)$ in equation (2.5). Specifically, the authors assume

$$\widehat{\psi}(m,n,g) \equiv \tau_{mn} + \sigma_{mn} exp^{-g\lambda} + \varepsilon, \tag{2.14}$$

and find the values of λ and (τ_{mn},σ_{mn}) for all $(m,n) \in [1,...,K]$ that minimize:

$$\sum_{mn}\sum_{g} \left(\widehat{\psi}(m,n,g) - \tau_{mn} - \sigma_{mn} exp^{-g\lambda}\right)^2. \tag{2.15}$$

Examples of fitted curves of using these values are given in the green lines of Figure 2.3d. Recall that λ is the date of admixture in generations. The estimates of σ_{mn} are used with the copy-vectors to help infer the proportions of ancestry from each source and the genetic make-up of each source, which in turn leads to new estimates of β_k for $k \in [1,...,K]$. In practice, GLOBETROTTER iterates between inferring the terms in Equation 2.15 and inferring the β_ks used to make the weights defined in Equation 2.10, typically using 5 such iterations.

The final results are the estimations of λ, τ_{mn} and δ_{mn} for all m,n, the proportions of DNA inferred from each source and a representation of each source group as a mixture of DNA from the surrogate groups. Moreover, GLOBETROTTER uses bootstrap re-sampling to infer confidence intervals for the dates. To do so in a admixed population with n_r individuals, each bootstrap re-sample generates n_r pseudo-individuals. For each chromosome of each pseudo-individual, the painting samples from that chromosome of a single target individual is randomly sampled from amongst the n_r target individuals. After n_r such pseudo-individuals are generated, the admixture dates and proportions are inferred as described above for that bootstrap re-sample. Typically 100 bootstrap re-samples are used to generate the final confidence intervals. The authors suggest concluding "no admixture" for any cases where these bootstrap re-samples contain an estimated date of 1, which is indicative of no detectable admixture, or $>$400, which is a date too old for GLOBETROTTER to reliably detect and hence an unclear admixture signal.

GLOBETROTTER also attempts to fit two dates of admixture by replacing Equation 2.15 with a similar equation containing the sum of two exponential distributions with rates equal to the two admixture dates (see a curve example in Figure 2.3e). The authors show this to be theoretically appropriate in the case of multiple dates of admixture, and similar inference to that described above is performed to infer the dates, sources and proportions of admixture for both events. Additional steps can also be used to test for $>$2 sources intermixing at one time.

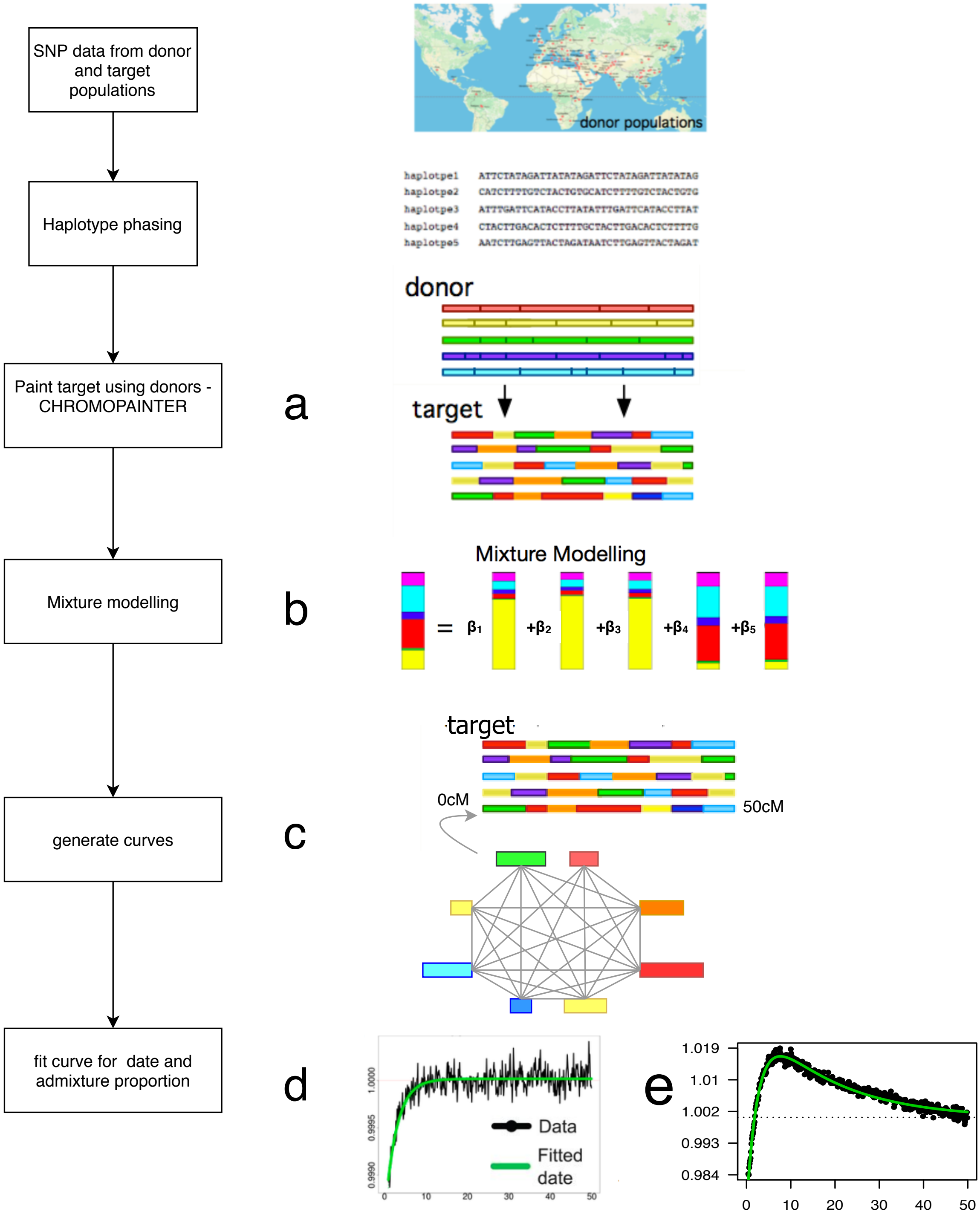
SNP data from donor and target populations
Haplotype phasing
Paint target using donors - CHROMOPAINTER
Mixture modelling
generate curves
fit curve for date and admixture proportion
donor populations
haplotpe1 ATTCTATAGATTATATAGATTCTATAGATTATATAG
haplotpe2 CATCTTTTGTCTACTGTGCATCTTTTGTCTACTGTG
haplotpe3 ATTTGATTCATACCTTATATTTGATTCATACCTTAT
haplotpe4 CTACTTGACACTCTTTTGCTACTTGACACTCTTTTG
haplotpe5 AATCTTGAGTTACTAGATAATCTTGAGTTACTAGAT
donor
a
target
Mixture Modelling
b
= β1 +β2 +β3 +β4 +β5
target
c
0cM
50cM
d
1.0000
0.9995
0.9990
0
10
20
30
40
50
Data
Fitted date
e
1.019
1.01
1.002
0.993
0.984
0
10
20
30
40
50

Figure 2.3: Steps in inferring admixture event from genotype data including haplotype phasing, chromosome painting to paint the target haplotypes using donor haplotype based on genetic similarity (corresponding colors), modelling target groups as mixtures of surrogate populations using copy vector, building coancestry curve – all possible chunk pairs corresponding to contributing donors (nodes) are compared to measure genetic distance, and curve fitting the observed coancestry curve (black) is fitted with expected coancestry curve (green) for admixture date.

Chapter 3

fastGLOBETROTTER Method

In this section, I describe a new algorithm that can more efficiently and accurately infer admixture events relative to GLOBETROTTER. In particular, the algorithm reduces complexity by 1) a sub-sampling method that preferentially selects informative pairs of chunks from all possible pairs when constructing the co-ancestry curve, 2) merging donors that appear to be genetically related and 3) novel code optimization. I also describe other improvements over the original method, such as removing non-admixture related signals from the co-ancestry curves and a new jackknifing method for inferring confidence intervals.

3.1 Sub-sampling informative chunks for building co-ancestry curves

As with GLOBETROTTER, fastGLOBETROTTER takes as input sampled realizations of the "painting" of each haploid genome of a target individual. This painting is inferred under the CHROMOPAINTER HMM model as described as $\vec{Y}$ in Section 2.2.2. Recall that a "chunk" refers to a contiguous DNA segment within a haploid genome of the target individual that is painted entirely by a single donor haploid. (Note that each SNP is assigned to a single chunk, while each chunk typically contains multiple SNPs.) Each painting sample is therefore the set of all such chunks across the entire haploid genome of the target individual, with this set changing across different random samples from the HMM model. In the original GLOBETROTTER algorithm, all pairs of chunks separated by < 50 cM within and between all painting samples of a target individual's two haploids are compared when tabulating the counts that two chunks are copied from a particular surrogate pair over various genetic distances (Figure 3.1:Left). Recall that these counts are used to generate "coancestry curves" that capture the decay of linkage disequilibrium attributable to admixture (e.g. Figure 2.3).

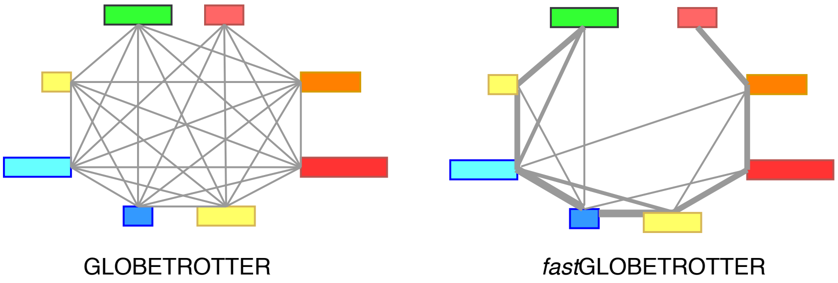

Figure 3.1: Schematic of building co-ancestry curves. Left – original GLOBETROTTER and Right – fastGLOBETROTTER. Nodes are chunks colored according to the donor it is painted by, and edges link chunk pairs used to build the coancestry curves. In fastGLOBETROTTER, chunk pairs separated by short distances are sampled more frequently than chunk pairs separated by larger distances. Increasing edge widths indicates an increased probability of sampling the connected chunk pair.

In contrast, fastGLOBETROTTER samples only a subset of chunk pairs, using a sampling distribution that preferentially chooses the chunk pairs that are the most informative for admixture when constructing the co-ancestry curves. I hypothesize that the shorter the distance between chunks, the more informative they are. This is based on the fact that for dates of admixture >7 generations ago, nearly all of the linkage disequilibrium between chunks that is attributable to genuine admixture decays to ~ 0 by 30cM. Therefore, I exclude chunk pairs separated by $\geq$ 30cM, as can be done in the original GLOBETROTTER. To save computational time, rather than considering all pairs of chunks separated by $<$ 30cM, I instead only consider a subset (e.g. $1/30$th) of the total chunk pairs. To do so, I use a sampling distribution whereby chunk pairs separated by shorter distances are preferentially considered over chunk pairs separated by longer distances (Figure 3.1:Right). As I demonstrate via simulations, this not only saves computational time but can improve the accuracy of inferred dates by reducing the random noise introduced by fitting chunk pairs that are separated by long distances and hence contain little or no information about the admixture event.

3.1.1 Algorithm

1. Divide a target chromosome into B bins of size XcM (Figure 3.2a)

2. Find which of the C total chunks fall into bin G_i for all $i \in [1,...,B]$. A chunk will be put into bin G_i if the midpoint of the chunk falls within the range of bin G_i (Figure 3.2b). Let N_i be a vector of size B that stores the number of chunks within each bin G_i, such that $\sum_{i=1}^{B} N_i = C$.

3. For each bin G_i, the program will compare the chunks in this bin to the chunks in bin G_{i+1}, where the distance $D_{i \to i+1}$ between G_i and G_{i+1} is X. The program then compares G_i with G_{i+2} (i.e. with distance $D_{i \to i+2} = 2X$) between them) and so on, until reaching the last bin n with $D_{i \to n} \leq K$, where K is the maximum allowed distance between chunks (e.g. 30cM; arrows in Figure 3.2c), or until reaching the end of the chromosome.

4. To do the comparison in 3, we do the following:

   4a. For each i and j , where $i < j$, calculate $Y_{ij} = N_i * N_j * M_{ij}$, which is the number of samplings of chunk pairs from bin i and j to be performed (i.e. with one chunk sampled from bin i and the other chunk sampled from bin j). M_{ij} is a scalar that is derived from the sampling distributions (see Section 3.1.2), which allows us to sample a different proportion of the total chunk pairs in bins i and j. For example, if $M_{ij} = 1$, an equivalent number of chunk pairs will be sampled as in the original GLOBETROTTER. Alternatively, one could make $M_{ij} < 1$ while having M_{ij} larger for smaller values of $D_{i \to j}$ relative to larger values of $D_{i \to j}$, meaning closer chunk pairs are preferentially sampled more than distant chunk pairs.

   4b. To compare chunks in G_i and G_j, the program randomly samples Y_{ij} chunk pairs without replacement, with one chunk from G_i and one chunk from G_j.

5. 5. Do step 4 for all pairs of bins (G_i, G_j) across the chromosome separated by $\leq K$cM.

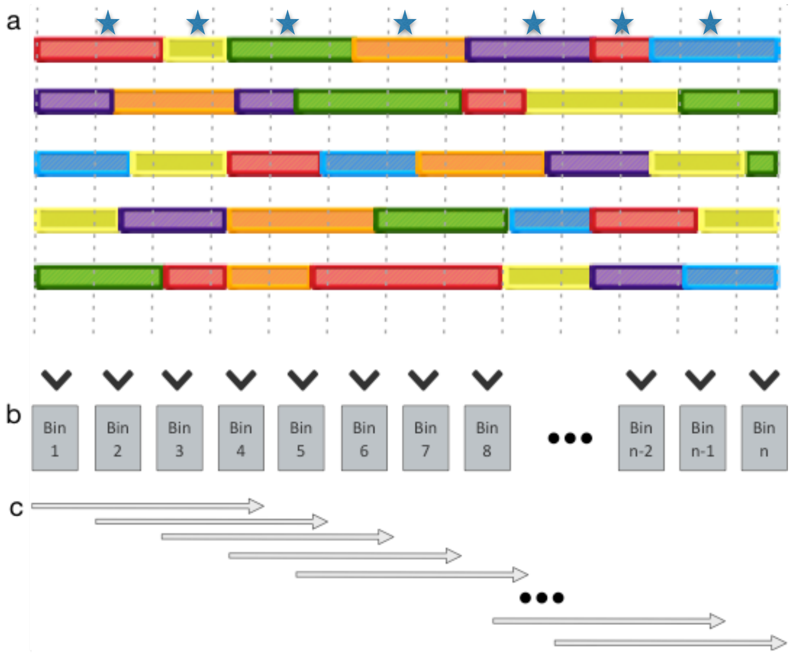

Figure 3.2: Illustration of fastGLOBETROTTER's chunk sampling algorithm a) Painting samples generated from CHROMOPAINTER, where chunks of SNPs are colored according to which donor population they are "painted by" (i.e. a close genetic match to). The grey dotted lines represent the range of bins that divide chromosome. b) Chunks are collected into each bin if the midpoint (star) of the chunk is in between the bin's range. c) Arrows – sliding windows indicate the included bins to compare chunk pairs.

3.1.2 Optimization for sampling distributions

To define the scalar M_{ij} that determines the number of chunk pairs to sample from bins G_i and G_j separated by some genetic distance $D_{i \to j}$, I introduce six different sampling distributions (Figure 3.3), with parameters defined so that the total number of chunk-pairs sampled is $1/30$th ($\sim 3\%$) that sampled by the original GLOBETROTTER when comparing all chunks within a distance of KcM:

1. Constant M_{ij} - the number of chunk pairs sampled for any two bins i and j with $D_{i \to j} < K$ are sampled equally regardless of the distance between them. In this case we use M_{ij}= 0.03, in order to reduce the number of overall sampled chunk pairs to $\sim 3\%$ of the total possible chunk pairs used in the original

GLOBETROTTER, for a fair comparison across all sampling distributions described below.

2. Linear decay - the number of chunk pairs sampled from bins G_i and G_j separated by distance $D_{i \to j}$ decrease linearly, according to the equation $M_{ij} = -0.001(j-i)+0.06$. This focuses on sampling more chunk pairs from closer bins while linearly decrease the number of samples as G_i is further from G_j. The slope and intercept values are chosen so that the total number of sampled chunk pairs is equal to ∼3% of the total possible chunk pairs.

3. Exponential decay - the number of chunk pairs sampled from two bins exponentially decreases as the distance between bins G_i and G_j increases. In particular $M_{ij} = exp^{-\gamma D_{i \to j}}/c$, where c is an arbitrary number to control the total number of chunk pairs sampled to be ∼3% of the total possible chunk pairs. An increasing γ denotes an increased focus on chunk pairs sampled from nearby bins. Here I tried $\gamma = 0.15$, 0.1, 0.05, and 0.03, with corresponding c equal to 5, 8, 10, and 19, respectively.

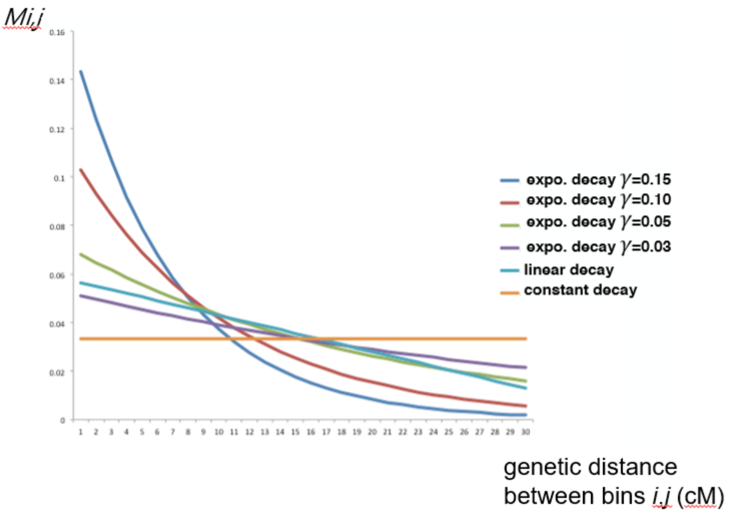

Figure 3.3: Sampling distributions, where the area under all curves is identical to allow a fair comparison, i.e. to fix the total number of compared chunk pairs, in this case ∼3% of all possible chunk pairs separated by 30 cM.

We validate the best sampling distributions for fastGLOBETROTTER by applying each to several simulated datasets in Chapter 4.

3.2 Introduce sub-sampling algorithm to null individual analysis

In this "null individual" step, GLOBETROTTER builds a "null" coancestry curve that reflects the relative probability that two chunks separated by some genetic distance that come from different target individuals are matched to a particular pair of surrogates.

Hellenthal et al 2014 describe a "null individual" analysis that aims to eliminate LD decay signals in the coancestry curves that are not attributable to admixture, hence providing more reliable date estimates. This is done by building a "null" coancestry curve using chunk pairs where each chunk is from the painting sample of a different target individual, this "null" coancestry curve should be unrelated to the admixture event because such chunk pairs on different individuals cannot fall within a single block of DNA inherited intact from an admixing source group. GLOBETROTTER scales the coancestry curve $\widehat{\psi}(m,n;g)$ described in Section 2.2.4 by the "null" coancestry curve before inferring dates and proportions of admixture. I implement an algorithm similar to that described in section 3.1.1-3.1.2 to the null individual analysis, by adding a check of whether the randomly sampled chunks are from the same individual or not. If they are from the same individual, then we resample again until they are from different individuals. Therefore, I replace two more steps in section 3.1.1 as follows:

- Step 2. Let P_{null} be a vector of size equal to number of total chunks C which keeps – for each chunk – the index of the individual to which that chunk belongs

- Step 4b. To compare chunks in G_i and G_j, the program randomly samples chunk pairs, with one chunk from G_i (call this chunk a_i) and one chunk from G_j (call this chunk b_j). I then consider the donors copied in these two chunks when building coancestry curves only if $P_{null}(a_i) \neq P_{null}(b_j)$. This process is repeated until it reaches Y_{ij} comparisons.

3.3 Code optimizing to speed up calculations

The computational time of the original GLOBETROTTER algorithm can be described as:

$$O[NC(B+M)(SL+J^2I+J^2I_j^2+GJ^2K^2)+C[\min(N;100)]^2(L+I_j^2)], \qquad (3.1)$$

where N is the number of the target population individuals, C the number of chromosomes, B the number of bootstrap re-samples (to infer uncertainty in the date estimation), M the number iterations of inferring dates and inferring source groups and admixture contributions, S the number of painting samples, L the maximum number of SNPs across chromosomes, J number of donor populations, I the maximum number of chunks across chromosomes and individuals, I_j the maximum number of chunks across chromosomes and individuals that are copied from a single donor population j, G the number of bins, and K the number of surrogates.

The first term $NC(B+M)$ is the computation cost of GLOBETROTTER reading in the input file and performing bootstrapping and mixing iterations. Next, SL is the cost of the step that assigns a unique label to each chunk across all painting samples for a given chromosome, and $J^2I+J^2I_j^2$ is the cost of comparing all possible chunk pairs to calculate the coancestry curves. GJ^2K^2 is the cost of GLOBETROTTER retaining information in bins that vary according to genetic distance in order to do the re-weighing step described in Section 2.2.4. And finally, $C[\min(N;100)]^2(L+I_j^2)$ is attributable to a so-called "null individual" analysis that aims to eliminate LD decay signals in the coancestry curves that are not attributable to admixture. In this "null individual" step, for each chromosome C and $\min(N,100)$ target individuals, GLOBETROTTER compares every possible pair of chunks on two different individuals that are separated by a user-supplied maximal distance in cM. The memory requirement of GLOBETROTTER are $O(NG\tilde{K}^2)$, where $\tilde{K}$ is the number of surrogates (out of K total) that have $\beta_k > 0.005$ as estimated in Section 2.2.3. In practice $\tilde{K}$ is often perhaps a factor of 10 smaller than the number of donors J.

Generally, the co-ancestry curve calculation is dependent on the number of target individuals N and the number of chromosomes C (term $SL + J^2 I + J^2 I_j^2 + GJ^2K^2$ in Equation 3.1). It is possible to speed up this calculation by performing some calculations once across all C.

Recall the equation to create the co-ancestry curve in the Equation 2.11, where GLOBETROTTER tabulates the counts $\phi_r(i, j; g)$ of chunk pairs separated by distance g in target individual r, where one chunk in the pair is painted by donor i and the other by donor j. This is done for all chunks pairs $\chi_{a,b;g}$ separated by distance g in painting samples a,b of target individual r. This step has cost $NC(J^2 I + J^2 I_j^2)$ in Equations 3.1 and 3.2. Later, this raw co-ancestry curve $\phi_r(i, j; g)$ is adjusted using a re-weighting to obtain the adjusted curve $\phi_r^*(g)$ in Equation 2.12. The computational cost of this re-weighting step across all individuals is represented in the term $NC(GJ^2K^2)$ in Equation 3.1 and is often the most time consuming part of GLOBETROTTER given typical values of J, K, I, I_j and G. To save memory, GLOBETROTTER does this computation once per chromosome per individual. Instead this can be done once per individual, by summing the $\phi_r(i, j; g)$ across chromosomes prior to the re-weighting. However, the trade-off for doing so is that the $\phi_r(i, j; g)$ must be stored across all target individuals r, in order to avoid having to re-read in the painting samples generated by CHROMOPAINTER that GLOBETROTTER takes as input. This increases the memory component of GLOBETROTTER for this step from $N\tilde{K}^2G$ to NJ^2G. I have implemented an option into fastGLOBETROTTER where users are informed of the different memory costs of using this approach, which requires only a computationally quick estimate of the β_ks.

This adjustment drops the C term multiplied by GJ^2K^2 from Equation 3.1 to give the following new computational complexity:

$$O[(B+M)(NC(SL+J^2I+J^2I_j^2)+NGJ^2K^2)+C[min(N;100)]^2(L+I_j^2)] \quad (3.2)$$

3.4 Combine donors to minimize computational complexity

Both the computation time and memory depend on the number of donors J, with the computational increase in section 3.3 giving a memory increase of order J^2. In addition to reducing time, it is compulsory to efficiently minimize the memory used by this method if we aim to handle large datasets. To do both, we can reduce the number of donors J, because this is usually the largest memory contributor amongst all terms when using GLOBETROTTER. To do this, we propose a method to combine donors that share a similar genetic background and merge them into a new group. In order to measure the genetic similarity of each pair of donors, we make use of the "copy vector" from Equation 2.8, whose elements $\hat{f}_{kj}$ contain the average genome-wide proportion of DNA by which surrogate group $k \in [1,...,K]$ copy from donor group j under CHROMOPAINTER. We define $\hat{f}_{\cdot i} = \hat{f}_{1i},...,\hat{f}_{ki}$ and create a correlation matrix R_{ij} for all $i, j \in [1,...,J]$:

$$R_{ij} = corr(\hat{f}_{\cdot i}, \hat{f}_{\cdot j})$$

We apply Pearson's correlation defined as

$$R_{ij} = \frac{\sum_{k=1}^{K}(\hat{f}_{ki} - \bar{\hat{f}}_{\cdot i})(\hat{f}_{kj} - \bar{\hat{f}}_{\cdot j})}{\sqrt{\sum_{k=1}^{K}(\hat{f}_{ki} - \bar{\hat{f}}_{\cdot i})^2 \sum_{k=1}^{K}(\hat{f}_{kj} - \bar{\hat{f}}_{\cdot j})^2}}, \tag{3.3}$$

where $\bar{\hat{f}}_{\cdot i} = \sum_{k=1}^{K} \hat{f}_{ki}/K$.

Since donors that exhibit high correlations reflect a similar pattern of genetic contribution to recipient groups, it is natural to combine them as a new group. If donor i and j are combined as donor m, we define a new vector of donor matching $\hat{f}_{\cdot m}$ with element $\hat{f}_{\cdot m} = 0.5 * (\hat{f}_{\cdot i} + \hat{f}_{\cdot j})$.

Figure 3.4 illustrates an example of copy vectors derived from painting individuals from 51 Europeans population (c1-c51) using 162 donor groups. It is possible that we can reduce the number donors, for example, if Sweden and Norway contribute similar relative amounts to each recipient group, then they are combined as

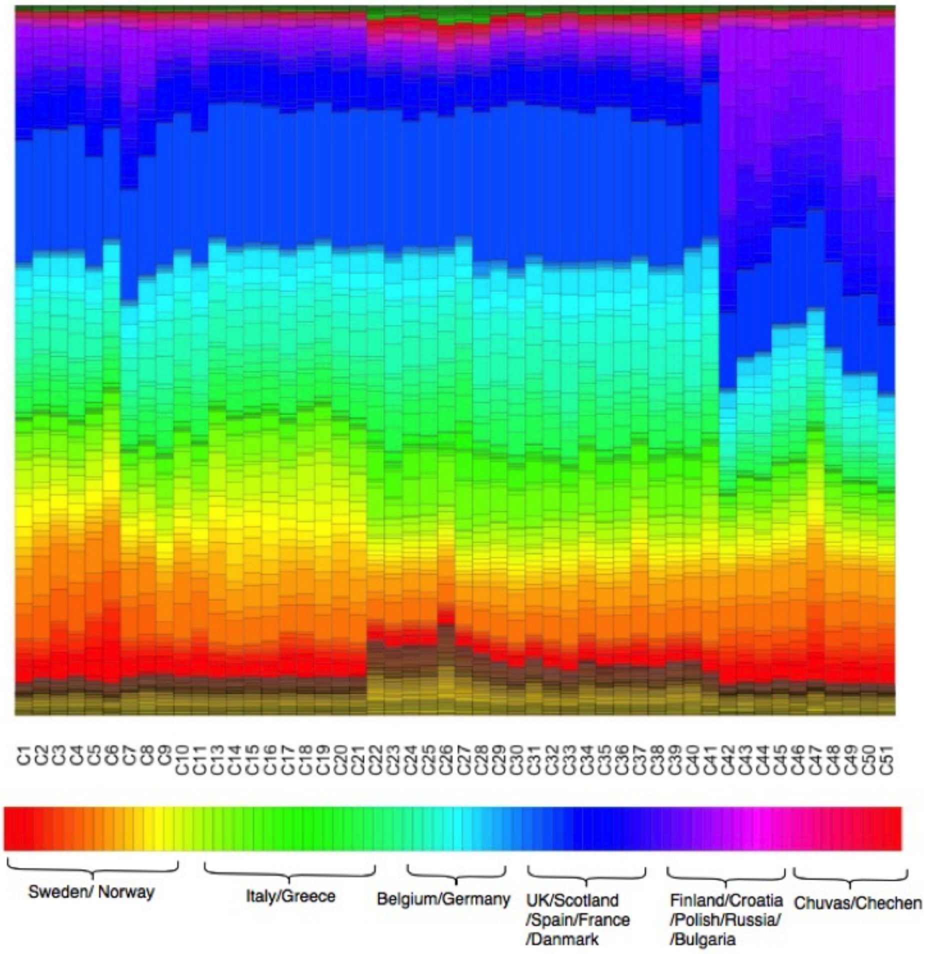

Figure 3.4: Example of copy vector where recipient individuals from European populations c1-c51 (x axis) are painted by donor populations that include c1-c51 plus 162 other populations (y axis). The below legend are donor populations and the correspondent colours that give the color code for each recipient population. (This dataset is described in detail in Chapter 5.)

a new donor group. In practice, I tried merging donor groups i, j where $R_{ij} > 0.95$. This relatively large value should ensure that the new copy vectors, while containing fewer elements (less colors in Figure 3.4), still preserve the main patterns of differentiation between recipient groups.

3.5 Removing non-admixture related signals in the co-ancestry curve

As noted previously, GLOBETROTTER performs a "null individual analysis" that mitigates signals in the LD decay curve that are not due to genuine admixture but may arise instead from e.g. bottleneck events experienced by the target population (or other factors that lead to signals of strong genetic drift). However, another potential issue that this procedure does not account for, but which may also occur in drifted groups, is that many chunks may be atypically long under the CHROMOPAINTER analysis. To cope with this, GLOBETROTTER ignores any chunk pairs separated by $\leq$1cM when fitting co-ancestry curves, as such within-population (or "background") LD may extend to such lengths in drifted groups. However, the 1cM threshold is arbitrary; related methods (e.g. ALDER [19]) try to automatically identify the threshold of minimal distance between segments to use. A particular concern is that the presence of many atypically long chunks over 1cM can lead to a small number of chunk pairs separated by genetic distances just above 1cM relative to that expected (see Figure 3.5a). This coancestry curve pattern looks similar to that expected under multiple distinct pulses of admixture involving different groups admixing at different times (see Figure 2.3e), and hence can lead to inaccurate admixture inference.

To cope with this issue, we propose a method for automatically detecting whether the left end of the coancestry curve is affected by long chunks, and then removing this part of the curve prior to model fitting. For example, Figure 3.5a shows the co-ancestry curve of a Finnish group, showing the scaled probability that two chunks separated by XcM have both chunks most recently related ancestrally to surrogate HB:welsh. While these curves should be monotonically decreasing (see Equation 2.2), there is an increase in scaled probability at the left end of the curve. It appears that the LD decay due to admixture does not begin until 4-5 cM, at which point the curve begins to decay as expected.

To find the portion of the left-end part of the coancestry curve to remove, we first analyse the coancestry curve for the highest contributing surrogate group m,

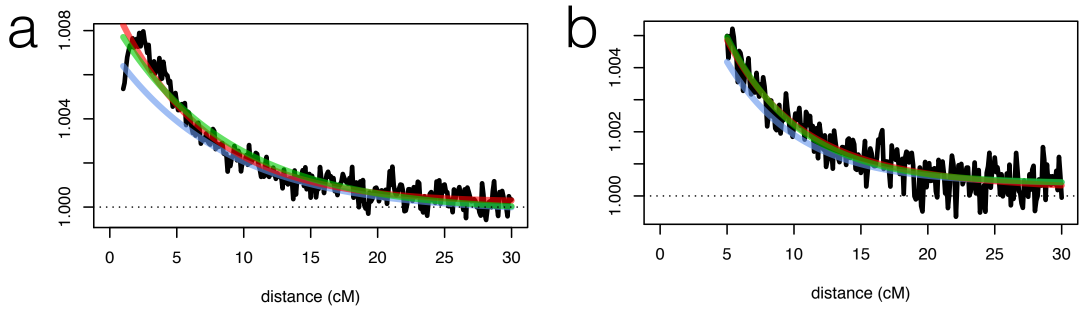

Figure 3.5: Co-ancestry curves from GLOBETROTTER null individual analysis on a Finnish group a) noticeable unwanted signal in the beginning of the curve and its fitted admixture date (green line). b) desirable co-ancestry curve after removing unwanted signal fitted for admixture date.

i.e. the group m where $\beta_m = \max_{k=1}^{K}(\beta_k)$ with β_ks estimated using Equation 2.9. The co-ancestry curve involving only surrogate m, $\widehat{\psi}(m,m;g)$, is likely to be the most informative curve given its high contribution, and hence the most suitable for detecting this non-admixture related signal. In the scenarios I considered, such as the European analysis of Chapter 5, adjusting for this maximal curve seemed to fix this trend in all other curves as well, though this may not extend to every scenario. Roughly, the aim is to search for the peak of this co-ancestry curve and remove the area prior to the the peak, given that $\widehat{\psi}(m,m;g)$ should be monotonically decreasing with increasing g for all m. To do so, we use a sliding window of fixed size M that moves across the x-axis (genetic distance) of the curve, within each window calculating the slope of a straight line fit to the data within that window. In practice, we use multiple different window sizes in order to make this method applicable to all types of curve regardless of their level of noise. Amongst all window size, we report the maximum of bin number where a negative slope is first reported i.e. possibly where the left-end part to be removed is located, and remove the left-end area from the coancestry curve.

More formally, these are the steps of the algorithm:

1. For a windows of size $M = 3$ bins, we move as a sliding window along the co-ancestry curve $\widehat{\psi}(m,m;g)$

   a. For all bins $i,...,i+M$ that fall within window M, we apply a linear model to fit the curve $\widehat{\psi}(m,m;i,...,i+M)$ with a straight line $Y = SX + c$ to obtain a slope S

b. If S is negative, let $V_M = i + ((M+1)/2)$ and go to Step 2

c. Move along the curve by $i = i+1$ and repeat step 1a abd 1b

2. Repeat step 1 for M =5, 7, 9, 10 and 11 bins

3. Remove $\widehat{\psi}(m,m;1,...,max(V_M))$ from $\widehat{\psi}(m,m;g)$

I tested this procedure in different drifted populations i.e. Finnish, Melanesian and Icelandic, and visually assessed whether the increasing left part of the curve was removed for all $\widehat{\psi}(m,m;g)$.

3.6 Introduce delete-m Jack-knife method

The objective of this section is to provide an additional option for measuring the confidence of the inferred admixture date. The algorithm GLOBETROTTER uses bootstrap re-sampling of individuals, and it is not always possible to perform the bootstrap resampling, for example, when inferring admixture in a single individual. Therefore, I implemented an alternative jackknifing procedure to fastGLOBETROTTER, which instead drops one chromosome at a time and estimates the dates using data from the other 21 chromosomes. This gives 22 estimated values, which can then be used to give confidence intervals for the inferred admixture date using previously derived jack-knifing formulas as in Busing et al.,1999 [56].

Chapter 4

Simulations

In this chapter, I describe simulations I used to validate fastGLOBETROTTER. I measured the efficiency in terms of inference accuracy and computational time used by fastGLOBETROTTER compared to the original GLOBETROTTER [21].

4.1 Simulation details

4.1.1 European Simulations

The first set of simulated admixed individuals I considered were those used to mimick admixture found in the UK population [57]. In particular, the authors made simulated individuals that were mixtures of real sampled individuals from Italy and northern Germany, assuming a single pulse of admixture between these two populations occurring $\lambda = 40$ generations ago. The proportion of admixture α from northern Germany varied from α=0.1, 0.25 and 0.5. Briefly, this simulation process, described in Price et al., 2009 [22], starts by (1) sampling a centimorgan (cM) genetic distance g from an exponential distribution with rate $\lambda/100$. Then the authors (2) generated the first g cM of a simulated chromosome as the first g cM of a randomly sampled real chromosome from population A or B (i.e. Italy or northern Germany), with A or B chosen according to probabilities α and $1 - \alpha$, respectively. The next g cM is then generated by repeating these two steps, and this process is repeated until the end of the chromosome is reached. To simulate a single haploid genome, the 22 haploid autosomes have to be generated in this way. If there are N simulated individuals, then the steps are repeated $2N$ times, with $N = 25$ in these simulations. I name this simulation “POBI” which is short for “People of the British Isles”, the project for which these simulations were generated (www.peopleofthebritishisles.org/). In summary, the admixture parameters of these simulations were:

- POBI1 ($\alpha = 0.10$, $\lambda = 40$, $N = 25$ samples)
- POBI2 ($\alpha = 0.25$, $\lambda = 40$, $N = 25$ samples)
- POBI3 ($\alpha = 0.50$, $\lambda = 40$, $N = 25$ samples)

The Italy and north Germany samples are taken from Multiple Sclerosis data, from which Leslie et al 2015 used 6,209 sampled individuals from ten countries in continental Europe typed at ≈500,000 SNPs as part of the Wellcome Trust Case Control Consortium 2 (WTCCC2) study [58]. As in the POBI paper, a set of 51 surrogate populations, defined based on the clusters used in [57] comprising these 6,209 individuals, were used as proxies to the admixing sources. Included among these clusters were the two (from Italy and north Germany) containing the individuals used to generate the simulated haploids, though the 40 individuals used to simulate were removed from these two clusters prior to analysis. The “target populations” consisted of the simulated individuals from POBI1, POBI2 and POBI3. All remaining 6,169 individuals from the 51 surrogate populations were also used as donors when painting each surrogate and target individual using CHROMOPAINTER. If successful, the admixture inference from GLOBETROTTER and fastGLOBETROTTER for these three simulations would identify that two ancestral groups genetically related to Italy and northern Germany intermixed at time λ, with inferred admixture proportions related to the true proportion α. G.Hellenthal provided the CHROMOPAINTER output for these simulations used in the POBI paper, which was then input into each of GLOBETROTTER and fastGLOBETROTTER. However, I also generated new CHROMOPAINTER output for these simulations as noted below.

4.1.2 Cross-continent Simulations

The simulated datasets in this section were generated using 95 worldwide human populations (Figure 4.1) described in Hellenthal et al., 2014 [21] to assess the performance of GLOBETROTTER when it was first introduced. Each admixed genome was simulated to mimic admixture occurring λ = 7, 30 or 150 generations ago (recent, moderate and ancient admixture), or approximately 200 to 4400 years ago assuming a generation time of 28 years, with varying admixture proportions α = 0.05, 0.20 and 0.50. The simulation process is analogous to that described for the POBI dataset. The data were again simulated using real data, simulating mixing between combinations of the Brahui from Pakistan, Yoruba from Nigeria, French from France, Han from China, and Colombians from Colombia. The pairs of populations mixed and sample sizes N per simulation are as follows:

- Brahui–Yoruba (N = 20 samples)
- Yoruba–French (N = 20 samples)
- Brahui–Han (N = 20 samples)
- Colombian–Han (N = 7 samples)
- French–Brahui (N = 20 samples)

With the different combinations of λ and α, this results in 5×3×3 = 45 simulations in total.

In contrast to POBI, here all individuals from the simulating populations were not used as surrogates or donors, i.e. the entire populations were excluded from downstream analysis, when testing the simulated populations for admixture. Instead the remaining $95-2=93$ populations were used as surrogates for each scenario above. The authors stated that these simulations represent a wide range of difficulties, from easier to more challenging problems. For example, it is reasonably easier to detect admixture between more distantly related populations (e.g. Yoruba intermixing with French) than closely related populations (e.g. French intermixing with Brahui). In addition, lower admixture proportions α and older dates of admixture λ are each likely to be more difficult to detect than more recent admixture with proportions nearer to 50/50%. And also, the number of samples N represents the content of admixture signal, making Colombian–Han a more challenging case. The dates and proportions used in these simulations partially reflect the real genetic history of modern populations, as inferred in Hellenthal et al., 2014.

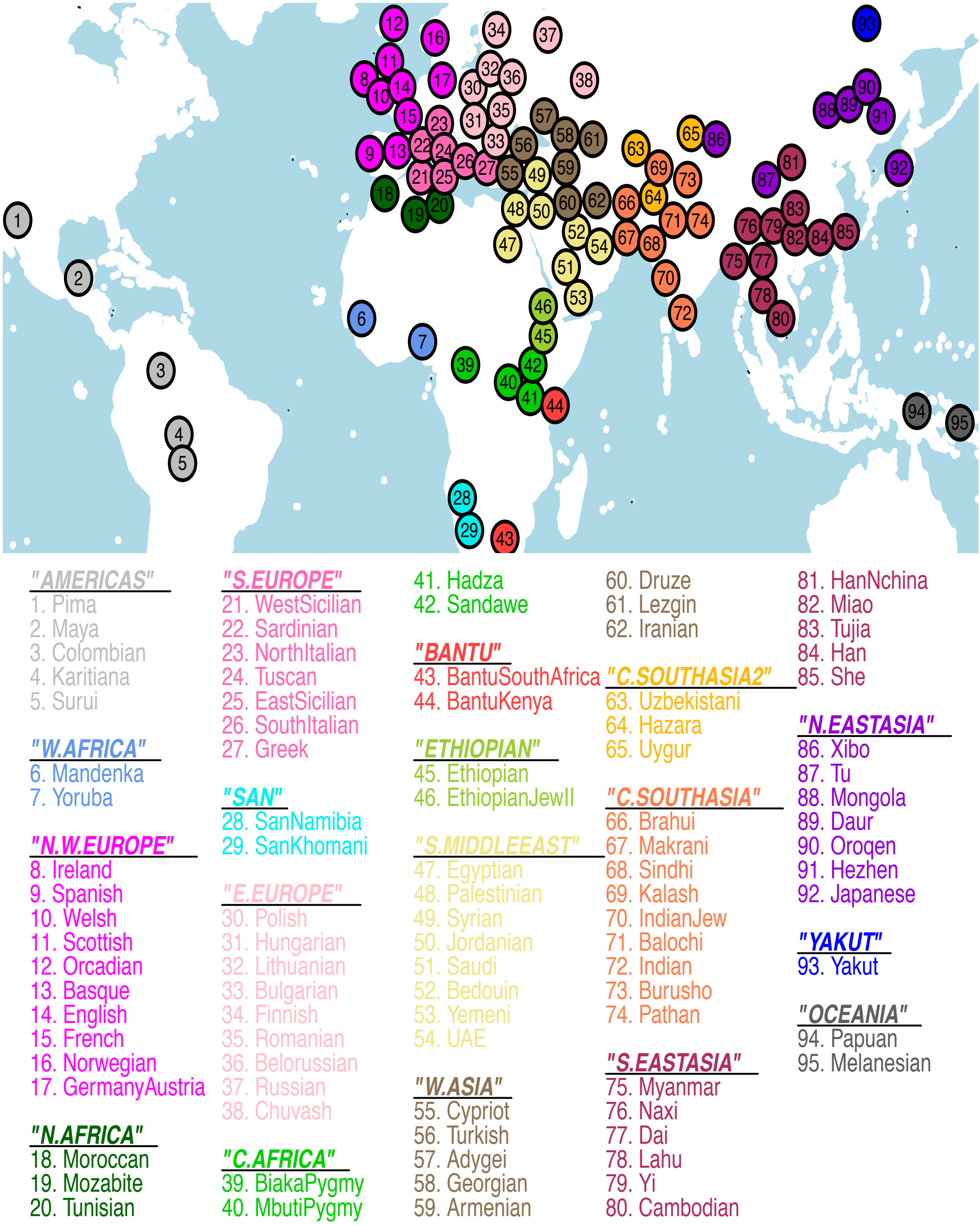

Figure 4.1: The geographical location of 95 world populations taken from Hellenthal et al, 2014 [21], with colors according to sub-continental regions listed at bottom.

4.2 The assessment of suitable sampling distributions

According to Section 3.1.2, most of the admixture signal in the coancestry curves has decayed by ~ 30cM. Therefore, rather than recycling the results from the GLOBETROTTER paper, which fit the coancestry curves over the range of [1,50]cM when inferring dates (i.e. considered all pairs of chunks separated by 1-50cM), I re-ran GLOBETROTTER only fitting the curves over the range [1,30]cM, which is the same range I fit with fastGLOBETROTTER. I evaluated the performance of fastGLOBETROTTER's various sampling distributions (described in Section 3.1.2) that sub-sample from the number of total of possible chunk pairs that fall within this range. The goal of this section is to compare and evaluate of how different sampling distributions improve the performance of fastGLOBETROTTER in terms of inferring admixture dates and proportions, as well as evaluating fastGLOBETROTTER's improvement in computational efficiency over GLOBETROTTER. I compared the six different sampling distributions shown in Figure 3.3 by initially analysing the POBI dataset (Section 4.1.1) to optimize the sampling distributions, as this is a considerably difficult case involving two relatively genetically similar European populations intermixing.

GLOBETROTTER and fastGLOBETROTTER have a built-in means of assessing uncertainty in date estimation, by bootstrap re-sampling of individuals. For simplicity, this bootstrapping fixes the inferred admixture proportions and sources from the original analysis on the full (non-bootstrapped) data. Thus GLOBETROTTER and fastGLOBETROTTER do not have a means of assessing uncertainty in estimated admixture proportions. Therefore, in order to assess consistency of proportion estimates, I ran CHROMOPAINTER on the POBI dataset 50 times to construct 50 different sets of painting samples, inferring dates and admixture proportions on each set of painting samples using fastGLOBETROTTER/GLOBETROTTER. The results are illustrated by box plots summarizing the inferred dates and proportions (Figure 4.2–4.8).

In fastGLOBETROTTER I controlled the total number of chunks sampled in all cases to be the same; as a result the computational time used was approximately 10 minutes for each sampling equation. This is a factor of 7-10 reduction compared to the time it took to run the original GLOBETROTTER that uses all pairings of chunks separated by 30cM. We can see that inferred dates of the exponential decay equation ($\gamma = 0.1$ in Figure 4.6) lie nicely in between the true date as boxes are bal-

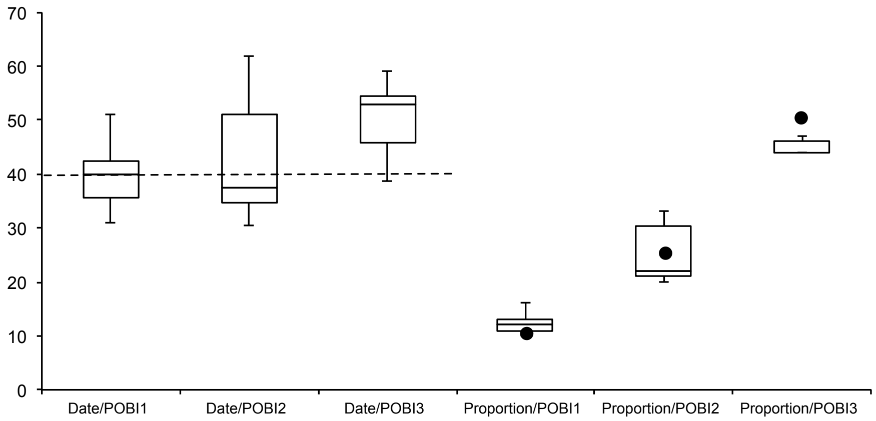

Figure 4.2: Box plot of original GLOBETROTTER inferred dates in generations (first 3 boxes) and admixture proportions in percentage (last 3 boxes) using 50 sets of painting samples of POBI1, POBI2 and POBI3. The dashed line and dots represent the true dates (λ = 40) and proportions (α = 10%, 25% and 50%), respectively.

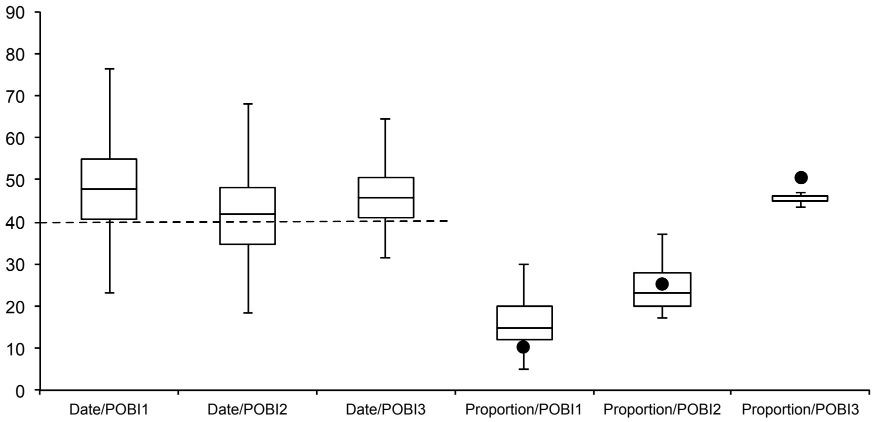

Figure 4.3: Box plot of fastGLOBETROTTER inferred dates in generations (first 3 boxes) and admixture proportions in percentage (last 3 boxes) using "constant equation" on 50 sets of painting samples of POBI1, POBI2 and POBI3. The dashed line and dots represent the true dates (λ = 40) and proportions (α = 10%, 25% and 50%), respectively.

anced around 40, meaning that it is more accurate when performing date inference. We also can see that there is a deviation from the true date in POBI1 and POBI3 when using the constant equation (Figure 4.3). This also occurred in the rest of the cases including linear decay (Figure 4.4). While there are some deviations in the mean inferred date in POBI3 when using the exponential decay fraction with γ=

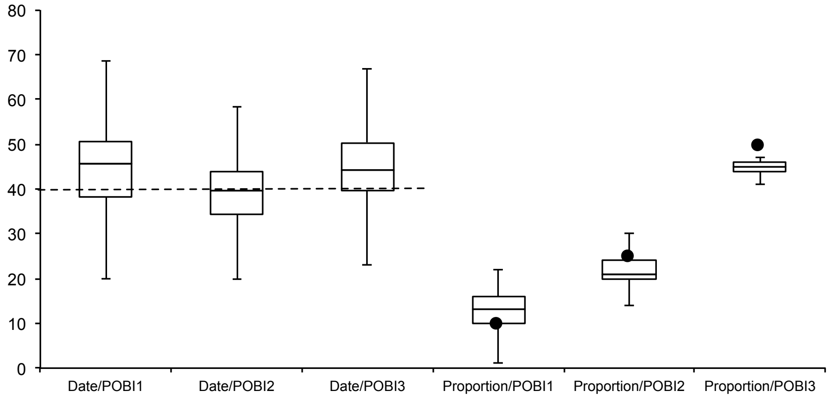

Figure 4.4: Box plot of fastGLOBETROTTER inferred dates in generations (first 3 boxes) and admixture proportions in percentage (last 3 boxes) using "linear decay equation" on 50 sets of painting samples of POBI1, POBI2 and POBI3. The dashed line and dots represent the true dates (λ = 40) and proportions (α = 10%, 25% and 50%), respectively.

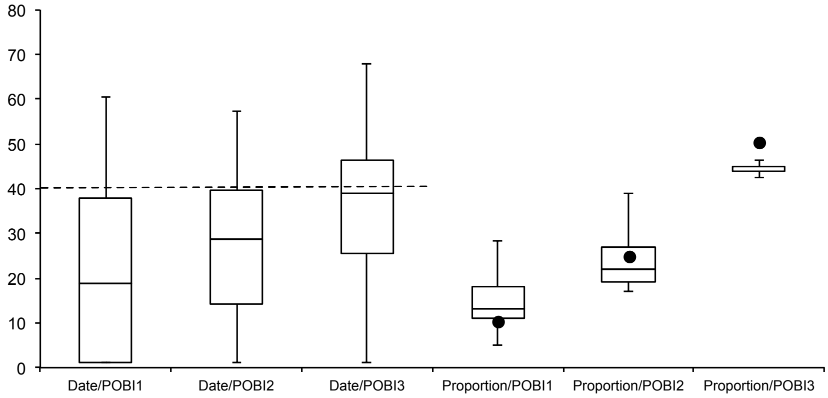

Figure 4.5: Box plot of fastGLOBETROTTER inferred dates in generations (first 3 boxes) and admixture proportions in percentage (last 3 boxes) using "exponential decay equation with γ = 0.15" on 50 sets of painting samples of POBI1, POBI2 and POBI3. The dashed line and dots represent the true dates (λ = 40) and proportions (α = 10%, 25% and 50%), respectively.

0.05, 0.03 (Figure 4.7–4.8), fastGLOBETROTTER performs worst in the case γ= 0.15 (Figure 4.5) where all date inferences are notably deviated from the truth.

The admixture proportions in most of the cases were well estimated, as we can see that the size of the boxes are small and located near the true proportions. An exception is POBI3 in all cases (Figure 4.2–4.8) where the inferred proportions are always underestimated from the truth (0.50). However, I am plotting the admixture

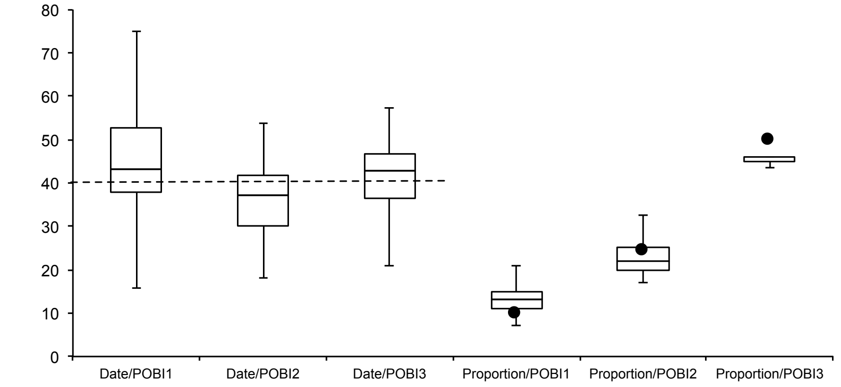

Figure 4.6: Box plot of fastGLOBETROTTER inferred dates in generations (first 3 boxes) and admixture proportions in percentage (last 3 boxes) using "exponential decay equation with $\gamma = 0.10$" on 50 sets of painting samples of POBI1, POBI2 and POBI3. The dashed line and dots represent the true dates (λ = 40) and proportions (α = 10%, 25% and 50%), respectively.

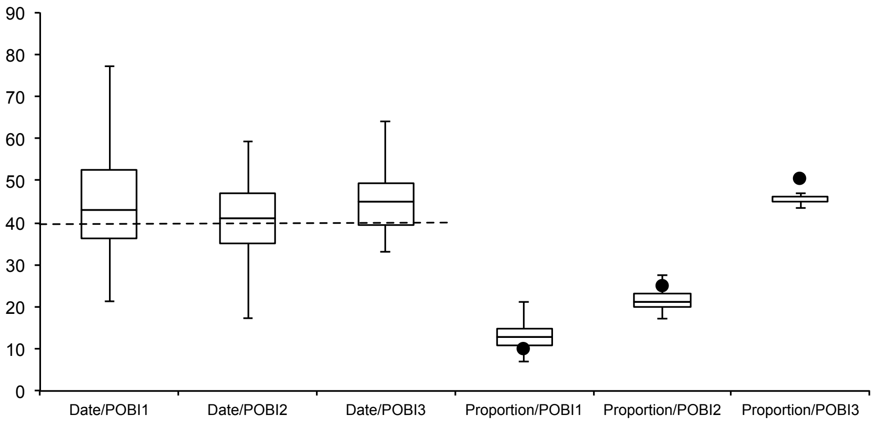

Figure 4.7: Box plot of fastGLOBETROTTER inferred dates in generations (first 3 boxes) and admixture proportions in percentage (last 3 boxes) using "exponential decay equation with $\gamma = 0.05$" on 50 sets of painting samples of POBI1, POBI2 and POBI3. The dashed line and dots represent the true dates (λ = 40) and proportions (α = 10%, 25% and 50%), respectively.

proportion of the minority contributing source here, which necessarily will have $\alpha \leq 50\%$, so this is somewhat expected. From the result shown in this section, I selected the condition of exponential decay with $\gamma = 0.1$ as the best fraction system for our further analysis.

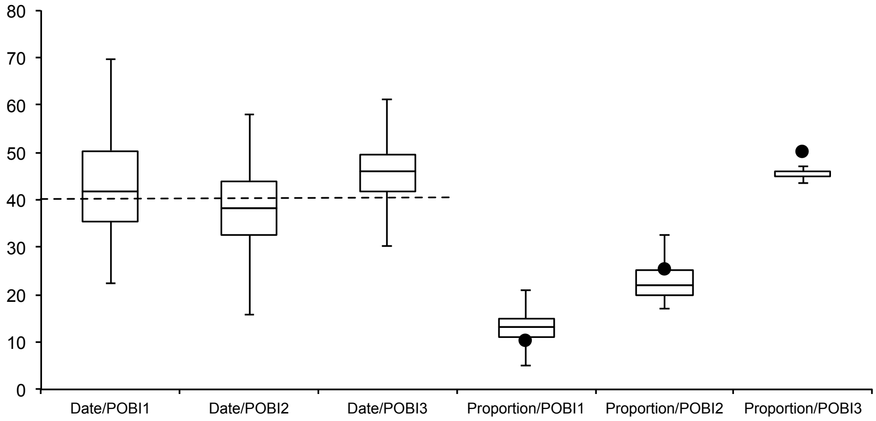

Figure 4.8: Box plot of fastGLOBETROTTER inferred dates in generations (first 3 boxes) and admixture proportions in percentage (last 3 boxes) using "exponential decay equation with $\gamma = 0.03$" on 50 sets of painting samples of POBI1, POBI2 and POBI3. The dashed line and dots represent the true dates (λ = 40) and proportions (α = 10%, 25% and 50%), respectively.

4.3 Validating the performance of fastGLOBETROTTER

In this section, I applied fastGLOBETROTTER with the selected $\gamma = 0.1$ to the second set of simulations (Section 4.1.2). For each of these 45 simulations, I performed GLOBETROTTER/fastGLOBETROTTER analyses to compare the efficiency in terms of time and date inference accuracy using the same computing unit at the Department of Computer Science, UCL. The cluster includes 4x3.5GHz processors with 16GB RAM. The date inference from 100 bootstrap resamples are summarized as a box plot, ordered for each simulation by true admixture date λ = 7, 30 and 150 generations, with results for the true admixture proportions α = 0.05, 0.2, 0.5 given consecutively within each λ.

The Yoruba–French result in Figure 4.9 suggests that the accuracy of fastGLOBETROTTER inference is equivalent to GLOBETROTTER, and it is close to true dates for most of the cases. We can see a little improvement in the case λ = 150, α = 0.05 (the hardest case amongst all cases in this simulation), where fastGLOBETROTTER delivers closer estimations to the truth while GLOBETROTTER tends to underestimate the date.

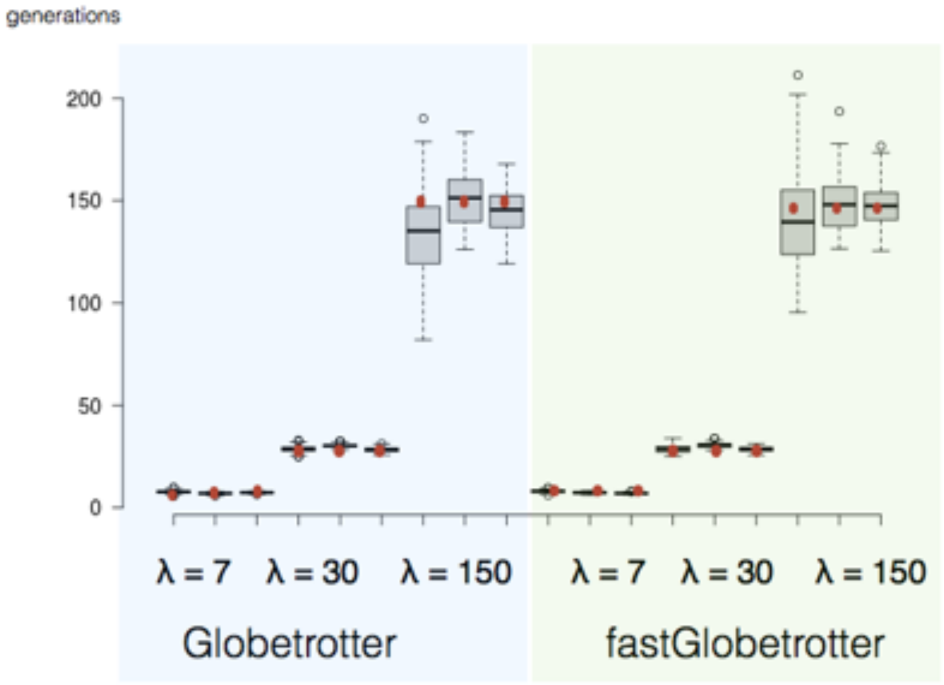

Figure 4.9: Box plot of date inference across 100 bootstrap re-samplings of Yoruba–French simulations. The red dots represent the true dates. For each date (λ), results are given for three different admixture proportions (left-to-right): $\alpha = 0.05, 0.2, 0.5$.

We also see similar performance in Brahui–Han simulation (Figure 4.10), most of cases were inferred correctly and equivalently between fastGLOBETROTTER and GLOBETROTTER. In the case λ = 150, α = 0.05, we see that fastGLOBETROTTER overestimates the dates while GLOBETROTTER infers closer to the truth. Moreover, there is an improvement in the case λ = 150, α = 0.2 for fastGLOBETROTTER.

GLOBETROTTER and fastGLOBETROTTER inference on Brahiu–Yoruba simulation (Figure 4.11) are close to the truth for λ = 7, 30 cases. They both underestimate the date for all cases where λ = 150. Interestingly in the case λ = 150, α = 0.05, the inferences tend to be only $\sim$100 generations for GLOBETROTTER, while fastGLOBETROTTER are around $\sim$130 generations.

In the French–Brahui simulations (Figure 4.12), we see that GLOBETROTTER and fastGLOBETROTTER fail to infer admixture in the case λ = 150, α = 0.05 (blue *) and they both underestimate the dates in the case α = 0.5. However,

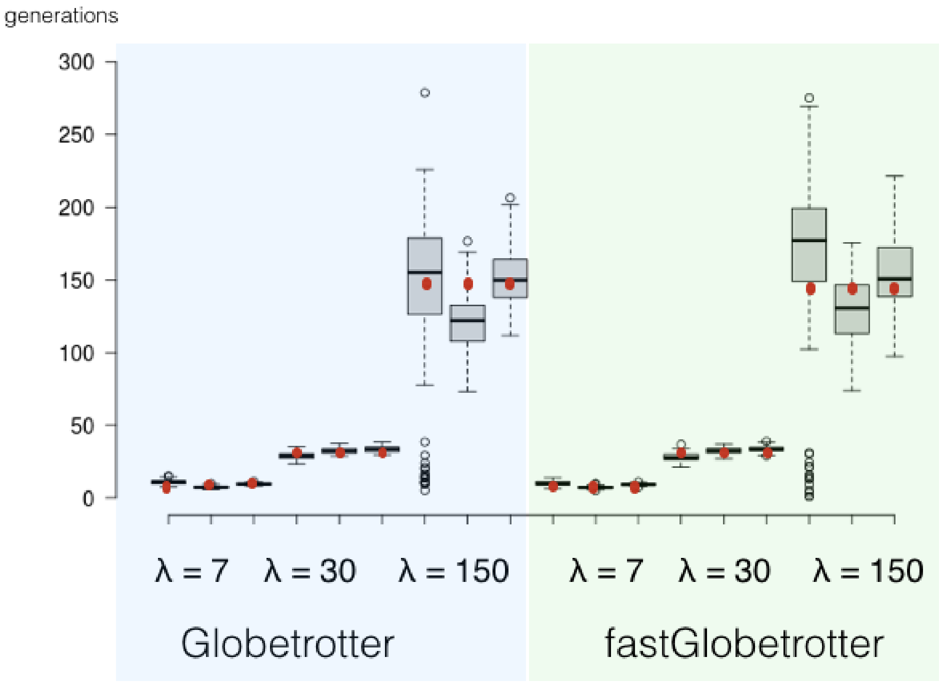

Figure 4.10: Box plot of date inference across 100 bootstrap re-samplings of Brahui–Han simulations. The red dots represent the true dates. For each date (λ), results are given for three different admixture proportions (left-to-right): α = 0.05, 0.2, 0.5.

there is an improvement of fastGLOBETROTTER in the case where λ = 150, α = 0.2.

We see the similar pattern in the last simulation Colombian–Han (Figure 4.13) in that GLOBETROTTER and fastGLOBETROTTER fail to infer admixture in the case λ = 150, α = 0.05 (blue *). GLOBETROTTER underestimates the date in the case λ = 150, α = 0.2 while fastGLOBETROTTER tends to infer better. Moreover, They both have inferred dates closer the truth in α = 0.5, with fastGLOBETROTTER often showing improved inference relative to GLOBETROTTER.

The poor inference of GLOBETROTTER and fastGLOBETROTTER for older dates suggests that such cases, for which the coancestry curve decays more rapidly with increasing genetic distance, are challenging to characterize even when up-weighting nearby segments. To cope with this, the authors of the original GLOBETROTTER suggest that if the default analysis gives date estimates >55 generations, then the algorithm should be re-run fitting only chunks separated by 1-5cM.

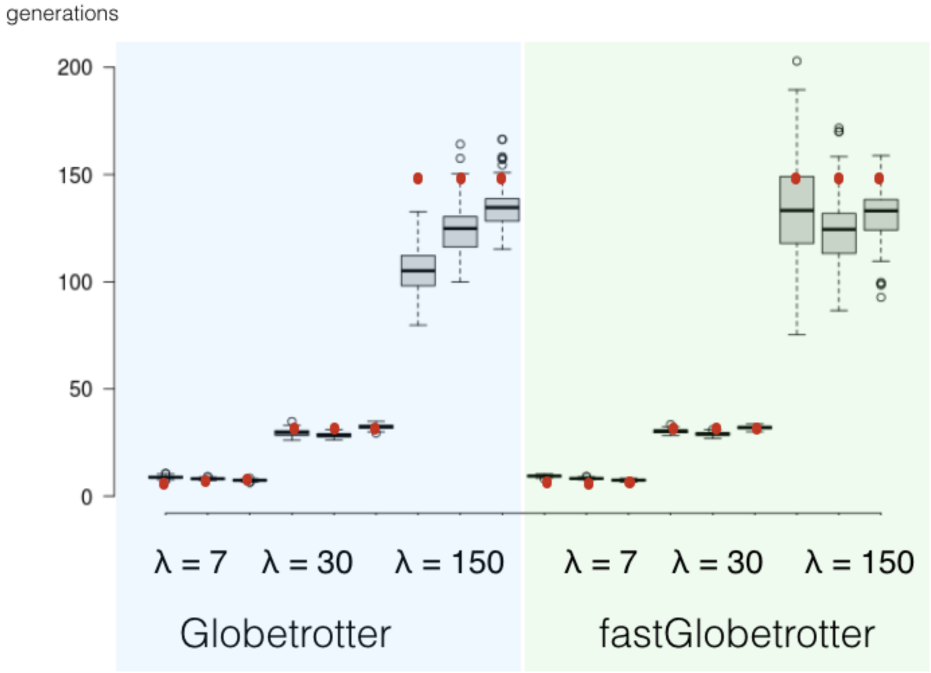

Figure 4.11: Box plot of date inference across 100 bootstrap re-samplings of Brahui–Yoruba simulations. The red dots represent the true dates. For each date (λ), results are given for three different admixture proportions (left-to-right): $\alpha = 0.05, 0.2, 0.5$.

I explored this by re-running GLOBETROTTER and fastGLOBETROTTER using Colombian–Han and French–Brahui simulations with λ = 150 shown in Figure 4.14. The date inference in Colombian–Han in all cases (α = 0.05, 0.2, 0.5) are improved in both GLOBETROTTER and fastGLOBETROTTER. However, in the hardest case (α = 0.05), both models still inferred "no admixture" (blue * in Figure 4.14), reiterating how a small amount of admixture among similar sources is challenging to characterize with these sample sizes.

simulation	true proportion	GLOBETROTTER	fastGLOBETROTTER
Brahui-Yoruba,α=0.05,λ =7	0.05	0.06	0.05
Brahui-Yoruba,α=0.20,λ =7	0.2	0.19	0.19
Brahui-Yoruba,α=0.50,λ =7	0.5	0.48	0.48
Brahui-Yoruba,α=0.05,λ =30	0.05	0.06	0.06
Brahui-Yoruba,α=0.20,λ =30	0.2	0.2	0.2
Brahui-Yoruba,α=0.50,λ =30	0.5	0.47	0.47

simulation	true proportion	GLOBETROTTER	fastGLOBETROTTER
Brahui-Yoruba,α=0.05,λ =150	0.05	0.07	0.09
Brahui-Yoruba,α=0.20,λ =150	0.2	0.17	0.18
Brahui-Yoruba,α=0.50,λ =150	0.5	0.49	0.47
Yoruba-French,α=0.05,λ =7	0.05	0.09	0.09
Yoruba-French,α=0.20,λ =7	0.2	0.26	0.26
Yoruba-French,α=0.50,λ =7	0.5	0.46	0.46
Yoruba-French,α=0.05,λ =30	0.05	0.08	0.08
Yoruba-French,α=0.20,λ =30	0.2	0.26	0.26
Yoruba-French,α=0.50,λ =30	0.5	0.47	0.47
Yoruba-French,α=0.05,λ =150	0.05	0.1	0.11
Yoruba-French,α=0.20,λ =150	0.2	0.24	0.23
Yoruba-French,α=0.50,λ =150	0.5	0.47	0.47
Colombian-Han,α=0.05,λ =7	0.05	0.3*	0.28*
Colombian-Han,α=0.20,λ =7	0.2	0.36*	0.41*
Colombian-Han,α=0.50,λ =7	0.5	0.35*	0.32*
Colombian-Han,α=0.05,λ =30	0.05	0.27*	0.16*
Colombian-Han,α=0.20,λ =30	0.2	0.34*	0.35*
Colombian-Han,α=0.50,λ =30	0.5	0.32*	0.29*
Colombian-Han,α=0.05,λ =150	0.05	0.45*	0.43*
Colombian-Han,α=0.20,λ =150	0.2	0.47*	0.44*
Colombian-Han,α=0.50,λ =150	0.5	0.42	0.36*
Brahui-Han,α=0.05,λ =7	0.05	0.12	0.15
Brahui-Han,α=0.20,λ =7	0.2	0.3	0.3
Brahui-Han,α=0.50,λ =7	0.5	0.49	0.49
Brahui-Han,α=0.05,λ =30	0.05	0.16*	0.13
Brahui-Han,α=0.20,λ =30	0.2	0.3	0.3
Brahui-Han,α=0.50,λ =30	0.5	0.5	0.49
Brahui-Han,α=0.05,λ =150	0.05	0.3*	0.2*
Brahui-Han,α=0.20,λ =150	0.2	0.29	0.28
Brahui-Han,α=0.50,λ =150	0.5	0.48	0.48

simulation	true proportion	GLOBETROTTER	fastGLOBETROTTER
French-Brahui,α=0.05,λ =7	0.05	0.07	0.08
French-Brahui,α=0.20,λ =7	0.2	0.19	0.21
French-Brahui,α=0.50,λ =7	0.5	0.48	0.48
French-Brahui,α=0.05,λ =30	0.05	0.07	0.06
French-Brahui,α=0.20,λ =30	0.2	0.2	0.2
French-Brahui,α=0.50,λ =30	0.5	0.48	0.49
French-Brahui,α=0.05,λ =150	0.05	0.4*	0.28*
French-Brahui,α=0.20,λ =150	0.2	0.22	0.27
French-Brahui,α=0.50,λ =150	0.5	0.46	0.47

Table 4.1: Genetic proportions inferred from GLOBETROTTER and fastGLOBETROTTER based on 1 run applied to 45 simulations. The asterisk (*) indicates the inferences that deviate from the truth more than 0.1.

In summary, these results suggest that the older admixture events are more difficult to detect than the more recent events, as we can see from all simulations. Also, the closer of admixture proportions to 0.5, the more accurately admixture signals are detected in the target individuals. In particular, for some of the cases where α = 0.05, dates tend to be either over or underestimated relative to cases where α = 0.2,0.5. The level of genetic similarity between ancestral sources also affects the methods to identify and characterize admixture; in particular in the French–Brahui simulations, both fastGLOBETROTTER and GLOBETROTTER fail to make inference in some cases.

The genetic proportion inference from both methods is shown in Table 4.1. Due to the algorithm limitation that does not allow bootstrap re-sampling for proportion, the values shown in the table are derived from only one run of each simulation. fastGLOBETROTTER and GLOBETROTTER perform equally well in most of the cases. However, There are some inaccurate inferences (indicated by asterisks) occurring in Colombian-Han simulation. This might due to the genetic relatedness between Colombian and Han, and also the low number of target individuals that also has an impact the performance of GLOBETROTTER and fastGLOBETROT-

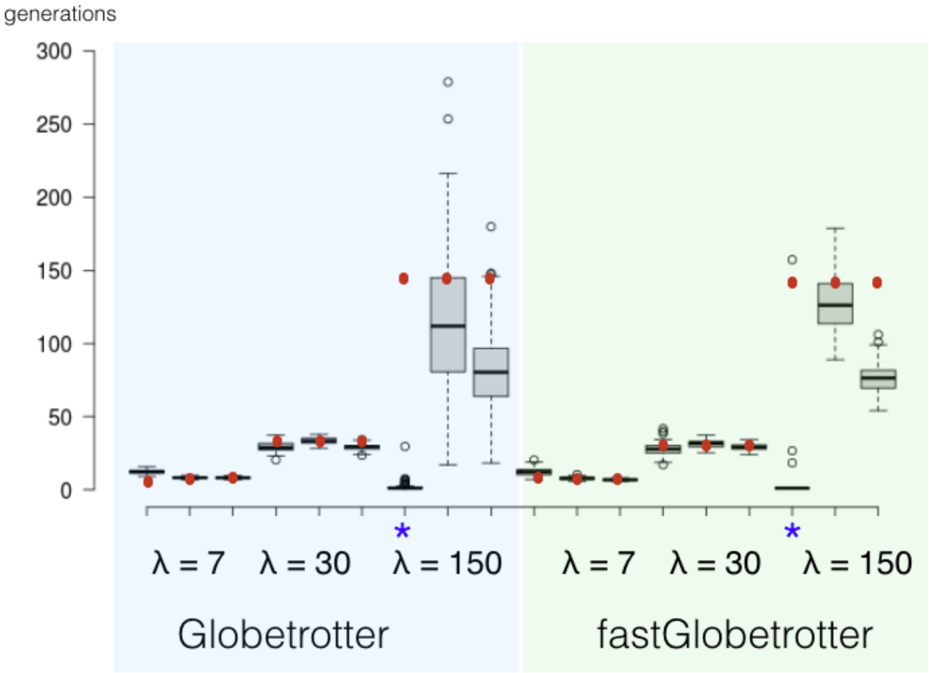

Figure 4.12: Box plot of date inference across 100 bootstrap re-samplings of French–Brahui simulations. The red dots represent the true dates. For each date (λ), results are given for three different admixture proportions (left-to-right): α = 0.05, 0.2, 0.5. Blue asterisks (*) represent no admixture inference.

TER ($N = 7$ in this case). Apart from this, there are 2 more inaccurate proportions from other 2 difficult cases i.e. French-Brahui, α=0.05, λ =150 and Brahui-Han, α=0.05, λ =150. In Section 4.4, we explore how inference can be improved by increasing number of targets.

The improvements of fastGLOBETROTTER over GLOBETROTTER in many mentioned cases is not surprising, as the sub-sampling algorithm down-weights chunks farther apart from each other that do not capture any information about the admixture event but can add random noise to the inference.

4.4 Increasing number of target samples to increase power of detection

To validate the performance of fastGLOBETROTTER when applied to a larger number of target individuals, I simulated additional individuals for the French–Brahui simulations with α = 0.5 and λ = 150. Previously we see that fastGLO-

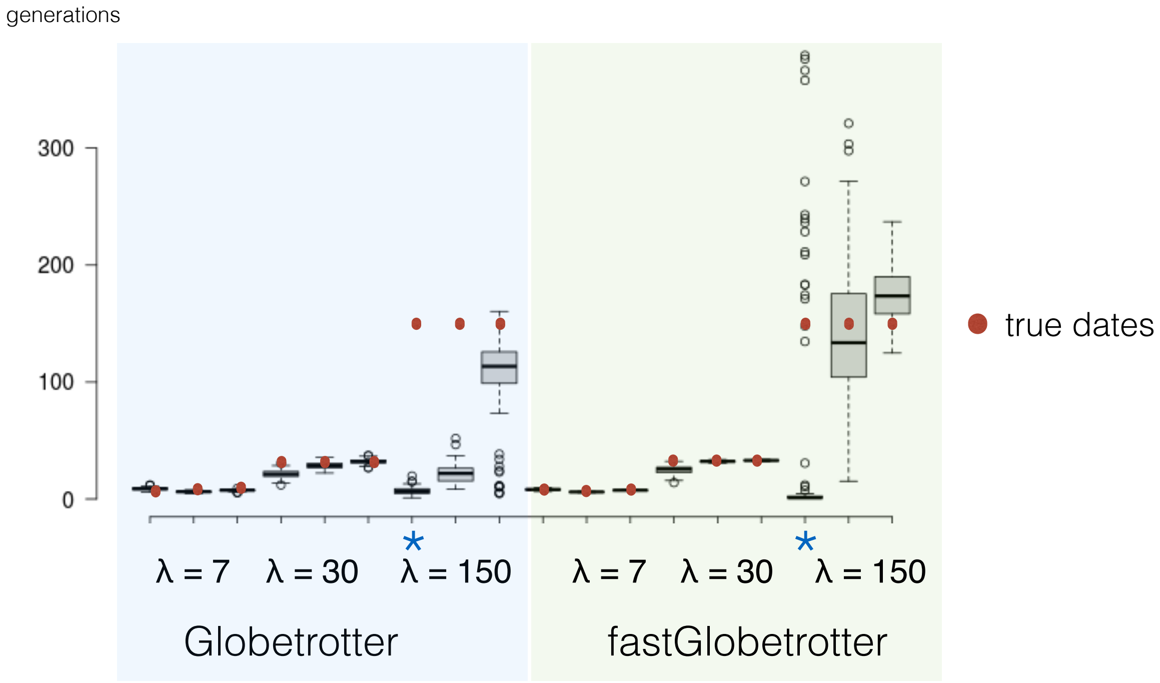

Figure 4.13: Box plot of date inference across 100 bootstrap re-samplings of Colombia–Han simulations. The red dots represent the true dates. For each date (λ), results are given for three different admixture proportions (left-to-right): α = 0.05, 0.2, 0.5. Blue asterisks (*) represent no admixture inference. (Note that GLOBETROTTER severely underestimates the date for $\lambda = 150/\alpha = 0.2$ in this example, though it performed better in previous applications (Hellenthal et al 2014). Similarly, fastGLOBETROTTER also underestimated the date in this simulation for other runs (results omitted) for unclear reasons. We note in general that coancestry curves are noisy for this simulation; future work will explore how this can be used to identify such problematic cases.)

BETROTTER and GLOBETROTTER infer admixture dates around 70 generations based on 20 simulated target samples in this scenario, hence both substantially underestimating the true dates. I simulated more target individuals, analysing 50 and 100 target individuals, to assess whether this improves the date inference. The box plot of date inference using fastGLOBETROTTER and GLOBETROTTER, again using 100 bootstrap resamples of individuals, for these simulations are shown in Figure 4.15.

The results suggest that with a higher number of target individuals, both fastGLOBETROTTER and GLOBETROTTER infer dates closer to the truth, even when fitting chunks separated by 1-30cM rather than 1-5cM. However, fastGLO-

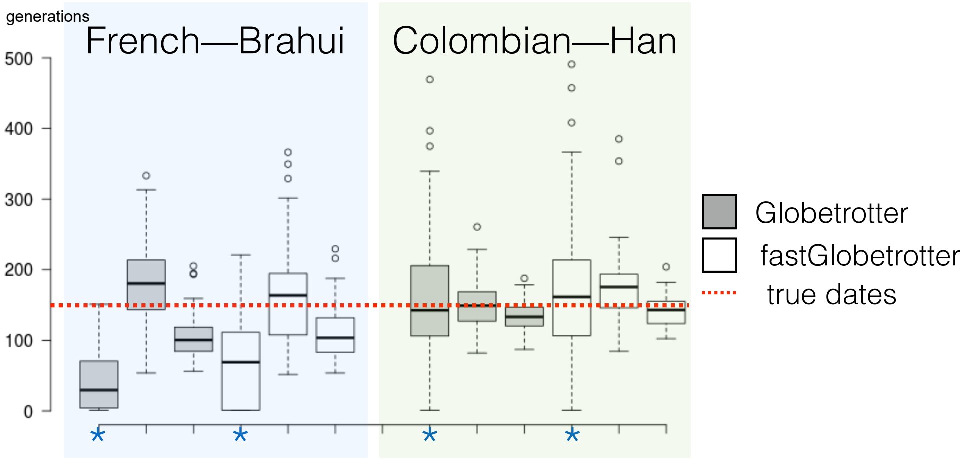

Figure 4.14: Box plot of date inference by fitting only chunks separated by 1-5cM across 100 bootstrap re-samplings of French–Brahui and Colombia–Han simulations. The red dashed line represent the true date. The results are given for three different admixture proportions (left-to-right): α = 0.05, 0.2, 0.5. Blue asterisks (*) represent no admixture inference.

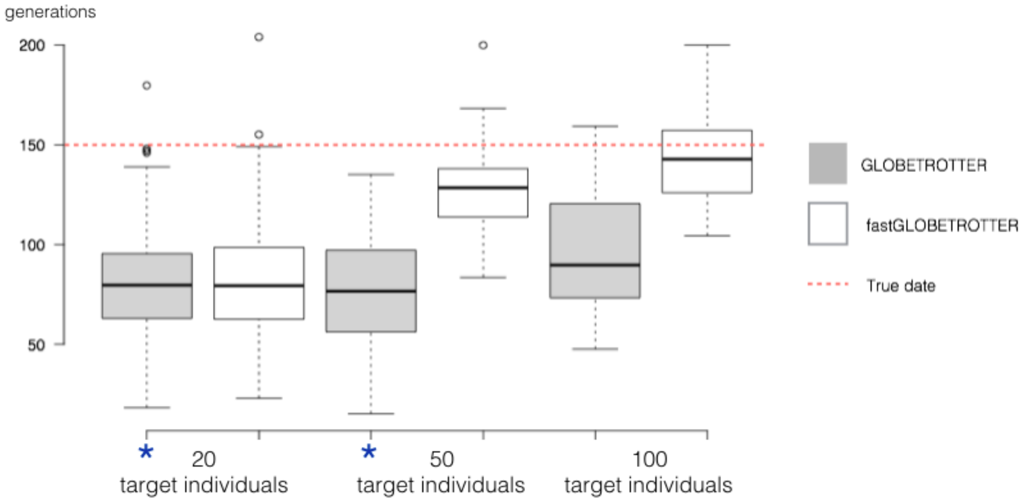

Figure 4.15: Box plot of date inference across 100 bootstrap re-samplings of French–Brahui simulations with 20, 50 and 100 target individuals, α=0.5, and λ=150. The x-axis gives the number of target individuals analysed. Blue asterisks (*) represent no admixture inference.

BETROTTER outperforms GLOBETROTTER in the cases of 50 and 100 target individuals, with inference substantially closer to the true dates (red dash line).

4.5 Improvement in computational time

I summarized the time used by fastGLOBETROTTER and GLOBETROTTER on each simulation using the same computing unit with CPU speed 3.5 GHz and 16 GB of RAM, as shown in Table 4.5.

Simulation	GLOBE-TROTTER(mins)	fastGLOBE-TROTTER(mins)	Time saved (folds)
POBI1	101.2	8.4	12.0
POBI2	91.9	8.1	11.4
POBI3	78.0	10.2	7.7
Brahui-Yoruba,α=0.05,λ=7	136.2	12.3	11.0
Brahui-Yoruba,α=0.20,λ=7	159.0	16.1	9.9
Brahui-Yoruba,α=0.50,λ=7	267.4	20.2	13.2
Brahui-Yoruba,α=0.05,λ=30	140.5	11.6	12.2
Brahui-Yoruba,α=0.20,λ=30	171.6	16.5	10.4
Brahui-Yoruba,α=0.50,λ=30	233.8	20.8	11.2
Brahui-Yoruba,α=0.05,λ=150	277.1	49.5	5.6
Brahui-Yoruba,α=0.20,λ=150	308.0	57.5	5.4
Brahui-Yoruba,α=0.50,λ=150	375.1	42.5	8.8
Yoruba-French,α=0.05,λ=7	338.0	34.6	9.8
Yoruba-French,α=0.20,λ=7	298.9	27.6	10.8
Yoruba-French,α=0.50,λ=7	255.4	43.7	5.8
Yoruba-French,α=0.05,λ=30	350.4	35.3	9.9
Yoruba-French,α=0.20,λ=30	310.5	31.4	9.9
Yoruba-French,α=0.50,λ=30	268.1	50.3	5.3
Yoruba-French,α=0.05,λ=150	419.1	35.4	11.8
Yoruba-French,α=0.20,λ=150	432.4	38.7	11.2
Yoruba-French,α=0.50,λ=150	424.6	73.2	5.8
Colombian-Han,α=0.05,λ=7	143.0	32.1	4.5
Colombian-Han,α=0.20,λ=7	155.4	26.8	5.8
Colombian-Han,α=0.50,λ=7	152.8	27.6	5.5
Colombian-Han,α=0.05,λ=30	131.8	30.9	4.3

Table 4.2 continued from previous page

Simulation	GLOBE-TROTTER(mins)	fastGLOBE-TROTTER(mins)	Time saved (folds)
Colombian-Han,α=0.20,λ=30	142.8	33.7	4.2
Colombian-Han,α=0.50,λ=30	146.1	28.0	5.2
Colombian-Han,α=0.05,λ=150	131.4	29.9	4.4
Colombian-Han,α=0.20,λ=150	147.4	32.5	4.5
Colombian-Han,α=0.50,λ=150	179.5	36.8	4.9
Brahui-Han,α=0.05,λ=7	563.2	77.0	7.3
Brahui-Han,α=0.20,λ=7	380.6	93.4	4.1
Brahui-Han,α=0.50,λ=7	554.5	89.7	6.2
Brahui-Han,α=0.05,λ=30	383.9	80.4	4.8
Brahui-Han,α=0.20,λ=30	354.0	85.2	4.2
Brahui-Han,α=0.50,λ=30	356.9	88.1	4.0
Brahui-Han,α=0.05,λ=150	421.5	100.3	4.2
Brahui-Han,α=0.20,λ=150	475.6	110.0	4.3
Brahui-Han,α=0.50,λ=150	464.5	103.0	4.5
French-Brahui,α=0.05,λ=7	151.4	30.5	5.0
French-Brahui,α=0.20,λ=7	143.0	28.6	5.0
French-Brahui,α=0.50,λ=7	131.3	16.8	7.8
French-Brahui,α=0.05,λ=30	147.3	29.3	5.0
French-Brahui,α=0.20,λ=30	148.9	25.3	5.9
French-Brahui,α=0.50,λ=30	144.2	24.4	5.9
French-Brahui,α=0.05,λ=150	257.2	54.7	4.7
French-Brahui,α=0.20,λ=150	253.3	48.1	5.3
French-Brahui,α=0.50,λ=150	258.3	58.2	4.4
French-Brahui,α=0.50,λ=150,50 targets	578.3	130.2	4.4
French-Brahui,α=0.50,λ=150,100 targets	1053.2	210.5	5.0

Table 4.2: Computational time used by GLOBETROTTER and fastGLOBETROTTER applied to simulations

According to Table 4.2, fastGLOBETROTTER is 4-12 times faster than GLO-

BETROTTER across all simulations. The fold change in speed depends on the number of donors *J* and number of surrogates *K* in Equation 3.1. While these analyses used only the sub-sampling approach of fastGLOBETROTTER described in Section 3.1 to speed up inference, we can further improve this by implementing the code optimization described in Section 3.3 and/or combining donors as described in Section 3.4. I have implemented three modes of fastGLOBETROTTER, which the user can toggle, as follows:

- Mode 1 includes the sub-sampling algorithm (Section 3.1) and the code optimization described in Section 3.3
- Mode 2 includes the sub-sampling algorithm (Section 3.1), code optimization (Section 3.3) and the protocol for combining donors described in Section 3.4
- Mode 3 includes only the sub-sampling algorithm (Section 3.1)

I applied these 3 modes on POBI data and the computation time results plotted in Table 4.3. The results suggest that all 3 modes of fastGLOBETROTTER infer dates accurately (close to the truth of 40 generations) and the time used by each mode is approximately 10-14 times faster than GLOBETROTTER. In this case there is little difference in time savings between the different modes, for this number of target individuals and donor set.

	inferred date (95% CI)				computational time (minutes)			
	GLOBE-TROTTER	mode1	mode2	mode3	GLOBE-TROTTER	mode1	mode2	mode3
POBI 1	40 (15-70)	39 (7-64)	31 (15-54)	39 (12-44)	195.7	16.7	14.9	18.6
POBI 2	45 (32-62)	43 (25-56)	48 (25-56)	40 (20-56)	184.0	15.4	14.8	17.4
POBI 3	53 (35-71)	45 (23-69)	46 (20-59)	44 (23-69)	164.8	11.8	10.8	14.2

Table 4.3: POBI date inference and computational time (minutes) used in GLOBETROTTER and 3 modes of fastGLOBETROTTER.

I also tested these 3 modes on the French-Brahui simulations with α=0.50, λ=150 and number of target samples = 20, 50, 100 individuals, with the computation time plotted in Figure 4.16 (The corresponding accuracy using mode 3 of fastGLOBETROTTER is in Figure 4.15.)

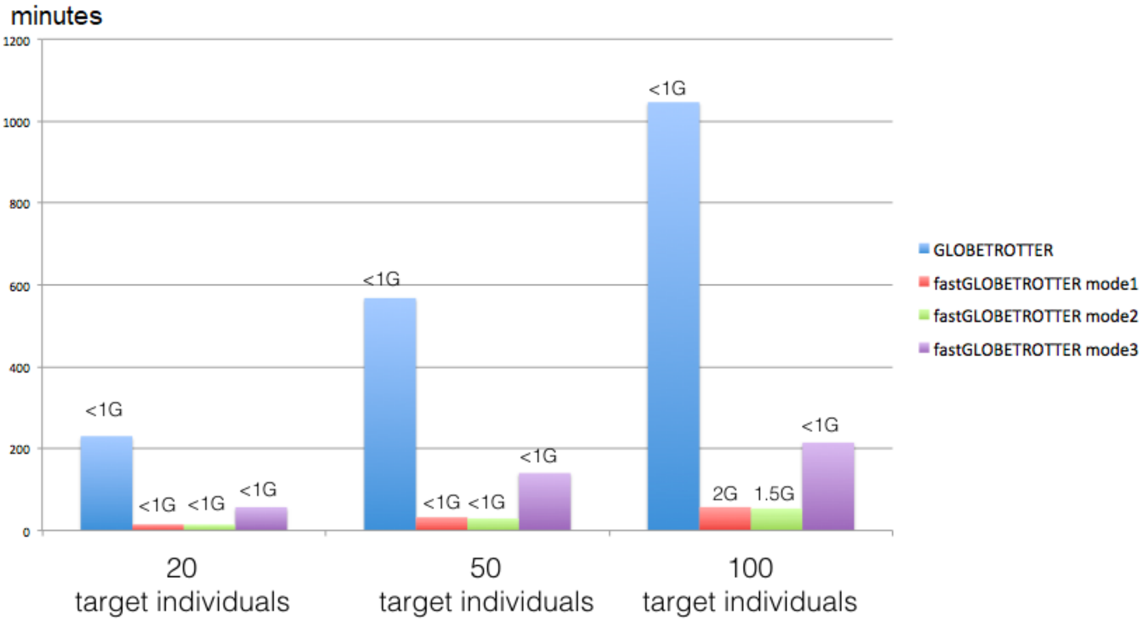

Figure 4.16: Bar plot summarizing computational time used by GLOBETROTTER and fastGLOBETROTTER mode 1,2,3 applied to the French-Brahui simulations with α=0.50, λ=150 and different number of target individuals. The number above each bar indicates the memory (RAM) used by each approach.

These results indicate that modes 1,2 are ~4-5 fold faster than mode 3, and ~20 times faster than GLOBETROTTER across different number of targets. However, as stated in Section 3.3, the code optimization requires more memory to store data. Here it requires 2G of memory in the case where the number of target individuals is 100. In contrast, mode 2, which additionally merges some donor groups to speed up the analysis and improve memory, requires 1.5G but with little improvement in computation time.

4.6 Testing for jack-knife technique for date inference CI

As described in Section 3.6, I performed jackknife re-sampling on the French–Brahui and Colombian–Han simulations by dropping one chromosome at a time, yielding 22 estimates across the 22 chromosomes. The results are plotted against the original bootstrap re-sampling of both GLOBETROTTER and fastGLOBETROT-

TER in Figure 4.17 and 4.18, respectively.

The date inferences from the jack-knife technique in both simulations aligned well with bootstrap re-sampling, especially in the case of recent admixture, i.e. $\lambda = 7$ and 30. However, we can see the differences in the case $\lambda = 150$ of French–Brahui, in that jack-knife technique gives the date around 10 generations more than that of bootstrap re-sampling of fastGLOBETROTTER, which is closer to the true date (150 generations). The jack-knife technique on Colombian–Han simulation gives the admixture date similar to bootstrap re-sampling of fastGLOBETROTTER. This suggests that the jack-knife technique performs similar to bootstrap re-sampling, providing a faster alternative to bootsrapping – e.g. a ≈5-fold increase in speed relative to doing 100 bootstrap re-samples.

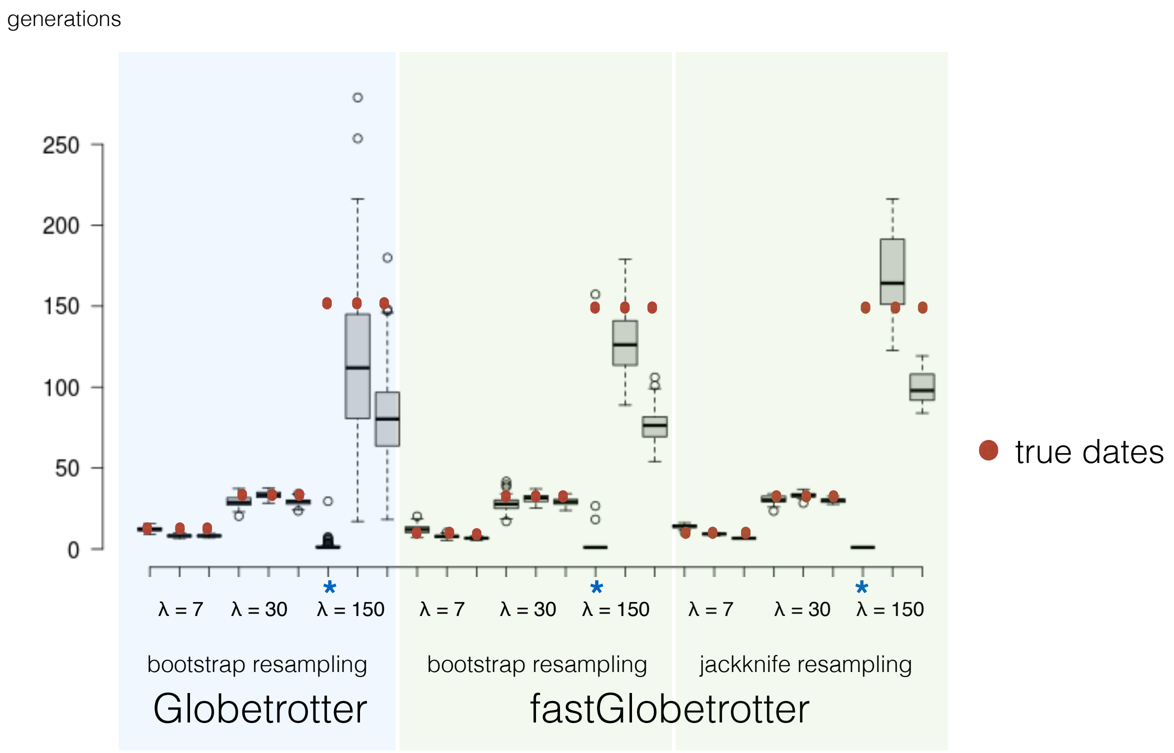

Figure 4.17: Box plot of date inference across 100 bootstrap re-samplings (the first two panels) and 22 jackknife re-samplings (the last panel) for the French–Brahui simulations. The red dots represent the true dates. For each date (λ), results are given for three different admixture proportions (left-to-right): α = 0.05, 0.2, 0.5. Blue asterisks (*) represent no admixture inference.

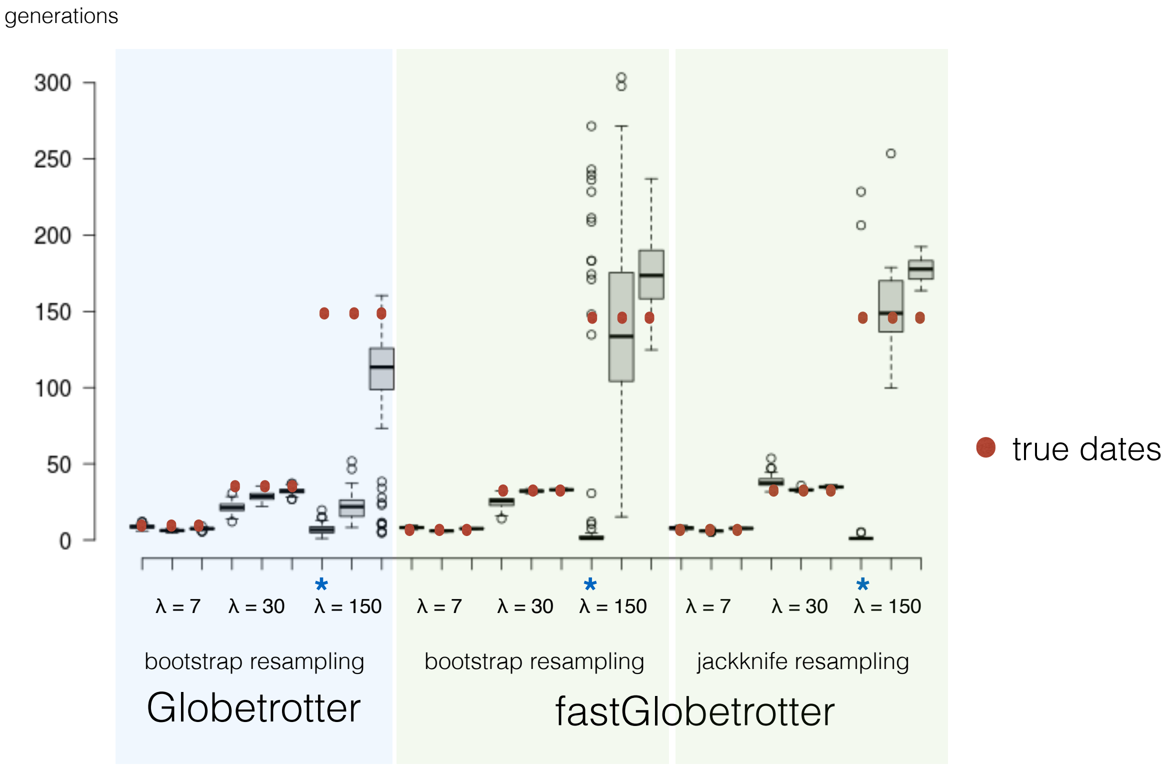

Figure 4.18: Box plot of date inference across 100 bootstrap re-samplings (the first two panel) and 22 jackknife re-samplings (the last panel) for the Colombian–Han simulations. The red dots represent the true dates. For each date (λ), results are given for three different admixture proportions (left-to-right): $\alpha = 0.05$, 0.2, 0.5. Blue asterisks (*) represent no admixture inference.

4.7 Simulation of bottlenecks following admixture

In this section, I explored the effect of bottlenecks on fastGLOBETROTTER inference as previously tested with GLOBETROTTER using simulations in Hellenthal et al., 2014. In particular, if the target population experiences a strong bottleneck following the admixture event, this can lead to relatively high LD between SNPs separately by large distances, which in turn mimics the signal expected under admixture. To test how fastGLOBETROTTER performs in this scenario, I used the coalescence-based simulated data from Hellenthal et al., 2014 described in Figure 4.19. Here they simulated Pop1-Pop4 in blue to represent the coalescence history of African groups, Pop5-Pop7 in orange to represent Western Eurasian groups and Pop8-Pop11 in green to represent East Asian groups. Specifically, the split at 2500 generations ago and subsequent bottleneck in Pop5-Pop11 mimics the“out-of-Africa" event, and the split between Pop4-Pop7 and Pop8-Pop11 at 1000 generations ago mimics the split between Western Eurasia populations and East Asian populations with a subsequent bottleneck in the East Asian groups.

They then simulated an admixed population comprised of 100 haploids sampled at present-day from Pop8 ("East Asia") and 150 haploids from Pop2 ("Africa"), which gives admixture proportions of 40% and 60%, respectively, and reflects a relatively small population size of only 250 haploids. Then for each haploid of the next generation, i.e. the first generation after admixture, they randomly sampled two pairs of (distinct) parent haploids from this pool of size 250 and composed the new haploid genomes as a mosaic of these two parent haploids, with switches in the mosaic based on the HapMap Phase 2 genetic map. For the simulations I test here, 100 haploids were generated in this manner, representing a reduction in population size from 250 to 100. For all subsequent $\lambda - 1$ generations, 100 new haploids were each composed of a mosaic of chunks from two distinct haploids randomly sampled with replacement from the previous generation. After λ generations, 50 haploids were randomly sampled to form 25 individuals for subsequent analysis. They simulated three different dates of admixture λ as noted below:

- Simulation PopG with 60% Pop2 + 40% Pop8 at $\lambda = 45$
- Simulation PopH with 60% Pop2 + 40% Pop8 at $\lambda = 20$
- Simulation PopI with 60% Pop2 + 40% Pop8 at $\lambda = 10$.

I applied these simulations to GLOBETROTTER and fastGLOBETROTTER using the "Null individual analysis" configuration, which is designed to eliminate signals in the coancestry curve that are not due to admixture (and hence be robust to bottleneck effects). The targets are the described admixed simulations PopG-PopI and donors are Pop2, Pop4, Pop9, Pop10, Pop11, plus six additional admixed populations (called PopA-PopF) described in Hellenthal et al 2014. The inference result is summarized in Table 4.4

			GLOBETROTTER inference				fastGLOBETROTTER inference			
	true date	true %Pop8	non-Null date (95% CI)	%Pop8	Null date (95% CI)	%Pop8	non-Null date (95% CI)	%Pop8	Null date (95% CI)	%Pop8
PopG	45	0.4	25(22-27)	0.44	46(40-50)	0.45	30(28-33)	0.45	46(41-50)	0.45
PopH	20	0.4	17(15-19)	0.48	19(17-22)	0.39	17(15-18)	0.39	19(18-20)	0.39
PopI	10	0.4	10(8-11)	0.41	10(9-12)	0.4	10(9-10)	0.40	10(9-11)	0.4

Table 4.4: Inferred dates and proportions of admixture on bottleneck simulations compared between GLOBETROTTER and fastGLOBETROTTER.

fastGLOBETROTTER's inference is similar to GLOBETROTTER in terms of date (95% CI) and proportion estimates in Null individual analysis, suggesting that it is equally robust to strong bottleneck effects.

4.8 Simulations to mimic the admixture in Europeans

In this section, I perform new simulations designed to understand admixture signals we detect in our real analysis of European data described in Chapter 5. In particular I considered four scenarios:

1. "One-date simulation mixing Denmark and Morocco:" $N = 50$ simulated individuals, $\lambda = 100$ and 200 generations ago, $\alpha = 0.8$ (from Denmark), generated by mixing genetic variation data from 162 individuals from Denmark cluster c49 (see Chapter 5) with that of 25 individuals from Morocco.
2. "One-date simulation mixing Denmark and Evenk:" $N = 50$, $\lambda = 100$ and 200, $\alpha = 0.8$ (from Denmark), derived by mixing 162 individuals from the Denmark Group c49 with 12 individuals from Evenk.
3. "Multiple-date simulation mixing Denmark and Evenk I:" First date, $N_1 = 50$, $\lambda_1 = 40$, $\alpha_1 = 0.8$ (from Denmark), derived by mixing 162 individuals from the Denmark Group c49 with 12 individuals from Evenk. Then the second date involves mixing the resulting admixed individuals with 134 individuals from Germany cluster c38 with $\lambda_2 = 10$, $\alpha_2 = 0.8$ (from resulting admixed in-

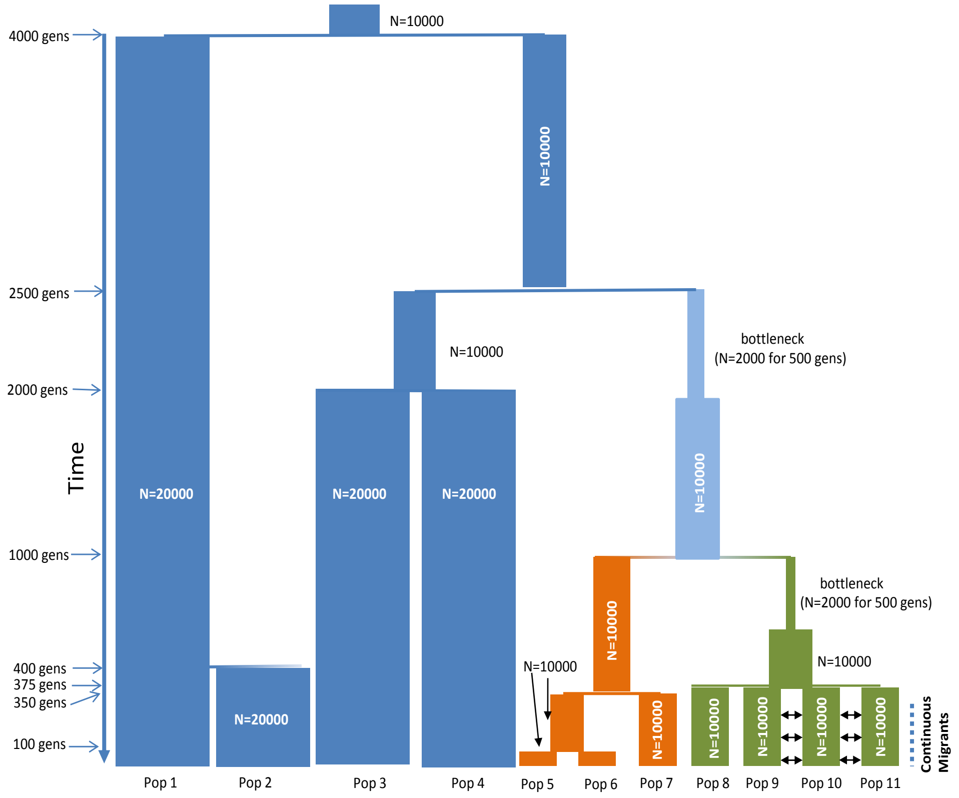

Figure 4.19: Simulated history for Pop1-Pop11 taken from Hellenthal et al., 2014 [21]. The simulations were generated using the coalescent-based software MaCS [59]. Pop1-Pop4 (blue) represent the coalescence history of African groups while Pop5-Pop7 (orange) and Pop8-Pop11 (green) represent Western Eurasian and East Asian groups, respectively. λ generations on the y-axis denotes the split time between groups. The bottleneck is shown in the figure when there is the decrease in population size N in a period of time.

dividuals) to create $N_2 = 50$ individuals as illustrated in Figure 4.20. Note the earlier admixture date should be $\lambda_1 + \lambda_2 = 50$ here, and the overall proportion of Evenk ancestry is $(1.0 - \alpha_1) * \alpha_2 = 0.16$.

4. "Multiple-date simulation mixing Denmark and Evenk II:" First date, $N_1 = 50$, $\lambda_1 = 10$, $\alpha_1 = 0.8$ (from Denmark), derived by mixing 162 individuals from the Denmark Group c49 with 12 individuals from Evenk. Then the second date involves mixing the resulting admixed individuals with 134 individuals from Germany cluster c38 with $\lambda_2 = 40$, $\alpha_2 = 0.8$ (from resulting admixed individuals) to create $N_2 = 50$ individuals as illustrated in Figure 4.21. Again, note the earlier admixture date should be $\lambda_1 + \lambda_2 = 50$ here, and the overall proportion of Evenk ancestry is $(1.0 - \alpha_1) * \alpha_2 = 0.16$.

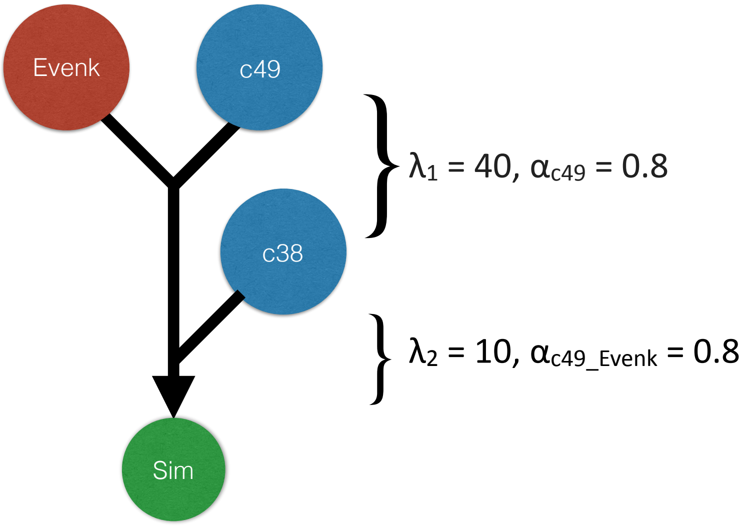

Figure 4.20: Illustration of simulation 3 with multiple-date admixture between European (c49) and Siberian sources (Evenk) at $\lambda_1 = 40$, and resulting Siberian–European admixed samples (c49_Evenk) with other European (c38) at $\lambda_2 = 10$.

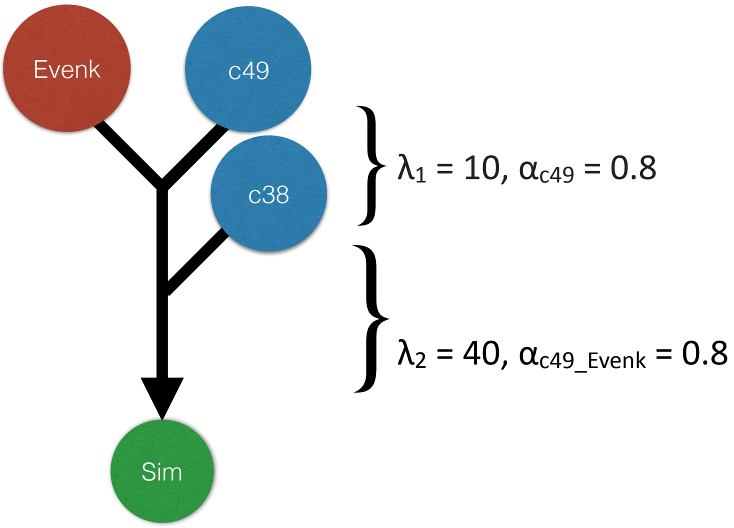

Figure 4.21: Illustration of simulation 4 with multiple-date admixture between European (c49) and Siberian sources (Evenk) with $\lambda_1 = 10$, and resulting Siberian–European admixed samples (c49_Evenk) with other European (c38) at $\lambda_2 = 40$.

The first simulation is designed to mimic the intermixing of groups related to North Africa (represented by Morocco) and Europe as we later report for many European populations in Spain, France, Belgium, and Germany. The second, third and fourth simulations further assess our model's ability to distinguish gene flow between Siberian-related groups and Scandinavian populations, which we later report for populations in Finland, Norway and Sweden. I considered two scenarios here, one with a single date of admixture (simulation 2) and one with multiple dates (simulation 3 and 4), to understand fastGLOBETROTTER's ability to detect multiple dates of admixture. The scenarios for Siberian gene flow considers cases where the fist admixture happens 50 generations ago, followed by a second admixture event from one of the original admixing sources at 10 generations ago (Simulation 3) versus 40 generations ago (simulation 4).

For each simulated dataset, I used CHROMOPAINTER to represent each of the simulated haplotypes as a mosaic of phased haploids taken from a reference of real individuals sampled from 160 worldwide populations (as described in Section 5.2) except individuals from the populations used in the simulations, i.e. HB:evenk and HB:moroccan. I used the estimated switch and emission rates from the main analysis, described in Section 5.2.1, and performed admixture inference in the same manner of Section 5.2.2. The date point estimates and 95% confidence intervals are provided in Table 4.5.

simulations	conclusion	inferred date (95% CI)	prop1	source1	source2
simulation 1					
λ=100	1-date	118(108-126)	0.08	HB:mozabite HB:mandenka	HB:german HB:norwegian
λ=200	1-date	197(178-220)	0.47	HB:norwegian HB:sannamibia	HB:tsi MS:NIreland
simulation 2					
λ=100	1-date	102(99-104)	0.19	HB:yakut HB:dolgan	HB:german HB:norwegian
λ=200	1-date	173(148-195)	0.35	HB:german HB:yakut	HB:german HB:norwegian
simulation 3					
λ_1=40, λ_2=10	1-date	45(45-46)	0.22	HB:yakut HB:dolgan	HB:german HB:norwegian
simulation 4					
λ_1=10, λ_2=40	1-date	51(49-52)	0.35	HB:yakut HB:belorussian	HB:german HB:english

Table 4.5: fastGLOBETROTTER inference applied to simulations mimicking admixture events in Europeans. "Conclusion" gives fastGLOBETROTTER's admixture conclusion, in the case a single date of admixture between two sources. "prop1" refers to the proportion of ancestry contributed by source1. Each source's genetic make-up is described by the proportions of recent ancestry they are inferred to share with each surrogate group; here I show the labels and proportions of the two surrogate groups with the highest such ancestry proportion contributions to each source.

Table 4.5 suggests that fastGLOBETROTTER can accurately detect the both European-Siberian and European-North Africa admixture with 50 individuals, accuratenly dating admixture in both cases of simulation 1 and when λ=100 in Simulation 2. There is a slight underestimation of the true date in the case λ=200 of simulation 2, though – as mentioned in Section 4.3 – this can likely be fixed by only considering chunk pairs separated by $<$5cM when building coancestry curves.

For both of the multiple-admixture simulations 3 and 4, fastGLOBETROTTER can only detect a single date of admixture. This suggests that fastGLOBETROTTER may fail to infer multiple pulses of admixture in cases where one of the original admixing sources (in this case European) intermixes again with the previously admixed group. This is not surprising, as multiple pulses of admixture that occur relatively close in time are difficult in theory to disentangle from a single admixture event with date between these two pulses. For simulation 3, the inferred date of 45 generations falls between the recent (10 generations) and older (50 generations) true admixture dates. In contrast, for simulation 4 the inferred date (51 generations) matches the initial admixture event, suggesting that the recent event (which is only 10 generations later) has little effect on admixture LD patterns. Meanwhile the simulation 3 results suggest that very recent admixture (in this case 10 generations ago) may lead to an excess of large (in this case European) segments that decreases fastGLOBETROTTER's inferred admixture date. I revisit these simulations when interpreting results of my application to European samples in Chapter 5.

4.9 Summary

In summary, I demonstrated the performance of the new algorithm, fastGLOBETROTTER, compared to GLOBETROTTER, using the simulated datasets described in this chapter. The performance of the methods were evaluated from 1) computational time 2) date inference and 3) proportion inference (the latter only in the POBI dataset, which is a very challenging scenario of admixture between two European groups). fastGLOBETROTTER is faster than GLOBETROTTER by a factor of 4-20 times depending on alternative modes used. While fastGLOBETROTTER preserves the accuracy of inference on most of the simulations tested, it also improves date inferences in some cases.

I also demonstrated that increasing the number of admixed target individuals can provide more precise estimates of admixture dates. In this scenario, fastGLOBETROTTER outperformed GLOBETROTTER by converging closer to the true dates faster as the number of targets increases, while using less computational time. This suggests that this technique is suitable for exploiting large scale cohorts to provide fast and accurate admixture inference.

Chapter 5

Genetic admixture in European populations

In this chapter, I describe the application of fastGLOBETROTTER to study European admixture using a large, previously published dataset that includes $\sim$6,000 individuals from Europe. This study shows the power of fastGLOBETROTTER to elucidate unearthed admixture events when the large number of individuals is considered.

5.1 Introduction

The initial migrations of the ancestors of modern Eurasians out of Africa began around 50,000-70,000 years ago [60]. Following this, there was a split to basal West Eurasian and East Asian groups and later basal West Eurasians were intermixed with Neanderthals [61]. After the Last Glacial Maximum that induced a major selection pressure in Western Eurasia [62], the population of West European Hunter-Gatherers (WHG) emerged around 20,000 years ago [63]. Later around 10,000 years ago in the Neolithic period, there was the arrival of agricultural-based people known as Early European Farmers (EEF) arrived from the Anatolian steppe, and spread across the continent. During the Bronze Age, the dispersal of a population related to the Ancient North Eurasians (ANE) who were related to Upper Palaeolithic Siberians from the Pontic-Caspian steppe further markedly affected the genetic make-up of Europeans [64]. The modern European populations are therefore believed to be made up of different levels of ancestry related to these WHG, EEF and ANE sources, which highlights the complex history of population interactions across the continent [61]. Moreover, with the European continent being bordered by Asia to the east and the Mediterranean Sea separating itself from North Africa to the south, European populations are associated with numerous admixture events induced by Roman, Norse, Arab and Turkish expansions to the continent.

To characterize the genetic demography of modern European populations, I tested several present-day European populations for admixture, attempting to identify and date any admixture events and quantify the populations' ancestry compositions using a genome-wide single nucleotide polymorphism (SNP) dataset composed of 6,209 Europeans and over 4,000 worldwide reference samples.

5.2 Admixture in 6,209 Europeans from 10 countries

The dataset used in this study includes 3 cohorts:

1. 1,471 individuals from 95 worldwide human populations genotyped using the Illumina 660W array from Hellenthal et al., 2014 [21] described in Figure 4.1. The individuals from these populations were treated as a reference to describe admixture in Europe.
2. 927 individuals from 65 worldwide human populations genotyped using the Illumina 550 described in Busby et al., 2015 [65]. The individuals from this dataset were also treated as reference samples.
3. 8,124 European samples from the International Multiple Sclerosis Genetics Consortium (IMSGC) [58] genotyped on the Illumina Human 660-Quad chip. Samples were collected from the United Kingdom (UK), North Ireland, Italy, Spain, France, Germany, Belgium, Poland, Denmark, Sweden, Finland and Norway. 1,915 samples from the UK and North Ireland were used as references, as the history of this region has already been characterized using GLOBETROTTER in Leslie et al., 2015 [57], while the remaining 6,209 Europeans were used as target individuals to detect admixture (Figure 5.1). These individuals were multiple sclerosis patients that were sampled from

various hospitals within these countries. Due to data access restrictions, hospital records of these people are not available; instead only the country of sampling is provided. Nonetheless we expect individuals to cluster based on geography within this dataset, as illustrated by previous analyses of these samples that were able to use hospital sampling information [57].

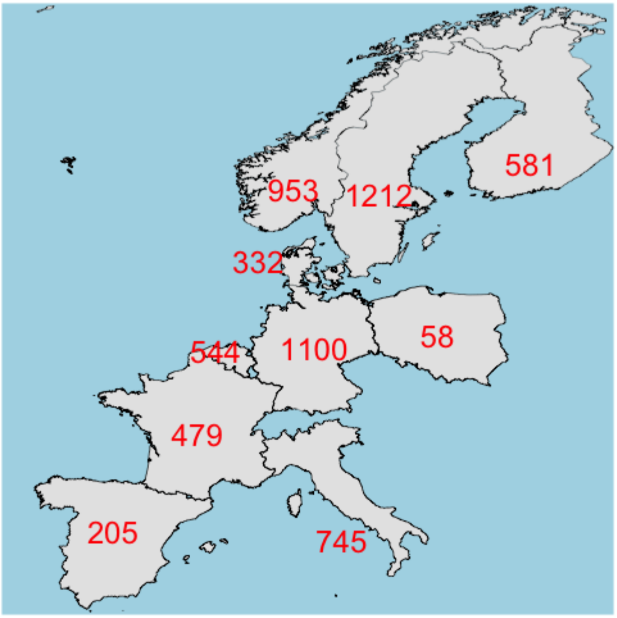

Figure 5.1: The 6,209 target European samples from Multiple Sclerosis data mapped based on their country of origin.

All datasets and genetic maps were based on build 36 of the human genome. The datasets above were merged using PLINK [66], which subsequently turned out a total of 10,522 individuals. Only SNPs with more than 1% of minimum allele frequency and less than 10% of missingness were retained, resulting in 477,417 autosomal SNPs. I split the dataset in binary PLINK format by chromosome and simultaneously phased all individuals using SHAPEIT [53], extracting the most likely pairs of haplotypes per individual (--output-max switch) and using a multi-core calculation of up to 8 CPUs for maximum speed (--thread 8 option).

5.2.1 Inference of population structure

I followed the framework for inferring population structure from haplotype data recommended in Leslie et al., 2015 [57]. This framework uses CHROMOPAINTER to paint the recipient haplotypes as a mosaic of pieces of the other donor haplotypes.

This painting process also summarizes the genome-wide relationship of haplotypes shared among individuals in the form of "copying vectors" that are comprised of either the total expected count of shared DNA segments among individuals, or the total expected length of shared DNA segments among individuals. Population structure can subsequently be inferred by clustering individuals based on the similarity of their copying vectors, which is done using the program fineSTRUCTURE [36].

Specifically, I ran CHROMOPAINTER to form the two phased haploids from each of the 10,522 individuals as a mosaic of those from all other 10,521 individuals. To do so, I ran CHROMOPAINTER initially to estimate the genome-wide average switch *Ne* (-n flag) and global emission rates θ (-M flag) in the Hidden Markov model of CHROMOPAINTER (Section 2.2.2) using a subset of individuals (1 individuals out of every 10) and chromosomes (4, 10, 15, 22) with 10 iterations of the expectation-maximization (EM) algorithm, i.e. using (excluding input/output file names for brevity):

```
    ChromoPainterv2 -g <infile> -r <recomrates> -t <individualfile>
-a 1 10 -s 0 -i 10 -in -iM -o <outfile>
```

I then averaged the estimated values of *Ne* and θ across chromosomes, and yielded 52.82727348 and 0.000134461, respectively. Next, I ran CHROMOPAINTER on each chromosome separately using these fixed values, and painted each individual against all the others, with the following command:

```
    ChromoPainterv2 -g <infile> -r <recomrates> -t <individualfile>
-a 0 0 -s 0 -i 10 -n 52.82727348 -M 0.000134461 -o <outfile>
```

This produced a matrix of copying vectors that provides the total expected count of DNA segments for which each individual shares a most recent ancestor with each other individual. I input this matrix into the program fineSTRUCTURE to cluster the European individuals into genetically homogeneous clusters. I performed 5 million iterations of Markov Chain Monte Carlo (MCMC) to infer the number of clusters and cluster assignments of all individuals, and sampled these inferred values at every 10,000 iterations using the following command:

```
    finestructure -X -Y -x 0 -y 5000000 -z 10000 <chunkcounts>
```

```
<mcmcfile>
```

This gave 500 MCMC samples, with each one having an inferred number of clusters and cluster assignment. I then used fineSTRUCTURE to single out the MCMC sample among these 500 with the highest posterior probability overall, after which a further 10,000 steps are taken to find a solution that improves the posterior under a greedy approach. This generated a final number of 91 inferred clusters. Next a tree is constructed that hierarchically merges these clusters, one-at-a-time, under a greedy approach until only two clusters remain. These steps were accomplished using the following command:

```
    finestructure -X -Y -x 10000 -m T -t 100000 <chunkcounts>
<mcmcfile> <treefile>
```

The flag -t indicates the number of pairwise comparisons of clusters to consider when merging at each level of the tree; the value I input was large enough that all such possibilities ($91\mathit{choose}2$) were considered. While fineSTRUCTURE inferred 91 clusters in total, in practice clusters that merge early in the tree, which typically have subtle genetic differences, are often combined in order to increase the sample size of the final classifications (Leslie et al., 2015 [57]). Based on a visual inspection of the tree, I classified individuals into 86 major groups (c1-c86) used in subsequent analyses. I performed principal components on the matrix of copying vectors excluding reference individual using the `princomp()` function in R to visualize the categorized both by country (Figure 5.2) and by the 86 clusters (Figure 5.3). The cluster name, total samples, region and number by population are given in Table A.1.

5.2.2 Dating of admixture events in the 86 clusters

I detected admixture across Europe used these 86 clusters to detect admixture across Europe using these 86 clusters on fastGLOBETROTTER, and to inferred the proportion and dates of any such admixture in these groups us-ing 162 reference populations as donors under the fastGLOBETROTTER analysis. To do so, when testing each cluster as a target, I performed an additional CHROMOPAINTER analysis that painted the cluster individuals in each cluster against all individuals from 162

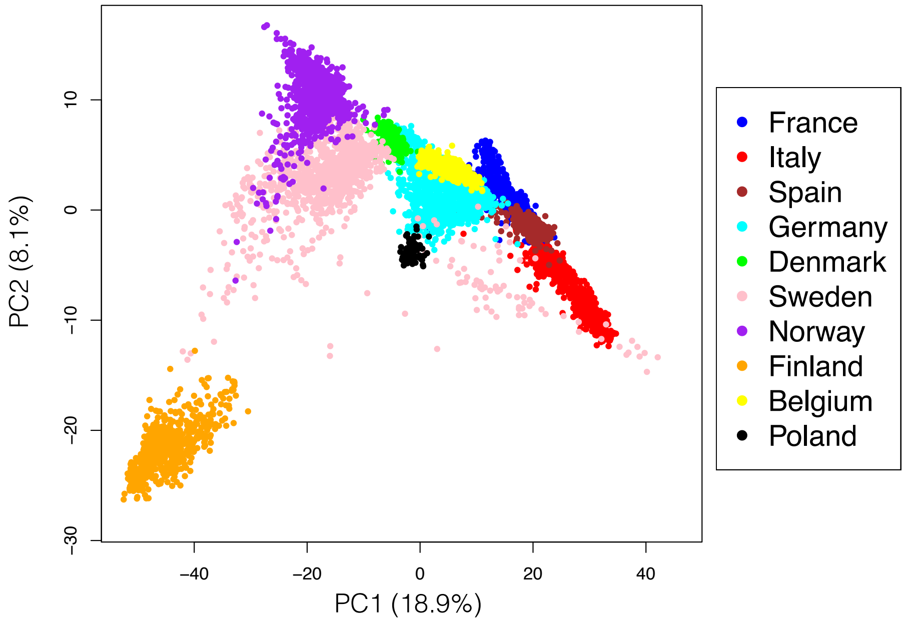

Figure 5.2: PCA of European individuals colored by country of origin. The percentage on each axis is the variance explained by that PC.

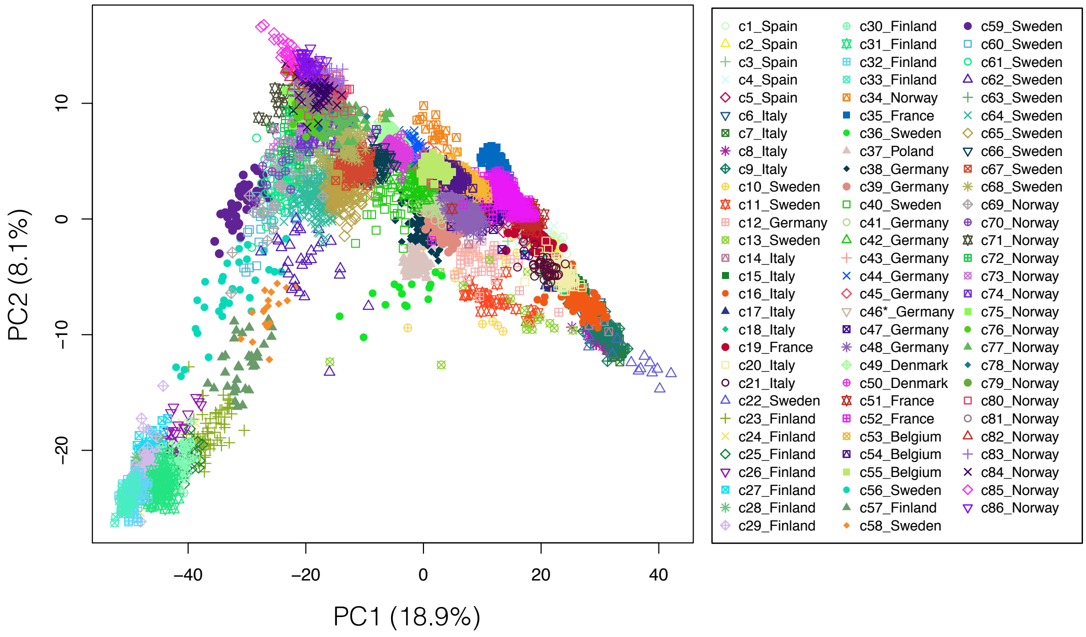

Figure 5.3: PCA of European individuals colored by groups derived from fineSTRUCTURE clustering. The legend shows the 86 groups (c1-c86) and their most common country of origin among samples within each group. The percentage on each axis is the variance explained by that PC.

world populations. Under this painting, I applied the switch "-s 10" to generate 10 painting samples per each European target haploid:

```
ChromoPainterv2 -g <infile> -r <recomrates> -f <donorfile>
-a 0 0 -s 10 -n 52.82727348 -M 0.000134461 -o <outfile>
```

The donorfile (-f flag) defines which sampled groups are to be painted (i.e. each of 86 clusters as a recipient) and which are to be painted against (i.e. 162 world populations as donors). Analogous to the analyses of Hellenthal et al., 2014 [21] and Leslie et al., 2015 [57], I used these painting samples and the paintings generated from CHROMOPAINTER that were used in the fineSTRUCTURE analysis to identify, date and describe the admixture with the following command:

```
R < fastGLOBETROTTER <paramfile> <paintingsamplefile>
<recomrates> <mode> --no-save > -o <outfile>.
```

In the paramfile, I designated the donors and surrogate groups as the 162 reference populations defined by population label, and the target group as each European cluster. For each European cluster c1-c86, I ran fastGLOBETROTTER to estimate admixture dates and proportions, using 100 bootstrap re-samples of individuals' chromosomes to infer the 95% CI for the actual admixture date(s). I summarize the results in Figure 5.4 for the 83 clusters for which fastGLOBETROTTER inferred a single of admixture between two or more sources (i.e. the "one-date" or "one-date, multiway" conclusion described in Chapter 6); the remaining three clusters that inferred multiple dates of admixture (i.e. "multiple-date") are not shown.

5.2.3 Admixture events in Europeans

As seen in Figure 5.4, fastGLOBETROTTER detected admixture in all clusters from multiple disparate source groups. Sources related to present-day individuals sampled from Africa, Asia, Siberia, and West Asia contributed in different proportions to shape the modern day European genetic make-up. There is also evidence of potentially "local" admixture events, where the intermixing sources were genetically similar to one another and to sampled individuals from the region where the cluster individuals now reside.

Finland

Finnish individuals appear the most genetically distinct from other Europeans in the PCA plot of Figure 5.2, thereby suggesting unique ancestry relative to other

Europeans and/or relatively strong isolation effects from other Europeans. Across all 11 clusters, inference from fastGLOBETROTTER indicates admixture between a source with Siberian/East Asian-related ancestry (Russian-like) and a source with Northwest European (UK-like) ancestry dated to a range of times spanning 180-550 CE (Figure 5.5 , 5.4). This is consistent with the geography of Finland, which is the easternmost of our European samples and nearest to Russia and Siberia [67, 68].

Sweden

The admixture events in Sweden can be divided into two main categories. Some clusters (c62-68) show admixture between Northwestern European (UK/German-like) and Northeastern European (Polish-like) sources dated 800-1100 CE. In contrast, other clusters show admixture between Northwestern European (UK/German-like) and Northeastern European with Siberian/East Asian ancestry (Russian-like). Specifically, the PCA plot of the latter clusters (c56, c59, c60, c61) shows relatedness to Finnish samples suggesting that these Sweden clusters might derive the Siberian/East Asian signal from migration of Finns to Sweden. Interestingly, the ancestry inferred in c22 shows a distinctive admixture event; one ancestry source from Armenians and another source from Northwestern Europeans (UK/German-like), the admixture date is as recent as 1300 CE. This is suggesting that c22 might be descendants of the Europeans in Sweden and the migrating Armenian-like group, this is also supported by the PCA in Figure 5.6 in that c22 is the most distant group among all Swedish groups.

Norway

Like Sweden, we can see the same trend in Norway clusters, with inference suggesting a source related to Northwestern Europeans carrying South Central European ancestry (UK/German-like) intermixed with a source related to Northwestern Europeans carrying additional Siberian/East Asian ancestry (Norwegian-like). The admixture dates among these clusters vary from 400-1400 CE.

Denmark

Denmark is the only country in Scandinavia with no ancestry related to present-day Siberians. Instead, the two Danish clusters show admixture between Northwestern European (UK/German-like) and Northeastern European (Polish-like) sources dated to around 830-900 CE. It is observed that there is a small amount of Armanian/Iran ancestry in the c49 cluster. My results coincide with a previous genetic study of a large Danish cohort in terms of date of admixture and inferred ancestry [69].

Poland

The detected admixture in Poland occurred around 830 CE and indicates that a source related to local Northeast Europe (Croatian-like) intermixed with a source related to Northwest/South-Central Europe (UK/Toscani-like).

Germany

The history of admixture in German groups shows two disparate types of admixture event in the region. The first, shown in clusters c38, c39, c41, c41, c42 suggests intermixing between sources related to Northwest Europe (UK-like) and Northeast Europe (Polish-like) dated around 1250 CE.

The second, shown in clusters c43-c48, suggests intermixing between a source related to Northwest Europe (UK-like) and South-Central Europe (Westsicilian-like) and a source related to North African/Armenia/Iran dated 500-700 CE. Significantly, the PCA plot aligns clusters c38, c39, c41, c41, c42 near Poland while clusters c43-c48 are aligned more with Belgium and France suggesting different admixture events, whose signals are discussed below.

Belgium

The Belgium samples presented in this study seem genetically homogenous. Even though they are divided into three clusters, fastGLOBETROTTER infers very similar admixture events in each, with a single event occurring around 600 CE between sources related to Northwest Europe (UK-like) and South-Central Europe (South-Italian-like). The latter group, which reflects the minority contributing source, also interestingly shows some shared ancestry with present-day peoples from North African/Armenian and Iranian along with South-Italian-like ancestry. Our results add further details to the previous genetic study of 189 Belgians by Van den Eynden et al., 2018 [70]], which reported the evidence of “recent” migration from Southern Europe and Northern Africa in Belgian but did not provide an estimated admixture date.

France

In France, fastGLOBETROTTER suggests that clusters c19 and c52 experienced admixture events between a source related to populations from Northwest Europe (UK-like) and a source related to populations from North Africa (Moroccan-like). In contrast, clusters c35 and c51 experienced admixture from sources related to North West Europe (UK-like) and South Central-Europe (South-Italian-like). The admixture dates are around 400-550 CE across all clusters.

Italy

The Italian clusters show the most diverse admixture signals among our European groups, suggesting multiple ancestral sources from within and outside Europe contributing to Italian genetics. Admixture related to European ancestry suggest contributions primarily from sources related to South Central Europe (Greek-like) and Northwest Europe (French-like). The most abundant source of non-European ancestry relates to West Asia/Armenia/Iran, with additional contributions related to North and Sub-Saharan Africa. The timing of the admixture events and the sources involved differ between clusters (c14-18, c20-21) that fall closer to Central Europe on the PCA and clusters (c6-9) that fall on the edge of the PCA, perhaps (consistent with the PCA) reflecting different admixture histories between geographically Northern and Southern Italy, respectively. Admixture dates in “Northern Italy” clusters (c14-18, c20-21) are much older at around 400-500 CE, relative to admixture dates of 700-1000CE in “Southern Italy” (c6-9). Similar findings were reported in a recent paper by Raveane et al., 2019 [71] using different Italian samples.

5.2.4 Summary of admixture inference in Europe

Consolidating the signals across all clusters divides them broadly into two categories. The first category includes clusters predominantly consisting of individuals sampled from countries north of the Baltic Sea (Finland, Norway, Sweden), which infer some proportion of their ancestry is related to present-day groups from East Asia and Siberia. Meanwhile clusters predominantly consisting of individuals sampled from countries south of the Baltic Sea (Belgium, France, Germany, Spain, Italy) have little trace of E.Asian/Siberian-like ancestry, and instead have some ancestral components matching those of individuals from present-day West Asia, North Africa and/or Sub-Saharan Africa.

Admixture signal from East Asian/Siberia ancestry in Scandinavians

The orange part of Figure 5.14 shows the signal of East Asian/Siberia ancestry in some Scandinavian clusters. The oldest signal is present in Finnish cluster c27 and dated to around 64 generations ago or $\approx$180 CE. In general, the proportions of ancestry matching to these E.Asian/Siberian sources are highest in Finland (e.g. clusters c32, c25), the easternmost Scandinavian country closest geographically to Siberia, and then steadily decrease in Finnish, Swedish and Norwegian clusters with more recent admixture dates (e.g. clusters c59, c60, c61, c85, c81, c77). This is consistent with the Finnish individuals descended from early intermixing between European and E.Asian/Siberian-like sources. This is consistent with results reported by Saag et al., 2019 [67], who found Siberian admixture in ancient DNA samples from Estonian in Late Bronze Age graves dating $\approx$2,500 years ago. While their reported date is more ancient than that inferred by fastGLOBETROTTER, this does not preclude multiple episodes of intermixing involving an East Asian-like source in the region by the Iron Age.

Our simulations 3 and 4 in Section 4.8 illustrate how multiple dates of admixture between two sources, where one of the original admixing sources subsequently intermixes with the previously admixed group, may be inaccurately described by fastGLOBETROTTER as a single admixture event, sometimes with an inferred date somewhere between the dates of the two admixture events. In our results here, when moving geographically from east to west (i.e. Finland to Norway), clusters show decreased dates and decreasing proportions of ancestry related to E.Asia/Siberia. These two observations are consistent with a scenario where a source related to northwest Europeans initially intermixed with a source related to East Asians around (or perhaps older than) 180 CE, with this intermixing perhaps occurring geographically nearer to Finland than to Norway or Sweden. Subsequently, this admixed group could have migrated west, intermixing with other unadmixed Europeans, which could lead to both decreased date estimates and decreased E.Asian/Siberian proportions of ancestry as we observe (Fig 5.14). Such a pattern was observed in simulation 3 where two episodes of admixture were simulated at 50 and 10 generations ago, for which fastGLOBETROTTER concluded a single admixture event with an inferred date at $\approx$45 generations. In contrast, for simulation 4 where the simulated dates were 50 and 40 generations, i.e. the second admixture event was not much more recent than the first admixture event, fastGLOBETROTTER inferred a single admixture event at $\approx$50 generations, i.e. matching the older admixture event. This suggests that recent admixture within Europe may be decreasing fastGLOBETROTTER's inferred date for an older event involving an East Asian-like source. Increased sample sizes from these areas, which may allow fastGLOBETROTTER to correctly identify and date multiple pulses of admixture, and/or additional data from ancient human remains may shed light on whether this is indeed the case.

In a simple analysis to mitigate effects of any such recent intermixing, a Masters student (Matthew Greenfield) that I co-supervised re-analyzed these same 86 clusters while removing all surrogate groups except the UK. Excluding European reference populations as surrogates may act to mask more recent admixture if indeed the source of this admixture was a more local European-like population. We retained the UK in this analysis to assist in detecting the admixture we initially inferred, i.e. admixture between sources related to present-day European and E.Asia/Siberia populations. However, we hope that the present-day UK is sufficiently diverged from any European-like group that subsequently intermixed with this European-E.Asian/Siberian admixed population. Results of this analysis are shown in grey dots in Figure 5.14. As expected under a scenario where indeed recent admixture was affecting our initial inference, inferred dates under this new analysis are consistently older with the shift in the range of admixture date from 180-1150 CE to 20-1000 CE. Furthermore, we can see that there are increasing proportions from E.Asian/Siberian in the new analysis, likely because some of the European groups included as surrogates in my original analysis carried some East Asian ancestry.

Admixture signal from West Asian/North African ancestry in Europeans

The green part of Figure 5.14 shows a signal of West Asian/African ancestry in some European clusters. In Italy and Spain, the proportions of ancestry related to these groups vary across clusters with no notable trend in inferred dates, perhaps reflecting multiple waves of migration from these geographically nearby regions into Italy/Spain over several generations. Notably some French clusters (c19, c52), all three Belgian clusters c53-c55 and some clusters containing Germans (c43, c45, c46, c47, c48) infer the same admixture date of 550-700 CE, showing various degrees of ancestry sharing with groups related to modern-day Greece, Cyprus, Morocco, Turkey and Armenia. While the historical event driving this signal is unclear, an intriguing possibility is that it relates to the Roman Empire, which covered all of present-day Belgium, Germany, France, Turkey, North Africa and elsewhere prior to its decline and eventual fall in 530CE [72]. In particular individuals with ancestry related to the ancestry of people found in North Africa, West Asia and South Europe today could have moved across the empire during this time, and intermixed with people living in or around present-day Belgium, France and Germany, perhaps with this intermixing occurring in part after the fall of the Roman Empire as our inferred date suggests. Other German clusters (c38, c41, c42) do not appear to have this W.Asian/African signal, instead exhibiting ancestral signals related to Eastern European groups (Figure 5.10), perhaps reflecting geography and the political history of a divided Germany.

5.3 Admixture in a Greek cohort of 631 individuals

In this section, I study genetic admixture in a Greek population. The data primarily comprises 748 individuals from the “TEENAGE (TEENs of Attica: Genes and Environment)” cohort [73,74], which consists of randomly sampled Greek students aged 13–15 attending public secondary schools located in Athens. DNA samples from TEENAGE were genotyped using Illumina HumanOmniExpress BeadChips, yielding 831,665 SNPs. The data was merged with the set of world-wide reference populations described in Section 5.2 including 1) 94 populations (excluding Greece) from Hellenthal et al., 2014 [21], 2) 65 populations from Busby et al., 2015 [65], and 3) 12 European populations from the Multiple Sclerosis cohort. To reduce complexity, a maximum of 100 individuals per population were included [58], which turned out 4,224 individuals in total. Quality control was performed using PLINK [66], taking only SNPs with allele frequency $>$1% and missingness $<$10%, which gave 281,079 SNPs across 22 chromosomes. I phased all individuals in the merged data using SHAPEIT [53].

5.3.1 Subpopulations in Greek population

In this analysis, I applied a similar framework to that described in my analysis of Section 5.2.1. Initially, I performed 10 iterations of expectation-maximization (EM) algorithm to infer the genome-wide average switch N_e and global emission rates θ in CHROMOPAINTER's Hidden Markov model ($-n$ flag and $-M$ flag, respectively), starting with default values. For computational simplicity, for this E-M step I painted only one out of every ten individuals, and only chromosomes 1, 8, 15 and 20. Each individual was painted using all other 4,223 individuals as donors. I averaged inferred values for each parameter, weighting the averages by chromosome length, and then averaged across all painted individuals. This produced values of 152.4843 and 0.0005527 for N_e and θ, respectively.

Next I ran CHROMOPAINTER to paint each of the 4,224 recipient individuals using all other 4,223 individuals while fixing the estimated values of N_e and θ. The matrix of inferred number of haplotype segments that each individual matches to every other (i.e. the "coancestry matrix") was input into fineSTRUCTURE to cluster individuals based on haplotype sharing patterns. I ran fineSTRUCTURE for 5 million iterations of MCMC, while sampling every 10,000 iterations and consequently obtained 500 MCMC samples, each with an inferred number of clusters and cluster assignment for each individual. Taking the MCMC sample with the highest posterior probability, fineSTRUCTURE then used 10,000 hill-climbing steps to find a solution with a higher posterior. After this hill-climbing step, the 631 Greek individuals were assigned to 45 clusters. Starting from the final cluster assignments, fineSTRUCTURE merged clusters one by one under its greedy approach until only two clusters remained. Based on visual inspection of the resulting tree, I classified the Greek individuals into ten main clusters (c1-c10). I excluded 117 Greeks that were not assigned to any of these ten clusters (Figure 5.15). In justification of the fineSTRUCTURE clusters, I show a visualization of principal components of the CHROMOPAINTER's copy vectors using *princomp*() in R in Figure 5.16. The Greek individuals are widely spread across this PCA, with some Greek clusters appearing close to Middle East populations (e.g. Jordanian) and others to Europeans (e.g. Poland).

5.3.2 Admixture inference in Greek clusters

I applied fastGLOBETROTTER to study admixture in the ten Greek clusters by treating each cluster as a target population and the other 171 (non-Greek) reference populations as surrogates for the admixing sources that these Greek populations are descended from. To do so, I performed additional CHROMOPAINTER analyses to obtain 10 painting samples per each Greek target haploid ("-s 10" similar to Section 5.2.2) where each Greek haploid is painted using only the 171 reference populations (and non-Greek populations) as donors. Finally, I used these painting samples, along with the coancestry matrix generated by CHROMOPAINTER for the fineSTRUCTURE analysis. In my fastGLOBETROTTER analysis, I entered default values and performed 100 bootstrap re-samples to generate 95% confidence intervals for inferred dates. Results for each Greek cluster are summarized in Figure 5.17 and Table A.3.

The admixture events detected by fastGLOBETROTTER in the Greek clusters c2-6, c9-c10 show evidence of contributions from sources related to European (Hungarian-like) populations versus those related to Near East (Lebanese-like) populations, with a clear split between them on each side of ancestral sources (Figure 5.17). The European ancestry in these Greek clusters typically matches that of

Polish, North Italian and Romanian individuals, while the Near East ancestry corresponds to that of Lebanese, South Italian, African, and Armenian individuals. This result is in line with many autosomal DNA studies [75, 76] showing that Greeks are genetically closest to Italians and Romanians. Also, recent studies by Kovacevic et al., 2014 [77] and Lazaridis et al., 2017 [78] show that modern Greeks have ancestry related to Near Eastern and East European populations. However, none of these studies indicated the trace of African ancestral sources that fastGLOBETROTTER detects in Greek clusters c1-c4 (Figure 5.17). There is particularly strong evidence of admixture between African and Near East sources in c8, which appears close to African populations in the PCA (Figure 5.16). Cluster c7 is the only group that shows no evidence of intermixing with Near East groups, while inferring only a 1% contribution from North Africa.

The inferred admixture dates in these Greek clusters range from 700-1300 CE. My analyses here broadly support the findings from the study by Hellenthal et al., 2014 [21] that inferred an admixture date in Greeks at around 718-1138 CE, with 37% of ancestry inherited from a Polish-like source and the remaining 63% from a Cypriot-like source. This indicates that the intermixing happened later than the fall of the Roman Empire (and later than detected admixture described in Section 5.2) that had ruled Greece until ∼300 CE. This timing is compatible with the settlement of either the Slavs or Byzantines in Greece around 700-1000 CE. However, clusters c1, c7, c8 have little evidence of northeast Europe-like ancestry and exhibit more recent inferred dates around 700-800 years ago. These clusters infer admixture involving different sources that suggest quite disparate histories, i.e. related to Southern Europe versus Armenia/Iran/Near East (c1), Southern versus Central Europe (c7) and North Africa versus South Asia/Armenia/Iran/Europe (c8). These analyses are complicated by the use of modern groups themselves who exhibit varying levels of recent admixture from different sources. Further studies using ancient DNA samples that are unaffected by recent admixture events may provide more insight into how and when the present-day Greek genetic pool was formed.

5.4 Summary

In summary, fastGLOBETROTTER analyses of European data in this chapter suggest that there have been admixture events among different European populations at different points in time, which have shaped the genetic make-up in current European populations in addition to much older events of intermixing between early farmers, hunter-gatherers and steppe peoples. These admixture events form one piece of the jigsaw reflecting the whole complex genetic history of European populations. My findings show high levels of admixture between different European populations and also from non-European ancestral sources, such as admixture from an East Asia/Siberian-like source into Finland groups in 180 CE and admixture from a West Asia/North African-like source dating back to the fall of the Roman empire in 550 CE.

This also showcases how efficiently fastGLOBETROTTER can handle large-scale data. For example, mode 1 of fastGLOBETROTTER managed to infer admixture dates on the largest cluster containing 212 target individuals, 162 donor populations and 470K SNPs in 73 minutes with 13G of RAM.

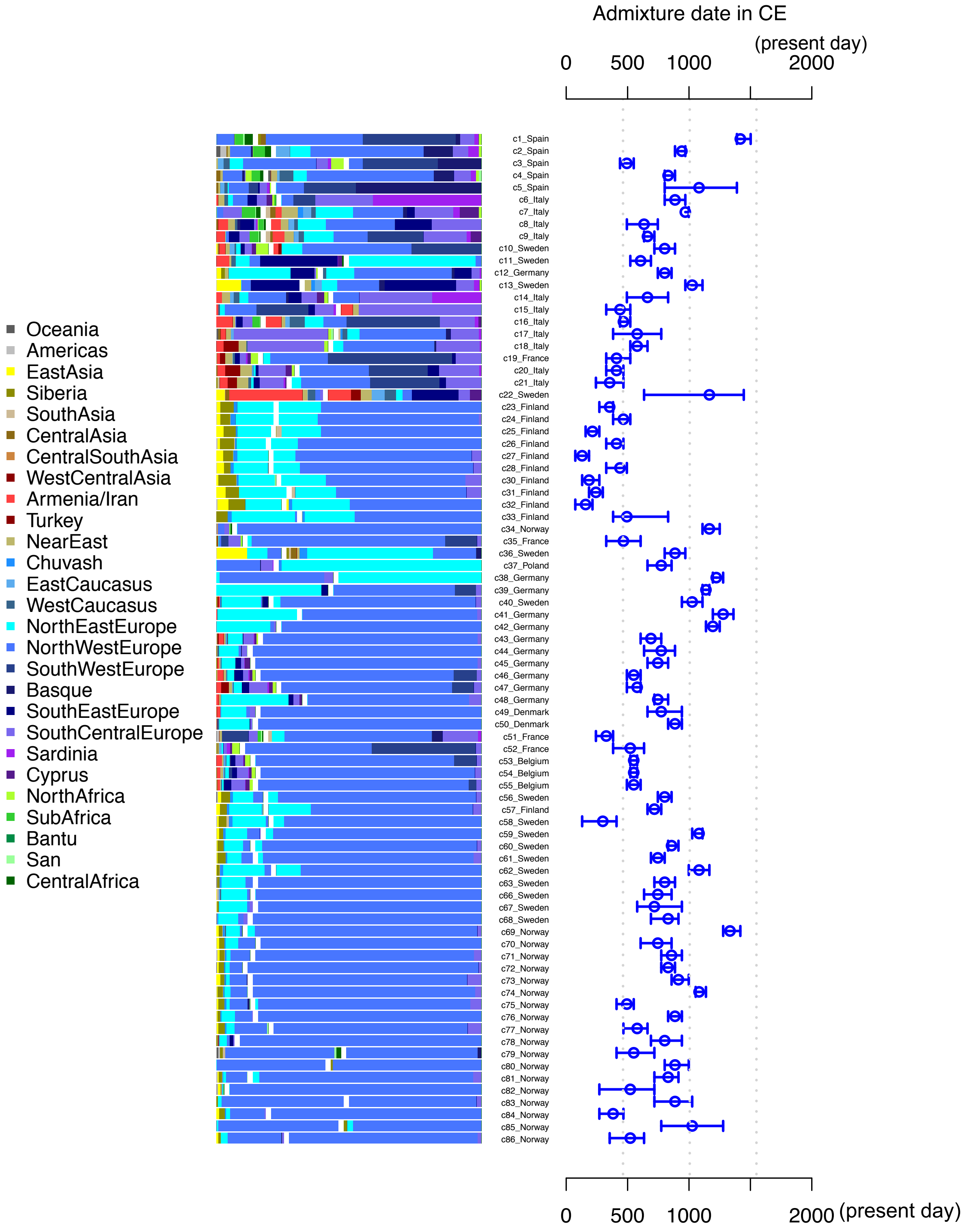

Figure 5.4: Summarized admixture inference of European clusters c1-c86, for the most strongly signalled admixture event in cases where > 2 admixing sources were detected. Bar plots give the inferred genetic make-up of each source, as summarized by their proportion of recent ancestry sharing with reference individuals from the labeled regions (colors). White bars in the barplot separate the two inferred admixing sources and are placed to show the inferred proportions of admixture from each source. To the right of each barplot is the cluster label, followed by the country label most represented among individuals in that cluster. Inferred dates (circles = point estimate, lines = 95% CIs) are given on the right, converted to years using the formula $1950 - (28 \times (\lambda - 1)$ with 28 years per generation.

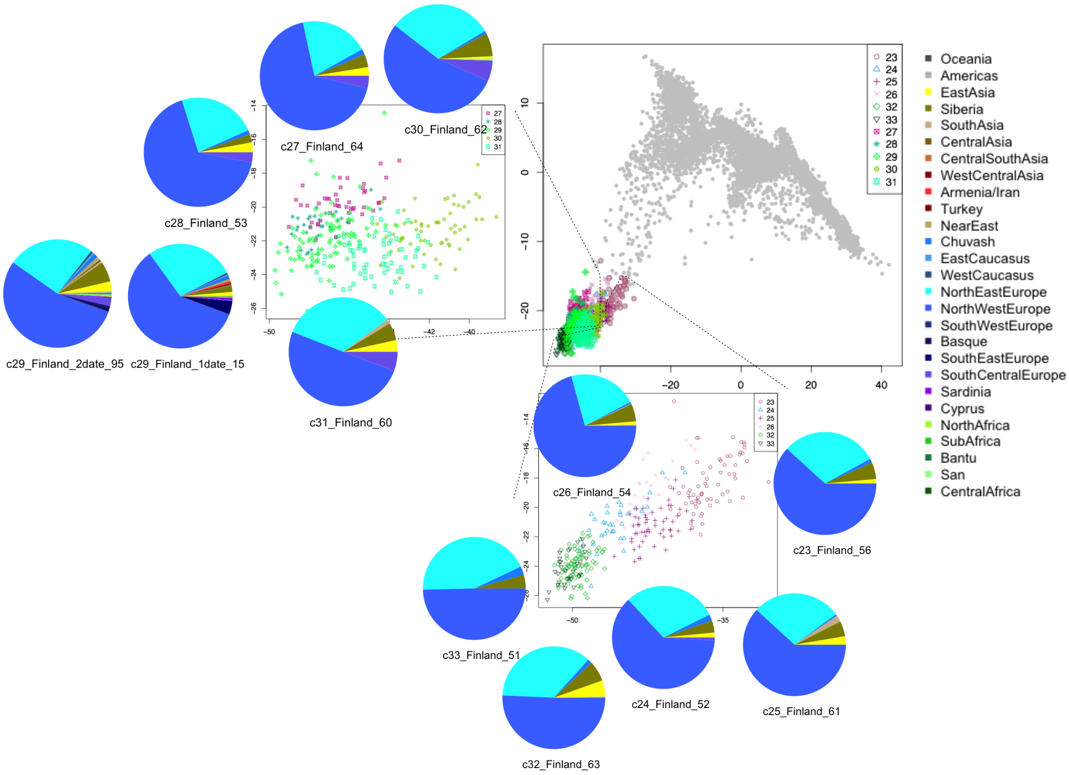

Figure 5.5: Pie illustrate the proportions of DNA that individuals in each Finnish cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

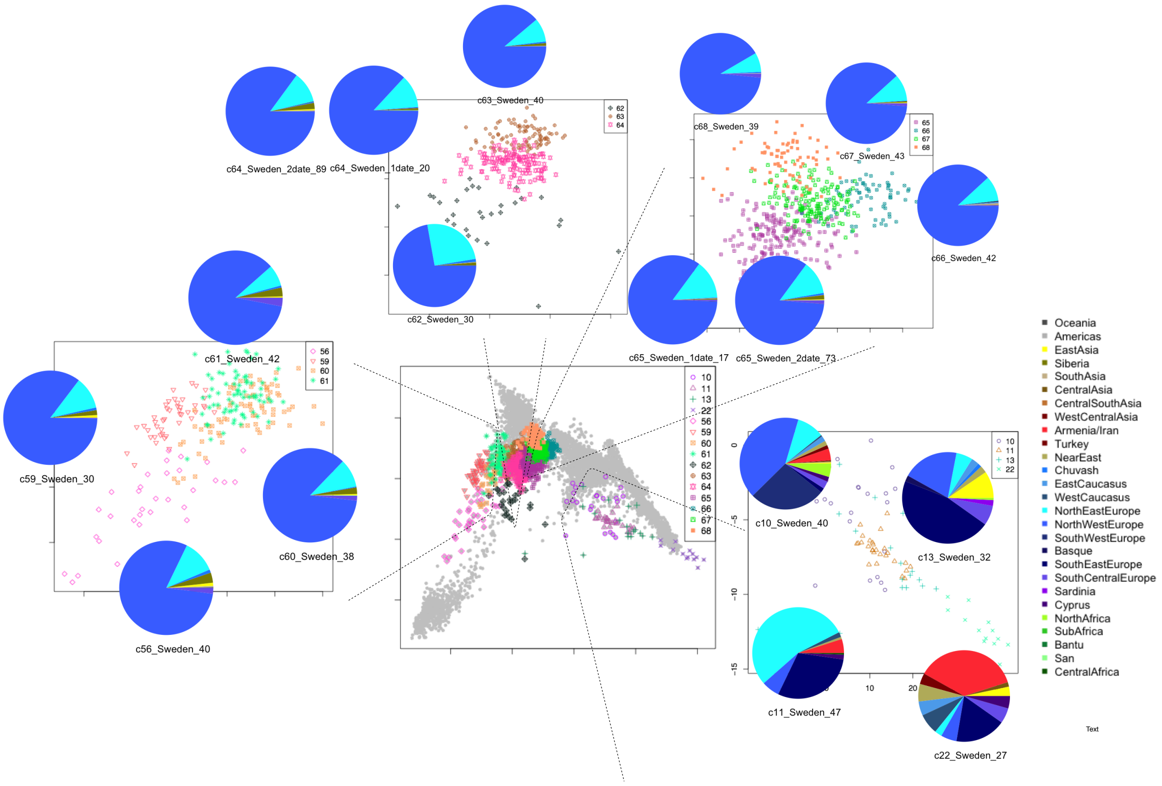

Figure 5.6: Pie illustrate the proportions of DNA that individuals in each Swedish cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

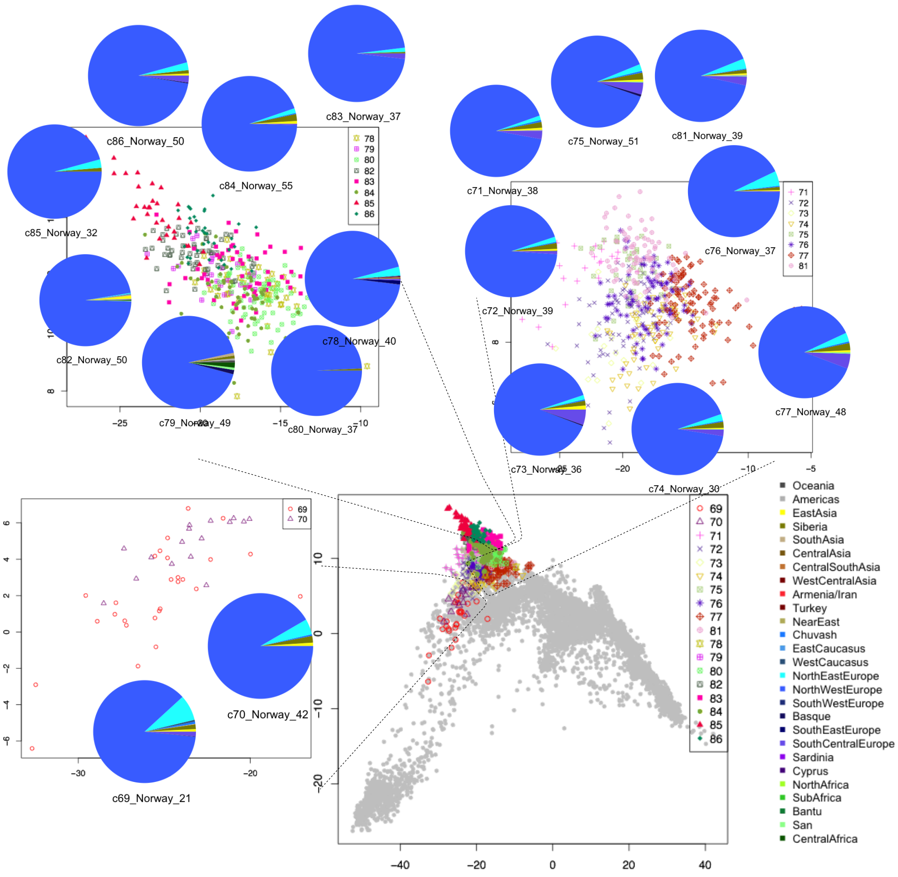

Figure 5.7: Pie illustrate the proportions of DNA that individuals in each Norwegian cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

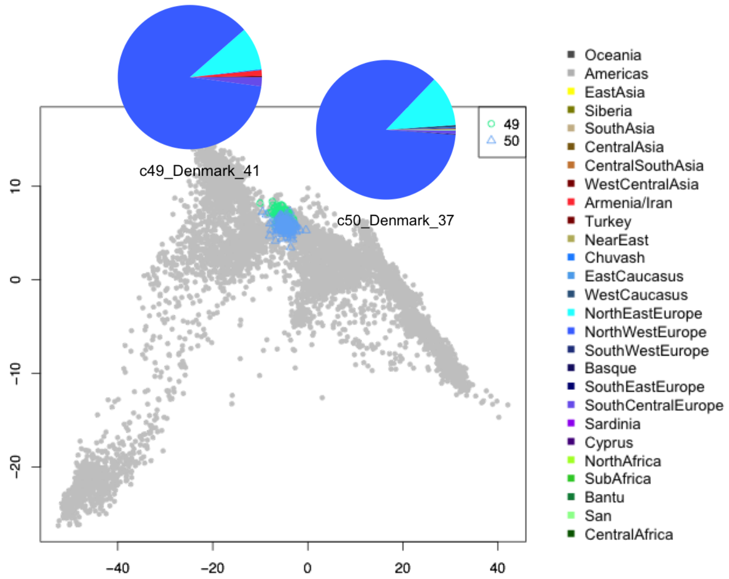

Figure 5.8: Pie illustrate the proportions of DNA that individuals in each Danish cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

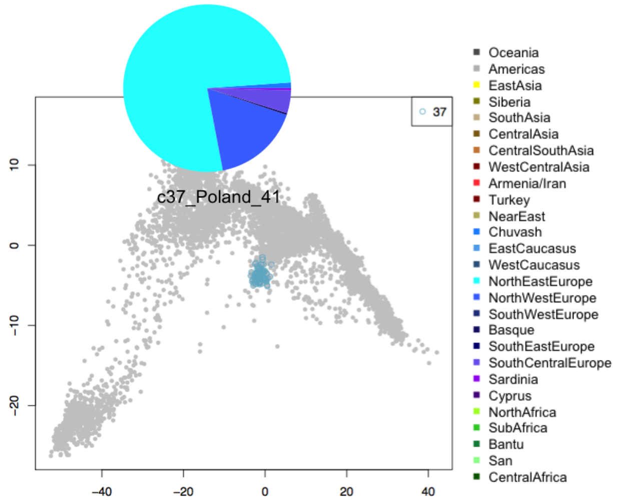

Figure 5.9: Pie illustrate the proportions of DNA that individuals in a Polish cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below the pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. The cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

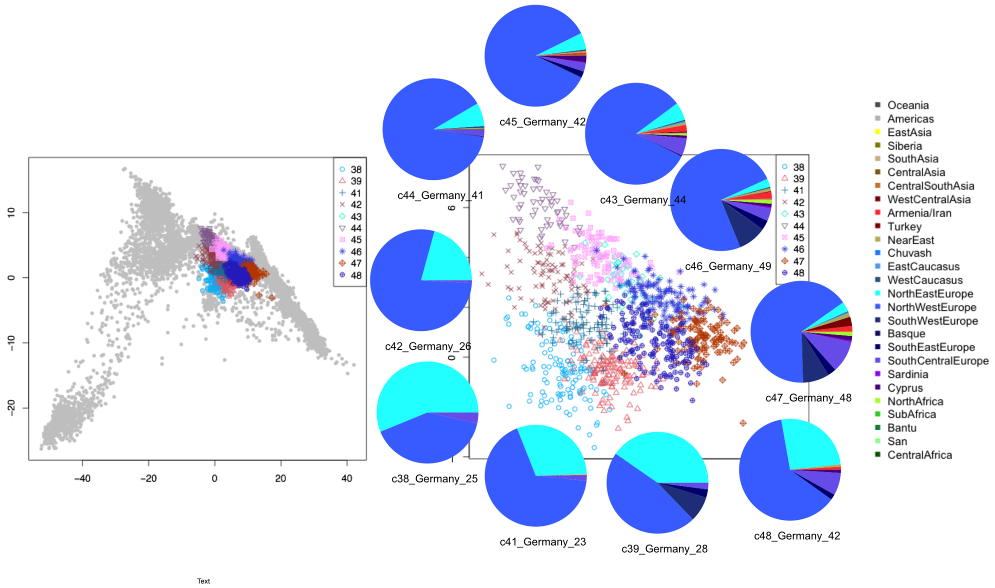

Figure 5.10: Pie illustrate the proportions of DNA that individuals in each German cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

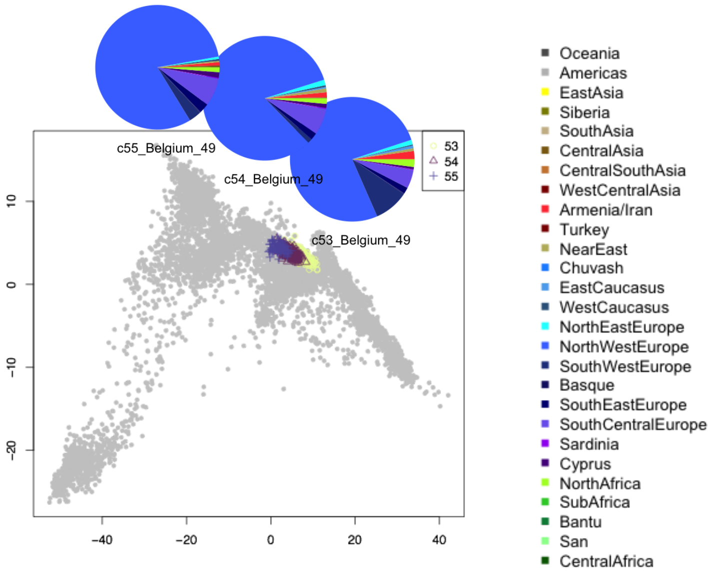

Figure 5.11: Pie illustrate the proportions of DNA that individuals in each Belgian cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

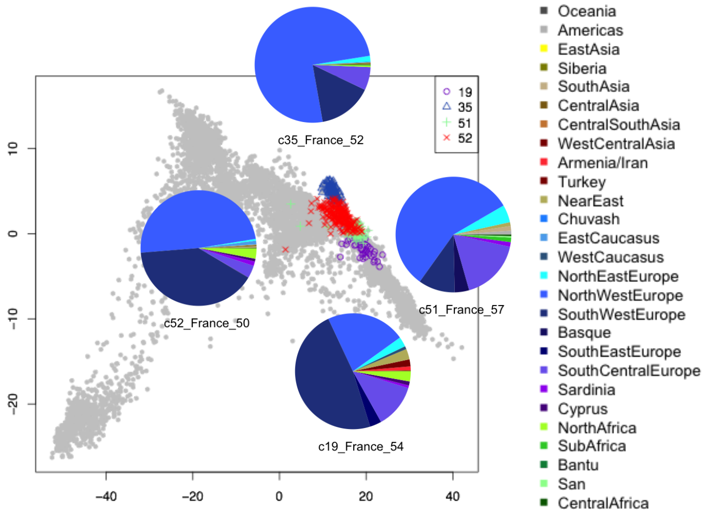

Figure 5.12: Pie illustrate the proportions of DNA that individuals in each French cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

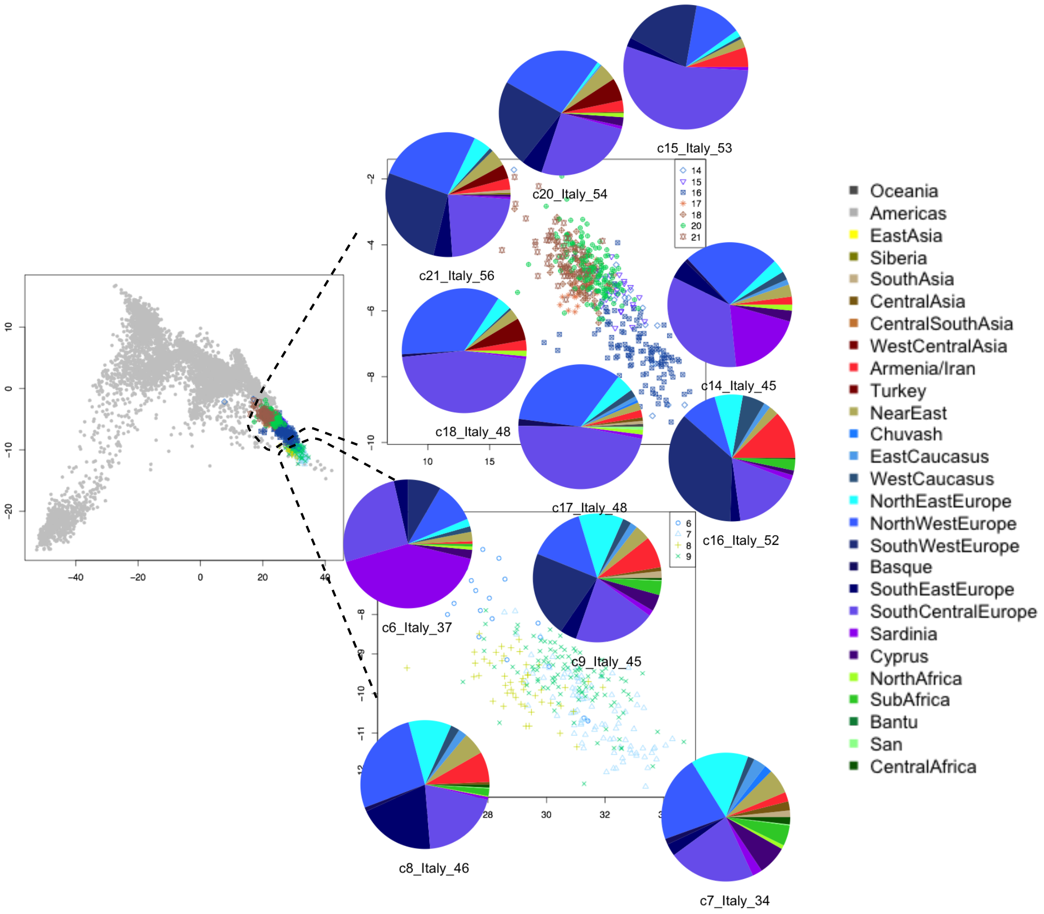

Figure 5.13: Pie illustrate the proportions of DNA that individuals in each Italian cluster (pie) are inferred, on average, to match individuals from each major geographic region (color). The labels below each pie separated by "_" represents cluster, major population in cluster, and admixture date in generations. Each cluster is depicted on the PCA plots as a unique color/symbol, with inset maps highlighting particular clusters.

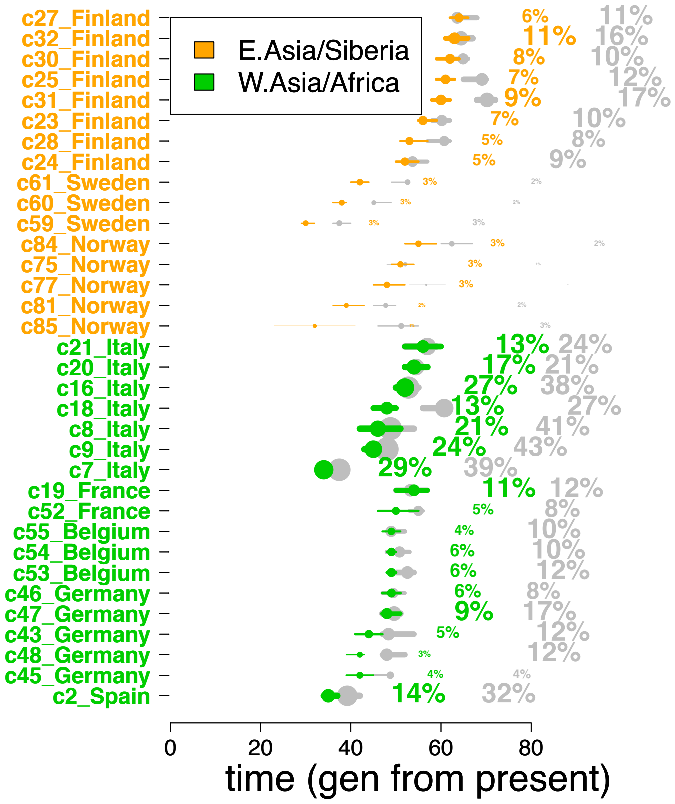

Figure 5.14: Traces of East Asian/Siberia (orange) or West Asian/North Afican (green) ancestry found in European clusters. The y-axis gives the cluster label and the label of the country whose individuals are most represented in that cluster, while the x-axis is the inferred date of admixture (dots = point estimate, lines = 95% CIs) in generations from the present. The size of the dots (dates) are proportional to the percentages to the right of the date, which give the total amount of ancestry each cluster shares with (orange) East Asia and Siberian reference populations or (green) West Asian, North African and Sub-Saharan African reference populations. Grey dots, CIs and percentages are corresponding inference from another study by Matthew Greenfield which used the same target clusters but excluded European reference groups (besides the UK) as donors and surrogates.

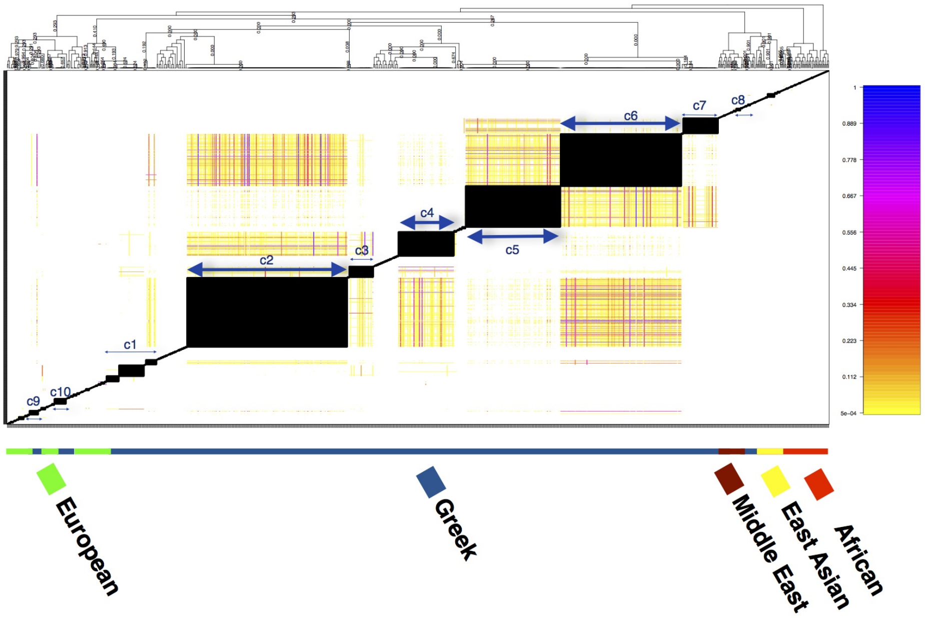

Figure 5.15: fineSTRUCTURE analysis on Greek and other reference samples sorted according to the inferred tree on the top, which also provides a confidence value for each branch. The heatmap shows, for each pair of individuals, the proportion of 500 MCMC samples where those two individuals are found in the same cluster (key at right). All proportions = 1 are in black. The Greek samples (blue) are indicated at the bottom bar, along with Africans (red), East Asians (yellow), Middle Easterners (brown) and Europeans (green). The Greek clusters c1-c10 (blue arrows and labels) are defined according to the proportion of co-occurrence and clustering in the fineSTRUCTURE tree.

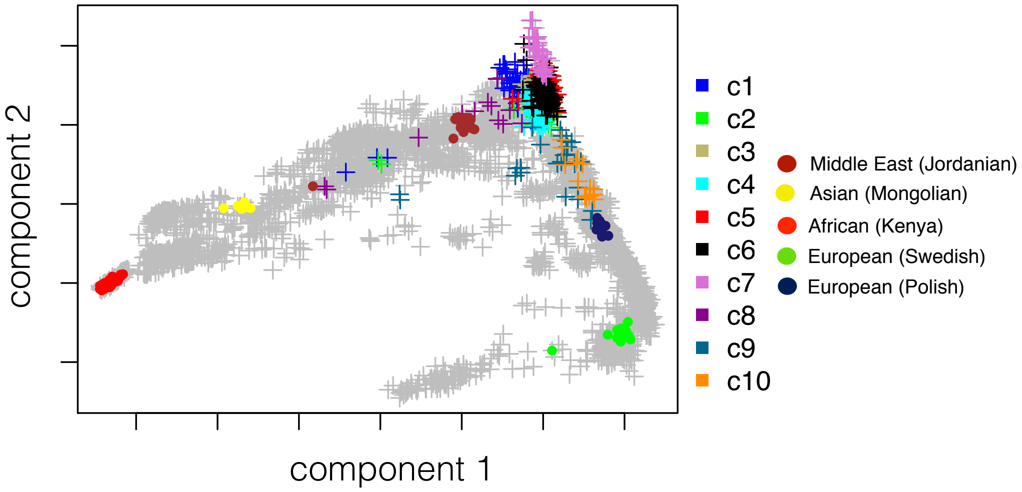

Figure 5.16: PCA of Greek individuals (crosses) colored by clusters c1-c10 and individuals from world reference populations (grey). The circles indicate sampled populations from Africa (red), East Asia (yellow), Middle East (brown) and Europe (blue and green).

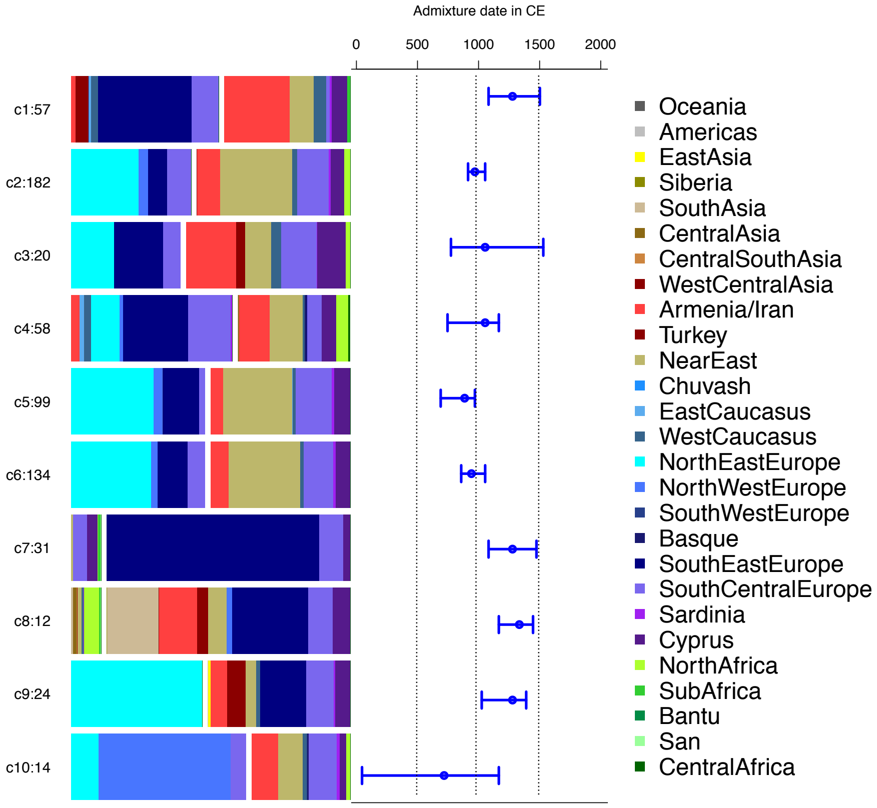

Figure 5.17: Admixture events inferred in Greek clusters c1-c10, with the most strongly signalled admixture event shown in cases where > 2 admixing sources were detected. Bar plots give the inferred genetic make-up of each source, as summarized by their proportion of recent ancestry sharing with reference individuals from the labeled regions (color legend at right). White bars in the barplot separate the two inferred admixing sources and are placed to show the inferred proportions of admixture from each source. To the left of each barplot is the cluster label and number of samples in each cluster separated by a colon ":". Inferred dates (circles = point estimate, lines = 95% CIs) are given in the middle, converted to current era (CE) using the formula $1950-(28 \times (\lambda - 1)$ with 28 years per generation.

Chapter 6

fastGLOBETROTTER instruction and tutorials

In this chapter, I provide user instructions for fastGLOBETROTTER, detailing the set-up, parameter options, input file formats and commands to execute the program. The second section of this chapter provides a step-by-step example of a real data analysis, beginning from haplotype phasing and proceeding to chromosome painting, admixture dating and interpreting results.

6.1 User Instruction

6.1.1 Introduction

fastGLOBETROTTER is a program for inferring and dating admixture in populations. It follows much of the protocol described in GLOBETROTTER [21], but provides faster and more accurate inference and is suitable for large-scale data (e.g. more individuals and larger numbers of SNPs). To describe each admixing event within a target population, fastGLOBETROTTER uses genetic information from multiple sampled reference groups or "surrogates" that may be related to ancestral sources of the target population. In particular fastGLOBETROTTER identifies whether the target population descends from multiple sources that intermixed at one or more times in the past, with each such source described as a mixture of the sampled surrogates provided by the user. The date(s) of admixture is determined by modelling how linkage disequilibrium (LD) among genetic segments inherited from the surrogates decays versus genetic distance.

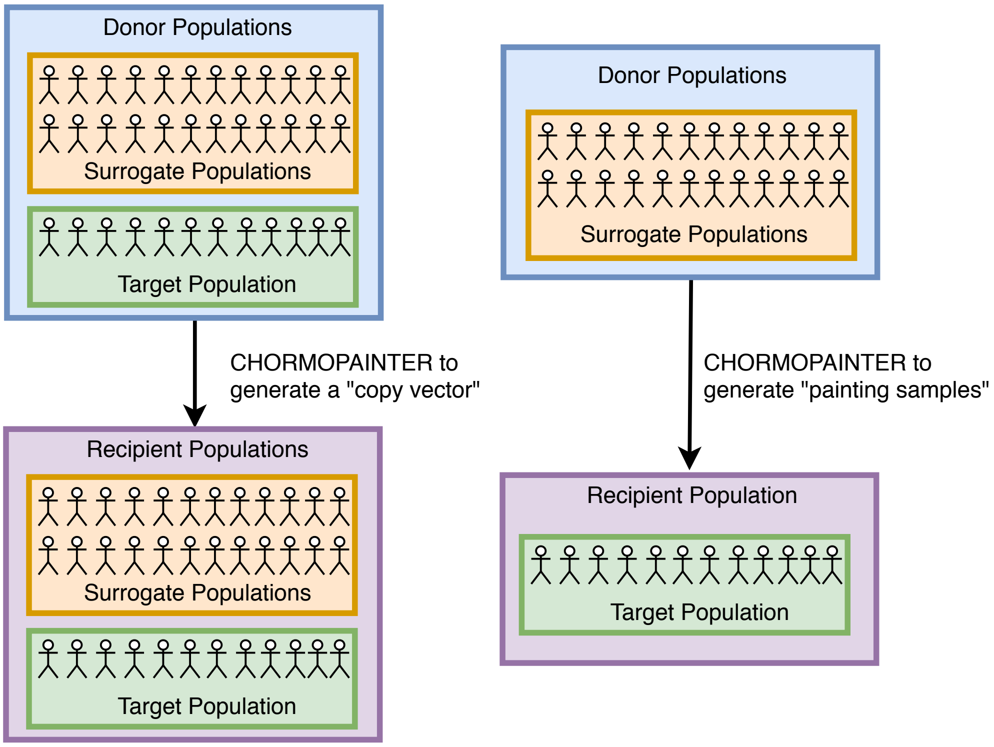

Figure 6.1: Sample arrangement used in CHROMOPAINTER/fastGLOBETROTTER analysis. Left – "all-versus-all analysis" where a copy vector is built by painting recipients (all surrogate and target individuals) conditional on using donors as surrogate and target. Right – "painting samples" are produced by painting all target individuals as a mixture of donor individuals (only surrogate groups are used as donor not the target).

In this user manual, "target population" refers to the putatively admixed population, "surrogate" refers to the sampled populations that represent potential sources of ancestry in the target. "Donors" and "recipients" are CHROMOPAINTER terminology that are used to indicate roles in painting process, "donors" are the sampled populations used to describe haplotype patterns in the "recipients" which can be target and surrogate populations (i.e. by using CHROMOPAINTER to paint each surrogate and target individual against a set of "donor" individuals). The steps to perform fastGLOBETROTTER inference are similar to those for GLOBETROTTER, with all details provided below.

6.1.2 Preparing inputs for fastGLOBETROTTER

Prior to fastGLOBETROTTER, users need to perform two CHROMOPAINTER analyses to obtain two files i.e. 1) a copying vector and 2) painting samples (Figure 6.1).

1. "a copying vector" file – a CHROMOPAINTER *XXX.chunklengths.out* output file produced by painting all surrogate and target individuals conditional on a set of donor individuals. (In practice, e.g. if testing multiple targets using the same dataset, CHROMOPAINTER is run allowing all individuals to copy from each other by using the '-a' switch in CHROMOPAINTER.)

2. "painting samples" file – a CHROMOPAINTER *XXX.samples.out* output file produced by painting all target individuals conditional on a set of donor individuals. The set of donor individuals should ideally be the same set as that used in (1), though – most critically – the target population should not be included among the donors, as doing so will mask admixture signals (see Section 6.2.3 and 6.2.4 for the details of this step).

6.1.3 Getting started

fastGLOBETROTTER was developed mainly in R, though accompanied by computing functions in C. To install the program, first extract the files in the .tar ball and then compile using the following command:

```
    R CMD SHLIB -o fastGLOBETROTTERCompanion.so
fastGLOBETROTTERCompanion.c -lz
```

Note that you must have "zlib" installed (e.g. sudo apt-get install zlib1g-dev). You must also have the package "nnls" installed in R (i.e. install.packages("nnls")). Note also that in order to run fastGLOBETROTTER on your machine, you may need to change the line in *fastGLOBETROTTER.R* that reads dyn.load("fastGLOBETROTTERCompanion.so") to include the pathway directory, i.e. dyn.load("/directorypath/fastGLOBETROTTERCompanion.so"). The fastGLOBETROTTER command line is as follows:

```
    R < fastGLOBETROTTER.R [parameter_infile]
[painting_samples_filelist_infile] [recom_rate_filelist_infile]
[running_mode] --no-save > [screen_output]
```

There are 4 required command paramters which are:

1. **parameter_infile** contains a description of all the parameters to use in fastGLOBETROTTER (see Section 6.1.4.1).
2. **painting_samples_filelist_infile** contains a list of the *XXX.samples.out* files (e.g. one file per chromosome) from a CHROMOPAINTER analysis of the target population conditional on the donors (see Section 6.2.4).
3. **recom_rate_filelist_infile** contains a list of the recombination rate files used when running CHROMOPAINTER. The file order must identical to the chromosome order in **painting_samples_filelist_infile**.
4. **running_mode** is used to specify four different fastGLOBETROTTER modes:

   "1" - run at maximum speed, though requiring greater memory (RAM). This mode is the fastest mode of fastGLOBETROTTER where it uses sampling algorithm and code optimization to reduce computational complexity. It is relatively 20-fold faster than the original GLOBETROTTER. The drawback of this mode is that it consumes more memory than other modes, however, it is most desirable mode if users can afford the required resource.

   "2" - run at maximum speed and require less memory than mode 1. fastGLOBETROTTER's mode 2 employs sampling algorithm, donor merging algorithm and code optimization making it more flexible for users in that it requires less memory than mode 1 while still runs at maximum speed (20x). However, the inferred ancestral sources may be inaccurate in some cases. This mode is suitable for

   "3" - run at less speed but require less memory. This mode only uses sampling algorithm to reduce complexity making it memory efficient but 3-4 times slower than mode 1 and 2. This mode is suitable for users who have limited resource.

"mem" - calculate memory (RAM) required by fastGLOBETROTTER in Mode 1,2 and 3. This mode should be execute prior fastGLOBETROTTER analysis, so users can select the most suitable mode based on resource availability.

6.1.4 Input Files

6.1.4.1 parameter_infile

The parameter file (see example in "*tutorial/paramfile.txt*") contains 20 rows. Below is a description of the parameters in each of these rows, which need to be formatted (and ordered) as shown in bold type, with brackets containing allowed values and followed by a description.

- **prop.ind: [0,1]** - indicate whether ("1") or not ("0") to infer admixture proportions, dates and sources (if "0", this information will be read from previously made fastGLOBETROTTER files specified by save.file.main)
- **bootstrap.date.ind: [0,1,2]** - "1" to perform bootstrap re-sampling to infer confidence intervals around date estimates, "2" to perform jackknife resampling and "0" for no action
- **null.ind: [0,1]** - indicate whether to standardize by a "NULL" individual when performing inference (recommended; this is used for inferring p-values for evidence of admixture and is also appropriate when "target" population has likely undergone bottleneck effects and in general testing for consistency)
- **input.file.ids: [input.filename1]** - pathway and name for file containing id labels for all samples, for the CHROMOPAINTER analysis run to make the *XXX.samples.out* files
- **input.file.copyvectors: [input.filename2]** - pathway and name for file containing copy vectors for all surrogate and target populations
- **save.file.main: [output.filename1]** - pathway and name (prefix) for main output file
- **save.file.bootstraps: [output.filename2]** - pathway and name (prefix) for inferred date bootstrap/jackknife output file
- **copyvector.popnames: [pop 1 pop 2 ... pop k]** - names of all *k* popula-

tions used as donors; i.e. that both surrogate and target populations copied from when running CHROMOPAINTER (NOTE: Any painted segments in the *XXX.samples.out* files that select the target population as a donor will be ignored, even if you include the target population in this line of the input file.)

- **surrogate.popnames: [pop 1 pop 2 ... pop j]** - names of all *j* surrogate populations, i.e. used to describe admixture in target.popname
- **target.popname: [pop rec]** - name of target population
- **num.mixing.iterations: [0,1,...,5,...]** - number of iterations of date and proportion/source estimation to perform; "0" specifies to only infer proportions of ancestry relating the target to each surrogate, and to not try and infer/date admixture events (only used when prop.ind: 1)
- **props.cutoff: [0.0,...,1.0]** - at each iteration, remove any surrogates that contribute $\leq$ this value to the mixture describing the target population
- **bootstrap.num: [0,1,...]** - number of bootstrap/jackknife re-samples (only used when bootstrap.date.ind: 1 or 2)
- **num.admixdates.bootstrap: [1,2]** - number of dates to fit when performing bootstrap/jackknife resampling (only used when bootstrap.date.ind: 1 or 2)
- **num.surrogatepops.perplot: [1,...]** - will plot this number squared of coancestry curves for each page of the curves output file (only used when prop.ind: 1)
- **curve.range: [lower.lim upper.lim]** - lower and upper bounds of x-axis (i.e. cM distance between DNA segments) to fit dates to when generating coancestry curves
- **bin.width: [e.g. 0.1]** - width of x-axis bins (in cM) when generating coancestry curves
- **xlim.plot: [lower.lim upper.lim]** - lower and upper bounds (in cM) of x-axis to plot for coancestry curves (only used when prop.ind: 1)
- **prop.continue.ind: [0,1]** - indicate whether you are continuing proportion estimation from those in a previous file (in which case the previous file will be read from save.file.main and output files will add the suffix "_continue")

- **haploid.ind: [0,1]** - indicate whether individuals are haploid (“1”) or diploid (“0”)

input.file.ids - This file should match exactly the donor input file used in ChromoPainterv2 (‘-t’ switch) when generating the *XXX.samples.out* files. An example of **input.file.ids** is provided in “*tutorial/individual.txt*”. Each row is ordered to match each row of the genotype input file (’-g’) run using CHROMOPAINTER. There are three columns per row, with the first column giving the individual identifier, the second column giving the individual’s population label and the third column an indicator for whether the individual is not included in the analysis (use “0” to specify NOT to include the given individual). For example, consider a file with the following 7 individuals:

IND1	Pop1	0
IND3	Pop1	1
IND2	Pop1	1
IND4	Pop2	1
IND5	Pop2	0
Pop4Ind1	Pop4	1
IND7	Pop1	1

Here we only include individual IND3, IND2, IND4, Pop4Ind1, IND7 while IND1 and IND5 are excluded from the analysis. Each population label specified in **copyvector.popnames**, **surrogate.popnames** and **target.popname** of the file “parameter_infile” MUST be in column 2 of at least one row of the file **input.file.ids**. An exception, incorporated to make things more flexible for the user, is if all **surrogate.popnames** and **target.popname** labels missing from column 2 of input.file.ids are specified in the row labels of input.file.copyvectors, and similarly all **copyvector.popnames** missing from column 2 of **input.file.ids** are specified in the column labels of **input.file.copyvectors**. In other words, the column names and row names of input.file.copyvectors must contain the individual identifiers and/or the population labels.

It is critical that the order of individuals in **input.file.ids** corresponds to the donor indices used in the *XXX.samples.out* files used in the analysis. Each painting sample (row) of each *XXX.samples.out* file gives a number D for each SNP, corresponding to the row of the CHROMOPAINTER haplotype input file ('-g' switch) that contains the donor haplotype copied at that SNP (where the first haplotype in this input file is assigned $D = 1$). The row containing the label for this donor individual in **input.file.ids** MUST be row D/p (with any decimal values of D/p rounded up to the nearest integer) where $p = 1, 2$ is the ploidy of the organism.

input.file.copyvectors - This file should contain the *XXX.chunklengths.out* file from the corresponding CHROMOPAINTER analyses, in the same format and combined across all chromosomes and individuals. Each row is an individual (or population), and the columns give the total amount of genome-wide DNA that the given individual (or population) is inferred to copy from every "donor" individual (or population) in the corresponding CHROMOPAINTER analyses. An example of **input.file.copyvectors** is provided in "*tutorial/copyvector.txt*". The first row of **input.file.copyvectors** lists the column labels reflecting the donor individuals and/or populations. The first column in this first row is "recipient", and the remaining columns of this first row must contain one of the labels specified in **copyvector.popnames** of "parameter_infile". The remaining rows of **input.file.copyvectors** list the "recipient" individual (or population) label in the first column, with the remaining columns containing the total amount (or proportion) of genome-wide DNA that the given recipient individual (or population) copies from each donor label provided in the first row.

6.1.4.2 painting_samples_filelist_infile

This file contains a list of file locations and names of *XXX.samples.out* output files from CHROMOPAINTER, specifying one file per line (see *tutorial/samplefile.txt*).

6.1.4.3 recom_rates_filelist_infile

This file is a list of file locations and names of the recobination rate files used when running CHROMOPAINTER to make the *XXX.samples.out* files. The chromosome order must correspond to the order listed in the painting_samples_filelist_infile (See *tutorial/recomfile.txt*).

6.1.5 Output

There are four output files from fastGLOBETROTTER:

6.1.5.1 [output.filename1].txt

This file summarizes the inferred admixture proportions, dates and sources. The first line lists our "best-guess" conclusion for admixture in the target population. The conclusion can be:

- **uncertain** - admixture is detected but difficult to describe (technical details: combined fit quality for two events "fit.quality.2events" < 0.985)
- **one-date** - a single date of admixture between two sources (combined fit quality for two events ≥ 0.985; two-date score "maxScore.2events" <0.35; fit-quality for a single event "fit.quality.1event" ≥ 0.975)
- **one-date-multiway** - a single date of admixture between more than two sources (combined fit quality for two events ≥ 0.985; two-date score < 0.35; fit quality for a single event < 0.975)
- **multiple-dates** - two (or more) distinct dates of admixture between two or more sources (combined fit quality for two events ≥ 0.985; two-date score ≥ 0.35)
- **no admixture**- admixture is undetected or unclear

Note that these are just guidelines, and that we highly recommend careful visual exploration of the inferred coancestry curves provided in the .pdf output file to see how well e.g. one versus two events fit the data.

The next lines ("1-DATE FIT EVIDENCE, DATE ESTIMATE, SINGLE BEST- FITTING DONORS") provide the fastGLOBETROTTER inferred date, proportions and "best-guess" sources of admixture for a single event or multiway admixture when assuming only a single date of admixture (i.e. this information is particularly appropriate when the "best-guess" conclusion is "one-date" or "one-date-multiway"), as well as measures of "goodness-of-fit" for these events. Specifically:

- **gen.1date** - inferred date of admixture in generations from present
- **proportion.source1** - inferred proportion of admixture from the minority

contributing source

- **maxR2fit.1date** - the goodness-of-fit (R2) for a single date of admixture, taking the maximum value across all inferred coancestry curves
- **fit.quality.1event** - the fit of a single admixture event
- **fit.quality.2events** the fit of the first and the second admixture event
- **bestmatch.event1.source1** - the single "best-guess" surrogate population that matches the inferred minority contributing source
- **bestmatch.event1.source2** - the single "best-guess" surrogate population that matches the inferred majority contributing source
- **proportion.event2.source1** - inferred proportion of admixture from the minority contributing source for the second, less strongly signaled event (appropriate for "one-date-multiway")
- **bestmatch.event2.source1** - the single "best-guess" surrogate population that matches the inferred minority contributing source for the second, less strongly signaled event (appropriate for "one-date-multiway")
- **bestmatch.event2.source2** - the single "best-guess" surrogate population that matches the inferred majority contributing source for the second, less strongly signaled event (appropriate for "one-date-multiway")

The next lines of output ("2-DATE FIT EVIDENCE, DATE ESTIMATES, SINGLE BEST-FITTING DONORS") give 2 inferred admixture dates, proportions and "best-guess" sources of admixture when assuming two distinct dates of admixture (date1 and date2). These lines should be considered only if the "best-guess" conclusion is "multiple-dates". In particular:

- **gen.2dates.date1** - inferred date of admixture (in generations from present) for the first (most strongly signaled) event, when assuming two dates
- **gen.2dates.date2** - inferred date of admixture (in generations from present) for the second event, when assuming two dates
- **maxScore.2events** - the additional goodness-of-fit (R2) explained by adding a second date versus assuming only a single date of admixture, taking the maximum such value across all inferred coancestry curves

- **proportion.date1.source1** - inferred proportion of admixture from the minority contributing source for the first date's event (when assuming two dates)
- **bestmatch.date1.source1** - the single "best-guess" surrogate population that matches the inferred minority contributing source for the first date's event (when assuming two dates)
- **bestmatch.date1.source2** - the single "best-guess" surrogate population that matches the inferred majority contributing source for the first date's event (when assuming two dates)
- **proportion.date2.source1** - inferred proportion of admixture from the minority contributing source for the second date's event (when assuming two dates)
- **bestmatch.date2.source1** - the single "best-guess" surrogate population that matches the inferred minority contributing source for the second date's event (when assuming two dates)
- **bestmatch.date2.source2** - the single "best-guess" surrogate population that matches the inferred majority contributing source for the second date's event (when assuming two dates)

The next lines of output ("1-DATE FIT SOURCES, PC1") present the fastGLOBETROTTER inferred composition of each admixing source in the most strongly signaled event, when assuming only a single date of admixture. In particular every two consecutive rows describe the inferred genetic composition of one admixing source (i.e. where each source is described as a mixture of the sampled surrogate groups), giving both the proportion of DNA contributed by that source (first column), and fastGLOBETROTTER's inferred mixture coefficients to describe each source (remaining columns - these should sum to 1 for each source). For instance, consider the following output:

```
###########################
### 1-DATE FIT SOURCES, PC1:
proportion BantuKenya Mandenka BantuSouthAfrica
```

```
0.29 0.26 0.31 0.43
proportion Han Japanese Balochi Druze Sardinian Ireland English
0.71 0.01 0.01 0.02 0.03 0.03 0.16 0.74
###########################
```

implies that fastGLOBETROTTER has inferred the first admixing source, which contributes 29% of the DNA of the target population, to be best represented genetically as a mixture of (0.26, 0.31, 0.43) times the copy vectors of surrogate labels (BantuKenya, Mandenka, BantuSouthAfrica), respectively. And fastGLOBETROTTER has inferred the second admixing source, which contributes 71% of the DNA of the target population, to be best represented genetically as a mixture of (0.01, 0.01, 0.02, 0.03, 0.03, 0.16, 0.74) times the copy vectors of surrogate labels (Han, Japanese, Balochi, Druze, Sardinian, Ireland, English), respectively. Similar source proportion and mixing coefficient inference is given next for the less strongly signaled event when assuming a single date of admixture (i.e. 1-DATE FIT SOURCES, PC2"), which is particularly appropriate when the best-guess conclusion is "one-date-multiway".

Following this is the analogous inference for the first date's event when assuming two dates of admixture ("2-DATE FIT SOURCES, DATE1-PC1"), and the second date's event when assuming two dates of admixture ("2-DATE FIT SOURCES, DATE2-PC1"), which is particularly appropriate when the "best-guess" conclusion is "multiple- dates".

6.1.5.2 [output.filename1]_curves.txt

This output file, with prefix specified by **save.file.main** in "parameter_infile" and suffix "_curves.txt", provides the numerical data of the coancestry curves for every pairwise combination of surrogate populations inferred to match > **props.cutoff** of the total ancestry of the target population. It is generated only if **prop.ind: 1** in "parameter_infile".

6.1.5.3 [output.filename1].pdf

This output file, with prefix specified by **save.file.main** in "parameter_infile" and suffix ".pdf", visualizes plots of the coancestry curves for every pairwise combination of surrogate populations using data stored in **[output.filename1]_curves.txt**. The given surrogate pairing is specified in each plot's title. The x-axis gives genetic distance in cM between pairs of segments (see Section 6.1.4.1 for a description of how the bins and range for these cM distance values are specified). The y-axis gives the (weighted) probability of copying from the first and second surrogate populations listed in the title at a pair of DNA segments separated by the corresponding x-axis (i.e. cM distance) value.

6.1.5.4 [output.filename2].txt

This output file, with prefix specified by **save.file.bootstraps** in "parameter_infile" and suffix ".txt", gives the inferred dates and goodness-of-fit (R2) values for bootstrap/jackknife re-samples of individuals' DNA (see above for details of column values).

6.1.6 Inferring admixture using fastGLOBETROTTER

To run fastGLOBETROTTER under these settings type:

```
    R < fastGLOBETROTTER.R tutorial/paramfile.txt tutorial/
samplefile.txt tutorial/recomfile.txt mem --no-save > output.out
```

This will take a moment to finish and will output (in "output.out") the calculated memory required by each mode of fastGLOBETROTTER. This will assist users to decide a suitable mode according to their computational resources. We suggest considering from mode 1, 2 and 3 respectively. The example of the command when mode 1 is selected is as:

```
    R < fastGLOBETROTTER.R tutorial/paramfile.txt tutorial/
samplefile.txt tutorial/recomfile.txt 1 --no-save > output.out
```

Once the program finishes, the output files will be produced in the identified directory.

6.1.7 Computational time and memory

Assume N target population individuals, C chromosomes, B bootstrap/jackknife resamples, M mixing iterations, S painting samples, L SNPs (maximum across all chromosomes), J donor populations, K surrogate groups, $I(\leq SL)$ total "chunks" (i.e. the maximum number of "chunks" across chromosomes and individuals), $I_j(\leq I)$ "chunks" copied from donor population j (the maximum number of "chunks" across chromosomes and individuals copied from a single donor population) and G grid points over which the coancestry curves are estimated (i.e. G= (**curve.range(upper.lim)-curve.range(lower.lim)**)**/bin.width** in section 6.1.4.1). Then the computational complexity of fastGLOBETROTTER at the maximum speed is

$$O[(B+M)(NC(SL+J^2I+I_j^2)+NGJ^2K^2)+C[min(N;100)]2(L+I_j^2)]$$

The maximum required memory for Mode 1 and 2 is $O(NGJ^2+NGK^2)$, while Mode 3 stores $O(NGK^2)$.

6.2 Tutorial

The aim of this tutorial is to walk users through the steps in inferring an admixture event which include 2 programs, CHROMOPAINTER [36] and fastGLOBETROTTER. We will begin with a simulated target population and attempt to infer admixture using the world populations as donors. In the last section, we will interpret the inference from the output file. Suppose we want to infer the unknown admixture event using 100 target individuals simulated as mixtures of present-day French people (contributing ≈70% of the DNA) intermixing with present-day Yoruban individuals (contibuting ≈30%) 30 generations ago. For simplicity, we will focus only on analysing chromosomes 21-22. We include the following groups (explored in Hellenthal 2014) listed in the Table below in our analysis.

Population	Region	Number of individuals
Balochi	Central South Asia	21
Balochi	Central South Asia	21
BantuKenya	Africa	11
BantuSouthAfrica	Africa	8
Druze	West Asia	42
EastSicilian	South Europe	10
English	Northwest Europe	6
Hadza	Central Africa	3
Han	Southeast Asia	34
Ireland	Northeest Europe	7
Japanese	Northeast Asia	28
Makrani	Central South Asia	22
Mandenka	West Africa	22
Maya	America	21
Naxi	Southeast Asia	8
Russian	East Europe	25
Sardinian	South Europe	28
Saudi	South Middle East	10
Scottish	Northwest Europe	6
She	Southeast Asia	10
Sindhi	Central South Asia	23
Surui	America	8
Syrian	South Middle East	16
Turkish	West Asia	17
UAE	South Middle East	9
Uygur	Central South Asia	10
Yemeni	South Middle East	4
FrenchYoruba	Simulated	100

Table 6.1: List of population region and number of samples included in the example.

6.2.1 Preparing input for CHROMOPAINTER

From the SNP data, we use PLINK [66] to prepare, QC and split the data into chromosomes (if necessary). Then we perform haplotype phasing using SHAPEIT [53], using the following command:

```
    shapeit --input-bed [bedfile] [bimfile] [famfile] -M
[genticmapfile] --output-max [phasedfile] [samplefile]
```

We then have to reformat the SHAPEIT output files into CHROMOPAINTER input format. This can be done using the perl script *"impute2chromopainter2.pl"* included in the fastGLOBETROTTER package or downloadable at `https://people.maths.bris.ac.uk/~madjl/finestructure/toolsummary.html`

```
     perl impute2chromopainter2.pl [phasedfile] [genticmapfile]
[output]
```

After this step, we have our input files ready for CHROMOPAINTER as included in the folder /tutorial, e.g. where *AllFrenchYoruba30gen30propchrXX.txt* file is a haplotype-phased file and *Chromxxx.recomrates* is its corresponding recombination rate file.

```
     tar -xzvf ChromoPainterv2.tar.gz
```

We then compile:

```
     gcc -o ChromoPainterv2 ChromoPainterv2.c -lm -lz
```

As mentioned in the fastGLOBETREOTTER User Instructions Section 6.1.2, there are 2 fastGLOBETROTTER inputs required from CHROMOPAINTER: (1) a copy vector file and (2) painting sample files. Here are the steps to obtain those files.

6.2.2 Estimating parameters required by CHROMOPAINTER

Before performing chromosome painting, we need to estimate 2 parameters, the switch (n) and mutation (M) rates, that are required in CHROMOPAINTER. To do this, we use Expectation-Maximization (EM), applying ChromoPainterv2 only to a subset of the data (i.e. only a subset of individuals and chromosomes) to save computational time. The command for estimating n and M on chromosome 21 is:

```
     ChromoPainterv2 -g AllFrenchYoruba30gen30propchr21.txt
-r Chrom21.recomrates -t individual.txt -a 1 10 -s 0 -i 10
-in -iM -o output_estimateEM_Chr21
```

Here, “-a 1 10” specified to only paint individuals 1-10 out of all included individuals to save computational time, “-s 0” specifies that no painting samples are needed in this step (as these take up storage and will not be used), “-i 10” – specifies to use 10 EM iterations to estimate parameters, and “-in -iM” – specifies that the switch (n) and mutation (M) rates should be estimated with EM.

We might apply the command to other chromosomes as well e.g. 1, 5, 10, 15, 20, etc. After completing the n and M estimation, we apply a perl script "*ChromoPainterv2EstimatedNeMutExtractEM.pl*" included in this package to summarize these inferred values across individuals. We also have to modify some parameters in the perl script such as output file prefix/suffix, the chromosomes analysed, and the number of SNPs per chromosome (which is used to weight the inference from each chromosome – see included file *ChromoPainterv2EstimatedNeMutExtractEM.pl*). The command is as below;

```
    perl ChromoPainterEstimatedNeExtractEM.pl
```

From the output of this perl script, we obtain n = 392.48 and M = 0.0004162 and we use these values throughout CHROMOPAINTER analyses.

6.2.3 Generating the copy vector input file

To construct the copy vectors, we use the estimated parameters (n, M) and run ChromoPainterv2 again, this time we paint all individuals using all individuals as donors on chromosome 21:

```
    ChromoPainterv2 -g AllFrenchYoruba30gen30propchr21.txt
-r Chrom21.recomrates -t individual.txt -a 0 0 -s 0 -n 392.48
-M 0.0004162 -o AllFrenchYoruba30gen30propchr21_AllvALL
```

The command "-a 0 0" specifies that we paint each individual in *individual.txt* using every other individual as a donor (Figure 6.1–Left). After the analysis is done for all included chromosomes, the *XXX.chunklengths.out* files from this analysis can be combined across chromosomes and into a single file using a perl script *"ChromoPainterOutputSum.pl"*. We also have to identify chromosome number in the perl script.

```
    perl ChromoPainterOutputSum.pl AllFrenchYoruba30gen30propch
_AllvALL.chunklengths.out
```

6.2.4 Generating the painting samples files

Here we create a population file that specifies which populations are recipients (the target population) and which are donors (likely all other populations). In this case, all reference populations are set as “D” (donors) and FrenchYoruba is set as as “R” (recipient; see *tutorial/popfile.txt*). The command for generating painting samples for chromosome 21 is as follows:

```
    ChromoPainterv2 -g AllFrenchYoruba30gen30propchr21.txt
-r Chrom21.recomrate -t individual.txt -f popfile.txt 0 0 -s
10 -n 392.48 -M 0.0004162 -o AllFrenchYoruba30gen30propchr21
_DonorvTarget
```

This is similar to the previous command except we remove “-a”, and we use the *popfile.txt* to determine which populations are to be painted and which are to be used as donors. In this case, each FrenchYoruba individual will be painted using all individuals from other populations as donors. (Figure 6.1–Right) Note that we do not set FrenchYoruba as a donor (in contrast to Section 6.2.3), because the painting samples used to date admixture events consider only segments matched to different surrogates (segments matching to other target individuals will be discarded, so that using your own population as a donor removes a lot of data). Lastly, stating “-s 10” means that 10 painting samples per each haploid genome of each target (recipient) individual will be produced and we apply this command to all other chromosomes in the same way.

6.2.5 Inferring admixture events using fastGLOBETROTTER

In this section, we will apply fastGLOBETROTTER to infer the admixture history of the target population (FrenchYoruba) using 26 reference populations as surrogates the admixing sources. We install fastGLOBETROTTER by unzipping and extracting the fastGLOBETROTTER.tar.gz file using the command:

```
    tar -xzvf fastGLOBETROTTER.tar.gz
```

and compiling the program:

```
    R CMD SHLIB -o fastGLOBETROTTERCompanion.so
fastGLOBETROTTERCompanion.c -lz
```

Note you may need add a new path to a line in *fastGLOBETROTTER.R* that reads dyn.load(“fastGLOBETROTTERCompanion.so”) to dyn.load(“/directorypath/fastGLOBETROTTERCompanion.so”). To run fastGLOBETROTTER, you have to prepare three files:

- a parameter file, *paramfile.txt*, that lists the fastGLOBETROTTER parameters, including which population is the target, and which populations should be specified as donors and surrogates, etc.
- a painting samples file, *samplefile.txt*, that collects the file locations of *AllFrenchYoruba30gen30propchrXX_DonorvTarget.samples.out.gz* of all included chromosomes (these files can be in gzipped format).
- a recombination rate file, *recomfile.txt*, that collects the file locations of *ChromXX.recomrates* of all included chromosomes.

In the parameter file, we specify "prop.ind: 1" to infer and date admixture, and "bootstrap.date.ind: 1" to perform bootstrap re-samples to infer confidence intervals around the inferred date (using "bootstrap.num: 20" such re-samples). To account for linkage disequilibrium patterns that may not be due to genuine admixture (e.g. if the target population has experienced a strong bottleneck since admixture), we suggest setting "null.ind: 1". We set "input.file.ids:" as *individual.txt*; this file must be the same file as we use in ChromoPainterv2 above. The output filename (save.file.main:) is defined according to the user's preference here we use *AllFrenchYoruba30gen30prop.main.txt*. We also specify the "input.file.copyvectors:" made in Section 6.2.3 as *copyvector.txt*. "copyvector.popnames:" contains the list of donor populations used in ChromoPainterv2, which we also replicate as our "surrogate.popnames:" (which is typically done in practice).

As mentioned, fastGLOBETROTTER allows users to select a suitable "running_mode" that depends on user's memory (RAM) availability. The fastest mode is "Mode 1" which consumes more memory than other modes, whereas "Mode 2" is faster while requiring less memory than "Mode 1." but the ancestral inference may not be accurate. "Mode 3" normally requires minimal memory, but is 2-4 times slower than Modes 1 and 2. To select the desired Mode, users can calculate RAM required for each mode by performing this command with the "mem" setting:

```
    R < fastGLOBETROTTER.R tutorial/paramfile.txt tutorial/
samplefile.txt tutorial/recomfile.txt mem -no-save >
```

```
AllFrenchYoruba30gen30prop.out
```

The result of the memory test will be printed in AllFrenchYoruba30gen30prop.out. Users then can choose which Mode suits their available resources. Below is an example of a command line when Mode 1 is selected.

```
    R < fastGLOBETROTTER.R tutorial/paramfile.txt tutorial/
samplefile.txt tutorial/recomfile.txt 1 -no-save >
AllFrenchYoruba30gen30prop.out
```

When the analysis is finished, the final inference will be printed in *AllFrenchYoruba30gen30prop.main.txt*. In this example, the admixture conclusion is "one date" of admixture at around 26 generations. The inferred sources are as follow;

```
###########################
### 1-DATE FIT SOURCES, PC1:
proportion BantuKenya Mandenka BantuSouthAfrica
0.29 0.26 0.31 0.43
proportion Han Japanese Balochi Druze Sardinian Ireland English
0.71 0.01 0.01 0.02 0.03 0.03 0.16 0.74
###########################
```

Out of the two sources contributing to the target, one contributes 29% of the total admixture proportion and is most genetically similar to the sampled BantuSouthAfrica. The other source contributes 71% and is most genetically similar to the sampled English. This inference is close to the true admixture we simulated in terms of the date, ancestral sources and admixture proportions. The *AllFrenchYoruba30gen30prop.main.pdf* file provides the plots of fitted "coancestry curves" for each surrogate pair inferred to describe >0.5% of the target population's haplotype patterns.

From these plots we can assess whether the inference proposed by the fastGLOBETROTTER model is supported by the data. Importantly, if a curve involving a pair of surrogates is increasing, this suggests that the given pair of surrogates

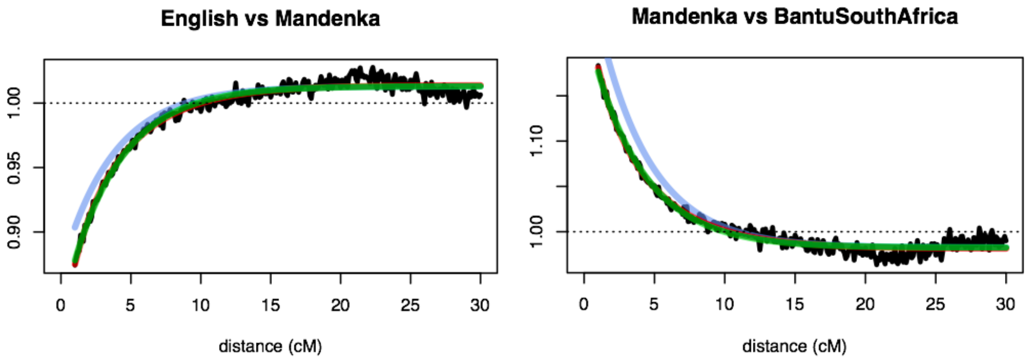

Figure 6.2: Coancestry curves showing side of ancestral sources. Left – increasing trend suggesting opposite ancestral source (English and Mandenka), Right – decreasing trend suggesting same ancestral source (Mandenka and BantuSouthAfrica).

represent different ancestral source (e.g. English and Mandenka in Figure 6.2–Left, with the former representing the “French” source and the latter the “Yoruba” source). Conversely, if the curve is decreasing, this suggests the surrogates are each representing the same ancestral source (e.g. Mandenka and BantuSouthAfrica in Figure 6.2–Right, with each representing the “Yoruba” source).

Chapter 7

Conclusion and Future Directions

7.1 Conclusion

In this thesis, I have developed a new haplotype-based method for fast and accurate inference of genetic admixture in populations using genotype data. The speedy calculation of this method allows population geneticists to analyse ever larger data to unearth previously unseen admixture events. Such approaches will have increasing utility as researchers collect large-scale datasets from precise geographic regions (e.g. Genomics England [79], China Kadoorie Biobank [80]) containing thousands of individuals with relatively homogeneous ancestral histories. I conclude each chapter in the following.

In Chapter 1, I described admixture events and how genetic variation patterns in descendants from such events can be used to infer features of the original admixture event, including information about the admixing sources, proportions and timings. I outlined different types of DNA that have been used to infer admixture, explaining the particular benefits of using densely typed genome-wide autosomal data. I also briefly summarized some of the most widely-used previous methods used to detect admixture, highlighting the strengths and limitations of each. Much of this summary was published in a review article: Wangkumhang & Hellenthal, 2019 [81].

In Chapter 2, I provided details of the intuition and theory behind the current state-of the-art method for inferring admixture events, GLOBETROTTER [21] and its accompanying method for local ancestry inference, CHROMOPAINTER [36]. I pointed out that – despite its increased accuracy over previous approaches (Hellenthal et al 2014) – the computational burden of the GLOBETROTTER inference, particularly for the coancestry curve building step, precludes it from being easily applicable to large-scale datasets with many hundreds of individuals.

In Chapter 3, I presented a new approach, fastGLOBETROTTER, that builds upon the GLOBETROTTER framework but is faster and more accurate. I described how fastGLOBETROTTER incorperates a new sampling method that upweights the contribution of DNA segment pairs that are more informative for admixture, as well as new code optimizatinon and donor merging steps to further speed up inference. These three techniques are important steps that provide a trade-off between accuracy, speed and memory-efficiency. I also developed and tested an additional technique to remove non-admixture related signals in the co-ancestry curve that can lead to misleading inference in target populations affected by relatively strong genetic drift (e.g. due to bottlenecks). Additionally, I provided an alternative technique, jack-knifing, to obtain confidence intervals for inferred admixture dates, which is of particular interest when making inference on single individuals (e.g. Chacon-Duque et al 2018) where bootstrap resampling is not suitable.

In Chapter 4, I validated the performance of fastGLOBETROTTER relative to GLOBETROTTER in terms of the accuracy of admixture dating and the computational efficiency. To do this, I carefully simulated multiple datasets including easy and hard cases to detect admixture. I found that the relative computational increase of fastGLOBETROTTER over GLOBETROTTER is 4-20 fold, while providing very similar inference accuracy in most cases. Interestingly, there are improvements in some cases, which conceivably may be due to noise reduction by only fitting the most informative sections of the co-ancestry curve. I also explored the ability of fastGLOBETROTTER to differentiate multiple pulses of admixture where the admixed group subsequently admixes again with one of the original sources. I concluded that fastGLOBETROTTER may detect only a single pulse of admixture in such settings, with an inferred date falling between the two pulses.

In Chapter 5, I applied fastGLOBETROTTER to a previously published dataset consisting of >6,000 European individuals sampled from Finland, Sweden, Norway, Denmark, Germany, Poland, France, Spain and Italy. I applied fineSTRUCTURE to cluster these individuals into groups of relative genetic homogeneity, with these clusters strongly correlating with country though showing multiple sub-groups within each country. Then using a set of world populations as a reference, I applied fastGLOBETROTTER to each European cluster, identifying admixture in all 86 clusters that suggested multiple admixture events occurring within the past two millennia. In particular fastGLOBETROTTER found evidence for admixture with sources that were genetically related – at least in part – to different non-European ancestry sources. This included admixture identified in clusters from Finland, Norway and Sweden dated to 180-1350 CE involving a source genetically related to present-day peoples of Siberia and East Asia. Another interesting signal was admixture dated to ≈550 CE in clusters from Belgium, France and Germany, involving a source related to peoples of W.Asia and N.Africa. East Asian-like ancestry previously has been identified in ancient DNA samples that are over 2.5kya found in areas near Finland [67,68] ; our results could be picking up intermixing involving this same source at later times. The signal of N.African and W.Asian ancestry in present-day peoples from Belgium, France and Germany is to my knowledge previously unknown. Its origin is unclear, though it is consistent with the movement of peoples across the Roman Empire, which at its height covered present-day Belgium, France, Turkey, and parts of Germany and North Africa. While our inferred admixture date is after the fall of the Roman Empire, intermixing between W.Asian and European-like sources may have occurred after the fall and/or we may be picking up signals of a previously admixed group (i.e. descendants of admixture prior to the fall of the Roman Empire) that intermixed later with a European-like source (see our Chapter 4 simulations with multiple pulses of admixture). This study provides further evidence that large-scale genetic data potentially can uncover previously hidden admixture events. Combining this genetic evidence with inference from other fields such as archaeology will further our understanding of human history.

In Chapter 6, I provide instructions and a tutorial for using fastGLOBETROTTER. As fastGLOBETROTTER has been developed as a distributed software freely available to academics, the documents are meant to assist other users in performing all steps of analysis, from phasing to chromosome painting, admixture detection and interpreting results. As part of this tutorial, I provide a step-by-step example from real data analysis.

7.2 Future perspectives

A future direction for fastGLOBETROTTER that may be fruitful is to make the methods more appealing and easy for users, given there are multiple steps involved in all analysis steps. To do so, one might consider automating the pipeline as described in the tutorials section. This might include reformatting input/output for SHAPEIT to CHROMOPAINTER and on to fastGLOBETROTTER, which currently requires a lot of different, specfic configurations. This future work would build up and strengthen the applicability of my software. Further speed-ups of GLOBETROTTER would be desirable, as SHAPEIT is relatively faster in completing the haplotype phasing step (e.g. offering multi-CPU computation) and CHROMOPAINTER can be accelerated by splitting individuals into subsets and making local inference separately on a high performance computational cluster. Parallelizing GLOBETROTTER is not as potentially possible, given each target individual is first analyzed separately before being combined again at each step of mixing proportion and date inference.

To further explore the genetic admixture signals in Europe unearthed here (Chapter 5), fastGLOBETROTTER can be applied to other datasets to test the hypothesis of a potential “Roman-like” genetic legacy (i.e. an ancestral signal related to present-day populations from South Central Europe, North African, and Armenian/Iran as found in this thesis). In particular testing other areas formerly under Roman control, such as Portugal, Luxamburg, Netherlands, Austria and Switzerland (e.g. using the POPRES dataset [82] and Novembre et al, 2008 [29]), and comparing these to other areas not under Roman control. For example, assuming there was a single admixture event between W.Asian/N.African-like and European-like sources in southern Europe, with this admixed group then moving north and mixing with other Europeans, we might expect to see a cascading effect of decreasing dates (in generations ago) and proportions of W.Asian/N.African-like ancestry the further away geographically from Italy. We see a similar observation in Figure 5.14 of decreasing E.Asian/Siberian-like ancestry from Finland to Norway, consistent with Siberian-like and European-like sources intermixing in the East and then spreading west in Scandinavia. Selection effects of these admixture events could also be an interesting area of future work, e.g. testing whether some parts of the genome of Finns retain unusually high proportions of East Asian-like ancestry, indicative of adaptive alleles being transmitted by East Asians to Europeans and retained over time.

Appendix A

Supplementary Data

Table A.1: The reference (donor) populations ("HB:" refers to **H**ellenthal and **B**usby dataset, "MS:" refers to **M**ultiple **S**clerosis dataset) and 86 fineSTRUCTURE clusters (c1-86) used in the analysis. Cluster name refers to a unique short name given to each cluster, the corresponding number of samples *N* and region are given in the next 2 columns, "*N* by country" shows the top 3 major populations in each cluster followed by its number of samples.

cluster	*N*	region	*N* by country
MS:UK	1854	NorthWestEurope	MS:UK 1854
MS:NIreland	61	NorthWestEurope	MS:NIreland 61
HB:yukagir	4	Siberia	HB:yukagir 4
HB:yoruba	21	SubAfrica	HB:yoruba 21
HB:yi	10	EastAsia	HB:yi 10
HB:yemeni	9	NearEast	HB:yemeni 9
HB:yakut	25	EastAsia	HB:yakut 25
HB:xibo	9	EastAsia	HB:xibo 9
HB:westsicilian	10	SouthCentralEurope	HB:westsicilian 10
HB:welsh	4	NorthWestEurope	HB:welsh 4
HB:velamas	9	SouthAsia	HB:velamas 9
HB:uzbekistani	15	CentralAsia	HB:uzbekistani 15
HB:uygur	10	CentralAsia	HB:uygur 10
HB:upcaste	5	SouthAsia	HB:upcaste 5
HB:ukrainian	20	NorthEastEurope	HB:ukrainian 20
HB:uae	14	NearEast	HB:uae 14
HB:tuva	13	Siberia	HB:tuva 13
HB:tuscan	8	SouthCentralEurope	HB:tuscan 8
HB:turkmen	10	WestCentralAsia	HB:turkmen 10
HB:turkishs	20	Turkey	HB:turkishs 20
HB:turkishn	20	Turkey	HB:turkishn 20
HB:turkishe	23	Turkey	HB:turkishe 23
HB:turkish	19	Turkey	HB:turkish 19

Table A.1 continued from previous page

cluster	*N*	region	*N* by country
HB:tunisian	12	NorthAfrica	HB:tunisian 12
HB:tujia	10	EastAsia	HB:tujia 10
HB:tu	10	EastAsia	HB:tu 10
HB:tsi	98	SouthCentralEurope	HB:tsi 98
HB:tharus	2	SouthAsia	HB:tharus 2
HB:tamilnadu	2	SouthAsia	HB:tamilnadu 2
HB:tajik	15	WestCentralAsia	HB:tajik 15
HB:syrian	16	NearEast	HB:syrian 16
HB:surui	5	Americas	HB:surui 5
HB:spanish	34	SouthWestEurope	HB:spanish 34
HB:southitalian	18	SouthCentralEurope	HB:southitalian 18
HB:sindhi	24	CentralAsia	HB:sindhi 24
HB:siciliane	10	SouthCentralEurope	HB:siciliane 10
HB:she	10	EastAsia	HB:she 10
HB:selkup	10	Siberia	HB:selkup 10
HB:scottish	6	NorthWestEurope	HB:scottish 6
HB:saudi	19	NearEast	HB:saudi 19
HB:sardinian	28	Sardinia	HB:sardinian 28
HB:sannamibia	5	San	HB:sannamibia 5
HB:sankhomani	30	San	HB:sankhomani 30
HB:sandawe	28	CentralAfrica	HB:sandawe 28
HB:sakd	4	SouthAsia	HB:sakd 4
HB:russian	25	NorthEastEurope	HB:russian 25
HB:romanian	16	SouthEastEurope	HB:romanian 16
HB:polish	17	NorthEastEurope	HB:polish 17
HB:piramalaikallar	8	SouthAsia	HB:piramalaikallar 8
HB:pima	14	Americas	HB:pima 14
HB:pathan	22	CentralAsia	HB:pathan 22
HB:papuan	17	Oceania	HB:papuan 17
HB:palestinian	46	NearEast	HB:palestinian 46
HB:oroqen	9	EastAsia	HB:oroqen 9
HB:orcadian	15	NorthWestEurope	HB:orcadian 15
HB:norwegian	18	NorthWestEurope	HB:norwegian 18
HB:northossetian	15	WestCaucasus	HB:northossetian 15
HB:northitalian	12	SouthCentralEurope	HB:northitalian 12
HB:nogay	16	WestCentralAsia	HB:nogay 16
HB:nihali	2	SouthAsia	HB:nihali 2
HB:nganassan	10	Siberia	HB:nganassan 10
HB:naxi	8	EastAsia	HB:naxi 8
HB:naga	4	EastAsia	HB:naga 4
HB:myanmar	3	SouthAsia	HB:myanmar 3
HB:muslim	5	SouthAsia	HB:muslim 5
HB:mozabite	29	NorthAfrica	HB:mozabite 29

Table A.1 continued from previous page

cluster	*N*	region	*N* by country
HB:moroccan	25	NorthAfrica	HB:moroccan 25
HB:mordovian	15	NorthEastEurope	HB:mordovian 15
HB:mongolian	19	EastAsia	HB:mongolian 19
HB:miao	10	EastAsia	HB:miao 10
HB:melanesian	10	Oceania	HB:melanesian 10
HB:meghawal	1	SouthAsia	HB:meghawal 1
HB:meena	1	SouthAsia	HB:meena 1
HB:mbutipygmy	13	CentralAfrica	HB:mbutipygmy 13
HB:maya	21	Americas	HB:maya 21
HB:mawasi	1	SouthAsia	HB:mawasi 1
HB:mandenka	22	SubAfrica	HB:mandenka 22
HB:malayan	1	SouthAsia	HB:malayan 1
HB:makrani	25	CentralAsia	HB:makrani 25
HB:maasai	97	SubAfrica	HB:maasai 97
HB:luhya	94	SubAfrica	HB:luhya 94
HB:lithuanian	10	NorthEastEurope	HB:lithuanian 10
HB:lezgin	18	EastCaucasus	HB:lezgin 18
HB:lebanese	5	NearEast	HB:lebanese 5
HB:lambadi	1	SouthAsia	HB:lambadi 1
HB:lahu	8	EastAsia	HB:lahu 8
HB:kyrgyz	16	CentralAsia	HB:kyrgyz 16
HB:kurumba	4	SouthAsia	HB:kurumba 4
HB:kurmi	1	SouthAsia	HB:kurmi 1
HB:kurd	6	Armenia/Iran	HB:kurd 6
HB:kumyk	14	EastCaucasus	HB:kumyk 14
HB:kshatriya	7	SouthAsia	HB:kshatriya 7
HB:koryake	5	Siberia	HB:koryake 5
HB:kol	16	SouthAsia	HB:kol 16
HB:ket	2	Siberia	HB:ket 2
HB:karnataka	8	SouthAsia	HB:karnataka 8
HB:karitiana	11	Americas	HB:karitiana 11
HB:kanjar	5	SouthAsia	HB:kanjar 5
HB:kalash	23	CentralAsia	HB:kalash 23
HB:jordanian	20	NearEast	HB:jordanian 20
HB:japanese	28	EastAsia	HB:japanese 28
HB:irish	7	NorthWestEurope	HB:irish 7
HB:iranian	20	Armenia/Iran	HB:iranian 20
HB:indianjew	8	CentralSouthAsia	HB:indianjew 8
HB:indian	1	CentralSouthAsia	HB:indian 1
HB:hungarian	19	NorthEastEurope	HB:hungarian 19
HB:hezhen	8	EastAsia	HB:hezhen 8
HB:hazara	22	CentralAsia	HB:hazara 22
HB:hannchina	10	EastAsia	HB:hannchina 10

Table A.1 continued from previous page

cluster	*N*	region	*N* by country
HB:han	34	EastAsia	HB:han 34
HB:hakkipikki	3	SouthAsia	HB:hakkipikki 3
HB:hadza	3	CentralAfrica	HB:hadza 3
HB:greek	20	SouthEastEurope	HB:greek 20
HB:gond	4	SouthAsia	HB:gond 4
HB:germanyaustria	4	NorthWestEurope	HB:germanyaustria 4
HB:german	30	NorthWestEurope	HB:german 30
HB:georgian	20	WestCaucasus	HB:georgian 20
HB:french	28	SouthWestEurope	HB:french 28
HB:finnish	2	NorthEastEurope	HB:finnish 2
HB:evenk	12	Siberia	HB:evenk 12
HB:ethiopiant	5	NorthAfrica	HB:ethiopiant 5
HB:ethiopiano	7	NorthAfrica	HB:ethiopiano 7
HB:ethiopianjew	11	NorthAfrica	HB:ethiopianjew 11
HB:ethiopiana	7	NorthAfrica	HB:ethiopiana 7
HB:english	8	NorthWestEurope	HB:english 8
HB:egyptian	12	NorthAfrica	HB:egyptian 12
HB:dusadh	7	SouthAsia	HB:dusadh 7
HB:druze	42	NearEast	HB:druze 42
HB:dolgan	7	Siberia	HB:dolgan 7
HB:dhurwa	1	SouthAsia	HB:dhurwa 1
HB:dharkar	8	SouthAsia	HB:dharkar 8
HB:daur	9	EastAsia	HB:daur 9
HB:dai	10	EastAsia	HB:dai 10
HB:cypriot	12	Cyprus	HB:cypriot 12
HB:croatian	19	NorthEastEurope	HB:croatian 19
HB:colombian	7	Americas	HB:colombian 7
HB:chuvash	17	Chuvash	HB:chuvash 17
HB:chukchi	5	Siberia	HB:chukchi 5
HB:chenchu	4	SouthAsia	HB:chenchu 4
HB:chechen	20	EastCaucasus	HB:chechen 20
HB:chamar	10	SouthAsia	HB:chamar 10
HB:cambodian	10	SouthAsia	HB:cambodian 10
HB:buryat	15	EastAsia	HB:buryat 15
HB:burya	2	Siberia	HB:burya 2
HB:burusho	25	CentralAsia	HB:burusho 25
HB:bulgarian	31	SouthEastEurope	HB:bulgarian 31
HB:brahui	25	CentralAsia	HB:brahui 25
HB:brahmin	11	SouthAsia	HB:brahmin 11
HB:biakapygmy	21	CentralAfrica	HB:biakapygmy 21
HB:bhunjia	1	SouthAsia	HB:bhunjia 1
HB:bengali	1	SouthAsia	HB:bengali 1
HB:belorussian	9	NorthEastEurope	HB:belorussian 9

Table A.1 continued from previous page

cluster	*N*	region	*N* by country
HB:bedouin	45	NearEast	HB:bedouin 45
HB:basque	24	Basque	HB:basque 24
HB:bantusouthafrica	8	Bantu	HB:bantusouthafrica 8
HB:bantukenya	11	Bantu	HB:bantukenya 11
HB:balochi	24	CentralAsia	HB:balochi 24
HB:balkar	19	WestCaucasus	HB:balkar 19
HB:armenian	35	Armenia/Iran	HB:armenian 35
HB:altai	13	Siberia	HB:altai 13
HB:adygei	17	WestCaucasus	HB:adygei 17
HB:abhkasian	20	WestCaucasus	HB:abhkasian 20
c1	39	SouthWestEurope	Spain 24 France 10 Germany 3
c2	96	SouthWestEurope	Spain 88 France 5 Belgium 2
c3	46	SouthWestEurope	Spain 28 France 17 Belgium 1
c4	77	SouthWestEurope	Spain 54 France 21 Italy 1
c5	9	SouthWestEurope	Spain 5 France 4
c6	24	SouthCentralEurope	Italy 23 Belgium 1
c7	96	SouthCentralEurope	Italy 84 France 6 Germany 4
c8	48	SouthCentralEurope	Italy 45 Germany 2 Belgium 1
c9	110	SouthCentralEurope	Italy 103 Belgium 3 Germany 2
c10	23	NorthWestEurope	Sweden 10 France 4 Germany 3
c11	27	NorthWestEurope	Sweden 13 Germany 9 Norway 1
c12	37	NorthWestEurope	Germany 15 Sweden 8 Italy 7
c13	18	NorthWestEurope	Sweden 8 Germany 3 Spain 2
c14	18	SouthCentralEurope	Italy 14 Belgium 1 Denmark 1
c15	30	SouthCentralEurope	Italy 30
c16	116	SouthCentralEurope	Italy 106 France 6 Germany 2
c17	26	SouthCentralEurope	Italy 26
c18	98	SouthCentralEurope	Italy 96 France 2
c19	41	SouthWestEurope	France 22 Italy 18 Belgium 1
c20	142	SouthCentralEurope	Italy 141 Germany 1
c21	47	SouthCentralEurope	Italy 46 France 1
c22	13	NorthWestEurope	Sweden 7 Germany 6
c23	64	NorthEastEurope	Finland 58 Sweden 6
c24	34	NorthEastEurope	Finland 30 Sweden 4
c25	58	NorthEastEurope	Finland 57 Sweden 1
c26	27	NorthEastEurope	Finland 25 Sweden 2
c27	45	NorthEastEurope	Finland 45
c28	32	NorthEastEurope	Finland 32
c29	89	NorthEastEurope	Finland 83 Sweden 6
c30	62	NorthEastEurope	Finland 59 Sweden 3
c31	63	NorthEastEurope	Finland 56 Sweden 7
c32	82	NorthEastEurope	Finland 80 Sweden 2
c33	30	NorthEastEurope	Finland 30

Table A.1 continued from previous page

cluster	*N*	region	*N* by country
c34	23	NorthWestEurope	Norway 8 Belgium 6 Germany 3
c35	77	SouthWestEurope	France 76 Denmark 1
c36	16	NorthWestEurope	Sweden 11 Germany 2 Norway 1
c37	84	NorthEastEurope	Poland 54 Germany 17 Sweden 8
c38	134	NorthWestEurope	Germany 128 Denmark 3 Poland 2
c39	146	NorthWestEurope	Germany 136 Sweden 5 France 3
c40	71	NorthWestEurope	Sweden 46 Norway 11 Denmark 8
c41	116	NorthWestEurope	Germany 113 Sweden 2 Belgium 1
c42	93	NorthWestEurope	Germany 82 Denmark 7 Sweden 4
c43	43	NorthWestEurope	Germany 43
c44	55	NorthWestEurope	Germany 52 Denmark 2 Belgium 1
c45	96	NorthWestEurope	Germany 92 Belgium 2 Denmark 1
c46	118	NorthWestEurope	Germany 111 Belgium 5 France 2
c47	122	NorthWestEurope	Germany 105 France 7 Belgium 4
c48	157	NorthWestEurope	Germany 151 Belgium 2 France 2
c49	162	NorthWestEurope	Denmark 153 Norway 7 Sweden 2
c50	155	NorthWestEurope	Denmark 140 Sweden 12 Norway 2
c51	87	SouthWestEurope	France 84 Norway 1 Spain 1
c52	206	SouthWestEurope	France 191 Belgium 7 Germany 5
c53	189	NorthWestEurope	Belgium 181 France 5 Germany 2
c54	204	NorthWestEurope	Belgium 203 Germany 1
c55	107	NorthWestEurope	Belgium 107
c56	32	NorthWestEurope	Sweden 23 Norway 9
c57	34	NorthEastEurope	Finland 22 Sweden 12
c58	18	NorthWestEurope	Sweden 15 Finland 2 Denmark 1
c59	44	NorthWestEurope	Sweden 44
c60	77	NorthWestEurope	Sweden 77
c61	73	NorthWestEurope	Sweden 71 Norway 2
c62	30	NorthWestEurope	Sweden 28 Finland 1 Norway 1
c63	90	NorthWestEurope	Sweden 88 Norway 2
c64	140	NorthWestEurope	Sweden 140
c65	212	NorthWestEurope	Sweden 211 Norway 1
c66	74	NorthWestEurope	Sweden 70 Denmark 4
c67	154	NorthWestEurope	Sweden 154
c68	78	NorthWestEurope	Sweden 77 Norway 1
c69	27	NorthWestEurope	Norway 25 Sweden 2
c70	16	NorthWestEurope	Norway 16
c71	21	NorthWestEurope	Norway 21
c72	80	NorthWestEurope	Norway 76 Sweden 4
c73	28	NorthWestEurope	Norway 28
c74	67	NorthWestEurope	Norway 67
c75	37	NorthWestEurope	Norway 37
c76	92	NorthWestEurope	Norway 91 Sweden 1

Table A.1 continued from previous page

cluster	*N*	region	*N* by country
c77	87	NorthWestEurope	Norway 78 Sweden 6 Denmark 3
c78	42	NorthWestEurope	Norway 41 Sweden 1
c79	27	NorthWestEurope	Norway 26 Sweden 1
c80	88	NorthWestEurope	Norway 86 Sweden 2
c81	79	NorthWestEurope	Norway 78 Sweden 1
c82	53	NorthWestEurope	Norway 53
c83	53	NorthWestEurope	Norway 52 Sweden 1
c84	58	NorthWestEurope	Norway 58
c85	32	NorthWestEurope	Norway 32
c86	36	NorthWestEurope	Norway 36

Table A.2: fastGLOBETROTTER inference on 86 Europeans clusters. The column "N" refers to number of samples in each cluster; "*N* by country" are the top 2 major populations in each cluster followed by its number of samples; "conclusion" – type of admixture event inferred from fastGLOBETROTTER; "date(95% CI)" – admixture date and its CI from 100 bootstrap resamplings; "prop1" – admixture proportion contributed from "source1" or surrogates from the first side.

cluster	*N*	*N* by country	conclusion	date (95% CI)	prop1	source1	source2
c1	39	Spain 24 France 10	1-date	18(15-19)	0.14	HB:welsh, HB:yoruba, HB:biakapygmy	HB:spanish,MS:UK,HB:scottish
c2	96	Spain 88 France 5	1-date-multiway	35(34-37)	0.21	HB:welsh, HB:mandenka, HB:maya	HB:norwegian,HB:welsh,HB:basque
c3	46	Spain 28 France 17	1-date-multiway	51(49-53)	0.49	HB:welsh, HB:scottish, HB:croatian	HB:basque,HB:french,HB:spanish
c4	77	Spain 54 France 21	1-date-multiway	39(37-40)	0.18	HB:welsh, HB:moroccan, HB:mandenka	HB:welsh,MS:NIreland,HB:basque
c5	9	Spain 5 France 4	1-date	30(19-40)	0.21	HB:welsh, HB:french, HB:tsi	HB:basque,HB:spanish,HB:welsh
c6	24	Italy 23 Belgium 1	1-date	37(34-40)	0.23	HB:greek, HB:welsh, HB:tuscan	HB:sardinian,HB:southitalian,HB:french
c7	96	Italy 84 France 6	1-date	34(33-34)	0.17	HB:southitalian, HB:mandenka, HB:tsi	HB:welsh,HB:northitalian,HB:croatian
c8	48	Italy 45 Germany 2	1-date-multiway	46(42-51)	0.19	HB:greek, HB:armenian, HB:maasai	HB:germanyaustria,HB:southitalian,HB:bulgarian
c9	110	Italy 103 Belgium 3	1-date	45(43-46)	0.17	HB:greek, HB:armenian, HB:mandenka	HB:french,HB:southitalian,HB:lithuanian
c10	23	Sweden 10 France 4	1-date-multiway	40(37-43)	0.2	HB:moroccan, HB:croatian, HB:tuscan	HB:german,HB:french,HB:welsh
c11	27	Sweden 13 Germany 9	1-date-multiway	47(44-50)	0.49	HB:greek, HB:croatian, HB:kurd	HB:lithuanian,HB:croatian,HB:russian
c12	37	Germany 15 Sweden 8	1-date-multiway	40(38-42)	0.39	HB:ukrainian, HB:croatian, HB:bulgarian	MS:UK,HB:scottish,HB:lithuanian
c13	18	Sweden 8 Germany 3	1-date-multiway	32(29-34)	0.32	HB:greek, HB:hezhen, HB:welsh	HB:romanian,HB:welsh,HB:tsi
c14	18	Italy 14 Belgium 1	1-date-multiway	45(39-51)	0.43	HB:welsh, HB:irish, HB:southitalian	HB:westsicilian,HB:sardinian,HB:welsh
c15	30	Italy 30	1-date-multiway	53(50-57)	0.46	HB:french, HB:welsh, HB:northitalian	HB:tsi,HB:armenian,HB:southitalian
c16	116	Italy 106 France 6	1-date-multiway	52(50-53)	0.17	HB:armenian, HB:tsi, HB:mandenka	HB:french,HB:english,HB:lithuanian
c17	26	Italy 26	1-date-multiway	48(41-55)	0.45	HB:northitalian, HB:kurd, HB:mozabite	HB:welsh,HB:tsi,MS:NIreland
c18	98	Italy 96 France 2	1-date-multiway	48(45-50)	0.43	HB:northitalian, HB:turkishe, HB:tsi	HB:english,HB:northitalian,HB:scottish
c19	41	France 22 Italy 18	1-date	54(50-57)	0.16	HB:moroccan, HB:turkish, HB:greek	HB:french,HB:english,HB:welsh
c20	142	Italy 141 Germany 1	1-date	54(52-57)	0.3	HB:northitalian, HB:tsi, HB:turkishe	HB:french,HB:welsh,HB:northitalian
c21	47	Italy 46 France 1	1-date	56(52-60)	0.27	HB:turkishe, HB:tsi, HB:northitalian	HB:french,HB:welsh,HB:northitalian
c22	13	Sweden 7 Germany 6	1-date-multiway	27(17-46)	0.41	HB:armenian, HB:georgian, HB:hezhen	HB:greek,HB:armenian,HB:cypriot
c23	64	Finland 58 Sweden 6	1-date	56(55-59)	0.22	HB:russian, HB:lithuanian, HB:koryake	HB:welsh,HB:german,HB:russian

Table A.2 continued from previous page

cluster	*N*	*N* by country	conclusion	date (95% CI)	prop1	source1	source2
c24	34	Finland 30 Sweden 4	1-date	52(50-55)	0.22	HB:russian, HB:lithuanian, HB:chuvash	HB:welsh,HB:german,HB:scottish
c25	58	Finland 57 Sweden 1	1-date	61(59-63)	0.21	HB:russian, HB:nganassan, HB:daur	HB:welsh,HB:norwegian,MS:NIreland
c26	27	Finland 25 Sweden 2	1-date	54(52-57)	0.19	HB:russian, HB:lithuanian, HB:nganassan	HB:welsh,HB:german,MS:UK
c27	45	Finland 45	1-date	64(62-66)	0.2	HB:russian, HB:oroqen, HB:chuvash	MS:UK,HB:german,HB:russian
c28	32	Finland 32	1-date	53(51-57)	0.2	HB:russian, HB:oroqen, HB:nganassan	HB:welsh,HB:german,MS:UK
c29	89	Finland 83 Sweden 6	multiple-dates(1st)	15(43788)	0.28	HB:norwegian,HB:belorussian,HB:russian	HB:german,HB:welsh,HB:russian
			multiple-dates(2nd)	95(83-102)	0.28	HB:russian,HB:dolgan,HB:nganassan	HB:german,HB:welsh,HB:russian
c30	62	Finland 59 Sweden 3	1-date	62(59-64)	0.23	HB:russian, HB:dolgan, HB:nganassan	HB:welsh,MS:NIreland,HB:norwegian
c31	63	Finland 56 Sweden 7	1-date-multiway	60(58-62)	0.27	HB:russian, HB:dolgan, HB:oroqen	MS:UK,HB:norwegian,HB:russian
c32	82	Finland 80 Sweden 2	1-date-multiway	63(61-66)	0.25	HB:russian, HB:oroqen, HB:nganassan	HB:welsh,HB:norwegian,MS:NIreland
c33	30	Finland 30	1-date	51(39-55)	0.31	HB:russian, HB:mordovian, HB:dolgan	MS:UK,HB:russian,HB:norwegian
c34	23	Norway 8 Belgium 6	1-date	27(24-29)	0.06	MS:NIreland, HB:tsi, HB:hadza	HB:welsh,MS:NIreland,HB:norwegian
c35	77	France 76 Denmark 1	1-date	52(47-57)	0.1	HB:southitalian, HB:spanish, HB:tsi	MS:UK,HB:french,HB:norwegian
c36	16	Sweden 11 Germany 2	1-date	37(34-40)	0.25	HB:daur, HB:lithuanian, HB:welsh	HB:polish,HB:welsh,HB:croatian
c37	84	Poland 54 Germany 17	1-date	41(38-45)	0.22	HB:welsh, MS:NIreland, HB:tsi	HB:belorussian,HB:croatian,HB:mordovian
c38	134	Germany 128 Denmark 3	1-date	25(23-26)	0.45	MS:UK, HB:german, HB:welsh	HB:polish,HB:belorussian,HB:lithuanian
c39	146	Germany 136 Sweden 5	1-date	28(27-29)	0.43	HB:polish, HB:belorussian, HB:romanian	HB:german,MS:UK,HB:french
c40	71	Sweden 46 Norway 11	1-date	32(29-35)	0.2	HB:ukrainian, HB:belorussian, HB:romanian	MS:UK,HB:german,HB:welsh
c41	116	Germany 113 Sweden 2	1-date	23(20-26)	0.31	HB:polish, HB:ukrainian, HB:lithuanian	HB:german,MS:UK,HB:welsh
c42	93	Germany 82 Denmark 7	1-date	26(24-28)	0.23	HB:polish, HB:ukrainian, HB:lithuanian	HB:german,MS:UK,HB:welsh
c43	43	Germany 43	1-date	44(41-47)	0.16	HB:polish, HB:ukrainian, HB:armenian	MS:UK,HB:german,HB:welsh
c44	55	Germany 52 Denmark 2	1-date	41(37-46)	0.12	HB:ukrainian, HB:lithuanian, HB:tsi	HB:welsh,MS:UK,HB:german
c45	96	Germany 92 Belgium 2	1-date	42(39-45)	0.13	HB:ukrainian, HB:cypriot, HB:romanian	MS:UK,HB:german,HB:welsh
c46	118	Germany 111 Belgium 5	1-date	49(47-51)	0.15	HB:romanian, HB:armenian, HB:moroccan	MS:UK,HB:german,HB:french
c47	122	Germany 105 France 7	1-date	48(47-51)	0.23	HB:turkish, HB:belorussian, HB:northitalian	HB:german,MS:UK,HB:french
c48	157	Germany 151 Belgium 2	1-date	42(39-43)	0.33	HB:polish, HB:hungarian, HB:croatian	MS:UK,HB:german,HB:english
c49	162	Denmark 153 Norway 7	1-date	41(35-45)	0.15	HB:polish, HB:tsi, HB:kurd	MS:UK,HB:welsh,HB:german

Table A.2 continued from previous page

cluster	*N*	*N* by country	conclusion	date (95% CI)	prop1	source1	source2
c50	155	Denmark 140 Sweden 12	1-date	37(35-39)	0.14	HB:polish, HB:finnish, HB:croatian	MS:UK,HB:welsh,HB:german
c51	87	France 84 Norway 1	1-date-multiway	57(55-60)	0.18	HB:spanish, HB:tsi, HB:southitalian	MS:UK,HB:northitalian,HB:welsh
c52	206	France 191 Belgium 7	1-date-multiway	50(46-55)	0.09	HB:moroccan, HB:tsi, HB:sardinian	HB:french,HB:welsh,HB:english
c53	189	Belgium 181 France 5	1-date	49(48-50)	0.13	HB:armenian, HB:northitalian, HB:greek	MS:UK,HB:german,HB:english
c54	204	Belgium 203 Germany 1	1-date	49(48-50)	0.15	HB:greek, HB:northitalian, HB:moroccan	MS:UK,HB:german,HB:welsh
c55	107	Belgium 107	1-date	49(47-51)	0.14	HB:northitalian, HB:greek, HB:cypriot	MS:UK,HB:german,HB:welsh
c56	32	Sweden 23 Norway 9	1-date	40(38-42)	0.18	HB:russian, HB:norwegian, HB:lithuanian	MS:UK,HB:norwegian,HB:welsh
c57	34	Finland 22 Sweden 12	1-date	43(41-45)	0.18	HB:russian, HB:lithuanian, HB:koryake	HB:welsh,MS:UK,MS:NIreland
c58	18	Sweden 15 Finland 2	1-date	58(54-64)	0.2	HB:russian, HB:lithuanian, HB:belorussian	MS:UK,HB:welsh,HB:german
c59	44	Sweden 44	1-date	30(29-32)	0.17	HB:norwegian, HB:belorussian, HB:russian	MS:UK,HB:welsh,HB:norwegian
c60	77	Sweden 77	1-date	38(36-39)	0.13	HB:belorussian, HB:norwegian, HB:russian	MS:UK,HB:german,HB:norwegian
c61	73	Sweden 71 Norway 2	1-date	42(40-44)	0.14	HB:norwegian, HB:belorussian, HB:koryake	MS:UK,HB:welsh,HB:norwegian
c62	30	Sweden 28 Finland 1	1-date	30(27-33)	0.21	HB:belorussian, HB:russian, HB:mordovian	MS:UK,HB:german,HB:welsh
c63	90	Sweden 88 Norway 2	1-date	40(37-43)	0.14	HB:belorussian, HB:norwegian, HB:russian	MS:UK,HB:german,HB:norwegian
c64	140	Sweden 140	multiple-dates(1st)	20(17-25)	0.1	HB:lithuanian,HB:russian,HB:belorussian	MS:NIreland,HB:welsh,HB:norwegian
			multiple-dates(2nd)	89(78-106)	0.13	HB:russian,HB:mordovian,HB:norwegian	HB:lithuanian,HB:norwegian,HB:welsh
c65	212	Sweden 211 Norway 1	multiple-dates(1st)	17(13-22)	0.37	HB:russian,HB:norwegian,HB:belorussian	HB:orcadian,HB:chukchi,HB:croatian
			multiple-dates(2nd)	73(68-84)	0.15	HB:croatian,HB:mordovian,HB:norwegian	HB:lithuanian,HB:norwegian,HB:welsh
c66	74	Sweden 70 Denmark 4	1-date	42(38-46)	0.13	HB:polish, HB:lithuanian, HB:ukrainian	MS:UK,HB:german,HB:welsh
c67	154	Sweden 154	1-date	43(35-48)	0.14	HB:belorussian, HB:ukrainian, HB:tsi	MS:UK,HB:german,HB:welsh
c68	78	Sweden 77 Norway 1	1-date	39(36-44)	0.12	HB:croatian, HB:lithuanian, HB:norwegian	MS:UK,HB:german,HB:norwegian
c69	27	Norway 25 Sweden 2	1-date	21(18-23)	0.1	HB:mordovian, HB:russian, HB:lithuanian	HB:norwegian,HB:welsh,MS:UK
c70	16	Norway 16	1-date	42(38-47)	0.15	HB:norwegian, HB:russian, HB:mordovian	MS:UK,HB:norwegian,HB:welsh
c71	21	Norway 21	1-date-multiway	38(35-41)	0.13	HB:norwegian, HB:russian, HB:nganassan	HB:welsh,HB:norwegian,MS:NIreland
c72	80	Norway 76 Sweden 4	1-date	39(37-41)	0.1	HB:norwegian, HB:russian, HB:nganassan	MS:UK,HB:norwegian,MS:NIreland
c73	28	Norway 28	1-date	36(33-38)	0.12	HB:norwegian, HB:russian, HB:oroqen	HB:welsh,HB:norwegian,MS:NIreland
c74	67	Norway 67	1-date-multiway	30(28-31)	0.12	HB:norwegian, HB:russian, HB:koryake	MS:UK,HB:norwegian,HB:welsh

Table A.2 continued from previous page

cluster	*N*	*N* by country	conclusion	date (95% CI)	prop1	source1	source2
c75	37	Norway 37	1-date	51(49-54)	0.13	HB:norwegian, HB:welsh, HB:lithuanian	HB:welsh,HB:norwegian,MS:NIreland
c76	92	Norway 91 Sweden 1	1-date	37(35-39)	0.14	HB:norwegian, HB:belorussian, HB:russian	MS:UK,HB:norwegian,HB:german
c77	87	Norway 78 Sweden 6	1-date-multiway	48(45-52)	0.2	HB:norwegian, HB:welsh, HB:polish	HB:welsh,MS:NIreland,HB:norwegian
c78	42	Norway 41 Sweden 1	1-date	40(35-44)	0.07	HB:belorussian, HB:greek, MS:NIreland	MS:UK,HB:german,MS:NIreland
c79	27	Norway 26 Sweden 1	1-date-multiway	49(43-54)	0.48	HB:norwegian, HB:hadza, HB:selkup	HB:welsh,HB:orcadian,HB:basque
c80	88	Norway 86 Sweden 2	1-date-multiway	37(33-40)	0.42	HB:welsh, MS:NIreland, HB:orcadian	HB:norwegian,HB:nganassan,
c81	79	Norway 78 Sweden 1	1-date-multiway	39(36-43)	0.12	HB:norwegian, HB:russian, HB:nganassan	HB:welsh,HB:norwegian,MS:NIreland
c82	53	Norway 53	1-date	50(43-59)	0.03	HB:kurumba, HB:naxi, HB:she	HB:norwegian,HB:welsh,MS:NIreland
c83	53	Norway 52 Sweden 1	1-date-multiway	37(32-43)	0.49	HB:norwegian, HB:russian, HB:nganassan	HB:welsh,MS:NIreland,HB:tsi
c84	58	Norway 58	1-date-multiway	55(52-59)	0.19	HB:norwegian, HB:koryake, HB:oroqen	MS:UK,HB:welsh,HB:norwegian
c85	32	Norway 32	1-date-multiway	32(23-41)	0.47	HB:welsh, MS:UK, MS:NIreland	HB:norwegian,HB:welsh,HB:german
c86	36	Norway 36	1-date-multiway	50(46-56)	0.26	HB:welsh, HB:norwegian, MS:NIreland	HB:ukrainian,MS:Nireland,HB:russian

Table A.3: fastGLOBETROTTER inference on 10 Greek clusters. The column “cluster” shows 10 Greek sub-clusters; “N” refers to number of samples in each cluster; “conclusion” – type of admixture event inferred from fastGLOBETROTTER; “date(95% CI)” – admixture date and its CI from 100 bootstrap resamplings; “prop1” – admixture proportion contributed from source1 or surrogates from the first side. “% within source1” % of contribution from each surrogate in source1.

cluster	*N*	conclusion	date(95% CI)	prop1	% within source1	% within source2
c1	57	1-date-multiway	25(17-32)	0.46	HB:armenian(50%) HB:syrian(15%) HB:cypriot(12%)	HB:romanian(63%) HB:northitalian(11%)
c2	182	1-date	36(33-38)	0.44	MS:Poland(34%) HB:croatian(19%) HB:northitalian(16%) HB:romanian(15%)	HB:lebanese(42%) HB:southitalian(13%) HB:armenian(10%)
c3	20	1-date	33(16-43)	0.4	MS:Poland(26%) HB:bulgarian(23%) HB:romanian(21%) HB:croatian(13%) HB:northitalian(12%)	HB:armenian(24%) HB:cypriot(17%) HB:southitalian(15%) HB:lebanese(10%)
c4	58	1-date-multiway	33(29-44)	0.41	HB:jordanian(21%) HB:armenian(20%) HB:cypriot(12%) HB:moroccan(10%)	HB:bulgarian(24%) HB:northitalian(17%) HB:romanian(15%)
c5	99	1-date	39(36-46)	0.49	HB:lebanese(47%) HB:southitalian(16%) HB:cypriot(11%)	MS:Poland(38%) HB:romanian(27%) HB:croatian(23%)
c6	134	1-date	37(33-40)	0.49	MS:Poland(34%) HB:croatian(24%) HB:romanian(22%) HB:northitalian(13%)	HB:lebanese(47%) HB:southitalian(10%) HB:cypriot(10%) HB:tsi(10%)
c7	31	1-date-multiway	25(18-32)	0.11	HB:southitalian(35%) HB:cypriot(33%) HB:tsi(11%)	HB:romanian(76%) HB:bulgarian(10%)
c8	12	1-date-multiway	23(19-29)	0.11	HB:ethiopiano(23%) HB:ethiopianjew(12%) HB:moroccan(11%) HB:makrani(10%)	HB:romanian(31%) HB:meghawal(21%) HB:armenian(12%)
c9	24	1-date	25(21-34)	0.48	MS:Poland(56%) HB:belorussian(36%)	HB:romanian(32%) HB:turkishe(13%) HB:southitalian(12%) HB:cypriot(10%)
c10	14	1-date	45(29-69)	0.36	HB:tsi(18%) HB:armenian(18%) HB:jordanian(16%)	MS:Germany(39%) HB:welsh(15%) MS:Belgium(11%)

Bibliography

[1] J. J. Hublin, A. Ben-Ncer, S. E. Bailey, S. E. Freidline, S. Neubauer, M. M. Skinner, I. Bergmann, A. Le Cabec, S. Benazzi, K. Harvati, and P. Gunz. Author Correction: New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, 558(7711):E6, 06 2018.

[2] D. Richter, R. Grun, R. Joannes-Boyau, T. E. Steele, F. Amani, M. Rue, P. Fernandes, J. P. Raynal, D. Geraads, A. Ben-Ncer, J. J. Hublin, and S. P. McPherron. The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature*, 546(7657):293–296, 06 2017.

[3] K. Harvati, C. Roding, A. M. Bosman, F. A. Karakostis, R. Grun, C. Stringer, P. Karkanas, N. C. Thompson, V. Koutoulidis, L. A. Moulopoulos, V. G. Gorgoulis, and M. Kouloukoussa. Apidima Cave fossils provide earliest evidence of Homo sapiens in Eurasia. *Nature*, 571(7766):500–504, 07 2019.

[4] S. J. Armitage, S. A. Jasim, A. E. Marks, A. G. Parker, V. I. Usik, and H. P. Uerpmann. The southern route ”out of Africa”: evidence for an early expansion of modern humans into Arabia. *Science*, 331(6016):453–456, Jan 2011.

[5] M. Balter. Was North Africa the launch pad for modern human migrations? *Science*, 331(6013):20–23, Jan 2011.

[6] C. J. Bae, K. Douka, and M. D. Petraglia. On the origin of modern humans: Asian perspectives. *Science*, 358(6368), 12 2017.

[7] T. Rito, D. Vieira, M. Silva, E. Conde-Sousa, L. Pereira, P. Mellars, M. B. Richards, and P. Soares. A dispersal of Homo sapiens from southern to eastern

Africa immediately preceded the out-of-Africa migration. *Sci Rep*, 9(1):4728, Mar 2019.

[8] A. Tigano and V. L. Friesen. Genomics of local adaptation with gene flow. *Mol. Ecol.*, 25(10):2144–2164, 05 2016.

[9] H. Ellegren and N. Galtier. Determinants of genetic diversity. *Nat. Rev. Genet.*, 17(7):422–433, 07 2016.

[10] I. Nasidze, D. Quinque, M. Ozturk, N. Bendukidze, and M. Stoneking. MtDNA and Y-chromosome variation in Kurdish groups. *Ann. Hum. Genet.*, 69(Pt 4):401–412, Jul 2005.

[11] M. A. Jobling and C. Tyler-Smith. The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.*, 4(8):598–612, Aug 2003.

[12] F. Balloux. The worm in the fruit of the mitochondrial DNA tree. *Heredity (Edinb)*, 104(5):419–420, May 2010.

[13] C. A. Winkler, G. W. Nelson, and M. W. Smith. Admixture mapping comes of age. *Annu Rev Genomics Hum Genet*, 11:65–89, 2010.

[14] I. Pugach, R. Matveyev, A. Wollstein, M. Kayser, and M. Stoneking. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol.*, 12(2):R19, 2011.

[15] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, Aug 2003.

[16] S. Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, Jun 2012.

[17] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, Nov 2012.

[18] J. E. Pool and R. Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, Feb 2009.

[19] P. R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, and B. Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–1254, Apr 2013.

[20] P. Moorjani, N. Patterson, J. N. Hirschhorn, A. Keinan, L. Hao, G. Atzmon, E. Burns, H. Ostrer, A. L. Price, and D. Reich. The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.*, 7(4):e1001373, Apr 2011.

[21] G. Hellenthal, G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, Feb 2014.

[22] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.*, 5(6):e1000519, Jun 2009.

[23] C. McHugh, L. Brown, and T. A. Thornton. Detecting Heterogeneity in Population Structure Across the Genome in Admixed Populations. *Genetics*, 204(1):43–56, Sep 2016.

[24] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, Nov 2008.

[25] O. Lao, T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, M. Balascakova, J. Bertranpetit, L. A. Bindoff, D. Comas, G. Holmlund, A. Kouvatsi, M. Macek, I. Mollet, W. Parson, J. Palo, R. Ploski, A. Sajantila, A. Tagliabracci, U. Gether, T. Werge, F. Rivadeneira, A. Hofman, A. G.

Uitterlinden, C. Gieger, H. E. Wichmann, A. Ruther, S. Schreiber, C. Becker, P. Nurnberg, M. R. Nelson, M. Krawczak, and M. Kayser. Correlation between genetic and geographic structure in Europe. *Curr. Biol.*, 18(16):1241–1248, Aug 2008.

[26] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet.*, 2(12):e190, Dec 2006.

[27] G. McVean. A genealogical interpretation of principal components analysis. *PLoS Genet.*, 5(10):e1000686, Oct 2009.

[28] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, Aug 2006.

[29] J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.*, 40(5):646–649, May 2008.

[30] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, Jun 2000.

[31] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19(9):1655–1664, Sep 2009.

[32] H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.*, 28(4):289–301, May 2005.

[33] A. Raj, M. Stephens, and J. K. Pritchard. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589, Jun 2014.

[34] P. Gopalan, W. Hao, D. M. Blei, and J. D. Storey. Scaling probabilistic models of genetic variation to millions of humans. *Nat. Genet.*, 48(12):1587–1590, 12 2016.

[35] J. P. Huelsenbeck, P. Andolfatto, and E. T. Huelsenbeck. Structurama: bayesian inference of population structure. *Evol. Bioinform. Online*, 7:55–59, 2011.

[36] D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genet.*, 8(1):e1002453, Jan 2012.

[37] L. van Dorp, D. Balding, S. Myers, L. Pagani, C. Tyler-Smith, E. Bekele, A. Tarekegn, M. G. Thomas, N. Bradman, and G. Hellenthal. 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.*, 11(8):e1005397, Aug 2015.

[38] D. Reich, K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. Reconstructing Indian population history. *Nature*, 461(7263):489–494, Sep 2009.

[39] M. Lipson, P. R. Loh, A. Levin, D. Reich, N. Patterson, and B. Berger. Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol. Biol. Evol.*, 30(8):1788–1802, Aug 2013.

[40] J. K. Pickrell and J. K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.*, 8(11):e1002967, 2012.

[41] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

[42] L. L. Cavalli-Sforza and A. W. F. Edwards. PHYLOGENETIC ANALYSIS: MODELS AND ESTIMATION PROCEDURES. *Evolution*, 21(3):550–570, Sep 1967.

[43] B. Pasaniuc, N. Zaitlen, G. Lettre, G. K. Chen, A. Tandon, W. H. Kao, I. Ruczinski, M. Fornage, D. S. Siscovick, X. Zhu, E. Larkin, L. A. Lange, L. A. Cupples, Q. Yang, E. L. Akylbekova, S. K. Musani, J. Divers, J. Mychaleckyj, M. Li, G. J. Papanicolaou, R. C. Millikan, C. B. Ambrosone, E. M.

John, L. Bernstein, W. Zheng, J. J. Hu, R. G. Ziegler, S. J. Nyante, E. V. Bandera, S. A. Ingles, M. F. Press, S. J. Chanock, S. L. Deming, J. L. Rodriguez-Gil, C. D. Palmer, S. Buxbaum, L. Ekunwe, J. N. Hirschhorn, B. E. Henderson, S. Myers, C. A. Haiman, D. Reich, N. Patterson, J. G. Wilson, and A. L. Price. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.*, 7(4):e1001371, Apr 2011.

[44] E. R. Chimusa, M. Daya, M. Moller, R. Ramesar, B. M. Henn, P. D. van Helden, N. J. Mulder, and E. G. Hoal. Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PLoS ONE*, 8(9):e73971, 2013.

[45] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*, 93(2):278–288, Aug 2013.

[46] W. Jin, R. Li, Y. Zhou, and S. Xu. Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *Eur. J. Hum. Genet.*, 22(7):930–937, Jul 2014.

[47] C. Churchhouse and J. Marchini. Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet. Epidemiol.*, 37(1):1–12, Jan 2013.

[48] R. Kaeuffer, D. Reale, D. W. Coltman, and D. Pontier. Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity (Edinb)*, 99(4):374–380, Oct 2007.

[49] M. Salter-Townshend and S. Myers. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, 212(3):869–889, Jul 2019.

[50] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, 78(4):629–644, Apr 2006.

[51] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81(5):1084–1097, Nov 2007.

[52] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5(6):e1000529, Jun 2009.

[53] O. Delaneau, B. Howie, A. J. Cox, J. F. Zagury, and J. Marchini. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.*, 93(4):687–696, Oct 2013.

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

[55] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[56] F.M.T.A. Busing, E. Meijer, and R.V.D. Leeden. Delete-m Jackknife for Unequal m. *Statistics and Computing.*, pages 3–8, Apr 1999.

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

[58] S. Sawcer, G. Hellenthal, M. Pirinen, C. C. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, S. Edkins, E. Gray,

D. R. Booth, S. C. Potter, A. Goris, G. Band, A. B. Oturai, A. Strange, J. Saarela, C. Bellenguez, B. Fontaine, M. Gillman, B. Hemmer, R. Gwilliam, F. Zipp, A. Jayakumar, R. Martin, S. Leslie, S. Hawkins, E. Giannoulatou, S. D'alfonso, H. Blackburn, F. Martinelli Boneschi, J. Liddle, H. F. Harbo, M. L. Perez, A. Spurkland, M. J. Waller, M. P. Mycko, M. Ricketts, M. Comabella, N. Hammond, I. Kockum, O. T. McCann, M. Ban, P. Whittaker, A. Kemppinen, P. Weston, C. Hawkins, S. Widaa, J. Zajicek, S. Dronov, N. Robertson, S. J. Bumpstead, L. F. Barcellos, R. Ravindrarajah, R. Abraham, L. Alfredsson, K. Ardlie, C. Aubin, A. Baker, K. Baker, S. E. Baranzini, L. Bergamaschi, R. Bergamaschi, A. Bernstein, A. Berthele, M. Boggild, J. P. Bradfield, D. Brassat, S. A. Broadley, D. Buck, H. Butzkueven, R. Capra, W. M. Carroll, P. Cavalla, E. G. Celius, S. Cepok, R. Chiavacci, F. Clerget-Darpoux, K. Clysters, G. Comi, M. Cossburn, I. Cournu-Rebeix, M. B. Cox, W. Cozen, B. A. Cree, A. H. Cross, D. Cusi, M. J. Daly, E. Davis, P. I. de Bakker, M. Debouverie, M. B. D'hooghe, K. Dixon, R. Dobosi, B. Dubois, D. Ellinghaus, I. Elovaara, F. Esposito, C. Fontenille, S. Foote, A. Franke, D. Galimberti, A. Ghezzi, J. Glessner, R. Gomez, O. Gout, C. Graham, S. F. Grant, F. R. Guerini, H. Hakonarson, P. Hall, A. Hamsten, H. P. Hartung, R. N. Heard, S. Heath, J. Hobart, M. Hoshi, C. Infante-Duarte, G. Ingram, W. Ingram, T. Islam, M. Jagodic, M. Kabesch, A. G. Kermode, T. J. Kilpatrick, C. Kim, N. Klopp, K. Koivisto, M. Larsson, M. Lathrop, J. S. Lechner-Scott, M. A. Leone, V. Leppa, U. Liljedahl, I. L. Bomfim, R. R. Lincoln, J. Link, J. Liu, A. R. Lorentzen, S. Lupoli, F. Macciardi, T. Mack, M. Marriott, V. Martinelli, D. Mason, J. L. McCauley, F. Mentch, I. L. Mero, T. Mihalova, X. Montalban, J. Mottershead, K. M. Myhr, P. Naldi, W. Ollier, A. Page, A. Palotie, J. Pelletier, L. Piccio, T. Pickersgill, F. Piehl, S. Pobywajlo, H. L. Quach, P. P. Ramsay, M. Reunanen, R. Reynolds, J. D. Rioux, M. Rodegher, S. Roesner, J. P. Rubio, I. M. Ruckert, M. Salvetti, E. Salvi, A. Santaniello, C. A. Schaefer, S. Schreiber, C. Schulze, R. J. Scott, F. Sellebjerg, K. W. Selmaj, D. Sexton, L. Shen, B. Simms-Acuna, S. Skidmore, P. M. Sleiman, C. Smestad, P. S. S?rensen, H. B. S?ndergaard, J. Stankovich, R. C. Strange, A. M. Sulonen, E. Sundqvist, A. C. Syvanen, F. Taddeo, B. Taylor, J. M. Blackwell, P. Tienari, E. Bramon, A. Tourbah, M. A. Brown, E. Tronczynska, J. P. Casas, N. Tubridy, A. Corvin, J. Vickery, J. Jankowski, P. Villoslada, H. S. Markus, K. Wang, C. G. Mathew, J. Wason, C. N. Palmer, H. E. Wichmann, R. Plomin, E. Willoughby, A. Rautanen, J. Winkelmann, M. Wittig, R. C. Trembath, J. Yaouanq, A. C. Viswanathan, H. Zhang, N. W. Wood, R. Zuvich, P. Deloukas, C. Langford, A. Duncanson, J. R. Oksenberg, M. A. Pericak-Vance, J. L. Haines, T. Olsson, J. Hillert, A. J. Ivinson, P. L. De Jager, L. Peltonen, G. J. Stewart, D. A. Hafler, S. L. Hauser, G. McVean, P. Donnelly, and A. Compston. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, Aug 2011.

[59] G. K. Chen, P. Marjoram, and J. D. Wall. Fast and flexible simulation of DNA sequence data. *Genome Res.*, 19(1):136–142, Jan 2009.

[60] P. Mellars. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature*, 439(7079):931–935, Feb 2006.

[61] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prufer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J. M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. Busby, F. Cali, M. Churnosov, D. E. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J. M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Ku?inskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Nakkalajarvi,

D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Paabo, J. Kelso, D. Reich, and J. Krause. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, Sep 2014.

[62] S. Beleza, A. M. Santos, B. McEvoy, I. Alves, C. Martinho, E. Cameron, M. D. Shriver, E. J. Parra, and J. Rocha. The timing of pigmentation lightening in Europeans. *Mol. Biol. Evol.*, 30(1):24–35, Jan 2013.

[63] E. R. Jones, G. Gonzalez-Fortes, S. Connell, V. Siska, A. Eriksson, R. Martiniano, R. L. McLaughlin, M. Gallego Llorente, L. M. Cassidy, C. Gamba, T. Meshveliani, O. Bar-Yosef, W. Muller, A. Belfer-Cohen, Z. Matskevich, N. Jakeli, T. F. G. Higham, M. Currat, D. Lordkipanidze, M. Hofreiter, A. Manica, R. Pinhasi, and D. G. Bradley. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat Commun*, 6:8912, Nov 2015.

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

[65] G. B. Busby, G. Hellenthal, F. Montinaro, S. Tofanelli, K. Bulayeva, I. Rudan, T. Zemunik, C. Hayward, D. Toncheva, S. Karachanak-Yankova, D. Nesheva, P. Anagnostou, F. Cali, F. Brisighelli, V. Romano, G. Lefranc, C. Buresi, J. Ben Chibani, A. Haj-Khelil, S. Denden, R. Ploski, P. Krajewski, T. Hervig, T. Moen, R. J. Herrera, J. F. Wilson, S. Myers, and C. Capelli. The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Curr. Biol.*, 25(19):2518–2526, 10 2015.

[66] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, Sep 2007.

[67] L. Saag, M. Laneman, L. Varul, M. Malve, H. Valk, M. A. Razzak, I. G. Shirobokov, V. I. Khartanovich, E. R. Mikhaylova, A. Kushniarevich, C. L. Scheib, A. Solnik, T. Reisberg, J. Parik, L. Saag, E. Metspalu, S. Rootsi, F. Montinaro, M. Remm, R. Magi, E. D'Atanasio, E. R. Crema, D. Diez-Del-Molino, M. G. Thomas, A. Kriiska, T. Kivisild, R. Villems, V. Lang, M. Metspalu, and K. Tambets. The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr. Biol.*, 29(10):1701–1711, May 2019.

[68] T. C. Lamnidis, K. Majander, C. Jeong, E. Salmela, A. Wessman, V. Moiseyev, V. Khartanovich, O. Balanovsky, M. Ongyerth, A. Weihmann, A. Sajantila, J. Kelso, S. Paabo, P. Onkamo, W. Haak, J. Krause, and S. Schiffels. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat Commun*, 9(1):5018, 11 2018.

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

[70] J. Van den Eynden, T. Descamps, E. Delporte, N. H. C. Roosens, S. C. J. De Keersmaecker, V. De Wit, J. R. Vermeesch, E. Goetghebeur, J. Tafforeau, S. Demarest, M. Van den Bulcke, and H. Van Oyen. The genetic structure of the Belgian population. *Hum. Genomics*, 12(1):6, 02 2018.

[71] A. Raveane, S. Aneli, F. Montinaro, G. Athanasiadis, S. Barlera, G. Birolo, G. Boncoraglio, A. M. Di Blasio, C. Di Gaetano, L. Pagani, S. Parolo, P. Paschou, A. Piazza, G. Stamatoyannopoulos, A. Angius, N. Brucato, F. Cucca, G. Hellenthal, A. Mulas, M. Peyret-Guzzon, M. Zoledziewska, A. Baali, C. Bycroft, M. Cherkaoui, J. Chiaroni, J. Di Cristofaro, C. Dina, J. M. Dugoujon, P. Galan, J. Giemza, T. Kivisild, S. Mazieres, M. Melhaoui, M. Metspalu, S. Myers, L. Pereira, F. X. Ricaut, F. Brisighelli, I. Cardinali, V. Grugni, H. Lancioni, V. L. Pascali, A. Torroni, O. Semino, G. Matullo, A. Achilli, A. Olivieri, and C. Capelli. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Sci Adv*, 5(9):eaaw3492, Sep 2019.

[72] B. O. Bengtsson and M. P. Nilsson. Strange history: the fall of Rome explained in Hereditas. *Hereditas*, 151(6):132–139, Dec 2014.

[73] I. Ntalla, M. Giannakopoulou, P. Vlachou, K. Giannitsopoulou, V. Gkesou, C. Makridi, M. Marougka, G. Mikou, K. Ntaoutidou, E. Prountzou, A. Tsekoura, and G. V. Dedoussis. Body composition and eating behaviours in relation to dieting involvement in a sample of urban Greek adolescents from the TEENAGE (TEENs of Attica: Genes and Environment) study. *Public Health Nutr*, 17(3):561–568, Mar 2014.

[74] K. Panoutsopoulou, K. Hatzikotoulas, D. K. Xifara, V. Colonna, A. E. Farmaki, G. R. Ritchie, L. Southam, A. Gilly, I. Tachmazidou, S. Fatumo, A. Matchan, N. W. Rayner, I. Ntalla, M. Mezzavilla, Y. Chen, C. Kiagiadaki, E. Zengini, V. Mamakou, A. Athanasiadis, M. Giannakopoulou, V. E. Kariakli, R. N. Nsubuga, A. Karabarinde, M. Sandhu, G. McVean, C. Tyler-Smith, E. Tsafantakis, M. Karaleftheri, Y. Xue, G. Dedoussis, and E. Zeggini. Ge-

netic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat Commun*, 5:5345, Nov 2014.

[75] O. Lao, T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, M. Balascakova, J. Bertranpetit, L. A. Bindoff, D. Comas, G. Holmlund, A. Kouvatsi, M. Macek, I. Mollet, W. Parson, J. Palo, R. Ploski, A. Sajantila, A. Tagliabracci, U. Gether, T. Werge, F. Rivadeneira, A. Hofman, A. G. Uitterlinden, C. Gieger, H. E. Wichmann, A. Ruther, S. Schreiber, C. Becker, P. Nurnberg, M. R. Nelson, M. Krawczak, and M. Kayser. Correlation between genetic and geographic structure in Europe. *Curr. Biol.*, 18(16):1241–1248, Aug 2008.

[76] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, Nov 2008.

[77] L. Kovacevic, K. Tambets, A. M. Ilumae, A. Kushniarevich, B. Yunusbayev, A. Solnik, T. Bego, D. Primorac, V. Skaro, A. Leskovac, Z. Jakovski, K. Drobnic, H. V. Tolk, S. Kovacevic, P. Rudan, E. Metspalu, and D. Marjanovic. Standing at the gateway to Europe–the genetic structure of Western balkan populations based on autosomal and haploid markers. *PLoS ONE*, 9(8):e105090, 2014.

[78] I. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, N. Rohland, S. Pfrengle, A. Furtwangler, A. Peltzer, C. Posth, A. Vasilakis, P. J. P. McGeorge, E. Konsolaki-Yannopoulou, G. Korres, H. Martlew, M. Michalodimitrakis, M. Ozsait, N. Ozsait, A. Papathanasiou, M. Richards, S. A. Roodenberg, Y. Tzedakis, R. Arnott, D. M. Fernandes, J. R. Hughey, D. M. Lotakis, P. A. Navas, Y. Maniatis, J. A. Stamatoyannopoulos, K. Stewardson, P. Stockhammer, R. Pinhasi, D. Reich, J. Krause, and G. Stamatoyannopoulos. Genetic origins of the Minoans and Mycenaeans. *Nature*, 548(7666):214–218, 08 2017.

[79] Mark Caulfield, Jim Davies, Martin Dennys, Leila Elbahy, Tom Fowler, Sue Hill, Tim Hubbard, Luke Jostins, Nick Maltby, Jeanna Mahon-Pearson, and et al. The National Genomics Research and Healthcare Knowledgebase, Dec 2017.

[80] Z. Chen, J. Chen, R. Collins, Y. Guo, R. Peto, F. Wu, L. Li, L. Li, Z. Chen, J. Chen, R. Collins, F. Wu, R. Peto, Z. Chen, G. Lancaster, X. Yang, A. Williams, L. Yang, Y. Chang, Y. Guo, G. Zhao, Z. Bian, L. Wu, C. Hou, Z. Pang, S. Wang, Y. Zhang, K. Zhang, S. Liu, Z. Zhao, S. Liu, Z. Pang, W. Feng, S. Wu, L. Yang, H. Han, H. He, X. Pan, S. Wang, H. Wang, X. Hao, C. Chen, S. Lin, X. Hu, M. Zhou, M. Wu, Y. Wang, Y. Hu, L. Ma, R. Zhou, G. Xu, B. Dong, N. Chen, Y. Huang, M. Li, J. Meng, Z. Gan, J. Xu, Y. Liu, X. Wu, Y. Gao, N. Zhang, G. Luo, X. Que, X. Chen, P. Ge, J. He, X. Ren, H. Zhang, E. Mao, G. Li, Z. Li, J. He, G. Liu, B. Zhu, G. Zhou, S. Feng, Y. Gao, T. He, L. Jiang, J. Qin, H. Sun, L. Liu, M. Yu, Y. Chen, Z. Hu, J. Hu, Y. Qian, Z. Wu, L. Chen, W. Liu, G. Li, H. Liu, X. Long, Y. Xiong, Z. Tan, X. Xie, and Y. Peng. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol*, 40(6):1652–1666, Dec 2011.

[81] P. Wangkumhang and G. Hellenthal. Statistical methods for detecting admixture. *Curr. Opin. Genet. Dev.*, 53:121–127, 12 2018.

[82] M. R. Nelson, K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, P. Vollenweider, J. R. Oksenberg, S. L. Hauser, H. A. Stirnadel, J. S. Kooner, J. C. Chambers, B. Jones, V. Mooser, C. D. Bustamante, A. D. Roses, D. K. Burns, M. G. Ehm, and E. H. Lai. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.*, 83(3):347–358, Sep 2008.