The Peopling of Europe

A Genetic Perspective

A Thesis Submitted for the Degree of *Doctor of Philosophy*

George B.J. Busby

Somerville College and the Department of Zoology

University of Oxford

Trinity 2012

'Man still bears in his bodily frame the indelible stamp of his lowly origin'

Charles Darwin, *The Descent of Man*, **1871**

To those who donated and admixed my chromosomes

Declaration

This thesis comprises one published paper as a chapter and two chapters that contribute towards manuscripts that are in preparation. All work presented here is my own, and where I have benefited from other people's research, I have declared this in the text. In addition, on page 141, I have listed all individuals that have helped, either by contributing DNAs, or through technical help in the laboratory or with analytical software.

In accordance with the Regulations and with permission from the Director of Graduate Studies in the Department of Zoology at the University of Oxford I present this thesis as a series of self-contained chapters in the style of scientific journal articles. Chapter 2 and parts of chapter 3 are based on the following paper:

Busby, G.B.J., et al (2012) The Peopling of Europe and the Cautionary Tale of Y Chromosome Haplogroup R-M269 *Proc. Roy. Soc. B.* 279 **1730** (884-892)

Abstract

The Peopling of Europe: a genetic perspective

George B.J. Busby

Somerville College and the Department of Zoology, University of Oxford

A thesis submitted for the degree of *Doctor of Philosophy*, Trinity 2012

Following their dispersal out of Africa, humans colonised all continents of the world save one, Antarctica. Whilst Europe was initially peopled soon after this exodus, paleoclimatic, archaeological, and historical evidence suggest that successive waves and migrations of people have contributed to the population resident in Europe today. I therefore examined the impact of past events on the European population through the analysis of DNA sampled both from contemporary Europeans, and from worldwide populations pertinent to its history. I genotyped and analysed data from the Y chromosomes of over 2,000 haplogroup R-M269 European men from over 30 different populations and, in combination with comparable datasets gathered from the literature, show that there it is not possible to assign a date to the origin of this lineage in Europe, and thus that any conclusion as to the ancient or recent spread of this lineage in Europe is unfounded. I also show that commonly used Y chromosome lineage dating techniques based on STR variation are biased by the markers used and conclusions based on such dates should be viewed with a large amount of caution. I next use genome-wide SNP data from 1,550 individuals from 95 worldwide populations to explore the population structure of Europe and present an analysis of the detailed structure of Europe in a novel analytical framework using *ChromoPainter* and *fineSTRUCTURE*. Admixture analysis based this data reveals distinct genomic inputs to peripheral European populations, from North Africa, Sub-Saharan Africa, the Middle East, and East Asia, and provides dates for this admixture within the last 1,000 years that correspond to the emergence and decline of empires and kingdoms in these regions of Europe. This novel analysis highlights the importance of recent historical events on European population structure, but also suggests a degree of ancient structure across European populations. Taken together, these analyses demonstrate the substantial effects of both ancient and recent migrations and mixture on the contemporary genetic structure of Europe.

Contents

List of Figures

List of Tables

List of Abbreviations

g	the genetic distance between admixture chunks
α	the proportion of admixture
ϑ	the linearity of an STR; the per site mutation rate parameter in *ChromoPainter*
λ	the rate parameter of an exponential distribution
μ	mutation rate
ASD	Average Squared Distance
BATWING	Bayesian Analysis of Trees With Internal Node Generation
BCE	Before Common Era
CE	Common Era
cM	centimorgan
D`	Duration of linearity
DNA	Dioxyribose Nucleic Acid
G-SMM	Generalised Stepwise Mutational Model
GWAS	Genome Wide Association Study
HGDP	Human Genome Diversity Panel
IBS	Identity By State

kb	kilobase
km	kilometre
kya	thousand years ago
LBK	*Linearbandkeramik*
LD	Linkage Disequilibrium
LGM	Last Glacial Maximum
Mb	Megabase
MCMC	Markov Chain Monte Carlo
mtDNA	mitochondrial DNA
Mya	Million years ago
nnls	non-negative-least-squares regression
PC	Principle Component
PCA	Principle Component Analysis
POPRES	Population Resource Database
PWC	Pitted Ware Culture
R	range of alleles that an STR can take
RFLP	Restriction Fragment Length Polymorphism
S-SMM	Simple Stepwise Mutational Model
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat or Microsatellite
T	Time Since the Most Recent Common Ancestor
TMRCA	Time Since the Most Recent Common Ancestor
TRB	*Trichterbecher Kultur* or Funnel Beaker Culture
YCC	Y Chromosome Consortium
YHRD	Y-chromosome Haplotype Research Database

1. Introduction

Europe has a complex and diverse evolutionary history. The first modern humans are believed to have entered Europe some 50,000 years ago. During the following millennia, the European continent underwent a series of environmental and climatic fluctuations, which caused much of the flora and fauna of the continent to move and migrate in order to maintain viable populations in the habitats to which they had evolved. Humans were no different. These factors, together with successive waves of new peoples from the Middle East, and movements of people within Europe during historical time, have combined to produce the genetic variation present in Europe today.

This thesis uses two complementary genetic lines of enquiry to investigate the present population structure of Europe, each of which can be used to consider human history within a different timeframe. In chapter 2, through a large analysis of European Y chromosomes, I attempt to understand where and when the most frequent lineage in Europe, R1b1b2-M269, entered this continent, and critically assess current hypotheses of its origins. In so doing, I highlight the fundamental importance that dating genetic lineages has to our ability to discriminate competing ideas of when they may have emerged. Short tandem repeats (STRs) have been the Y chromosome marker of choice for looking at diversity and calibrating this variation to time. Dating using STRs is, however, far from satisfactory and a number of underlying assumptions are rarely taken into account in studies where confident dating estimates are made. Given the utility of STRs in Y chromosome studies and the certainty with which STR dating results are often presented, it seems appropriate to take a look at these underlying as-

sumptions and to investigate whether they are indeed sound. Chapter 3 explores the issues surrounding Y chromosome dating using STRs.

Whilst the analysis of uniparental markers has become the norm in studies of human migration, chiefly due to the ease at which they can be typed and the large amount of data which is already available for comparison, there are limitations to the conclusions one can draw from analyses based on a single locus. The financial and technical barriers that have previously restricted evolutionary studies to the analysis of a small number of markers in a large number of individuals have recently begun to be broken down, primarily by the advent of next generation sequencing and the reduced cost of mass genotyping platforms. The second theme of this thesis then, is to use this large-scale genotyping data to address specific historical questions relating to ancient population movements. As the costs of producing these types of data continue to decrease, the ability to type many individuals at high resolution cheaply becomes a real possibility for biologists interested in historical evolutionary questions, and it is likely that in time these huge datasets will become the norm in evolutionary studies. The future will almost certainly involve analysing multiple markers across the entire genome - if not the analysis of complete genomes - and it is my aim to use novel methods here to attempt this in a European context. In chapter 4 and chapter 5 I explore genome-wide population structure and admixture in European populations and in so doing assesses how and when past human migrations may have affected the current people of Europe.

This introduction will outline the background, both historical and genetic, to these themes, starting with the initial peopling of Europe, covering what we know from archaeological and historical sources about the migrations that have influenced this continental population. I also cover the evolution of human evolutionary genetics, from the first original studies on sparse datasets, to the current 21st century research using genomic technologies. First, however, we must start at the beginning, with the archaeological evidence for the very first Europeans.

1.1. A Non Genetic Perspective

The Colonisation of Europe

The current earliest fossil of a hominin in Europe is purported to be a mandible found in the Sierra de Atapuerca in Spain dating to the Early Pleistocene, at least some 1.1 million years ago (Mya) (Carbonell et al., 2008; Bermúdez de Castro et al., 2011). It is likely that the continent has been inhabited ever since with successive species of *Homo*. Given that the very first evidence of human lithic technologies occurred in Ethiopia, approximately 2.5Mya (Semaw et al., 1997), it is likely that these early European hominins arrived in Western Europe as a result of dispersal out of Africa (van der Made and Mateos, 2010), possibly via Western Asia (Dennell and Roebroeks, 2005). It appears then that migrations have occurred from the very beginning of hominid history. These expansions may have been linked to environmental change or to the movement of other mammals (van der Made and Mateos, 2010), or to both. Alternatively, cultural and technological advances may have allowed early hominins to disperse independently of faunal or climatic change (van der Made and Mateos, 2010, and references therein). Much of what we know about the distributions of pre-Holocene hominins is gleaned from the few fossils we have from around the world which are open as much to discussion about their interpretation, as to theories about the processes behind the patterns of these distributions. However, what is clear is that, as a genus, *Homo* has always been on the move.

If we fast-forward one million years - a time during which several species of protohumans, such as *H. erectus*, *H. heidelbergensis*, and *H. neanderthalensis*, either expanded from Africa, or perhaps evolved outside of Africa only to later die out, the current understanding is that anatomically modern humans (*H. sapiens*) evolved in an area of East Africa, and from there dispersed once more. The timing and mode of this dispersal is hotly debated, but it is likely to have occurred at some point after 100 thousand years ago (kya), perhaps as recently as 60kya (Mellars and Cunliffe, 1994; Mellars, 2006a),

either through the Levant or across the Arabian Gulf into Arabia and round into Asia (Hoffecker, 2009). Given that anatomically modern humans are observed in the fossil record in south-east Asia well before they are in Europe, it is likely that the spread into Europe occurred after expansions into, although not necessarily via, Asia. There is consensus that the European continent was exclusively inhabited by *H. neanderthalensis*, Neandertals, 60kya but that by 30kya anatomically modern humans were abundant (Hoffecker, 2009). Two competing models of this transition persist, and are of relevance here in passing, as they resemble population changes that have occurred on numerous occasions since, and highlight fundamental differences between methods of population change. The first is that modern humans replaced Neandertals, with Neandertals having no significant exchange of genes and culture with the colonising species. The alternative is that there was some, perhaps substantial, genetic and cultural exchange between the two species and that Neandertals were assimilated into the modern human population (Trinkaus, 2007). These two models present testable, alternative expectations as to the morphology and archaeology of the two species, but importantly, also to the genetics of modern human populations.

Indeed, genetics has begun to shine light on this fundamental question of our origins. Recent analysis and publication of a Neandertal genome has shown that non-Africans share between 1 and 4% of their genome with Neandertals, which is not the case for Africans, suggesting that a small amount of inbreeding occurred between modern humans and Neandertals after the former's exit from Africa (Green et al., 2010). However, note that Green and colleagues compared the draft Neandertal genome to genomes from a San (southern African) and a Yoruban (western African) individual only. Comparison of X chromosomes from over 6,000 humans with the Neandertal X has also shown there to be a significant Neandertal component in non-Africans, further suggesting admixture between the two human species soon after the *H. sapiens*'s exodus from Africa (Yotova et al., 2011). Additionally, another recently sequenced ancient hominin genome (Krause et al., 2010; Reich et al., 2010), from Denisova in Russia, appears to have a more recent shared common ancestor with Neandertals than mod-

ern humans. Intriguingly, there are also traces of gene flow between the Denisovan and modern humans from Melanesia, but not Africa or Eurasia, suggesting separate admixture between Denisovans and humans to that between Neandertals and Eurasians (Reich et al., 2010).

The earliest archaeological evidence of modern humans in Europe is from artefact assemblages found in the region of Bulgaria and Poland (Kozlowski, 2007), which resemble Bohunician assemblages found in sites in the Near East and date to between 48 and 40kya (Richter et al., 2008). However, there is little evidence of these types in more westerly parts of Europe (Hoffecker, 2009). The Proto-Aurignacian is a more widespread group of assemblages that are believed to be a proxy for modern humans, due to similarities with another Near Eastern industry known as the Ahmarian (Bar-Yosef et al., 2000) dating to around 45kya, and which have associated skeletal remains. The earliest skeletal remains in Europe, found in Romania, date to this period and are possibly linked to the Proto-Aurignacian technology (Trinkaus, 2007). Both of these industries are associated with movements of early humans from the Near East via the Balkans to continental Europe (Hoffecker, 2009). The Aurignacian industry is believed to have developed in south central Europe some 40kya, possibly in response to a climatic cold spell , and spread rapidly across the continent by 32kya at the very latest (Churchill and Smith, 2000). There would have been a period of anywhere between 2,000 and 10,000 years of co-habitation between modern *H. sapiens* and Neandertals, but archaeologically the level of assimilation between them is still unclear (Hoffecker, 2009). So, by approximately 30kya, a population of modern humans was present across Europe. Soon after this time the continent was plunged into the last Ice Age, the most recent of a succession of ice ages that characterised the Pleistocene and which had a hugely important effect on the distributions of flora and fauna in Europe today (Mix et al., 2001; Hofreiter and Stewart, 2009).

Europe Occupied

There is an increase in the presence of technologies in the archaeological record dating to the second half of the Upper Paleolithic. Specifically: blades; bone, antler and ivory work; and art as a symbolic activity, are increasingly observed (Mithen, 2003). Of particular note are the cave paintings of western Europe, produced over a 20,000 year period from c.32-12kya (Mithen, 2003). Following the Aurignacian, increasingly more complex technologies, such as the Gravettian (around 33-20 kya), another pan-European technology, defined by bladelets and stone projectiles, persist before a fragmentation, presumably due to the retreat caused by the ice, to more regionally-specific industries, such as the Solutrean (c.24-18kya) of western Europe and the Epigravettian (c.20-10 kya) in Italy and perhaps also in central and Eastern Europe (Mellars and Cunliffe, 1994). Before the beginning of the current epoch, the Holocene, there is a final Late Glacial industry, the Magdalenian (from 22-12kya), which is characterised by a large amount of artwork and blades, and concentrated in mid-latitudinal western Europe, stretching as far as Poland and northern England by 12kya, following the expansion of populations after the Last Glacial Maximum (LGM). Estimates of the population size of western Europe, based on archaeological sites, range from 4,000 individuals during the Aurignacian, 17,000 during the LGM, with an explosion occurring with the Magdalenian, creating a total population size in western Europe of around 64,000 individuals around 12kya (Bocquet-Appel and Demars, 2000).

The effect of the last Ice Age on the current distribution of plants and animals in Europe cannot be over-emphasised. This final massive episode of climatic change removed our ability to study human genetic population structure *prior* to the recolonisation of Europe because the genetic slate was effectively wiped clean (Gamble et al., 2005). So these evolving technologies must be viewed within the context of their interaction with climate change, and the associated ecological, behavioural and cultural changes that were needed to mitigate its effects. Indeed, they may well have

evolved *in response to* the increased pressures of this climate change (Banks et al., 2008). It also provides an upper limit for the age of populations in northern Europe: we know that the ice penetrated as far down as Scandinavia and the British Isles, with much of the region between the ice front and southern regions arid, cold and uninhabited (Banks et al., 2008).

From foraging to farming

Following the LGM, some 26-19kya, when the world's ice sheets were at their largest (Clark et al., 2009), the ice steadily retreated pole-ward. The Mesolithic period followed this last Ice Age (Mithen, 2003) and it is at this point in history that the foundations of the current European population were laid. A brief interstadial period from 13 to 11.5kya, known has the Younger Dryas (Gamble et al., 2004), temporarily halted the expansion of hunter-gatherers back into central and northern Europe, but at around 10kya the rate of climate change in Europe was at its peak and this point marks the beginning of the Mesolithic proper (Mithen, 2003) and the establishment of modern humans in Europe. The continuing retreat of the glaciers and the warming climate contributed to the changing environment across Europe and the amelioration of the continent to human occupation (Mithen, 1994). The human population was rapidly growing as people moved north from southern refuges in Iberia, Italy and the Balkans, where they had weathered the Ice Age (Gamble et al., 2005). These people would have been organised into groups and networks of hunter-gatherers that would have stretched across large regions at a population density of 0.005-0.5 people/km^2 (Mithen, 1994). Small bands of pioneering hunter-gatherers would have led the re-colonisation of the north exploring the new environment for fresh resources (Housley et al., 1997). These Mesolithic hunter-gatherers, concentrated mainly in western Europe (Mithen, 2003), would certainly have been descendants of the Paleolithic humans described above, and through a foraging lifestyle would have made a living by dispersal, mobility and diversification (Whittle, 1996).

Driven by the improving climate (Blockley and Pinhasi, 2011), it was around this time that developments began in the region of south-western Asia known as the Fertile Crescent, which would change human world history forever. The Neolithic period hurriedly followed the Mesolithic and is generally regarded as the era when farming began, but when stone tools were still used and ceramic pottery was invented (Whittle, 1996; Anthony, 2007). The origins of agriculture can be traced to this region, defined as the area to the west of the Syrian desert and to the north and east of the Tigris and Euphrates rivers, and from the Nile in the south to the Taurus mountains of Anatolia to the north (Barker, 2006). Agriculture and the domestication of wild animals arose independently in up to nine separate regions around the world between 11kya and 2kya (Diamond and Bellwood, 2003). In the context of Europe, Vere Gordon Childe first proposed the notion of the ``Neolithic Revolution'' (Childe, 1925, 1942), believing the origins of agriculture to be in the Fertile Crescent, based on the presence there now of the wild relatives of barley, wheat and goats (Barker, 2006). Childe's theory was that farming spread through Europe not through trade but migration using the term ``Revolution'' to describe the upheaval that this change brought (Childe, 1925, 1942; Barker, 2006). This idea has largely been vindicated with increasingly detailed and sophisticated archaeobotanical and genetic studies pointing towards the origins of European domesticates to this region and time (Brown et al., 2009), and has been of particular interest in archaeology (Clark, 1965; Ammerman and Cavalli-Sforza, 1971; Pinhasi et al., 2005; Blockley and Pinhasi, 2011). Ammerman and Cavalli-Sforza subsequently promoted the ``wave of advance'' model of Neolithic expansion (Ammerman and Cavalli Sforza, 1984), based initially on the archaeological work of Ammerman (Ammerman and Cavalli-Sforza, 1971), but also later in relation to human genetic variation (Menozzi et al., 1978; Cavalli-Sforza et al., 1994; Cavalli-Sforza, 1997). The wave of advance model asserts that, because farming necessarily leads to an increase in population growth, agriculture spread across Europe by local migratory movement of people at the front of the wave of advance (Cavalli-Sforza et al., 1994). Archaeologists also believe that the Neolithic was unlikely to have spread from south-western

Asia without accompanying people (Mithen, 2003; Barker, 2006).The wave of advance model is likely to be too simplistic a representation of a process that involved farmers moving towards pockets of resources (Housley et al., 1997), and that was both spasmodic and lengthy (Sherratt, 1994b; Whittle, 1996; Pinhasi et al., 2005; Blockley and Pinhasi, 2011). Nevertheless, agriculture did spread from south-western Asia, and disregarding for the moment whether colonisation or acculturation was the dominant method of change, it is against this backdrop of broad cultural movement that I shall now discuss the human population history of Europe.

The Spreading Neolithic

The Neolithic spread from south-western Asia through Anatolia and Turkey to the Balkans arriving between 7,000 and 6,000 BCE[1](Whittle, 1996; Mithen, 2003). This is attested by the presence of farming settlements in the south-east European archaeological record, such as Nea Nikomedia in Greece (Mithen, 2003), and the existence of Impressed Ware pottery (Barker, 2006). This pottery, also known as Cardial Ware due to its decoration with shells of the cockle *Cardium edule*, hints towards a role for the Mediterranean coast in its dispersal (Whittle, 1996; Barker, 2006). This seafaring theory is further reinforced by the presence of the earliest definite indications of agropastoralism in the Mediterranean being on the island of Cyprus (Barker, 2006), the earliest Impressed Ware sites being found along the Dalmatian and south-east Italian coasts (Spataro, 2011), and the existence of early Neolithic sites on Crete and in southern Italy (Mithen, 1994) which must have had to have been reached across the Adriatic Sea. The Neolithic also brought increasingly sophisticated stone tools, art in the form of figurines, deliberate burials and the development of long-term settlements (Whittle, 1996).

The agricultural frontier stretched up to the northern Balkans, in the region of the

[1]For the remainder of the introduction dates will be represented as either BCE (Before the Common Era) or CE (Common Era): the international equivalents of BC (Before Christ) and AD (Anno Domini), respectively.

lower Danube and then appears to have remained there for a thousand years (Barker, 2006). Although Cardial Ware, and thus farming, was present in Iberia and Mediterranean France, having slowly spread from southern Italy (Sherratt, 1994b), western and northern Europe was still largely inhabited by hunter-gatherers (Mithen, 1994). In 5,500 BCE a novel culture of farmers, named *Linearbandkeramik* (LBK) after their banded ceramic pottery, emerged from northern Hungary, with a distribution spreading to the Carpathian mountains in the east, north to the Baltic, and west as far as modern day Paris (Barker, 2006). Its distribution roughly matched that of the agriculturally productive loess soils in Europe (Sherratt, 1994a; Whittle, 1996; Barker, 2006). Settlements began to contain several house units, including farmsteads (Barker, 2006). There is still debate as to whether the LBK was spread by migrating farmers or whether it was the result of acculturation by the foragers already present in central Europe (Whittle, 1996; Barker, 2006), however genetic analysis of ancient DNA from putative LBK farmers suggests the former, having found genetic continuity neither with contemporary Europeans nor Mesolithic foragers (Haak et al., 2005) but perhaps with present day Middle Eastern populations (Haak et al., 2010).

In northern Europe following a millennium of co-existence between the farmers in the south and the foraging hunter-gatherers in the north, the beginning of the Neolithic is associated with the appearance of the *Trichterbecher Kultur*, or Funnel Beaker Culture (TRB), c.4,000 BCE (Whittle, 1996). In Scandinavia, TRB, itself a mixed foraging-farming culture (Barker, 2006), existed in parallel with one of the last hunter-gatherer populations in Europe, known as the Pitted Ware Culture (PWC). A recent genetic study on ancient DNA from TRB and PWC burials suggests that there is no continuity between contemporary Scandinavians and European hunter-gathers (PWC), agreeing with the replacement of hunter-gatherers by farming populations (Malmström et al., 2009). So, while sporadic in both time and space, farming spread throughout Europe, through an as yet unknown combination of acculturation and colonisation. Almost as soon as the first farmers were sowing seeds in Britain, two more equally important developments were happening in south-east Europe and the Middle East: the invention

of metallurgy and the beginnings of urban civilisation.

Copper, Bronze, Trade and the Dawn of Western Civilisation

Copper-working began in the Balkans sometime around 5,000 BCE, with central, western and northern Europe remaining essentially oblivious to the development (Sherratt, 1994a). Although much of central Europe outside of the Danubian basin was still covered by foragers (Mithen, 2003), the Balkan smiths learnt how to construct ornaments and jewellery (Sherratt, 1994a), some of which were exchanged across early trade networks (Anthony, 2007). At this time, in the Caucasian and Russian Steppe to the east of Europe the horse was being domesticated (Anthony, 2007). The creation of trade led to social developments as well. Now it was possible to become a decorated chief, replete with copper ornamentation and, perhaps most importantly, power over other members of the community (Anthony, 2007). Objects of social prestige were increasingly used as grave goods, with copper being used mainly for competitive display, rather than for any major functional developments such as tools or weapons (Sherratt, 1994a). Meanwhile, in south-western Asia, an independent metallurgy evolved, more sophisticated in fashion, involving alloys and complex castings that would lead to the production of bronze (Sherratt, 1994a). The emergence of urban society in Mesopotamia c.3,500 BCE further accelerated the development of strong alloys (Sherratt, 1994a).

In southern Iraq the first cities were founded from the villages and towns of farmers, from c.5,000 BCE in a region known as Sumer, where agriculturists learnt irrigation techniques that allowed food to be grown for large numbers of people (Wood, 1992). With the origins of Indo-European languages traced to around this time (Gray and Atkinson, 2003), and the presence of the first cuneiform writing in the archaeological record a little later, c.3,000 BCE (Beckwith, 2006), the evidence points towards this as the moment that urban civilisation began. It is believed that people from the Eurasian steppes who are associated with the spread of the Indo-European language and the do-

mestication of the horse, also invented the wheel, and thus the wagon and chariot and the ability to move materials over long distances (Beckwith, 2006; Anthony, 2007). Whether and when Indo-European languages were spread with farmers from Anatolia (Renfrew, 2000b; Gray and Atkinson, 2003), or from a separate Kurgan expansion (Gimbutas, 1973) is still under debate, but the vast majority of European, and many central Asian languages, have the same lexical routes.

In the Atlantic zone of Europe, to the west and north, from 2,800-1,800 BCE, a mobile and broad culture grew out of the Rhine valley that is attested by the appearance of individual burials in round mounds, small stone and metal ornaments, and handleless drinking cups called Bell-Beakers (Sherratt, 1994a). The Bell Beaker Phenomenon represents the emergence of new maritime links, and the development of leather artefacts, in particular belts, which were a sign of the elite, and the development of social stratification in western Europe (Sherratt, 1994a). Importantly, these developments led to the origins of long distance trade routes, across Europe and into south western Asia, the Caucasian steppes and beyond. In the Aegean, the ancient civilisations of the Minoans in Crete and Mycenaean Greece, some of whose palaces still stand today, expanded during the maturing Bronze Age from around 2,000-1,000 BCE (Wardle, 1994). Descendants of these first civilisations are the ancient Greeks of the Homeric epics in the era of Troy and later Alexander the Great. Indeed, the evolution of these, and other Indo-European civilisations such as the Anatolian Hittites (c.1,650 BCE), combined with the greater mobility provided by the horse and the wheel (Anthony, 2007), and the growing spirituality and religion associated with the Indo-Europeans (Beckwith, 2006) led to the beginnings of ``heroic'' warfare and the subjugation of many under an elite (Wardle, 1994). The picture at this time, then, is of an introgression of new language and technologies from the European periphery, which because of an increase in trade, may have been associated with the movement of people.

The Moving Peoples of Europe

Before the first millennium BCE, the different cultural, social and technological advances that had shaped Europe and its periphery had largely arisen through the gradual diffusion of both people and ideas. At this point in history, different civilisations in central Eurasia, south-west Asia and the eastern Mediterranean began to expand and come together under a common cause. The Scythians, fierce nomadic steppe warriors from central Iran raided the eastern periphery of Europe, but also developed the hugely important trade system that became the Silk Road (Beckwith, 2006). Following the collapse of the once thriving Mycenaean and Cretan Minoan civilisations at around 1,200BCE, the Phoenicians, a seafaring Semitic people from the Levant, jostled for trading supremacy with the city-states of ancient Greece in the eastern Mediterranean (Cunliffe, 1994a). From 800 BCE, populations within the increasingly urbanised and expanding Greece, began leaving home to set-up in the new Greek colonies of the Mediterranean (Cunliffe, 1994a). The western Mediterranean was further opened to the extensive trade networks of these colonies and trading posts. Southern Italy and Sicily became known as *Magna Graecia*, and the Greeks traded with western European ``barbarians'' through colonies in coastal towns such as Marseille (Cunliffe, 1994a). The Phoenicians established direct trading links with north Africa and the Tartessos in Spain and at this time built an important settlement in Carthage in modern-day Tunisia (Cunliffe, 1994a). When their founding cities in the Levant were overrun by the Babylonians in 573 BCE, it was to Carthage they fled, forming a strategic alliance with the Etruscans of northern Italy (Cunliffe, 1994a). Whilst not the first Italians, the Etruscans were skilled metallurgists and with iron weapons established a hegemony that at one time stretched all the way across the Italian peninsular which included a small Latin city on the south bank of the Tiber called Rome (Roberts, 2007). The influence of the mysterious Etruscans was short lived but important nevertheless in accessing Greek tradition and the growth of Rome as a city (Roberts, 2007).

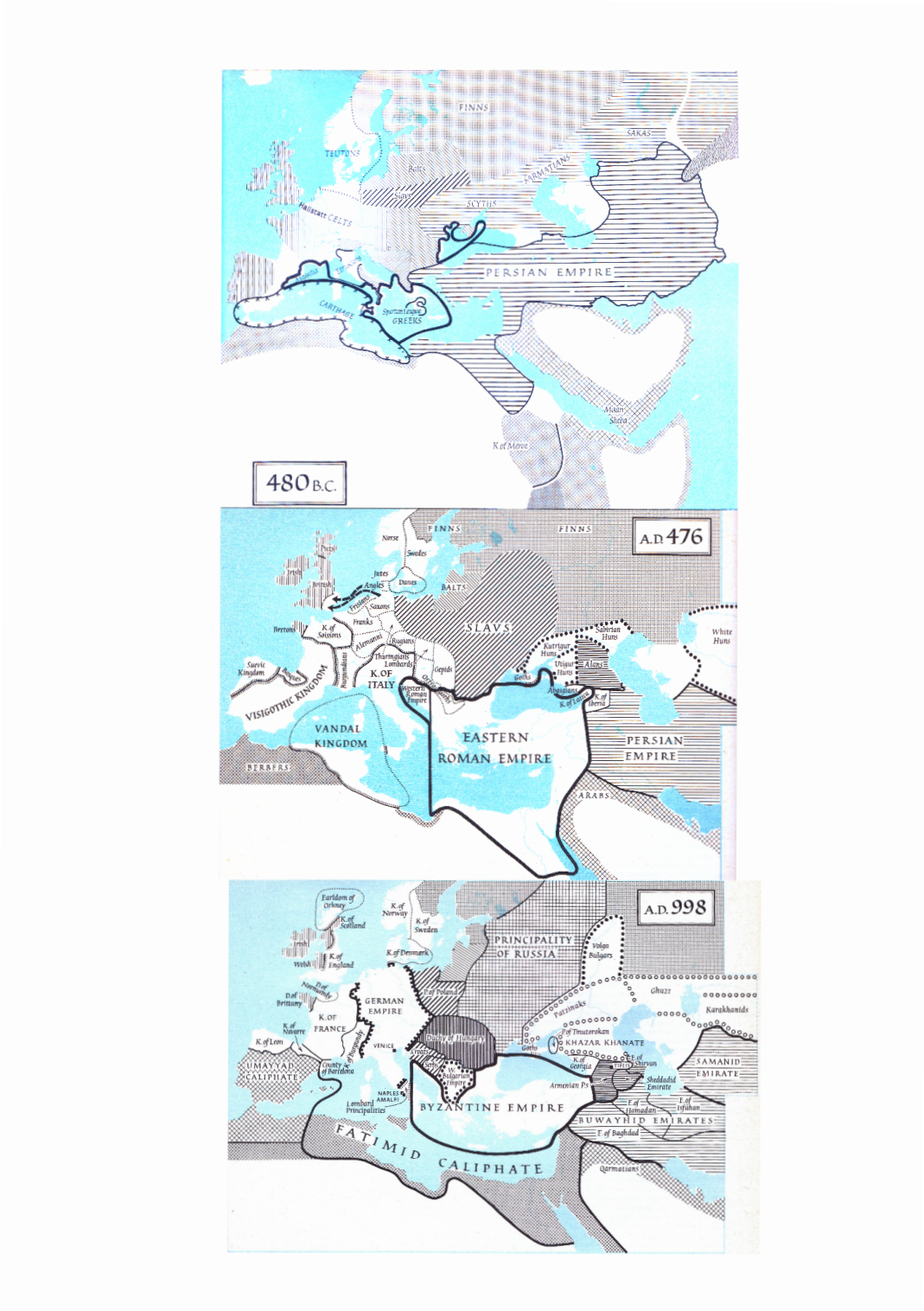

Figure 1.1. Three historical snapshots of European peoples. The top frame depicts Europe in 480BCE, during the time of the Ancient Greeks. The second frame is approximately 1,000 years later, as Europe begins to segment after Roman hegemony. The bottom frame shows Europe approximately 1,000 years ago, at the height of the Arab Conquests (after McEvedy, 1967a,b).

In Iberia, the Tartessians avidly traded with the Carthaginians (Phoenicians) and Greeks, as is attested by the large quantities of colourful eastern Mediterranean pots and ornaments that can be found in the archaeology of this area (Cunliffe, 1994a). Within Iron Age central Europe, the *Hallstat* zone, a region and people identified by the distinctive burial of its elite, was spread across the central European Seine, Rhine and Danube rivers. Evidence of both Mediterranean trade with the Greeks and Etruscan trade over the Alps has been found at these burial sites (Cunliffe, 1994a). The western *Hallstat* collapsed c. 500 BCE and in its place emerged an elite centred on a new and strikingly different and eclectic art form, known as Celtic or *La Tène* (Cunliffe, 1994a). North of this central European/Mediterranean contact zone were the late Bronze Age hill-forts of west-central Europe, the village economies of the northern-European plain, and the exchange networks of the Atlantic fringe (Cunliffe, 1994a).

Later in the first millennium, around 400 BCE, the city of Rome in Italy began to shake off Etruscan domination and began the path towards becoming an empire (Cunliffe, 1994a). Notwithstanding the Celtic migrations, Rome spread both north and south along the Italian peninsula, enveloping the declining Italian cities of *Magna Graecia* in the process (Cunliffe, 1994a). In the latter part of the fifth century BCE, ostensibly due to over-population, people from the old *Hallstat* region in central Europe migrated both east along the Danube to Hungary, and south toward Italy, in what have become known as the "Celtic migrations". This involved a number of different tribes moving to different regions in Europe (Cunliffe, 1994a). Although the La Tène art form spread across the vast majority of Europe at this time, there is no evidence to suggest that it was spread west through migration (Cunliffe, 1994a). In 206 BCE, the Romans took Cádiz in southern Iberia and with it the Carthaginian trade network and tightened their increasingly strong grip on both the peninsula and the Mediterranean coastline (Cunliffe, 1994a). In time, Rome would unlock the enormous potential of barbarian Europe as the Republic expanded ever more into the further reaches of Europe (Cunliffe, 1994a).

At its heart, the Roman Republic was interested in obtaining the wealth, raw materials

and manpower to maintain itself (Cunliffe, 1994b). Northern barbarian hordes, in the form of the *Teutones* and *Cimbri* antagonised the Romans but were eventually decimated by the armies of Rome around 100 BCE. Over the following century Rome would conquer and engulf the Belgic and Celtic tribes of Gaul (France) and most of Europe south and west of the Rhine-Danube frontier (Cunliffe, 1994a).

The Roman Empire co-existed with the northern barbarian tribes for a number of centuries through a complex interrelationship based on trade and diplomacy (Cunliffe, 1994a; Todd, 1994). However, starting in the third century CE, the non-Roman armies of the north began to attack the Empire (Todd, 1994). The Goths, who were not a single homogenous people, but a loose group of varied ethnic elements formed in the Pontic steppe north of the Black Sea, repeatedly attacked the Balkan and Italian Roman Empire, eventually becoming a paid federate (Todd, 1994). Having migrated from the Pontic Steppe, the western Goths, or Visigoths, settled in Aquitaine in Gaul and spread into northern Iberia(Todd, 1994). Following a rapid thrust westward in 370 CE, steppe nomads known as Huns, subjugated the Goths in the east, known as the Ostrogoths, for a period, although in the sixth century CE the two gothic kingdoms were linked to Rome: the Visigoths in Iberia and the Ostrogoths in Italy (Todd, 1994). Following over a millennium of Mediterranean and Ancient Near Eastern cultural dominance in Europe, the Germanic migrations from northern and eastern Europe marked a shift to a more central Eurasian cultural complex (Beckwith, 2006). At this time too, the Angles, Saxons and Jutes crossed the English Channel, and the Vandals, a people from north of the Rhine, marched south through Gaul and Spain, establishing a kingdom in the north African city of Carthage (Beckwith, 2006). The Vandals were later overpowered by the Byzantines of the eastern Roman Empire (Todd, 1994). The Franks, a Germanic people from east of the Rhine River gradually spread their control over Gaul from c.500 CE (Todd, 1994; Beckwith, 2006).

Major invasions from western Russia occurred in the sixth century AD. The Avars appeared from the Caucasus and attacked the people of the Black Sea (Todd, 1994). The dominance of the Avars, however, lasted a short while as during the middle of the

seventh century CE, the Slavs became the dominant force in east and central Europe (Todd, 1994). The Slavs, likely an amalgam of Baltic peoples to the north and Germanic peoples from the west, moved south and west to fill the void left by the Franks and other Germanic groups (Todd, 1994). Although the Franks withheld them to the east of the Rhine, the Slavs migrated south through Bohemia, Poland, eastern Germany and into the Balkans (Todd, 1994). Their occupation of Greece was ended by the Byzantines in the ninth century, but so widespread was their dispersal that numerous Slav states emerged over the eighth and ninth centuries (Todd, 1994).

In 637 CE, five years after the death of Muhammed, his Arab followers defeated the Persians in south-western Asia (Beckwith, 2006). With victories over the Byzantines in Syria and further successes in the Near East, the Arabs captured Egypt in 640 CE and went on to conquer North Africa in the west and destroyed the Sasanid armies to the east (Beckwith, 2006). Over the following century the Arabs would gain control over much of central Eurasia (Beckwith, 2006), and from North Africa plundered Spain in 711 CE, moving up as far as Poitiers in 734 CE, where they were beaten by Charlemagne and the Franks in Tours, and retreated south of the Pyrenees (Beckwith, 2006). The Arabs would be a major presence in Iberia and the Mediterraean for the next eight centuries, including forays into southern Italy during the ninth century, until their final expulsion from Spain in 1492 CE.

Whilst the migrations of people mentioned above perhaps imply large scale movements of people, the actual level of migration, in terms of numbers of people is an active research question. Historians have been able to largely falsify suggestions that historic migrations involved large-scale replacement (Heather, 2009). The two major alternatives to large scale replacement are models based on wave of advance type movements, and elite replacement. I have covered wave of advance model elsewhere, but the elite transfer models are gaining traction with historians, because they have the potential to explain various different scenarios in recent human history (Heather, 2009). The idea here is that a small population intrudes on the resident population of an area, but because they quickly assume power and control over a population, they

are able to effect change on a large scale. The effect of these various models on current genetic structure is uncertain, but it is important to note here, that there is no real historical evidence of mass migration across large distances, involving large-scale population replacement.

1.2. Human Population Genetics

In 1900 Karl Landsteiner discovered the ABO blood system and that most individuals belonged to one of four different blood groups (Landsteiner, 1900). Populations from different parts of the world tended to show different frequencies for different blood groups and, given that these differences were heritable, geneticists soon realised that this information could be used to infer something about the ancestry of these populations. Following this first attempt to understand genetic variation, in its broadest sense, research of this sort really took off in the 1940s and 50s when it was discovered that there was a huge amount of variation within many of the proteins in human blood. For example haemoglobin, the molecule in blood that carries oxygen, was shown to vary by a single amino acid in some people, and whilst those with one type were medically normal, people with the alternative type had sickle cell anaemia (Pauling et al., 1949; Ingram, 1957). It was possible to observe these differences by applying an electric field to the proteins: the single amino acid difference giving the two proteins a different charge which allowed them to be separated across an electric field (Cavalli-Sforza et al., 1994). This process of electrophoresis was further developed and used to investigate the variation of many different proteins present in blood (figure 1.2). Studying the differences in these so called ``classical markers'' paved the way for the use of genetics to study variation and population history.

From these beginnings, the field of human evolutionary genetics has grown exponentially over the last 20 years. Initially, the loci of choice in human genetic studies were the haploid uniparental markers, first mitochondrial DNA (mtDNA; e.g. Cann et al.

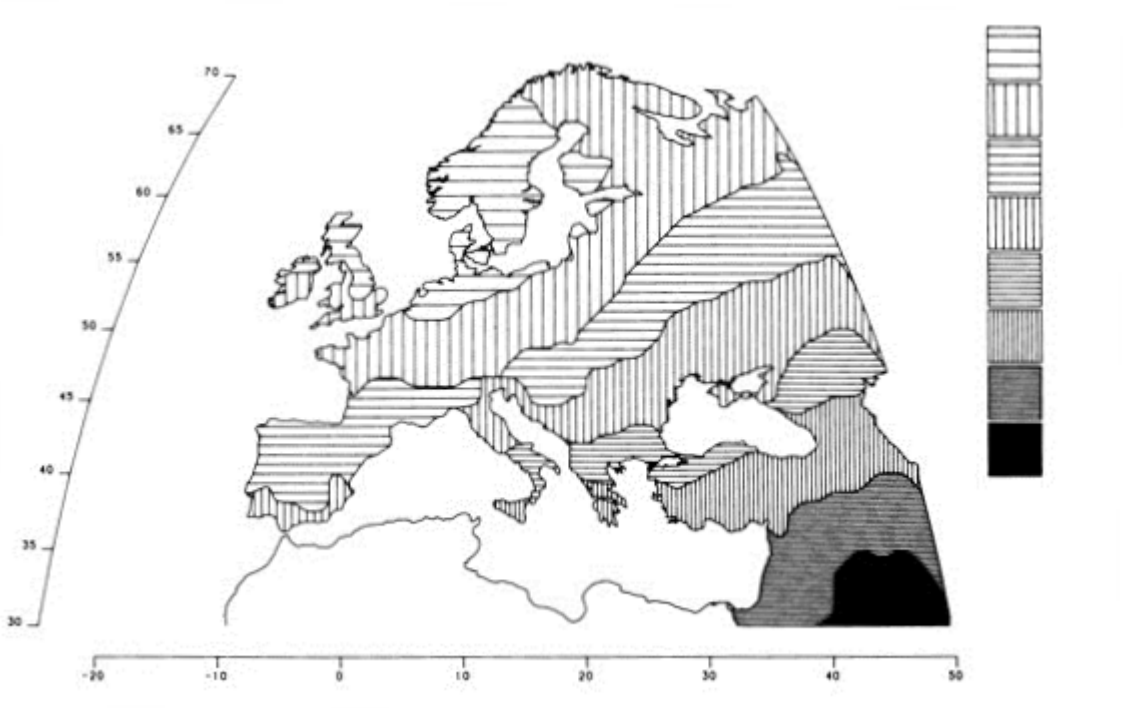

Figure 1.2. The first principle component of variation in 95 classical polymorphisms (Cavalli-Sforza, 1997).

(1987)), which is much more abundant than nuclear DNA, and then the Y chromosome (Semino et al., 1996). Both of these markers undergo no recombination when passed down from generation to generation and so any differences between two Y chromosomes or mtDNA sequences should be down to mutation alone. The Y chromosome is passed from father to son only, whilst mtDNA is passed from a mother to all of her children. Whilst investigating these two systems can only give a snapshot of the maternal or paternal history of a population, they have nevertheless been employed with great effect to elucidate an understanding of the evolutionary history of human populations. Modern methods, both technical and statistical, are emerging that are making the study of the whole human genome a possibility by taking into account multiple loci and incorporating information on recombination and selection.

The Y Chromosome

The Y chromosome, once believed to be ``rich in junk, poor in useful attributes, reluctant to socialise with its neighbours and with an inescapable tendency to degenerate'' (p598; Jobling and Tyler-Smith, 2003), has provided geneticists with a valuable resource with which to study our evolutionary origins. Its effectiveness as a genetic record of the past stems from our knowledge of its deep phylogeny, which is now based on approximately 600 biallelic markers (Karafet et al., 2008). The latest Y chromosome tree displays these biallelic markers, which are more usually known as single nucleotide polymorphisms (SNPs), along its branches, making it possible to group Y chromosomes on the basis of the identity of these SNPs. The assumption is that, given the short evolutionary timescale within which we are investigating, any mutation will have occurred only once, and so individuals with the same nucleotide at a given position, are perceived to be more closely related to each other than those that do not, and can therefore be grouped together. The assumption of a single mutational event largely holds true (although see Sanchez et al. (2004) and Adams et al. (2006)), and it is further possible to infer the ancestral and derived state of each SNP by observing the base present at a particular locus in different parts of the tree. Thus, individual Y chromosomes can be assigned to a haplogroup depending on which SNPs are derived and which ancestral. The Y Chromosome Consortium (YCC) has agreed upon a common nomenclature that is updated as new SNPs are discovered, and which is used to label the branches and terminal groups in such a way that the hierarchical nature of the tree is shown in the name of the haplogroup to which a sample belongs (Hammer, 2002; Karafet et al., 2008).

Global Distribution of Y Chromosome Haplogroups

One of the most striking patterns observed from some of the first detailed studies of the global distribution of Y chromosomes is that there is geographical clustering of haplogroups (Jobling and Tyler-Smith, 2003). Figure 1.3 shows the global distribution

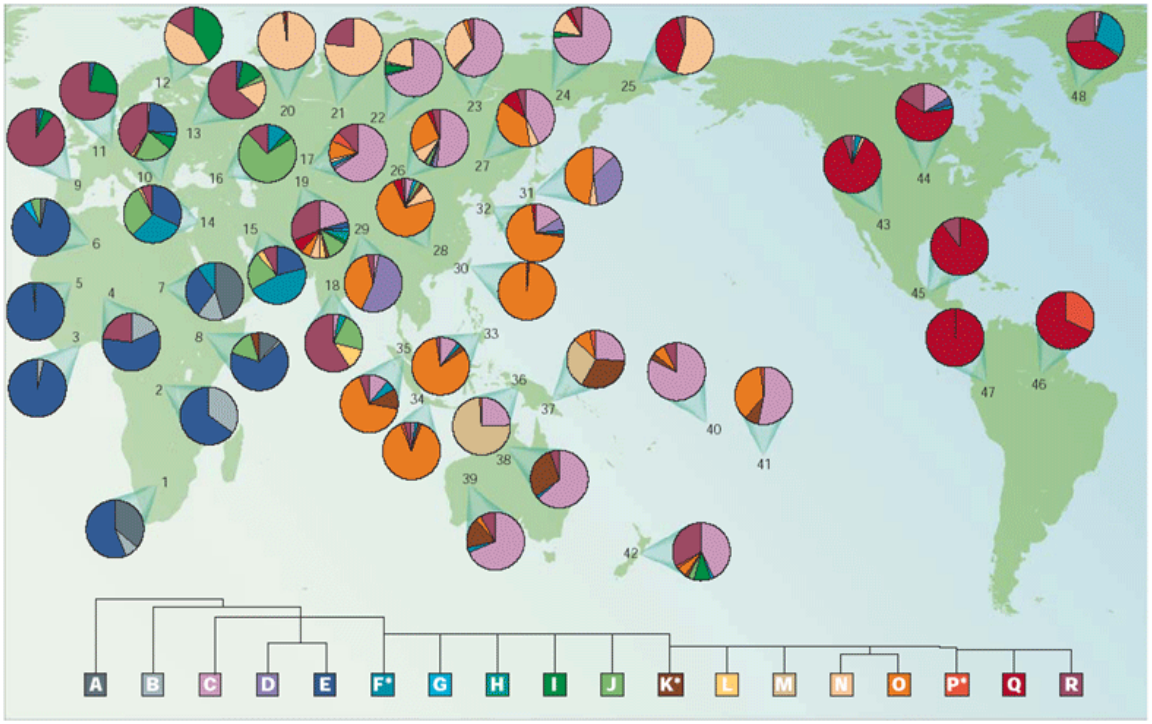

Figure 1.3. The global distribution of Y chromosome haplogroups (Jobling and Tyler-Smith, 2003).

of Y chromosome haplogroups together with the outline Y chromosome tree at the bottom of the figure. The pie charts in this figure show that the frequencies of close populations are more similar to each other than to those of populations that are further away. Moreover, the frequency of haplogroups appears to decay as one moves from one continental extreme to another. For example, haplogroup E (dark blue) is almost ubiquitous in western Africa, but drops to frequencies of just over 50% in the south of Africa, around 50% in eastern Africa, around 20% in southern Europe, and is almost non-existent in Asia and the Americas. On a global scale, haplogroup frequencies differ in a definite pattern.

Y Chromosome Haplogroups in Europe

The very first Y chromosome census of Europe was performed by Semino and colleagues in 1996 (Semino et al., 1996), and included only two Y chromosome polymorphisms, p12f2 and 49a,f which were detected by restriction enzyme analysis and

are known as restriction fragment length polymorphisms (RFLP). In around 3,000 individuals from across Europe, and a few from Asia and Africa, they showed that one specific allele, the 8 kilo-base (kb) RFLP of the p12f2 system, was more frequent in Near Eastern and southern Mediterranean populations and non-European individuals, with the frequency declining from south-east to north-west Europe (figure 1.4; Semino et al. (1996)). The 49a,f RFLP is a collection of many different fragments, the combinations of which are equivalent to haplotypes. Within this system, a particular haplotype (Ht 15) was more frequent in north western Europe and dropped off in frequency towards central Europe and was not present past the Balkans or in the Near East, Asia or Africa (Semino et al., 1996). This was the first time a cline was observed in the distribution of a Y chromosome marker, and the authors argued that the correlation with the known archaeological route of agriculture out of the Middle East supported the demic movement of farming out from the south western Asia during the Neolithic (Ammerman and Cavalli Sforza, 1984).

This work was updated in 2000 to use 22 Y chromosome markers in 1007 individuals (Semino et al., 2000). Major clines across Europe were observed for the first time for the SNP-defined haplogroup M173, excluding individuals with the downstream M17 SNP - the common nomenclature for such a haplogroup is M173(xM17) - from western to eastern Europe, and for SNP-defined lineage M17 which was also clinal from higher frequencies in north-eastern Europe to lower frequencies in central Europe, and absent in western and south-western Europe. (Haplogroup M173(xM17) is now referred to as haplogroup R1(xR1a)). High frequencies of a combined haplogroup containing SNPs M201, M172 and M89(exM9,M170,M69), which roughly equates to what we now call haplogroups G and J, were observed in the Middle East which decreased as one moves north-westerly into Europe. Rosser and colleagues, using a more complicated nomenclature based on different markers, essentially showed the same relationship, with clinal distributions of Haplogroup 1 (equivalent to M173(exM17)), Haplogroup 3 (equivalent to M17) and Haplogroup 2, which appeared to spread from the Middle East (Rosser et al., 2000).

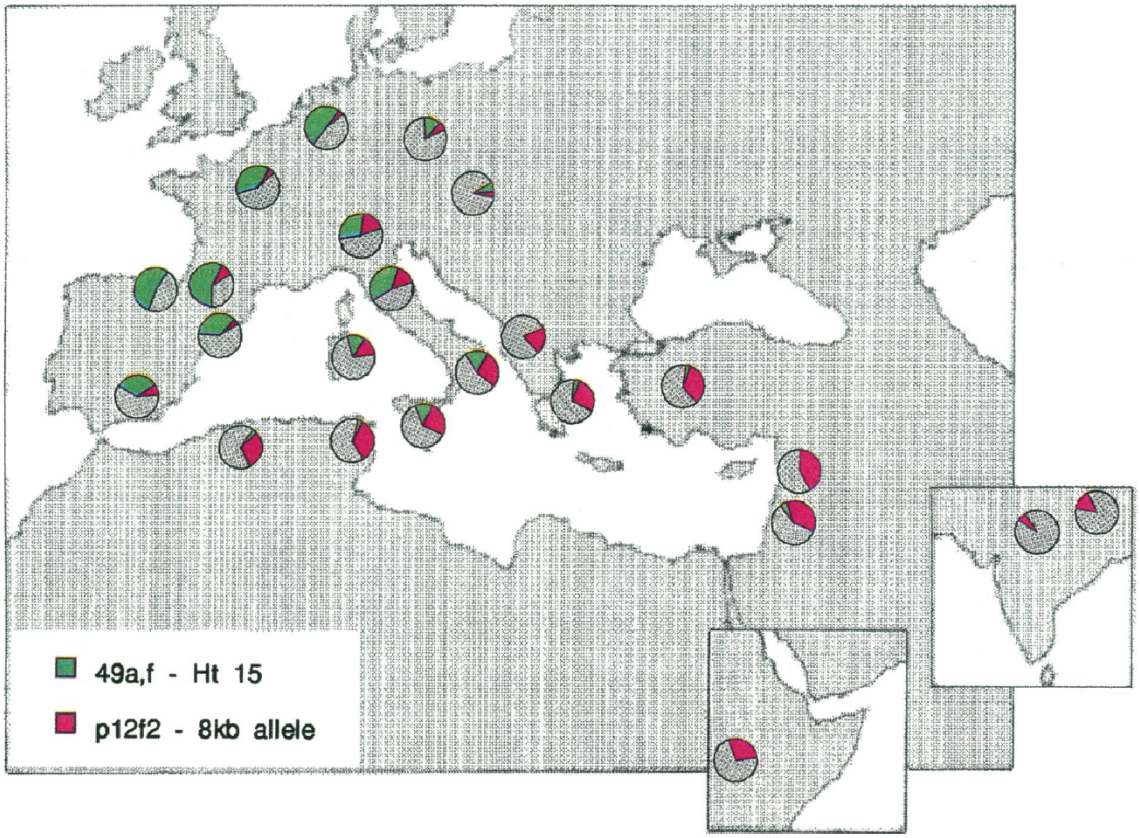

Figure 1.4. A first attempt to explore Y chromosome frequency differences in Europe (Semino et al., 1996).

The presence of these genetic clines was putatively linked to ancient population movements that archaeologists had long recognised in the archaeological record across Europe. The key debate related to these findings in Europe concerns whether farming was spread through the demic-diffusion of ``new'' people across the underlying genetic substratum of the initial Paleolithic people of Europe, as suggested by nuclear (Chikhi et al., 1998) and some Y chromosome data (Rosser et al., 2000; Semino et al., 2000; Chikhi et al., 2002; Balaresque et al., 2010; Busby et al., 2012), or was the spread of agriculture achieved through the diffusion of ideas alone, as may be inferred from mtDNA data (Richards et al., 1996), where clines are not observed. Ideally, researchers should integrate information from multiple loci (Wilson et al., 2001), such as Cavelli-Sforza's original 95 classical markers (Cavalli-Sforza et al., 1994). However, clines produced by summarising multivariate datasets, such as those from principle components analysis (PCA), can cause ambiguity as they may simply be the result of

isolation by distance, and thus reflect continuous population structure rather than a specific migratory event (Novembre and Stephens, 2008). Further debate within the demic-diffusion camp centres around the relative contributions to the current Y chromosome landscape of the migrating farmers, from a small yet important proportion (Rosser et al., 2000; Semino et al., 2000), to almost complete replacement (Balaresque et al., 2010).

The Neolithic Revolution (see page 8) is a central issue to archaeology and one which geneticists feel they have the tools to investigate. Was the cultural change from itinerant hunting and gathering to sedentary farming accompanied by a movement of people, or by the movement of the idea alone? Is the current European gene pool a remnant of the very first modern humans who strode into Europe some 40,000 years ago, or can most Y chromosomes trace their ancestry to a group of farmers who domesticated plants and animals in the Near East a mere 10,000 years ago? As mentioned on on page 21, frequencies of Y chromosome types show a clinal pattern in Europe, as do nuclear but not mtDNA markers (Soares et al., 2010), which correlate with archaeological patterns of expansion. The next step then, is to try to date Y chromosome lineages across Europe to test whether these clines could be the result of either Paleolithic or Neolithic human expansions. In order to address this and other questions relating to the age of specific haplogroups, it is necessary to use a different type of genetic marker. Y chromosome Single Tandem Repeats (STRs) are polymorphic microsatellites that evolve much more quickly than SNPs and can therefore be utilised to estimate diversity within Y chromosome lineages (Roewer et al., 1992; Jobling and Tyler-Smith, 1995, 2003). Individual Y chromosomes can be assigned both a haplogroup based on the presence of derived SNPs, as well as a haplotype, which is a list of allele sizes for an individual for a given set of STRs. The faster mutating STRs can be used to investigate the diversity within a lineage under the assumption that the greater the variation in haplotypes of the individuals derived at a particular SNP, the older this lineage is likely to be.

Dating Y Chromosome Lineages

When a large number of STRs are used (more than 15), most individual Y chromosomes within a population can be distinguished, and as such a lot of effort has been expended by the forensic community in characterising these polymorphisms (Roewer et al., 1992; Jobling and Tyler-Smith, 2003; Gusmao et al., 2006; Ballantyne et al., 2010). From a population genetics point of view, STRs are extremely useful in establishing the diversity of lineages, and the putative ages of them. The accuracy of this endeavour hinges upon correctly calibrating the mutation rates to the differences between Y chromosome alleles to produce a value for the time for the most recent common ancestor (TMRCA) between the chromosomes in question.

Mutation rates can be estimated based on the observed mutations in father/son pairs (Heyer et al., 1997; Kayser et al., 1997, 2000; Gusmão et al., 2005; Onofri et al., 2009; Ballantyne et al., 2010). It is also possible to calculate an ``evolutionary'' mutation rate by estimating the diversity present in a population with a known time of origin (Forster et al., 2000; Zhivotovsky et al., 2004, 2006). Unfortunately, when mutation rates are calculated in this way, they tend to give rates that are up to three times slower than those based on observation alone. This is of enormous significance when trying to establish the age of a lineage based on STRs, as estimates of lineage age will differ by a factor of up to 3, depending on which rate is used.

The mutation model assumed is also a factor to consider when using STRs to make inferences about population history. The most basic model is the single-step symmetric stepwise mutation model (S-SMM), where STRs mutate to one motif length shorter or longer with equal probability (Ohta and Kimura, 1973; Kimura and Ohta, 1978). In the generalised stepwise mutation model (G-SMM) the length change can also be multi-step and be directionally asymmetrical (Di Rienzo et al., 1994; Amos and Rubinstzein, 1996). A microsatellite ``molecular clock'' was established, using the G-SMM, once it was shown that the average squared distance (ASD) between two alleles drawn from the same population was equivalent to μT; that is, the product of mutation

rate, μ, and the time in generations to the most recent common ancestor, T (Goldstein et al., 1995a,b). This allowed the estimation of the average coalescent time, and therefore the TMRCA of alleles within a population, with the TMRCA between two lineages being $2\mu T$ (Goldstein et al., 1995a,b). Recent work has further argued that ASD, which is free of any assumptions of demographic history, can offer accurate insights into divergence times between populations for genome-wide STRs (Sun et al., 2009).

Alternative STR dating methods are available. Now almost ten years old, Bayesian Analysis of Trees With Internal Node Generation, or BATWING (Wilson et al., 2003), applies a Markov Chain Monte Carlo (MCMC) procedure to generate a series of genealogical trees with associated demographic parameter values consistent with the data (Shi et al., 2010). Posterior estimates with confidence intervals of several parameter estimates are produced, including: the effective population size of the sample; the TMRCA of the sample; population growth rates; and various tree parameters. BATWING works primarily with population-level data incorporating SNP and STR data, but has recently been used to date a lineage (Balaresque et al., 2010). Other methods involving the use of haplotype networks and the rho statistic, founder analysis, and the comparison of long Y chromosome sequences are also used. Chapter 3 of this thesis explores Y chromosome dating in further detail.

mtDNA

Mitochondrial DNA has long been used to elucidate the demographic history of our species (Cann et al., 1987; Ingman et al., 2000; Simoni et al., 2000; Torroni et al., 2001; Behar et al., 2008). Phylogeographic analysis of European mtDNA has suggested that Europe was populated roughly 50kya, in line with archaeological evidence, and the comparatively homogenous nature of mtDNA haplogroups in Europe appears to point towards an older origin for European mtDNA haplogroups, than those for European Y chromosomes (Richards et al., 1996; Soares et al., 2010). Compared to the Y chromosome, mtDNA variation is less geographically structured, but nevertheless

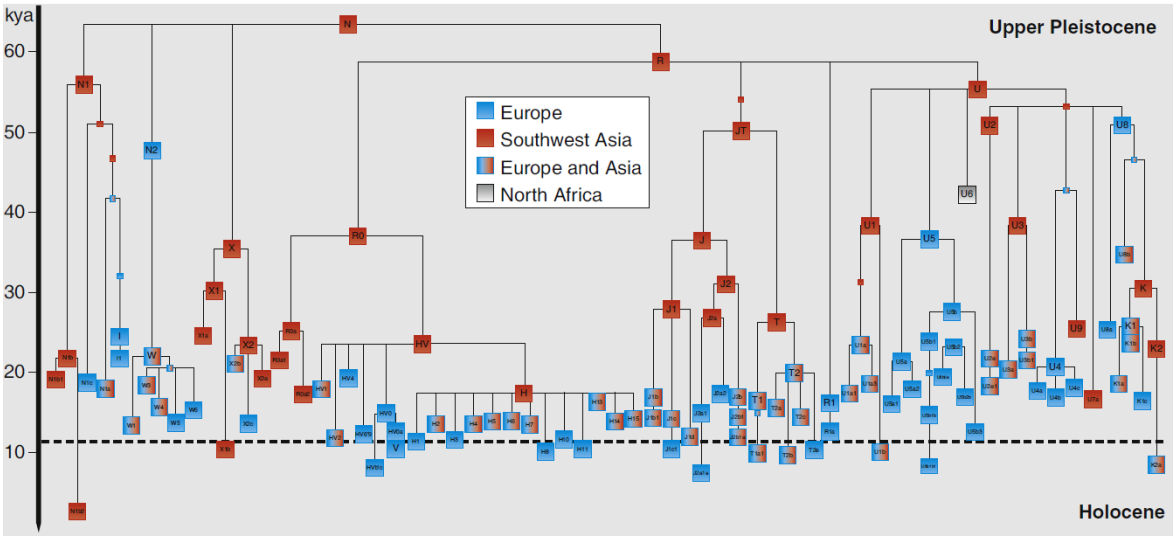

Figure 1.5. A mtDNA phylogenetic tree with revised dating of the European branches (Soares et al., 2010). **The young age of the majority of European lineages suggests that the re-colonisation of Europe after the LGM was the key event in shaping the mtDNA landscape.**

provides important clues as to the route and timing of human expansion out of Africa.

Global human mtDNA variation is represented by a phylogenetic tree, analogous to the Y chromosome, although the nomenclature system is complicated by the fact that the branches all relate to the Cambridge reference sequence, a European, such that the basal African branches are labelled L (a recent paper has sought to correct this complexity Behar et al., 2012). The system was developed first from work on Asian mtDNA (Torroni et al., 1994), with various rules that provide a basis for the incorporation of new data, which in practise means that the resultant tree has a root depicted by the letter L, and the two major non-African haplogroups evolving from the L3 branch, and all African mtDNA represented by branches L0-L3 (Torroni et al., 2006). The tree however agrees with expectations that humans evolved and spread out of Africa, with the most recent common ancestor of all human mtDNA occurring within the last 200ky (Torroni et al., 2006; Soares et al., 2010), and the most recent common ancestor of the L3 clade some 70kya (Soares et al., 2010).

European mtDNA lineages comprise a broad range of haplogroups (figure 1.5; Richards et al. (1996); Soares et al. (2010)). Recent dating of these lineages, com-

bined with an understanding of their distribution in Europe, suggests that they may have been spread from south-western Europe following the warming after the LGM (Richards et al., 1996; Finnilä et al., 2001; Torroni et al., 2001; Achilli et al., 2004; Soares et al., 2010). Increasingly, whole mitochondrial genomes are being sequenced and analysed, which increases the resolution of the geographic structure of mtDNA haplogroups. However, doubt has recently been cast of this endeavour with mtDNA variation being shown to correlate with climate, undermining the key assumption of neutrality of this locus (Balloux, 2009). Moreover, as with the Y chromosome, analyses based on single loci should be viewed with caution, and given the key role of mtDNA in metabolism, which is clearly a potential target for natural selection (Balloux et al., 2009), it is necessary to use more than just these two loci to study human evolutionary history.

Autosomal DNA

Although, technically, the study of the differences in classical markers between populations is in some sense analogous to the study of autosomal DNA, research in human evolutionary genetics has only recently begun to use data from multiple autosomal loci. Early studies on proteins investigated the outcome of genetic mutation: the differences in amino acids. However, with the completion of the first draft of the human genome almost ten years ago (Lander et al., 2001; Venter et al., 2001), there has been an explosion in our ability to find and type useful genetic markers in many people at the most fundamental level of the DNA bases themselves. A major challenge was deciding how to use this wealth of information in meaningful ways.

The Discovery of Genetic Variation and Association

The International HapMap Project is the major collaborative undertaking to develop a haplotype map of the human genome by identifying variants across multiple worldwide populations (The International HapMap Consortium, 2003, 2005, 2007, 2010).

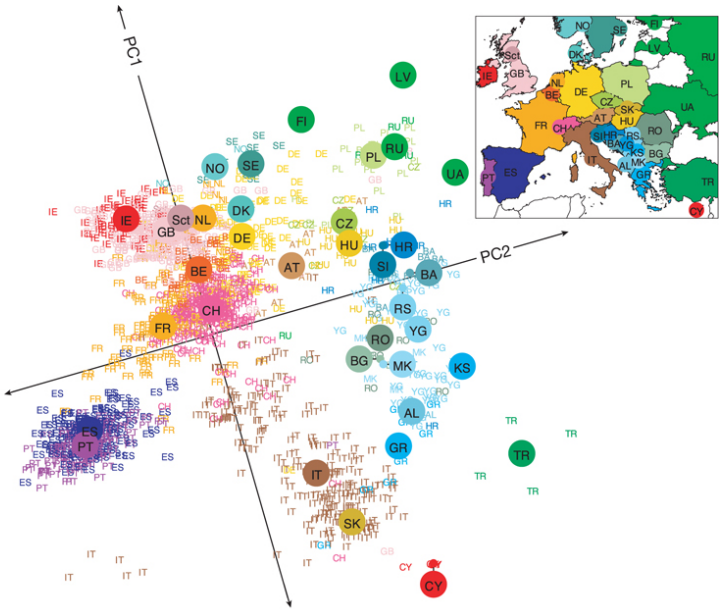

Figure 1.6. Principle Components Analysis of European populations resembles a geographic map of Europe (Novembre et al., 2008).

At the most recent count, over 3 million SNPs have been identified based on the HapMap population set of 1397 individuals from 11 geographically diverse populations (The International HapMap Consortium, 2007). This enterprise supplies the grist to the mill of population genetic researchers, most of whom are interested in the variation underlying disease, such as the Wellcome Trust Case Control Consortium (WTCCC, 2007). Genome Wide Association Studies (GWAS) take genotype information and apply it for this very use. The aim here is to type many thousands of individuals, who are either medically normal or who have a disease, in order to identify alleles that are overly represented in one group as compared to another. This approach has identified key genetic factors for several diseases (McCarthy et al., 2008). Integral to GWAS is the ability to remove the confounding effects of underlying population structure. GWAS studies typically use hundreds of thousands of SNPs spaced across the genome in an attempt to locate regions that differ between cases and controls. However, at this genomic resolution, the genetic structure of European populations has been shown to mirror geography (figure 1.6; Novembre et al. (2008); Lao et al. (2008)), so false positives may occur in a GWAS if the latent genomic structure of populations

is not suitably taken into account.

There has therefore been a significant amount of research into the genomic structure of populations, the majority of which has used as its reference the Human Genome Diversity Panel (HGDP). The introduction of the HGDP has helped endeavours to map genetic diversity on a global scale (Cann et al., 2002; Conrad et al., 2006; Jakobsson et al., 2008; Li et al., 2008). The HGDP is a set of 1064 publicly available samples collected from 52 indigenous populations across the world and represent a unique resource for geneticists interested in understanding the variation within and amongst global populations. Rosenberg and colleagues (Rosenberg et al., 2002) were among the first to use the panel to investigate global autosomal population structure through the analysis of 377 STRs. This paper also applied the program *structure* (Pritchard et al., 2000), which has proved to be a popular analytical tool for understanding population structure. *structure* is a model-based clustering algorithm that assigns individuals into a user-defined number of groups, without the use of *a priori* defined population labels (Pritchard et al., 2000). One particularly useful aspect of *structure* is that it estimates the probability of each individual of belonging to k ancestral groups and so, by looking at the distribution of these probabilities, it is possible to visualise admixture within an individual. Subsequently, our understanding of human history has become increasingly refined through the use of the HGDP together with an increasing availability of other geographically labelled populations. For example, studies have confirmed that when multiple markers are considered, linkage disequilibrium (LD) is lower in putatively older populations from Africa compared to non-African populations (Conrad et al., 2006; Jakobsson et al., 2008; Auton et al., 2009). The fine scale structure of closely related populations has also been inferred (Li et al., 2008) and there is growing evidence that geographic populations are genetically distinguishable.

PCA, a method of reducing the variability in datasets to orthogonal axes of variation, has also been applied to great effect (e.g. figure 1.6; Novembre et al. (2008); Lao et al. (2008)). PCA allows multivariate and multi-dimensional genome-scan data to be visualised and for co-varying axes of variation to be explored. If appropriate source

populations are included, it can also hint at admixture between two populations. For example Reich et al. (2009) used PCA on genome-scan data to show that Siddi Indians have African ancestry, consistent with their known historical involvement with the slave trade.

Using Genomic Data to Investigate Human History

Whilst the initial thrust of genome scan data was to investigate medical and health issues, the growing availability of data from geographically-labelled populations is allowing molecular anthropologists to approach questions of our origins at a new level of detail. Recent papers investigating human evolution and exploiting either of the twin pillars of population structure analysis, *structure* and PCA, include studies on: Jewish diaspora (Behar et al., 2010); northern Europe (Nelis et al., 2009); continental Europe (Bauchet et al., 2007); India (Reich et al., 2009); and continental Africa (Tishkoff et al., 2009).

The use of genome-wide data to investigate human history is still a nascent field. With the costs of genotyping large numbers of individuals, on a scale equivalent to Y chromosome and mtDNA studies, falling on a yearly basis, and statistical innovation also increasing, there is huge potential to uncover new and hitherto unappreciated strings to the human story.

1.3. Thesis Aims

The overall aim of this thesis is to investigate the population history of Europe through the analysis of Y chromosome and autosomal data. Central to any inferential link between genomes and history is the accurate use of a molecular clock. A corollary of this work therefore, is a critical review of this process in relation to Y chromosome data. My specific objectives are:

- To investigate the distribution and spatial variability of the main European Y chromosome haplogroup R1b1b2-M269 and its sublineages
- To explore the relationship between the choice of STR and the associated date for a haplogroup
- To investigate the fine structure of Europe with a genome-wide dataset in order to infer common ancestry across populations
- To date genome-wide historical admixture in Europe and to assess the genomic legacy of past human migrations

2. The Peopling of Europe and the Cautionary Tale of Y Chromosome Lineage R-M269

Recently, the debate on the origins of the major European Y chromosome haplogroup R1b1b2-M269 has reignited, and opinion has moved away from Paleolithic origins to the notion of a younger Neolithic spread of these chromosomes from the Near East. In this chapter I address this debate by investigating frequency patterns and diversity in the largest collection of R1b1b2-M269 chromosomes yet assembled. My analysis reveals no geographic trends in diversity, in contradiction to expectation under the Neolithic hypothesis, and suggests an alternative explanation for the apparent cline in diversity recently described. I further investigate the young, STR-based Time to the Most Recent Common Ancestor estimates proposed so far for R-M269 related lineages and find evidence for an appreciable effect of microsatellite choice on age estimates. As a consequence, the existing data and tools are insufficient to make credible estimates for the age of this haplogroup and conclusions about the timing of its origin and dispersal should be viewed with a large degree of caution.

2.1. Introduction

Since the first attempts to use biological variation in humans to aid our understanding of early human migrations, the peopling of Europe has been a major research focus (Menozzi et al., 1978; Cavalli-Sforza et al., 1994). Following the development of agriculture in the Fertile Crescent some 10kya (Diamond and Bellwood, 2003; Blockley and Pinhasi, 2011) this technology spread from the Near East westward into Europe, causing a major cultural transition from itinerant hunter-gathering, to sedentary farming which led to dramatic population growth (Gamble et al., 2005; Collard et al., 2010), during what has become known as the "Neolithic transition" (Childe, 1925, 1942; Cunliffe, 1994c). Within this archaeological framework, the debate rages about the relative contributions to modern European populations of the first people of Europe and those who migrated into it with the Neolithic transition, both in terms of their genetic legacy, and as to the processes of migration and succession (Chikhi et al., 2002; Capelli et al., 2003, 2006; Francalacci and Sanna, 2008; Battaglia et al., 2008; Gallagher et al., 2009; Rowley-Conwy, 2009; Francalacci et al., 2010). The true scenario is undoubtedly multi-faceted and complex. Both early work on ``classical markers" using principal components analysis (PCA), and more recent studies using the Y chromosome have shown that in Europe genetic variation is distributed along a southeast-northwest gradient. Such observations have been suggested to support a model of demic diffusion for the Neolithic transition in Europe, i.e. that the spread of agriculture also involved an associated movement of people from the Near East (Cavalli-Sforza et al., 1994; Rosser et al., 2000; Semino et al., 2000; Novembre et al., 2008).

New work (Balaresque et al., 2010; Morelli et al., 2010; Myres et al., 2011) has addressed the Neolithic transition in Europe, by focusing on the main western European Y chromosome haplogroup R1b1b2-M269 (hereafter referred to as R-M269). This lineage had hitherto received little recent attention in this context, although previous work suggested that the broader R-M173 clade (excluding the R1a-M17 sub-lineage)

and Haplogroup 1 (derived at SNP 92r7) are likely to have spread into Europe during the Paleolithic (Rosser et al., 2000; Semino et al., 2000; Wilson et al., 2001) and therefore unlikely to have been carried into Europe with the migrating farmers. Balaresque and colleagues (Balaresque et al., 2010) used 840 Y chromosomes within haplogroup R-M269 to show that although this haplogroup is characterised by a strong frequency cline from high in the west to low in the east, with the associated cline in haplotype diversity (measured as mean microsatellite variance) is in the opposite direction. They posited that this correlation could be explained by a more recent dispersal of this lineage from the Near East coinciding with the Neolithic transition in Europe. The lineage was estimated to be approximately 6,000 years old in various populations, which was argued to be consistent with this model. This result, as noted in their introduction "indicates that the great majority of the Y chromosomes of Europeans have their origins in the Neolithic expansion" (p2; Balaresque et al., 2010).

Myres et al. (2011) described several new SNP mutations downstream of R-M269 which show strong geographic structuring in a much larger sample of 2,043 R-M269 chromosomes. They highlight an essentially European specific clade, defined by the presence of SNPs M412 (also known as S167) and L11 (S127), which is clinal from high frequencies (>70%) in Western Europe, decreasing eastward. This study showed that the distributions of several downstream SNPs exhibit striking frequency patterns and appear to spread from different areas of highly localised frequencies some of which were also observed by Cruciani et al. (2011). Myres et al. (2011) estimated coalescence times for the R-S116 haplogroup in different populations in Europe, and suggested, in broad agreement with Balaresque et al. (2010), that the R-M269 haplogroup may have spread with the Neolithic, and more specifically with the *Linearbandkeramik* (LBK), a Neolithic agricultural industry that spread throughout northern Europe, from Hungary to France around 7,500 years ago.

The current uncertainty surrounding STR mutation rates shows that despite these recent studies, there can still be no consensus on when and where the R-M269 haplogroup originated and spread in Europe. Even if invoking the origins of the European

Y chromosome gene pool "must be viewed cautiously especially when such an argument is based on just a single incompletely resolved haplogroup" (Myres et al., 2011), it is of profound interest to try to understand how the vast majority of western European men (>100 million) carry Y chromosomes that belong to the R-M269 Y chromosome haplogroup.

Consequently, I have addressed these issues with a novel large R-M269 dataset, both on its own and in combination with compatible data from the most recent comprehensive survey (Myres et al., 2011). I show that the fundamental relationship between mean STR variance and longitude, which is the basis of the recent claim of support for the Neolithic hypothesis (Balaresque et al., 2010), does not hold for our larger and geographically broader sample. I also explain how this previous analysis may have resulted in this spurious association. Finally, I explore the spatial distribution of genetic diversity associated with the R-M269 European-specific sub-lineage, defined by SNP S127, showing an essentially homogenous background of microsatellite variation at several different sub-lineage levels, based on a common set of 10 STRs typed across over 2,000 R-M269 chromosomes.

Whilst acknowledging uncertainty, researchers usually report the age of Y-chromosome lineages based on differences between individuals across multiple STRs, often using average squared distance (ASD) or related summary statistics (Goldstein et al., 1995b; Zhivotovsky et al., 2004; Sengupta et al., 2006) as unbiased estimators of coalescence time, T. I investigated how ASD changes in our dataset based on different sets of STRs. Contrary to common belief, estimates of ASD, and therefore T, vary widely when different subsets of STRs are used with the same sample. Whilst recent evidence has increased support for the Neolithic spread of R-M269, I conclude that at the present time it is not possible to make any credible estimate of divergence time based on the sets of Y-STRs used in recent studies. Furthermore, I show that it is the properties of Y-STRs, not the number used *per se*, that appear to control the accuracy of divergence time estimates, attributes which are rarely, if ever, considered in practise.

2.2. Materials and Methods

Ethics Statement

All males sampled gave informed consent, following ethical approval by the ethics committees at the various Universities where the samples were collected.

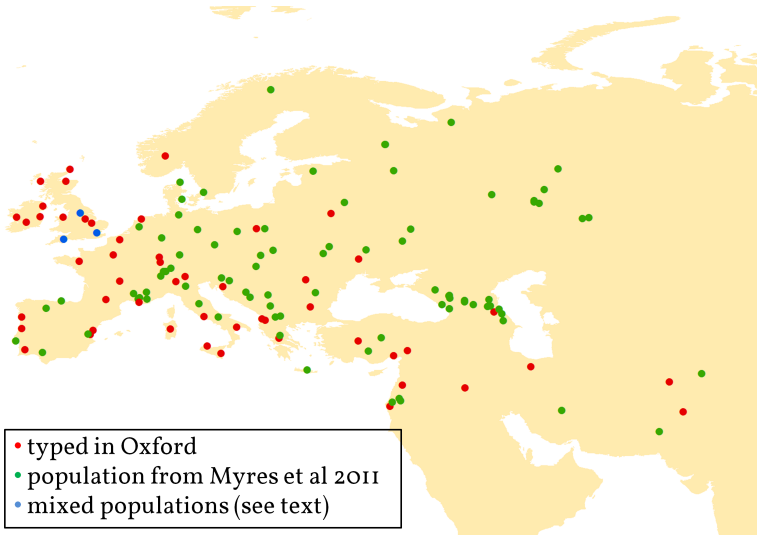

Figure 2.1. Populations used to generate frequency maps. New populations from the current study are shown in red, those taken from Myres et al. (2011) are in green. The blue dots represent populations where data was combined from the two studies.

DNA Samples and Populations

I assembled a dataset of 2,486 R-M269 Y chromosomes from across Europe, the Near East and Western Asia, from a total population of 6,503, which included both novel and previously published Y chromosomes. To assess the frequency distribution of R-M269 and various sub-haplogroups in Europe and Asia, I combined my data with that of Myres et al. (2011), which gave a combined set of 4,529 R-M269 chromosomes from a total sample of 16,298 from 172 different populations[1] (figure 2.1). Populations with

[1] Full data tables containing information on all 172 populations included in this chapter can be found on the CD attached to this thesis and online at: http://rspb.royalsocietypublishing.org/content/279/1730/884/suppl/DC1

a total size of 30 or above were used to build the frequency maps. Variance was calculated only for those populations where haplotypes were available for at least 10 individuals within the relevant haplogroup. Three English populations were generated by combining data between the two studies, where they came from the same area and the resultant population was greater than 30: North-West England (ENG-NW); South-West England (ENG-SW); and South-East England (ENG-SE). Two additional populations were made by combining populations within the Myres et al dataset: South-East Denmark (DEN-SE) and Switzerland (SWI-SC).

PCR and minisequencing primers

The frequencies of the following SNPs, whose phylogeny is shown in figure 2.2, were ascertained: S127/L11 (rs9786076), S21/U106 (rs16981293), S116 (rs34276300), S145/M529 (rs11799226), and S28/U152 (rs1236440). The following PCR primers were used: M269: F-5`-GTG GAT TCT GTT ACA TGG TAT CAC AA-3` and R-5`-TCC AAG GTG CTG GGA TTA CAC-3`; S127: F-5`-CCT GTG GGC ATT TGT AAG AGA-3` and R-5`-CCT GGA GAG AGC AAG GAT TG-3`; S21: F- 5`-CCA CAT GCT GTC CTC TCT CA-3` and R-5`-GGG GAA GGC AGG TAT TCA GA-3`; S116: F- 5- GAT GCA ATG GGA TAA CAG TCA G-3`and R- 5`TGA CTT CAG ATC CAG TGC TCA T-3`; S145: F- 5`-AAA CCC TCC TCA GCA ACA GT-3` and R-5`-AAT TTA TTA GTC TTG ATT CCC AAT TTT-3`; S28: F-5`-GAA ACA TTC CAC GCT TGA GG-3` and R-5`-AAT GGT AGT TTA ATG GGA GTA GCA-3`.

The SNaPshot® Multiplex System (Life Technologies Corp., Carlsbad, CA, USA) primer extension protocol was then used with the following minisequencing primers: for M269: 5`-GGA ATG ATC AGG GTT TGG TTA AT-3`; for S127: 5`-$(T)_{21}$AAC AGA CAG AAC CAA AAG TTC TTC-3`; for S21 5`-$(T)_{26}$CAG AGA AGA AGC AAT TGA ACC C-3`; for S116 5`-$(T)_{31}$GCT AAT GTA TCT GCT GCA CTG-3`; for S145 5`-$(T)_{33}$ATA ACA ACC GCT CTC TCA GAC A-3`;and for S28: 5`-$(GACT)_4$ TAC ATT ACT TTG AGA AGT ATG G-3`. M269 was typed as ancestral if T was present or derived if a C was present; in-

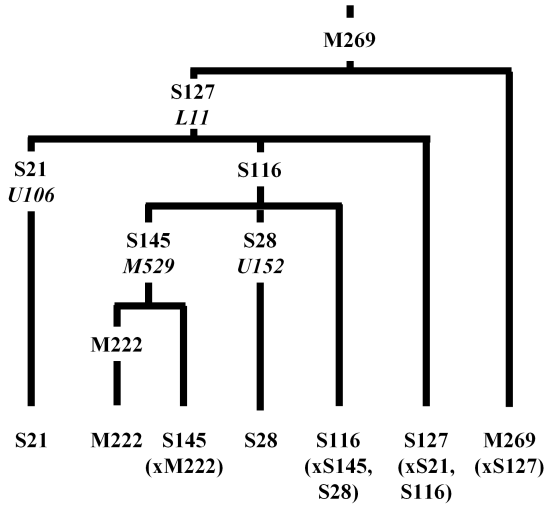

Figure 2.2. Y chromosome tree showing the relationship between the SNPs downstream of R-M269 typed in this study. Alternative nomenclature is shown for some SNPs in italics.

dividuals were ancestral at S127 (reverse minisequencing primer) if an A was present and derived if G was present; S21 (reverse minisequencing primer) G ancestral, A is derived; S116 is C when ancestral and A when derived; S145 C is ancestral, G is derived; M222 (reverse minisequencing primer) C is ancestral T is derived; and S28 (also reverse minisequencing primer) C is ancestral, T is derived.

For the majority of the individuals typed in this study (2,289) the following 10 STRs were available: DYS19; DYS389I; DYS389b (subtracting the alleles scored at DYS389I from the DYS389II locus); DYS390; DYS391; DYS392; DYS393; DYS437; DYS438; and DYS439, either being previously published or having been typed in Oxford using the Yfiler kit (Life Technologies Corp. , Carlsbad, CA, USA; Mulero et al. (2006)) or the Promega Powerplex assay (Promega Corp. , Madison, WI, USA; Krenke et al. (2005)). For the samples I used that had previously been published inWeale et al. (2002), only 5 STRs were available and so the remaining 5 were typed with an internally designed and verified multiplex using primers from Butler et al. (2002) for DYS391, DYS437, DYS389I and II and DYS439, and primers from Gusmao and Alves (2005) for DYS438. DYS391 calls were used to check for consistency with the original Weale et al haplo-

types. Three of the Weale et al populations were not typed further for these STRs (114 individuals). Individuals typed using the Yfiler kit (1,035) were used to investigate the effect of STR selection on ASD calculations.

Analysis

Maps of SNP frequencies were displayed using ArcMap GIS (version 9.2; ESRI). Interpolation was performed using the Inverse Distance Weighting procedure. Latitudes and longitudes for all populations were based on the highest resolution sampling centre associated with the samples.[2]

The *R* statistical package (R Development Core Team, 2011) was used to calculate the median STR variance (the variance in the number of repeats within a locus averaged across all loci) between all individuals within a population following 1,000 bootstrap replicates with replacement over individuals to identify confidence intervals. Regression analysis was performed in R to compare average STR variance with latitude and longitude for the R-M269, R-M269(xS127) and R-S127 haplogroups.

I investigated how ASD estimates change within our sample when using different combinations of STRs based on two separate criteria: mutation rate, μ, and observed linearity, $\vartheta(R)$ (table 2.1). I used the observed μ calculated recently (Ballantyne et al., 2010) to rank the fifteen STRs on a scale of speed, and separately calculated ASD based on the seven fastest, and seven slowest rates. My second criterion was based on the estimated duration of linearity, D`, of different groups of STRs. Duration of linearity is an estimate of the divergence time after which ASD ceases to increase linearly with time. For STRs mutating under a strict stepwise model, Goldstein and colleagues showed that ASD initially increases linearly with time, but that this linearity is constrained by the maximum number of repeats an STR can take, R (Goldstein et al., 1995b). D` is approximated using $\vartheta(R)$ (which is a simple transformation of R) and

[2] Full information on the samples can be found on the CD attached to this thesis and online at: http://rspb.royalsocietypublishing.org/content/279/1730/884/suppl/DC1

Table 2.1. 15 Y-STRs with mutation rates, range of allele and estimation of the duration of linearity shown. All STRs investigated in this study are shown with their mutation rates, μ, estimated from Ballantyne et al. (2010),and range of alleles observed in a worldwide dataset (YHRD: Willuweit and Roewer (2007)).$\vartheta(R)/2\mu$ is a proxy for D' (linearity; see text).

Y-STR	μ (95%CI)	R	$\vartheta(R)/2\mu$
DYS448	3.94×10^{-4} (1.41×10^{-5} - 2.11×10^{-3})	11	25,381
DYS392	9.7×10^{-4} (1.43×10^{-4} - 3.23×10^{-3})	15	19,244
DYS438	9.56×10^{-4} (1.37×10^{-4} - 3.18×10^{-3})	12	12,465
DYS390	1.52×10^{-3} (3.85×10^{-4} - 4.09×10^{-3})	13	9,211
DYS393	2.11×10^{-3} (6.21×10^{-3} - 5.0×10^{-3})	12	5,648
DYS439	3.84×10^{-3} (1.63×10^{-3} - 7.54×10^{-3})	15	4,861
DYS437	1.53×10^{-3} (3.54×10^{-4} - 4.1×10^{-3})	9	4,357
DYS635	3.85×10^{-3} (1.63×10^{-3} - 7.55×10^{-3})	14	4,221
DYS456	4.94×10^{-3} (2.35×10^{-3} - 8.97×10^{-3})	14	3,289
DYS389II	3.83×10^{-3} (1.61×10^{-3} - 7.49×10^{-3})	12	3,111
DYS391	3.23×10^{-3} (1.26×10^{-3} - 6.65×10^{-3})	10	2,554
DYS458	8.36×10^{-3} (4.8×10^{-3} - 1.34×10^{-2})	14	1,944
DYS19	4.37×10^{-3} (1.98×10^{-3} - 8.23×10^{-3})	10	1,888
Y-GATA-H4	3.22×10^{-3} (1.28×10^{-3} - 6.62×10^{-3})	8	1,630
DYS389I	5.51×10^{-3} (2.72×10^{-3} - 9.74×10^{-3})	8	953

μ, and the effective population size (*Ne*; equations 3 and 4 inGoldstein et al. (1995b)). Greater values of $\vartheta(R)/2\mu$ yield increased estimates of D. Using STRs with greater values of $\vartheta(R)/2\mu$ should allow linearity to be assumed further into the past, and ASD calculated from these STRs should is less likely to be underestimated as a result of saturation. Table A.1 on page 179 shows the different groups of STRs used and associated values of μ, R, $\vartheta(R)/2\mu$ and ASD.

2.3. Results

To investigate the origins of the R-M269 lineage in Europe, I analysed a large dataset of 4,529 R-M269 chromosomes (2,486 of which have not previously been published at such detailed resolution) from several populations across Europe, the Near East and Western Asia (figure 2.1). Within Europe, I observed a northwest-southeast frequency cline for R-M269, similar to those seen previously, (Capelli et al., 2006; Balaresque

et al., 2010) from high frequencies in Western Europe to lower frequencies in the east.

Within haplogroup R-M269 I genotyped a newly characterised SNP, S127 (equivalent to L11), whose distribution in Europe and the Near East, together with that of R-M269 and R-M269(xS127) is shown in figure 2.3. The distributions of R-M269 and R-S127 are broadly overlapping, but the frequency of R-S127 drops off around the Balkans, reaching extremely low values further to the east and outside of Europe. Conversely, R-M269(xS127) shows higher frequencies in eastern populations. Frequency maps showing three geographically-localised R-S127 sub-haplogroups, R-S21, R-S145 and R-S28, are shown in figure 2.4.

I next calculated STR diversity for each population for the whole R-M269 lineage, and for the R-S127 and R-M269(xS127) sub-haplogroups and investigated the relationship between average STR variance and longitude and latitude in exactly the same fashion as Balaresque et al. (2010). I provide estimates of uncertainty for these values by bootstrapping over individuals and report the median of the observed variance values and its 95% confidence interval (figure 2.3). I normalised latitude and longitude and performed a linear regression between these values and the median microsatellite variance for the three R-M269 sub-haplogroups. There was no correlation with latitude (figure B.1), and contrary to Balaresque et al, we did not find any significant correlation between longitude and variance for any haplogroup.

Relationship between ASD and linearity

Microsatellite-based-ASD has been shown to increase linearly with time (Goldstein et al., 1995b) and has been used as an unbiased estimator of mean coalescence time, given that it approximates to $2\mu T$ (μ: average mutation rate per generation; T: average coalescent time in generations: Zhivotovsky et al., 2004; Sengupta et al., 2006; Myres et al., 2011) . It would be expected that using different sets of STRs should not dramatically alter the estimation of T: as μ changes, ASD should similarly change, with

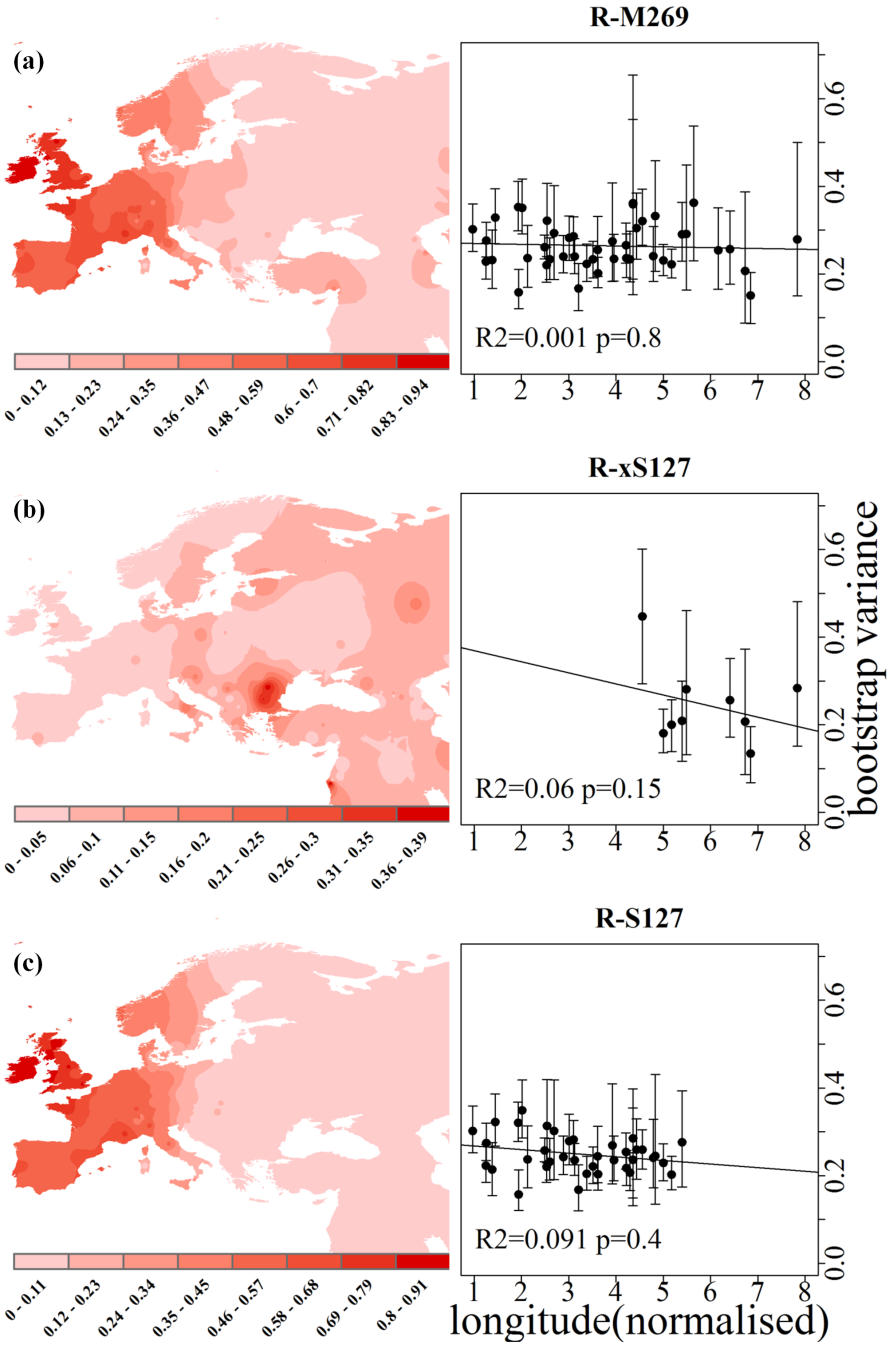

Figure 2.3. Frequency distributions and variation of Y chromosome haplogroups R-M269, R-S127,R-M269(xS127) in Europe. The three panels show contour maps based on the frequencies of the different haplogroups found across Europe and Western Asia: **(a)** R-M269 **(b)** R-M269(xS127), and **(c)** R-S127. Keys relating the colour contours to SNP frequency are shown underneath each map. The maps on the left are based on the frequencies of the SNPs in all populations marked on the map. The plots on the right hand side show the relationship between longitude and bootstrap variance based on 10 STRs for all populations with at least 10 individuals carrying that SNP. The R^2 and associated p-values are shown for the correlations in the plots.

T staying constant. Table 2.1 shows estimates of the duration of linearity based on observed mutation rates estimated recently (Ballantyne et al., 2010) and range estimated from the YHRD (Y Haplogroup Reference Database: Willuweit and Roewer, 2007). The ASD for R-S127 was calculated by comparing the haplotypes based on 15-STR haplotypes that were available for individuals from two of its two major subhaplogroups, R-S21 (141 chromosomes) and R-S116 (717). Figure 2.5(a) is a plot of T (estimated as ASD/2μ) for several different sets of STRs with different characteristics (table A.1 on page 179). To further explore the correlation between T and STR selection, I calculated T in the same way described above based on chromosomes belonging to the two deepest branches of the Y chromosome phylogeny, AxA1 and B (figure 2.5(b); Batini et al. (2011)). The purpose of this analysis was to assess whether any trends observed in the ASD calculations for R-S127 were generalisable for a completely different set of haplotypes, dating a different part of the tree. As a comparison, ASD was calculated from the same STR subsets is shown for the R-S127 on the same plot.

Re-analysis of the Balaresque dataset

The Balaresque et al. (2010) dataset presents genotype data only to the resolution of SNP R-M269. Only 10% of the Anatolian Turkish R-M269 samples used in this study were R-S127 derived. Similarly, out of 91 R-M269 Y chromosomes from two populations in Turkey, Myres et al. (2011) found only 8 R-S127 derived chromosomes (9%). The majority of these samples then, potentially belong to a different major clade within R-M269, the R-M269(xS127) sub-haplogroup, whilst the remaining populations are likely to be mostly R-S127 derived. Removing these Turkish populations from the Balaresque et al. (2010) data and repeating the regression removes the significant correlation (R^2=0.23, p=0.09; figure 2.6(b)). These populations are therefore intrinsic to the significant correlation.

Balaresque et al used haplotypes downloaded from the online Ysearch database[3],

[3] www.ysearch.com

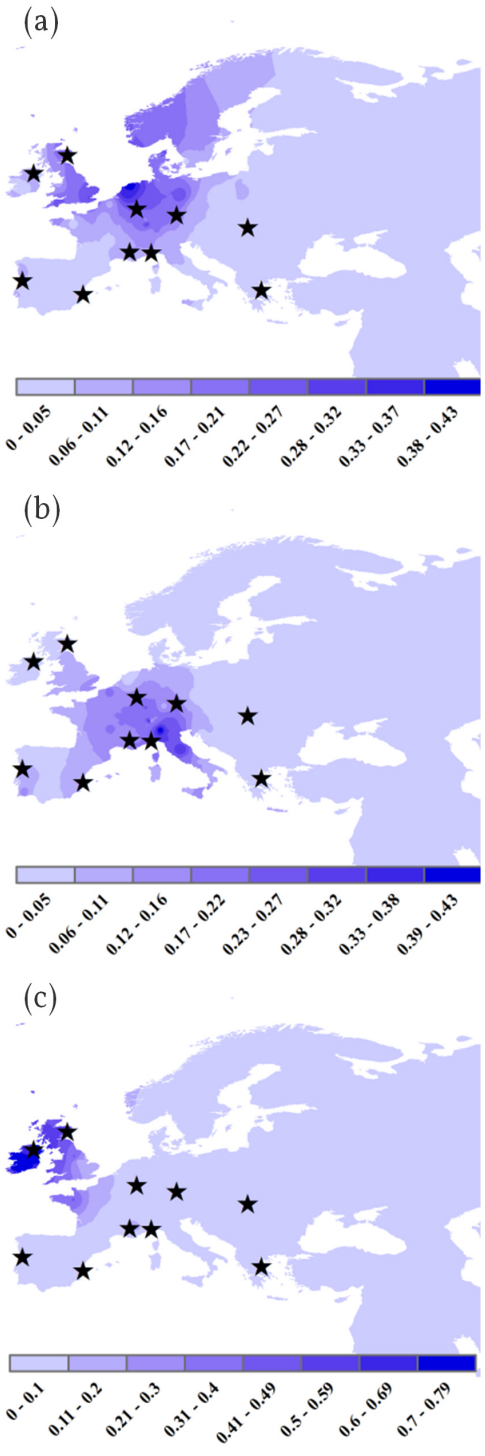

Figure 2.4. Contour maps of three R-M269 sub-haplogroups with putative centres of Neolithic expansion represented with stars (Bocquet-Appel et al., 2009) **(a)** R-S21 **(b)** R-S28; and **(c)** R-S145Keys relating the colour contours to SNP frequency are shown underneath each map.

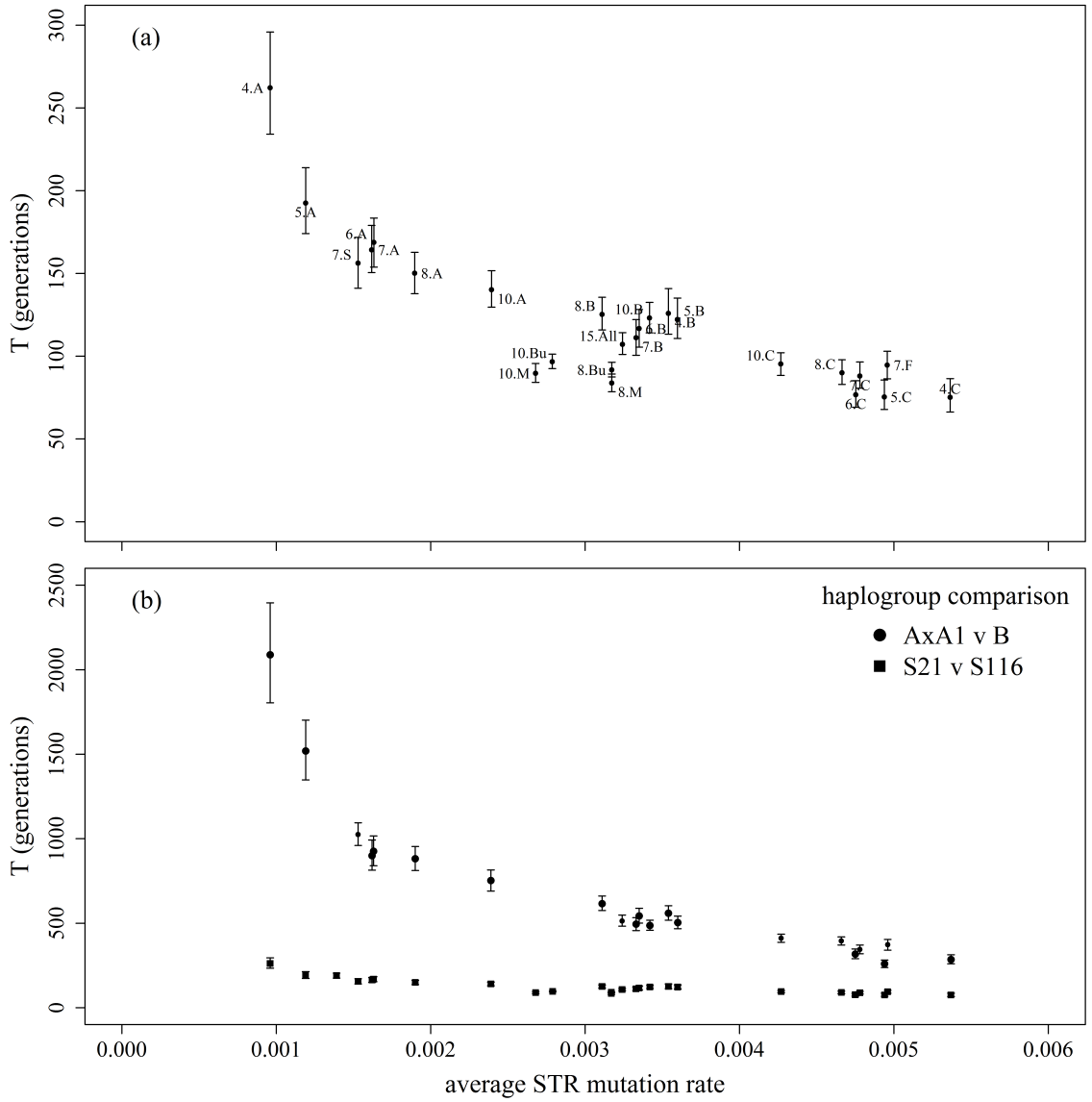

Figure 2.5. Relationship between Time to the Most Recent Common Ancestor and mutation rate for various STR subsets. Panel **(a)** shows estimated of T for the R-S127 haplogroups. Points are labelled with the subset of STRs used to calculate T. Panel **(b)** shows the same data but this time together with estimates of T based on comparisons of Y chromosome haplogroups A(xA1) and B. The labels for the STR groups are defined in table A.1.

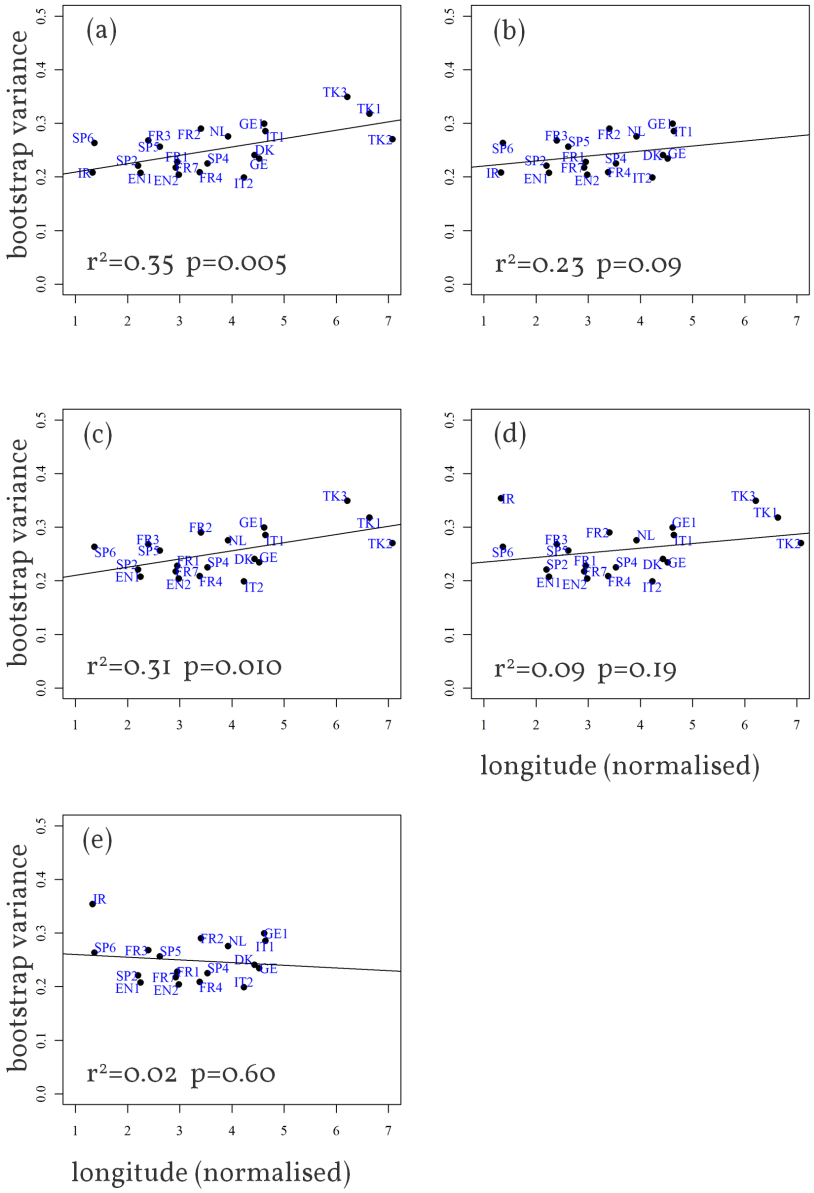

Figure 2.6. Reanalysis of the Balaresque et al. (2010) R-M269 samples. **(a)** the original correlation from Balaresque et al. (2010); **(b)** the same data as **(a)** but without the three Turkish populations; **(c)** the same data as **(a)** but without the Irish population **(d)** the Balaresque dataset with Irish variance replace with that calculated from Moore et al. (2006); **(e)** the same data as **(d)** but without the Turkish populations.

which is a repository for genetic genealogists to upload and compare haplotypes (P. Balaresque *pers. comm.*). However, 17-STR haplotypes, including the 9 STRs used in Balaresque et al's analysis, are available for 681 Irish R-M269 derived individuals in Moore et al. (2006), which is, in fact, the study which Balaresque et al. (2010) use to estimate R-M269 frequency in Ireland. A subset of the Moore et al. (2006) samples were re-analysed in the current study for SNPs downstream of R-M269, and the original haplotype data are used here to calculate variance. To test if the Ysearch haplotypes were representative of the Irish R-M269 in Moore et al. (2006), I independently re-sampled the Moore et al. (2006) dataset 10,000 times, selecting sub-samples of 75 haplotypes from which I estimated the variance using the same 9 STRs used in the Balaresque et al. (2010) paper. The median variance of these 10,000 repetitions was 0.354 with a 95% CI of (0.285-0.432). The lowest variance value out of the 10,000 samples was 0.242, which is still higher than the figure observed in the Balaresque et al. (2010) Ysearch sample (0.208). I therefore believe my current estimate of Irish R-M269 variance to be a more robust representation of the true variance than that estimated from Ysearch by Balaresque et al. (2010). Note that although the positive correlation between longitude and variance was still present after removing only the Irish and retaining the Balaresque et al. (2010) Turkish populations, if we replace the variance calculated by Balaresque et al. (2010) with that calculated from our repetitions, then the correlation is no longer significant, independent of whether or not we remove the Turkish samples (figure 2.6).

2.4. Discussion

Here I have confirmed with the broadest analysis to date, that the spatial distribution of Y chromosome haplogroup M269 can be split by R-S127 into European and Western Eurasian lineages. Contrary to the results of Balaresque et al. (2010), there is no relationship between diversity and longitude (figure 2.3) for R-M269. The presence of two sets of populations in the Balaresque paper appear to be causal to the observed rela-

tionship: the underestimated diversity of the Irish population and the inclusion of the Turkish chromosomes, the majority of which potentially belong to the non-European clade R-M269(xS127). When these elements are properly taken into account, together or on their own, the correlation no longer exists. This correlation is the central tenet of the hypothesis that R-M269 was spread with expanding Neolithic farmers.

Morelli et al. (2010) found STR motifs that split R-M269 into eastern and western lineages. I observed that 71% of the Myres et al. (2011) R-M269(xS127) chromosomes for which STR information is available have the eastern motif (DYS393-12/DYS461-10), whilst 80% of the Myres et al. (2011) R-S127 chromosomes have the western motif (DYS393-13/DYS461-11). No R-S127 chromosomes displayed the eastern motif, whilst 5% of R-M269(xS127) chromosomes displayed the western motif (all of which were either L23 (S141) or M412 (S167) derived). In both cases, however, these motifs differed to those suggested by Morelli et al. (2010) by having one less repeat at the DYS461 locus. The dichotomy observed by Morelli et al. (2010) based on a two STR motif is therefore corroborated, at least in part, by the presence of this SNP.

Dating of Y chromosome lineages is notoriously controversial (Di Giacomo et al., 2004; Zhivotovsky et al., 2004; Zhivotovsky and Underhill, 2005; Gusmão et al., 2005; Zhivotovsky et al., 2006), the major issue being that the choice of STR mutation rate can lead to age estimates that differ by a factor of three (i.e. the evolutionary (Zhivotovsky et al., 2004) versus observed (genealogical) mutation rates (Kayser et al., 2000; Ballantyne et al., 2010)). Interestingly, despite the fact that Myres et al. (2011) and Balaresque et al. (2010) used different STR mutation rates and dating approaches, their TMRCA estimates overlap: 8,590-11,950, years using a mutation rate of 6.9×10^{-4} per generation; and 4,577-9,063 years using an average mutation rate of 2.3×10^{-3}, respectively. Separately, Morelli et al. (2010) calculated the TMRCA based only on Sardinian and Anatolian chromosomes and estimated the R-M269 lineage to have originated 25,000-80,700 years ago, based on the same evolutionary mutation rate (Zhivotovsky et al., 2006) as Myres et al. (2011). This disparity is discussed in more detail in chapter 3.

In seeking to find a suitable set of STRs with which to estimate the average coalescence time, T, of sub-haplogroup R-S127, I have shown that not all STRs are of equal use in this context. I concentrated on estimating the duration of linearity, D` using different sets of STRs. The analyses suggest that the D` of an STR is key to its ability to uncover deep ancestry. Duration of linearity refers to the length of time into the past over which ASD and T continue to be linearly related for a specific STR. Goldstein et al. (1995b) showed that D` is affected by two properties of the STRs used to calculate ASD: the mutation rate and range of possible alleles that the STR can take. When I manipulated the choice of STR marker based on $\vartheta(R)/2\mu$ (a surrogate for D`), I found that different sets of STRs gave different values for T. It is clear then, that coalescence estimates explicitly depend on the STRs that one uses.

My analysis confirms that this phenomenon is not specific to the R-M269 haplogroup. Figure 2.5(b) shows that STRs with high D` produce larger estimates of T for comparisons involving haplogroups A(xA1) and B. What is clear is that estimates of T implicitly depend on the STRs that are selected to make this inference. Whilst researchers take into account STR mutation rates when estimating divergence time with ASD, commonly used STRs do not have the specific attributes that allows linearity to be assumed further into the past. The majority of haplogroup dates based on such sets of STRs may therefore have been systematically underestimated.

2.5. Conclusion

The distribution of the main R-S127 sub-haplogroups, R-S21, R-S145 and R-S28, show markedly localised concentrations (figure 2.4). If the R-M269 lineage is more recent in origin than the Neolithic expansion, then its current distribution would have to be the result of major population movements occurring since that origin. For this haplogroup to be so ubiquitous, the population carrying R-S127 would had to have displaced most of the populations present in western Europe after the Neolithic agri-

cultural transition. Alternatively, if R-S127 originated prior to the Neolithic wave of expansion, then either it was already present in most of Europe before the expansion, or the mutation occurred in the east, and was spread before or after the expansion, in which case one would expect higher diversity in the east closer to the origins of agriculture, which is not what is observed. The maps of R-S127 sub-haplogroup frequencies for R-S21, R- S145 and R-S28, show radial distributions from specific European locations (figure 2.4). These centres have high absolute frequencies: R-S21 has a frequency of 44% in Friesland, and R-S28 reaches 25% in the Alps; and in the populations where they are at the highest frequency, the vast majority of R-S127 belong to that particular sub-lineage. For example, half of all R-M269 across Southern Europe is R-S28 derived, and around 60% of R-M269 in Central Europe is R-S21 derived. At the sub-haplogroup level then, R-M269 is split into geographically localised pockets with individual R-M269 sub-haplogroups dominating, suggesting that the frequency of R-M269 across Europe could be related to the growth of multiple, geographically-specific sub-lineages that differ in different parts of Europe.

A recent analysis of radiocarbon dates of Neolithic sites across Europe (Bocquet-Appel et al., 2009) reveals that the spread of the Neolithic was by no means constant, and that several ``centres of renewed expansion" are visible across Europe, representing areas of colonisation, three of which map intriguingly closely to the centres of the sub-haplogroups foci (figure 2.4). Future work involving spatially explicit simulations, together with accurate measures of Y chromosome diversity, are needed to investigate how the current distribution of sub-haplogroups may have been produced. In this context, recent work by Sjödin and François (2011) rejected a Palaeolithic dispersion for R1b-M269 using spatial simulations based on Balaresque dataset. Nevertheless, additional work is still necessary as these authors were not aware of the limitation of the Balaresque dataset presented here and did not fully explore the impact of the different molecular characteristics of the investigated loci on their analysis.

Age estimates based on sets of Y-STRs carefully selected to possess the attributes necessary for uncovering deep ancestry (for example from the almost 200 recently char-

acterised by Ballantyne et al. (2010)), and from whole Y chromosome sequence comparisons will provide robust dates for this haplogroup in the future. For now I can offer no date for the age of R-M269 or R-S127, but believe that these STR analyses suggest the recent age estimates of R-M269 (Balaresque et al., 2010) and R-S116 (Myres et al., 2011) are likely to be younger than the true values, and the homogeneity of STR variance and distribution of sub-types across the continent are inconsistent with the hypothesis of the Neolithic diffusion of the R-M269 Y chromosome lineage.

3. An Exploration of Y Chromosome Dating

In the process of researching chapter 2 I observed that STR choice had a very definite effect on the age estimates of Y chromosome lineage R-M269. As we can see from figure 2.5, this phenomenon was not just restricted to R-M269, and a similar pattern is seen for ASD-based dates for the coalescence between Y haplogroups A(xA1) and B, which represents a deep division of the Y chromosome tree. I therefore decided to further explore Y chromosome dating with a large, recently published dataset of Y-STRs that had been typed on the HGDP sample. In this short chapter I describe the background to Y-STR dating and highlight the inconsistencies in the literature surrounding the date of Y chromosome lineage R-M269. Using the approach taken in chapter 2, I compare ASD-based dates for two nodes in the Y chromosome tree with dates generated recently from sequence data. I also show that an alternative Y chromosome dating method, BATWING, also produces dates that differ when different STRs are used. Whilst these analyses show that there is a clear problem with the current approach to dating Y chromosome lineages with STRs, at this time I am unable to offer a viable solution. However, the results presented here, in addition to those in chapter 2, highlight an important limitation of the use of these markers in human evolutionary studies, and I conclude by suggesting that future studies using Y-STRs to date lineages should proceed with caution.

3.1. Introduction

Traditionally, researchers have used a combination of Y chromosome single nucleotide polymorphisms (SNPs) to characterise lineages within populations, and microsatellites (or short tandem repeats, STRs) to estimate the diversity within these lineages (Jobling et al., 2004). This has proved a fruitful line of enquiry as genotyping many samples for SNPs and STRs has become increasingly cheap, thus enabling large population-scale studies to be undertaken. Typing novel populations for these commonly-used markers has the added benefit of allowing data from multiple studies, including those published several years apart, to be combined as and when new data are available. The literature abounds with collections of populations of Y chromosomes typed for similar markers which have been a treasure-trove for human evolutionary genetic analyses (e.g. Semino et al., 1996; Thomas et al., 1999; Capelli et al., 2001; Kayser et al., 2001; Rootsi et al., 2004; Behar et al., 2004; Moore et al., 2006; Capelli et al., 2006). Y chromosome SNPs can be used to discover the frequencies of haplogroups within populations, which are then compared across large geographical areas to make inferences about the evolutionary history of these populations. Inherent to this phylogeographic approach is the ability to quantify the amount of time a particular lineage has been present in an area, for which fast-evolving markers, STRs, are used. This approach has yielded many studies based on populations across the world and has greatly aided our understanding of human migration and history over the last 150,000 years.

However, the ease with which these data have been produced has overshadowed objective assessments about how well analytical methods should be expected to perform with these markers. For example, it is now common for investigators to type Y chromosomes with the popular Y-filer (Mulero et al., 2006) and Powerplex (Krenke et al., 2005) kits, which were designed to type the European minimal haplotype for forensic applications, and to then use these STRs to address evolutionary questions. Whilst this appears to be an efficient strategy as one can type a number of STRs quickly and in a

single reaction, almost no effort has been exercised to test whether these STRs are in fact suitable for the investigating the diversity and timing of events over evolutionary time.

Since 2004, over 200 Y chromosome STRs with varying molecular characteristics have been available to researchers (Kayser et al., 2004; Ballantyne et al., 2010). Despite Kayser and colleagues explicitly stating that the knowledge of the variability of these STRs could lead to an informed choice for the appropriate STRs for the hypothesis in question (p1194: Kayser et al., 2004; Ballantyne et al., 2010), almost all evolutionary studies use a subset of the same core ~20 STRs. In practise this makes sense, as it allows investigators to pool data from different laboratories and publications. However, given that the mutability of STRs can differ widely, it must be more appropriate to select the STRs that are most suitable to the hypothesis to be tested rather than those that can simply be typed easily. Indeed, in light of this, Ballantyne and colleagues recently published a novel set of rapidly mutating STRs that, because they will accrue mutations more quickly over time, have the potential to be of greater use to the forensic community by discriminating close relatives more easily (Ballantyne et al., 2010, 2012). Conversely, slower mutating STRs will accumulate similar numbers of mutations over longer periods of time and are therefore useful for providing information about the relationships between different populations and lineages of Y chromosomes which occur over greater timescales (Kayser et al., 2004). Thus, mutation rate is an important factor to consider when selecting appropriate loci for evolutionary studies. However, it is not the only important criterion critical to satisfactory STR choice. Loci which evolve according to a simple stepwise mutation model (S-SMM), such as STRs, can exhibit homoplasy, where identical alleles at the same locus in two different chromosomes may not be identical by descent (Kimura and Ohta, 1978). Estoup and colleagues describe how homoplasy is likely to be problematic when STR loci have not only high mutation rates, but also strong allele size constraints, and where population size is large (Estoup et al., 2002), which echoes theoretical work originally introduced by Goldstein et al. (1995a). Additionally, no regard tends to be given as to

whether an STR is structurally simple or complex. Given that structurally simple loci mutate in a simple way, fragment size is a reliable guide to allele size (Kayser et al., 2004) which is not the case for structurally complex loci where a different underlying allelic structure cannot be ruled out for fragments of similar sizes. Whilst structurally complex STRs may in fact be more informative than simple ones, this is only the case when the underlying allelic structure is known, which would then allow haplotypes with similar allele sizes to be segregated into groups dependent on this underlying structure. So, without an appreciation of the components underlying complex STRs, issues with identifying truly similar allele sizes remains. Large-scale structural rearrangements have also been observed. For example, Y chromosome STR DYS19, which was the first Y-STR to be used to date an evolutionary event (Underhill et al., 1996), can be deleted, duplicated and triplicated (Balaresque et al., 2009), suggesting that the perceived allele typed in one individual, may be a different allele in another. When investigating evolutionary questions then, it seems appropriate to discriminate against STRs that evolve quickly, that have a small range of observed alleles, or that have complex repeat structures: considerations which are rarely contemplated in practise.

Different methods: different dates

Compounding the issue of potentially poor STR choice is the observation that different Y-STR dating methods often give different dates. For example, three recent studies have all attempted to date the common western European haplogroup R-M269 (table 3.1; Balaresque et al., 2010; Morelli et al., 2010; Myres et al., 2011).

The dates in table 3.1 are the result of three separate analyses. The methods, mutation rates and data used all differ. Morelli et al. (2010) estimate the coalescence time of R-M269 using the Bayesian estimator, Bayesian Analysis of Trees With Internal Node Generation, or BATWING, and the evolutionary mutation rate (further description of this and other methods can be found below). Balaresque et al. (2010) estimate the TM-

Table 3.1. Three estimates for the coalescence time of Y chromosome haplogroup R-M269.

STR dating method	n	n STRs	μ	T (95% CI)	reference
BATWING	142	10	6.9×10^{-4}	32,600 (25,000-80,700)	Morelli et al. (2010)
BATWING	940	9	2.5×10^{-3}	6,512 (4,577-9,063)	Balaresque et al. (2010)
ASD	245	10	6.9×10^{-4}	10,270 (8,590-11,950)	Myres et al. (2011)

RCA of R-M269, using a slightly different implementation of the same method, giving an estimate that is roughly a fifth of the Morelli et al. (2010) estimate, essentially because they use a mutation rate that is just under a quarter as fast. Myres et al. (2011) use the same mutation rate as Morelli et al. (2010), but a different method and find that the TMRCA of R-M269 to be a third of the Morelli et al. (2010) estimate. Viewed together these results show that TMRCAs calculated in this way are highly dependent on both STR and mutation rate that is used. In this example then, it is difficult to conclude what the true age of Y chromosome lineage R-M269 in Europe is.

Mutation rate controversy

STR mutation rates have long been known to vary for different STRs and have been estimated through the observation of mutations across many father-son pairs (Heyer et al., 1997; Kayser et al., 2000, 2001; Gusmão et al., 2005; Onofri et al., 2009; Ballantyne et al., 2010). These *genealogical mutation rate* estimates rely on a small number of mutations being observed across a large number of meioses. Another way to estimate mutation rates is to count the mutations present on the branches of a haplotype tree or network in a population with a known time of origin. For example, Forster and colleagues, assuming an age of 20,000 years for the origin of Native American Y chromosome defined by mutation DYS199 T, used this approach with slowly evolving STRs to estimate an *evolutionary mutation rate* of 2.6×10^{-4} per 20 year generation, an order of

magnitude slower than the average genealogically estimated rate of 2.5 x 10^{-3} (Forster et al., 2000). The authors argue that the discrepancy is likely due to the average genealogical mutation rate being systematically over-estimated due to the presence of fast evolving STRs in its calculation, and that the genealogical mutation rate, estimated at that time in a modern German population, was based on a longer generation time (>30 years) than was likely over evolutionary time (Forster et al., 2000). Developing this concept, Zhivotovsky et al. (2004) produced another evolutionary mutation rate, termed the *effective mutation rate*, estimated this time by using STR variation within SNP-defined haplogroups in populations with well documented short-term history (e.g. Bulgarian gypsies carrying the M82 `Indian' Y chromosome SNP), and arrived at an even slower rate of 6.9 x 10^{-4} per 25 year generation, which is 3.5 times slower than the equivalent genealogical estimate. From its publication this rate has never been universally accepted (Di Giacomo et al., 2004; Luca et al., 2005; Zhivotovsky and Underhill, 2005; Zhivotovsky et al., 2006), and although Zhivotovsky and colleagues presented locus-specific mutation rate estimates for ten STRs (note that the standard deviation around the 6.9 x 10^{-4} estimate is 0.57 x 10^{-3}) researchers have since had the option of two different rates, often using both, the choice of which can essentially cover any historical or ancient scenario (e.g.table 3.1). This variety of rates does little to add credibility to any lineage-based date estimated from Y chromosome STRs.

Dating Y chromosomes

To perform the "gold standard" method for genetic dating, one compares several sequences of DNA of known length and counts the number of mutations observed along the sequence. Then, using either the average neutral mutation rate, 1.1-3.0 x 10^{-9} per base per generation (Nachman and Crowell, 2000; Kondrashov, 2003) or the sequence specific mutation rate (e.g. for the Y chromosome: Thomson et al., 2000; Xue et al., 2009), one can calculate the TMRCA in generations of the sequences in question by finding the product of the proportion of mutations along the sequence and the muta-

tion rate. In an attempt to date the root of the Y chromosome tree, Cruciani and colleagues recently used this approach to date 200kb of DNA sequence from 7 Y chromosomes spread across the Y chromosome tree dating the whole tree to 142kya (Cruciani et al., 2011). However, due to the high cost of sequencing large stretches of DNA in multiple individuals, Y-STRs are usually employed to date divergence times in evolutionary studies, particularly as large datasets of common STRs from different populations are available. Several methods use STR variation to estimate the TMRCA within populations or between lineages, for example by using ASD (Goldstein et al., 1995a), reduced-median networks and the rho statistic (Bandelt et al., 1999), and BATWING (Wilson et al., 2003). Using microsatellite data and coalescent modelling, Cox showed that the rho statistic is largely confounded by the demographic history of a population, and inaccurate dates are found when populations have undergone a change in size, or undergone an extreme bottleneck or constricted founder effect (Cox, 2007). A popular method, BATWING, employs *a priori* defined distributions of historical demographic parameters, such as population size and expansion rates, together with the Bayesian paradigm to assess what the true values of these parameters are and to find the TMRCA, given the data (Wilson et al., 2003). Consequently, although broad flat prior distributions can be given to the input parameters, the results obtained will always depend on assumptions made about the (generally unknown) demographic history of a population. Conversely, ASD has been shown to be unbiased to demographic history (Goldstein et al., 1995a,b). This metric of genetic differentiation increases linearly with time and estimates of T can be estimated by the simple relationship in equation (3.1).

$$T = ASD \times \frac{1}{2\mu} \tag{3.1}$$

Goldstein et al. (1995a) also showed that the duration of time into the past over which this relationship remains linear, D`, can be estimated from the mutation rate, μ, of the STR and the range of alleles which are observed in the population, R: equation (3.2).

Using equation (3.2) then, one can assess the estimated length of time over which an STR should theoretically remain linear and choose appropriate STRs to the question that they hope to answer.

$$D` = \frac{R^2 - 1}{6} \times \frac{1}{2\mu} - (2N + 1) \tag{3.2}$$

The aim of this chapter is to demonstrate the effect of STR choice on Y chromosome dating. To achieve this I use ASD, calculated between different nodes in the Y chromosome tree, to generate estimates of T based on different groups of STRs. As a comparison, I run BATWING on a single population, with the same groups of STRS, and show that an appreciation of the the molecular characteristics of Y-STRs is an integral part of using STRs as molecular clocks.

3.2. Materials and Methods

Samples

I used recently published data from the Human Genome Diversity Project (HGDP) for which Y-SNP and 76 Y-STRs are available for 671 males from 52 worldwide populations(Shi et al., 2010).[1] Individuals were assigned to specific Y chromosome lineages A-R. I calculated ASD by comparing lineages beneath a node of interest (as in chapter 2 and Busby et al. (2012)). Figure 3.1 shows the outline of the Y chromosome tree with the main Y haplogroups marked as well as the number of individuals belonging to that haplogroup in the HGDP dataset.

The HGDP dataset contained several STRs where two alleles were observed in some individuals. The method I used to calculate ASD requires that each individual has a

[1] HGDP data downloaded from the Centre Études Polymorphisms Historiques website, http://www.cephb.fr/en/hgdp/ on 23/06/2011

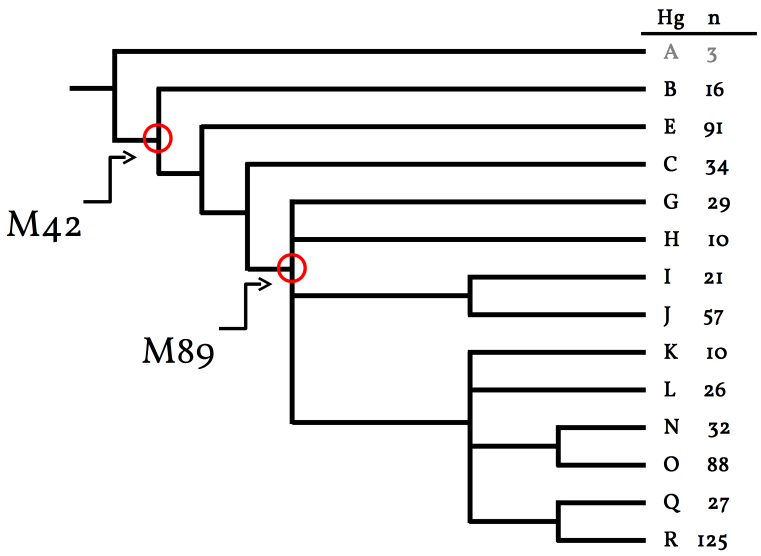

Figure 3.1. Outline of the Y chromosome tree indicating the relationships between the major haplogroups. Two nodes, named for the SNPs that define them, are shown: M42, which separates haplogroup A and B from the rest of the tree; and M89, which separates haplogroup A,B,E and C from the lower groups in the tree.

single integer allele value at each locus. Because it is not possible to identify which of the duplicated alleles is equivalent across individuals, I removed eight STRs immediately from the dataset where duplicated alleles were present in all individuals. These were: DYF387 S1; DYF399 S1; DYF403 S1; DYF404 S1; DYS385; DYS389; DYS472; DYS526. For three other STRs, DYS525, DYS589 and DYS636, mutation rates that had been calculated in the same way across all STRs were not available and so these STRs were also dropped. Any half-alleles were coded as missing data as these also would lead to complications with estimating ASD. Note that the removal of these loci will cause an underestimation in the ASD that would be calculated, as the differences between individuals at these haplotypes will be smaller than the truth. I then removed 136 individuals with more than 5% missing data across the remaining 65 STRs. All of the remaining 531 individuals were included in the H952 HGDP dataset (Rosenberg, 2006) and are unrelated. This final dataset is presented in table A.2 on page 185.

STR mutation rates

To effectively use STRs as molecular clocks, accurate mutation rates must be known. STR mutation rates can be calculated by genotyping many father-son pairs for a given STR and counting the number of times a mutation is observed across all observed meioses in the father-son pairs. Ballantyne et al. (2010) calculated mutation rates using a hierarchical Bayesian binomial model on data collected in this way for almost 200 STRs including most of those present in the current dataset. Their method allows for positive mutation rates to be estimated in the absence of an observed mutation by assuming that the mutation rate for each individual STR can be considered as a realisation of the mutation underlying any STR (Ballantyne et al., 2010). These rate estimates were chosen as they include 65 out of the 67 STRs available in the HGDP dataset, thus allowing the vast majority of the HGDP data to be used with similarly calculated mutation rates, as well as allowing mutation rates to be estimated for STRs even when no mutations were observed in father-son pairs (table A.2).

ASD and T

For the two nodes highlighted in figure 3.1, I calculated ASD for each STR separately using equation (3.3) (Goldstein et al., 1995a; Sun et al., 2009). For example, to calculate ASD for the M42 node (marked in figure 3.1), for each locus separately, I calculated the squared distance between each individual within haplogroup B with each individual in haplogroup E and then computed the average of these squared distances. Bootstrapped estimates of the 95% confidence interval associated with this value of ASD were estimated by repeating the ASD calculation 1000 times with replicate lineages taken from the originals with replacement. In the same way, I then calculated ASD and confidence intervals for each individual in haplogroup B with each individual in haplogroup C. This was repeated for the two nodes shown in figure 3.1.

$$ASD = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (A_i - B_j)^2 \times \frac{1}{n_A \times n_B} \tag{3.3}$$

The average ASD per STR was calculated from the average of these comparisons, giving a final value for the ASD per STR for each node. For each node ASD was converted into T using equation 3.1 and the mutation rate for the STR for which the ASD had been calculated. In equation 3.1, n_A is the number of individuals in population A, n_B is the number of individuals in population B. A_i is the value of the *i*th allele in populations A, and B_j is the value of the *j*th allele in population B.Thus, for each node, 65 realisations of T were estimated, based on each of the 65 STRs that were used.

Duration of linearity

Knowledge of the range of possible alleles an STR takes and its mutation rate allows for estimates of the time over which ASD remains linear with time, D`, to be estimated: equation (3.2) and Goldstein et al. (1995a). I was interested in investigating how the relationship between values of T, estimated from ASD, differed depending on the values of D`, estimated from equation 2. To calculate D`, I used the Ballantyne et al. (2010) mutation rates and estimated range from the current dataset. The HGDP populations were chosen to represent worldwide diversity, so using these data for R should give a reasonably accurate, if conservative, estimate of this parameter. Table A.2 on page 185 shows the estimated length of linearity, D`, for each of the 65 STRs used in the analysis.

ASD Analysis

Previous studies using ASD to estimate T have tended to use an average value of T calculated across the STRs available. I therefore selected several groups of STRs arranged by different criteria: by mutation rate, making a group of 20 fast and 20 slow STRs; by

D`, with a groups of 20 STRs with the longest D` and 20 STRs with the shortest D`. I also used found the average T based on all 65 STRs as well as the average T for the Y-filer and PowerPlex groups of STRs. All analyses were performed using homemade scripts written with the *R* statistical package (R Development Core Team, 2011).

BATWING

I used BATWING (Wilson et al., 2003) to assess how estimates of TMRCA differed using this method. This method applies a Markov Chain Monte Carlo (MCMC) procedure to generate a series of genealogical trees with associated demographic parameter values consistent with the data (Shi et al., 2010). Posterior estimates of the parameter estimates, including TMRCA and confidence intervals are produced. For this analysis, I was only interested in observing how the relative dates changed using different sets of STRs, rather than finding the true value of the T for a particular population. Whilst BATWING has been used to date lineages in the past (for example, see the estimates in table 3.1), the method was originally developed to date the coalescence time of haplotypes in a population. I therefore chose the largest HGDP population for which data were available, the Bedouin (24 males), and ran BATWING with a constant population size for the same sets of STRs used to calculate ASD. I used individual gamma priors based on observed mutation rates, as in Shi et al. (2010), together with a broad initial population size prior [gamma(1,0.0001)] and a uniform theta prior [uniform(0,100)]. STRs were chosen on the basis of the estimated D` as calculated using equation (3.2). Following the approach of Shi et al. (2010), I ran BATWING for $1x10^6$ generations, sampling every 200 generations with a random seed.

3.3. Results

Different STRs give different dates

To mimic traditional multi-STR based estimates of T, I first selected groups of STRs based on their mutation rate and D`. I averaged T across four groups of 20 STRs made up of the twenty fastest and slowest STRs and the twenty STRs with the highest expected D` and 20 with the lowest D`. I also made groups of STRs containing the 13 Y-filer STRs and 8 PowerplexY STRs present in the dataset, as well as an average T based on all 65 STRs (figure 3.2).

As an alternative method to ASD, I used BATWING to investigate whether a model-based technique of finding T behaves in a similar way to ASD. Figure 3.3 shows the results of a simple BATWING run on the HGDP Bedouin using the same sets of STRs for which ASD was calculated. These STRs were chosen to represent the extreme ends of the mutation rate and linearity spectra.

3.4. Discussion

Using two different date estimation methods, I have shown that STR choice fundamentally effects the estimated coalescence time of lineage-based and population-based analyses. This is not a trivial observation. In the majority of Y chromosome studies that have been published over the last 15 years, STRs have generally been used to produce some estimate of diversity or variation, and this has tended to be linked to time, in some way. Of course, it is of deep interest to try to use the tools available to provide meaningful estimates of the time that a population has inhabited an area, and this will continue to be a fruitful line of enquiry in human evolutionary studies (Jobling, 2012). However, the analysis provided here shows that not every tool is up to the job, and that future work in this field should take into account the molecular

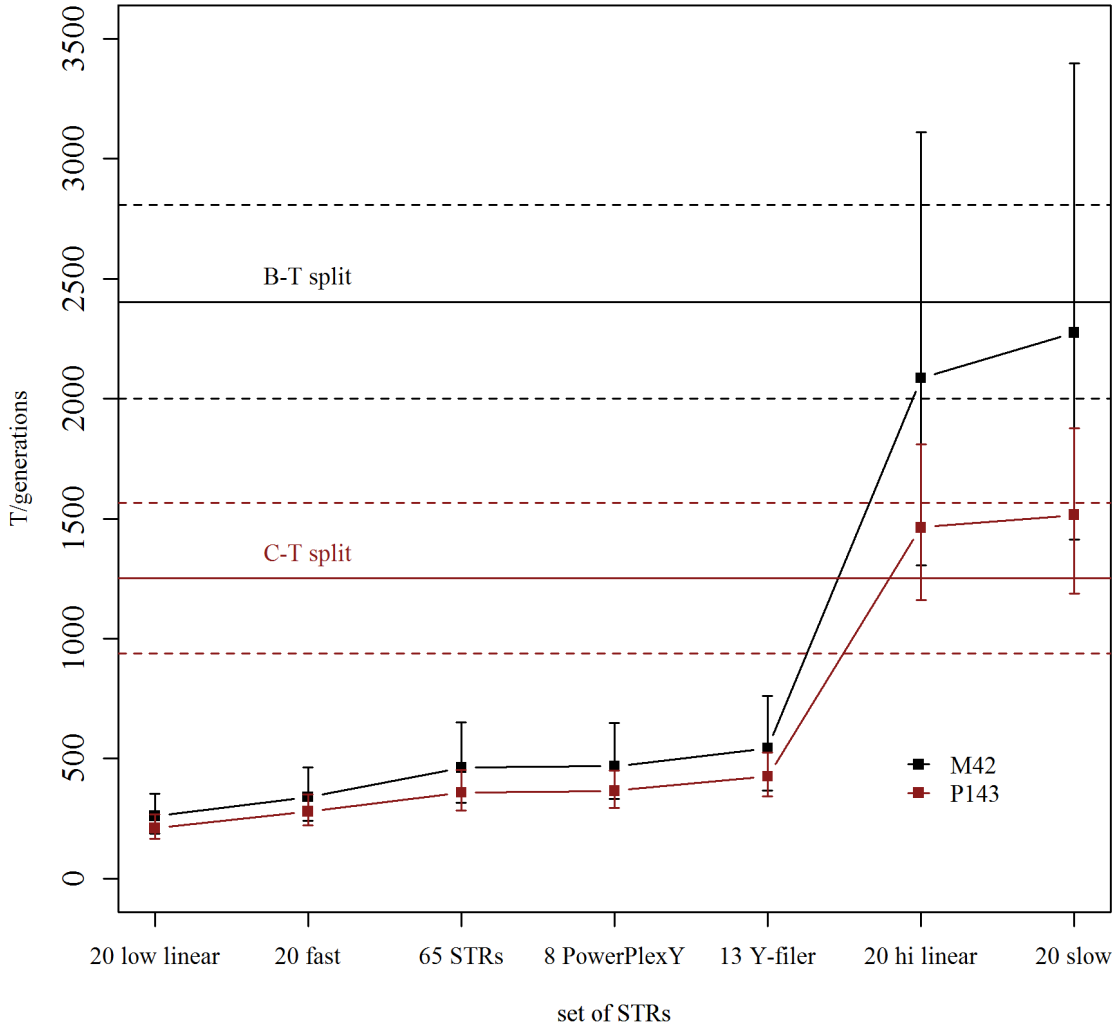

Figure 3.2. ASD-based estimates of T based on different sets of STRs. The plot shows the average T with 95% confidence intervals based on different sets of STRs. For reference, recent estimates from sequence data of divergence times (Cruciani et al., 2011) for splits in the Y chromosome tree equivalent to the nodes used in the present analysis are shown as horizontal lines.

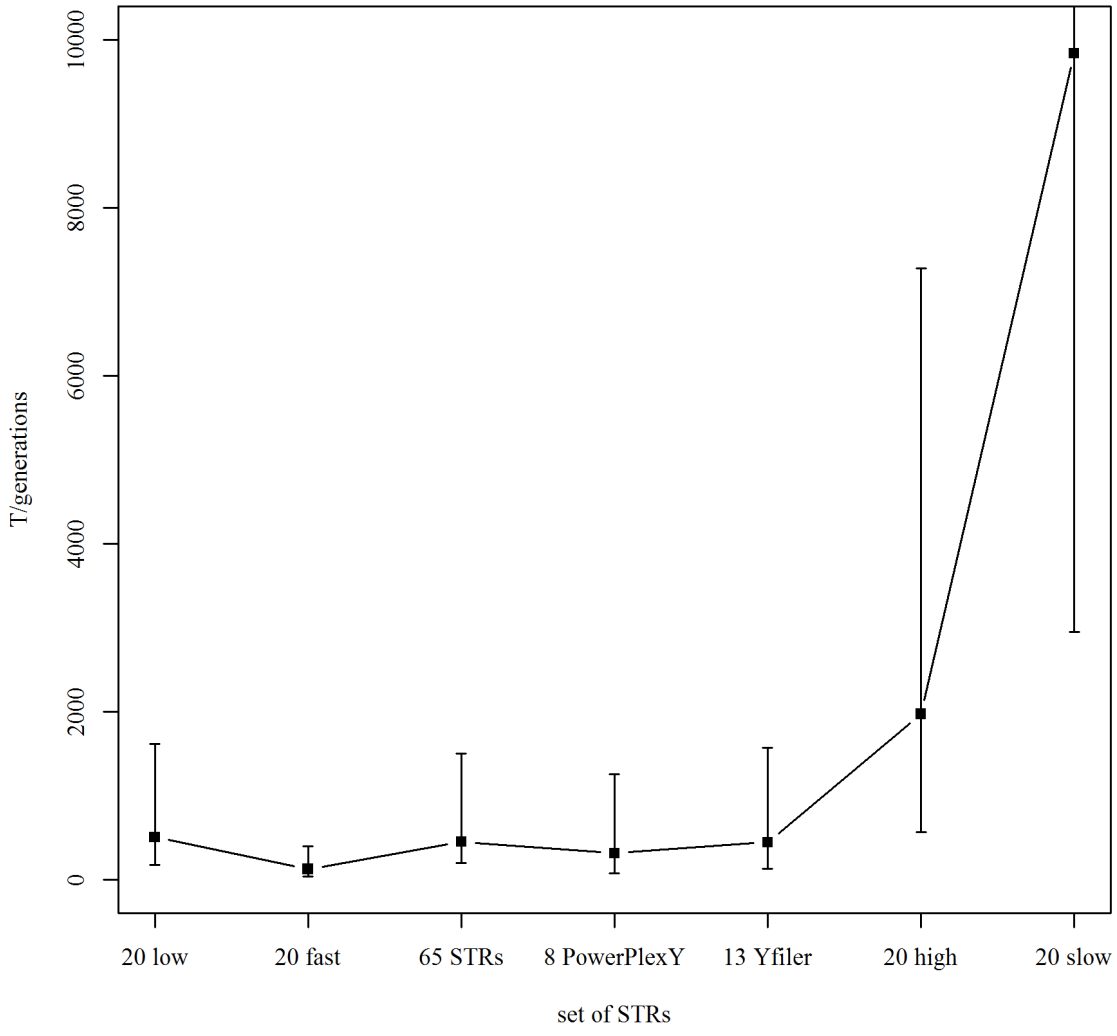

Figure 3.3. Seven runs of *BATWING* to date the origins of the HGDP Bedouin population using the same groups of STRs as the ASD experiment. The sets of STRs were chosen to cover a range of average mutation rates and linearities. The black points represent the values of T with 95% confidence intervals.

characteristics of STRs, such as their mutation rate and linearity, if confidence in the conclusions of these studies is sought.

4. The Genome-wide Fine Structure of Europe

The history of people in Europe has long been studied through successive developments in western archaeology, ancient history and most recently, population genetics. Whilst early genetic analyses helped to elucidate broad patterns of variation, they failed to provide sufficient power to test the support for different models of human evolutionary history. For example, protein polymorphisms (Cavalli-Sforza et al., 1994) and some Y chromosome haplogroups (e.g. Rosser et al., 2000; Semino et al., 2000; Di Giacomo et al., 2004; Adams et al., 2008; Battaglia et al., 2008; Balaresque et al., 2010; Busby et al., 2012) show clinal distributions from east to west Europe, but the *processes* behind these patterns are unclear. These studies have insufficient power to segregate competing hypotheses, ostensibly because the relatively small number of loci used to make these observations leads to low statistical power when addressing the timing of dispersal and mixing events. This chapter and the next address the peopling of Europe from a genome-wide perspective. Studies that have used autosomal SNP data have tended to introduce vast datasets of up to a million markers, only to reduce the number actually used to around 200,000 to limit the effect of linkage disequilibrium (LD). However, LD holds a lot of information about the evolutionary history of populations, so it would appear more appropriate to try to use this information. By using haplotypes, as opposed to genotypes, the following two chapters aim to glean further insight into the population history of Europe from dense genotype data.

4.1. Introduction

The recent advent and decreasing cost of mass sequencing technologies have allowed researchers the platform to genotype many thousands of polymorphic markers in hundreds or thousands of individuals. By providing data on genetic variation across the genome, it is possible to produce a detailed picture of the average ancestry across the genomes of *individuals* within and between populations, and therefore gain a deeper, more nuanced understanding of human evolutionary history from a whole genome perspective. This is fundamentally different from most previous evolutionary studies using uniparental markers which necessarily only provide information on the ancestry of a *population* through the assessment of the number and identity of different lineages within that population.

The availability of public single nucleotide polymorphism (SNP) array datasets from populations from around the world has led to a surge of research exploring human evolutionary history from a genome-wide perspective. In particular, the Human Genome Diversity Panel, (HGDP; Rosenberg et al., 2002), and the Population Research database, (POPRES; Nelson et al., 2008) are now routinely combined with newly typed datasets to provide a detailed picture of genetic variation which can take into account the different ancestries across different parts of the genome. For example, in line with historical understanding, principle component analysis (PCA) of genome-wide autosomal data has shown that almost all contemporary Jewish populations have a shared Levantine ancestry (Behar et al., 2010). In contrast, genome-wide SNP data have also provided fresh insight into unknown human history. For example, a large analysis of Indian populations picked up previously unidentified population structure, suggesting that contemporary Indian populations are a mixture of two geographically distinct ancestral populations, from north and south India (Reich et al., 2009; Metspalu et al., 2011). The putative northern Indian ancestral population was shown to be genetically more similar to Europeans, Middle Easterns and Central Asians, and the southern ancestral population being most similar to modern day Andamans (Reich

et al., 2009).

Several studies have attempted to explore genomic history in a European context. Research has tended to concentrate on PCA (Patterson et al., 2006) and *structure*-like analyses (Pritchard et al., 2000; Alexander et al., 2009), comparing particular geographic regions of Europe with populations of potential ancestral interest. Initial studies showed that in Europe, over and above the high genetic similarity of individuals, genetic variation follows expectations of an isolation by distance model, with geography being the best predictor of genetic similarity: when averaged over many loci across the whole genome, individuals tend to be most similar to their neighbours and this similarity decreases with increasing distance between populations (Novembre et al., 2008; Novembre and Stephens, 2008; Lao et al., 2008; Nelis et al., 2009). However, when populations are incorporated from outside of Europe, a more detailed analysis of the historical relationships between peripheral European populations can be undertaken. For example, Moorjani and colleagues assessed the level and origin of African admixture in southern European populations identifying a small amount of distinct African admixture in almost all of the southern, but not northern, European populations tested (Moorjani et al., 2011). In a study focusing on the Caucasus, Yunusbayev and colleagues showed that despite their geographical proximity, there is a lack of genetic continuity between eastern European and Caucasian populations, suggesting limited gene-flow between these two regions, perhaps because of geographical barriers such as the Black Sea (Yunusbayev et al., 2011).

In addition to the exploration of shared axes of genomic variation and ancestral components, differences in genome-wide genetic diversity have also been observed, with broadly defined southern European populations exhibiting greater genetic diversity than populations in the north of Europe (Auton et al., 2009). These observations can be aligned to several distinct hypotheses of human evolutionary history, including the initial peopling of Europe, the subsequent recolonisation of northern latitudes from southern refugia after the last Ice Age, or greater recent admixture from non-European peoples in historical times in southern populations compared to those from

the north. Whilst Europe-wide studies have hinted at explanations for the observed diversity and variation, none has so far systematically tested these competing hypotheses.

Recent advances in genomic analytical tools now make this a possibility. Several methods now exist to convert genome-wide SNP scan genotype data into phased haplotypes (Scheet and Stephens, 2006; Browning and Browning, 2007; Kong et al., 2008; Howie et al., 2009; Delaneau et al., 2011), which allow researchers the possibility to reconstruct chromosomes in individuals and thus to exploit information on linkage disequilibrium across the genome. Previously, researchers would typically have trimmed ~500k SNP datasets to ~200k SNPs, in order to remove any bias introduced from potential linkage between markers. With the steady improvements in chromosome phasing, however, this information can now be incorporated into population genetic models thereby unlocking more of the information in the data. Detailed genetic maps can be used to assign recombination probabilities to different parts of chromosomes which can further define haplotypic recombination chunks within chromosomes. The pairing of haplotype data and a genetic map thus opens up the possibility of ancestry estimation along whole chromosomes, but also, importantly, the amount of recombination that has occurred along particular chromosomes. This latter development is of primary note as it leads to the potential to estimate the number of generations (i.e. recombination events) that have occurred since the incorporation of particular genetic material into a population and therefore the ability to provide a timescale for particular admixture events.

In this chapter I will investigate the population structure of Europe using haplotype data. Initially I confirm with the current genotypic dataset that the relationship between genetics and geography previously observed in Europe is present. As mentioned above, to move from a dataset of binary SNP genotypes to a dataset of phased haplotypes it is necessary to employ a phasing algorithm with a reference set of individuals of known phase. To achieve this, I use *IMPUTE2* (Howie et al., 2009), and the next section of this chapter explores the consistency of this program in a sub-

set of the data. I follow this analysis with the use of these haplotypes and a new analytical framework using the programs *ChromoPainter* and *fineSTRUCTURE* (Lawson et al., 2012). I present the fine scale relationships between European and a sample of world populations. A second Eurasian-focused analysis presents the detailed genomic relationships between Europe and its surrounding populations. Finally, I confirm that current haplotypic diversity can be represented by a north-south gradient in Europe, with northern European populations having lower diversity. In chapter 5, I will build on these baseline observations and investigate the evidence for admixture in European populations.

4.2. Materials and Methods

Samples

DNA was collected from unrelated volunteers after getting informed consent in accordance with the guidelines of the ethical committees of the various institutions involved. A list of colleagues who shared DNA for this study is given on page 141 towards the end of this thesis. Two hundred and twenty-four individuals from 14 populations were genotyped on Illumina 660W whole genome scan SNP chips at the Wellcome Trust Centre for Human Genetics (WTCHG) in Oxford. Of these, eight had call rates of less than 0.6 and were removed as failures, presumably because of poor quality and/or insufficient quantity of DNA. I then applied the following quality control procedures. First I dropped any individuals with a call rate of less than 0.985 (n=4) and then any SNP with a call rate of less than 0.985 across the remaining individuals. The data were explored using PCA as implemented by the *smartpca* program in the in the EIGENSOFT package (Patterson et al., 2006) to highlight any individuals whose first two eigenvalues differed qualitatively from those of the majority of individuals from the same population. Hierarchical clustering in the *R* statistical program (R Development Core Team, 2011) based on an IBS (identical by state) distance matrix produced

in *PLINK* (Purcell et al., 2007) was used to further test for mislabelled individuals, by highlighting individuals that clustered away from other individuals from their population. These analyses indicated that four individuals were potentially mislabelled: 2 Norwegians and 2 Spanish. Investigation of the raw chip data showed that these four samples were run on the same chip and it was thus assumed that the labels had been mixed up at some point between delivering the DNA to WTCHG and obtaining the raw results. These four individuals were therefore relabelled. The 224 newly genotyped samples were typed in three batches at the WTCHG. Boxplots of the first five eigenvalues across individuals grouped by batch (figure B.2), and PCA visualisation of the samples labelled by batch and population (figures B.3 and B.4), showed no significant difference across the batches or provenance, suggesting there was no effect of batch to the variation observed .

Table 4.1. Populations, sample sizes, genotyping platform, and provenance of samples used in the genomic analyses.

continent	region		population	n	platform	source
Africa	south		SanKhomani	30	550	Henn et al 2011
``	``		BantuSouthAfrica	8	650Y	Li et al 2008
``	``		SanNamibia	5	650Y	Li et al 2008
``	central		BiakaPygmy	21	650Y	Li et al 2008
``	``		MbutiPygmy	13	650Y	Li et al 2008
``	east		Hadza	3	550	Henn et al 2011
``	``		Sandawe	28	550	Henn et al 2011
``	``		BantuKenya	11	650Y	Li et al 2008
``	west		Mandenka	22	650Y	Li et al 2008
``	``		Yoruba	21	650Y	Li et al 2008
``	north	east	Egyptian	12	610	Behar et al 2010
``	``	``	Ethiopian	19	610	Behar et al 2010
``	north	west	Mozabite	29	650Y	Li et al 2008

continent	region		population	n	platform	source
"	"	"	Moroccan	25	660W	this study; Behar et al 2010
"	"	"	Tunisian	12	660W	this study
Europe	south	west	Basque	24	650Y	Li et al 2008
"	"	"	French	28	650Y	Li et al 2008
"	"	"	Spanish	34	660W	this study; EthnoAncestry; Behar et al 2010
"	north	west	English	6	660W	EthnoAncestry
"	"	"	Ireland	7	660W	EthnoAncestry
"	"	"	Scottish	6	660W	EthnoAncestry
"	"	"	Welsh	4	660W	EthnoAncestry
"	"	"	Orcadian	15	650Y	Li et al 2008
"	"	"	Norwegian	18	660W	this study
"	north	central	GermanyAustria	4	660W	EthnoAncestry
"	"	"	Finnish	2	660W	EthnoAncestry
"	south	central	Sardinian	28	650Y	Li et al 2008
"	"	"	Tuscan	8	650Y	Li et al 2008
"	"	"	EastSicilian	10	660W	this study
"	"	"	NorthItalian	12	660W	this study
"	"	"	SouthItalian	18	660W	this study
"	"	"	WestSicilian	10	660W	this study
"	north	east	Belorussian	8	610	Behar et al 2010
"	"	"	Chuvash	17	610	Behar et al 2010
"	"	"	Lithuanian	10	610	Behar et al 2010
"	"	"	Russian	25	650Y	Li et al 2008
"	"	"	Polish	16	660W	this study
"	south	east	Cypriot	12	610	Behar et al 2010
"	"	"	Hungarian	20	610	Behar et al 2010
"	"	"	Romanian	14	610	Behar et al 2010
"	"	"	Bulgarian	18	660W	this study
"	"	"	Croatian	19	660W	this study
"	"	"	Greek	20	660W	this study; EthnoAncestry

continent	region		population	n	platform	source
Asia	north	west	Daghestani	20	660W	this study
``	``	``	Armenian	16	610	Behar et al 2010
``	``	``	Georgian	20	610	Behar et al 2010
``	``	``	Lezgin	18	610	Behar et al 2010
``	``	``	Adygei	17	650Y	Li et al 2008
``	central	west	Syrian	16	660	Behar et al 2010
``	``	``	Turkish	17	610	Behar et al 2010
``	``	``	Bedouin	45	650Y	Li et al 2008
``	``	``	Druze	42	650Y	Li et al 2008
``	``	``	Jordanian	20	660W	Behar et al 2010
``	``	``	Palestinian	46	650Y	Li et al 2008
``	``	``	Iranian	20	610	Behar et al 2010
``	south	west	Saudi	10	660	Behar et al 2010
``	``	``	Yemeni	7	610	Behar et al 2010
``	``	``	UAE	14	660W	this study
``	south	central	Uzbekistani	15	610	Behar et al 2010
``	``	``	Balochi	24	650Y	Li et al 2008
``	``	``	Brahui	25	650Y	Li et al 2008
``	``	``	Burusho	25	650Y	Li et al 2008
``	``	``	Hazara	22	650Y	Li et al 2008
``	``	``	Kalash	23	650Y	Li et al 2008
``	``	``	Makrani	25	650Y	Li et al 2008
``	``	``	Pathan	22	650Y	Li et al 2008
``	``	``	Sindhi	24	650Y	Li et al 2008
``	south	east	Indian	13	610	Behar et al 2010
``	``	``	Cambodian	10	650Y	Li et al 2008
``	``	``	Myanmar	3	610	M.Metspalu (pers comm)
``	far	east	Daur	9	650Y	Li et al 2008
``	``	``	Mongola	10	650Y	Li et al 2008
``	``	``	Uygur	10	650Y	Li et al 2008
``	``	``	Xibo	9	650Y	Li et al 2008
``	``	``	Yakut	25	650Y	Li et al 2008
``	east		Dai	10	650Y	Li et al 2008
``	``		Han	34	650Y	Li et al 2008
``	``		HanNchina	10	650Y	Li et al 2008
``	``		Hezhen	8	650Y	Li et al 2008

continent	region	population	n	platform	source
``	``	Japanese	28	650Y	Li et al 2008
``	``	Lahu	8	650Y	Li et al 2008
``	``	Miao	10	650Y	Li et al 2008
``	``	Naxi	8	650Y	Li et al 2008
``	``	Oroqen	9	650Y	Li et al 2008
``	``	She	10	650Y	Li et al 2008
``	``	Tu	10	650Y	Li et al 2008
``	``	Tujia	10	650Y	Li et al 2008
``	``	Yi	10	650Y	Li et al 2008
``	``	Melanesian	10	650Y	Li et al 2008
``	``	Papuan	17	650Y	Li et al 2008
America		Colombian	7	650Y	Li et al 2008
``		Karitiana	14	650Y	Li et al 2008
``		Maya	21	650Y	Li et al 2008
``		Pima	14	650Y	Li et al 2008
``		Surui	8	650Y	Li et al 2008
TOTAL		95	1550		

I combined these samples with 40 individuals from various populations genotyped by Ethnoancestry[I] on the same Illumina 660W chips as well as published data from Illumina chips containing overlapping SNPs (table 4.1). These included individuals from the HGDP (Li et al., 2008), and populations selected from Behar et al. (2010) to increase the coverage of Middle-Eastern, North African, and Eurasian regions, and African populations from Henn et al. (2011), to give a broader coverage of Africa than the HGDP popuations alone. Whilst the main focus of this study was Europe and the populations presumed through geographical proximity to have an influence on the

[I] Data from Ethnoancestry, a genetic testing company, were kindly shared by Dr Jim Wilson of the University of Edinburgh. All data were anonymous and supplied only with a country of origin.

history of Europe, I included these extra populations from around the world in case interesting anomalies showed up. From this combined dataset, SNPs with a call rate of less than 0.985 were again dropped, followed by individuals who had a call rate of less than 0.985 for the remaining SNPs. Individuals dropped at this stage were all from the Henn dataset (14 Hadza and 1 Khomani San). The final combined dataset contained 1,550 individuals from 95 populations typed for 475,855 autosomal SNPs (table 4.1). Figure B.6 is a map depicting the geographic locations of each population and can be found on page 196

Principal Components Analysis

The *smartpca* program in the EIGENSOFT package (Patterson et al., 2006) was used to identify the principal components (PCs) of autosomal variation in the sample. To remove SNPs in linkage disequilibrium (LD), the full SNP dataset was pruned using the *--indep-pairwise* option in *PLINK* (Purcell et al., 2007). Using the approach of Behar et al. (2010), the data was pruned using 200 SNP windows. For each pair of SNPs in the window, the pairwise LD correlation was calculated for all of the genotypes at the two SNPs and one of the pair was removed if the correlation (R^2) was greater than 0.4. The window was then moved on 25 SNPs and repeated. The pruned dataset comprised 221,404 autosomal SNPs. The genotypic structure within Europe was explored by including the European subset of populations (table 4.1). To explore relationships between European populations and west Asian populations, a second Eurasian analysis was performed including all European populations together with populations from the Caucasus: the Adygei, Armenians, Daghestanis, Georgians and Lezgin; and two Asian populations: the Iranians and Turkish. No outliers were removed in these analyses and for all populations containing more than 20 individuals, a subset of 20 random individuals was used.

Phasing

IMPUTE2 (Howie et al., 2009) was used to phase the full ~475,000 SNP genotype dataset. *IMPUTE2* uses a set of reference haplotypes to estimate the haplotypes present in a collection of individuals genotyped across the 22 autosomes separately. I used all of the Phase III HapMap individuals (The International HapMap Consortium, 2010), based on the same NCBI build 36 (hg18; HapMap Phase II) coordinates as the SNPs on the Illumina chips used in this study and phased using *PHASE* software (Stephens and Scheet, 2005). I performed a series of tests to explore the consistency of the *IMPUTE2* phasing algorithm.

When phasing whole chromosomes using *IMPUTE2* , chromosomes should be split into chunks no longer than 7Mb. First, I performed two independent runs with a 7Mb region of chromosome 22 containing 1,030 SNPs using the default settings. I used the ``best-guess'' haplotypes output from the algorithm, which are haplotypes estimated from the results of 20 iterations of the MCMC algorithm.

Mean switch error rate between the two runs

I calculated the average switch error rate between the two runs. To do this, I assumed that the first run gave the ``true'' haplotypes. A comparison was made between one of these true haplotypes and the second run haplotypes to estimate the switch error rate. The first SNP from the true haplotype was compared to the first SNP from one of the second run haplotypes. If it was the same, then the second SNP was compared, if this was again the same, then the third SNP was compared, and so on, until a different SNP was encountered. At this point the comparison switched to the other haplotype from the second run and a switch error was recorded. This process was continued until the end of the chromosome was reached. So, for each individual, I was able to calculate the number of times along the 1,030 SNP true haplotype that the haplotypes from the second run switched. To control for the total amount of potential switches

that could occur, this value was divided by the total number of heterozygous SNPs, minus one, in an individual. I found no appreciable difference in the switch error rate between different individuals or populations and the median switch error rate for the complete dataset was ~5%, which is in line with differences found between different phasing methods reported by Browning and Browning using simulated datasets of a similar size and SNP distribuition (Browning and Browning, 2007). Furthermore, the presence of these ``long-range'' errors (approximately one error every 0.35 Mb) in the data should not qualitatively effect the results of the chromosome painting method which inherently accounts for such errors (see below).

Input file order

In the analysis described above, I noticed that the consistency between the two runs improved depending on the position in the input file an individual came, with the mean switch error rate showing a general decrease when individuals were ordered by input file order. I therefore included a randomisation step in all further analyses, where the input order of individuals was randomised for each separately-phased chromosome chunk.

Comparisons with a different phasing approach

To further test the consistency of the phased haplotypes, I compared the HGDP individuals phased with *IMPUTE2*, with the same individuals phased using an alternative methodology, *fastPHASE* (Scheet and Stephens, 2006) [2]. For this analysis I phased chromosome 22 in all individuals with *IMPUTE2*, and kept only the HGDP individuals for further analysis. I split the unphased data into 7Mb chunks, each chunk starting 1Mb downstream of the previous chunk, thus forming a buffer that was used to stitch

[2] HGDP haplotypes phased with *fastPHASE* were kindly shared by Joseph Pickrell at the University of Chicago.

the chromosomes back together. For reasons outlined above, for each chunk, the order of individuals in the input file was randomised to account for any input order bias. Following phasing, the individuals were un-randomised and the 7Mb chunks were stitched back together to produce full haplotypes for each individual for the whole of chromosome 22. I ran *ChromoPainter* (see below) on both sets of phased haplotypes splitting the individuals into major geographic regions: Europe, Africa and the Americas. Due to resource limitations, Asian individuals were discarded at this stage and not painted. In this context, the output of *ChromoPainter* can be viewed simply as a matrix stating the amount of genome shared between each individual and every other individual in the dataset. I compared both the chunk count and chunk length matrices between the two runs. For each population, the mean chunk length copied from every population was estimated to obtain a population average for the mean chunk size copied from every other population. For all populations, this value was larger in the *IMPUTE2* dataset. That is, the average length of chromosome 22 chunks copied by each population was longer in the *IMPUTE2* dataset, suggesting that longer haplotypic blocks were present in the *IMPUTE2* phased data. I also looked at which donor individual each recipient copies from maximally, for each of, the number of chunks, the length of chunks and the average length of chunks. In almost all cases an individual from a given population was more likely to copy most from another individual from the same population in the *IMPUTE2* phased dataset, compared to the the *fastPHASE* dataset. Even when no *a priori* population label is given, one would expect an individual to copy most from another individual from the same population, at least in the HGDP dataset whose populations have been shown to be clearly separated when analysed by PCA in the past (Li et al., 2008).

These tests indicate that the within *IMPUTE2* switch error rate is comparable to rates between different phasing methods, and that the *IMPUTE2* phased haplotypes perform at least as well with the chromosome painting method as our expectations based on analysis using the *fastPHASE* haplotypes. I therefore proceeded with *IMPUTE2* phasing of the full dataset and used these haplotypes for all further analyses.

Chromosome Painting

I used a novel inferential framework for investigating population structure from haplotype data (Lawson et al., 2012). Initially, haplotypic chromosomes are "painted" sequentially using the *ChromoPainter* algorithm[3]. Exploiting the model of Li and Stephens (2003), which explicitly relates LD to the underlying recombination process, *ChromoPainter* uses an approximate method to reconstruct each "recipient" individual as a series of recombination chunks from all of the other "donor" individuals. The aim of this approach is to identify, at each SNP as we move along the genome, the closest relative genome among the members of the other sampled populations. Because recombination occurs along chromosomes over time, the identity of the closest relative will change depending on the admixture history between individual genomes. The *ChromoPainter* algorithm provides, for a given haplotype, the identity of the donor haplotype at each SNP along a chromosome. Here, painting refers to the application of a different label to the donor haplotypes such that, conceptually, each donor is represented by a colour, with each haplotype therefore made up as a series, or mosaic, of colours based on the identity of the donor at each SNP along a chromosome. Donors may be coloured individually or coloured in groups based on *a priori* defined labels, such as the population that they come from. This result is then efficiently summarised by decomposing each of the chromosomes into a series of chunks, and identifying the number and length of these chunks donated by each colour (which may be either an individual or a population). This is termed the *coancestry matrix*, and displays the pairwise relationships between all individuals in the analysis. This colouring provides a rich yet simple summary of the information on the historical relationships between individuals and populations which I utilise for the remainder of this thesis.

Practically, to aid computation time, I split the data into groups of populations, and painted each chromosome of each population separately on a high performance computing cluster at the Oxford Supercomputing Centre. In all of the downstream ana-

[3]Downloaded from www.paintmychromosomes.com

lyses except the genomic diversity estimation, I used the results of one analysis where each individual was conditioned on every other individual in the whole dataset. *ChromoPainter* outputs several summary matrices. I used the pairwise chunk count matrix for the *fineSTRUCTURE* analysis, and the chunk length matrix for the admixture analyses in chapter 5.

Before running *ChromoPainter* on the complete dataset, it is necessary to estimate two nuisance parameters from the data: *Ne* and ϑ. These are not *Ne* and ϑ in the traditional sense, but refer to parameters from the Li and Stephens (2003) copying model that are used to estimate the recombination rate distribution underlying the model. I note the process of estimating them here and their values for reproducibility. I used the expectation-maximisation (E-M) algorithm option within the *ChromoPainter* program that iterates over the data to find the local optimum values of these parameters given the data. *Ne* is the "recombination scaling constant" and is directly related to the effective population size and is used by *ChromoPainter* to convert the values of the genetic distances between SNPs taken from the genetic map to the population-scaled values of these distances required by the algorithm. I used the human genome build 36 genetic distance estimates from the HapMap website[4]. ϑ is the per site mutation rate parameter and is used by *ChromoPainter* to allow for imperfect copying between haplotypes. I jointly estimated *Ne* and ϑ by choosing 8 populations from across the world and applying 10 iterations of the full E-M algorithm on a selection of 5 chromosomes by using the *--in* and *--iM* flags. As this process is laborious, even on a high-performance computing cluster, to aid computation time I selected a subset of populations chosen to represent a broad sample of the world: Armenian; Brahui; Karitiana; HanNchina; Melanesian; SanNamibia; Saudi; SouthItalian, and haplotypes from a selection of five chromosomes: 2; 5; 8; 15; 22. This gave a value of 282 for *Ne* and 0.000610 for ϑ, which were used in all further *ChromoPainter* analyses using the *--n* and *--M* flags respectively. The default settings were used for all other parameters.

The outputs of the *ChromoPainter* runs were combined first by chromosome then by

[4] Downloaded from www.hapmap.org

population using the associated program *ChromoCombine*[5]. The final output is a matrix with recipient individuals as rows, and donor individuals as columns, or, more intuitively, a matrix where each individual's *copying vector* is a row. A copying vector can be viewed as a vector that describes the proportion of DNA (either in chunk counts or in length of chunks) that each recipient individual copies from each potential donor individual: i.e. in this case, every other individual in the dataset. Whilst it is possible to run *ChromoPainter* to produce population-level output, in the current study both donors and recipients are individuals.

fineSTRUCTURE

fineSTRUCTURE is a model based Bayesian clustering algorithm that efficiently uses the output of *ChromoPainter* (or indeed any coancestry matrix) to identify population structure (Lawson et al., 2012). Using each individual's copying vector from the *ChromoPainter* chunk count matrix, *fineSTRUCTURE* compares the copying vectors of all individuals and clusters them on the basis of similarity in these vectors. Similarity between the copying vector of individuals implies similar ancestry between those individuals.

For the initial run of *fineSTRUCTURE*, the algorithm was run twice and convergence was assessed by comparing the cluster membership of the individuals output by the two runs. Starting with all individuals as a single cluster, I ran *fineSTRUCTURE* for 10 million MCMC iterations, thinning to only include a single posterior sample for every consecutive 1,000 iterations. At each iteration, a series of splits and merges are performed on random samples of individuals, such that clusters with higher partition probability are kept at the end of each iteration (see supplementary material in Lawson et al., 2012). To produce a tree relating the clusters, I ran the maximum *a posteriori* (MAP) state from the output of this first step with 1 million iterations of the tree-building model and a very large value for the maxtreestates (-t) option of 10 million, to

[5]Downloaded from www.paintmychromosomes.com

ensure that a large number of trees were considered at each iteration. To find the tree, *fineSTRUCTURE* starts from the MAP state and successively merges clusters, choosing the merge giving the highest probability for the merged group at each step. This results in a bifurcating tree relating the clusters together (Lawson et al., 2012). Bipartition certainties for the nodes of the trees were estimated by *fineSTRUCTURE*.

I performed a second analysis with *fineSTRUCTURE* using the concept of *superindividuals*. Superindividuals look like (re-weighted) normal individuals, but cannot be split and do not contribute to parameter inference and can thus be considered as a copying vector containing the average values of the individuals within them. This allows them to be included in the algorithm without additional computational cost and they exist primarily to provide chunks copied to (and from) the remaining population. Therefore, to aid computation time, individuals belonging to groups that are not of interest can be combined into a single copying vector. For example, the SubSaharanAfrica superindividual is a single copying vector that represents all the individuals from the SubSaharan African branch of the tree (figure B.10). The processing time of the algorithm is directly related to the number of individuals included in the analysis, so reducing the total number of individuals speeds up computation time. Furthermore, the *fineSTRUCTURE* algorithm uses a prior that assumes that all populations are equally distant from each other, which in the current analysis is not true: the European populations will be more closely related to each other than to the African populations. The result of this is that not all substructure is identified in one run. Using the results of the first *fineSTRUCTURE* tree (figure 4.3), I forced individuals from similar broadly defined continental regions into superindividuals. This resulted in 8 superindividuals: SubSaharanAfrica, EastAsia, CentralAsia, the Pacific, Americas, NorthAfrica, the MiddleEast and SouthCentralAsia (figures B.7-B.10 in the Appendix show the make-up of each superindividual). 579 Eurasian individuals remained in the analysis. Broadly speaking, the names of these superindividuals represent the continental area from which the vast majority of individuals within them come from. The one exception is the SubSaharanAfrica superindividual which, as well as all sub-

Saharan African individuals, contained a group of north African individuals who clustered with these sub-Saharan African individuals, and so are included within the SubSaharanAfrica superindividual. The CentralAsia superindividual refers to a group of clusters containing Uzbekistan, Hazara and Uygur individuals, whilst SouthCentralAsia is a group of clusters containing individuals from populations from Pakistan and India. All trees and heatmaps were plotted using the *R* programming language (R Development Core Team, 2011), on both the full dataset as a whole, and with major non-European continental regions collapsed.

Genomic Diversity

To investigate using *ChromoPainter* as a method of assessing genomic diversity across Europe, I ran *ChromoPainter* on subsets of the data. Recall that *ChromoPainter* reconstructs the chromosomes from each recipient as a mosaic of all available donors. Therefore, to assess diversity within different regions of Europe, I selected a sample of individuals from within a particular European region, for example south-west Europe, and ran only these individuals as recipients against themselves as donors. This had the effect of reconstructing each individual's chromosomes only from those other individuals within the same geographic region. In this way, the resultant copying vector for each individual represents the number of distinct recombination chunks copied from others within the same geographic region. So, the greater an individual's total chunk count when compared only to others within its geographic region, the greater the amount of total chunks available within this region, and so the greater the diversity of genomes within a region. Understandably, this metric will be greatly influenced by the number of chromosomes being analysed. To control for this, I randomly selected 45 individuals from each of 5 European regions. I sampled each region three times and saw no appreciable difference between the separate sub-samples. For comparison, I also selected 3 independent sub-samples from eight regions of the world, and found the mean European diversity by averaging across all

European regions. In all analyses, I grouped populations based on the *fineSTRUCTURE* clusters observed in the whole-world analysis (figure B.6).

4.3. Results

European population structure

Figure 4.1 shows two plots of the first two eigenvectors of variation determined by *smartpca* for the European populations only, using the pruned set of SNPs. Individuals in both plots are labelled by their population and are coloured by population in the upper plot, and by geographic region of Europe in the lower plot. As previously observed (Novembre et al., 2008; Lao et al., 2008) the main two axes of variation broadly represent geographical axes, with genetics ``mirroring geography''.

In both figure 4.1(b) and figure 4.2(b) individuals are coloured by geographic region, which are largely self-explanatory. The NorthCentral European region contains two populations with very low sample sizes, GermanyAustria (n=4) and Finland (n=2), which were included together to differentiate them from other populations containing greater numbers of individuals.

Figure 4.1(a) shows individuals coloured by population. Whilst the majority of individuals cluster with the rest of their population, there are some notable exceptions. The first PC splits north-eastern Europe from southern Europe, with the Chuvash, from Russia, pulled out to the top of the plot. One Chuvash clusters with the Russians, however. The Lithuanians, Belorussians and Poles form a close-knit cluster just south of the Russians, with the two Finnish individuals also appearing here. One Norwegian appears in this part of the plot (as clearly shown in figure 4.1(b)), with the rest of the Norwegians forming a tight cluster with the Orcadians and four British populations, which are unable to be split in this analysis. Three French individuals and two of the GermanyAustria population are also clearly clustered with this north-western

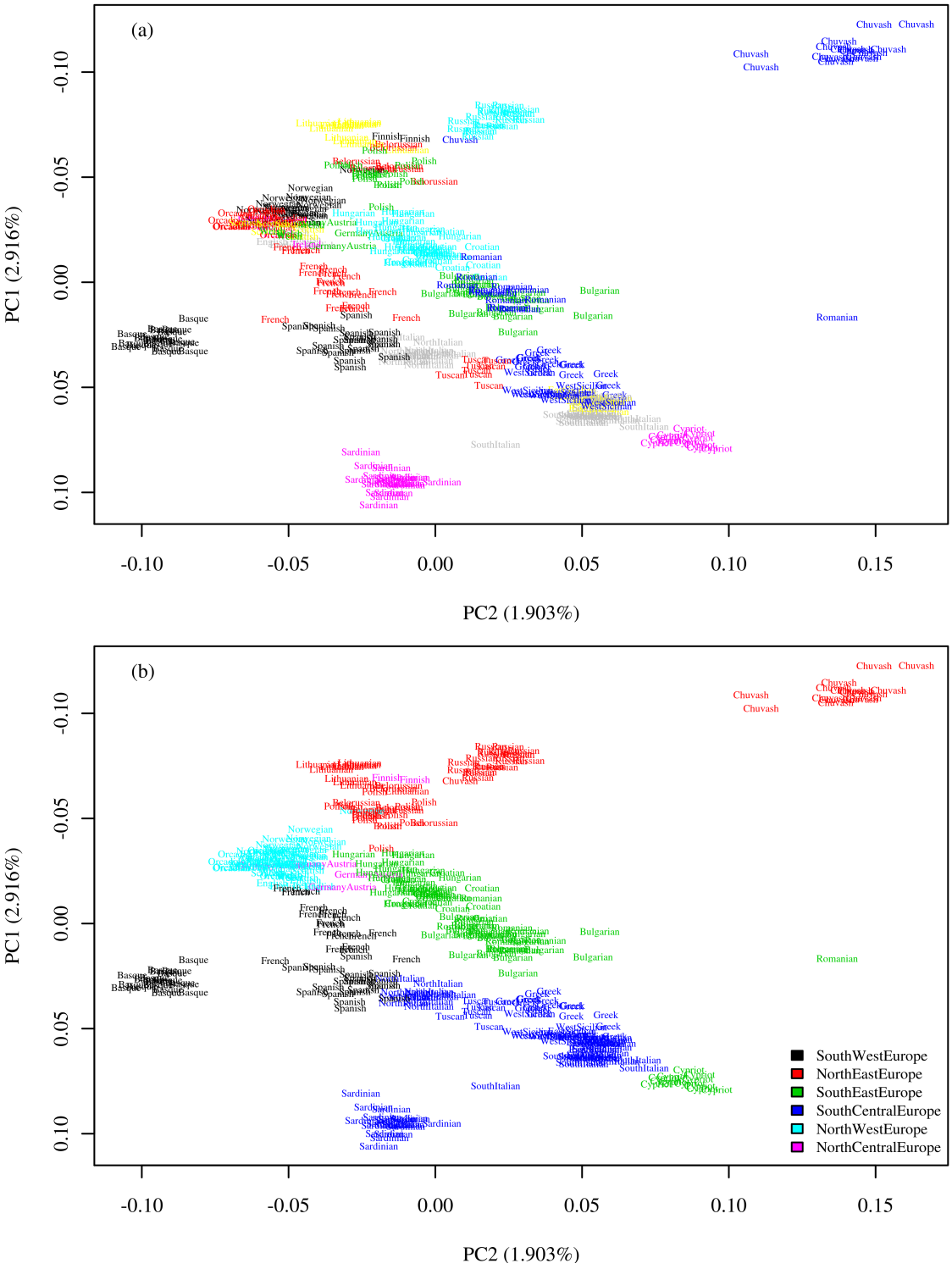

Figure 4.1. Plots showing the results of PCA on European populations. Individuals are labelled by population. Percentage of variation explained by each component is listed in parentheses after the axes labels. In figure 4.1(a) individuals are coloured by population. In figure 4.1(b) individuals are coloured by European region.

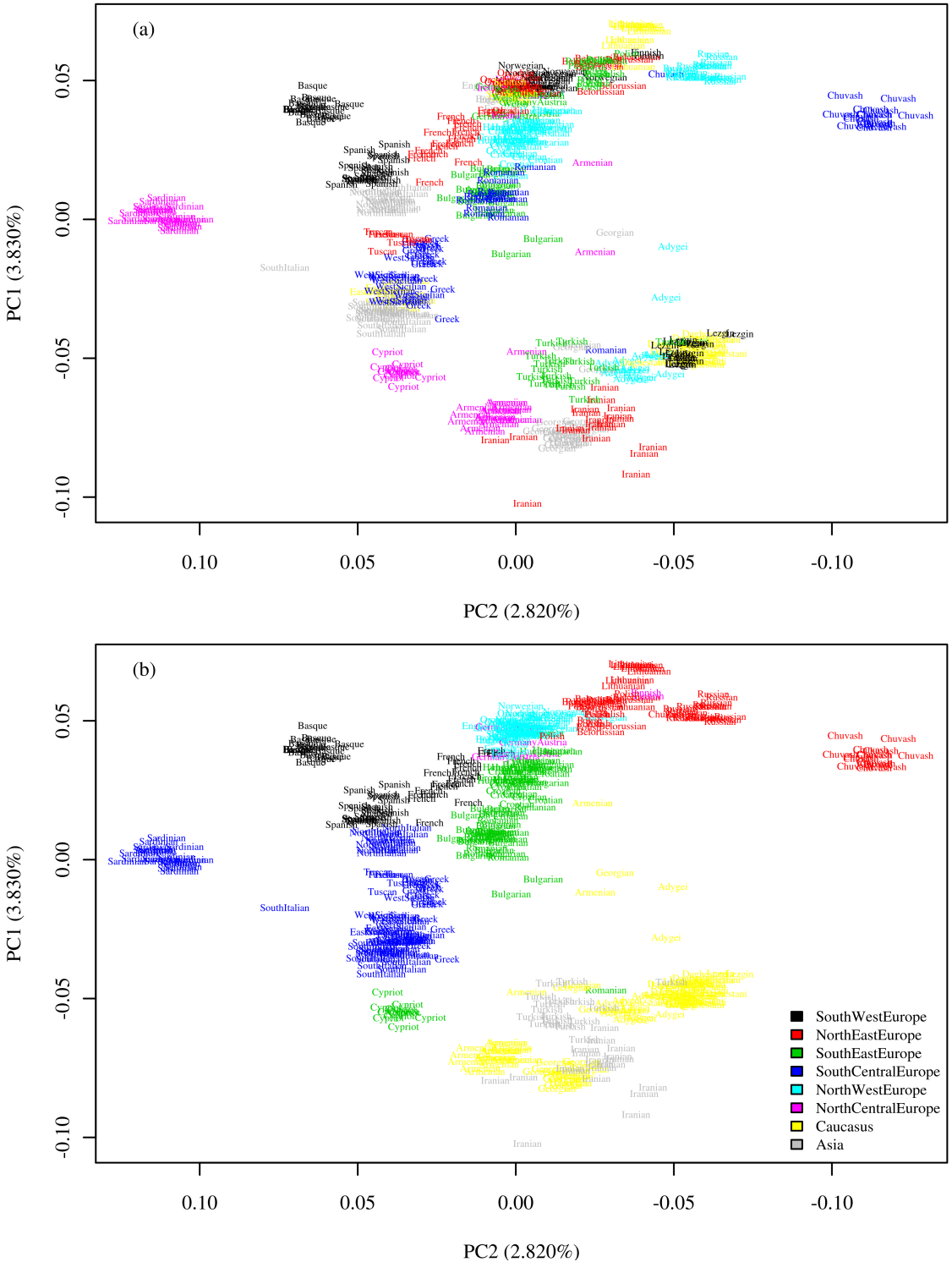

Figure 4.2. Plots showing the results of PCA on Eurasian populations. Individuals are labelled by population. Percentage of variation explained by each component is listed in parentheses after the axes labels. In figure 4.2(a) individuals are coloured by population. In figure 4.2(b) individuals are coloured by European region. Population and region colours for individuals in figures 4.1 and 4.2 are the same.

European group. The rest of the French appear midway between the north-western group and the Spanish individuals, with the Basques clearly differentiated at the extreme end of PC2, which corresponds to a roughly east-west axis. The Sardinians are also clearly differentiated from the rest of Italy, which the increased resolution of this data showing that the spread along an axis between the NorthItalians and Spanish samples on the one hand and Cypriot, presumably more Middle-Eastern, with the Sicilians and SouthItalians on the other. Interestingly, the Balkans are split into three groups: the Greeks associate with the Tuscans, whilst the Croatians and Hungarians are interspersed and appear closer to the north-western group of populations, with the Bulgarians and Romanians occurring along a parallel, but different axis to the Italians, but pulled down more towards the eastern end of PC2. One Romanian is a clear outlier on its own.

The Eurasian PCA in figure 4.2 incorporate Caucasian and south west Asian populations and shows additional detail to the relationships between the individuals. The Caucasus, Turkish and Iranian populations replace the southern European populations at the pole of the east-west axis of variation. Interestingly, the Turkish individuals split the Caucasian groups and align between the Armenians and Georgians on the one hand, and the Adygei, Lezgin and Daghestan individuals on the other, even though the Caucasian populations are geographically more proximate to each other than Turkey. The northern European populations show similarity to the additional populations on PC2, but are clearly differentiated on PC1, with the non-continental European populations at the opposite pole to the Chuvash, Russian, Poles, Lithuanians, Belorussians, Norwegians, British, and Basques.

fineSTRUCTURE analysis

The initial *fineSTRUCTURE* analysis split the world into 201 clusters (figures B.6-B.10). Figure 4.3 shows the relationship between the Eurasian clusters found with this initial *fineSTRUCTURE* run. Each leaf of the tree represents a cluster and each cluster is

labelled with the population(s) where individuals come from in that cluster, and the number of individuals from that population in that cluster in square brackets. As the focus of this study is Europe, I have not investigated the structure of non-European populations further, but present these clusters in the Appendix for the interested reader (figures B.6-B.10). The non-European populations (superindividuals) are included in the trees for reference only. The first tree (figure 4.2) is generally as expected by geography: north west European populations cluster close to south west populations and north east close to south east. South central populations, from Italy, Sicily, Sardinia and Greece cluster away from the rest of Europe. *fineSTRUCTURE* largely produced clusters containing the majority of individuals from a given population, and in some cases merged populations, indicating similar ancestry. For example, in agreement with the results of the PCA (figure 4.2), Belorussia, Poland and Lithuania form one cluster, as do all of the British Isles populations,together with three French and two GermanyAustria individuals. The Orcadian sample is split in two and, together with the British populations and Norwegians, form a north-west European group of clusters. The Basques, French and Spanish form another group, with 7 Spanish individuals clustering with the majority French cluster, and the remaining Spanish forming a cluster on their own. Moving up the tree, a clear eastern European and Balkan group of clusters forms, with the Croatians and Hungarians merged into a single group, together with a single Armenian and Romanian and two GermanyAustria individuals. The Bulgarians and Romanians merge to form a cluster, together with a small cluster of two Bulgarians, two Armenians and a Georgian. The outlying Romanian individual in figure 4.3 appears in a cluster with Iranians and a Turkish individual further up the tree, which can also be seen in the PCA plot that includes the Iranian and Turkish individuals in figure 4.3. The Belorussian, Lithuanian and Polish cluster is joined by the Finnish-Russian cluster, which in agreement with the PCA in figure 4.3, also includes a Chuvash individual. The rest of the Chuvash form a cluster on their own away from this north-eastern and western European group of clusters.

The south central European populations, from Italy, Sicily, Sardinia and Greece form

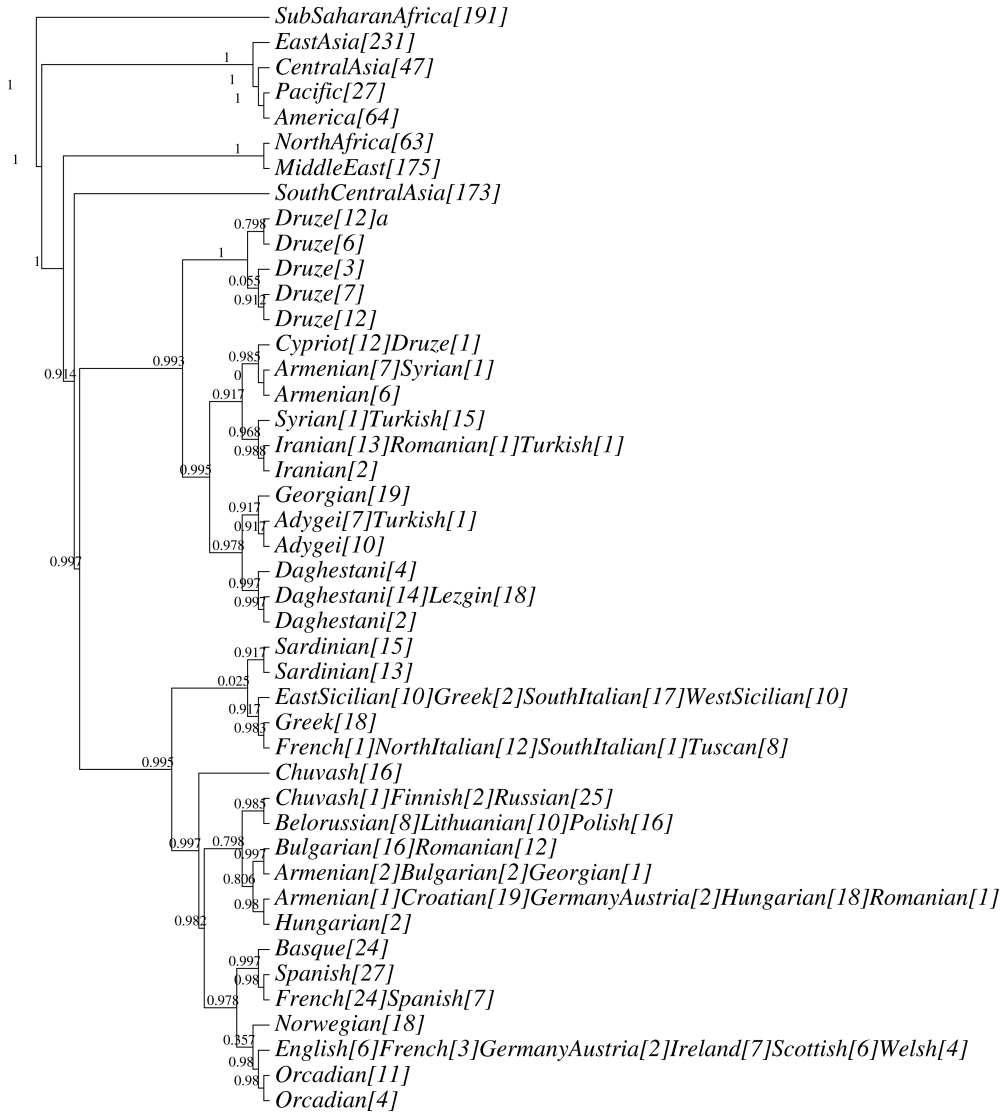

Figure 4.3. A tree relating the *fineSTRUCTURE* Eurasian clusters from the initial run. Each leaf of the tree represents a cluster of individuals with similar ancestry. The labels represent the populations from which the individuals come from in each cluster, with the number of individuals from each popuations in clusters shown in square brackets. Structure within the superindividuals shown at the top of the plot is not shown. Certainty values are given above each node (see section 4.2).

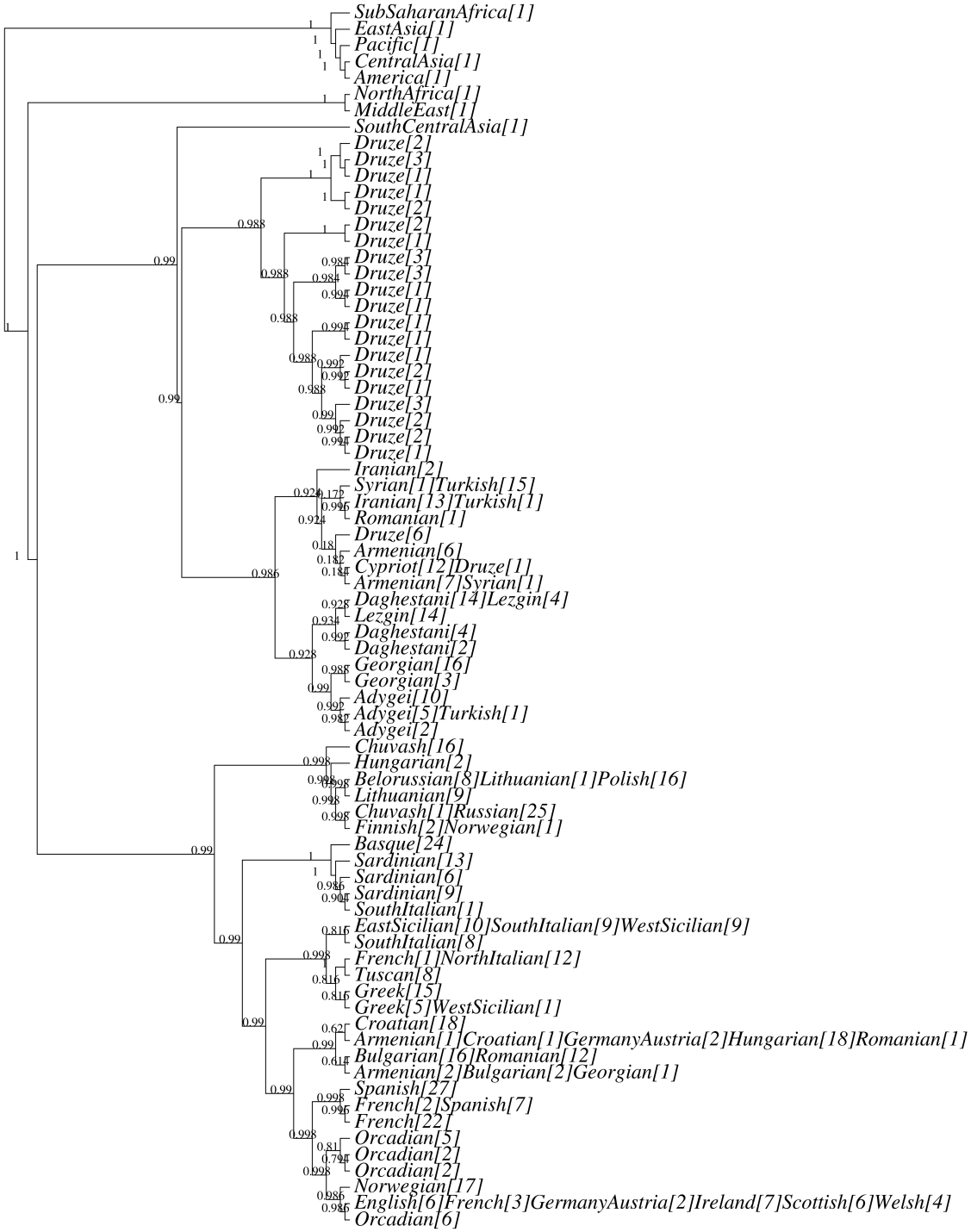

Figure 4.4. A tree relating the *fineSTRUCTURE* clusters from the second run using superindividuals in the place of non-Eurasian individuals. Nomenclature is as figure 4.3. Certainty values are given above each node (see section 4.2).

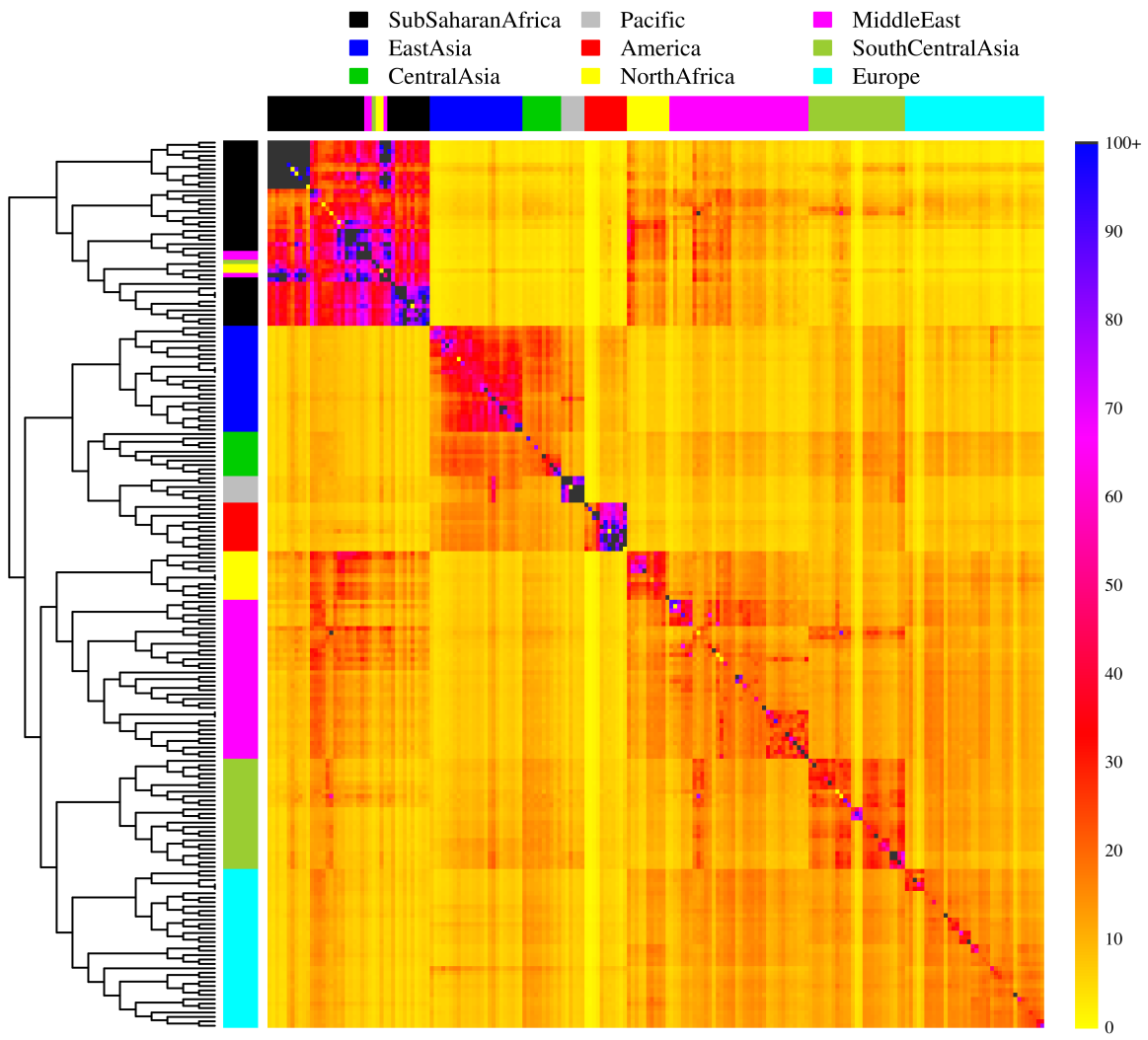

Figure 4.5. Results of the *ChromoPainter* and *fineSTRUCTURE* analysis on the whole-world dataset of 1,550 individuals from 95 populations For each cluster as a row, the columns show the number of chunks copied between populations. The colours in the plot indicate the number of chunks copied by each cluster and a key to these colours is shown on the right of the plot. The clusters are grouped according to the output of *fineSTRUCTURE*, with a tree representing the relationships between clusters on the left hand side. The colour bar along the top and left side of the heatmap indicate the continental group that the clusters belong to.

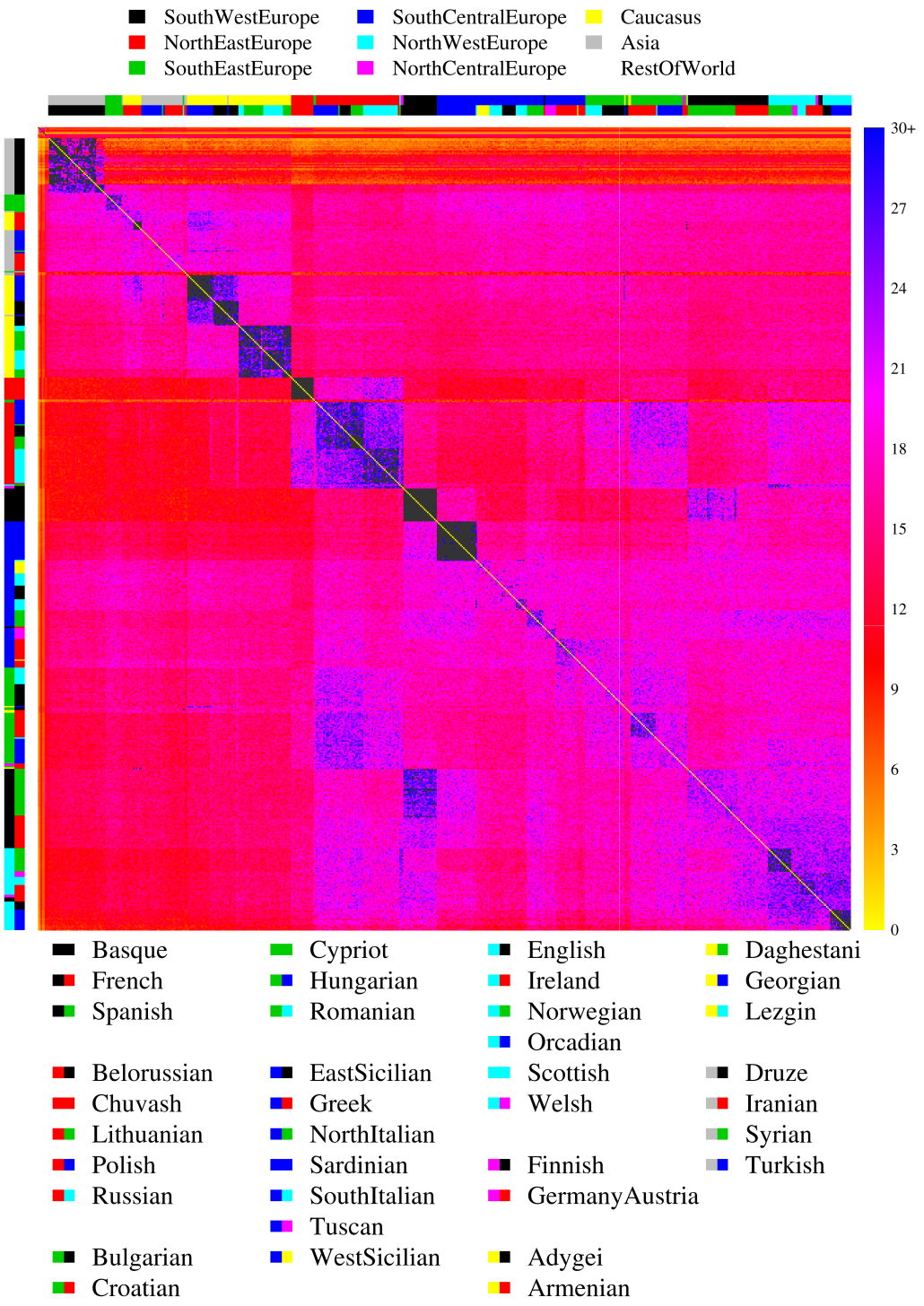

Figure 4.6. Genome copying proportions for European and Eurasian clusters only, with remaining populations collapsed into superindividuals This heatmap displays individuals as rows and columns. The colours in the plot indicate the number of chunks copied by each individual and a key to these colours is shown on the right of the plotEach individual is labelled with two colours, the first identifies the region, top legend, the second identifies the population within that region, as indicated in the legend at the bottom. The superindividuals occupy the first 8 rows at the top and left of the figure and do not have colour labels.

a group of clusters equally distant from the north-eastern and western populations. A NorthItalian and Tuscan cluster contains single SouthItalian and French individuals, and interestingly, as seen in figure 4.3, show greater affinity to the Greeks, than to the SouthItalians, Sicilians and Sardinians, who are their geographic neighbours. This analysis found two distinct clusters in the Sardinians, which cannot be seen in the PCA, whilst both the EastSicilians and WestSicilians form a cluster with the SouthItalians and two Greeks. Further up the tree, a clear group of Caucasian clusters from Daghestan and Georgia, and the Lezgin and Adygei, group together, as well as Anatolian populations, from Turkey, Iran and Armenia. The Cypriot individuals cluster with the Armenians, well away from the most of the rest of the European populations.

The second *fineSTRUCTURE* run reveals a deeper insight to the genomic structure of Europe by identifying inbreeding (drift) within clusters (the darker colour on the diagonal) and the admixture relationships between clusters. This analysis found 77 clusters in total, which includes each of the 8 continent level superindividuals. The resulting tree is shown in figure 4.4. Broadly, the tree has similarities to figure 4.3 based on the full analysis, but there are some revealing and interesting differences. The Orcadians are further split into four clusters, 3 of which group together away from the a cluster of 6 Orcadians and the British and Norwgian individuals, who form the same clusters as before. The certainty values show that the support for the split between the group of Orcadian clusters and the Norwegian-British group is high, although the split within the Orcadians has less certainty. A new cluster comprising the 7 Spanish individuals that were previously grouped with France and 2 French individuals now splits off from a larger French only cluster. Interestingly, the Basque cluster has moved from this south west European group to a different part of the tree and is now placed next to the group of Sardinian clusters, with high certainty.

More detail appears in the south-east European group of Balkan clusters. The majority of the Croatians now form a cluster on their own, leaving most of the rest of their former cluster, the Hungarians, a Croatian, a Romanian and 2 GermanyAustria indi-

viduals to form a separate cluster. The rest of the Romanians and Bulgarians remain unsplit, with a small cluster of Armenians, Bulgarians and a Georgian forming their own cluster. Both of these new splits have lower certainty, but there is strong support for the higher level groupings.

Further structure can also be seen in the south central group of clusters. A separate Greek and WestSicilian cluster splits off from the main Greek cluster and the Tuscans all split off from the other NorthItalians and French individual, who form their own cluster. A cluster of 7 SouthItalians splits off from the Sicilians and SouthItalians that were previously clustered together. Further structure within Sardinia is found, with a solitary SouthItalian, clearly visible from the PCA on its own.

The north east European group of clusters have move away from their former position and are now to be found as an outgroup to the rest of the European groups. All splits at this level have high certainty, with 9 of the Lithuanians now forming their own cluster, whilst the Polish and Belorussians and one Lithuanian remaining together. The Finnish and a single Norwegian are now split from the Russians. Two Hungarian individuals, who were formerly in the main south-east European group of clusters have moved up in the north-east European group of clusters, with the Chuvash maintaining their position on the edge of this group.

The Caucasian and Asian groups show similar evidence of increased structure, with further splitting of the four main groups of clusters representing the Daghestanis and Lezgin, the Georgians and Adygei, Iranians and Turkish and the Armenians and Cypriots. Several populations now show sub-population structure, such as the Daghestani, the Adygei, the Georgians and the Armenians. Several individuals from the Turkish and Iranian populations occur in clusters mostly containing individuals from other populations. Many clusters are found in the Druze, perhaps because of the high relatedness between individuals in this closed society (Lawson et al., 2012).

Figure 4.5 and figure 4.6 are heatmaps of the copying vectors from the *ChromoPainter* and *fineSTRUCTURE* analysis for the whole world and Eurasia respectively. Figure 4.5

depicts a 201 x 201 matrix with each row representing the mean proportion of chunks copied by a given recipient cluster, from each donor cluster. The tree shows the full relationship between the clusters found by *fineSTRUCTURE (B.6)*. The order of populations is the same as that in the full tree found in the Appendix. A coloured key is shown to indicate the continental identity of the individuals in each cluster. In the plot, the darker blue colours indicate higher copying amounts, and lighter yellows indicate fewer chunks copied. The clustering in figure 4.5 shows that within Europe, individuals within a geographic population tend to cluster with other individuals from that population and that geographically proximate populations tend to cluster together, such that clusters can be grouped into clades based on broad geographic regions of Europe: north-west Europe, south-west Europe and so on.

The genomic diversity of European regions

The diversity of each European population is shown in table 4.2, controlled for sample size. By this metric, the diversity of African populations is much greater than the rest of the world, with all non-African populations being relatively similar, apart from the Americas which have the lowest diversity. As previously observed from genome-wide haplotype data (Auton et al., 2009), southern European populations are more diverse than northern European populations.

4.4. Discussion

The study has used a combination of PCA and novel, haplotype-based, *ChromoPainter/-fineSTRUCTURE* analyses to explore the genetic relatedness of European populations. Recent theoretical work has shown that PCA can be used to infer genealogical relationships such as admixture between samples (McVean, 2009), but generally only in the specific situation where admixed populations are projected onto the PCs of two

Table 4.2. Genomic Diversity of European and select World Regions, controlled for number of individuals.

Region	Mean Diversity	Mean +/- SE
South West Europe	29,680	(29,588-29,771)
South Central Europe	31,120	(31,021-31,220)
South East Europe	29,367	(29,265-29,470)
North East Europe	27,452	(27,341-27,563)
North West Europe	26,774	(26,586-29,962)
Caucasus	31,547	(31,329-31,765)
Europe	28,879	(28,760-29,585)
Africa	45,550	(45,269-45,830)
East Asia	29,702	(29,455-29,948)
America	15,477	(14,249-16,706)
Middle East and North Africa	31,772	(31,035-32,510)
Central Asia	31,363	(30,923-31,802)

putative source populations (Wall et al., 2009; Auton et al., 2009; Behar et al., 2010; Bray et al., 2010; Bryc et al., 2010; Moorjani et al., 2011). PCA is an excellent non-parametric tool for displaying genetic population structure, but similar PC structures can be the result of different population genetic processes (Novembre and Stephens, 2008). Therefore we are stuck with the notion that whilst visually aesthetic, maps of genomic PCs cannot distinguish the *process* behind the structure of the current genetic landscape, for example by differentiating migration from isolation. Furthermore, whilst PCA is routinely used to suggest historical relationships (e.g. admixture) between populations , there is no explicit framework to link PC plots to actual events in human history.

Analyses incorporating haplotype structure aim to address this predicament by implicitly including information on recombination from the data. For example, by viewing an individual's genotype as a mixture of chromosome chunks of designated length, copied from distinct populations, the *ChromoPainter/fineSTRUCTURE* paradigm necessarily aids insight not only into the genetic relatedness of populations, but also, potentially to the historical timescale of these events. This idea is developed more in chapter 5, but in the current context, these analyses provide a fresh appreciation of the relatedness between populations by simultaneously visualising drift (on the di-

agonal) and admixture (across the rows) with the heatmaps. The *fineSTRUCTURE* trees further provide an approximate guide to similarity and the historical relationships between different European peoples by displaying the similarities of the clusters.

Populations from similar parts of Europe tend to cluster together on the basis of the comparison of their copying vectors. This suggests that it is not just the similarity of their DNA that groups them together, but also, importantly, it is the the similarity in their dissimilarity with other populations. This can clearly be seen with the north west European populations at the bottom of the heatmaps in figure 4.5 and figure 4.6. In figure 4.5, the two Orcadian clusters at the very bottom have similar colours to the rest of the north west European cluster, but do not share the similar purples of the rest of the clusters from this region, who all copy relatively more from the surrounding clusters. Note how the British and Norwegian clusters copy more from north east European clusters than the Orcadians. Similarities in the copying vectors of the south west European clusters can also be seen, with relatively more copying from the south central clusters than the north west Europeans. There is also clear copying from North Africa visible to the left hand end of the south west European clusters, in particular in the cluster containing only Spanish individuals. Both Y chromosome analysis (Adams et al., 2008; Capelli et al., 2009) and previous genome-wide studies (Auton et al., 2009; Moorjani et al., 2011) have identified a genetic link between southern Europe and (North) Africa, with the latter dating the incorporation of North African DNA into south western Europe as approximately 55 generations ago. Here, it is possible to see that both the Spanish cluster and all of the south Central (i.e. Italy and Greece) clusters also copy DNA from north Africa, with south Central populations copying from the Middle-Eastern populations included in the study. Interestingly, in only one of the south-eastern European clusters, containing Armenians, Bulgarians and a Georgian is there a similar proportion of Middle-Eastern DNA shared, despite the closer proximity of the Balkans to the Middle East. Furthermore, the tree in figure 4.4 shows the south east European clusters to be closer, genetically, to the western European clusters, and depicts the south central European group of clusters as an out-

group to this western block. The north eastern European clusters are equally different from the rest of Europe. This branching pattern, with south central Europe and north eastern Europe peripheral to the core western and south-eastern group of clusters, suggests that whilst genotypic variation appears to follow expectations based on an isolation by distance model - that is, that genetic similarity between individuals decreases as distance between them increases - when haplotype information is taken into account, this is potentially not the case. Further work incorporating explicit tests of isolation by distance between groups will aid with understanding whether these models are supported by the haplotype-based methods presented here.

North eastern Europe, whilst clearly showing some affinity with north western and south eastern European clusters, shows a distinct lack of genome copying from non-European and southern European clusters. Despite the populations involved in this group, Lithuania, Poland, Belorussia and Russia, being on the periphery of Europe, they appear to have had little genetic influence from outside of Europe, at least in relation to the other peripheral populations in southern Europe. Recent PCA of Caucasian populations (Yunusbayev et al., 2011), as well as the analysis presented here (figure 4.2) show there to be little continuity between the western edge of Asia and the eastern periphery of Europe. Poor climate, the vastness of the Eurasian steppe, and lack of proximity of potentially admixing populations, may all be factors affecting this. Whilst the southern borders of Europe seem to have be porous, the eastern and especially north eastern seem to have been less so. One caveat is the lack of western Siberian populations in the current analysis, which may provide a link between eastern Europe and East Asia.

The analysis presented here reaffirms the observation that southern Europe is more genomically diverse than northern Europe (Auton et al., 2009). Together with the indications from the *fineSTRUCTURE* analysis that both the southern European populations copy more from non-European populations, and the northern European populations copy mainly from within Europe, this suggests that the increased diversity could be due to the incorporation of more non-European DNA into southern popula-

tions. The next chapter develops this idea by assessing the evidence for admixture in European populations.

5. The History of Genomic Admixture in Europe

Around the time of the collapse of the Roman Empire, Europe and western Asia were plunged into a millennium of upheaval (Heather, 2009). Successive waves of raiding Germanic peoples from north-eastern Europe, and nomads from the central Eurasian Steppe plundered the central European plain for slaves and booty (Beckwith, 2006; Heather, 2009). The fall of the Roman Empire notwithstanding, in Mediterranean Europe too, the eastern Roman or Byzantine Empire clung on and maintained some control of the eastern seas from its capital in Constantinople. From the 8thcentury several empires expanded across Eurasia - in all directions - and for the first time in history, the great empires came into contact with each other. But it is over the course of this millennium, roughly from 500-1,500CE that the basis of current world societies was lain (Roberts, 2007). Empires and people came and went, but perhaps most importantly, civilisation and culture developed in ways that had never previously been known. It is at this time that we see the development of the group identity that would later become the basis for political cohesion and the formation of states and countries. In this chapter I will argue that this progression was important in producing contemporary European populations, and as such, leaves a genetic legacy in the differences we see in European populations today.

To do this, I will attempt to align observations based on the contributions of genomic admixture between current European and world populations with estimates

of the time at which this mixture happened, within a general historical timeframe. Whilst we will perhaps never know whether particular events had a specific impact on the genes of Europeans today, the analyses presented here offer persuasive evidence of the importance of this era to the formation of Europe's current populations and provide, I hope, a fresh synthesis between historic and genomic data.

5.1. Introduction

In chapter 4, I outlined the measures made to produce phased haplotype data from genome-wide chip data, and characterised the fine-scale genomic structure of Europe by comparing European populations both to themselves, and to a worldwide sample of populations. In this chapter I develop this broad base by attempting to understand in greater detail the relationships between European populations. The *ChromoPainter/fineSTRUCTURE* analysis showed that European populations can be clustered into groups based on the similarities of their copying-vectors. But *how* was this structure established?

Various models have been proposed to explain how Europe was colonised. The European peninsula has been inhabited by modern humans since at least 40kya, with archaeology suggesting earlier dates from sites in more eastern parts of Europe (Kozlowski, 2007; Richter et al., 2008; Hoffecker, 2009). The Paleolithic people who left these clues to their existence must have originally come from further east, as that is the only overland route from Africa, and this axis of movement, from the east to the west, must have been the major direction of flow of people initially into the continent. As I noted in the Introduction, the last ice age was effective in removing people (as well as many other species of fauna and flora) from the expanse of the northern European plain (Gamble et al., 2005), and following the LGM, the majority of Europe was re-populated from the east again (Cavalli-Sforza et al., 1994), and also potentially from southern European refugia (Soares et al., 2010). The spread of Neolithic cultures

over the following millenia, again from the east, may or may not have brought associated peoples (Childe, 1925, 1942; Ammerman and Cavalli Sforza, 1984; Semino et al., 2000; Barker, 2006; Balaresque et al., 2010; Busby et al., 2012), but the focus of archaeological and genetic research into the formation of Europe's population has in the past centred on the various contributions from Paleolithic and Neolithic populations from this recolonisation after the ice.

Implicitly, then, the discussion of the genetic history of Europe has traditionally been framed by the antagonism between these two opposing factions of people: the Paleolithic hunter-gatherers and the Neolithic farmers. Thus, the genetic structure that we see in Europe today has generally been explained by the varying contributions of these two groups, with the differences in the distribution of Y chromosome haplogroups (Rosser et al., 2000; Semino et al., 2000; Battaglia et al., 2008; Balaresque et al., 2010; Busby et al., 2012), mtDNA lineages (Richards et al., 1996; Torroni et al., 2001; Richards, 2003; Soares et al., 2010; Malyarchuk et al., 2010), and autosomal variation (Novembre et al., 2008; Lao et al., 2008; Auton et al., 2009) couched, to a greater or lesser degree, in terms relating to the interaction between these two groups of people. However, this interpretation necessarily assumes a strong genetic continuity in European populations, and explicitly requires that the DNA from contemporary populations offers a direct ancestral line through the people that have always lived in that area. The analysis of ancient DNA has so far given mixed results in this context (Haak et al., 2005; Gilbert et al., 2008; Malmström et al., 2009; Haak et al., 2010; Skoglund et al., 2012). However, greater sample sizes will in the future allow researches to objectively test the legitimacy of this assumption (Stoneking and Krause, 2011), and currently the evidence is does not suggest that continuity is widespread (Sampietro et al., 2007; Pinhasi and von Cramon-Taubadel, 2012). For example, the recent publication and analysis of the 5,000 year old Tyrolean Iceman, shows that his autosomal variation is similar to contemporary Sardinians, and not individuals from northern Italy or other Alpine populations, which could be expected given the location where he was found in the Italian/Swiss Alps . Moreover, his Y chromosome belongs to a

rare haplogroup (G2a4) that is currently found in appreciable frequencies only in the Mediterranean islands of the Tyrrhenian Sea (Keller et al., 2012). Although based on a single individual, this analysis does little to help the case that contemporary populations are *a priori* representative of the ancient people and cultures of their locale.

History and demography further suggest that, far from populations remaining static since the end of the Neolithic, there have been periods of flux and movement in Europe over the last 3,000 years. The Medieval period in Europe involved several migrations of 'Germanic' peoples in and around northern Europe and western Asia (Heather, 2009), and the contraction of the broad pan-Mediterranean power structures following the collapse of the western Roman empire released swathes of Europe to the control of smaller units made up of indigenous and wandering peoples (Roberts, 2007; Heather, 2009). Rapidly emerging infectious disease also had the potential to disrupt the underlying genetic structure of human populations. For example, the bubonic plague, in the epidemic known as the Black Death, is estimated to have killed 30-50% of the European population in the five year period from 1357 to 1351 (Benedictow, 2004). Recent comparative demographic research has indicated that the increase in sedentary lifestyle associated with the evolution of complex society and civilisation led to an increase in reproductive variance, particularly in males, with powerful elite males potentially fathering hundreds of children (Betzig, 2012). This conclusion gives added credence to the claim that Genghis Khan's Y chromosome is now shared by 8% of Asian men (Zerjal et al., 2002), providing a demographic explanation to this intriguing genetic result. It is therefore at least possible that European (and world) populations could have been affected by neoteric events within the last thousand years or so.

One reason why more recent events have not garnered a lot of attention in this context is that it has only recently been possible to identify the finer genome-wide structure in human populations. Studies using uni-parental markers in the past have struggled to assist detailed interpretations of recent human history, because differences across Europe tend to be measured as small variations in haplogroup frequency, which can

only loosely be linked to historical events (Jobling, 2012). Conversely, the analysis of hundreds of thousands of genome-wide markers led to two studies that were able to segregate even closely related European populations based on their DNA (Novembre et al., 2008; Lao et al., 2008). Some of this variation is undoubtedly due to natural selection (Coop et al., 2009; Moskvina et al., 2010), but human history, both ancient and recent, will also have played a part. Genome-wide data can be used to explore human history through the identification of admixture between different populations. In Europe, if the original peopling processes of movement from the Middle-East are solely to explain current population structure, then the expectation is that northern and western European populations should share less DNA from non-European populations than the more southern and eastern populations. Moreover, under this hypothesis, recent historical events will have had little impact on the underlying genetic differences between European populations. Recent research has challenged this view. Auton and colleagues showed that southern and south-western populations share the highest proportion of haplotypes with African (Yoruban) populations, which, they suggest, relates to recent gene-flow from Africa after the initial migrations into Europe from the Middle-East (Auton et al., 2009). However, the authors were unable to assess when this admixture may have happened. Further evidence of recent admixture in southern European populations is provided by Moorjani and colleagues, who identified a small amount of African ancestry in these populations, and were able to date the admixture to 55 generations ago, or the end of the Roman Empire (Moorjani et al., 2011).

In this chapter, I aim to address the proposition that recent events may have had an effect on the genetic structure of some or all of Europe's populations. To do this, I again use haplotype-based methods to explore the history of admixture in European populations. First, it is necessary to establish whether different parts of Europe obtain genetic elements from different parts of the world. Having achieved this, I use a novel analytical pipeline to understand the existence, proportion and date of admixture in European populations. Finally, I show that by accounting for recent admixture

in several populations, it is possible to isolate the pre-admixture genomic profile of a population, and suggest that prior to the recent admixture events that are identified, certain European populations were more similar than they appear today.

5.2. Materials and Methods

Comparisons of genome-wide copying

I first investigated whether countries in different regions of Europe copied differentially from different parts of the world. To do this I used the chunklengths coancestry matrix from my original *ChromoPainter* run (chapter 4) where each individual was conditioned on every other individual. Practically, this meant that I collapsed all non-Eurasian donor individuals into eight independent *superindividuals* based on the *fineSTRUCTURE* tree used in chapter 4 (figure B.6). These superindividuals conform well to different geographic regions of the world and were labelled such. By comparing the differences in chunklengths copied from each superindividual, or world region, by each European country, it should be possible to assess if certain countries copy more from different non-European regions. I further collapsed the copying vector for each population by averaging the chunklengths copied from all Caucasus' clusters into a ninth region, and all European clusters into a tenth region, which resulted in a ten element copy-vector for each recipient individual, where each element recorded the proportion of genome copied from each of the ten different world regions. To compare across populations, I found the mean of these copying vectors across all European populations, and compared each population separately to the European mean, standardising the mean to 0 and standard deviation of 1. I removed the Orcadians, Sardinians and Basques from all subsequent analyses, as these populations represent drifted groups, and hence will copy a far greater amount of DNA from their own population, which in turn will adversely affect the European averages. I also removed the Cypriots due to their position away from the rest of Europe in the *fineSTRUCTURE* tree.

Estimating and Dating Admixture

I next investigated the amount and origin of admixture in European populations. To do this, I utilised several programs kindly shared by Dr Garrett Hellenthal[1] that use the output of the *ChromoPainter* algorithm to assess the sources and timing of admixture in human populations (Hellenthal, Busby, Band, Capelli, Falush and Myers, in prep)[2]. This analysis attempts to assess the evidence for admixture in a population and to recreate the admixing sources. If an admixture event is inferred to have taken place, the analysis can be used to recreate the copying vectors of the two admixture source populations, the proportion that each source contributes to the admixture event, and an estimate of the time of the event in generations. The full mathematical derivation of this method forms the core of the manuscript in preparation with Dr Hellenthal and other colleagues. Here, for brevity and simplicity, I outline a summary of the analytical pipeline that is used. In this example, I will refer to the pipeline used to analyse one of the simulated populations where an admixture event was simulated to have occurred 30 generations in the past between the Yoruba and the Brahui, with the Yoruba contributing 20% of the DNA to the simulated population and the remaining 80% coming from the Brahui. The analysis was carried out in exactly the same way for all the other simulated and true populations.

Initially, fresh runs of *ChromoPainter* were required where, for each population separately, recipient individuals were only allowed to copy from donor populations other than their own, i.e. they were not allowed to copy from other individuals within their own population. Self-copying will potentially mask subtle signals of admixture between closely related populations. For each individual in a population, *ChromoPainter* outputs "painting samples", which show, for each SNP along each chromosome, the population that the individual copies from at that given SNP. These painting samples are, in fact, summarised by *ChromoPainter* to produce the chunkcount

[1] the author of *ChromoPainter*

[2] This manuscript is in preparation and involves a global investigation of admixture using similar populations to those presented here.

and chunklength coancestry matrices used elsewhere. I used the default value of 10 which produces 10 painting samples for each individual. Figure 5.1(a) shows a schematic of part of an admixed Yoruban/Brahui simulated haplotype (top chromosome), where African (in this case Yoruban) chunks have been coloured yellow, and Asian (Brahui) chunks coloured red. The middle plot shows the results of the painting analysis: i.e. a painting sample. This painting sample shows the identity of the donor at each SNP along the haplotype as coloured by the region which that donor comes from (see legend for identity of different colours). In this example each thin vertical line is a SNP. This sample is then modified by re-weighting each SNP by the genome-wide ancestry of the donor populations. For example, in this simulated population, African and Asian populations donate a greater proportion to the simulated genomes than European and American populations (i.e. contribute more chunklengths in the chunklengths matrix), and so are given a greater weight. More intuitively, because a given population, say the San, are typically inferred to have 60% San ancestry, 30% other African ancestry, and various smaller percentages from elsewhere, a segment of a recipient individual's genome copied from the San will similarly not be 100% San. The chunklength copying vectors can thus be used to clean up the original painting sample. The painting after this inference step is shown on the third, bottom chromosome in 5.1(a).

I used the chunklength coancestry matrix from the the original analysis (i.e. from chapter 4 where self-copying *was* allowed) to produce copying vectors for each population. (The alternative would be to re-run *ChromoPainter* separately for every population in the analysis, which is computationally very exhaustive, and for which resources were unavailable.) In order to obtain the initial admixing proportions to weight the raw paintings, a non-negative-least-squares (nnls) regression was performed on each recipient population chunklength copying vector modelled as the response, with the chunklength copying vectors of each donor population as predictors. This has the effect of identifying the donor populations with highly correlated chunklength copying vectors to the recipient population under analysis. The coef-

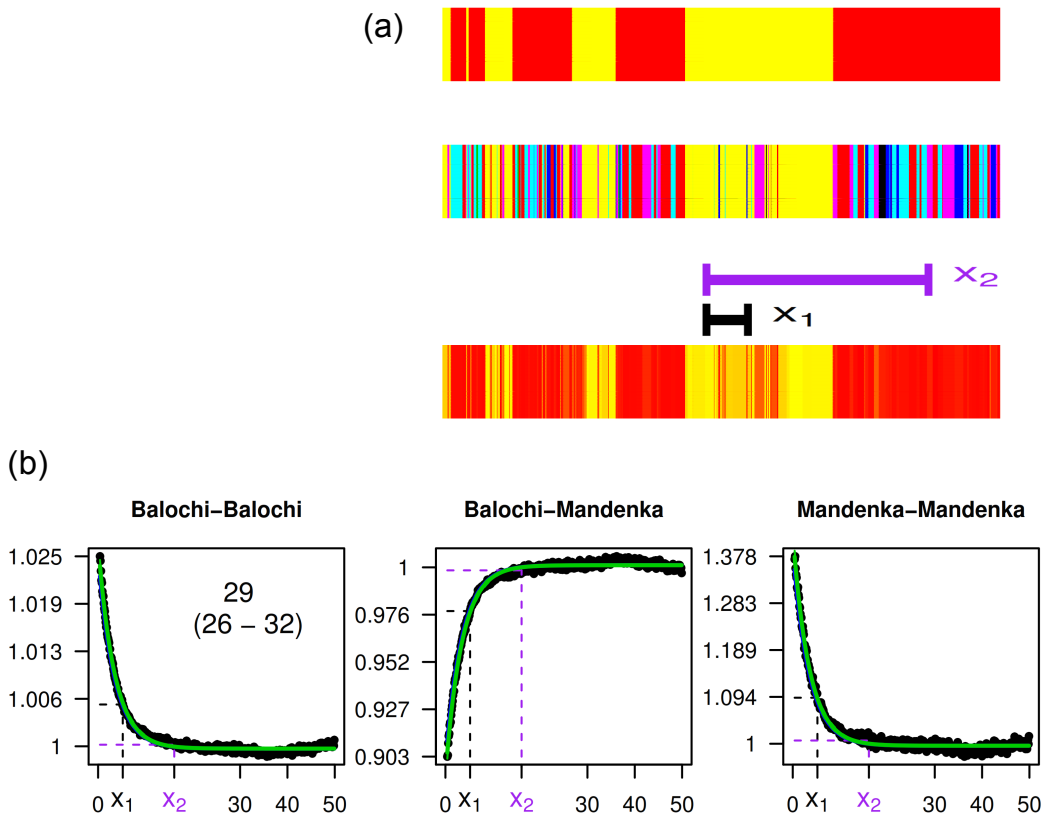

Figure 5.1. A summary of chromosome painting and dating approach (figure prepared by Dr Garrett Hellenthal, and is taken from Hellenthal, Busby et al (in prep)) **(a)** Top to bottom: haplotype simulated with admixture event occurring 30 generations ago with 20% Yoruba (yellow) and 80% Brahui (red) contributions; ChromoPainter's inferred ancestry for the simulated haplotype (yellow=Africa, green=America, red=Central-South-Asia, blue=East-Asia, cyan=Europe, pink=Middle-East, black=Oceania); inference after reweighting each segment using genome-wide ancestry decomposition as fit via linear modeling. **(b)** Each plot gives the normalized probability of copying from a segment of the left source in the title to a segment of the right source in the title versus the genetic distance in cM between the two segments (i.e. "coancestry curve", black lines). The green lines depict an exponential distribution fitted to all curves simultaneously, with rate of decay given (with 95% CI) in the left-most plot. Probabilities for segments separated by genetic distances x_1 and x_2 in (a) are highlighted.

ficients of this regression were restricted to be ≥ 0 and summed to 1. These are the weights referred to above that are used produce the cleaner painting samples used for the next step.

The 10 painting samples were then used to generate "admixture decay curves" for each European population investigated. Each painting sample is composed of two haploid chromosomes, which are defined as a mosaic of chunks (a chunk being a segment of contiguous DNA copied from the same donor). Separate curves are generated for each pair of donor populations by summing the number of chunks separated by a given genetic distance, *g,* for each pair of chunks along a haploid chromosome. For example in figure 5.1(a) two genetic distances, x_1 and x_2, are shown along the re-weighted painted chromosome, with both the chunks separated by x_1 being African (yellow), and the chunks separated by distance x_2 being African and Asian (red). (Note that in this example, the haplotypes are coloured by region, but in reality each donor population has a different colour.) The number of times a pair of chunks is separated by a given distance (e.g. x_1) across the entire genome, for a given pair of donor populations, is then calculated, and this is repeated for a range of genetic distances *g,* from 1 to 50 centimorgans (cM) in incremental bin sizes of 0.1cM. Chunk pairs separated by less than 1cM were ignored, as such distances can capture signals more prone to influence from within-population linkage disequilibrium patterns.

To account for phasing "switch errors", a common source of error in phasing (Browning and Browning, 2007), this procedure is repeated both within and between haploids within a sample. These tabulated counts of copying from one donor population to another across the range of g are referred to as admixture decay curves and are found for all pairs of donor populations for each recipient population in turn. This gives one a measure of how donor chunks are distributed along the chromosome at specific genetic distances.

The tabulated admixture curves are then scaled again using the estimated admixture proportions calculated above from the regression of the chunklength copying vec-

tors, and are further divided by the value expected under the assumption that the two chunks in the pair are independent, which gives the final "weight-scaled admixture decay curves" for each pair of populations. These values, which are still tabulated across all values of *g*, for each pair of donor populations, represent the probability of copying from one source population to another across the range of genetic distances. Three such curves are shown in figure 5.1(b) for an analysis involving a Yoruba/Brahui simulated haplotype. In the case of the simulated data, neither the Brahui nor the Yoruba were included as potential donor populations. This mimics the real data analysis, where the sampled donor populations are at best imperfect matches of the original admixing source groups, for example due to the sampling scheme or the original sources being extinct. Here, the curves calculated within and between the Balochi and Mandenka (the two populations most similar to the Brahui and Yoruba, respectively) are shown. Intuitively, the Balochi-Balochi (or Mandenka-Mandenka) curves are showing us that Balochi (or Mandenka) chunks are present in the admixed population, at distances that decay exponentially. The Balochi-Mandenka curve tells us that these chunks are often found close to each other, but have been broken up over time such that a large number of chunks donated from the two populations are observed at close genetic distances. It is the decay of this breakdown over genetic distance that is used to date the admixture event.

The maximum likelihood estimate of the rate parameter λ of an exponential distribution, simultaneously fit to curves from all donor populations is shown. Following arguments stated by Falush et al. (2003), in the case of a single instantaneous admixture event occurring at time λ generations ago in the direct ancestors of the population being studied, we expect the curves informative for this event to be exponentially distributed with rate equivalent to λ (Hellenthal, Busby, et al, in prep). This value of λ therefore represents the date since admixture in generations. A bootstrap resampling procedure is used to find the 95% confidence intervals around this date for each population.

In effect, this process allows one to identify the distribution of chunks along a recip-

ient population's chromosome from all important source populations in a sample, as all important populations will produce curves such as those in figure 5.1(b). Furthermore, this distribution is then used to date when these chunks mixed into a population. Populations with no history of admixture in a recipient population will not produce curves with a clear exponential decay.

Next, we can use the information in the curve 'heights', which we define for each curve as its y-axis value at x=0 minus its y-axis value at x=∞ to inform us about both (1) which pops are representing the same admixing source group and (2) which pops best represent each of the two sources. Intuitively, for two populations A and B, the information for (1) comes from the observation that if the curve for population A and population B has positive 'height', then population A and population B most likely contribute to the same admixing source. In contrast, if the curve for population A (e.g. Balochi) and population B (e.g. Mandenka) has negative "height", then these populations most likely contribute to different admixing source. This can clearly be seen in the three curves in figure 5.1(b), where the central curve has a negative height compared to the two "self" curves. To get at (2), the populations whose "self" curves (e.g. the Balochi-Balochi curve) have the largest heights, especially relative to the noise in the curve at large genetic distances (beyond which we expect any admixture signal to have been completely diminished) are generally speaking identified as the best representatives of their admixing source group. In practice, a PCA is performed on a pairwise matrix of donor population curve heights, after which identification of (1) and (2) is accomplished by using a further regression that simultaneously infers both the proportion of admixture contributed by each source and the copying vector of each source population, the latter defined as mixtures (with mixing values estimated by the regression) of the present-day donor population copying vectors.

The admixture proportions are finally re-estimated as α multiplied by the first admixing source plus 1-α multiplied by the second admixing source. These new admixing proportions replace those initially estimated, and the tabulated scaled admixture decay curves are re-weighted and this process is repeated for five iterations.

Assessing the evidence for admixture

Several metrics collected during the process outlined above are then used to assess the evidence for admixture. These metrics are based on those recorded through the use of the above analytical procedure on the simulated populations described below. Broadly speaking, the assessment metrics concentrate on two key criteria: the first relates to how well the admixture decay curves describe any signals in the data, and the second relates to how well the eigen decomposition describes the variation in the curve matrix. For the former, the "coefficient of determination", i.e. the R^2 value, is calculated from the linear regression for each of the weight-scaled admixture decay curves and the exponential distribution with rate λ. Recall that λ is the estimate of the date of admixture. This reflects how well each pairwise population curve correlates with the distribution represented by the inferred date. I use the largest value of R^2across all curves to assess the presence of a clear date. The idea here is to assess whether the signal, as identified by the admixture decay curves, is consistent across different pairs of populations.

The second major metric relates to the proportion of variance explained, or "fit quality" of the first eigenvector of the PCA of curve heights matrix. Here, the fit quality of the largest eigenvector of this PCA, FQ_1, represents how well a single event describes the height matrix. A high value of FQ_1suggests that the populations involved can clearly be segregated to different sides of the admixture event. These metrics can therefore be used to assess not only how well the PCA decomposition explains the curves, but also whether there is evidence of more than one admixture event.

For my analysis, and based on thresholds taken from the analysis of simulated populations, I characterized admixture in the following way:

1. no admixture - max$(R^2) < 0.7$
2. two source admixture event - max$(R^2) \geq 0.7$ and $FQ_1 \geq 0.975$
3. complex admixture event - max$(R^2) \geq 0.7$ and $FQ_1 < 0.975$

A complex admixture event describes an event where a clear signal of admixture is obtained, but where this admixture may involve multiple sources or may have occurred at multiple dates.

Simulations

Before using this analysis pipeline on the real data, the programs were tested on some simulated data. Simulations were performed by Dr Garrett Hellenthal and involved generating artificial admixture events at 7, 30 and 150 generations in the past between several different populations across the world with several different admixing proportions (see table 5.1 for the full list of simulations). In a similar approach to Price et al. (2006) and Moorjani et al. (2011), simulated populations were generated by stitching together segments of real modern chromosomes from each of the two admixing populations, for example the Brahui and the Yoruba mentioned above, in proportions equivalent to those expected from events where a given population contributes 5, 20 or 50% to that event. So, in the case of the Brahui-Yoruba simulations, first a centimorgan genetic distance x was sampled from an exponential distribution with rate λ corresponding to the time in generations since the admixture event. Then the x cM of the simulated haploid was composed of the x cM of a real data haploid from the Brahui with probability α (the simulation admixing proportion), otherwise it was composed of the x cM of a real data haploid from the Yoruba. To limit the chance of multiple simulated pseudo-individuals copying from the same real data individual at any area of the genome, wherever possible the new haploid sampled was selected from the pool of haploids in the Brahui or Yoruba for which no other previously simulated haploid had copied at the same location. When this was not possible, a haploid was selected at random from the Brahui or Yoruba. A new genetic distance was sampled from the same exponential distribution (λ) and this process was repeated until the entire simulated haploid was generated. Simulated populations were generated for a range of dates, proportions and populations (table 5.1), as well as several no-admixture simulations where populations were generated by mixing two identical or closely related

populations 1,000 generations ago. In total 60 different scenarios were simulated and the metrics described above were identified from their analysis (Hellenthal, Busby, et al. , in prep).

I analysed the results of each simulated population run through the programs outlined above to find the *inferred* date of the admixture event, the proportion of admixture from each source, and the copying vector of each source.[3] These inferred results were then compared to the true sources and dates. As well as recreating the copying vector of the two inferred source populations, it was also possible to reconstruct the inferred admixed population, by using the proportion of admixture estimated by the model and the copying-vectors of the sources. For example, if one source was estimated to contribute 20% to the admixture event, and the second source 80%, the admixed population can be reconstructed by multiplying the first source's copying vector by 0.2, and combine it with the second source copying vector multiplied by 0.8. This *inferred* combined population copying vector can then be compared to the simulated population as an additional test of the ability of the model to recreate the simulated admixture population.

Before the correlations were made, a further standardisation step was required because, as mentioned above, during the analysis of the individuals from the simulated populations, each was barred from copying any genome from either of the two true source populations. So, for example, in the case of the Brahui-Yoruba simulations, when admixture was estimated in the simulated individuals from this "population", the individuals were allowed to copy neither from the Brahui nor the Yoruba, and so these populations were excluded during the curve generation step. Furthermore, when the linear model was used to fit the simulated copying vectors as a mixture of all real copying vectors, the true admixing populations (in this case Brahui and Yoruba) were not included either. In effect, this means that the inferred source copying vectors will differ from the true source copying vectors, because the copying vectors

[3]The results of the analyses of simulated populations were generated and shared by Dr Garrett Hellenthal.

from the true source populations were estimated from the full all individual versus all individual analysis where individuals were allowed to copy from other individuals in their population (i.e. self-copying *was* allowed). To account for this, when finding the correlations, I removed the true source populations from the copying vectors for both the true and inferred sources, and re-calibrated the copying vectors as a proportion of total remaining chunklengths. Thus, for each inferred and true source/population, I compared the proportions of genome copied across all populations as donors, except the two true donors originally used to generate the simulated population.

European Populations

I explored admixture in 16 European populations. An associated study (Hellenthal, Busby, et al, in prep) using the same methods and a very similar dataset found no evidence of significant admixture in the Finnish, GermanyAustrian or any of the northwestern European populations. I therefore did not reanalyse these populations and present the results only for these populations from the alternative analysis prepared for a different study (Hellenthal, Busby, et al, in prep). The remaining populations were all on or close to the periphery of Europe. Additionally, these populations represent the key contact zone between the Mediterranean and Africa on the one side, and eastern Europe and Asia on the other. In each of the 16 European populations I initially assessed the evidence for admixture using the metrics outlined on page 115. For those populations that showed evidence for admixture, I found: the date of admixture; the copying vector of the two sources; and the proportion each source contributed to the event.

Removing the effect of recent historical admixture

The preceding analysis allowed me to reconstruct the copying vectors of the two admixing sources in situations where a significant admixture event had occurred. As a

final step, I compared these copying vectors to each other. In each case, a European-like source was found to contribute the largest proportion to the admixture event. I plotted the copying vectors, grouped by European and non-European sources, to explore the similarities between the pre-admixture European populations.

5.3. Results

Comparisons of the copying of different European population

Figure 5.2 shows comparisons between the European average amount of genome copied from different world-wide regions, for different European populations. The horizontal line at 0 on the y-axis represents the mean across all European populations, and the difference between this for each population is shown, measured in standard deviation units. Absolute values of these comparisons are shown in table A.3. Generally, populations from the same geographic region copy similarly from different parts of the world. North-western European populations copy marginally more than average from within Europe, as well as from SouthCentralAsia, CentralAsia and America. North-eastern European populations also copy more from within Europe than the average, and copy more from the Central and East Asians and America world regions. The southern-central populations copy more widely: from the Caucasus, MiddleEast, NorthAfrica, SouthCentralAsia and Sub-Saharan Africa. Of the south-eastern populations, Croatia and Hungary copy more from within Europe, whilst Bulgaria and Romania copy from the Caucasus and MiddleEast, SouthCentralAsia and the Pacific.

However, there are also clear exceptions to these general patterns. Recall that the Finnish and GermanyAustria populations were grouped together as North-Central Europeans on the basis of geography and their small sample sizes. The Finnish appear more much more similar to the Russians than to GermanyAustria, with GermanyAustria showing a similar copying pattern to north-western European populations.

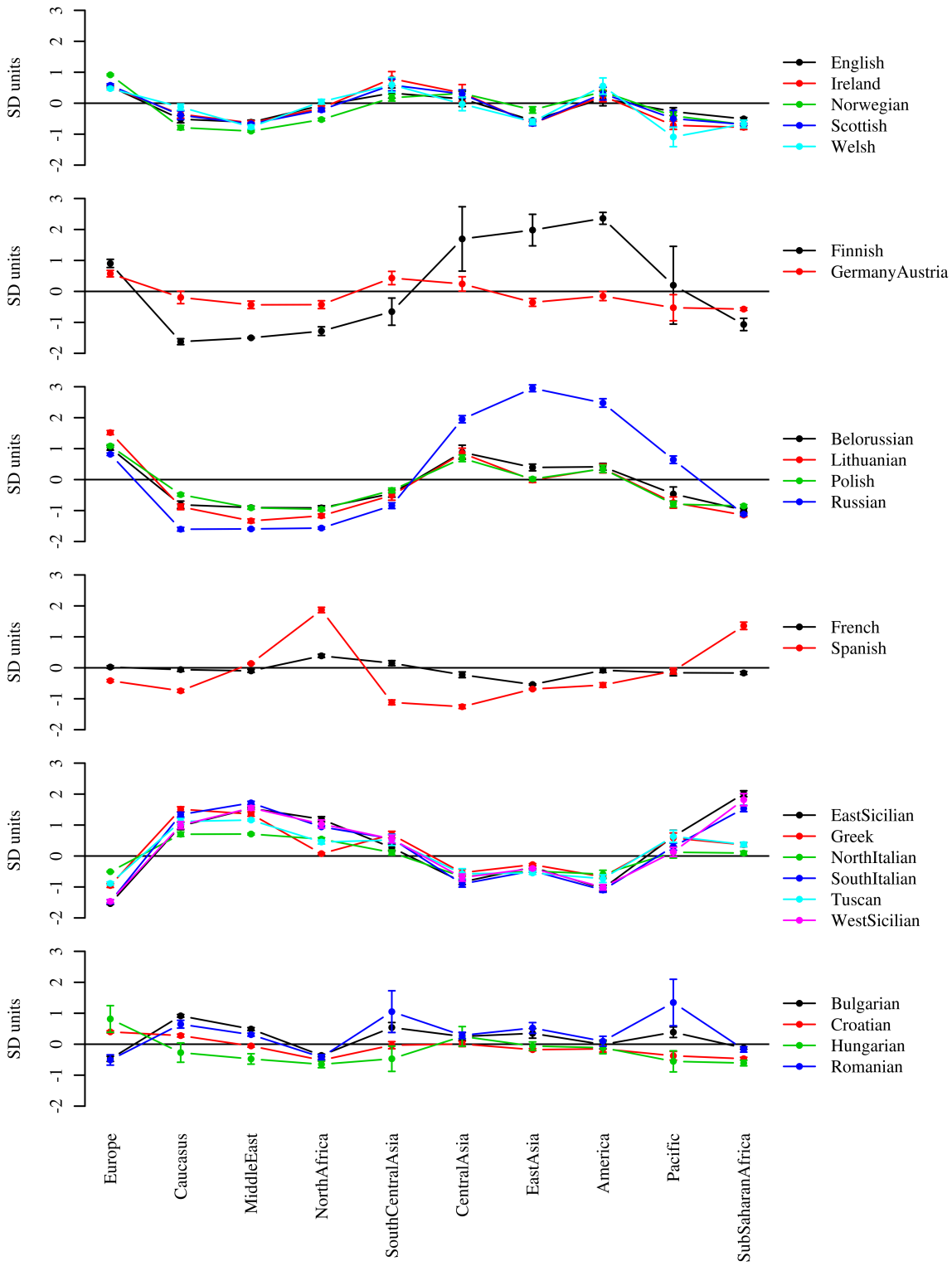

Figure 5.2. The proportion of genome copied by European populations from 10 worldwide regions. The populations are grouped by European region. The mean difference from the European average is shown for each populations. The x-axis shows the donor world regions, and the y-axis displays difference from the mean in units of standard deviation for each European country. Lines are shown to aid visualisation. Error bars are mean standard error.

France, too, shows a similar copying profile to the north-western European populations, unlike the Spanish, who clearly copy more from North and Sub-Saharan Africa. The Russians copy more from Central and East Asians, America, and the Pacific populations than the rest of the north-eastern European populations. The NorthItalians and Tuscan copy less than the rest of south-central Europe from the Caucasus, Middle-East and NorthAfrica, as well as significantly less from Sub-Saharan Africa. The Greeks also copy little more than the European average from NorthAfrica, but do appear to copy to similar levels as the rest of south-central Europe from Sub-Saharan Africa.

Admixture analysis of simulated populations

For each of the simulated populations, I found the correlation between the copying vector of the two inferred sources and the copying vectors from the true sources, as well as the copying vector from the inferred admixed population, recreated from the admixture proportions, and the true admixed populations (table 5.1).

Table 5.1. Results of the admixture analysis on simulated populations. For each simulation the two admixing source populations are shown as Source A and Source B. The true and inferred dates and admixing proportions shown for comparison with confidence intervals for these inferences. The correlations refer to how well the inferred source population copying vectors correlate with the true source population copying vectors.

Source A	Source B	Date Estimates				Proportion Estimates				Correlations / R^2		
		truth	infer	2.5%	97.5%	truth	infer	2.5%	97.5%	A	B	admixed
Brahui	Han	7	8	6	10	0.05	0.06	0.04	0.18	0.97	1	0.998
Brahui	Han	7	7	6	8	0.20	0.20	0.14	0.30	0.967	1	0.997
Brahui	Han	7	7	6	8	0.50	0.48	0.39	0.56	0.995	1	0.997
Brahui	Han	30	31	28	35	0.05	0.06	0.04	0.22	0.981	1	0.999
Brahui	Han	30	32	29	34	0.20	0.21	0.15	0.31	0.968	1	0.997

Source A	Source B	Date Estimates				Proportion Estimates				Correlations / R^2		
		truth	infer	2.5%	97.5%	truth	infer	2.5%	97.5%	A	B	admixed
Brahui	Han	30	29	26	31	0.50	0.49	0.40	0.57	0.995	0.999	0.998
Brahui	Han	150	60	8	102	0.05	0.19	0.06	0.42	0.546	1	1
Brahui	Han	150	136	116	160	0.20	0.23	0.13	0.40	0.995	0.997	0.999
Brahui	Han	150	168	148	184	0.50	0.49	0.40	0.58	0.991	0.996	0.995
Brahui	Yoruba	7	8	6	10	0.05	0.06	0.04	0.16	0.971	1	0.998
Brahui	Yoruba	7	7	6	9	0.20	0.19	0.14	0.27	0.987	1	0.997
Brahui	Yoruba	7	7	6	8	0.50	0.47	0.39	0.55	0.994	0.999	0.994
Brahui	Yoruba	30	31	27	34	0.05	0.07	0.04	0.18	0.979	1	0.998
Brahui	Yoruba	30	29	26	32	0.20	0.20	0.15	0.28	0.988	1	0.997
Brahui	Yoruba	30	30	27	33	0.50	0.46	0.38	0.54	0.994	1	0.995
Brahui	Yoruba	150	74	18	106	0.05	0.10	0.01	0.39	0.916	0.999	1
Brahui	Yoruba	150	133	116	149	0.20	0.19	0.12	0.35	0.995	0.996	0.999
Brahui	Yoruba	150	146	128	164	0.50	0.45	0.36	0.57	0.994	0.99	0.995
Colombian	Han	7	7	4	9	0.05	0.16	0.04	0.55	0.623	1	1
Colombian	Han	7	7	5	9	0.20	0.27	0.16	0.51	0.99	1	1
Colombian	Han	7	7	5	9	0.50	0.42	0.32	0.55	0.997	0.999	1
Colombian	Han	30	25	20	32	0.05	0.10	0.05	0.57	0.833	1	1
Colombian	Han	30	31	26	38	0.20	0.28	0.16	0.52	0.976	0.999	1
Colombian	Han	30	30	26	33	0.50	0.44	0.32	0.55	0.999	0.999	1
French	Brahui	7	12	8	15	0.05	0.08	0.03	0.33	0.971	0.999	0.999
French	Brahui	7	8	6	10	0.20	0.22	0.13	0.45	0.999	0.999	0.999
French	Brahui	7	7	5	9	0.50	0.46	0.31	0.57	1	0.998	0.999
French	Brahui	30	29	20	37	0.05	0.11	0.04	0.35	0.961	0.999	0.999
French	Brahui	30	27	23	33	0.20	0.23	0.13	0.45	0.997	0.999	0.999
French	Brahui	30	33	28	37	0.50	0.47	0.33	0.57	0.999	0.996	0.999
Yoruba	French	7	8	7	10	0.05	0.08	0.04	0.36	0.811	0.996	0.994
Yoruba	French	7	7	6	8	0.20	0.27	0.20	0.39	0.989	0.994	0.991

Source A	Source B	Date Estimates				Proportion Estimates				Correlations / R^2		
		truth	infer	2.5%	97.5%	truth	infer	2.5%	97.5%	A	B	admixed
Yoruba	French	7	7	6	9	0.50	0.46	0.38	0.56	0.998	0.994	0.993
Yoruba	French	30	32	27	36	0.05	0.12	0.06	0.37	0.735	0.995	0.993
Yoruba	French	30	29	26	32	0.20	0.28	0.19	0.40	0.993	0.994	0.992
Yoruba	French	30	31	28	33	0.50	0.47	0.38	0.56	0.998	0.994	0.992
Yoruba	French	150	86	30	106	0.05	0.19	0.05	0.57	0.354	0.995	0.995
Yoruba	French	150	124	104	145	0.20	0.21	0.15	0.45	0.988	0.996	0.994
Yoruba	French	150	151	135	170	0.50	0.45	0.37	0.55	0.994	0.994	0.992

In almost all cases, the admixture modelling process recreated the copying vector of the source with a high degree of fidelity (R^2>0.9). The linear model had more difficulty finding the proportions of older events. The dates and proportions estimated by the linear model were also all within the true values used to create the simulated populations (table 5.1). These combined results suggest that the linear model can confidently recreate the sources of admixture events.

Admixture analysis of European populations

Table 5.2 shows the assessment of admixture for the European populations. No admixture was observed in all north western and north central European populations, as well as one north eastern population, Polish, and the two most northern south central European populations: NorthItalian and Tuscan. Of the remaining populations, the majority showed evidence for a single admixture event between two source populations, with the Spanish, French, SouthItalians, Lithuanians, and Belorussians showing evidence of more complex, multi-way admixture.

Figure 5.3 shows the admixing proportions, dates and admixing sources of all European populations that show evidence for admixture. The confidence intervals

Table 5.2. Evidence for admixture in European populations. Populations marked with an asterisk were not analysed by the author and were analysed and shared by Dr Garrett Hellenthal.

Region	Population	R^2	FQ_1	admixture
NorthCentralEurope	Finnish*	0.43	0.9	no-admixture
	GermanyAustria*	0.20	0.99	no-admixture
NorthEastEurope	Belorussian	0.88	0.96	complex-event
	Lithuanian	0.77	0.97	complex-event
	Polish	0.59	0.98	no-admixture
	Russian	0.96	1	one-event
NorthWestEurope	Norwegian*	0.65	0.94	no-admixture
	English*	0.19	0.99	no-admixture
	Ireland*	0.14	0.96	no-admixture
	Scottish*	0.16	0.91	no-admixture
	Welsh*	0.14	0.96	no-admixture
SouthCentralEurope	NorthItalian	0.58	0.99	no-admixture
	Tuscan	0.54	1	no-admixture
	SouthItalian	0.92	0.95	complex-event
	WestSicilian	0.93	0.99	one-event
	EastSicilian	0.92	0.99	one-event
	Greek	0.79	1	one-event
SouthEastEurope	Bulgarian	0.88	1	one-event
	Romanian	0.92	1	one-event
	Hungarian	0.76	1	one-event
	Croatian	0.80	0.99	one-event
SouthWestEurope	French	0.70	0.9	complex-event
	Spanish	0.96	0.85	complex-event

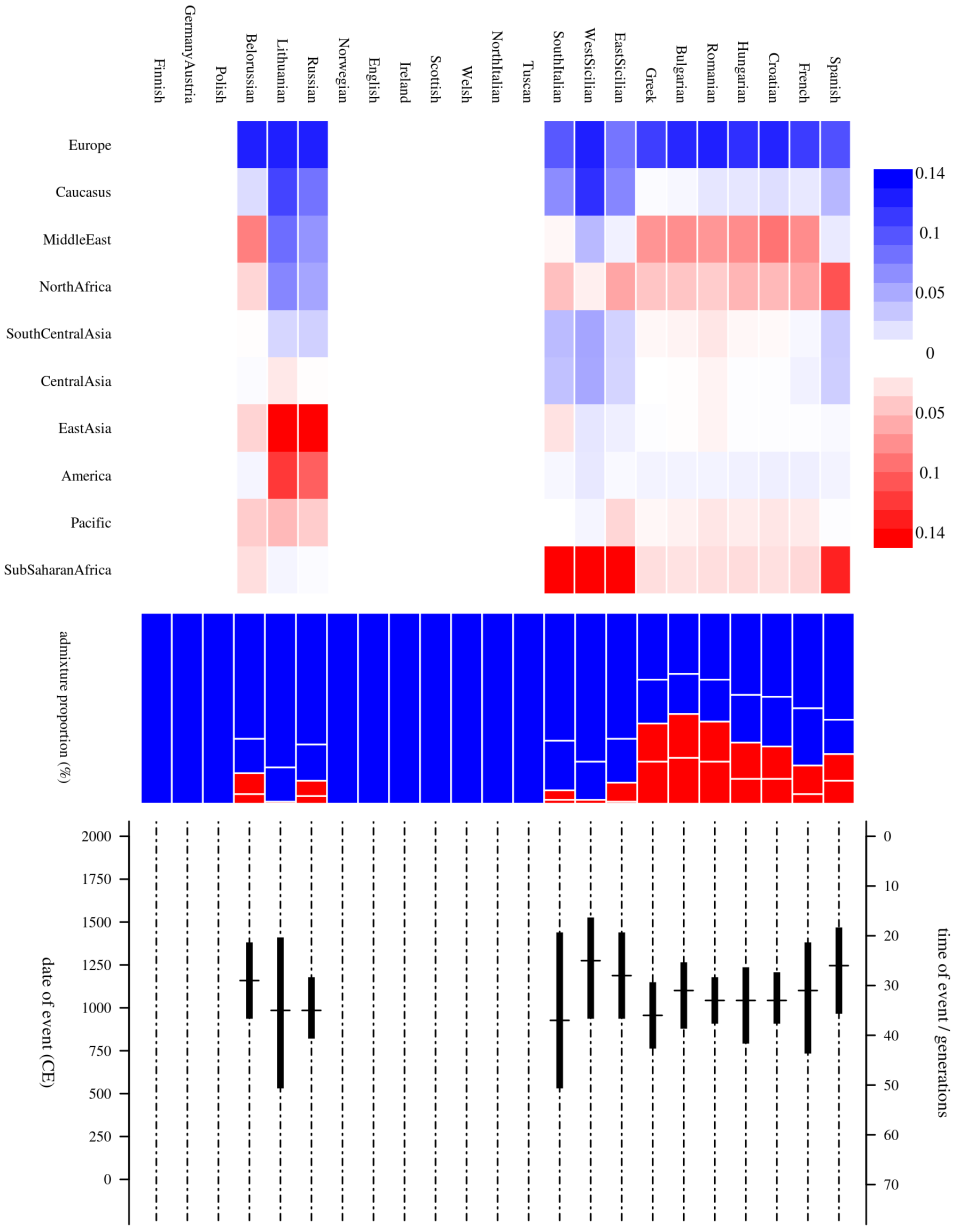

Figure 5.3. Dates, proportions and sources of admixture in European populations. European countries are shown on the x-axis of all panels. The top panel shows a comparison of the two source copying vectors. Here, the more blue a region's contribution to the admixture event of the population, the more it contributes to the major admixing source. The world regions are shown on the y-axis. The scale shows the difference in proportion that is donated by each source region. The more red the region's contribution, the more it contributes to the minor source. The middle panel gives the proportions of the two sources. The major source, which was always the more European source, is shown in blue and the minor source is in red. The point estimate of the proportion is shown where blue meets red, and the confidence region around the proportion is shown with the two white lines. The bottom panel gives, for all European populations where admixture was identified, the date and confidence interval of this admixture event.

around the proportions are represented by the two white lines either side of the central white line in the bars. The sources are shown summarised into the same 10 world regions as before. To colour the sources, I subtracted the copying vector of the minor source from that of the major source. All positive values are coloured in blue of varying intensities, and all negative values are coloured red. The effect of this is to show the relative contributions of each of the ten world regions to the sources on either side of the event.

In all cases, the admixture analysis produced a European-like source on the one hand and a non-European source on the other. For those populations where admixture was observed, similar dates were found for the date of admixture. The Russians and Lithuanians get an identical date for admixture from an East Asian source at approximately 1,000CE. The Russians appear to have had a greater proportion for admixture from this source. The Belorussians show a later date with a more Middle Eastern source admixing around 1,200CE, although all three admixed north-eastern European population dates have overlapping confidence intervals. The SouthItalians and Sicilians all appear to be admixed between a strongly sub-Saharan African group and a European group. The non-European source of the SouthItalian and EastSicilian admixture also seems to have a North African component that is missing from the WestSicilians. In all three populations the proportion of admixture from the non-European source is very small. The date of 800CE in the SouthItalians is earlier than the Sicilians at around 1,200CE, although again the confidence intervals are overlapping. Interestingly, the Spanish also have a clearly sub-Saharan African and North African admixing source dating to the same period around 1,200CE which appears to have given a greater proportion of DNA to the current Spanish population. The Greeks and Balkan (south east Europe) populations have similar admixture sources and dates, but vary in the proportion of admixture they obtain from a Middle-Eastern like source population with a date around 1,000CE. On the basis of this analysis the Greeks appear much more similar to their geographically proximate Balkan neighbours than the Italians.

Figure 5.4 shows the copying vectors of the European sources in the same format as figure 5.2. This time, because the admixture programs work at the population level, there are no standard error bars and I have plotted the difference in the more European admixing sources compared to the European averages. The top four plots contain, as before, populations plotted by European region. The bottom two plots show the same populations grouped as all of eastern Europe, and south central and south west Europe, respectively. Note the high concordance between the copying vectors of these populations in the lower two plots, in stark contrast to those in figure 5.2. Figure 5.5 shows exactly the same populations, but this time with the non-European admixing sources compared with the European average. The Lithuanians and Russian non European sources have a clear spike from East Asia, suggesting that these sources copy ~50 standard deviations more from East Asia than the average European population. The Spanish, and to some extent French, non European source copy from North Africa and sub-Saharan Africa to a much greater extent than the European average. The Greek non-European admixing source differs from the SouthItalian and Sicilian source mainly due to the lack of sub-Saharan African copying, with the WestSicilian showing a large difference to EastSicilian in the amount of sub-Saharan DNA copied by its non-European source. Again the bottom two plots show the populations grouped together, this time the fifth plot shows the south eastern populations with Greeks, French and Belorussians, and the bottom plots shows the south central and Spanish non-European copying vectors. Whilst there is some similarity in the southern populations, the south eastern and other populations show similar trends but with differing amounts of copying from Middle Eastern and Central Asian populations. Interestingly, the French show a similar trend to the south eastern populations with a markedly large amount of North African DNA in their non-European copying source. Note also the inflated copying from SouthCentral Asia and Central Asia in the Romanian non-European source, as compared to the other south eastern populations.

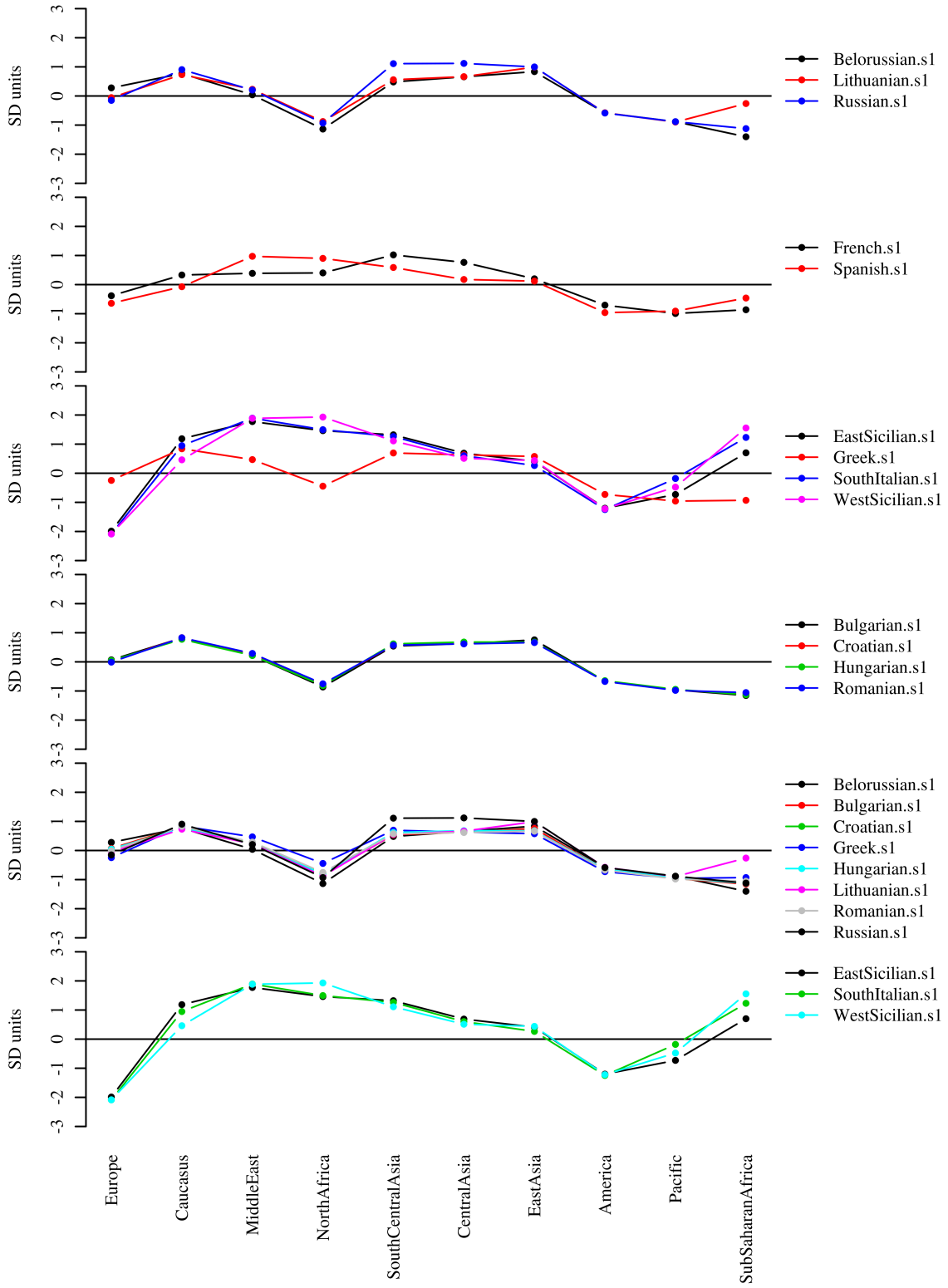

Figure 5.4. The proportion of genome copied by European source populations. The x-axis shows the donor world regions, and the y-axis displays difference from the mean in units of standard deviation for each European country.

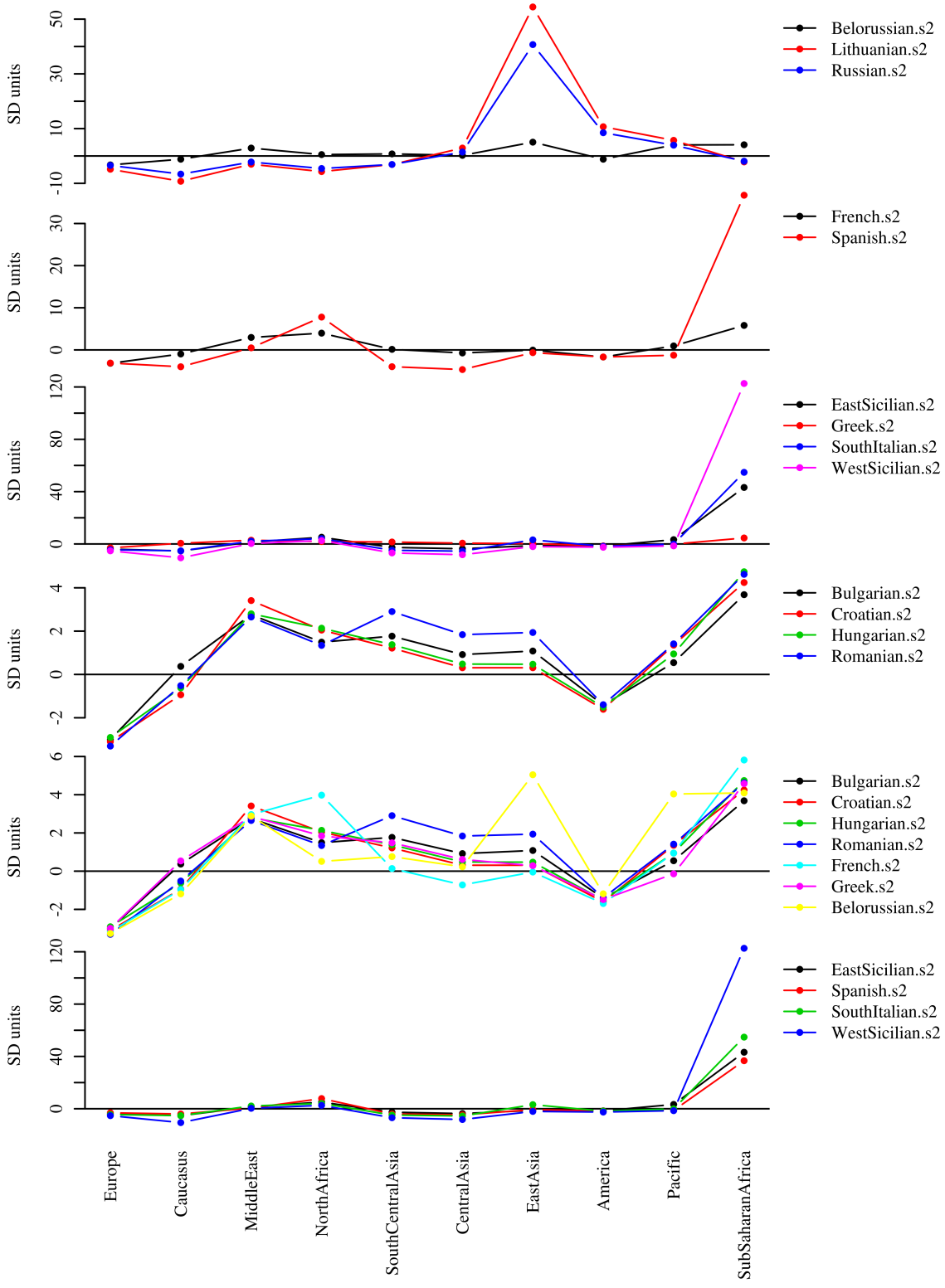

Figure 5.5. The proportion of genome copied by non-European admixing sources. The x-axis shows the donor world regions, and the y-axis displays difference from the mean in units of standard deviation for each European country.

5.4. Discussion

In this chapter I have explored the ancestral relationships between European populations by investigating the presence of signals of genome-wide admixture. Admixture is by no means a universal attribute in the populations included in this study, which covers the majority of Europe. In particular, no evidence for admixture was found in all of the British and Scandinavian populations, as well as GermanyAustria. This may not be the result of a lack of admixture in these populations, but rather a lack of power in the current method to highlight admixture in closely related populations with small sample sizes. Future work on north west European populations, for example that proposed by the Peopling of the British Isles project Winney et al. (2012), with increased sample sizes and better coverage than the current study, should provide a better analysis of the fine-grained picture of admixture in these populations. For the moment however, it is still important to note that when viewed with similar sample sizes to other European populations and from a world-wide perspective, north west European populations copy less DNA from outside of Europe than the more southern and eastern populations.

With respect to the south western, south central and eastern European populations, recent admixture appears to be a common characteristic. As previously observed from studies on uniparental markers, southern Europe has significant levels of African admixture (Adams et al., 2008; Capelli et al., 2009; Cerezo et al., 2012). Previous genomic studies have also shown there to be African admixture in southern European populations (Auton et al., 2009; Moorjani et al., 2011), with Moorjani and colleague dating the admixture to approximately 55 generations ago. Here, using a broader range of southern European populations, including five from the Italian peninsular, I have shown that there are putative differences in the origin and amount of DNA copied in these populations from Africa. Spain had a long and famous period of subjugation under the Moors, lasting almost 700 years (Roberts, 2007). This is attested in the genes of Spanish people today through the presence observed here of a clear North African compon-

ent in the non-European admixing source (figures 5.3 and 5.5). The date proposed here of 26 generations is roughly a half of that provided by Moorjani et al. (2011) dating towards the latter part of the Islamic conquest rather than the beginning as previously suggested.

One intriguing advance of the method used here is its ability to dissect the contributions from different yet closely related populations. In this way, it is possible to show that the African admixture in Sicily and Southern Italy (in addition to a proportion from Spain) comes mainly from sub-Saharan, rather than North, Africa. The point estimate of the admixture event in the SouthItalians is perhaps earlier than that in the Sicilian populations, although this is not significantly so, at around 1,000 CE as compared to 1,250CE , and in all populations is proportionally very small. The occurrence of sub-Saharan African DNA in these populations at this time is certainly thought-provoking, and perhaps relates again to the Islamic conquest of the Mediterranean, but this time to the slaves that were brought with the Muslims (Metcalfe, 2009), rather than the Moors or Saracens themselves, as appears to be the case in Spain. The Islamic influence in Italy and Sicily at this time was markedly different to that in Spain. The establishment of Muslim rule of Sicily was slow and piecemeal (Metcalfe, 2009), and in Italy also, the Muslims were unable to conquer the peninsular in the same fashion as they did in Spain. Perhaps this disparity is why there is a marked difference in the North African contribution to Spain as compared to Italy.

The south eastern European populations from the Balkans also show evidence for admixture. Generally speaking, these populations can be split into two groups based on their copying vectors (figure 5.2), with the Hungarians and Croatians copying less from outside of Europe than the Bulgarians and Romanians. Interestingly, when the copying vectors of the European admixing source of these populations are compared, they are remarkably similar (figure 5.4) and differences between contemporary Balkan populations presumably being due to a combination of the increased Asian component in the Bulgarian and Romanian admixing sources (figure 5.5), and the difference in the proportions of admixture observed in these populations (figure 5.3).

The date of admixture in these populations all cluster around 1,100CE, a tumultuous period in Balkan history, when the Byzantine Empire consolidated the Greek and Turkish archipelagos to the south, and the Kingdoms of Poland and Germany lay claim to much of the land to the north (Roberts, 2007). Figure 5.4 shows that, in fact, the European sources of admixture in both south and north eastern European populations at this time were remarkably similar. If we view these European admixing sources as the indigenous population in these areas at this time, then this result suggests a level of homogeneity across eastern Europe at this time. Moreover, the clear presence of an East Asian component in the Lithuanian and Russian non-European admixing sources further suggests that the current differentiation between populations in the east of Europe is due to a difference in the make-up of the sources of *recent* admixture, within the last thousand years.

Developing this idea further, it is possible to highlight three, or perhaps four, different patterns of admixture within the current European population. The first is shown in the north western European populations. The shape of the copying vectors of these populations in figure 5.2, whilst similar across these populations, is different from other European populations and is characterised by greater copying from within Europe and Central Asia than other European populations. The second pattern is Mediterranean in origin and here is characterised by the copying vectors of the European admixing sources of Southern Italy and Sicily (figure 5.4). The third pattern is pan-eastern European. A putative fourth pattern involves the European admixing source of the French, which could conceivably be grouped with the northern Italian populations. Strong identification of this component requires further exploration of the copying vectors of this part of Europe, but its presence does, at least in theory, seem plausible given the location of the three other components.

The picture that begins to emerge then, is one of greater similarity in the recent past across large parts of Europe, with respect to the relationship between populations and putative non-European sources. The current analysis aims to explain contemporary genomic heterogeneity by highlighting the effects of recent admixture on peripheral

European populations from different non-European sources. Recent events are therefore important because it appears to be recent admixture that differentiates European populations, albeit on top of a somewhat structured substrate. As I mentioned in the introduction to this chapter, this observation, whilst intuitive, has had little support in previous work, because of previous authors' obsessions with Neolithic and Paleolithic contributions to Europe. Whilst the analysis presented here clearly shows that structure was present in Europe prior to these more recent admixture events, it also provides a fresh appreciation of the importance of historical events in the shaping of the European gene pool.

6. The Peopling of Europe: a synthesis

In this thesis I have aimed to provide fresh insight into the Peopling of Europe. Whilst achieving this has required the use of two different genetic analytical paradigms, the Y chromosome and autosomal SNP data, there are, I believe, two major threads that link the work together. The first is the, perhaps obvious, observation that European populations are structured genetically. Of course, this is no new insight: Cavalli-Sforza and colleagues showed that European population genetic structure is clinal with classical markers (Cavalli Sforza and Feldman, 1976; Cavalli-Sforza et al., 1994). However, the work presented here shows that this is the case at increasingly finer resolution for both genetic systems employed. This greater granularity also allows for more complicated structures to be observed beneath the overarching theme of "clinality". In chapter 2, we saw that whilst Y chromosome haplogroup R-M269 is classicly clinal in its distribution, the variation displayed by several sub-haplogroups is not. So for this locus, the emerging historical picture needs more than the spread from the Neolithic transition to explain its current structure. Future work investigating the putatively causal demographic scenarios to these patterns will aid a better understanding of the genetic history of Europe. The fine-scale autosomal structure of Europe observed in chapter 4, also implies that there is more information in the genomes of Europeans than that which can be explained by PCA maps alone. When incorporating haplotypic information, it is possible to interrogate genomes to understand the deeper relationships between populations, and to begin to understand where individuals and populations have got their DNA from. The next few years, with greater and greater access to genetic resequencing data, will surely provide some surprises as we delved deeper

into the genetic records of human history.

The second major theme identified in this thesis relates to understanding how this structure has arisen. This requires an historical perspective and necessarily also requires an appreciation of the time over which variation has arisen. The work on the Y chromosome in chapters 2 and 3 addresses the contentious issue of dating Y chromosome lineages. The Y chromosome has, I believe, a future in human evolutionary studies, but this must surely embrace the genomics age. For example, the analysis of whole Y chromosomes will greatly aid our understanding of male-mediated gene flow, and the exploration of patterns of their variation, based on high-quality sequence data, together with whole mtDNA sequences, will allow for fresh and new insight into the different demographic roles of men and women in different human societies.

The admixture analysis presented in chapter 5 builds on the *fineSTRUCTURE* analysis of chapter 4 and shows that it is now possible to recreate the admixing sources of different human populations. As this methodology develops, in tandem with the generation of more detailed datasets, it becomes increasingly possible to recreate the genomes of the past. Developments in the sequencing of DNA from ancient skeletons and fossils will also give us the ability to calibrate these inferences. It is not absurd to imagine that in the future we will not only be able to reconstruct the pedigrees and phylogenies of different human populations, but also the genomes of their ancestors. Indeed work in this area has already begun (Young, 2011). This will open a window into our past like no other, and allow us to gain a profound understanding of our evolutionary past.

Greater Significance

Over and above the inherent interest that comes from investigating human origins, what is the importance of the work presented here? Understanding the baseline genetic structure of populations can help medical geneticists. For example, in a recent

comparison of British men with coronary heart disease, individuals belonging to Y chromosome haplogroup I had a ~50% increased incidence of the disease compared to individuals from other haplogroups (principally haplogroup R-M269: Charchar et al., 2012). Although there is no suggestion of a causal link between carrying a haplogroup I chromosome and heart disease, this result suggests that there is some benefit, in terms of identifying populations that are at increased risk, from understanding the structure of Y chromosome variation. The genomics approach suggested above will aid with establishing whether there is a molecular basis for such links. Moreover, understanding the distribution of Y chromosomes at increasing granularity could help with prediction of these particular diseases in the population moving forward.

The work presented here on autosomal SNPs will also aid medical study in the future. Understanding the relatedness, underlying structure, and admixture history of human populations is important information to incorporate into genome wide association studies (Price et al., 2008). The explosion in sequencing technology means that within the next few years it will be financially tractable to compare many thousands of genomes for any particular study. Indeed, recent studies have compared many thousands of individuals in a GWAS context (e.g. Sawcer et al., 2011). Elucidating the underlying, latent genetic structure of populations will greatly aid this endeavour and the quest to find important risk alleles.

Infectious disease epidemiology is another field that may benefit from the research and methods presented in this thesis. Understanding the interaction between human genomes and our pathogens is a major research focus (Hill, 2012). For example, identifying signatures of natural selection in populations that have been exposed to a disease for many years can help researchers understand the molecular targets of such diseases (Andersen et al., 2012). The *ChromoPainter/fineSTRUCTURE* approach provides a methodology whereby the evolutionary history of both human and pathogen population can be tracked in analogous fashions, that will greatly aid the understanding of how an why diseases spread. A corollary of the application of the methodologies used in chapters 4 and 5 is that they demonstrate that sophisticated algorithms can reduce

genomic scale data into anthropologically meaningful units that can be used to understand population history. As the scale of genomic datasets continues to increase, this framework provides the ability to condense such data into manageable pieces.

Future years will undoubtedly herald a greater understanding of the story of human history, and genetics will be an integral key to a fully interdisciplinary approach to this endeavour. The idea that we will increasingly be able to penetrate the story of the past using the information from the DNA within us is both exciting and daunting, because of the potential explanations that these stories may offer to the varied and diverse human populations that we see today. Moreover, because our genes offer a window through time, we will not only be able to piece together the recent history of Europe, but also look further back through evolutionary time to the very dawn of our species in Africa and beyond. But perhaps we should not be too surprised that the tiny molecule of DNA within all of us could have such expository power, as even eighty years before its discovery Darwin was typically precise when he observed that "Man still bears in his bodily frame the indelible stamp of his lowly origin".

Acknowledgements

Human genetics is by its nature a collaborative endeavour and there is a huge number of people without whom this project would not be possible. In order to compile a database of samples from across Europe, I have had to contact a great many people who have shared their DNA samples with us. The technical nature of genome-wide analysis also requires teamwork, and I have received a great deal of help with this in that regard. A full list appears at of my collaborators appears after this section.

Specifically, however, a few people require extra mention: Francesca Brisighelli, Jim Wilson, Paula Sanchez-Diz, Eva Ramos-Luis, Mark Thomas, Dan Bradley, Walter Bodmer, Bruce Winney, Ellen Royrvik, Leonor Gusmao, Gianmarco Ferri, Conrado Martinez-Cadensas and Marielle Vennemann all gave me access to already partially genotyped samples. FB is technically the best lab worker I have ever met, and taught me a lot of new protocols and tricks in the laboratory. Jim Wilson requires much further credit for providing additional genome chip data, as well as characterising most of the R-M269 SNPs that are the basis for my Y chromosome work. His enthusiasm for the Y chromosome is infectious and he was always a fountain of new thoughts and ideas. He is also a great person to know at a conference.

Further assistance in the lab was gratefully received from Rory Bowden, and assistance with technical aid and computing has come from Garrett Hellenthal, Rune Lynsgoe, Eleni Giannoulatou and Mihai Duta at the Oxford Super Computer. In particular, Garrett has been a huge help and was always willing to lend a hand which I have very much appreciated. I would have been marooned on an island of incomprehension in

an R-infested sea of Unix without him. Thanks also to Simon Myers for useful initial discussions on the analysis of large scale SNP data, and for agreeing to be part of my supervisory team. I would also like to thank my examiners, Drs Bruce Winney and Agnar Helgarson, for many constructive comments on the final copy of the thesis.

Thanks to my friends on William Allen Crescent, Michelle, Nick and Tom for housing me during the week and for helping trim back the sails after a hard day at sea. Thanks to the other permanent member of the Capelli group, Sarah Marks, who in return for all my amazing jokes never shied away from helping me out whenever she could.

Thanks also to my supervisor Cristian Capelli for choosing such a diverse and stimulating project and for all of our fruitful discussions, many of which involved debating whilst pointing at maps of Europe in his office. He was always encouraging and often excited about our work, and provided a fantastic infrastructure within which to work. It is always a pleasure to debate and explore the genetic history of Europe with you over a beer in the pub. I am also extremely grateful for the funding of my DPhil which came from a BBSRC studentship and a scholarship from Somerville College.

Finally, I thank my wife Angel, who has put up with so many weeks apart and has weathered it with all her grace and wit. Always sympathetic to my need to work and the peaks and troughs that have accompanied these last years, I honestly could not have done this without you.

Contributors

Table 6.1. List of the all collaborators who have contributed to the work presented in this thesis. The majority of indviduals who contributed to the Y chromsome work did so by sharing DNA. For the genomic analyses, DNA was again shared, but I was also helped by several key collaborators with the statistical and technical analysis of genome-wide data.

name	chapter	contribution
Francesca Brisighelli	2	initially genotyped Italian and Greek samples
Paula Sanchez-Diz	2	genotyped French samples
Eva Ramos-Luis	2	genotyped French samples
Conrado Martinez-Cadensas	2	provided and initially genotyped Spanish samples
Mark G. Thomas	2	provided and initially genotyped English samples
Daniel G. Bradley	2	provided and initially genotyped Irish samples
Leonor Gusmao	2	provided and initially genotyped Portuguese samples
Bruce Winney	2	provided and initially genotyped English samples
Walter Bodmer	2	provided English samples

name	chapter	contribution
Marielle Vennemann	2	provided and initially genotyped German samples
Valentina Coia	2	provided Italian samples
Francesca Scarnicci	2	provided Sicilian samples
Sergio Tofanelli	2,4	provided and initially genotyped Daghestani, Sardinian, Corsican samples
Giuseppe Vona	2	provided Daghestani samples
Rafal Ploski	2	provided Polish samples
Carla Vecchiotti	2	provided Italian samples
Tatijana Zjemunik	2,4	provided Croatian samples
Igor Rudan	2	provided Croatian samples
Sena Karachanak	2,4	provided Bulgarian samples
Draga Toncheva	2,4	Provided Bulgarian samples
Gianmarco Ferri	2	provided and initially genotyped Albanian samples
Cesare Rapone	2	provided Ukrainian samples
Tor Hervig	2	provided Norwegian blood samples
Torolf Moen	2	provided Norwegian blood samples
James F. Wilson	2,4	genotyped Scottish samples, provided genotyped British samples
Paolo Anagnostou	2,4	provided Greek samples
Francesco Cali	4	provided Sicilian samples
Valentino Romano	4	provided Sicilian samples
Gerard LeFranc	4	provided Moroccan and Tunisian samples
Rene Herrera	4	provided UAE samples
Garrett Hellenthal	4,5	provided technical assistance with QC, phasing, running *ChromoPainter* and *fineSTRUCTURE*
Simon Myers	4,5	provided assistance in planning analyses

name	chapter	contribution
Daniel Lawson	4	provided assistance in implementing and running *fineSTRUCTURE*
Cristian Capelli	2,3,4,5	helped conceive and plan all four chapters

References

Achilli, A., C. Rengo, C. Magri, V. Battaglia, A. Olivieri, R. Scozzari, F. Cruciani, M. Zeviani, E. Briem, V. Carelli, P. Moral, J. Dugoujon, U. Roostalu, E. Loogväli, T. Kivisild, H. Bandelt, M. Richards, R. Villems, A. Silvana Santachiara-Benerecetti, O. Semino, and A. Torroni (2004). The molecular dissection of mtDNA haplogroup h confirms that the Franco-Cantabrian glacial refuge was a major source for the european gene pool. *American Journal of Human Genetics 75*(5), 910--918.

Adams, S., T. King, E. Bosch, and M. Jobling (2006). The case of the unreliable SNP: recurrent back-mutation of y-chromosomal marker p25 through gene conversion. *Forensic science international 159*(1), 14--20.

Adams, S. M., E. Bosch, P. L. Balaresque, S. J. Ballereau, A. C. Lee, E. Arroyo, A. M. López-Parra, M. Aler, M. S. G. Grifo, and M. Brion (2008, December). The genetic legacy of religious diversity and intolerance: Paternal lineages of christians, jews, and muslims in the iberian peninsula. *The American Journal of Human Genetics 83*(6), 725--736.

Alexander, D. H., J. Novembre, and K. Lange (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research 19*(9), 1655 --1664.

Ammerman, A. and L. Cavalli Sforza (1984). *The Neolithic Transition and the Population Genetics of Europe.* Princeton, US: Princeton University Press.

Ammerman, A. J. and L. L. Cavalli-Sforza (1971). Measuring the rate of spread of early farming in europe. *Man 6*(4), 674--688.

Amos, W. and D. Rubinstzein (1996). Microsatellites are subject to directional evolution. *Nature genetics 12*(1), 13--14.

Andersen, K. G., I. Shylakhter, S. Tabrizi, S. R. Grossman, C. T. Happi, and P. C. Sabeti (2012, March). Genome-Wide scans provide evidence for positive selection of genes implicated in lassa fever. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*(1590), 868--877.

Anthony, D. W. (2007). *The Horse The Wheel and Language: How Bronze-Age Riders from the Eurasian Steppes shaped the Modern World*. Princeton, US: Princeton University Press.

Auton, A., K. Bryc, A. R. Boyko, K. E. Lohmueller, J. Novembre, A. Reynolds, A. Indap, M. H. Wright, J. D. Degenhardt, R. N. Gutenkunst, K. S. King, M. R. Nelson, and C. D. Bustamante (2009, February). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Research 19*(5), 795--803.

Balaresque, P., G. R. Bowden, S. M. Adams, H. Leung, T. E. King, Z. H. Rosser, J. Goodwin, J. Moisan, C. Richard, A. Millward, A. G. Demaine, G. Barbujani, C. Previderè, I. J. Wilson, C. Tyler-Smith, and M. A. Jobling (2010, January). A predominantly neolithic origin for european paternal lineages. *PLoS Biol 8*(1), e1000285.

Balaresque, P., E. Parkin, L. Roewer, D. Carvalho-Silva, R. Mitchell, R. van Oorschot, J. Henke, M. Stoneking, I. Nasidze, J. Wetton, P. de Knijff, C. Tyler-Smith, and M. Jobling (2009). Genomic complexity of the Y-STR DYS19: inversions, deletions and founder lineages carrying duplications. *International Journal of Legal Medicine 123*(1), 15--23.

Ballantyne, K. N., M. Goedbloed, R. Fang, O. Schaap, O. Lao, A. Wollstein, Y. Choi, K. van Duijn, M. Vermeulen, S. Brauer, et al. (2010). Mutability of Y-Chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. *The American Journal of Human Genetics 87*, 341--353.

Ballantyne, K. N., V. Keerl, A. Wollstein, Y. Choi, S. B. Zuniga, A. Ralf, M. Vermeulen, P. de Knijff, and M. Kayser (2012, March). A new future of forensic y-chromosome ana-

lysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Science International. Genetics 6*(2), 208--218. PMID: 21612995.

Balloux, F. (2009, September). The worm in the fruit of the mitochondrial DNA tree. *Heredity*.

Balloux, F., L. L. Handley, T. Jombart, H. Liu, and A. Manica (2009, July). Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. *Proceedings of the Royal Society B: Biological Sciences 276*(1672), 3447--3455.

Bandelt, H., P. Forster, and A. Röhl (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution 16*(1), 37--48.

Banks, W. E., F. d'Errico, A. T. Peterson, M. Vanhaeren, M. Kageyama, P. Sepulchre, G. Ramstein, A. Jost, and D. Lunt (2008). Human ecological niches and ranges during the LGM in europe derived from an application of eco-cultural niche modeling. *Journal of Archaeological Science 35*(2), 481–491.

Bar-Yosef, O., O. Bar-Yosef, and D. Philbeam (2000). The middle and early upper paleolithic in southwest asia and neighboring regions. In *The Geography of Neandertals and Modern Humans in Europe and the Greater Mediterranean*, pp. 109*156. Cambridge, MA, USA: Peabody Museum of Archaeology and Ethnology.

Barker, G. (2006). *The Agricultural Revolution in Prehistory: why did foragers become farmers?* Oxford, UK: Oxford University Press.

Batini, C., G. Ferri, G. Destro-Bisol, F. Brisighelli, D. Luiselli, P. Sánchez-Diz, J. Rocha, T. Simonson, A. Brehm, V. Montano, N. E. Elwali, G. Spedini, M. Eugenia D'Amato, N. Myres, P. Ebbesen, D. Comas, and C. Capelli (2011, April). Signatures of the preagricultural peopling processes in sub-Saharan africa as revealed by the phylogeography of early y chromosome lineages. *Molecular Biology and Evolution EPub ahead of print*. PMID: 21478374.

Battaglia, V., S. Fornarino, N. Al-Zahery, A. Olivieri, M. Pala, N. M. Myres, R. J. King, S. Rootsi, D. Marjanovic, D. Primorac, R. Hadziselimovic, S. Vidovic, K. Drobnic, N. Durmishi, A. Torroni, A. S. Santachiara-Benerecetti, P. A. Underhill, and O. Semino (2008, December). Y-chromosomal evidence of the cultural diffusion of agriculture in southeast europe. *European Journal of Human Genetics 17*(6), 820--830.

Bauchet, M., B. McEvoy, L. Pearson, E.E. Quillen, T. Sarkistan, K. Hovhannesyan, R. Deka, D. Bradley, and M. Shriver (2007, May). Measuring european population stratification with microarray genotype data. *The American Journal of Human Genetics 80*(5), 948--956.

Beckwith, C. I. (2006). *Empires of the Silk Road: A History of Central Eurasia from the Bronze Age to the Present*. Princeton, US: Princeton University Press.

Behar, D., M. van Oven, S. Rosset, M. Metspalu, E.-L. Loogväli, N. Silva, T. Kivisild, A. Torroni, and R. Villems (2012, April). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *The American Journal of Human Genetics 90*(4), 675--684.

Behar, D., R. Villems, H. Soodyall, J. Blue-Smith, L. Pereira, E. Metspalu, R. Scozzari, H. Makkan, S. Tzur, D. Comas, J. Bertranpetit, L. Quintana-Murci, C. Tyler-Smith, R. Wells, and S. Rosset (2008). The dawn of human matrilineal diversity. *American Journal of Human Genetics 82*(5), 1130--1140.

Behar, D. M., D. Garrigan, M. E. Kaplan, Z. Mobasher, D. Rosengarten, T. M. Karafet, L. Quintana-Murci, H. Ostrer, K. Skorecki, and M. F. Hammer (2004, March). Contrasting patterns of y chromosome variation in ashkenazi jewish and host non-Jewish european populations. *Human Genetics 114*(4), 354--365.

Behar, D. M., B. Yunusbayev, M. Metspalu, E. Metspalu, S. Rosset, J. Parik, S. Rootsi, G. Chaubey, I. Kutuev, G. Yudkovsky, et al. (2010). The genome-wide structure of the jewish people. *Nature 466*(7303), 238–242.

Benedictow, O. (2004). *The Black Death 1346–1353: The Complete History*. Boydell Press.

Bermúdez de Castro, J., M. Martinón-Torres, A. Gómez-Robles, L. Prado-Simón, L. Martín-Francés, M. Lapresa, A. Olejniczak, and E. Carbonell (2011). Early pleistocene human mandible from sima del elefante (TE) cave site in sierra de atapuerca (Spain): a comparative morphological study. *Journal of Human Evolution 61*(1), 12--25.

Betzig, L. (2012). Means, variances, and ranges in reproductive success: comparative evidence. *Evolution and Human Behavior* (0).

Blockley, S. and R. Pinhasi (2011, January). A revised chronology for the adoption of agriculture in the southern levant and the role of lateglacial climatic change. *Quaternary Science Reviews 30*(1-2), 98--108.

Bocquet-Appel, J. and P. Demars (2000). Population kinetics in the upper palaeolithic in western europe. *Journal of Archaeological Science 27*(7), 551--570.

Bocquet-Appel, J., S. Naji, M. Linden, and J. Kozlowski (2009). Detection of diffusion and contact zones of early farming in europe from the space-time distribution of 14C dates. *Journal of Archaeological Science 36*(3), 807--820.

Bray, S. M., J. G. Mulle, A. F. Dodd, A. E. Pulver, S. Wooding, and S. T. Warren (2010). Signatures of founder effects, admixture, and selection in the ashkenazi jewish population. *Proceedings of the National Academy of Sciences 107*(37), 16222 --16227.

Brown, T., M. Jones, W. Powell, and R. Allaby (2009, February). The complex origins of domesticated crops in the fertile crescent. *Trends in Ecology & Evolution 24*(2), 103--109.

Browning, S. R. and B. L. Browning (2007, November). Rapid and accurate haplotype phasing and Missing-Data inference for Whole-Genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics 81*(5), 1084--1097.

Bryc, K., C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C. D. Bustamante, and H. Ostrer (2010, May). Colloquium paper: Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences 107*(Supplement_2), 8954--8961.

Busby, G. B. J., F. Brisighelli, P. Sánchez-Diz, E. Ramos-Luis, C. Martinez-Cadenas, M. G. Thomas, D. G. Bradley, L. Gusmão, B. Winney, W. Bodmer, M. Vennemann, V. Coia, F. Scarnicci, S. Tofanelli, G. Vona, R. Ploski, C. Vecchiotti, T. Zemunik, I. Rudan, S. Karachanak, D. Toncheva, P. Anagnostou, G. Ferri, C. Rapone, T. Hervig, T. Moen, J. F. Wilson, and C. Capelli (2012, March). The peopling of europe and the cautionary tale of y chromosome lineage R-M269. *Proceedings of the Royal Society B: Biological Sciences 279*(1730), 884 --892.

Butler, J. M., R. Schoske, P. M. Vallone, M. C. Kline, A. J. Redd, and M. F. Hammer (2002, September). A novel multiplex for simultaneous amplification of 20 y chromosome STR markers. *Forensic Science International 129*(1), 10--24.

Cann, H., C. De Toma, L. Cazes, M. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. Ferrara, J. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. Herrera, X. Huang, J. Kidd, K. Kidd, A. Langaney, A. Lin, S. Mehdi, P. Parham, A. Piazza, M. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. Weber, H. Greely, M. Feldman, G. Thomas, J. Dausset, and L. Cavalli-Sforza (2002). A human genome diversity cell line panel. *Science 296*(5566), 261--262.

Cann, R., M. Stoneking, and A. Wilson (1987). Mitochondrial DNA and human evolution. *Nature 325*(6099), 31--36.

Capelli, C., V. Onofri, F. Brisighelli, I. Boschi, F. Scarnicci, M. Masullo, G. Ferri, S. Tofanelli, A. Tagliabracci, L. Gusmao, A. Amorim, F. Gatto, M. Kirin, D. Merlitti, M. Brion, A. B. Verea, V. Romano, F. Cali, and V. Pascali (2009, January). Moors and saracens in europe: estimating the medieval north african male legacy in southern europe. *European Journal of Human Genetics 17*(6), 848--852.

Capelli, C., N. Redhead, J. K. Abernethy, F. Gratrix, J. F. Wilson, T. Moen, T. Hervig, M. Richards, M. P. Stumpf, P. A. Underhill, et al. (2003). A y chromosome census of the british isles. *Current Biology 13*(11), 979–984.

Capelli, C., N. Redhead, V. Romano, F. Cali, G. Lefranc, V. Delague, A. Megarbane, A. E. Felice, V. L. Pascali, P. I. Neophytou, et al. (2006). Population structure in the mediterranean basin: A y chromosome perspective. *Annals of human genetics 70*(2), 207--225.

Capelli, C., J. F. Wilson, M. Richards, M. P. Stumpf, F. Gratrix, S. Oppenheimer, P. Underhill, V. L. Pascali, T. M. Ko, and D. B. Goldstein (2001). A predominantly indigenous paternal heritage for the austronesian-speaking peoples of insular southeast asia and oceania. *The American Journal of Human Genetics 68*(2), 432–443.

Carbonell, E., J. M. Bermúdez de Castro, J. M. Parés, A. Pérez-González, G. Cuenca-Bescós, A. Ollé, M. Mosquera, R. Huguet, J. van der Made, A. Rosas, R. Sala, J. Vallverdú, N. García, D. E. Granger, M. Martinón-Torres, X. P. Rodríguez, G. M. Stock, J. M. Vergès, E. Allué, F. Burjachs, I. Cáceres, A. Canals, A. Benito, C. Díez, M. Lozano, A. Mateos, M. Navazo, J. Rodríguez, J. Rosell, and J. L. Arsuaga (2008, March). The first hominin of europe. *Nature 452*(7186), 465--469.

Cavalli-Sforza, L. (1997). Genes, peoples, and languages. *Proceedings of the National Academy of Sciences of the United States of America 94*(15), 7719--7724.

Cavalli Sforza, L. and M. Feldman (1976). Evolution of continuous variation: direct approach through joint distribution of genotypes and phenotypes. *Proceedings of the National Academy of Sciences of the United States of America 73*(5), 1689--1692.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza (1994). *The history and geography of human genes*. Princeton University Press.

Cerezo, M., A. Achilli, A. Olivieri, U. A. Perego, A. Gómez-Carballa, F. Brisighelli, H. Lancioni, S. R. Woodward, M. López-Soto, □. Carracedo, C. Capelli, A. Torroni, and A. Salas (2012, March). Reconstructing ancient mitochondrial DNA links between africa and europe. *Genome Research*.

Charchar, F. J., L. D. Bloomer, T. A. Barnes, M. J. Cowley, C. P. Nelson, Y. Wang, M. Denniff, R. Debiec, P. Christofidou, S. Nankervis, A. F. Dominiczak, A. Bani-Mustafa, A. J.

Balmforth, A. S. Hall, J. Erdmann, F. Cambien, P. Deloukas, C. Hengstenberg, C. Packard, H. Schunkert, W. H. Ouwehand, I. Ford, A. H. Goodall, M. A. Jobling, N. J. Samani, and M. Tomaszewski (2012, March). Inheritance of coronary artery disease in men: an analysis of the role of the y chromosome. *The Lancet 379*(9819), 915--922.

Chikhi, L., G. Destro-Bisol, G. Bertorelle, V. Pascali, and G. Barbujani (1998). Clines of nuclear DNA markers suggest a largely neolithic ancestry of the european gene pool. *Proceedings of the National Academy of Sciences of the United States of America 95*(15), 9053--9058.

Chikhi, L., R. A. Nichols, G. Barbujani, and M. A. Beaumont (2002). Y genetic data support the neolithic demic diffusion model. *Proceedings of the National Academy of Sciences 99*(17), 11008.

Childe, V. (1925). *The Dawn of European civilisation*. London: Routledge and Kegan Paul.

Childe, V. (1942). *What happened in history*. Harmondsworth, UK: Penguin Books.

Churchill, S. and F. Smith (2000). Makers of the early aurignacian of europe. *American Journal of Physical Anthropology Suppl 31*, 61--115.

Clark, J. (1965). Radiocarbon dating and the spread of farming economy. *Antiquity 39*(153), 45--48.

Clark, P. U., A. S. Dyke, J. D. Shakun, A. E. Carlson, J. Clark, B. Wohlfarth, J. X. Mitrovica, S. W. Hostetler, and A. M. McCabe (2009, August). The last glacial maximum. *Science 325*(5941), 710--714.

Collard, M., K. Edinborough, S. Shennan, and M. Thomas (2010). Radiocarbon evidence indicates that migrants introduced farming to britain. *Journal of Archaeological Science 37*(4), 866--870.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, N. A. Rosenberg, and J. K. Pritchard (2006, October). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics 38*(11), 1251--1260.

Coop, G., J. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R. Myers, L. Cavalli-Sforza, M. Feldman, and J. Pritchard (2009). The role of geography in human adaptation. *PLoS Genetics 5*(6).

Cox, M. (2007, May). Extreme patterns of variance in small populations: Placing limits on human Y-Chromosome diversity through time in the vanuatu archipelago. *Annals of Human Genetics 71*(3), 390--406.

Cruciani, F., B. Trombetta, C. Antonelli, R. Pascone, G. Valesini, V. Scalzi, G. Vona, B. Melegh, B. Zagradisnik, G. Assum, G. D. Efremov, D. Sellitto, and R. Scozzari (2011, June). Strong intra- and inter-continental differentiation revealed by y chromosome SNPs m269, u106 and u152. *Forensic Science International: Genetics 5*(3), e49--e52.

Cruciani, F., B. Trombetta, A. Massaia, G. Destro-Bisol, D. Sellitto, and R. Scozzari (2011). A revised root for the human y chromosomal phylogenetic tree: The origin of patrilineal diversity in africa. *American Journal of Human Genetics 88*(6), 814--818.

Cunliffe, B. (1994a). The impact of rome on barbarian society 140BC-AD300. In B. Cunliffe (Ed.), *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Cunliffe, B. (1994b). Iron age socities in western europe and beyond, 800-140 BC. In B. Cunliffe (Ed.), *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Cunliffe, B. (1994c). *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Delaneau, O., J. Marchini, and J. Zagury (2011, December). A linear complexity phasing method for thousands of genomes. *Nat Meth advance online publication*.

Dennell, R. and W. Roebroeks (2005). An asian perspective on early human dispersal from africa. *Nature 438*(7071), 1099--1104.

Di Giacomo, F., F. Luca, L. O. Popa, N. Akar, N. Anagnou, J. Banyko, R. Brdicka, G. Barbujani, F. Papola, G. Ciavarella, F. Cucci, L. Stasi, L. Gavrila, M. G. Kerimova, D. Kovatchev, A. I. Kozlov, A. Loutradis, V. Mandarino, C. Mammi′, E. N. Michalodimitrakis, G. Paoli, K. I. Pappa, G. Pedicini, L. Terrenato, S. Tofanelli, P. Malaspina, and A. Novelletto (2004, August). Y chromosomal haplogroup j as a signature of the post-neolithic colonization of europe. *Human Genetics 115*(5), 357--371.

Di Rienzo, A., A. Peterson, J. Garza, A. Valdes, M. Slatkin, and N. Freimer (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America 91*(8), 3166--3170.

Diamond, J. and P. Bellwood (2003). Farmers and their languages: The first expansions. *Science 300*(5619), 597--603.

Estoup, A., P. Jarne, and J. Cornuet (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology 11*(9), 1591--1604.

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

Finnilä, S., M. Lehtonen, and K. Majamaa (2001). Phylogenetic network for european mtDNA. *American Journal of Human Genetics 68*(6), 1475--1484.

Forster, P., A. Röhl, P. Lünnemann, C. Brinkmann, T. Zerjal, C. Tyler-Smith, and B. Brinkmann (2000). A short tandem repeat-based phylogeny for the human y chromosome. *American Journal of Human Genetics 67*(1), 182--196.

Francalacci, P., L. Morelli, A. Useli, and D. Sanna (2010). The history and geography of the y chromosome SNPs in europe: An update. *Journal of Anthropological Sciences 88*, 207--214.

Francalacci, P. and D. Sanna (2008). History and geography of human y-chromosome in europe: A SNP perspective. *Journal of Anthropological Sciences 86*, 59--89.

Gallagher, A., M. Gunther, and H. Bruchhaus (2009, March). Population continuity, demic diffusion and neolithic origins in central-southern germany: The evidence from body proportions. *HOMO - Journal of Comparative Human Biology 60*(2), 95--126.

Gamble, C., W. Davies, P. Pettitt, L. Hazelwood, and M. Richards (2005). The archaeological and genetic foundations of the european population during the late glacial: Implications for 'agricultural thinking'. *Cambridge Archaeological Journal 15*(2), 193--223.

Gamble, C., W. Davies, P. Pettitt, and M. Richards (2004). Climate change and evolving human diversity in europe during the last glacial. *Philosophical Transactions of the Royal Society B: Biological Sciences 359*(1442), 243--254.

Gilbert, M. T. P., D. L. Jenkins, A. Gotherstrom, N. Naveran, J. J. Sanchez, M. Hofreiter, P. F. Thomsen, J. Binladen, T. F. G. Higham, R. M. Yohe, R. Parr, L. S. Cummings, and E. Willerslev (2008, May). DNA from Pre-Clovis human coprolites in oregon, north america. *Science 320*(5877), 786--789.

Gimbutas, M. (1973). The beginning of the bronze age in europe and the Indo-Europeans 3500-2500 B.C. *Journal of Indo-European Studies* (1), 163--214.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman (1995a). An evaluation of genetic distances for use with microsatellite loci. *Genetics 139*(1), 463.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman (1995b). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences 92*(15), 6723.

Gray, R. and Q. Atkinson (2003). Language-tree divergence times support the anatolian theory of Indo-European origin. *Nature 426*(6965), 435--439.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, N. F. Hansen, E. Y. Durand, A. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S.

Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, □. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo (2010, May). A draft sequence of the neandertal genome. *Science 328*(5979), 710 --722.

Gusmao, L. and C. Alves (2005). Y chromosome STR typing. In A. Carracedo (Ed.), *Forensic DNA Typing Protocols*, Volume 297 of *Methods in Molecular Biology*, pp. 67--81. Totowa, New Jersey: Humana Press.

Gusmao, L., J. M. Butler, A. Carracedo, P. Gill, M. Kayser, W. R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, et al. (2006). DNA commission of the international society of forensic genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic science international 157*(2-3), 187–197.

Gusmão, L., P. Sánchez-Diz, F. Calafell, P. Martín, C. Alonso, F. Álvarez-Fernández, C. Alves, L. Borjas-Fajardo, W. Bozzo, M. Bravo, J. Builes, J. Capilla, M. Carvalho, C. Castillo, C. Catanesi, D. Corach, A. D. Lonardo, R. Espinheira, E. F. d. Carvalho, M. Farfán, H. Figueiredo, I. Gomes, M. Lojo, M. Marino, M. Pinheiro, M. Pontes, V. Prieto, E. Ramos-Luis, J. Riancho, A. S. Góes, O. Santapa, D. Sumita, G. Vallejo, L. V. Rioja, M. Vide, C. V. d. Silva, M. Whittle, W. Zabala, M. Zarrabeitia, A. Alonso, A. Carracedo, and A. Amorim (2005). Mutation rates at y chromosome specific microsatellites. *Human Mutation 26*(6), 520--528.

Haak, W., O. Balanovsky, J. J. Sanchez, S. Koshel, V. Zaporozhchenko, C. J. Adler, C. S. I. Der Sarkissian, G. Brandt, C. Schwarz, N. Nicklisch, V. Dresely, B. Fritsch, E. Balanovska, R. Villems, H. Meller, K. W. Alt, A. Cooper, and the Genographic Consortium (2010, November). Ancient DNA from european early neolithic farmers reveals their near eastern affinities. *PLoS Biol 8*(11), e1000536.

Haak, W., P. Forster, B. Bramanti, S. Matsumura, G. Brandt, M. Tanzer, R. Villems, C. Renfrew, D. Gronenborn, K. W. Alt, and J. Burger (2005, November). Ancient DNA from the first european farmers in 7500-Year-Old neolithic sites. *Science 310*(5750), 1016--1018.

Hammer, M. (2002). A nomenclature system for the tree of human Y-Chromosomal binary haplogroups. *Genome Research 12*(2), 339--348.

Heather, P. (2009). *Empires and Barbarians: migration, development and the birth of Europe*. London, UK: Macmillan.

Henn, B. M., C. R. Gignoux, M. Jobin, J. M. Granka, J. M. Macpherson, J. M. Kidd, L. Rodríguez-Botigué, S. Ramachandran, L. Hon, A. Brisbin, A. A. Lin, P. A. Underhill, D. Comas, K. K. Kidd, P. J. Norman, P. Parham, C. D. Bustamante, J. L. Mountain, and M. W. Feldman (2011, March). Hunter-gatherer genomic diversity suggests a southern african origin for modern humans. *Proceedings of the National Academy of Sciences*.

Heyer, E., J. Puymirat, P. Dieltjes, E. Bakker, and P. De Knijff (1997). Estimating y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human Molecular Genetics 6*(5), 799--803.

Hill, A. V. S. (2012, March). Evolution, revolution and heresy in the genetics of infectious disease susceptibility. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*(1590), 840--849.

Hoffecker, J. F. (2009). The spread of modern humans in europe. *Proceedings of the National Academy of Sciences 106*(38), 16040.

Hofreiter, M. and J. Stewart (2009). Ecological change, range fluctuations and population dynamics during the pleistocene. *Current Biology 19*(14), 584–594.

Housley, R., C. Gamble, M. Street, and P. Pettitt (1997). Radiocarbon evidence for the late-glacial human recolonisation of northern europe. *Proceedings of the Prehistoric Society 63*, 25--54.

Howie, B. N., P. Donnelly, and J. Marchini (2009, June). A flexible and accurate genotype imputation method for the next generation of Genome-Wide association studies. *PLoS Genetics 5*(6), e1000529.

Ingman, M., H. Kaessmann, S. Pääbo, and U. Gyllensten (2000). Mitochondrial genome variation and the origin of modem humans. *Nature 408*(6813), 708--713.

Ingram, M. (1957). Gene mutations in human haemoglobin: the chemical difference between normal and sickle-cell haemoglobin. *Nature 180*, 326--328.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton (2008, February). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature 451*(7181), 998--1003.

Jobling, D. M., M. Hurles, and C. Tyler-Smith (2004). *Human Evolutionary Genetics: Origins, Peoples and Disease* (1 ed.). Garland Science.

Jobling, M. (2012). The impact of recent events on human genetic diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*(1590), 793--799.

Jobling, M. and C. Tyler-Smith (1995). Fathers and sons: The y chromosome and human evolution. *Trends in Genetics 11*(11), 449--456.

Jobling, M. A. and C. Tyler-Smith (2003). The human y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics 4*(8), 598–612.

Karafet, T. M., F. L. Mendez, M. B. Meilerman, P. A. Underhill, S. L. Zegura, and M. F. Hammer (2008, February). New binary polymorphisms reshape and increase resolution of the human y chromosomal haplogroup tree. *Genome Research 18*(5), 830--838.

Kayser, M., A. Caglià, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, P. Schneider, R. Szibor, J. Teifel-Greding, G. Weichhold,

P. De Knijff, and L. Roewer (1997). Evaluation of y-chromosomal STRs: a multicenter study. *International Journal of Legal Medicine 110*(3), 125--133.

Kayser, M., R. Kittler, A. Erler, M. Hedman, A. Lee, A. Mohyuddin, S. Mehdi, Z. Rosser, M. Stoneking, M. Jobling, A. Sajantila, and C. Tyler-Smith (2004). A comprehensive survey of human y-chromosomal microsatellites. *American Journal of Human Genetics 74*(6), 1183--1197.

Kayser, M., M. Krawczak, L. Excoffier, P. Dieltjes, D. Corach, V. Pascali, C. Gehrig, L. F. Bernini, J. Jespersen, E. Bakker, L. Roewer, and P. de Knijff (2001, April). An extensive analysis of Y-Chromosomal microsatellite haplotypes in globally dispersed human populations. *The American Journal of Human Genetics 68*(4), 990--1018.

Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke, S. Brauer, C. Krüger, M. Krawczak, M. Nagy, T. Dobosz, R. Szibor, P. De Knijff, M. Stoneking, and A. Sajantila (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human y chromosome, as revealed by direct observation in father/son pairs. *American Journal of Human Genetics 66*(5), 1580--1588.

Keller, A., A. Graefen, M. Ball, M. Matzas, V. Boisguerin, F. Maixner, P. Leidinger, C. Backes, R. Khairat, M. Forster, B. Stade, A. Franke, J. Mayer, J. Spangler, S. McLaughlin, M. Shah, C. Lee, T. T. Harkins, A. Sartori, A. Moreno-Estrada, B. Henn, M. Sikora, O. Semino, J. Chiaroni, S. Rootsi, N. M. Myres, V. M. Cabrera, P. A. Underhill, C. D. Bustamante, E. E. Vigl, M. Samadelli, G. Cipollini, J. Haas, H. Katus, B. D. O'Connor, M. R. Carlson, B. Meder, N. Blin, E. Meese, C. M. Pusch, and A. Zink (2012, February). New insights into the tyrolean iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications 3*, 698.

Kimura, M. and T. Ohta (1978, June). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences 75*(6), 2868 --2872.

Kondrashov, A. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Human Mutation 21*(1), 12--27.

Kong, A., G. Masson, M. L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P. I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D. F. Gudbjartsson, H. Stefansson, and K. Stefansson (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics 40*(9), 1068--1075.

Kozlowski, J. (2007). Rethinking the human revolution: new behavioural and biological perspectives on the origin and dispersal of modern humans. *The significance of blade technologies in the period 50-35 kya BP for the Middle-Upper Palaeolithic transition in central and Eastern Europe*, 317--328.

Krause, J., Q. Fu, J. M. Good, B. Viola, M. V. Shunkov, A. P. Derevianko, and S. Pääbo (2010, March). The complete mitochondrial DNA genome of an unknown hominin from southern siberia. *Nature 464*(7290), 894--897.

Krenke, B., L. Viculis, M. Richard, M. Prinz, S. Milne, C. Ladd, A. Gross, T. Gornall, J. Frappier, A. Eisenberg, C. Barna, X. Aranda, M. Adamowicz, and B. Budowle (2005). Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Science International 148*(1), 1--14.

Lander, E., L. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. Levine, P. McEwan, K. McKernan, J. Meldrim, J. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims,

R. Waterston, R. Wilson, L. Hillier, J. McPherson, M. Marra, E. Mardis, L. Fulton, A. Chinwalla, K. Pepin, W. Gish, S. Chissoe, M. Wendl, K. Delehaunty, T. Miner, A. Delehaunty, J. Kramer, L. Cook, R. Fulton, D. Johnson, P. Minx, S. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. Gibbs, D. Muzny, S. Scherer, J. Bouck, E. Sodergren, K. Worley, C. Rives, J. Gorrell, M. Metzker, S. Naylor, R. Kucherlapati, D. Nelson, G. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, M. Hong, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. Davis, N. Federspiel, A. Abola, M. Proctor, B. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. McCombie, M. De La Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. Brown, C. Burge, L. Cerutti, H. Chen, D. Church, M. Clamp, R. Copley, T. Doerks, S. Eddy, E. Eichler, T. Furey, J. Galagan, J. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. Johnson, T. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. Kent, P. Kitts, E. Koonin, I. Korf, D. Kulp, D. Lancet, T. Lowe, A. McLysaght, T. Mikkelsen, J. Moran, N. Mulder, V. Pollara, C. Ponting, G. Schuler, J. Schultz, G. Slater, A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. Wolf, K. Wolfe, S. Yang, R. Yeh, F. Collins, M. Guyer, J. Peterson, A. Felsenfeld, K. Wetterstrand, R. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. Cox, M. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. Evans, M. Athanasiou, R. Schultz, A. Patrinos, and M. Morgan (2001). Initial sequencing and analysis of the human genome. *Nature 409*(6822), 860--921.

Landsteiner, K. (1900). Zur kenntnis der antifermentativen, lytischen und agglutinierenden wirkungen des blutserums und der lymphe. *Zentralblatt Bakteriologie 27*, 357--362.

Lao, O., T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, M. Balascakova,

J. Bertranpetit, L. A. Bindoff, and D. Comas (2008, August). Correlation between genetic and geographic structure in europe. *Current Biology 18*(16), 1241--1248.

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

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers (2008, February). Worldwide human relationships inferred from Genome-Wide patterns of variation. *Science 319*(5866), 1100--1104.

Li, N. and M. Stephens (2003, December). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics 165*(4), 2213--2233. PMID: 14704198.

Luca, F., M. Basile, F. Di Giacomo, and A. Novelletto (2005). Independent methods for evolutionary genetic dating provide insights into y-chromosomal STR mutation rates confirming data from direct father-son transmissions. *Human Genetics 118*(2), 153--165.

Malmström, H., M. T. P. Gilbert, M. G. Thomas, M. Brandström, J. Storå, P. Molnar, P. K. Andersen, C. Bendixen, G. Holmlund, and A. Götherström (2009, November). Ancient DNA reveals lack of continuity between neolithic Hunter-Gatherers and contemporary scandinavians. *Current Biology 19*(20), 1758--1762.

Malyarchuk, B., M. Derenko, T. Grzybowski, M. Perkova, U. Rogalla, T. Vanecek, and I. Tsybovsky (2010, April). The peopling of europe from the mitochondrial haplogroup u5 perspective. *PLoS ONE 5*(4), e10285.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn (2008, January). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics 9*(5), 356--369.

McEvedy, C. (1967a). *The Penguin Atlas of Ancient History*. Norwich, UK: Fletcher & Son Ltd.

McEvedy, C. (1967b). *The Penguin Atlas of Medieval History*. Norwich, UK: Fletcher & Son Ltd.

McVean, G. (2009, October). A genealogical interpretation of principal components analysis. *PLoS Genetics 5*(10), e1000686.

Mellars, P. (2006a, August). Going east: New genetic and archaeological perspectives on the modern human colonization of eurasia. *Science 313*(5788), 796--800.

Mellars, P. and B. Cunliffe (1994). The upper paleolithic revolution. In *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza (1978, September). Synthetic maps of human gene frequencies in europeans. *Science 201*(4358), 786--792.

Metcalfe, A. (2009). *The Muslims in Medieval Italy*. Edinburgh, UK: Edinburgh University Press.

Metspalu, M., I. Romero, B. Yunusbayev, G. Chaubey, C. Mallick, G. Hudjashov, M. Nelis, R. Mägi, E. Metspalu, M. Remm, R. Pitchappan, L. Singh, K. Thangaraj, R. Villems, and T. Kivisild (2011, December). Shared and unique components of human population structure and Genome-Wide signals of positive selection in south asia. *The American Journal of Human Genetics 89*(6), 731--744.

Mithen, S. J. (1994). The mesolithic age. In B. Cunliffe (Ed.), *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Mithen, S. J. (2003). *After the Ice*. London, UK: Phoenix.

Mix, A. C., E. Bard, and R. Schneider (2001, February). Environmental processes of the ice age: land, oceans, glaciers (EPILOG). *Quaternary Science Reviews 20*(4), 627--657.

Moore, L., B. McEvoy, E. Cape, K. Simms, and D. Bradley (2006). A y-chromosome signature of hegemony in gaelic ireland. *American Journal of Human Genetics 78*(2), 334--338.

Moorjani, P., N. Patterson, J. Hirschhorn, A. Keinan, L. Hao, G. Atzmon, E. Burns, H. Ostrer, A. Price, and D. Reich (2011). The history of african gene flow into southern europeans, levantines, and jews. *PLoS Genetics 7*(4).

Morelli, L., D. Contu, F. Santoni, M. B. Whalen, P. Francalacci, and F. Cucca (2010, April). A comparison of Y-Chromosome variation in sardinia and anatolia is more consistent with cultural rather than demic diffusion of agriculture. *PLoS ONE 5*(4), e10419.

Moskvina, V., M. Smith, D. Ivanov, D. Blackwood, D. Stclair, C. Hultman, D. Toncheva, M. Gill, A. Corvin, C. O'Dushlaine, D. W. Morris, N. R. Wray, P. Sullivan, C. Pato, M. T. Pato, P. Sklar, S. Purcell, P. Holmans, M. C. O'Donovan, M. J. Owen, and G. Kirov (2010, July). Genetic differences between five european populations. *Human Heredity 70*(2), 141--149. PMID: 20616560.

Mulero, J., C. Chang, L. Calandro, R. Green, Y. Li, C. Johnson, and L. Hennessy (2006). Development and validation of the AmpFℓSTR® yfiler™ PCR amplification kit: A male specific, single amplification 17 Y-STR multiplex system. *Journal of Forensic Sciences 51*(1), 64--75.

Myres, N. M., S. Rootsi, A. A. Lin, M. Järve, R. J. King, I. Kutuev, V. M. Cabrera, E. K. Khusnutdinova, A. Pshenichnov, B. Yunusbayev, O. Balanovsky, E. Balanovska, P. Rudan, M. Baldovic, R. J. Herrera, J. Chiaroni, J. Di Cristofaro, R. Villems, T. Kivisild, and P. A. Underhill (2011). A major y-chromosome haplogroup r1b holocene era founder effect in central and western europe. *European Journal of Human Genetics 19*(1), 95--101.

Nachman, M. and S. Crowell (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics 156*(1), 297--304.

Nelis, M., T. Esko, R. Mägi, F. Zimprich, D. Toncheva, S. Karachanak, T. Piskáčková, I. Balaščák, L. Peltonen, E. Jakkula, K. Rehnström, M. Lathrop, S. Heath, P. Galan, S. Schreiber, T. Meitinger, A. Pfeufer, H. Wichmann, B. Melegh, N. Polgár, D. Toniolo, P. Gasparini, P. D'Adamo, J. Klovins, L. Nikitina-Zake, V. Kučinskas, J. Kasnauskiene, J. Lubinski, T. Debniak, S. Limborska, A. Khrunin, X. Estivill, R. Rabionet, S. Marsal, A. Juliá,

S. Antonarakis, S. Deutsch, C. Borel, H. Attar, M. Gagnebin, M. Macek, M. Krawczak, M. Remm, and A. Metspalu (2009). Genetic structure of europeans: A view from the north-east. *PLoS ONE 4*(5).

Nelson, M. R., 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 (2008, September). The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *American Journal of Human Genetics 83*(3), 347--358. PMID: 18760391 PMCID: 2556436.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. Boyko, A. Auton, A. Indap, K. King, S. Bergmann, M. Nelson, M. Stephens, and C. Bustamante (2008). Genes mirror geography within europe. *Nature 456*(7218), 98--101.

Novembre, J. and M. Stephens (2008, April). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics 40*(5), 646--649.

Ohta, T. and M. Kimura (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research 22*(2), 201--204.

Onofri, V., L. Buscemi, and A. Tagliabracci (2009). Evaluating y-chromosome STRs mutation rates: A collaborative study of the Ge.F.I.-ISFG italian group. *Forensic Science International: Genetics Supplement Series 2*(1), 419--420.

Patterson, N., A. L. Price, and D. Reich (2006). Population structure and eigenanalysis. *PLoS Genetics 2*(12), e190.

Pauling, L., H. Itano, S. Singer, and I. Wells (1949). Sickle-cell anaemia, a molecular disease. *Science 110*, 543--548.

Pinhasi, R., J. Fort, and A. J. Ammerman (2005, November). Tracing the origin and spread of agriculture in europe. *PLoS Biol 3*(12), e410.

Pinhasi, R. and N. von Cramon-Taubadel (2012). A craniometric perspective on the transition to agriculture in europe. *Human Biology 84*(1), 45--66.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006, August). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics 38*(8), 904--909. PMID: 16862161.

Price, A. L., M. E. Weale, N. Patterson, S. R. Myers, A. C. Need, K. V. Shianna, D. Ge, J. I. Rotter, E. Torres, K. D. Taylor, et al. (2008). Long-range LD can confound genome scans in admixed populations. *The American Journal of Human Genetics 83*(1), 132–135.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000, June). Inference of population structure using multilocus genotype data. *Genetics 155*(2), 945--959.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly, and P. Sham (2007, September). PLINK: a tool set for Whole-Genome association and Population-Based linkage analyses. *American Journal of Human Genetics 81*(3), 559--575. PMID: 17701901 PMCID: 1950838.

R Development Core Team (2011). R: a language and environment for statistical computing. vienna (Austria): r foundation for statistical computing.

Reich, D., R. Green, M. Kircher, J. Krause, N. Patterson, E. Durand, B. Viola, A. Briggs, U. Stenzel, P. Johnson, T. Maricic, J. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. Shunkov, A. Derevianko, J. Hublin, J. Kelso, M. Slatkin, and S. Pääbo (2010). Genetic history of an archaic hominin group from denisova cave in siberia. *Nature 468*(7327), 1053--1060.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh (2009, September). Reconstructing indian population history. *Nature 461*(7263), 489--494.

Renfrew, C. (2000b). *Time Depth in Historical Linguistics*. Cambridge, UK: The McDonald Institute for Archaeological Research.

Richards, M. (2003, October). The neolithic invasion of europe. *Annual Review of Anthropology 32*(1), 135--162.

Richards, M., H. Corte-Real, P. Forster, V. Macaulay, H. Wilkinson-Herbots, A. Demaine, S. Papiha, R. Hedges, H. Bandelt, and B. Sykes (1996). Paleolithic and neolithic lineages in the european mitochondrial gene pool. *American Journal of Human Genetics 59*(1), 185--203.

Richter, D., G. Tostevin, and P. Škrdla (2008). Bohunician technology and thermoluminescence dating of the type locality of Brno-Bohunice (Czech republic). *Journal of Human Evolution 55*(5), 871--885.

Roberts, J. (2007). *The New Penguin History of the World* (5th ed.). London, UK: Penguin Books.

Roewer, L., J. Arnemann, N. Spurr, K. Grzeschik, and J. Epplen (1992). Simple repeat sequences on the human y chromosome are equally polymorphic as their autosomal counterparts. *Human Genetics 89*(4), 389--394.

Rootsi, S., T. Kivisild, G. Benuzzi, H. Help, M. Bermisheva, I. Kutuev, L. Barać, M. Peričić, O. Balanovsky, A. Pshenichnov, et al. (2004). Phylogeography of y-chromosome haplogroup i reveals distinct domains of prehistoric gene flow in europe. *The American Journal of Human Genetics 75*(1), 128–137.

Rosenberg, N., J. Pritchard, J. Weber, H. Cann, K. Kidd, L. Zhivotovsky, and M. Feldman (2002). Genetic structure of human populations. *Science 298*(5602), 2381--2385.

Rosenberg, N. A. (2006, November). Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics 70*(6), 841--847.

Rosser, Z., T. Zerjal, M. Hurles, M. Adojaan, D. Alavantic, A. Amorim, W. Amos, M. Armenteros, E. Arroyo, G. Barbujani, G. Beckman, L. Beckman, J. Bertranpetit, E. Bosch, D. Bradley, G. Brede, G. Cooper, H. Corteal-Real, P. De Knijff, R. Decorte, Y. Dubrova,

O. Evgrafov, A. Gilissen, S. Glisic, M. Golge, E. Hill, A. Jeziorowska, L. Kalaydjieva, M. Kayser, T. Kivisild, S. Kravchenko, A. Krumina, V. Kucinskas, J. Lavinha, L. Livshits, P. Malaspina, S. Maria, K. McElreavey, T. Meitinger, A. Mikelsaar, R. Mitchell, K. Nafa, J. Nicholson, S. Norby, A. Pandya, J. Parik, P. Patsalis, L. Pereira, B. Peterlin, G. Pielberg, M. Prata, C. Previdere, L. Roewer, S. Rootsi, D. Rubinsztein, J. Saillard, F. Santos, G. Stefanescu, B. Sykes, A. Tolun, R. Villems, C. Tyler-Smith, and M. Jobling (2000). Y-chromosomal diversity in europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics 67*(6), 1526--1543.

Rowley-Conwy, P. (2009, November). Human prehistory: Hunting for the earliest farmers. *Current Biology 19*(20), R948--R949.

Sampietro, M., O. Lao, D. Caramelli, M. Lari, R. Pou, M. Martí, J. Bertranpetit, and C. Lalueza-Fox (2007). Palaeogenetic evidence supports a dual model of neolithic spreading into europe. *Proceedings of the Royal Society B: Biological Sciences 274*(1622), 2161--2167.

Sanchez, J. J., M. Brión, W. Parson, A. J. Blanco-Verea, C. Børsting, M. Lareu, H. Niederstätter, H. Oberacher, N. Morling, and A. Carracedo (2004, March). Duplications of the y-chromosome specific loci p25 and 92R7 and forensic implications. *Forensic Science International 140*(2-3), 241--250.

Sawcer, S., G. Hellenthal, M. Pirinen, C. C. A. Spencer, T. International Multiple Sclerosis Genetics Consortium, S. L. Wellcome Trust Case Control Consortium 2, The Hauser, G. McVean, P. Donnelly, and A. Compston (2011, August). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature 476*(7359), 214--219. PMID: 21833088.

Scheet, P. and M. Stephens (2006, April). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics 78*(4), 629--644. PMID: 16532393.

Semaw, S., P. Renne, J. Harris, C. Feibel, R. Bernor, N. Fesseha, and K. Mowbray (1997). 2.5-million-year-old stone tools from gona, ethiopia. *Nature 385*(6614), 333--336.

Semino, O., G. Passarino, A. Brega, M. Fellous, and A. Santachiara-Benerecetti (1996). A view of the neolithic demic diffusion in europe through two y chromosome-specific markers. *American Journal of Human Genetics 59*(4), 964--968.

Semino, O., G. Passarino, P. Oefner, A. Lin, S. Arbuzova, L. Beckman, G. De Benedictis, P. Francalacci, A. Kouvatsi, S. Limborska, M. Marcikiae, A. Mika, B. Mika, D. Primorac, A. Santachiara-Benerecetti, L. Cavalli-Sforza, and P. Underhill (2000). The genetic legacy of paleolithic homo sapiens sapiens in extant europeans: A y chromosome perspective. *Science 290*(5494), 1155--1159.

Sengupta, S., L. Zhivotovsky, R. King, S. Mehdi, C. Edmonds, C. Chow, A. Lin, M. Mitra, S. Sil, A. Ramesh, M. Rani, C. Thakur, L. Cavalli-Sforza, P. Majumder, and P. Underhill (2006). Polarity and temporality of high-resolution y-chromosome distributions in india identify both indigenous and exogenous expansions and reveal minor genetic influence of central asian pastoralists. *American Journal of Human Genetics 78*(2), 202--221.

Sherratt, A. (1994a). The emergence of elites: Earlier bronze age europe, 2500-1300 BC. In B. Cunliffe (Ed.), *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Sherratt, A. (1994b). The transformation of early agrarian europe: The later neolithic and copper ages 4500-2500 BC. In B. Cunliffe (Ed.), *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Shi, W., Q. Ayub, M. Vermeulen, R.-g. Shao, S. Zuniga, K. van der Gaag, P. de Knijff, M. Kayser, Y. Xue, and C. Tyler-Smith (2010, February). A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol Biol Evol 27*(2), 385--393.

Simoni, L., F. Calafell, D. Pettener, J. Bertranpetit, and G. Barbujani (2000). Geographic

patterns of mtDNA diversity in europe. *American Journal of Human Genetics 66*(1), 262--278.

Sjödin, P. and O. François (2011, June). Wave-of-Advance models of the diffusion of the y chromosome haplogroup r1b1b2 in europe. *PLoS ONE 6*(6), e21592.

Skoglund, P., H. Malmström, M. Raghavan, J. Storå, P. Hall, E. Willerslev, M. T. P. Gilbert, A. Götherström, and M. Jakobsson (2012, April). Origins and genetic legacy of neolithic farmers and Hunter-Gatherers in europe. *Science 336*(6080), 466--469.

Soares, P., A. Achilli, O. Semino, W. Davies, V. Macaulay, H. J. Bandelt, A. Torroni, and M. B. Richards (2010). The archaeogenetics of europe. *Current Biology 20*(4), R174–R183.

Spataro, M. (2011, February). A comparison of chemical and petrographic analyses of neolithic pottery from south-eastern europe. *Journal of Archaeological Science 38*(2), 255--269.

Stephens, M. and P. Scheet (2005, March). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics 76*(3), 449--462. PMID: 15700229 PMCID: PMC1196397.

Stoneking, M. and J. Krause (2011, August). Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics 12*(9), 603--614.

Sun, J., J. Mullikin, N. Patterson, and D. Reich (2009). Microsatellites are molecular clocks that support accurate inferences about history. *Molecular Biology and Evolution 26*(5), 1017--1027.

The International HapMap Consortium (2003, December). The international HapMap project. *Nature 426*(6968), 789--796.

The International HapMap Consortium (2005, October). A haplotype map of the human genome. *Nature 437*(7063), 1299--1320.

The International HapMap Consortium (2007, October). A second generation human haplotype map of over 3.1 million SNPs. *Nature 449*(7164), 851--861.

The International HapMap Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature 467*(7311), 52--58.

Thomas, M., N. Bradman, and H. Flinn (1999). High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human y-chromosome. *Human Genetics 105*(6), 577--581.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman (2000, June). Recent common ancestry of human y chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America 97*(13), 7360 --7365.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, and S. M. Williams (2009, April). The genetic structure and history of africans and african americans. *Science 324*(5930), 1035--1044.

Todd, M. (1994). Barbarian europe, AD 300-700. In B. Cunliffe (Ed.), *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Torroni, A., A. Achilli, V. Macaulay, M. Richards, and H. Bandelt (2006, June). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics 22*(6), 339--345.

Torroni, A., H. Bandelt, V. Macaulay, M. Richards, F. Cruciani, C. Rengo, V. Martinez-Cabrera, R. Villems, T. Kivisild, E. Metspalu, J. Parik, H. Tolk, K. Tambets, P. Forster, B. Karger, P. Francalacci, P. Rudan, B. Janicijevic, O. Rickards, M. Savontaus, K. Huoponen, V. Laitinen, S. Koivumäki, B. Sykes, E. Hickey, A. Novelletto, P. Moral, D. Sellitto, A. Coppa, N. Al-Zaheri, A. Santachiara-Benerecetti, O. Semino, and R. Scozzari (2001). A signal, from human mtDNA, of postglacial recolonization in europe. *American Journal of Human Genetics 69*(4), 844--852.

Torroni, A., M. Lott, M. Cabell, Y. Chen, L. Lavergne, and D. Wallace (1994). mtDNA and the origin of caucasians: Identification of ancient caucasian- specific haplogroups, one of which is prone to a recurrent somatic duplication in the d-loop region. *American Journal of Human Genetics 55*(4), 760--776.

Trinkaus, E. (2007). European early modern humans and the fate of the neandertals. *Proceedings of the National Academy of Sciences of the United States of America 104*(18), 7367--7372.

Underhill, P., L. Jin, R. Zemans, P. Oefner, and L. Cavalli-Sforza (1996). A pre-Columbian y chromosome-specific transition and its implications for human evolutionary history. *Proceedings of the National Academy of Sciences of the United States of America 93*(1), 196--200.

van der Made, J. and A. Mateos (2010). Longstanding biogeographic patterns and the dispersal of early homo out of africa and into europe. *Quaternary International 223-224*, 195--200.

Venter, J., M. Adams, E. Myers, P. Li, R. Mural, G. Sutton, H. Smith, M. Yandell, C. Evans, R. Holt, J. Gocayne, P. Amanatides, R. Ballew, D. Huson, J. Wortman, Q. Zhang, C. Kodira, X. Zheng, L. Chen, M. Skupski, G. Subramanian, P. Thomas, J. Zhang, G. Gabor Miklos, C. Nelson, S. Broder, A. Clark, J. Nadeau, V. McKusick, N. Zinder, A. Levine, R. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. Heiman, M. Higgins, R. Ji, Z. Ke, K. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. Merkulov, N. Milshina, H. Moore, A. Naik, V. Narayan, B. Neelam, D. Nusskern, D. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, and C. Xiao (2001). The sequence of the human genome. *Science 291*(5507), 1304--1351.

Wall, J., K. Lohmueller, and V. Plagnol (2009). Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution 26*(8), 1823--1827.

Wardle, K. (1994). The palace civilisations of minoan crete and mycenean greece 2000-1200 BC. In B. Cunliffe (Ed.), *The Oxford Illustrated History of Prehistoric Europe*. Oxford, UK: Oxford University Press.

Weale, M., D. Weiss, R. Jager, N. Bradman, and M. Thomas (2002). Y chromosome evidence for Anglo-Saxon mass migration. *Molecular Biology and Evolution 19*(7), 1008--1021.

Whittle, A. (1996). *Europe in the Neolithic: The creation of new worlds*. Cambridge World Archaeology. Cambridge, UK: Cambridge University Press.

Willuweit, S. and L. Roewer (2007, June). Y chromosome haplotype reference database (YHRD): update. *Forensic Science International: Genetics 1*(2), 83--87.

Wilson, I., M. Weale, and D. Balding (2003). Inferences from DNA data: Population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society. Series A: Statistics in Society 166*(2), 155--188.

Wilson, J., D. Weiss, M. Richards, M. Thomas, N. Bradman, and D. Goldstein (2001). Genetic evidence for different male and female roles during cultural transitions in the british isles. *Proceedings of the National Academy of Sciences of the United States of America 98*(9), 5078--5083.

Winney, B., A. Boumertit, T. Day, D. Davison, C. Echeta, I. Evseeva, K. Hutnik, S. Leslie, K. Nicodemus, E. C. Royrvik, S. Tonks, X. Yang, J. Cheshire, P. Longley, P. Mateos, A. Groom, C. Relton, D. T. Bishop, K. Black, E. Northwood, L. Parkinson, T. M. Frayling, A. Steele, J. R. Sampson, T. King, R. Dixon, D. Middleton, B. Jennings, R. Bowden, P. Donnelly, and W. Bodmer (2012). People of the british isles: preliminary analysis of genotypes and surnames in a UK-control population. *European Journal of Human Genetics 20*(2), 203--210.

Wood, M. (1992). *In Search of the First Civilisations*. Reading, UK: Random House Group.

WTCCC (2007, June). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*(7145), 661--678. PMID: 17554300.

Xue, Y., Q. Wang, Q. Long, B. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, D. MacArthur, M. Quail, N. Carter, H. Yang, and C. Tyler-Smith (2009). Human y chromosome Base-Substitution mutation rate measured by direct sequencing in a Deep-Rooting pedigree. *Current Biology 19*(17), 1453--1457.

Yotova, V., J. Lefebvre, C. Moreau, E. Gbeha, K. Hovhannesyan, S. Bourgeois, S. Bédarida, L. Azevedo, A. Amorim, T. Sarkisian, P. Avogbe, N. Chabi, M. H. Dicko, E. S. K. Santa Amouzou, A. Sanni, J. Roberts-Thomson, B. Boettcher, R. J. Scott, and D. Labuda (2011, January). An x-linked haplotype of neandertal origin is present among all non-African populations. *Molecular Biology and Evolution*.

Young, S. (2011, October). Rebuilding the genome of a hidden ethnicity. *Nature News*.

Yunusbayev, B., M. Metspalu, M. Järve, I. Kutuev, S. Rootsi, E. Metspalu, D. M. Behar, K. Varendi, H. Sahakyan, R. Khusainova, L. Yepiskoposyan, E. K. Khusnutdinova, P. A. Underhill, T. Kivisild, and R. Villems (2011). The caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Molecular Biology and Evolution*.

Zerjal, T., R. Wells, N. Yuldasheva, R. Ruzibakiev, and C. Tyler-Smith (2002). A genetic landscape reshaped by recent events: Y-chromosomal insights into central asia. *American Journal of Human Genetics 71*(3), 466--482.

Zhivotovsky, L. A. and P. A. Underhill (2005, March). On the evolutionary mutation rate at y-chromosome STRs: comments on paper by di giacomo et al. (2004). *Human Genetics 116*(6), 529--532.

Zhivotovsky, L. A., P. A. Underhill, C. Cinnioğlu, M. Kayser, B. Morar, T. Kivisild, R. Scozzari, F. Cruciani, G. Destro-Bisol, G. Spedini, et al. (2004). The effective mutation rate at y chromosome short tandem repeats, with application to human population-divergence time. *The American Journal of Human Genetics 74*(1), 50–61.

Zhivotovsky, L. A., P. A. Underhill, and M. W. Feldman (2006, December). Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Mol Biol Evol 23*(12), 2268--2270.

A. Supplementary Tables

Table A.1. Table showing the groups of STRs with varying linearity used in chapter 2

HG comparison	code in figure	ASD			μ	R	description	STRs
		median	2.5%	97.5%	mean	mean		
S21vS116	4.A	0.503	0.450	0.568	9.60E-04	12.75	4 STRs with highest ϑ(R)	DYS390; DYS392; DYS438; DYS448
S21vS116	4.B	0.891	0.802	0.998	3.54E-03	12.60	4 STRs with middle values of ϑ(R)	DYS19; DYS389I; DYS458; Y-GATA-H4
S21vS116	4.C	0.806	0.711	0.927	5.37E-03	10.00	4 STRs with lowest ϑ(R)	DYS389II; DYS437; DYS439; DYS456
S21vS116	5.A	0.458	0.414	0.509	1.19E-03	12.60	5 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS438; DYS448
S21vS116	5.B	0.879	0.797	0.972	3.60E-03	12.80	5 STRs with middle values of ϑ(R)	DYS389II; DYS437; DYS439; DYS456; DYS635
S21vS116	5.C	0.744	0.670	0.845	4.94E-03	10.00	5 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS391; DYS458; Y-GATA-H4

HG comparison	code in figure	ASD			μ	R	description	STRs
		median	2.5%	97.5%	mean	mean		
S21vS116	6.A	0.551	0.502	0.599	1.63E-03	13.00	6 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS438; DYS439; DYS448
S21vS116	6.B	0.782	0.706	0.859	3.35E-03	12.67	6 STRs with middle values of ϑ(R)	DYS389II; DYS393; DYS437; DYS439; DYS456; DYS635
S21vS116	6.C	0.729	0.657	0.810	4.75E-03	10.33	6 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS458; Y-GATA-H4
S21vS116	7.A	0.531	0.487	0.579	1.62E-03	12.43	7 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448
S21vS116	7.B	0.740	0.670	0.814	3.33E-03	12.29	7 STRs with middle values of ϑ(R)	DYS389II; DYS391; DYS393; DYS437; DYS439; DYS456; DYS635
S21vS116	7.C	0.841	0.772	0.923	4.78E-03	10.86	7 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS456; DYS458; Y-GATA-H4
S21vS116	7.F	0.938	0.856	1.021	4.96E-03	12.43	7 STRs with the fastest mutation rates	DYS19; DYS389I; DYS389II; DYS439; DYS456; DYS458; DYS635
S21vS116	7.S	0.478	0.431	0.525	1.53E-03	11.00	7 STRs with the slowest mutation rates	DYS390; DYS392; DYS393; DYS437; DYS438; DYS448; Y-GATA-H4
S21vS116	8.A	0.569	0.523	0.617	1.90E-03	12.63	8 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448; DYS635

HG comparison	code in figure	ASD			μ	R	description	STRs
		median	2.5%	97.5%	mean	mean		
S21vS116	8.B	0.779	0.721	0.843	3.11E-03	12.38	8 STRs with middle values of ϑ(R)	DYS389II; DYS390; DYS391; DYS393; DYS437; DYS439; DYS456; DYS635
S21vS116	8.C	0.839	0.774	0.912	4.66E-03	11.25	8 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS456; DYS458; DYS635; Y-GATA-H4
S21vS116	8.Bu	0.582	0.555	0.611	3.17E-03	11.50	8 shared STRs with the Myres et al 2010 dataset	DYS19; DYS389I; DYS389II; DYS390; DYS391; DYS392; DYS393; DYS439
S21vS116	8.M	0.532	0.498	0.566	3.17E-03	11.50	8 shared STRs with the Myres et al 2010 dataset	DYS19; DYS389I; DYS389II; DYS390; DYS391; DYS392; DYS393; DYS439
S21vS116	10.A	0.671	0.621	0.726	2.39E-03	12.70	10 STRs with highest ϑ(R)	DYS389II; DYS390; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448; DYS456; DYS635
S21vS116	10.B	0.841	0.780	0.906	3.42E-03	12.50	10 STRs with middle values of ϑ(R)	DYS389II; DYS390; DYS391; DYS393; DYS437; DYS438; DYS439; DYS456; DYS458; DYS635
S21vS116	10.C	0.813	0.755	0.872	4.27E-03	11.40	10 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS437; DYS439; DYS456; DYS458; DYS635; Y-GATA-H4

HG comparison	code in figure	ASD			μ	R	description	STRs
		median	2.5%	97.5%	mean	mean		
S21vS116	10.Bu	0.539	0.516	0.564	2.79E-03	11.60	10 STRs present in this dataset	DYS19; DYS389I; DYS389II; DYS390; DYS391; DYS392; DYS393; DYS437; DYS438; DYS439
S21vS116	10.M	0.480	0.451	0.512	2.68E-03	11.50	10 STRs present in the Myres et al 2010 dataset	DYS19; DYS388; DYS389I; DYS389II; DYS390; DYS391; DYS392; DYS393; DYS439; DYS461
S21vS116	15.All	0.695	0.655	0.741	3.24E-03	11.80	all 15 STRs	DYS19; DYS389I; DYS389II; DYS390; DYS391; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448; DYS456; DYS458; DYS635; Y-GATA-H4
AvB	n/a	4.067	3.573	4.551	9.60E-04	12.75	4 STRs with highest ϑ(R)	DYS390; DYS392; DYS438; DYS448
AvB	n/a	3.960	3.668	4.270	3.54E-03	12.60	4 STRs with middle values of ϑ(R)	DYS19; DYS389I; DYS458; Y-GATA-H4
AvB	n/a	3.062	2.790	3.367	5.37E-03	10.00	4 STRs with lowest ϑ(R)	DYS389II; DYS437; DYS439; DYS456
AvB	n/a	3.702	3.286	4.117	1.19E-03	12.60	5 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS438; DYS448
AvB	n/a	3.975	3.673	4.285	3.60E-03	12.80	5 STRs with middle values of ϑ(R)	DYS389II; DYS437; DYS439; DYS456; DYS635
AvB	n/a	2.698	2.468	2.929	4.94E-03	10.00	5 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS391; DYS458; Y-GATA-H4

HG comparison	code in figure	ASD			μ	R	description	STRs
		median	2.5%	97.5%	mean	mean		
AvB	n/a	3.325	2.974	3.690	1.63E-03	13.00	6 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS438; DYS439; DYS448
AvB	n/a	3.687	3.433	3.973	3.35E-03	12.67	6 STRs with middle values of ϑ(R)	DYS389II; DYS393; DYS437; DYS439; DYS456; DYS635
AvB	n/a	2.942	2.705	3.207	4.75E-03	10.33	6 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS458; Y-GATA-H4
AvB	n/a	3.169	2.871	3.461	1.62E-03	12.43	7 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448
AvB	n/a	3.341	3.105	3.572	3.33E-03	12.29	7 STRs with middle values of ϑ(R)	DYS389II; DYS391; DYS393; DYS437; DYS439; DYS456; DYS635
AvB	n/a	3.239	3.041	3.487	4.78E-03	10.86	7 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS456; DYS458; Y-GATA-H4
AvB	n/a	3.491	3.249	3.752	4.96E-03	12.43	7 STRs with the fastest mutation rates	DYS19; DYS389I; DYS389II; DYS439; DYS456; DYS458; DYS635
AvB	n/a	3.723	3.449	4.038	1.53E-03	11.00	7 STRs with the slowest mutation rates	DYS390; DYS392; DYS393; DYS437; DYS438; DYS448; Y-GATA-H4
AvB	n/a	3.644	3.333	3.979	1.90E-03	12.63	8 STRs with highest ϑ(R)	DYS390; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448; DYS635

HG comparison	code in figure	ASD			μ	R	description	STRs
		median	2.5%	97.5%	mean	mean		
AvB	n/a	3.890	3.655	4.141	3.11E-03	12.38	8 STRs with middle values of ϑ(R)	DYS389II; DYS390; DYS391; DYS393; DYS437; DYS439; DYS456; DYS635
AvB	n/a	3.701	3.487	3.944	4.66E-03	11.25	8 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS456; DYS458; DYS635; Y-GATA-H4
AvB	n/a	3.826	3.544	4.139	2.39E-03	12.70	10 STRs with highest ϑ(R)	DYS389II; DYS390; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448; DYS456; DYS635
AvB	n/a	3.671	3.449	3.910	3.42E-03	12.50	10 STRs with middle values of ϑ(R)	DYS389II; DYS390; DYS391; DYS393; DYS437; DYS438; DYS439; DYS456; DYS458; DYS635
AvB	n/a	3.341	3.153	3.514	4.27E-03	11.40	10 STRs with lowest ϑ(R)	DYS19; DYS389I; DYS389II; DYS391; DYS437; DYS439; DYS456; DYS458; DYS635; Y-GATA-H4
AvB	n/a	3.452	3.271	3.671	3.24E-03	11.80	all 15 STRs	DYS19; DYS389I; DYS389II; DYS390; DYS391; DYS392; DYS393; DYS437; DYS438; DYS439; DYS448; DYS456; DYS458; DYS635; Y-GATA-H4

Table A.2. Information on STRs used in chapter 2. Two values for R are quoted, one from the HGDP and the other from YHRD (Y Haplotype Research Database). Mutation rates quoted are those from Ballantyne et al 2010.

STR	mutations	meioses	motif length	complexity	R(HGDP)	R(YHRD)	μ
DYS472	0	1952	3	simple	1	NA	0.000446
DYS487	3	1914	3	simple	7	NA	0.00177
DYS508	6	1947	4	simple	7	NA	0.00303
DYS570	24	1981	4	simple	11	NA	0.0124
DYS525	1	2115	4	simple	5	NA	0.000978
DYS579	0	2158	4	simple	1	NA	0.000394
DYS583	0	2133	4	simple	3	NA	0.000399
DYS488	0	1979	3	simple	7	NA	0.00044
DYS531	1	2165	4	simple	6	NA	0.001
DYS537	3	1942	4	simple	5	NA	0.00238
DYS568	1	1950	4	simple	5	NA	0.00108
DYS522	1	2175	4	simple	6	NA	0.00104
DYS578	1	2089	4	simple	4	NA	0.000995
DYS580	0	2128	4	simple	7	NA	0.000405
DYS533	10	2285	4	simple	6	NA	0.00501
DYS590	0	2183	5	simple	3	NA	0.000391
DYS594	1	2038	5	simple	7	NA	0.00103
DYS617	0	2087	3	simple	6	NA	0.000413
DYS505	2	2163	4	simple	6	NA	0.00151
DYS638	2	2020	4	simple	5	NA	0.00104
DYS641	0	2171	4	simple	4	NA	0.00039
DYS476	1	2182	3	simple	4	NA	0.00094
DYS492	0	2173	3	simple	4	NA	0.000392
DYS540	5	2121	4	simple	4	NA	0.0033

STR	mutations	meioses	motif length	complexity	R(HGDP)	R(YHRD)	μ
DYS480	0	2186	3	simple	3	NA	0.000391
DYS485	1	2133	3	simple	9	NA	0.000404
DYS572	4	2173	4	simple	6	NA	0.00207
DYS490	0	2162	3	simple	9	NA	0.000395
DYS495	3	2158	3	simple	8	NA	0.00209
DYS567	0	2116	4	simple	6	NA	0.000408
DYS494	0	2186	3	simple	5	NA	0.000389
DYS565	5	2160	4	simple	5	NA	0.00209
DYS575	1	2167	4	simple	5	NA	0.000391
DYS481	11	2147	3	simple	13	NA	0.00497
DYS569	2	2099	4	simple	2	NA	0.00158
DYS576	33	2282	4	simple	10	NA	0.0143
DYS497	3	2189	3	simple	5	NA	0.00149
DYS511	3	2163	4	simple	5	NA	0.00152
DYS554	2	2180	4	simple	6	NA	0.000941
DYS618	0	2169	3	simple	6	NA	0.000395
DYS556	2	2086	4	simple	6	NA	0.00159
DYS573	2	2101	4	simple	5	NA	0.00041
DYS643	2	2328	5	simple	9	NA	0.0015
DYS491	0	2109	3	simple	5	NA	0.000409
DYS530	0	2163	4	simple	3	NA	0.000394
DYS549	8	2239	4	simple	6	NA	0.00455
DYS640	2	2119	4	simple	3	NA	0.000398
DYS388	1	4029	3	simple	7	NA	0.000425
DYS19	30	11596	4	complex	5	10	0.00437
DYS389AB	29	9615	4	complex	7	8	0.00551
DYS389CD	34	9585	4	complex	5	12	0.00383
DYS390	24	11099	4	complex	8	13	0.00152
DYS391	30	11038	4	complex	4	10	0.00323

STR	mutations	meioses	motif length	complexity	R(HGDP)	R(YHRD)	μ
DYS392	6	10992	3	complex	9	15	0.00097
DYS393	9	9585	4	simple	4	12	0.00211
DYS437	8	6141	4	complex	4	9	0.00153
DYS438	3	6316	5	simple	6	12	0.000956
DYS439	34	6279	4	complex	5	15	0.00384
DYS448	1	2960	6	complex	9	11	0.000394
DYS456	16	3000	4	simple	6	14	0.00494
DYS458	27	2999	4	simple	9	14	0.00836
DYS635	18	3652	4	complex	9	14	0.00385
Y_GATA_H4	16	3838	4	complex	6	8	0.00322
DYS636	1	403	0	simple	4	NA	NA
DYS589	0	403	0	simple	7	NA	NA

Table A.3. Detail of genome coying proportions from chapter 5. Included are the European average and standard deviation copying proportions for each of the ten world regions. The average copying proportion across all individuals in each European populations is also shown.

	America	Caucasus	Central Asia	East Asia	Europe	Middle East	North Africa	Pacific	South Central Asia	SubSaharan Africa
Europe Mean	0.054	0.141	0.099	0.048	0.0253	0.112	0.121	0.042	0.103	0.027
Europe SD	0.008	0.014	0.012	0.014	0.044	0.017	0.022	0.004	0.012	0.005
Belorussian	0.057	0.135	0.107	0.050	0.284	0.100	0.101	0.041	0.103	0.022
Lithuanian	0.056	0.135	0.107	0.047	0.302	0.094	0.096	0.040	0.102	0.021

	America	Caucasus	Central Asia	East Asia	Europe	Middle East	North Africa	Pacific	South Central Asia	SubSaharan Africa
Polish	0.056	0.139	0.106	0.047	0.286	0.100	0.100	0.040	0.103	0.023
Russian	0.067	0.127	0.116	0.069	0.277	0.090	0.088	0.044	0.100	0.022
English	0.055	0.138	0.101	0.042	0.266	0.104	0.119	0.041	0.107	0.025
Ireland	0.056	0.140	0.103	0.042	0.266	0.104	0.117	0.040	0.109	0.023
Norwegian	0.056	0.136	0.103	0.045	0.280	0.100	0.109	0.041	0.106	0.024
Scottish	0.056	0.140	0.102	0.042	0.268	0.104	0.116	0.041	0.108	0.024
Welsh	0.057	0.142	0.100	0.042	0.264	0.102	0.122	0.039	0.108	0.024
Finnish	0.067	0.127	0.114	0.062	0.280	0.091	0.094	0.043	0.101	0.022
Germany-Austria	0.054	0.142	0.102	0.044	0.268	0.107	0.112	0.041	0.107	0.024
EastSicilian	0.049	0.154	0.093	0.044	0.190	0.137	0.146	0.044	0.107	0.037
NorthItalian	0.051	0.151	0.094	0.043	0.228	0.124	0.132	0.043	0.106	0.028
SouthItalian	0.049	0.157	0.092	0.043	0.192	0.140	0.141	0.043	0.108	0.035
Tuscan	0.051	0.155	0.095	0.043	0.214	0.131	0.130	0.044	0.108	0.029
WestSicilian	0.049	0.154	0.094	0.044	0.193	0.137	0.142	0.043	0.108	0.036
Bulgarian	0.054	0.153	0.102	0.049	0.229	0.121	0.113	0.044	0.108	0.026
Croatian	0.054	0.147	0.100	0.045	0.261	0.113	0.110	0.041	0.105	0.025
Greek	0.051	0.159	0.095	0.045	0.211	0.134	0.122	0.044	0.109	0.029
Hungarian	0.054	0.141	0.102	0.046	0.276	0.107	0.107	0.041	0.102	0.024
Romanian	0.055	0.150	0.102	0.051	0.228	0.118	0.112	0.047	0.111	0.026
French	0.054	0.143	0.098	0.043	0.247	0.112	0.129	0.042	0.106	0.026
Spanish	0.052	0.136	0.089	0.042	0.231	0.116	0.160	0.042	0.099	0.034

B. Supplementary Figures

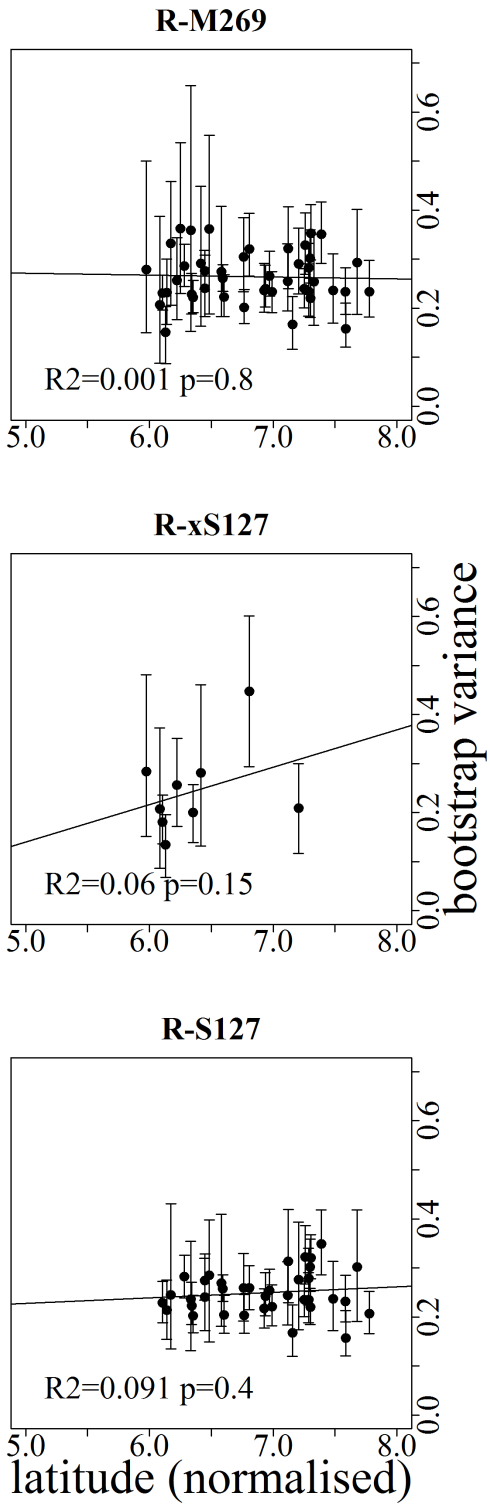

Figure B.1. The relationship between latitude and STR variance for 3 Y chromosome haplogroups. As with longitude, the plots show that there is no significant correlation between latitude and STR variance for the whole of the R-M269 haplogroup, as well as the two subhaplogroups: R-M269xS127 and R-S127.

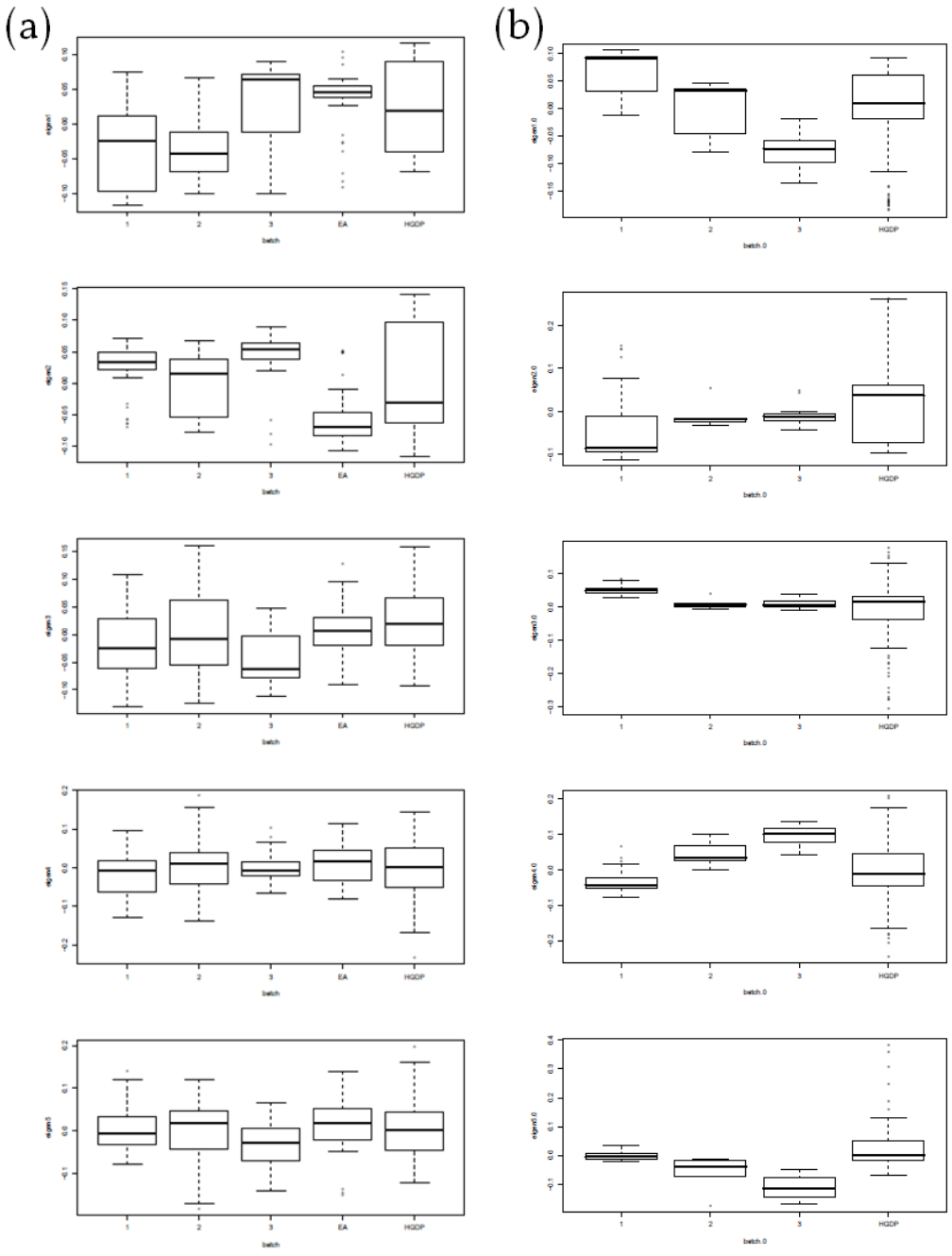

Figure B.2. Boxplots of the first five eigenvalues, grouped by batch and origin, of the samples in chapters 4 and 5, split into (a) European individuals and (b) non-European individuals. Although there seems to be some difference between the different batches, the following two figures (figuresB.3 and B.4) show that there is dues to the non-random assignation of populations to batches.

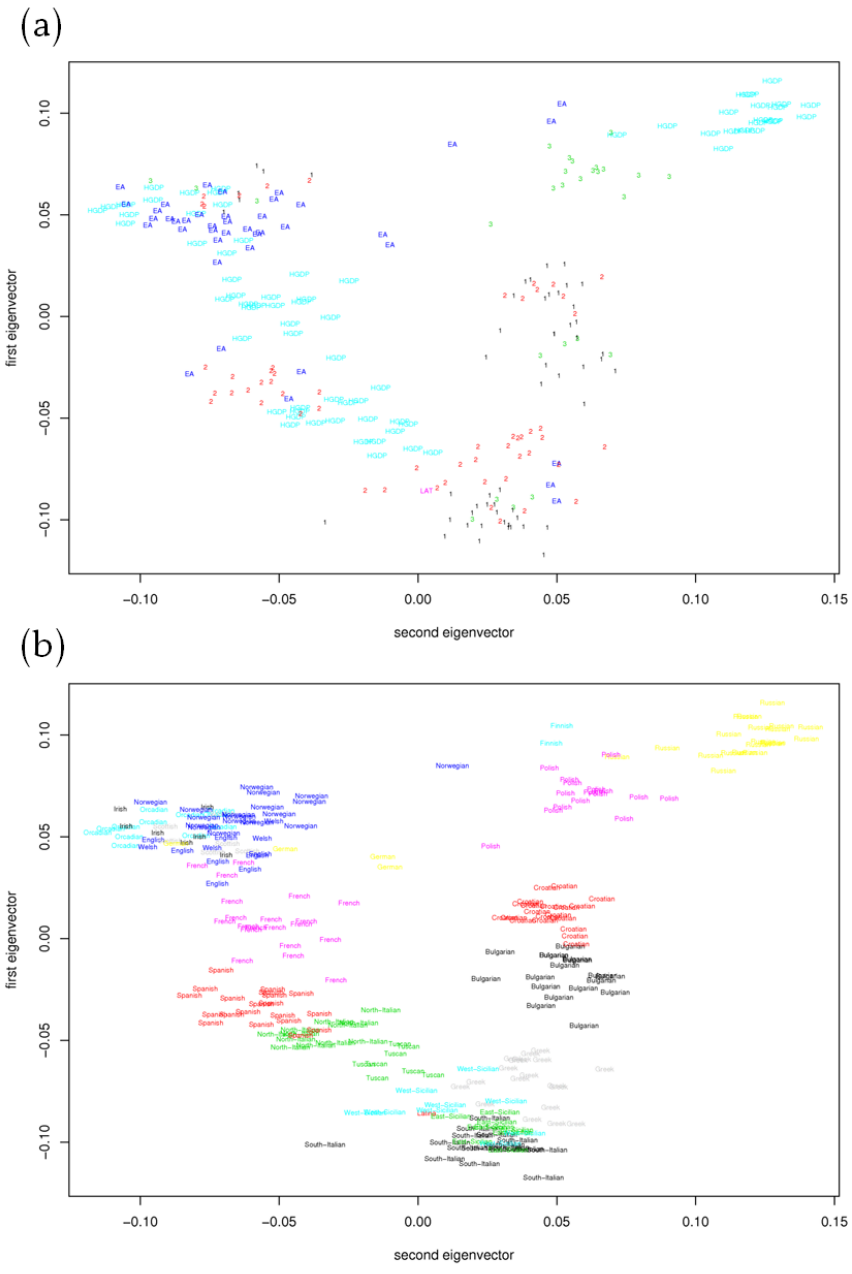

Figure B.3. PCA of European individuals used in chapters 4 and 5, labelled by (a) batch/origin and (b) population.

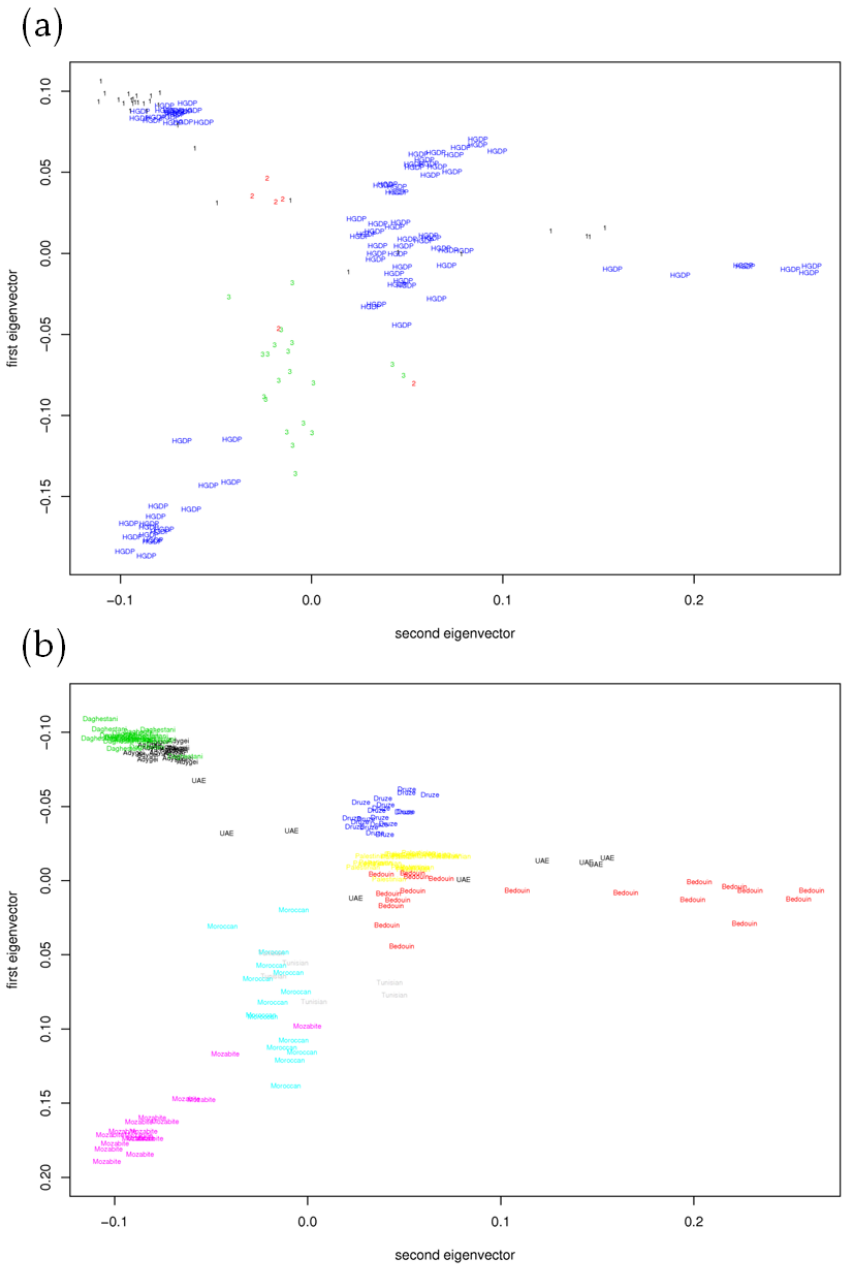

Figure B.4. PCA of North African and Middle Eastern individuals used in chapters 4 and 5 labelled by (a) batch/origin and (b) population.

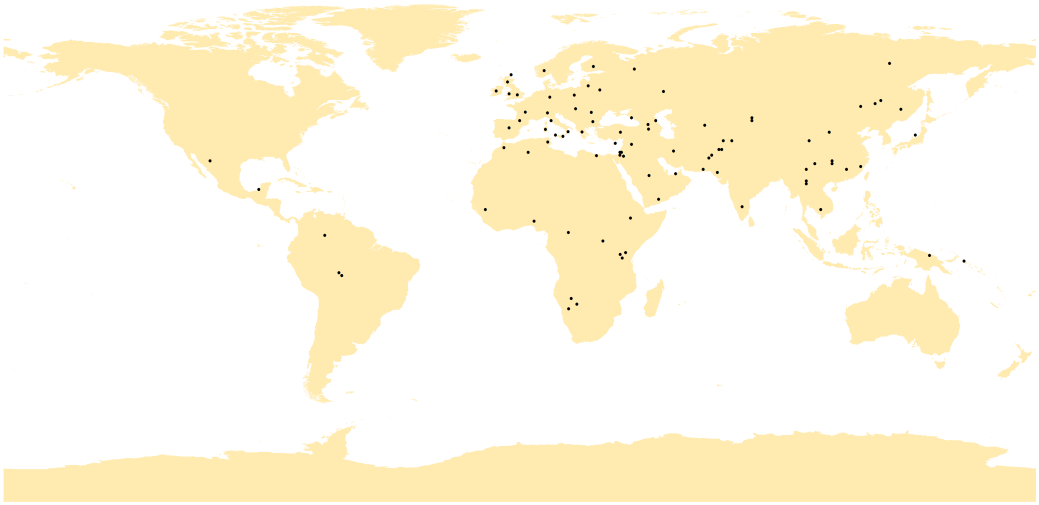

Figure B.5. Locations of samples used in genomic analysis in chapters 4 and 5

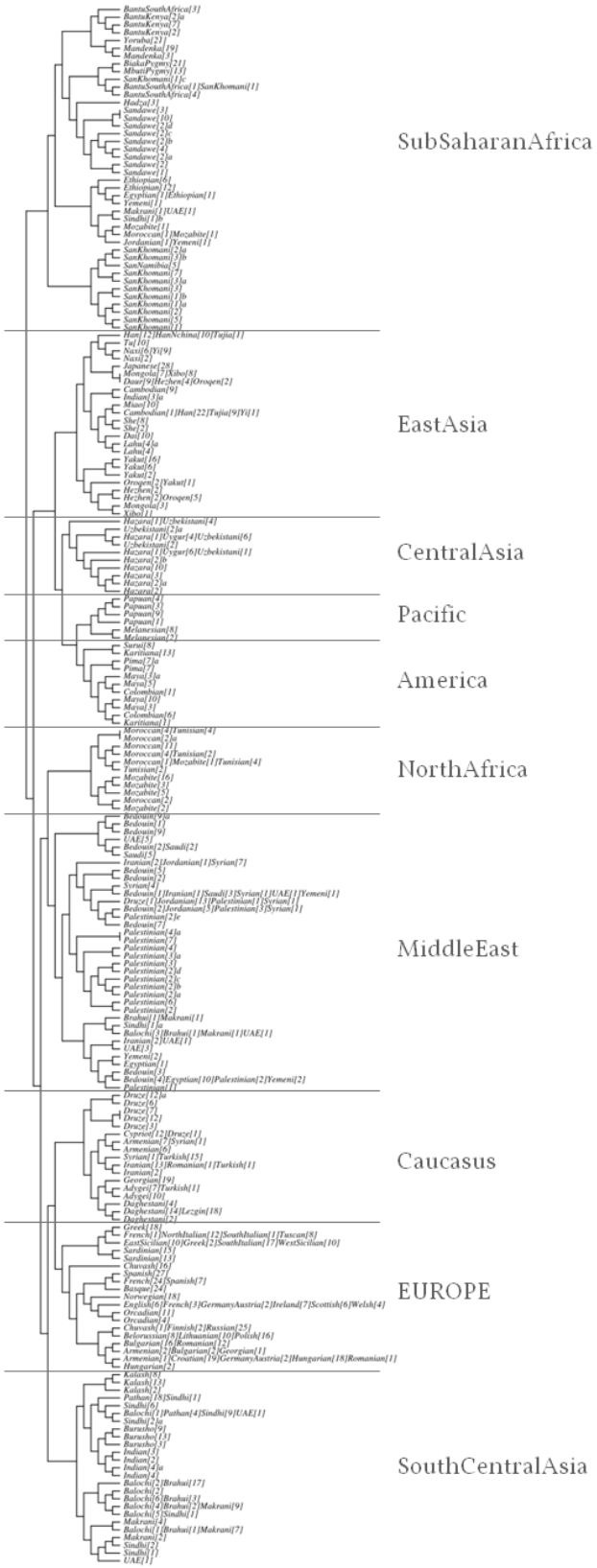

Figure B.6. The full *fineSTRUCTURE* tree based on the analysis of all individuals in chapter 4. Each leaf represents a *fineSTRUCTURE* cluster and is named for the populations and numbers of individuals from each population in a given cluster.

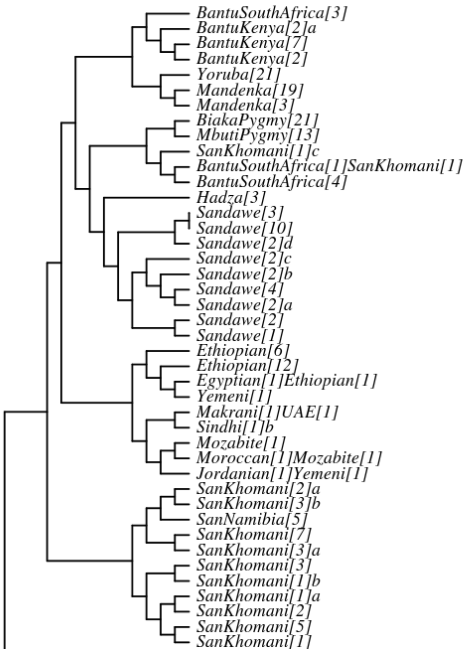

Figure B.7. The Sub-Saharan Africa section of the full *fineSTRUCTURE* tree shown in figure B.6

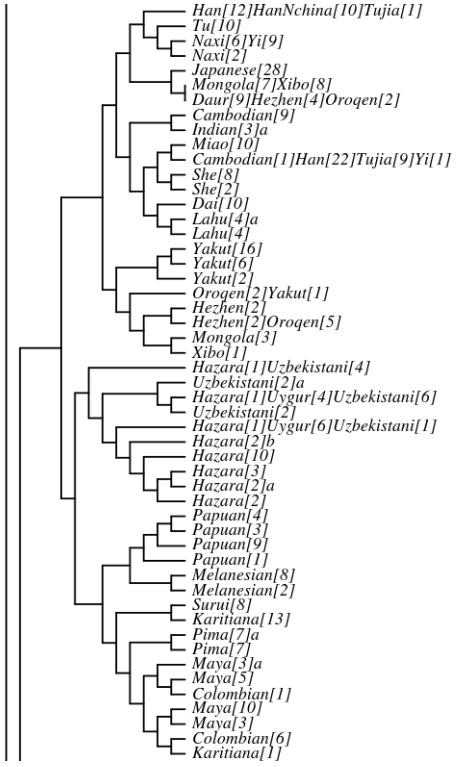

Figure B.8. The East and Central Asian section of the full *fineSTRUCTURE* tree shown in figure B.6

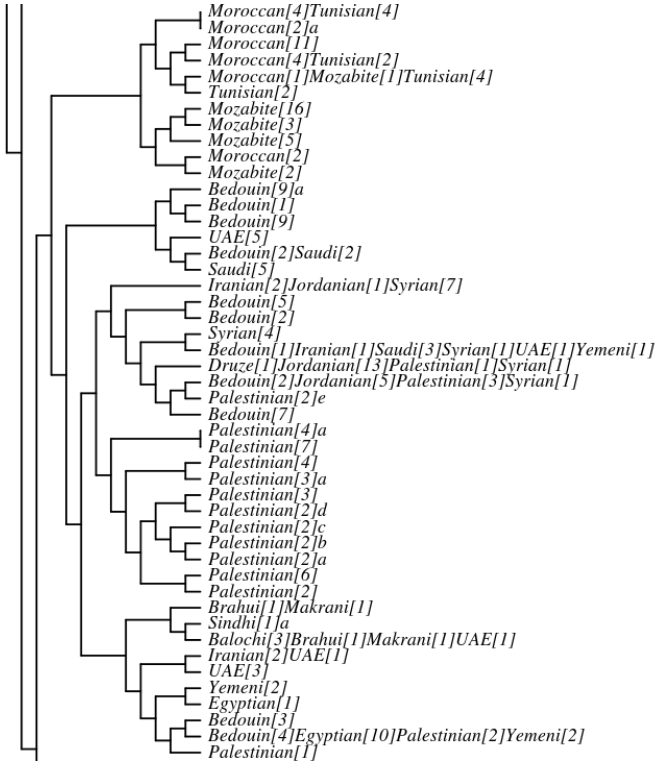

Figure B.9. The Middle Eastern and North African section of the full *fineSTRUCTURE* tree shown in figure B.6

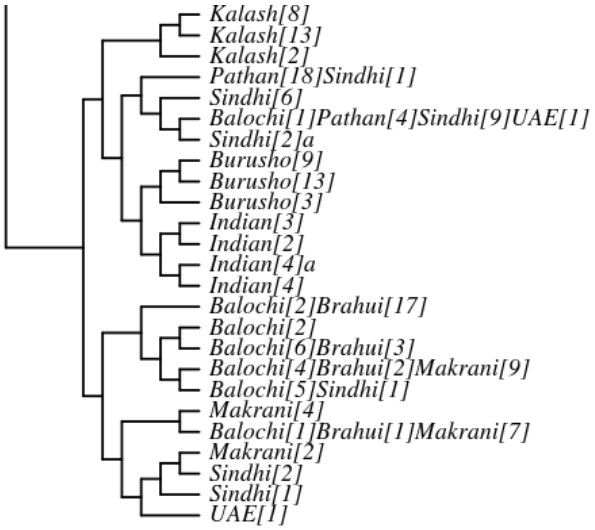

Figure B.10. The Central South Asian section of the full *fineSTRUCTURE* tree shown in figure B.6