Using genealogical trees to examine admixture between modern humans and Neandertals

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2016

Anna Frangou

Department of Statistics
24-29 St Giles
Oxford
OX1 3LB

afrangou@stats.ox.ac.uk

Acknowledgements

To Simon Myers, for his brilliance, ideas, and attention to detail, meaning I have produced a thesis far beyond what I thought possible. To Gil McVean for his comments, reassurance, and support. To Elspeth Garman, for her utterly endless patience, encouragement, and generosity. To my parents, for supporting me in so many ways, and having unwavering faith in my ability. To Luke Miller and Daniel Straulino, for behaving like five year olds with me in the office, making going to work something I genuinely looked forward to. To Jo Humphreys, Ellie Pearce, Reyhaneh Esmaielbeiki, and Hannah Edwards, for understanding me, pointing me in the right direction, and talking and laughing with me the whole way through. Genuinely couldn't have done it without you. You may all never read it, but it's done.

Abstract

This thesis uses genealogical trees to identify, date, and quantify patterns of admixture between Neandertals and individual modern human populations, using a combination of high quality data and parametric methodology. Previous methods on this subject have either approximated features of trees, or inferred them indirectly. Here, genealogical trees are used directly to understand the admixture process between humans and Neandertals by extending a recently developed method named *CEPHi*: Coalescent Estimation of Population History. *CEPHi* uses recombinationally cold regions of the human genome to build genealogical trees specifying the relationships between individuals in two input populations (one Neandertal, one human), including estimated population size histories, split times, and coalescence and mutation times.

Using *CEPHi*, a Neandertal-human population split time of ∼712,000 years in the past is estimated, as well as uncovering loci introduced by Neandertal-human admixture, revealing distinct bimodal distributions of estimated coalescence times between non-African and Neandertal haplotypes. A Neandertal population history is inferred, from the time of their split with humans up to ∼50,000 years ago (the fossil age), showing this archaic species to have suffered a bottleneck at this time, consistent with leaving Africa, followed by a further reduction to extinction.

Contrasting African-Neandertal and Eurasian-Neandertal analyses are used to define admixture using genealogical trees, and test our procedures in *CEPHi* via coalescent-based simulations. This region-level definition of admixture is used to specify sets of introgressed coldspots across 13 modern human populations. These sets are compared between pairs of populations, revealing information about the possible timing of interactions between Neandertals and modern humans, and sharing of admixture events between human groups, especially with respect to the split time between European and Asian populations. Online sets of introgressed regions for each of the four continents in our dataset are provided: African, American, Asian, and European.

Finally, in order to investigate the variation in time of contact between Neandertals and individual human populations, a novel method is described and implemented which dates admixture between individual human populations and Neandertals, using information from genealogical trees. Dates of admixture are estimated as ∼50-60,000 years in the past in European populations, and ∼80-90,000 years in the past in Asian populations, suggestive of potentially somewhat distinct histories between European and Asian populations. This method can be applied to date any set of introgressed regions, including those shared between particular populations, enabling a clearer picture of the joint evolutionary history of modern humans, Neandertals, and other archaic species.

Contents

Glossary

bp Base pair(s). 12

cM centiMorgan(s): a unit of recombinant frequency. 29

kb Kilobase(s). 14

kya Thousand years ago. 1

kyr cal BP Thousand calendar years before present. 3

LD Linkage disequilibrium: A measure of association between genomic positions. 19

MLE Maximum Likelihood Estimation. 24

MRCA Most Recent Common Ancestor. 16

mtDNA Mitochondrial DNA. 8

sd Standard deviation(s). 12

SE Standard error(s). 37

SNP Single Nucleotide Polymorphism. 15

List of Figures

List of Tables

CHAPTER 1

Interactions between *Homo sapiens* and *Homo neanderthalensis*: the knowns and unknowns

1.1 The human-Neandertal evolutionary story

Neandertals (*Homo neanderthalensis*) - alongside their sister species, Denisovans (*Homo altai*) - are our closest known evolutionary relative. There exists extensive fossil and archaeological evidence for Neandertals across Eurasia, whereas for Denisovans there is currently a finger phalanx and two molars, identifying three individuals from Denisova Cave in Siberia (Meyer *et al.* [2012]; Sawyer *et al.* [2015]). We focus on Neandertals throughout this work, which are thought to have evolved into their classic - as opposed to prototypical - form in Eurasia, perhaps from

a stem population of *Homo heidelbergensis* (Stringer [2016]). *Homo heidelbergensis* may have evolved in Africa and exited more than 700kya (thousand years ago), and in which case can be considered a potential common ancestor of both *Homo sapiens* and *Homo neanderthalensis*. Alternatively, the species may have existed only in Europe, and therefore be ancestral only to *Homo neanderthalensis*; it may even be that *Homo heidelbergensis* is too young a species to be the ancestor to humans and Neandertals (Meyer *et al.* [2016]). Debate certainly surrounds this topic (Stringer [2012, 2016]), but there exists substantial archaeological evidence for a significant Neandertal presence across Eurasia, stretching as far east as Siberia, between ∼400-30kya (Finlayson *et al.* [2006]; Mellars [2004]; Stringer [2012]). Classic Neandertal fossils so far discovered show their range to extend from western Europe, through Spain (Finlayson *et al.* [2008]), Italy (Mallegni *et al.* [1987]), Croatia (Wolpoff *et al.* [1981]), and east to Iraq (Trinkaus [1978]), Uzbekistan (Glantz *et al.* [2008]), Russia (Faerman *et al.* [1994]), and most recently extending into southern Siberia (Prüfer *et al.* [2014]). The current status of the archaeological record shows their extinction to have been diachronous, starting in northern Europe and moving south, with evidence to say that the last Neandertals survived in some southern and Atlantic refugia in Iberia (Finlayson [2004]), the Balkans (Higham *et al.* [2006]), and possibly the Levant (Belmaker and Hovers [2011]), up to 28kya. Uncertainty surrounds the date of the last Neandertal presence, however, with Higham *et al.* [2014] unable to reproduce such recent date estimations, and concluding the most recent date estimation for Neandertal to be 39,260ya. A subset of these fossils have so far been sequenced, and we use two Neandertal genomic datasets in this thesis. The draft Neandertal genome used for a single analysis in Chapter 2 is constructed from 3 fossils found in Vindija, Croatia (Green *et al.* [2010]), and the high coverage Neandertal genome used for all remaining analyses is taken from a toe bone found in Denisova Cave in the Altai mountains, Siberia (Prüfer *et al.* [2014]).

Homo sapiens first appears in the fossil record by ∼195kya, in Omo, Ethiopia (McDougall *et al.* [2005]). The earliest anatomatically modern human fossils found outside Africa are from the Skhul and Qafzeh caves in Israel, and are dated to between ∼80-120 thousand years old (Shea [2008]). Their presence is often thought to be the result of an initial migration out of Africa along the Nile Corridor and into the Levant, which may not have resulted in humans populating the

globe, due to climatic changes forcing population retraction (Petraglia and Rose [2009]) (this region is significant, however, in that it is thought to be one of the first possible points of contact between humans and Neandertals). Fossil evidence also suggests that there may have been another early route for humans out of Africa ~125kya, along the Bab-El-Mandeb Straits at the mouth of the Red Sea (Armitage *et al.* [2011]), and a recent study suggests that an early human migration from Africa around 120kya is evident in the genomes of modern Papuans and Philippine Negritos (Pagani *et al.* [2016]). Regarding the global expansion of *Homo sapiens*, previous research has postulated the existence of a second wave of human migration out of Africa between 65-40kya, thought to have resulted in the colonisation of Europe and Asia (Stringer [2000]), and more recent evidence supports this: a cluster of papers which sequenced and analysed a total of 787 human genomes from more than 270 populations agree that the vast majority of DNA in modern humans across the globe come from a major dispersal out of Africa somewhere between 50-80kya (Malaspinas *et al.* [2016]; Mallick *et al.* [2016]; Pagani *et al.* [2016]).

Humans then colonised Europe late into their global walk, somewhere between 45-35kya (Stringer [1996]). We know that populations of Neandertals were still in existence at this time, although the extent of their range isn't fully known. Given the nature of the Neandertal depletion, it is possible that the more northern Neandertal populations had already disappeared by the time modern humans reached those areas. Contraction of the Neandertal range was coeval with the arrival of modern humans, however we cannot assume a causal link. There has been substantial debate on this point in the archaeological literature, with climate, competition, and disease being suggested as possible explanations, to which we now turn.

The climate between 60-30kya was characterised by an intermediate-sized ice sheet, with warm interstadial periods being punctuated by cool Heinrich events, where the massive discharge of icebergs into the North Atlantic quickly cooled the ocean and land temperatures. It is thought by some that these cool periods, or perhaps the quick fluctuation between warm and cool periods - requiring fast adaptation - may have severely depleted Neandertal populations (Finlayson and Carrion [2007]), while human populations may have been better equipped to deal with

such climatic oscillations. Some refute this idea, citing the specifics of climatic events and the respective species ranges during these times, as well as Neandertals having survived multiple previous ice ages. Banks *et al.* [2008], for example, state that the ranges of both humans and Neandertals probably increased during Greenland Interstadial 8 (GI8: 38.6-36.5kyr cal BP (thousand calendar years before present)) - a warm period which followed Heinrich Event 4 (H4: 43.3-40.2kyr cal BP) - thereby allowing the potential distribution of the two species to overlap, making contact and competition between them possible. Specifically, the authors suggest humans would have had the opportunity to exploit Neandertal refugia, thus reaching the conclusion that competition - rather than climate - was the more direct cause of the Neandertal decline. Opinions on the events essentially range from believing Neandertals met a violent end at the hands of *Homo sapiens* (Lowe *et al.* [2012]), to considering a more indirect outcompetition of Neandertals by modern humans (Banks *et al.* [2008]), to their succumbing to the cold of the last ice age (Hublin and Roebroeks [2009]). Others takes a middling viewpoint (for example Stringer [1996]), suggesting that the demise of the Neandertal was likely due to a mixture of climate and competition; that humans may have advanced upon Neandertal habitat during an interstadial period, and with the onset of a Heinrich event, the double challenge of climate and competition may have favoured the better technologically prepared humans.

Aside from these potential determinants of Neandertal extinction, we know that modern humans and Neandertals interbred (Green *et al.* [2010]; Prüfer *et al.* [2014]), making disease also a potential contributor. An expanding population of humans exiting Africa and roaming into Neandertal territory may have had both negative and positive consequences for populations of both species; hybridisation may have conferred both protective and deleterious alleles upon both. Neandertals are thought to have lived in small groups of 15-30 individuals (Davies and Underdown [2006]), which may have been individually more susceptible to disease spread, however, living in small groups such as these may also have prevented the spread of disease on a larger scale. Candidates for transmission from humans to Neandertals include *Helicobacter pylori*, a bacterium causing stomach ulcers, and herpes simplex 2, which causes genital herpes (Houldcroft and Underdown [2016]). Humans may have benefited from Neandertal alleles protecting them against bacterial sepsis, and encephalitis contracted through Siberian forest

ticks (Houldcroft and Underdown [2016]). Differentiation with regard to disease susceptibility has recently been cited with regard to cooking with fire, thought to have been practised by both species. Hubbard *et al.* [2016] suggest that due to a nonsynonymous mutation in the aryl hydrocarbon receptor gene (*AHR*) in humans, that our response to environmental pollutants - including those produced when cooking meat - makes us less susceptible than Neandertals were to respiratory and reproductive diseases caused by polycyclic aromatic hydrocarbons.

It may be that one potent factor severely affected the overall population size of Neandertals, or that some mixture of intermediate level factors caused their eventual extinction. For example, given the probable size of Neandertal populations, it is possible that interbreeding essentially subsumed the Neandertal population into the human populations with which they had contact. Although the location and extent of human-Neandertal interactions is not known in any detail, we can still use archaeological and anthropological evidence to consider whether this amount of interbreeding was likely. Assuming the two species had sufficient geographical overlap, there were still a number of conditions that need to be met in order for interbreeding to have occurred to a large enough extent that we now see traces of it in modern human genomes. These include genetic, physiological, and perhaps cognitive compatibility.

Neandertals probably had the same karyotype as humans, making the two species (as well as Denisovans) likely to be chromosomally compatible (Meyer *et al.* [2012]). Morphologically, Neandertals are distinct from humans in a number of ways, typically exhibiting robust features such as thick femurs and patellas, and large deltoid and pectoral muscles, suggesting considerable strength. Cranofacially they differ from modern humans (see Figure 1.1), with a backward-sloping chin and forehead, a prominent brow ridge and projecting mid-face, as well as a lateral widening but flattening of the parietal lobes. Endocranial volume is very similar for the two species, ranging for both between ∼1200-1800cm^3 (Pearce *et al.* [2013]), but once body size is accounted for, the standardized endocranial volume is significantly lower in Neandertals (mean of ∼1134cm^3 as compared with 1373cm^3 in humans). Standardised endocranial volume is a common measure of cognitive capacity, however, there remains substantial debate regarding the cognitive abilities of Neandertals, with some attributing to them abilities equal to those of

(a) A highly detailed reconstruction created of *Homo sapiens* (left) and *Homo neanderthalensis* (right), by Dutch artist Alfonis Kennis, part of a recent exhibition at the Natural History Museum, London.

(b) A visual comparison of human (left) and Neandertal skulls. Absolute endocranial volume for both ranges between $1200 - 1800\text{cm}^3$, but Neandertal standardised endocranial volume is lower. Note the flattened parietal lobes, backward-sloping forehead and chin, and heavy brow ridge in the Neandertal.

Figure 1.1: Comparing *Homo sapiens* and *Homo Neanderthalensis*. Images taken from the Natural History Exhibition 'Britain: One Million Years of the Human Story'.

modern humans (Zilhão *et al.* [2010]), and others challenging this (Roebroeks *et al.* [2012]).

This debate centres around aspects of culture, including the creation and use of tools, and self-adornment, thought to be indicative of cognitive ability. It is known, for example, that Neandertals possessed their own material tool culture - Mousterian - during the Middle Palaeolithic, before transitioning to one more complex - Châtelperronian - in the Upper Palaeolithic. Whether this more complex tool culture was an independent progression, or a result of interaction or observation of *Homo sapiens* is debated, and fuels the ongoing discussion regarding the cognitive capacity of Neandertals. In the Grotte du Renne in central France, there exists an ornament layer containing stone blades, bone awls, and the pierced teeth of foxes and marmots used for creating pendants. This layer lies beneath another containing objects typical of *Homo sapiens*, as well as a fragment from a Neandertal skull (Hublin *et al.* [2012]), leading some to conclude they are Neandertal- created objects (Caron *et al.* [2011]), however, there remains debate about whether these layers are scrambled, fuelling controversy. Others claim the layers are retained, but that Neandertals were introduced to such technology by humans (Hublin *et al.* [2012]), given the contemporaneous nature of human entry into Europe and the increased complexity of Neandertal culture (45-40kya). It can be argued that even the capacity to take on a tool culture which used a wider range of materials and tool types demonstrates an equally high cognitive ability. Neandertals may also have buried their dead (Rendu *et al.* [2014]), created art (Heyes *et al.* [2016]; Pike *et al.* [2012]; Rodríguez-Vidal *et al.* [2014]), and practiced ritualistic behaviour (Jaubert *et al.* [2016]), all indicative of an advanced social structure, abstract thought, and symbolic expression, although there is some controversy surrounding these conclusions.

Pearce *et al.* [2013] approach this question from a different angle, concluding that the Neandertal brain was more heavily devoted to motor control of their large bodies, as well as to the visual cortex - having adapted to lower light levels in the northern territories that they occupied - while the human brain allowed better social cognition. Putative social differences such as this may have placed an obstacle in the path of levels of interbreeding higher than the ∼1-2% currently reported in the literature (Prüfer *et al.* [2014]).

1.2 The strength of archaeology lies in the dating

A spectrum of evidence from the archaeological sciences, supplemented by the cognitive sciences, suggests there was likely some significant interaction between humans and Neandertals somewhere between 120,000 and 30,000 years ago, in the Middle East, Europe, and potentially further east into Asia. The nature of these interactions could have taken many forms, which may have varied from place to place. A major strength of archaeology with regard to detailing the interactions between humans and Neandertals lies in the precision it permits when dating fossils. This is hugely useful for dating stratigraphic layers containing Neandertal and human tools and fossils, and we use the fossil age of the Altai Neandertal toe bone in Chapter 3 and beyond for our analyses. Accurate dating is also crucial when putatively hybrid fossils are discovered; sufficient temporal and spatial fossil variation alongside correct dating would bring the field close to outlining the admixture history of these two species. In the meantime, we are able to combine the accurate inference of the age of (sometimes putatively) hybrid fossils with the study of their ancient genomes, to provide more insights into Neandertal-human admixture.

1.2.1 Evidence of human-Neandertal admixture from hybrid fossils

Importantly, there exists some potentially very direct archaeological evidence of human-Neandertal interbreeding, in the form of hybrid fossils. The skeleton of a 4 year old boy from Lagar Velho in Portugal was found in 1999, and dates to 24.5kya (Duarte *et al.* [1999]). This is approximately 4,000 years after the last known Neandertal existed (Finlayson *et al.* [2006]), and is therefore thought to be the product of a persistent hybrid population. It displays intermediate morphology in the cranium, mandible, dentition, and postcranium. There is, however, debate on the interpretation of this skeleton, others believing it to be an example of a robust modern human (Tattersall and Schwartz [1999]). This is the most recent and therefore a very significant example of a possible human-Neandertal hybrid.

Riparo Mezzena - a rock-shelter in the Monti Lessini region of northern Italy - revealed a 40-30,000 year old mandible, the chin of which displays morphology similar to that of both

Neandertals and modern humans, neither fully receding (as with Neandertals) nor fully projecting (as with modern humans) (Condemi *et al.* [2013]). The individual's mtDNA is Neandertal, meaning there is a maternal Neandertal ancestor at some point in the past, but it remains impossible to be certain whether this specimen is a hybrid with one *Homo sapiens* and one *Homo neanderthalensis* parent.

We also see specimens which are morphologically human or very close to human, but on sequencing are shown to contain some proportion of Neandertal sequence. One example of this is Oase 1, a Romanian individual found in the Petera cu Oase who died between 37-42kya (Fu *et al.* [2015]). Similar to modern humans, Oase 1 exhibits a rounded cranium and protruding chin. His Neandertal features include a very wide ramus (the vertical part of the mandible or lower jaw) and large distal molars (Trinkaus *et al.* [2003]). His genome contains 6-9% Neandertal DNA, interestingly substantially more than the 1-2% reported for modern humans existing now (Prüfer *et al.* [2014]). Some segments of Neandertal ancestry are longer than 50cM, indicating a very recent Neandertal ancestor about 5 generations before he existed. Another is a 45,000 year old human from the banks of the Irtysh river near Ust'-Ishim in Western Siberia (Fu *et al.* [2014]), who is reported to be 2.3% Neandertal admixed, close to estimates for modern human populations. However, the stretches of Neandertal regions in his genome are substantially longer, suggesting admixture occurred somewhere between 7-13,000 years before he existed, and coinciding with the expansion of modern humans in Europe.

Archaeological evidence of hybrid fossils are themselves significant evidence for the existence of admixture between humans and Neandertals, but they are strengthened considerably when combined with genetic analyses to investigate the extent and timing of the admixture events. We note however that it is not clear how the now widely used sequenced genomes taken from the ancient individuals relate to the source and timings of the Neandertal DNA we find in modern human genomes. We return to the topic of the timing of admixture events in Chapter 5 when we use a novel method to date admixture across modern human populations from 4 continents. We now turn to reviewing the evidence that has been presented regarding human-Neandertal admixture from the field of genetics.

1.3 Ancient-modern admixture: a genetic perspective

Over the last few years - and given the recent innovations regarding the sequencing of ancient genomes - conversation on the subject of Neandertal-human interactions has shifted from being a heavily archaeological question to a genetic one, and the theory receiving the most attention is that of the two hominid groups interbreeding: a process called admixture.

The occurrence of admixture between our species and others has an effect on our depiction of the evolutionary route of humans. Four main models describe the origin and journey of our species, shown in Figure 1.2. First is the Recent African Origin model (RAO) (Stringer and Andrews [1988]), which states that modern humans left Africa and gradually populated the globe, outcompeting or otherwise eradicating other hominid populations as they were encountered. Second is the African Hybridization and Replacement model (Mellars *et al.* [2007]), which differs from the RAO model only in that it allows for a small amount of hybridization with other hominids during population expansion. Third is the Assimilation model (Aiello [1993]), which lies close to the African hybridization and replacement model, but places a greater emphasis on interbreeding over replacement during the expansion of *Homo sapiens*. Last is the Multiregional model (Wolpoff *et al.* [2001],) which states that there is no recent African origin of modern humans, and instead that the first hominids of the genus *Homo*) left Africa ∼1.8mya, and that this variable but single species eventually resulted in *Homo sapiens*.

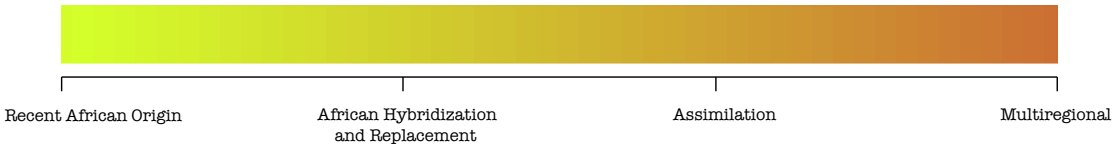


Figure 1.2: The continuum of modern human evolutionary models. Moving from left to right, the amount of interaction and mixing between *Homo sapiens* and other hominid species increases.

Debate has progressed from pitching the two models at the extremes of this spectrum as direct competitors of one another to an implicit consideration of the Assimilation model - essentially a middle-way description. This shift can be attributed in large part to the publication of the original Neandertal and Denisovan genomes (Green *et al.* [2010]; Reich *et al.* [2010]) reporting

evidence of admixture between these populations and modern humans. Some work countering this claim also exists, which we discuss below. First, we briefly describe the results of previous analyses of ancient DNA before whole genome analysis of archaic populations became possible.

1.3.1 Mitochondrial DNA (mtDNA)

Before the Neandertal genome became available for analysis, detecting admixture used mtDNA, and following that on short segments of nuclear DNA. Analysis of mtDNA predominantly involved simple pairwise comparisons between Neandertal and modern human sequences. Between 1997 and 2008, ∼17 regions of mtDNA from various Neandertal specimens were sequenced and analysed in this way. Typically, hypervariable regions in the control region of the mitochondria were used, as they display the most polymorphism in modern human populations, and can therefore reasonably be assumed would reveal divergence if it is present. Taken from Green *et al.* [2008], Figure 1.3 summarises the consensus conclusion from these papers: that humans form a monophyletic group to the exclusion of Neandertals, and therefore that admixture did not impact the mtDNA of either species.

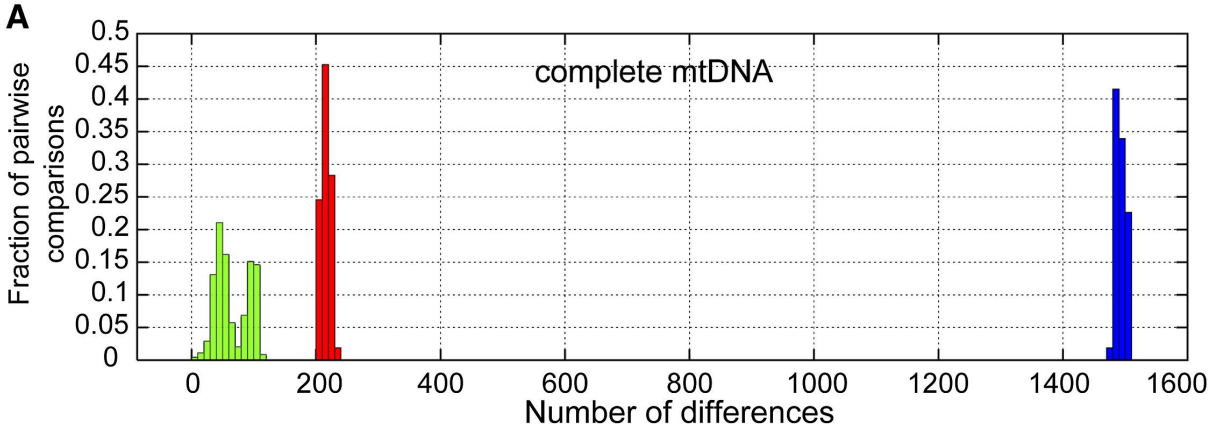


Figure 1.3: Reproduced from Green *et al.* [2008], we can see the extent of mtDNA sequence differences between humans from 53 populations (green), between humans and Neandertals (red), and between humans and chimpanzees (blue). The separation of humans and Neandertals led the authors to conclude that no admixture had occurred between the two species.

However, a lack of evidence for admixture at a single locus does not severely reduce the possibility of it having occurred, and the lack of Neandertal mtDNA in humans is in fact only informative

about a lack of interbreeding between Neandertal females and human males (although of course it is possible that Neandertal mtDNA in humans has been lost through processes such as drift), as opposed to a lack of interbreeding between Neandertal males and human females. Before the first publication of a full ancient genome in 2010 (Green *et al.* [2010]), two papers sequenced regions of Neandertal nuclear DNA. One (Green *et al.* [2006]) was subsequently criticised for a large amount of modern human contamination (Wall and Kim [2007]). The other successfully sequenced 65,250bp (base pairs) (Noonan *et al.* [2006]), and worked from the premise that if admixture from Neandertals into modern humans had occurred, that it would manifest itself in an abundance of low frequency derived alleles in those populations with which Neandertals interbred, matching those in Neandertals. No site in their human dataset showed a low frequency derived allele that matched the Neandertal, and it was reported to be unlikely that admixture had occurred between the two species, thereby strengthening conclusions from mtDNA.

However, the ability to sequence the entirety of the nuclear genome of ancient individuals - to which we now turn - has transformed our understanding of admixture.

1.3.2 Landmark genomes

Two pairs of landmark papers sequenced and investigated the genomes of Neandertals and Denisovans (a sister hominid of the Neandertal, sequenced by Reich *et al.* [2010]). Green *et al.* [2010] provided a 1.3× coverage draft shotgun-sequenced version of the Neandertal genome from 3 bones from Vindija cave in Croatia, with a 52× coverage version following later taken from a toe phalanx found in Denisova cave in the Altai mountains in Siberia (Prüfer *et al.* [2014]). The equivalents for the Denisovan individual are Reich *et al.* [2010] and Meyer *et al.* [2012], sequenced from a distal manual phalanx of a juvenile also found in Denisova cave.

Since the publication of these archaic genomes, a variety of approaches for discerning whether admixture occurred between modern humans and ancient populations have been used. In Green *et al.* [2010], the D-statistic - a measure of the relative proximity of one individual to two others - was first used. It makes use of phylogenetic logic, stating that if two populations have admixed to the same extent with a specified archaic population (whether zero or complete admixture), that

they will share the same number of derived alleles with that archaic population (an outgroup - usually the chimpanzee - is used as a reference point). In this case, D will be at or close to zero (thresholds vary but ordinarily, deviations within 2 standard deviations are deemed insignificant). Neandertals are shown to be equally related to any combination of non-Africans or Africans (for example, D(ASN, CEU, Neandertal, chimpanzee) $= -0.53 \pm 0.46\%$, <1.2 sd (standard deviations) from 0%, or $p = 0.25$), but significantly closer to any non-African than African individual (for example D(YRI, CEU, Neandertal, chimpanzee)$= 4.57 \pm 0.39\%$, >11sd from 0% or $p < 10^{-12}$). Further comparisons were made within and between Africa and Eurasia using additional genomes from a French, Han Chinese, Papuan, Yoruban, and San individual, and the same pattern is shown: no within-continent comparison shows skews in $D(|Z| < 2\text{sd})$, all between-continent comparisons show skews in $D(|Z| > 7\text{sd})$. Similarly, in Reich *et al.* [2010] (the first publication of the Denisovan genome), extensive use of the D-statistic shows Melanesians to be significantly more similar to the Denisovan as compared with Eurasians or Africans, as well as confirming the previously shown proximity to the Neandertal between Africans and non-Africans. Prüfer *et al.* [2014] further consolidated this result, reporting French and Han Chinese individuals to show significantly greater proximity to the Neandertal as compared with the Dinka or San (both sub-Saharan populations). Denisovans are also shown to be closer to Oceanians (Aboriginal New Guineans, Australians, and Philippinos) than Eurasians.

Prüfer *et al.* [2014] also employ the enhanced D-statistic which uses only positions where sub-Saharan Africans carry the ancestral allele, thereby enriching for mutations originating in the archaic population. This shows Europeans (French, Sardinian) to be more divergent from the Altai Neandertal than East Asians (Dai, Han, Karitiana, Mixe) and Oceanians, and East Asians to be further than Oceanians, a result previously reported (Wall *et al.* [2013]).

The greater distance between Africans and Neandertals as compared with that between non-Africans and Neandertals is certainly clear from use of this statistic. Limitations of the D-statistic include the use of single individuals to represent a population, as well as simply being a measure of the genetic proximity between individuals. It does not, for example, distinguish between admixture and other demographic scenarios (such as ancient African substructure,

discussed further below) as the cause of observed patterns.

A second line of evidence in Green *et al.* [2010] looked to detect putative signals of gene flow from Neandertals to humans - the rationale being that introgression may be revealed via a genomic region with low divergence to the Neandertal (<60%) and high divergence to other humans. Low divergence from the Neandertal sequence is a necessary but not sufficient condition for Neandertal origin; high divergence from other humans ensures it is not simply a region of low mutation rate. To test this, 2,825 >50kb regions of African ancestry, and 2,797 regions of European ancestry were selected from the human reference genome. European regions with low divergence to Neandertals were shown to have a high divergence to other present-day humans (∼140% of the genome-wide average), while African regions with low divergence to Neandertals diverge much less from other present-day humans (∼35% of the genome-wide average). For clarity, human-Neandertal divergence is reported as a proportion of human-chimpanzee divergence, and normalised by the average genome-wide absolute human-Neandertal divergence, making the differences between Neandertals and individuals of different ancestry more visible. This cutoff gives 32 regions, 30 of which are European (94%), and only two of which are African. The authors highlight that reads from three Neandertal bones were used to construct the Neandertal sequence, making the final genome hexaploid, weakening power to detect low-divergence regions. Furthermore, given the use of the human reference genome, the regions used to represent Europeans and Africans are by definition different and therefore not directly comparable.

However, in Prüfer *et al.* [2014], this analysis was extended, using the high coverage Altai Neandertal genome and 13 phased human genomes, classified as European (Sardinian, French), Eastern (Karitiana, Han, Dai, Mixe), Oceanian (Papuan, Australian), and African (Yoruban, Mbuti). Low divergence between a human and the Altai Neandertal was shown to be associated with high divergence between that individual's two haplotypes (i.e. heterozygosity), and these regions are seen almost exclusively in non-African populations. The same is true for the Denisovan, but applies only to Oceanian populations. Interestingly, in the Altai Neandertal analysis, the Oceanian populations look most similar to the Neandertal and have the highest intra-individual heterozygosity, followed by East Asians, and then Europeans. This is an

emergent pattern in similar analyses.

Lastly, the search for evidence of admixture in Green *et al.* [2010] used sequence data from modern humans alone. As with the previous method, the aim was to identify candidate regions for gene flow, but this time by locating those regions with considerably higher diversity outside Africa than inside. It is known that genetic diversity outside Africa is largely a subset of that within it, and therefore regions which flout this expectation may be indicative of Neandertal admixture, explicable through Neandertal introgression. Specifically, this was done by identifying alleles at ∼1.2 million SNPs (Single Nucleotide Polymorphisms) in 24 European Americans (CEU), 24 East Asians (ASN) and 23 Africans (YRI). The genome was split into 50kb regions, and a statistic $S_T = \frac{T_{OOA}}{T_{AFR}}$ was calculated for each, where T_{OOA} is the time to the root of the Out of Africa samples (including both CEU and ASN) and T_{AFR} the African equivalent. Regions where six consecutive 50kb windows had a value of S_T in the top 0.5% of regions across the genome - i.e. where the time to root was highest in the Europeans relative to the Africans - were chosen as candidates for gene flow from Neandertals. For each of these regions, the OOA and AFR samples were combined to produce a genealogical tree, and mutations were identified as cosmopolitan (shared by OOA and AFR), or divergent (present only in OOA). Four types of 'tag' SNPs are labelled: Derived Match (DM), 'derived' in that the Neandertal does not match the chimpanzee allele at this position, 'match' in that the Neandertal carries the divergent SNP, and Derived Non-match (DN) meaning the Neandertal carries the non-chimpanzee allele which is cosmopolitan (Ancestral Match and Ancestral Non-match are defined equivalently); an excess of DM SNPs is evidence of Neandertal gene flow. It was shown that Neandertals match the divergent OOA clade at 133/166 tag SNPs, as well as 10/12 putative admixture regions containing an excess of divergent over cosmopolitan sites, consolidating previous results. This analysis is not repeated in the most recent publication (Prüfer *et al.* [2014]).

It is concluded that the reported results can be best explained via the occurrence of admixture between humans and archaic populations. However, the authors also highlight the fact that ancient substructure in African populations may also explain the greater genetic similarity between Eurasians and Neandertals. We turn to this less frequently addressed possibility in the

next section.

1.3.3 Admixture or ancient African substructure?

Admixture between modern humans and Neandertals is often invoked as the most parsimonious explanation for the observations made by various genetic studies. A competing hypothesis has received less attention: that of population substructure in ancient Africa. If African populations exhibited significant genetic differentiation when the ancestors of Neandertals left Africa, and this substructure persisted until the ancestor of modern humans exited Africa from the same root population, we would expect to observe two things. Firstly, a greater genetic similarity between non-Africans and Neandertals as compared with Africans and Neandertals, and secondly, some variation in proximity between Africans and Neandertals. We demonstrate the respective scenarios of admixture and ancient substructure in Figure 1.4.

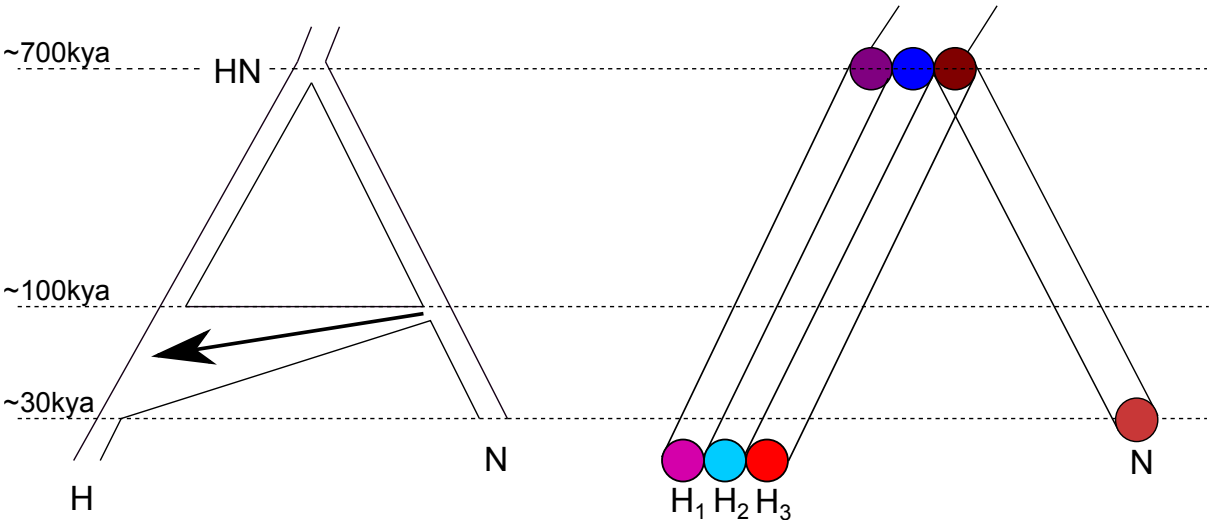


Figure 1.4: Two possible scenarios of human and Neandertal coevolution. The left diagram shows an admixture scenario, the right shows persistent population substructure in Africa. Left: humans and Neandertals may have speciated ∼700kya (see Chapter 3), their most recent common ancestor (MRCA) being HN (Human-Neandertal). We use 100kya and 30kya as bounds for admixture, represented by the large arrow. Right: a structured African population ancestral to humans and Neandertals (HN), represented by coloured circles. The ancestors of Neandertals leaving Africa came from one subsection of the HN population (dark red). Substructure remained in Africa for an extended period, at least until the human exit. Humans exiting Africa (H_3) also came from the red population, explaining their greater proximity to Neandertals than that of other populations (H_1, H_2). An accurate representation of African substructure is not attempted.

The data is not yet available for elucidation of the nature and extent of ancient population struc-

ture across Africa using samples from ancient humans; it is nevertheless interesting to observe the structure present in modern Africa. Tishkoff *et al.* [2009], for example, applied Principal Components Analysis (PCA) and a clustering program (*STRUCTURE* - Pritchard *et al.* [2000]) to a dataset containing 1,327 nuclear microsatellites and indels from 2,432 Africans from 121 geographically diverse populations. Fourteen ancestral population clusters were identified, which correlate well with self-described ethnicity and shared cultural and linguistic properties, demonstrating a relatively high current level of African population substructure.

Some subtle but significant differences among African populations as regards their proximity to Neandertals also exist, as measured by the D-statistic. Wang *et al.* [2013] applied the D statistic to 38 individuals from 8 sub-Saharan African populations: 4 hunter-gatherer (Pygmy, San, Hadza, Sandawe) and 4 other (Yoruba in Ibadan, Nigeria (YRI), African ancestry in SW US (ASW), Luhya in Webuye, Kenya (LWK), and Maasai in Kinyawa, Kenya (MKK)) groups. Overall they found some significant variation in African populations with regard to proximity to the Neandertal, amongst this (a) that non-African populations share significantly more derived alleles with the Neandertal than African populations, (b) that MKK, ASW, and Sandawe all share significantly more derived alleles with the Neandertal than the remaining African populations, and (c) that ASW shares significantly more derived alleles with the Neandertal than Pygmy, LWK, and YRI. Three potential explanations are offered for the observed variation. The first is that gene exchange has occurred between Africans since introgression from Neandertal-like genomes, but not into the population that gave rise to the out of Africa population, causing Africans to show more similarity to one another than to non-African populations, but to still display some structure. The second is that African-archaic admixture happened before admixture between archaic populations and non-Africans, meaning the mixing populations were genetically more similar, producing low D-statistics (although this may not account for the large differences in proximity seen between non-Africans and Africans with regard to the Neandertal). The third is that there has been recent introgression into sub-Saharan African populations from non-Africans. The final possibility is examined using *ADMIXTURE* to assign each individual a proportion of admixture from each of three (African (AFR), Asian (ASN), European (EUR)) populations. They find a high correlation between the amount of non-African

admixture in African populations and the D-statistics, concluding that this supports the recent admixture of non-Africans with Africans.

There is further evidence supporting a hypothesis of ancient back migration of Eurasian populations into Africa. Recently, Llorente *et al.* [2015] sequenced a 4,500 year old hunter gatherer from Mota cave in the Ethiopian highlands, and found his genome to be suggestive of migration into Eastern Africa by Middle Eastern farmers, around 3,500 years ago. Mota is most closely related to the Ari, a modern Ethiopian highland population, as shown via a comparison of a 250kb region of the genome (sequenced to 12.5× coverage) to 40 African and 81 European populations. Significantly, Mota lacks 4-7% of the DNA that the Ari and all other Africans carry. This region of DNA in Mota most closely matches that of modern Sardinians and a German prehistoric farmer, strongly supporting a European backflow into Africa. Other papers fall in agreement with regard to a ∼3,000 year old west Eurasian backflow in to Africa, which has left strong signatures in the the whole of Eastern Africa (Pagani *et al.* [2012]) and, to a lesser extent, Southern Africa (Pickrell *et al.* [2014]). We return to this in Chapter 4, where we investigate sets of Neandertal- introgressed regions visible in African populations.

Africa is a continent which should be given high priority with regard to sequencing ancient populations, despite the environmental limitations on the preservation of remains. Perhaps given this current lack of genomic data for ancient African populations, some subsequent studies have taken a different approach, using simulations to investigate the possibility of ancient substructure accounting for current observed genetic patterns (Durand *et al.* [2011]; Eriksson and Manica [2012]).

Eriksson and Manica [2012]) generated synthetic human and Neandertal genomes from a structured ancestral population, concluding that such structure could have produced values of the D-statistic seen in the Neandertal draft genome. This paper was the first to build a model of the shared history of modern humans and Neandertals incorporating spatial population structure in Africa, and thus explicitly considering ancient population structure in Africa as an explanation. Human and Neandertal populations were separated into a number of connected but discrete demes which can only mix with their neighbours, rather than assuming one panmictic popu-

lation. Combinations of parameter values (such as the number of individuals in a deme, the proportion of migrants exchanging demes, and the population growth rate) gave a time to most recent common ancestor T_{MRCA} under the model. Combinations producing an average T_{MRCA} most closely matching those calculated from 51 real populations from the HGDP-CEPH panel (Cann *et al.* [2002]) were used to generate simulated Neandertal and human genomes for each of these demographic scenarios. D-statistics were then calculated for Europe and Africa, Asia and Africa, and Europe and Asia, and were shown to fall close to those estimated in that paper. This led the authors to conclude that the excess polymorphism shared between Eurasians and Neandertals is compatible with various scenarios in which no admixture occurred, and may in fact be strongly linked to the strength of population structure in ancient populations.

Eriksson and Manica [2012]) support ancestral population structure using one summary of the data: the D-statistic. This is, however, not a common conclusion, and other aspects of the data would need to show agreement. In fact there exists significantly more evidence suggesting admixture explains the observations, while directly addressing ancient structure. This includes both Sankararaman *et al.* [2012] and Yang *et al.* [2012], who found substructure to be incompatible with particular aspects of the data.

Sankararaman *et al.* [2012] distinguished between the competing hypotheses by inferring the date of most recent gene flow between the two species by measuring linkage disequilibrium (LD) in the genomes of present-day Europeans. The admixture hypothesis is supported if this date is significantly more recent than speciation, and indeed the study reports that the last gene-flow from Neandertals into Europeans most likely occurred 37-86kya. The paper examines LD patterns in 59 West Africans, 60 European Americans, and 60 East Asians using a statistic, $\bar{D}(x)$, which gives the excess rate of occurrence of derived alleles at two SNPs compared with the expected frequency if they were independent. This is given as $\bar{D}(x) = \frac{\sum_{(i,j)\in S(x)} D(i,j)}{|S(x)|}$, where $D(i,j)$ denotes the classic measure of LD at SNPs i, j, making $\sum_{(i,j)\in S(x)} D(i,j)$ the sum of LD for all pairs of SNPs, and $|S(x)|$ all pairs of SNPs at genetic distance apart x (binned by this distance because the probability of recombination varies dependent upon it). SNPs were chosen under a particular ascertainment scheme that maximises the chances of identifying admixture

SNPs in amongst non-admixture SNPs, requiring that SNPs are (a) derived in the Neandertal, (b) polymorphic in humans, and (c) have a derived allele frequency in humans of <0.1.

All pairs of SNPs under this scheme and at most 1cM apart were used, and an exponential curve fitted to the decay of LD. The curve for Europeans and East Asians shows longer range LD (and a slower exponential decay) where the Neandertal carries the derived allele compared to where it carries the ancestral allele. In West Africans, LD is shorter range (decaying more quickly) where the Neandertal carries the derived allele. In a non-admixed scenario, where the Neandertal carries the ancestral allele, the mutation occurred on the human lineage and is therefore more recent than if both carry the derived allele, in which case the mutation occurred before the Neandertal-human split. Given this, we expect LD in that population to be higher where the Neandertal carries the ancestral allele and less where he carries the derived, because the mutation is younger. This pattern is observed in the West African populations, requiring no invocation of admixture. However, in Europeans and East Asians, high LD is seen where the Neandertal carries the derived allele, suggestive of admixture. The scale of LD at these pairs of SNPs thus provides information about the date of gene flow, and is estimated, after modelling the impact of errors in the recombination map, to be 1,597 generations in Europeans, far subsequent to the 5,000 generation cutoff given above. The authors therefore conclude that recent admixture has indeed occurred between Neandertals and non-Africans.

Additionally, Yang *et al.* [2012] used a doubly conditioned site frequency spectrum ($dcsfs$) to distinguish between the admixture and population structure hypotheses: the conditions being that the Neandertal must carry the derived allele, and one randomly selected African must carry the ancestral allele. This enriches for sites shared only between Neandertals and non-Africans. Given these conditions, the site frequency spectra for non-Africans shows an excess of rare alleles which was present in all simulated admixture scenarios but not in ancient structure scenarios, leading to the conclusion that admixture is likely to have occurred, whilst not ruling out additional ancient population structure.

We must take seriously the probability of a complex evolutionary history of humans and archaic species - possibly far more complex than a single or small number of admixture events from

archaic populations. Although multiregional theory has essentially been dismissed over the past decade, there is nevertheless a progressive shift away from the Recent African Origin model (which doesn't allow for admixture) towards the theory of 'leaky replacement', thus allowing a more complex history of the human species and its close relatives. It is likely that this will become evermore intricate as more ancient genomes are sequenced and their relationships with human populations examined. This is already apparent in Prüfer *et al.* [2014], where an additional archaic individual from the east is inferred to have contributed to the Denisovan genome, and will no doubt be added to over time.

1.4 Admixture: the knowns and the unknowns

There is significant evidence that admixture between Neandertals and modern humans has occurred, using both archaeological and genetic sources and methodology: we see probable hybrid fossils (Condemi *et al.* [2013]; Duarte *et al.* [1999]; Fu *et al.* [2015]) alongside genetic signatures that reveal non-African human populations to contain 1-2% Neandertal DNA (Prüfer *et al.* [2014]; Sankararaman *et al.* [2014]; Vernot and Akey [2014]). Similar genetic evidence exists for the Denisovan, with Melanesian and Australian Aboriginal populations having 3-5% Denisovan DNA (Meyer *et al.* [2012]), but given it is a species classified from a single finger bone (Mednikova [2013]) - although two additional molars have since been found (Sawyer *et al.* [2015]) - there is significantly less archaeological evidence. Intriguingly, a 400,000 year old femur found in the Sima de los Huesos in Spain was shown to have mtDNA closer to the Denisovan than the Neandertal (Meyer *et al.* [2014]), although recent investigation into the nuclear DNA of individuals in the Sima de los Huesos shows two specimens to be more closely related to Neandertals as opposed to Denisovans (Meyer *et al.* [2016]).

Despite the knowledge that our species interbred with other ancient humans, the details of mixing remain uncertain. We know little surrounding the specifics of where and when the populations met one another and interacted to significant enough an extent for their DNA to be visible in modern humans across Eurasia. It may be that there was one main meeting

place in the Middle East when populations of humans left Africa, and migrated and expanded across the continents, it may also be that there were multiple subsequent meetings in Asia and Europe. We are able to move somewhere closer to understanding the process by comparing sets of admixed regions (created via a new method) across populations and observing how much overlap exists. If sets are very similar, this naturally points to introgressed regions having entered the population at a time before the Asian-European human population split. If very different, perhaps the story of admixture has been more complicated. Using these sets, we are also able to deduce from where any admixture seen in African populations may have come.

Little is also known about both human and Neandertal population sizes throughout the history of both species. Learning about this means we can understand more about possible population dynamics. There is some limited archaeological and genetic evidence to show that Neandertal population sizes were likely small (Prüfer *et al.* [2014]), and we are able to significantly contribute to this line of enquiry by showing their changing sizes in times more recent than the Neandertal-human MRCA, using a new method based on the creation of genealogical trees. This allows us to consider questions surrounding the nature of the human-Neandertal split. We are also able to demonstrate and discuss the decline of the Neandertal species using information about their effective population sizes.

Significantly, we are able to contribute to the field by using an orthogonal approach to inferring the date of admixture between Neandertals and multiple individual human populations, giving us a more detailed picture of admixture across the globe. At this point, there is little work on this, with only Sankararaman *et al.* [2014] using modern human genomes to investigate this question (although some work exists which uses more ancient human genomes, as discussed in Chapter 5). Notably our method is in complement to those employed which leverage patterns of LD for inference, as we use regions of the genome which rarely recombine throughout our work (which we refer to as 'recombinationally cold regions'), as we explain below.

In addition, the rate of the molecular clock is a vital tool for genetic analysis, as it directly and significantly affects date estimations for speciation and population splits, and we address this in our work. We are able to combine information about the dates of admixture in par-

ticular human populations with information regarding the sets of admixed regions in each of these populations, in order to consider the accuracy of two frequently used mutation rates: 0.5×10^{-9}bp/yr, 1×10^{-9}bp/yr (Scally and Durbin [2012]). Lastly, we are able to contribute to discussion regarding whether admixture was uni or bi-directional, and whether additional uncharacterised interactions between ancestors of these species are likely to have occurred.

1.5 In this thesis

A diverse set of approaches have previously been applied to investigate admixture between modern and archaic human populations. A majority of these methods have either approximated features of trees, or inferred them indirectly. In the following chapters, we use genealogical trees directly to understand significant aspects of the admixture process between humans and Neandertals. We build these trees in regions of the genome which rarely recombine and are thus termed 'recombinationally cold'. This removes the modelling complications of recombination, and by removing noise, increases power to find evidence of admixture. Genealogical trees allow us to capture all possible information (all coalescence and mutation times) about the relationships between large-sample populations of humans and Neandertals. From these we are able to directly visualise admixture events and precisely identify the descendants of admixture.

In Chapter 2, we use a method which employs the use of direct pairwise sequence comparisons between humans and Neandertals, and which makes no strong assumptions about the nature of the demographic history of the two species, thus attempting to take a broad view of the data. We explore the distribution of coalescence times between human and Neandertal lineages, both through simulations and using real data. This then motivates our use of a newly developed method - *CEPHi* - in Chapter 3, to build genealogical trees between two input populations in order to jointly infer split times and size histories between these populations, allowing us to examine the variation in these factors across African and non-African populations. In Chapter 4, we examine the large amounts of information stored in these genealogical trees. We first explore possible demographic scenarios by performing coalescent simulations for two situations: the first

with admixture from Neandertals into a European population (GBR), the second lacking any admixture for an African population (YRI). We then provide a region-specific definition of admixture, based on the respective coalescence times of African and non-African populations with the Altai Neandertal. This definition is then employed to identify population-specific sets of introgressed regions, comparisons between which reveal the variation and interdependence across African and Eurasian populations with regard to the Altai Neandertal.

Finally, in Chapter 5, we date admixture across 14 human populations using a new method which employs information about the placement of two SNP types on genealogical trees in putatively admixed regions. We create distributions of the number of descendants of these SNPs, the shape of which we use to infer admixture dates using Maximum Likelihood Estimation (MLE). This gives us new insight into the evolutionary relationships between Neandertals and multiple human populations. We discuss our findings and their implications in Chapter 6.

CHAPTER 2

Using recombination coldspots to search for ancient admixture

2.1 Introduction

We begin this thesis by exploring a new starting point to the study of admixture between ancient and modern individuals. In this chapter, we develop and apply a set of exploratory methods that use regions of the human genome that are recombinationally very cold, to the extent that we can assume they have not recombined since the speciation of humans and Neandertals circa 6-700,000 years ago (Prüfer *et al.* [2014], Chapter 3). By effectively filtering out the confounding effects of recombination, we expect haplotypes that have been introgressed from Neandertals into modern humans to be intact and thus relatively large. We detail this starting point - which

is maintained throughout this thesis - in the Methods section below.

This approach directly complements those using linkage disequilibrium (LD) to study admixture, as we rely on the absence rather than the presence of recombination from which to make inferences. Methods using patterns of LD employ the following rationale; we present a simple case here. Take two populations A and B and allow for one interbreeding (admixture) event between two individuals from the respective populations at two unlinked loci, x and y. The offspring from this mating will contain genomic regions in perfect LD and identifiable as having been inherited from one or other parent. During gamete production in the offspring, meiotic recombination during prophase exchanges regions of DNA between chromosomal pairs, occurring repeatedly in each generation, breaking down LD. Methods utilising this information model this decay using an exponential distribution, with the rate parameter dependent on the number of generations since admixture (n), and the genetic distance (d) in Morgans between the loci (x and y) in question.

As presented in Loh *et al.* [2013], with time zero defined as zero generations following an instantaneous admixture event, we can write admixture LD at time zero (D_0) as:

$$D_0 = \alpha\beta\delta(x)\delta(y) \tag{2.1}$$

where α and β are the respective proportions of the admixing populations A and B, where:

$$\delta(x) = P_A(x) - P_B(x) \tag{2.2}$$

$P_A(x)$ is the frequency of a particular allele at locus x in population A, and similarly $P_B(x)$ and $P_A(y)$. The exponential modelling assumption implies that n generations post-admixture, we have:

$$D_n = e^{-nd}\alpha\beta\delta(x)\delta(y) \tag{2.3}$$

The aggregation of pairwise measurements of LD across SNPs allows for the production of LD curves plotting estimated LD against genetic distance. The rate of decay of this exponent is the age of admixture: where admixture is older, the fitted exponential is steeper, because LD quickly reduces with SNPs that are at an increasing distance away from one another.

This observation has been used in a number of studies to investigate admixture, both for events between modern human populations (Hellenthal *et al.* [2014]; Loh *et al.* [2013]; Price *et al.* [2009]) and for ancient events (Sankararaman *et al.* [2012]). In the latter study, the scale of LD between introgressed SNP pairs was calculated to be 1,597 generations in American Europeans. This led the authors to conclude that admixture explains the presence of regions similar to the Neandertal in these populations, rather than ancient African substructure (which would have required the presence of LD on the scale of 5,000 generations).

Alternatively, methods may make direct pairwise sequence comparisons between individuals, such as the D-statistic, which has been used to show Neandertals to be significantly closer to non-Africans than to Africans (Green *et al.* [2010]), and Denisovans to be significantly closer to modern Melanesians, Australian Aboriginals, and other Southeast Asian islanders than to European and East Asian populations (Meyer *et al.* [2012]), as mentioned in Chapter 1. We detail this rationale and use of this statistic in the Methods section of this chapter, and use it to demonstrate that these same patterns exist in the set of recombination coldspots we use for analysis throughout, providing support for the basis of our analyses.

Differing approaches come with various benefits and disadvantages. Those directly employing the use of ancient DNA have to combat the inherent problems with sequencing ancient genomes. These include the fact that bone samples contain very small percentages of DNA endogenous to the species of interest, short reads which are more difficult to map to a reference genome, as well as contamination and deamination. Additionally, we have very few high coverage ancient genomes to use for analysis, and must be cognisant of that fact that the relationship between the sampled ancient individuals and modern human populations will almost certainly vary relative to that between other unsampled individuals from other locations and modern humans.

By contrast, methods investigating LD patterns do not require the use of ancient genomes,

instead simply using high quality datasets from a large number of individuals from modern human populations, thereby sidestepping the inherent problems that come with using ancient DNA. However, these methods may require reliance upon one or many simplifying assumptions to be made about human demographic history, such as a lack of population structure, selection, or drift being present in the population in question, which may be difficult to justify or test.

While recognising the limitations of using ancient data, we use recombinationally cold regions of the human genome ('coldspots') as a starting point from which to explore admixture between various modern human populations and Neandertals, with the intention of reducing noise in both the Neandertal and human data, and increasing inferential power.

2.2 Methods

Given the set of genomic regions we define as recombinationally cold, we first perform coalescent simulations to investigate the effect of admixture between humans and Neandertals on these distributions between the two species. Secondly, we examine the distributions of estimated coalescence times between a set of humans and the draft Neandertal genome (Green *et al.* [2010]), using the number of mutations on the human genome as a proxy. Thirdly, we calculate values of the D-statistic (Green *et al.* [2010]) using the high coverage Altai Neandertal genome (Prüfer *et al.* [2014]) in combination with a set of 13 human genomes published by Sankararaman *et al.* [2014]. Analyses in this chapter are exploratory and consist of pairwise comparisons between human and Neandertal sequences; in all later chapters, we coanalyse hundreds of human sequences alongside those of the Altai Neandertal for deeper investigation.

2.2.1 Defining and identifying recombinationally cold regions

We used four genetic (recombination) maps to robustly identify recombinationally cold regions (coldspots) in the human genome. Each map corresponds to a different human population: European (CEU) (Frazer *et al.* [2007]), West African (YRI) (Frazer *et al.* [2007]), Icelandic

(deCODE) (Kong *et al.* [2002]), and African American (AA) (Hinch *et al.* [2011]). We define 'cold' to mean having a recombination rate of ≤0.2cM/Mb (centiMorgans/Megabase) across all maps (although it is often much lower), meaning in each region we expect to see ≤0.2 recombination events per lineage, per kb in 2,500,000 years. The speciation time of humans and Neandertals is thought to be ∼600,000 years (Prüfer *et al.* [2014] give a range of 550-765kya when using a mutation rate of 0.5×10^{-9}bp/yr), meaning we can reasonably assume there to have been close to zero recombination events in these regions, for most sampled sequences, since speciation. This is because on a single lineage, a recombination rate of 1cM/Mb is equal to 1 recombination every 100 generations, or 2500 years (assuming a generation time of 25 years), per Mb. This is equal to 1 recombination every 100 generations per 2.5my, per kb. At a recombination rate of 0.2cM/Mb, we expect 0.2 recombination events per 2.5my, per kb, per lineage, equal to 5 recombination events per 2.5my, per 25kb, per lineage, equal to ∼1 recombination event per 500ky, per 25kb, per lineage. We undercut this slightly and impose a minimum region size of 21kb. Disallowing any region shorter than this increases our confidence that the number of mutations in that region will be informative about the coalescence time between the two species: the shorter the regions used, the greater the effect of stochasticity on the number of mutations, making them less reliably indicative of coalescence times. Using this, then, a region of the human genome identified as a signal of admixture should represent an intact Neandertal haplotype from the past. With regard to mutation rate, we expect ∼ $\frac{1}{1000}$bp/my sequence differences between humans and chimpanzees, and a similar mutation rate between humans and Neandertals. According to Scally and Durbin [2012], the previously standard mutation rate of 1×10^{-9}bp/yr, may be halved in more recent evolution, and this rate variation is taken into account for our simulations.

2.2.2 Coalescent simulation of admixture scenarios varying four parameters

We first simulate pairs of human and Neandertal sequences under a set of demographic scenarios in order to investigate what effect the proportion of admixture between humans and Neandertals at some point in our evolutionary history may have had on (a) the distribution of coalescence

times between human and Neandertal haplotypes, and (b) the amount of sequence divergence between the two species, across recombinationally cold regions of the genome.

We vary four parameters: admixture fraction between 1-5%, time of admixture between 40-100kya, Neandertal effective population size between 5,000-10,000, and mutation rate between 0.5-1$\times 10^{-9}$bp/yr. Human effective population size is fixed at 10,000 individuals, generation time at 25 years, and we assume speciation times of 6.5 million years and 600,000 years for the human-chimpanzee and human-Neandertal splits respectively. The chosen parameter values are supported in the literature. For those parameters with relatively accepted values, we use these, such as a human-Neandertal split time of 600kya (Prüfer *et al.* [2014]; Scally and Durbin [2012]), a human-chimpanzee split time of 6.5mya (Noonan *et al.* [2006]), and effective population sizes for humans and Neandertals of 10,000 (Noonan *et al.* [2006]). We vary Neandertal effective population size between 5,000 and 10,000 individuals because the signal of admixture may look very different if it were substantially smaller than the standard 10,000. (Specifically, with a smaller Neandertal effective population size, faster coalescence between the Neandertal haplotypes would occur, and if admixture between humans and Neandertals had taken place, the Neandertal haplotypes would also coalesce faster with human haplotypes. This would lead to an earlier peak in the β distribution.) We do not attempt to model the Out of Africa bottleneck for the modern human population in this analysis. As with Neandertal effective population size, a reducing human effective population size would quicken coalescences between individuals, again leading to an earlier peak in the β distribution. We choose the range of admixture fractions based on those in Green *et al.* [2010], where 1-4% of European/East Asian genomes are reported to be from Neandertals; this also encompasses more recent estimates of Neandertal introgression in modern humans of 1-2% (Prüfer *et al.* [2014]). We vary the mutation rate between the values given to reflect the range in the literature. Until the publication of Scally and Durbin [2012], human mutation rate was as standard 1×10^{-9}bp/yr. The revised mutation rate - determined using trio data - is theorised to be due to the 'hominoid slowdown': a reduced mutation rate for hominoids (humans, our ancestors, and great apes) as compared with their ancestors. Lastly, we choose admixture times which bookend the likely range: 100kya being approximately when humans may have first exited Africa (Mallick *et al.* [2016]), and 40kya being a time by which

humans had likely encountered pockets of Neandertal habitation (Fu *et al.* [2014]; Mallick *et al.* [2016]; Pagani *et al.* [2016]).

The simulations provide, for each of a set of 9,245 coldspots (as exist across the genome), four parameters: a human-Neandertal coalescence time (C_{hn}), a human-chimpanzee coalescence time (C_{hc}), a number of pairwise sequence differences between the human and chimpanzee (N_i^1), and a number of pairwise sequence differences between the human and Neandertal, of the human-chimpanzee differences (N_i^2).

To do this we first initialise a number of fixed values: the human-Neandertal speciation time (HN), the human-chimp speciation time (HC), and a generation time (g) of 25 years (for humans, reasonable and frequently used generation times are between 25 and 30 years (Moorjani *et al.* [2016]). In this chapter we use 25, and for all future chapters we use 28 as it is a midpoint of this range - results for future chapters are, however, all easily rescalable for generation length). For each region, we find the number of positions at which sequence is available for each of the three species (human, Neandertal, chimpanzee) in our actual data. We assign this region as 'admixed' or 'nonadmixed' with probability dependent on the admixture fraction; an admixture fraction of 0.1 means a region has a 10% chance of being assigned as an admixed region. Depending on this assignment, we simulate the two coalescence times listed above. The human-Neandertal coalescence time (in years) is generated by sampling from an exponential distribution with rate:

$$\lambda = \frac{g}{2N_e} \tag{2.4}$$

where N_e here is Neandertal effective population size and g is generation time as before. In the case of admixture, this is added to the time of admixture, else it is added to the time of the human-Neandertal split, to give an estimated coalescence time. The human-chimpanzee coalescence time is simulated in the same way whether or not admixture is said to have occurred for this genomic region, by adding the time since the human-chimpanzee split to a random exponential variable with rate as given in equation 2.4. We then simulate the number of human mutations contributed by two branches of the tree: from the human-Neandertal most recent

common ancestor (MRCA) to the present, and between the human-Neandertal MRCA and the chimpanzee. Both variables are independent and Poisson distributed. We simulate the former by sampling randomly from a Poisson distribution with rate:

$$\lambda = l\theta C_{hn} \tag{2.5}$$

where l is the region length (with data for all three species), and θ the mutation rate. We then simulate the latter by again sampling randomly from a Poisson distribution, this time with rate:

$$\lambda = l\theta(2C_{hc} - C_{hn}), \tag{2.6}$$

For each of 140 admixture scenarios (the total number of combinations from varying the admixture fraction, time of admixture in years, mutation rate, and Neandertal effective population size as described above), we iterate over each of 9,245 coldspots. This allows us to explore the effect of admixture on the distribution of estimated coalescence times between human and Neandertal sequences in recombination coldspots. We plot coalescence times and human-Neandertal base mismatches against the number of coldspots to highlight the distributions of both parameters for a given demographic scenario.

2.2.3 Bimodality of human-Neandertal estimated coalescence times

We consider the fraction of derived mutations seen in a human sequence that also differ from the Neandertal sequence. To do this, we use the chimpanzee genome (as ancestral), a single human haplotype, and the homologous regions of the Neandertal to calculate the proportion of human-Neandertal sequence differences in each region, relative to the total number of sequence differences between the human and chimpanzee sequences in that region (a normalising constant). This proportion will vary across regions. This measure contains information on the coalescence time between a human and Neandertal for each region, and we can then investigate the distribution of this measure across all coldspots.

We filter out the confounding effects of recombination by selecting coldspots. By removing recombination as a variable, we aim to more reliably estimate the distribution of times at which humans and Neandertals coalesce. We also avoid relying heavily on the accurate ascertainment of recombination rates across the genome (other than that they are low), unlike methods more directly using LD patterns, for example.

For this analysis, we use the draft Neandertal genome (Green *et al.* [2010]), the reference human genome (Kent *et al.* [2002b], build 24/hg18), and an experimentally phased German genome MP1 (Suk *et al.* [2011]), alongside the chimpanzee genome (panTro4, Feb 2011), which is used as an outgroup (all except the Neandertal are from the UCSC Genome Browser (Kent *et al.* [2002b])). We label the German haplotypes hap1(MP1) and hap2(MP1). We label the ancestry of the human reference genome (either European or African) with two categories: conservative and aggressive, as detailed in Appendix A. This gives a total of six human sequences used for this analysis: hap1(MP1), hap2(MP1), ref-afr-CONS, ref-eur-CONS, ref-afr-AGG, ref-eur-AGG.

Considering a single region i, we define T_i^1 to be the total time to the most recent common ancestor ($\mathrm{T_{MRCA}}$) of, or minimum coalescence time between, humans and chimpanzees (HC taken to be 6.5mya), and T_i^2 to be the $\mathrm{T_{MRCA}}$ of a human and Neandertal lineage (HN - for example 600kya). We then define the coalescence time ratio of the human and Neandertal sequences as:

$$\alpha_i = \frac{T_i^2}{T_i^1} \tag{2.7}$$

We cannot know this coalescence time ratio, instead we estimate it using the number of observed mutations. So, as an estimator of α_i, we calculate:

$$\beta_i = \frac{2N_i^2}{N_i^1} \tag{2.8}$$

where N_i^2 is the number of derived mutations on the human lineage not shared with the Neandertal, and N_i^1 is the number of differences between the human and the chimpanzee (halved to

make β_i an estimate of α_i rather than of $\frac{\alpha_i}{2}$), both measurable quantities (Figure 2.1). Given that the process of mutations occurring is a Poisson process, we know that the number of mutations in any defined stretch of time is distributed binomially, with the number of trials being the number of mutations, and the probability of the mutation falling before the HN ancestor being α_i, meaning that given N_i^1:

$$N_i^2 \sim Binom(N_i^1, \frac{\alpha_i}{2}) \tag{2.9}$$

Thus β_i is an unbiased estimator of α_i, conditional on $N_i^1 > 0$. Here, we explore the distribution of β_i across loci. If admixture has occurred, we expect to find regions of the genome that are candidates for introgression from Neandertals to humans. This will result in the coalescence ratio of this region, α_i, and hence β_i, being reduced. More simply, this means that the putatively introgressed regions will look more similar between the human and Neandertal sequences than other regions. Therefore, we search for an excess of low β_i values.

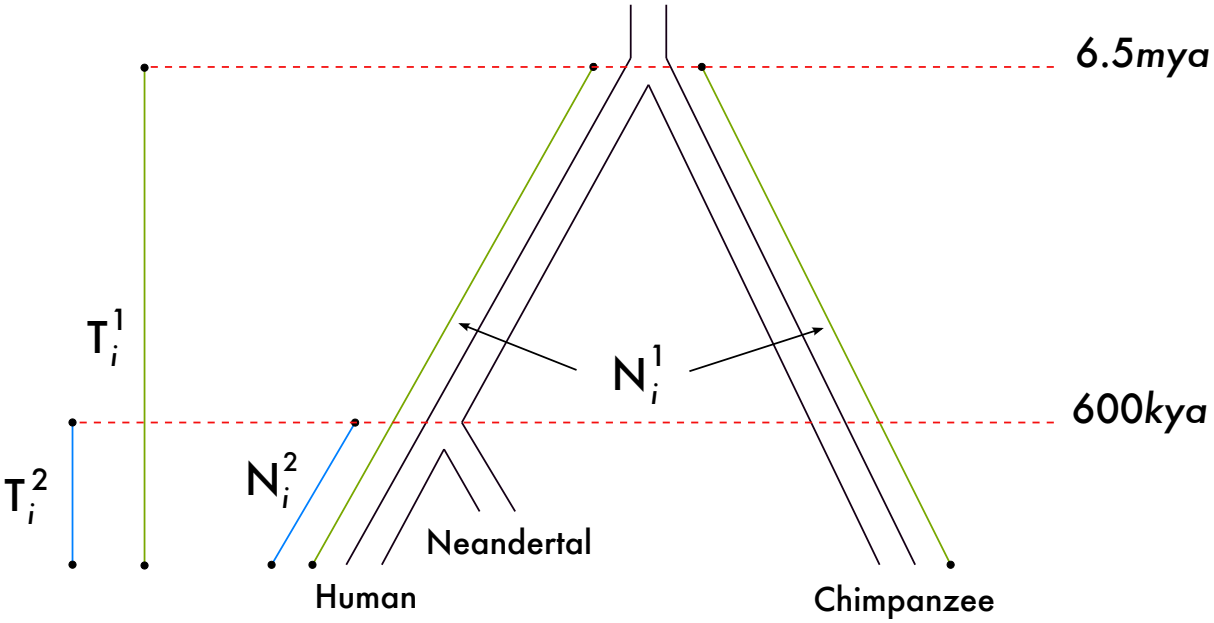


Figure 2.1: Looking for a bimodal distribution of mutations. For this region, i, T_i^1 is equal to the T_{MRCA} of the human and chimpanzee sequences, T_i^2 the T_{MRCA} of the human and Neandertal sequences. The coalescence ratio α_i of this region is given by $\frac{T_i^2}{T_i^1}$, estimated by the proportion of mutations $\beta_i = \frac{2N_i^2}{N_i^1}$ where N_i^2 is the number of human-Neandertal sequence differences, and N_i^1 is the number of human-chimpanzee sequence differences.

2.2.4 Calculating the *D*-statistic across coldspots

The *D*-statistic gives a good initial indication of the relative proximity of two haplotypes to a third. We use it to ascertain that our coldspots show intra and intercontinental patterns of sequence divergence from and similarity to the Neandertal genome reflective of those in Green *et al.* [2010] and Prüfer *et al.* [2014]: Asian and European genomes have been shown to exhibit significantly greater similarity to the Neandertal genome as compared with African genomes, but show little difference in similarity to the Neandertal between themselves. With these observations also true for our coldspots, we have support for using these regions to search for admixture. This statistic assumes both infinite sites (that only a single mutation has happened on the genealogy of the four haplotypes under consideration), and that the chimpanzee carries the ancestral allele. However, it remains robust to these modelling departures as it is essentially a test of symmetry between a set of three haplotypes. Incidentally, it also has the potential to be extended to a larger sample, for example by averaging counts across pairs.

We calculated values of the *D*-statistic in our coldspots (Green *et al.* [2010]; Prüfer *et al.* [2014]). This statistic uses homologous sequences from four individuals: in this case two modern humans (H_1, H_2), an ancient hominid: the Neandertal (N), and an outgroup: the chimpanzee (C). It compares the proximity of H_1 and H_2 to the ancient hominid, using the outgroup as ancestral at each position. Where A denotes the ancestral allele, and B the derived allele, D is defined as:

$$D(H_1, H_2, N, C) = \frac{\sum_{i=1}^{n} [ABBA_i - BABA_i]}{\sum_{i=1}^{n} [ABBA_i + BABA_i]} \tag{2.10}$$

where $ABBA_i$ is a count of the number of SNPs (of which n is the total) in a genomic region i where H_2 matches N, and where N carries the derived allele. $BABA_i$ is the equivalent for H_1. The difference between the respective sums of these counts is normalized by the total number of SNPs in that region.

The null hypothesis of the *D*-statistic is that the two modern human populations form a clade,

with the ancient hominid being more distantly related. Phylogenetically this is the case, in that humans and Neandertals split from one another before human populations differentiated. This means we expect at many sites to see a genealogy matching the phylogeny, where a mutation has occurred since the split with Neandertals but before the split between the populations H_1 and H_2 are sampled from: $BBAA$. Once we restrict to sites where the Neandertal carries the derived allele, under the null hypothesis of no admixture, we expect two genealogies at equal rates: $ABBA$ and $BABA$. This is because without admixture, H_1 and H_2 go back to divergence from Neandertal without coalescence, and there is an equal probability that a mutation has occurred on either of the human branches before their coalescence. However, where admixture has occurred between Neandertals and the population H_1 is sampled from, for example, we expect to see the $BABA$ genealogy significantly more frequently than the $ABBA$ genealogy.

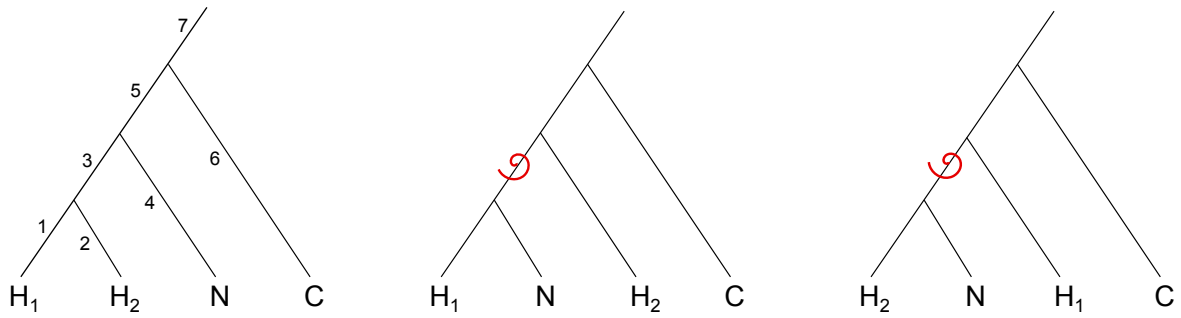


Figure 2.2: Understanding the D-statistic. From left to right we have three gene genealogies. The left shows the phylogeny we expect, with human 1 (H1) coalescing first with human 2 (H2), followed by the Neandertal (N), and lastly the chimpanzee (C). Mutations can happen on any one of 7 branches, labelled 1-7. A mutation on branch 1 will only be present in H1, and on branch 2, 4, or 6, will be present on only H2, N, and C respectively. A mutation on branch 3 will be present in H1 and H2, on branch 5 in H1, H2, and Nean, and on branch 7, in all four. The chimpanzee allele is considered the ancestral allele. The D-statistic considers only those positions where the Neandertal differs from the chimpanzee, therefore carrying the derived allele, and where the human alleles are biallelic, with one human matching the Neandertal. If the human alleles are different and neither match the Neandertal, i.e. the three individuals are triallelic, the position is discarded from analysis. This leaves us concerned only with positions at which either the middle or right hand gene genealogies apply: where chimpanzee is ancestral (A), Neandertal is derived (B), and either H1 and the Neandertal match (centre diagram), or H2 and the Neandertal match (rightmost diagram). A single mutation occurring on branch 3 in the latter two diagrams is represented by the red swirl, and creates these allele patterns with which we are concerned. If H1 matches the Neandertal, we map these to the leftmost tree to give the pattern BABA (using the order H1, H2, N, C), if H2 matches the Neandertal, we get the pattern ABBA.

The human dataset consists of 13 genomes from individuals from 5 continents (as released with

the publication of Prüfer *et al.* [2014] and detailed in Appendix A): 2 Europeans (1 French, 1 Sardinian), 2 Asians (1 Dai, 1 Han Chinese), 4 Africans (1 Mandenka, 1 Mbuti, 1 San, 1 Yoruban), 2 Americans (1 Karitiana, 1 Mixe), and 3 Australasians (1 Papuan, 2 of unknown location on the Australasian continent). These are used in conjunction with the high coverage Neandertal genome from the Denisova cave in Siberia (Prüfer *et al.* [2014]. We used the most recent build of the chimpanzee reference genome (panTro4, Feb 2011) from the UCSC Genome Browser (Kent *et al.* [2002b]) as the outgroup for all analyses. We calculated the value of this statistic for all pairs of individuals, totalling 156 comparisons (two per population pairing as we calculate values for each Neandertal haplotype).

We aligned all four sequences across our set of 8881 recombination coldspots at all human SNP positions within those regions. *Vcftools* (Danecek *et al.* [2011]) was used to select only those SNPs within coldspots. SNPs not passing quality controls (with a Phred score of <40) were removed. We then removed from all analyses positions where we lacked chimpanzee base information, or that were triallelic across sequences. For each individual analysis, we removed any positions lacking Neandertal or human information. We filtered for those positions where the Neandertal base is derived (differing from that of the chimpanzee which we assume to be ancestral), and selected those positions where H_1 and H_2 are biallelic. We calculated D across all regions cumulatively, scoring a homozygous site as 1 when it matched the Neandertal haplotype, and a heterozygous site 0.5. The standard error (SE) for each D-statistic was estimated using a block-jackknife, removing a set of 10 coldspots at a time and recomputing the D-statistic), and a Z-score obtained by dividing D by this estimated SE. The Z-score is tested against an $N(0, 1)$ null distribution; significant results are defined as those $\geq$2 SE from the mean.

2.3 Results

We first report some summary statistics surrounding recombinationally cold regions of the genome that we use throughout analysis. We then describe the effect, within coalescent simulations, of changing parameter sets and thus varied demographic scenarios on the visible signature

of admixture between humans and Neandertals, through coalescence time and derived mutation distributions. Thirdly, we examine the distribution of human-Neandertal sequence differences in real data using derived mutations as a proxy for coalescence times. Lastly we calculate the D-statistic, using the high coverage Neandertal genome and a set of 13 continentally varied human genomes (both from Prüfer *et al.* [2014]) to show that the pattern of proximity between geographically differentiated human haplotypes and Neandertals seen in Green *et al.* [2010] is reflected in human coldspots.

2.3.1 Coldspots summary information

Our coldspots have a maximum recombination rate of 0.2cM/Mb and cover 373,614,245 bases: >10% of the human genome. We summarise the data available for analyses for both the bimodal and D-statistic analyses in Figures 2.3 and 2.4. For the bimodal analysis (Figure 2.3, left hand plot), we have 136,303,737 (36.5%) positions with all species data: the reference genome, both MP1 haplotypes, the Neandertal, and the chimpanzee. All human sequences have close to full coverage, with the reference human genome covered at 367,385,424 positions (98.3%), hap1(MP1) at 367,382,877 (98.3%), and hap2(MP1) at 367,383,007 (98.3%). The equivalent Neandertal and chimpanzee numbers are lower, with 189,421,711 (50.7%) and 261,624,467 (70%) positions covered respectively. Mean coldspot size is ∼40,000bp, the minimum being 21,001bp (Figure 2.4). Average sequence coverage of the Neandertal genome for our coldspots was 1.7×. For the D-statistic analysis (Figure 2.3, right hand plot), we have 4,340,950 SNPs in total, covered completely for both the Altai Neandertal and all human haplotypes, and 92.89% by the chimpanzee genome (4,032,223bp). We show the coldspots to be representative of the genome in our D-statistic calculations in Section 2.3.4.

2.3.2 Investigating the effect of demographic scenarios on admixture signal

We simulated a variety of demographic scenarios, and generate distributions for the set of coalescence times (α), and the counts of derived mutations present in Neandertals that aren't

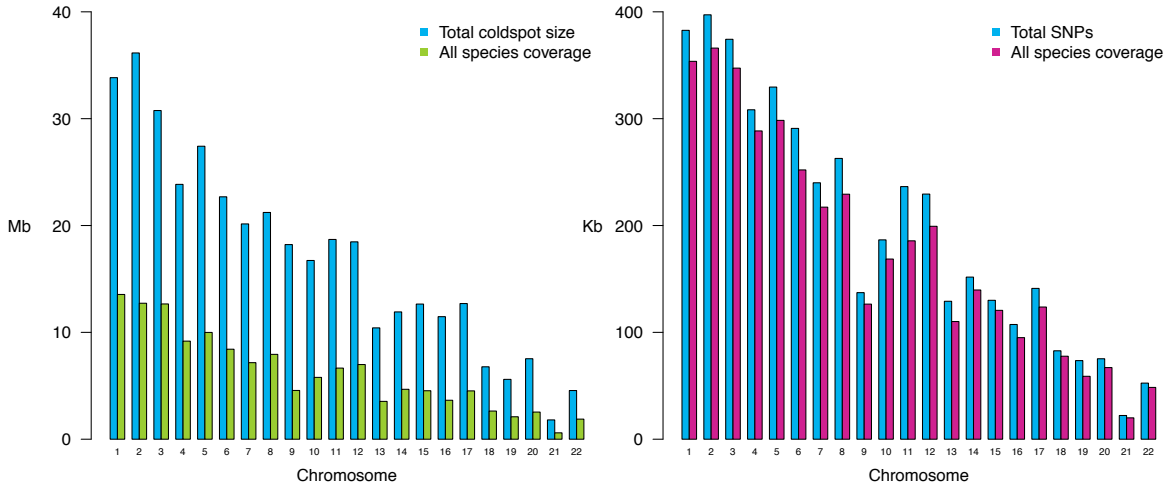


Figure 2.3: Coldspot size and data coverage for all analyses in this chapter. The left hand plot gives the combined size of the cold regions per chromosome, and the amount of that total for which we have full data, when using the draft Neandertal genome, the chimpanzee, and the human reference and German genomes for analysis. This therefore applies for the bimodal analysis. The right hand plot applies to the D-statistic analysis, showing the total number of SNPs used per chromosome, and the number of that total for which we have chimpanzee, Altai Neandertal, and human data (excluding Karitiana).

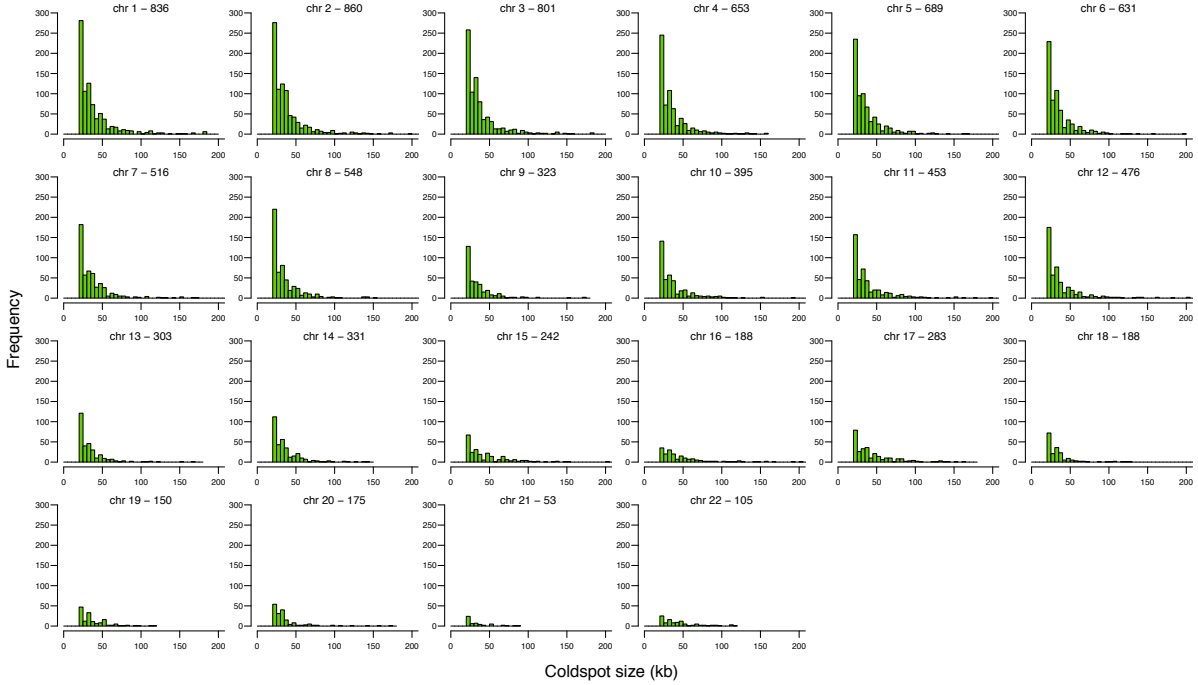


Figure 2.4: Showing the number and distribution of coldspot sizes across chromosomes. The total number of coldspots for each chromosome is given in each subplot title.

shared with humans (β), as described in the methods section. Figure 2.5 shows the effects of changing admixture fraction on the resulting α and β distributions, and Figure 2.6 shows the

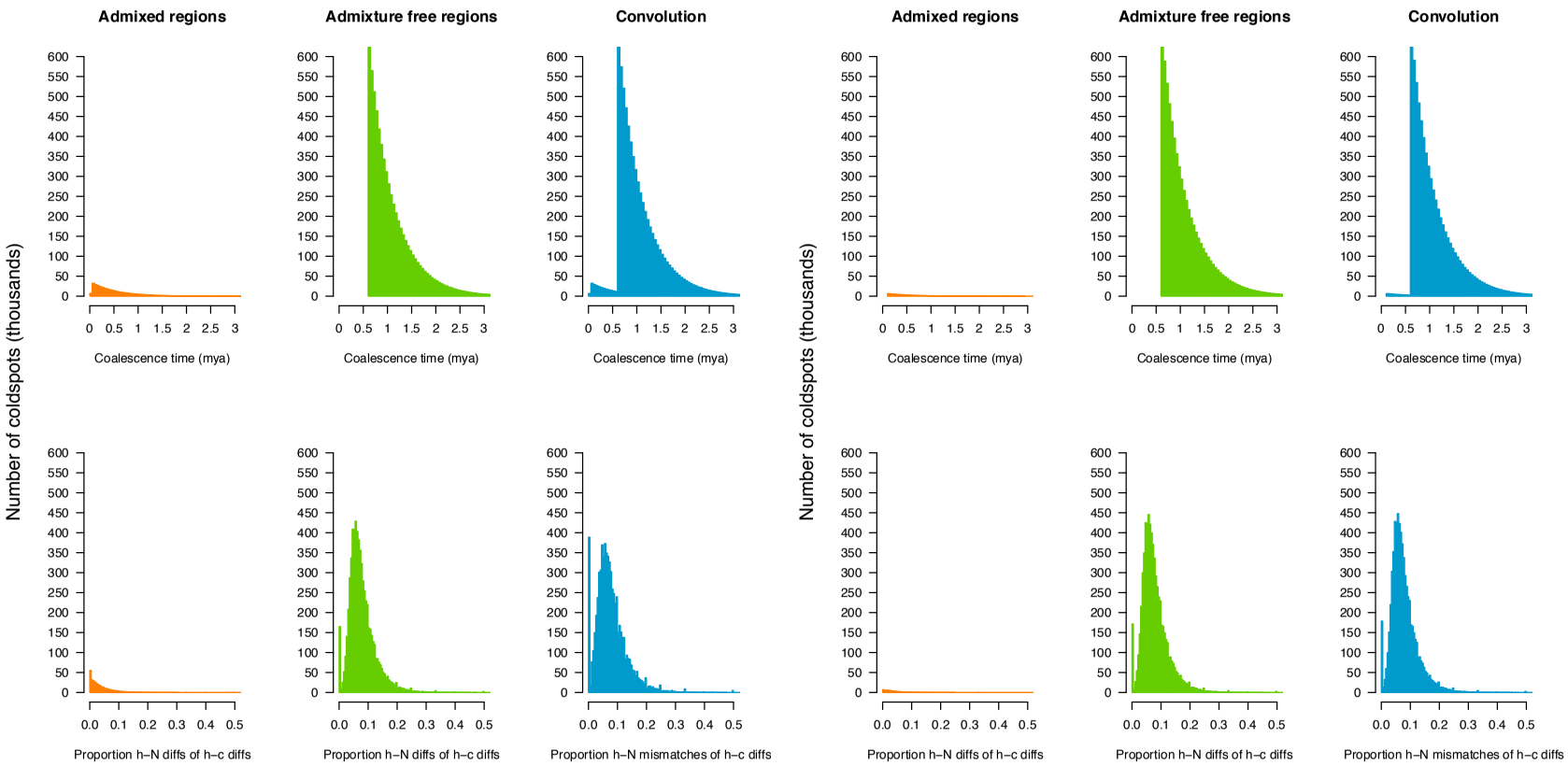
effect of changing the mutation rate. The first scenarios use parameters $f = 0.05$, $d = 40$kya, $\theta = 1 \times 10^{-9}$bp/yr, and $N_{e_n} = 10,000$, and $f = 0.01$, $d = 100$kya, $\theta = 1 \times 10^{-9}$bp/yr, and $N_{e_n} = 10,000$. We use a haplotype from the German genome (hap1(MP1)), as we have data across all coldspots for this genome.

The first row of plots in Figure 2.5 shows that a reduction in the admixture fraction from 5% to 1% shrinks the peak in the top left (orange) plot in each of subplots (a) and (b); when the admixture fraction is at 1% (close to published estimates, such as Prüfer *et al.* [2014]), the peak is barely visible. The β distributions on the bottom row relatively closely match their corresponding α distributions in the row above. This is also true for each of the middle plots on the bottom row - showing the distribution of β for the admixture-free regions; we see a peak at the point of speciation, and a sharp falloff after the peak is reached. However, this distribution of β sits closer to zero than seen in the corresponding α distributions, and when the two distributions (admixed and non-admixed) are convolved (right hand plots), the signal of admixture is swallowed by the admixture-free regions.

Thus, we can make two observations. The first is that in realistic scenarios, we expect it to be difficult to use a single human sequence to find admixed regions. However, the estimated coalescence time distribution does show a marked difference, and it is possible we may gain power by using many human sequences to detect admixture. We take forward this notion in all future chapters. The second is that comparing divergence between two sequences from two populations that are admixed and non-admixed respectively may aid power, something that we test in this chapter using the D-statistic.

Notably, we can see a spike at zero in the simulated data for the β distribution, also present in the real. It is driven mainly by the admixture-free regions, and therefore does not represent a signal of admixture.

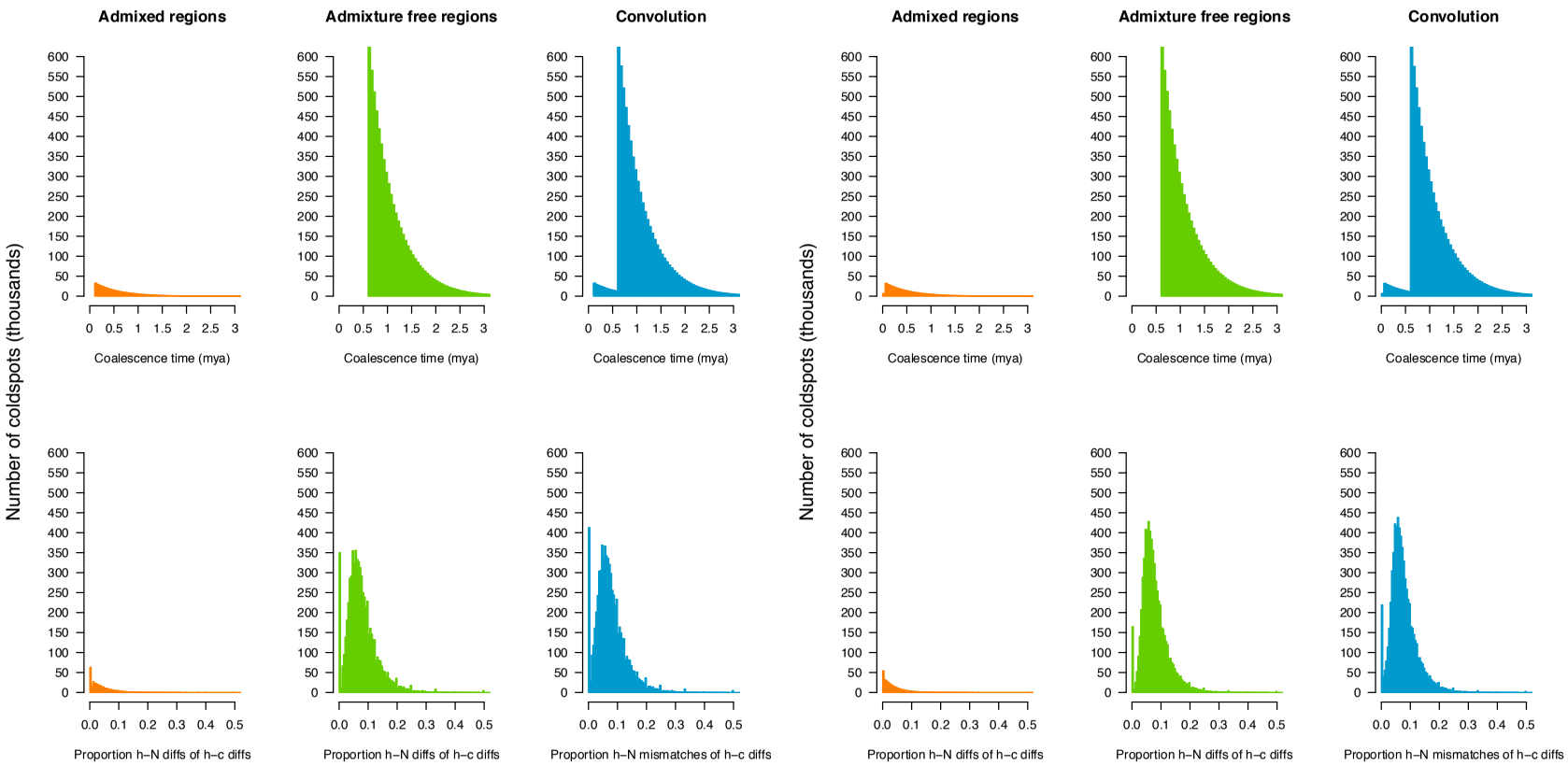
Figure 2.6 shows the effect of altering the mutation rate on the distributions of β. In the first set of plots (a), we see a pattern similar to that seen in Figure 2.5, where the admixed regions increase both the number of regions which either match exactly, or are very similar between humans and Neandertals, as shown in the spike at zero and a heavier weight just past zero.



(a) 5% admixture, 40kya, $\theta = 1 \times 10^{-9}$bp/yr

(b) 1% admixture, 100kya, $\theta = 1 \times 10^{-9}$bp/yr

Figure 2.5: Comparing simulated distributions of α (coalescence times for each region, top row) and β (human-Neandertal differences of derived human alleles, bottom row) for two sets of parameter values, each the result of 1000 sets of 9,245 genomic regions (a) where admixture might be most probable, with parameter values $f = 0.05$, $d = 40$kya, $\theta = 1 \times 10^{-9}$bp/yr, and $N_{e_n} = 10,000$, and (b) to look for an effect of varying the amount of admixture and the time at which it occurs, using the parameter set $f = 0.01$, $d = 100$kya, $\theta = 1 \times 10^{-9}$bp/yr, and $N_{e_n} = 10,000$. Each subplot is separated into three columns: the first gives those regions classified as admixed as per the simulations, the second the admixture-free regions, and the third the convolution of the two distributions. Only regions with greater than 20 human-chimpanzee differences are included.



(a) 5% admixture, 40kya, $\theta = 1 \times 10^{-9}$bp/yr

(b) 5% admixture, 100kya, $\theta = 0.5 \times 10^{-9}$bp/yr

Figure 2.6: Comparing simulated distributions of α (coalescence times for each region, top row) and β (human-Neandertal differences of derived human alleles, bottom row) for two sets of parameter values, each the result of 1000 sets of 9,245 genomic regions (a) with parameter set $f = 0.05$, $d = 40$kya, $\theta = 1 \times 10^{-9}$bp/yr, and $N_{e_n} = 10,000$, and (b) to look for an effect of varying the time of admixture and the mutation rate, using the parameter set $f = 0.05$, $d = 100$kya, $\theta = 0.5 \times 10^{-9}$bp/yr, and $N_{e_n} = 10,000$. Each subplot is separated into three columns: the first gives those regions classified as admixed as per the simulations, the second the admixture-free regions, and the third the convolution of the two distributions. Only regions with greater than 20 human-chimpanzee differences are included.

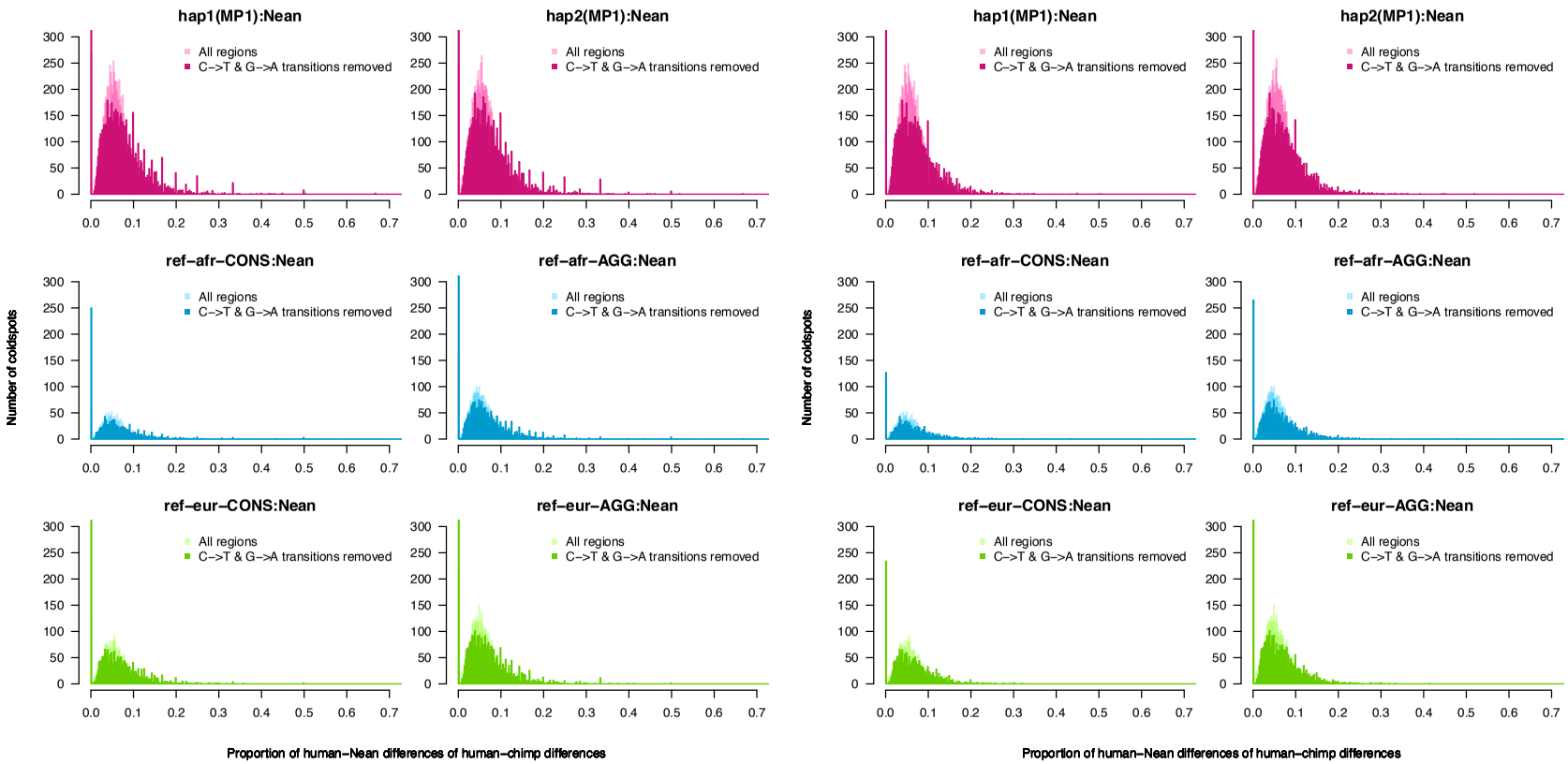
In the second set of plots (b), we see the effect of halving the mutation rate, as the spike at zero and lefthand weight are both increased. A lower mutation rate increases the number of regions which are very similar between the human and Neandertal; admixture is harder to see in settings of reduced diversity. Variation of θ across regions, therefore, could potentially be a confounding factor.

2.3.3 Human-Neandertal sharing of derived alleles: looking for a bimodal distribution of β

In Figure 2.7 we present distributions of allele differences between humans and the draft Neandertal constituted of data from all coldspots, excluding only those where there is insufficient data for the three genomes with which we are concerned. As stated in our methods section, this is equal to β, the estimator for α - the set of coalescence times for the coldspots, which are not calculable, because we cannot know actual rates of mutation, recombination, and selection, as well as other factors affecting the number of differences between sequences. The lefthand set of 6 plots shows the distribution of β for the six human genomes, and the righthand plot is equivalent but using only regions with $\geq$20 human-chimpanzee differences. For each set of plots we superimpose the equivalent distributions removing all C$\longrightarrow$T transition SNPs and their exact equivalent, G$\longrightarrow$A transitions, as these are the most common effects of deamination. We provide this split to eradicate any significant possibility of deamination affecting our results.

Figure 2.7, shows us that of the derived human alleles in human coldspots, the proportion of human and Neandertal differences is most commonly about 5-6%. This is as expected: human-Neandertal speciation time (600kya) is just below 5% of the evolutionary time between humans and chimpanzees, who speciated ~6.5mya (Prüfer *et al.* [2014]), and thus divergence time between humans and Neandertals is expected to be moderately more than 600kya.

It is evident that there are a substantial number of coldspots at zero across analyses, meaning that in these coldspots, at those positions where the human and chimpanzee differ, the human sequence matches the Neandertal sequence exactly. To eradicate the possibility of this resulting simply from there being very few differences between the human and chimpanzee in these regions



(a) All regions

(b) Regions with ≥20 human-chimpanzee differences

Figure 2.7: Distribution of the fraction of positions in each coldspot differing between humans and Neandertals, of all human-chimpanzee differences. The first row gives distributions for the MP1 haplotypes 1 and 2, the second for the African regions of the human reference genome, and the third for the European parts of the reference genome. For the second and third rows, the first column gives those coldspots defined as African or European using a conservative estimate, the second column using an aggressive estimate. All regions are included in (a), and only regions with ≥20 human-chimpanzee differences are included in (b). Both subplots (a and b) contain two distributions per plot, the first with all regions fitting the definition, and the overlay removing all transitions.

(thus making it more likely for the human and Neandertal sequences to subsequently match at all positions), we conditioned on requiring at least 20 human-chimpanzee differences in a region and replotted our results. The spike at zero reduces a small amount after implementing this condition in this way, but it does not disappear (Figure 2.7). Thus, we see a suggestion of a bimodal distribution, which we expect to see with admixture; those regions matching the Neandertal sequence may be a result of introgression from Neandertals. Both the spike at zero and the small shoulder visible in some comparisons may indicate admixture. However, these peaks are visible in both European and African sequences. This may mean we are seeing admixture in African sequences, it may also be explicable simply as a signal seen in any comparison between a human and Neandertal, possibly representing chance similarity in certain regions.

2.3.4 *D*-statistic: Examining recombinationally cold regions of the human genome using a set of human genomes from five continents

The asymmetry between African and non-African individuals with regard to their genetic proximity to Neandertals as shown in Green *et al.* [2010] and Prüfer *et al.* [2014] is clearly visible in our recombinationally cold regions of the human genome, as shown in Figure 2.8. Each analysis has two results, one for each of the Neandertal haplotypes. As we would expect - given the low levels of heterozygosity of the Altai Neandertal - results differ minimally within each pair.

When comparing intracontinental populations we see almost no significant results, as we would expect, except some marginally significant differences where the San are involved. Notably, a lack of significant differences between African individuals may be indicative of a lack of visible substructure within the African continent for these populations, at least across these coldspots.

Importantly, all comparisons between non-African and African individuals show the former to be significantly more similar to the Neandertal. The Australasians being significantly closer to the Neandertal than are the Africans can be thought of as consistent with the results from Meyer *et al.* [2012], which suggest that Melanesian populations are more closely related to the Denisovan than are other human populations - this may also be the case with Neandertals. We note that there is a significant data reduction for the Karitianan individual in these coldspots,

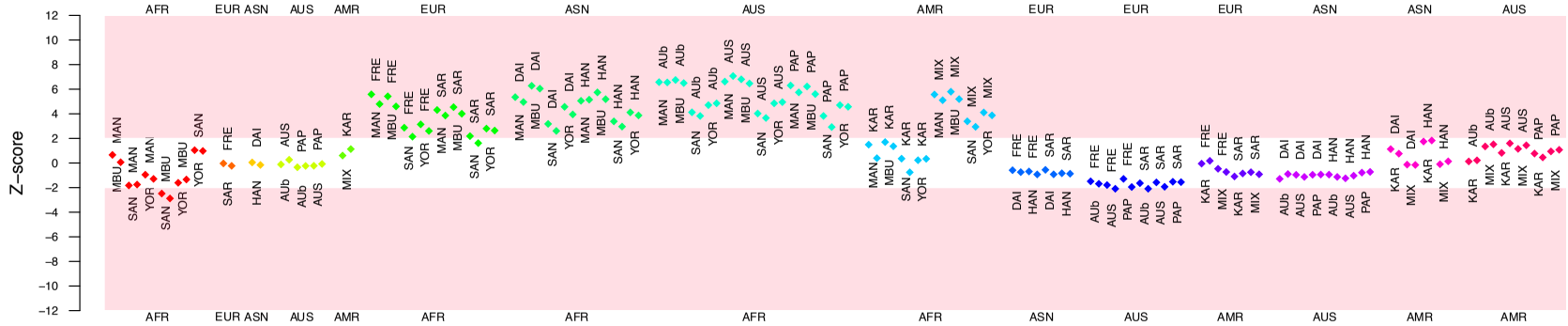


Figure 2.8: *D*-statistics in recombination coldspots. All pairs of populations are included, continents are shown across the top and bottom for each set. From right to left: all intracontinental comparisons, all intercontinental comparisons including AFR, and all intercontinental comparisons excluding AFR. The white region between (-2,2) shows values of Z that do not deviate further than 2 standard errors from 0, those in the pink regions show a significant difference with regard to proximity to the Neandertal. A full table detailing for each combination of haplotypes: the ABBA and BABA counts, D, the standard error (SE), and the Z-score, is available in Appendix A.

as can be seen in the table detailing this statistic in Appendix A, making results less conclusive, although expected patterns remain in most cases. Lastly, between non-African continents we see very few significant differences. These results are thoroughly consistent with previous publications using the D-statistic, supporting our use of human coldspots in search of admixture between humans and Neandertals.

2.4 Discussion

We have here made an initial investigation into whether it is possible to study signals of admixture in modern human populations using cold regions of the human genome - alongside their homologous counterparts in other species - as a starting point. This allows us to assume that these coldspots have not been affected by potentially confounding factors such as recombination, reducing noise in our data and increasing inferential power.

Our use of cold regions can be separated from methods looking to detect ancient admixture which rely on genetic maps detailing linkage disequilibrium. These maps are typically at 3kb resolution, which although reasonably detailed, may still suffer from biases due to error. Additionally, it is important to consider the evolution of hot and coldspots since our MRCA with chimpanzees, and also with Neandertals in this context. Recombination hotspots have had significant attention over the past few years and it is known that they evolve relatively quickly (Lesecque *et al.* [2014]; Myers *et al.* [2010]). Recombination coldspots, by contrast, have had very little attention, and given that these seem to surround genic regions, it may be true that cold regions are more resolutely cold than recombining regions are fixed at a specific rate. It may be important to ask whether, as the position of hotspots evolves, how often they interfere with cold regions, and more generally the strength of conservation for coldspots.

With regard to the value of using recombination coldspots for investigating admixture between modern and ancient human populations, our coalescent simulations showed the effects of varying admixture fraction and mutation rate on coalescence time (α) and derived allele frequency (β) distributions between individual human and Neandertal sequences.

When creating β distributions from real data, we saw distributions very similar to the simulated data. Although we could not reasonably conclude from this that admixture certainly exists in the coldspots we used for analysis, it could not be ruled out. However, single human sequences may be insufficient to clearly show the presence of admixture in the context of both methods. Further work using these methods could include using a large set of phased human genomes to search for a bimodal distribution of estimated coalescence times between modern and ancient sequences.

Lastly, and importantly, use of the D-statistic has shown us that a simple indicator of potential admixture - a human sequence being more similar to a Neandertal sequence than is a second human sequence - is present in human coldspots. This supports our use of recombinationally cold regions of the human genome when searching for and detailing aspects of admixture. This motivates our next chapter, where we employ a new method developed in the group, which combines the phased genome of the Altai Neandertal alongside a set of 14 human genomes from the 1000 Genomes Project, to jointly and directly infer the coalescence times between the two species by building genealogical trees across recombinationally cold regions of the human genome.

CHAPTER 3

Divergence times, coalescence times, and population histories

3.1 Introduction

In the previous chapter, we first explored the distribution of sequence differences between the draft Neandertal genome (Green *et al.* [2010]) and a number of high quality human genomes (Kent *et al.* [2002b]; Suk *et al.* [2011]) in recombinationally cold regions of the human genome. The number of mutations between human and Neandertal sequences in a region was used as a proxy for a coalescence time, defined as the time to the most recent common ancestor (T_{MRCA}) of a pair of samples. This allowed an initial exploration of the distribution of estimated coalescence times between geographically distinct humans, and Neandertals. In a subsequent analysis, using

the high coverage Altai Neandertal genome (Prüfer *et al.* [2014]) and a set of 13 continentally varied human genomes released by Sankararaman *et al.* [2014], we noted the power gained by comparing multiple human sequences from African and non-African groups with the Neandertal with regard to finding evidence of admixture. This motivated our employment in this chapter of a novel method entitled *CEPHi* - Coalescent Estimation of Population History - to infer these region-specific coalescence times directly across a set of 14 human populations with regard to the high coverage Altai Neandertal genome. This archaic genome has been sequenced to $\sim 52\times$ coverage from material taken from the proximal phalanx of an adult's fourth or fifth toe found in Denisova cave in the Altai mountains, Siberia, in 2010 (Mednikova [2011], Prüfer *et al.* [2014], see Appendix B for details), providing data of comparable quality to modern human genomes.

Created by Marie Forest, Simon Myers, and Jonathan Marchini, *CEPHi* was previously used to examine population split times and histories between modern human populations. For this work, it has been adapted and extended (by Marie, on request from Simon and me) to take a Neandertal dataset to infer these same estimates between the Altai Neandertal and various modern human populations, as well as to produce genealogical trees across regions of the genome, allowing for a nonzero Neandertal fossil age. These adaptations are explained in the Methods section below. *CEPHi* enables us to investigate a number of aspects of the evolutionary history of *Homo sapiens* and *Homo neanderthalensis*, using recombinationally cold regions of the human genome and their homologous Neandertal counterparts.

Firstly, estimating a single population split time (T_D) between a Neandertal and various human populations provides an initial indication of whether admixture is visible between any two groups. Although a simple split model is not a 'correct' model of admixture if it has occurred, it is essentially a parametric method of assessing the relative similarity of sequences from two populations to the Neandertal. The split time between two populations can be defined as the amount of time that has elapsed since these populations arose from a shared ancestral group, and can be thought of as a minimum coalescence time. T_D is uniquely defined for a pair of populations in the case where separation into two groups happens instantly and with no subsequent migration. In the current context, the group from which Neandertals and humans emerged may

have been constituted of *Homo heidelbergensis*, a potential candidate for the ancestor of both species (Buck and Stringer [2014]). We see evidence of the presence of *Homo heidelbergensis* in Europe by ∼400kya, most prominently from the Sima de los Huesos (Pit of Bones) in the Atapuerca mountains in northern Spain (Meyer *et al.* [2014]). Neandertals are found in their most modern form by ∼250-200kya (Marra *et al.* [2015]), and the first anatomically modern humans (AMH) are dated to ∼200kya in Omo Kibish, Ethiopia (McDougall *et al.* [2005]). The present archaeological record does not provide sufficient information to detail the *Homo* evolutionary tree with full confidence, and the sequencing of archaic hominid genomes has only recently seen big advances in quality and coverage. Over the coming years it is likely that further sequencing of archaic genomes will reveal relationships between different species, potentially including more detailed information about migrations and population splits. However, the nature of the separation of the ancestral population of humans and Neandertals is unknown, including both its location and timeframe. For the purposes of this initial analysis, we assume that the split between the ancestors of humans and Neandertals was instantaneous and complete, with no migration occurring between the separated populations. This is likely a reasonable model when considering the split between African human and Neandertal populations.

The split time between two populations can be estimated in various ways. The two most recently used in this context are given in Prüfer *et al.* [2014]. The first, employing a method previously seen in Green *et al.* [2010] and Meyer *et al.* [2012], finds polymorphic sites in a Yoruban (YRI) human from Nigeria (used as an example of a non-admixed human population), and takes the proportion of these that are derived in the archaic hominid, using the chimpanzee alleles as ancestral. The greater this proportion, the more recent the split time between the Yoruban (YRI) and the archaic hominin, because it can be assumed that shared derived alleles occurred before the YRI-archaic split, lengthening the branch before the split, so pushing the split time forward. This essentially assumes the presence of infinite sites (that the probability of recurrent mutation is negligible and therefore ignored) which could affect results, so the analysis is restricted to transversions (purine$\longleftrightarrow$pyrimidine mutations) where the mutation rate is lower, so reducing the probability of recurrent mutations. In Green *et al.* [2010], this method gave an estimate for the human-Neandertal split of 270-440kya, and for Meyer *et al.* [2012], the same

method produced an estimate of 170-440kya, both using a mutation rate of 1×10^{-9}bp/yr, and human-chimpanzee speciation times of 5.6-8.3mya and 6.5mya respectively. Using a halved mutation rate from Scally and Durbin [2012] of 0.5×10^{-9}bp/yr, Meyer *et al.* [2012] and most recently, Prüfer *et al.* [2014] gave human-Neandertal split times of 410-700kya and 550-765kya respectively.

The second approach for estimating the split time between two populations, found in Prüfer *et al.* [2014], uses a modified version of the Pairwise Sequentially Markovian Coalescent (PSMC) (Li and Durbin [2011]) to infer the distribution of coalescence times between a sub-Saharan African individual (San, Mandenka) and an archaic individual (Neandertal, Denisovan), each sample represented by a single haplotype. It is modified by incorporating a sudden split time before which no coalescence is permitted. Sub-Saharan African haplotypes were taken from experimentally phased individuals; these are not available for archaic individuals. Instead, regions of the genome of an archaic individual that are closely related and therefore coalesce with one another recently are used to approximate a haplotype. This method resulted in an estimate of 553-589kya for the Neandertal-African split, falling in line with the estimates using the halved mutation rate above.

Notably, both methods just discussed calculate split time estimates using single individuals to represent each population. Another commonly employed method of investigating the amount of differentiation between individuals is by examining simple sequence divergence. Where population split times give a minimum coalescence time between two populations, sequence divergence gives an average coalescence time between two populations or individuals, thus providing earlier dates. At its simplest, sequence divergence is reported as an average proportion of the branch length between the two species in question (for example humans and Neandertals) and an ancestral species (the chimpanzee). Specific applications vary with regard to which regions of the genome are used. The first analysis of the Denisovan genome employed this method (Reich *et al.* [2010]), using the regions of the human reference genome that are of African ancestry, giving a mean coalescence time estimate of 12.38% of the human-chimpanzee split (6.5mya), equal to 804kya, in a manner similar to our analysis in Chapter 2. In Green *et al.* [2010], estimates

were reported at 825kya: 12.7% of the human-Neandertal-chimpanzee sequence divergence. The later higher coverage Denisova genome (Meyer *et al.* [2012]) gave very similar human-Denisovan sequence divergence estimates of ∼800kya, equivalent to 12.35% of the human-chimpanzee divergence (again assuming a human-chimpanzee speciation time of 6.5mya and a mutation rate of 1×10^{-9}bp/yr). A number of estimates based on mitochondrial sequences date the split to anywhere between 300-800kya (Endicott *et al.* [2010]; Green *et al.* [2008]; Krause *et al.* [2010]; Soares *et al.* [2009]), but should be viewed with caution as they represent only one locus, and so are susceptible to stochastic tree variability.

CEPHi, by contrast, allows for population split times to be calculated using a large number of individuals (samples) for the pair of populations in question. For Neandertals, the population is necessarily one individual, made up of two haplotypes. The 1000 Genomes human populations however, vary between 14 and 100 individuals, with only one population (IBS) falling below 55 samples. Given that the variation within a population is therefore taken into account, we should expect to obtain more comprehensive and hopefully reliable estimates of population split times. We estimate these split times across all human populations included in the 1000 Genomes dataset, detailed in Appendix B.

Secondly, we examine the changes in size that pairs of human and Neandertal populations have experienced in their ancestral population (before T_D, looking forwards in time) and separately (after T_D). We note the importance of considering population size when looking very far back in the past - assuming a constant population size is unrealistic and may confound parameter estimation. Inferring population histories in general allows us to see firstly the general shape of a population's past, for example when major expansions and bottlenecks occurred. We may then be able to match this to demographic and climatic processes, such as migrations, invasions, and glaciations. Finding the effects of commonly cited events such as a bottleneck in out of Africa populations, for example, will also serve to validate the method. Prüfer *et al.* [2014] used PSMC to estimate archaic species' population size histories alongside those of humans from various populations. It shows both the Altai Neandertal and the Denisovan individual declining gradually to zero between 500kya-50kya (assuming the more recently estimated, slower mutation

rate of 0.5×10^{-9}bp/yr), a time when human populations were shown to be expanding. This is far before humans exited Africa, as represented by a bottleneck for human populations at ∼125kya - and seen more severely in out of Africa populations (although interestingly this is still clearly present in the Dinka, Mbuti, Mandenka, Yoruba, and San: all African populations). A decline of the archaic populations at this time may be the result of direct conflict between the two species, competition for resources, climatic fluctuations, or a subsuming of the archaic populations into the larger human population, as discussed in Chapter 1. *CEPHi* can be differentiated from PSMC in a number of ways. Firstly we use haplotypic rather than genotypic information for the archaic species. Secondly we use population variation data rather than single individuals to infer the size changes over time for a particular population. Thirdly, the population history is inferred jointly alongside split times. Lastly, given this joint inference, we are able to examine the population histories before T_D of the ancestral population of humans and Neandertals directly, and to observe how population size changes relate to the separation between the species, which can give insights into the separation process itself.

In addition to inferring population split times and histories, *CEPHi* also produces genealogical trees detailing estimates of all coalescence times between the set of human and Neandertal haplotypes for each of the recombinationally cold regions input for analysis. In contrast to population split times, a coalescence time is not uniquely defined for a pair of populations - it varies as we move across the genome, and between individuals. It is defined as the time that has elapsed since the existence of the most recent common ancestor (T_{MRCA}) of a collection of copies at a locus (haplotypes in this case). The genealogical trees thus provide a very fine-grained and visual description of the relationship between a human and Neandertal population in that cold region. A comparison of these coalescence times across populations allows us to investigate differences in relatedness between the Neandertal and various human populations at each of the loci examined. An example of this is to compare the times at which the two Neandertal haplotypes coalesce with the human haplotypes. In many cases, the Neandertal haplotypes may coalesce with one another first, so giving one coalescence time with the human population, whereas in others they may intermingle with the human haplotypes. Both the number of coalescences with humans we see in a region, as well as the corresponding coalescence times,

can give information about admixture. Where we see a short coalescence time between the Neandertal haplotypes and a particular human haplotype (as compared with another), we can putatively assign this region as admixed. A non-admixed region should show a late coalescence time with the Neandertal across human populations, somewhere closer to the split time of the two populations. On this basis, we can find a set of regions in one or many populations which may be indicative of introgression. Sets of introgressed regions have been produced by Sankararaman *et al.* [2014] and Vernot and Akey [2014] using different methods, and we compare these with ours in Chapter 4. As far as we are aware, this is the first exploration of genealogical trees in the study of the evolutionary history of humans and Neandertals.

We perform our analyses using *CEPHi* with haplotype datasets from 14 human populations divided into 4 continents from the 1000 Genomes project detailed in Appendix B. We allow human, Neandertal, and joint human-Neandertal predecessor population sizes to vary across 10 epochs to investigate the differences in T_D between each population and the Altai Neandertal, their respective population size histories, and T_{MRCA} across recombinationally cold regions of the genome.

3.2 Methods

CEPHi uses the structured coalescent with mutation and population size changes, and without recombination or migration. This means we assume that two populations (one human, one Neandertal) originate from an instant and complete split from an ancestral population at time T in the past, after which there was no contact between them (looking forwards in time). The intention is to find this divergence time (T_D) and the respective population size histories that are most likely, given the haplotypic sequence data from these populations.

The model does not involve recombination, meaning we use only recombinationally cold regions for analysis, as classified by a set of four genetic maps specified in Chapter 2. This is computationally advantageous, and a further benefit is that admixed regions should be more clearly visible when comparing unrecombined haplotype stretches. The precise number of regions anal-

ysed varies slightly between populations due to filtering procedures differentially affecting human populations, but totals ∼8,500 regions for each population. These are of minimum length 21kb, and mean length 39kb, covering ∼10% of the genome.

CEPHi uses SNP information within these cold regions to jointly calculate the full likelihood of these parameters (population split time T_D and population size history N_{ij}) using importance sampling, and does this for multiple samples; this combination of capabilities does not exist in other available methods. Instead, these often look to infer divergence times and population histories using summary statistics (for example Becquet and Przeworski [2007]; Lopes *et al.* [2009]) rather than calculating full likelihoods, due to the long computation times required for this when using large datasets. These methods therefore lose information available in the data. The loss of information in *CEPHi* by excluding recombination (and therefore reducing the size of the genome for analysis) is mitigated by using multiple samples and regions.

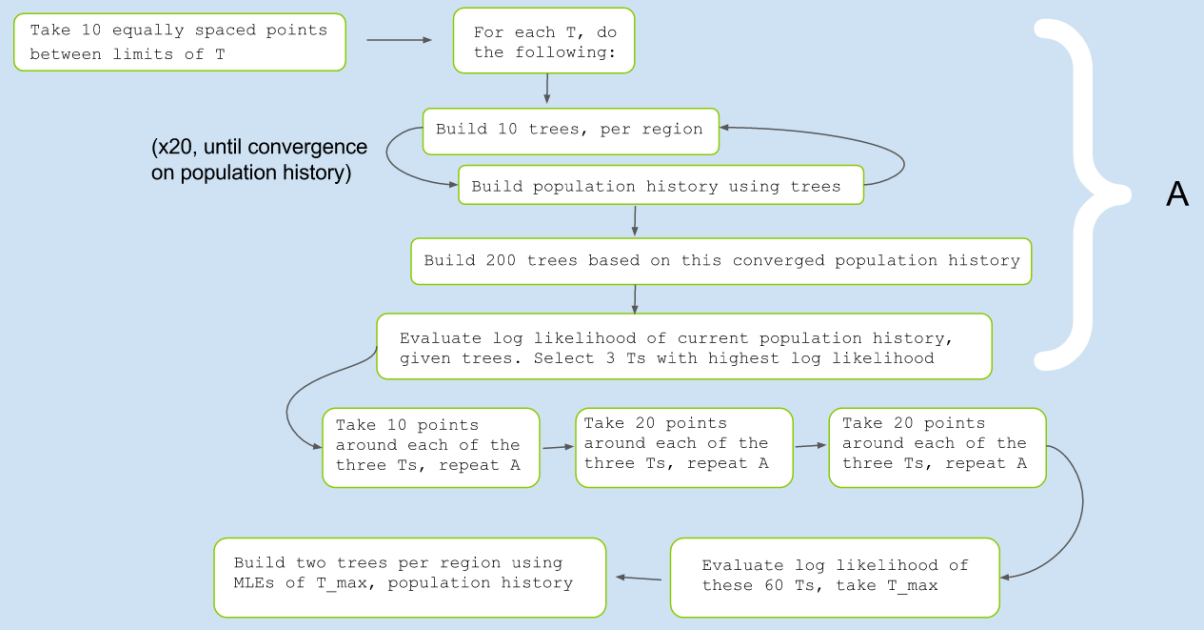


Figure 3.1: Summarising *CEPHi*'s algorithm. Split times, T, are taken, and an EM algorithm used to cyclically (×20) build genealogies using population histories, and vice versa, until a population history for each individual population and their joint ancestral population is converged upon. Further genealogies based on these histories are then built and the likelihood of the histories is calculated given the genealogies. Up to this point constitutes a subset of the algorithm, highlighted here as 'A'. The search algorithm selects those split times with the highest likelihoods, and takes sets of Ts surrounding these, repeating A. This is done repeatedly, until a T_{max} is reached. At this point, *CEPHi* produces two genealogies per coldspot based on the current population histories and split times. These trees are used for downstream analyses in further chapters.

In Figure 3.1 and the bullet points below, we outline the nested algorithms used to produce the maximum likelihood estimates of the population split times (T) and histories (N_{ij}), and then expand on some of the more relevant details:

1. Select 10 T values equally spaced across a plausible range supplied as input (for example 0.05-2.5, equating to 28k-1.4mya for a generation time of 28 years and effective population size of 10,000 - any generation time and effective population size can be selected as output from *CEPHi* is scaled), and for each T, run the following algorithm:

   (a) Using the θ ($= 4N\mu$, where N is population size, and μ is a per base pair per generation mutation rate) supplied as input, build 10 trees for each region using an adapted version of Stephens and Donnelly's Importance Sampler (Stephens and Donnelly [2000]). This choice of θ does not affect later results (as it is simply scaled according to the variation observed in the human population being analysed), though μ must be specified, and has the impact of scaling T and N_{ij} proportionally.

   (b) From these trees, build a population history for all epochs, and select updated parameters to maximise the estimated expected log-likelihood of all trees jointly, via a Monte Carlo Expectation Maximisation (MCEM) algorithm. Trees from different genomic regions are treated as independent.

   (c) This new set of parameters is then used with importance sampling to produce a new set of trees in steps (a) and (b). These two steps are then repeated until the population history converges for this particular T, or for a fixed number of iterations. We fixed this to iterate for 20 iterations to ensure the most complete maximum likelihood estimate search for the population split time, T.

2. Once a population history is fixed for a given T, build a larger number of trees ($\sim$200, a user-defined number) using this population history, and evaluate the log-likelihood at each T.

3. Of the 10 current T values (T_{max}), select those three which gave the highest estimated log-likelihood, and choose a further 10 equally spaced points surrounding each T_{max}, and

replay steps 1(a)-3 for the newly selected set of Ts.

4. Repeat the last step twice, using 20 equally spaced points around each T_{max}.

5. At this point, take 20 equally spaced points surrounding the three values of T with the highest estimated log-likelihood and evaluate the likelihoods at these 60 values of T.

6. Build a small number of trees per region (1-5, given as input) using the MLEs of both T and the population history, which are then output, along with all coalescence times between all lineages within each of these trees.

3.2.1 Building genealogical trees in *CEPHi*

At three points in the above algorithm, genealogical trees for a particular genomic region are built. In each case, this is done using an adapted version of Importance Sampling (IS) (Stephens and Donnelly [2000]). The general function of IS in this scenario is to allow the simulation of genealogical trees that fit the sampled sequences (because the space of possible trees knowing only sample size is unwieldy). However, Stephens and Donnelly [2000] created importance sampling for use within a single population, and in this case we are dealing with a structured population, always with two populations. Apart from the joint analysis of many regions, there are several adaptations to the version of importance sampling used in *CEPHi*. The first is that as genealogies are built, the time of the split between the populations T is taken into account. If the time of the next event happens in the next epoch, or after time T, then the event is not performed, and we instead move into the next epoch, or past time T (where the populations have converged). A second adaptation is seen in a restriction of the possible mutation events - which may only happen in one population if this mutation is not shared with the other. If it is, this event must wait until the populations have converged at T. Thus, events occurring in each of the two populations are not independent of one another before T. Thirdly, the Neandertal sequence is not present in the analysis until a specified time (when we believe the species went extinct). Finally, the importance sample is adaptive, and uses a method of choosing each allowed coalescence or mutation event which depends on the population history.

Under the structured coalescent with mutation, then, a genealogy is built by first simulating the time until the next event. The method works an epoch at a time, so this time will either fall inside the current epoch, or will fall in the next epoch. It will also fall either before or after time T. We will assume for now that it occurs within the current epoch (i) and before time T. In the coalescent with variable population size, the time in units of $2N$ generations until the next event (t_e) is sampled from the exponential distribution as given in Equation 3.1.

$$t_e \sim Exp\frac{\left(k(\theta + (k-1)\frac{N}{N_{ij}})\right)}{2} \tag{3.1}$$

where $\theta = 2N\mu$ (μ =population mutation rate per bp per generation, N=(arbitrary) reference effective population size), k = number of lineages remaining in the population under consideration, and N_{ij} = size of population j in epoch i.

This comes from standard coalescent theory, where the waiting time to a mutation event in a genealogical tree, backwards in time, is exponentially distributed with rate $\frac{k\theta}{2}$ (the 2 is conventional and of no biological significance), and the waiting time to a coalescence event is distributed exponentially with rate $\binom{k}{2}$, and rate $\binom{k}{2}\frac{N}{N_{ij}}$ in units of $2N$ generations if the current population size is N_{ij}. Given that they are independent, it follows that the waiting time until either a mutation or a coalescence event is distributed exponentially with rate $\binom{k}{2} + \frac{k\theta}{2} = \frac{k(\theta+(k-1))}{2}$, simply a sum of the respective probabilities of each event type. Here this is adjusted for effective population size.

At this point we take the set of sequences available for a particular region and, looking backwards in time, ask what the next event could be. The assumption of infinite sites is used, meaning that only one mutation has occurred at any SNP position, turning the base from ancestral to derived. Some sequences will be able to coalesce with others (if they match), and some to mutate (if they uniquely differ from another sequence at some position, and only one sequence carries the mutant type at this position). From this subset of sequences which may be part of the next event, one is selected at random and the event performed. Under the coalescent with variable population size, a mutation event occurs with probability P_m, given in Equation 3.2,

otherwise the event is a coalescence, P_c (Equation 3.3).

$$P_m = \frac{\theta}{\left(\theta + (k-1)\frac{N}{N_{ij}}\right)} \tag{3.2}$$

$$P_c = \frac{(k-1)\frac{N}{N_{ij}}}{\left(\theta + (k-1)\frac{N}{N_{ij}}\right)} \tag{3.3}$$

From above, we can see in the standard coalescent that with probability $\frac{\binom{k}{2}}{\binom{k}{2}+\frac{k\theta}{2}} = \frac{k-1}{\theta+(k-1)}$, the next event will be a coalescence, and with probability $\frac{\theta}{\theta+(k-1)}$, the next event will be a mutation. Equations 3.3 and 3.2 differ from these terms only in that they allow for the population size to be taken into account.

In *CEPHi*, importance sampling is used to introduce a proposal distribution Q, which assigns importance weights to each event according to this distribution. This ensures that the tree built fits with the sampled sequences used in the analysis. A toy example of this tree building process is shown in Figure 3.2. In the sampler, probabilities in Equations 3.2 and 3.3 are not used directly, but instead all possible sequences that could mutate and all those that could coalesce have the same probabilities of having their event performed. The relative probabilities of each class of event (a mutation or a coalescence) are weighted in the proposal distribution so that, if the current inferred population history were the truth, Equations 3.2 and 3.3 approximately hold within each population and epoch, on average across all regions. The appropriate weights are determined by simulation of datasets under the current inferred population history, and calculation of event probabilities with the importance sampler.

A tree, having been built, is given a likelihood under the current model, weighted via importance sampling (Stephens and Donnelly [2000]). These weights are averaged across trees built within each region, then multiplied across regions (because regions are assumed to be independent) to obtain an estimate of the likelihood for the current value of T.

This estimated likelihood, given the current value of T, is then recorded, and the algorithm continues, selecting the value of T with the highest log-likelihood, moving a jump away from it,

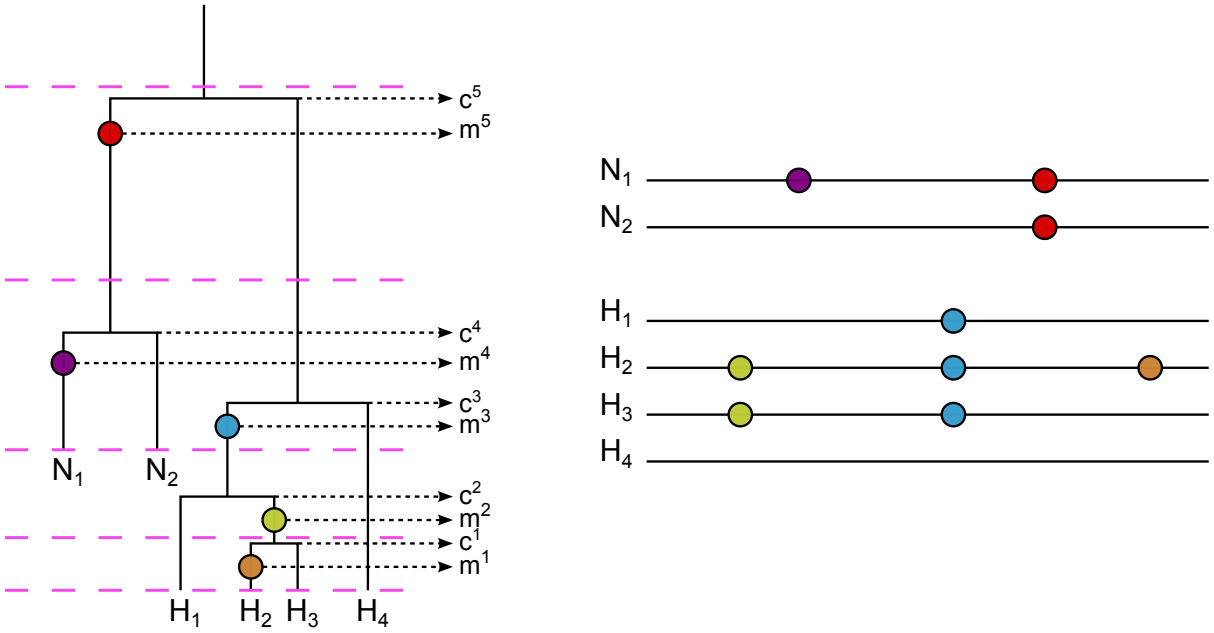


Figure 3.2: Building trees in *CEPHi*. A possible genealogy of a genomic region of four human ($H_1 - H_4$) and 2 Neandertal chromosomes ($N_1 - N_2$) is given on the left, created using the sequences on the right. Circles at different positions/in different colours indicate mutations ($m^1 - m^5$, coalescences occur $c^1 - c^5$. Dashed magenta lines indicate epochs, getting longer as we look backwards in time.

and repeating the process. This continues progressively until a maximum likelihood estimate of T is found as described above.

3.2.2 The relationship between genealogies and population histories

It is simplest to describe the relationship between genealogical trees and population histories once both the maximum likelihood estimate of the split time (T_D) and the population size history are known. This means we have a fixed split time, and delineated epochs with prescribed population sizes for the human and Neandertal populations, both before and after the population split.

Maximising the likelihood of the population histories for a particular human population alongside Neandertals requires the use of genealogical trees. Within each analysis, all trees are split into the same set of epochs as chosen initially, extending the epoch model of Li and Durbin [2011]. Epochs are shorter at the beginning (looking backwards in time) as we will see more coalescent events at the tips of the tree, and so time is compressed in these earlier epochs to

account for this. Within any individual epoch, the likelihood for an effective population size within this epoch is influenced by the product of the probability of all the events that occurred during that epoch, from all the genealogical trees. Each population is evaluated individually. The population size ratio q_{ij} in epoch i, for population j, is defined as $q_{ij} = \frac{N}{N_{ij}}$, where N and N_{ij} are as above. The relative coalescence rate in epoch i for population j is proportional to this ratio, so that larger population sizes N_{ij} yield slower coalescence rates.

The maximum likelihood estimate for q_{ij} is found using an Expectation Maximisation (EM) algorithm. These are used when a model depends on unobserved variables - i.e. data - which in this case are the trees. The E step calculates the expected log-likelihood of new parameters averaging across these unobserved variables, conditioned on the observed data (sequences). The expectation is taken conditional on the current parameters, to yield a Q function, which is approximated by importance sampling (IS) in *CEPHi*. The M step finds the value of a parameter or set of parameters which maximises the Q-function. The maximum likelihood estimate of q_{ij} is given by Equation 3.4, the value of which is then given to the next iteration of the EM algorithm until convergence:

$$q_{ij} = \frac{C}{\sum_{c=1}^{C} \binom{n_c}{2} t_c + \binom{n_{C+1}}{2} \left(\tau_2 - \left(\tau_1 + \sum_{c=1}^{C} t_c\right)\right)} \tag{3.4}$$

where C = the total number of coalescence events in an epoch across trees and regions, t_c = the time between coalescence events, n_c = the number of lineages remaining immediately before the c^{th} coalescence event, and τ_1 and τ_2 are the lower and upper bounds of the epoch under consideration respectively.

3.2.3 Calling SNPs in the Altai Neandertal

The dataset required for input into *CEPHi* consists of all variation found between humans and Neandertals. This therefore includes all human SNPs, all Neandertal SNPs, and all fixed differences between the two species. Human SNP positions are taken from the recently re-released 1000 Genomes haplotypes, encompassing all SNPs identified in 1092 individuals from 14 human

populations from 4 continents, detailed in Appendix B; a substantial set of human variation data, frequently employed for population genetic analysis. We used the most recent version of these data available at the time of analysis, subsequent to the genotypes having been rephased to a higher quality than previously by Olivier Delaneau and Jonathan Marchini. The data is publicly available here: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference.

By contrast, Neandertal SNP positions were called from scratch using two methods: the 'HaplotypeCaller' in Harvard's Genome Analysis ToolKit (or GATK, DePristo *et al.* [2011]), and a hard filtering procedure to call SNPs directly from the raw bam files utilising *Rsamtools*, an R package (Morgan *et al.* [2014]). We then combined these two SNP sets to produce a reliable and conservative set of Neandertal SNPs.

We chose the GATK's HaplotypeCaller over its alternative, the UnifiedGenotyper, because the HaplotypeCaller has a higher true positive (0.9923 vs 0.9843) and lower false positive (0.0012 vs 0.0016) rate for SNP calling (DePristo *et al.* [2011]). This is because when the HaplotypeCaller identifies a region of possible variation, it performs local *de novo* assembly to resolve any misaligned reads before calling SNPs. The UnifiedGenotyper instead relies on the original aligner and simply walks along the genome, calling variation at individual positions where it is found without any subsequent realignment, leading to a higher rate of erroneous calls.

The HaplotypeCaller requires reference SNP sets to refine calls. We could either use only human reference sets, only Denisovan, or a combination of the two. Future analyses could include the use of chimpanzee SNP sets, to reduce reference bias. We ran HaplotypeCaller with all three options on the Altai Neandertal's chromosome 1, and the transition to transversion ratio - an indication of the likely accuracy of the resulting SNP set - varied a small amount between these runs. Across the human genome, the Ti/Tv ratio is ∼2-2.1, meaning that twice as many transitions (within purine A⟷G or within pyrimidine C⟷T base changes) occur as transversions (between purine and pyrimidine base changes A⟷C/T, G⟷C/T). This does not map on to absolute numbers, however, because there are twice as many possible transversions as transitions that can occur. However, this ratio varies even between populations in the 1000 Genomes dataset. The human sets alone gave a ratio of 2.046, the Denisovan set gave

2.111, and a combination of the human and Denisovan sets gave 2.066. Given their similarity, we used the human SNP sets, these being the most comprehensive and reliable. We therefore input the following sets of human SNPs in line with the GATK's recommendations for human data (all sets were input in variant call format (vcf), and reference data in fasta format): the human reference genome build 37, Omni 2.5 genotypes for 1000 Genomes samples, 1000 Genomes high confidence phase 1 SNPs, Hapmap genotypes and sites, and the latest dbSNP release.

Clearly we expect some discrepancy between the true set of SNPs present in the Neandertal, and those we discovered using the GATK's HaplotypeCaller, due to both false positives and false negatives. In light of this, we performed an independent search for all Neandertal SNPs ourselves using the Neandertal sequence files directly with the *Rsamtools* package, and applying our own filters. We downloaded the recently published high coverage Altai Neandertal genome (sequenced by Svante Pääbo's group in the Department for Evolutionary Genetics) in bam format from the Max Planck Institute of Evolutionary Anthropology's website: http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/bam/, as used in Prüfer *et al.* [2014]. Several common preprocessing steps had already been applied to these. These include marking PCR duplicates, indel realignment, and base quality score recalibration. PCR duplicates are molecules that have been sequenced repeatedly, and must be removed from analysis as they provide no further evidence of a SNP or indel, only new molecules (i.e. from a different chromosome or cell) do this. Placement of indels may make it appear that a SNP is present where it is not, and indel realignment adjusts for this. Base quality score recalibration (BQSR) adjusts for biases in the score assignments from the sequencer, but the quality of assignments given by the Improved Base Identification System, *Ibis* (Kircher *et al.* [2009]), is deemed sufficiently unbiased so as not to require readjustment.

This meant we could simply hard filter on the following aspects of the Neandertal data. Firstly we required that for analysis inclusion, each read must have a whole-read mapping quality (MAPQ) or 'Phred' score of at least 30, translating to a ∼99.9% confidence that it is aligned to the human reference genome correctly. We applied the same quality filter (30) to the sequencing quality of each individual base within a read. We removed 3 bases from both ends of each read

to partially account for deamination (the degradation of ancient DNA which, through a large number of C⟶T transitions occurring, can lead to unreliable sequencing, especially at the ends of reads). Once individual reads were filtered in this way, we filtered on coverage using a read pileup, requiring between 10 and 150 reads for inclusion in the dataset. Higher than this could indicate repeat regions where alignment loses reliability, lower and we do not have sufficient information to reliably call SNPs. All bases outside these requirements were removed from both the Neandertal dataset due to being unreliably called.

We made comparisons between the SNP sets called respectively by GATK and the hard filtering method, before combining these to produce a single comprehensive set. Table 3.1 shows the high rate of matching between the two Neandertal SNP call sets across the recombinationally cold regions of the human genome.

Us ↓ GATK ⟶	HomRef	Het	HomAlt	Total
HomRef	**332084728**	51370	27128	332163226
Het	9075	**57850**	471	67396
HomAlt	3648	6917	**344022**	354587

Table 3.1: Assessing the agreement of position categorisation by the GATK HaplotypeCaller and our hard filtering method, throughout our recombinationally cold regions. HomRef = homozygous reference, Het = heterozygous, HomAlt = homozygous alternative.

It is clear that we have a great deal of agreement between the labelling of positions by GATK's HaplotypeCaller and through our hard filtering method. Our set is smaller due to more stringent criteria, and of those that we call, we have 99.98% (332084728/332163226) agreement on homozygous reference alleles, 85.84% (57850/67396) agreement on heterozygous alleles, and 97.02% (344022/354587) agreement for homozygous alternative alleles. The remaining categories have very low numbers, the highest being those that we call homozygous reference which are classified as heterozygous by GATK (0.015% of our homozygous reference calls). These are due to the more conservative criteria we require in order to call a heterozygous SNP: we require coverage of at least 10×, at least 5 base calls of the primary allele and 3 of the secondary; we also require a heterozygous site to have a ratio of ≤80:20 ratio of two alleles, a higher ratio is classified as homozygous for the majority base. A position with, for example 4 calls of the primary alleles and 3 of the second could fairly be classified as a heterozygous position with

GATK but is not classified as such under our criteria. This same logic also applies to those we call as homozygous alternative which GATK classifies as a heterozygous position. The HaplotypeCaller also does not account for deamination, so the ends of reads are not masked, which may increase the number of heterozygous positions being called here in comparison with hard filtering. Lastly, we see a number of positions we call as homozygous reference being classified as homozygous alternative by GATK; this is however a tiny proportion (0.008%) of the number we call as homozygous reference.

In addition to the sets from GATK's HaplotypeCaller and our own SNP set, we also force-called all human SNP positions from the 1000 Genomes dataset using GATK's UnifiedGenotyper to ensure we had information for the Neandertal at each of these positions. We clarify disagreement by referring to counts tables produced using the bam files, and implementing criteria such as requiring a smaller than 80/20 ratio to call a heterozygous position. For clarity and replicability, we reproduce a version of the algorithm used to produce this Neandertal SNP set below.

1. Load GATK HaplotypeCaller output for the coldspot and create GATK-specific masks
   (a) Mask positions classed as indels by GATK
   (b) Mask positions classed as trialleles by GATK
2. Load Altai Neandertal bam file for the coldspot and create hard filtering-specific masks
   (a) Mask positions with MAPQ (whole read quality) <30
   (b) Mask positions with BASEQ (individual base quality) <30
   (c) Trim 3bp of both ends of each individual read to account for deamination
   (d) Mask positions with coverage falling outside 10-150× range
   (e) Mask positions classed as indels by hard filter: where more than 2 reads are indels
   (f) Mask positions classed as trialleles by hard filter: where a third allele has more than 2 reads
3. Combine GATK and hard filtering masks and apply
4. Of the remaining positions, mark as heterozygous, homozygous reference, or homozygous alternative (due to filters, there will be only two alleles to consider at this point)
   (a) Where the ratio exceeds 80/20, class as homozygous reference or homozygous alternative
   (b) Otherwise class as heterozygous

5. Add those classed as homozygous reference to the previous hard-filter mask of the same type
6. Load forced calls from GATK at all human SNP positions
   (a) Find all those which are not in the current SNP set, and are not masked positions
   (b) Where there is only one allele with more than 2 calls, class as homozygous reference or homozygous alternative accordingly
   (c) Where there are 2 alleles each with more than 2 calls, class as heterozygous
   (d) Where there are more than 2 alleles each with more than 2 calls, class as a triallele and add to the corresponding mask

In summary, combining these datasets resulted in a conservative set of Altai Neandertal positions which fulfils the following criteria:

- That all positions used to call a SNP or fixed difference are of high quality (Phred≥30).
- That no positions have <10× or >150× coverage, to prevent low reliability through little information or repeat regions.
- That no positions contain indels as categorised by either GATK's HaplotypeCaller or through the hard filtering. An indel position is defined in the hard filtering as a position with more than 2 reads of an indel in the Neandertal.
- That no positions are triallelic as declared either by GATK's HaplotypeCaller or through the hard filtering. Triallelic is defined in our hard filtering as having more than 2 reads of a third allele in the Neandertal. We use all human, Neandertal, and human-chimp ancestor calls for this classification.
- That each position called as a heterozygous SNP in the Altai Neandertal has a minimum of 3 reads of each allele present, after whole read and individual base pair quality filters have been applied.

Our SNP set for the Altai Neandertal contains 91,138 homozygous reference positions, 247,805 heterozygous positions, and 455,022 homozygous alternative positions.

3.2.4 Phasing the Neandertal genome

We phase the forced calls made at all human 1000 Genomes SNP positions for the Neandertal genome, using the GATK's UnifiedGenotyper. At 0.2% of these positions, we have no Neandertal information - these positions are excluded from analysis. We then phased the Altai Neandertal at these sites using *ShapeIt2* (Delaneau *et al.* [2013]), and the 1000 Genomes human reference panel containing 2184 haplotypes. Given this reference panel, we were able to phase those SNPs which are also human SNPs. We then reinserted all remaining Neandertal variation into the dataset. Sites where the Neandertal is homozygous are trivially phased automatically. As there is no human reference by which to phase the remaining heterozygous Neandertal-only sites, SNP positions present in the Altai Neandertal but not in humans were assigned randomly to haplotypes. This is expected in some cases to make recent admixture, if present, less clear, but not to influence estimates of, for example, Neandertal population size. This resulted in a set of data equivalent to that of the human SNP data from the 1000 Genomes dataset detailing the alleles of the Neandertal haplotypes and legend files to accompany these.

We indirectly checked the phasing accuracy (the true phase is unknown) of our Neandertal dataset produced as described above by inserting missing data into our dataset and testing how well it was imputed. We deleted 3% of positions from the phased Neandertal haplotypes of chromosome 21 produced by *ShapeIt2* (Delaneau *et al.* [2013]) across the cold regions in this chromosome. Using 'prephasing' mode (as we are inputting already-phased haplotypes), we imputed these missing positions using *Impute2* (Howie *et al.* [2009]). We then compared the genotypic information at these imputed sites with the true Neandertal genotypes at these positions. The success rate was reassuringly high, as is shown in the below table. Given that genotypic inference is accurate, we can have confidence that there is sufficient similarity between the human and Neandertal genomes for the related problem of haplotypic phasing to be well performed. Improvements to this phasing method can be made for future analyses, as we address in the discussion.

It is important to note that we have called SNPs using one archaic individual. Ideally we would use multiple Neandertals to match the human samples, but we are limited by the rate and

	Total SNPs	SNPs correctly inferred	% SNPs correctly inferred
HomRef	20484	20070	97.98%
Het	93	69	74.20%
HomAlt	1515	1210	79.87%
Total	22092	21349	96.64%

Table 3.2: Genotyping accuracy of *ShapeIt2* for chromosome 21

amount of sequencing of ancient genomes that has been completed so far.

3.2.5 Combining the human and Neandertal datasets within *CEPHi*

Combining the Neandertal and human SNPs sets required the masking of any position where either the Neandertal reads were insufficient as per the filter outlined above, or the ancestral (chimpanzee) information was not available. Where a SNP exists in the Neandertal but is not found segregating in humans, we set each human as carrying the human reference allele.

After these filters and masks have been applied, a final filter is implemented before *CEPHi* can be used. This removes SNPs that fail the 3-gamete test, which tests for the presence of repeat mutation or recombination. Essentially it removes a small number of SNPs involved in incompatible pairs with all combinations (0,1), (1,0), and (1,1) in existence. The presence of this pattern is not compatible with a recombination-free scenario under the infinite sites model of mutation; a genealogical tree cannot be drawn to include these sequences. This test is done for all individuals.

3.2.5.1 Correcting for heterozygosity in coldspots using *CEPHi*

In adapting *CEPHi* for use with Neandertals, one of the most significant alterations comes in the form of adjusting the population-scaled mutation rate, θ, to account for properties of our cold regions relative to the rest of the genome. A number of features influence the value of θ. First, the application of filters to account for incomplete chimpanzee and Neandertal coverage, and failure of the 3-gamete test, causes removal of sites, reducing variation relative to

the genome as a whole in a manner equivalent (at least in the first two cases) to a reduction in the mutation rate. Secondly, the underlying heterozygosity in the human population in the cold regions used for analysis may differ from the average genome-wide heterozygosity commonly used. To account for both sets of features, a correction factor (θ_{adj}) was implemented, which adjusts the 'genome-wide' value of θ by a factor θ_{adj} to represent the final diversity of the cold regions analysed, post all filters. This factor is given by:

$$\theta_{adj} = \frac{\frac{V_r}{n_r l_r}}{\frac{V_g}{n_g l_g}} \tag{3.5}$$

where V = mean number of pairwise SNP differences across all samples in the human population being considered, n = number of individuals in that population, and l=number of base pairs being considered (r = cold regions, g = whole genome). This correction is performed separately for each human population, and values are detailed in the results section below. The correction accounts for filtering-based reductions in observable diversity in our cold regions, and operationally results in a scaling up by a factor $\frac{1}{\theta_{adj}}$ of both estimated split times, and population sizes, for a particular choice of mutation rate (μ) input to the program.

3.3 Results

We present results from two sets of analyses using *CEPHi*, which vary only in the range of possible split times input into the software. The first set provides human-Neandertal split times and population size histories, and uses the range 0.08-2.5 (scaled by $2N$ generations), equating to absolute times of 44,800-1,400,000 years ago. These boundaries were chosen to provide significant space to find human-Neandertal split times of the highest maximum likelihood without restriction. Importantly, the lower bound is also a close match to the fossil age of the Altai Neandertal of ~50,000 years (Mednikova [2011]) with a little additional space. The second set uses the range 0.04-0.09, equating to absolute times of 22,400-50,400 years ago; as close to zero

as was permitted by the software, with an upper bound just above the given fossil age. Allowing early human-Neandertal coalescences in this way enabled us to produce distributions of coalescence times across cold regions of the human genome, and genealogical trees for individual cold regions to show region-specific relationships between all human (from within a particular population) and Neandertal haplotypes.

3.3.1 Correcting for sequence diversity in coldspots

On applying the filters mentioned in the section above, a number of SNPs are removed from each population's dataset before analysis. This is due to (a) a lack of coverage in the chimpanzee or Neandertal (we have human information at all positions by definition), (b) because a SNP is deemed at least triallelic, or (c) because a SNP fails the three gamete test and is therefore thought to have undergone recombination.

Given the levels of SNP removal in each dataset, we present the correction factor (θ_{adj}) used for each population analysis in Figure 3.3. This varies per population, and is calculated as given in Equation 3.5. It gives a measure of the heterozygosity contained within the recombinationally cold regions used in that population (post filter application), normalised by the population-specific heterozygosity present across the genome. The corresponding θ then used in *CEPHi* refers to the final mutation rate for that population, after the correction factor has been applied. The larger the correction factor, the lower the final θ for that population.

It is clear that θ_{adj} and therefore θ are very similar across populations, and also result in values of μ close to, but slightly lower than the revised human mutation rate of 0.5×10^{-9}bp/yr given by Scally and Durbin [2012]. This is because we see lower heterozygosity in coldspots. The Iberian population (IBS) shows a slightly higher rate than other populations, likely due to being constituted of significantly fewer samples (14) - all others contain at least 55 samples.

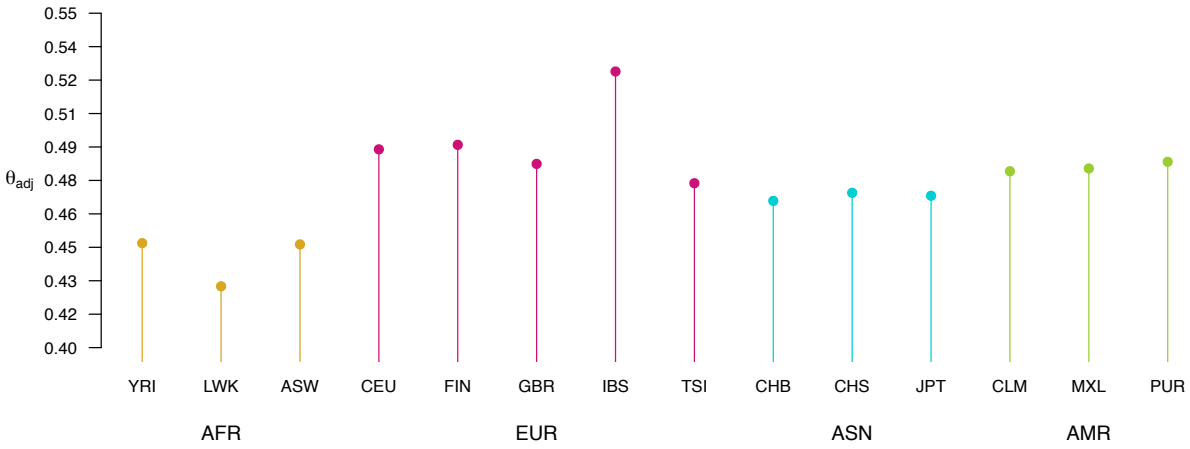


Figure 3.3: Plotting θ_{adj} - a per population correction factor which reflects heterozygosity in recombinationally cold regions as compared with genome-wide heterozygosity. The higher the θ_{adj}, the lower the resulting θ used for that population. Populations are grouped and coloured by continent.

3.3.2 Deconstructing the human-Neandertal evolutionary scenario

We first present results from a model searching for split times between each of 14 human populations and the Altai Neandertal. This model uses the structured coalescent (we have two populations per analysis), without migration or recombination, and with mutation. Importantly, it permits variable population size, where variation in size of the joint human-Neandertal population, the individual human populations, and the Neandertal population is permitted between epochs (of which there are 10) but remains constant within them.

3.3.2.1 Population split times

We first present the searches (Figure 3.4) for the maximum likelihood estimate of T_D for each population alongside the Neandertal. For each population, we performed searches between 44,800-1,400,000 years ago.

The searches are all highly unimodal, and the search space is very well explored, providing confidence that the resulting human-Neandertal population split times are stably estimated. Most notably, we see much later estimated split times in African (between 535-713,000 years) as compared with European, Asian, and American populations (between 313-478,000 years).

Split times between Neandertals and humans from putatively non-admixed populations can be thought of as realistic absolute times at which the predecessor population of these resulting species (or sub-species) bifurcated. Using YRI as perhaps the best current example of this, we estimate the date at which humans and Neandertals split to be 712,936 years ago. By contrast, in the non-African populations, we cannot take these to be equivalent dates. In fact what these dates show is some average of the YRI-Neandertal split time, and the date (or range of dates) at which admixture occurred between Neandertals and that modern human population. For GBR, for example, we see a date of 348,264, just under half of the YRI-Neandertal split time. We list the human-Neandertal population split time estimates in Table 3.3.

Continent	Population	T_D	Years in past
Africa	YRI	1.273	712,936
	LWK	1.270	710,976
	ASW	0.957	535,696
Europe	CEU	0.574	321,244
	FIN	0.560	313,460
	GBR	0.622	348,264
	IBS	0.716	401,100
	TSI	0.573	321,160
Asia	CHB	0.567	317,352
	CHS	0.854	478,464
	JPT	0.604	337,988
America	CLM	0.710	397,600
	MXL	0.600	336,140
	PUR	0.632	353,668

Table 3.3: Split times between 14 1000 Genomes human populations and the Altai Neandertal. Population acronyms (second column) are given in Appendix B, T_D is the split time scaled by $2N$ generations (third column), and the final column gives the human-Neandertal split time in years.

Figure 3.5 allows an easy comparison of the resulting maximum likelihood estimates of T_D inferred from this analysis. As we have seen, there is a clear distinction between the African and non-African populations, the split time with Neandertals being much earlier in the former. We include two of the most recent estimates of YRI-Neandertal split times on Figure 3.5, from Meyer *et al.* [2012] and Prüfer *et al.* [2014]. These both use a mutation rate of 0.5×10^{-9}bp/yr. It is clear that our African-Neandertal split times fall within both these ranges, and are closer to that of the most recent from Prüfer *et al.* [2014]. Our non-African populations fall well below

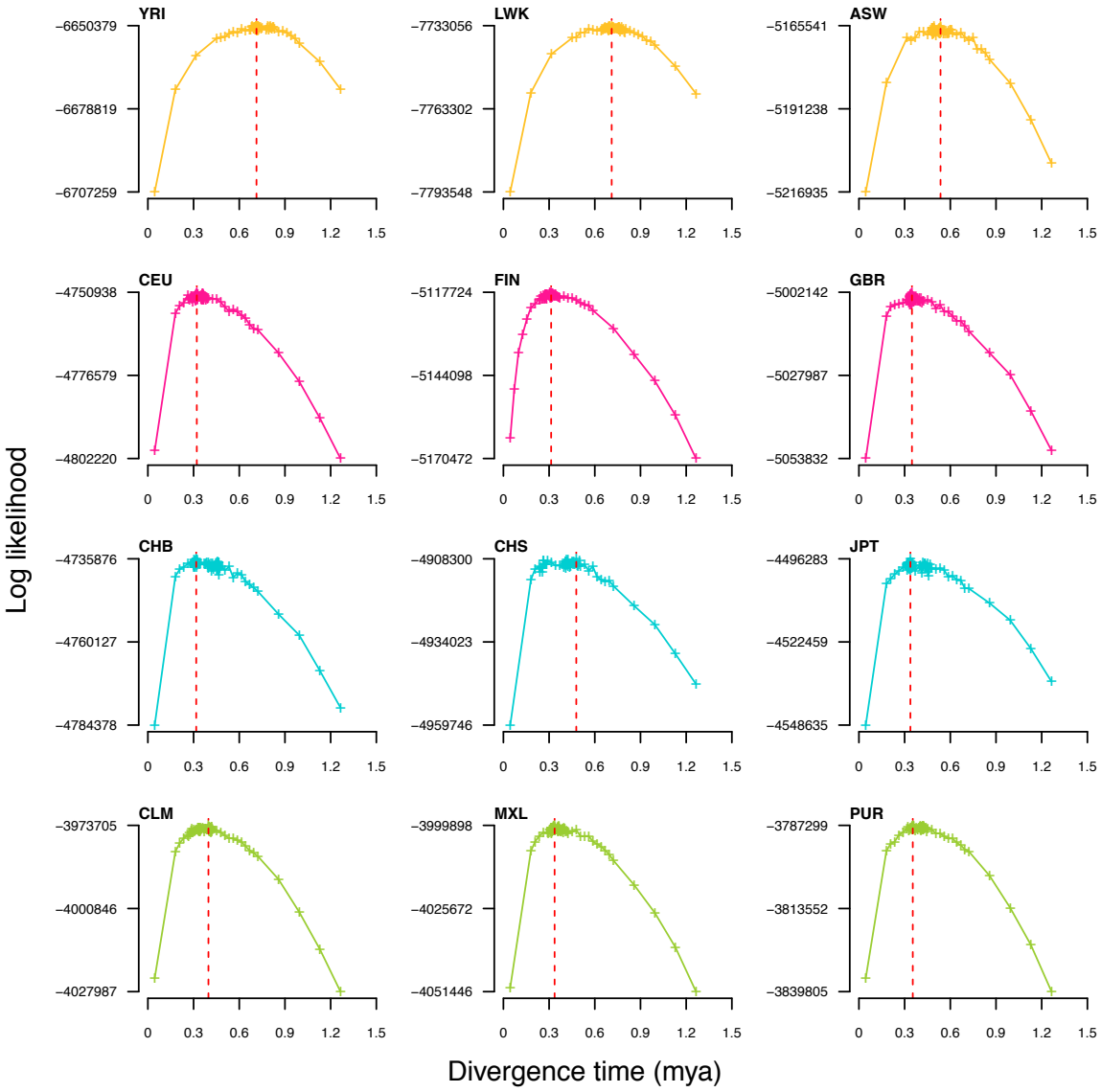


Figure 3.4: Maximum likelihood searches for split times (T_D) for individual populations across four continents (Africa, Europe, Asia, America: descending) using a model with variable population size. The vertical dashed red line highlights population-specific MLE of human-Neandertal T_D.

these estimates as is to be expected, likely explained by admixture having occurred in these populations.

The Han Chinese population from southern China (CHS) show a population split time with Neandertals that is perhaps slightly earlier than expected. Figure 3.4 shows us the slightly more bimodal surface between 0.2 and 0.4mya, and the high likelihood for a split time with

Neandertals closer to that of the other non-African populations. ASW (a population with African ancestry in the southwest US) have the most recent split time with Neandertals, which is to be expected as they are likely to have had recent admixture with European populations. We also note the slightly higher split times between American populations and the Altai Neandertal. This is likely to be due again to some admixture with individuals of African descent in these regions.

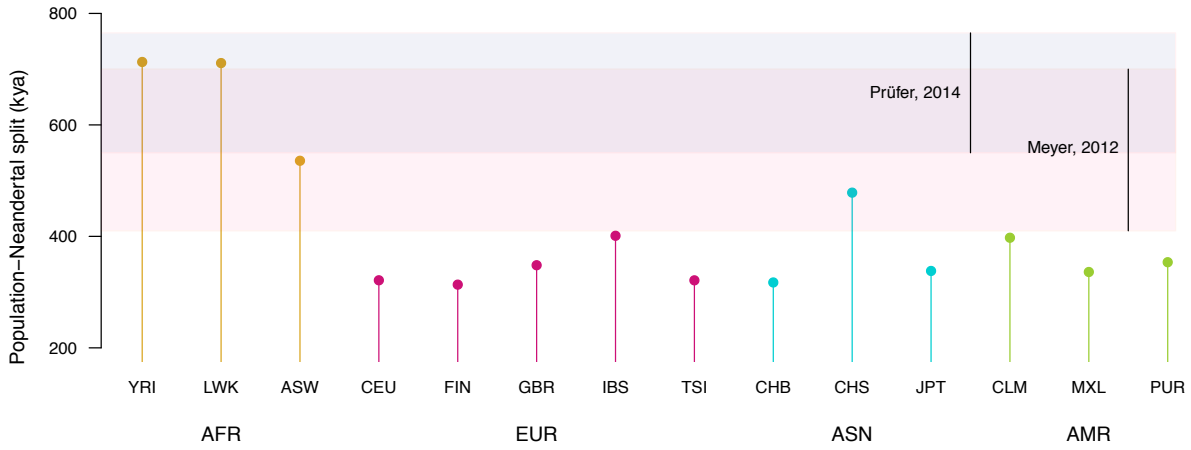
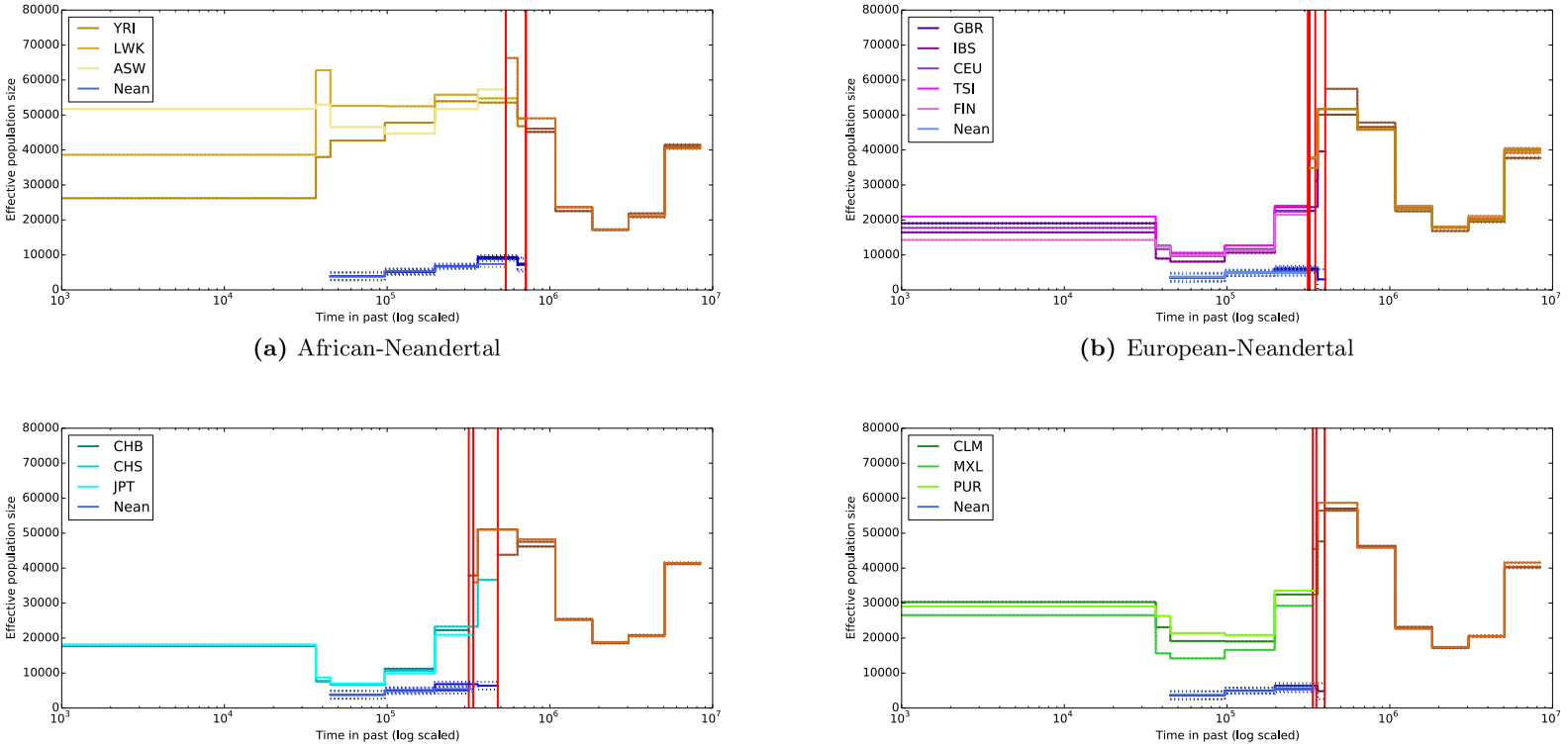


Figure 3.5: Comparing *CEPHi* population split times between the Altai Neandertal and 14 1000 Genomes human populations, grouped by continent (Africa, Europe, Asia, America). Equivalent Neandertal-human population split time range estimates from Prüfer *et al.* [2014] (pale blue) and Meyer *et al.* [2012] (pale pink) using the Yoruban (YRI) population and a mutation rate of 0.5×10^{-9}bp/yr given for comparison.

3.3.2.2 Population histories

Inferred population histories for individual continental populations from the second *CEPHi* analysis are shown in Figure 3.6. By allowing population size to vary, we can see how human and Neandertal populations may have expanded or contracted at various times between our boundary dates of 44,800-1,400,000 years ago. Effective population size scales inversely with coalescence rate, so in an epoch where, for example, a human population is small, we will see a greater number of coalescences between haplotypes.

Before each population split time (shown with a vertical red line) and once the populations have coalesced with one another to become a single population, we see very consistent population size



(a) African-Neandertal

(b) European-Neandertal

(c) Asian-Neandertal

(d) American-Neandertal

Figure 3.6: Population size histories for individual populations grouped by continent. Time (on the x-axis) is log-scaled and we look backwards in time as we move left to right. 10^5 is, for example, 100,000 years ago. For each plot, the vertical red lines highlight the population split times for each human population with the Altai Neandertal. The brown lines to the right show the population size change of the joint human-Neandertal population before the population split time. The coloured lines to the left of the vertical red line give the population size changes of each human population listed in that plot (see the legends at the top left of each plot), and of the Altai Neandertal populations for each of these human populations (shown in dark blue).

changes between populations and even continents; this varies very little. After the split time (looking forwards in time), we see significant human population bottlenecks in all non-African cases, as we would expect: populations moving out of Africa will have likely been small compared to the numbers remaining in Africa. By contrast, all African populations show a reasonably steady population size, with some reduction. Notably, we also see a very sudden, dramatic drop in Neandertal effective population size following the human-Neandertal split (looking forwards in time), then a continuing steady reduction before the Neandertal goes extinct. At the time of Neandertal extinction, we see an increase in the human population sizes across the non-African board.

3.3.3 Genealogical trees and coalescence times

CEPHi produces highly detailed information about the T_{MRCA} within each cold region between all the individuals in the human population under consideration, and the Altai Neandertal. We are interested in when each of the Neandertal haplotypes coalesces with a human haplotype. In most cases, the two Neandertal haplotypes will coalesce with one another first, and later coalesce with a human haplotype, after most (or all) human haplotypes have coalesced with one another, given the restriction in the method that does not allow a coalescence between populations until T_D. This is because most genealogical trees for a particular genomic region will simply reflect the phylogenetic relationship between humans and Neandertals. Sometimes, however, we will see a Neandertal haplotype coalesce with a human haplotype before it coalesces with its own second haplotype. One of the most important pieces of information we garner from these trees is the time at which the Neandertal coalesces with the first human across regions, and therefore how these times are distributed. We therefore reran *CEPHi*, now using an early split time of <50,400 years ago, to allow coalescence between humans and Neandertals at all times younger than the Neandertal fossil age. Where these times are unexpectedly recent - given a speciation over 600kya - we can label a region as putatively admixed. Comparing these distributions across populations can allow us to see whether populations and even continents are distinct from one another in this regard, and we do so in Figure 3.7.

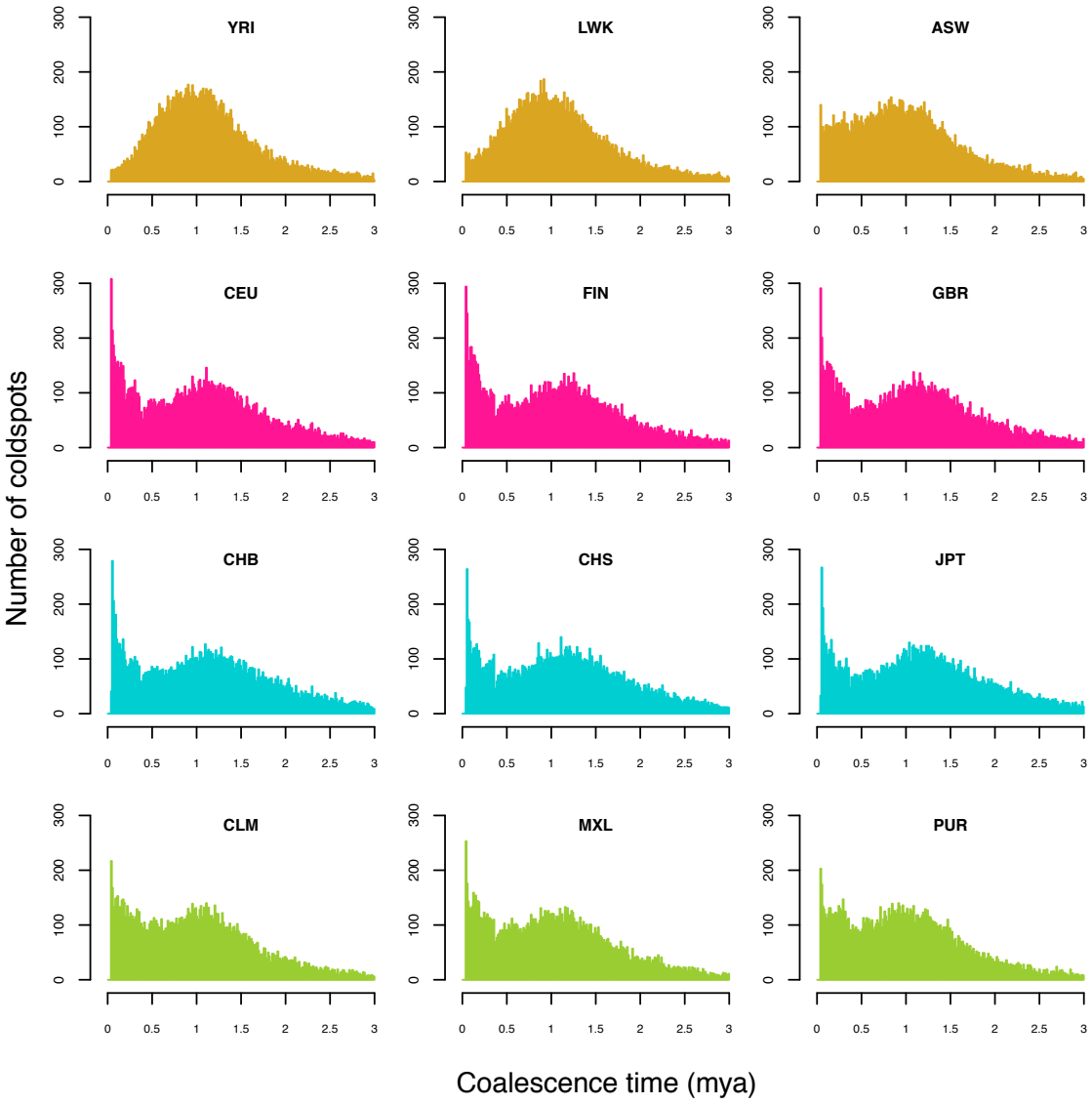


Figure 3.7: Estimated coalescence time distributions between each human population and the Altai Neandertal for a population split time range $T < 50{,}400$kya: forced early population split times. One continent per row (African, European, Asian, American: descending).

Again we see a clear distinction shown between the African distributions on one hand, and the European, Asian, and American distributions on the other. The African populations show a very clear unimodal peak around the split time with Neandertals. This is particularly clear in YRI and LWK. ASW (an African population with recent European admixture living in the United States) also shows this peak, but in addition, we see a substantial number of regions

with more recent coalescence times in this population. Given that admixture events are not explicitly modelled by *CEPHi*, the presence of admixture might produce a distribution similar to that seen in the non-African populations, with large numbers of recent coalescences, as well as a peak in the number of regions around the original split time between Neandertals and humans. LWK is likely to have had a small amount of recent admixture, ASW more, while YRI represents a non-admixed population; we get a sense of the blurred picture admixture might provide with regard to these distributions. Strikingly, we see strong but narrow secondary peaks close to zero in each of the non-African populations, seen most prominently in the European and Asian populations, but also in the American populations. In reality, the x-axis extends much further, in some cases to tens of millions of years, however we trim it to clearly display the distribution of coalescence times before and around the population split times.

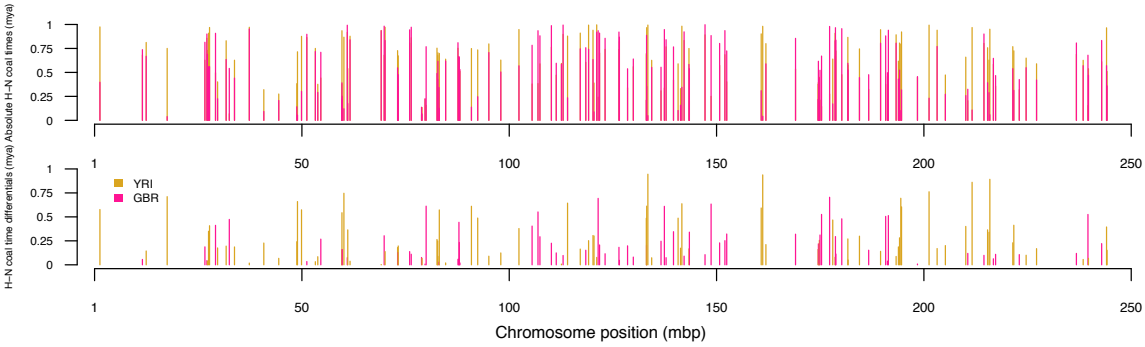


Figure 3.8: Comparing the minimum coalescence times of the Neandertal and human haplotypes from one African (YRI) and one European (GBR) population across chromosome 5. Top: Absolute coalescence times with the Neandertal for YRI (dark yellow) and GBR (pink) are shown for comparison. Bottom: Difference in coalescence times between Neandertal-YRI and Neandertal-GBR; dark yellow where the Neandertal-YRI coalescence time is longer, and pink where the Neandertal-GBR coalescence time is longer. Regions are filtered to show only those where coalesce with the Neandertal is $\leq$1 million years ago for both populations.

Figure 3.8 makes a comparison between the coalescence times presented in Figure 3.7 for an African (YRI) and a European (GBR) population, for chromosome 5 containing $\sim$400 regions per population in this comparison. Figure 3.8 displays only those regions for each population that coalesces with a Neandertal haplotype at a maximum of 1 million years ago, so the sets differ slightly. It is clear that overall, YRI frequently show coalescence times with the Altai Neandertal across chromosome 5 that are significantly later than those seen for the same regions between GBR and the Neandertal. Where very short coalescence times are seen in the non-

African population (in a range between the extinction of the Neandertal and a point in the past at which Neandertals and humans may have encountered one another), and substantially longer coalescence times are seen between the African human and Neandertals (for example, greater than 600kya), we can putatively assign these regions as admixed. We explore this further in Chapters 4 and 5.

In Figures 3.9 and 3.10 we present genealogical trees for an example region from chromosome 1 for two contrasting populations, YRI and GBR. The genealogy in Figure 3.9 shows a single coalesence time between YRI and Neandertals, given a Neandertal-Neandertal coalescence time at 1,658,609ya (looking backwards in time). In contrast, Figure 3.10 - a genealogy of the same region - shows an intermingling of human with Neandertal haplotypes, meaning the two Neandertal haplotypes coalesce with humans separately, before their respective lineages coalesce together later. These coalescence times are also much more recent. This region comes from the left hand side of the GBR distribution in Figure 3.7. In Chapters 4 and 5 we take investigation of these regions further, including dating the mixing in all non-African groups, and examining the direction of admixture between human and Neandertal populations.

3.4 Discussion

Using *CEPHi*, we have jointly inferred the population split times between the Altai Neandertal and 14 human populations, as well as the changing size of these populations over time. We have also begun to explore the distribution of coalescence times between Neandertal and human haplotypes. This information gives us a number of insights into these species' evolutionary history.

Firstly, we see a much earlier split time between African populations and Neandertals than we do for non-African populations. This reflects previous investigations into the relationships between African and non-African populations with Neandertals (Green *et al.* [2010]; Prüfer *et al.* [2014]), which show evidence of a closer relationship between non-Africans and Neandertals than between Africans and Neandertals, with the genome of non-Africans harbouring somewhere

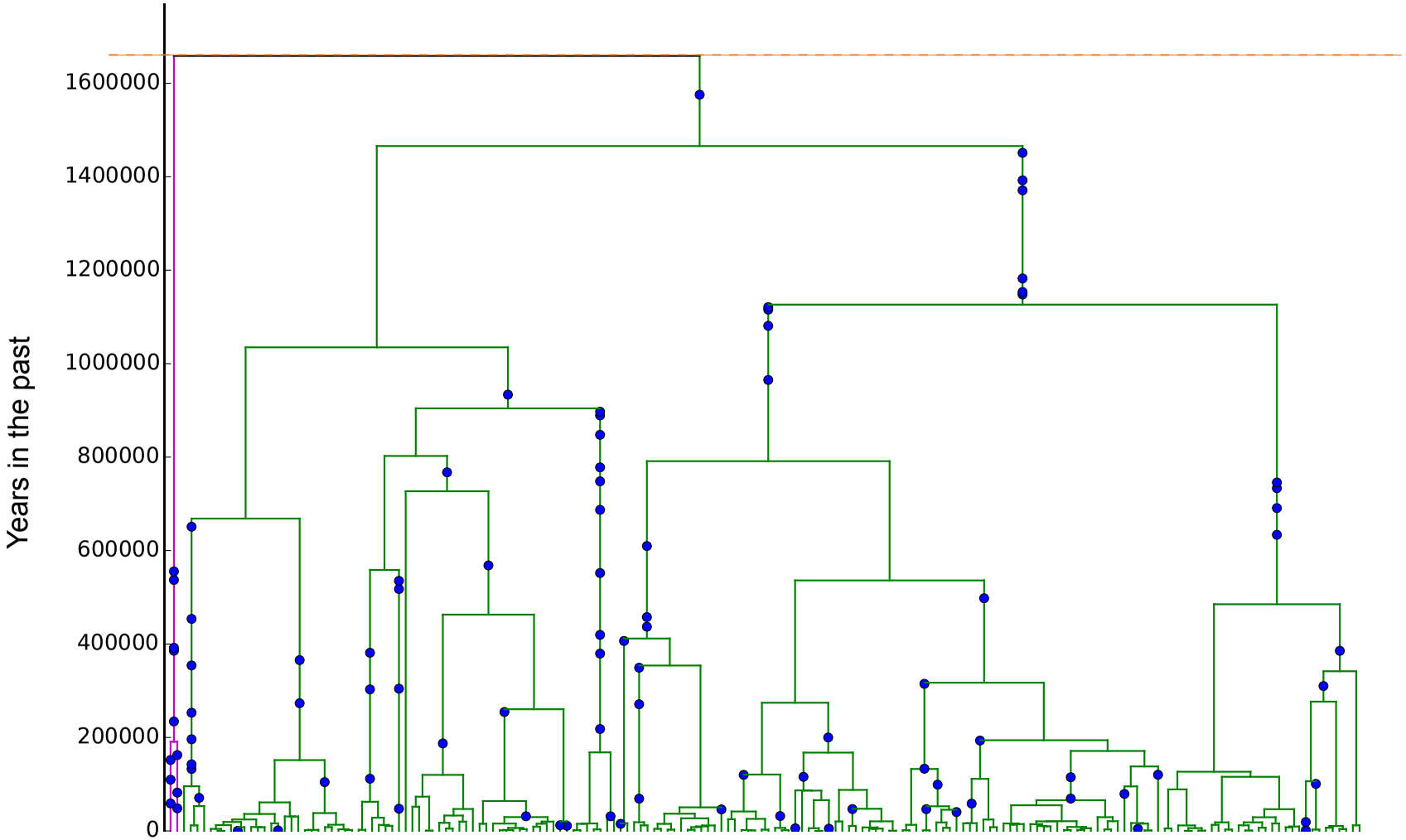


Figure 3.9: Comparing genealogical trees between an African (YRI - this plot) and a non-African (GBR) population. Chromosome 1: 10010344-10034056 (hg19 positions). Branches representing Neandertal haplotypes are shown in pink, human in green, and those where the populations join in black. Blue dots indicate mutations. YRI shows a single coalescence time with the Neandertal haplotypes at 1,658,609ya (horizontal dotted line).

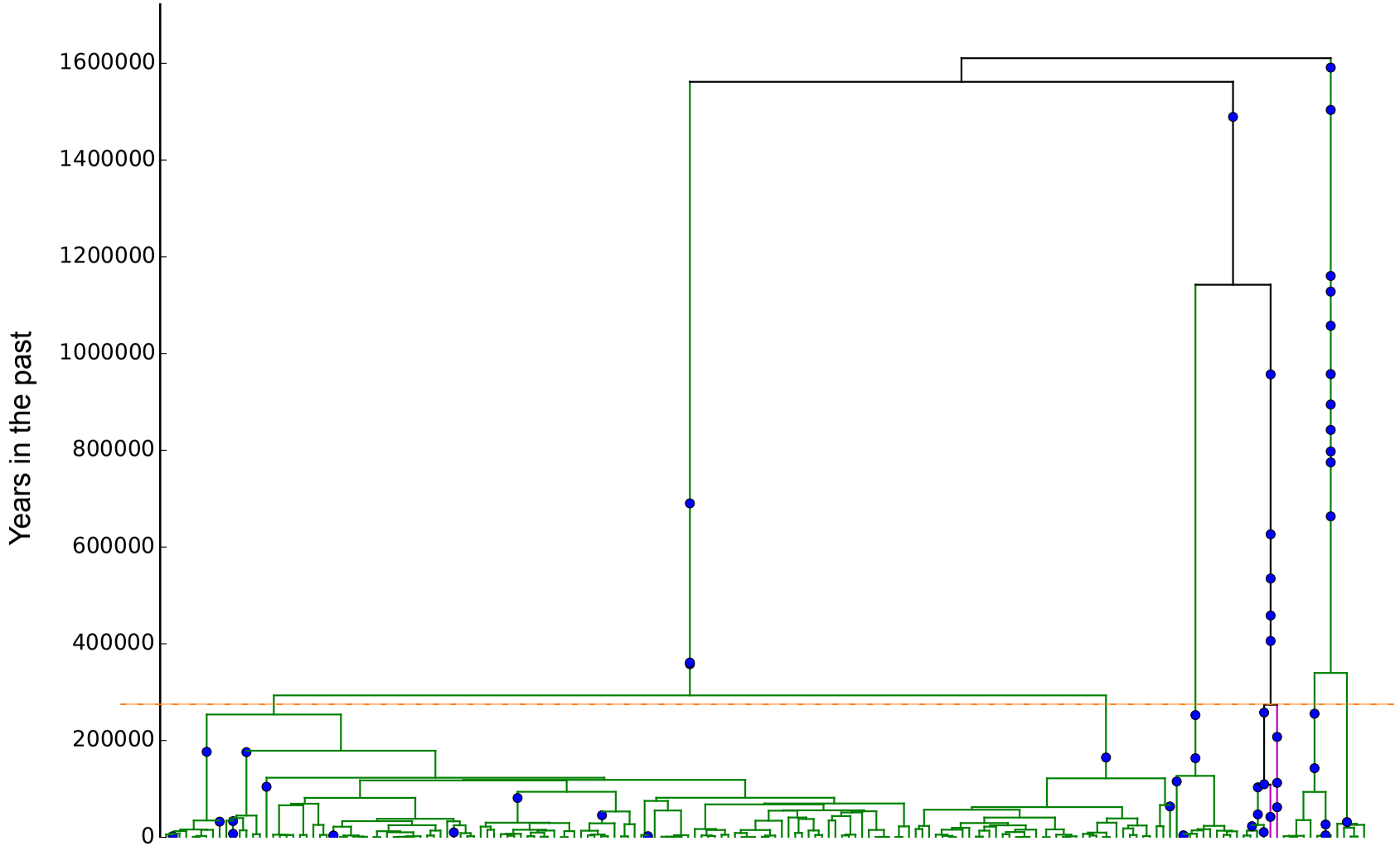


Figure 3.10: Comparing genealogical trees between an African (YRI) and a non-African (GBR - this plot) population. Chromosome 1: 10010344-10034056 (hg19 positions). Branches representing Neandertal haplotypes are shown in pink, human in green, and those where the populations join in black. Blue dots indicate mutations. GBR shows early coalescence of human haplotypes with the two Neandertal haplotypes, meeting them at 108,906 & 273,873ya (horizontal dotted line at the latter time). We see four human descendants of admixture in this region.

between 1-4% Neandertal DNA. This distinction is drawn very clearly between Africans and non-Africans.

Average split time for the American populations is slightly higher if we omit CHS (the T_D of which may actually be closer to CHB and JPT, as can be seen in the MLE searches in the results section) and IBS (for which we have significantly fewer haplotypes): later than the population split times we see for other non-Africans in this analysis, but earlier than those for Africans, which we can attribute to the recently admixed nature of these populations (for which the Americans are exceptional in this set of populations). Puerto Ricans (PUR), for example, are a mixture of European (Spanish, Italian), Native American, and West African ancestry - the product of colonial movements such as the Conquistadors and the American slave trade. The history is very similar for both the Mexican (MXL) and Columbian (CLM) populations. We know that these recent admixture events may be present in the American populations analysed here (Hellenthal *et al.* [2014]), and because the model applied does not account for admixture events in the recent past, it can result in a later population split time.

Overall, the population split time estimates show African populations to be further differentiated from the Altai Neandertal than non-African populations are. The smooth nature of the maximum likelihood curves shows, in general, a confident ascertainment of the split time between these populations and Neandertals. If admixture has occurred between Neandertals and non-African populations, we cannot use their respective MLEs as true split times. Instead, the MLE reflects some average of the original split time between each population and Neandertals (stretching back to the YRI-Neandertal split time), and more recently, the pulses of admixture that have occurred - this may be more than one. Demographically, we can envision a small proportion of a large expanding human population admixing with small pockets of Neandertals across Eurasia, potentially in pulses as the human population expanded into different areas.

Our split time estimate for Neandertals and Africans (YRI specifically) of just past 700kya falls very much in line with a recent estimate from Prüfer *et al.* [2014] of 550-765kya, and is slightly later than their second estimate of 553-589kya, although this latter range was calculated for the San and Mandenka populations from Africa, neither of which was included in our analyses. It

also falls at the top end of the estimate by Meyer *et al.* [2012] of 410-700kya. Agreement with these publications as a result of having used a novel and distinct method provides confidence in these results.

The population size histories for the Neandertal-human population pairings reveals some strong consistencies, particularly with regard to the sizes of the joint populations before the two populations split (looking forwards in time), which is shown to have dropped and then increased in size very similarly across all populations and continents. Equally, at each of the respective split times, Neandertal population sizes drop substantially in all cases, as we would expect when a subset of their ancestral population exited Africa, before slowly declining until the species suffers extinction. This is consistent with the idea that human and Neandertal populations were separated by the predecessor of the Eurasian Neandertal leaving Africa, and the predecessor of modern humans remaining. European, Asian, and American human populations also experience a bottleneck at their respective split times, but one that is less severe than that of the Neandertals. By contrast, African populations have a reasonably consistent (although somewhat decreasing) effective population size, as is expected. The Out of Africa bottleneck is visible particularly in the European and Asian populations, continuing the population declines from the respective split times. Human population size stays consistently higher than that of the Neandertal until the Neandertal goes extinct at ∼50 thousand years ago, as defined by our model. In comparison to Prüfer *et al.* [2014], who used a modified version of PSMC (with an added split time) to infer the population histories of both the Altai Neandertal and Denisovan individual, we see some overlap, in that there is continuing decline in the human populations since before the Out of Africa bottleneck, and that human populations are larger than those of archaic hominids. African effective population sizes are reasonably high until the present, followed by American, and European and Asian, as we would expect.

The coalescence time distributions reinforce the idea that non-African populations have most likely admixed with Neandertals, as we begin to specify the relationships between human and Neandertal haplotypes across recombinationally cold regions of the genome. We see significant weight on the left hand sides of non-African distributions, highlighting the fact that large

numbers of regions show short coalescence times between human and Neandertal haplotypes for these populations. This is most clearly seen in Eurasian populations, and still very clearly visible in Americans. The fact that the Yoruban (YRI) and Luhya from Kenya (LWK) populations show very little weight at this end of the distribution, and that the more recently admixed African population from the south west United States (ASW) shows a small proportion of regions with shorter coalescence times with the Altai Neandertal compounds the idea that this is as a result of human-Neandertal admixture.

The genealogical trees created by *CEPHi* hold a wealth of information. We have begun to explore the distribution of coalescence times in each population between the two Neandertal lineages and any of the human lineages. The peak in the number of regions at the split time between humans and Neandertals, especially visible in YRI, can be viewed as a base distribution for a non-admixed population from which we can understand distributions from other populations. All non-African populations show a visible set of regions coalescing before this split time, which may be indicative of admixture. Details of the genealogical trees will allow us to explore these regions further.

In the following chapter, we provide a region-specific definition of admixture using a comparison of coalescence times between African-Neandertal genealogies with those linking Neandertals and non-Africans. This allows us to identify sets of population-specific introgressed regions, the features of which we then detail and compare. We also highlight the presence of the two Neandertal haplotypes coalescing separately (intermingling) with human lineages as an intriguing signal. In the final chapter, we introduce and apply a novel method which dates admixture across European and Asian populations using information about SNP placement on genealogical trees, alongside corroborative simulations.

CHAPTER 4

Extracting signals of admixture from genealogical trees

4.1 Introduction

In the previous chapters, we used SNP information contained in recombinationally cold regions of the human genome to search for bimodal distributions of estimated coalescence times between human and Neandertal haplotypes. By building genealogical trees between these two-population haplotype sets, we inferred split times and population size histories, leading us to conclude that using this novel method, we see highly visible signs of admixture between non-African populations and Neandertals in human evolutionary history.

A substantial amount of information is contained in the thousands of genealogical trees produced

by *CEPHi*, and in this chapter, we focus our attention on analysing this in order to visualise, explore, and quantify admixture and introgression events between the two species. Genealogical trees showing the specific relationships between a set of haplotypes from *Homo sapiens* and *Homo neanderthalensis* allow us not only to observe the distribution of coalescence times between human and Neandertal haplotypes (as addressed to some extent in Chapter 3), but also to examine Neandertal introgression within and between particular populations. First, we explore some methods which have been used previously to detail introgression from Neandertals in modern humans.

4.1.1 Neandertal introgression in modern human populations

Two recent publications have provided sets of Neandertal-introgressed regions in modern human populations: Sankararaman *et al.* [2014] and Vernot and Akey [2014]. We detail their methodologies and results.

Sankararaman *et al.* [2014] applied a Conditional Random Field (CRF) across the genome to identify individual SNPs and regions as of Neandertal origin, and therefore putative regions of introgression in individuals from the 1000 Genomes Project. This CRF combines characteristics of a test haplotype to infer whether an allele at a SNP has been introgressed from Neandertals into modern humans. A CRF is similar to a Hidden Markov Model (HMM); the ancestry of SNPs and thus regions is the unobserved state, the observed being the haplotype sequences of the test populations. The emission functions couple the unobserved ancestral state to the observed features at SNPs across the test haplotype, to give some probability of a particular position having Neandertal ancestry, given its observed allele. There are two classes of emission functions. The first involves two indicator variables which both use information about the alleles across African, non-African, and Neandertal populations. The first of these indicator variables is marked as '1' if three conditions are satisfied: (a) the human carries a derived allele, which (b) is also found in at least one of the Neandertal haplotypes, and (c) all Yorubans from Ibadan, Nigeria (YRI) carry the ancestral allele. If all three of these conditions are not satisfied, the position is marked '0'. SNPs fulfilling these conditions are more likely to have

Neandertal ancestry. The second indicator variable for a given SNP is marked 1 in the test haplotype if the allele is derived, not present in Neandertals, and polymorphic in YRI; it is marked 0 otherwise. Here, SNPs marked 1 are less likely to have Neandertal ancestry. The second class of emission functions uses non-overlapping haplotype regions of 100kb in which there are multiple SNPs. Where the divergence of the test haplotype to the Neandertal is less than the minimum divergence of the test haplotype to the set of YRI haplotypes, it is more likely to be of Neandertal ancestry than if the conditions are reversed. Where heterozygous sites were present in the Neandertal or test haplotype, the derived allele was chosen, thus minimizing the distance between the test and Neandertal haplotypes (this could potentially lead to overestimation of Neandertal introgression). The transition function was given using probabilities from a standard Markov process of admixture between two populations, and the CRF parameters were estimated through simulation using three populations (Neandertal, GBR, and YRI), with admixture occurring 1,900 generations ago at 3%. The distribution of unobserved ancestral states comes from the feature functions and their associated parameters, given the observed data. This then allows inference of the marginal probability of Neandertal ancestry at each SNP along a test haplotype. Those regions with a marginal probability of >90% are then described.

In summary, this method uses individual SNPs and sets of SNPs to locate haplotype regions in the test population which are likely to have been introgressed from Neandertals. This is based on their matching the Neandertal and differing from YRI (a reference population presumed not to have admixed with Neandertals) at both an individual SNP and region level, and which are of length of ∼0.05cM, a length consistent with gene flow from ∼2000 generations ago (37-86kya), a condition taken from Sankararaman *et al.* [2012].

Applying this method results in a set of 4,437 regions inferred to have originated from Neandertals that covers over a third of the human genome, regions having a median length of 129kb. According to this paper, an individual European contains between 1.07-1.2% Neandertal genome, an East Asian between 1.37-1.4%, an American between 1.05-1.22%, and an African between 0.08-0.34%, notably an order of magnitude less than the former three continents. These

percentages are smaller than those reported in Green *et al.* [2010], but similar to those in more recent work, such as Prüfer *et al.* [2014]. A higher proportion in East Asian as compared with European populations falls in line with some previous publications (Meyer *et al.* [2012]; Wall *et al.* [2013]), and Sankararaman *et al.* [2014] suggest this may be due to smaller population sizes following introgression in East Asians, making selection against introgressed deleterious Neandertal alleles less efficient, which is supported in Juric *et al.* [2015].

A second genome-wide study, Vernot and Akey [2014], used a previously developed summary statistic from Plagnol and Wall [2006], S^*, to identify 15Gb of introgressed sequence (totalling 600Mb of the Neandertal genome) in 379 Europeans and 286 East Asians from the 1000 Genomes dataset. S^* uses modern human genomes - without an archaic genome - to detect haplotypes that are (a) highly divergent within human populations (and therefore with a high T_{MRCA}), and (b) long (∼50kb), and therefore potentially indicative of admixture.

Originally, S^* was used simply to test whether there is evidence for introgression at a locus (Plagnol and Wall [2006]). In Vernot and Akey [2014], it is used to detect haplotypes that are divergent within a population of humans, which contain variants in strong LD, and which do not match a reference population assumed to have no introgression. These factors together mean these haplotypes may be the result of introgression. In this paper, S^* was extended so that it is calculated across individuals, and a hypothesis test added to determine whether any individuals have a large enough value of S^* to reject the null that they do not contain introgressed sequence. We outline its use below.

For every 50kb window, a group of 20 European and East Asian individuals is randomly selected. S^* is calculated separately for each individual across a set of SNPs, J, a subset of the set of all SNPS in a region, V_i. The set of SNPs J all satisfy the conditions that they are variant in a population, and ancestral in the reference population of YRI. Each SNP, j, is labelled 0, 1, or 2, in each individual, i, depending on its genotype in that individual. Each individual and its corresponding genotype labels is taken and a score is given to each SNP dependent upon its LD with the remaining SNPs in J. A score is assigned to a SNP by creating a matrix of distances for an individual, SNP j in individual i's score being the sum of the genotype differences between

it and SNP j in all other individuals, which is termed $d(j, j+1)$).

If a SNP's score exceeds 5, and therefore there is too little LD between all pairs including that SNP across individuals, a score of minus infinity is assigned. Where the score is between 1 and 5, a penalty of -10,000 is assigned, as LD for that SNP is low. Where the distance is 0, and therefore that SNP is in perfect LD, a score of 5,000 is awarded, in addition to the distance in base pairs ($bp(j, j+1)$): this rewards SNPs in high LD more highly if they are further from one another, to help find regions in strong LD.

These scores are summed in each individual to give $S(J)$, given as:

$$S(J) = \sum_{j \in J} \left\{ \begin{array}{ll} -\infty, & d(j, j+1) > 5 \\ -10000, & d(j, j+1) \{1, ..., 5\} \\ 5000 + bp(j, j+1), & d(j, j+1) = 0 \\ 0, & j = max(J) \end{array} \right\} \tag{4.1}$$

S^* is defined for each individual as $S_i = max_{j \subseteq V_i} S(J)$: this highlights SNP pairs with the highest linkage score. A hypothesis test at the 99% level is used to determine whether the S^* score is significant, those deemed so are labelled 'tag SNVs' (Single Nucleotide Variants). Introgressed haplotypes are then located by requiring at least 80% of all tag SNVs to be present on a single haplotype, and the limits of that haplotype delimited by those tag SNVs furthest apart.

Although the method does not require ancient genomes, the candidate regions of introgression were compared to the Altai Neandertal genome and shown to match it significantly more often than would be expected by chance.

We note briefly here that the respective sets of introgressed regions reported in these two papers overlap substantially; this is quantified and explored more thoroughly in the results section below, in relation to our inferred set of introgressed regions.

4.1.2 Adaptive introgression in modern humans, and admixture into Neandertals

Previous to and surrounding the time of publication of Sankararaman *et al.* [2014] and Vernot and Akey [2014] were a number of smaller studies of introgression between Neandertals and humans, most of which study introgression at single loci. Although we do not address the possibility of selection in our work, we note here that some of these studies are examples of potential adaptive introgression, and, interestingly, some categories of functionality tend to arise repeatedly, including both immunity and pigmentation.

Taken from the two studies above (Vernot and Akey [2014] and Sankararaman *et al.* [2014]), Dannemann *et al.* [2016] report that 12% of the top 1% of genes seen in the largest number of Eurasians are related to immunity, which includes 3 toll-like receptors on chromosome 4, TLR1, TLR-6, and TLR-10. TLRs are key components of innate immunity, providing a first line of defence against bacteria, fungi, and parasites. These are thought to be the result of introgression as opposed to Incomplete Lineage Sorting (ILS) due to the presence of the relevant haplotypes almost exclusively in non-Africans, a recombination rate of 1.5-2.4cM/Mb - higher than the genome-wide average, and low diversity within the introgressed haplotypes.

In addition, the first two regions classified as introgressed were $STAT2$ and the OAS gene cluster (Mendez *et al.* [2012, 2013], both on chromosome 12. $STAT2$ is an innate immunity gene which plays a role in interferon signalling pathways, and the OAS cluster encodes proteins involved in the innate immune response to viral infection, and they are of length 8.6kb and 185kb respectively. Both are classed as introgressed because the T_{MRCA} between the Neandertal and non-African haplotypes is substantially shorter than estimates of the divergence time of the two species: 80ky and 125ky respectively. The regions are also both in strong linkage disequilibrium in non-Africans, and $STAT2$ is found more frequently in Papua New Guineans, leading the authors to conclude it may have been positively selected in Melanesians.

With regard to pigmentation, Ding *et al.* [2013] found a ∼200kb region on chromosome 3 (3p21.31) which contains 18 genes, three of which (HYAL1, 2, 3) encode hyaluronoglucosaminidases, which are involved in the cellular response to UVB. It is found at high frequencies in Asian

populations (49.4% in JPT, 66.5% in the Han Chinese), and very low in European populations (0.57% in CEU, 3.7% in FIN). Classification of this region as introgressed required that the reconstruction of a phylogenetic tree connecting modern and archaic human lineages (including both the high coverage Neandertal and Denisovan) showed a cluster between some modern humans with Neandertal and Denisovan before joining the remaining lineages. A weak latitudinal gradient in the frequency distribution of this region is suggestive of adaptations to UV as humans migrated throughout Asia. Additionally, a SNP found at 70% frequency in Europeans while being absent in Asians, BNC2, and which is associated with freckling and skin pigmentation (Jacobs *et al.* [2013]), is included in both sets of introgressed regions from Vernot and Akey [2014] and Sankararaman *et al.* [2014].

There are also some so far stand alone studies relating to particular functions, such as increased susceptibility to Type 2 diabetes in Mexicans and Native Americans (SIGMA [2013]), a beneficial hypoxia response at high altitude in Tibetan highland populations (Huerta-Sánchez *et al.* [2014]), and increased lipid catabolism in Europeans as compared with Asians (Khrameeva *et al.* [2014]). Each had a particular method of identifying the region as potentially introgressed. SIGMA [2013] cite two lines of evidence with regard to a 5-SNP risk haplotype, 73kb long, containing 4 missense substitutions and one synonymous amino acid substitution, seen at very low frequency in European and African samples, at ∼10% frequency in East Asians, and up to ∼50% frequency in Native Americans. The four missense mutations are all found in SLC16A11, a solute carrier. The risk of Type 2 Diabetes was shown to be increased by ∼20% for each haplotype copy. The first was that the human risk haplotype has a T_{MRCA} of 250,000 years with the Neandertal sequence, as compared with a T_{MRCA} of 677,000 years between non-risk haplotypes and the Neandertal, and a clade forms between the Neandertal and the affected modern humans. The second was that the haplotype is of unexpectedly long length given the expectation of recombination over 9,000 generations since the human-Neandertal split. Huerta-Sánchez *et al.* [2014] report a very strong example of selection at the ∼33kb region surrounding the *EPAS1* locus in Tibetan highlanders which confers a severe reduction or lack of haemoglobin level increase in high altitudes: an advantageous trait. The region is concluded to be introgressed from Denisovans because (a) there is significantly lower divergence between the

Denisovan haplotype and that which is most common in Tibetans than is expected of human-Denisovan comparisons, (b) the amount of divergence between Tibetans and Denisovans makes sense with the likely time since introgression, and (c) it shows highly significant values of both the D-statistic (see Chapter 2 for explanation and examples) and S^*. Lastly, Khrameeva *et al.* [2014] found a 3× increase of D-statistic values in genes involved in lipid catabolism across 23 genomic regions, in all European populations as compared with Asians and Africans. This in itself is suggestive of introgression, and the differences in lipid concentrations which occur as a result of these differing metabolic speeds between populations may confer an advantage in cold temperatures which would have had to be endured by migrants of western Asia and Europe.

It is certainly interesting to consider the biological functions of the aspects of their genome that humans have obtained from Neandertals, who existed in Eurasia substantially before we did, and who therefore may have had some beneficial adaptations to conditions present there, despite their small effective population size. The converse of this is that modern humans are likely to have also gained a larger number of aspects of the Neandertal genome which impact negatively. For example, Sankararaman *et al.* [2014] noted deserts of Neandertal ancestry which correlate with high gene density as measured using the B-statistic. The same study showed Neandertal ancestry on the X chromosome to be 20% of that seen on the autosomes; it is known that a large proportion of genes involved in male fertility exist on the X chromosome, and given that genes in introgression deserts were shown to be disproportionately expressed in testes tissue, it was concluded that Bateson-Dobzhansky-Muller incompatibilities (where two alleles have evolved differently in separate populations and are incompatible when brought back together through hybridization) causing a reduction in male hybrid fertility may have contributed to negative selection on Neandertal introgressed alleles. However, Juric *et al.* [2015] suggest that selection against Neandertal alleles in humans is due to a larger effective population size in humans as compared with Neandertals, where selection is more effective. They suggest that alleles which had been effectively neutral in Neandertals and segregating at high frequency, were weakly deleterious in humans and were then selected against. Harris and Nielsen [2016], in line with this conclusion, used simulations to suggest that a high mutational load in Neandertals due to their small effective population size, then introgressed into modern humans with a much

larger effective population size and therefore subject to stronger purifying selection, may explain the dearth of introgression in genic regions in humans, rather than requiring the invocation of epistatic incompatibilities. We touch on this matter in this chapter, because we also note a negative correlation between recombinationally cold regions and genetic material introgressed from the Neandertal.

Most studies on the topic of admixture between humans and archaic species have focused on the existence of introgression from Neandertals or Denisovans into humans. However, very recently, a paper was published (Kuhlwilm *et al.* [2016]) describing evidence for human introgression into the ancestors of the Altai Neandertal, and not into either of two European Neandertals (El Sidrón, Vindija), or the Denisovan, suggesting that this introgression occurred after the divergence between these Neandertal lineages approximately 110kya (Kuhlwilm *et al.* [2016]). *ARGWeaver* (Rasmussen *et al.* [2014]) was used to build ancestral recombination graphs between six African genomes from the San, Mbuti, and Yoruban populations, and the two archaic genomes, at particular regions. These regions showed an archaic haplotype that was similar to an African haplotype and divergent from the second archaic individual. This resulted in Ancestral Recombination Graphs (ARGs) where the archaic haplotype(s) coalesced with the African subtrees earlier than with the other archaic individual. The age distribution of these coalescence times was markedly different between the Neandertal and Denisovan, with significantly more Neandertal than Denisovan haplotypes coalescing with the African populations at more recent dates, potentially suggestive of admixture. A potential confounder of this conclusion is the Denisovan containing introgressed material from an unknown archaic (Meyer *et al.* [2012]), thus making it more divergent from human populations. However, this is more likely to affect older haplotypes than those reported here as evidence of introgression from humans into archaic humans, which range between 100-230ky old, so may be separable. We address this intriguing result later in the chapter.

4.1.3 In this chapter

Before we explore the genealogical trees detailing the relationships between the Altai Neandertal and 14 modern human populations from the 1000 Genomes Project, we first use three-population simulations between the Altai Neandertal, an African population (YRI), and a European population (GBR), to ascertain that *CEPHi* accurately calls admixture where it exists, and does not call admixture where it does not exist. With this confirmed, we use the trees to give a definition of an admixed region based on the comparison of the minimum estimated coalescence time between the Altai Neandertal and an African population, and the Altai Neandertal and a non-African population. Using this to define sets of introgressed regions for each of 14 human populations from the 1000 Genomes Project, we are able to find clues as to the sequence of migration and admixture events in our species' evolutionary past. First, we quantify the total amount of introgression per population. We then make comparisons between individual and continental populations, focusing on identifying shared regions between European and Asian populations, as this allows us to examine the question of when admixture occurred between Neandertals and modern humans. Sets of shared regions within Eurasians imply that some admixture likely happened before the European and Asian human population split; those regions that are specific to a continent or population may be the result of separate instances of admixture. We also search for the source of introgression seen in African populations by comparing their introgression sets with Asian and European populations.

We then compare our sets of introgressed regions with those from two previously published sets from Sankararaman *et al.* [2014] and Vernot and Akey [2014], obtained using methods distinct from one another and from ours. Furthermore, again using information from the genealogical trees, we are able to provide insights into the direction of introgression. A recent paper concluded that there is evidence for introgression from humans into Neandertals (Kuhlwilm *et al.* [2016]). Given that there is substantial overlap between our admixed regions and those identified as having introgressed into Neandertals, we address this question by briefly examining some properties of the genealogies for these regions.

4.2 Methods

Here we describe the methodology we employ in this chapter to explore the information contained in the genealogical trees produced by *CEPHi* in Chapter 3. In summary, we first use simulations to consolidate our conclusions in the previous chapter. We then provide a definition of admixture which uses estimated coalescence times from genealogical trees linking modern humans with the Altai Neandertal. This allows us to define sets of population-specific and continent-specific introgressed regions. We compare these sets within and between continents, as well as examining the overlap between them and two sets of Neandertal-introgressed regions published in 2014. Additionally, we provide online lists of functional elements contained in our introgressed region sets.

4.2.1 Simulating admixture and non-admixture scenarios

We first use simulations to test the accuracy of *CEPHi* at (a) estimating speciation times between humans and Neandertals when admixture has not occurred, and (b) correctly calling admixture where it exists. We simulate segregating sites data with the coalescent simulator *scrm* (Sequential Coalescent with Recombination Model) (Staab *et al.* [2015]), for input into *CEPHi*. *Scrm* is open-source software which allows for the fast simulation of long genomic sequences under an accurate representation of the coalescent with recombination, using a large range of user-specified parameters making up specific demographic scenarios, including (but not limited to) genomic region length, mutation and recombination rates, and the number of segregating sites. We define a demographic scenario with three populations: one African (YRI), one European (GBR), and one Neandertal. To construct this scenario, we use output from two previous runs of *CEPHi* - one between YRI and GBR (provided by Marie Forest), the second between YRI and the Altai Neandertal (taken from Chapter 3). Starting at the present and looking backwards in time, we use population size histories and population split times from the YRI-GBR analysis for YRI and GBR up to and including the first population split 79,741 years ago. Looking backwards in time from this split, all split times and population sizes are

Epoch (yrs ago)	YRI	GBR	YRI-GBR	Nean	HN Anc
0-5,158	25,010	29,666	-	-	-
-10,537	22,610	23,669	-	-	-
-21,994	26,017	17,858	-	-	-
-34,449	28,644	13,107	-	-	-
-44,800	33,314	8,936	-	-	-
-47,992	33,314	8,936	-	3,871	-
-78,722	38,738	6,316	-	3,871	-
-79,740	30,681	18,903	-	3,871	-
-96,750	-	-	42,690	3,871	-
-196,279	-	-	47,825	5,075	-
-360,657	-	-	53,937	6,687	-
-632,139	-	-	53.552	8,902	-
-712,936	-	-	46,799	7,128	-
-1,080,512	-	-	-	-	45,168
-1,821,032	-	-	-	-	22,533
-3,044,056	-	-	-	-	17,139
-5,063,968	-	-	-	-	20,878
-500,000,000	-	-	-	-	41,461

Table 4.1: Simulations: describing epochs and population sizes corresponding to the 3-population scenario given in Figure 4.1. Epoch time in the left hand column indicates the upper bound of the epoch. YRI and GBR are individual populations existing more recently than the YRI-GBR split. YRI-GBR indicates the joint population further back in time from this split. HN Anc = population ancestral to humans and Neandertals. Dashes indicate lack of existence of that population in that epoch.

taken from the YRI-Neandertal analysis. This includes (comprehensively) human population sizes (now a combined population of YRI and GBR), Neandertal population sizes, the human-Neandertal split time of 712,936 years, and the ancestral human-Neandertal population sizes (now a combined population) are taken from the YRI-Neandertal analysis. All timings of population splits and sizes are specified in Table 4.1, and Figure 4.1 describes this demographic scenario visually.

We use these timings and population sizes directly. Furthermore, we match our simulated data to the data from the 1000 Genomes Project in a number of ways. Firstly, we simulate the same number of haplotypes per population as exists in this data (totalling 356: 176 YRI, 178 GBR, 2 Neandertal). We use a mutation rate (θ) that varies according to region size, and which equates to a μ of 0.5×10^{-9}bp/yr (so for a region of 30000bp, because $\theta = 4N\mu$, we halve this to result in a θ of 7.5 - using a generation time of 25 - to match our real data). We also match the data used in our *CEPHi* analyses by performing the simulations across 8881 regions (the same number as exist in the YRI-Neandertal analysis), and by specifying regions lengths matching

those across each of the 8881 regions in the set.

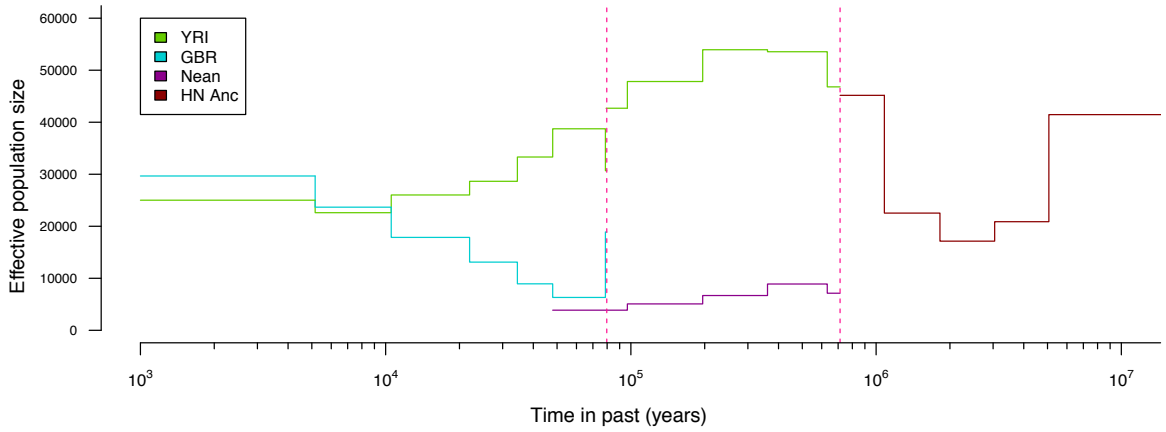


Figure 4.1: Simulating an admixture scenario with three populations: YRI, GBR, and the Altai Neandertal. We combine the output of two previous runs of *CEPHi*: YRI-Neandertal and YRI-GBR, to produce a 3-population scenario. From the present day, looking backwards in time, YRI (green) and GBR (blue) exist as separate populations up until 79,741 years ago (lefthand vertical dashed red line), where they become a joint population (YRI-GBR). The Neandertal (purple) comes into existence at 44,800 years ago, and at 712,936 years ago (righthand vertical dashed red line), joins with the YRI-GBR human population to form the (brown) human-Neandertal ancestral population (HN Anc). YRI and GBR individual population sizes and the YRI-GBR split time come from the YRI-GBR analysis. YRI-GBR joint population sizes, Neandertal population sizes, ancestral human-Neandertal population sizes, and the human-Neandertal population split time come from the YRI-Neandertal analysis. This demographic scenario was input into *scrm*, as is detailed in the main text, to produce segregating sites data for each of these three populations. This was separated into two population pairings of YRI-Neandertal and GBR-Neandertal and run through *CEPHi*, to investigate whether this simulated demographic scenario closely matches reality (as represented by datasets created from 1000 Genomes Project human data, and the Altai Neandertal genome).

Each simulation creates a set of segregating sites across the specified number of haplotypes for a single region, and iterates over the 8881 regions. We use an island model with three islands, each containing one population: YRI, GBR, and Neandertal. Recombination is set to zero as we use recombinationally cold regions within which we do not expect recombination to occur. We move 2% (an upper bound of admixture so far reported in Prüfer *et al.* [2014]) of the GBR population into the Neandertal population at 44,800 years in the past. The full command line used is given in Appendix C.

We then input the simulated segregating sites data from this scenario into *CEPHi*. *CEPHi* uses two populations, so we split the haplotypes into two analyses: the first between YRI and Neandertal (a no-admixture scenario), the second between GBR and Neandertal (an admixture

scenario). We use matching command lines for *CEPHi* as used for the analyses in Chapter 3 under two conditions for each. The first has a wide boundary for population split times (between 44,800 and 1,400,000 years ago), the lower end of which includes the age of the Neandertal fossil), the second with a much smaller and more recent boundary for population split times (between 22,400 and 50,400 years ago), which allows haplotypes to coalesce very early - more recently than the real population split times - thereby enabling us to identify potentially admixed regions. To explore robustness, we run *CEPHi* three times: once with 12 equally sized epochs, once with 20 equally sized epochs, and once with 20 epochs with shorter epochs recently in the past (to account for the fact that more coalescences happen here relative to further back in the past).

We then compare population split times, population size histories, and coalescence time distributions between the real and simulated data.

4.2.2 Defining admixture

Coarsely, admixture refers to two distinct (by some definition) populations interbreeding. Here, we use genealogical trees to define it more precisely using coalescence times between humans and the Altai Neandertal, which allows us to identify genomic regions introgressed from the Neandertal into one or many human individuals and populations. We use heatmaps to compare the coalescence times between non-African populations (European, Asian, American) and the Altai Neandertal, against those of an African population from Yoruba in Nigeria (YRI), which is thought to contain negligible to no admixture with Neandertals (Dobon *et al.* [2015]). From these heatmaps, and the inflexion point on the coalescence time distributions in Figure 3.7 from Chapter 3, we specify a definition of admixture for single genomic regions. Using this definition, we plot the coalescence time distributions per population, separating regions into admixed and non-admixed, according to the pairwise values of coalescence times.

4.2.3 Comparing sets of introgressed regions between populations

We first present the total amount of introgression from the Altai Neandertal present in each of 13 populations (YRI is excluded from this as it is used as a reference set), both as a total across the genome, and the amount per chromosome. This is given as a set of regions which are present in any individual (or number of individuals) across the genome in a population.

We then compare the sets of admixed and non-admixed regions between European and Asian populations, using two heatmaps. This not only shows which introgressed regions are seen in the respective continents and which are seen only in one, but also gives us information regarding the time at which admixture may have occurred between Neandertals and Eurasians. A large amount of sharing between sets of admixed regions is suggestive of admixture between humans and Neandertals before the European and Asian population split. Sets of introgressed regions not shared between Europeans and Asians by contrast are suggestive of separate admixture events or other evolutionary processes. We group all regions together for five European populations (GBR, CEU, TSI, FIN, IBS) and three Asian populations (CHB, CHS, JPT), and then match regions between the two continents to produce a set of shared regions. Each region within a continental set is defined as admixed if it is defined as such in at least one of the populations, and non-admixed if no population defines it to be admixed. After removing any regions with coalescence times with the Neandertal above 2mya, we then define two subsets. The first is constituted of those regions defined as admixed in either European, or Asian, or both continental populations. The second is constituted of those regions not defined as admixed in either European, Asian, or both continental populations. Note that these subsets are not fully mutually exclusive; if, for example, a region is defined as admixed in Europe and not in Asia, this will be plotted in both heatmaps. We plot the minimum estimated coalescence time between the continental population and Neandertal against the equivalent for the remaining continent.

Building on these heatmaps, we then take only those admixed regions in any of the five European populations, and plot the minimum coalescence time with the Neandertal, against the coalescence time with the Neandertal in the Asian group, disregarding whether these are clas-

sified as admixed in any of the individual Asian populations. We also give the complementary version. This further highlights the admixed regions shared between the two continents, and those which are specific to each.

We then compare sets of admixed regions between all pairs of individual populations. We do this by using our definition of an admixed region: that the mean of the minimum coalescence times with the Altai Neandertal across two trees in a particular population is $\leq$300kya, and the equivalent mean for YRI is >600kya. We reduce comparison sets between a pair of populations (for example GBR and CHB) to only those that both analyses include (slightly different sets are used in *CEPHi* due to filtering procedures). For each population in turn we then take a simple proportion of the number of admixed regions in that population which are also classed as admixed in the second population, of the total number of admixed regions in the first population. This gives us an asymmetrix matrix of proportions of shared admixed regions, dependent upon the initial size of the first population's set.

We then perform tests of correlation between sets of admixed regions, for every pair of populations. We construct 2×2 contingency tables with the first population as rows, and the second as columns, using 'admixed' and 'non-admixed' for both. We calculate the odds ratio for the population pair to give a measure of enrichment of one set of admixed regions, given the other; essentially this number gives us a measure of how much similarity we see between two sets of admixed regions. We then perform Fisher's Exact tests to test the significance of the association, where a p-value <0.05 indicates there is a significant dependence at the 95% level between the sets of admixed regions between populations.

We address the small amount of Neandertal admixture seen in the African populations (LWK and ASW), and are able to infer, from the pattern of shared regions between Eurasian and African populations, information about instances of back migration from Asia and Europe into Africa.

4.2.4 Examining anomalous regions with intermingled Neandertal haplotypes

We investigate regions, whether admixed or not, in which the genealogy contains two coalescences between YRI and the Neandertal. Normally, we expect a single coalescence time between the Altai Neandertal and the set of modern humans in the genealogical tree. This is because the Neandertal has high levels of homozygosity, meaning its two haplotypes ordinarily coalesce together before meeting the modern human haplotypes, in any population. Where we see two coalescence times in a non-African population and a single coalescence time in YRI, we might expect this to be explicable via admixture. Where two coalescence times exist in YRI, the explanation is less clear. To explore this, we first take counts of the number of trees we see in a GBR-YRI comparison with one and two coalescence times for each population. We then take the group with two coalescences in both GBR and YRI as it is a substantial number and examine the timings of the first and second coalescences. This allows us to address the possibility of admixture from humans into Neandertals, as was recently put forward in Kuhlwilm *et al.* [2016], as well as the possibility of a complex separation event between humans and Neandertals.

4.2.5 Comparing our introgressed regions with previously published sets

We compare the set of introgressed regions we have inferred for each of the populations to those reported in Vernot and Akey [2014] and Sankararaman *et al.* [2014], as discussed in the introduction to this chapter. These include all European and all Asian populations.

A chromosome plot displays our admixed regions from GBR as compared with our recombinationally cold regions across the whole genome, alongside both sets of GBR admixed regions from Vernot and Akey [2014] and Sankararaman *et al.* [2014].

We then compare the sets of admixed regions using heatmaps, where we plot the GBR-Neandertal coalescence time against the YRI-Neandertal coalescence time across regions which satisfy the criterion of being admixed in either or both sets. Specifically, we take our set of recombinationally cold regions, and comparing with each of the sets from Vernot and Akey [2014] and Sankararaman *et al.* [2014] in turn, we create four sets: those that are classified as admixed

in both, those that are admixed in neither, those that are admixed in our set and not theirs, and those that are admixed in their set and not ours. This allows us to observe the type of regions each of the methods classifies as admixed, and those which are not included by respective methods.

We then show, per population, the proportion of our admixed regions that are covered by the admixed regions of Vernot and Akey [2014] and Sankararaman *et al.* [2014] respectively, with barplots showing the total number of admixed regions, the number overlapped by the comparison set by >50%, >75%, and by 100%.

4.3 Results

4.3.1 Simulating admixture and non-admixture scenarios

We simulated segregating sites data across a set of 8881 genomic regions from a three population scenario involving two human populations (GBR, YRI) and a single Neandertal individual, as described in Section 4.2.1. We divided this data into two groups (YRI-Neandertal and GBR-Neandertal) for analysis with *CEPHi* analyses, and were then able to compare the results with those from our real data (taken from the 1000 Genomes Project and presented in the previous chapter). For clarity, then, for each of YRI and GBR, we compare *CEPHi* results from Chapter 3 (our real data) with *CEPHi* results from data simulated using *scrm*.

We present the parameters used to simulate the segregating sites data (the 'simulated data truth'), the *CEPHi* results using this simulated data (the 'simulated data *CEPHi*'), and the *CEPHi* results using real data (the 'real data *CEPHi*) together in Figures 4.2 and 4.3. For each population pairing, we ran *CEPHi* three times using the simulated data: once with 12 epochs to match our analyses in the previous chapter, and twice with 20 (one with *CEPHi*-defined epoch sizes, one with shorter epochs more recently), to match the GBR-YRI analysis which provided us with the YRI population size histories in the more recent history.

In Figures 4.2 and 4.3, we see the population size histories and split times for YRI-Neandertal

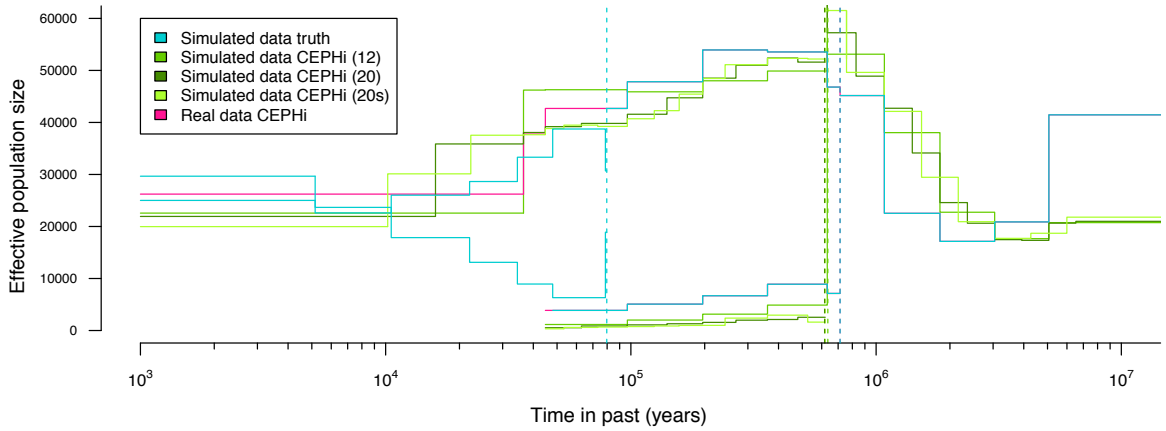


Figure 4.2: Comparing *CEPHi* population size histories from real and simulated datasets (as defined in Section 4.2.1 and Table 4.1: YRI. In turquoise we plot the population sizes and split times used to generate the simulated data for reference, as given in Table 4.1 and Figure 4.1. In green we show the results using simulated data, output from *CEPHi* across three separate runs (midgreen: 12 epochs, yellow-green: 20 epochs, dark green: 20 epochs with shorter recent epochs (20s)). In magenta we show the population size histories and split times from *CEPHi* using real data (matching the turquoise from the YRI-GBR split looking backwards in time). The vertical lines give population split times with the same colour schemes: the YRI-GBR split is shown in turquoise at 79,741ya, and the YRI-Neandertal split at 712,936ya; the YRI-Neandertal split times from *CEPHi* using three sets of simulated data are given in the corresponding shades of green; YRI-Neandertal split time from the *CEPHi* analysis with real data at 712,936ya in magenta, matching the turquoise.

and GBR-Neandertal respectively. All values of these parameters are those inferred by *CEPHi*, other than those of the 'Simulated data truth' which gives the parameters we input into *scrm* to create our simulated data.

We can see in Figure 4.2 that population histories match closely between the simulated truth, the *CEPHi* results using the simulated data, and the *CEPHi* results using the real data. Split times match reasonably closely, with YRI-Neandertal split times inferred to be 635,516ya, 620,144ya, and 616,672ya for the three simulated sets of data respectively (12 epochs, 20 epochs, and 20 epochs with shorter recent epochs), where the date is 712,936ya in the simulated truth. Neandertal population sizes are shown to be slightly smaller in the simulated datasets in comparison to the simulated truth, but population sizes match closely overall, confirming that *CEPHi* works well.

Figure 4.3 again shows very similar population histories between the simulated truth and the simulated data as analysed by *CEPHi*. We see a large difference in the YRI-Neandertal and

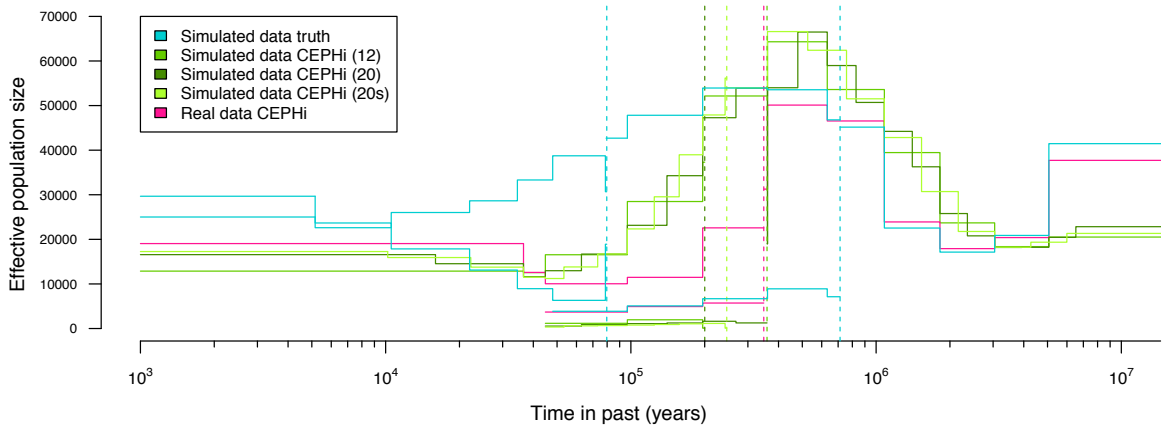
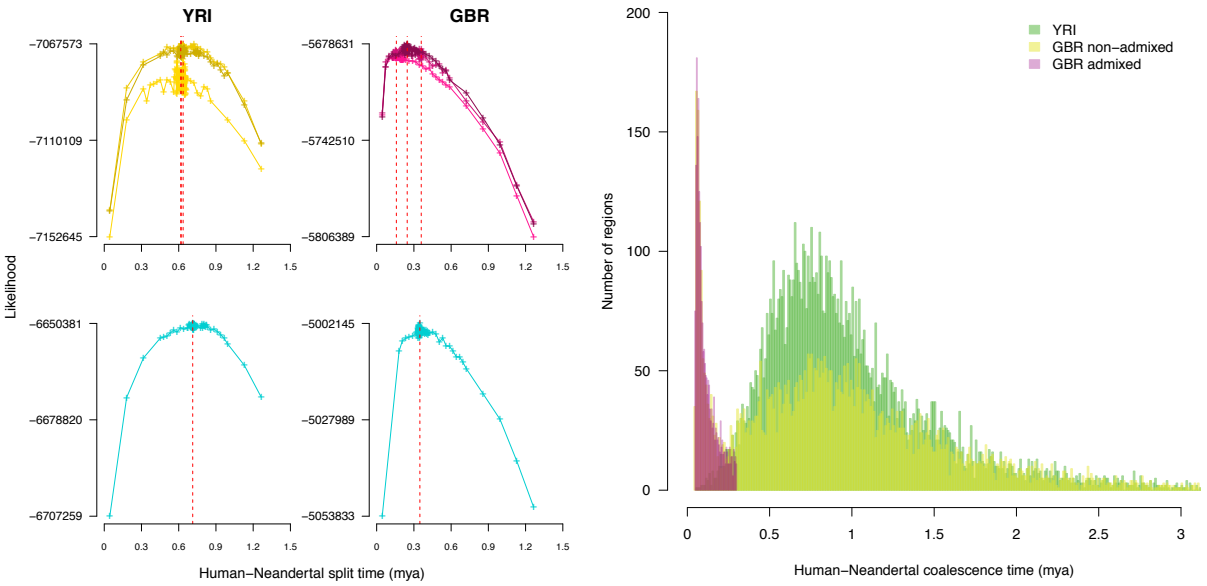


Figure 4.3: Comparing *CEPHi* population size histories from real and simulated datasets (as defined in Section 4.2.1 and Table 4.1: GBR. In turquoise we plot the population sizes and split times used to generate the simulated data for reference, as given in Table 4.1 and Figure 4.1. In green we show the results using simulated data output from *CEPHi* across three separate runs (midgreen: 12 epochs, yellow-green: 20 epochs, dark green: 20 epochs with shorter recent epochs (20s)). In magenta we show the population size histories and split times from *CEPHi* using real data. The vertical lines give population split times with the same colour schemes: the YRI-GBR split is shown in turquoise at 79,741ya, and the YRI-Neandertal split at 712,936ya; the GBR-Neandertal split times from *CEPHi* using three sets of simulated data are given in the corresponding shades of green; GBR-Neandertal split time from the *CEPHi* analysis with real data at 348,264ya in magenta.

GBR-Neandertal population split times, which is explicable through admixture, as was simulated to exist in the GBR dataset. The GBR-Neandertal split time is seen at 348,264ya, and the simulated GBR-Neandertal split times are seen at 200,088ya (12 epochs), 359,072ya (20 epochs), and 246,008ya (20 epochs with shorter epochs more recently). The difference between these split times and the YRI-Neandertal split time of 712,936ya simply shows that admixture is present in the GBR datasets, both real and simulated, and we see some average of the true human-Neandertal split date and the admixture time, just as we did in Chapter 3. Overall, comparisons between the GBR and YRI analyses for the simulated and real data analysed by *CEPHi* match well.

In Figure 4.4 we compare the MLE searches for population split times between YRI-Neandertal and GBR-Neandertal using real and simulated data. It is clear that the searches follow similar paths in both the YRI and GBR cases, and we can see the variation in split times inferred using the simulated data for GBR. We also present the estimated coalescence times between GBR and the Altai Neandertal using simulated data when the split time is set to zero, separating



(a) MLE searches for simulated and real data

(b) Human-Neandertal coalescence time distributions

Figure 4.4: Subplot (a) shows the maximum likelihood estimate searches for simulated population split times in comparison to those estimated by *CEPHi* using real data. The MLE searches for simulated data (with differing numbers and separations of epochs) are shown on the top row, and for comparison, the MLE searches using real data are given on the bottom row. For the simulated data (top row), the three analyses with different epoch numbers and lengths are given: pale yellow/purple for 20 epochs with shorter recent epochs, mid yellow/purple for 12 equally spaced epochs, and dark yellow/purple for 20 equally spaced epochs. MLEs for split times for each analysis are given by vertical dotted lines. Subplot (a) is separated into YRI (left column) and GBR (right column). Subplot (b) shows human-Neandertal coalescence time distributions for simulated YRI and GBR data. YRI-Neandertal coalescence times are shown in grass green, and we separate the GBR-Neandertal coalescence time distribution into those regions we define as non-admixed and admixed: non-admixed are shown in pale green, admixed in purple.

admixed regions from non-admixed regions, and including the full distribution of YRI minimum coalescence times with the Altai Neandertal for comparison. We find the same clear separation in distribution as is seen in the real data in Figure 4.8.

Figure 4.5 shows the comparison of GBR-Neandertal coalescence times with YRI-Neandertal coalescence times for the simulated data. As is seen with the real data in Figures 4.6 and 4.7, we have a set of regions surrounding the diagonal: the population split time used for the simulations is seen clearly as this set is centred around 700kya (the estimated split date using real data is 712kya). The admixed regions form a set that is significantly larger than in the equivalent heatmap for the real data. This may be because we simulated an upper bound on admixture that has so far been reported in the literature (between 1-2%).

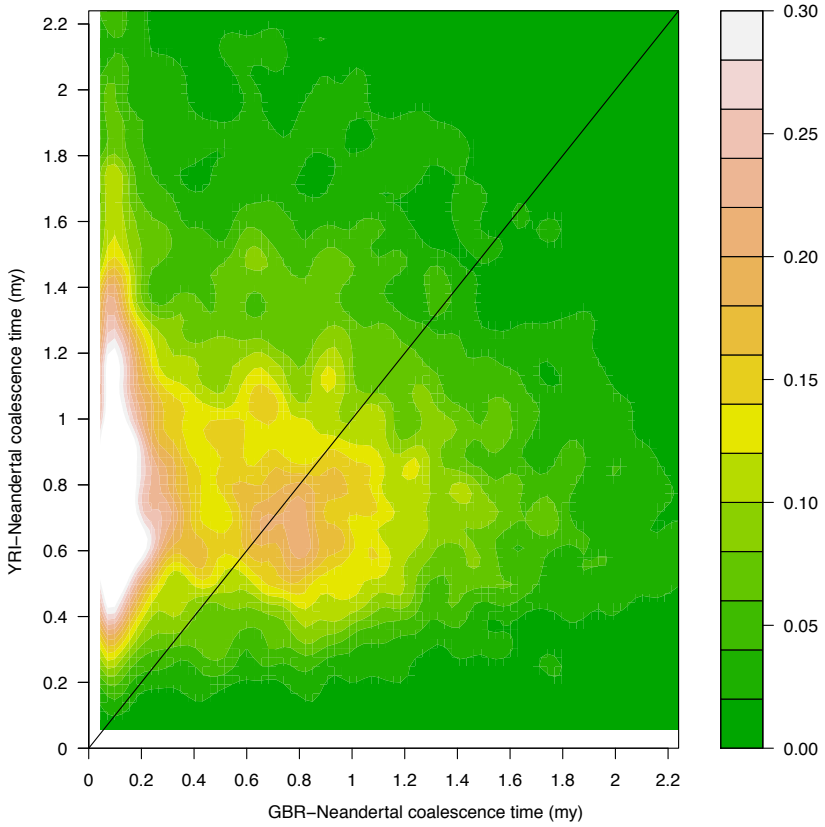


Figure 4.5: Comparing GBR-Neandertal and YRI-Neandertal coalescence times: simulated data. Coalescence time with the Neandertal (millions of years) of all regions in GBR is plotted against the YRI-Neandertal coalescence time for that region, to highlight combinations of coalescence times that are seen frequently. Dense regions (as measured by the sidebar to the right of the plot) are seen on the diagonal and most notably on the left hand side.

If we were to define admixture using this plot, we might look for regions with a non-African-Neandertal coalescence time of $\leq$300kya, matching analyses completed with real data, however our second condition, that the YRI-Neandertal coalescence time be >600kya might be reduced to $\sim$300-400kya.

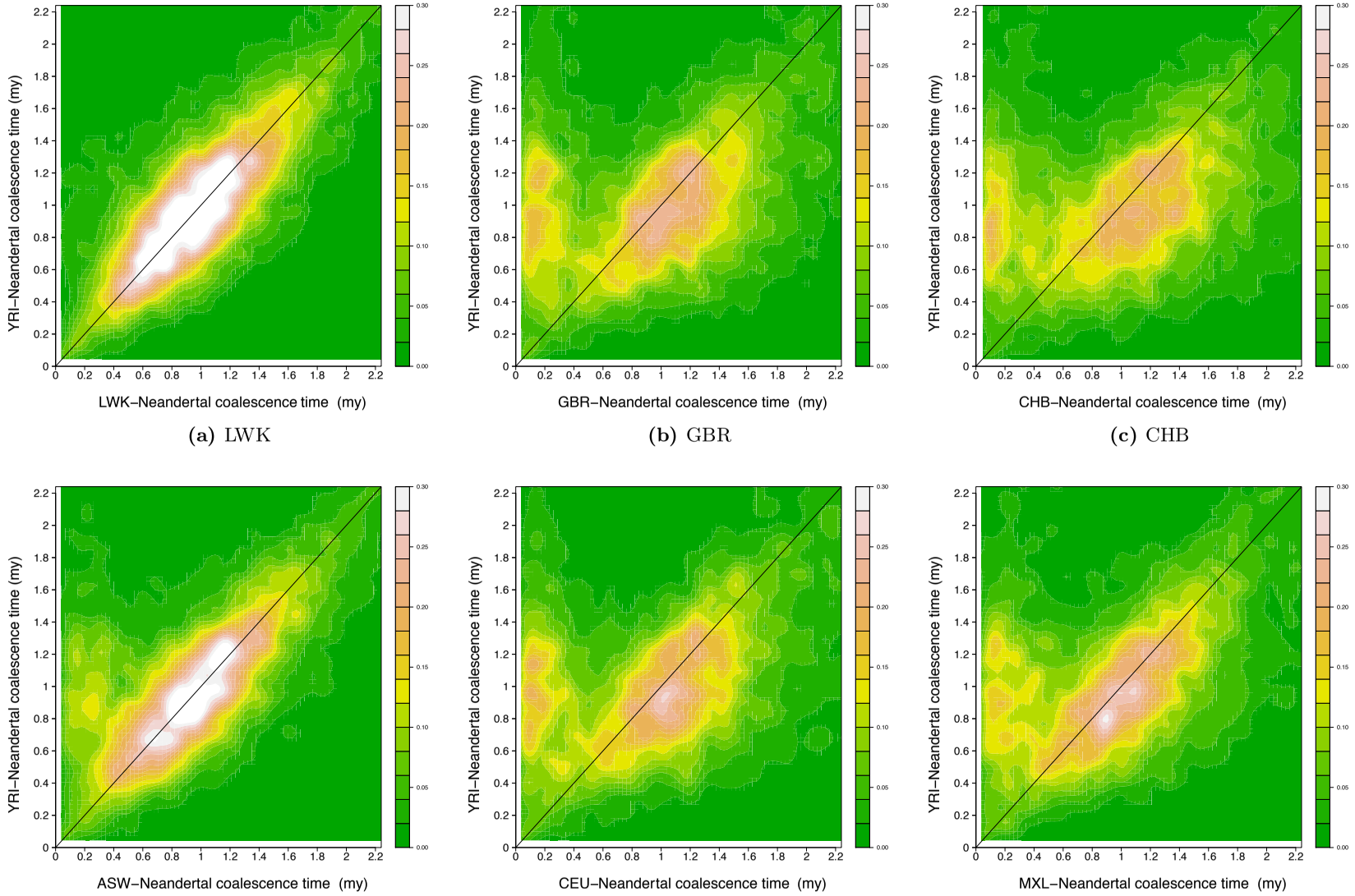
4.3.2 Defining admixture using pairs of estimated coalescence times

The heatmaps in Figures 4.6 and 4.7 provide us with a visual depiction of admixture. Figure 4.6 uses the Yoruba from Ibadan, Nigeria as a reference population which we assume to contain no introgressed regions from Neandertals. Thus, we expect a coalescence time between a genomic region in YRI and the same region in the Neandertal to be at least at long as the

split time between humans and Neandertals, which we believe to be ∼700kya. We pair each of these YRI-Neandertal coalescence times with the equivalent coalescence time for that region between another human population and the Neandertal; we are comparing actual coalescence time estimates for individual regions between populations, and the sharing of admixture events at the haplotype level. Those regions on or close to the diagonal show a matching or similar coalescence time with the Neandertal for both human populations. Above the diagonal are those regions showing a shorter coalescence time between the non-YRI population and the Neandertal than we see between YRI and the Neandertal. We suggest that the regions to the extreme left of the plot, separated from the main body of regions, are likely to be indicative of admixture from Neandertals into the relevant human population.

Figure 4.6 shows there to be significant portions of the genome displaying short non-African-Neandertal coalescence times with correspondingly long YRI-Neandertal coalescence times, strongly suggestive of admixture between Neandertals and all non-African populations considered here. LWK is thought to be one of the least admixed African populations after YRI, and our analysis supports this - there is substantial correlation between the coalescence times for LWK-Neandertal and YRI-Neandertal, forming a single body along the diagonal. This correlation, although present, is much weaker in non-African populations, and is indeed slightly right-shifted. Importantly, for each of the non-African populations, we see patches of high intensity falling far to the left of the diagonal, indicating large numbers of regions where the coalescence times are far shorter between the Neandertal and the haplotypes from the population in question than they are between the Neandertal and YRI haplotypes. Although varying in extent and intensity between populations, we can identify these putatively admixed regions to typically fall between ∼0-300,000 years on the x axis, and greater than 600,000 years on the y axis. We use this reference to these heatmaps to define admixture as follows, where all other regions are defined as non-admixed:

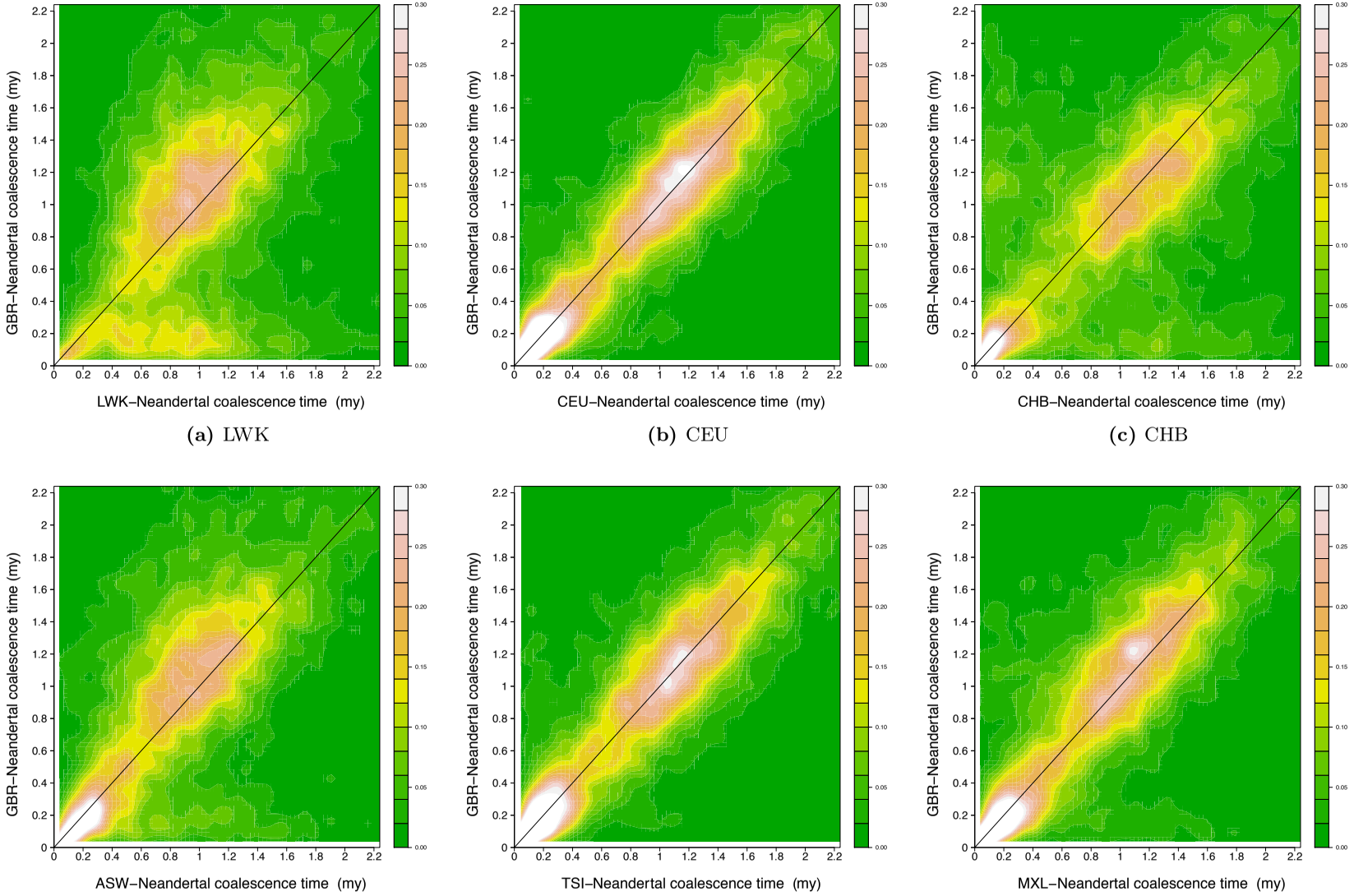
- For a single region in a population to be classified as admixed, we require that averaging over two trees, it shows a coalescence time between non-African populations and Neandertals of $\leq$300,000 years, and a coalescence time of $>$600,000 years between Yorubans



(a) LWK **(b)** GBR **(c)** CHB

(d) ASW **(e)** CEU **(f)** MXL

Figure 4.6: Visualising admixture with heatmaps using YRI as the baseline population (which we assume contains negligible to no introgression from Neandertals). Coalescence time with the Neandertal (millions of years) of all regions in the named population is plotted against the YRI-Neandertal coalescence time for that region, to highlight combinations of coalescence times that are seen frequently. Shared non-admixed regions are seen on the diagonal in all cases, and introgressed regions along the y-axis. African populations are shown in the first column, European in the second, and Asian and American in the third.



(a) LWK **(b)** CEU **(c)** CHB

(d) ASW **(e)** TSI **(f)** MXL

Figure 4.7: Visualising admixture with heatmaps using GBR as the baseline population (which contains a set of admixed regions from Neandertals). Coalescence time with the Neandertal (millions of years) of all regions in the named population is plotted against the GBR-Neandertal coalescence time for that region, to highlight combinations of coalescence times that are shared. Clear bimodality is seen: sets of shared admixed regions in the bottom left corner, and sets of shared non-admixed regions further along the diagonal. African populations are shown in the first column, European in the second, and Asian and American in the third.

(YRI) and Neandertals. We use the minimum coalescence time with Neandertals for each tree to calculate this mean.

In addition, we note that there to be a large number of regions falling below the diagonal in all non-African cases. This is seen most clearly in Asian populations. Having more regions with a longer coalescence time between the Neandertal and, for example, the Chinese population from Shanghai (excepting those regions on the far left that we believe to be the result of admixture), than is seen between the Neandertal and the YRI population is perhaps immediately counter to expectation, but may be explicable with recourse to these populations' histories. Firstly, we know that the human population experienced a severe bottleneck on exiting Africa, as seen in Chapter 3. For an individual population, a bottleneck pushes forward the time at which individuals coalesce with one another, as we see fewer lineages further back in the past. Conversely, fewer lineages being present further back in the past mean that bottlenecks push backward the times at which individuals from one population coalesce with those of another. In the case of Asian populations it is thought that there has been a second bottleneck after the out of Africa event (Conrad *et al.* [2006]; Keinan *et al.* [2007]), meaning we see the result of this to a larger extent in the coalescence times between Asian populations and the Neandertal than we do for the European populations.

Figure 4.7 gives a second set of equivalent plots using GBR as the reference population, in order to highlight the fact that many regions that look potentially admixed in Figure 4.6 are held in common between various non-African populations, shown by the hot region in the bottom left corner of each plot. As may be expected, there are visibly more regions shared between GBR and other European populations compared with GBR and non-European populations (shown here are CHB and MXL). Notably, there is again a smaller set of regions shared between GBR and both African population: LWK and ASW. This is delineated more specifically below.

Figure 4.8 highlights the distributional differences between admixed and non-admixed regions, as per our definition. We show the distribution of coalescence times between the Altai Neandertal and YRI in pink at the back of each plot for an easy comparison. The regions classified as non-admixed in the population in question are given in bright blue, and admixed regions are

given in purple. We see clearly the distinct distributions for non-admixed and admixed regions across all populations. We note the close similarity between the YRI and LWK distributions, as well as the much larger number of admixed regions in the non-Africans compared with the two African populations (ASW, LWK).

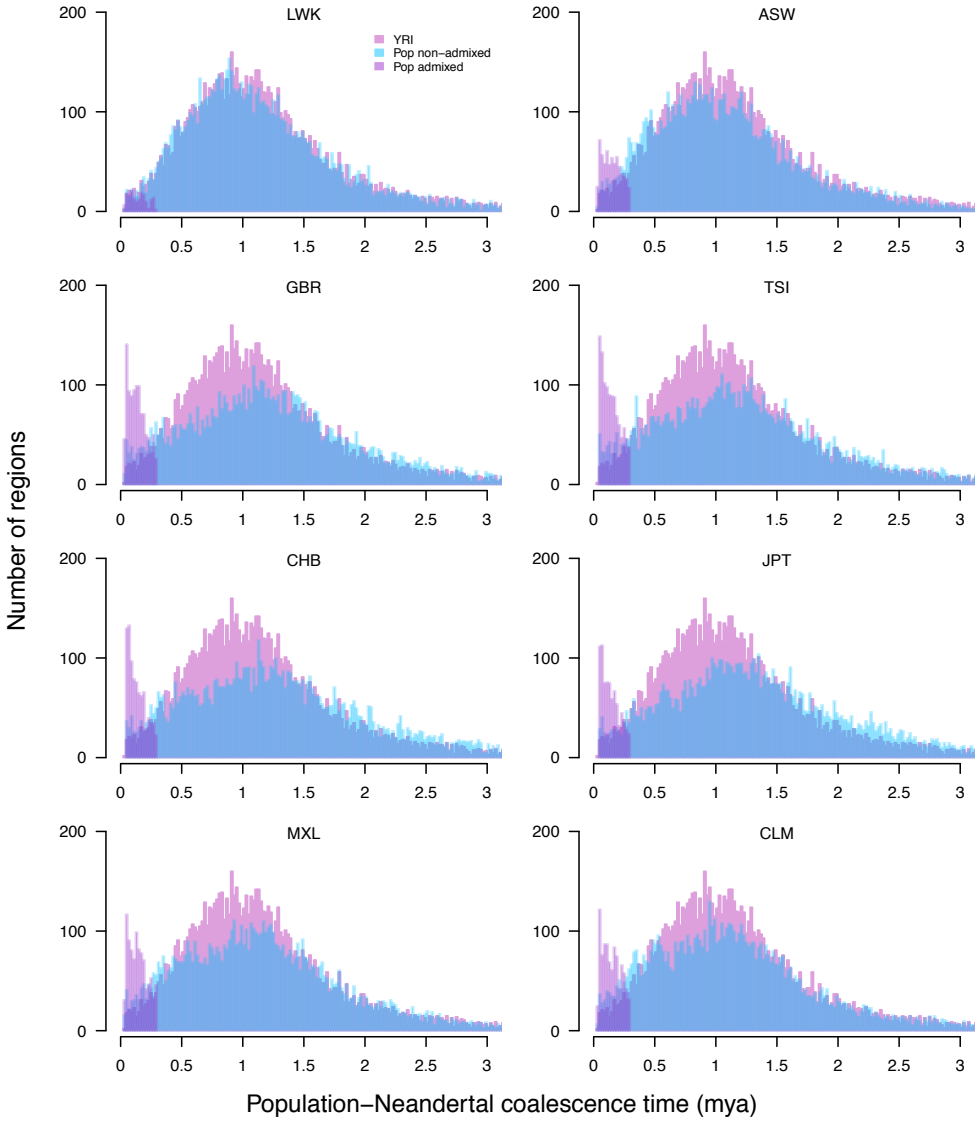


Figure 4.8: Coalescence time distributions of non-admixed and admixed regions as estimated by *CEPHi*, moving through the continents from Africa, Europe, Asia, and America, from top to bottom. We plot separately the regions we deem to be admixed and non-admixed using our definition given earlier, for a set of 8 populations. The remaining populations (CEU, FIN, IBS, CHS, PUR) show distributions very similar to those from their continent. Each plot gives the distribution of estimated coalescence times between YRI and Neandertal in pink for ease of comparison. Each population's distribution of estimated non-admixed coalescence times with the Neandertal is shown in bright blue, and the distribution of estimated admixed regions for that population in purple.

4.3.3 Comparing sets of introgressed regions between populations

In Appendix C, we give details enabling online access to the sets of regions we have identified as introgressed from the Altai Neandertal into each of the continental human populations studied here. Below, we summarise this in terms of quantity per population, and distribution across the genome.

4.3.3.1 The total amount of introgressed material within populations

In Figure 4.9 we summarise the total amount of introgressed material per population, by chromosome. Note that these are absolute amounts and are therefore affected by the number of individuals in each population. IBS, for example, has only 14 individuals, meaning the absolute amount of introgressed material is significantly lower than other populations, but probably not due to a different evolutionary history than other European populations. By contrast, LWK has a large number of haplotypes, but a much smaller amount of introgressed material: in this case we will more quickly attribute this to significantly less admixture from Neandertals into this population. We then separate total absolute amounts of Neandertal introgressed sequence into per chromosome amounts, per population.

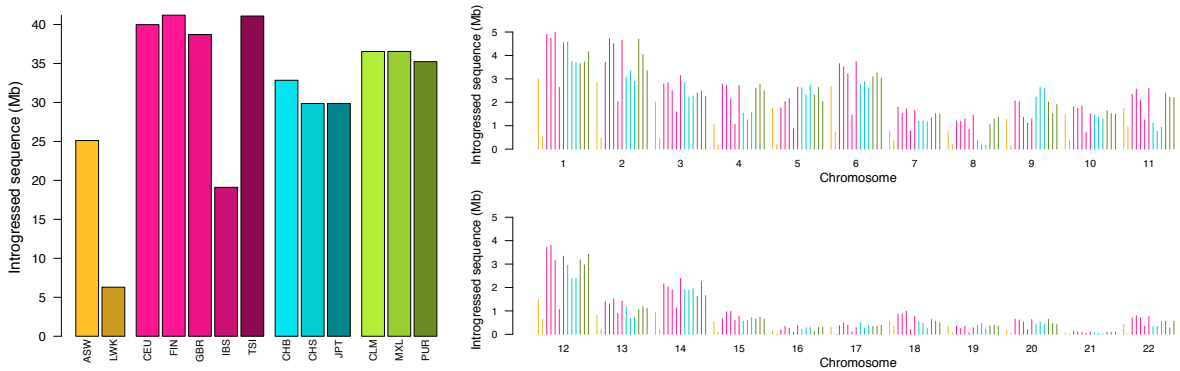
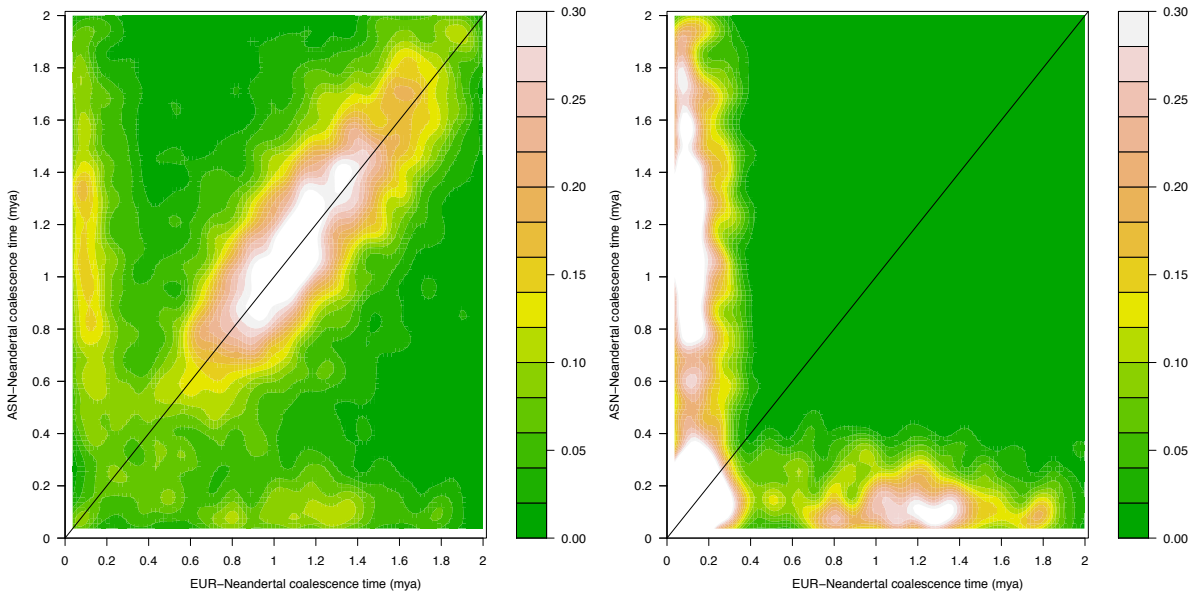


Figure 4.9: Amount of introgressed sequence per population. The left hand barplot shows the total Mb of genetic material introgressed from the Neandertal per population (not including YRI as Yorubans are used as the reference population which we assume contains no Neandertal sequence). The right hand plots show this total separated into chromosomes. Populations are in the same order as the left hand plot.

4.3.3.2 Comparing introgression in Asians and Europeans

Figure 4.10 shows a comparison the sets of admixed and non-admixed regions between European and Asian populations, using two heatmaps. From subplot (a), it is clear that a large number of non-admixed regions (in total 8036 regions), defined as such for either the European (7050) or Asian (7503) continental populations, are also defined as such for both continental populations (6517), as we see a large hotspot across the diagonal, meaning coalescence times with the Neandertal are largely shared or similar (this is also evident in Figure 4.7). We also see small patches of regions along the y and x axes, showing us regions which are defined as non-admixed in Europeans but not Asians (986), and vice versa (533), respectively. Subplot (b) in Figure 4.10 shows the respective coalescence times with Neandertals for admixed regions in European and Asian populations. We see a large number of regions (840) at the bottom left of the plot, representing regions defined as admixed in both European and Asian populations. Those seen vertically on the left hand side of the plot are admixed in European populations but not in Asian (986), and those seen horizontally along the base of the plot are admixed in Asian populations and not European (533). Thus, we see substantial sharing of a large number of admixed regions, but important separation of others, suggesting that there have been instances of admixture between humans and Neandertals which took place before the Eurasian population split, as well as potential instances of admixture after that split, where we see different sets of regions for the respective continental populations.

We then reduce the plotted regions to only those that are classified as admixed in either European or Asian populations respectively. The first of the two plots in Figure 4.11 shows those regions defined as admixed in any of the five European populations and the corresponding Asian coalescence time. We again take the minimum coalescence time with the Altai Neandertal and discard the remaining times. We do the equivalent for ASN populations: starting with all regions defined as admixed in any of the ASN populations and plotting the corresponding coalescence times for EUR populations. Interestingly, we see a larger number of regions defined as admixed across Europe (1842) as compared with Asia (1372), totalling 64Mb and 47Mb respectively. Furthermore, the minimum coalescence time in European populations is 36,534



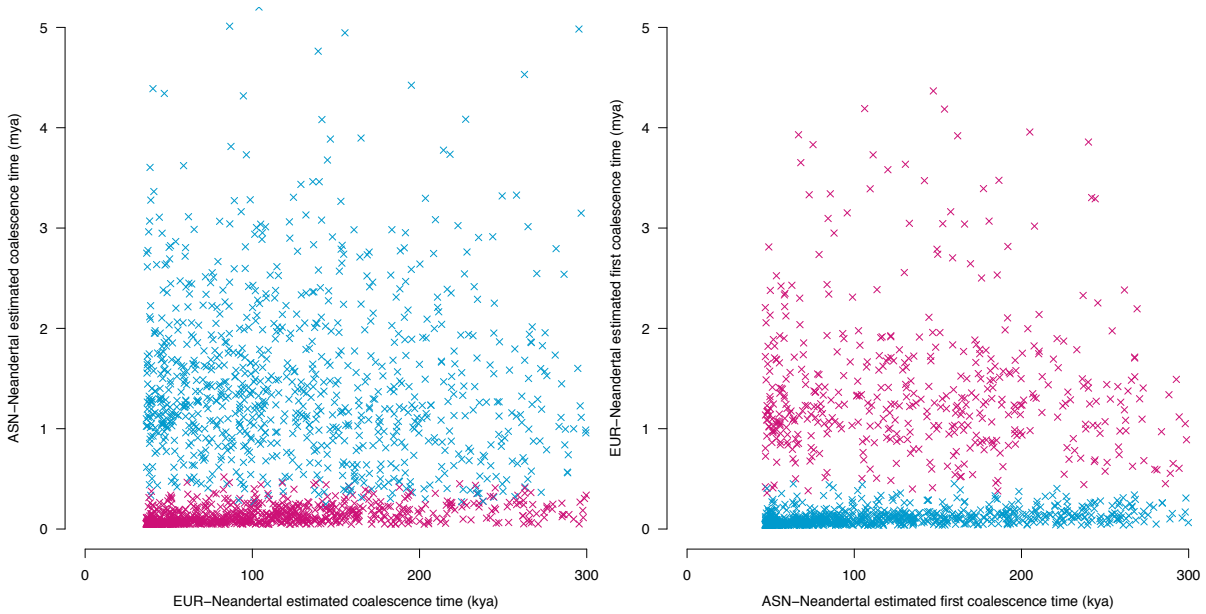
(a) Non-admixed **(b)** Admixed

Figure 4.10: Comparing admixed and non-admixed regions between European and Asian populations. Each heatmap compares the minimum estimated coalescence times with Neandertals seen in an individual region, between European and Asian populations. Subplot (a) shows all regions defined as non-admixed in either European, or Asian, or both continental populations. Subplot (b) shows all regions which are defined as admixed in either European, or Asian, or both continental populations.

years ago, where for Asian populations it is 46,687 years ago.

4.3.3.3 Comparing introgession in individual populations

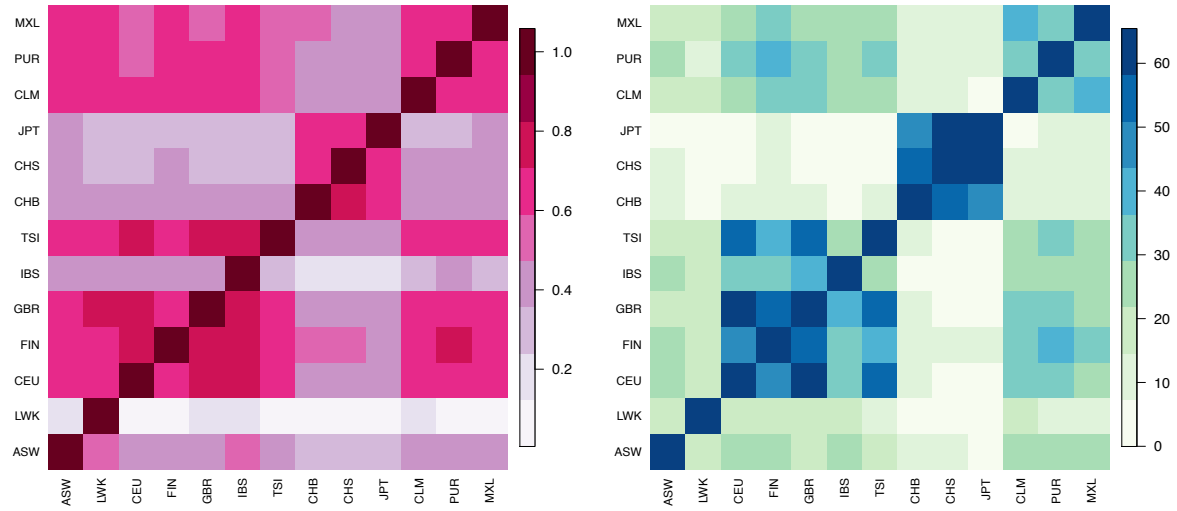
Figure 4.12 shows two comparisons of sets of admixed regions across populations. We compare each set of admixed regions with all other sets from the remaining 12 populations, to ascertain the proportion defined as admixed that is shared with all other populations. To do this, we first define regions in a population as admixed in the normal way: by taking the mean of the minimum coalescence time with the Altai Neandertal, and the mean of the minimum coalescence time with the Altai Neandertal for YRI for the corresponding region. Where the former is $\leq$300,000 years ago, and the latter >600,000 years ago, a region is define as admixed. We then select only those regions shared between each of the populations in the pair; we only use regions used in all of three *CEPHi* runs with the Neandertal: YRI, and each of the populations in the pair. We then plot the proportion of the regions admixed in both populations as a proportion of



(a) All regions classified as admixed in Europe **(b)** All regions classified as admixed in Asia

Figure 4.11: Comparing minimum continent-Neandertal estimated coalescence times between Asia and Europe. Subplot (a) shows the corresponding minimum Asian-Neandertal coalescence times for the set of regions classified as admixed in Europe by our definition (1842 regions): pink crosses represent regions classed as admixed in both European and Asian populations, blue are admixed only in Europeans. Subplot (b) shows the converse (1372 regions): pink crosses represent regions classed as admixed in Asian and European populations, pink are admixed only in Asians. Note the differing scales of the x and y axes. No coalescence times are seen before 22,400 years ago as this is the minimum coalescence time used in *CEPHi*.

the total number of regions used in the analysis for that population. Thus, the matrix is not symmetric (as there are different numbers of regions used for each population) but there is high correlation on either side of the diagonal as the number of regions does not vary greatly (except for LWK which has very few admixed regions, making the proportion it shares with other populations very high, and the proportion each of the remaining populations shares with it very low). Overall we see high correlations between populations within a continent: this is clear within American, European, and Asian populations. We also see larger shared proportions between African and European populations than between African and Asian populations, highly suggestive of back migration from Europe into Africa. We note that LWK shares the largest proportion of its admixed regions with GBR. Substantial proportions of admixed regions are also shared between Asian and European populations, which reflects Figure 4.10. We note enrichment of Spanish (IBS) regions in Puerto Ricans (PUR), and an apparent enrichment of



(a) Proportion of shared admixed regions

(b) Odds ratios and Fisher's exact test of shared admixed regions

Figure 4.12: Comparing sets of admixed regions for all pairs of populations. Subplot (a) shows the proportion of admixed regions shared between each pair of populations, as a proportion of the total number of admixed regions in the first population. Thus, the matrix is asymmetric: coming from the y-axis population (for example FIN), you see the proportion of admixed regions in that population that are also seen in the adjoining x-axis population (say LWK), of the total number of admixed regions in the x-axis population (LWK). So in this case, the shared admixed regions between FIN and LWK is high because those regions make up a large proportion of the total set of admixed regions in LWK. Starting from the y-axis for the second population in the first comparison (LWK) and moving across to FIN on the x-axis will yield a different proportion, as it takes into account the total number of admixed regions for the second population. So of those admixed regions shared between LWK and FIN, this is a small proportion, because FIN have a much larger set of admixed regions than do LWK. Indeed, because LWK has very few admixed regions in total, meaning that its horizontal line shows low proportions across all populations, because the y-axis populations include many more admixed regions which LWK does not have. By contrast, the vertical line for LWK shows relatively high proportions, because the sets of admixed regions in the x-axis populations encompass more of the admixed regions in LWK. Subplot (b) shows the odds ratios calculated for each pair of populations (a symmetric matrix). Every comparison is statistically significant at the 95% level, using Fisher's exact test, showing significant enrichment of each admixed set with that of another population.

Finnish (FIN) regions in the Han Chinese (CHB, CHS). We suggest that we see a stronger correlation when starting with the Finnish regions than when starting with the Chinese regions because the Finnish population has a larger total number of introgressed regions: we expect an Asian influence in Finns, likely from Siberia.

The odds ratios show a very similar pattern. Fisher's exact tests for all population comparisons were highly significant ($p << 0.05$), showing strong similarity between pairs of admixed regions. This is again grouped clearly by continent, with the strongest set enrichments seen

within European populations, within Asian populations, and within American populations. African populations show a stronger enrichment with European populations than Asian, again supporting the notion that back migration from Europe explains the existence of introgressed regions in African populations.

4.3.4 Examining anomalous regions with intermingled Neandertal haplotypes

Within the set of coldspots we use for analysis, we see either one or two coalescence times between a population of modern humans and the Altai Neandertal haplotypes, per region. The vast majority contain one coalescence, meaning the two Neandertal haplotypes have coalesced with one another before meeting the human haplotypes back in time. Where there are two, at least one of the Neandertal haplotypes coalesces with some human haplotypes separately, before later coalescing together. Table 4.2 shows the joint subsets of regions used in the YRI-Neandertal and GBR-Neandertal analyses in *CEPHi*, with one and two coalescences with the Neandertal. As we might expect, given the level of homozygosity in the Neandertal, the majority of regions in both GBR and YRI show a single coalescence with the Neandertal, as the Neandertal haplotypes will normally coalesce with one another recently in the past before meeting any human haplotypes. We also see a substantial set with two coalescences for GBR, and one for YRI; this is also in line with expectations as these are likely to be regions of admixture. The regions of particular interest are those with two coalescences in YRI - regardless of whether there are one or two coalescences in GBR. This is contrary to expectation, as we expect YRI to coalesce with the Neandertal relatively far back in the past, given its position as a negligibly admixed population.

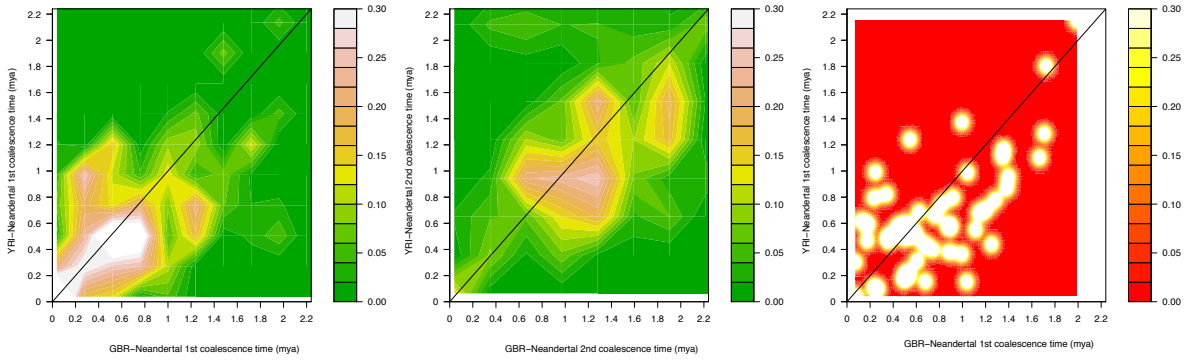
Thus there are two intriguing region subsets containing two coalescence times between YRI and the Neandertal. We investigate those with two coalescences in both YRI and GBR as it is a substantial subset containing 454 regions; there are only 61 regions with two coalescences for YRI and one for GBR. For each of these regions, we plot the first estimated coalescence times of GBR and YRI with the Altai Neandertal in Figure 4.13 (left hand plot) alongside the equivalent for the second coalescence times (central plot).

Number of coalescences	YRI-Nean 1	YRI-Nean 2
GBR-Nean 1	7582	61
GBR-Nean 2	732	454

Table 4.2: Taking all regions used in both the GBR-Neandertal and YRI-Neandertal analyses in *CEPHi* (regions are filtered before analysis so the resulting sets differ slightly), we show the set of comparable regions (those present in both analyses) with one and two coalescences with the Altai Neandertal across GBR and YRI. For example, the top left shows the number of regions with a single coalescence between the 178 GBR haplotypes and the two Altai Neandertal haplotypes, and a single coalescence between the 176 YRI haplotypes and the two Altai Neandertal haplotypes. A single coalescence in this sense means the two Neandertal haplotypes have coalesced with one another before they coalesce with any human haplotypes.

We see a clear bimodal distribution between the first coalescence times with the Altai Neandertal for GBR and YRI. One set of regions shows very recent coalescence times below 300kya for both GBR and YRI, the second (and larger) set is seen between ∼300kya and ∼700kya. Overall, then, we note that the large majority of regions coalesce with the Neandertal in both populations before the human-Neandertal population split of ∼712kya. We also note some potentially admixed regions on the left of the plot where GBR-Neandertal coalescence times are typically recent and YRI-Neandertal coalescence times further in the past. The area in the bottom left which shows very recent coalescence times in both YRI and GBR may reflect admixture into YRI which also exists in GBR (as we saw for LWK earlier in the chapter), although we know this to be rare. Alternatively, it may be that the Neandertal is highly diverged in these regions, either by chance, or through admixture with an undetermined population, preventing it from coalescing with itself until further back in the past. Local variation in the rate of mutation, as well as the possibility of selection require consideration before drawing strong conclusions, but it is possible that this may instead reflect admixture from humans into Neandertals. The second coalescence events for these regions appear typical of non-admixed regions as a whole (Figure 4.13) - with a centrepoint matching those seen in Figure 4.6 (at ∼1mya for YRI and ∼1.1mya for GBR). We see, then, putative evidence of some uncharacterised interactions between ancestral groups more recently than the main separation of the two hominid groups. This might take several forms, which further investigation is required in order to distinguish.

For example, it might reflect ancient population substructure in this time period. A ghost



(a) 1st coalescences **(b)** 2nd coalescences **(c)** Kuhlwilm's

Figure 4.13: Comparative coalescence times for 454 regions with intermingled Neandertal haplotypes in both GBR and YRI (row 2 of Table 4.2). Subfigure (a) shows the comparative first coalescence time for each of the 454 regions, and subfigure (b) the equivalent for the second coalescence times. Subfigure (c) shows the first coalescence times with the Neandertal for the 49 regions overlapping our coldspots that were reported in Kuhlwilm *et al.* [2016].

population far back in the past on the lineage leading to modern humans, and before the later Out of Africa event, for example, may have admixed with the population leading to the Neandertal lineage, causing this thorough mixing of modern human and Neandertal haplotypes across African and non-African human populations. This mixing must have been ancient, given the deep coalescence times we predominantly observe. For this to be the case, we expect one of two coalescence times to be recent, and the second to be far in the past. This is because we assume admixture to occur only once in any one region, so a single Neandertal haplotype should be similar to a set of modern humans, with the second haplotype coalescing with this clade further back in the past. Looking at subplot (b) in Figure 4.13, we see that they fall within the normal range of coalescence times that we see across regions in Figure 4.6, starting at ∼600kya, with a central point around at 1mya in YRI and slightly later (∼1.1mya) in GBR. Potentially the regions with very young first coalescence times are compatible with admixture from humans into Neandertals, because as explained above, under this scenario we would expect one of the two Neandertal haplotypes to coalesce recently with a set of human haplotypes (and the second to coalesce later, as we assume an admixture event to occur only once in a region), but we suggest that the majority whose first coalescence time falls later than this, are not.

Related to this is a recently published paper which suggests the existence of admixture in this

direction (Kuhlwilm *et al.* [2016]). The authors identified 162 regions displaying this, of which 50 fall within our coldspot set (and 6 within our set of admixed regions across populations). We investigate the set of coalescence times with Neandertals for GBR and YRI within these regions. Of these 50 regions, 49 show two coalescence times with both GBR and YRI; a huge enrichment. For admixture from humans into Neandertals to have occurred, we expect a young first human-Neandertal coalescence time, and an old second coalescence time. We plot the first coalescence times in GBR and YRI with the Altai Neandertal in subplot (c) of Figure 4.13. It is clear that very few of these first coalescence times are young, the majority of them falling later than 400kya, and a large number falling later than the human-Neandertal split time. In fact, most fall within the equivalent of the largest region in subplot (a) of Figure 4.13, between ∼300kya and ∼700kya. These regions clearly have an interesting history, and we suggest, given the structure of their genealogies - and particularly the majority of first coalescences sitting between ∼300kya and ∼700kya - that they may be indicative of a complex speciation event between humans and Neandertals. Further work regarding this could first locate these regions on the genome to ascertain whether they are evenly distributed across the genome: if not, this may strengthen conclusions that they have an interesting history; potentially demonstrative of a complex speciation event between humans and Neandertals, perhaps with ancestral populations of the two species mixing before evolution into later forms had occurred (although discussion so far is based on the Altai Neandertal alone). We could also apply our dating method presented in Chapter 5 to these regions to examine when this admixture may have occurred.

4.3.5 Comparing our introgressed regions with previously published sets

We compare sets of admixed regions obtained using three different methods. The approaches of Vernot and Akey [2014] and Sankararaman *et al.* [2014] are described in detail in the introduction to this chapter. One of the main differences between our search for admixed regions, and these previous publications is that we only look in recombinationally cold regions of the human genome, so our resulting set is significantly smaller. However, we see significant overlap between the sets of regions captured. There are also some differences, which we now highlight.

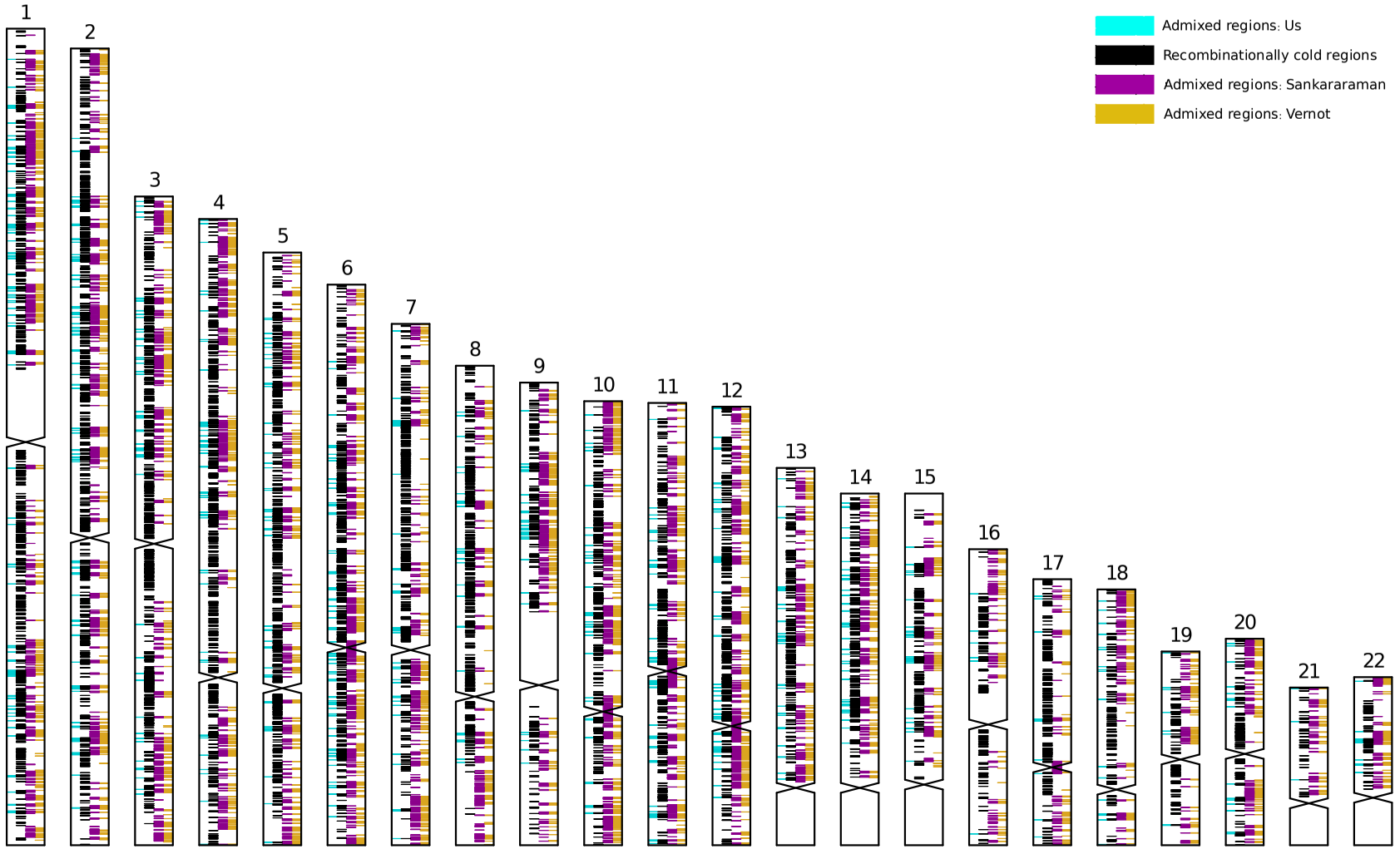
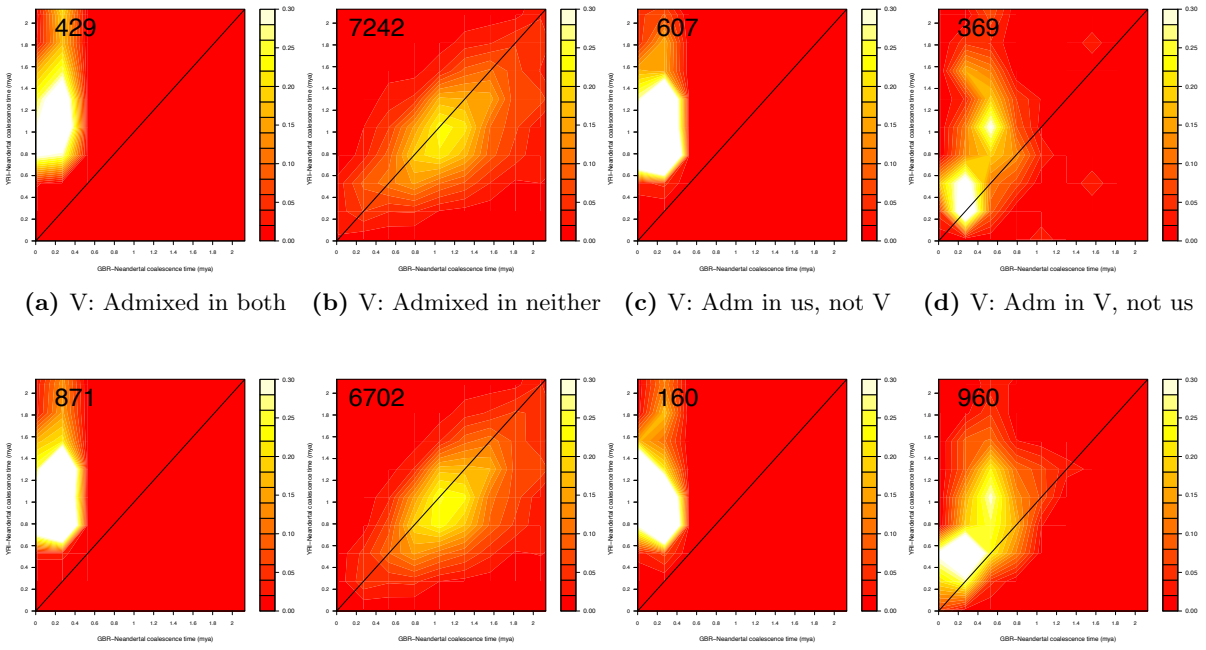


Figure 4.14: Comparing regions of introgressed material across studies for GBR. Each chromosome shows in four columns (from left to right) (a) the regions we have inferred to be the result of Neandertal admixture (b) recombinationally cold (<0.2cM/Mb) regions of the human genome (c) regions declared admixed by Sankararaman *et al.* [2014] (d) regions declared admixed by Vernot and Akey [2014]. Chromosomes are shown with the p arm to the base. Crosses indicate centromeres.

Figure 4.14 shows a comparison of our admixed regions (cyan) alongside those of Vernot and Akey [2014] (gold) and Sankararaman *et al.* [2014] (purple) in GBR. We also show the recombinationally cold regions (black) that we searched within. Firstly, it is clear that the sets of Vernot and Akey [2014] and Sankararaman *et al.* [2014] line up very closely. Secondly, we note that our admixed regions are very often met with a corresponding admixed region in both of the remaining sets. Thirdly, we note that admixture deserts seem to be similar between sets. Aside from the notably empty p arms of chromosomes 13, 14, 15, 21, and 22, Vernot and Akey [2014] cite $8q$ and $17q$ as the main deserts in EUR, and Sankararaman *et al.* [2014] cite the end of $3p$, the beginning of $4q$, the beginning of $5q$, and the middle to end of $7q$. Lastly, on inspection, we confirm the appearance of a negative correlation between coldspots and regions of introgression in the admixture sets from Vernot and Akey [2014] and Sankararaman *et al.* [2014], as discussed in the introduction.

Finally, we compared the regions called as admixed between our set and those of Vernot and Akey [2014] and Sankararaman *et al.* [2014]. Taking each of their sets in turn, we take our cold regions, and of these, find those which are called as classified as admixed in both, those which are classified as admixed in ours and not theirs, those which are classified as non-admixed in ours and admixed in theirs, and those which are classified as non-admixed in both sets. For each of these sets, we present heatmaps of the mean first coalescence times across two trees between GBR and the Altai Neandertal, and between YRI and the Altai Neandertal, in Figure 4.15

There are 8829 coldspots in total in the YRI-GBR comparison, 1109 of which we classify as admixed, and 7720 as non-admixed, using our definition. Of this 7720, Vernot and Akey [2014] classify 7242 as non-admixed, and Sankararaman *et al.* [2014] 6702. Vernot and Akey [2014] classify 980 of 8829 as admixed. Of our 1109 admixed regions, those classified as also admixed in Vernot and Akey [2014] at all number 500, and reduces to 429 when we ask for at least a 25% overlap of our region by theirs. This leaves 607 which do not. There remain only 369 regions in our coldspot set which Vernot and Akey [2014] class as admixed and which we do not. For Sankararaman *et al.* [2014], 1967 regions of 8829 are classed as admixed at all, this reducing to 1831 when we ask for at least a 25% overlap of their regions across ours. Of this 1831, we class



(a) V: Admixed in both **(b)** V: Admixed in neither **(c)** V: Adm in us, not V **(d)** V: Adm in V, not us

(e) S: Admixed in both **(f)** S: Admixed in neither **(g)** S: Adm in us, not S **(h)** S: Adm in S, not us

Figure 4.15: Comparing our introgressed regions with those from Vernot and Akey [2014] and Sankararaman *et al.* [2014]. The top row shows comparisons made with Vernot and Akey [2014], the bottom row with Sankararaman *et al.* [2014]. From right to left in each row we show (a) those regions classified as admixed in both our set and their set, and where their region overlaps ours by at least 25%, (b) those regions classified as non-admixed in both our set and their set, (c) those regions classified as admixed in our set and not their set, and where their region overlaps ours by at least 25%, and (d) those regions classified as admixed in their set and not in our set. The number in the top left corner of each subplot gives the number of regions plotted.

871 of these as admixed, and 960 as not.

It is clear that there is substantial classificatory similarity within our set of coldspots with both sets. The large majority of regions are classified as non-admixed in each paired comparison. Of those that we do not classify as admixed but the comparator set does, the coalescence times are often young in YRI, and we see corresponding hot regions in the bottom left hand corner of the plots in the fourth column. We also see, particularly in Vernot and Akey [2014], some regions where the coalescence time in GBR is older than we allow in our definition, making these regions debatably admixed.

In Figure 4.16, we contrast the sets of admixed regions by examining the overlap of our admixed regions by the sets produced by Vernot and Akey [2014] and Sankararaman *et al.* [2014]. We

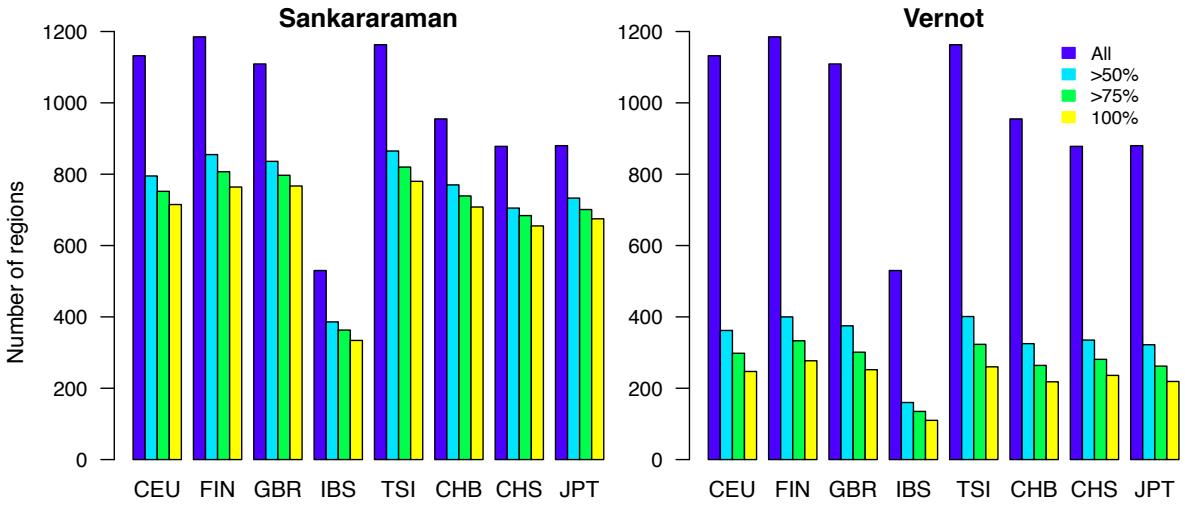


Figure 4.16: The overlap of our set of admixed regions with two previously published sets from Vernot and Akey [2014] and Sankararaman *et al.* [2014]. We include all European and Asian populations. From left to right the bars represent (1) the total number of admixed regions for that population, (2) the number overlapped by the comparison set by more than 50%, (3) the number overlapped by more than 75%, and (4) the number overlapped completely.

give the number of our regions classed as admixed which are overlapped by 50%, 75%, and 100%. The numbers are absolute, and the set produced by Sankararaman *et al.* [2014] is larger than that of Vernot and Akey [2014], explaining the greater coverage of our admixed regions by the former set.

4.4 Discussion

In this chapter, we have investigated the significant amount of information contained within the genealogical trees built in *CEPHi* which link human populations with the Altai Neandertal. We see that, relative to YRI, there is an overwhelmingly clear signal of introgression from Neandertals into each of the 13 human populations we use, with this signal being significantly reduced, but not absent, in the two African populations (LWK and ASW) - or at least heavily enriched in African ancestry as is the case with ASW.

Our simulations show that simulated segregating sites data using a demographic scenario with 2% admixture from Neandertals into a non-African population 45,000 years in the past, and none

into an African population, produces genealogical trees from *CEPHi* connecting Neandertals and modern humans with population split times, size histories, and sets of admixed regions matching those of our real data. This thereby supports the notion that admixture occurred between Neandertals and non-African human populations at $\leq$2% somewhere around 45kya (Prüfer *et al.* [2014]). We see strong similarities between the population size histories of YRI, GBR, and Neandertals from the simulated truth and that inferred by *CEPHi*, as well as likelihood searches with clear maxima for split times between humans and Neandertals matching that of the real data, and clear distributional splits between admixed and non-admixed regions in GBR, showing we can successfully pick apart admixed regions from non-admixed regions in the data. Further work with regard to these simulations could compare the per region coalescence times in the true genealogical trees produced by the coalescent simulator (*scrm*) to the coalescence times in the genealogical trees produced by *CEPHi* when using that simulated data. Clear matching of coalescence times between admixed (and non-admixed) regions would further strengthen support for the accuracy of *CEPHi* with regard to inferring parameters for interactions between ancient and modern human species.

Moving to real data, we provide a per-region definition of admixture. Across two genealogical trees, we take (a) the mean of the minimum coalescence times between the haplotypes in the test human population and the Altai Neandertal, and (b) the mean of the minimum coalescence times between the haplotypes in YRI and the Altai Neandertal. We define an individual region as admixed in the test human population where (a) is $\leq$300kya, and (b) is $>$600kya. This definition comes from the clear delineation between two region types seen in Figure 4.6 where YRI:Neandertal coalescence times are compared with non-African:Neandertal coalescence times. Across non-African populations, we see a set of regions along the y-axis, fulfilling the criteria (a) and (b) given above. These are likely indicative of admixture as they represent the qualitative differences we expect between African and non-African populations as regards their relationship to the Neandertal. When this definition is applied to the sets of coalescence times across all recombinationally cold regions of the genome, we find distinct distributional separation for admixed and non-admixed regions in all groups, including the African populations LWK and ASW, which have notably less admixture as compared with the European and Asian

populations. The bimodality is very clear (far more substantial and distinct, for example, than peaks at zero seen in Chapter 2), and thus we are able to separate those regions which seemingly have Neandertal ancestry from those which do not.

When comparing sets of admixed regions between populations, we see some interesting patterns. Firstly that there is both substantial correlation and separation between sets of admixed regions when populations are grouped by continent, implying that some admixture likely took place before the Eurasian split ∼40kya, and also that there may have been admixture into both populations since the split. Individually, we see high correlations between populations, especially intracontinentally, but also intercontinentally, further compounding the idea that older instances of admixture likely took place. Interestingly, we see closer links for LWK and ASW with European as opposed to Asian populations, suggestive of European back migration into Africa. Incomplete overlap between sets of introgressed regions from different populations may reflect sampling biases, genetic drift (including the extinction of rare haplotypes), and selection, but detailed investigation beyond the scope of this thesis is required for conclusions to be drawn. Further work might also include consideration of the influence of basal Eurasian admixture into European populations. Basal Eurasians are a lineage hypothesised to have split from the ancestors of all other Eurasians, and which contains little to no Neandertal ancestry (Lazaridis *et al.* [2016]).

On comparing our sets of introgressed regions with those from previous publications, we see substantial agreement overall, showing that our definition of admixture has good overlap with other approaches. However, due to differing methods, all methods locate a set of introgressed regions not found in others. Because we condition on specific coalescence times between the Neandertal and non-African and African populations, the regions defined as introgressed by Vernot and Akey [2014] and Sankararaman *et al.* [2014] which we do not include, are all shown to have recent coalescence times with the Neandertal in YRI, or late coalescences with GBR, and therefore may not be the result of admixture from humans into Neandertals. Further investigation into regions classified as admixed by these previous publications which we do not call as admixed may reveal further insight.

CEPHi is designed to find regions that have been introgressed from Neandertals into humans, but the publication of Kuhlwilm *et al.* [2016] led to consideration of the possibility of regions that may have been introgressed from humans into Neandertals. Our genealogical trees are highly useful here, as their structure can be indicative of historical events. For example, where two Neandertal haplotypes coalesce together recently, and subsequently coalesce with a subset of human haplotypes, this is suggestive of introgression from Neandertals into humans, because only a subset of humans are descendants of this introgression. By contrast, where the Neandertal haplotypes are intermingled, with one coalescing recently with human haplotypes, and the second coalescing much further back in time, this is potentially suggestive of introgression in the other direction; the set of humans is more similar to each other than it is to the second Neandertal haplotype. Thus, we investigated 454 regions with two coalescence times (and therefore where at least one haplotype is more similar to a set of human haplotypes than to the second Neandertal haplotype) in the genealogical trees between both GBR and YRI with Neandertals, and found that the set of regions from Kuhlwilm *et al.* [2016] that overlap our coldspots may not be compatible with introgression from humans into Neandertals. We suggest these regions likely have an unusual evolutionary history, and may represent a complex separation event between humans and Neandertals, with admixture occurring between the separating populations constituted of the human and Neandertal predecessors after the initial population split.

In the next chapter, we introduce a new method to date admixture between Neandertals and modern human populations. To do this, we employ information about SNP placement on genealogical trees to classify two SNP types which are indicative of admixture from Neandertals into humans.

CHAPTER 5

Dating admixture with SNP placement on genealogical trees

5.1 Introduction

As we uncover more about our species' history from both genetic and archaeological sources, it becomes predictably and increasingly complex from both an intra- and inter-species standpoint. Looking within *Homo sapiens*, for example, we see evidence of unexpected groups having contributed to modern human populations, including a pre-Columbian ghost population to South American indigenous groups, and a Siberian population to the ancestors of American-Indians and Europeans (Raghavan *et al.* [2014]; Skoglund *et al.* [2015]).

Between species (although this terminology is of course debatable), admixture from Neandertals

and Denisovans has been detected in various non-African human populations (Meyer *et al.* [2012]; Prüfer *et al.* [2014]; Sankararaman *et al.* [2014]; Vernot and Akey [2014]). Unknown archaic genetic material has also been found in the Denisovan (Meyer *et al.* [2012]), and from this it seems probable that alongside the increasingly complex history of humans with regard to Neandertals and Denisovans, that other archaic species will be included in the history of our species. Significantly, these discoveries have introduced the characterisation of our evolution as reticulate - a networked phylogeny, or 'braided stream' - rather than more simply bifurcating, as has been the traditional assumption.

In this final chapter, we focus on dating admixture between the Altai Neandertal and the 14 human populations in the 1000 Genomes dataset (see Appendix B for details of these datasets). Dating admixture between *Homo sapiens* and *Homo neanderthalensis* across human populations - and thereby geographic regions - further contributes to the painting of a more detailed picture of the evolutionary path of humans and our cousin species.

Previously to this, one paper (Sankararaman *et al.* [2012]) looked to date admixture between modern-day humans and Neandertals, using a method which distinguishes between the hypotheses of admixture versus ancient African population structure (see Chapter 1 for a fuller discussion of these possibilities). The admixture hypothesis is supported if this date is significantly more recent than speciation, and indeed the authors report the last gene flow from Neandertals into Europeans most likely occurred within the broad range of 37-86kya. This is done by calculating a measure of LD, $\bar{D}$, given as:

$$\bar{D}(x) = \frac{\sum_{(i,j)\in S(x)} D(i,j)}{|S(x)|}, \tag{5.1}$$

where $\sum_{(i,j)\in S(x)} D(i,j)$ is the sum of the standard measure of LD, D, for all pairs of SNPs (i,j), and $|S(x)|$ is the set of all pairs of SNPs at genetic distance apart x, where $0.02\text{cM} \leq x \leq 1\text{cM}$ (pairs of SNPs are binned by genetic distance because the probability of recombination is dependent upon it). SNPs are chosen according to an ascertainment scheme that searches for SNPs of Neandertal origin, requiring that SNPs (a) contain at least one derived allele in the Neander-

tal, (b) are polymorphic in humans, and (c) have a derived allele frequency in humans of <0.1 (those SNPs for which an excess of derived alleles is shared between humans and Neandertals as compared with humans and Denisovans are more frequently seen below this cutoff). Given a time of admixture - an LD-generating event - regions that have been introgressed from the Neandertal will be broken up by recombination exponentially with probability proportional to $1-e^{-t_{GF}x}$, where t_{GF} is the date of introgression. LD is plotted across genetic distance to give a decay curve. The rate parameter of this curve (t_{GF}) is equal to the number of generations in the past that admixture occurred between two groups. The steeper the exponential decay, the older the admixture. This is because where admixture is old and there has therefore been more time for recombination within the admixed region, there is a faster reduction in probability that the ancestry of a pair of SNPs remains the same as genetic distance increases. This curve is solved for t_{GF} (with the inclusion of a bias correction (α) to account for random errors in the genetic maps) using ordinary least squares to give a single admixture date. This method is of course dependent upon accurate fine-scale genetic maps for inference, which may contain biases at the scale of tens of kilobases, therefore potentially affecting analyses. Sankararaman *et al.* [2012] implement a correction to account for this. They used simulations to model random errors in the genetic maps, and show that they can obtain a corrected date from the uncorrected date using $t_{GF} = \alpha(e^{\frac{\lambda}{\alpha}} - 1)$.

$\bar{D}$ was calculated for 59 West Africans, 60 Europeans, and 60 East Asians, for all pairs of SNPs between 0.02cM and 1cM apart, incrementing in steps of 10^{-3}cM, and an exponential curve fitted to the LD decay. In West Africans, LD decays more quickly where the Neandertal carries the derived allele. This is expected in a scenario lacking admixture, because for ascertainment conditions (a) and (b) to be fulfilled, the mutation must have occurred further back in time (in fact before the human-Neandertal population split) than if the Neandertal carries the ancestral allele (where the mutation must have occurred on the human lineage). In this non-admixed scenario, then, an older mutation carries lower LD than a younger one, as we see in West Africans. By contrast, Europeans and East Asians show a slower exponential decay (and therefore longer range LD) at SNPs where the Neandertal carries the derived allele compared to where it carries the ancestral. This is expected in an admixture scenario where the mutation has occurred on

the Neandertal lineage and subsequently been introgressed into the human lineage.

Admixture is dated to between 37-86kya in European populations. A date estimate is not supplied for East Asian populations, due to the difficulty in applying the bias correction to account for random errors in the relevant genetic map. However, before correction, t_{GF} is given as 1,253-1,287 generations BP, similar to the uncorrected estimate of 1,159-1,183 generations BP for Europeans, which - it is noted by the authors - is suggestive but not conclusive of the idea that Neandertal material in Europeans and East Asians may stem from the same gene flow event.

Distinguishing admixture LD - linkage that has been generated through introgression, from non-admixture LD - that is present due to shared ancestry, incomplete lineage sorting, or bottlenecks and genetic drift in human populations since the Neandertal-human split, is required when using patterns of LD to date ancient admixture. For more recent admixture events, admixture LD will likely stretch over much longer distances than non-admixture LD, but for ancient events, stretches of non-admixture LD can be similar in length to stretches of admixture LD. This is dealt with in Sankararaman *et al.* [2012] by using the ascertainment scheme described above to minimise the signal from non-admixture LD.

Additionally, Seguin-Orlando *et al.* [2014] used a Hidden Markov Model (HMM) to date admixture to 54kya between Neandertals and the genome of a 36.2-38.7 thousand year old male from the Middle Don River in Russia, called Kostenki 14 (K14). A small number of studies using ancient humans have in fact looked to date admixture between these individuals and Neandertals, the majority relying on patterns of LD. One example of this involves the 45,000 year old Ust'-Ishim human from Western Siberia (Fu *et al.* [2014]). Admixture between Neandertals and this individual's ancestors is dated at 50-60kya, 232-430 generations before he lived. An ascertainment scheme is employed in order to find SNPs that are informative for Neandertal introgression: (a) Africans (YRI and LWK from the 1000 Genomes Project) are fixed for the ancestral allele, (b) the Altai Neandertal carries the derived allele, and (c) at least one derived allele is seen in the remaining 1000 Genomes haplotypes. A genetic map is used to calculate the average covariance over all pairs of SNPs in 0.001cM bins, and an exponential function

with rate parameter nd is fitted, where n is the number of generations since admixture, and d is the genetic distance in Morgans. A block jackknife is used - removing one chromosome at a time from the analysis - to compute the standard error of the admixture date estimate, resulting in an estimate of 51,728-57,740 years, using a generation time of 29 years. Ust'-Ishim is reported to be 2.3±0.3% Neandertal admixed (up to 3.08% in Fu *et al.* [2015]) and, using the same generation time, a mutation rate of 0.43×10^{-9}bp/yr is inferred, corroboratively close to the estimates from human pedigrees (Scally and Durbin [2012]). Incidentally, this supports our use of the mutation rate of 0.5×10^{-9}bp/yr throughout this work, and makes our admixture date estimates almost directly comparable.

The same authors expanded their approaches to dating admixture using a slightly younger Romanian individual named Oase 1 (Fu *et al.* [2015]), dated to have died between 37-42kya. Interestingly he displays some Neandertal morphological characteristics, as do his contemporaries, Oase 2 and 3 from the same cave in the Carpathian mountains. Similar to modern humans, Oase 1 exhibits a rounded cranium and protruding chin, and similar to Neandertals, he has a very wide ramus (vertical part of the mandible or lower jaw) and large distal molars (Trinkaus *et al.* [2003]), making him an early modern human (EMH), and potentially part of one of the first human European populations.

Fu *et al.* [2015] show Oase 1 to contain substantially more Neandertal genome than other humans (whether modern or early): between 6-9%, with some very large segments longer than 50cM, indicating a very recent Neandertal ancestor. The SNP ascertainment scheme looked to locate positions that almost always differ between sub-Saharan Africans (YRI) and Neandertals, and to then remove any positions where the Dinka matched the Neandertal, as the Dinka are reported to have negligible to no introgression from Neandertals (Prüfer *et al.* [2014]). On this scheme, Oase 1 contains 3,746 putative Neandertal alleles (compared to 1,586 and 1,121 for Ust'-Ishim and Kostenki) of 78,055 across 6 modern humans. This results in 7.3% of the Oase 1 genome derived from Neandertals (after an adjustment using the French genome to implement the assumption that the French individual has 2% Neandertal admixture). To acquire the date of admixture, as above, an exponential function was fitted to the curve of genetic distance

between all pairs of SNPs in 0.001cM windows that fulfil these criteria, and solved for the rate parameter which is equal to the date of admixture in generations, to give an estimate of 8.1±5.5 back from Oase 1's age. Generation time is not supplied in the paper, but reasonably assuming a generation time of 25-30 years, for example, this gives a date range between approximately 75-400 years before he lived. As this uses SNPs from regions of Neandertal admixture of all size, this date reflects an average of all admixture dates, but a mixture of two exponential distributions could not be fitted.

An alternative method for dating admixture in this paper used simulations to determine the length of introgressed regions after successive generations of recombination. In order to estimate the most recent date of admixture, only the largest putatively introgressed regions were used; these regions also provide more power as they contain more Neandertal SNPs. The number of generations was varied from 1-10, as well as the definition of 'largest' Neandertal regions (to include only the largest, or the largest and second largest, and so on). The actual size of introgressed regions matched those that had experienced 4-6 generations of recombination, meaning that Oase 1 may have had a great-great Neandertal grandparent, and certainly had a great-great-great-great Neandertal grandparent.

In this chapter, we take an orthogonal approach to estimating the date of admixture between humans and Neandertals as compared with those used in for example Sankararaman *et al.* [2012] and Fu *et al.* [2014], as we do not rely on LD to be informative. Genetic maps specify the probability of recombination per generation between all pairs of SNPs present in the genome they represent. However they may contain errors and bias, as noted in Sankararaman *et al.* [2012], and can be unreliable on the scale of tens of kb, potentially confounding inference. Corrections may be applied, such as in Sankararaman *et al.* [2012] where an error rate α was estimated by comparing the distribution of a set of crossovers in a pedigree to what would be expected given a genetic map. However there may be differences between a modern-day genetic map and that which would have been accurate (at least on average) between the time of admixture and the present, and these are not yet available. We instead introduce a method that relies on a lack of recombination in genomic regions, thereby sidestepping the potential confounders involved in

heavier reliance upon genetic maps. Having constructed human-Neandertal genealogical trees across recombinationally cold regions of the genome, we select those that we classify as being introgressed, and deduce admixture dates based on the position of two SNP types within trees across regions, and the corresponding distributions of descendants for these SNPs. Using this, we are able to date admixture at both a region and population level, across modern European and Asian populations using Maximum Likelihood Estimation: this is explained in detail in the Methods section. Given this population- and region-specific admixture dating, we are able to make some inferences about the nature of human-Neandertal mixing over the course of our evolutionary history. We detail this method fully in the next section.

5.2 Methods

We introduce a method of dating ancient admixture using information about the distribution of descendants of two SNP types on genealogical trees.

5.2.1 Some background coalescent theory

From standard coalescent theory, we can specify the expected Site Frequency Spectrum (SFS) X under neutrality of selection, with no recent population size changes and no gene flow, as follows:

$$\mathrm{X} = (x_1, x_2, ..., x_{n-1}) \tag{5.2}$$

where x_i is the number of alleles present i times across the sample of n individuals; x_1 gives the number of alleles present in only one of n individuals. We expect $x_i = \frac{\theta}{i}$, where $\theta = 4\mathrm{N}\mu$.

The expected number of mutations on a genealogy depends on two things: the mutation rate μ (the number of mutations per site, per generation) and the total length of the genealogy, L. We can work out the length of the genealogy by adding the total length of the tree when we

have 2 lineages (a timespan which we call t_2) and the total length of the tree when we have 3 lineages (t_3).

We know that $\mathrm{E}[t_2] = 2N$ and $\mathrm{E}[t_3] = \frac{2N}{3}$ from standard coalescent theory. This is because there are $\binom{n}{2}$, or $\frac{n(n-1)}{2}$ pairs of lineages (we choose one lineage n, then a second, $(n-1)$, and we are only interested in the set of unordered lineage pairs, so we divide by 2). For any individual pair of lineages, the probability that they share the same parental lineage is $\frac{1}{2N}$, thus multiplying these together gives us the probability of any pair of lineages having the same parent lineage ($\mathrm{P}(Pa)$), whilst there are n lineages. So:

$$\mathrm{P}(Pa) = \frac{n(n-1)}{2}\frac{1}{2N} \tag{5.3}$$

The inverse of this quantity is the expected length of time for which there will remain n lineages. So, the expected length of the coalescent interval during which there remain n lineages is:

$$\mathrm{E}[t_n] = \frac{4N}{n(n-1)} \tag{5.4}$$

giving us $\mathrm{E}[t_2] = 2N$ and $\mathrm{E}[t_3] = \frac{2N}{3}$. The contribution of this interval to the entire length of the tree is given as:

$$n\mathrm{E}[t_n] = \frac{4N}{n-1} \tag{5.5}$$

The total expected length of the tree $\mathrm{E}[L]$ is given as:

$$\begin{aligned} \mathrm{E}[L] &= \sum_{n=2}^{k} n\mathrm{E}[t_n] \\ &= 4N \sum_{n=1}^{k-1} \frac{1}{i} \end{aligned} \tag{5.6}$$

And therefore the expectated number of mutations, $E[Mu]$, on the genealogy is given as:

$$\begin{aligned} \mathrm{E}[Mu] &= \mu \mathrm{E}[L] \\ &= 4N\mu \sum_{n=1}^{k-1} \frac{1}{i} \end{aligned} \tag{5.7}$$

Substituting $\theta = 4N\mu$, we result in:

$$\mathrm{E}[Mu] = \theta \sum_{n=1}^{k=1} \frac{1}{i} \tag{5.8}$$

The above defines an SFS describing the number of descendants we expect to observe from a set of SNPs. It is highly applicable in our case because we assume a set of simplifying assumptions: that we observe selectively neutral alleles segregating in a population that does not have recent population size changes, that no gene flow is present (we assume it is sufficiently rare for this assumption to apply), and, as a function of the fact that admixture is rare, that each region we classify as putatively admixed corresponds to admixture from a single Neandertal lineage at a single locus.

We deduce our expected SFS from regions we have classified as non-admixed, based on our definition from Chaper 4: where the mean of the first coalescence times across two trees for that region between the population of interest and the Neandertal is $\leq$300kya, and the corresponding mean of the first coalescence times across two trees for that same region between YRI and the Neandertal is >600kya. The remaining regions not fulfilling these criteria are classed as non-admixed.

So, from standard coalescent theory, if we take a timepoint t in the past, we expect to see a particular distribution of the number of descendants from a set of SNPs across the branches existing at that time, as given by X. This varies dependent upon t, and as t increases, we see fewer lineages because of coalescent events. Intuitively then, we understand that the SFS flattens as we move backwards in time, as we find mutations which exist in lineages with multiple descendants and so appear further to the right of the spectrum. Thus, we can say

something about the age of the admixture, given the descendant distribution across a set of putative 'admixture SNPs', in comparison to our empirical descendant distribution collected from non-admixed regions.

Thus, if admixture from Neandertals into humans happened at time t in the past, the number of carriers of a single admixed haplotype (from one lineage introduced at this time t), has the same distribution as the number of descendants of any lineage at time t in the past. In order to create this descendant distribution, we define an ascertainment scheme for two SNP types that we expect to be indicative of admixture: type I and type II. We depict the placement of these SNP types on genealogical trees in Figure 5.1.

- Type I: SNPs that are homozygous in the Altai Neandertal and segregate in the human population of interest. Where admixture has occurred, these are usually expected to be found, because we expect some humans in our sample to be descendants of this admixture event, and others to be cladistically further separated.
- Type II: SNPs that are heterozygous in the Altai Neandertal and segregate in the human population of interest. Where admixture has occurred, these are sometimes expected to be found: the same rationale applies as with type I SNPs, however we expect them less frequently because the Neandertal is highly homozygous.

For both type I and type II SNPs, we assume each SNP of either type is carried by all descendants of introgression in a genomic region. This is because the amount of introgression is small (1-2%), and provided the number of lineages is not large, it would be unusual for more than one lineage to be introgressed. So any SNP inherited along this lineage and shared with the Neandertal we can assume to be present in all descendants of introgression. We are also able to assume either of these SNP types is not carried by any human haplotypes other than those descended from admixture. This is because Neandertal homozygosity is high, and their population sizes small, making the coalescence time of the two Neandertal haplotypes usually relatively small. This makes it unlikely that the second Neandertal haplotype will not coalesce with any other human haplotype before it coalesces into this group.

Type I and type II SNPs, although both indicative of admixture, can be differentiated from one another in some senses. Type I SNPs are more common than type II SNPs, because type I SNPs do not require heterozygosity in the Neandertal; we rarely see an intermingling of Neandertal haplotypes among the human haplotypes, typically coalescing with itself relatively quickly. Type I SNPs are often found further back in time than type II SNPs, again due to a low level of heterozygosity in the Neandertal: the ordinarily quick coalescence of the Neandertal haplotypes means that type II SNPs are more likely to be found at a more recent time. Thus, the presence of type II SNPs amongst a set of human haplotypes indicates a recent Neandertal introgression. Despite these differences, we expect the descendant numbers of type I and type II SNPs to match as we expect both to directly indicate the number of descendants of admixture in that region.

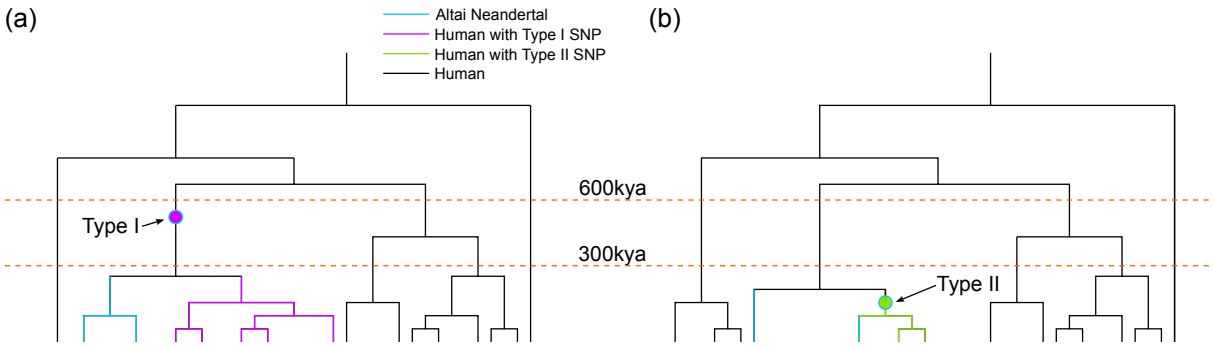


Figure 5.1: Depicting the two SNP types used to date admixture with toy genealogies. Subfigure (a) shows Type I SNPs: these are present in both Neandertal haplotypes (two blue lineages) and are segregating within the human population (purple lineages are humans with a Type I SNP, black lineages are humans without). Subfigure (b) shows Type II SNPs: these are segregating between Neandertal haplotypes as well as segregating in the human population (green lineages are humans with a Type II SNP, black without). Alternative configurations fulfilling the criteria for each SNP type are possible, as long as the conditions in (a) or (b) are fulfilled, and the coalescence event between the human and Neandertal haplotypes containing the SNP of interest occurs $\leq$300kya. Note that although this coalescence is required to occur more recently than 300kya for both SNP types, the mutation event leading to either SNP type is not, and can be either further in the past or more recently than 300kya, with no upper or lower limit.

A tree may contain any number of type I and type II SNPs. For a tree containing a nonzero number of type I SNPs, we record the minimum number of descendants we see across that set. We do this because a type I SNP can be sat deep in the past, and reflect human-Neandertal coalescences which occur further back in time than the human-Neandertal split, thereby including individuals who do not contain a real signal of admixture, and being a false signal of admixture.

For this reason, we think of type I SNPs as providing an upper bound to the number of descendants of the human-Neandertal introgression event, and therefore take the minimum of this set of descendant numbers. Note that although it is theoretically possible for this set of SNPs to contain only SNPs that occurred before the human-Neandertal population split, that this is rare in reality, because coldspots are sufficiently long ($\geq$21kb) for this to be statistically unlikely in a region where introgression occurred. For example, given 500,000 years of potential mutations since introgression, the expected number of mutations in 21kb is $500{,}000 \times 1 \times 10^{-9} \times 21{,}000$, which gives 5.25, making the probability of no mutation $e^{-5.25}$, or 0.005, i.e. very small. For the set of type II SNPs in a pair of trees, we record both the minimum and maximum number of descendants across the set.

CEPHi provides a user-specified number of possible genealogical trees for each introgressed region in a population, we output two per region. For each pair of trees in that region (built from the same haplotype information), we record the number of type I and type II SNPs present, the minimum number of descendants contained in the set of type I SNPS, and the minimum and maximum number of descendants contained in the set of type II SNPs. We do this for each pair of trees across all admixed regions in that population. These are our introgression frequencies; how many individuals contain Neandertal introgression in this region. This descendant distribution is typically decaying; many lineages with few descendants, and fewer with many. We use this resulting distribution of descendants to infer a date of admixture from Neandertals into modern humans using Maximum Likelihood Estimation, detailed below.

Based on the presence of type I and type 2 SNPs, we define four different sets of admixed regions. These are each subsets of the initial set of introgressed regions defined by comparative coalescence times between the population in question, YRI, and the Altai Neandertal. We name them t1, t2, G, and GG, where the admixed set t1 is the largest, and GG the smallest; they consist of the following:

- t1 = contains all admixed regions where a type I SNP exists.
- t2 = contains all admixed regions where a type II SNP exists.

- G (gold) = contains all admixed regions where there are nonzero type II SNPs, and the maximum number of descendants of type II SNPs matches the minimum number of descendants of type II SNPs.
- GG (double gold) = contains all admixed regions where there are nonzero type I and II SNPs, and the minimum number of descendants of type I SNPs matches both the maximum and minimum number of descendants of type II SNPs.

5.2.2 Data filtering

Before separating the data into admixed and non-admixed sets, we first apply some filters. For each analysis we are comparing two populations: YRI and a second, admixed population. We first reduce the set of regions to only those that were included in both *CEPHi* runs for those respective populations. We then remove regions in which the total number of SNPs does not differ by more than twofold from the expected number of SNPs in any given genomic region of that size, using Watterson's estimator (the number of SNPs divided by the size of the genomic region).

We then separate the data into two: admixed and non-admixed regions. We mark regions as admixed using the classification from Chapter 4: where the mean of the first coalescence times with a Neandertal across two trees is $\leq$300kya in the population in question, and for that same region the mean of the first coalescence times for YRI with the Neandertal is >600kya.

To help remove noise from our data, we remove regions from each analysis where any type II SNP is present in YRI; the YRI population may have some admixture through back migration and this could confound analysis.

5.2.3 Using Maximum Likelihood Estimation to infer admixture dates

We then use Maximum Likelihood Estimation to infer the date of admixture as indicated by a particular set of admixed regions. We have four ways of classifying sets of admixed re-

gions, as given above, all of which are subsets of our admixed regions as defined by having a short (≤300kya) coalescence time of an individual of the population in question and the Altai Neandertal, and a long (>600kya) coalescence time between a YRI haplotype and the Altai Neandertal.

For a particular population, we call a set of admixed regions i=1,2,...,n. For each region i, we count an observed number of individuals carrying a likely Neandertal SNP, k_i. Across all n regions, we obtain the distribution of descendants carrying putative Neandertal admixture. Given our ascertainment criteria for recombinationally cold regions, we can assume regions are independent, and therefore can write down the probability of the set of k_i we observe, if admixture happened at time t. This is given as:

$$L(k_1, k_2, ...k_N; t) = \prod_i^N P(k_i \text{ descendants} \mid t) \tag{5.9}$$

where $\prod_i^N P(k_i \text{ descendants} \mid t)$ is simply the joint probability of each observed number of descendants, given admixture occurred, at time t. For each human population, we calculate this likelihood across a set of 995 possible admixture times between 5,000 and 1,000,000 years ago in 1,000 year intervals.

To calculate $P(k_i \text{ descendants} \mid t)$, we must estimate, for each population, the distribution of the number of descendants there are today, from a single lineage at time t ago. To do this, we use the empirical distributions at time t for trees built for those regions classified as non-admixed; we use two trees per region for this. The time with the highest likelihood gives the maximum likelihood estimate of the admixture time between a specific population and the Altai Neandertal.

We additionally calculate these likelihoods conditional on different maximum numbers of descendants, between 20 and 100 in intervals of 10. This essentially removes the descendant frequency above some value. We do this by calculating the empirical distribution of descendants as before, but cutting the table down to that number of descendants and normalising the counts of descendants. It is important to investigate robustness to the cutoff used (and

thereby calculate likelihoods at a range of descendant cutoffs) because most of the descendant distribution clusters on the left hand side of the distribution, meaning the tail is expected to carry little information. Moreover, the right hand tail of the distribution might be enriched for false positive regions which may have incorrectly been inferred as representing admixture. As mentioned above, in this approach to dating admixture, we assume neutrality of regions. Although there have been suggestions of purifying selection acting upon Neandertal introgression (Sankararaman *et al.* [2014]), it is likely that there are very few positively selected loci amongst the introgressed regions, and these may tend to be in the right hand tail of the descendant distribution, thereby not affecting the model fit for lower cutoff values.

We first perform simulations to demonstrate the utility of our method, and we then apply it to analyse the 1000 Genomes data across 14 human populations.

5.2.4 Using coalescent simulations to demonstrate a new method of dating admixture

We first use simulations to validate our method, which resemble those performed in Chapter 4. We use *scrm* (Sequential Coalescent with Recombination Model) (Staab *et al.* [2015]) to simulate segregating sites across a set of regions that is matched in number and region size to those in our real data, for three populations (YRI, GBR, Neandertal), of sample sizes 88, 89, and 1 individuals respectively (haplotype numbers are twice this). We then separate the output into two population pairings: GBR-Neandertal and YRI-Neandertal, and run it through *CEPHi*, allowing early coalescences between all haplotypes (command lines used are in Appendix D), to produce genealogical trees depicting the relationships between those haplotypes. This set of simulations differs from the first in that Neandertal effective population size (N_e) is set to be very large (10,000,000) between time zero (present) and the time at which the Neandertal went extinct (50kya), rather than being set to 1. Because *scrm* does not allow extinction, we use a very large number to prevent coalescence between the Neandertal haplotypes more recently than 50kya.

5.2.5 Dating admixture from Neandertals across 14 modern human populations

We first present the empirical descendant distributions across a range of times (5, 25, 50, 75, 100, 150, 1000kya). We next give the range of admixture date estimates we obtain for each population using these empirical distributions. For each set of admixed regions and across all populations under consideration, we plot the maximum likelihood estimate of the admixture date between the Altai Neandertal and the population in question at 10 descendant cutoffs between 20 and 100; a cutoff of 20 descendants uses information from all regions inferred to have $\leq$20 descendants from admixture. We then present the inferred descendant distributions for all 4 sets of admixed regions.

Lastly we show the log likelihood curves for each population, with 95%, 99%, and 99.9% Confidence Intervals (CIs). We calculated the CIs using the likelihood ratio (LR) statistic which is approximately distributed as chi-squared when sample size is large. Our sample sizes are equal to twice the number of trees across the number of admixed regions, in the order of hundreds or - in most cases - thousands. An approximate 95% CI, for example, for our parameter (time of admixture), includes all the possible values of this parameter for which the log likelihood function is a distance away from by no more than 1.92 units (from standard lookup tables). This comes from the definition of the LR, which, for the 95% interval is given as $2log\frac{L(\hat{\theta};x)}{L(\theta;x)} \leq 3.84$, where θ=a particular time in the past, $\hat{\theta}$=time in the past with the maximum log likelihood, and x is our data. This definition of the LR is equal to $2\big(l(\hat{\theta};x) - l(\theta;x)\big)$. Thus, by rearrangement we calculate $l(\theta;x) - l(\hat{\theta};x) \geq -\frac{-3.84}{2}$. Equivalent values for 99% and 99.9% CIs widen these bounds.

5.3 Results

We first present the results from a set of coalescent simulations to demonstrate the efficacy of the method. We then give our inferred dates of admixture using data from 14 modern human populations from the 1000 Genomes Project.

5.3.1 3-population simulations: admixed and non-admixed scenarios

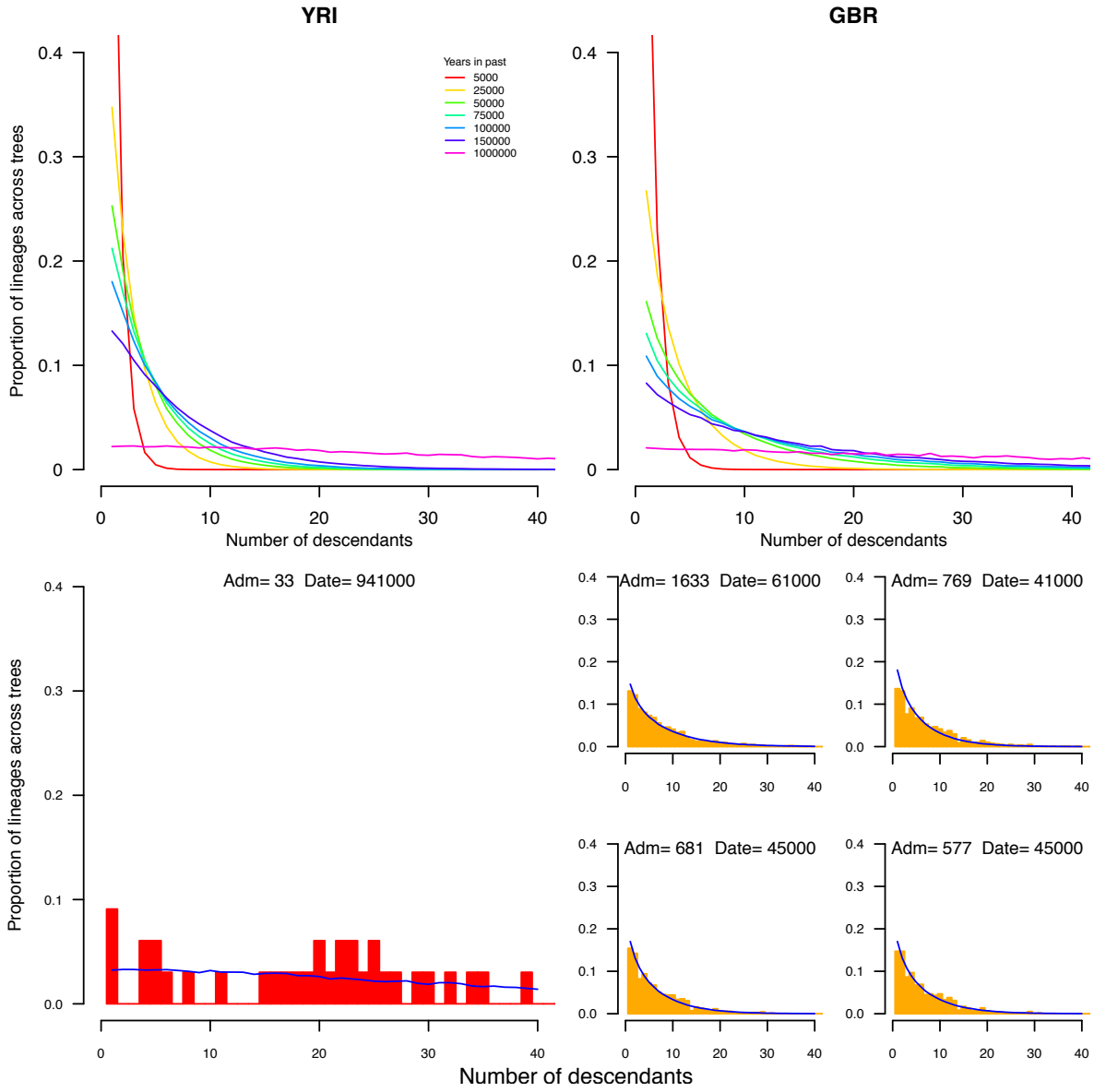


Figure 5.2: Simulations: Empirical and admixed descendant distributions using a cutoff of 40 descendants. Left column: YRI, right column: GBR. Top row: Empirical descendant distributions using regions classified as non-admixed, from 5,000-1,000,000 years ago. The bottom left quadrant gives the admixed descendant distribution for YRI, using the t1 set of admixed regions. Plots for YRI using t2, G, and GG are not shown as there were insufficient regions in each set (<5). In the bottom right quadrant we show admixed descendant distributions using t1, t2, G, and GG sets for GBR (from right to left, top to bottom). Blue lines indicate the empirical descendant distribution for the inferred admixture date. 'Adm' gives the number of admixed regions plotted.

In Figure 5.2, we show comparisons between the empirical distribution of descendants across

non-admixed regions between 5,000-1,000,000 years ago, and the descendant distributions we obtain from four sets of admixed regions. We first note in the empirical distributions (top row) that the distributions are as expected, with more recent times showing a sharper decay in the number of descendants, therefore having a greater proportion of lineages with fewer descendants. As we move backwards in time, through the occurrence of coalescence events, the distribution flattens as lineages gather more descendants. The empirical distribution for GBR shows a quicker flattening as we move backwards in time, likely due to the lower variation in the GBR population (it is generally accepted that variation outside Africa is a subset of that within it). The lower half of Figure 5.2 gives the descendant distributions for 4 sets of differentially defined admixed regions for YRI and GBR. We see very few admixed regions for YRI, and in fact we only present the figure for the t1 set, as there were too few regions in other sets to be informative (two in t2, zero in G and GG). This set gives a date of 941kya. For GBR, estimated admixture times with the Altai Neandertal are 61kya and 41kya for the t1 and t2 sets, and at exactly 45kya - the actual simulated time of admixture - for the gold (G) and double gold (GG) sets.

We simulated admixture occurring between GBR and the Altai Neandertal at 45kya, using 3-population coalescent simulations (YRI, GBR, Altai Neandertal). No admixture was simulated between the Neandertal and YRI populations. The population split time used for GBR and the Altai Neandertal is 348,264 years ago, and between YRI and the Neandertal is 712,936 years ago (both as given in Chapter 3). Figure 5.3 shows the inferred admixture times obtained through Maximum Likelihood Estimation, for YRI and GBR, across four sets of admixed regions.

It is clear that this method infers admixture dates correctly, both when using the distribution of descendants from admixed regions where type II SNPs are present, and where we implement the gold and double gold standards of specifying a set of admixed regions. Each of these three subsets of admixed regions are shown to be robust to different descendant number cutoffs, with all points placed very closely. For example, for the t2 set of admixed regions, the inferred admixture date is 43kya when using a descendant cutoff of 20 and 30, and 41kya when using all remaining descendant number cutoffs. For set G, all cutoffs from and inclusive of 40 descendants

give a date of 45kya, as is the case for set GG. We see more variation in estimated admixture dates when using all admixed regions containing type I SNPs, as this set will include some older type I SNPs, thereby overestimating the number of descendants of a region. This effect is increased as we increase the descendant number cutoff. For example, using the t1 set for GBR, we see date estimates of 69kya with a 70 descendant cutoff, 90kya with 90 descendants, and 139kya with 100 descendants. This is also the case in YRI, where the admixture date is 894kya at a cutoff of 50, 973kya at 70, and 989kya at 100. Comprehensive sets of date estimations are given in Appendix D.

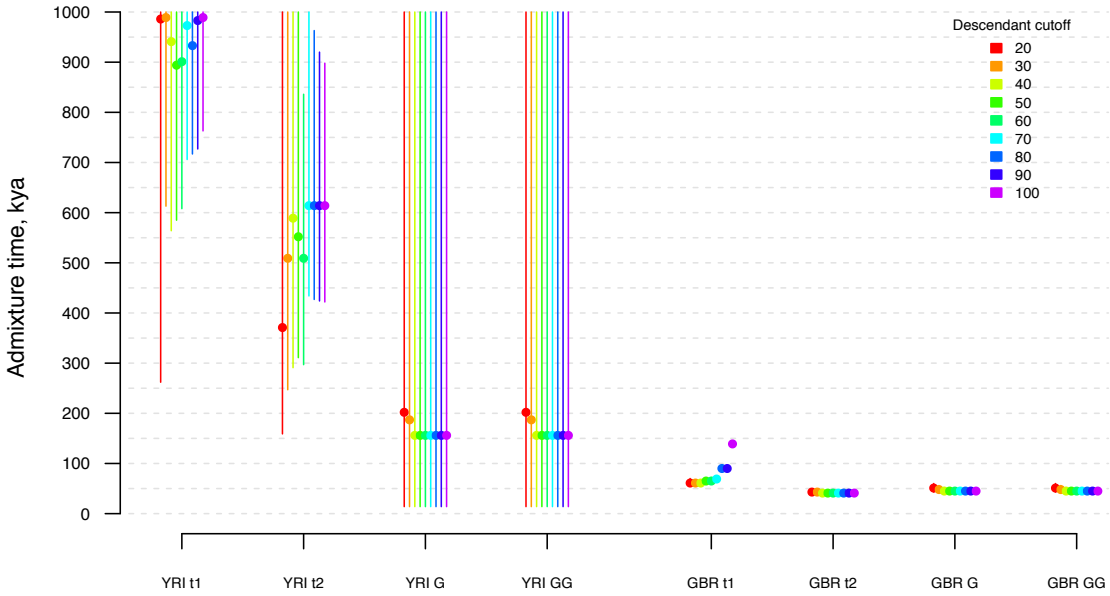


Figure 5.3: Simulating admixture between GBR and the Altai Neandertal 45kya: range in inferred MLE of admixture times for YRI and GBR, dependent on the number of descendants used in the analysis. YRI t1 = admixed regions containing type I SNPs, t2 = admixed regions containing type II SNPs, G = adimxed regions with a matching maximum and minimum number of descendants using type II SNPs, GG = using the G condition, additionally matching the (minimum) number of descendants using type I SNPs. Vertical lines indicate 95% confidence intervals, plotted for all points.

Where no admixture was simulated - between YRI and the Altai Neandertal - we see much later estimated dates of admixture. Where the population under consideration is other than YRI, the definition of an admixed region uses the YRI-Neandertal coalescence time as a reference point, being required to be long (>600kya). For YRI, we define an admixed region as having a short (≤300kya) inferred mean coalescence time with the Neandertal (when averaged across the two trees), with GBR also showing a short (≤300kya) inferred mean coalescence time with the

Neandertal (when averaged across the two trees). We see significantly later dates when using the t1 and t2 sets of admixed regions than when using G or GG, and very wide confidence intervals for YRI, where they are plotted but barely visible for GBR due to being encouragingly tight; this is likely due to there being very few SNPs used in the YRI case, and means we cannot read significance into the variation in dates for YRI. For example, for type I SNPs in GBR, we see 3,566 admixed regions, with a mean of 21 SNPs per region, and a median and maximum of 18 and 141 SNPs respectively. YRI by contrast has 142 admixed regions, a mean of 10 SNPs per region, and a median and maximum of 9 and 31 SNPs respectively. Because each set of admixed regions (t1, t2, G, GG) is smaller than the last, these numbers reduce when using sets other than t1. For G and GG for YRI, we result in fewer than 5 admixed regions to use for likelihood calculations, whereas for GBR the equivalent numbers are 680 and 577 respectively.

In Figure 5.4 we show the maximum likelihood searches for admixture times across all four sets of admixed regions, for YRI (first row) and GBR (second row), for 40 descendants. For GBR, admixture is conclusively inferred at, or close to, the time at which it occurred in the simulations; we see very peaked curves. For YRI, likelihoods remain high for times further back in the past, and 1mya is always within each CI, consistent with the fact that YRI has no admixture in these simulations.

5.3.2 Dates of Neandertal admixture across 14 modern human populations

We applied this method to real data, using the same set of 14 modern human populations from the 1000 Genomes Project as have been used throughout, and which are detailed in Appendix B.

Figure 5.5 presents the empirical descendant distributions for YRI, LWK, GBR, and CHB, two African populations, one European, and one Asian. These are created from sets of non-admixed cold regions, as explained in the Methods section. We use individual populations from each continent that are representative of that continent in this figure; in fact there is very little difference between distributions within continents. It is evident that these distributions match the simulated non-admixed distributions very closely, and we also see the same differences

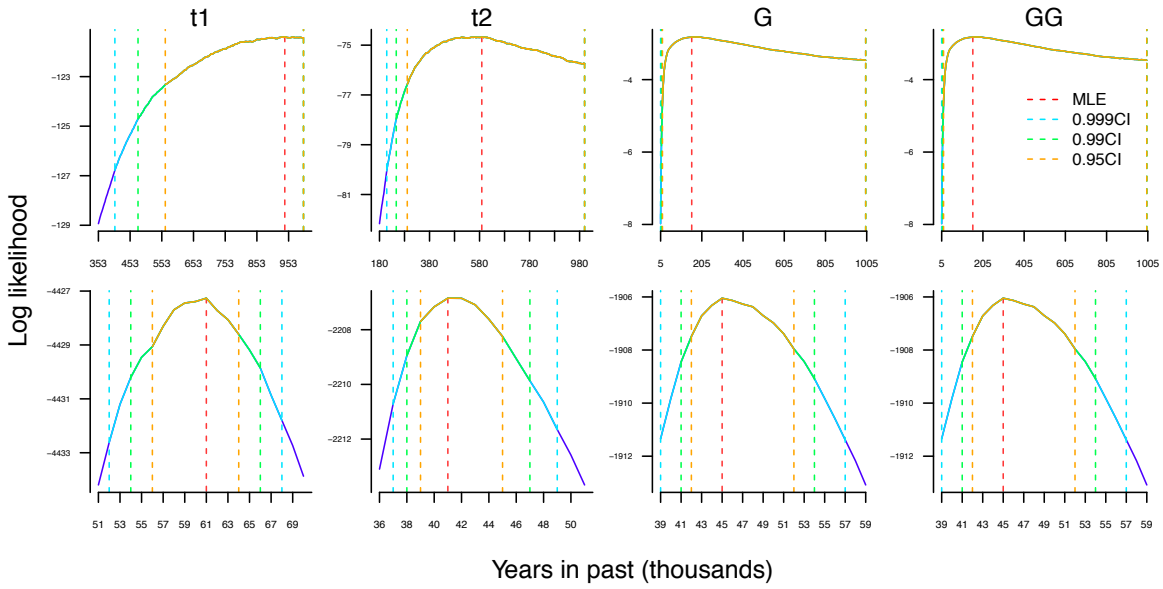


Figure 5.4: Simulations: Log likelihood curves for YRI (top row) and GBR (bottom row) using 40 descendants. From left to right there are four columns: t1 draws the likelihood curve using those admixed regions containing type I SNPs, t2 the equivalent for admixed regions containing type II SNPs only, G uses admixed regions where the number of descendants is nonzero and has a matching number of descendants for the minimum and maximum for type II SNPs, GG uses admixed regions where this is true, with this number also matching the number of descendants for type I SNPs. The orange, green, and blue dashed lines indicate the 95%, 99%, and 99.9% confidence interval boundaries. The red dashed lines show the maximum likelihood estimate of admixture time. Plots show the region at the peak of the maximum likelihood search.

between YRI and GBR: the distributions being steeper in YRI. GBR matches closely to CHB, which is to be expected, given previous results and their likely somewhat similar evolutionary history. We note the existence of upticks at the left hand of the distributions for GBR and CHB, probably representative of an excess of singletons in these populations.

Taking GBR as an example population, this method infers an admixture date with the Altai Neandertal between 51-68kya, varying dependent upon which set of admixed regions is used. We refer back to Figure 3.10 in Chapter 3 which shows a genealogical tree built between GBR and the Altai Neandertal for chromosome 1, between positions 10010344-10034056: a region we classified as admixed. In this region, for GBR, we have 6 type I SNPs (seen on the black branches above the orange dotted line), and 2 type II SNPs (seen on the black branch below the orange dotted line). As seen on the tree, both SNP types show 4 human descendants of admixture in this population, to give a real example of the workings of the method: the

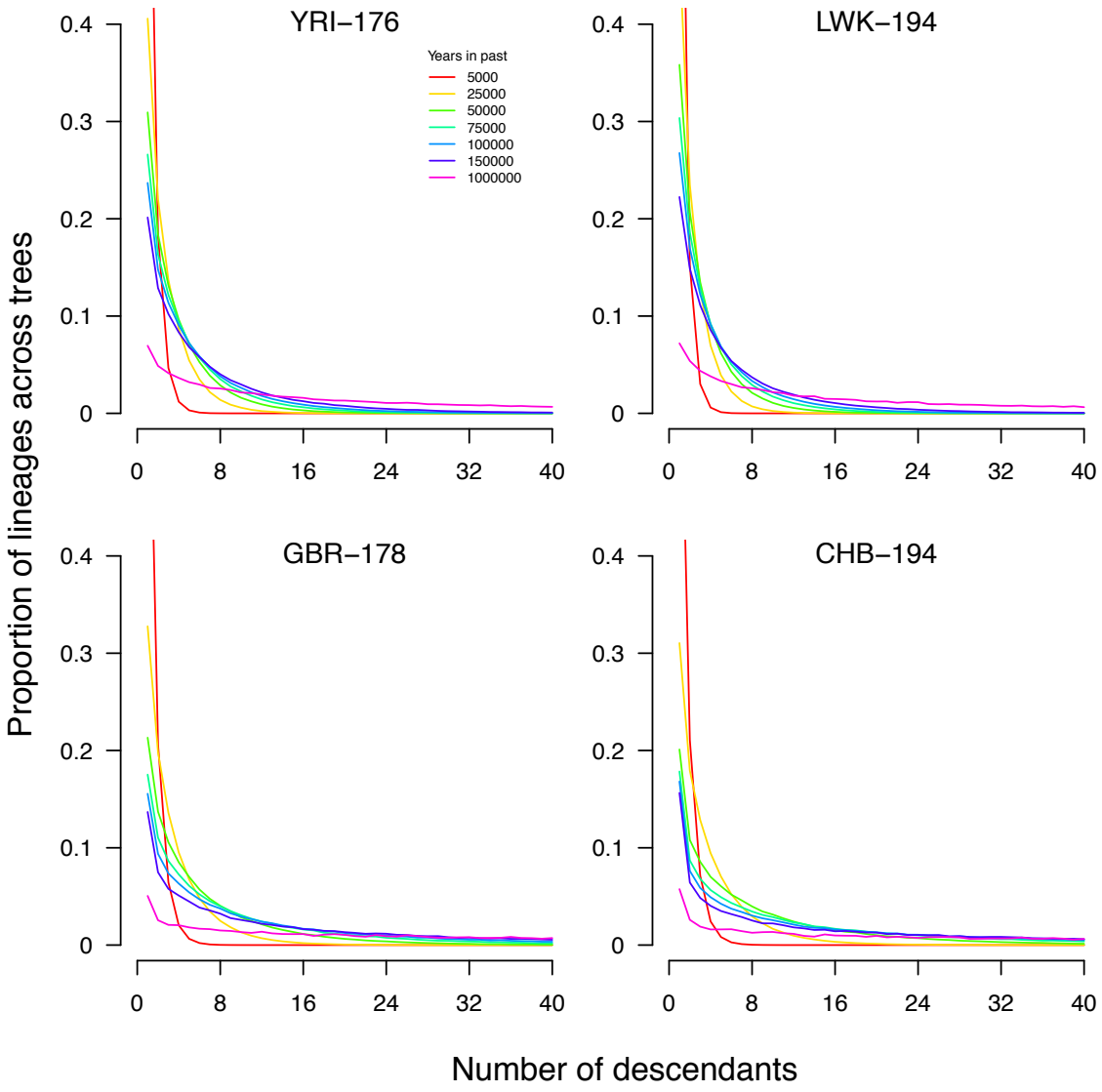


Figure 5.5: Empirical descendant distributions across 4 populations (YRI, LWK, GBR, CHB) from 3 continents (AFR, EUR, ASN). Created from non-admixed regions, these distributions are used as a reference point from which to match the descendant distributions from the admixed regions, done per population as explained in the Methods section. We plot a selection of times between 5,000 and 1,000,000 years in the past. Each subplot title gives the number of haplotypes present in that population.

placement of type I and type II SNPs and their descendant numbers.

Figure 5.6 gives the range of admixture dates inferred across ten descendant-number cutoffs between 20 and 100. Within each population we show the range for each of four sets of admixed regions: t1, t2, G, and GG.

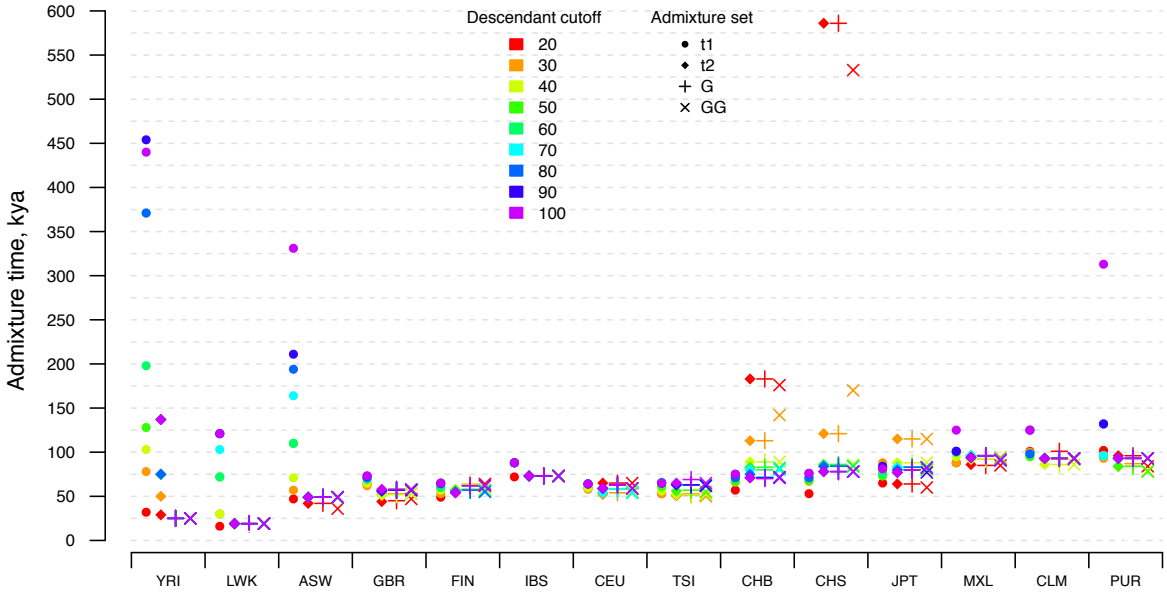


Figure 5.6: Admixture times inferred for 14 modern human populations, for four subsets of admixed regions, across a set of descendant-number cutoffs between 20 and 100. Continents are Africa, Europe, Asia, Americas, from left to right. Within each population are the maximum likelihood estimates of admixture time between the Altai Neandertal and that population for subsets t1, t2, G, and GG.

Firstly, we note that for YRI, ASW, and to some extent for LWK, we see large ranges stretching between young and old inferred admixture dates for the t1 set, and also in t2 in YRI. For the t1 sets in all three cases, admixture dates increase with the number of descendants included in the calculation as we expect, and as occurred in the above simulations. For YRI, admixture is defined as mentioned above: we require estimated coalescence times between the Neandertal and YRI of $\leq$300kya and between the Neandertal and GBR of $\leq$300kya. These estimates reduce to very recent admixture dates in YRI for the G and GG sets, and this occurs across the t2, G, and GG sets for LWK and ASW; the resulting admixture dates for ASW are slightly older. The Luhya of Kenya are a Bantu speaking group have sub-Saharan, Nilo-Saharan, North African/Middle Eastern, and Coptic ancestry (Dobon *et al.* [2015]), ASW carry Native American and European ancestry alongside African (Bryc *et al.* [2015]), while YRI are almost entirely sub-Saharan (Dobon *et al.* [2015]). LWK, notably, has the most recent admixture dates of any population, likely revealing introgression of Neandertal haplotypes through back-migration of north or non-African populations (Hervella *et al.* [2016]; Llorente *et al.* [2015]), as seen in Chapter 4. The five European populations show a tight set of ranges of inferred admixture

times, predominantly between 50-75kya.

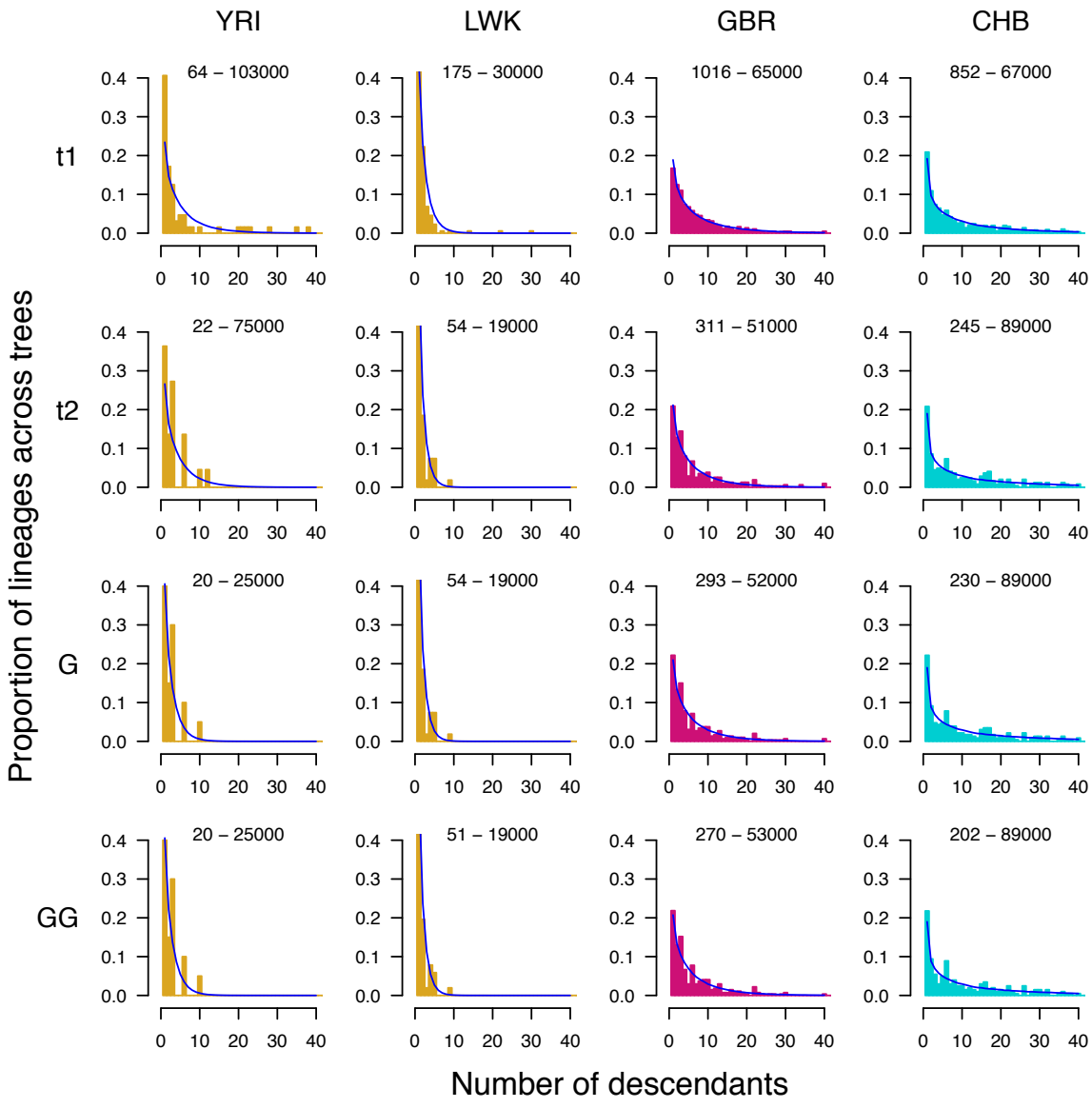


Figure 5.7: Descendant distributions for admixed regions from four populations (YRI, LWK, GBR, and CHB), one from each of the continents (in columns from left to right). Rows from top to bottom give the distributions for each set of admixed regions: t1, t2, G, and GG. Numbers at the top of each subplot give the total number of admixed regions shown in the plot, and the inferred admixture date. The blue line shows the distribution of regions expected from that admixture date, taken from the empirical distributions in Figure 5.5.

The three Asian populations (CHB, CHS, JPT) show an intriguing result. We note firstly that there is a greater dependency on the cutoff used in comparison to the European populations, even when using the G and GG sets. Two reasonably significant anomalies are clear in CHB,

CHS, and to some extent JPT, for the t2, G, and GG admixture sets with 20 and 30 descendants for CHB and CHS, and for 30 descendants in JPT, with somewhat to very much older dates of admixture being inferred. Interestingly, and contrary to the pattern seen across other populations, admixture dates decrease as the descendant number cutoff increases for t2, G, and GG. This is also in contrast to the patterns seen for each t1 set for CHB, CHS, and JPT, which follow the pattern seen in other populations: the admixture date increases slightly with descendant number cutoff. When an older date is inferred at lower descendant number cutoffs, this is because the distribution of descendants for the set of admixed regions is flatter, so we have fewer lineages with small numbers of descendants. We suggest firstly that using only lineages up to those with 20 or 30 descendants is simply too few and therefore unreliable - this is supported in the 95% confidence intervals we calculate for each cutoff, given in Appendix D. Thus, we use 40 for descriptions from hereonin. An older admixture date for Asians as compared with Europeans may be due to more effective purging of introgressed Neandertal sequence due to a larger effective population size in Europeans. Equally, it may be as a result of genetic drift in Asians, for the same reason.

The American populations (MXL, CLM, PUR) show later admixture times in comparison to the majority of the European and Asian populations. The fact that these populations are recently admixed, being a combination of European, Native American, and African populations, invalidates our coalescent modelling assumptions of no admixture other than that from Neandertals. For this reason we omit them from discussion.

Figure 5.7 presents descendant distributions of four sets of admixed regions across four human populations: one African, one European, and one Asian, using YRI as a reference population. The barplots give the actual distribution of the admixed regions in each of these populations, and the blue curves give the distribution we find at the inferred admixture date, taken from our empirical distribution of non-admixed regions. Figure 5.8 gives the corresponding maximum likelihood estimate searches for the date of admixture for each of these populations, with 3 sets of confidence intervals: 95%, 99%, and 99.9%. A full list of admixture dates and their corresponding 95% confidence intervals are given in Appendix D.

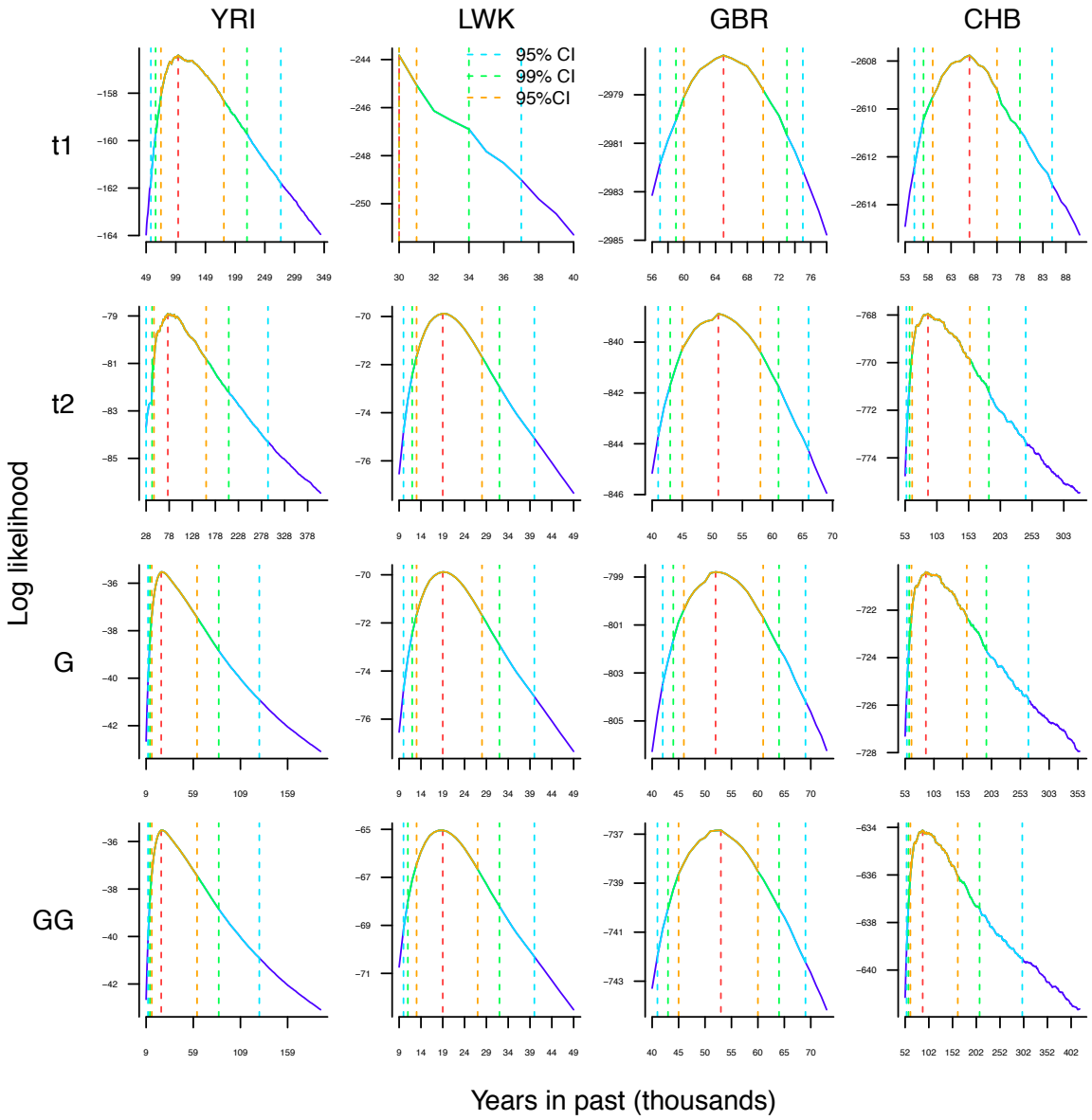


Figure 5.8: Maximum Likelihood Estimation searches for 4 of 14 1000 Genomes populations: YRI, LWK, GBR, CHB (columns from left to right). Rows from top to bottom use the different sets of admixed regions: t1, t2, G, and GG. Confidence intervals are given by the blue, green, and orange dashed lines, corresponding to 99.9%, 99%, and 95% levels respectively. The red dashed line pinpoints the admixture date with the highest log likelihood. Note that for LWK t1, the point to the left of that of the MLE lies at [29,000,-350.97], i.e. has a far lower likelihood than the MLE, and so is not included in the plot. Plots show the region at the peak of the maximum likelihood search.

In Figure 5.7, as we move from t1 through to GG, the set of admixed regions becomes smaller, and should move closer to the true date of admixture. We see dates just over 50kya for GBR across the t2, G, and GG sets for a descendant cutoff of 40, similar to other European pop-

ulations. 95% confidence intervals for GBR are 45-61kya across these three sets, using a 40 descendant cutoff. CHB shows a notably later date of admixture than that for GBR. The three latter groups of admixed regions (t2, G, and GG) for CHB each infer a date of 89kya (95% CIs between 63-163kya), using a descendant cutoff of 40. This is a notable result; significantly different admixture dates between European and Asian populations suggest there may have been separate interactions between these populations and Neandertals, rather than a single period of interaction in the Middle East before the Eurasian population split, thought to have occurred somewhere between 37-45kya (Fu *et al.* [2014]; Seguin-Orlando *et al.* [2014]). The Luhya from Webuye, Kenya show a very recent date of 19kya (13-28kya 95% CI for a cutoff of 40 descendants using t2), which is potentially indicative of populations back migrating into Africa from Eurasia. Notably we have very few admixed regions for LWK (51-175), but the 95% CIs across all four sets give a range of 13-31kya, so our inference is not detrimented significantly by this limitation. Back migration from Eurasia may make this date difficult to interpret accurately, as this is not modelled, but the inferred date should be at least as old as back-migration, and likely less than or equal to the date of real admixture.

5.4 Discussion

In this final chapter we have introduced a new method for dating ancient admixture using SNP information from genealogical trees linking Neandertals with modern humans. We have confirmed that we can infer accurate date estimates for admixed populations using coalescent simulations, and have thus applied the method to 14 modern human populations from 4 continents, with some corroborative and some intriguing results.

A previous paper (Sankararaman *et al.* [2012]) gave a date estimate of 37-86kya for Europeans, and our range (across all descendant cutoffs) is corroboratively similar: 38-96kya (95% CIs), and if we exclude IBS due to it containing only 14 samples, the range changes slightly to 38-80kya (95% CIs). (Given the consistency of results we see with sample sizes of 55 and upwards, we might recommend avoiding very small sample sizes for use with this method, although

further tests would be required to confidently define a minimum number.) Our method enables us to infer dates for individual populations, and thereby have a reasonably fine-grained set of results. As we look across continental admixture dates, we see that both Han Chinese populations (from Beijing and Shanghai respectively) reveal some interesting questions. The inferred admixture dates for these populations are significantly later than those of any of the five European populations. The Han Chinese from Beijing (CHB) show dates ranging between ∼59 and 163kya (95% CIs), the Han Chinese from Shanghai (CHS) ∼61-162kya, where the Europeans show date ranges closer to ∼45-75kya (using a descendant number cutoff of 40 across admixed sets); so substantially more recent.

These differences may be explicable through separate admixture events between Neandertals and various European and Asian populations. We know from Chapter 4 that there are a substantial number of shared regions between Asian and European populations, which are most parsimoniously explained through some admixture having occurred between Neandertals and modern humans before the European-Asian human split somewhere before ∼37kya (Seguin-Orlando *et al.* [2014]). This of course does not prevent further interactions between the cousin species subsequent to this population split, and the more recent times inferred for European populations may be indicative of further instances of admixture with Neandertals subsequent to this population split. Further examination of the influence of the basal Eurasians on individual European populations may contribute to the differences seen with regard to Neandertal ancestry across Eurasian populations (Lazaridis *et al.* [2016]). One benefit of this method is that we can apply our method to date any set of introgressed regions, so a future step could be to date those regions which are specific to European populations, and those specific to Asian populations. With regard to the African populations, the very recent dates for admixture into the Luhya of Webuya, Kenya, for example, suggest these introgressed regions are likely the result of backmigration into Africa from Europe (as we saw in Chapter 4), as Neandertals were long extinct by this time. The distribution dating admixture into YRI is of poor fit, as we saw in Figure 5.5, so we can use LWK to better date back migration. Again, dating those regions classified as introgressed in both GBR and LWK would provide further insights.

With the introduction of a new method, we find a number of details and future lines of enquiry which would serve to further investigate this method and its applications. For example, with regard to the simulations, we simulated a 3-population scenario, leaving YRI as an non-admixed population, and introducing 2% admixture into GBR from Neandertals at 45kya. Additional simulations using other Eurasian populations, and across other admixture dates could be used to further corroborate the utility of the method. We are somewhat limited in this instance by the YRI-GBR split time (∼79kya) and the Neandertal extinction date (∼44kya), so the bounds to work within are relatively narrow, but simulating admixture at 20-79kya at 5kya intervals, for example, could potentially corroborate the accuracy of the date inference further.

Secondly, our t1 sets of admixed regions inevitably include SNPs at a range of dates: some very recent, and some deep in the past. Although we expect very few old type I SNPs, they may be present, and are less likely to be indicative of admixture (and certainly are not if they are older than 200ky, the age of *Homo sapiens*). The influence of these SNPs inevitably increases as we increase the descendant number cutoff, because where we accept SNPs with larger descendant numbers, we are often moving further back into the past. It may be worth investigating the effect of removing very old type I SNPs from analysis. Another way to streamline our dataset might be to filter our region set to consist of only those containing more than 20 SNPs. We can see the very wide confidence intervals for the admixture date estimates for YRI in Figure 5.3, and this change may increase reliability of these date estimates, making confidence intervals narrower.

However, we find substantial utility in this method. Having isolated sets of regions shared between populations in Chapter 4, for example, we are able to date the admixture events leading to their initial introgression from Neandertals. Specifically, we could date those admixed regions which are shared among European and Asian populations to show us which regions were introgressed into modern humans before the Eurasian population split. We could then date sets of regions unique to Europeans and Asians separately, to investigate whether these regions were introgressed significantly later. This approach could potentially be taken to a population-specific level. Furthermore, applying the same principle to regions shared between African and non-

African populations, we are able to date back migration into Africa from various sources. We could also make these comparisons intra-continentally to map the transmission of Neandertal regions through the human population. This would enable a more intricate painting of the probably complex picture of human-Neandertal interactions.

Furthermore, we could potentially apply a variant of the method introduced here to search for regions which are introgressed from humans into the Altai Neandertal, as addressed in Chapter 4 with regard to Kuhlwilm *et al.* [2016]. In these regions, we firstly (and similarly) expect to see recent coalescence times between Neandertals and modern humans. We would also expect to see many type II SNPs, because introgression would in most cases affect only one Neandertal haplotype. A large number of type II SNPs, as with a large number of type I SNPs in the current analysis, means we would see variation in the number of descendants of type II SNPs. For this reason, we might expect to see old dates for t1 and t2 SNP sets, as we do with t1 sets in the current analysis, and we would certainly expect sensitivity to the cutoff for t1 and t2 sets for the same reasons. Adapting this method to investigate introgression from humans to Neandertals however could provide significant further insights.

Extension of the current analysis could be performed by using datasets with a larger number of populations. The simplest extension would involve using the Phase 3 1000 Genomes data, which is constituted of 26 populations, 12 of which are additional to the dataset used here, including 4 additional African populations, 7 additional Asian populations, and 1 additional American population. Additionally, we could employ the HGDP-CEPH dataset representing 51 populations from Africa, Europe, the Middle East, South and Central Asia, East Asia, Oceania, and the Americas (Cann *et al.* [2002]). This adds value in that it contains data from the Oceanian continent, which has not been used in our analyses, as well as checking for consistency of conclusions across the continents already explored. In addition, simply by significantly increasing the number of samples in the analyses, and providing data across a larger and more varied set of human populations, employing the 787 geographically varied human genomes very recently released (Malaspinas *et al.* [2016]; Mallick *et al.* [2016]; Pagani *et al.* [2016]) would inevitably increase the intricacy of the global picture of admixture history

between Neandertals and modern human populations.

CHAPTER 6

Discussion

In this thesis, we have sought to identify gaps in our knowledge regarding admixture between modern humans and Neandertals, and we have successfully examined some major points on this subject. We have done this using an adaptation of a new method - *CEPHi* - which builds genealogical trees in recombinationally cold regions of the genome linking populations of modern humans and Neandertals with mutation and coalescence events, thus allowing for the inference of various pieces of information surrounding their interactions.

Before our main analyses, we first showed that a set of ∼9,000 human coldspots with a recombination rate of ≤0.2cM/Mb present the same patterning of similarity to the Neandertal that is seen genomewide (Green *et al.* [2010]; Prüfer *et al.* [2014]): non-African individuals are

significantly closer genetically to Neandertals than are African individuals (Chapter 2). This provided us with support for using coldspots throughout our analyses, using the assumption that these coldspots are likely to have retained intact introgressed regions from Neandertals since the occurrence of admixture between the cousin species.

From Chapter 3, we see that when the Eurasian Neandertal predecessor (potentially *Homo heidelbergensis*) split from the modern human lineage, this population suffered a strong bottleneck, as we also see with modern humans on their African exit. This split may therefore coincide with an early Neandertal exit from Africa. Related to this, in Chapter 4 we found unusual regions of the genome showing the Neandertal haplotypes intermingled with both Africans (YRI) and Europeans (GBR), which may represent a complex separation between modern humans and Neandertals. Nevertheless, we can state that the population ancestral to Neandertals, having evolved into *Homo neanderthalensis*, never reached a large effective population size in Eurasia, and in fact following its original severe bottleneck, slowly decreased until the species became extinct. Given that the species had such a small effective population size, it may be that they were fundamentally maladapted and therefore inherently unable to persist, with factors such as competition, climate, and disease contributing to their demise. However, it is overwhelmingly clear that admixture also occurred between Neandertals and modern humans to a large enough extent that its signatures are highly visible in modern human non-African genomes. To a much lesser extent, we see Neandertal introgression into some African genomes, which has a more recent origin (Chapter 5), and so must have occurred through recent back migration of European populations into Africa (Chapter 4).

In contrast to the late population split time we see between YRI and the Altai Neandertal (712kya), we see definitively more recent split times between non-African populations and the Neandertal (∼300-500kya). These represent some average of the YRI-Neandertal split time and the time of admixture between non-African and Neandertals. We are able to strongly support the conclusion that admixture occurred via clearly bimodal distributions of coalescence times between non-Africans and the Altai Neandertal (Chapters 3 and 4); the heavier peak of the distribution further back in the past being representative of non-admixed regions, and

the sharper and more recent peak being constituted of those regions that were admixed from Neandertals into the human population.

The genealogical trees from *CEPHi* provide a wealth of information regarding the interactions between humans and Neandertals, and we investigated a substantial amount of this in Chapter 4. Having confirmed the accuracy of *CEPHi* with regard to inferring admixture in populations where it is present, we defined admixture using heatmaps comparing the estimated coalescence times of African-Neandertal genealogies with those from genealogies connecting non-Africans and Neandertals. This definition was used to produced sets of introgressed regions across thirteen human populations from four continents, revealing that there exists more of the Neandertal genome in any European population (CEU, FIN, GBR, TSI - we exclude IBS here as it contains very few samples) than in any Asian population in our set (CHB, CHS, JPT).

Significant intracontinental correlation is evident between sets of introgressed regions across populations, accompanied by a strong sharing even between continents, with European-related groups being the predominant source of the small amounts of admixture seen in African populations. When considering the overlap between Asians and Europeans as grouped continental populations, substantial sets of shared and non-shared introgressed regions were evident, with the shared regions having correlated coalescence times, strongly suggestive of their being the result of the same admixture events. This is indicative of admixture events having occurred before the Eurasian population split. While differences between sets of introgressed regions may indicate further events following the split, they might also be explicable through drift: the random and differential loss of introgressed regions between populations. Alternatively, continent-specific regions may also be a result of introgression from other archaic human groups, or the differential purging of Neandertal material between Asian and European populations, due to differences in effective population size between the continents (Juric *et al.* [2015]).

There are a multitude of further avenues that could be explored regarding the set of introgressed regions we produced using *CEPHi*. For example, we could further investigate regions that are shared between populations by comparing the respective sets of introgressed SNPs seen in individual coldspots. Where we see matching or highly similar haplotypes across populations,

for example, we can increase our certainty that these came from the same Neandertal-human admixture event.

Additionally, where a region we infer to be introgressed from Neandertals is shown to span an entire coldspot, or to reach one end, it may be interesting to examine how far the introgressed region extends. It may also be informative to assess the extent of linkage disequilibrium with regard to introgressed regions, both within individual populations, as well as between populations and continents. Finding associations between introgressed regions, or indeed between introgressed and non-introgressed regions, may be biologically and evolutionarily informative, potentially highlighting linked functionality or other dependencies. Selection of Neandertal-introgressed regions has been approached by some recent studies (Huerta-Sánchez *et al.* [2014]; Khrameeva *et al.* [2014]; SIGMA [2013]), and genealogical trees in admixed regions could be used to further explore this, as trees provide mutation times. Where the age of a mutation is young, but the derived allele is unexpectedly common in a population - as could be revealed with the use of site frequency spectra - positive selection may be inferred; trees built for individual populations allow for fine-grained investigation of selection across modern humans. Additionally, it may be informative to assess the gene and functional region density in recombination coldspots and introgressed regions, given the paucity of Neandertal sequence in genic regions (Harris and Nielsen [2016]; Sankararaman *et al.* [2014]), and the association of Neandertal DNA with various health problems in European populations (Simonti *et al.* [2016]).

Previous sets of introgressed regions were shown to have significant overlap with ours. We have also detailed those regions not captured by each method, using heatmaps comparing African and non-African coalescence times with the Altai Neandertal (Chapter 4). Notably, the set from Sankararaman *et al.* [2014] overlaps with ours to a larger extent than the set from Vernot and Akey [2014], which appears conservative. With regard to both sets, we find a proportion of our introgressed regions to be unique to our approach, and which appear to show admixture evidence. It may be helpful to investigate the criteria of the methods of Vernot and Akey [2014] and Sankararaman *et al.* [2014] to further understand why a portion of regions we classify as admixed are filtered from their sets. A separate proportion are not seen by us, and show

deeper Neandertal coalescences with the European samples. A next step might be to explore genomic localisation of these regions with regard to the location of recombination coldspots, as well as to investigate local variation in mutation rates and selection coefficients, but it may also be helpful to provide alternative and more aggressive definitions of admixture, for example widening the bounds of our definition to require a non-African coalescence with the Neandertal of $\leq$400kya, and an African-Neandertal coalescence of >500kya. Equally, we could simply require that the non-African coalescence time with the Neandertal is less than that of the equivalent African-Neandertal coalescence time, which may then encompass all admixed regions. Lastly, we could investigate the result of dropping our criterion of using recombinationally cold regions completely, and scanning the whole genome for sequence introgressed from the Neandertal.

On investigating a set of genealogical trees built in regions with two coalescence times in both the YRI-Neandertal and GBR-Neandertal analyses, where we find an intermingling of Neandertal haplotypes amongst the human haplotypes, we presented some initial evidence for a complex speciation event between humans and Neandertals. It is particularly interesting that the set of regions concluded to be a result of introgression from humans into Neandertals by Kuhlwilm *et al.* [2016] which overlap our coldspot set, all but one display two coalescence times with the Altai Neandertal in both YRI and GBR. Notably, we find a much larger set of similarly behaving regions, making up $\sim$5% of all coldspots, and there is certainly more information contained in trees built in regions which fall counter to expectation. Further examination of this set of trees, the equivalent trees in other human populations, and those with two coalescence times with Neandertals in YRI and not in the non-African population, may for example reveal new insights. A starting point to this may be to stratify genealogical trees by the number of coalescence times seen between humans and Neandertals, and potentially other features, such as their shape, as well as examining the distribution of these regions along the genome.

In the final chapter, we introduced a new method to date admixture between Neandertals and individual human populations. This method uses an ascertainment scheme to enrich for two SNP types which have entered the population via admixture events between Neandertals and humans. The descendant distribution from these SNP sets in admixed regions are matched

via Maximum Likelihood Estimation to an empirical descendant distribution given by non-admixed regions, calculated separately for each population. By introducing and utilising this method, we have been able to infer admixture dates between populations and Neandertal in finer detail than has thus far been seen (Sankararaman *et al.* [2012]) - with population-specific dates across two continents - and which estimates Asian admixture dates to be somewhat further in the past than those of Europeans. For further investigation, sets of introgressed regions shared between populations can be dated using this method to understand more intricately the interrelationships between different populations with regard to the Neandertal, as well as the timings of introgression across populations and continents.

We rely on a set of modelling assumptions throughout this thesis, as is standard in population genetics. These include the fact that we expect negligible to no recombination in our set of coldspots, that generation time is 28 years in the human and Neandertal populations we analyse, and that we can assume a mutation rate of 0.5×10^{-9}bp/yr (although we do also employ a rate that is twice this in Chapter 2, in order to explore the possible range), as per Scally and Durbin [2012]. These are all frequently used rates based on fine-scale recombination maps (for example Hinch *et al.* [2011]) and *de novo* estimation of mutation rates (Scally and Durbin [2012]). Corroboratively, recent work also supports a human generation time over the last 45,000 years of 26-30 years (Moorjani *et al.* [2016]). Indeed we can contribute to this literature as our work supports, for example, use of the slower mutation rate (0.5×10^{-9}bp/yr): we obtain dates of introgression in European populations that if halved, would be more recent than the extinction time of Neandertals of ∼30kya (Higham *et al.* [2014]). Given that previous work using *CEPHi* gives a YRI-GBR population split time of ∼79kya (as used for simulations in Chapter 4), we can have confidence that the genealogical trees from our analyses with the Altai Neandertal are well calibrated, providing accurate pictures of the history between the populations. This allows us to support, with confidence, our estimation of a YRI-Neandertal population split of ∼712kya, and the inferred dates of admixture across Eurasian populations.

Additionally, we can likely make some small improvements to our data. Firstly, we can more closely examine the phasing of the Neandertal genome, extending assessment of our phasing of

chromosome 21 to the rest of the genome, to give a more comprehensive measure of the potential accuracy of phasing using *ShapeIt2*. Furthermore, given the reference set of haplotypes for phasing was human, only these (and not Neandertal-specific) SNPs could be phased. This was the majority, but the remaining Neandertal-only SNPs were assigned to a haplotype randomly, and it may be useful to check the effect of alternative assignments. Additionally, singletons in the 1000 Genomes haplotypes are called with low confidence, and we could test the effect of removing these from analysis. We can also increase the accuracy of *CEPHi* by force calling each human population at Neandertal-specific SNPs. Currently we refer to the reference human genome, but there may be an excess of shared Neandertal-SNPs in some human populations over others, and including this information would further increase the accuracy of genealogical trees.

It is important to remember that we have substantially more high quality, geographically varied modern human whole genomes available for analysis than for any archaic species. Current analyses using archaic species ordinarily use a single ancient genome as representative of an archaic population; we use the Altai Neandertal genome from 50,000 years in the past, from the Altai mountains in Siberia, and this is currently the only high coverage Neandertal genome available. Recently, two additional nuclear high coverage genomes for Denisovans there have been published (Sawyer *et al.* [2015]), and as more high quality ancient genomes become available, analysis and inference will improve through simply having and using more information. This will allow for more detailed characterisation of the Neandertal (and Denisovan) in and of themselves, with regard to their variation: SNP calling, for example, will benefit substantially from multiple samples, as well as estimates of effective population size. Furthermore, data from a set of Neandertals will allow for spatial and temporal comparisons between the archaic species and populations of modern humans. Looking even further into the future, if sequencing techniques improve sufficiently, it may be possible for other more ancient genomes from species such as *Homo heidelbergensis* or *Homo erectus* to be sequenced. This will allow for possibilities regarding the influence of other species in the evolutionary trajectory of humans to be investigated more thoroughly (Meyer *et al.* [2012]).

Notably, this line of thought also applies to human data. The 1000 Genomes Project has provided us with substantial amounts of high quality human data with which to work on questions surrounding admixture between modern humans and Neandertals. An obvious way to extend the analyses contained within this thesis would be to use the newly available and large sets of data from a triad of papers recently released which contain 787 high quality genomes from geographically diverse populations (Malaspinas *et al.* [2016]; Mallick *et al.* [2016]; Pagani *et al.* [2016]). The analyses could also be repeated using the Denisovan genomes (Meyer *et al.* [2012]; Sawyer *et al.* [2015]), results from which will be significantly more interesting given the newly available Papuan genomes (among others). It would be relatively straightforward to ascertain a set of admixture dates between the Denisovan individuals and various human populations, and ideal to build genealogical trees in *CEPHi* which include human, Neandertal, and Denisovan haplotypes. Together, information from a large set of modern human genomes, alongside additional archaic genomes, will allow the painting of a more intricate picture of the interactions that have taken place between modern and ancient humans, across our joint evolutionary history.

Appendix A: Chapter 2 Supplement

Data preparation

Draft Neandertal genome

In Chapter 2 we use the draft Neandertal genome (Green *et al.* [2010]), made publicly available by Svante Pääbo's group at the Max Planck Institute of Evolutionary Anthropology in Leipzig, Germany, in May 2010 (http://www.eva.mpg.de/Neandertal/data.html). This genome is constituted of sequence taken from 6 fossilized Neandertal bones from different locations. Three are from Vindija cave, Croatia: Vi33.16 (54.1% genome coverage), Vi33.25 (46.6%) and Vi33.26 (45.2%). Additional single bones are from Neander Valley, Germany: Feld1 (0.1%), El Sidrón cave, Asturias, Spain: Sid1253 (0.1%), and Mezmaiskaya, Altai Mountains, Russia: Mez1 (2%).

The genome was sequenced using the Illumina platform, and reads were mapped to the human reference genome (NCBI Build 36.1/hg18, released March 2006) using a custom mapper which accounts for characteristics of ancient DNA which are not seen in modern DNA, such as shorter read length (typically ∼30-50bp) and deamination. Average genomic coverage is 1.3×, while the level of contamination by human DNA is estimated at ∼1.4%.

We selected the relevant regions using the Rsamtools package in R (Morgan *et al.* [2014]). From these, we created a single consensus sequence for each region via a number of steps.

Firstly, we used *Phred* quality scores to remove reads with low quality. A *Phred* score Q is logarithmically related to the probability of an error P for that base call, so $Q = -10 \log 10P$, meaning a *Phred* score of 20 makes the probability of an incorrect base call 1 in 100, and therefore that the base call accuracy is 99%, and a score of 30 gives an estimated accuracy of 99.9%, etc. Complete short reads with a *Phred* (MAPQ) score $\leq$20 resulted in their removal, and individual reads below the same threshold were also removed. Secondly, we discarded 3 regions which had coverage over 100× because they might represent variation between copy

number repeats, meaning the reads that map to this location in the reference are actually from duplicated sites in the sample. Thirdly, we removed any centromeric regions from analysis using centromere coordinates for the relevant genome version (build 36/hg18, Mar 2006) from the UCSC genome browser.

We further filtered on read reliability by accounting for deamination that has occurred over the thousands of years since the death of the individual. This causes bases to be mutated and read as others, giving us inaccurate sequence information. Various combinations of base replacements can occur, termed transitions when they occur between two purines or two pyrimidines, and transversions when they happen between purines and pyrimidines. The most common is the C to T (C$\longrightarrow$T) transition, which occurs when hydrolysis of a cytosine residue converts cytosine to uracil. When the DNA containing these lesions is used as a template for PCR, the base opposite the U is read as an A, changing the pair from C$\longrightarrow$G to T$\longrightarrow$A. Because deamination impacts single-stranded DNA, C$\longrightarrow$T transitions are more common at 5' ends of reads, and G$\longrightarrow$A at 3' ends. Given that deamination most severely affects the ends of reads, we removed 3bp from both ends of all reads. For the remainder of our sequences, we perform analyses with and without positions where these transitions and transversions may have occurred, which is detailed in Chapter 2 where relevant.

For a very large proportion of positions in a region, all bases at a particular position agreed. However, there was occasional disagreement. This predominantly represents sequencing errors, but also Neandertal polymorphism. Where two non-matching bases existed, we selected one base at random. Where there were more than two bases and at least one non-matching bases, we took the majority base if it existed, and a random base from the possible majority choices in the case of a tie.

We detail the construction and use of the high coverage Altai Neandertal genome in Appendix B.

Human reference genome

We use the human reference genome in Chapters 2, 3, and 4. Build GRCh36/hg18 (Mar 2006) is used in Chapter 2, and build GRCh37/hg19 (Feb 2009) in Chapters 2, 3, and 4; the large majority of analyses employing this later release. Both builds have high coverage of $\sim$20$\times$. This genome is made up of 11 individuals contributing unequal portions, one individual accounting for $\sim$60% of the genome. This individual is African-American, and we exploited this fact in Chapter 2, as it allowed us to effectively compare the Neandertal with two distinct modern human populations of African and European descent respectively.

Once coldspots were identified, we downloaded the human reference sequence from the UCSC genome browser for those regions (build 36/hg18, Mar 2006) (Kent *et al.* [2002a]). The ancestry of this genome has previously been inferred using a genome library of Bacterial Artificial Chromosomes (BACs) (Green *et al.* [2010]), provided to us by David Reich. For this library, BACs of 50-150kb were produced, and the ancestry assigning software HAPMIX (Price *et al.* [2009]) was run on each, followed by an HMM on those BACS from an African-American individual (RPCI-11) to obtain local ancestry. All BACS then fall into one of three categories: 'African', 'European', and 'East Asian', ancestry having been defined as the geographic region in which a

clone's ancestor lived 1,000 years ago, inferred based on its genetic proximity to other individuals from that region today. Both conservative and aggressive ancestry estimates were inferred, and we perform subsequent analysis at both levels, and compare conclusions.

Thus, for each coldspot, the human reference sequence was categorised by ancestry, giving us the ability to make comparisons between the distributions of the estimator of proportional mutational distance to Neandertal of the African and European regions. Regions assigned as East Asian were too few to be informative and were discarded.

Experimentally phased German genome

We added to our data a recently published experimentally phased German genome (Suk *et al.* [2011]). This genome comes from a 51 year old individual (MP1) in good health, and was resolved into haplotypes using a fosmid pool-based approach, and sequenced using the SOLiD platform. Coverage is 47×, with 98.1% of autosomes having coverage >5×.This increased the amount of European sequence data we had approximately threefold.

Additional modern human genomes

We employed 13 modern human genomes for analysis on their publication with Prüfer *et al.* [2014]. They are detailed below and used in in Chapter 2. The Australian genomes are Aboriginal, and although their exact source is unknown, they are shown to form clades with Papuans, and show no evidence of recent admixture with Europeans or East Asians. Coverage is between 24-32×.

Population	Country	Continent	Sample ID	Sex
Mbuti	Congo	Africa	HGDP00456	M
Yoruba	Nigeria	Africa	HGDP00927	M
San	South Africa	Africa	HGDP01029	M
Mandenka	West Africa	Africa	HGDP01284	M
French	France	Europe	HGDP00521	M
Sardinian	Italy	Europe	HGDP00665	M
Han	China	Asia	HGDP00778	M
Dai	China	Asia	HGDP01307	M
Karitiana	Brazil	America	HGDP00998	M
Mixe	Mexico	America	MIXE0007	M
Australian	Unknown	Australasia	WON,M	M
Australian	Unknown	Australasia	BUR,E	F
Papuan	Papua	Australasia	HGDP00542	M

Table 1: Modern human genomes used in Chapter 2

Chimpanzee genome

We used panTro4, published in February 2011 and available on the UCSC genome browser (Kent *et al.* [2002a]). It is approximately 6× coverage, and was initially sequenced and analysed by the Chimpanzee Sequencing and Analysis Consortium (Consortium *et al.* [2005]).

D-statistics for all pairs of modern humans

Table 2 below shows the complete set of calculated *D*-statistics. H1 and H2 indicate the first and second humans respectively, which are specified for each calculation. BABA gives the number of matches of H1 to the Neandertal, ABBA the number of matches of H2 to the Neandertal. $N1$ and $N2$ refer to Neandertal haplotypes. D is the value of the statistic itself, SE the standard error, and Z the Z-score, as described in section 2.2.4 of Chapter 2. The table is split into subsections for each continental comparison.

Table 2: *D*-statistics for all pairs of individuals.

D(H1, H2, Nean, Chimp)	BABA	ABBA	*D*	SE	Z
AFR-AFR					
D(MAN,MBU,N1,C)	15623.0	15290.5	0.0108	0.0163	0.6583
D(MAN,MBU,N2,C)	15457.0	15426.0	0.0010	0.0161	0.0622
D(MAN,SAN,N1,C)	15815.0	16801.5	-0.0302	0.0166	-1.8238
D(MAN,SAN,N2,C)	15974.5	16926.5	-0.0289	0.0165	-1.7523
D(MAN,YOR,N1,C)	13845.0	14331.5	-0.0173	0.0182	-0.9507
D(MAN,YOR,N2,C)	13864.5	14513.5	-0.0229	0.0177	-1.2898
D(MBU,SAN,N1,C)	15083.5	16406.0	-0.0420	0.0169	-2.4827
D(MBU,SAN,N2,C)	15053.0	16587.5	-0.0485	0.0168	-2.8795
D(MBU,YOR,N1,C)	15094.5	15936.0	-0.0271	0.0169	-1.6081
D(MBU,YOR,N2,C)	15203.0	15899.0	-0.0224	0.0167	-1.3401
D(SAN,YOR,N1,C)	16600.0	15977.5	0.0191	0.0183	1.0441
D(SAN,YOR,N2,C)	16786.0	16198.5	0.0178	0.0182	0.9802
EUR-EUR					
D(FRE,SAR,N1,C)	10728.0	10746.0	-0.0008	0.0241	-0.0348
D(FRE,SAR,N2,C)	10795.0	10910.5	-0.0053	0.0231	-0.2303
ASN-ASN					
D(DAI,HAN,N1,C)	9801.5	9781.0	0.0010	0.0227	0.0462
D(DAI,HAN,N2,C)	9834.5	9904.5	-0.0035	0.0226	-0.1571
AUS-AUS					
D(AUS,AUb,N1,C)	9140.5	9186.0	-0.0025	0.0214	-0.1161
D(AUS,AUb,N2,C)	9244.0	9139.0	0.0057	0.0216	0.2648
D(PAP,AUb,N1,C)	10171.5	10357.5	-0.0091	0.0261	-0.3470
D(PAP,AUb,N2,C)	10262.0	10386.5	-0.0060	0.0255	-0.2361
D(PAP,AUS,N1,C)	10261.5	10381.0	-0.0058	0.0256	-0.2258
D(PAP,AUS,N2,C)	10387.5	10427.0	-0.0019	0.0253	-0.0750

Table 2: *D*-statistics for all pairs of individuals.

D(H1, H2, Nean, Chimp)	BABA	ABBA	*D*	SE	Z
AMR-AMR					
D(KAR,MIX,N1,C)	346.0	301.5	0.0687	0.1134	0.6058
D(KAR,MIX,N2,C)	369.5	291.5	0.1180	0.1042	1.1314
EUR-AFR					
D(FRE,MAN,N1,C)	16244.5	13412.5	0.0955	0.0171	5.5824
D(FRE,MAN,N2,C)	16090.0	13649.0	0.0821	0.0171	4.7915
D(FRE,MBU,N1,C)	18056.5	14909.5	0.0955	0.0176	5.4205
D(FRE,MBU,N2,C)	17755.0	15082.5	0.0814	0.0177	4.6002
D(FRE,SAN,N1,C)	17994.0	16244.0	0.0511	0.0178	2.8712
D(FRE,SAN,N2,C)	17838.0	16556.0	0.0373	0.0174	2.1423
D(FRE,YOR,N1,C)	16457.5	14420.0	0.0660	0.0210	3.1402
D(FRE,YOR,N2,C)	16314.0	14643.5	0.0540	0.0207	2.6059
D(SAR,MAN,N1,C)	16536.0	13987.5	0.0835	0.0194	4.3101
D(SAR,MAN,N2,C)	16469.0	14249.5	0.0723	0.0187	3.8599
D(SAR,MBU,N1,C)	18199.5	15234.5	0.0887	0.0195	4.5422
D(SAR,MBU,N2,C)	18054.0	15530.5	0.0751	0.0188	4.0063
D(SAR,SAN,N1,C)	17873.0	16414.5	0.0425	0.0195	2.1854
D(SAR,SAN,N2,C)	17707.5	16644.0	0.0310	0.0191	1.6186
D(SAR,YOR,N1,C)	16833.0	14875.0	0.0618	0.0221	2.7983
D(SAR,YOR,N2,C)	16706.5	14908.5	0.0569	0.0215	2.6393
ASN-AFR					
D(DAI,MAN,N1,C)	16764.0	13812.0	0.0965	0.0180	5.3547
D(DAI,MAN,N2,C)	16615.0	13930.5	0.0879	0.0177	4.9558
D(DAI,MBU,N1,C)	17886.0	14552.0	0.1028	0.0164	6.2773
D(DAI,MBU,N2,C)	17869.5	14683.5	0.0979	0.0162	6.0475
D(DAI,SAN,N1,C)	17880.0	15883.5	0.0591	0.0186	3.1789
D(DAI,SAN,N2,C)	17800.5	16193.5	0.0473	0.0181	2.6077
D(DAI,YOR,N1,C)	16713.0	14113.5	0.0843	0.0185	4.5676
D(DAI,YOR,N2,C)	16418.5	14216.5	0.0719	0.0182	3.9494
D(HAN,MAN,N1,C)	16680.5	13705.0	0.0979	0.0194	5.0573
D(HAN,MAN,N2,C)	16889.0	13823.0	0.0998	0.0194	5.1489
D(HAN,MBU,N1,C)	17926.0	14572.5	0.1032	0.0180	5.7424
D(HAN,MBU,N2,C)	17746.5	14749.5	0.0922	0.0178	5.1883
D(HAN,SAN,N1,C)	17714.5	15714.5	0.0598	0.0177	3.3857
D(HAN,SAN,N2,C)	17672.5	15924.5	0.0520	0.0176	2.9615
D(HAN,YOR,N1,C)	16947.0	14351.5	0.0829	0.0202	4.1138
D(HAN,YOR,N2,C)	16791.5	14368.0	0.0778	0.0201	3.8686
AUS-AFR					
D(AUb,MAN,N1,C)	17115.0	13423.0	0.1209	0.0184	6.5735
D(AUb,MAN,N2,C)	17298.0	13658.5	0.1176	0.0179	6.5508
D(AUb,MBU,N1,C)	18575.5	14624.0	0.1190	0.0176	6.7491
D(AUb,MBU,N2,C)	18681.0	14886.5	0.1130	0.0174	6.4986
D(AUb,SAN,N1,C)	18552.5	15901.0	0.0770	0.0187	4.1225
D(AUb,SAN,N2,C)	18742.5	16266.0	0.0707	0.0185	3.8261
D(AUb,YOR,N1,C)	17162.5	14182.5	0.0951	0.0202	4.7069
D(AUb,YOR,N2,C)	17211.0	14159.0	0.0973	0.0200	4.8529
D(AUS,MAN,N1,C)	17305.0	13666.5	0.1175	0.0177	6.6186

Table 2: *D*-statistics for all pairs of individuals.

D(H1, H2, Nean, Chimp)	BABA	ABBA	*D*	SE	Z
D(AUS,MAN,N2,C)	17330.5	13552.0	0.1224	0.0173	7.0714
D(AUS,MBU,N1,C)	18747.5	14740.0	0.1197	0.0176	6.8037
D(AUS,MBU,N2,C)	18675.5	14938.0	0.1112	0.0172	6.4657
D(AUS,SAN,N1,C)	19006.5	16330.0	0.0757	0.0188	4.0244
D(AUS,SAN,N2,C)	18849.5	16450.5	0.0680	0.0186	3.6613
D(AUS,YOR,N1,C)	17320.5	14257.5	0.0970	0.0200	4.8505
D(AUS,YOR,N2,C)	17416.5	14378.0	0.0956	0.0193	4.9493
D(PAP,MAN,N1,C)	17817.5	14039.0	0.1186	0.0188	6.3048
D(PAP,MAN,N2,C)	17641.5	14189.5	0.1084	0.0189	5.7428
D(PAP,MBU,N1,C)	19058.5	15104.0	0.1158	0.0186	6.2197
D(PAP,MBU,N2,C)	18799.0	15302.0	0.1025	0.0183	5.6045
D(PAP,SAN,N1,C)	19455.0	16709.5	0.0759	0.0198	3.8255
D(PAP,SAN,N2,C)	19177.0	17088.5	0.0576	0.0197	2.9235
D(PAP,YOR,N1,C)	17684.0	14703.0	0.0920	0.0196	4.7039
D(PAP,YOR,N2,C)	17676.5	14793.0	0.0888	0.0194	4.5721
AMR-AFR					
D(KAR,MAN,N1,C)	653.0	482.5	0.1501	0.1005	1.4926
D(KAR,MAN,N2,C)	635.0	582.5	0.0431	0.1098	0.3925
D(KAR,MBU,N1,C)	887.0	632.0	0.1678	0.0986	1.7022
D(KAR,MBU,N2,C)	883.5	671.0	0.1366	0.1004	1.3612
D(KAR,SAN,N1,C)	590.0	544.5	0.0401	0.1132	0.3542
D(KAR,SAN,N2,C)	539.5	636.5	-0.0824	0.1091	-0.7557
D(KAR,YOR,N1,C)	649.5	610.0	0.0313	0.1384	0.2265
D(KAR,YOR,N2,C)	645.0	588.5	0.0458	0.1346	0.3402
D(MIX,MAN,N1,C)	17142.5	13964.5	0.1022	0.0183	5.5680
D(MIX,MAN,N2,C)	17086.0	14201.0	0.0922	0.0181	5.1044
D(MIX,MBU,N1,C)	18284.5	14947.0	0.1004	0.0173	5.7961
D(MIX,MBU,N2,C)	18056.5	15136.5	0.0880	0.0169	5.1957
D(MIX,SAN,N1,C)	18248.5	16140.5	0.0613	0.0181	3.3885
D(MIX,SAN,N2,C)	18241.0	16467.5	0.0511	0.0174	2.9426
D(MIX,YOR,N1,C)	16930.0	14478.0	0.0781	0.0190	4.1100
D(MIX,YOR,N2,C)	16881.5	14604.5	0.0723	0.0187	3.8737
EUR-ASN					
D(FRE,DAI,N1,C)	11735.0	12069.5	-0.0141	0.0240	-0.5859
D(FRE,DAI,N2,C)	11642.5	12056.0	-0.0174	0.0235	-0.7412
D(FRE,HAN,N1,C)	11431.5	11800.5	-0.0159	0.0231	-0.6877
D(FRE,HAN,N2,C)	11378.0	11878.0	-0.0215	0.0231	-0.9294
D(SAR,DAI,N1,C)	11749.5	12099.0	-0.0147	0.0264	-0.5556
D(SAR,DAI,N2,C)	11832.5	12388.5	-0.0230	0.0249	-0.9215
D(SAR,HAN,N1,C)	12076.5	12600.0	-0.0212	0.0254	-0.8336
D(SAR,HAN,N2,C)	12056.5	12587.0	-0.0215	0.0247	-0.8699
EUR-AUS					
D(FRE,AUb,N1,C)	11854.5	12746.0	-0.0362	0.0243	-1.4888
D(FRE,AUb,N2,C)	11950.0	12952.5	-0.0403	0.0238	-1.6885
D(FRE,AUS,N1,C)	11875.5	12941.0	-0.0429	0.0239	-1.7930
D(FRE,AUS,N2,C)	11871.0	13096.5	-0.0491	0.0234	-2.0932
D(FRE,PAP,N1,C)	12471.0	13319.5	-0.0329	0.0255	-1.2895
D(FRE,PAP,N2,C)	12237.0	13510.0	-0.0494	0.0253	-1.9510

Table 2: *D*-statistics for all pairs of individuals.

D(H1, H2, Nean, Chimp)	BABA	ABBA	*D*	SE	Z
D(SAR,AUb,N1,C)	12010.5	13072.5	-0.0423	0.0258	-1.6419
D(SAR,AUb,N2,C)	11884.0	13181.5	-0.0518	0.0247	-2.0927
D(SAR,AUS,N1,C)	12231.0	13256.0	-0.0402	0.0255	-1.5760
D(SAR,AUS,N2,C)	12146.0	13362.0	-0.0477	0.0247	-1.9324
D(SAR,PAP,N1,C)	12915.5	14005.0	-0.0405	0.0266	-1.5239
D(SAR,PAP,N2,C)	12925.0	14001.0	-0.0400	0.0257	-1.5541
EUR-AMR					
D(FRE,KAR,N1,C)	523.5	530.5	-0.0066	0.1312	-0.0505
D(FRE,KAR,N2,C)	576.0	550.0	0.0230	0.1298	0.1778
D(FRE,MIX,N1,C)	11765.5	12019.0	-0.0107	0.0227	-0.4695
D(FRE,MIX,N2,C)	11798.5	12181.0	-0.0160	0.0218	-0.7311
D(SAR,KAR,N1,C)	327.0	473.0	-0.1825	0.1667	-1.0947
D(SAR,KAR,N2,C)	345.5	474.0	-0.1568	0.1846	-0.8493
D(SAR,MIX,N1,C)	12027.0	12470.5	-0.0181	0.0242	-0.7477
D(SAR,MIX,N2,C)	12030.0	12549.0	-0.0211	0.0230	-0.9163
ASN-AUS					
D(DAI,AUb,N1,C)	11102.5	11785.0	-0.0298	0.0231	-1.2883
D(DAI,AUb,N2,C)	11255.0	11716.0	-0.0201	0.0224	-0.8958
D(DAI,AUS,N1,C)	11387.5	11891.0	-0.0216	0.0225	-0.9601
D(DAI,AUS,N2,C)	11364.0	11947.0	-0.0250	0.0221	-1.1308
D(DAI,PAP,N1,C)	11566.5	12072.5	-0.0214	0.0224	-0.9549
D(DAI,PAP,N2,C)	11597.0	12089.5	-0.0208	0.0222	-0.9381
D(HAN,AUb,N1,C)	11333.5	11843.5	-0.0220	0.0233	-0.9433
D(HAN,AUb,N2,C)	11322.0	11938.0	-0.0265	0.0234	-1.1298
D(HAN,AUS,N1,C)	11516.5	12212.5	-0.0293	0.0236	-1.2429
D(HAN,AUS,N2,C)	11679.0	12249.0	-0.0238	0.0231	-1.0302
D(HAN,PAP,N1,C)	12189.5	12668.5	-0.0193	0.0247	-0.7808
D(HAN,PAP,N2,C)	12275.0	12719.0	-0.0178	0.0246	-0.7221
ASN-AMR					
D(DAI,KAR,N1,C)	437.0	341.5	0.1226	0.1086	1.1289
D(DAI,KAR,N2,C)	409.0	350.0	0.0777	0.1021	0.7610
D(DAI,MIX,N1,C)	10857.0	10919.0	-0.0028	0.0218	-0.1307
D(DAI,MIX,N2,C)	10933.5	11007.5	-0.0034	0.0214	-0.1574
D(HAN,KAR,N1,C)	456.5	287.5	0.2271	0.1299	1.7477
D(HAN,KAR,N2,C)	428.5	273.0	0.2216	0.1211	1.8300
D(HAN,MIX,N1,C)	10818.0	10865.5	-0.0022	0.0226	-0.0970
D(HAN,MIX,N2,C)	10892.0	10826.0	0.0030	0.0223	0.1360
AUS-AMR					
D(AUb,KAR,N1,C)	450.5	439.0	0.0129	0.1003	0.1288
D(AUb,KAR,N2,C)	464.5	445.5	0.0208	0.0958	0.2178
D(AUb,MIX,N1,C)	12539.5	11789.0	0.0308	0.0229	1.3471
D(AUb,MIX,N2,C)	12661.0	11836.5	0.0337	0.0222	1.5178
D(AUS,KAR,N1,C)	420.0	340.0	0.1052	0.1281	0.8216
D(AUS,KAR,N2,C)	447.0	447.0	0.1625	0.1017	1.5974
D(AUS,MIX,N1,C)	12717.0	12067.5	0.0262	0.0228	1.1496
D(AUS,MIX,N2,C)	12833.5	12034.5	0.0321	0.0223	1.4397
D(PAP,KAR,N1,C)	593.5	512.0	0.0737	0.0959	0.7680

Table 2: *D*-statistics for all pairs of individuals.

D(H1, H2, Nean, Chimp)	BABA	ABBA	*D*	SE	Z
D(PAP,KAR,N2,C)	549.5	504.0	0.0431	0.0949	0.4549
D(PAP,MIX,N1,C)	12998.0	12437.0	0.0221	0.0234	0.9426
D(PAP,MIX,N2,C)	13090.5	12461.0	0.0246	0.0229	1.0743

Appendix B: Chapter 3 Supplement

High coverage Neandertal genome

On its release in March 2013, we began employing a high coverage version of the Neandertal genome (Prüfer *et al.* [2014]). This is used for parts of Chapter 2, and is the only Neandertal genome used from Chapter 3 onwards. All sequence data comes from a single toe phalanx from the fourth or fifth toe of a female Neandertal from Denisova cave in the Altai mountains in Siberia (at the meeting of Russia, China, Mongolia, and Kazakhstan), which is approximately 50,000 years old, slightly older than the Denisovan finger bone taken from the same cave, which is dated to 50,000±2,300 years old. It is most closely related to the Mezmaiskaya Neandertal, followed by the three Vindija Neandertal fossils (Vi33.16, Vi33.25, Vi33.26).

The genome was sequenced using the Illumina platform, and reads were mapped to the human reference genome (GRCh37/hg19, released Feb 2009). Coverage is ∼52×, and contamination is estimated at ∼1%. We include details of our use of this genome in Chapter 3.

1000 Genomes SNP data

In Chapters 3, 4, and 5 we use the phase 2 1000 Genomes data (Via García *et al.* [2012]), rephased using ShapeIt2 by Jonathan Marchini and Olivier Delaneau (Delaneau *et al.* [2013]). This dataset consists of SNP and indel data for 1092 individuals from 14 populations split into 4 continents: Africa, Europe, Asia, and America. The data are constituted of 36,820,992 SNPs (using build GRCh37/hg19), from which we use those that fall within a collection of recombinationally cold regions defined in section 2.2.1 of Chapter 2.

Population acronyms for this dataset are used throughout this thesis, and a reference list for these is given in Table 3.

Ref	Population	Continent	Number of individuals
YRI	Yoruba in Ibadan, Nigeria	Africa	88
LWK	Luhya in Webuye, Kenya	Africa	97
ASW	African Ancestry in Southwest US	Africa	61
GBR	British from England and Scotland	Europe	89
FIN	Finnish from Finland	Europe	93
IBS	Iberian populations in Spain	Europe	14
CEU	Utah residents (CEPH) with N. & W. Eur. ancestry	Europe	85
TSI	Toscani in Italia	Europe	98
CHB	Han Chinese in Beijing, China	Asia	97
CHS	Han Chinese South	Asia	100
JPT	Japanese in Toyko, Japan	Asia	89
MXL	Mexican Ancestry in Los Angeles, CA	Americas	66
CLM	Colombian in Medellin, Colombia	Americas	60
PUR	Puerto Rican in Puerto Rico	Americas	55

Table 3: Human populations from 1000 Genomes Project

Other SNP sets

The SNP sets listed below were used to call SNP genotypes in the high coverage Neandertal genome as described in Chapter 3, using the HaplotypeCaller in the Genome Analysis Toolkit (GATK) from the Broad Institute (DePristo *et al.* [2011]). These sets come as standard in the GATK bundle, are considered to add up to a very complete set of human SNPs, and can be found here: https://www.broadinstitute.org/gatk/download/. Versions used in our analysis are given as filenames in parentheses.

- 1000 Genomes SNP variation data: as above (1000G_phase1.snps.high_confidence.b37.vcf).
- Hapmap genotypes and sites (hapmap_3.3.b37.vcf).
- Latest release of dbSNP (dbsnp_137.b37.vcf), alongside a subset of these sites excluding those identified by the 1000 Genomes project, which is used to evaluate Ti/Tv values at novel sites.
- Omni 2.5 genotypes for 1000 Genomes samples (1000G_omni2.5.b37.vcf).

Appendix C: Chapter 4 Supplement

Command line for *scrm* used to simulate a 3 population scenario

./scrm-1.5.0/scrm 356 1 -T -L -t $i -I 3 176 178 2 -r 0 $j -en 0 1 2.50100786945518 -en 0 2 2.9666892 -en 0 3 0.0001 -en 0.004605932 1 2.261052 -en 0.004605932 2 2.3669581 -en 0.004605932 3 0.0001 -en 0.009408523 1 2.601788 -en 0.009408523 2 1.7858201 -en 0.009408523 3 0.0001 -en 0.01963763 1 2.864409 -en 0.01963763 2 1.3107920 -en 0.01963763 3 0.0001 -en 0.03075889 1 3.331466 -en 0.03075889 2 0.8936999 -en 0.03075889 3 0.0001 -en 0.04 1 3.331466 -en 0.04 2 0.8936999 -en 0.04 3 0.3871809 -en 0.04285012 1 3.873893 -en 0.04285012 2 0.6316645 -en 0.04285012 3 0.3871809 -en 0.07028823 1 3.068115 -en 0.07028823 2 1.8903380 -en 0.07028823 3 0.3871809 -ej 0.07119732 2 1 -en 0.08638462 1 4.269029 -en 0.08638462 2 4.26902 -en 0.08638462 3 0.3871809 -en 0.1752491 1 4.782507 -en 0.1752491 2 4.782507 -en 0.1752491 3 0.5075720 -en 0.3220153 1 5.393798 -en 0.3220153 2 5.393798 -en 0.3220153 3 0.6687181 -en 0.5644103 1 5.355230 -en 0.5644103 2 5.355230 -en 0.5644103 3 0.8902034 -en 0.63655 1 4.679918 -en 0.63655 2 4.679918 -en 0.63655 3 0.7128860 -ej 0.63655 3 1 -eN 0.9647433 4.516866 -eN 1.625922 2.253350 -eN 2.717907 1.713967 -eN 4.5214 2.087848 -eN 500 4.146150 -eps 0.04017857 2 3 0.98 –print-model

$i = mutation rate θ, where $\theta = 4N\mu$
$j = region length in bp

-en = population size
-eN = all population sizes
-ej = speciation event
-eps = admixture event
-T = Newick format genealogies
-L = print TMRCA and local tree length for each segment

Please see https://github.com/scrm/scrm/wiki/Command-Line-Options for further details.

Accessing introgressed regions inferred by *CEPHi*

We created an account at http://www.pantherdb.org/ where we have uploaded the sets of introgressed regions inferred in Chapter 3. We give one list per continent (African, European, Asian, and American). To log in, please use the username 'AFDPhilThesis' and the password 'introgression'. Choose a list from the workspace, select 'Homo sapiens', and select the type of output preferred. The pie charts are particularly helpful, allowing for different types of biological information to be selected, with further detail provided for each.

Appendix D: Chapter 5 Supplement

Command line for *scrm* used to simulate a 3 population scenario

./scrm-1.5.0/scrm 356 1 -T -L -t $i -I 3 176 178 2 -r 0 $j -en 0 1 2.50100786945518 -en 0 2 2.9666892 -en 0 3 0.0001 -en 0.004605932 1 2.261052 -en 0.004605932 2 2.3669581 -en 0.004605932 3 1000 -en 0.009408523 1 2.601788 -en 0.009408523 2 1.7858201 -en 0.009408523 3 1000 -en 0.01963763 1 2.864409 -en 0.01963763 2 1.3107920 -en 0.01963763 3 1000 -en 0.03075889 1 3.331466 -en 0.03075889 2 0.8936999 -en 0.03075889 3 1000 -en 0.04 1 3.331466 -en 0.04 2 0.8936999 -en 0.04 3 0.3871809 -en 0.04285012 1 3.873893 -en 0.04285012 2 0.6316645 -en 0.04285012 3 0.3871809 -en 0.07028823 1 3.068115 -en 0.07028823 2 1.8903380 -en 0.07028823 3 0.3871809 -ej 0.07119732 2 1 -en 0.08638462 1 4.269029 -en 0.08638462 2 4.26902 -en 0.08638462 3 0.3871809 -en 0.1752491 1 4.782507 -en 0.1752491 2 4.782507 -en 0.1752491 3 0.5075720 -en 0.3220153 1 5.393798 -en 0.3220153 2 5.393798 -en 0.3220153 3 0.6687181 -en 0.5644103 1 5.355230 -en 0.5644103 2 5.355230 -en 0.5644103 3 0.8902034 -en 0.63655 1 4.679918 -en 0.63655 2 4.679918 -en 0.63655 3 0.7128860 -ej 0.63655 3 1 -eN 0.9647433 4.516866 -eN 1.625922 2.253350 -eN 2.717907 1.713967 -eN 4.5214 2.087848 -eN 500 4.146150 -eps 0.04017857 2 3 0.98 –print-model

Please see Appendix C and https://github.com/scrm/scrm/wiki/Command-Line-Options for details of parameters.

Example command line for *CEPHi* once data is split into pairs of populations

python cephi.py -f simscephi1/ -m optim T pop sizes per epoch -n 100 –correctBias –Nean 0.04 –nbrIterPSPE 20 –ncpu 20 -o sims-3pop-YRINean-allchrs –theta 7.5 7.5 -T 0.04 0.09 –variableTheta 30.0 –epochFormula None 10 12 –nbrTreesPSPE 20 –saveTree 2 > screens/trace-sims-3pop-YRINean-allchrs.txt

Inferred admixture dates from simulated data for YRI and GBR

YRI

Cutoff	t1	t2	G	GG
20	986	371	202	202
30	989	509	187	187
40	941	589	156	156
50	894	552	156	156
60	901	509	156	156
70	973	614	156	156
80	933	614	156	156
90	983	614	156	156
100	989	614	156	156

GBR

Cutoff	t1	t2	G	GG
20	61	43	51	51
30	61	43	48	48
40	61	41	45	45
50	65	41	45	45
60	65	41	45	45
70	69	41	45	45
80	90	41	45	45
90	90	41	45	45
100	139	41	45	45

Table 4: Inferred admixture dates from simulated data for YRI and GBR. The left hand table gives dates for YRI, and the right hand table for GBR. Cutoff gives the descendant number cutoff. Sets of admixed regions are t1, t2, G, and GG, as defined in the main text. All numbers are in thousands of year ago (kya).

Inferred admixture dates from 1000 Genomes data

t1 cutoff	YRI	LWK	ASW	GBR	FIN	IBS	CEU	TSI	CHB	CHS	JPT	MXL	CLM	PUR
20	23 32 58	14 16 20	43 47 55	57 64 76	45 49 59	62 72 79	53 60 69	47 53 60	46 57 82	44 53 67	53 65 88	77 88 101	90 101 117	92 102 116
30	49 78 127	30 30 31	48 57 60	57 62 69	49 53 60	88 88 93	53 58 63	50 53 58	56 66 78	59 67 85	70 88 111	81 88 97	91 100 108	85 93 101
40	74 103 180	30 30 31	71 71 73	60 65 70	53 57 62	88 88 93	56 60 64	53 56 62	59 67 73	61 68 80	65 74 86	88 95 104	87 95 102	90 96 103
50	86 128 209	72 72 76	110 110 111	65 69 74	55 60 64	88 88 93	60 64 68	57 61 66	62 67 73	63 68 77	67 73 84	93 100 107	90 95 103	89 96 102
60	130 198 297	72 72 76	110 110 111	66 70 76	57 60 66	88 88 93	60 64 68	62 65 69	66 71 78	67 71 82	71 76 88	93 100 106	91 98 104	92 96 104
70	257 371 564	103 103 107	164 164 166	66 70 76	59 64 69	88 88 93	60 64 67	62 66 69	66 71 77	67 71 80	74 81 88	94 101 107	91 98 104	92 96 104
80	251 371 532	120 121 125	194 194 195	66 71 77	60 64 69	88 88 93	61 64 68	62 65 69	67 71 76	67 71 80	76 83 89	95 101 109	91 98 104	132 132 133
90	325 454 650	120 121 125	211 211 212	68 73 78	61 65 70	88 88 93	61 64 67	62 65 69	69 75 81	69 76 83	76 84 88	95 101 108	125 125 126	132 132 133
100	321 440 621	120 121 125	331 331 333	69 73 80	63 65 71	88 88 93	61 64 67	65 65 71	69 75 81	69 76 82	76 81 88	125 125 128	125 125 126	313 313 315

t2 cutoff	YRI	LWK	ASW	GBR	FIN	IBS	CEU	TSI	CHB	CHS	JPT	MXL	CLM	PUR
20	19 29 58	13 19 28	32 42 53	38 44 54	44 57 80	61 73 96	50 65 88	40 51 67	81 183 656	60 586 1000	47 64 120	63 86 117	76 92 130	78 96 130
30	26 50 94	13 19 28	47 49 57	44 53 60	49 57 75	61 73 96	45 52 62	43 51 60	66 113 242	75 121 683	79 115 175	77 92 119	78 92 108	72 84 105
40	45 75 158	13 19 28	47 49 57	45 51 58	51 58 72	61 73 96	49 58 66	45 52 59	64 89 155	70 86 162	69 88 124	78 92 108	76 86 101	71 84 101
50	45 75 152	13 19 28	47 49 57	48 56 63	50 57 67	61 73 96	50 58 65	50 56 65	65 83 120	68 84 113	65 80 97	86 96 116	78 93 106	71 84 100
60	45 75 150	13 19 28	47 49 57	48 56 62	49 56 65	61 73 96	49 54 64	55 63 72	65 81 104	69 86 105	68 83 98	85 96 113	78 93 105	78 93 105
70	45 75 150	13 19 28	47 49 57	48 56 62	49 55 64	61 73 96	49 54 63	55 63 70	66 80 96	67 78 94	69 83 96	85 96 111	78 93 104	77 93 105
80	45 75 149	13 19 28	47 49 57	48 56 62	49 55 63	61 73 96	51 59 66	55 63 70	65 75 90	69 84 94	68 80 93	85 94 109	78 93 104	77 93 105
90	124 137 276	13 19 28	47 49 57	48 56 62	52 54 68	61 73 96	51 59 66	55 63 70	64 71 86	68 78 91	68 77 91	85 94 109	78 93 104	77 93 105
100	124 137 274	13 19 28	47 49 57	53 58 63	52 54 68	61 73 96	51 59 66	65 65 76	64 71 84	67 78 90	70 77 94	85 94 109	78 93 104	77 93 105

G cutoff	YRI	LWK	ASW	GBR	FIN	IBS	CEU	TSI	CHB	CHS	JPT	MXL	CLM	PUR
20	15 25 65	13 19 28	32 42 54	39 45 57	45 57 89	61 73 98	53 65 103	40 52 68	73 183 664	67 586 1000	47 64 130	63 85 120	78 101 135	78 96 131
30	15 25 63	13 19 28	47 49 60	46 53 63	51 64 79	61 73 98	46 54 66	44 53 62	67 113 270	75 121 734	80 115 191	76 92 117	78 92 109	72 87 106
40	15 25 63	13 19 28	47 49 60	46 52 61	52 64 75	61 73 98	51 58 68	46 52 61	64 89 160	69 86 162	70 88 129	77 92 108	76 86 102	71 84 102
50	15 25 63	13 19 28	47 49 60	50 58 65	51 58 70	61 73 98	51 58 66	51 57 67	63 83 116	67 85 114	65 80 101	86 96 116	78 93 106	71 84 101
60	15 25 63	13 19 28	47 49 60	50 57 64	50 57 67	61 73 98	51 58 66	56 63 74	64 80 104	69 86 107	69 83 101	85 96 113	78 93 105	78 93 106
70	15 25 63	13 19 28	47 49 60	50 57 63	50 57 66	61 73 98	51 58 65	56 63 73	64 72 94	67 78 95	70 83 98	85 96 110	78 93 105	78 93 106
80	15 25 63	13 19 28	47 49 60	50 57 63	50 57 65	61 73 98	53 63 68	56 63 72	64 71 86	69 84 94	69 83 95	85 96 109	78 93 105	78 93 106
90	15 25 63	13 19 28	47 49 60	50 57 63	52 62 70	61 73 98	53 63 68	56 63 72	63 71 83	68 78 92	68 80 93	85 96 109	78 93 105	78 93 106
100	15 25 63	13 19 28	47 49 60	53 58 64	52 62 70	61 73 98	53 63 68	65 69 77	62 70 82	67 78 90	70 80 95	85 96 109	78 93 105	78 93 106

GG cutoff	YRI	LWK	ASW	GBR	FIN	IBS	CEU	TSI	CHB	CHS	JPT	MXL	CLM	PUR
20	15 25 65	13 19 27	28 36 47	39 47 60	47 64 97	61 73 97	53 65 114	40 52 68	57 176 719	62 533 1000	41 60 104	63 85 121	74 92 128	69 84 110
30	15 25 63	13 19 27	47 49 52	47 53 65	48 57 74	61 73 97	47 55 67	42 50 60	70 142 348	80 170 803	78 115 179	77 94 121	76 92 108	67 79 98
40	15 25 63	13 19 27	47 49 52	45 53 60	51 61 74	61 73 97	48 56 65	45 52 60	63 89 163	66 86 152	67 88 124	79 92 110	75 86 101	66 79 95
50	15 25 63	13 19 27	47 49 52	50 57 66	50 57 69	61 73 97	48 54 65	48 55 63	64 82 123	65 84 114	64 80 100	80 92 107	77 93 106	66 78 94
60	15 25 63	13 19 27	47 49 52	50 57 64	49 56 66	61 73 97	48 54 64	53 63 70	66 82 113	66 84 105	68 83 102	79 92 105	77 93 106	69 93 101
70	15 25 63	13 19 27	47 49 52	49 57 64	49 56 65	61 73 97	48 54 63	53 62 69	65 80 97	65 78 94	70 83 99	79 91 104	77 93 105	69 93 101
80	15 25 63	13 19 27	47 49 52	49 57 64	49 55 65	61 73 97	51 59 66	53 62 69	64 72 91	67 78 94	68 83 95	79 91 104	77 93 105	69 93 101
90	15 25 63	13 19 27	47 49 52	49 57 64	52 62 70	61 73 97	51 59 66	53 62 69	64 71 87	66 78 91	68 77 94	79 91 104	77 93 105	69 93 101
100	15 25 63	13 19 27	47 49 52	53 58 65	52 62 70	61 73 97	51 59 66	65 65 75	64 71 85	66 78 90	70 82 96	79 91 104	77 93 105	69 93 101

Table 5: Inferred admixture dates and 95% confidence intervals across 14 human populations, using t1, t2, G, and GG sets of admixed regions (from top to bottom). Cutoff is the descendant number cutoff. All numbers are given in thousands of years ago (kya).

References

Aiello, L. (1993). The fossil evidence for modern human origins in Africa: A revised view. *American Anthropologist*, **95**(1), 73–96. 10

Armitage, S., Jasim, S., Marks, A., Parker, A., Usik, V., and Uerpmann, H. (2011). The southern route out of Africa: Evidence for an early expansion of modern humans into Arabia. *Science*, **331**(6016), 453–456. 3

Banks, W., d'Errico, F., Peterson, A., Kageyama, M., Sima, A., and Sánchez-Goñi, M. (2008). Neanderthal extinction by competitive exclusion. *PLoS One*, **3**(12), e3972. 4

Becquet, C. and Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**(10), 1505–1519. 56

Belmaker, M. and Hovers, E. (2011). Ecological change and the extinction of the Levantine Neanderthals: implications from a diachronic study of micromammals from Amud Cave, Israel. *Quaternary Science Reviews*, **30**(21), 3196–3209. 2

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., and Mountain, J. L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, **96**(1), 37–53. 151

Buck, L. T. and Stringer, C. B. (2014). *Homo heidelbergensis*. *Current Biology*, **24**(6), R214–R215. 51

Cann, H. M., De Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., *et al.* (2002). A human genome diversity cell line panel. *Science*, **296**(5566), 261–262. 19, 158

Caron, F., d'Errico, F., Del Moral, P., Santos, F., and Zilhão, J. (2011). The reality of Neandertal symbolic behavior at the Grotte du Renne, Arcy-sur-Cure, France. *PLoS One*, **6**(6), e21545. 7

Condemi, S., Mounier, A., Giunti, P., Lari, M., Caramelli, D., and Longo, L. (2013). Possible interbreeding in late Italian Neanderthals? New data from the Mezzena jaw (Monti Lessini, Verona, Italy). *PloS One*, **8**(3), e59781. 9, 21

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, **38**(11), 1251–1260. 111

Consortium, T. C., Analysis, *et al.* (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**(7055), 69–87. 171

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., *et al.* (2011). The variant call format and vcftools. *Bioinformatics*, **27**(15), 2156–2158. 37

Dannemann, M., Andrés, A. M., and Kelso, J. (2016). Introgression of Neandertal-and Denisovan-like haplotypes contributes to adaptive variation in human toll-like receptors. *The American Journal of Human Genetics*, **98**(1), 22–33. 91

Davies, R. and Underdown, S. (2006). The Neanderthals: a social synthesis. *Cambridge Archaeological Journal*, **16**(02), 145–164. 4

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, **10**(1), 5–6. 68, 176

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**(5), 491–498. 63, 177

Ding, Q., Hu, Y., Xu, S., Wang, J., and Jin, L. (2013). Neanderthal introgression at chromosome 3p21. 31 was under positive natural selection in East Asians. *Molecular Biology and Evolution*, page 260. 91

Dobon, B., Hassan, H. Y., Laayouni, H., Luisi, P., Ricaño-Ponce, I., Zhernakova, A., Wijmenga, C., Tahir, H., Comas, D., Netea, M. G., *et al.* (2015). The genetics of East African populations: a Nilo-Saharan component in the African genetic landscape. *Scientific Reports*, **5**. 99, 151

Duarte, C., Maurício, J., Pettitt, P. B., Souto, P., Trinkaus, E., Van Der Plicht, H., and Zilhão, J. (1999). The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *Proceedings of the National Academy of Sciences*, **96**(13), 7604–7609. 8, 21

Durand, E., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*. 18

Endicott, P., Ho, S. Y., and Stringer, C. (2010). Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. *Journal of Human Evolution*, **59**(1), 87–95. 53

Eriksson, A. and Manica, A. (2012). Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences*, **109**(35), 13956–13960. 18, 19

Faerman, M., Zilberman, U., Smith, P., Kharitonov, V., and Batsevitz, V. (1994). A Neanderthal infant from the Barakai Cave, western Caucasus. *Journal of Human Evolution*, **27**(5), 405–415. 2

Finlayson, C. (2004). *Neanderthals and modern humans: an ecological and evolutionary perspective*, volume 38. Cambridge University Press. 2

Finlayson, C. and Carrion, J. (2007). Rapid ecological turnover and its impact on Neanderthal and other human populations. *Trends in Ecology & Evolution*, **22**(4), 213–222. 3

Finlayson, C., Pacheco, F., Rodríguez-Vidal, J., Fa, D., López, J., Pérez, A., Finlayson, G., Allue, E., Preysler, and Cáceres, J. (2006). Late survival of Neanderthals at the southernmost extreme of Europe. *Nature*, **443**(7113), 850–853. 2, 8

Finlayson, C., Fa, D. A., Espejo, F. J., Carrión, J. S., Finlayson, G., Pacheco, F. G., Vidal, J. R., Stringer, C., and Ruiz, F. M. (2008). Gorham's Cave, Gibraltar - The persistence of a Neanderthal population. *Quaternary International*, **181**(1), 64–71. 2

Frazer, K., Ballinger, D., Cox, D., Hinds, D., Stuve, L., Gibbs, R., Belmont, J., Boudreau, A., Hardenbol, P., Leal, S., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861. 28

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L., Aximu-Petri, A., Prüfer, K., de Filippo, C., *et al.* (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, **514**(7523), 445–449. 9, 31, 132, 134, 155

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., *et al.* (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature*. 9, 21, 133

Glantz, M., Viola, B., Wrinn, P., Chikisheva, T., Derevianko, A., Krivoshapkin, A., Islamov, U., Suleimanov, R., and Ritzman, T. (2008). New hominin remains from Uzbekistan. *Journal of Human Evolution*, **55**(2), 223–237. 2

Green, R., Krause, J., Ptak, S., Briggs, A., Ronan, M., Simons, J., Du, L., Egholm, M., Rothberg, J., Paunovic, M., *et al.* (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**(7117), 330–336. 12

Green, R., Malaspinas, A., Krause, J., Briggs, A., Johnson, P., Uhler, C., Meyer, M., Good, J., Maricic, T., Stenzel, U., *et al.* (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, **134**(3), 416–426. 11, 53

Green, R., Krause, J., Briggs, A., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M., *et al.* (2010). A draft sequence of the Neandertal genome. *Science*, **328**(5979), 710. 2, 4, 10, 12, 14, 15, 27, 28, 30, 33, 35, 38, 45, 49, 51, 52, 80, 89, 160, 168, 169

Harris, K. and Nielsen, R. (2016). The genetic cost of Neanderthal introgression. *Genetics*, **203**(2), 881–891. 93, 163

Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. *Science*, **343**(6172), 747–751. 27, 83

Hervella, M., Svensson, E., Alberdi, A., Günther, T., Izagirre, N., Munters, A., Alonso, S., Ioana, M., Ridiche, F., Soficaru, A., *et al.* (2016). The mitogenome of a 35,000-year-old *Homo sapiens* from Europe supports a Palaeolithic back-migration to Africa. *Scientific Reports*, **6**. 151

Heyes, P. J., Anastasakis, K., de Jong, W., van Hoesel, A., Roebroeks, W., and Soressi, M. (2016). Selection and use of manganese dioxide by Neanderthals. *Scientific Reports*, **6**. 7

Higham, T., Ramsey, C., Karavanić, I., Smith, F., and Trinkaus, E. (2006). Revised direct radiocarbon dating of the Vindija G1 upper Paleolithic Neandertals. *Proceedings of the National Academy of Sciences*, **103**(3), 553–557. 2

Higham, T., Douka, K., Wood, R., Ramsey, C. B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., Barroso-Ruíz, C., *et al.* (2014). The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, **512**(7514), 306–309. 2, 165

Hinch, A., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C., Chen, G., Wang, K., Buxbaum, S., Akylbekova, E., *et al.* (2011). The landscape of recombination in African Americans. *Nature*, **476**(7359), 170–175. 29, 165

Houldcroft, C. J. and Underdown, S. J. (2016). Neanderthal genomics suggests a Pleistocene time frame for the first epidemiologic transition. *American Journal of Physical Anthropology*. 4, 5

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, **5**(6), e1000529. 68

Hubbard, T. D., Murray, I. A., Bisson, W. H., Sullivan, A. P., Sebastian, A., Perry, G. H., Jablonski, N. G., and Perdew, G. H. (2016). Divergent Ah receptor ligand selectivity during hominin evolution. *Molecular Biology and Evolution*. 5

Hublin, J. and Roebroeks, W. (2009). Ebb and flow or regional extinctions? On the character of Neandertal occupation of northern environments. *Comptes Rendus Palevol*, **8**(5), 503–509. 4

Hublin, J.-J., Talamo, S., Julien, M., David, F., Connet, N., Bodu, P., Vandermeersch, B., and Richards, M. P. (2012). Radiocarbon dates from the Grotte du Renne and Saint-Césaire support a Neandertal origin for the Châtelperronian. *Proceedings of the National Academy of Sciences*, **109**(46), 18743–18748. 7

Huerta-Sánchez, E., Jin, X., Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., *et al.* (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, **512**(7513), 194–197. 92, 163

Jacobs, L. C., Wollstein, A., Lao, O., Hofman, A., Klaver, C. C., Uitterlinden, A. G., Nijsten, T., Kayser, M., and Liu, F. (2013). Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. *Human Genetics*, **132**(2), 147–158. 92

Jaubert, J., Verheyden, S., Genty, D., Soulier, M., Cheng, H., Blamart, D., Burlet, C., Camus, H., Delaby, S., Deldicque, D., *et al.* (2016). Early Neanderthal constructions deep in Bruniquel Cave in southwestern France. *Nature*, **534**(7605), 111–114. 7

Juric, I., Aeschbacher, S., and Coop, G. (2015). The strength of selection against Neanderthal introgression. *bioRxiv*, page 030148. 89, 93, 162

Keinan, A., Mullikin, J. C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics*, **39**(10), 1251–1255. 111

Kent, W. *et al.* (2002a). The human genome browser. **12**(6), 996–1006. 169, 171

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002b). The human genome browser at UCSC. *Genome Research*, **12**(6), 996–1006. 33, 37, 49

Khrameeva, K., Bozek, K., He, L., Yan, Z., Jiang, X., Wei, Y., Tang, K., Gelfand, M., Prufer, K., Kelso, J., Paabo, S., Giavalisco, P., Lachmann, M., and Khaitovich, P. (2014). Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nature Communications*, **5**(3584), –. 92, 93, 163

Kircher, M., Stenzel, U., Kelso, J., *et al.* (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol*, **10**(8), R83. 64

Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., *et al.* (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, **31**(3), 241–247. 29

Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., and Pääbo, S. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, **464**(7290), 894–897. 53

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H. A., Lalueza-Fox, C., de la Rasilla, M., *et al.* (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, **530**(7591), 429–433. 94, 95, 102, 120, 121, 128, 158, 164

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., *et al.* (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature*, **536**(7617), 419–424. 127, 156

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., and Duret, L. (2014). The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. *PLoS Genet*, **10**(11), e1004790. 47

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357), 493–496. 52, 61

Llorente, M. G., Jones, E., Eriksson, A., Siska, V., Arthur, K., Arthur, J., Curtis, M., Stock, J., Coltorti, M., Pieruccini, P., *et al.* (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science*, **350**(6262), 820–822. 18, 151

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, **193**(4), 1233–1254. 26, 27

Lopes, J. S., Balding, D., and Beaumont, M. A. (2009). PopABC: a program to infer historical demographic parameters. *Bioinformatics*, **25**(20), 2747–2749. 56

Lowe, J., Barton, N., Blockley, S., Ramsey, C., Cullen, V., Davies, W., Gamble, C., Grant, K., Hardiman, M., Housley, R., *et al.* (2012). Volcanic ash layers illuminate the resilience of Neanderthals and early modern humans to natural hazards. *Proceedings of the National Academy of Sciences*. 4

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V.-C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J., Crawford, J. E., *et al.* (2016). A genomic history of Aboriginal Australia. *Nature*. 3, 158, 167

Mallegni, F., Piperno, M., and Segre, A. (1987). Human remains of *Homo sapiens neanderthalensis* from the Pleistocene deposit of Sants Croce Cave, Bisceglie (Apulia), Italy. *American Journal of Physical Anthropology*, **72**(4), 421–429. 2

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., *et al.* (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 3, 30, 31, 158, 167

Marra, F., Ceruleo, P., Jicha, B., Pandolfi, L., Petronio, C., and Salari, L. (2015). A new age within MIS 7 for the *Homo neanderthalensis* of Saccopastore in the glacio-eustatically forced sedimentary successions of the Aniene River Valley, Rome. *Quaternary Science Reviews*, **129**, 260–274. 51

McDougall, I., Brown, F., and Fleagle, J. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, **433**(7027), 733–736. 2, 51

Mednikova, M. (2011). A proximal pedal phalanx of a Paleolithic hominin from Denisova cave, Altai. *Archaeology, Ethnology and Anthropology of Eurasia*, **39**(1), 129–138. 50, 70

Mednikova, M. (2013). Distal phalanx of the hand of *Homo* from Denisova Cave stratum 12: A tentative description. *Archaeology, Ethnology and Anthropology of Eurasia*, **41**(2), 146–155. 21

Mellars, P. (2004). Neanderthals and the modern human colonization of Europe. *Nature*, **432**(7016), 461–465. 2

Mellars, P., Boyle, K., and Bar-Yosef, O. (2007). *Rethinking the human revolution: new behavioural and biological perspectives on the origin and dispersal of modern humans*. McDonald Inst of Archeological. 10

Mendez, F. L., Watkins, J. C., and Hammer, M. F. (2012). A haplotype at STAT2 introgressed from Neanderthals and serves as a candidate of positive selection in PNG. *The American Journal of Human Genetics*, **91**(2), 265–274. 91

Mendez, F. L., Watkins, J. C., and Hammer, M. F. (2013). Neandertal origin of genetic variation at the cluster of OAS immunity genes. *Molecular biology and evolution*, **30**(4), 798–801. 91

Meyer, M., Kircher, M., Gansauge, M., Li, H., Racimo, F., Mallick, S., Schraiber, J., Jay, F., Prüfer, K., de Filippo, C., *et al.* (2012). A high-coverage genome sequence from an archaic

Denisovan individual. *Science*. 1, 5, 12, 21, 27, 45, 51, 52, 53, 73, 75, 84, 89, 94, 130, 166, 167

Meyer, M., Fu, Q., Aximu-Petri, A., Glocke, I., Nickel, B., Arsuaga, J.-L., Martínez, I., Gracia, A., de Castro, J. M. B., Carbonell, E., *et al.* (2014). A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, **505**(7483), 403–406. 21, 51

Meyer, M., Arsuaga, J.-L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., de Castro, J. M. B., Carbonell, E., *et al.* (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature*, **531**(7595), 504–507. 2, 21

Moorjani, P., Sankararaman, S., Fu, Q., Przeworski, M., Patterson, N., and Reich, D. (2016). A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proceedings of the National Academy of Sciences*, page 201514696. 31, 165

Morgan, M., Pagès, H., Obenchain, V., and Hayden, N. (2014). *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import*. R package version 1.20.4. 63, 168

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, **327**(5967), 876–879. 47

Noonan, J., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Pääbo, S., Pritchard, J., *et al.* (2006). Sequencing and analysis of Neanderthal genomic DNA. *Science*, **314**(5802), 1113. 12, 30

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Romero, I. G., Ayub, Q., Mehdi, S. Q., Thomas, M. G., Luiselli, D., *et al.* (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *The American Journal of Human Genetics*, **91**(1), 83–96. 18

Pagani, L., Lawson, J., Jagoda, E., Mörseburg, A., Clemente, F., Hudjashov, G., DeGiorgio, M., Eriksson, A., Saag, L., Wall, J., *et al.* (2016). Geographical barriers, environmental

challenges, and complex migration events during the peopling of Eurasia. *Nature*. 3, 31, 158, 167

Pearce, E., Stringer, C., and Dunbar, R. (2013). New insights into differences in brain organization between Neanderthals and anatomically modern humans. *Proceedings of the Royal Society B: Biological Sciences*, **280**(1758). 5, 7

Petraglia, M. and Rose, J. (2009). *The evolution of human populations in Arabia: paleoenvironments, prehistory and genetics*. Springer. 3

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient West Eurasian ancestry in Southern and Eastern Africa. *Proceedings of the National Academy of Sciences*, **111**(7), 2632–2637. 18

Pike, A. W., Hoffmann, D., García-Diez, M., Pettitt, P. B., Alcolea, J., De Balbin, R., González-Sainz, C., de las Heras, C., Lasheras, J. A., Montes, R., *et al.* (2012). U-series dating of paleolithic art in 11 caves in Spain. *Science*, **336**(6087), 1409–1413. 7

Plagnol, V. and Wall, J. (2006). Possible ancestral structure in human populations. *PLoS Genetics*, **2**(7), e105. 89

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, **5**(6), e1000519. 27, 169

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**(2), 945–959. 17

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., *et al.* (2014). The complete genome sequence of a Neanderthal from the Altai mountains. *Nature*, **505**(7481), 43–49. 2, 4, 7, 9, 12, 13, 14, 15, 21, 22, 25, 28, 29, 30, 35, 37, 38, 40, 43, 45, 50, 51, 52, 53, 64, 73, 75, 80, 83, 84, 89, 98, 126, 130, 133, 160, 170, 176

Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford Jr, T. W., Orlando, L., Metspalu, E., *et al.* (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, **505**(7481), 87–91. 129

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genet*, **10**(5), e1004342. 94

Reich, D., Green, R., Kircher, M., Krause, J., Patterson, N., Durand, E., Viola, B., Briggs, A., Stenzel, U., Johnson, P., *et al.* (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**(7327), 1053–1060. 10, 12, 13, 52

Rendu, W., Beauval, C., Crevecoeur, I., Bayle, P., Balzeau, A., Bismuth, T., Bourguignon, L., Delfour, G., Faivre, J.-P., Lacrampe-Cuyaubère, F., *et al.* (2014). Evidence supporting an intentional Neandertal burial at La Chapelle-aux-Saints. *Proceedings of the National Academy of Sciences*, **111**(1), 81–86. 7

Rodríguez-Vidal, J., dErrico, F., Pacheco, F. G., Blasco, R., Rosell, J., Jennings, R. P., Queffelec, A., Finlayson, G., Fa, D. A., López, J. M. G., *et al.* (2014). A rock engraving made by Neanderthals in Gibraltar. *Proceedings of the National Academy of Sciences*, **111**(37), 13301–13306. 7

Roebroeks, W., Sier, M. J., Nielsen, T. K., De Loecker, D., Parés, J. M., Arps, C. E., and Mücher, H. J. (2012). Use of red ochre by early Neandertals. *Proceedings of the National Academy of Sciences*, **109**(6), 1889–1894. 7

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. *PLoS Genetics*, **8**(10), arXiv:1208.2238. 19, 27, 88, 130, 131, 132, 134, 155, 165

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*. 21, 22, 28, 50, 55, 87, 89, 91, 92, 93, 95, 102, 103, 121, 122, 123, 124, 125, 127, 130, 143, 163

Sawyer, S., Renaud, G., Viola, B., Hublin, J.-J., Gansauge, M.-T., Shunkov, M. V., Derevianko, A. P., Prüfer, K., Kelso, J., and Pääbo, S. (2015). Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proceedings of the National Academy of Sciences*, **112**(51), 15696–15700. 1, 21, 166, 167

Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, **13**(10), 745–753. 23, 29, 30, 52, 71, 133, 165

Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., Malaspinas, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., *et al.* (2014). Genomic structure in Europeans dating back at least 36,200 years. *Science*, **346**(6213), 1113–1118. 132, 155, 156

Shea, J. (2008). Transitions or turnovers? Climatically-forced extinctions of *Homo sapiens* and Neanderthals in the east Mediterranean Levant. *Quaternary Science Reviews*, **27**(23), 2253–2270. 2

SIGMA, C. (2013). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*. 92, 163

Simonti, C. N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D. S., Chisholm, R. L., Crosslin, D. R., Hebbring, S. J., Jarvik, G. P., Kullo, I. J., *et al.* (2016). The phenotypic legacy of admixture between modern humans and Neandertals. *Science*, **351**(6274), 737–741. 163

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., Salzano, F. M., Patterson, N., and Reich, D. (2015). Genetic evidence for two founding populations of the Americas. *Nature*, **525**(7567), 104–108. 129

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M. B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. *The American Journal of Human Genetics*, **84**(6), 740–759. 53

Staab, P. R., Zhu, S., Metzler, D., and Lunter, G. (2015). Scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, **31**(10), 1680–1682. 96, 143

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(4), 605–635. 57, 58, 60

Stringer, C. (1996). *African Exodus: the Origins of Modern Humanity*. Cape, London. 3, 4

Stringer, C. (2000). Palaeoanthropology: coasting out of Africa. *Nature*, **405**(6782), 24–27. 3

Stringer, C. (2012). The status of *Homo heidelbergensis* (Schoetensack 1908). *Evolutionary Anthropology: Issues, News, and Reviews*, **21**(3), 101–107. 2

Stringer, C. (2016). The origin and evolution of homo sapiens. *Phil. Trans. R. Soc. B*, **371**(1698), 20150237. 2

Stringer, C. and Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science*, **239**(4845), 1263–1268. 10

Suk, E., McEwen, G., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D., McLaughlin, S., Peckham, H., *et al.* (2011). A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Research*, **21**(10), 1672–1685. 33, 49, 170

Tattersall, I. and Schwartz, J. H. (1999). Hominids and hybrids: The place of Neanderthals in human evolution. *Proceedings of the National Academy of Sciences*, **96**(13), 7117–7119. 8

Tishkoff, S., Reed, F., Friedlaender, F., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J., Awomoyi, A., Bodo, J., Doumbo, O., *et al.* (2009). The genetic structure and history of Africans and African Americans. *Science*, **324**(5930), 1035–1044. 17

Trinkaus, E. (1978). Dental remains from the Shanidar adult Neanderthals. *Journal of Human Evolution*, **7**(5), 369IN1377–376IN7382. 2

Trinkaus, E., Milota, Ş., Rodrigo, R., Mircea, G., and Moldovan, O. (2003). Early modern human cranial remains from the Peştera cu Oase, Romania. *Journal of Human Evolution*, **45**(3), 245–253. 9, 133

Vernot, B. and Akey, J. M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science (New York, NY)*. 21, 55, 87, 89, 91, 92, 95, 102, 103, 121, 122, 123, 124, 125, 127, 130, 163

Via García, M., Consortium, . G. P., *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 2012, vol. 491, p. 56-65*. 176

Wall, J. and Kim, S. (2007). Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genetics*, **3**(10), e175. 12

Wall, J. D., Yang, M. A., Jay, F., Kim, S. K., Durand, E. Y., Stevison, L. S., Gignoux, C., Woerner, A., Hammer, M. F., and Slatkin, M. (2013). Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics*, **194**(1), 199–209. 13, 89

Wang, S., Lachance, J., Tishkoff, S. A., Hey, J., and Xing, J. (2013). Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from non-African populations. *Genome Biology and Evolution*, **5**(11), 2075–2081. 17

Wolpoff, M., Hawks, J., Frayer, D., and Hunley, K. (2001). Modern human ancestry at the peripheries: a test of the replacement theory. *Science*, **291**(5502), 293–297. 10

Wolpoff, M. H., Smith, F. H., Malez, M., Radovčić, J., and Rukavina, D. (1981). Upper Pleistocene human remains from Vindija Cave, Croatia, Yugoslavia. *American Journal of Physical Anthropology*, **54**(4), 499–545. 2

Yang, M., Malaspinas, A., Durand, E., and Slatkin, M. (2012). Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Molecular Biology and Evolution*, **29**(10), 2987–2995. 19, 20

Zilhão, J., Angelucci, D. E., Badal-García, E., dErrico, F., Daniel, F., Dayet, L., Douka, K., Higham, T. F., Martínez-Sánchez, M. J., Montes-Bernárdez, R., *et al.* (2010). Symbolic use

of marine shells and mineral pigments by Iberian Neandertals. *Proceedings of the National Academy of Sciences*, **107**(3), 1023–1028. 7