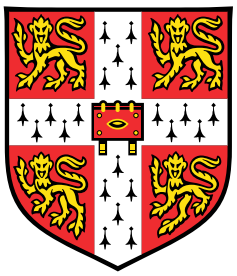
Investigating the genetic diversity, population structure and archaic admixture history in worldwide human populations using high-coverage genomes



Ruoyun Hui

Supervisor: Dr Aylwyn Scally

Department of Genetics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Peterhouse July 2019

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Ruoyun Hui
July 2019

I dedicate this thesis to the courageous, the affectionate, and the silenced voices, who preserve our compassion towards fellow human beings through public and private resistance.

Acknowledgements

I am most grateful to my supervisor, Aylwyn Scally, who offered invaluable guidance and support throughout my PhD research, but at the same time encouraged my own explorations, even in seemingly irrelevant directions. Many thanks to Chris Tyler-Smith, Yali Xue, and Anders Bergström for hosting me at the Wellcome Trust Sanger Institute and offering insightful feedback. I also want to thank Laurits Skov and Richard Durbin for the opportunity to collaborate on an ingenious model. I am indebted to my examiners, Chris Jiggins and Stephan Schiffels, for insightful discussions during the viva voce and valuable suggestions on improving this thesis. I would like to thank Gates Cambridge whose generous funding made it possible for me to pursue the degree.

I also want to thank my family, especially my parents, who have always been supportive when I choose to research on something they never truly grasp, instead of earning a decent wage while staying closer to home. I am indebted to many friends, old and new, near and far, for caring for and believing in me. A very special thank you to Gustavo Nicolas Paez for his continuous understanding and support, and for convincing me not to acknowledge all the red pandas on Twitter whose cuteness offered immediate relief of stress.

Finally, my sincere gratitude goes to my sensei and senpai at two dojo, Cambridge University Kendo Club and Kenyukan Suffolk. You have taught me much about resilience, integrity and courage that are indispensable in research and in life.

Abstract

I present the analysis on 929 high-coverage (>30x) genomes from the Human Genome Diversity Project (HGDP) panel, a collection of cell lines from 54 populations across the world. Some data processing steps were necessary for downstream analysis, including lifting over resources on a different reference genome assembly, annotating the genome, and statistical phasing. Genome-wide genetic diversity conforms with previous studies using SNP arrays and microsatellites, yet haplotype information reveals fine scale structures and recent demographic history that vary between populations.

This dataset also provides a valuable opportunity to explore the diversity and distribution of archaic segments in modern human populations. I implemented a hidden Markov model to detect such segments, based on patterns of allele-sharing with sequenced archaic genomes and a sub-Saharan African control panel. I also compared several variants of the model and different training methods using simulated data. Applying the model on the HGDP dataset using two Neanderthal genomes and one Denisova genome, I detected variations in the level of archaic ancestry across continental regions, populations, and individuals within each population. I further compared Neanderthal and Denisovan segments regarding their lengths, genomic distribution, divergence to the archaic genomes, nucleotide diversity, and haplotype networks to shed light on the structure of the admixture events. Neanderthal segments from all non-African populations appear largely homogeneous after accounting for the recent demographic history of modern human populations, which is consistent with a single admixture event that happened before they diverged from each other. In contrast, a distinct separation exists between Denisovan haplotypes recovered from Oceania and those from East/South Asia, whilst the complicated structure in the latter cannot be explained by a single source of gene flow. Therefore I propose that more than one episode of admixture with different Denisova groups occurred in the ancestral population of present-day East Asian, South Asian and American populations after the separation from the ancestors of present-day Oceanians, and that a separate admixture event occurred between the ancestors of Oceanians and the Denisova population.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Genetic structure of human populations

1.1.1 From classical to DNA makers

Our understanding of the genetic variation in humans dates back to a century ago. Since the early 20th century, researchers have noticed that the frequencies of blood groups differ between populations [2, 3]. Along with other classical markers such as human leukocyte antigen (HLA), serum protein and enzyme variants, polymorphism at protein level that follows Mendelian inheritance was used to study population structure (e.g. [4, 5]) as well as the relationship between populations (e.g. [6, 7]). Combining different markers, Cavalli-Sforza *et al.* identified three major modern human groups consistent with most people's expectation at that time: an Asiatic group (including indigenous Americans and Oceanians), a European group and an African group [6]. Studies using blood group and protein polymorphism also revealed that the majority of genetic variation exists between individuals within the same population or ethnic group rather than between groups [6, 8, 9], calling an end to many attempts for racial classification based on genetic variations.

Despite the ease of measurement, protein polymorphism only reflects a small fraction of the variations at the DNA level, and the markers are likely to be under natural selection. Since the 1980s, technological advances have enabled direct assaying at the DNA level, first through restriction fragment length polymorphism (RFLP) [10–12], then with the wide application of polymerase chain reaction, through microsatellites (or short tandem repeats, STR) [13, 14], and lastly through single nucleotide polymorphism (SNP) by microarrays or direct sequencing [38, 15, 16]. A pioneering study by Cann *et al.* examined RFLP in

mitochondrial DNA (mtDNA) in 147 individuals to obtain a population tree that clearly supports an African origin of all modern humans ("Out of Africa" or "Recent African Origin" model) [12], in contrast to the "multiregional" model that argues for a genetic continuity between archaic hominins and modern humans inhabiting the same geographical regions [17]. This was corroborated by later mtDNA studies that dated the most recent common ancestor (MRCA) of modern humans in Africa to roughly 200k years ago, and detected signals of historical bottlenecks and expansions in some populations but not others [18–20]. The Y chromosome was also targeted in studies using microsatellites and SNP markers; similarly, the MRCA is placed parsimoniously in Africa where the highest genetic diversity is found, and signs of population growth were detected in non-African populations [21–23]. The effective population size for modern humans was estimated to be around 10,000 in both mtDNA and Y chromosome studies, also inconsistent with the "multiregional" model [19, 21]. Early studies using DNA markers on the recombining nuclear genome usually included only one or a few loci near protein-coding regions; although an African origin was still supported, they yielded highly variable estimates for the time to MRCA and expansion patterns [24]. Subsequent studies selected genomic regions that are more likely to be neutral, and also detected recent population expansions [25, 26].

1.1.2 The Human Genome Diversity Project

The Human Genome Diversity Project (HGDP) was envisioned in the early 1990s when DNA markers continued to exhibit increasing power in revealing the origin and history of modern humans. It is an international collaboration initiated by researchers at the Morrison Institute of Stanford University to document the genetic variation of the human species worldwide [27]. The project was in part a response to the Human Genome Project, which only aimed to sequence a single consensus human genome. Instead, the organizers of HGDP believed that it is crucial to explore the genetic variations in diverse human populations to understand the evolutionary history and genetic structure in our species, especially at a time when many isolated human populations were being rapidly absorbed into neighbouring populations [27]. The study of genetic variations can also help to understand fundamental processes of mutation and adaptation, as well as factors contributing to diseases.

The project initially planned to sample 500-700 populations, with 25-150 individuals from each. The DNA would be preserved through immortalized cell lines stored in repositories available to future researchers, along with the analysis result from genetic markers, but commercial use would be prohibited [27]. However, political, ethical, legal and social

controversies have impeded the project since the beginning [28]. Opponents were concerned about, for example, whether the discussion on genetic variation would fuel scientific racism, how informed consent could be given for unknown research purposes over indeterminate time, what rights the populations should have apart from the individuals being sampled, and whether the indigenous populations would be exploited under biocolonialism in such an initiative largely led by western scientists. Despite the effort from the HGDP to address these concerns and adopt an ethical guideline, it faced strong oppositions from indigenous activists groups. Eventually, the plan to collect phenotypic and medical data had to be abandoned, which greatly limited the medical implications of the collection.

The project was stalled until 1997, when a committee convened by the US National Research Council of the National Academy of Sciences finally allowed it to proceed. Sampling continued despite a serious lack of funding, thanks to researchers donating their collections from separate projects. The final collection consists of 1,064 lymphoblastoid cell lines from 1,050 individuals sampled from 52 populations around the world, deposited at Centre d'Etude du Polymorphisme Humain (CEPH) at the Foundation Jean Dausset in Paris [29]. Laboratories that require DNA samples from the panel are obliged to share their results in the CEPH database.

By the time the HGDP collection was established, researchers had realized the importance of integrating evidence from independent loci across the genome. Analysing 377 autosomal microsatellite loci on genomes from the HGDP panel, Rosenberg *et al.* found that model-based genetic clustering is able to recover population origins at the continental level without prior information, but on a smaller scale the genetic clusters only correspond to the pre-defined populations in America and Oceania, where smaller effective population sizes caused genetic drift to occur more rapidly [30]. The genetic structure is reported to be less pronounced in Eurasia. They also identified genetic isolates, Kalash being a most notable example. When more microsatellite markers were genotyped, the observation that heterozygosity decreases linearly with distance from Africa was interpreted as a serial founder effect, a model where a series of expansions happened as modern humans spread from a single source in Africa, each time drawing a small number of founders from the previous population [31]. Microarray technology subsequently enabled genotyping as many as a million SNP loci efficiently. 650k common SNPs in unrelated HGDP individuals led to more clearly separated regional clusters and made it possible to observe fine-scale population structures and admixture patterns [32]. For example, a north-south cline was found in East Asia, and European populations which were indistinguishable in microsatellite studies now

form separate clusters. Many populations in the Middle East and Central/South Asia were found to be admixed, which could be explained by either recent admixture or shared ancestry. It became clear that the genetic diversity in human populations is shaped by a combined process of serial founder effect, isolation by distance, long-range migration and gene flows [33].

1.1.3 Genome sequencing era

As high-throughput sequencing technologies greatly reduced the sequencing cost, whole genome sequencing (WGS) has fueled the next breakthrough in population genetics studies. WGS data not only overcome the ascertainment bias in microarray-based SNP discovery, but also facilitates the study on structural variants and fine-scale haplotype structure. New statistical methods have been developed to also exploit linkage disequilibrium patterns in the genome, greatly enhancing the resolution for fine-scale population structure and demographic history [34–37].

After the establishment of the HGDP collection, the HapMap Project [38] and the 1000 Genomes Project [16] as its continuation have contributed enormously to describing the genetic variations in major human populations. The final phase of the 1000 Genomes Project reconstructed genomes of 2,504 individuals from 26 populations, highlighting the internal substructure at continental level and a shared ancestry of all populations prior to 150-200k years ago [16]. But due to an interest in biomedical implication, only demographically large populations were represented. More recently, large-scale sequencing projects have included smaller and more isolated populations to address population diversity and history: the Simons Genome Diversity Project (SGDP) presented 300 genomes from 142 populations worldwide, including 2-4 HGDP individuals from each population [39]; the Estonian Biocentre Human Genome Diversity Panel included 483 individuals from 148 populations [40]. Both studies established the split times between populations from major geographical regions and refined estimates on the amount of archaic admixture, although they disagreed over whether an unknown early out-of-Africa lineage has contributed to the ancestry in Papuans.

Throughout these developments, the role of the HGDP collection remains central. To date, more than 100 investigators have requested DNA from the HGDP panel for genotyping or sequencing [41]. From the first study using autosomal microsatellites to exome and whole genome sequencing for a subset of the samples, the CEPH database became populated with increasingly detailed data on the indels, copy number variation, and single nucleotide polymorphism (SNP) of the samples, shedding light on not only population structure and

demographic history, but also the fundamental patterns of microsatellite variation, linkage disequilibrium, and runs of homozygosity in the human genome [30, 42, 31, 43, 15, 44, 32, 45–48, 39]. Many more regional or national biobanks and sequencing projects have come into being (e.g. [49–51]); nevertheless, the HGDP panel remains a widely consulted reference for ancestry mapping, demographic inference as well as functional analysis.

Recently, the Human Evolution team led by Dr. Chris Tyler-Smith at Wellcome Trust Sanger Institute (WTSI) have sequenced 801 unrelated individuals from the HGDP panel to high coverage (>30X), including 26 genomes also sequenced on the 10x Genomics platform to allow physical phasing. Together with 128 previously sequenced ones [39], these high-quality whole genome sequences improve our power to detect rare variants and delineate haplotype structure, which would hopefully benefit the human genetics community.

1.2 Archaic ancestry in modern human genomes

1.2.1 Relationship between modern humans and archaic hominins

Current views on speciation as a divergence process along a continuum imply that a concrete species boundary may not exist [52]. Reticulate evolution was believed to be more prevalent in prokaryotes through horizontal gene transfer, but accumulating genomic data suggest it is also more common than anticipated in eukaryotes [53]. Around 25% of plant species, for example, are estimated to undergo gene exchange with related species via hybridization [54]; ongoing or past gene flows between closely related groups have also been observed in Darwin's finches [55], cichlid fish [56], *Anopheles* mosquitoes [57] and *Heliconius* butterflies [58]. The incorporation of alleles from one group into another divergent group, termed introgression, may contribute to local adaptation or speciation [52].

The origin and history of our own species are also inevitably intertwined with archaic forms of humans. Ever since the discovery of the first Neanderthal fossil in the mid 19th century, there has been contention over their relationship with modern humans. Partially fueled by an emphasis on the similarity between Neanderthals and modern humans that had gained ground in the second half of the 20th century, the multiregional model proposed evolutionary continuity between Neanderthal and modern humans [17]; in contrast, the out-of-Africa model stated that when modern humans expanded into Eurasia, they replaced the Neanderthal populations living there. Genetic evidence since the 1990s consistently supported the latter, yet new discussion arose over whether a low level of Neanderthal ancestry had been assimilated into modern humans through interbreeding. Morphological comparisons revealed

that early modern human in Europe shared some anatomical features with the Neanderthals [59], whilst genetic studies detected deeply diverged haplotypes that are rare in sub-Saharan Africa [60, 61]. Access to the Neanderthal genomes finally provided unequivocal evidence for archaic admixture. Although the reconstructed Neanderthal mitochondrial genome suggested no contributions to the modern human mtDNA pool [62–64], a draft sequence of the Neanderthal nuclear genome was found to share more derived variants with present-day non-Africans than sub-Saharan Africans, suggesting gene flow from the Neanderthals into ancestors of present-day non-Africans [65]. The finding has been corroborated when high-coverage Neanderthal genomes became available [66, 67] and when more Neanderthal genomes were analysed [68, 69]. Additionally, genetic contributions from modern humans have also been found in Neanderthals from the Altai Mountains [68], and gene flow from an African modern human source into the Neanderthals has also been suggested based on higher mtDNA similarity between modern humans and Neanderthal than between modern human and Denisova [70, 71].

Another archaic hominin group, the Denisovans, has only been characterised genetically to date. A manual phalanx bone of a hominin was excavated in 2008 from Denisova Cave, where fossils of Neanderthals were also found. Subsequently named after the cave, its mtDNA suggests a deep divergence time of one million years from the ancestors of Neanderthal and modern humans [72], but the nuclear genome places it as a sister group to the Neanderthals and reveals its genetic contribution to Near Oceanians, aboriginal Australians and insular Southeast Asia [73, 47]. Small amounts of Denisovan ancestry have been subsequently detected in East Asians [74], indigenous Americans [66, 75] and South Asians [76] as well. More recently, even a first-generation hybrid between a Neanderthal and a Denisovan was discovered in the same cave [77], suggesting that interbreeding between different human groups could have been common. As African genomes were found to share more derived alleles with Neanderthals than Denisovans, it has also been suggested that Denisovans could have received gene flow from another hominin that diverged early from the lineage leading to modern humans, Neanderthals and Denisovans [66].

1.2.2 Methods for detecting archaic introgression

As mentioned in the previous section, speculations of archaic introgression and efforts to detect it predate the successful sequencing of the genomes of archaic hominins. Evans *et al.* found a positively selected haplotype of the *MCPH1* gene likely originated from a lineage separated from modern humans 1.1 million years ago [60]. Plagnol and Wall developed the

S^* statistics to systematically search for private SNPs in linkage disequilibrium (LD) over windows of a suitable length, which are likely to constitute an archaic haplotype [61].

Accurately identify localised introgression became possible with genomic data for both the source and the target groups. Although model-based statistical frameworks were available [78–80], they are usually computationally demanding or unrealistic in certain aspects [81]. Instead, most studies relied on the heterogeneity in summary statistics measuring population differentiation: introgressed regions are expected to show higher similarity to the source group relative to a sister group than unadmixed regions. F_{ST} [82] was a widely used measurement (e.g. [83, 84]), but has been shown to be confounded by local mutation rate and genetic diversity [85, 86]. Another popular choice is D-statistic [65, 73, 87], which was initially developed as a genome-wide measurement and has been shown to possess inherent bias when applied to small genomic regions [88]. More recent studies have proposed new summary statistics that derive from the above and other divergence measures [89, 88, 81].

When the genomes of archaic humans became available, D-statistic were frequently used to measure the genome-wide level of archaic ancestry based on the asymmetrical sharing of drift in a population tree [65, 73, 87, 47, 66], whilst principal component analysis has also been used to illustrate the different affinities to archaic genomes by projecting modern human genomes onto principal components defined by archaic hominins and chimpanzee [73, 74]. When it comes to inferring locations of individual archaic segments, most studies adopted a probabilistic approach that combines the pattern of sequence divergence, LD, and/or local recombination rate. Vernot and Akey trained a generalized linear model through simulations to determine the critical values for S^* according to local recombination rate and nucleotide diversity, before filtering these putative segments by their affinity to the archaic genomes [90, 91]. Sankararaman and others developed a conditional random field (CRF) based on allelic patterns, divergence to a panel of Yoruba haplotypes, and haplotype lengths, although the second feature was omitted when searching for Denisovan haplotypes due to a bias when the archaic ancestry is at the order of 1/1000 [1, 76]. Steinrücken *et al.* presented another method to infer the genealogy of haplotypes through explicit modeling of the demographic history between modern human and archaic populations [92].

Recently researchers have devoted more effort to detecting introgressed segments without access to the archaic reference. Browning and others improved the S^* method so that it can be applied to various sample sizes [93]. S^* score is also included together with other summary statistics in a logistic regression model, which was able to discover unknown archaic ancestry in West Africa [94]. Along a different line, Skov *et al.* used a hidden Markov model (HMM),

where the observed variable is the number of private SNPs in genomic windows as compared to an unadmixed outgroup, to detect introgression from a deeply diverged lineage [95]. I was also involved in evaluating this model, which will be described in more detail in Section 4.6. These reference-free methods are useful to study admixture from unknown "ghost" population and from Denisova, where the sequenced Altai Denisova genome has diverged from the actual source of admixture by hundreds of thousands of years [66].

1.2.3 Geographical structure of archaic introgression

The proportion of Neanderthal ancestry has been consistently found to be around 12-20% higher in modern East Asia and America compared to Europe [74, 47, 96, 1, 90]. Since the estimated time of Neanderthal admixture predates the split time between present-day Europeans and East Asians [97, 98], the difference cannot arise from completely independent gene flows. Assuming that most Neanderthal alleles are weakly deleterious in the genetic background of modern humans, Sankararaman *et al.* hypothesized that purifying selection could have been stronger in Europe than in East Asia, considering a smaller effective population size in the latter due to stronger historical bottlenecks, which led to more Neanderthal segments being removed in Europe [1]. However, simulation studies do not support this explanation [99, 100]. Alternative explanations involve either an additional pulse of Neanderthal gene flow into the ancestor of East Asians and Americans, or dilution of Neanderthal ancestry in Europe through admixture with another population carrying no or very low level of Neanderthal ancestry [99, 100]. Studies on the ancestry of early European farmers have identified a component from a lineage that split off from other non-Africans before their diversification and received little or no Neanderthal admixture, which could have caused the dilution of Neanderthal ancestry in Europe [101, 102], although this hypothetical "Basal Eurasian" lineage has not yet been associated with actual populations or sites. On the other hand, the discovery of an early modern human from approximately 40k years ago with a recent Neanderthal ancestor clearly supports that the admixture could have happened multiple times [103].

Beyond Eurasia, Oceania was found to harbour substantial Neanderthal ancestry, but studies disagree over whether the amount is almost the highest [76, 39] or lower than in Eurasia [91], possibly due to both methodological and sampling differences. In contrast, Denisovan ancestry is mostly concentrated in Oceania, although later studies also detected it in East Asia, South Asia, and America at a much-reduced level [74–76, 39]. No substantial Denisovan introgression has been reported in Europe or the Middle East so far.

Further effort to delineate the structure of archaic admixture events has examined not only the amount of archaic ancestry but also the relationship between archaic segments found in modern humans from different geographical regions. Modelling the pattern of the reciprocal match between archaic segments across regions, Vernot *et al.* suggested at least three pulses of Neanderthal gene flow, the first one into all non-Africans, the second into Europe, East Asia, and South Asia, and the last one into East Asia alone [91]. The same study only detected Denisovan ancestry in Near Oceania, therefore assuming a single pulse of Denisovan admixture. However, Browning *et al.* reported that in addition to a shared component in East Asia, South Asia, and Oceania, there is an additional Denisovan component uniquely present in East Asia [93]. Their detection method does not rely on the archaic genomes, hence separate clusters in the match scores between the inferred archaic haplotypes in modern genomes and the archaic genomes are interpreted as separate pulses of admixture. No such structure was detected in Neanderthal segments; thus the authors reached the conclusion of two waves of Denisovan admixture and a single wave of Neanderthal admixture, but had difficulty determining the sequence of them. More recently, Villanea and Schraiber simulated the joint frequency spectrum of Neanderthal segments between Europe and East Asia and found that a model with an original pulse of gene flow into the ancestral Eurasian population followed by additional pulses into both the European and East Asian population provides the best fit to the observed spectrum [104]. So far there have not been comprehensive surveys of archaic segments in populations worldwide, nor attempts to infer the admixture history from global genetic data.

1.2.4 Functional consequences of archaic introgression

The consequences of hybridisation and introgression have long been debated [105–107]. It could reinforce species barriers if the hybrids are generally less fit than the parental groups [108]; but recent studies also show evidence where the variants introduced are adaptive in the local environment, including mimetic wing patterns in *Heliconius* butterflies [58], beak shape in Darwin's finches [109], and insecticide resistance in *Anopheles* mosquitoes [110]. Both mechanisms have also been discussed in the case of archaic introgression in modern humans.

Once the archaic haplotypes entered the modern human population, they have been subject to not only genetic drift associated with demographic changes, but also natural selection on a different genetic background. Studies repeatedly found that archaic ancestry is unevenly distributed in modern human genomes: large regions totally depleted of archaic ancestry

also tend to be gene-rich regions [1, 76, 91]. In particular, both Neanderthal and Denisovan ancestry are found to be significantly reduced on the human X chromosome, consistent with reduced male fertility as an initial stage of reproductive isolation. This has been interpreted as negative epistatic interactions between modern human and archaic alleles, but subsequent analyses have pointed out that the small effective size of Neanderthal and Denisovan populations would lead to a higher mutational load, where natural selection is not strong enough to purge out many weakly deleterious alleles; but in the larger modern human population, they would become the targets of stronger purifying selection without the need to invoke genetic incompatibilities [111, 112, 92]. Neanderthal ancestry is also enriched in genomic regions with higher recombination rates where linkage to potentially deleterious alleles is weaker [113]. The selection against Neanderthal alleles was considered to have happened throughout a long time, based on a monotonous decline of Neanderthal ancestry in the past 45k years in Europe revealed by ancient genomes [114]; however, Petr *et al.* showed that the pattern could be an artefact when not accounting for later gene flows between western Eurasia and Africa, as more rigorous measurement results in nearly constant level of Neanderthal ancestry across time [115]. The lack of decline in Neanderthal ancestry also implies no significant dilution from the "Basal Eurasian" lineage. Their simulations suggest that the abrupt removal of Neanderthal alleles by negative selection mainly happened in the first few generations after the admixture.

Nevertheless, many archaic alleles still survived in the modern human genomes. The ancestors of Neanderthal and Denisova entered Eurasia hundreds of thousands of years before modern humans, where they encountered climate, pathological and ecological environments vastly different from those in Africa. Consequently, we would expect new adaptations to have developed during such a long time span. When modern human also expanded into Eurasia, some of the archaic alleles acquired through admixture might help them to adapt to the local environment despite the overall deleterious effect of archaic ancestry. In Eurasian genomes both Neanderthal and Denisovan alleles have been found in *HLA class I*, an important component in the immune system that is subject to balancing selection [116]. Another gene related to the immune system, *STAT2*, also has a haplotype closely matching that of the Neanderthal [117]. Perhaps the most notable example of adaptive introgression is the *EPAS1* gene, where the Denisovan haplotype that conveys an advantage to life at high altitude reaches a much higher frequency in Tibetans in comparison to the neighbouring Han Chinese population [118]. In addition, studies on the genomic distribution of archaic ancestry found that genes related to skin and hair colour are enriched for Neanderthal ancestry, [90, 1, 76], and those involved in fat metabolism are enriched for Denisovan ancestry [76].

Querying a GWAS database with Neanderthal alleles also suggests connections to smoking behaviour, optic disk size, type 2 diabetes, lupus, bone abnormalities, and celiac diseases [1, 119]. Some introgressed haplotypes also act as expression quantitative trait loci (eQTL) to genes involved in immune functions across multiple tissues, suggesting adaption at the level of gene regulation [119]. Two recent studies directly targeted the association between identified Neanderthal alleles and phenotypic records on large population samples [120, 121]; added to the list of traits affected by Neanderthal ancestry are height, sleeping patterns, mood, blood-clotting disorders, and skin lesions. Some of these phenotypes might reflect the environmental challenges faced by the archaic hominins, such as sunlight exposure.

1.3 Thesis overview

The HGDP dataset represents one of the most comprehensive collections of high-coverage WGS data from diverse human populations worldwide. In particular, this thesis focuses on describing the genetic structure and archaic introgression in the HGDP genomes, and their implications on the history of our species. Some other intriguing aspects of this dataset such as Y-chromosome haplogroups, structural variants, rare allele sharing, and population size history are being investigated by collaborators on the project. By including 10-20 unrelated genomes from most of the populations, the current project allows discussion at population and subpopulation level for the first time. In addition to fine-scale population structure and relationship, this dataset also provides an unprecedented opportunity to investigate the structure and diversity of archaic segments in modern human genomes, which might provide new clues to how many separate admixture events there were, when and where the admixture happened, how many archaic hominins contributed to the modern human gene pool, and many other unanswered questions.

After this introduction, Chapter 2 briefly describes the sequencing data and some preparatory procedures for downstream analysis. Chapter 3 explores genetic diversity and population structures in this dataset and their implications on the history of human populations. Following a review of previous methods to detect archaic segments in modern populations, Chapter 4 describes the implementation and evaluation of a hidden Markov model for this purpose, together with another reference-free method that I also contributed to. The result of applying the hidden Markov model to the HGDP dataset is discussed in Chapter 5, where the geographical variations in archaic segments form the basis to infer the structure of the admixture events. Finally, Chapter 6 concludes the thesis with a summary and future recommendations.

Chapter 2

Data preparation

2.1 Sequencing data

Using microsatellite markers, Rosenberg *et al.* excluded duplicates, closely-related and mislabeled samples to establish a subset of 952 unrelated (up to second-degree) samples among all 1,064 included in the HGDP panel [43]. 135 of them had been deep-sequenced previously as part of the Simons Genome Diversity Project [39], and 10 along with the publication of the high-coverage Denisova genome to measure archaic admixture [47]. In addition to genomes from the unrelated subset that have not been sequenced before, 22 samples were re-sequenced at the Wellcome Trust Sanger Institute (WTSI) to assess cross-platform reproducibility and batch effect, but only one genome with the best quality was included in the final dataset for each individual. The first 178 genomes were sequenced on Illumina HiSeq X using PCR-based libraries (the pilot dataset); the remaining ones were sequenced on the same platform using PCR-free libraries. All the reads have been mapped onto GRCh38 reference assembly. After excluding genomes with quality problems, the final dataset consists of 128 genomes from previous studies and another 801 sequenced at WTSI. The source, library type, accession number and population information of each sample are listed in Appendix A. 26 of these genomes from 13 populations were also sequenced with linked-reads using 10x Genomics, which allows physical haplotype phasing.

2.2 Lift-over genomes and resources

The Genome Reference Consortium regularly updates the human reference genome assembly. Between major releases, there can be substantial changes in genotypes and coordinates. The HGDP sequences were mapped to the most recent assembly, GRCh38, which was first released in 2009, but many resources are still based on earlier builds. Conversion from older reference genome builds to the new one (lift-over) becomes a routine task in order to incorporate external data into analyses.

The relationship between the old and new genomic coordinates is usually described in a chain file in the form of alignments. Several web-based or standalone tools are able to batch convert genomic coordinates according to the chain files; I used CrossMap [122] for most lift-over tasks as it directly handles VCF and BED files. If a position in the old build is deleted or merged in the new build, or in the case of VCF files, has a different genotype, the record will be ignored. Such unmapped records only constitute less than 0.1% of all records.

2.3 Annotation of genomic features

I consulted external databases to annotate features such as genomic function and ancestral state. The information was either added to the VCF files using *Vcfanno* [123] or queried from scripts when needed. Below is a summary of the resources:

- Ancestral alleles in the human genome were downloaded from Ensembl FTP server (http://www.ensembl.org/info/data/ftp/index.html). The ancestral states were inferred from an EPO alignment [124] of 12 primate species.
- *B* values use the reduction in genetic diversity to measure the strength of linked selection [125]. Values across the human genome on GRCh36 were downloaded from http://www.phrap.org/othersoftware.html and lifted-over to GRCh38.
- A list of known genes [126] was downloaded from UCSC web server (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/).
- A collection of GWAS SNP-trait associations were downloaded from GWAS Catalog (https://www.ebi.ac.uk/gwas/).

2.4 Statistical phasing

Some downstream analyses, such as archaic segment detection and local ancestry inference, rely on haplotype information. As the main HGDP dataset was not phased, haplotypes were estimated through statistical phasing. Most computational phasing methods nowadays adopt a hidden Markov model and update the haplotype frequencies in the population and the haplotype configuration of individuals iteratively. Of the two methods used here, SHAPEIT2 reduces the complexity by only sampling haplotypes consistent with an individual's genotype, and improves the accuracy by updating the transition probabilities locally from K closest haplotypes [127]. It is also capable of incorporating sequencing reads and family information. Eagle2 is a more recent development that losslessly stores the full haplotype structure using the positional Burrows-Wheeler transform (PBWT) and explores only the most likely phase paths [128].

Haplotype structure of the 26 genomes sequenced with 10x Genomics is retained using linked reads. Although the 10x strategy is not completely free of errors, its accuracy is much higher than statistical phasing [129]. I evaluated the error rate of different phasing strategies using these physically phased genomes as the gold standard.

2.4.1 Cohort-based phasing

I first tested phasing all the 929 HGDP genomes internally without a reference panel using SHAPEIT2. Phasing accuracy is commonly measured by switch error rate, which equals the number of switches needed to convert the phased result to the truth divided by the total number of possible switches. When comparing the statistically phased data and 10x sequencing data of the same genome, only phase-resolved heterozygous sites with compatible genotypes in both datasets were considered. Comparisons were only meaningful within the same phasing block, indicated by the same phase set ID outputted in 10x sequences. The number of possible switches in this case is the number of comparable heterozygous sites minus one.

Table 2.1 shows the switch error rate in 13 individuals whose 10x sequencing data were available at the time of analysis. Three rates were calculated: without any mask, with the 1000 Genome pilot mask and with the 1000 Genome strict mask. Switch error rate was usually reduced after masking out less accessible regions. The error rate is below 0.03 in most populations, but rises to around 0.05 in Mbuti and 0.11 in San population. The performance of cohort-based phasing depends heavily on the sample size in relation to the

population diversity. Both San and Mbuti are groups traditionally hunter-gather societies with a high genetic diversity, yet a small sample size (6 and 13 respectively) in the HGDP collection. Their isolated status also means that the algorithm cannot effectively learn about their haplotype frequency from other populations.

Table 2.1 Switch error rates in cohort-based phasing

Sample	Population	Region	Switch error rate			
			SHAPEIT2 no mask	SHAPEIT2 pilot mask	SHAPEIT2 strict mask	Eagle2 strict mask
HGDP00450	Mbuti	Africa	0.0550	0.0502	0.0495	0.0421
HGDP00460	Biaka	Africa	0.0334	0.0300	0.0283	0.0234
HGDP00547	PapuanSepik	Oceania	0.0322	0.0270	0.0250	0.0289
HGDP00551	PapuanHighlands	Oceania	0.0363	0.0307	0.0269	0.0317
HGDP00580	Druze	Middle East	0.0227	0.0183	0.0166	0.0160
HGDP00670	Sardinian	Europe	0.0213	0.0162	0.0143	0.0134
HGDP00774	Han	East Asia	0.0259	0.0218	0.0205	0.0216
HGDP00819	Han	East Asia	0.0315	0.0275	0.0260	0.0270
HGDP00930	Yoruba	Africa	0.0395	0.0368	0.0365	0.0235
HGDP01032	San	Africa	0.1135	0.1096	0.1103	0.0922
HGDP01043	Pima	America	0.0128	0.0071	0.0044	0.0065
HGDP01067	Sardinian	Europe	0.0250	0.0189	0.0182	0.0167
HGDP01081	Mbuti	Africa	0.0489	0.0451	0.0436	0.0376

I also tested cohort-based phasing in Eagle2 on the same data. Compared to SHAPEIT2, Eagle2 improves the accuracy slightly in most individuals (last column in Table 2.1) at only a fraction of its running time.

2.4.2 Using a reference panel

One way to increase the sample size is to use a phased reference panel that includes individuals from related populations. Table 2.2 shows the switch error rates on chromosome 9 when using Eagle2 to phase the HGDP dataset with genomes first from Phase 3 of the 1000 Genomes Project (1KG) [16], and then those from the African Genome Resource (AGR) as the reference panel, in comparison to cohort-based phasing without a reference panel. The AGR contains a total of 4,957 genomes, including all genomes from 1KG and additional ones sampled in sub-Saharan Africa (including samples from [130] and other unpublished genomes). The number of PBWT iterations was set to three in all runs. All error rates were calculated after applying the strict mask.

The use of a reference panel greatly reduced the error rate in all tested individuals, sometimes by more than two thirds. Between 1KG and AGR panels, denser sampling in Africa in the

Table 2.2 Switch error rates using Eagle2 with different reference panels

Sample	Population	Region	Switch error rate			
			no reference	1KG reference	AGR reference	merged with AGR
HGDP00450	Mbuti	Africa	0.0421	0.0168	0.0078	0.0686
HGDP00460	Biaka	Africa	0.0234	0.0062	0.0057	0.0567
HGDP00547	PapuanSepik	Oceania	0.0289	0.0078	0.0105	0.0430
HGDP00551	PapuanHighlands	Oceania	0.0317	0.0078	0.0116	0.0543
HGDP00580	Druze	Middle East	0.0160	0.0066	0.0038	0.0201
HGDP00670	Sardinian	Europe	0.0134	0.0052	0.0042	0.0168
HGDP00774	Han	East Asia	0.0216	0.0073	0.0053	0.0241
HGDP00819	Han	East Asia	0.0270	0.0098	0.0082	0.0291
HGDP00930	Yoruba	Africa	0.0235	0.0058	0.0039	0.0389
HGDP01032	San	Africa	0.0922	0.0390	0.0153	0.1241
HGDP01043	Pima	America	0.0065	0.0009	0.0027	0.0275
HGDP01067	Sardinian	Europe	0.0167	0.0065	0.0045	0.0206
HGDP01081	Mbuti	Africa	0.0376	0.0131	0.0070	0.0795

AGR led to huge improvement in African genomes, moderate improvement in Eurasian genomes, but slightly higher error rate in American and Oceanian genomes. Perhaps the highly diverse African haplotypes include some "false neighbours" of the haplotypes found in America and Oceania, where historical bottlenecks have caused strong genetic drift in comparison to other regions.

Despite the improvement in accuracy, a drawback of using a reference panel is that variants not present in the reference will be ignored in the current implementation of both SHAPEIT2 and Eagle2. In practice, out of over 3 million unphased variant records on chromosome 9 only 1.45 million and 1.50 million remained after phasing against 1KG and AGR as reference panel, respectively. Losing more than half of the variants is certainly a serious drawback for a dataset highlighting genetic diversity.

I first attempted to avoid excluding missing variants in the reference panel by merging the AGR and the HGDP genomes and performing cohort-based phasing on merged dataset. The algorithm will ignore the phasing information but still have access to unphased genotypes in AGR. However, the result turned out even worse than cohort-based phasing without a reference panel (last column in Table 2.2). The existing phasing in the AGR panel is more accurate than what is possible with cohort-phasing alone, possibly achieved via splitting the dataset into population groups prior to phasing and making use of family information. Therefore scaffold-based phasing (Section 2.4.5) becomes the preferred method to utilise a reference panel while retaining all genomic variants.

2.4.3 Read-aware phasing

A single sequencing read sometimes spans over adjacent variants. SHAPEIT2 can also incorporate such information to increase the probability of assigning the variants onto the same haplotype according to the quality of the read [131]. I tested whether including reads information can improve phasing accuracy when using a reference panel.

The in-house tool from the developers of SHAPEIT2 requires BAM files to extract phasing informative reads (PIRs). The HGDP mapped reads, however, were stored in CRAM format. Since the limit in storage space made it impractical to convert the CRAM files to BAM, I extracted PIRs from VCF files instead by looking up the phasing group information generated by the GATK pipeline [132]. In this way, some PIRs from BAM/CRAM files filtered out during the variant calling pipeline will be excluded; meanwhile, since every phasing group is represented once in the PIR records, read depths information is no longer available to assign different weights to the PIRs.

When running on a whole chromosome in read-aware mode, SHAPEIT2 always threw an exception for unclear reasons. I was able to circumvent the problem by dividing the chromosome into shorter segments to be processed. Chromosome 22 was divided into 283 segments each containing 600 PIR records, with 100 overlapping variant sites between adjacent segments; phased results of all segments were subsequently ligated together. Table 2.3 compares the switch error rates on chromosome 22 with and without reads information, both using AGR as the reference panel.

Table 2.3 Switch error rates in SHAPEIT2 read-aware phasing

Sample	**Population**	**Region**	**Switch error rate**	
			without reads info	with reads info
HGDP00450	Mbuti	Africa	0.0162	0.0245
HGDP00460	Biaka	Africa	0.0189	0.0349
HGDP00547	PapuanSepik	Oceania	0.0167	0.0275
HGDP00551	PapuanHighlands	Oceania	0.0154	0.0276
HGDP00580	Druze	Middle East	0.0196	0.0202
HGDP00670	Sardinian	Europe	0.0105	0.0125
HGDP00774	Han	East Asia	0.0124	0.0190
HGDP00819	Han	East Asia	0.0173	0.0251
HGDP00930	Yoruba	Africa	0.0129	0.0293
HGDP01032	San	Africa	0.0309	0.0423
HGDP01043	Pima	America	0.0091	0.0163
HGDP01067	Sardinian	Europe	0.0134	0.0166
HGDP01081	Mbuti	Africa	0.0178	0.0229

Curiously, including the reads information increased the error rate as measured by comparison to the 10x genomes. The limitation could be due to the loss of read quality and read depth information in the VCF files. I also checked if the genomes phased without read-based information are consistent with the PIR from phase group records, and the rate of inconsistency is 1-3%. There is also approximately 1% of disagreement between physically-phased 10x genomes, and the physical phasing information in the VCF files generated in the GATK pipeline, possibly caused by errors in both processes. Read-level information appears unable to further improve the current phasing performance.

2.4.4 Number of PBWT iterations in Eagle

By default, Eagle2 automatically chooses the number of PBWT iterations by the relative size of the target and reference panel [128]. In our case where the number of target genomes to be phased (929) is less than half of the reference (4,957), only one iteration will be performed, but more iterations might improve phasing accuracy. In subsequent iterations, the reference panel will be expanded to include the inferred target haplotypes from the previous iteration. Table 2.4 compares the error rates on chromosome 21 after one to four PBWT iterations.

Table 2.4 Switch error rates after various numbers of PBWT iterations in Eagle2

Sample	**Population**	**Region**	**Switch error rate**			
			1 iteration	2 iterations	3 iterations	4 iterations
HGDP00450	Mbuti	Africa	0.01148	0.00808	0.00755	0.00705
HGDP00460	Biaka	Africa	0.00977	0.00469	0.00420	0.00446
HGDP00547	PapuanSepik	Oceania	0.01623	0.01045	0.00976	0.00907
HGDP00551	PapuanHighlands	Oceania	0.02338	0.01104	0.00931	0.00919
HGDP00580	Druze	Middle East	0.00738	0.00461	0.00457	0.00419
HGDP00670	Sardinian	Europe	0.00501	0.00430	0.00404	0.00399
HGDP00774	Han	East Asia	0.00694	0.00610	0.00614	0.00605
HGDP00819	Han	East Asia	0.00827	0.00645	0.00616	0.00664
HGDP00930	Yoruba	Africa	0.00368	0.00350	0.00357	0.00368
HGDP01032	San	Africa	0.02053	0.01594	0.01549	0.01523
HGDP01043	Pima	America	0.00707	0.00372	0.00297	0.00215
HGDP01067	Sardinian	Europe	0.00638	0.00500	0.00470	0.00490
HGDP01081	Mbuti	Africa	0.01008	0.00684	0.00562	0.00558

The second round of iteration reduced error rates in all samples, sometimes by over a half. After the third iteration, the accuracy still improved in most cases but decreased slightly (by less than 2%) in HGDP00930 (Yoruba) and HGDP00774 (Han). A fourth iteration reduced over 10% of error only in HGDP01043 (Pima), but increased error rates in four

individuals. Perhaps the increase in error rate results from incorporating incorrect haplotypes from the previous iteration into the reference panel, which indicates that the algorithm is unable to extract any new information. Such saturation of information seems to happen faster in populations whose genetic diversity is better represented in the reference and target panels combined, either because the sample size is sufficiently large (Han and Yoruba), or because the level of variation is low (Pima, due to recent inbreeding). I used three PBWT iterations in later runs as it appears sufficient for most populations.

2.4.5 Scaffold-based phasing

To exploit the advantage of a reference panel and at the same time phase all the variants, I eventually turned to a scaffold-based method implemented as the genotype calling mode in SHAPEIT2 [133]. This mode is originally intended for genotype calling from low-coverage sequences when a reliable scaffold from microarray data is available. Here the phasing result with a reference panel can be used as the scaffold, with variants absent from the reference panel added later by SHAPEIT2 according to local linkage patterns.

I made minor changes to the provided pipeline to allow reading in the PL field in place of GL, treating chromosome names as strings, and recognising missing genotypes in different formats. The scaffold was obtained by phasing with AGR as the reference panel in Eagle2 after three PBWT iterations. Subsequently, the chromosomes were divided into windows each spanning 2,400 SNPs, with 200 overlapping SNPs between adjacent windows. Beagle (v4.1) [134] was used to estimate genotype likelihoods in each window. Finally, the variants were mapped onto the scaffold in SHAPEIT2's call mode, based on the posterior genotype likelihoods from Beagle. To avoid under- and overflow errors (which is also reported in [39]), I also ran SHAPEIT2 in the same windows as Beagle. On rare occasions when SHAPEIT2 still exited with error, the windows were enlarged on both ends by 500 SNPs at one time until the run finishes successfully. Beagle also imputes the missing genotypes, but I filtered them out in the end for consistency with the unphased dataset. For the same reason, Beagle was run in the gtgl mode rather than gl mode in the original pipeline to treat non-missing genotypes as fixed.

Table 2.5 lists the switch error rates in scaffold-based phasing measured against all 26 10x-sequenced genomes on chromosome 1. Since the phasing of singletons in the dataset is largely random (unless it happens to be more frequent in the reference panel), the error rates measured without singletons are also reported.

Table 2.5 Switch error rates using scaffold-based phasing

Sample	**Population**	**Region**	**Switch error rate**	
			with singletons	w/o singletons
HGDP00930	Yoruba	Africa	0.0092	0.0033
HGDP00931	Yoruba	Africa	0.0097	0.0027
HGDP00460	Biaka	Africa	0.0140	0.0044
HGDP00450	Mbuti	Africa	0.0144	0.0051
HGDP01081	Mbuti	Africa	0.0155	0.0052
HGDP00472	Biaka	Africa	0.0157	0.0064
HGDP01032	San	Africa	0.0374	0.0122
HGDP01029	San	Africa	0.0378	0.0131
HGDP01019	Karitiana	America	0.0066	0.0052
HGDP01043	Pima	America	0.0076	0.0039
HGDP01056	Pima	America	0.0081	0.0052
HGDP01013	Karitiana	America	0.0086	0.0047
HGDP00228	Pathan	Central South Asia	0.0124	0.0040
HGDP00224	Pathan	Central South Asia	0.013	0.0043
HGDP00946	Yakut	East Asia	0.0061	0.0033
HGDP00954	Yakut	East Asia	0.0088	0.0046
HGDP00774	Han	East Asia	0.0156	0.0049
HGDP00819	Han	East Asia	0.0167	0.0054
HGDP00670	Sardinian	Europe	0.0107	0.0033
HGDP01067	Sardinian	Europe	0.0115	0.0038
HGDP00562	Druze	Middle East	0.0067	0.0031
HGDP00580	Druze	Middle East	0.0125	0.0038
HGDP00549	PapuanHighlands	Oceania	0.0237	0.0117
HGDP00547	PapuanSepik	Oceania	0.0278	0.0115
HGDP00542	PapuanSepik	Oceania	0.0281	0.0119
HGDP00551	PapuanHighlands	Oceania	0.0281	0.0120

Although the San population still has the highest error rate, it is reduced to around 1.3% after excluding singletons. Phasing is also less accurate in the Papuan populations at an error rate around 1.2%, which roughly corresponds to on average one switch error per 160 kB if singletons are ignored. At this level, the error rate should not have a detectable impact on downstream analyses, such as identity-by-descent analysis and detection of archaic segments.

2.5 Conclusion

The human genetics community should be encouraged to update existing resources and release new sequences on the latest GRCh38 reference assembly, which better captures the genetic heterogeneity between human populations by providing alternative genomic loci. Although the discussion on phasing performance is largely technical, variations in error rates also reflects patterns of genetic diversity and population relationship: the reduction of error rates in Eurasia when African genomes were included as reference is consistent with the high diversity in the latter; the difficulty in phasing Oceanian genomes is likely due to their early divergence from Eurasia and not being represented in the 1000 Genomes Project. More comprehensive sampling or family-based studies in San and Papuan populations would be helpful to improve the accuracy in phasing and imputation.

Chapter 3

Diversity and structure of HGDP populations

3.1 Principal component analysis

The pattern of genetic diversity in human population is shaped by past demographic changes. Apart from historic and anthropological interest, understanding its global variation also lays the basis for studying evolutionary dynamics, identifying genetic associations, or in the scope of this thesis, comparing with the diversity in archaic segments (chapter 5).

Principal component analysis (PCA) was performed in the R package SNPRelate [135] using 929 HGDP genomes. Variants were pruned to exclude sites with a minor allele frequency lower than 0.05 or missing rate greater than 0.1, and those in linkage disequilibrium above 0.5 with others in a sliding window of maximum 500 kB long. Figure 3.1 shows the result grouped by geographical regions, and Figure 3.2 shows the break-down of populations in each region.

The first PC separates sub-Saharan African and non-African populations, whilst the second PC separates populations by their affinity to west Eurasians. Similar patterns have been reported in previous studies on worldwide human genetic structure [16, 39]. The third PC separates American and East Asian genomes, which occupy similar spaces in the first two PCs. On the third PC, all non-American populations also cluster with East Asian samples, highlighting pronounced genetic drift in America possibly from a strong founder effect and recent inbreeding.

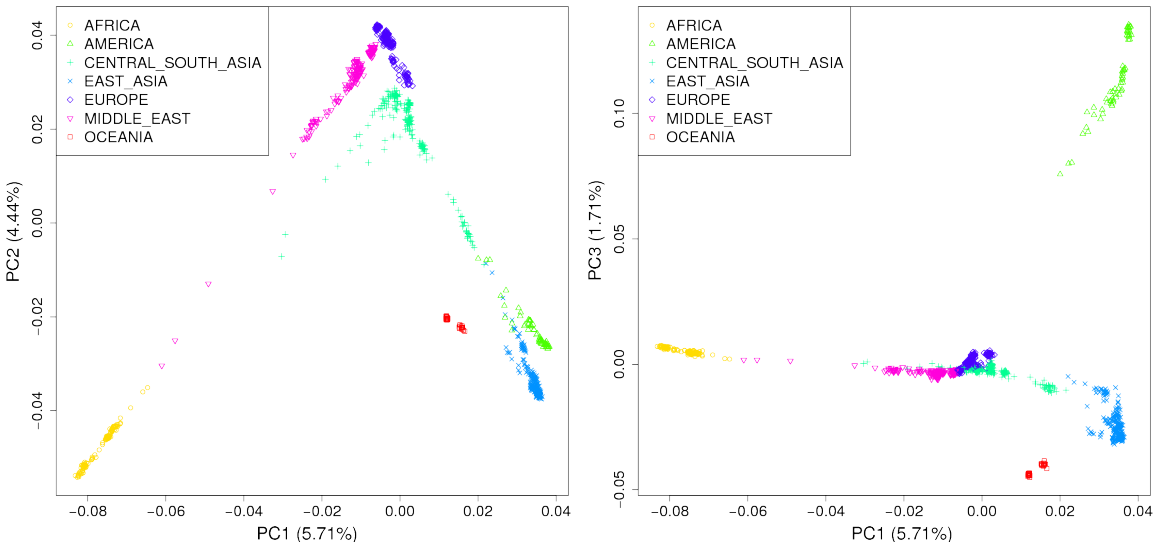


Fig. 3.1 Principal component analysis on HGDP genomes

Within Africa, there is a continuous cline between societies traditionally following a hunter-gatherer lifestyle (Biaka, San and Mbuti) and those following an agrarian lifestyle (Bantu, Mandenka and Yoruba) (Figure 3.2a), in accordance with the historical expansion of Bantu-speakers replacing local hunter-gatherer groups [136–138]. Some individuals from the Middle East (in particular the Mozabite population) lie between sub-Saharan Africans and the other Middle Eastern individuals (Figure 3.2g), possibly due to recent gene flows at various levels from Africa. Within Europe, there is a division between western (Sardinia, Bergamon and Basque) and eastern (Russian and Adygei) Europeans (Figure 3.2c). In particular, the Sardinians form a well-defined cluster at the end of western Europe. Previous studies have detected in them the highest genetic affinity to Neolithic farmers in Europe, which suggests that they were unaffected by the Bronze age migration that introduced steppe ancestry to Europe [139, 101, 140]. Central and South Asia exhibit connections with the Middle East, European and East Asian populations (Figure 3.2d). The European and Middle Eastern ancestry might reflect both historical and more recent admixtures, as the amount varies considerably within populations such as Makrani, Sindhi and Brahui. Uygur and Hazara show the highest genetic affinity to East Asians, bordering Yakut from East Asia and some Maya individuals from America (Figure 3.1 and 3.2d). The genetic diversity is relatively limited in Chinese populations, apart from some Xibo and Tu individuals that show affinity to Central/South Asian populations (Figure 3.2b). The Cambodian population, on the other hand, appear drifted slightly towards the Papuans. Within America, the Maya

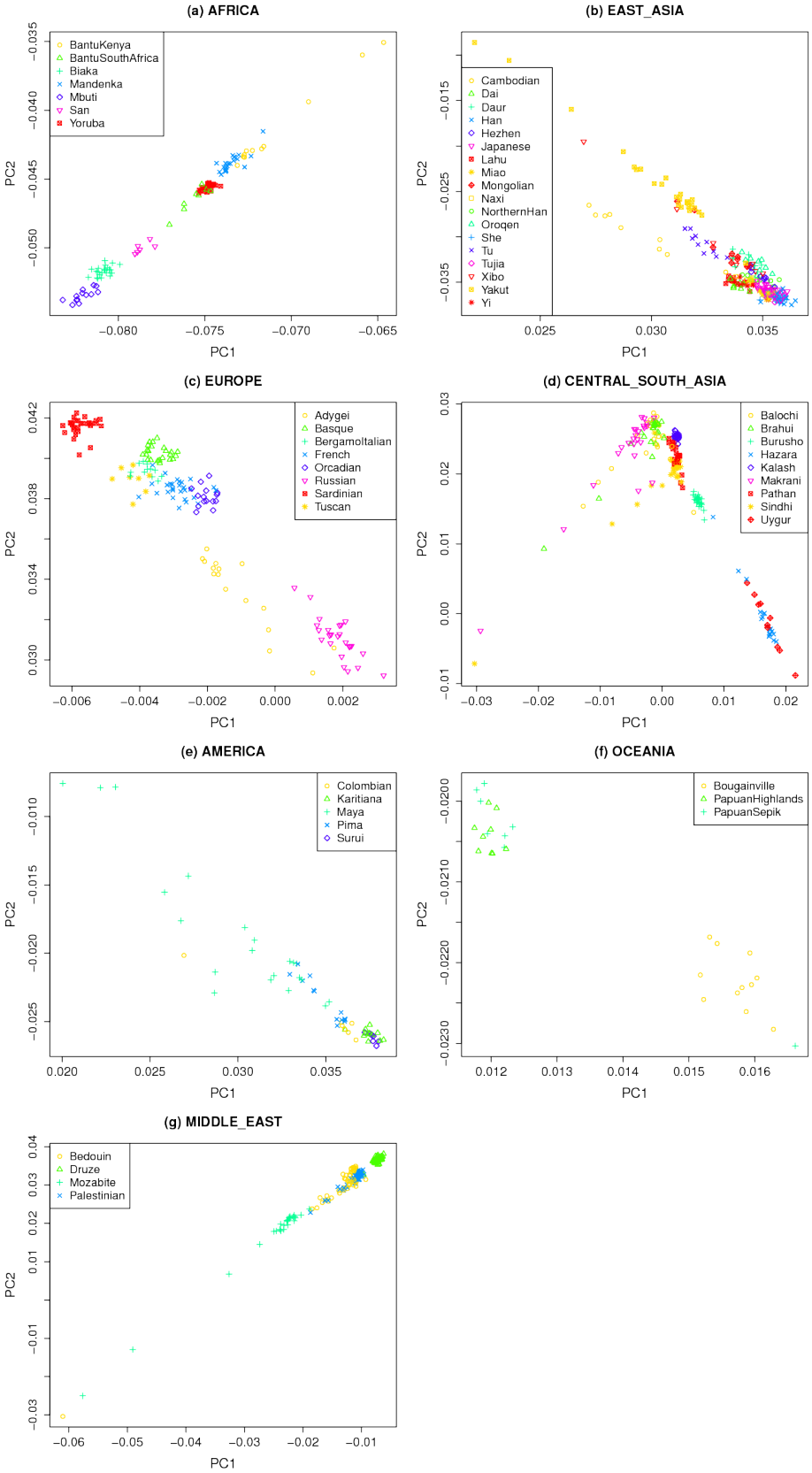


Fig. 3.2 Principal component analysis on HGDP genomes by region

population appears the most diverse, possibly due to recent European admixture (Figure 3.2e). In Oceania, there is a clear division between the population from Bougainville Island and the other two from New Guinea (except for one Papuan Sepik individual that clusters with the Bougainville population), consistent with historical migration from South East Asia into Bougainville (Figure 3.2f).

3.2 Heterozygosity and runs of homozygosity

Runs of homozygosity (RoH) are regions of the genome where long tracks of identical haplotypes have been inherited from both parents, when they share one or more ancestors in the recent past. It is a special case of identity-by-descent (IBD) within the same genome. The amount of RoH tracks therefore reflects the level of recent inbreeding in the population. In contrast, heterozygosity within the genome reflects the average time to coalescence between the two parental lineages, thus affected by the effective population size through a longer time span.

The RoH tracks were identified using the ROH extension in BCFtools, which implements a hidden Markov model to detect long homozygous stretches that are not likely to have arisen by chance given the allele frequency and local recombination rate [141]. Many populations in the dataset do not reach a sample of 20 individuals, the recommended minimum for estimating the allele frequencies; therefore I provided the allele frequencies not by population, but by geographical regions.

Figure 3.3 compares the total length of RoH tracks with the total number of heterozygous sites. The total lengths of RoH per individual are distributed similarly to the previous result using SNP array data [142]. African genomes are distinguishable by the highest level of heterozygosity and shortest lengths of RoH tracks; heterozygosity gradually decreases as the distance to Africa increases, with the lowest values found in some American genomes. The wide variations in America, Central/South Asia and the Middle East is consistent with a recent history of admixture. There is a strong linear negative correlation between the two statistics, yet the total length of RoH tracks in African (especially in traditionally hunter-gatherer groups of Mbuti and San) and American (especially the Amazonian populations of Surui and Karitiana) genomes is elevated in relation to their heterozygosity levels, suggesting a high degree of inbreeding in recent generations possibly caused by population decline.

The lengths of individual RoH tracks reflect the age of the shared ancestor(s), as recombination breaks down longer tracks through the generations. Figure 3.4 shows the distribution of

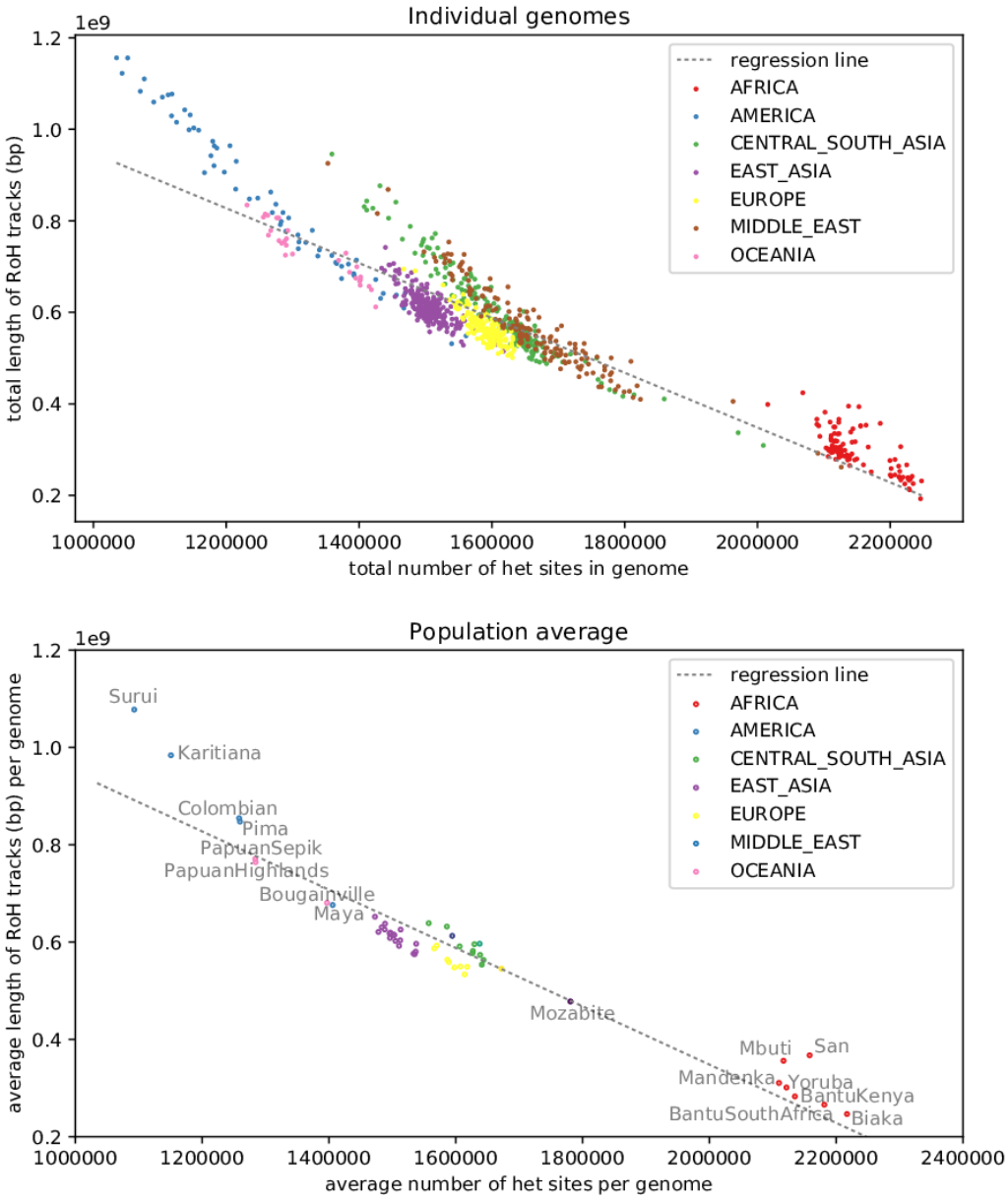


Fig. 3.3 Comparison between the total length of RoH tracks and heterozygosity in HGDP genomes, showing the elevated RoH lengths in relation to heterozygosity in Africa and America

individual RoH track lengths across geographical regions. The tracks are shortest in African genomes and longest in Oceanian and American genomes. The population decline appears more recent in the Papuan and American populations, but further back in history in the sub-Saharan African populations.

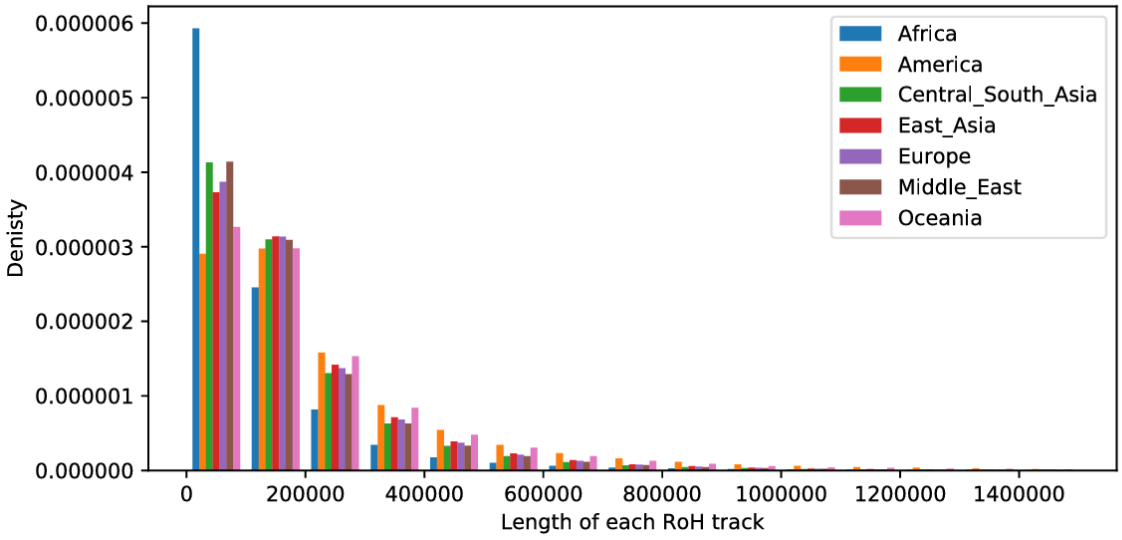


Fig. 3.4 Distributions of the lengths of individual RoH tracks by geographical regions

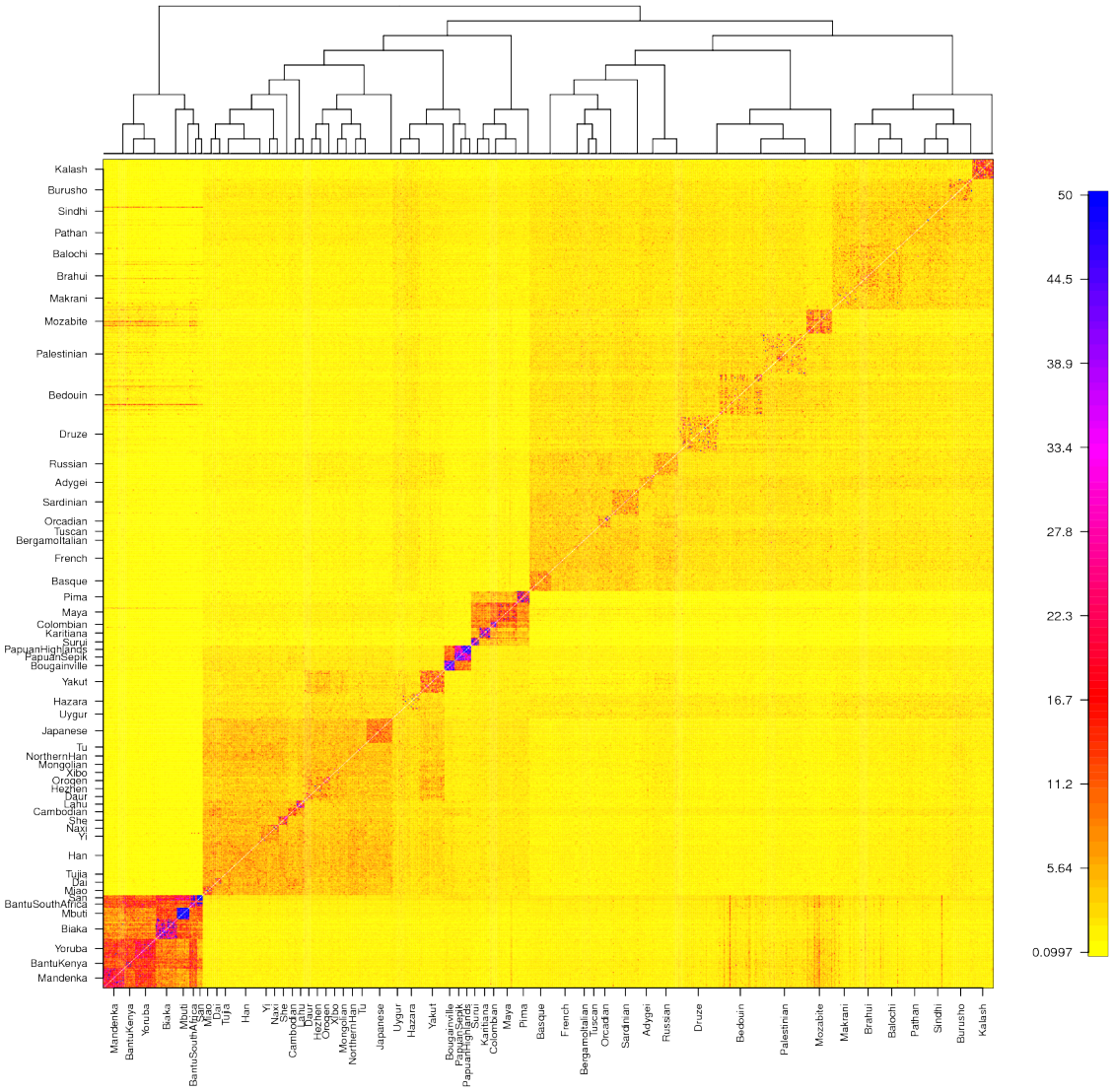
3.3 Haplotype structure

Whole genome sequences allow us to investigate genetic structure in more detail than what is allowed by sparse markers. By capturing all positions in a segment, it becomes possible to compare very similar haplotypes and model the local genealogies. Here I present the result from two methods that model shared haplotypes to infer relatedness using statistically phased genomes (section 2.4).

3.3.1 Coancestry matrix

The fineSTRUCTURE pipeline aims to combine non-parametric and model-based approaches to describe genetic structure[35]. In the first step, ChromoPainter identifies shared ancestry between pairs of genomes by searching for the nearest neighbours of a haplotype. Each local stretch of the haplotype in an individual is modelled using all other individuals as potential donors; hence the entire genome becomes a mosaic of chunks from other genomes. The

number of chunks received from another individual then serves as an asymmetric measure of relatedness in the coancestry matrix, which can be used as a basis for clustering and demographic modeling.



(Rows: recipients; columns: donors. The value is capped off at 50)

Fig. 3.5 Heat map showing the coancestry matrix (number of shared segments) on chromosome 22 between 929 genomes produced by ChromoPainter, along with the tree inferred from the matrix by FineSTRUCTURE

Figure 3.5 shows the coancestry matrix including all 929 genomes, with the population tree reconstructed from it on top. The pipeline was run in linked mode, which models linkage

between sites based on a genetic map. Due to the high computational cost, the analysis was not performed on other chromosomes. Genomes from the same population were enforced to form a clade in the clustering step. Most genomes from the same populations form distinct clusters, although the boundaries with neighbouring populations appear weak in Han, French, and some South Asian populations. Ancestry make-up also varies within populations, perhaps most notably in the Middle East, where some individuals show signs of recent admixture with sub-Saharan Africans but others do not. Since the ChromoPainter model distinguishes between donors (columns in the coancestry matrix) and recipients (rows in the coancestry matrix), asymmetries in the heat map could suggest directional gene flows. For example, the San population appears to have received genetic contribution from west African populations (Mandenka, Yoruba, and BantuKenya that was connected to the Bantu expansion), which is supported in previous studies [143, 137]; similarly, the Maya and Colombian populations appear to have received contribution from Amazonian groups (Surui and Karitiana), however, it could also be explained by the affinity of the latter to one founding lineage in the Americas [144, 145]. The structure and clustering pattern mostly agrees with previous results using 640k SNP markers [35].

It is worth noting that since ChromoPainter assigns a donor for every segment in the genome, not all shared chunks are equally similar to each other. The sampling scheme would also influence the coancestry matrix, although this should be less a problem as the HGDP collection covers most regions with multiple populations. The darker square in the lower left corner of Figure 3.5 does not suggest that two randomly drawn genomes from Africa are more similar than two drawn from elsewhere, only that an African genome is more similar to another random one from Africa than from other regions in the world. In other words, the tree generated from this coancestry matrix is a cladogram whose branch lengths are not meaningful. The quantitative relatedness within and between populations will be discussed in the next section.

3.3.2 Identical-by-descent segments

Segments from different genomes are termed identical-by-descent (IBD) if they exhibit identical haplotypes inherited from a common ancestor. The amount of IBD segments is a measurement of relatedness between individuals; similar to RoH tracks, the lengths of IBD segments also reflect how far back in time the common ancestor lived.

I used Refined IBD [146] to detect IBD segments in all HGDP genomes. In the first step, the algorithm searches for identical long haplotypes; the refinement step then evaluates

the likelihood under IBD and non-IBD models for each candidate segment. The average genetic lengths of IBD segments shared between genomes from pairs of populations are shown in Figure 3.6. Most IBD sharing happens within the same population, nevertheless Makrani, Brahui and Balochi populations in Pakistan also share substantial amount (> 80cM) of IBD segments between each other, suggesting recent migration in this region. The level of IBD sharing between Oroqen and Daur in China is also almost the same as that within each population. Given the proximity in their territory range, intermarriage might have been common in recent time.

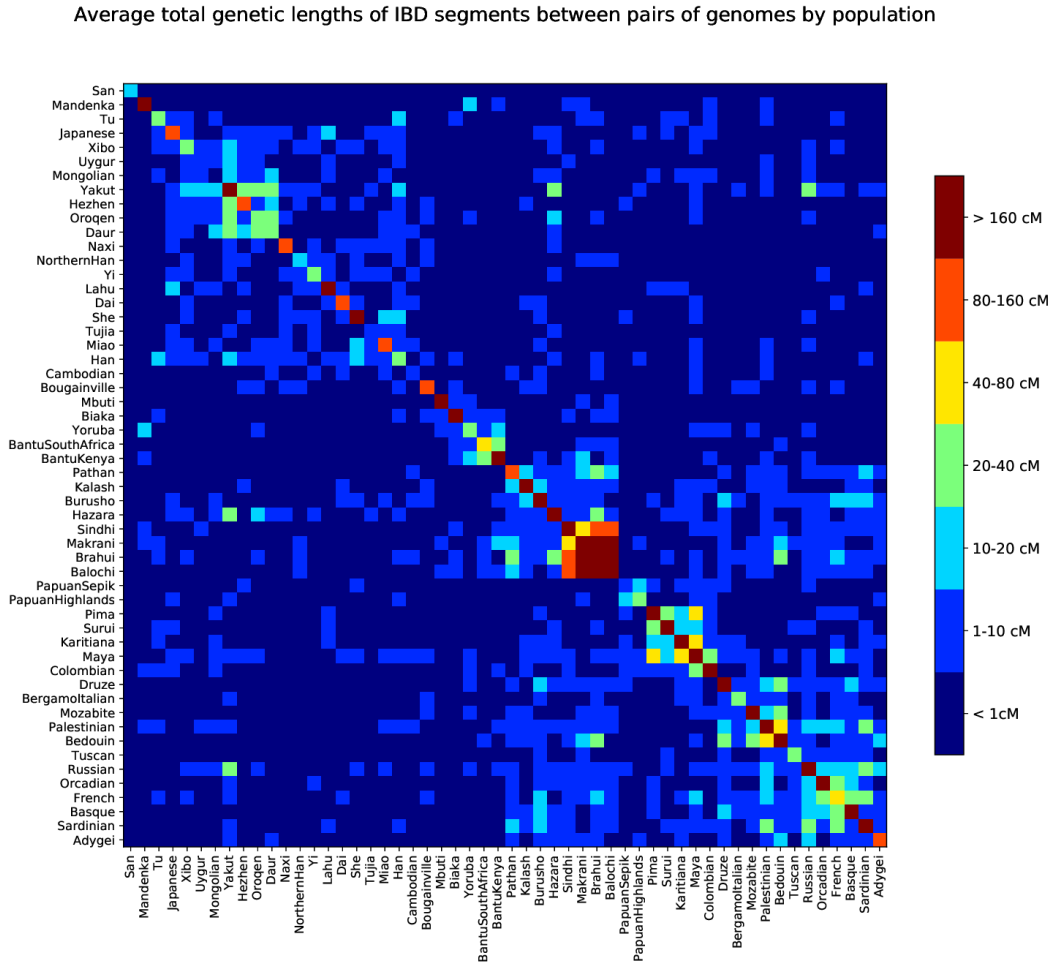


Fig. 3.6 Heat map showing the average lengths of IBD segments in two genomes from the same (diagonal positions) and different (non-diagonal positions) populations

Refined IBD also detects homozygous-by-descent (HBD) segments, which are analogous to RoH tracks but emphasizes the shared recent ancestor as the origin of homozygosity. In contrast to RoH tracks identified by BCFtools, HBD segments are required to be completely

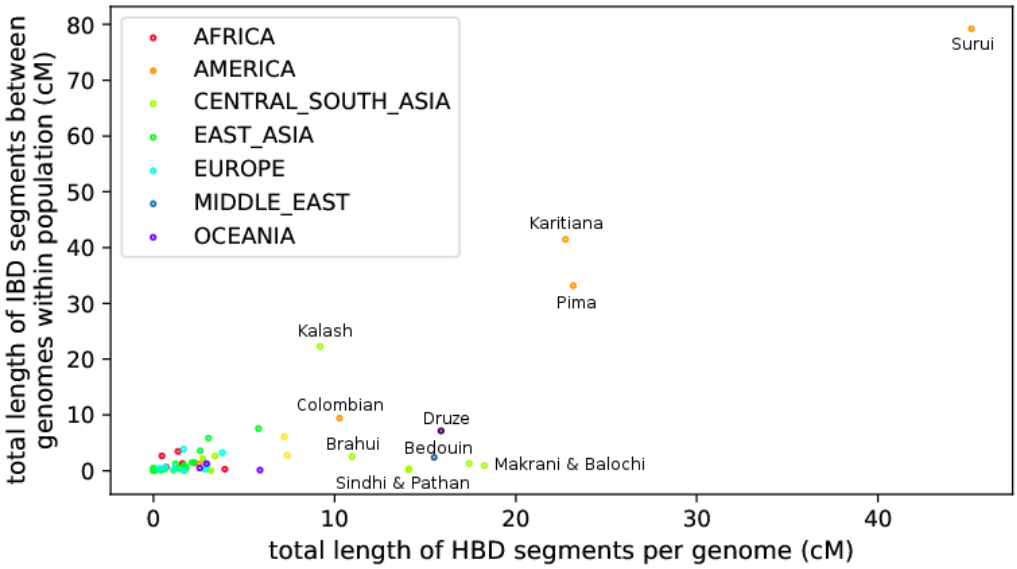


Fig. 3.7 Comparison between the total lengths of HBD segments per genome and the total lengths of IBD segments between genomes in each population

identical. Figure 3.7 compares the average length of HBD tracks per genome with the average length of IBD between genomes by populations. There is a strong positive correlation ($R^2 = 0.6723$), however some non-African populations contain relatively long HBD in the genomes in spite of the low amount of IBD segments in the population around the same level as in Africa. Since HBD primarily reflects the relationship between the parents in contrast to random individuals in the population that IBD measures, it suggests the presence of consanguinity where the parents are more closely related than random individuals from the population. The pattern is particularly notable in some populations from Central/South Asia, where marriage within the clan, tribe or caste is common [147]. Populations that deviate the most from the inbreeding trend - the Kalash population and indigenous American populations - all have long been isolated and underwent a recent decline in size. They might live in relatively homogeneous groups whose small size does not allow substantial population structures to develop.

3.4 Conclusion

To summarise, both model-free and model-based methods have been used to reconstruct the relationship between genomes from the HGDP panel. Although the major patterns of genetic diversity and structure have been described in previous studies using microsatellite

and SNP arrays [30, 32, 35], analysis using whole genome sequences adds more clarity and details, especially concerning recent inbreeding, migration and population size changes in some populations.

This part also demonstrates that population genetics methods can scale up to handle thousands of high-coverage genomes to meet the rapid development in the field. Many algorithms originally developed for sparse genetic markers or for small numbers of populations become cumbersome in the era of large-scale whole genome sequencing, but the detection of RoH and IBD segments remains efficient. More than a measurement of recent inbreeding and population size, these segments also have the potential to reveal recent demographic history (e.g. [142] and [148]), complementing methods based on the site frequency spectrum which are also scalable but have limited resolution for the recent past. More recently, new methods have also been developed to efficiently reconstruct local coalescent trees throughout the genome across thousands of samples or more [149, 150], which opens up the possibility to study adaptative and demographic histories at unprecedented resolution.

Chapter 4

Hidden Markov model for tagging archaic segments

4.1 Model motivation

Much effort has been devoted to identifying introgressed segments from archaic hominins in modern human genomes. Introgressed segments are expected to coalesce first with archaic lineages, therefore show higher affinity to the archaic haplotypes than to modern haplotypes that do not pass through the archaic populations (Figure 4.1). Sub-Saharan African haplotypes are usually used for comparison, as no substantial archaic ancestry from the Neanderthals or the Denisovans has been found there (although a recent study reports unidentified archaic ancestry in the Yoruba population [94]). The lengths of introgressed segments should also follow an exponential decay consistent with the admixture time. But in practice, detecting archaic haplotypes locally is complicated by the variation in branch lengths: the sequenced archaic genomes may not be a close reference to the source population that contributed to the gene flow, which increases the distance between an archaic haplotype in modern genome and the corresponding haplotype in the archaic reference genome; some non-African modern haplotypes also show deep coalescence with African haplotypes, considering the large effective size of the ancestral human population.

In the conditional random field (CRF) developed by Sankararaman *et al.*, two informative features of archaic ancestry - sharing a derived allele seen in the archaic genome but absent from the African genomes, and high divergence of the haplotype to African ones but low to the archaic ones - reflect the expected genealogy of the archaic segments, although the

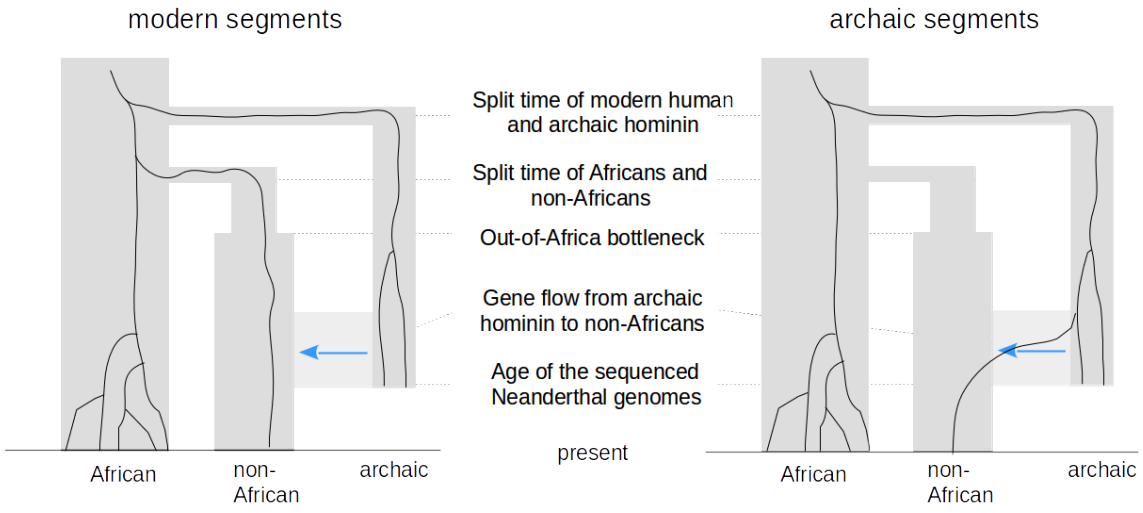


Fig. 4.1 Examples of the genealogies of modern and archaic segments on top of population demographic history

second feature was omitted when searching for Denisovan haplotypes due to a bias when the archaic ancestry is at the order of 1/1000; the third feature, controlled through the transition functions, fits the expected haplotype lengths following the estimated time of introgression [1, 76]. In comparison, Vernot *et al.* used a two-stage method based on the S^* statistics [61]: after excluding SNPs present in an African reference panel, the first stage identifies haplotypes containing SNPs in linkage disequilibrium over extended length, where the cutoff for S^* is determined by a generalized linear model trained on simulations to account for variations in local recombination and mutation rate; then the putative archaic segments are filtered by their affinity to the archaic genomes [90, 91]. More recent methods show that it is possible to identify introgressed segments from summary statistics that are informative about the underlying genealogy, even without any genetic information about the source population [95, 94].

With the quantity and diversity of the samples, the HGDP dataset provides an excellent opportunity to study the global variation of archaic segments. At the time when I started looking into archaic segments in the HGDP dataset, none of the implementations of the published methods to detect archaic introgression is publicly available; moreover, The focus here is to obtain an accurate collection of such segments introgressed from archaic sources whose genomes have been sequenced. It is also desirable to minimize differences in detection accuracy caused by different sample sizes and population histories. I therefore developed a

hidden Markov model (HMM) based on allelic sharing pattern between modern non-African genomes, African genomes, and the archaic genomes. The following sections describe the model in detail.

4.1.1 Hidden Markov models

An HMM consists of two layers: the invisible underlying states x are assumed to form a first-order Markov chain, where the probability of entering various states in the next time point depends only on the current state; events in the observed layer y are controlled by the hidden states through the emission probabilities. Additionally, a set of initial state probability describes the state at the first point. The emission probabilities, transition probabilities, and initial state probabilities specify an HMM. The aim is usually to decode the hidden states or to infer parameters of interest. Its inherent sequential structure, computational efficiency, as well as the ease to interpret has made HMM hugely popular in biological sequence analysis (see [151] for a detailed introduction and review).

4.1.2 Model setup

Our HMM runs on high-coverage haploid genomes (requiring phased data). It contains two hidden states, modern (0) and archaic (1). The observations are summarised from genotypes, whilst an optional recombination map can be used to adjust the transition matrix along the chromosome. I explored various implementations of this approach, discussed later in this chapter; the current section describes some features common to all of them.

Emission

We summarise the observed data by the pattern of allele sharing between sub-Saharan African, non-African and archaic genomes (Table 4.1), and HMM emission functions are defined as distributions involving these patterns. If a genetic segment entered the modern human population from an archaic hominin population relatively recently, it should share more derived variants with the archaic genomes. Since sub-Saharan Africa is generally assumed to have no or negligible archaic ancestry, at least not in common with non-Africans, an allele shared between African and non-African genomes would most likely have arisen on a lineage within the modern human population. Incomplete lineage sorting in modern segments might cause some African lineages to coalesce first with the archaic genomes, thus sharing a derived

Table 4.1 Informative allele-sharing patterns

Genotype*			Likely underlying state
African panel	non-African sample	Archaic	
1	0	1	modern
1	1	0	modern
0	1	1	archaic

* 0: ancestral allele in all genomes; 1: derived allele in at least one genome

allele unseen in a non-African genome. Such observation is less likely to occur in the archaic segments due to the small effective size of the archaic populations.

For convenience, the allele-sharing patterns are number-coded into emission types. In the emission matrix E, E_{ik} denotes the probability of emitting type $k(k \in \{1,2,3,\dots\})$ in state $i(i \in \{0,1\})$.

Transition

State changes may occur at points of recombination, which follows a Poisson process. Two modes are implemented in the model: in the absence of a genetic map, transition probabilities are read from user input and remain unchanged throughout the sequence; when a genetic map is provided, transition probabilities are calculated on-the-fly according to local recombination rate. In the latter case, the probability of entering the other state conditioned on the occurance of a recombination event also depends on the overall proportion of archaic ancestry (α). Therefore we have the following transition matrix:

$$T = \begin{bmatrix} 1-(1-e^{-dt})\cdot\alpha & (1-e^{-dt})\cdot\alpha \\ (1-e^{-dt})(1-\alpha) & 1-(1-e^{-dt})(1-\alpha) \end{bmatrix} \tag{4.1}$$

where $T_{ij}(i,j \in \{0,1\})$ is the probability to transit from state i to state j, d the genetic distance between adjacent sites, t the time since admixture in generations, and α the admixture proportion. α also defines the initial state probability $\pi = (1-\alpha, \alpha)$.

4.1.3 Viterbi decoding

The Viterbi algorithm is used to obtain the most likely sequence of hidden states in an HMM. The probability of the most likely sequence until position p can be computed recursively:

$$V_{1,i} = E_{iy_1} \cdot \pi_i$$

$$V_{p,i} = \max_{k \in \{0,1\}} (E_{iy_p} \cdot T_{ky_p} \cdot V_{p-1,k})$$

where i is the hidden state ($i \in \{0,1\}$). The value of the previous hidden state k that maximizes $V_{p,i}$ should be saved, so that the most likely sequence (Viterbi sequence) can be retrieved after reaching the end of the sequence.

4.1.4 Model training

A number of methods to estimate the model parameters were used in variations of the HMM (Sections 4.2.3 and 4.3). These training methods are briefly discussed below.

Maximum likelihood estimation from simulations

It is impossible to know the true underlying state of real-world genome sequences, but simulations can provide tagged data for model training. I generated coalescent simulations in msprime [152] with an African population, a non-African population, and one or more archaic populations that admixed with ancestors of modern humans. The background demographic model in Figure 4.1 is an example with only one archaic admixture event. The simulator keeps track of migration records, through which we can recover the true state of segments: if the lineage moves at any point from the non-African population to an archaic population in the local tree, the segment is tagged as archaic.

The maximum likelihood estimate for the emission matrix E is simply the probability of observing each type of emission within each state:

$$E_{ik} = \frac{\sum_{p=1}^{L} 1_{y_p=k, x_p=i}}{\sum_{p=1}^{L} 1_{x_p=i}}$$

where $\mathbf{1}$ is an indicator function:

$$\mathbf{1}_{x_p=i} = \begin{cases} 1 & \text{if } x_p = i \\ 0 & \text{otherwise} \end{cases}$$

Due to randomness in the coalescent process, the true proportion of archaic segments in the sequences usually differs from the admixture proportion specified in the demographic model. Adjusting the value of α according to the true underlying states could possibly improve the accuracy of inference. Since the true states are known, the initial state probabilities π are estimated from the actual proportion of archaic segments:

$$\pi_i = \frac{\sum_{p=1}^{L} 1_{x_p=i}}{L}$$

Although the transition probabilities can also be estimated from the simulated sequences, I calculated them from admixture time and and proportion ($\alpha = \pi_1$) following Equation 4.1, which is explicit about the recombination process.

Baum-Welch training

In the absence of training data, the Baum-Welch algorithm is a classical method to infer unknown parameters in an HMM from the observed sequences only. It makes use of the Expectation-Maximization (EM) algorithm. Starting from a set of parameters $\theta = (E, T, \pi)$, which can be chosen randomly or using prior information, the forward-backward algorithm is run once to obtain the probability of being in state i at position p, $\gamma_i(p)$. For a sequence of length L, let $\alpha_i(p) = P(Y_1 = y_1, \ldots, Y_p = y_p, X_p = i|\theta)$, the probability of being in state i at position p and observing all previous emissions including that at p, we have:

$$\alpha_i(1) = \pi_i \cdot E_{iy_1}$$

$$\alpha_i(p+1) = E_{iy_{p+1}} \sum_{j\in\{0,1\}} \alpha_j(p) \cdot T_{ji} \tag{4.2}$$

Similarly, let $\beta_i(p) = P(Y_{p+1} = y_{p+1}, \ldots, Y_L = y_l|X_p = i, \theta)$, the probability of observing all subsequent emissions excluding that at p, given the hidden state at p:

$$\beta_i(L) = 1$$

$$\beta_i(p) = \sum_{j\in\{0,1\}} \beta_j(p+1) \cdot T_{ij} \cdot E_{j,y_{p+1}} \tag{4.3}$$

Then the probability of being in hidden state i conditioned on all the emissions can be expressed as

$$\gamma_i(p) = P(X_p = i|Y, \theta) = \frac{\alpha_i(p)\beta_i(p)}{\sum_{j\in\{0,1\}} \alpha_j(p)\beta_j(p)}$$

and the probability of being in state i at position p and in state j at position $p+1$:

$$\xi_{ij}(p) = P(X_p = i, X_{p+1} = j|Y,\theta) = \frac{\alpha_i(p)T_{ij}\beta_j(p+1)E_{jy_{p+1}}}{\sum_{m\in\{0,1\}}\sum_{n\in\{0,1\}}\alpha_m(p)T_{mn}\beta_n(p+1)E_{ny_{p+1}}}$$

Subsequently, we can update the transition matrix and the emission matrix as follows:

$$T^*_{ij} = \frac{\sum_{p=1}^{L-1}\xi_{ij}(p)}{\sum_{p=1}^{L-1}\gamma_i(p)}$$

$$E^*_{ik} = \frac{\sum_{p=1}^{L}1_{y_p=k}\gamma_i(p)}{\sum_{p=1}^{L}\gamma_i(p)}$$

Note that the form of T does not preserve the constraints in Equation 4.1. An alternative approach is to apply Equation 4.1 and update the value of admixture time t instead of T with its maximum likelihood estimate. This possibility is not explored here.

In the original algorithm, the initial state probability π is also updated from the frequency spent in each state at time 1:

$$\pi^*_i = \gamma_i(1)$$

But regarding introgression, the first position of the chromosomes is indistinguishable from any random positions. Instead we update the probabilities with the expected time spent in each state:

$$\pi^*_i = \frac{\sum_{p=1}^{L}\gamma_i(p)}{\sum_{k\in\{0,1\}}\sum_{p=1}^{L}\gamma_k(p)}$$

The next iteration then starts with $\theta^* = (E^*, T^*, \pi^*)$, and repeats the steps above until convergence. The Baum-Welch algorithm is guaranteed to converge towards a local optimum $\theta^* = \arg\max_\theta P(Y|\theta)$.

Numerical optimisation of likelihood

The Baum-Welch algorithm as presented above assumes constant transition probabilities when updating the transition matrix T; it does not properly account for variations in recombination rate. Another training method is to maximize the likelihood through numerical optimisation. In this way, it is more convenient to perform constrained optimisation and to estimate parameters embedded in the expressions of E, T or π (such as admixture time t in the transition matrix T) from genomic sequences.

Using the forward algorithm, we have the likelihood of parameter θ as the sum of the partial probabilities at the last position of the sequence:

$$P(Y|\theta) = \sum_{i\in\{0,1\}} \alpha_i(L)$$

In practice, the log-likelihood function is used for numerical convenience. A limited-memory quasi-Newton algorithm that allows for bounded constraints, L-BFGS-B [153–155], is used for optimisation. Each run of optimisation finishes when the parameters converge. 1,000 independent runs are launched from randomly generated initial parameter sets, and the final parameter values along with the log-likelihood at the end of each run are recorded.

4.2 Site-wise vs informative-site-only models

4.2.1 Site-wise model

I first implemented an HMM that passes through each position in the genetic sequence. The emission types are encoded according to Table 4.2. Missing sites in real data (or masked ones in simulation) are coded as type 0, where the forward-backward equations (Equations 4.2 and 4.3 simplifies to

$$\alpha_i(p+1) = \sum_{j\in\{0,1\}} \alpha_j(p)\cdot T_{ji}$$

and

$$\beta_i(p) = \sum_{j\in\{0,1\}} \beta_j(p+1)\cdot T_{ij}$$

without the observation y_{p+1}. The 2×4 emission matrix E has 6 free parameters, along with 2 free parameters in the transition matrix if the genetic map is not used, and the amount of admixture that defines the initial state probabilities.

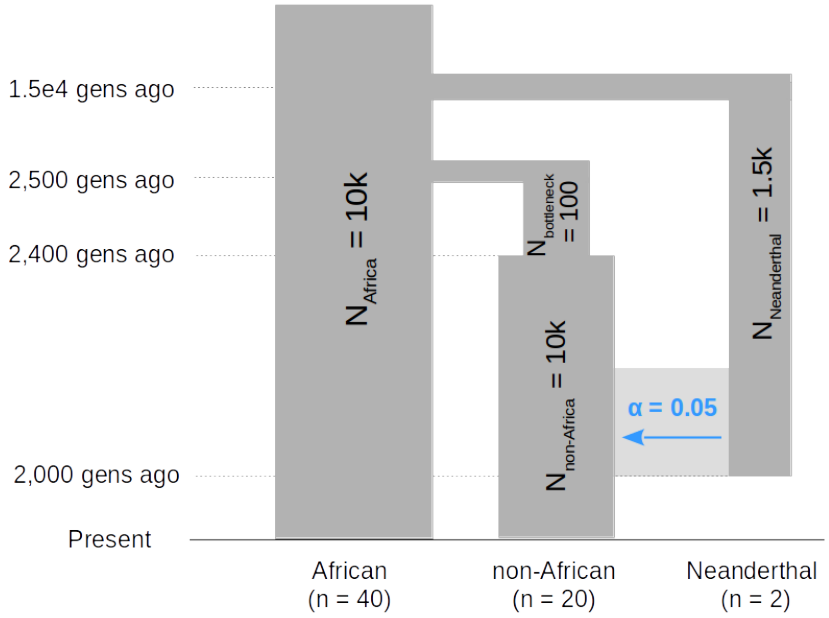
To estimate model parameters, we simulated 10 Mb regions in 40 African sequences (to match the pilot dataset), 20 non-African sequences, and 2 Neanderthal sequences. 2,000 generations ago, the non-African population in the simulation received gene flow from the Neanderthal population, represented by the lineages moving into the Neanderthal populations (backward in time) with a probability of 0.05. The split time between African and non-African populations was set to 2,500 generations ago, followed by a bottleneck of size 100 in the non-African population for 100 generations. Figure 4.2 shows the demographic

Table 4.2 Encoding of emission types in site-wise model

Genotype			Emission type k
African panel	non-African sample	Archaic	
1	0	1	1
1	1	0	2
0	1	1	3
All other combinations			4

* 0: ancestral allele in all genomes; 1: derived allele in at least one genome

diagram used in the simulation. The mutation rate and recombination rate are fixed at 1.25×10^{-8} /(site · generation) and 1.2×10^{-8} /(site · generation), respectively.



n: number of sampled sequences; N: effective population size; α: admixture proportion

Fig. 4.2 Demographic model used in simulations to train the site-wise HMM

The distribution of allele-sharing patterns indeed differs clearly between segments of archaic and modern origin (Figure 4.3), confirming that the choice of emission signal is informative about the underlying state. The MLE model parameters are listed in Box 4.1.

Run with the MLE parameters on simulations, Viterbi decoding recovers 99.75% of the archaic segments with a false discovery rate of 0.114. Overall, the model correctly assigns the

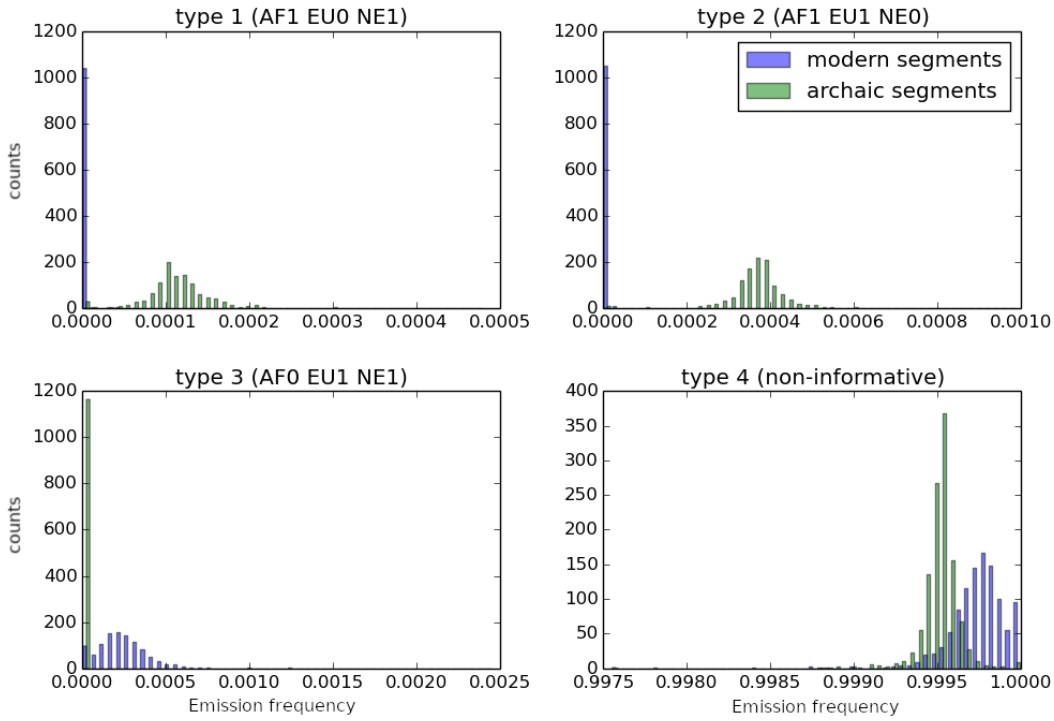


Fig. 4.3 Histograms comparing the frequency of emissions in archaic vs. modern segments

Box 4.1 Sitewise HMM parameters estimated from simulations

Initial state distribution (π)

$$\begin{bmatrix} 0.9709 & 0.0291 \end{bmatrix}$$

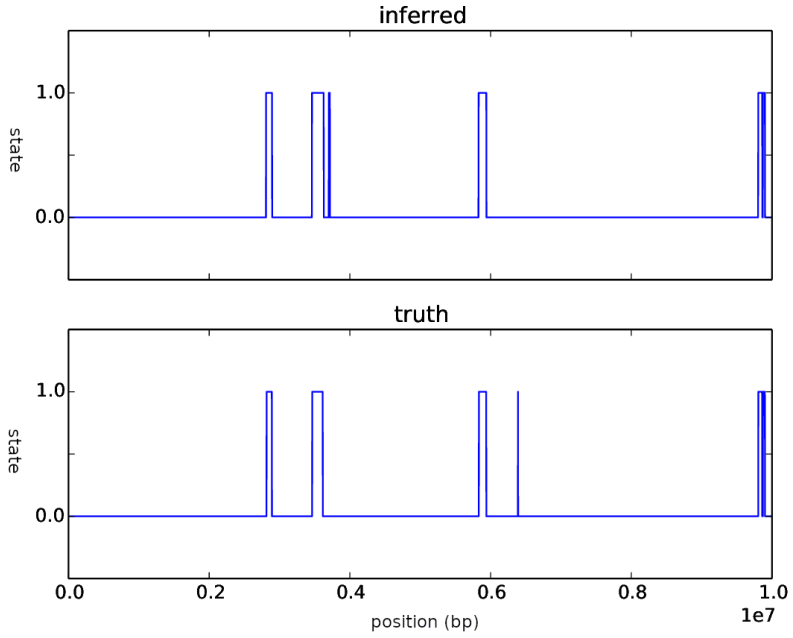
Transition probabilities (T)

$$\begin{bmatrix} 0.9999993 & 7.065\times10^{-7} \\ 2.359\times10^{-5} & 0.999976 \end{bmatrix}$$

Emission probabilities (E)

$$\begin{bmatrix} 0.0001081 & 0.0003494 & 1.7514\times10^{-6} & 0.9995407 \\ 4.836\times10^{-9} & 8.291\times10^{-9} & 0.0004507 & 0.9995493 \end{bmatrix}$$

hidden state in 99.45% of the sites. Figure 4.4 compares the inferred state on one non-African haplotype with the true state as an example.



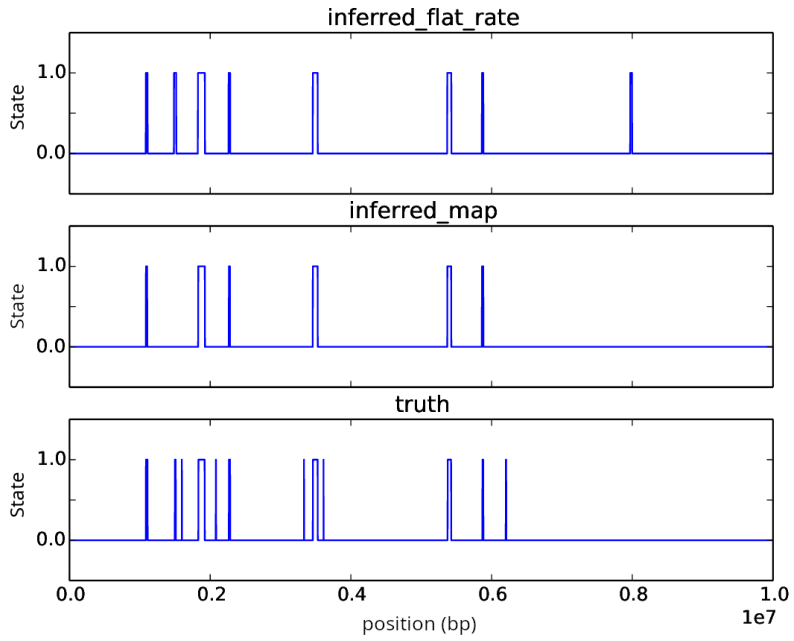
0: modern state; 1: archaic state

Fig. 4.4 Inferred and true state along one simulated non-African haplotype

Incorporating varations in recombination rate

As the recombination rate is not constant across the genome in reality, I also performed simulations with local recombination rate variations to evaluate the influence on the HMM inference. The demographic model remains unchanged, but the local recombination rate is read from the genetic map on chromosome 1. The HMM was run first without and then with the genetic map following 4.1.2, both using the MLE model (Box 4.1). When the genetic map is provided, the transition probabilities are not fixed but calculated from Equation 4.1 assuming that the time since admixture (t) is 2,000 generations. Figure 4.5 shows the inference result in the last 10 Mb of the chromosome. The performance of Viterbi decoding with and without the genetic map is summarised in Table 4.3.

Although the HMM misses out some true archaic segments when the genetic map is provided, this is accompanied by a 2/3 decrease in false discovery rate. Therefore it is desirable to



0: modern state; 1: archaic state

Fig. 4.5 Inferred and true state along one simulated non-African haplotype, comparing the effect with and without the genetic map

Table 4.3 Performance of HMM with and without genetic map on sequences simulated with variations in recombination rate

	Accuracy	Recall	False discovery rate	False omission rate
without genetic map	0.9908045	0.9908045	0.2246925	0.0012445
with genetic map	0.9948902	0.8928102	0.0729990	0.0031737

include recombination rate information. All runs on empirical data are performed with the genetic map.

Test runs

To test if the method produces sensible result on empirical sequencing data, I ran the HMM with the MLE parameters on two haploid genomes in the HGDP panel, one from the Bantu population in South Africa, the other from Inner Mongolia in China (4.6). The model detects much more Neanderthal segments in the Chinese genome than in the Bantu one, in agreement with little or no Neanderthal ancestry in sub-Saharan Africa. Furthermore, the locations of inferred Neanderthal segments mostly coincide with the peaks in a map describing the distribution of Neanderthal ancestry in East Asians from [1]. Minor mismatches can be expected, since the map from the previous study represents the average ancestry level across 572 East Asian haplotypes, which does not include the Mongolian population in China.

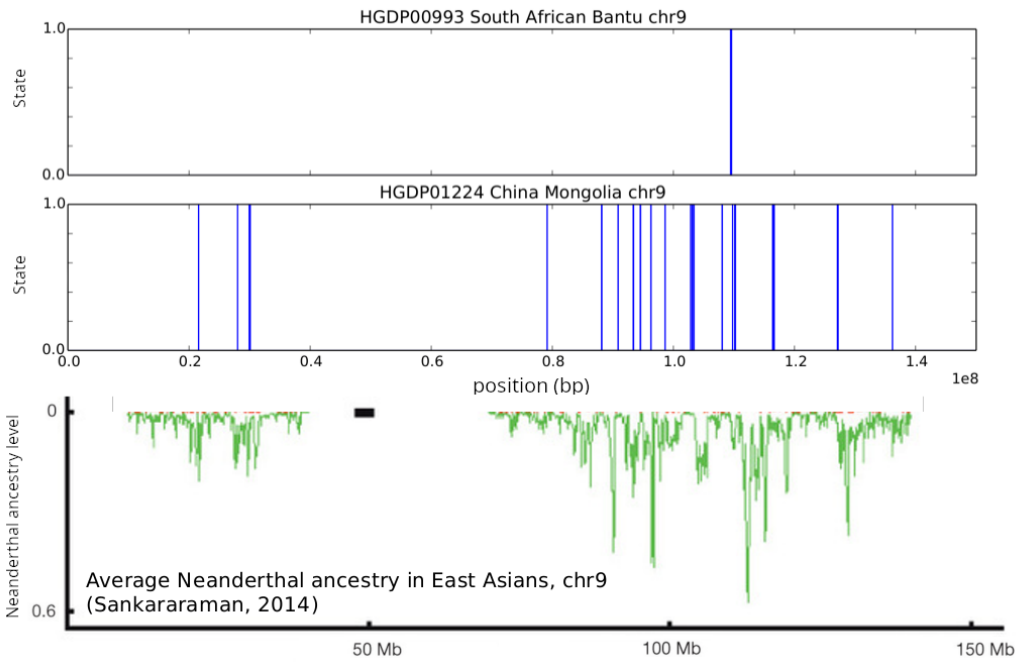


Fig. 4.6 Detected archaic segments in two genomes from the HGDP panel, in comparison to a published map of Neanderthal ancestry in East Asians [1]

Baum-Welch training

I also trained the model using the Baum-Welch algorithm on both simulated and sequencing data. Initial parameter values are from the MLE model (Box 4.1), and iteration terminates

when the change in log-likelihood falls below 1×10^{-5}. In some runs the parameters converged close to their starting values, and the inference results from the initial MLE model and the trained model are almost identical. However, the majority of the runs finished with a model that assigns a high proportion of the archaic state (sometimes up to 0.7). One such model is shown in Box 4.2. Whilst the emission probabilities in the modern state does not

Box 4.2 Site-wise HMM parameters with high archaic proportion from Baum-Welch training

Initial state distribution (π)

$$\begin{bmatrix} 0.7680 & 0.2320 \end{bmatrix}$$

Transition probabilities (T)

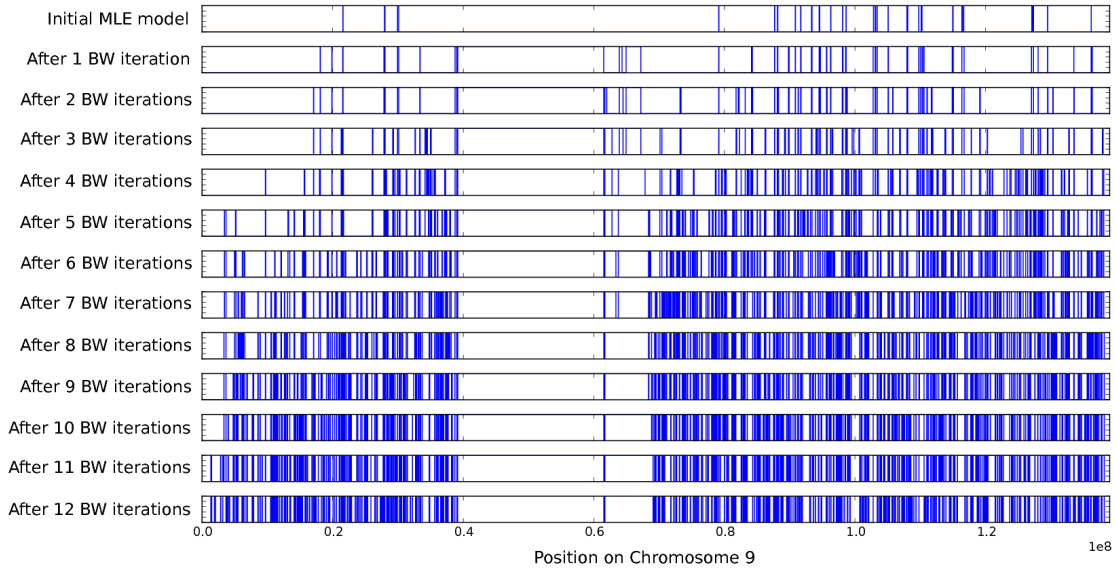
$$\begin{bmatrix} 0.99993 & 6.941\times10^{-5} \\ 2.627\times10^{-5} & 0.99997 \end{bmatrix}$$

Emission probabilities (E)

$$\begin{bmatrix} 0.0001738 & 0.0004589 & 2.817\times10^{-8} & 0.9994 \\ 1.927\times10^{-6} & 1.927\times10^{-6} & 2.283\times10^{-5} & 0.9998 \end{bmatrix}$$

change much from the MLE model, the archaic state becomes more permissive to the first two emission types expected to signify the modern state, coupled with an increased probability to enter the archaic state from the modern state (T_{01}). Figure 4.7 shows the archaic segments detected with the updated model after each Baum-Welch iteration, highlighting the gradual expansion of the archaic state. Such trajectory happens more often in real-world sequences, but the likelihood of such models always exceed those close to the MLE model, even in simulated dataset.

To better understand what features might have caused the training algorithm to expand the archaic state, I examined a genomic region in which some segments not sharing any derived alleles with the Neanderthal genome become tagged as archaic after several Baum-Welch iterations. The inference results using parameters after 5-8 Baum-Welch iterations are shown in Figure 4.8 along with local recombination rate and the density of different emission types in 1kb windows. The problematic region appears to contain longer tracks of missing and non-informative sites than the rest. Perhaps the HMM will become misled when the sequential structure is largely disturbed by missing segments, or when type 4 emission is dominant. I subsequently explored two variants of the HMM in hope of avoiding this issue with Baum-Welch training.



regions in blue: archaic state

Fig. 4.7 Inferred archaic segments on chromosome 9 of HGDP01224 using parameters after each Baum-Welch iteration, showing the expansion of the inferred archaic state

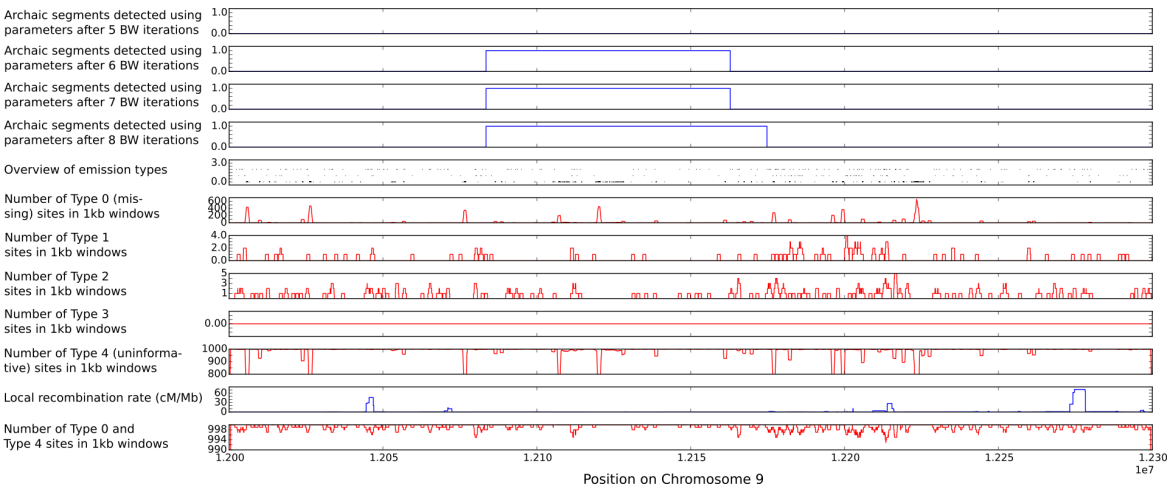


Fig. 4.8 Features along a genomic region with spurious archaic state, showing that the inferred archaic state tends to arise in regions with long tracks of missing and uninformative sites

4.2.2 Window-based model

One possibility to alleviate the effect of local clusters of missing and non-informative sites is to adopt a window-based approach, where the total number of sites belonging to each type in the window becomes the observed variable, and the state of the window is determined by the majority of the sites. To start with, I only considered the presence or absence of each type of sites. There exist 8 emission categories in total, corresponding to all the binary combinations of 3 types (Table 4.4). The same set of simulations from 4.2.1 is used to obtain the MLE model in Box 4.3. Here the window size is set to 10kB.

Table 4.4 Encoding of emission states in window-based HMM

presence (1) or absence (0)			**Emission state k**
Type 1 sites	Type 2 sites	Type 3 sites	
0	0	0	1
1	0	0	2
0	1	0	3
0	0	1	4
1	1	0	5
1	0	1	6
0	1	1	7
1	1	1	8

Box 4.3 MLE model in window-based HMM

Initial state distribution (π)

$$\begin{bmatrix} 0.9679 & 0.0321 \end{bmatrix}$$

Transition probabilities (T)

$$\begin{bmatrix} 0.99947 & 5.339\times10^{-4} \\ 0.02153 & 0.97847 \end{bmatrix}$$

Emission probabilities (E)

$$\begin{bmatrix} 0.6561 & 0.0491 & 0.2589 & 9.745\times10^{-4} & 0.0349 & 9.868\times10^{-6} & 4.261\times10^{-5} & 1.869\times10^{-6} \\ 0.7261 & 3.836\times10^{-5} & 1.963\times10^{-4} & 0.2736 & 6.770\times10^{-6} & 2.257\times10^{-5} & 4.288\times10^{-5} & 0 \end{bmatrix}$$

Figure 4.9 shows the Viterbi decoding result from five test runs on simulated sequences. The model infers many false archaic segments; the false discovery rate is as high as 0.2274.

Not satisfied with this model, I also tested a logistic regression model to predict the archaic/modern state of the windows. In this way, the number of sites in each type in the window, other than the mere presence/absence of them, becomes the predictor variables. To

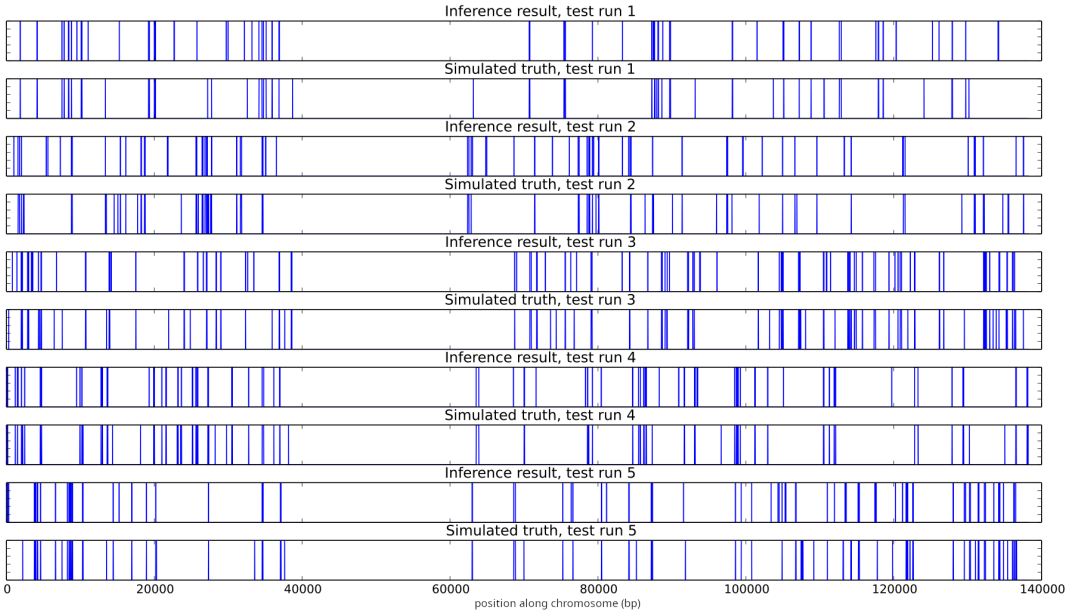


Fig. 4.9 Inferred vs. true archaic state along simulated chromosome 9 using window-based HMM

compensate for the lack of sequential structure, I also tried including the numbers from the previous and the next window. The model is able to recover more archaic segments with information from the neighbouring windows, yet the performance is still poorer than the site-wise HMM regarding the recall and false discovery rate (Table 4.5).

Table 4.5 Performance of logistic regression in window-based HMM

Included predictors	Recall	False discovery rate	False omission rate
current window only	0.6844	0.1676	0.0052
previous and current windows	0.7244	0.1809	0.0045
previous, current and next windows	0.7822	0.1579	0.0036

4.2.3 Informative-site-only model

Another variant of the HMM examines only the informative sites along the genome (emission type 1, 2 and 3 in Table 4.2. Transitions are only allowed before informative sites. Since the distance between the Markovian steps is no longer constant, transition probabilities have to be calculated on-the-fly, either using genetic distance extrapolated from a genetic map or, in the absence of such a map, using the physical distance and a constant per-site recombination

rate. In this way, non-informative and missing sites are invisible to the model and should not be mistaken as features of a state. Since the vast majority of the sequence no longer needs to be stored, this model also shows great improvement in memory and computational efficiency over the site-wise model.

To estimate model parameters, I simulated 20 haploid genomes with the real-world human genetic map. The number of African lineages was increased to 200 to match the complete HGDP dataset; otherwise the demographic model for simulation remains the same as in Figure 4.2. The MLE model parameters for the informative-site-only model is shown in Box 4.4. Since the genetic map was used following result from 4.2.1, the transition matrix is not time-homogeneous. The time since admixture (t) was fixed at 2,000 generations ago, the true value used in simulations. Viterbi decoding using this model recovered 90.74% of the archaic segments, with a false discovery rate of 0.0368.

Box 4.4 MLE model for informative-site-only HMM

Initial state distribution (π)

$$\begin{bmatrix} 0.9655 & 0.0345 \end{bmatrix}$$

Emission probabilities (E)

$$\begin{bmatrix} 0.1777 & 0.8208 & 1.336\times10^{-3} \\ 1.691\times10^{-3} & 2.015\times10^{-4} & 0.9981 \end{bmatrix}$$

Baum-Welch training

The Baum-Welch algorithm can also be used to update the emission probabilities (E) and the admixture proportion (α) in the new model. Since the informative sites are not evenly spaced, the algorithm no longer applies to transition probabilities, which will be calculated on-the-fly from admixture time t and the genetic distance anyway. The value of t has to be fixed. In practice, varying t while keeping the other parameters unchanged has little influence on the Viterbi decoding result: in a simulated sequence containing over 3 Mb true archaic segments, increasing t from 1,000 to 4,000 caused the HMM to detect 14 more archaic segments spanning 2,935 bp, 12 of which were only 1 bp long.

Under this implementation the iterations are able to move away from models similar to Box 4.2, and converge close to the MLE model (Box 4.4). On very rare occasions, however, Baum-Welch training led to another type of models with a permissive archaic state, an example of which is shown in Box 4.5. This time the archaic state becomes more tolerant towards the second type of emission, where a new mutation is shared between the African

panel and the non-African genome. It appears that once the non-informative sites are ingored, this type becomes the most common and sometimes becomes absorbed into the archaic state.

Box 4.5 Example of problematic model resulted from training the informative-site-only HMM

Initial state distribution (π)

$$\begin{bmatrix} 0.5864 & 0.4136 \end{bmatrix}$$

Emission probabilities (E)

$$\begin{bmatrix} 0.3609 & 0.6390 & 1.236\times10^{-4} \\ 1.304\times10^{-3} & 0.9602 & 0.03847 \end{bmatrix}$$

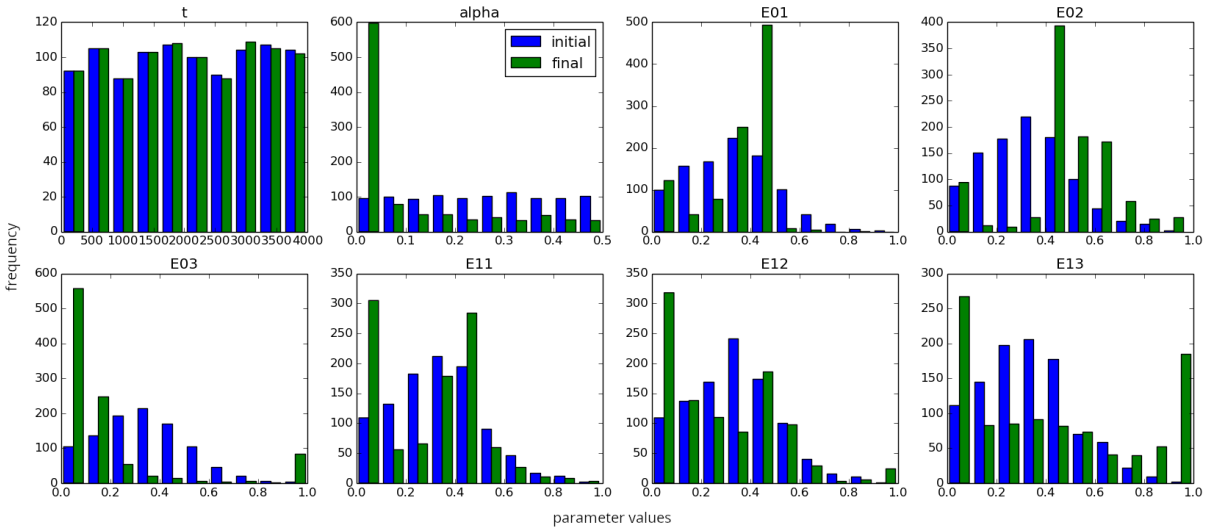
L-BFGS training

I considered training the model by numerical optimisation in hope of avoiding the expansion of the archaic state in Baum-Welch training. The L-BFGS-B algorithm was used to search for the parameters that maximizes the log-likelihood of the model, as described in 4.1.4. Since all the entries in the emission matrix E should be positive and sum up to 1 in each row, it is parameterised inside the optimisation function as:

$$E = \begin{bmatrix} \frac{p_0}{p_0+q_0+1} & \frac{q_0}{p_0+q_0+1} & \frac{1}{p_0+q_0+1} \\ \frac{p_1}{p_1+q_1+1} & \frac{q_1}{p_1+q_1+1} & \frac{1}{p_1+q_1+1} \end{bmatrix}$$

where p_0, q_0, p_1, q_1 are all positive numbers. Along with admixture proportion α (lower bound 0, upper bound 1) and admixture time t (lower bound 1), there are six parameters to be estimated. I randomly generated 1,000 initial models, and trained them with the same set of simulations as before. α was drawn from a uniform distribution between 0 and 1; t was drawn from a uniform distribution between 0 and 4000; all the entries in the emission matrix E were also drawn from a uniform distribution between 0 and 1, then rescaled to ensure values in each row add up to 1. The distribution of parameter values at the start and end of L-BFGS-B training is shown in Figure 4.10. The data provide little information to estimate t; the correlation coefficient between its initial and final values was over 0.99. In general, the more common state (modern state) becomes associated with the first two types of emission as expected; yet the emission probabilities in the rarer state (archaic state) varies between runs.

I ranked all the resulting models by their likelihood. The top four models out of 1,000 are all close to the model from Baum-Welch training in Box 4.5, but the next nine are close to



The state labels are swapped sometimes to ensure that 0 corresponds to the more common and 1 the rarer state; α always measures the proportion in the rarer state.

Fig. 4.10 Parameter value distributions before and after L-BFGS-B training

the MLE model in Box 4.4. In fact, the bimodal distribution of E_{13} values after L-BFGS training (Figure 4.10, last three panels) results from a combination of these two types of models. Figure 4.11 compares the result of Viterbi decoding using the MLE model and the best L-BFGS-B trained model. Unsurprisingly, the latter tags a massive amount of false archaic regions.

In conclusion, the HMM over only informative sites shows considerable improvement in both performance and efficiency over the HMM over all sites in the sequence. Both Baum-Welch training algorithm and numerical optimisation reject the problematic model in site-wise HMM (Box 4.2), yet another model in which the archaic state absorbs the second emission type is slightly favoured over the MLE model. Distinguishing an archaic and a modern state through unsupervised training proves difficult. To the best of my knowledge, all other existing methods to detect archaic segments with reference to the archaic genomes also learn model parameters from simulations with tagged unadmixed and introgressed segments. Nevertheless, with minor artificial inspection to exclude the improper models, both training methods are capable of fine-tuning the model parameters when one is uncertain about the demographic history.

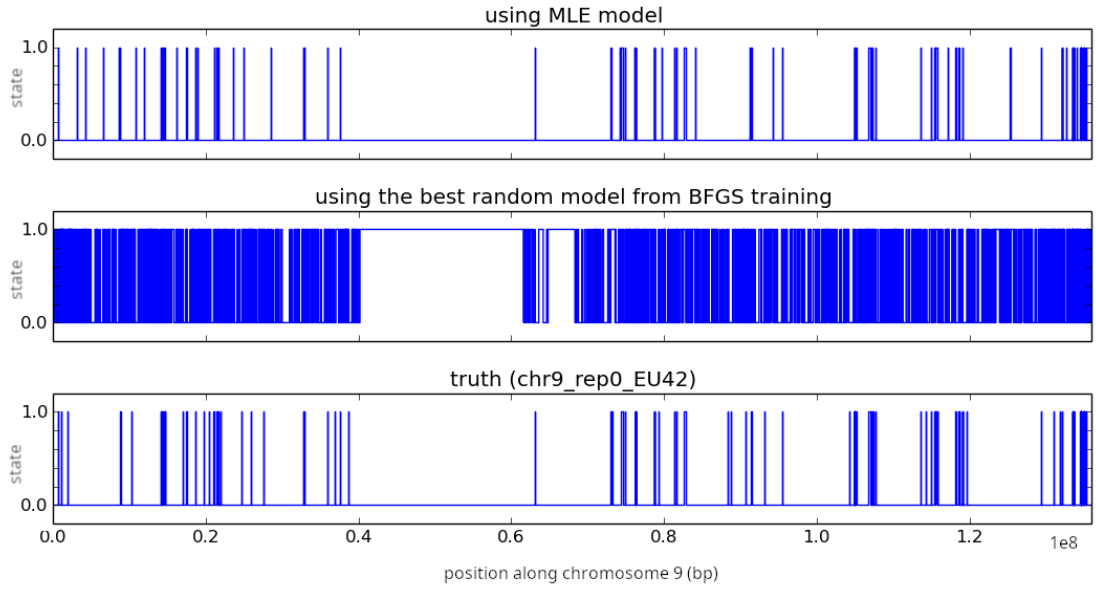


Fig. 4.11 Comparison of Viterbi sequences using the MLE model and the best model from L-BFGS-B training

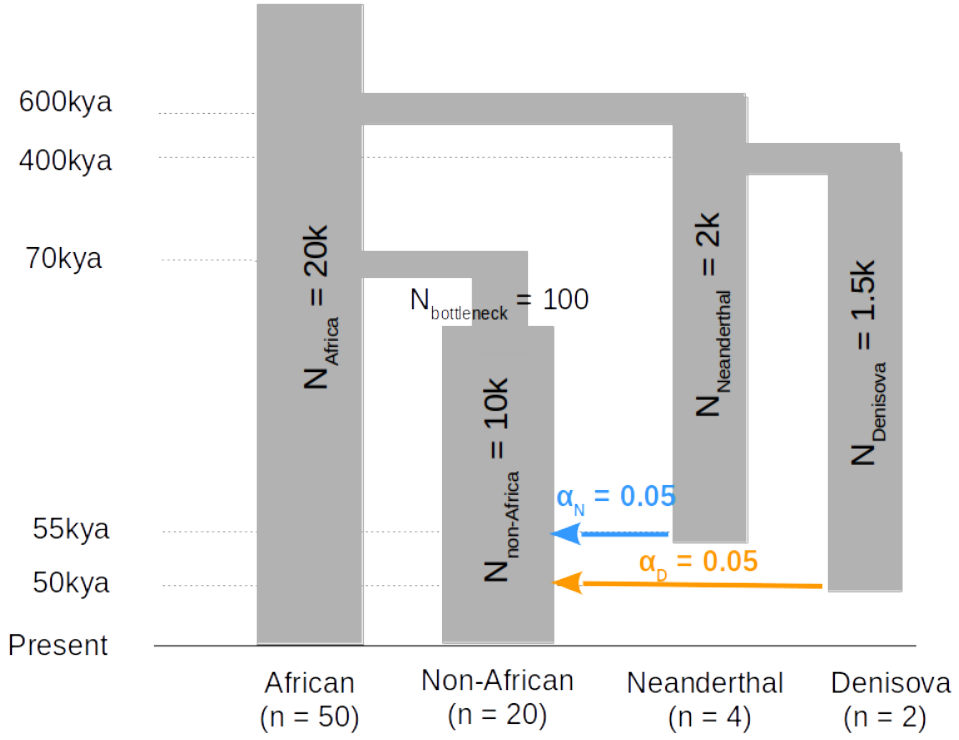
4.3 Three-state vs two-state model

A major challenge in applying the HMM to the HGDP dataset lies in distinguishing between Neanderthal and Denisovan segments. Their genomes share 87.9% of SNPs at ancestral positions and 97.7% at derived positions [73]. Although the amount of Denisovan ancestry is substantially lower than Neanderthal ancestry in most non-African populations (with the exception of Oceania) [47, 76], misclassifying them could have a major impact on analysis about the diversity in archaic segments. I compared two methods for differentiation: 1. implementing an HMM with three states; 2. running the two-state HMM independently with different archaic genomes, and assigning the most likely state according to posterior probabilities.

4.3.1 Three-state model

An obvious approach is to extend the HMM to include three hidden states, Neanderthal, Denisova, and modern. I simulated genetic sequences under a new demographic model, including a Denisovan population that diverged from the Neanderthal population 400k years ago [66] and gene flow from the Denisovan population into non-African population 50k years ago. The complete demographic model is illustrated in Figure 4.12. Some parameters are changed to reflect another newly available high-coverage Neanderthal genome from

Vindija Cave and updated inference about modern human demography. Similar to the case with only one archaic population, if a present-day lineage ever travels through the Neanderthal/Denisovan population in its coalescent history, the region is considered to be in the Neanderthal/Denisova state; otherwise the region is labelled as modern.



N: effective population size; n: number of sampled sequences; α: admixture proportion

Fig. 4.12 Demographic model used in simulations with Neanderthal and Denisovan admixture

With genetic information from four panels (sub-Saharan African, non-African, Neanderthal, and Denisova), there exist a total of 16 possible combinations of genotypes (including non-segregating sites where the derived allele is present in all of the panels, and none of the panels). The proportion of each type emitted from the three hidden states is listed in Table 4.6. Cases 0, 4, 8 and 15 are not informative about the relationship between the lineages, since they describe mutations on either one or all the lineages. Cases 3, 11 and 12 strongly support the modern states, thus can be combined as one category. Similarly, cases 10 and 13 rule out the Neanderthal state while allowing for modern or Denisova state; cases 9 and 14 are the counterpart that rule out the Denisova state while allowing for modern or Neanderthal state. Eight emission types remain after binning similar ones (Table 4.7). Using the new encoding for emission states, Box 4.6 shows the MLE model based on simulations. The transition

matrix is still calculated on-the-fly according to the genetic distance and the proportion of each state.

Table 4.6 Proportions of 16 allele sharing patterns emitted in unadmixed, Neanderthal and Denisova state in simulation

Index	**Derived allele***				**True simulated proportion in states**		
	A	E	N	D	Unadmixed	Neanderthal	Denisova
0	0	0	0	0	0.90361467	0.04470899	0.05167633
1	0	0	0	1	0.94287598	0.04808293	0.00904109
2	0	0	1	0	0.93314042	0.01457970	0.05227988
3	0	0	1	1	0.99941416	0.00046463	0.00012121
4	0	1	0	0	0.92298185	0.03274278	0.04427537
5	0	1	0	1	0.01219279	0.00293459	0.98487261
6	0	1	1	0	0.01359961	0.98206762	0.00433277
7	0	1	1	1	0.04651198	0.41047789	0.54301013
8	1	0	0	0	0.89010031	0.05175048	0.05814920
9	1	0	0	1	0.90724025	0.09076098	0.00199877
10	1	0	1	0	0.88961774	0.00374902	0.10663324
11	1	0	1	1	0.99954413	0.00029831	0.00015756
12	1	1	0	0	0.99987553	0.00007604	0.00004844
13	1	1	0	1	0.89487555	0.00120523	0.10391922
14	1	1	1	0	0.90909817	0.08827130	0.00263053
15	1	1	1	1	0.84278822	0.07481581	0.08239597

*A: Africa; E: non-Africa; N: Neanderthal; D: Denisova

Table 4.7 Encoding of emission types in the three-state HMM

New emission type	**Index from Table 4.6**
1	1
2	2
3	3, 11, 12
4	5
5	6
6	7
7	9, 14
8	10, 13

Testing this model on simulations, I am able to recover over 90% of both Neanderthal and Denisovan segments with the correct label (Table 4.8). However, the false discovery rate

Box 4.6 MLE model in the three-state HMM, $\alpha_N = \alpha_D = 0.05$

Initial state distribution (π)

$$\begin{bmatrix} 0.90 & 0.05 & 0.05 \end{bmatrix}$$

Emission probabilities (E)

$$\begin{bmatrix} 0.1155 & 0.1566 & 0.6719 & 6.648\times10^{-5} & 7.384\times10^{-5} & 7.210\times10^{-4} & 0.0273 & 0.0278 \\ 0.2567 & 0.1065 & 6.041\times10^{-3} & 6.969\times10^{-4} & 0.2322 & 0.2772 & 0.1173 & 3.359\times10^{-3} \\ 0.0409 & 0.3236 & 2.194\times10^{-3} & 0.1981 & 8.680\times10^{-4} & 0.3106 & 2.576\times10^{-3} & 0.1211 \end{bmatrix}$$

in both states is alarmingly high, indicating that many segments tagged as Neanderthal or Denisova are in fact modern. The problem became even worse when I lowered the amount of Neanderthal ancestry to 0.03 and Denisovan ancestry to 0.01, which should be more realistic in most Eurasian populations. Filtering out archaic segments that are shorter than 10kB only slightly decreases the false discovery rate (Table 4.8).

Table 4.8 Performance of MLE model in the three-state HMM

Model	False discovery rate (N, D)	Recall (N, D)	Incorrectly labelled archaic (N, D)
Equal admixture proportion ($\alpha_N = \alpha_D = 0.05$)	0.2729, 0.1922	0.9300, 0.9397	0.0052, 0.0084
Equal admixture proportion, segment lengths > 10kb	0.2394, 0.1537	0.9128, 0.9243	0.0049, 0.0075
Unequal admixture proportion ($\alpha_N = 0.03, \alpha_D = 0.01$)	0.3048, 0.5210	0.9287, 0.9188	0.0018, 0.0116
Unequal admixture proportion, segment lengths > 10kb	0.2585, 0.4593	0.9114, 0.9012	0.0016, 0.0115

4.3.2 Independent runs of two-state model

In a second approach, I run the two-state informative-sites-only HMM with the model from Box 4.4 twice, first with two Neanderthal genomes in the archaic panel, then with the Denisovan genome. To combine the result from independent runs, the HMM is configured to output posterior probabilities of the archaic state at each input site, rather than Viterbi sequences. All the sites from both runs are then merged, and the posterior probabilities of the sites not seen in one run (as some sites are informative regarding one archaic panel, but not the other) are extrapolated from neighbouring sites, assuming a linear change locally. If the probability of being in the archaic state is below 0.5 in both runs, the site will be tagged as modern; if the posterior probabilities of being in the Neanderthal and the Denisova state both exceed 0.8, the site will be tagged as ambiguous archaic; otherwise the archaic state

with a higher posterior probability will prevail. The final segments are formed by linking consecutive sites in the same state.

The performance evaluated on simulations with similar and unequal amount of admixture is shown in Table 4.9. The false discovery rate is greatly reduced, at the cost of binning a large amount of archaic segments into the ambiguous category. The true origin of segments ending up in the ambiguous category reflects the ratio of contribution from the two admixture events. Nevertheless, 90% of the archaic segments overall are still recovered. The probability to cross-label archaic segments is also low. Filtering out short archaic segments slightly reduced false positives, but also reduced recall rate. Most short archaic segments detected this way result from longer ones being broken down into Neanderthal, Denisova and ambiguous parts, instead of artefacts in the modern regions.

Table 4.9 Performance of running two-state model independently

Model	**False discovery (N, D, A)***	**Recall (N, D)**	**Recall including ambiguous (N, D)**	**Cross-labelled archaic (false N, false D)**	**Proportion in ambiguous (true N, true D)**
Equal admixture proportion ($\alpha_N = \alpha_D = 0.05$)	0.0426, 0.0672, 0.0132	0.2424, 0.2246	0.9049, 0.9150	0.0122, 0.0246	0.4668, 0.5200
Equal admixture proportion, segment lengths > 10kb	0.0383, 0.0629, 0.0098	0.2035, 0.1890	0.8459, 0.8614	0.0113, 0.0235	0.4675, 0.5226
Unequal admixture proportion ($\alpha_N = 0.03, \alpha_D = 0.01$)	0.0354, 0.0826, 0.0574	0.2113, 0.2189	0.8996, 0.9026	0.0021, 0.0403	0.7010, 0.2416
Unqual admixture proportion, segment lengths > 10kb	0.0154, 0.0527, 0.0484	0.1704, 0.1854	0.8383, 0.8482	0.0029, 0.0323	0.7076, 0.2440

In conclusion, neither method to distinguish Neanderthal and Denisovan segments is perfect. The three-state model is able to assign most archaic segments to their correct origin, but mislabels many modern regions as archaic. Running the two-state model independently reduces the false discovery rate, but a substantial amount of archaic segments end up in the ambiguous category. The latter also inevitably breaks down some archaic segments and assign the parts to different archaic bins, thus the distribution of segment lengths is disturbed (more details in 4.4.2). Nonetheless, a high false discovery rate of modern segments labelled as archaic is more concerning in most analyses about the distribution and diversity of archaic segments. The two-state model is also more flexible, where the criteria to assign the final state from posterior probabilities can be adjusted according to the need in different analyses, reflecting a trade-off between Type 1 and Type 2 errors (Chapter 5 uses different sets of result; 4.4.2 also uses other ad-hoc criteria). Short segments can also be excluded in subsequent analysis. Unless noted otherwise, this is the default model in subsequent analyses and discussions.

4.4 Model features

4.4.1 Detection of segments at various lengths

The HMM relies on shared genetic variants between two lineages to infer the hidden genealogy. However, novel mutations are not guaranteed to occur, especially in older and therefore shorter archaic segments. We would expect that the detection power is higher for longer archaic segments, whilst many short segments will be missed out. Figure 4.13 and Figure 4.14 explore the relationship in simulations with one and two archaic admixture sources respectively. When only one archaic source is present, 90% of archaic segments longer than 30kB are detected, and the recall rate is close to 100% in segments longer than 50kB. When separate admixtures with the Neanderthal and the Denisova are included, however, if measured by the total base pairs classified into the Neanderthal/Denisova bin, the recall rate starts to drop around 20kB (Figure 4.14a). This is caused by the necessity to establish three archaic categories in the two-state HMM: the longer the true segments, the more likely that parts of it will be classified into different categories. Figure 4.14b shows the result using a different measurement for recall rate: a segment is considered to be detected as long as at least 10kB of it is tagged with the correct state. Similar to the case with a single admixture event, approximately 80% of archaic segments longer than 40 kB can be detected.

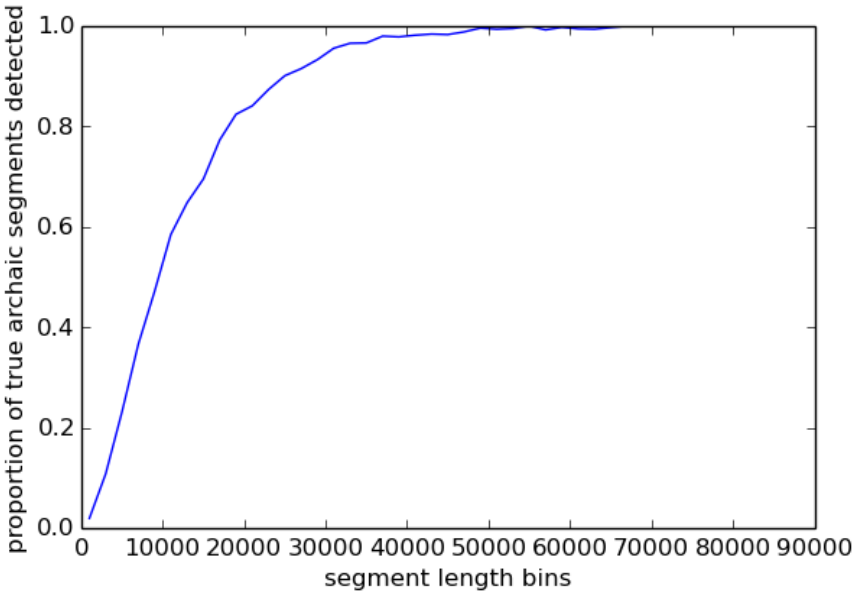
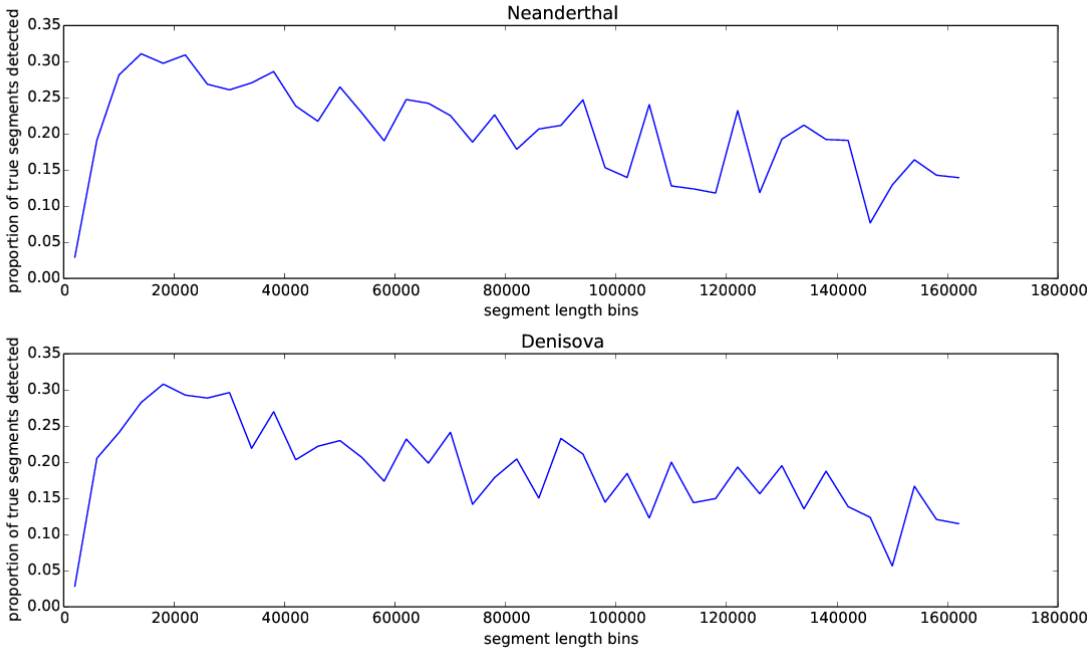
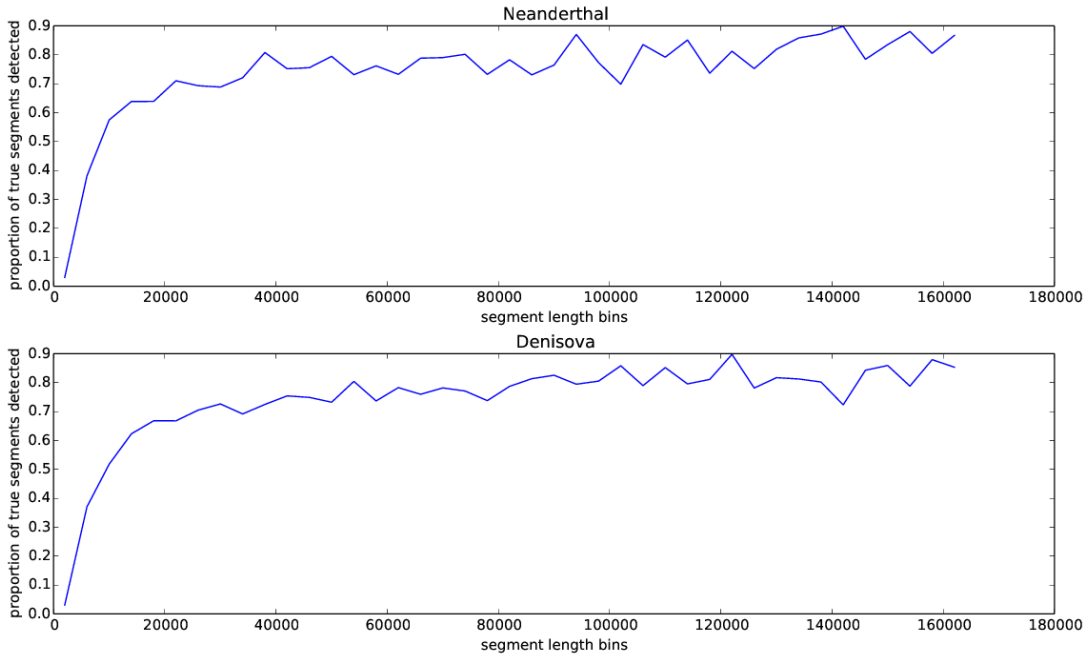


Fig. 4.13 Recall rate in archaic segments of different lengths, one admixture



(a) Measured by base pairs detected



(b) At least 10kB correctly detected

Fig. 4.14 Recall rate in archaic segments of different lengths, two admixtures

4.4.2 Inferring admixture time from segment lengths

After the archaic segments entered the modern human population, recombination breaks them down into shorter segments through the generations. The lengths of introgressed segments from the same source are exponentially distributed with a decay rate

$$\lambda = r \cdot t$$

where r is the per-unit recombination rate, and t the time since admixture. Although the effect of t on segment lengths should be reflected in the transition probabilities in the HMM, we have seen that it is very difficult to estimate t during model training. Here I explore another approach to estimate t from the lengths of recovered archaic segments in modern genomes.

One archaic admixture event

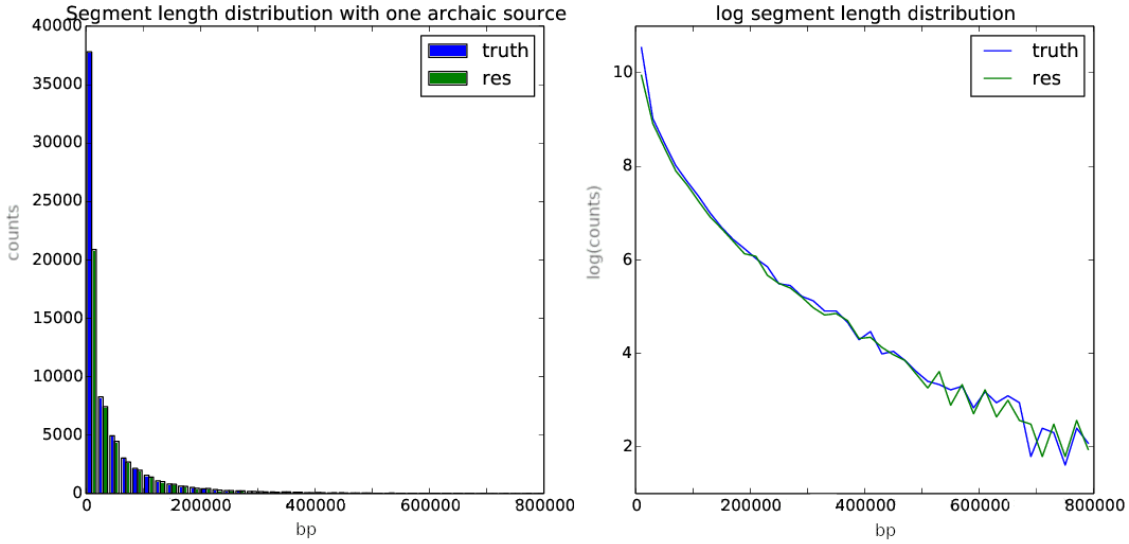
In simulations with one archaic admixture event, the distribution of the lengths of both true and detected archaic segments in base pairs roughly follows an exponential decay, despite an excess of segments shorter than 100kB and fluctuations towards very long segments (Figure 4.15a). As mentioned previously, the recall rate is lower in short segments (< 20kB) than in long ones. The exponential curve fits better when I examined the genetic lengths, other than physical lengths of the segments (Figure 4.15b).

Since the curve appears more noisy towards both extremes of the lengths, I used an iterative method based on the method of moments to estimate the decay constant from a truncated distribution [156]. Briefly, given a current value of λ_k, the value of λ_{k+1} is obtained as:

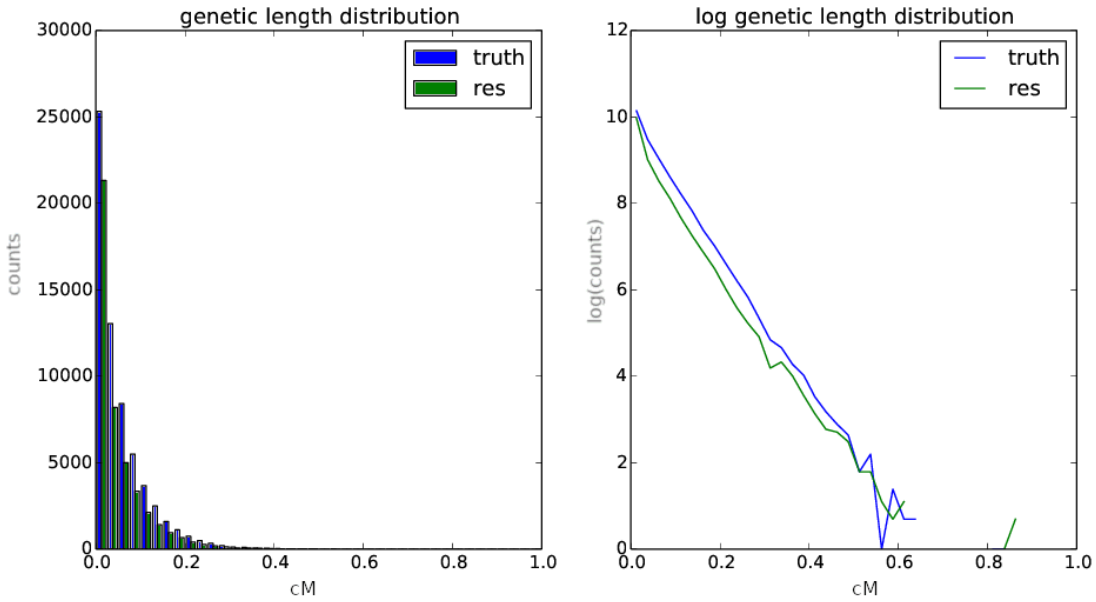
$$\lambda_{k+1} = \frac{\exp(-\lambda_k l_0) - \exp(-\lambda_k l_u)}{(\bar{l} - l_0)\exp(-\lambda_k l_0) - (\bar{l} - l_u)\exp(-\lambda_k l_u)}$$

where l_0 and l_u are the lower and upper length bounds, and $\bar{l}$ the mean segment length after the truncation. I chose $\lambda_0 = 1/(\bar{l} - l_0)$ as the initial value, and terminated the iteration once the difference between previous and updated λ falls below 1×10^{-4}. Assuming a average recombination rate of 0.01 per centiMorgan and a generation time of 29 years, the estimated value of λ can be converted to the time in years since admixture:

$$t = 29 \cdot \hat{\lambda} / 0.01$$



(a) Physical lengths (bp)



(b) Genetic lengths (cM)

Fig. 4.15 Distribution of true vs detected archaic segment lengths

The estimated admixture time from the genetic lengths of inferred archaic segments is around 51k years ago (Figure 4.16), which is about 7% lower than the true value used in the simulation (2,000 generations, namely 58k years). Perhaps because the recall rate is positively correlated with segment length across the entire range of lengths, the effect cannot be fully eliminated by excluding very short segments. The decay rate therefore still appears lower than the true value.

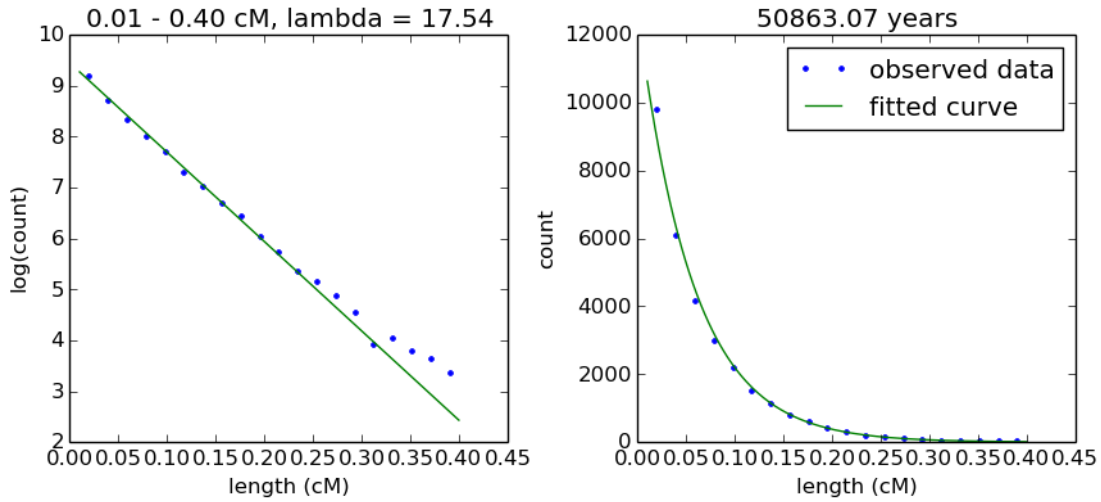


Fig. 4.16 Estimating the age of a single admixture event

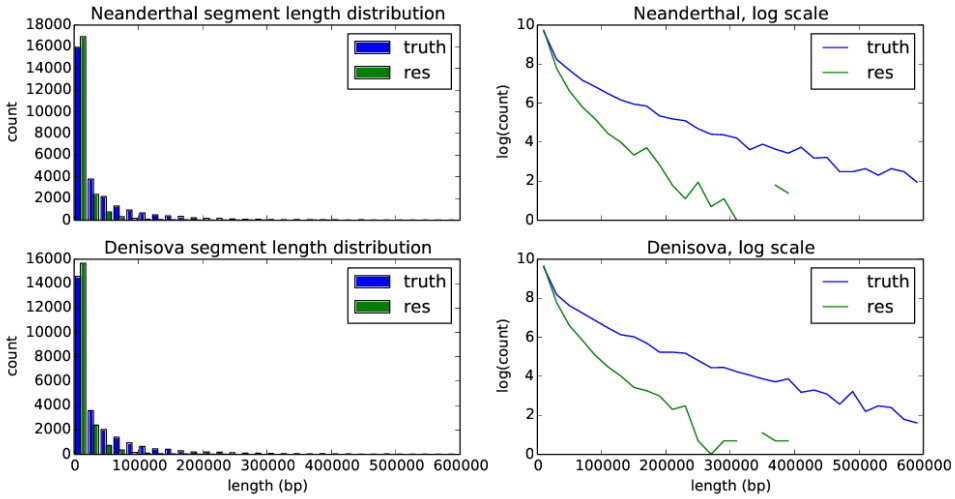
Two archaic admixture events

When two separate admixture events are present, the primary difficulty in dating each event lies in distinguishing between them. The criteria in 4.3.2 might no longer be appropriate, because comparisons of posterior probabilities will inevitably break down long segments. As shown in Figure 4.17a, this leads to a much sharper decay in estimated length distribution. Ignoring any segments containing ambiguous regions produces similar result (Figure 4.17b). But surprisingly, the Viterbi sequences from running the two-state model independently with the Neanderthal and Denisovan genomes generates a close match to the true distribution regarding segment lengths (Figure 4.17c). Fitting an exponential distribution on the genetic lengths infers an admixture time of 54,537 years ago with the Neanderthal, and 52,360 years ago with the Denisovans (Figure 4.18). The true values used in simulation are 55k years ago and 50k years ago, respectively. Both estimates are biased towards an intermediate value, probably confounded by segments from the other source. It is also worth noting that the segments detected in this way will surely contain segments from the other archaic source, yet

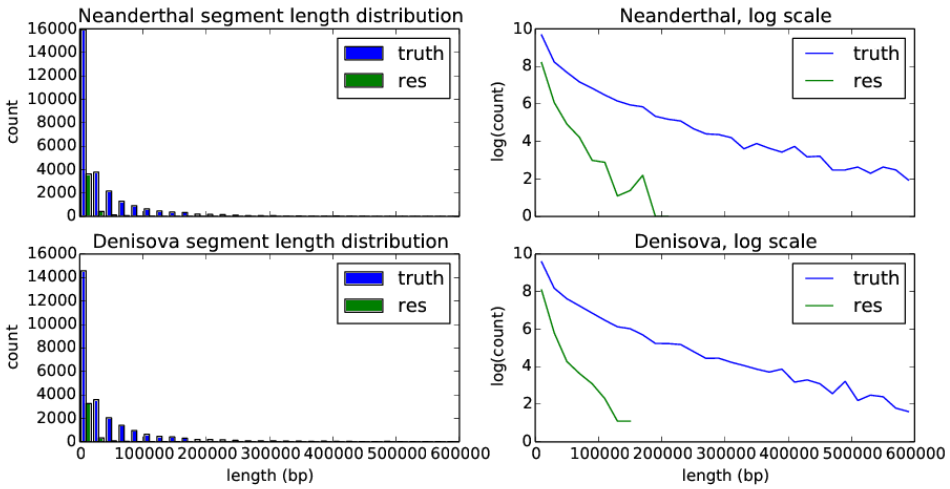
our purpose here is to capture the distribution of their lengths, rather than to obtain individual segments. Perhaps by trying to fit a geometric distribution of segment lengths, the HMM helps to approximate the true length distribution.

Afterwards, I explored if this method can accurately estimate the age or at least the relative age of the two admixture events when the admixture proportion is much higher from the Neanderthal than from the Denisovan source. I modified the admixture amount in the demographic model in Figure 4.12 to 0.02 from the Neanderthal and 0.01 from the Denisova. 100 non-African haplotypes were simulated to match the smallest sample size in Oceania. The time of Neanderthal admixture was fixed as 55k years ago, whilst the time of Denisovan admixture varied among 50k, 52k and 54k years ago in three sets of simulations. Figure 4.19a shows the fitted exponential curves and estimated time in each scenario. Although the length distribution of the true archaic segments is also shown for comparison, only the genetic lengths of the detected segments between 0.1 and 0.3 centiMorgans are used in fitting. Notably, the Neanderthal admixture is only correctly dated to be older in the scenario with the largest time gap (55k and 50k years ago). One explanation is that since Denisovan segments are relatively rare, a large proportion of the detected ones is likely to have come from Neanderthal in reality. The false assignment is more likely to happen in shorter segments, which do not contain enough Neanderthal-specific variants. As the model consistently detects more Denisovan segments of shorter lengths than there really are, the sharper decay is interpreted as an older admixture. The bias should also affect Neanderthal segments, but the effect will be more subtle considering the higher baseline level of genuine Neanderthal segments.

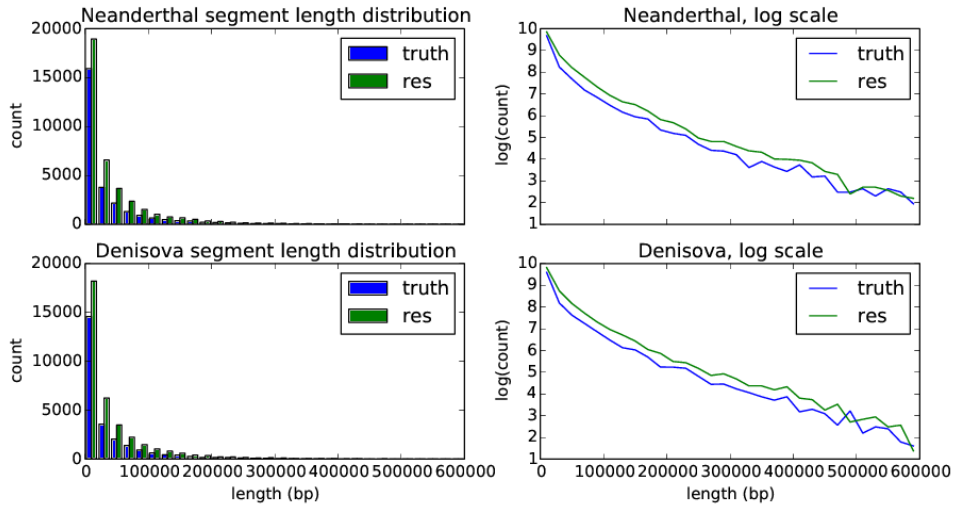
Since separate decoding proves problematic, I devised a different procedure to classify the segments into Neanderthal and Denisovan origins without breaking them down. Still starting from the segments tagged by separate runs of the two-state model, I first validated the labels on segments that do not overlap with any from the other source; in case of conflicts where putative segments from different sources overlap, the longer one will override with its label, and all other segments overlapping with it are removed; if the lengths of overlapping segments are equal, the posterior probabilities of the segments are used to break the tie; if the posterior probabilities are still the same, the label will be randomly chosen. These criteria greatly reduce the proportion of archaic segments assigned to the wrong origin (Table 4.10), but still fail to infer the correct sequence of the admixture events when the gap between them is shorter than 5k years (Figure 4.19b).



(a) Classified by posterior probabilities

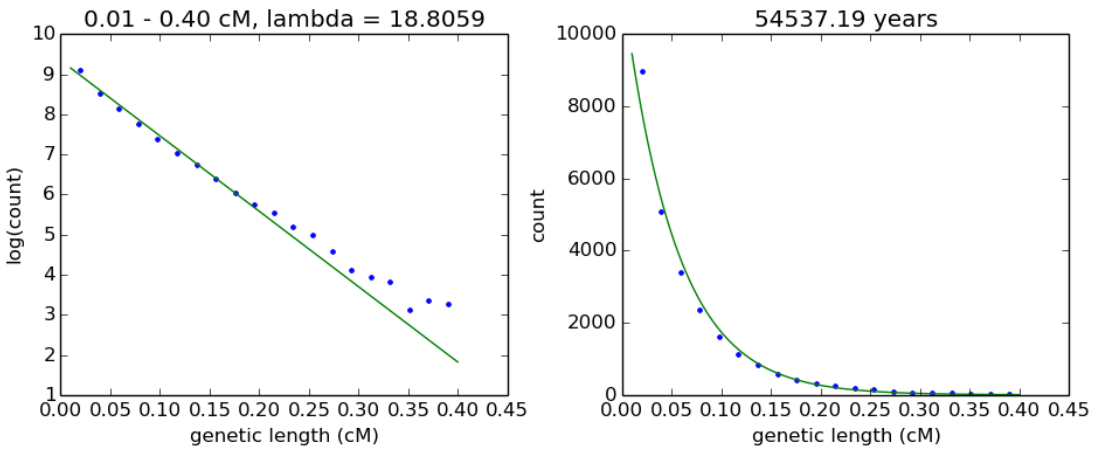


(b) Excluding segments containing ambiguous parts

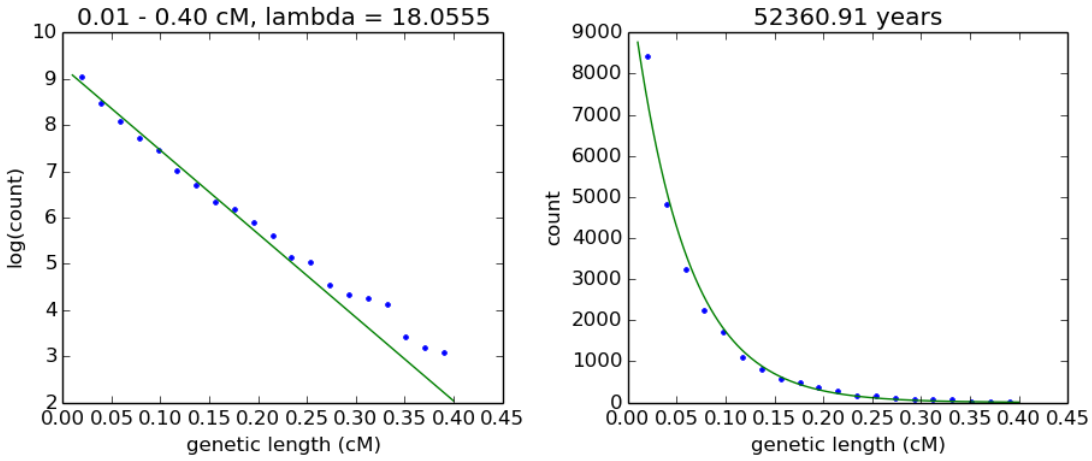


(c) Independent decoding from two-state models

Fig. 4.17 Distribution of true vs detected archaic segment lengths, two archaic admixture

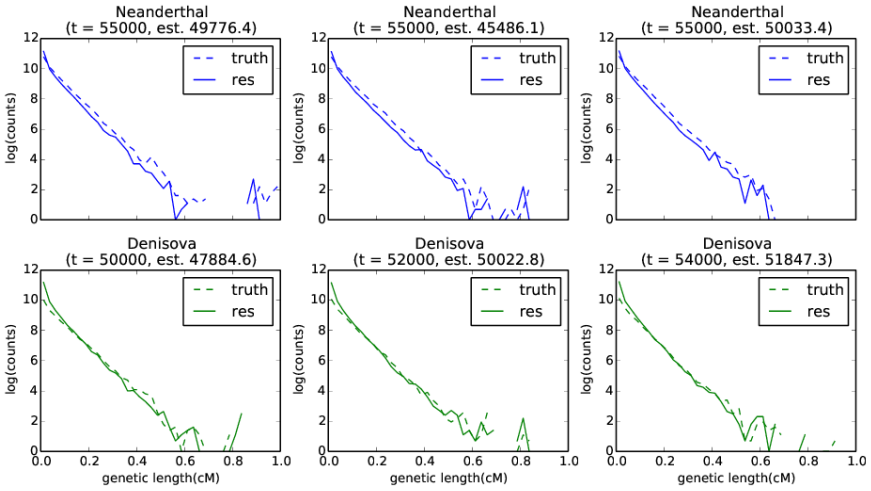


(a) Neanderthal (truth: 55kya)

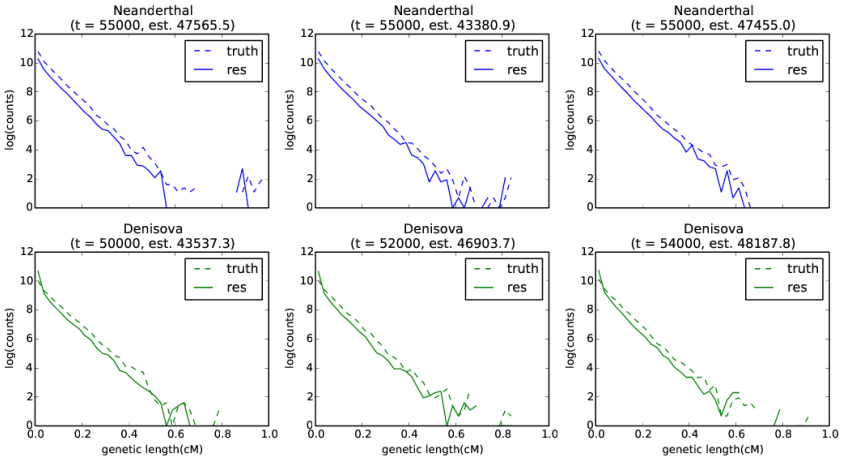


(b) Denisova (truth: 50kya)

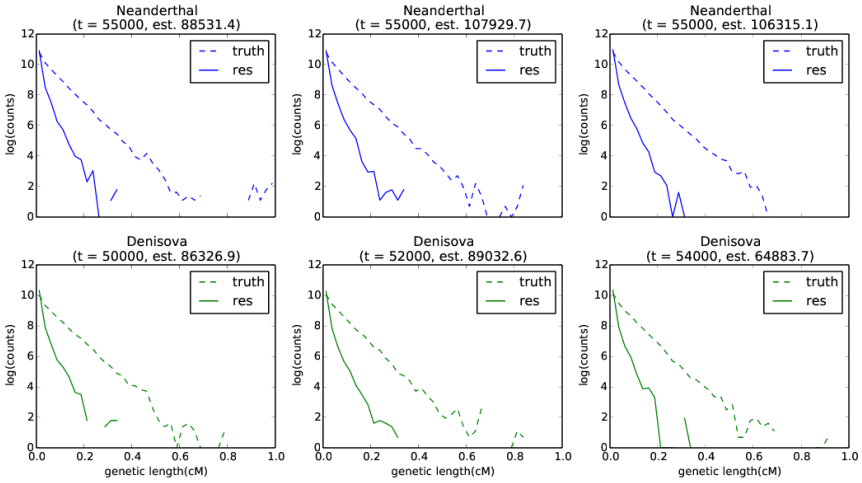
Fig. 4.18 Estimating the age of two archaic admixture events



(a) Separate runs for Neanderthal and Denisovan segments



(b) Classify overlapping segments by lengths and posterior probabilities



(c) Strict filters on segments by posterior probabilities

Fig. 4.19 Estimating the age of two archaic admixture events with different admixture proportions (Neanderthal 0.02, Denisova 0.01)

I also attempted filtering the segments according to the criteria in 4.3.2, where only segments whose posterior probability of being in the Neanderthal state is higher than 0.8 and of being in the Denisova state lower than 0.5 are tagged as Neanderthal, and vice versa. This method should have the lowest misclassification rate, at the cost of missing out many genuine archaic segments and breaking down long ones (Table 4.10). Indeed, the resulting segments are shorter than the true distribution, and the decay appears much sharper as long haplotypes are divided into shorter ones (Figure 4.19c). The upper bound of genetic length was lowered to 0.2 centiMorgan for age estimation. Although the Neanderthal admixture event is correctly estimated to predate the Denisovan admixture in all three scenarios, the observed length distribution no longer fits well with an exponential decay. The estimated time also deviates wildly from the true values.

Table 4.10 Performance of various procedures to tag Neanderthal and Denisovan segments when $t_N = 55k$ and $t_D = 50k$

Method	False discovery rate (N, D)	Recall (N, D)	cross-labelled archaic (N, D)
Separate decoding	0.3096, 0.5924	0.8198, 0.8326	0.2765, 0.5553
By length and post. prob.	0.0170, 0.2273	0.7899, 0.8208	0.0090, 0.1606
Strictly by post. prob.	0.0297, 0.0814	0.1260, 0.1363	0.0013, 0.0161

In conclusion, even with the assumption that each archaic gene flow results from a single episode without interval structure, dating the admixture events with the Neanderthal and the Denisova proves difficult when the amount of their contribution to the modern genome is vastly different, and the time gap between the two events is narrow. The performance might improve in scenarios where the sources of gene flow are more genetically distinct. An alternative approach independent of the HMM is to fit the exponential decay in linkage disequilibrium of derived alleles specific to an archaic source [1, 76].

4.5 Comparision with published methods

It is of interest to compare the performance of the HMM with other methods for detecting archaic segments. In particular, I looked into the S^* method [90, 91] which searches for long haplotypes in linkage disequilibrium unseen in the African panel, and a conditional random field (CRF) [1] which examines allele sharing, haplotype divergence, and local recombination rate.

Since the implementation of neither methods is publicly available, I compared the result of running CRF, S^* and the informative-site-only HMM on the same set of genomes instead.

The Neanderthal segments detected by CRF and S^* in individuals from the 1000 Genomes Project were downloaded from the authors' websites. I then ran the HMM on chromosome 1 of 544 individuals that were included in both studies and obtained the Viterbi sequences. To be consistent with the other two methods, only the high-coverage Neanderthal genome from the Altai cave [66] was used in the archaic panel.

Figure 4.20 shows the total amount of Neanderthal segments detected by each method on chromosome 1 and the relationship between them in a Venn diagram. The highest agreement is between HMM and S^*, where the shared regions constitute 72.84% of the total material recovered by the HMM and 82.43% of that recovered by S^*. It is worth noting that although the S^* score itself does not rely on the archaic genome, the reported segments in [91] have undergone subsequent filtering on their match score to the archaic genomes. In comparison, other pairwise comparisons only show around 40% of reciprocal agreement. Such pattern is consistent across seven Eurasian populations analyzed.

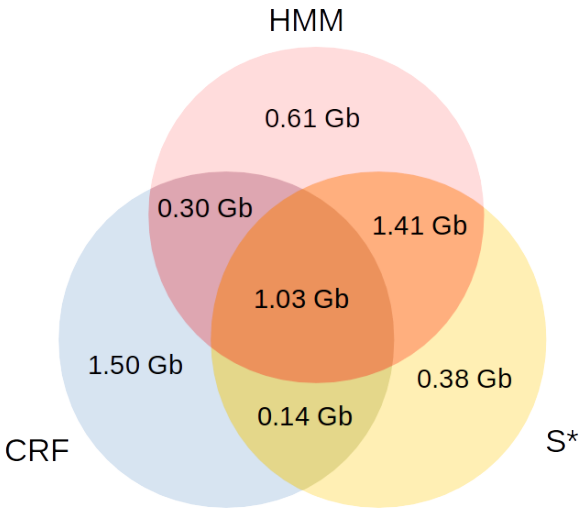


Fig. 4.20 Amount and relationship of Neanderthal segments from chromosome 1 of 544 genomes using three methods

Another criterion is to classify a segment detected by one method as a match if at least half of it is also reported by another method. Although the HMM does not preferentially detect or miss segments of particular lengths compared to the other two methods, I found that segments detected by the HMM but unreported by other methods tend to be shorter (4.21). Segments shorter than 50kB constitute 91.22% of those undetected by S^*, and 82.37% undetected by CRF. In other words, the HMM appears superior in detecting shorter introgressed segments than CRF and S^*. Perhaps shorter segments rarely show high divergence to African haplo-

types in CRF (the feature of haplotype divergence is removed from the CRF in a later study by the authors [76], on account that it is not appropriate when admixture proportion varies largely across samples), or produce high S^* scores, though a solid explanation would call for a simulation study. Most segments longer than 50kB are reported by all three methods. Figure 4.22 gives an example on the haplotype with the highest archaic fraction, HG00063.1. The broadscale pattern is shared across all methods.

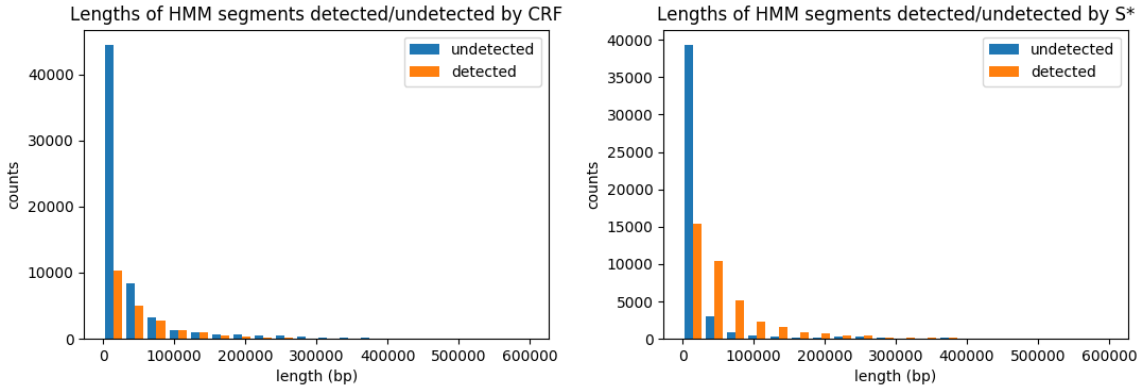


Fig. 4.21 Lengths of HMM result detected/undetected by other methods

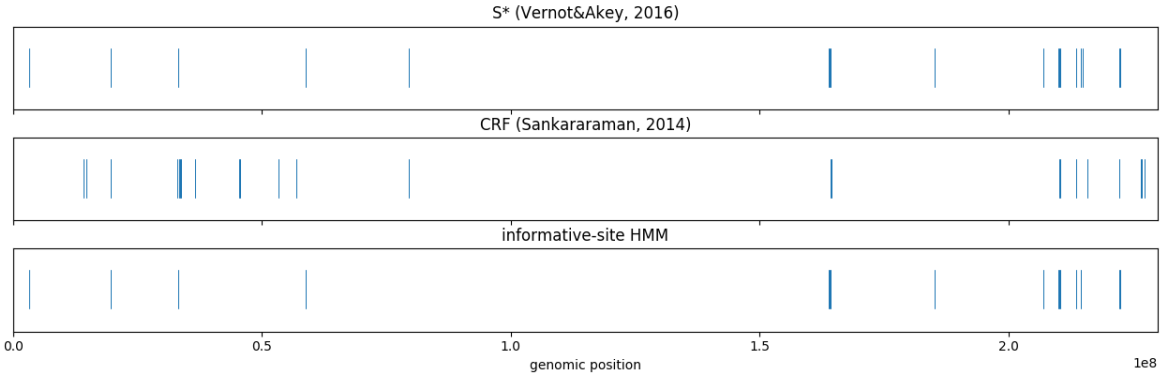


Fig. 4.22 Distribution of Neanderthal segments on HG00063, chromosome 1 using three methods

4.6 Reference-free HMM

At the time when I worked on the HMM described above, Laurits Skov also developed an HMM to detect archaic segments that treats the number of private SNPs in a genomic window with respect to an outgroup as the emission variable [95]. I eventually become involved

in testing it with simulations since the work flow was already in place for evaluating my own HMM. This model requires a specific demographic history, where the divergence time between the admixture source and the target population (the ingroup hereafter) is much deeper than the divergence time between the target population and an unadmixed outgroup, but does not require any genetic information about the admixture source. It is preferable to have a large recent population size in the outgroup, but a small size in the ancestral population right before the ingroup diverges from the outgroup. Thus most modern segments will coalesce quickly with some of the outgroup lineages once they are in the same population. Segments enriched for private SNPs are likely to have an archaic origin.

Here the model is mainly used to detect Denisovan segments in modern Papuan genomes, although the application is not limited to modern human. Oceanian populations are known to have a much higher level of Denisovan ancestry than populations from mainland Eurasia, suggesting that a major admixture happened after the population split [73]. Variants that arose in the ancestral human population or from Neanderthal gene flow can both be excluded by using all genomes from the 1000 Genomes Project as the outgroup. Therefore private variants in Papuan genomes are expected to result from admixture with either the Denisova or other archaic sources. Since the Altai Denisova genome is not a close proxy for the source of Denisovan gene flow [66], the reference-free method has the potential to recover more introgressed segments than methods that rely on an archaic reference genome.

Another advantage of the reference-free HMM is its interpretability. The emission and transition parameters, which can be estimated from genetic sequences using the Baum-Welch training algorithm, are informative about the demographic history. The Poisson mean of the number of mutations in each state corresponds to the average time till the first coalescence between lineages in the corresponding population and in the outgroup, hence establishing an upper bound for the population split time. For example, the mean number of private SNPs in the modern state can be expressed as:

$$\lambda_{\text{ingroup}} = \mu \cdot L \cdot t_{\text{ingroup}}$$

where μ is the mutation rate, and L the genomic window length. Thus the mean coalescent time between ingroup and outgroup lineages, t_{ingroup}, can be calculated from λ_{ingroup}. Similarly, the mean coalescent time between archaic and outgroup lineages can be estimated from the emission parameter in the archaic state.

The transition probabilities contain information about the time and amount of admixture. In the example of transitions between a modern (0) and an archaic (1) state,

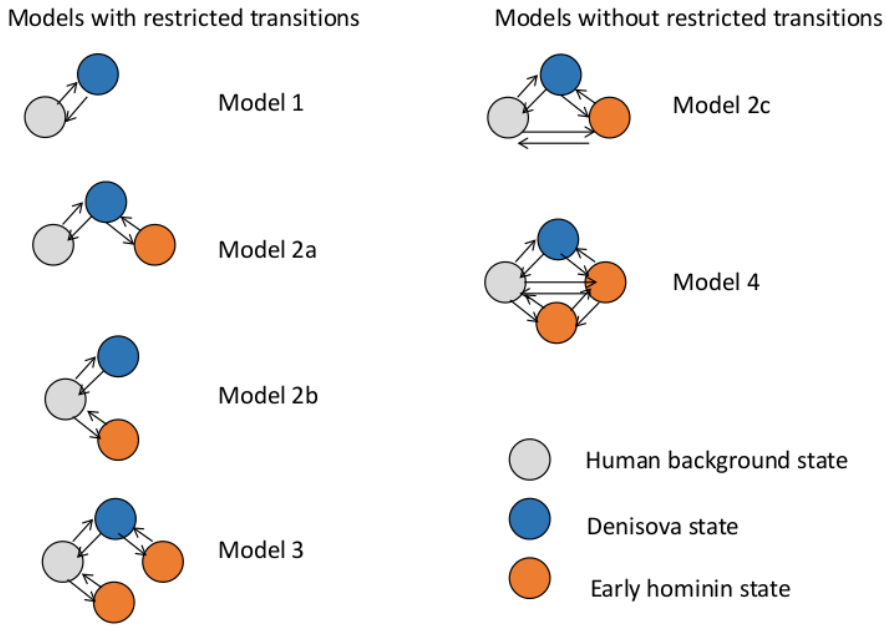
$$T_{01} = t_{\text{admix}} \cdot r \cdot L \cdot \alpha$$
$$T_{10} = t_{\text{admix}} \cdot r \cdot L \cdot (1 - \alpha)$$

where r is the recombination rate between adjacent genomic windows, t_{admix} the time since admixture, and α the admixture proportion. [95] evaluates the accuracy and robustness of demographic inference from emission and transition parameters in details. The following section describes my contribution to evaluating this reference-free HMM.

4.6.1 Simulation studies

Initially three types of possible hidden states are considered, a modern human state, a Denisova-related state, and one or two early hominin state(s). A previous study has proposed the presence of segments from a deeply diverged hominin lineage, possibly *Homo erectus*, in the Denisova genome [66]. Therefore a range of models with various topologies were proposed (Figure 4.23). Model 1 only includes an unadmixed human background state and a Denisova state, with bidirectional transitions between them. Model 2 adds an early hominin state, but transitions between this state and the modern human state is not allowed in Model 2a, implying that the early hominin segments are always embedded within the Denisovan segments. Similarly, Model 2b does not allow direct transitions between the early hominin state and the Denisovan segments, implying gene flow from an early hominin into the modern human population. Model 2c allows transitions between all pairs of states. Model 3 assumes two different early hominin states that correspond to introgression into the modern human and the Denisovan populations, respectively. Finally, Model 4 allows for transitions between all pairs among the Denisovan state, the modern human state, and two different early hominin states. These models have been trained on 14 Papuan genomes from the HGDP dataset, and a comparison of their likelihoods favours a three-state model that allows bi-directional transitions between the modern human and Denisova states and between the Denisova and early hominin states, but not direct transitions between modern human and early hominin states [157].

To verify the performance of the HMM and in particular, whether the assignment of the early hominin state is genuine, I simulated genetic sequences under three scenarios: where Oceanians did not receive any archaic gene flow (no gene flow), where Oceanians received



(Model 2a and 2b are topologically the same; they only differ in initial parameter values)

Fig. 4.23 Various models tested on Papuan genomes (from Laurits Skov)

gene flow from a Denisovan population (one gene flow), and where the Denisovan population contributing the gene flow into modern humans also received gene flow from an early hominin population (two gene flows). The demographic model used in coalescent simulations is shown in Figure 4.24, with corresponding changes in the presence or absence of gene flows. In some simulations, the accessibility mask from the 1000 Genomes Project and/or the genetic map were applied to incorporate missing data and varying recombination rate. Variations in local mutation rate were also tested by concatenating observations from separate simulations.

Figure 4.25 shows the log-likelihood of models with 1-3 hidden states on simulated sequences, with panels comparing the presence/absence of missing data and recombination rate variations. The parameters were estimated using Baum-Welch training on individual genomes. The HMM can detect the presence of archaic admixture reasonably well: with no archaic gene flow, adding one additional state on top of the null model (Model 0) does not improve the log-likelihood noticeably; whilst when the Denisovan gene flow is present (one/two gene flow), the likelihood increases substantially from Model 0 to Model 1. However, the HMM is less capable at detecting whether the Denisova received gene flow from an earlier-diverging archaic source. The addition of a third state only marginally increased the likelihood in scenarios with both one and two gene flows. Furthermore, the fitted emission probabilities

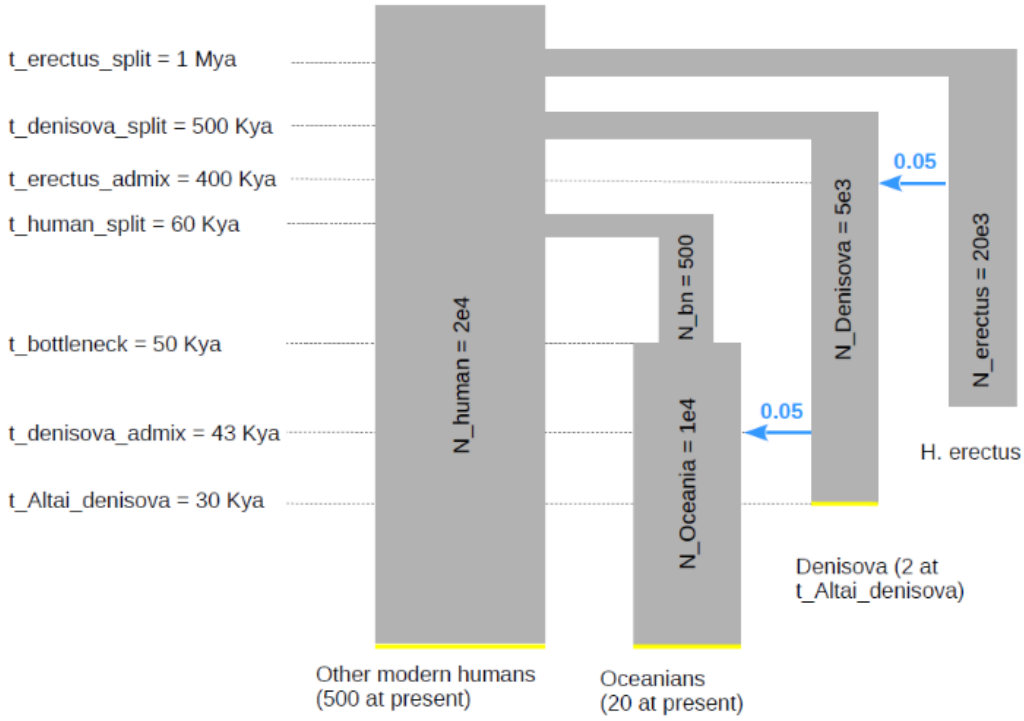
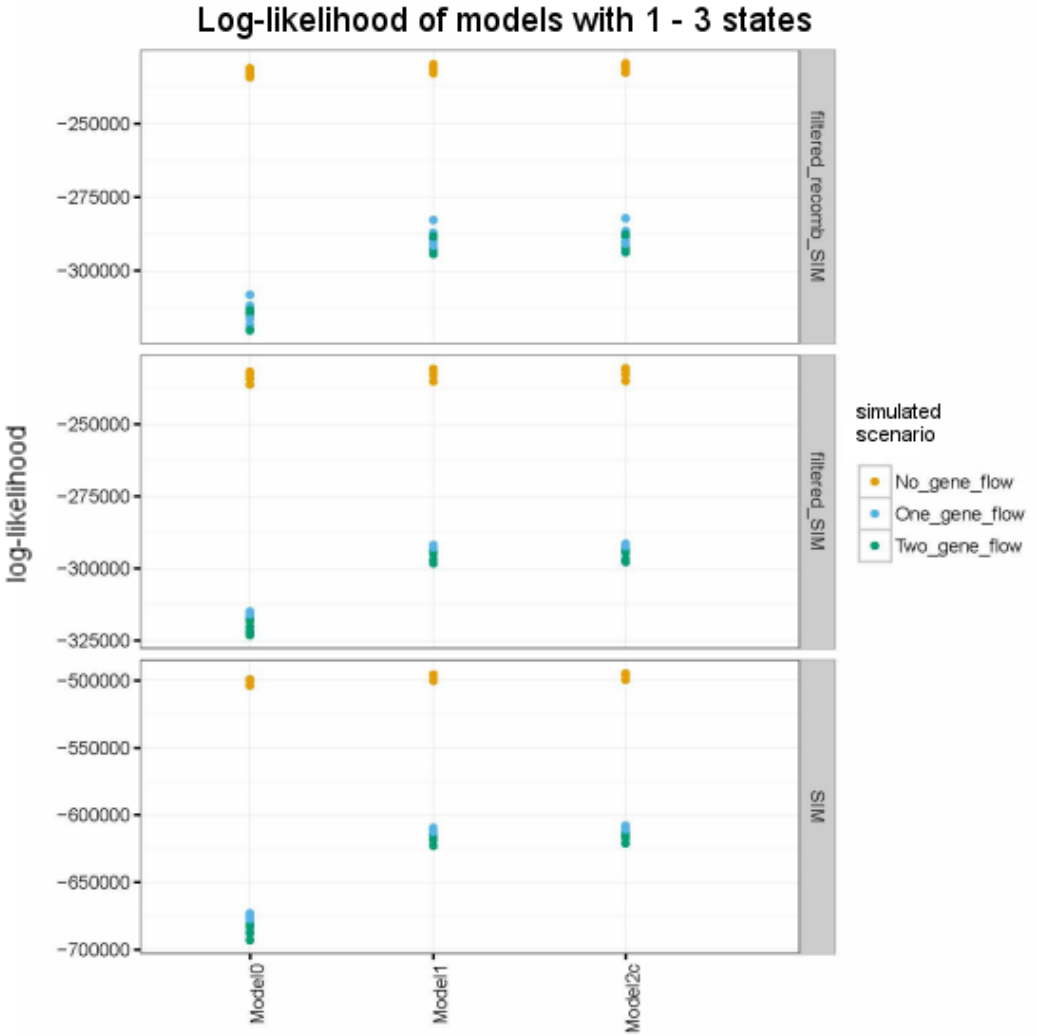


Fig. 4.24 Demographic model used in coalescent simulations for the reference-free HMM

in the added state did not describe an early hominin state; instead, segments in the previous modern human state became split into two states, one with an increased Poisson mean and the other a decreased one in comparison to the modern human state in a two-state model.



(Model 0 is the null model with only one state)

Fig. 4.25 Log-likelihood of models from Figure 4.23 fitted to simulations with different gene flow scenarios (from Laurits Skov)

The reason is most likely due to some deep coalescing modern human lineages. Even within the same population, the time until coalescence for two lineages is geometrically distributed, therefore such deep coalescence is bound to occur in some modern segments especially when the ancestral population size is large. Indeed, when I examined the true time until the first coalescence with sampled outgroup lineages from the simulated trees, the segments that

become classified into the two split modern human states show distinct trends (Figure 4.26). On the other hand, in our two gene flow simulation where the admixture proportion is set at 5%, genuine segments from early hominin only constitute an average 0.25% of the modern human genome. The HMM probably has limited capacity to distinguish it from the much more abundant noise of deep coalescing modern segments.

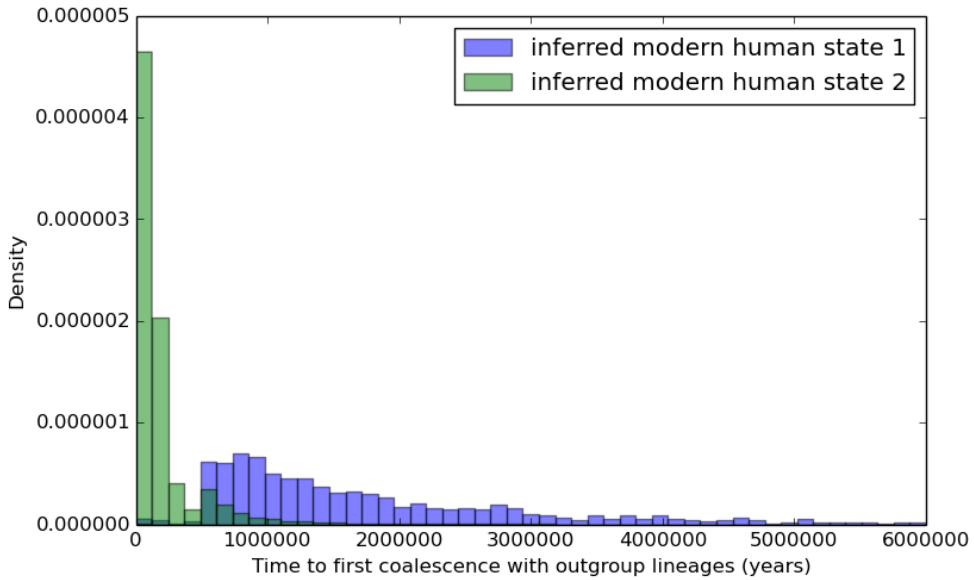


Fig. 4.26 Normalised histogram of the time till the first coalescence with outgroup lineages for segments in two modern human states, one gene flow scenario

In conclusion, the HMM can correctly infer the presence of an archaic state with reference to the change of likelihood between one-state and two-state models, but Baum-Welch training alone is not able to recover a possible earlier state in the genome, at least not at the admixture amount relevant in this example. Therefore the three-state model being favoured in model selection should not be interpreted as evidence supporting an early hominin state in the Denisovan genome.

4.6.2 Comparison with S^* score

Other methods have also been developed to detect archaic introgression without using an archaic reference genome. Most of them make use of the S^* statistic [90, 91, 93, 94], which screens for long haplotypes in linkage disequilibrium that are different from other haplotypes in the population and absent in a control panel. Vernot and Akey have presented two versions

of S^* for detecting specific archaic segments: the 2014 version operates on 20 individuals from one population [90], whilst the 2016 version operates on one individual at a time [91]. For a region in an individual genome, in the 2014 version, S^* is defined as the maximum score obtained by a subset of variants V in this region,

$$S^* = \max_{J \subseteq V} S_{2014}(J)$$

$$S_{2014}(J) = \sum_{j \in J} \begin{cases} -\infty, & d(j, j+1) > 5 \\ \text{-10000}, & d(j, j+1) \in 1 \ldots 5 \\ 5000 + bp(j, j+1), & d(j, j+1) == 0 \\ 0, & j = max(J) \end{cases}$$

where j and $j+1$ are adjacent variants, $d(j, j+1)$ is the sum of genotype distance between them in all individuals when the genotypes are encoded as 0, 1, and 2, and $bp(j, j+1)$ is their physical distance in base pairs.

Similarly in the 2016 version,

$$S^* = \max_{J \subseteq V} S_{2016}(J)$$

$$S_{2016}(J) = \sum_{j \in J} \begin{cases} \text{-10000}, & d(j, j+1) > 0 \\ 5000 + bp(j, j+1), & d(j, j+1) == 0 \\ 0, & j = max(J) \end{cases}$$

Since only one individual is considered, the genotype distance $d(j, j+1)$ will only take values 0, 1 and 2.

In both 2014 and 2016 versions, the authors fit a generalized linear model on coalescent simulation data to predict the null distribution of S^* under different demographic models, recombination rates, and variants numbers. They also compared the putative segments to archaic genomes to obtain match p-values, which were used to further filter the segments (2014 version) and distinguish between Neanderthal and Denisovan segments (2016). They also present a bound-constraint SVM to filter the segments when no archaic reference genome is available, which recovers around 30% of the sequence identified using the match p-values.

I would like to limit the discussion to reference-free methods. Because the behavior of the statistic is of more interest than a specific method, I only compared the two versions of S^* score to the posterior probabilities in the HMM. Figure 4.27 shows the result on a section of a simulated chromosome. S^* scores of 50kB windows with a step size of 20kB were

obtained. In the 2014 version, the population consists of 20 simulated haplotypes. In general, all three panels capture the longest true archaic segment, whereas HMM also clearly suggests the presence of a shorter one, albeit at a lower probability that is likely to be filtered out subsequently. S^*_{2016} appears more nuanced possibly because that unlike S^*_{2014}, when only one individual rather than a population is considered, the genotype distance does not track linkage disequilibrium accurately. The shorter segments on haplotype 0 are not captured by any of the three methods. Although subsequent statistical learning can surely improve the performance of S^*, the HMM produces more accurate and transparent result at this stage. Interestingly, in another recently published reference-free method, the authors report that the minimum distance between the focal and the reference haplotypes and the number of private SNPs to be most indicative features in their logistic regression model, whilst the weight of S^* score is very small [94]. S^* score is correlated with the density of private SNPs, which also indirectly affects the haplotype distances; but other criteria in calculating S^* do not add much new information.

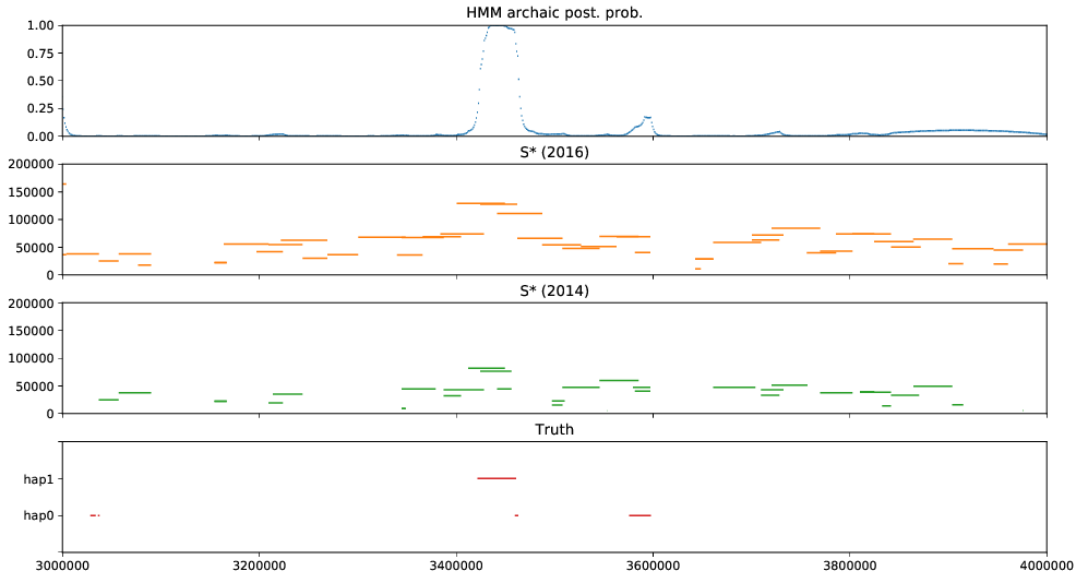


Fig. 4.27 Comparision of posterior probabilities in HMM and two versions of S^* scores on simulated sequence

4.7 Conclusion

The sequential architecture of the HMM makes it a natural choice for labelling genetic segments. In this chapter, I presented a HMM for detecting archaic introgression in modern

genomes based on allele sharing with the archaic genome and with an unadmixed modern outgroup. Model parameters can be learned from tagged simulation, using Baum-Welch training, or using numerical optimization of the likelihood. Although the latter two does not always converge towards parameter values consistent with the admixture history, they can be used to fine-tune the parameters when the demographic model used in simulations deviate slightly from the true population history. A HMM that only processes informative sites out-performs one that takes each site as a time step.

Because of the genetic similarity between the Neanderthal and the Denisovan populations, a three-state HMM with unadmixed, Neanderthal and Denisovan states has a high probability of labelling segments from one archaic source as the other. Instead, I showed that the two-state model can be run twice with the Neanderthal and Denisovan genomes in the archaic panel, respectively. The posterior probabilities produced in both runs can be used to assign archaic segments into Neanderthal, Denisovan or ambiguous origins.

I also explored ways to date the admixture events based on the distribution of archaic segment lengths, however the difficulty to distinguish short Neanderthal and Denisovan segments - especially when their amounts are unequal - makes it biased towards assigning an older date to the Denisovan gene flow. This might be a caveat for other studies using HMM and related graphical models as well.

The comparison with other methods that have been applied to detect archaic introgression in human genomes shows that the HMM has less bias towards missing out shorter segments. Compared to methods based on S^*, output from the HMM does not require further refining, which often involves a less transparent statistical learning process [90, 91]. The CRF is structurally very similar to the HMM, with the flexibility to incorporate an arbitrary number of observed features; nevertheless the HMM accesses the same set of features as the current setup of the CRF [76].

Finally I described my work with Laurits Skov on a reference-free HMM to detect archaic segments from the density of private alleles. This is another example that demonstrates the simplicity and interpretability of HMMs applied to genetic sequences.

Chapter 5

Archaic segments in HGDP genomes

5.1 Surveying archaic segments in diverse human populations

The HGDP dataset is the first high-coverage sequencing panel to represent worldwide ethnic groups, many of which are relatively small and isolated, at a decent sample size to capture genetic diversity both between and within the populations. Among a range of questions about demographic history that can be addressed with this dataset, the diversity of genetic segments from archaic hominins and its implication on the admixture process is particularly exciting: did East Asians and Europeans receive Neanderthal gene flows from separate sources? Were there more than one pulse of Denisovan or Neanderthal admixture? Does any region harbour unique archaic haplotypes that are not seen anywhere else? How many Neanderthal or Denisovan individuals are likely to have been involved in the admixture? Some of these questions have been addressed by previous studies, but their power is limited by only examining the largest populations [1, 91, 93] or mainly focusing on a certain geographical region [158]. High-coverage sequencing data also makes it possible to explore the full diversity between archaic haplotypes within and between populations.

Therefore in this chapter, I searched for Neanderthal and Denisovan segments in the HGDP dataset using the HMM described in Chapter 4, and explored their features including segment lengths, divergence to the archaic genomes, and the structure of diversity around the world. Differences between archaic segments recovered from various geographical regions would support different admixture histories; distinct components found within the same population would suggest that the admixture happened more than once. I also made the first attempt to

estimate the number of unique Neanderthal lineages introgressed into the modern human population by building a genealogy of the Neanderthal haplotypes found in modern humans.

5.2 Running the HMM on HGDP dataset

The two-state HMM as described in 4.3.2 was run twice on 929 phased genomes from the HGDP dataset with the genetic map to obtain the posterior probabilities of being in the Neanderthal and Denisovan state at all informative sites. All 104 genomes from sub-Saharan Africa were included in the African panel, but when extracting observations, I allowed the archaic allele to reach a maximum frequency of 0.01 (namely at most two copies) in this panel to account for possible back-migration into Africa. A previous study detected 2-3% of Eurasian ancestry in Mandenka and Kenya Bantu populations [130], which constitute 33 out of 104 sub-Saharan African genomes in the HGDP panel. Two high-coverage Neanderthal genomes, one from Denisova Cave [66] and the other from Vindija Cave [67], were used in the archaic panel in the Neanderthal run; whilst the Altai Denisova genome [47] was used in the Denisova run. All archaic genomes along with their respective filters were downloaded from the web server of the Max Planck Institute for Evolutionary Anthropology and lifted over from GRCh37 to GRCh38. In addition to the strict mask on modern genomes, I added a mask on low complexity regions [159]. All sites that do not pass the masks were ignored as non-informative in the HMM runs. I referred to the ancestral sequences from EPO alignment of 6 primates (obtained from http://www.ensembl.org/info/genome/compara/ancestral_sequences.html) to determine the ancestral state; and in case of unknown sites in this panel, assumed the genotype in chimpanzee (Pan_tro 3.0) to be ancestral. To keep the files at manageable sizes, only sites that are polymorphic in the HGDP dataset were retained after merging. In effect, this leaves out derived sites shared by all modern human genomes but not in the archaics, which are type 2 emission (Table 4.2) that supports assigning the modern state. Such sites should be very rare in the genome, as derived alleles shared by all modern humans will be over 200k years old.

In order to facilitate different types of downstream analysis, several different criteria have been employed to classify segments into Neanderthal, Denisovan and ambiguous origins from the posterior probabilities at informative sites. As described in 4.3.2, the posterior probabilities of being in the archaic state during the Neanderthal (p_N) and the Denisova (p_D) runs are compared at each informative site. If a site only shows up as informative regarding one archaic source but not the other, the absent posterior probability is estimated linearly from its adjacent sites. In the first set of results, Neanderthal segments span over sites where

$p_N > 0.5$ and $p_N > p_D$; similarly $p_D > 0.5$ and $p_D > p_N$ for Denisovan segments. However, if p_N and p_D both exceed 0.8, the segment is tagged as ambiguous. Results following these criteria ("basic" hereafter) are used to estimate the amount of archaic ancestry in modern genomes.

In analyses involving the genomic location and genotypes of archaic segments, a more stringent set of criteria is applied to better discriminate the archaic origin: the Neanderthal segments are required to have $p_N > 0.8$ and $p_D < 0.5$; similar rules apply for the Denisovan segments. This set of results ("strict" hereafter) also includes a collection of confident modern segments, where p_N and p_D are both lower than 0.1. The accessibility and low complexity region masks were applied on top in some analyses.

Lastly, the "separate" and "no-overlapping" sets of results are intended for comparing the lengths of archaic segments. Here the Neanderthal and Denisovan segments were first recovered from the respective posterior probabilities separately (the "separate" set), but those overlapping longer than 500 bp with any segments from the other archaic source are removed afterwards in the "no-overlapping" set.

The performance of some assigning criteria has been evaluated in Table 4.10, assuming an admixture proportion of 0.02 from the Neanderthal and 0.01 from the Denisova. In practice, the "basic" set assigns a lower proportion of segments to the ambiguous category when run on the HGDP genomes than in simulation studies. Perhaps this is because the Vindija Neanderthal genome is closer to the Neanderthal source of gene flow compared to the simulated scenario, allowing for better distinguishing power.

5.3 Geographical distribution of archaic ancestry

5.3.1 Amount of Neanderthal and Denisovan ancestry

Figure 5.1 compares the average amount of Neanderthal, Denisovan and ambiguous ancestry in the HGDP genomes (from the "basic" set of result) by geographical regions; Figure 5.2 and Figure 5.3 show the mean and standard deviation of the amount of Neanderthal and Denisovan ancestry in each population. The length of masked regions was excluded in all plots.

Very few archaic segments are detected in sub-Saharan African populations (Figure 5.1). This is consistent with the model set-up where sub-Saharan Africa serves as an archaic-free control panel, except for very few haplotypes that could have entered via back-migration

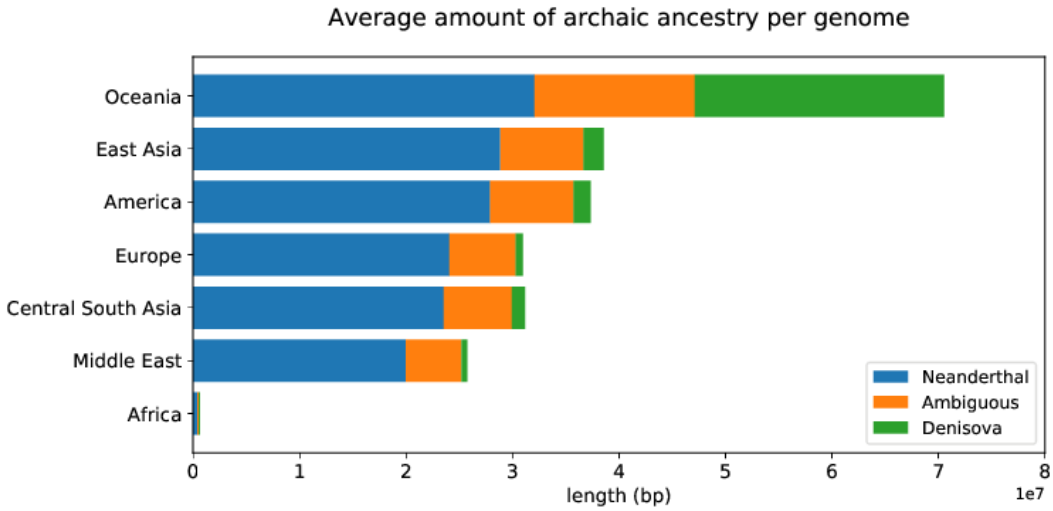


Fig. 5.1 Average amount of archaic ancestry per genome by geographical regions

from Eurasia. In accordance with previous studies, the amount of Neanderthal ancestry is higher in East Asia and America than in Europe and the Middle East. Subsequent gene flow from a "basal Eurasian" source, which is an outgroup to all non-African populations today and contains very little to no Neanderthal ancestry, has been proposed to explain a lower level of Neanderthal ancestry in Europe and the Middle East[101, 102], although other researchers attribute it to different strengths of negative selection [1] or an additional episode of gene flow from Neanderthal into East Asia [100, 104]. The highest amount of Neanderthal ancestry is found in Oceania, but it could be an artefact caused by some misclassified Denisovan segments. No prominent differences are observed between populations within the same geographic regions (Figure 5.2). The intra-population variance is higher in Middle Eastern populations (especially Mozabite and Bedouin), possibly reflecting recent admixture between sources with different levels of Neanderthal ancestry.

Denisovan segments are most abundant in Oceania, represented by three Papuan populations (Figure 5.1). It is also detectable at a much-reduced level in East Asia, America, and Central and South Asia, but negligible in Europe and the Middle East. Within Oceania, the population from Bougainville Island contains less Denisovan ancestry than the other two populations from New Guinea Island (Figure 5.3). Similar to the case with Neanderthal ancestry, I do not find any population to deviate from the broad regional pattern.

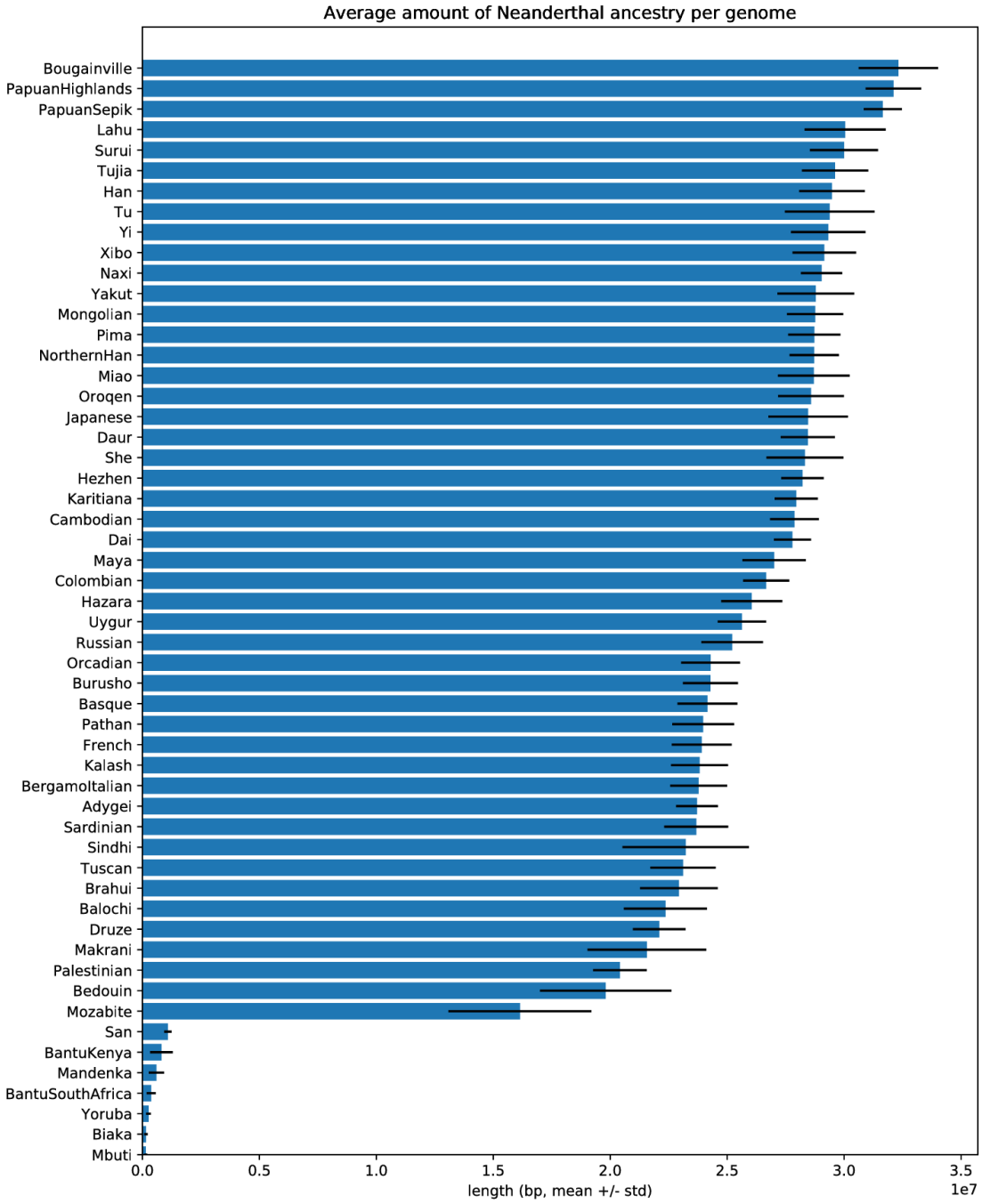


Fig. 5.2 Average amount of Neanderthal ancestry per genome by populations

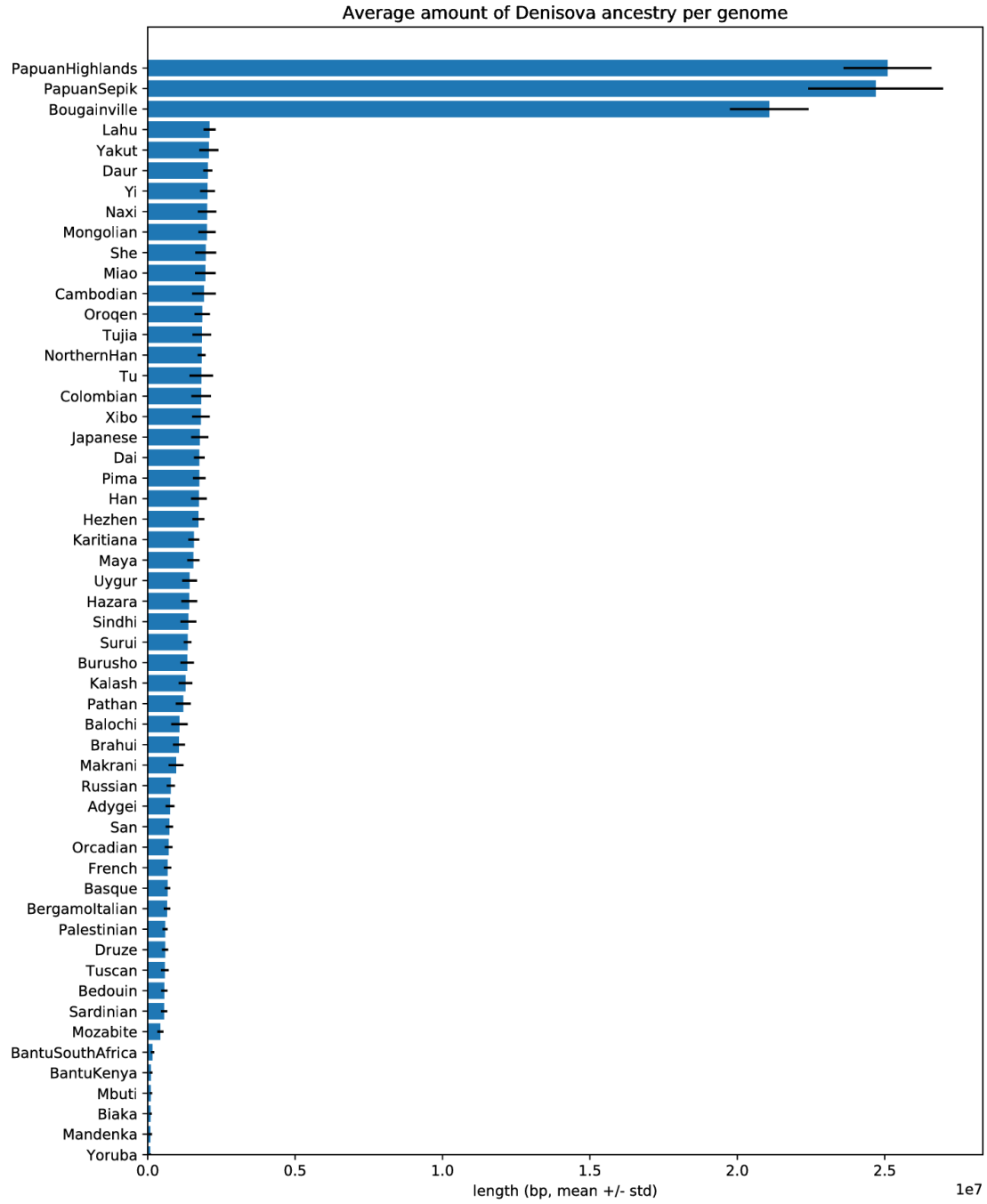


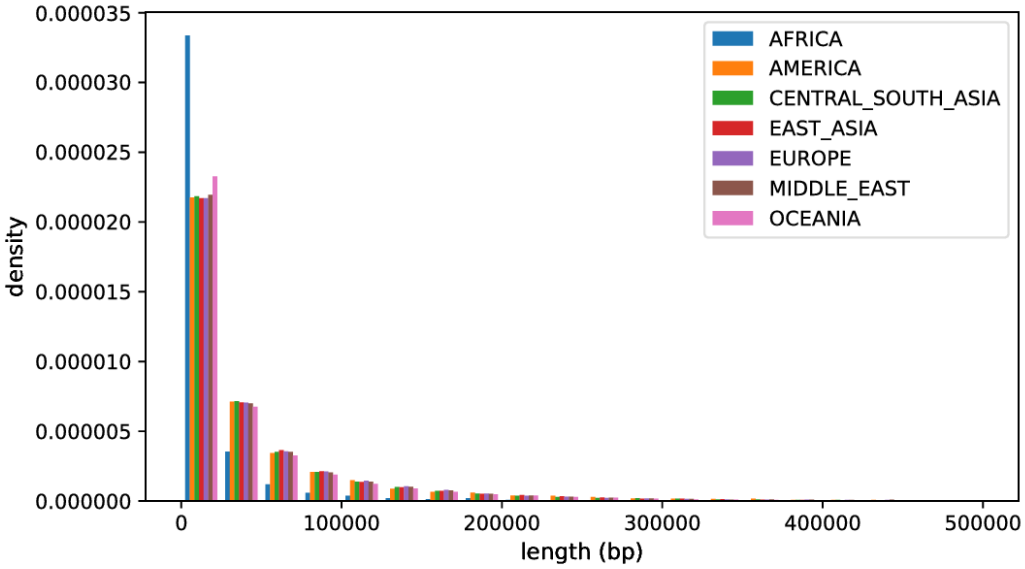
Fig. 5.3 Average amount of Denisovan ancestry per genome by populations

5.3.2 Lengths of Neanderthal and Denisovan segments

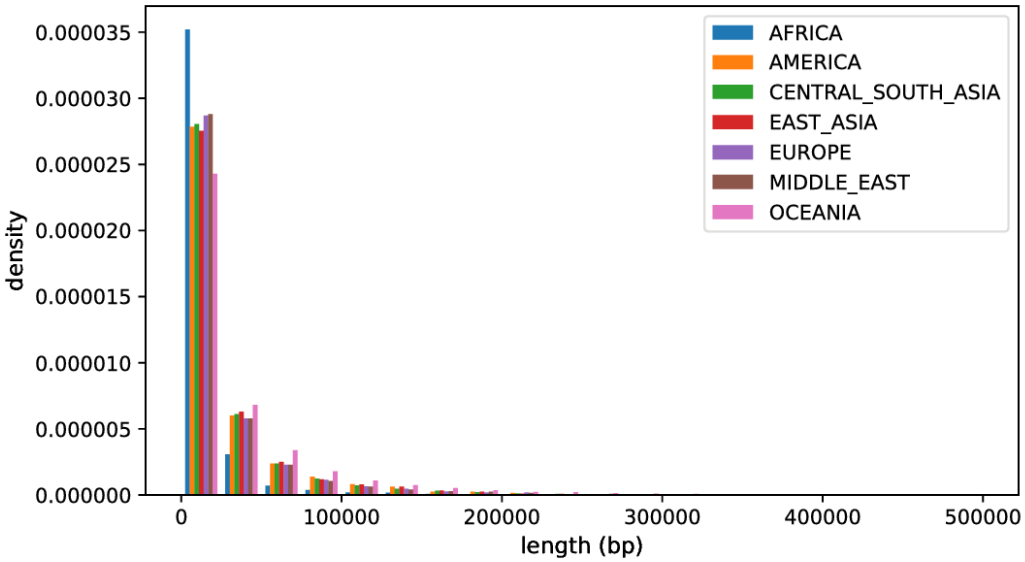
I compared the distribution of the lengths of introgressed segments across populations. Figure 5.4 shows the normalised distribution of Neanderthal and Denisovan segments as detected in separate runs without access to the other archaic genome (the "separate" set of results). Most archaic segments detected in Africa are short, which can be either very old segments from incomplete lineage sorting or false hits due to sequencing/mapping error. The total amount of such segments per genome is less than 0.42 Mb for Neanderthal segments and less than 0.14 Mb for Denisovan segments (Figure 5.1). Regarding Neanderthal segments, the length distribution is similar across all non-African populations, except for a slight excess of shorter segments in Oceania (Figure 5.4a). With Denisovan segments, there are more variations between regions, yet the widest difference is still between Oceania and the other non-African regions: Denisovan segments in Oceania appear longer than those outside of Oceania (Figure 5.4b).

It is worth noting that the difference in the relative amount of Denisovan ancestry compared to Neanderthal ancestry, which leads to different probabilities of misclassifying archaic segments around the world, is likely to bias the distributions. Since Denisovan segments are relatively rare outside of Oceania, a large proportion of the segments detected in non-Oceanians by running the HMM with the Denisovan genome will actually have a Neanderthal origin. The admixture with the Denisovan has been estimated to have happened more recently than the admixture with Neanderthal, with longer Denisovan fragments than Neanderthal fragments in the Oceanian genomes [76]. Indeed, in Figure 5.4b, the more Denisovan ancestry a region contains, the longer the detected Denisovan segments appear.

In theory, the lengths are informative about the admixture time, but simulation studies show that the estimates are also affected when the amount of contribution from various sources differs considerably (Section 4.4.2). Therefore although the Denisovan segments appear in general shorter than the Neanderthal segments, it might merely reflect that shorter true Neanderthal segments containing fewer distinguishing alleles are more likely to be incorrectly classified as Denisovan, rather than that the Denisovan segments entered earlier into modern human population. The effect should be more pronounced in regions where the true genetic contribution from Denisova is lower, as true Neanderthal segments that are mistaken as Denisovan ones dominate in number over true Denisovan segments. The average length of Neanderthal and Denisovan segments in Oceania, where the amount of ancestry from both sources is more similar (Figure 5.1), also appears more comparable to each other.

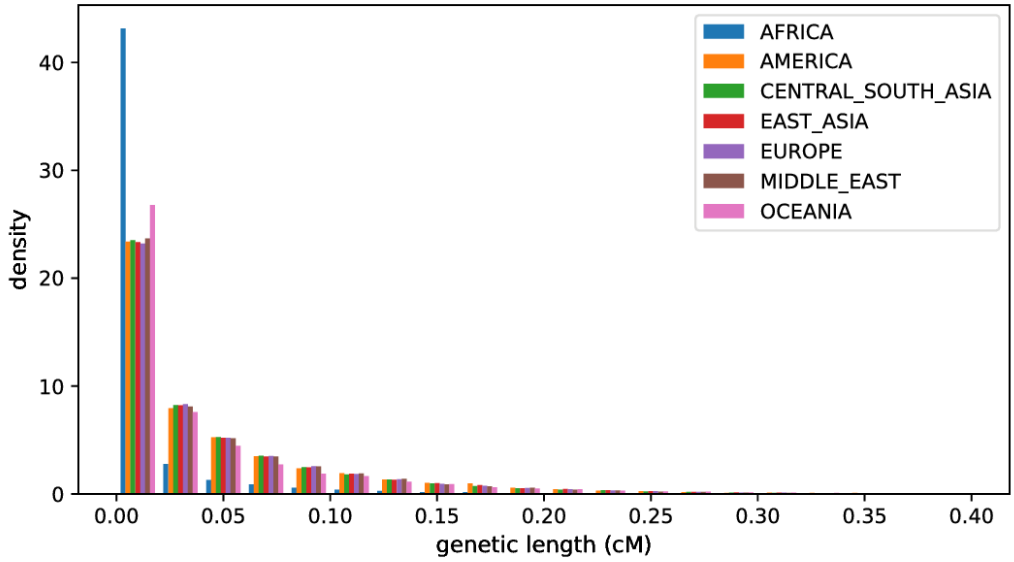


(a) Physical lengths of Neanderthal segments

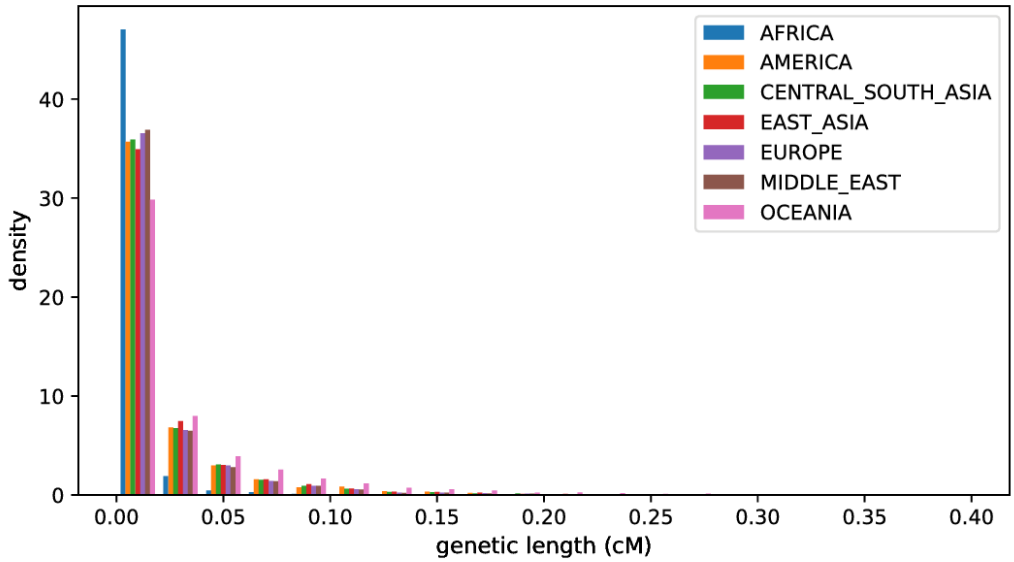


(b) Physical lengths of Denisovan segments

Fig. 5.4 Distribution of the physical lengths of introgressed segments by geographical regions ("separate")

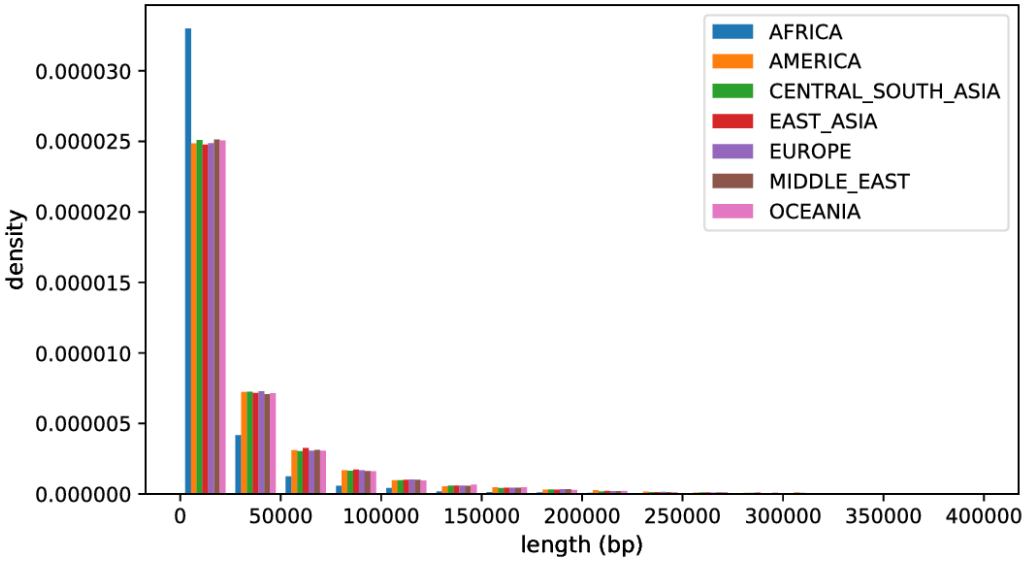


(a) Genetic lengths of Neanderthal segments

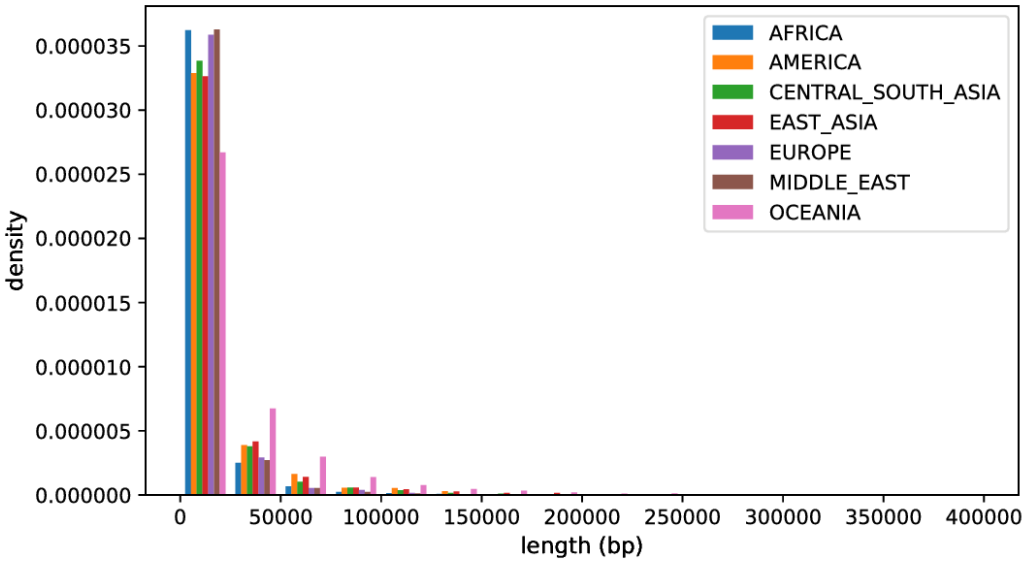


(b) Genetic lengths of Denisovan segments

Fig. 5.5 Distribution of the genetic lengths of introgressed segments by geographical regions ("separate")



(a) Physical lengths of Neanderthal segments



(b) Physical lengths of Denisovan segments

Fig. 5.6 Distribution of the physical lengths of introgressed segments by geographical regions ("no-overlapping")

Converting physical distances to genetic distances according to the genetic map did not change the above observations (Figure 5.5). Limiting the segments to those that do not overlap with any from the other source ("no-overlapping" set) also reproduced the same pattern (Figure 5.6). Because the lengths of Neanderthal segments are similar across mainland Eurasia and America, I do not find support for localised encounters with Neanderthals occurring at detectably different times in history.

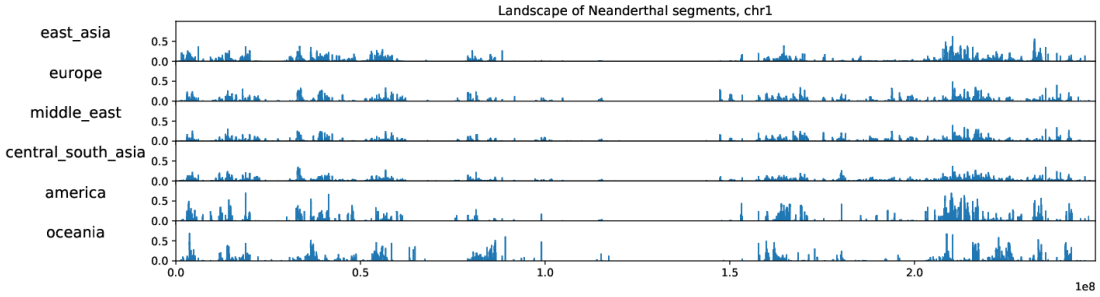
5.4 Genomic distribution of archaic segments

Previous studies have identified genomic hotspots and coldspots for archaic ancestry. Functional regions enriched for Neanderthal or Denisova alleles might have helped modern humans adapt to the environment in Eurasia; whilst regions enriched in genes in general contain reduced archaic ancestry, suggesting widespread selection against archaic introgression [1, 76, 90, 91]. Here I explore the distribution of archaic ancestry in the genome and its variation across geographical regions.

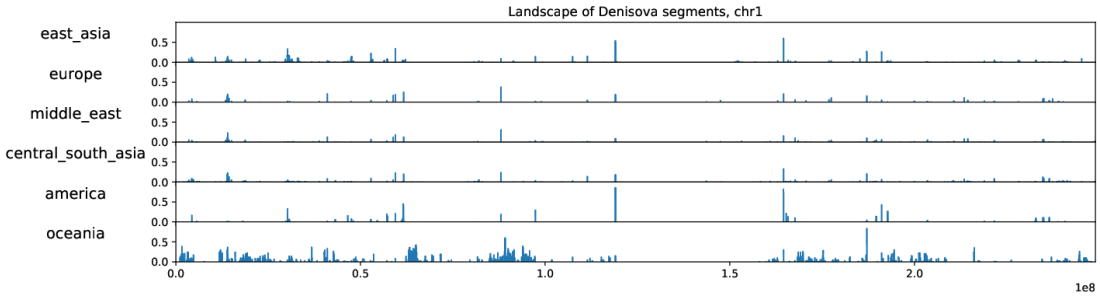
5.4.1 Variation across geographical regions

All archaic segments tagged according to the "strict" criteria are pooled by geographical regions to obtain the frequency of archaic ancestry along the genome. Figure 5.7 depicts the distribution of Neanderthal and Denisovan segments along chromosome 1 as an example.

We can observe that Neanderthal segments are distributed across the genome similarly in all non-African regions. The similarity appears higher between East Asia and America and between Europe and the Middle East, in agreement with the demographic history of modern human populations. America shows the wildest fluctuations in the frequency of Neanderthal segments, possibly due to a strong founder effect associated with the colonization of the Americas. Although the distribution in Oceania deviates from the other regions to a certain degree, the difference lies more in frequency rather than the presence/absence of archaic segments. Considering a divergence time of ~58k years [98], it is possible that random drift could account for the difference between Oceanian and Eurasian populations. To formally test whether separate episodes of gene flow are required to generate the observed regional difference would require more detailed simulations and model selection, and would be very contingent on demographic history. Such an attempt has been made in [104], where the authors favour multiple episodes of Neanderthal gene flow into both Europe and East Asia through fitting the joint fragment frequency spectrum of introgressed Neanderthal segments.



(a) Distribution of Neanderthal segments along chromosome 1



(b) Distribution of Denisovan segments along chromosome 1

Fig. 5.7 Distribution of archaic segments ("strict") along chromosome 1 by geographical region

The difference is more pronounced in the plot of Denisovan segments (Figure 5.7b). The amount of Denisovan segments is small outside Oceania, nevertheless similar peaks and gaps can be found across these regions; in contrast, many genomic regions enriched for Denisovan segments in Oceania do not show up in Eurasia or America at all, and vice versa.

Since it is impossible to distinguish close genomic coordinates by eye, Table 5.1 summarises the length of overlapping genomic regions throughout the genome covered by at least two archaic segments between pairs of geographical regions, regardless of the genotypes in the segments. Here I use $P(A|B)$ to denote the probability that a genomic region also shows up in geographical region A, conditioned on it showing up in geographical region B. The geographical structure appears stronger in the genomic distribution of Denisovan segments. Even if we ignore Europe and the Middle East, where Denisovan ancestry is scarce, there is less overlapping between America, East Asia, Central/South Asia and Oceania. Denisovan segments might be lost through genetic drift more easily than Neanderthal segments, considering their low frequency in most populations; thus it is not surprising that $P(\text{non-Oceania}|\text{Oceania})$ is much lower regarding Denisovan segments than Neanderthal segments. However, despite similar amount of Neanderthal and Denisovan ancestry in Ocea-

nia, $P(\text{Oceania}|\text{non-Oceania}$ is also lower regarding Denisovan segments than Neanderthal regions: for example, $P(\text{Oceania}|\text{East Asia})$ is 0.1878 for Denisovan regions, and 0.3000 for Neanderthal regions. Since the evidence so far is consistent with a single admixture event introducing the Neanderthal segments, the fact that Denisovan segments are less likely to be shared between Asia and Oceania suggests a possible second source (or even multiple sources) for Denisova admixture. Even if more than one episode of admixture indeed happened between modern human and Neanderthals, the process involving Denisova admixture should be more complicated and structured geographically.

Table 5.1 Intersection of genomic regions covered by at least two archaic segments between non-African populations, expressed as the probability to find a genomic region in the column label conditioned on finding it in the row label

Neanderthal

Geographic region	**Total length (Mb)**	**Conditional probability to be also found in**					
		America	CS Asia	E Asia	Europe	Middle East	Oceania
America	204.89	-	0.9117	0.9105	0.7235	0.6155	0.3579
CS Asia	671.83	0.2780	-	0.6099	0.6458	0.6056	0.2409
E Asia	525.81	0.3548	0.7793	-	0.5513	0.4900	0.3000
Europe	482.25	0.3074	0.8997	0.6011	-	0.7981	0.2399
Middle East	453.09	0.2783	0.8979	0.5687	0.8495	-	0.2261
Oceania	218.29	0.3359	0.7413	0.7226	0.5300	0.4693	-

Denisova

Geographic region	**Total length (Mb)**	**Conditional probability to be also found in**					
		America	CS Asia	E Asia	Europe	Middle East	Oceania
America	13.33	-	0.5761	0.8382	0.3202	0.2470	0.2309
CS Asia	55.28	0.1389	-	0.3665	0.2106	0.1773	0.1980
E Asia	56.28	0.1986	0.3600	-	0.1330	0.0852	0.1878
Europe	14.07	0.3035	0.8276	0.5321	-	0.6230	0.2146
Middle East	13.89	0.2370	0.7054	0.3451	0.6308	-	0.2091
Oceania	190.23	0.0162	0.0575	0.0556	0.0159	0.0153	-

5.4.2 Negative selection against archaic segments

Figure 5.7 also illustrates the uneven distribution of archaic ancestry in the modern human genome with respect to genomic position. In particular, previous studies find a depletion of Neanderthal and Denisovan ancestry in gene-rich regions as evidence for widespread incompatibility between archaic and modern human haplotypes[1, 76], or heavier genetic load in the Neanderthal population due to small effective population size [112]. I tested if the

same result can be reproduced in Neanderthal segments in the HGDP dataset. The genetic sequences used in this section is an earlier internal release (v0.3) with slightly different filtering and phasing from the final release, which are not expected to affect the results.

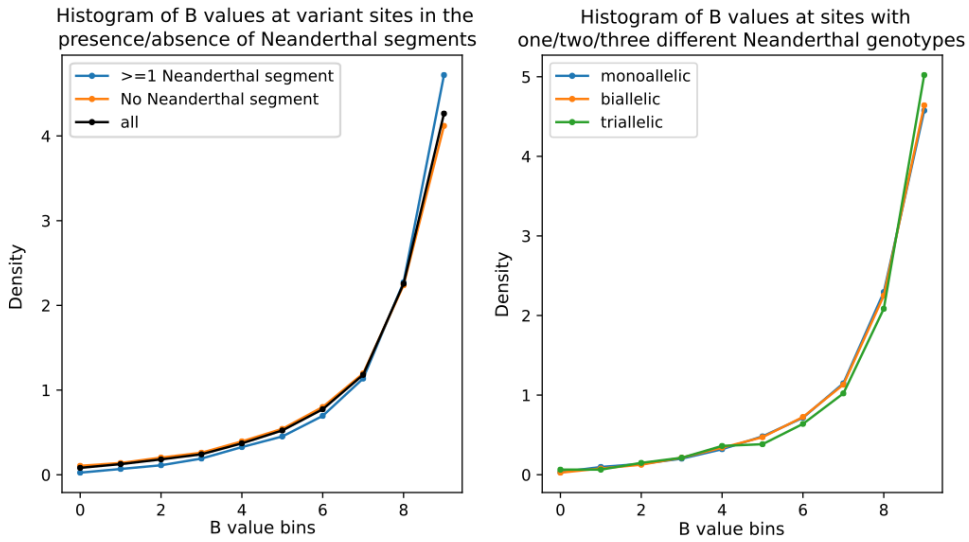
Most previous studies used B values to evaluate selection against introgressed segments [47, 1, 76, 68, 91]. Calculated from the fraction of neutral variations present at each site in the genome, B values serve as a measurement for the strength of linked selection (including both selective sweeps from long-term positive selection, and background selection from negative selection on deleterious variants) [125]. Smaller values indicate regions that are often functional or close to genes.

I compared the number of Neanderthal segments in all non-African samples spanning each polymorphic site to the corresponding B value of the site. Sites covered by at least one Neanderthal segment were further partitioned according to the number of distinct alleles found among all Neanderthal haplotypes recovered from modern genomes. If genomic evolution in the Neanderthal population is governed by a similar pattern of linked selection, the genetic diversity of Neanderthal haplotypes is also expected to decrease in functionally important regions, so that sites in regions with small B values should be less polymorphic.

Figure 5.8 compares the distribution of B values at polymorphic sites with and without Neanderthal ancestry, and the number of distinct Neanderthal alleles recovered. Comparing to sites totally free of Neanderthal ancestry, the B values of sites covered by at least one Neanderthal segment are slightly shifted towards 1 (neutral evolution). This trend is better depicted by the average number of Neanderthal segments covering sites across B value bins (Figure 5.9). Sites under less selective constraint are enriched for Neanderthal ancestry.

The distribution of B value is similar regardless of how many Neanderthal alleles are recovered (Figure 5.8). Most multiallelic sites also contains the modern human genotype. Perhaps genetic drift and negative selection have largely reduced the genetic diversity in Neanderthal haplotypes, and what remains is not a truthful representation of the genetic diversity in the Neanderthal population.

Although I was able to replicate the correlation between the strength of linked selection and the frequency of Neanderthal haplotypes, there are some caveats in deciding whether negative selection happened in modern human or the Neanderthal population. The B values are estimated from comparing five primate species[125], thus should reflect an evolutionary trend across tens of millions of years. However, the same constraint should also apply to more recent evolution in functional regions, causing the modern human-Neanderthal



(Sites spanned by Neanderthal haplotypes are weighted by the frequency of Neanderthal alleles)

Fig. 5.8 Distribution of *B* values at sites with and without Neanderthal ancestry

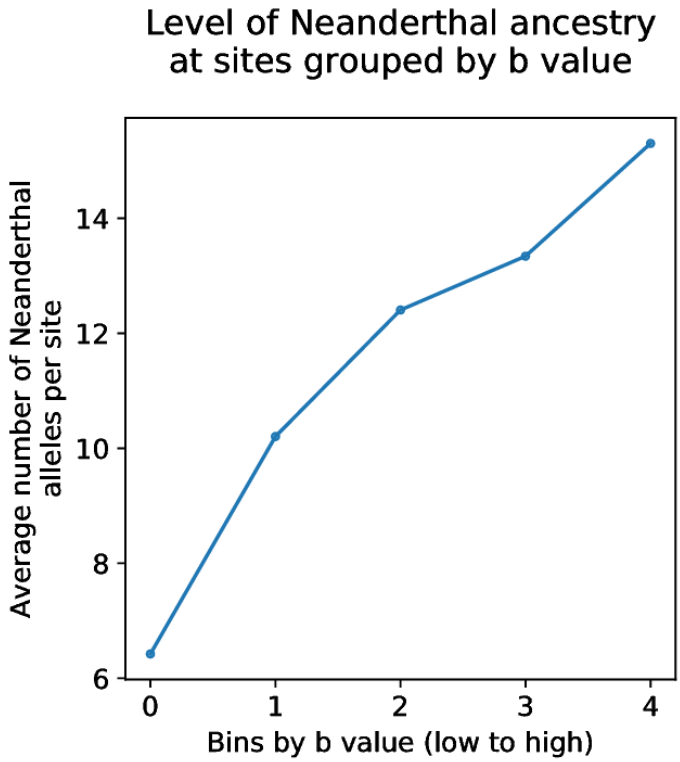


Fig. 5.9 Amount of Neanderthal ancestry at sites grouped by *B* values

divergence to decrease. If the divergence becomes too low, the HMM might have difficulty detecting Neanderthal haplotypes with no or very low divergence from the modern human haplotypes, reporting a decrease in Neanderthal ancestry when no selection was actually involved; alternatively, we might observe unchanged or even increased levels of Neanderthal ancestry if the divergence is high enough for the HMM to distinguish, but the functional effect is minimal under selective constraint. It has been estimated through simulation that natural selection is less effective in the Neanderthal population due to its small effective population size [111]. Ideally, access to more Neanderthal genomes or historical human genomes soon after the admixture will help to measure selection before and after the introgression.

5.4.3 Potential functional consequences of introgression

Archaic introgression has been found to influence a number of phenotypes in modern humans, including immune system components [116], autoimmune disorders [117], adaptation to high altitude [118], risk of depression and tobacco use [120], to name a few examples. To explore functional consequences of archaic alleles, I searched the GWAS database for effect alleles that are likely to come from introgression.

A total of 66,508 SNP-trait associations were downloaded from the GWAS Catalog [160] (https://www.ebi.ac.uk/gwas/, v1.0.1). They were filtered according to the following criteria: 1. the effect allele is a well-defined SNP; 2. the SNP position passes all the masks; 3. the SNP position is covered by at least one archaic segment in all non-African genomes; 4. the effect allele is not the ancestral allele; 5. the frequency of the effect allele is below 0.05 in Africans; 6. the effect allele is present in the archaic genome(s).

After merging records of the same variant, I found 66 associations where the effect allele has a likely Neanderthal origin (Table 5.2), but none remains after filtering for the Denisovan origin. The list contains some previously identified genes and loci [67], and many records are connected with previously reported traits, such as lipid metabolism [161], immune function and autoimmune diseases [162, 163], body shape, sleep pattern, and mood[121], however most loci do not overlap with previous studies. The difference is most likely due to the reliance on the existing GWAS database: none of the loci reported in two major studies directly searching for associations between Neanderthal ancestry and phenotype data ([120] and [121]) are included in the GWAS Catalog. In addition, most GWAS designs are not able to detect the effect of very rare alleles; together with a heavy bias towards populations with European ancestry in GWAS studies so far, most Denisova variants are thus excluded from the database.

Table 5.2 GWAS records where the effect allele is likely to have Neanderthal origin

chr	pos	gene	trait	SNP	freq in Africa	freq in Neand.	freq in introgres.
1	3734467	TP73, CCDC27, KIAA0495, LRRC47	Visceral adipose tissue/subcutaneous adipose tissue ratio	rs12562437-T	0.0	0.5	0.809
1	3734845	TP73, CCDC27, KIAA0495, LRRC47	Visceral fat	rs10910018-A	0.0	0.5	0.829
1	13866181	PRDM2	Left ventricular function change in anthracycline treatment	rs7542939-A	0.0	0.75	1.0
1	39370145	MACF1	Peripheral arterial disease (traffic-related air pollution interaction), Type 2 diabetes	rs2296172-G	0.01	1.0	1.0
1	39570256	MACF1, PABPC4, RP11-69E11.8	Type 2 diabetes, red cell distribution width	rs3768321-T	0.01	1.0	1.0
1	115135325	TSPAN2	Migraine without aura	rs12134493-A	0.014	1.0	1.0
1	150578448	MCL1	Eosinophil counts	rs34645101-C	0.0	1.0	1.0
1	159492591	OR10J1, OR10J5	Obesity-related traits	rs4325129-G	0.034	1.0	1.0
1	168134378	GPR161	Red cell distribution width	rs78320035-C	0.0	0.75	1.0
1	169129799	ATP1B1	QT interval	rs10919070-C	0.0	1.0	1.0
1	169552615	F5	Blood protein levels	rs6033-G	0.0	1.0	1.0
1	170225682	METTL11B	Atrial fibrillation	rs72700118-A	0.0	1.0	1.0
1	208821591	intergenic	Educational attainment	rs17013497-T	0.0	1.0	0.976
1	214143432	NR	Facial morphology (factor 15, philtrum width)	rs145984379-T	0.0	1.0	1.0
1	217544790	GPATCH2	Visceral adipose tissue adjusted for BMI	rs2059397-G	0.0	0.25	1.0
1	225402973	NR	IgG glycosylation	rs16844841-C	0.005	1.0	1.0
3	2144738	CNTN4, CNTN4-AS2, RPL21P17	Daytime sleep phenotypes	rs62246964-C	0.048	1.0	1.0

3	2222925	CNTN4	Middle childhood and early adolescence aggressive behavior	rs4685500-T	0.005	0.5	1.0
5	9552226	SEMA5A, SNHG18, SNORD123	Coronary artery disease	rs17263917-A	0.0	1.0	1.0
5	44068744	FGF10	Nonsyndromic cleft lip with cleft palate	rs10462065-A	0.0	0.75	1.0
6	7232156	RREB1	Red blood cell count	rs75757892-T	0.0	0.5	1.0
6	15176868	intergenic	Red blood cell count	rs9464759-C	0.019	1.0	0.9375
7	28149464	JAZF1	Height	rs1029534-T	0.048	1.0	1.0
7	28150327	JAZF1	Waist circumference adjusted for body mass index	rs1708299-A	0.048	1.0	1.0
7	36044919	EEPD1	Fibrinogen levels	rs2710804-C	0.034	1.0	1.0
7	45977063	IGFBP3	Sitting height ratio	rs1722141-A	0.024	1.0	0.963
7	46362671	IGFBP3	Hypospadias	rs7811653-A	0.0	1.0	1.0
7	95826533	DYNC1I1	Dementia and core Alzheimer's disease neuropathologic changes	rs3779483-T	0.0	1.0	1.0
7	128945562	IRF5	Systemic lupus erythematosus	rs35000415-T	0.0	1.0	1.0
7	128954129	TNPO3, IRF5	Systemic lupus erythematosus, primary biliary cholangitis, systemic sclerosis , primary biliary cholangitis	rs10488631-C	0.005	1.0	1.0
7	128956751	TNPO3, IRF5	Systemic lupus erythematosus	rs12539741-T	0.0	0.25	1.0
7	128977412	IRF5	Systemic lupus erythematosus, primary biliary cholangitis	rs12531711-G	0.0	1.0	1.0
7	129041008	IRF5, TNPO3	Sjögren's syndrome	rs17339836-T	0.0	1.0	1.0
8	14531972	SGCZ	Platelet aggregation	rs1903595-G	0.0	1.0	1.0
8	22172552	BMP1, SFTPC	Coronary artery disease	rs73225842-T	0.02	1.0	1.0
9	25452814	intergenic	RR interval (heart rate)	rs13300284-A	0.0	1.0	1.0
11	20194210	intergenic	Educational attainment	rs10500871-T	0.0	1.0	1.0
11	60993140	CD6	Multiple sclerosis	rs17824933-G	0.029	0.75	0.0
11	63147874	SLC22A9	Sex hormone levels	rs112295236-G	0.0	0.25	1.0
12	3283934	TSPAN9	Glomerular filtration rate (creatinine)	rs67551338-T	0.0	1.0	1.0

12	20446178	PDE3A	Systolic blood pressure (alcohol consumption interaction)	rs10841530-G	0.034	1.0	1.0
12	28447311	CCDC91	Height	rs11049611-T	0.01	1.0	1.0
12	40208138	MUC19, LRRK2	Crohn's disease	rs11175593-T	0.0	1.0	1.0
12	45574922	NR	Subjective well-being	rs75279353-A	0.0	0.75	1.0
12	62786149	PPM1H	Sense of smell	rs11174650-T	0.005	1.0	1.0
12	102119753	NUP37, C12orf48, PMCH	Height	rs2292303-C	0.0	1.0	1.0
12	103763724	STAB2	Sense of smell	rs3751196-A	0.0	1.0	1.0
13	38041119	TRPC4	Alanine aminotransferase (ALT) levels after remission induction therapy in actute lymphoblastic leukemia (ALL)	rs747 09575 -C	0.0	0.75	1.0
13	40220445	LINC00548	Primary sclerosing cholangitis	rs61954180-C	0.019	0.5	1.0
14	22533736	TRA	Narcolepsy	rs1154155-G	0.019	1.0	0.0
15	63175688	RAB8B	Social communication problems	rs17828380-C	0.0	0.75	0.922
16	24076855	PRKCB	Post bronchodilator FEV1/FVC ratio in COPD	rs9928486-C	0.034	1.0	1.0
16	24078450	PRKCB	Post bronchodilator FEV1/FVC ratio in COPD	rs12921419-G	0.048	1.0	1.0
16	24082016	PRKCB	Post bronchodilator FEV1/FVC ratio in COPD	rs35526040-T	0.034	1.0	1.0
16	24083656	PRKCB	Post bronchodilator FEV1/FVC ratio in COPD	rs12929627-T	0.034	1.0	1.0
16	24084053	PRKCB	Post bronchodilator FEV1/FVC ratio in COPD	rs34289708-C	0.034	1.0	1.0
17	897353	NXN, NR	Colorectal cancer	rs12603526-C	0.0	1.0	1.0
19	8081044	FBN3, ELAVL1	Blood protein levels	rs3848570-T	0.014	1.0	1.0
19	47898636	SULT2A1	Dehydroepiandrosterone sulphate levels	rs2637125-A	0.0	1.0	1.0
20	4151717	SMOX	Reticulocyte fraction of red cells	rs13043612-G	0.0	1.0	1.0
20	4245273	ADRA1D	Paneth cell defects in Crohn's disease	rs12481514-A	0.0	0.25	0.988

22	17116572	CECR6, IL17RA	Heschl's gyrus morphology	rs971768-A	0.0	1.0	1.0
22	25808585	NR	Trans fatty acid levels	rs8184969-T	0.01	1.0	1.0
22	25811943	NR	Trans fatty acid levels	rs5752223-T	0.0	0.5	1.0
22	32049959	SLC5A1	GLP-1 levels in response to oral glucose tolerance test (120 minutes)	rs17683011-G	0.0	0.5	0.889
22	49952394	IL17REL, LOC90834, MAPK11, MAPK12, MLC1, PANX2, PIM3, PLXNB2, PPP6R2, SELO, TRABD, TUBGCP6, ZBED4, ALG12, BRD1, C22orf34, MOV1 0L1, CRELD2, DENND6B, HDAC10	Acne (severe)	rs56091001-T	0.0	0.5	0.994

This list also illustrates some deleterious effect of Neanderthal haplotypes, with increased risks for various diseases and social difficulties. Unlike some of the cited studies, querying by genotype merely suggests the Neanderthal origin of these alleles, instead of negative or positive selection on them. The latter could be explored by studying the local linkage disequilibrium structure and comparing geographical distributions, but it is beyond the scope of this thesis.

5.5 Divergence of archaic segments to archaic genomes

To date, three archaic hominin genomes have been sequenced to high coverage: a 52-fold Neanderthal genome sequenced from a toe bone found in the Denisova Cave in the Altai Mountains in Siberia [66], a ~30-fold Denisova genome sequenced from a finger bone in the same cave [47], and another ~30-fold Neanderthal genome sequenced from a bone fragment found the Vindija Cave in Croatia [67]. The bone fragment from the Vindija Cave was radiocarbon dated to over 45.5k years before present [67]; the bones from the Denisova Cave have not been directly dated, but animal bones in the same layer were radiocarbon dated to >48-30k years before present [73]. None of the individuals whose genomes have been sequenced to high-coverage are from the populations that admixed with modern humans. Especially in the case of Denisovan admixture, the population split times between the admixture source and the Altai Denisova genome is estimated from haplotype divergence to be 276–403k years, in contrast to 77–114k years between the source of Neanderthal gene flow and the Altai Neanderthal genome [66] (and even closer to the Neanderthal genome from Vindija Cave [67], although no split time estimates were given). A detailed study of the archaic segments recovered from modern human genomes might shed some light on the distribution, population size or the structure of the source populations, of which we know very little.

In this section, I compared the divergence between archaic segments and the high-coverage archaic reference genomes to infer the relation of the sources of gene flow to available archaic genomes. If the same Neanderthal/Denisova source contributed to all surviving Neanderthal/Denisovan segments in non-African genomes, the divergence of archaic segments to any particular archaic reference genome is expected to remain similar across geographical regions; alternatively if certain regions admixed with a different Neanderthal/Denisova source, or received one or more additional episodes of gene flow on top of what is shared by all non-African populations, it might manifest as variations between geographical regions. In addition, if more than one episode of gene flow from genetically distinct Neanderthal/Denisova sources occurred in the same modern human population, the divergence measured in each archaic segment might form corresponding clusters.

The divergence is measured by the average number of nucleotide differences per base pair (Hamming distance), excluding masked regions and missing sites. All Neanderthal/Denisovan segments from the "strict" set in each genome are compared with two Neanderthal genomes from the Denisova and Vindija Caves and the Denisova genome from the Denisova Cave. To recover archaic-private variants that were not in the merged VCF files, I assumed that

all modern sites passing the strict mask but not present in the VCF files carry the reference allele; if these sites also pass the respective archaic sequencing accessibility mask and appear in the archaic GVCF files with alternative alleles, they also contribute to the difference count.

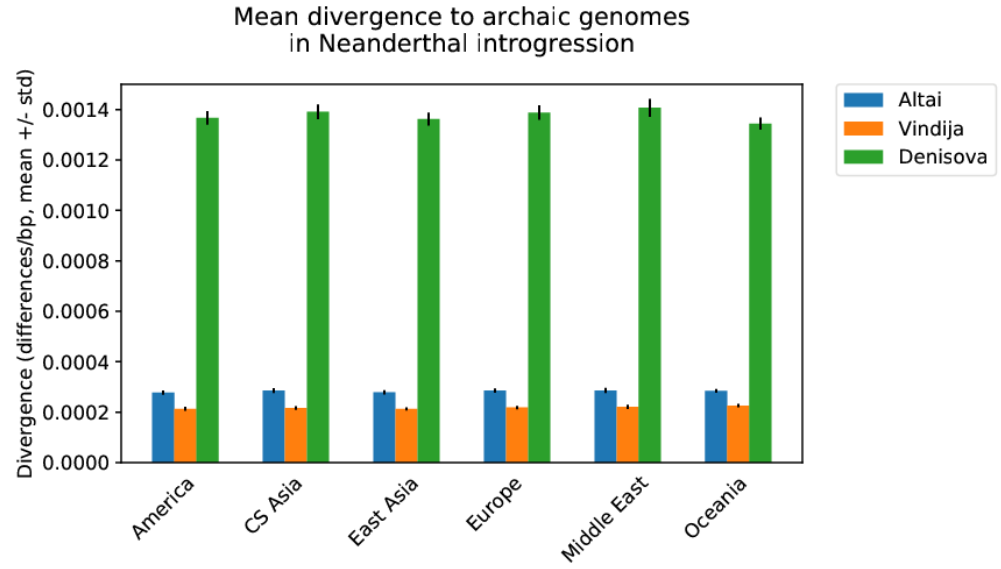
5.5.1 Genomewide divergence

The divergence in each modern individual is obtained by dividing the total number of nucleotide differences between its Neanderthal/Denisovan segments and the archaic genomes over the total length of Neanderthal/Denisovan segments that pass both the modern mask and the respective archaic genome mask. The individual values were then combined to calculate the population and regional average (Figure 5.10 and Figure 5.11).

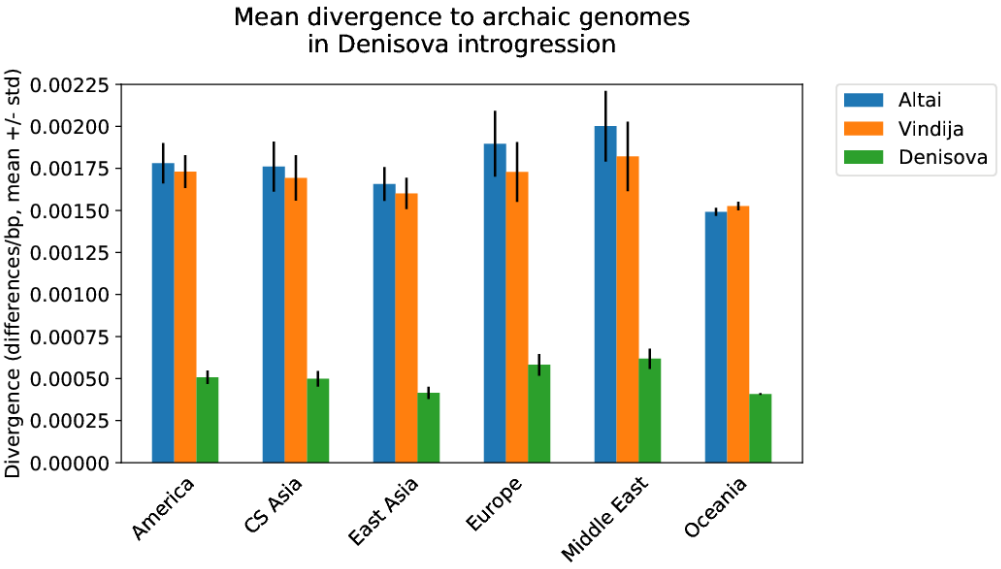
Assuming a generation time of 29 years, a mutation rate of 1.25×10^{-8} /(site $\cdot$ generation) and the tentative dating of ~50k years ago for the archaic genomes, the average divergence translates into a genetic split time of ~200k years between the introgressing Neanderthal and the Vindija Neanderthal genome, and ~500k years between the introgressing Denisova and the Altai Denisova genome. This is roughly in line with previously estimated genetic split time [47]. It should be noted that the split time between two lineages provides an upper bound for the population split time, and the margin can be wide of the ancestral population is large.

The divergence of Neanderthal segments to all three archaic genomes shows little variation across geographical regions (Figure 5.10a) and within each region (Figure 5.11). In agreement with previous studies, the segments are closer to the Neanderthal genome from Vindija Cave in northern Croatia than the one from Denisova Cave in Siberia [67, 69], suggesting that the admixture with Neanderthal might have happened closer to Central Europe than to Siberia.

I find wider variations in the divergence of Denisovan segments to archaic genomes (Figure 5.10b and Figure 5.11). The variance within each population or region is correlated to the amount of Denisovan ancestry: the more segments there are, the smaller the standard deviation. In regions with the highest levels of Denisovan ancestry, namely Oceania and East Asia, the detected Denisovan segments also show the highest affinity to the Altai Denisova genome. The highest divergence is found in Europe and the Middle East, where the presence of Denisovan ancestry is minimal; these Denisovan segments also show increased divergence to two Neanderthal genomes, therefore unlikely to be misclassified segments from Neanderthal. Perhaps this is caused by modern segments flanking authentic Denisovan segments



(a) Neanderthal segments



(b) Denisovan segments

Fig. 5.10 Average divergence of inferred Neanderthal and Denisovan segments in HGDP genomes to three archaic genomes across geographical regions

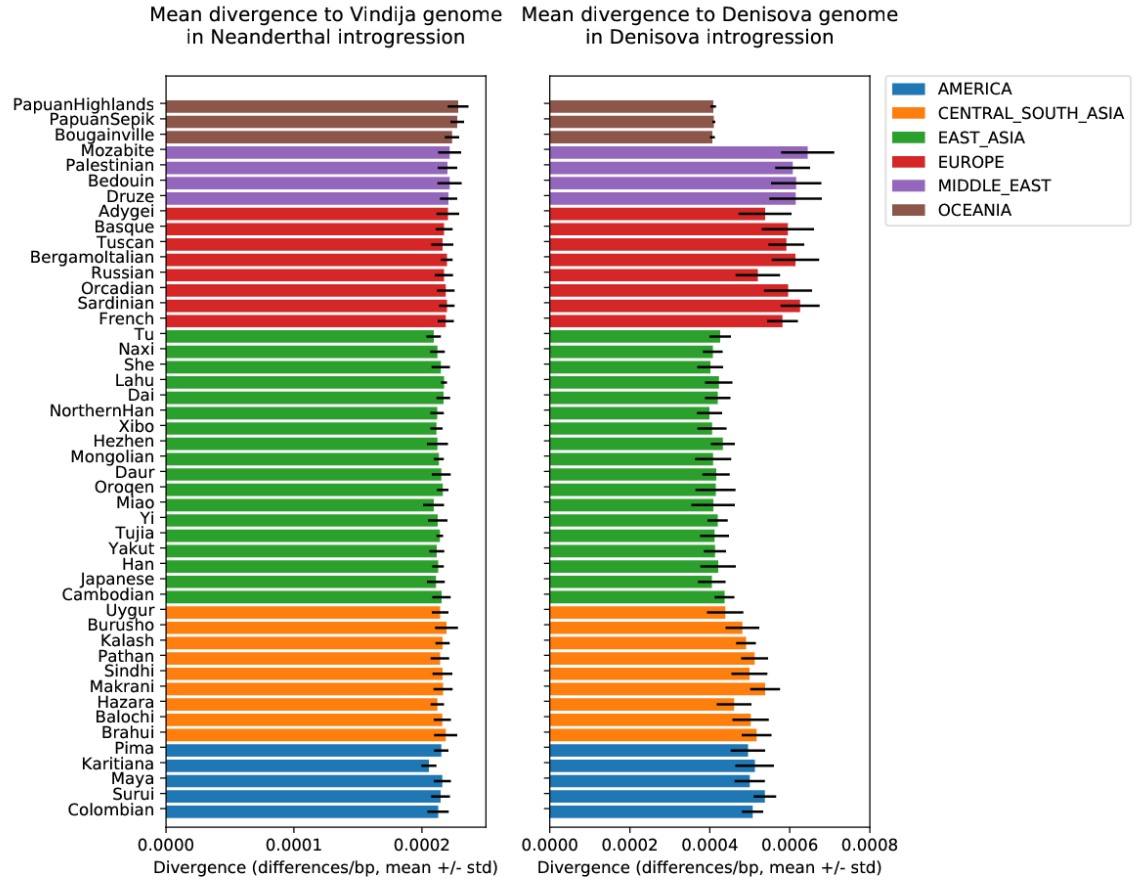


Fig. 5.11 Average divergence of inferred Neanderthal and Denisovan segments in HGDP genomes to their closest archaic genomes across populations

also classified as Denisova. Considering the overall scarcity of Denisovan ancestry in these parts of the world, caution should be taken when drawing conclusions.

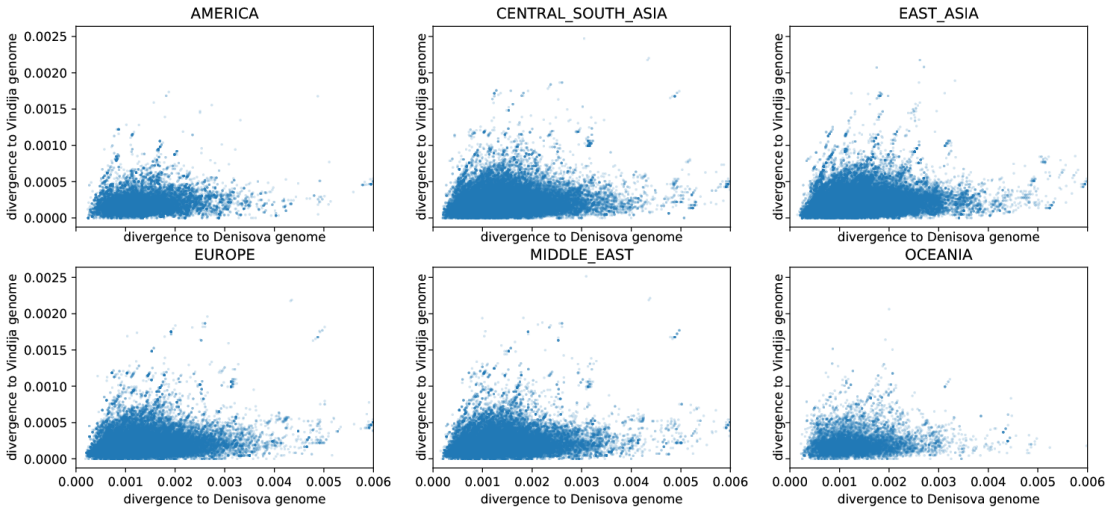
Overall, the average divergence per genome of archaic segments to Neanderthal genomes does not support distinct sources of Neanderthal gene flow into different regions of the world, whilst the geographical variations in the divergence between Denisovan segments in modern genomes and the Denisova genome could provide tentative evidence for different sources of Denisova gene flow. On the other hand, the genome-wide average can conceal the presence of multiple components within the same genome, which is discussed in the next section.

5.5.2 Divergence by segment

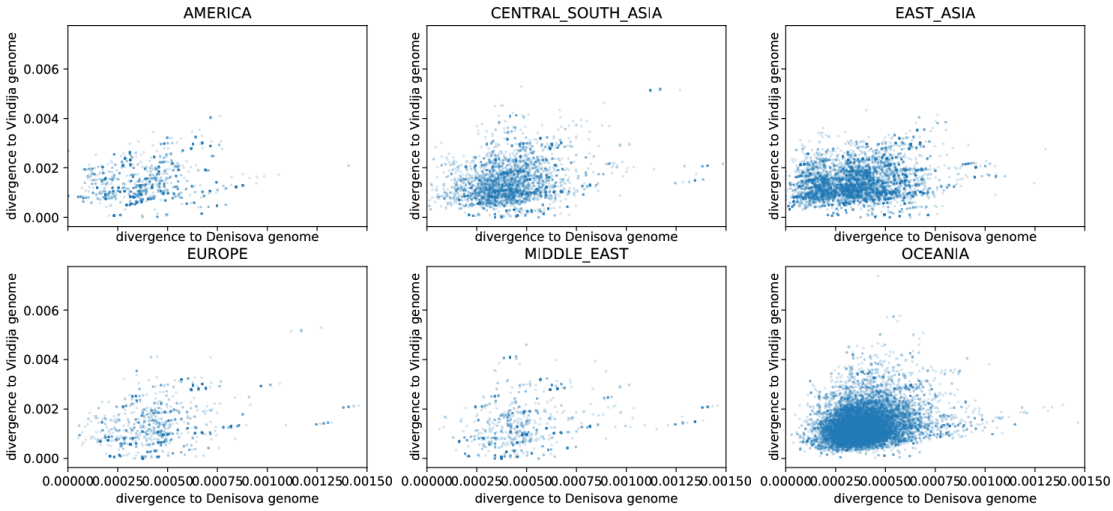
To explore whether more than one source population contributed successively to the Neanderthal and/or Denisovan ancestry in some modern human populations, I also plotted the divergence in each archaic segment. The same measurement of divergence as above was used, only without averaging in each genome. Figure 5.12 shows the divergence of each Neanderthal/Denisovan segment to the corresponding sequence in the Vindija Neanderthal genome and the Altai Denisova genome.

In Figure 5.12a, the overall pattern is almost identical in all six geographical regions, although differences in the level of Neanderthal ancestry and the sample size in each region cause the density to change. But in Figure 5.12b, the distribution in East Asia and Oceania follows visibly different shapes: the points in Oceania form a well-defined cluster; in East Asia, the pattern appears noisier, but there exist segments very close to the Altai Denisova genome (divergence less than ~0.0001) that are absent from the Oceania plot. This component is also potentially visible in America and Central/South Asia.

The result here corroborates the finding from [93] about an additional pulse of Denisova gene flow into East Asia. However, there are a number of disagreements stemming out from differences between our methods to detect archaic segments. The Sprime algorithm introduced in [93] is a reference-free method aimed at recovering putative archaic haplotypes, other than detecting particular archaic segments in each genome. It operates at the population level since the S^* score at its basis makes use of linkage-disequilibrium information. When multiple putative archaic haplotypes occur at overlapping genomic regions, only one with the highest score will be reported. In practice, this means that the haplotypes in their contour plot are not weighted by their frequency, and the rarer archaic haplotypes in the population will be missing if they overlap with other more frequent haplotypes. This is potentially the reason



(a) Neanderthal segments



(b) Denisovan segments

Fig. 5.12 Divergence of each archaic segment to Vindija Neanderthal and Altai Denisova across geographical regions

why the additional component in East Asia also show up in America and Central/South Asia in Figure 5.12b, while [93] did not detect it in South Asia or America.

Although both methods find the extra Denisova component in East Asia to be closer to the Denisova genome than the Oceania component, the divergence to archaic genomes is also measured differently in [93]: since Sprime only reports inferred archaic genotypes at certain positions, the match rate is defined as the proportion of such genotypes that exist in the archaic genome. Any differences outside the reported positions will be ignored.

To confirm if the additional Denisova component in East Asia detected here is similar to what was found in [93], I also ran Sprime on 223 HGDP genomes from East Asia, using 104 sub-Saharan African genomes as outgroup. Also following their measurement of divergence, I was able to reproduce the contour density plot in East Asia in Figure 4 of [93] (Figure 5.13). The highest peak in the lower right corner correspond to putative Neanderthal segments, and the two smaller peaks near the y-axis (excluding the peak near the origin, which are segments that match neither archaic genomes) correspond to putative Denisovan segments. According to [93], the peak with a higher match rate (over 0.8) to the Altai Denisova genome is exclusively found in East Asia. As a preliminary exploration, I collected all the segments from Sprime output whose match score to Altai Denisova falls between 0.8 and 0.9 and match score to Vindija Neanderthal does not exceed 0.1 into a "private East Asia" set. Segments whose match score to Altai Denisova falls between 0.4 and 0.7 and match score to Vindija Neanderthal does not exceed 0.1 were collected into a "shared" set. If a segment detected by HMM overlaps in genomic coordination by more than half of its length with the "private East Asia" segments in Sprime result, it was also grouped as private East Asia; similarly, some segments in the HMM result were labelled as shared when they overlap with the "shared" segments from Sprime result by more than half the lengths. The classification was entirely based on genomic locations without reference to genotypes. Figure 5.14 shows where these Denisovan segments detected by the HMM and overlapping with Sprime results fall onto the divergence plot in East Asia and Oceania. In East Asia, although there is a large space of overlapping between these two components, points overlapping with the "private East Asia" set from Sprime result indeed tend to fall around the area with low divergence to the Denisovan genome, consistent with the additional East Asian component identified previously when looking at HMM results alone. Fewer Denisovan segments identified by the HMM in Oceania overlap with either component reported by Sprime in East Asia, nevertheless, the shared component extends throughout the core cluster, whilst the component private to East Asia appears rarer and more restricted in its distribution. Therefore the HMM and Sprime

detect similar components of Denisovan introgression in East Asia that are not present in Oceania.

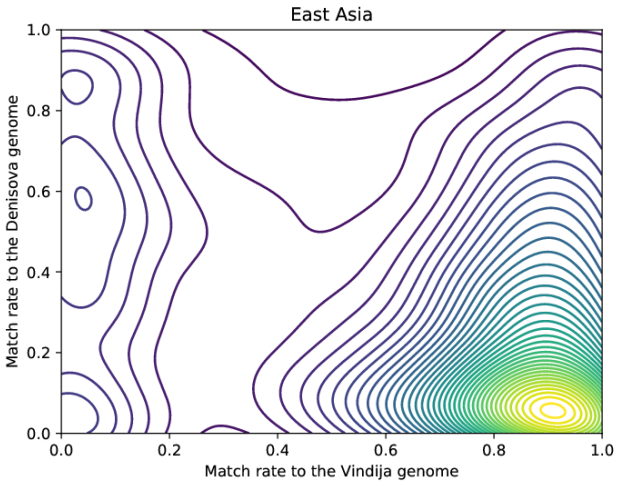


Fig. 5.13 Contour density plot showing the match score of archaic segments in East Asia detected by Sprime to Altai Denisova and Vindija Neanderthal genomes

The pattern of divergence between individual archaic segments and the archaic genomes does not provide evidence for additional local encounters with Neanderthals, but supports more than one admixture with Denisova in agreement with a previous study [93]. This additional component, however, does not appear to be restricted to East Asia; traces of it were also found in American and Central/South Asian genomes in the HGDP dataset. Moreover, a similarity in relation to known archaic genomes does not guarantee that the segments come from the same source at the same time, since different source populations might show identical relationships to the Altai Neanderthal and the Denisova individuals. Based on evidence from Section 5.6 and 5.7.3, it is plausible that the Denisovan ancestry in Oceania results from another admixture event separate from East Asia. The lack of distinct structure in the East Asia plot in Figure 5.12b might be due to variance in divergence from two components of admixture, or might reflect an even more complicated history of admixture, possibly from several source populations at various locations and times.

5.6 Nucleotide diversity within archaic segments

The intra- and inter-population diversity in archaic segments has been shaped by a unique demographic history that traces back to the archaic populations. The admixture process,

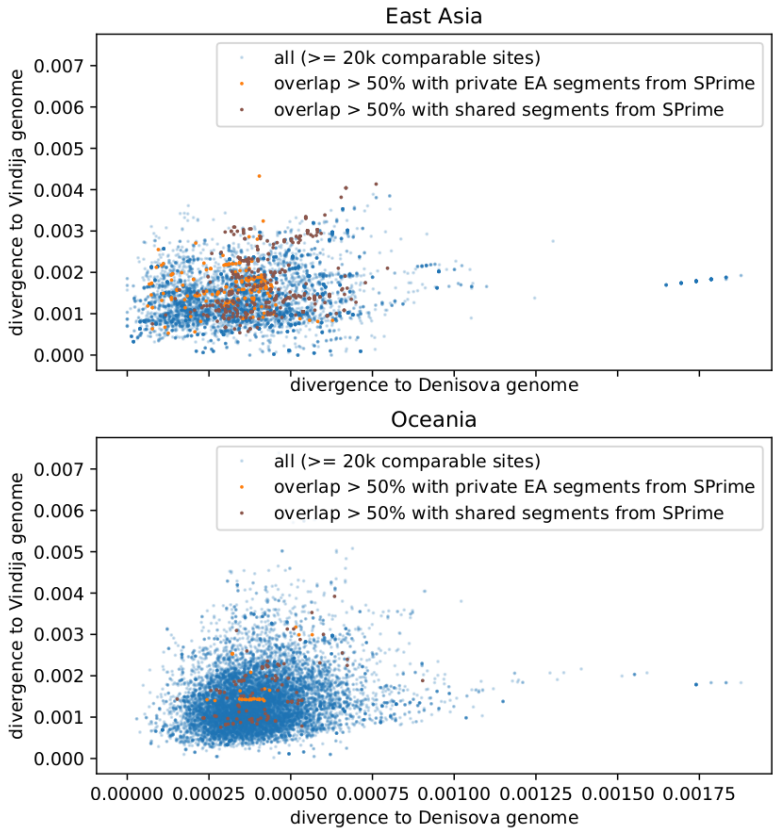


Fig. 5.14 Divergence of each archaic segment to Vindija Neanderthal and Altai Denisova in East Asia and Oceania, highlighting segments overlapping with private East Asia and shared components identified in Sprime. Segments reported by Sprime as private to East Asia also show up in the HMM result.

along with strong selection against hybrids in the next few generations [1, 113], could have imposed a sharp bottleneck on archaic haplotypes. Subsequently negative selection could further reduce genetic diversity and wipe out archaic lineages, although modeling has shown that its long-term influence is limited [115]. The encounters with both the Neanderthal and the Denisova population are estimated to have happened as modern humans first moved out of Africa into Eurasia, Oceania and the Americas in a series of expansions. Therefore, the archaic segments, once in the modern human population, also experienced recent population growth and subdivisions in the same way as the unadmixed part of modern human genomes.

By comparing the genetic diversity in archaic versus unadmixed regions of the genomes, I aim to explore the structure of the admixture event, the size of the bottleneck, and even the deeper demographic history in the archaic populations.

5.6.1 π and D_{XY}

Within one population, the expected number of nucleotide differences per site between two randomly drawn haplotypes is commonly known as nucleotide diversity (π). When comparing two populations, the same expected value between sequences randomly drawn from two populations (excluding all comparisons within the same population) is commonly known as absolute divergence (D_{XY}, also referred to as π_{XY}, π_B or d_{XY} in the literature) [164]. Based on the "strict" set of result, three sets of π in all populations and D_{XY} between all pairs of populations were obtained: values calculated from only the Neanderthal segments in the genomes, from only the Denisovan segments in the genomes, and from only the unadmixed (also referred to as "modern") segments of the genomes.

In this context, a "haplotype" refers to the collection of all Neanderthal (or Denisova or modern) segments located on the same haploid genome. Since introgressed segments typically span different genomic regions in different individuals, it is only meaningful to compare nucleotide differences in the overlapping regions of two haplotypes (Figure 5.15). π is calculated by averaging the values between all pairs of haplotypes within the population; but to limit computational cost when calculating D_{XY}, if the sample size of a population exceeds 10, only 20 haplotypes are randomly drawn for pairwise comparison with a maximum of 20 haplotypes from the other population. In practice, the number of 20 enables the calculation to finish within a reasonable time while causing little variation between repeated runs. The values of π and D_{XY} obtained from different partitions of the genomes are compared below, with an emphasis on geographical variations and the implications for the history of archaic admixture.

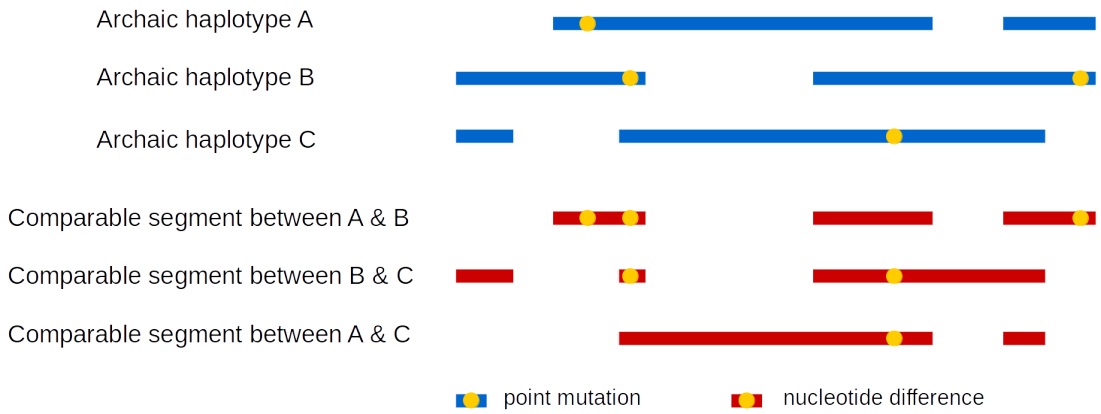


Fig. 5.15 Diagram showing comparable regions between three archaic haplotypes and their nucleotide differences

Neanderthal vs. unadmixed regions

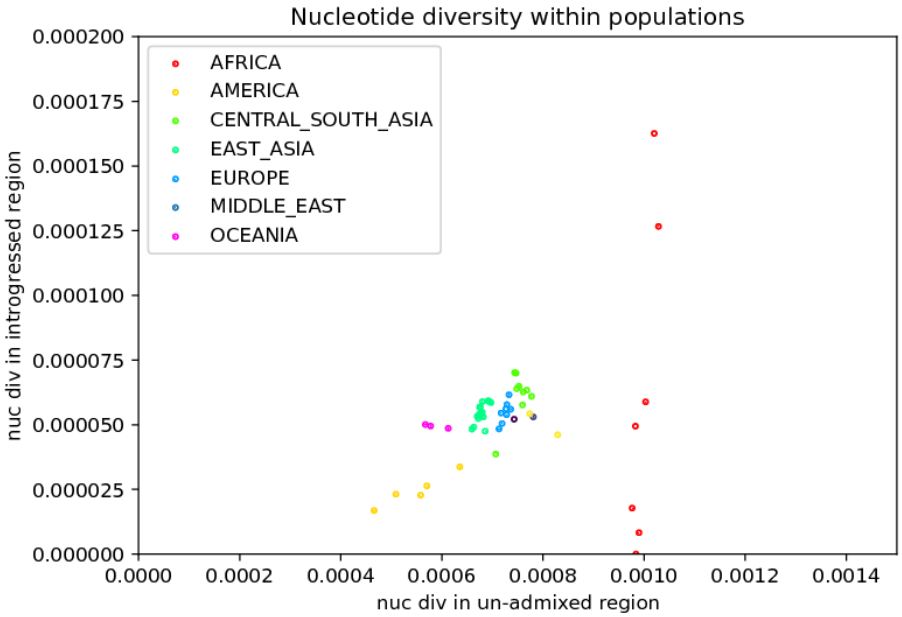
The nucleotide diversity within each population measured in Neanderthal segments (π_N) is plotted against the value measured in unadmixed segments (π_M) in Figure 5.16. The first plot highlights geographical divisions and includes sub-Saharan African populations; the second one shows non-African populations only, labeled with population names. In Figure 5.16a the sub-Saharan African populations exhibit high π_M but highly variable π_N, consistent with a small number of segments that are mostly misclassified as archaic. Overall, the diversity is higher in unadmixed modern human regions (x-axis) than in Neanderthal regions (y-axis), possibly because the historical effective population size is considerably larger in modern human than in Neanderthal prior to the admixture. Additionally, some structure could have already developed in modern human population outside Africa at the time of admixture, although it should not be deep enough to hinder the spread of Neanderthal ancestry across all non-African populations today. Assuming a mutation rate of 1.25×10^{-8} /(site $\cdot$ generation), a typical π_N at 0.00005 would translate to 2,000 generations of divergence, coinciding with the estimated time of the Neanderthal admixture at ~50-60k years ago [114]. A clear positive correlation between π_N and π_M ($R^2 = 0.1055$ with Africans, 0.5530 without) suggests that both have been influenced by similar demographic processes, namely the recent population history after the Neanderthal gene flow.

After excluding sub-Saharan African populations, which show high π_M but a wide range of π_N, Mozabite and three Papuan populations deviate more from the linear relationship (Figure 5.16b). Since the Papuan π_M is reasonably distributed around the same level as

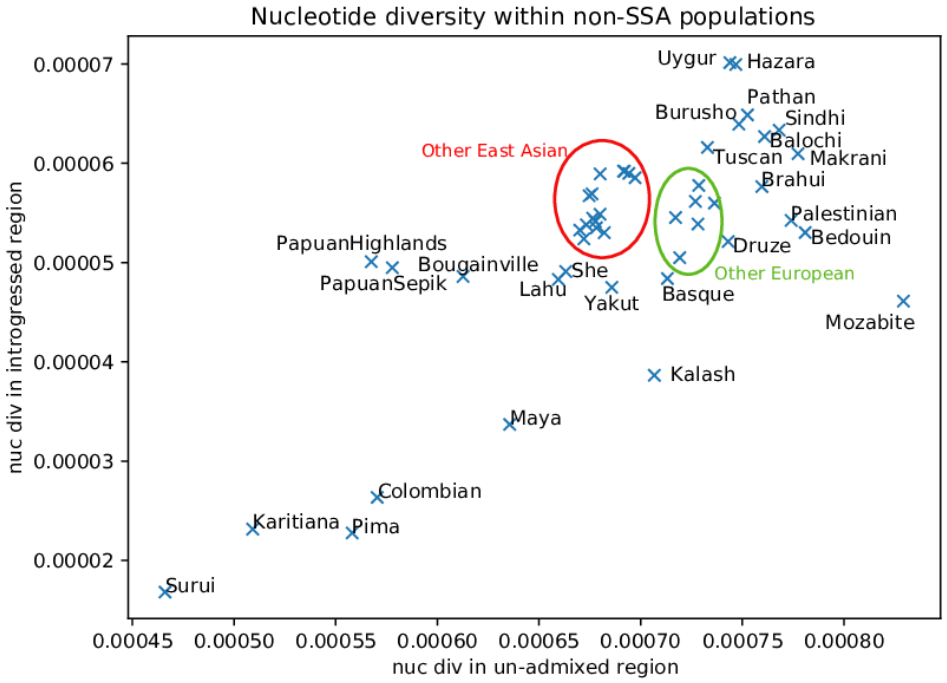
American populations, I suspect the shift is due to π_N being inflated by the presence of some Denisovan segments misclassified as Neanderthal ones. Such cases should be rare in the "strict" set of results, but their potentially high divergence to the authentic Neanderthal haplotypes (Figure 5.10) means even a tiny amount might raise the expected values. An alternative explanation is a small amount of ancestry from an unknown source related to the Neanderthals, which exists only in the Papuans. In the Mozabite population, a history of isolation followed by admixture with both sub-Saharan and Middle Eastern population [165] produces the highest genetic diversity outside sub-Saharan Africa in the modern genomic regions; meanwhile, the diversity in Neanderthal regions is hardly affected by the recent admixture, either because only the component from the Middle East carries substantial amount of Neanderthal ancestry, or because the Neanderthal haplotypes in admixing human populations were still very similar at that time.

Similarly, I also plotted the absolute divergence in Neanderthal regions (D_{XY-N}) against unadmixed regions (D_{XY-M}). Figure 5.17 shows the scatter plot, each panel highlighting all pairs containing a particular population. Comparison between populations also reveals greater genetic diversity in unadmixed regions (x-axis) than in Neanderthal regions (y-axis). The relationship involving any one population is largely linear, meaning the average divergence to all other populations in Neanderthal regions remains proportional to the value in unadmixed regions. It is unlikely, therefore, that certain populations received substantial Neanderthal gene flow from a genetically distinct source, in which case we would expect different D_{XY-N} values for populations with similar D_{XY-M}. However, the position of points involving different populations varies considerably worldwide, primarily due to variation in D_{XY-M} values corresponding to different demographic histories after Neanderthal admixture. The most pronounced outlier turns out to be Mozabite again, whose divergence to all other non-Africans in the unadmixed regions is elevated possibly due to its African ancestry. The range of D_{XY-M} seen in each population is consistent with recent demographic history. For example, pairs containing Papuan Sepik mostly fall into the high divergence range, except for two data points resulting from comparison to the other Papuan populations.

One striking observation in Figure 5.17 is that the trend established by all data points involving a certain population follows two distinct slopes by geographical division: all populations from East Asia, Oceania and America show a shallower slope, whilst all populations from Europe, the Middle East, Central Asia, and South Asia show a steeper slope. D_{XY} can be influenced by a number of factors, including admixture with other sources, effective sizes of the ancestral populations, in addition to the accuracy in detecting Neanderthal segments that



(a) All populations



(b) Excluding sub-Saharan Africans

Fig. 5.16 Intra-population nucleotide diversity in Neanderthal segments and unadmixed segments

could vary with different phasing error rate or different amount of other archaic ancestries, including Denisova. Nevertheless, the structure of Neanderthal gene flow alone (such as separate sources into different modern human populations or an additional episode in some modern human populations) cannot explain the pattern.

A correction has been proposed to account for the diversity in their shared ancestral population when the two populations under comparison are closely related, by deducting the average π in each population from D_{XY} [164]:

$$\delta_{XY} = D_{XY} - (\pi_X + \pi_Y)/2$$

δ_{XY} (also referred to as d_a, D_a or D_m) is an estimate of net nucleotide differences accumulated after the population split. The slopes in populations from the above two geographical regions become similar after applying the adjustment (Figure 5.18). As the diversity in the ancestral population is approximated by current day populations, δ_{XY} will become biased by recent demographic history. The parallel shift towards the right (higher divergence in unadmixed regions) now reflects recent bottlenecks, most prominent in Oceanian and American populations, where the ancestral diversity is not fully accounted for by their diversity at present. δ_{XY} in the Neanderthal regions is hardly affected, perhaps because their ancestral diversity is low across all populations. Therefore, the different slopes in Figure 5.17 most likely result from a smaller ancestral size between populations descending from an ancestral East Asian population than between populations in the European and Central/South Asia regions.

To facilitate comparison between D_{XY} in Neanderthal and unadmixed regions across populations, the values are colour-coded onto the upper-right and lower-left triangles of a matrix in a heat map (Figure 5.19). All D_{XY-M} and D_{XY-N} values were centred around the mean and scaled by the standard deviation to show variations on the same scale. A neighbour-joining tree was built with D_{XY-M} values using San as an outgroup (Figure 5.20), and the populations in Figure 5.19 are ordered according to the tree. The heatmap is generally symmetrical: the pattern in Neanderthal segments largely mirrors that in unadmixed segments, forming major clusters separating Oceanian, American, East Asian and European-Central/South Asian populations. The neighbour-joining tree built from D_{XY-N} only differ from the unadmixed tree by two swapping of adjacent branches: between Xibo and Mongolian, and between Tuscan and the French-Orcadian clade. No additional structure was detected in the Neanderthal segments that was not present in the unadmixed segments of the genomes.

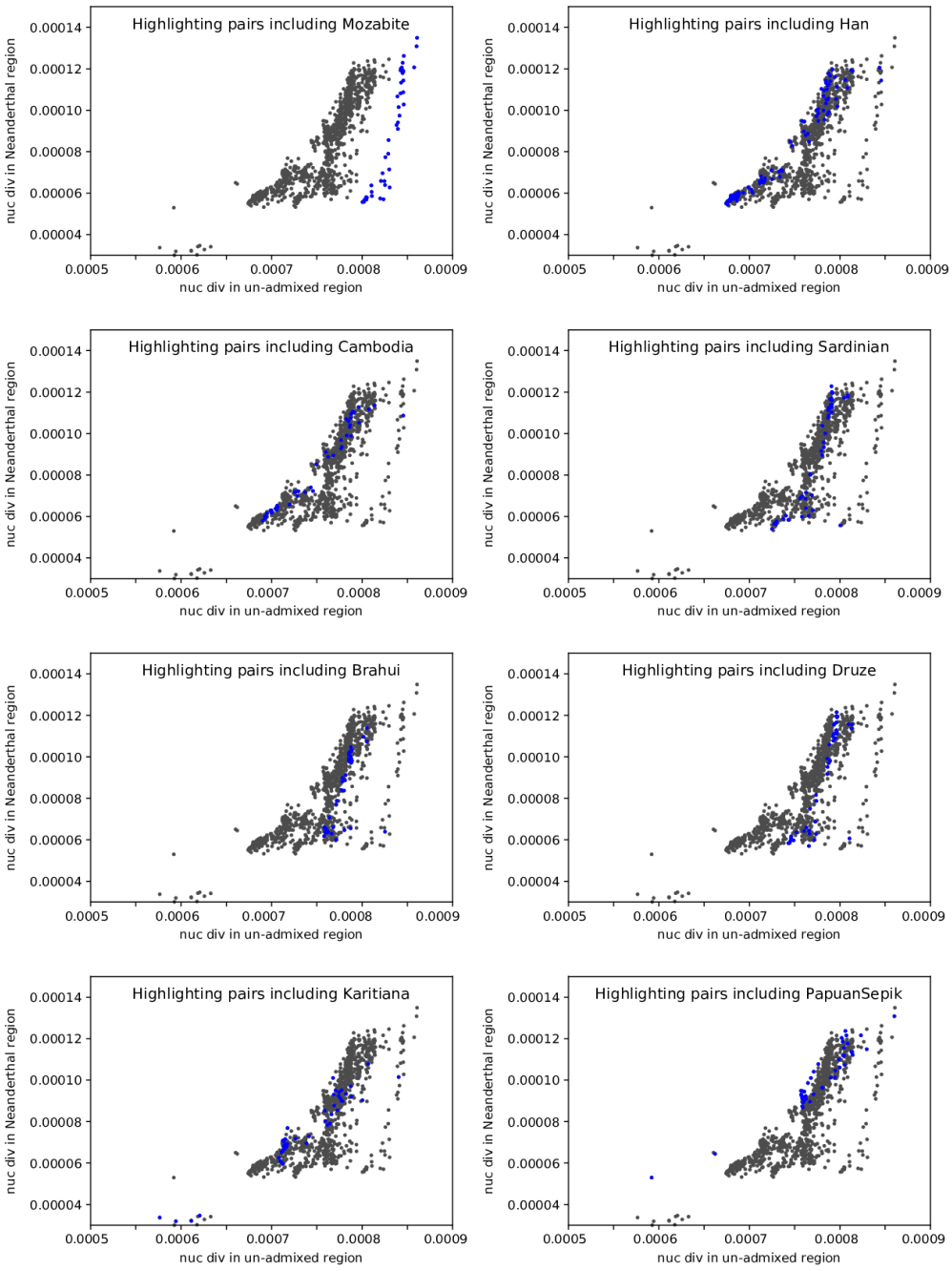


Fig. 5.17 Absolute divergence between pairs of populations in Neanderthal segments and unadmixed segments, highlighting pairs including 8 populations. The slope in European populations appears distinct from that in East Asian and American populations.

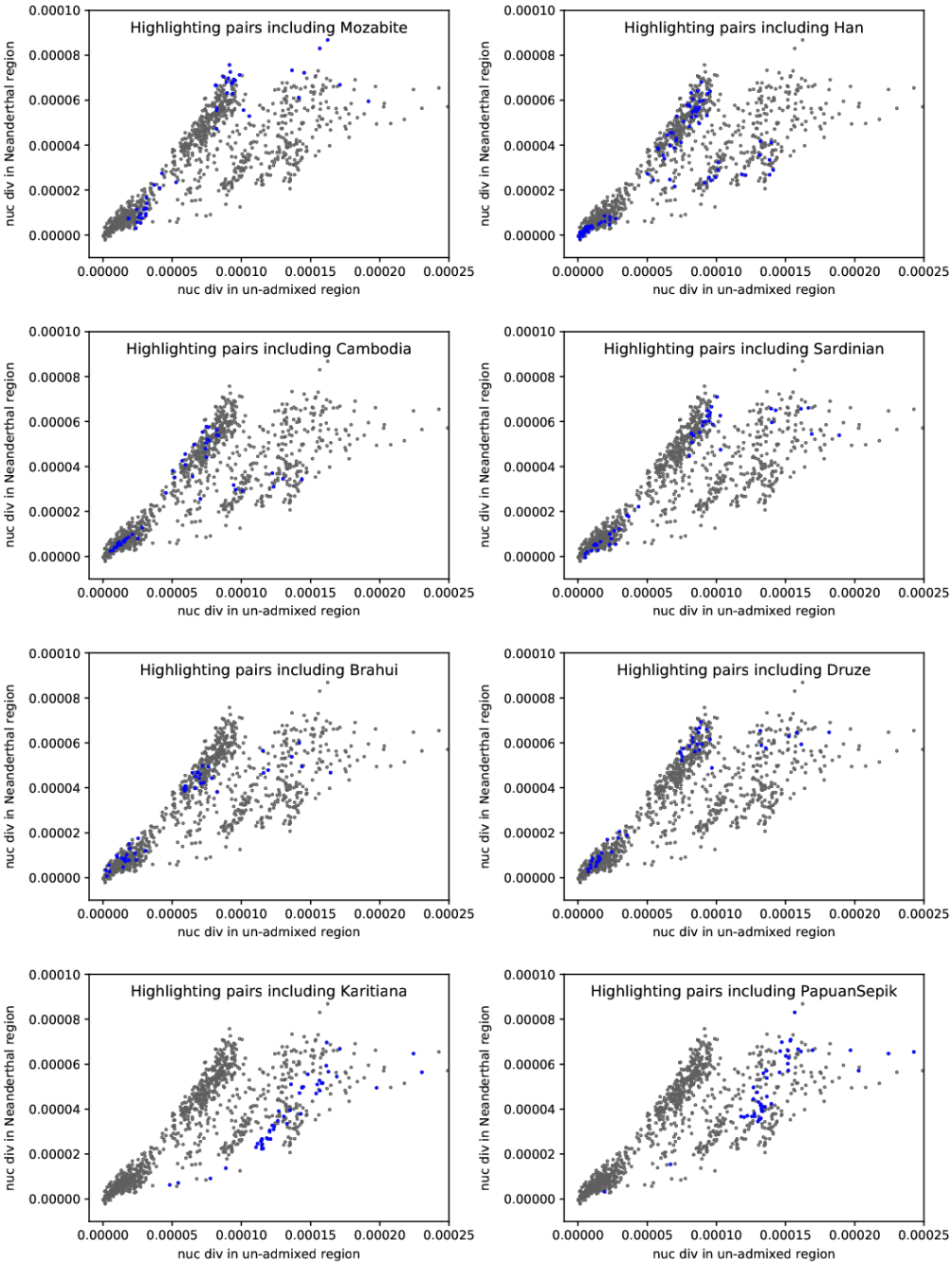


Fig. 5.18 Net nucleotide differences between pairs of populations in Neanderthal segments and unadmixed segments, highlighting pairs including 8 populations

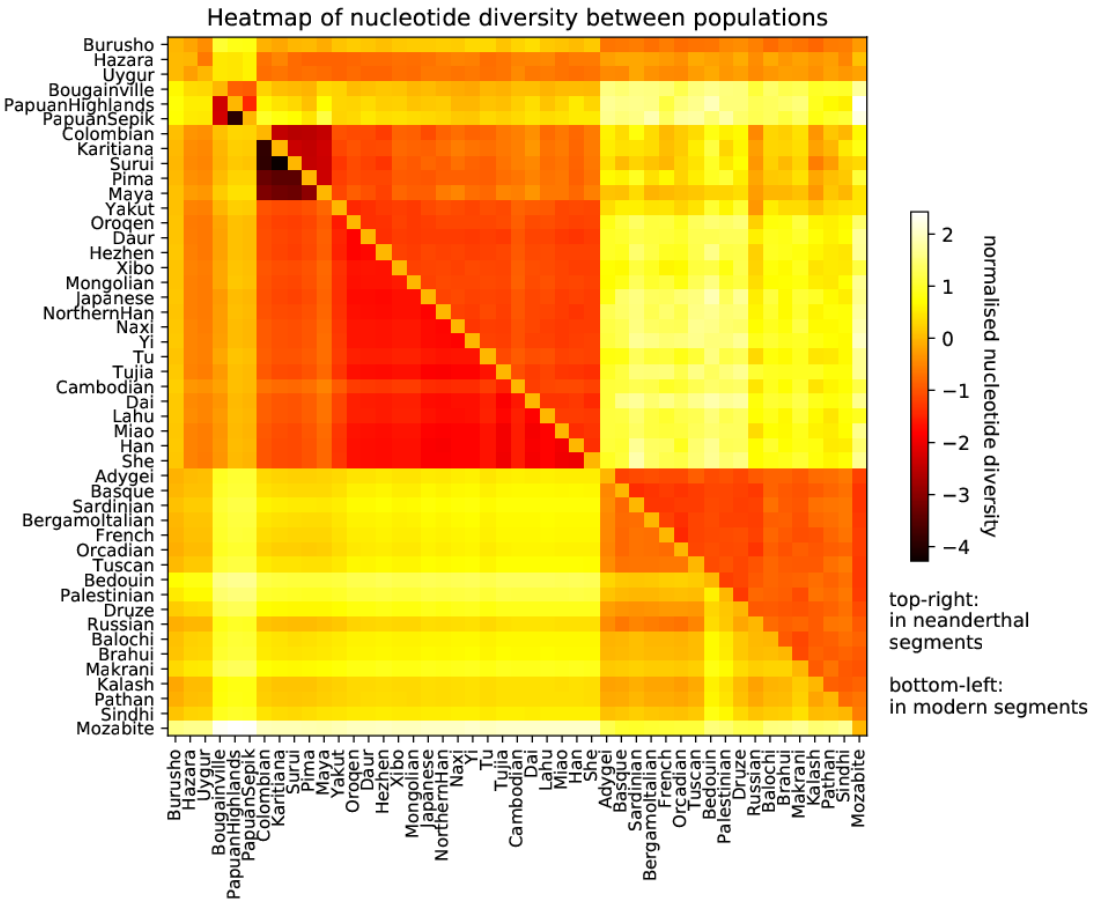


Fig. 5.19 Heat map comparing normalised D_{XY} measured in Neanderthal (top right) vs. unadmixed (bottom left) regions of the genome

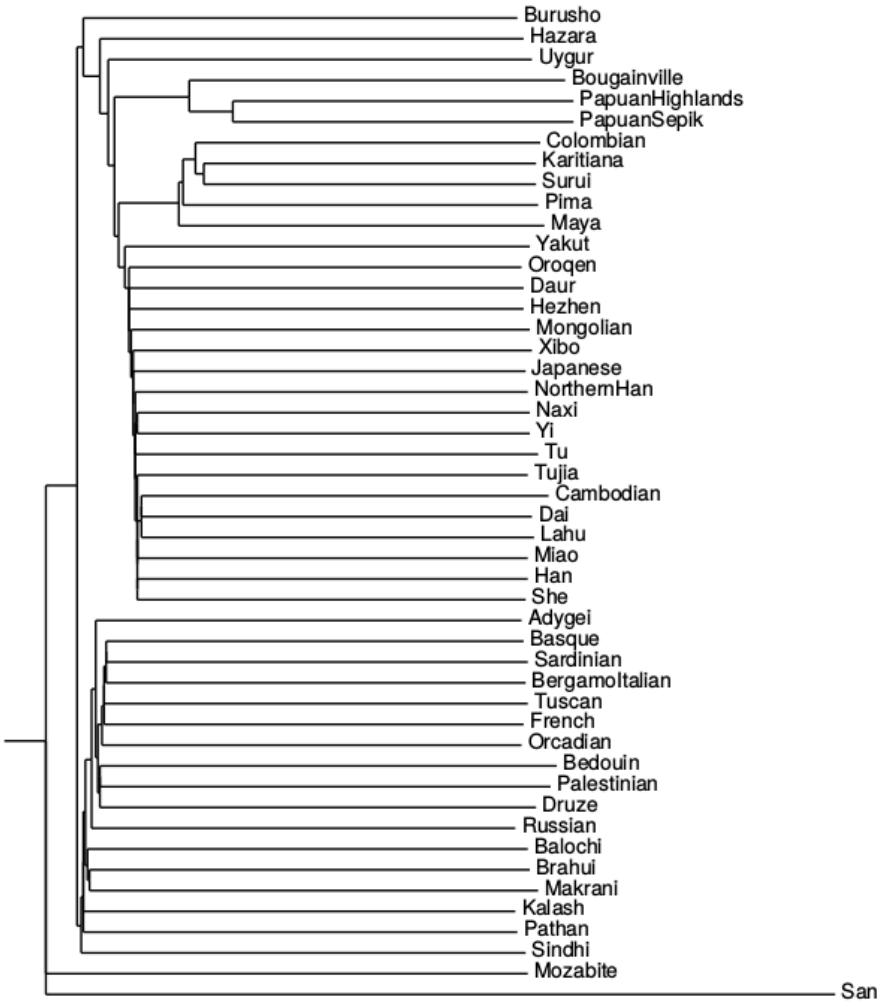


Fig. 5.20 Neighbour-joining tree built from D_{XY} measured in unadmixed regions of the genome, rooted by San as outgroup

The symmetry is broken on finer scales. If we look at the Neanderthal segments, D_{XY-N} between Mozabite and European/Middle Eastern populations appears lower than that between some Central/South Asian populations and the latter; in fact, it is almost as low as comparisons within Europe. But in the unadmixed segments, all Central/South Asian populations are closer to European/Middle Eastern ones than Mozabite. Most likely this is because the sub-Saharan African ancestry in Mozabite increases D_{XY-M} to Europe/Middle East, but the Neanderthal ancestry in Mozabite remains what was received from a shared source as Europe/Middle East, as no archaic admixture has been known to occur in sub-Saharan Africa. A similar process might have happened in Burusho, though it is less clear if that is a scaling artefact.

Notably, the cluster in the lower right corner, including populations from Europe, the Middle East and Central/South Asia (excluding three with high genetic affinity to East Asia), exhibits less genetic differentiation in Neanderthal segments. To get a better resolution in this area, I excluded the other populations and normalised the values again (Figure 5.21). The different pattern between D_{XY-M} and D_{XY-N} involving Mozabite and Europe/Middle East appears more striking on this scale. Now a European cluster can be distinguished, yet the relationship involving the Middle East and Central/South Asia is not well-defined. Many population pairs, such as Bedouin and Palestinian, and Sindhi and Makrani, show different relations to neighbouring groups in the Neanderthal and the unadmixed regions. A history of complicated population movement and admixture, especially when sources with no or very low level of Neanderthal ancestry (including sub-Saharan Africans and the proposed basal Eurasian lineage [101, 102]) were involved, could have contributed to the noise; in addition, a lower level of Neanderthal ancestry might also add to the variance seen in D_{XY-N}. Nevertheless, the fact that D_{XY-N} alone enables the reconstruction of the population tree with minimal change to Figure 5.20 indicates that the noise is localised.

In conclusion, the observed patterns of π and D_{XY} are consistent with our current understanding of the demographic history of modern human populations. Genetic diversity in Neanderthal segments is much lower than in unadmixed regions of the genome, but the pattern of its variation within and between populations largely mirrors the latter in a linear relationship. I found no evidence for distinct Neanderthal gene flows into the ancestors of one or more modern day populations. The most parsimonious explanation is that a single episode of admixture with Neanderthals occurred in the ancestral population of all non-Africans today; the diversity in Neanderthal segments is mainly shaped by the population history of modern human after the admixture event.

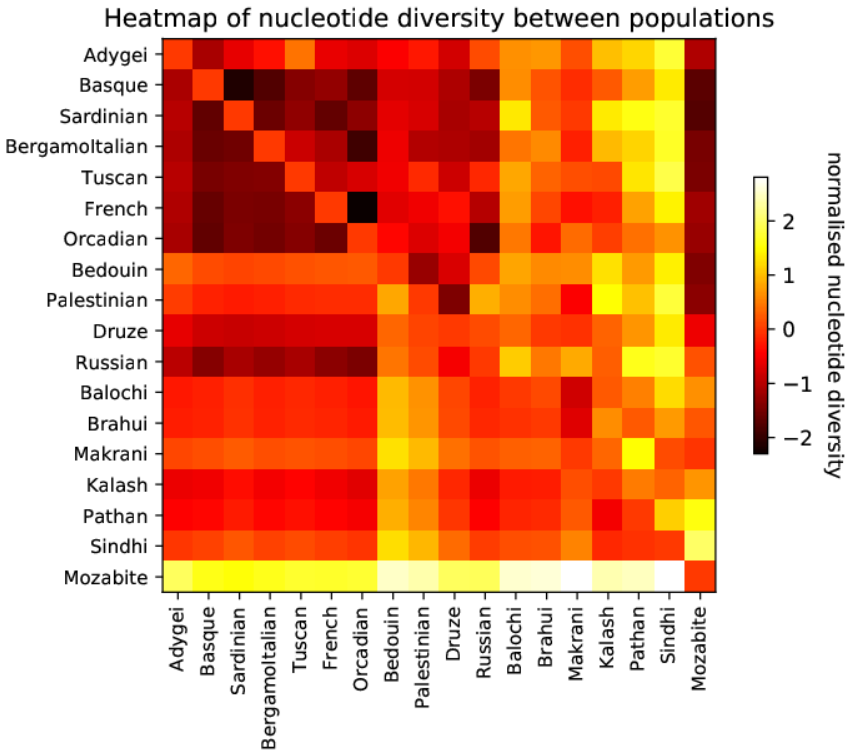


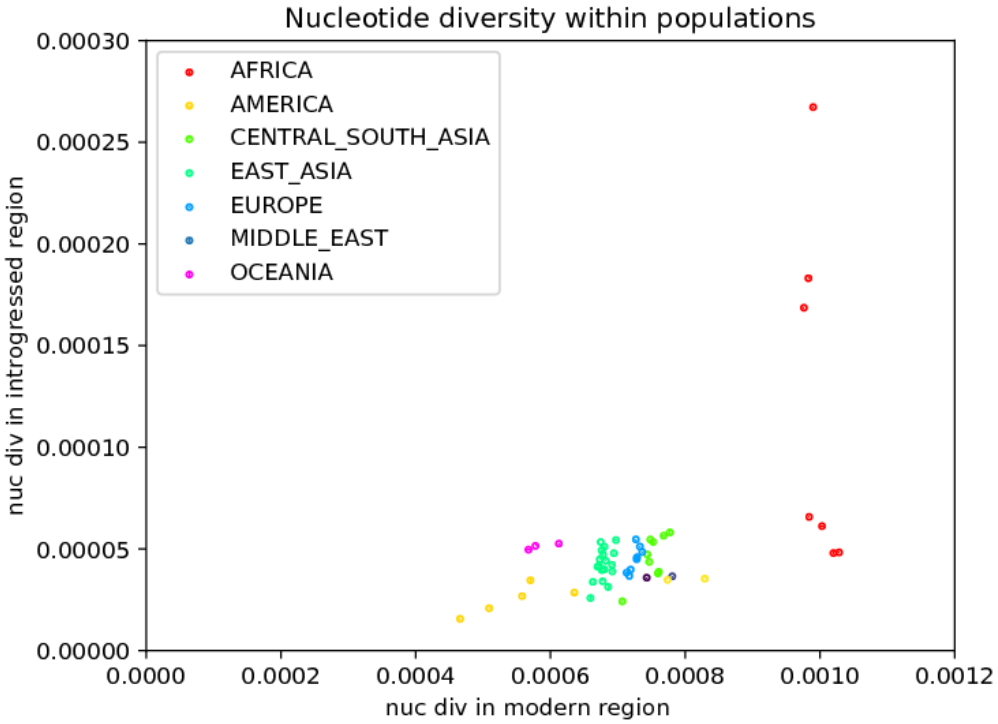
Fig. 5.21 Heat map comparing normalised D_{XY} measured in Neanderthal (top right) vs. unadmixed (bottom left) regions of the genome, showing a subset of the populations

Denisova vs. unadmixed regions

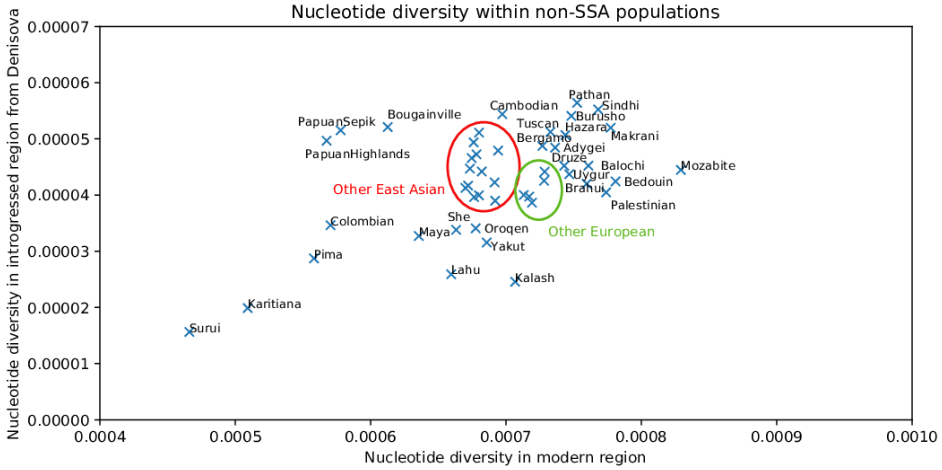
Similarly, I obtained the measurement of intra-population nucleotide diversity (π_D) and inter-population absolute divergence (D_{XY-D}) from recovered Denisovan segments in the HGDP genomes. The comparison between nucleotide diversity in Denisova (π_D) and in unadmixed modern human regions (π_M) is shown in Figure 5.22. The pattern in general is similar to Figure 5.16: a positive correlation between π_D and π_M ($R^2 = 0.3345$ with African populations, 0.1652 without), Mozabite and the Papuan populations being outliers at each end. The linear relationship is not as strong as in the case of Neanderthal, with many populations from Central/South Asia and the Middle East such as Kalash and Palestinian showing lower π_D than other populations with similar π_M values. π_D in the Papuans is higher than π_N in absolute value and also higher than π_D in adjacent populations. It is possible that similar to Figure 5.16b, misclassified Denisovan segments inflated π_D; however, the ratio of Neanderthal to Denisovan ancestry is many folds higher in East Asia, where π_D remains about the same level as π_N. If misclassification of archaic segments drives the increase in π, π_D should be much higher than π_N in East Asia, since assigning authentic Neanderthal segments to Denisova is much more likely to happen than the other way around. The increased values of π_D in Papuans, therefore, should reflect the true diversity in the Denisovan segments detected. Since the admixture time with Denisova is estimated to be more recent than with Neanderthal [76], fewer changes should have accumulated after the Denisovan segments entered modern human population. Hence the Denisovan segments in Papuans could trace back to a more diverse source than the Neanderthal segments, and also more diverse than in other regions of the world. A highly diverse source or multiple sources might have admixed with ancestors of Papuans, or many more Denisova lineages might have been lost from genetic drift in other populations.

Figure 5.23 and Figure 5.24 compare the inter-population absolute (D_{XY}) and net divergence δ_{XY} measured in the Denisova regions and unadmixed regions, highlighting pairs including the same populations as in Figure 5.17. Likewise, D_{XY-D} is positively correlated with D_{XY-M}. Adjusting for diversity in the ancestral population strengthens the linear relationship, and also makes the slopes in different population appear more similar.

However, the three Papuan populations clearly behave differently from the other populations. All the data points involving Papuans (except three comparisons between them) fall on the top of the plot (D_{XY-D}) above 0.00013; δ_{XY-D} above 0.0001), a region totally absent in the Neanderthal result (Figure 5.17 and Figure 5.18). In most other populations, points



(a) All populations



(b) Excluding sub-Saharan Africans

Fig. 5.22 Intra-population nucleotide diversity in Denisovan segments and unadmixed segments

representing the divergence to Papuans are disconnected from the others, causing a upward bent in the trend lines.

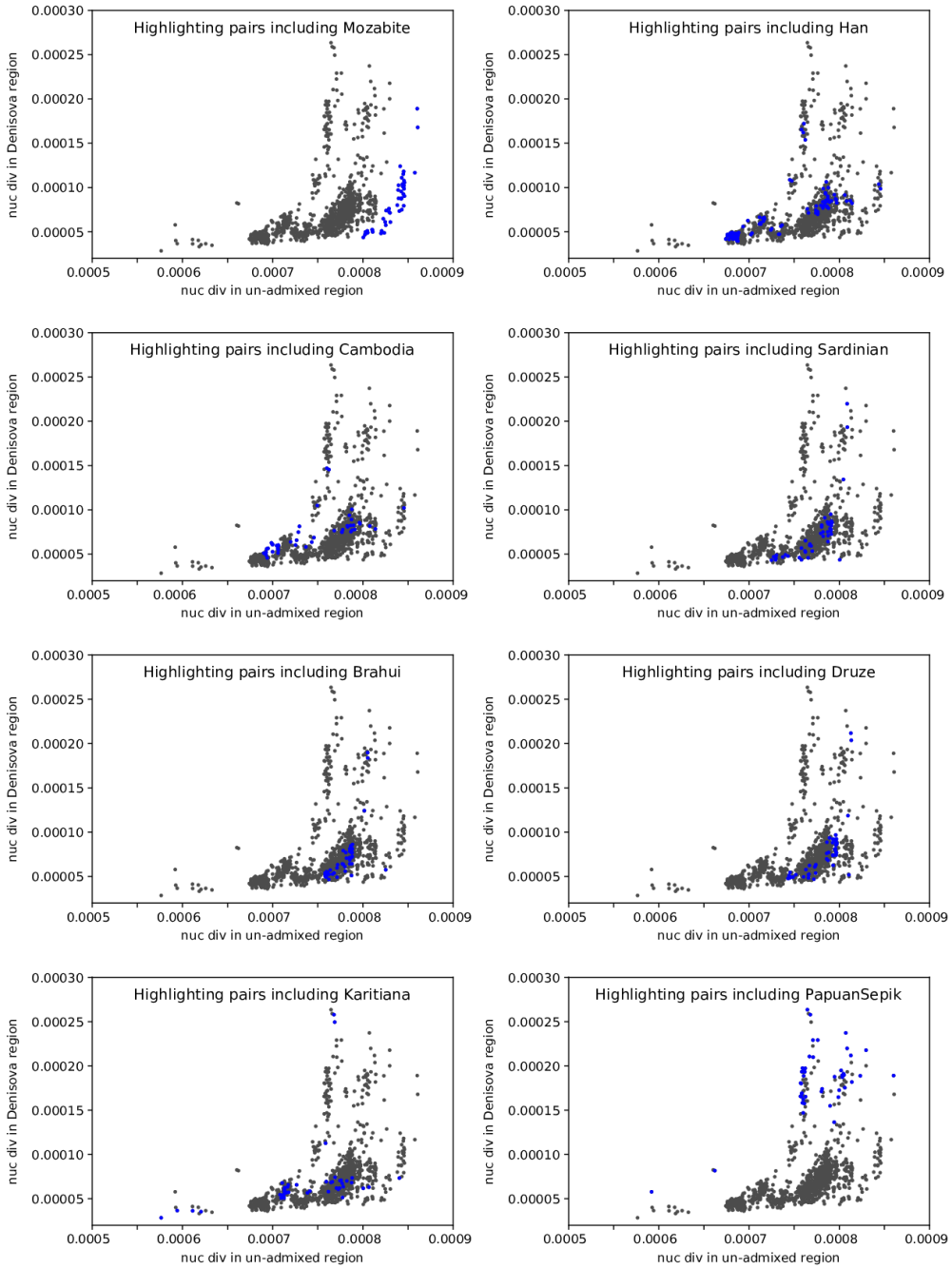


Fig. 5.23 Absolute divergence between pairs of populations in Denisovan segments and unadmixed segments, highlighting pairs including 8 populations

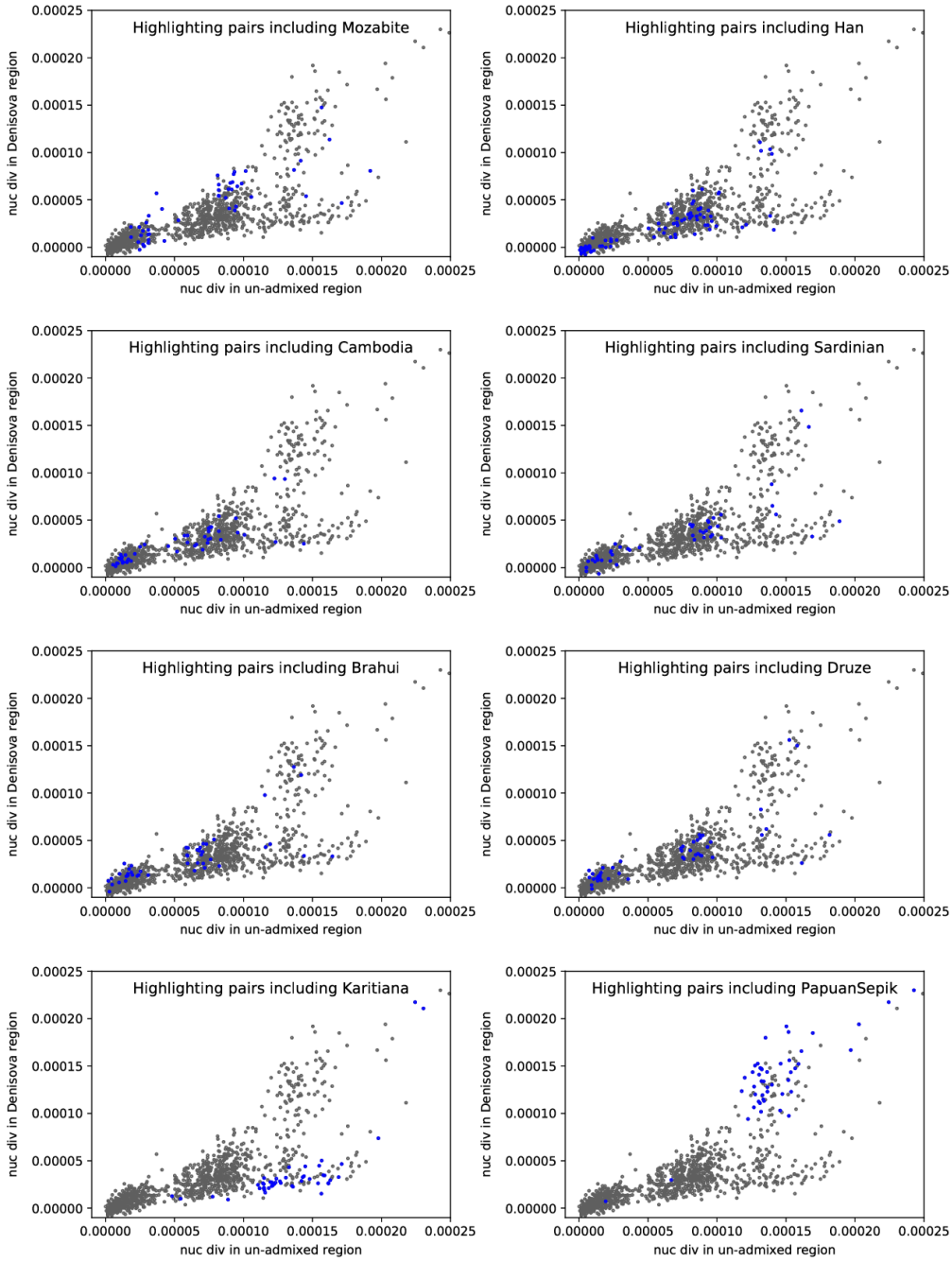


Fig. 5.24 Net nucleotide differences between pairs of populations in Denisovan segments and unadmixed segments, highlighting pairs including 8 populations

The contrast is more striking in the normalised heat map (Figure 5.25). The three Papuan populations form a sister clade to all East Asian and American populations in unadmixed regions of the genome, but exhibit very high divergence to all other populations if only the Denisova regions are concerned. Their D_{XY-D} to other populations also appears similar, with only a faint affinity to East Asian populations. In comparison to Papuan Highlands and Papuan Sepik, D_{XY-D} between Bougainville and non-Papuan ("mainland") populations appears lower. An unrooted neighbour-joining tree reconstructed from the D_{XY-D} matrix also places the Papuans along an extended branch diverging from all other populations (Figure 5.26). The extraordinarily high divergence of Denisovan segments in Papuans to those in all other populations strongly supports a separate source of Denisova gene flow into Oceania.

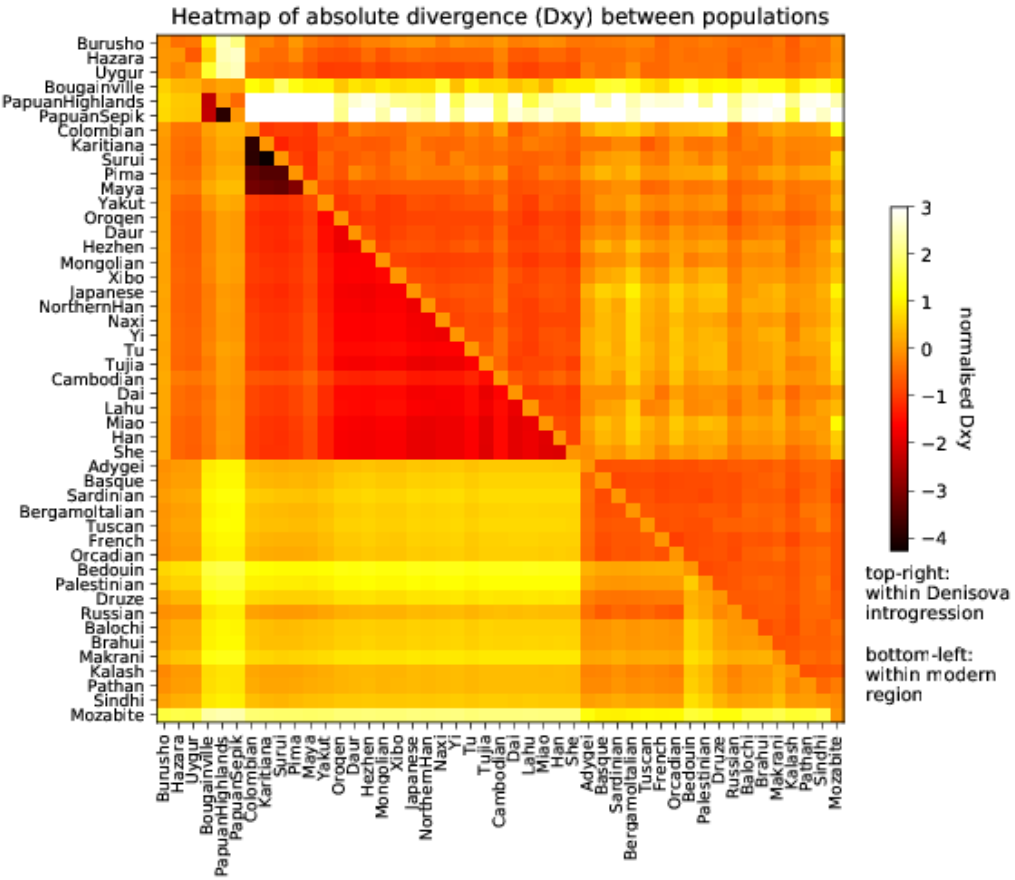


Fig. 5.25 Heat map comparing normalised D_{XY} measured in Denisova (top right) vs. unadmixed (bottom left) regions of the genome

In the rest of the heat map, Mozabite again shows higher affinity to European and Central/South Asian populations in Denisova regions than in unadmixed regions. The differentia-

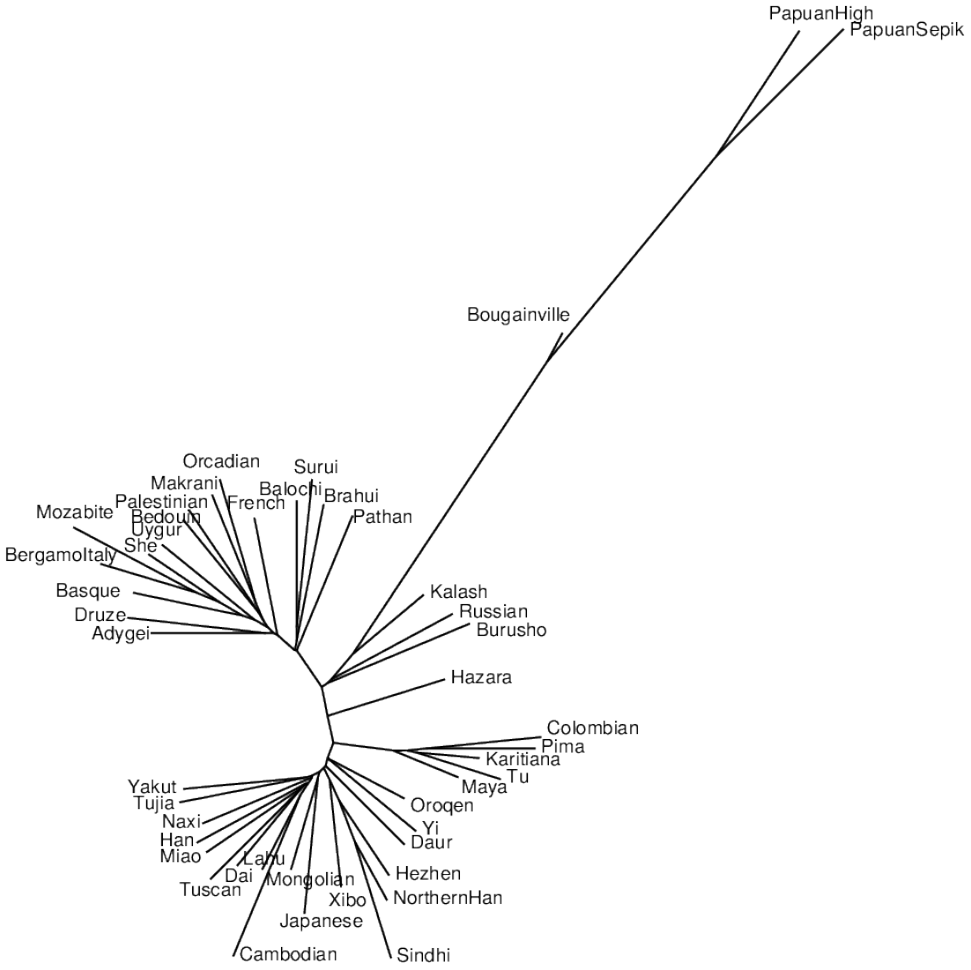


Fig. 5.26 Unrooted neighbour-joining tree built from D_{XY} measured in Denisova regions of the genome

tion within the East Asia and America clade appears low in Figure 5.25, so I plotted the heat map within this region after rescaling (Figure 5.27). Without over-interpreting the random fluctuations in individual cases, which is expected to be stronger on the Denisova side due to the scarcity of comparable segments, it is worth mentioning two types of pattern. Higher affinity to some groups in the Denisova regions than in the unadmixed regions could result from recent admixture with a source further diverged from these groups but low in Denisovan ancestry, similar to the case with Mozabite. This might have happened in Mongolia and Maya in relation to East Asian populations. The reverse trend, however, can be seen in Cambodia, whose divergence to other East Asian populations is higher in the Denisovan segments than in unadmixed regions (Figure 5.27). Scaling might have an effect, yet the Denisovan segments in Cambodia do not show higher affinity to those in East Asia over those in America, as the population history should suggest; meanwhile, they show higher affinity in comparison to other East Asian populations to the Denisovan segments in the Papuans (Figure 5.25). In the population tree built from D_{XY-D} (Figure 5.26), Cambodia also ends up on an extended branch in relation to other East Asian and American populations. Similar to the more prominent case of the Papuans, the evidence may suggest the presence of a different source of Denisovan ancestry in Cambodia, with a tentative connection to that in Oceania consistent. The possibility will be revisited in the discussion on haplotype network structure (Section 5.7.3).

In contrast to the tree built from D_{XY-N}, which accurately reflects the population tree (Figure 5.20) with few changes, the D_{XY-D} tree (Figure 5.26) deviates a lot from the population tree even if we ignore structures within the same geographical regions. For example, Tuscan does not cluster with the other European populations, but invades the East Asian clade instead; Tu appears within the American clade; She and Uygur are located amidst European and Middle Eastern populations. The relatively low level of Denisovan ancestry in most parts of the world certainly adds much noise to inter-population comparisons, and those remaining today also underwent strong drift since the Denisova admixture. Nevertheless, together with the uneven distribution of Denisovan ancestry worldwide, the pattern of D_{XY-D} suggests that the gene flow from Denisova did not sweep through all ancestors of non-Africans rapidly as happened in the admixture with Neanderthal; instead, Denisovan ancestry in many parts of the world was possibly acquired and defined gradually through complicated dynamics of drift, migration, and admixture in modern human populations.

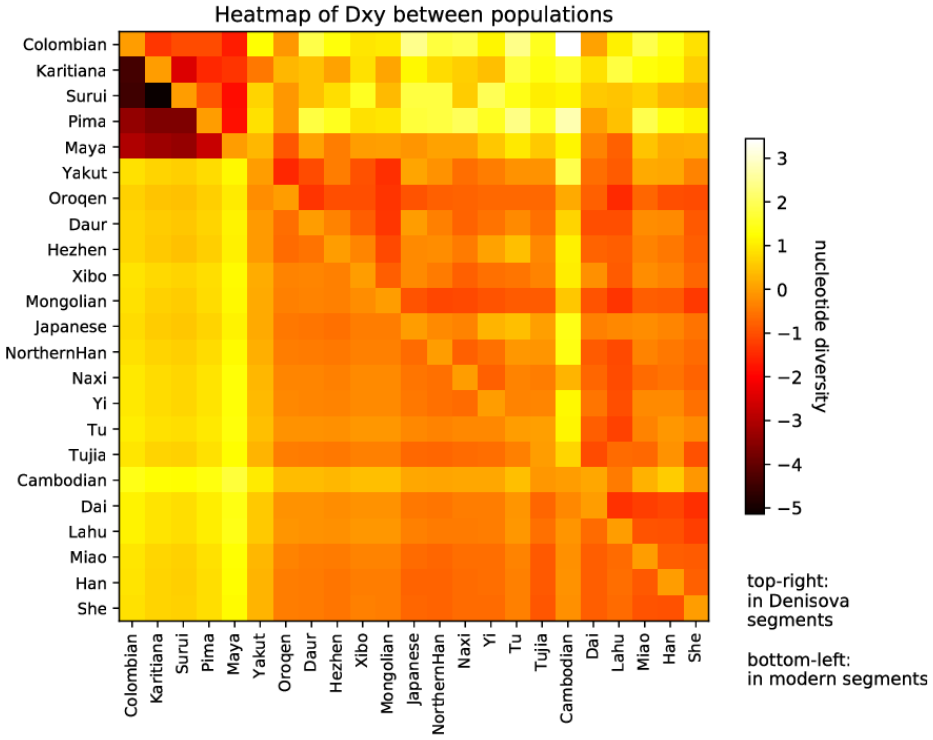
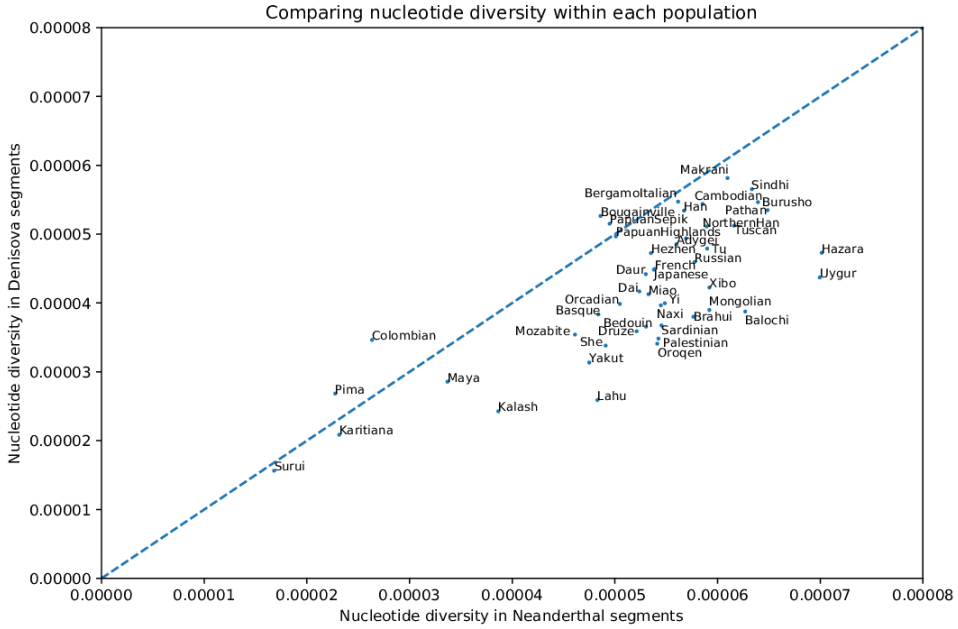


Fig. 5.27 Heat map comparing normalised D_{XY} measured in Denisova (top right) vs. unadmixed (bottom left) regions of the genome, showing the East Asia and America clade

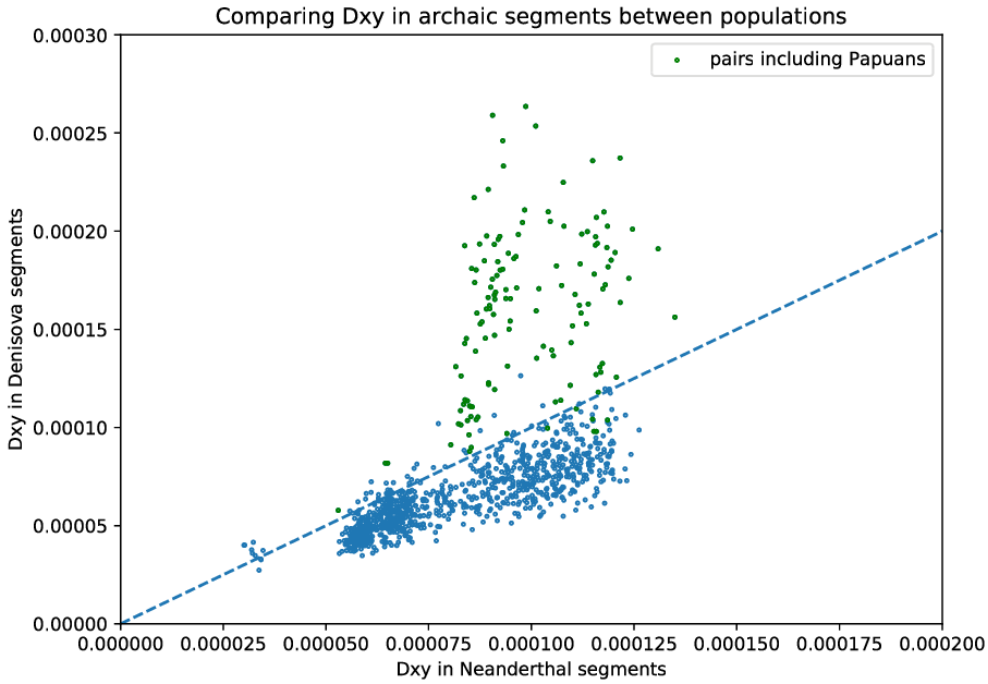
Neanderthal vs. Denisova regions

Finally, Figure 5.28 and Figure 5.29 directly compares intra- and inter-population divergence in Neanderthal and Denisovan segments, other than comparing each to values in the unadmixed regions. In general, π_N and π_D are strongly correlated, and π_D is lower than π_N in most populations, with the exception of two American populations and two Papuan populations. Assuming comparable genetic diversity in incoming sequences, higher nucleotide diversity could reflect that the admixture with Neanderthal happened longer ago; however the possibility of more than one episode of admixture or distinct population sizes in Neanderthal and Denisova readily breaks down the assumption. Figure 5.29 represents yet another visualisation of the distinct Denisovan ancestry in Papua: in all comparisons excluding Papuans, we observe again a strong correlation between D_{XY-D} and D_{XY-N}, with the former typically less than the latter; in contrast, almost all comparisons including Papuan populations show higher D_{XY-D} than D_{XY-N}, and deviate from the linear relationship defined by other points.



Dashed line: x = y

Fig. 5.28 Nucleotide diversity (π) in all non-African populations measured in Denisova vs. Neanderthal haplotypes of the genome



Dashed line: x = y

Fig. 5.29 Absolute divergence (D_{XY}) between all pairs of non-African populations measured in Denisova vs. Neanderthal regions of the genome

Overall, comparisons between the worldwide pattern of divergence in Neanderthal, Denisova and unadmixed regions of the genome expose striking differences between the structure of the two archaic admixture events. The intra- and inter-population diversity in Neanderthal segments mainly mirrors that in the unadmixed regions, suggesting that all populations out of Africa received a single shared episode of gene flow from the Neanderthals before they diverge from each other. In the Denisovan segments, the Papuan populations show higher nucleotide diversity within themselves, as well as considerably higher divergences to all other non-African populations. The Denisovan ancestry in Oceania therefore most likely result from a separate admixture event, with a genetically distinct Denisova population from the one(s) admixing with mainland populations. Moreover, the genetic structure in Denisovan segments does not always agree with the population history. In combination with the uneven distribution of Denisovan ancestry around the world (Figure 5.3), it suggests that the initial Denisova gene flow did not reach all non-African populations; instead, some regions acquired Denisovan ancestry through later admixtures with other modern human populations, which causes different relatedness as measured in the Denisova versus unadmixed regions.

5.6.2 Archaic site frequency spectrum

The site frequency spectrum (SFS, also known as allele frequency spectrum) is the distribution of the frequencies of derived alleles across polymorphic sites in a population. As a summary of the genetic variation at unlinked sites, it is an informative statistic frequently used in population genetics inference. The large amount of archaic sequences detected in the HGDP dataset means that multiple overlapping archaic haplotypes exist in many regions of the genome, making it possible to reliably obtain an SFS for derived alleles present in the archaic regions. Such SFS should have been shaped by the demographic history of the archaic populations before the admixture, and the demographic history of modern human populations afterwards.

Model-based demographic inference was performed using the software package fastsimcoal2, which uses efficient simulations to estimate the expected SFS under a specific demographic model, and fits the observed SFS using maximum composite likelihood [166]. I looked at the SFS in Neanderthal segments in non-African genomes and the SFS in Denisovan segments in Oceanian genomes. Because the number of comparable archaic alleles varies across sites, only sites with a sufficient number of archaic alleles were considered, and the allele frequencies were rounded to a fixed sample size.

The information content in the SFS is limited by both the sample size and the number of sites; when neither is ideal regarding archaic alleles, the inference accuracy might be poor for complex demographic models. Therefore I fitted a simple three-epoch, piecewise constant population model, where the most recent population size is fixed according to estimates for modern human populations, preceded by an introgression-related bottleneck of size N_{bot} that started at T_{bot} generations ago and lasted for 100 generations, and an ancestral population of size N_{anc} prior to the admixture (Figure 5.30). Table 5.3 summarises the inference results for the three parameters from all models.

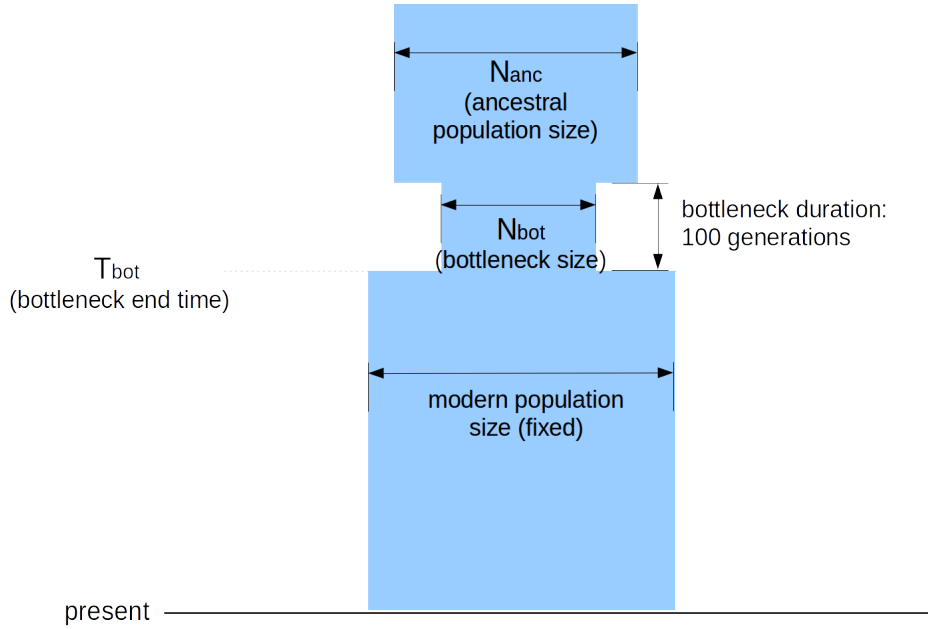


Fig. 5.30 Demographic model and parameters used in fastsimcoal2 analyses

Table 5.3 Summary of Fastsimcoal2 inference results

Scenario	**Estimated parameters**		
	T_{bot}	N_{bot}	N_{anc}
Non-African Neanderthal SFS	1102	50	17333
Non-African Neanderthal SFS, lower bound on T_{bot}	1673	48	23947
Oceanian Denisova SFS	395	89	11119
Oceanian modern SFS	727	371	21465

Neanderthal SFS in non-African genomes

First, I treated Neanderthal segments from all non-African individuals as in one population. Because the total number of archaic haplotypes changes along the genome, the observed allele frequency was used to calculate the expected number of allele counts in a haploid population of size 20. Only polymorphic sites spanned by at least 40 Neanderthal haplotypes (from the "strict" set of results) were examined. Sites whose ancestral state can be determined from neither the EPO multiple primates alignment nor the chimp genome were ignored, and so were those not passing the HGDP mask. Since linked selection could have reduced genetic diversity near functional regions, I also excluded sites whose B value falls below 0.8. Figure 5.31 shows the observed SFS in Neanderthal segments in non-African individuals against the expected SFS under neutral evolution. Compared to the neutral model, the SFS show an excess of singletons and high frequency variants that could have been shaped by a sharp bottleneck. Values of T_{bot}, N_{bot} and N_{anc} were estimated using fastsimcoal2.6

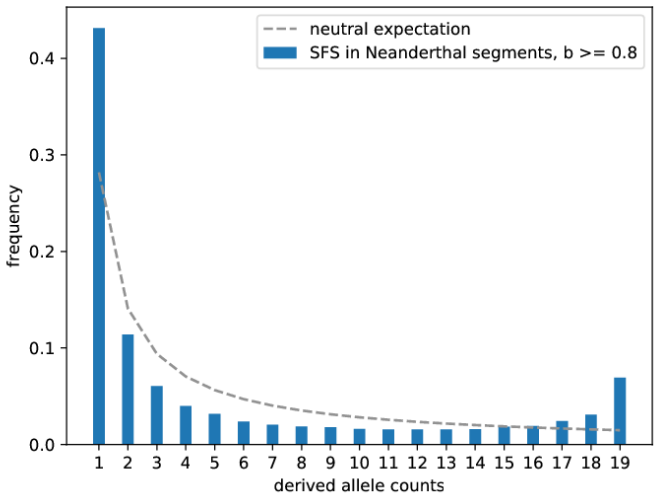


Fig. 5.31 Site frequency spectrum in Neanderthal segments outside of Africa rescaled to 20 haplotypes

following the demographic model in Figure 5.30. The present-day population size for all non-Africans was fixed at 10,000. I collected results from 100 independent runs each containing 100 expectation/conditional maximization cycles to find the estimation with the highest likelihood.

Initially, a bottleneck of size 50 was estimated to have ended 1,102 generations ago in an ancestral population of size 17,333. As expected from previous analyses, the small bottleneck

size reflects limited genetic diversity in extant Neanderthal lineages in modern genomes. The time of the bottleneck is much more recent than the estimated time of Neanderthal admixture (although the value coincides with estimates from [76] and [95]). If the time of admixture is indeed older, this bottleneck would correspond to a demographic event within modern humans. However, a number of factors not accounted for in the model, including linkage disequilibrium, recent population expansion, and selection against Neanderthal segments could make the bottleneck appear more recent.

To obtain a better estimation of N_{bot} incorporating prior knowledge about the time of Neanderthal admixture, I added a lower bound of 1,600 generations ago to T_{bot}. With this constraint, fastsimcoal2 inferred a bottleneck size of 48 ended 1,673 generations ago. The estimation of N_{bot} remains similar despite the change in T_{bot}. The results also suggest that the ancestral Neanderthal population could be much larger than a group size of 3,000 estimated from the heterozygosity of the sequenced Neanderthal genomes [66, 67], which might come from small inbred groups.

Denisova SFS in Oceanian genomes

All Denisovan segments from Oceania were also modelled as in one population. As only 28 genomes were available, the number of entries in the SFS and the minimum number of haplotypes were both set to 10. The present-day effective population size was fixed at 5,000. Figure 5.32 shows the observed SFS, with a similar excess of singletons and near-fixed alleles as the Neanderthal SFS. Based on this fastsimcoal2 inferred a bottleneck of size 89 happening 395 generations ago. This date is much more recent than the admixture time of 44-54k years ago estimated previously [76], and in fact falls within the range of lowland-highland division, the most prominent structure in the Papuan populations today [167]. If the bottleneck reflects an event within modern humans, such as the founding of separate Papuan populations, the signal should also appear in the unadmixed regions of the genome. Therefore I repeated the same inference using SFS from confident modern regions of the genome (the "strict" set of results). In this case, a milder bottleneck of size 371 was estimated at 727 generations ago, close to the estimated lower bound of divergence time between Papuans and aboriginal Australians [98]. It is not clear why the Denisova and the modern region SFS would suggest distinct events in the recent past. Perhaps the sharp division between Papuan populations means that Denisova haplotypes at a specific location mostly come from the same population, causing the Denisova SFS to be more affected by the structure within Oceania. Alternatively, the bottleneck signals might reflect the same

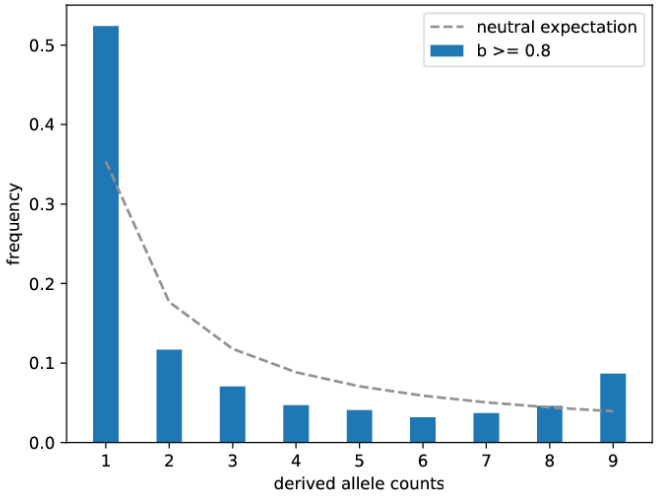


Fig. 5.32 Site frequency spectrum in Denisovan segments in Oceania rescaled to 10 haplotypes

underlying event, but the collection of Denisovan segments detected in HGDP genomes is insufficient for SFS-based inference.

Neanderthal JSFS in European and East Asian genomes

The joint site frequency spectrum (JSFS) is an extention of the SFS to multiple populations, which specifies the joint distribution of allele frequencies in different populations. Similarly, the JSFS is informative about the demographic history involving multiple populations, including population divergence time and subsequent migration between them. Since part of the Neanderthal ancestry in East Asia and in Europe have been proposed to come from separate admixture events [91, 104], I also examined the Neanderthal JSFS in European and East Asian genomes.

Polymorphic sites were subject to similar filters on passing the HGDP mask, known ancestral state and high B value as before. Only those covered by at least 16 Neanderthal haplotypes in Europe (out of 155 European genomes) and 20 Neanderthal haplotypes in East Asia (out of 223 East Asian genomes) were included, and the frequency was used to calculate the expected allele count in a population of 16 European haplotypes and 20 East Asian haplotypes. The resulting JSFS is shown in Figure 5.33. Similar to the one population case, an excess of singletons and near-fixed alleles appears in both populations. However, I observed that the marginal SFS in both populations no longer appear smooth (Figure 5.34). This could be

caused by insufficient number of polymorphic sites, since Neanderthal segments at the same genomic location are not likely to reach the required frequency in both Europe and East Asia, on top of the other filters. Fastsimcoal is unlikely to generate reliable result on such data; nevertheless, the excess of private alleles in each population might support separate Neanderthal admixture events in Europe and East Asia, although other lines of evidence do not.

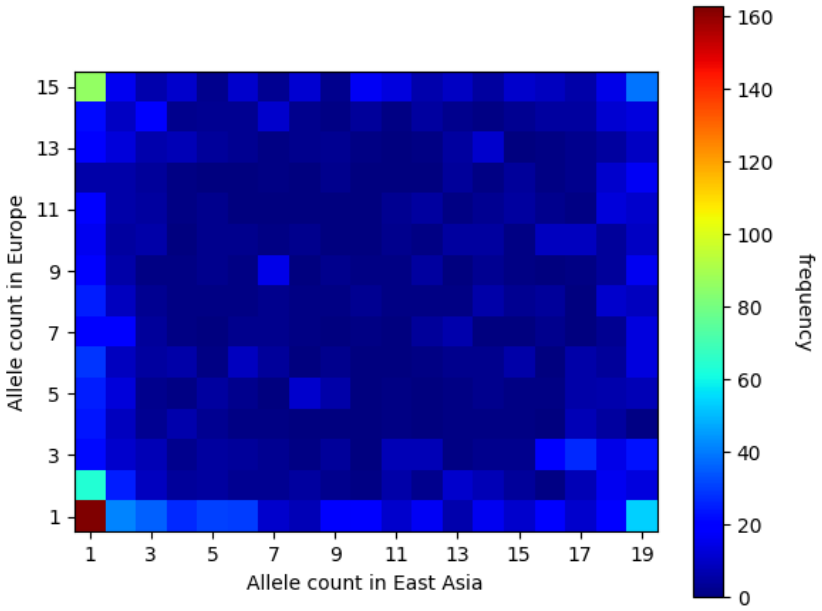


Fig. 5.33 Joint site frequency spectrum in Neanderthal segments rescaled to 16 European and 20 East Asian haplotypes

5.7 Archaic haplotype networks

If more than one episode of gene flow has occurred between modern human and genetically distinct Neanderthal/Denisova populations, archaic haplotypes from different source populations might be present even within the same genome. In addition to measuring the divergence to archaic genomes in each segment (discussed in Section 5.5.2), it is also of interest to construct the relationship between all the archaic segments located in the same genomic region. After controlling for recent demographic history in modern humans, distinct clusters could suggest different sources of gene flow. Such comparison is only made possible by the size and diversity of the HGDP panel.

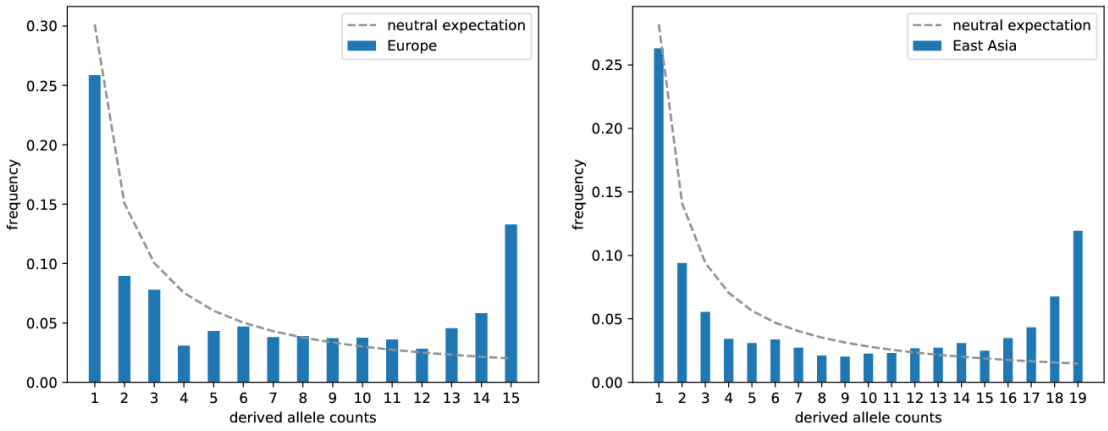


Fig. 5.34 Marginal site frequency spectrum in Neanderthal segments calculated from the joint site frequency spectrum in Figure 5.33

Assuming two archaic sequences descend from the same ancestral sequence at the time of admixture, after 2000 generations, one would only expect to observe one difference per 20kB given a mutation rate of 1.25×10^{-8} /(site $\cdot$ generation). Long regions covered by as many archaic haplotypes as possible are necessary to achieve a reasonable resolution. I searched for candidate regions in the genome by the following procedures.

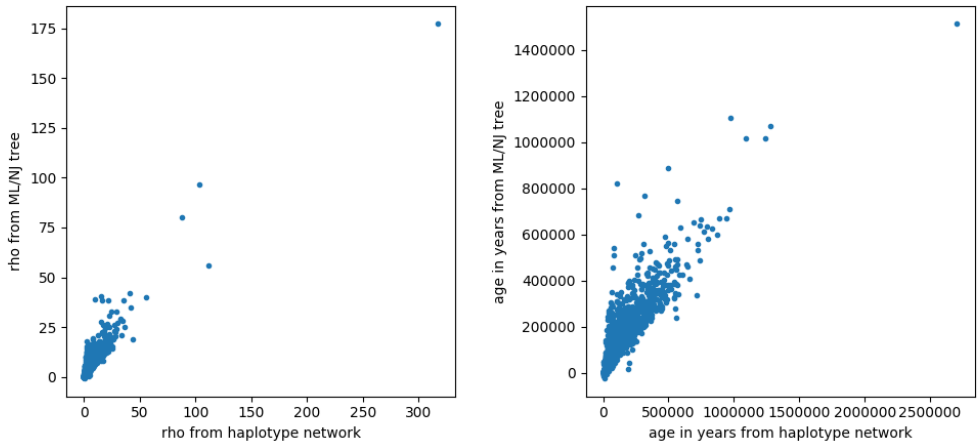
1. a multiple intersection of all Neanderthal/Denisovan segments (the "strict" set of result) on a particular chromosome was performed using multiIntersectBed from BEDTools [168] to obtain the total number of archaic haplotypes at any genomic interval;
2. I scanned through the list of intervals and added new intervals from merging adjacent ones if a subset of samples are present in both;
3. A score is assigned to each interval based on the length (L) and the number of samples (n):

   $$s = n^w \cdot L$$

   where w can be tuned to adjust the weight of including more haplotypes over extending the genomic interval; here I fixed it at 1, hence the score equals the total length of archaic sequences in the interval;
4. Intervals shorter than 50kB or with fewer than 5 samples were removed;
5. Non-overlapping intervals with the highest scores were collected following a greedy algorithm: within a minimal candidate set of intervals that do not overlap with any others, the interval with the highest score is selected and moved to the selected set,

and any other intervals that overlap with it removed from the candidate set; next to be selected is the interval with the highest score among the remaining ones in the candidate set; the process repeats until no interval remains in the candidate set, and the algorithm moves on to the next set of intervals.

I considered two methods to study the evolutionary relationship between aligned archaic sequences, namely phylogenetic tree building and haplotype network analysis. More work has been done on phylogenetic methods to incorporate complex molecular evolution models, but when very few base differences exist between haplotypes, haplotype networks have the advantage of allowing alternative links other than imposing a bifurcation tree with high uncertainty. Haplotype networks can also capture recombination events between the haplotypes, although it should be very rare for recombination to happen between two archaic haplotypes in the modern human population. The median joining network algorithm (implemented in the pegas package [169] in R) starts with a minimum spanning network, followed by adding consensus sequences (median vectors) of three closely located sequences which might represent unsampled sequences or extinct ancestral haplotypes [170]. Preliminary analysis also shows that the time to the most recent common ancestor (tMRCA) estimated from haplotype network analysis and maximum likelihood trees are highly correlated, although alternative links in the network tend to reduce tMRCA, especially when the sample size is large (Figure 5.35).



(Neighbour-joining tree was built when maximum likelihood tree took too long to compute)

Fig. 5.35 Comparison of tMRCA in number of mutations and in years measured in haplotype network and in phylogenetic tree

In each selected genomic region, polymorphic sites in all archaic haplotypes were retrieved to form a sequence alignment for haplotype network analysis. If a site has a singleton variant also present at a frequency above 0.01 in African genomes, it was deemed likely to have arisen from phasing errors and was ignored.

A total of 4,153 Neanderthal haplotype networks and 727 Denisova ones were reconstructed using Neanderthal/Denisovan segments from all non-African genomes. A few samples are shown in Figure 5.36 and Figure 5.37, and more are included in Appendix B. The size of the networks ranges between 3 and over 200 nodes. In some networks, haplotypes from the same geographical region group together (e.g. Figure 5.37a), but in others identical haplotypes can be found across distant regions (e.g. Figure 5.36b and 5.37b). The Neanderthal haplotypes do not obviously fall into separate clusters, as would be expected from multiple admixture events, whilst Denisova haplotypes in Oceania are often separated from those in other geographical regions. Some quantitative measurements, however, are necessary to extract patterns from this collection of networks.

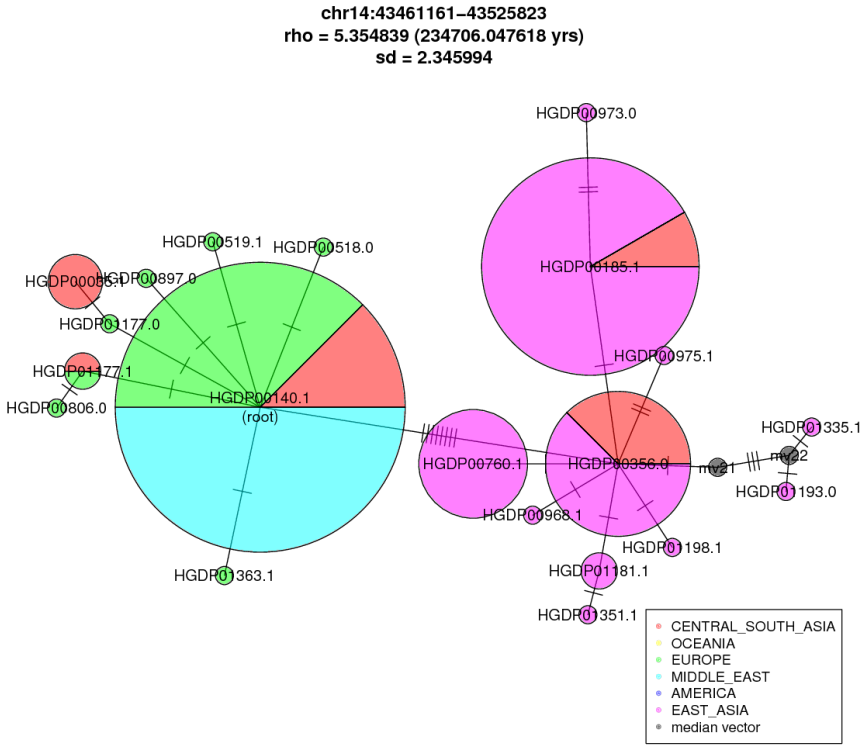
5.7.1 Age of archaic haplotype network

One outstanding question regarding the encounter between modern and archaic human populations is whether the diversity of archaic ancestry found in modern human derives from many archaic individuals, a few or even just one. To estimate the number of founding lineages contributing to extant haplotypes, I calculated the age (ρ) of each network (equivalent to tMRCA, or the height of a phylogenetic tree). The haplotype closest to the Vindija Altai genome in the Neanderthal haplotype network or closest to the Altai Denisova in the Denisova haplotype network was assumed to be the root node. In case of ties, the age of the network was calculated for all candidate root nodes, and the smallest value was retained. It is possible that the true founding haplotype in some networks is more distant but no longer exists, in which case the age here might be underestimated. Following [171], ρ is measured as the average shortest distance from all nodes to the root:

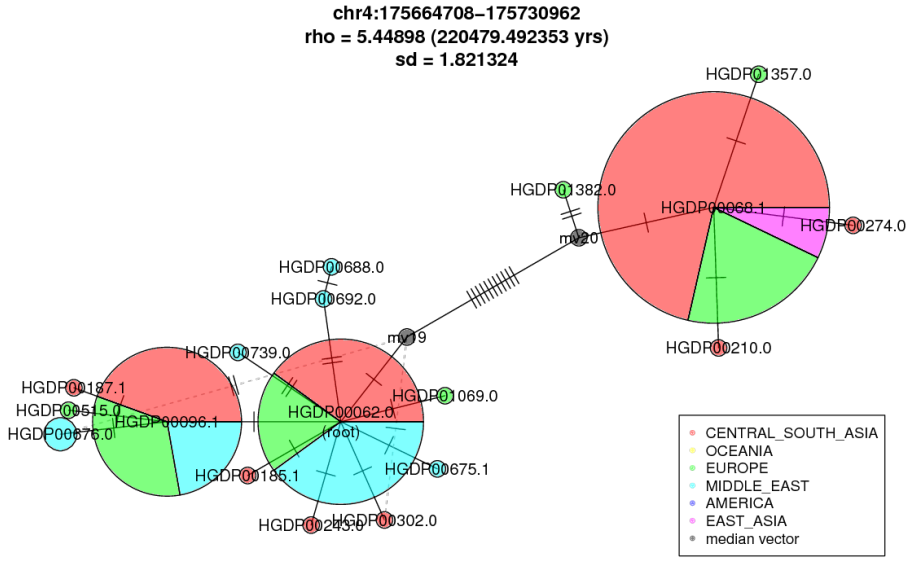
$$\rho = \frac{1}{n}\sum_{i=1}^{m} n_i l_i$$

and the variance

$$\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{m} n_i^2 l_i$$



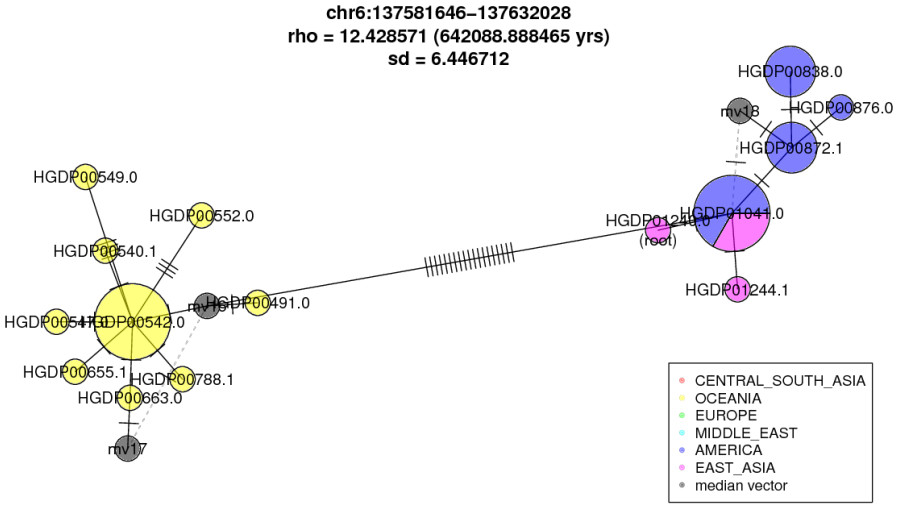
(a)



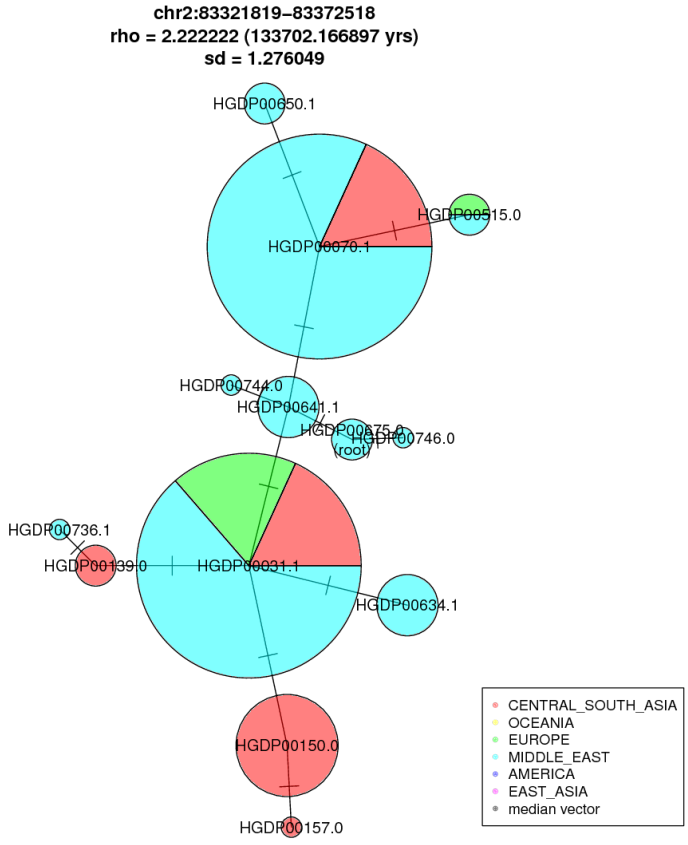
(b)

(Each circle represents a distinct haplotype, labelled by one sample name and coloured by the geographical origins of the samples, and the radius is proportional to the number of samples carrying that haplotype. The number of bars on the edges equals the number of mutations between haplotypes. Small grey circles labelled "mv" represents median vectors reconstructed in the median joining algorithm. Dashed lines: alternative links.)

Fig. 5.36 Examples of Neanderthal haplotype networks



(a)



(b)

(Each circle represents a distinct haplotype, labelled by one sample name and coloured by the geographical origins of the samples, and the radius is proportional to the number of samples carrying that haplotype. The number of bars on the edges equals the number of mutations between haplotypes. Small grey circles labelled "mv" represents median vectors reconstructed in the median joining algorithm. Dashed lines: alternative links.)

Fig. 5.37 Examples of Denisova haplotype networks

where n is the number of sequences, m the total number of edges, and n_i the number of samples whose shortest route to the root node passes through the ith edge. ρ can then be converted into time in years with the mutation rate and the number of comparable sites in the genomic region.

Figure 5.38 shows the distribution of Neanderthal and Denisova haplotype network age in years. Filtering networks based on the length of the genomic region passing the mask, average B value of the genomic region, the number of missing sites in archaic genomes, the number of sites skipped (singletons also present in Africa), or the total number of polymorphic sites does not alter the shape of the distribution visibly; nor do these values exhibit distinct distributions between the groups of largest and smallest networks.

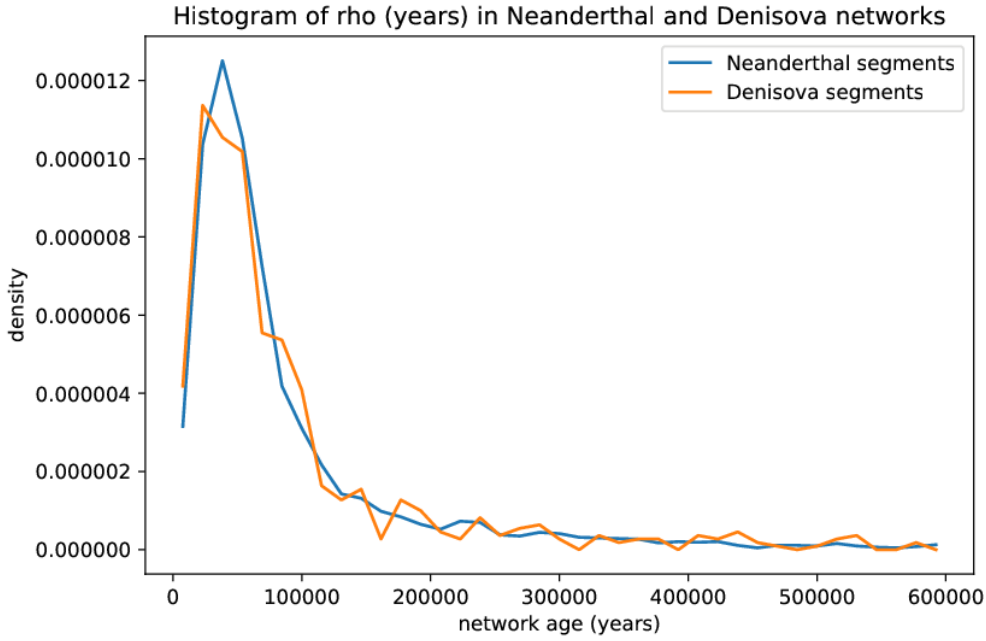


Fig. 5.38 Histogram of Neanderthal and Denisova haplotype network age

The age distribution from the two archaic sources is similar, though the Neanderthal curve appears smoother with a larger sample size. Both distributions reach the highest density below 50k years. The median in Neanderthal and Denisova haplotype networks is 55,613 and 55,070 years, respectively. However, we also observe networks hundreds of thousands of years old in the tail of the distributions: 11 Neanderthal and 3 Denisova networks are even estimated to be older than one million years.

It could be that these ages reflect high diversity in the introgressing haplotypes, perhaps from multiple divergent source populations. On the other hand, these age estimates are based on a small number of mutations and hence potentially subject to bias and error in the inference process. To explore this, I also constructed median joining networks on simulated haplotypes. The simplified demographic history is shown in Figure 5.39. The sample size in Eurasia (including East Asia, Central/South Asia, the Middle East, and Europe), Oceania, and America populations were specified to match the geographical origin of actual Neanderthal haplotypes observed in each genomic region. The number of introgressing haplotypes was measured as the number of surviving lineages 2,000 generations ago, namely (backward in time) at the end of the bottleneck associated with admixture and initial negative selection. In order to efficiently sample genealogies with few introgressing haplotypes, the duration and size of the bottleneck were arbitrarily selected. Since the genealogy is guaranteed to change across the genome, I set an upper limit other than a fixed value to the number of introgressing haplotypes. For each genomic region used to construct a haplotype network, coalescent trees were repeatedly simulated with a matching sample size, until the number of introgressing haplotypes became equal or less than the desired maximum. Genetic sequences of a length matched to the genomic region after filtering were then generated from the tree. In principle, the choices of bottleneck severity and ancestral Neanderthal size will influence the distribution of the coalescent trees retained; the demographic model chosen here is only meant for preliminary exploration.

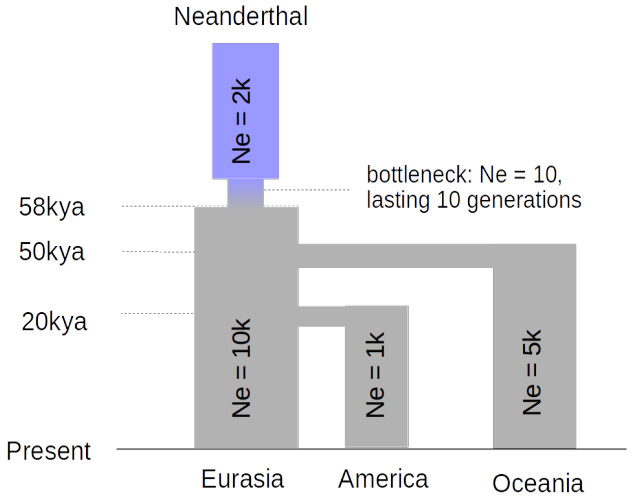


Fig. 5.39 Demographic model used in simulations exploring different numbers of founding Neanderthal haplotypes

Three sets of 4,153 genealogies with at most 1, 2 and 4 founding haplotypes at the time of admixture respectively were obtained, to match the 4,153 genomic regions used in Neanderthal haplotype network analysis. The same algorithm used on the real-world data was used to build haplotype networks for each simulated alignment dataset and estimate ρ. Since the underlying genealogy is known in simulations, the distribution of ρ is compared to the true tMRCA in Figure 5.40. The estimated age has greater dispersion than the true value in the case of only one founding haplotype; when more than one founding haplotypes are present, fewer older networks (between 100k and 400k years old) are reported, whilst more networks are estimated to be under 50k years of age. By allowing alternative links, the age estimate from haplotype networks can potentially underestimate the age of moderately old networks, but not the extremely old ones in the right tail of the distribution.

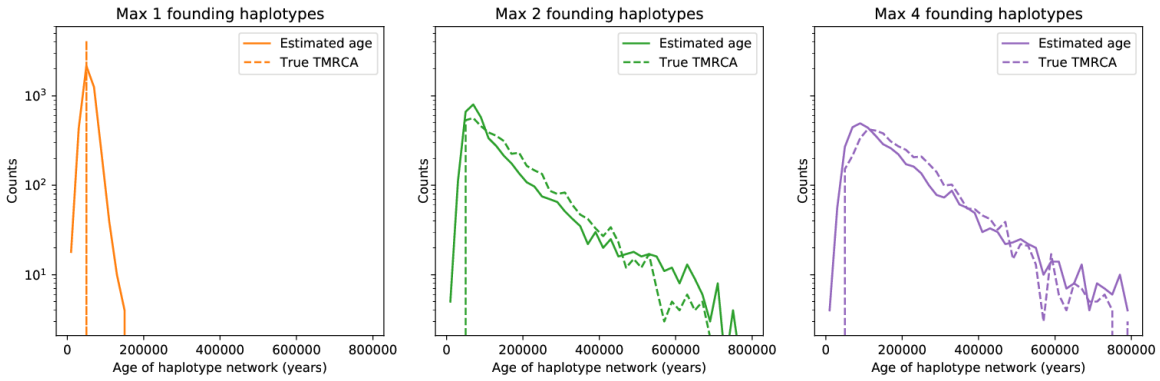


Fig. 5.40 Distribution of estimated haplotype network age and true tMRCA conditioned on the maximum number of introgressing Neanderthal haplotypes at 2,000 generations ago

The backward simulations started with the same number of samples as the real data, but in both cases some of them might carry the same haplotype. Figure 5.41 checks whether the number of unique haplotypes in simulations also matches that in real data. The number of unique haplotypes in simulated and in real data align along the identity line with a strong correlation in all three sets of simulations. This is a validation that the demographic model used in the simulations is a reasonable approximation to the true history.

Finally, Figure 5.42 compares the distribution of Neanderthal haplotype network age estimated from real data and three sets of simulated data, shown in log scale for a better resolution in the tail region. The observed distribution clearly differed from that produced by only one founding haplotype, which does not contain any networks older than 200k years; yet its overall shape appears shifted towards the left in comparison to the curves produced by a maximum of two and four haplotypes. The shift could result from selection against

Fig. 5.41 Comparison of the number of unique Neanderthal haplotypes in modern populations between observed data and simulations

introgressed haplotypes or sub-population structure not implemented in the simulations. The simulations with maximum two and four haplotypes generated very similar distributions, and it is difficult to determine which fits the observation better. Nevertheless, even two founding haplotypes appear sufficient to produce the number of large haplotype networks observed. In agreement with a very small bottleneck size estimated from the Neanderthal SFS (Section 5.6.2), the overall diversity of incoming Neanderthal sequences appears so limited that it could in principle have been contributed by a single individual. Many more Neanderthal individuals could have been involved in reality, contributing a reduced number of distinct haplotypes depending on the genetic diversity in the Neanderthal population. The combined effect of negative selection, genetic drift, and dilution could have further reduced the genetic diversity from the Neanderthal source to a very low level.

5.7.2 Number of founding lineages

The estimation from the age of the haplotype networks reflects the genome-wide average number of founding archaic haplotypes. In this section, I directly estimate the number of introgressing archaic haplotypes in each genomic region as the number of surviving lineages in the tree at the time of admixture.

For each of the 4,135 genomic regions, a maximum-likelihood tree was built from the sequence alignment and rooted by the haplotype closest to a San individual (HGDP00991). I chose to work with the tree structure rather than network here mainly for the ease to perform bootstrap analysis. The height at each node was determined by assigning height 0 to the

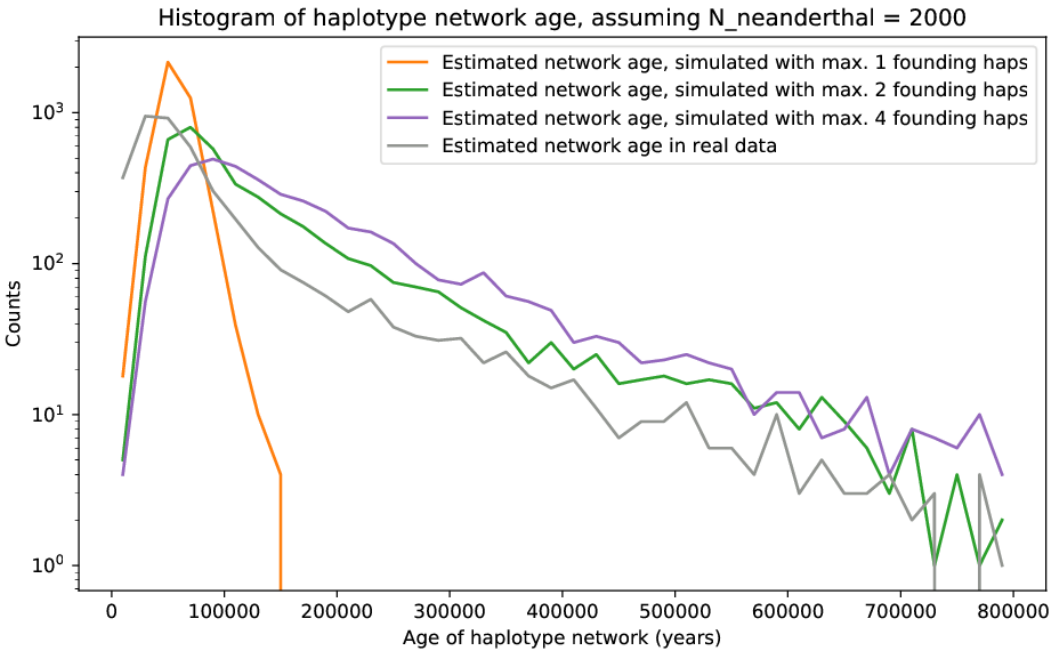


Fig. 5.42 Distribution of Neanderthal haplotype network age estimated from real and simulated data

tip farthest from the root, and positive heights to all other nodes, with the largest value at the root. Then I cut the tree at the height corresponding to the expected differences per basepair in the time since admixture - assumed to be 2,000 generations - and counted how many lineages remains connected to the root. To gauge the uncertainty, 1,000 non-parametric bootstraps were performed for each genomic region, and the same procedure was applied to the bootstrap trees to obtain the distribution of the number of lineages at the time of admixture.

Figure 5.43 is a box plot showing the dispersion of the number of founding haplotypes in 100 randomly sampled genomic regions, and Figure 5.44 shows the distribution of the mean number from all genomic regions. The number of haplotypes were mostly estimated to be small: in over 70% of the trees, the value of two standard deviations below the mean is lower than 2. However, there are also cases where more than 10 or even 20 lineages seem to exist at the time of introgression, with a narrow dispersion estimated from bootstrapping. A total of 17 genomic regions were estimated to have more than 20 founding Neanderthal haplotypes (Table 5.4); their haplotype networks (shown in Appendix B.3) indeed exhibit complicated radiating structures from one or two core haplotypes. It could mean that initially around a dozen (or more, depending on their relatedness) Neanderthal individuals were involved in the

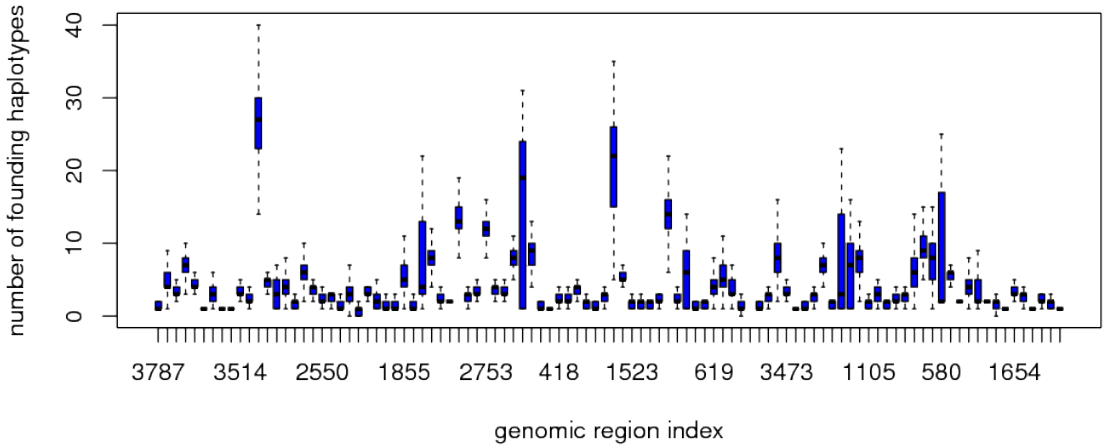


Fig. 5.43 Boxplot showing the number of founding Neanderthal lineages from 1,000 bootstraps in 100 out of 4,135 genomic regions

interbreeding, but except for very few regions of the genome, the majority of Neanderthal lineages were subsequently lost through genetic drift and negative selection, so that only a handful of them contributed to the present diversity of the Neanderthal segments in modern humans.

Table 5.4 Genomic regions with more than 20 founding Neanderthal haplotypes

chrom	start	end	number of haps	number of unique haps	rho±sd	rho±sd (years)	number of lineage (mean±sd)
9	110937947	111009555	220	75	6.0591±3.81415	223448±140659	22.02±3.395
6	66848428	66915134	163	68	2.1779±0.34909	89369±14325	20.54±5.666
12	113841761	113933611	150	85	2.9400±0.31640	80829±8699	23.75±4.832
5	58495484	58599651	198	99	2.5808±0.40432	78919±12364	26.95±5.374
1	217246438	217327394	200	89	2.3250±0.59293	74454±18987	25.49±8.721
19	56087294	56138132	228	67	1.1096±0.10359	70634±6594	25.79±5.941
10	62941526	62993549	209	52	1.1866±0.07060	70208±4177	22.54±2.512
12	114080296	114130307	204	69	1.3039±0.08852	66588±4521	20.27±11.773
2	13827358	13919411	125	73	2.1360±0.16531	60265±4664	20.10±4.068
1	216557045	216647205	218	85	2.1468±0.15512	58609±4235	24.36±7.341
1	32911081	32992108	211	68	1.4218±0.26828	57412±10833	20.78±6.449
4	28482612	28545486	191	64	1.2775±0.08042	54515±3432	21.27±3.558
12	20849814	20933980	151	63	1.7086±0.21727	52310±6652	21.67±5.702
9	126565708	126646783	190	66	1.3895±0.16172	46613±5425	21.93±7.082
1	212466385	212548707	196	78	1.3010±0.04391	43196±1458	26.19±3.870
9	94515802	94565893	148	42	0.5338±0.01082	38531±781	23.01±2.213
1	33403459	33454985	263	63	0.6692±0.02455	34713±1273	20.62±3.131

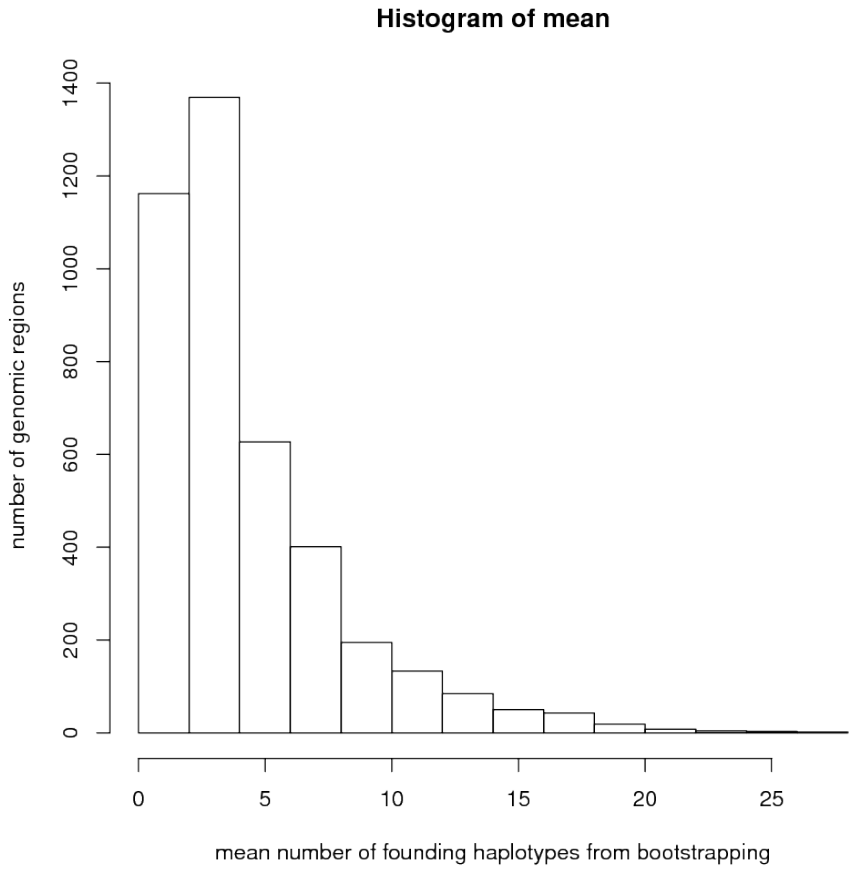


Fig. 5.44 Histogram showing the distribution of the mean number of founding lineages estimated from 1,000 bootstraps in 4,135 genomic regions

5.7.3 Geographical separation

Deep splits in the haplotype network could reflect multiple sources of archaic gene flow. To measure the divergence between geographic regions, I define two regions as completely separate in a network if none of the nodes containing haplotypes from one region has a closest neighbour containing haplotypes from the other region, and vice versa; namely the two regions form distinct "clades" within the network. For each pair of geographical regions, Table 5.5 displays the total numbers of comparable haplotype networks of Neanderthal and Denisova introgressed haplotypes, where comparable means that they contain at least two haplotypes from each region, and the numbers of such networks showing complete separation between pairs of geographical regions.

Table 5.5 The number of completely separated haplotype network between pairs of regions out of the total number of comparable networks

Neanderthal haplotype networks, separated / total					
	Central/South Asia	East Asia	Europe	Middle East	Oceania
America	254 / 862	184 / 878	221 / 618	245 / 506	159 / 263
Central/South Asia	-	228 / 2064	133 / 2190	139 / 2032	261 / 690
Eeast Asia	-	-	338 / 1356	429 / 1162	187 / 714
Europe	-	-	-	81 / 1932	249 / 448
Middle East	-	-	-	-	230 / 393

Denisova haplotype networks, separated / total					
	Central/South Asia	East Asia	Europe	Middle East	Oceania
America	6 / 36	8 / 59	4 / 15	3 / 9	10 / 12
Central/South Asia	-	4 / 88	3 / 42	3 / 32	37 / 39
East Asia	-	-	5 / 27	5 / 14	35 / 40
Europe	-	-	-	2 / 31	7 / 9
Middle East	-	-	-	-	6 / 8

The structure in modern human populations is expected to cause divergence in the archaic segments even if they descended from the same source of gene flow. Indeed, the proportion of completely separated Neanderthal haplotype networks between geographical regions is largely consistent with the history of population splits and migration within modern humans: the Middle East, Central/South Asia and Europe show the lowest level of separation, reflecting continuity of movement; in contrast, Oceania is separated from America, Europe, and the Middle East in more than half of the networks, but shows a closer relationship to

East Asia. Much fewer Denisova haplotype networks are available for comparison, yet here there is a deep split between Oceania and all other non-African regions: they appear completely separated in almost all the haplotype networks. Perhaps the most striking contrast to the networks of Neanderthal segments is between Oceania and East Asia, and between Oceania and Central/South Asia, where nearly 90% of networks exhibit complete separation. With a sample size of about 40, it is unlikely to observe complete separation in around 90% of the networks by chance, and in Neanderthal haplotype networks the proportion is 26.19% and 55.58% respectively. Fisher's exact test confirms that the distribution of fully separated networks between Neanderthal and Denisova haplotypes is highly significant in these two pairs (Table 5.6). Another pair that shows near-significant difference is between Central/South Asia and East Asia, but in this case they are better connected in Denisova networks than in Neanderthal networks. Overall, if we take the Neanderthal networks as representative of a single-source scenario, the strong separation between Oceania and other regions in the Denisova haplotype network provides evidence for different source populations of Denisova gene flow in and out of Oceania.

Table 5.6 p-values from Fisher's exact test on different distributions of separated/connected networks in Neanderthal vs. Denisova haplotypes between pairs of regions

	Central/South Asia	East Asia	Europe	Middle East	Oceania
America	0.1320	0.2419	0.5908	0.5067	0.1375
Central/South Asia	-	0.0534	0.7398	0.4802	3.075×10^{-13}
East Asia	-	-	0.6523	1	5.207×10^{-15}
Europe	-	-	-	0.3801	0.3100
Middle East	-	-	-	-	0.4793

Looking more closely at the networks involved, it is notable that in the only two networks where Denisova haplotypes in Oceania are not fully separated from those in Central/South Asia, Oceania, and East Asia are also connected. Out of the five networks where Oceania and East Asia are not separated, three involve Denisova haplotypes from Cambodia bridging them: in two cases the Cambodian haplotype is the sole connection, in another one a haplotype from Lahu is also involved (Appendix B.4). One reasonable interpretation is that Southeast Asia serves as a genetic corridor to connect Oceania and East Asia. But in light of the increased genetic diversity between Denisovan segments in Cambodia and many other East Asian populations (described in Section 5.6.1), it could also suggest another independent component of Denisova gene flow in Cambodia, whose source population is closer to the source in Oceania.

5.8 Conclusion

The amount of archaic segments inferred by the HMM in HGDP genomes is consistent with previous studies. I also detected negative selection against Neanderthal segments, and identified risk alleles in modern human that are likely to have been acquired through the admixture with Neanderthal.

Through analysing Neanderthal and Denisovan segments, I found distinct structure in these two admixture events. Various lines of evidence are consistent with a single Neanderthal source population contributing to Neanderthal ancestry in all non-African populations: Neanderthal segments recovered from modern genomes worldwide show considerable overlapping in genomic locations, similar divergence to the archaic genomes, and similar distribution of segment lengths. The genetic diversity in Neanderthal segments within and between populations mostly mirrors that in the unadmixed part of the genomes, supporting a shared Neanderthal gene flow into the ancestral population of all non-Africans today before subsequent population divergence. In contrast, Denisovan segments detected in Oceania show deep divergence to those in Eurasia mainland and the Americas in genomic distribution, divergence to archaic genomes, inter- and intra-population nucleotide diversity, and positions in the haplotype network. The evidence strongly supports that separate admixture events with Denisova happened in the ancestors of modern Oceanians, and the ancestors of modern Eurasians and Americans. Additionally, the pattern of divergence between Denisovan segments in East Asia and archaic reference genomes suggests more than one episode of gene flow from distinct Denisova populations into the common ancestors of modern East Asian, South Asian and American populations.

The genetic evidence opens up space for hypotheses on not just the dynamics of the admixture events, but also the distribution and structure of the archaic populations. The admixture with Neanderthal should have happened only once when the effective population size of the out-of-Africa population was still small and relatively homogeneous. Therefore, the range of Neanderthal populations might have been limited to Europe and Central/West Asia. On the other hand, multiple episodes of gene flow from Denisova into ancestors of modern-day Asians and Americans, and another separate episode into ancestors of modern-day Oceanians could suggest a wider presence of Denisova populations in Asia, but in smaller and more isolated groups. As modern humans expanded through Asia, they could have multiple encounters with local Denisova populations. The ancestors of Oceanians should have split off before their admixture with a local Denisova group, which did not interbreed with the other modern human populations. This could explain the distinct Denisovan segments found

in Papuan populations. A large part of the theory, of course, still awaits corroboration from archaeological or ancient DNA evidence.

Chapter 6

Conclusions and recommendations for future work

6.1 Conclusions

The endeavour to deep-sequence unrelated individuals in the HGDP panel aims to provide a resource for the human genetics community, especially at a time when many populations are poorly represented in biomedical research [172]. The structure and diversity in this dataset reaffirm our understanding of population history, meanwhile highlighting the presence of recent inbreeding in isolated populations and recent admixture in the Middle East and part of Central Asia. I have also shown that current resources and methods allow accurate haplotype estimation in most populations (switch error rate < 1% excluding singletons), which has a significant bearing on biomedical studies.

This thesis also describes a hidden Markov model (HMM) to detect Neanderthal and Denisovan segments in modern human genomes with reference to the archaic genomes. Although several probabilistic graphical models have been used previously for this purpose, I presented an extensive exploration over variations of the model, including site-wise/window-based observation and various criteria to distinguish sources of archaic introgression, as well as an evaluation of the performance and features of the model. Themes observed here, such as the relationship between segment lengths and detection rate, might also be relevant to other HMM.

Running this HMM on the HGDP dataset provided a large collection of archaic segments for comparisons at subpopulation, population, and regional levels. Contributing to the ongoing

discussion over the number of separate admixture events between archaic and modern humans [93, 104], I have shown that the genomic distribution, divergence to the archaic genomes and geographical diversity of the archaic segments consistently suggest one shared episode of Neanderthal introgression into the ancestral population of all present-day non-Africans, but separate episodes into Oceania and Eastern Eurasia. To the best of my knowledge, this is also the first attempt to estimate how many archaic individuals contributed to the gene pool of modern humans. The diversity in Neanderthal haplotypes suggests that at least a dozen Neanderthal individuals could have contributed to the admixture initially, but many haplotypes have been subsequently lost through drift and negative selection so that only a few founding haplotypes are necessary to explain the average diversity in Neanderthal segments today.

6.2 Future directions

The potential of the high-coverage HGDP dataset undoubtedly extends beyond the scope of this thesis and calls upon numerous researchers with diverse expertise. I will limit the recommendations below to themes that have been touched upon in previous chapters.

1. More work is needed to characterize the genetic variations in sub-Saharan Africa and Oceania. The switch error rates in statistical phasing (Table 2.5) are highest in San and two Papuan populations, the former being the most diverse population in the panel, the latter deeply diverged from the other large populations without substantial recent inbreeding. The presence of unique Denisova haplotypes in Cambodia suggests that another region with potentially underexplored genetic history is Southeast Asia, which is not well represented in the HGDP panel either. A better understanding of genetic variation in these regions not only serves the historical and anthropological interest, but also promotes equal benefits from biomedical research; in this process, it is also crucial to actively engage the local communities at all stages.

2. Detailed demographic simulation and formal tests might reveal more insights about the admixture events and the adaptive history of archaic alleles. In particular, forward-time simulations have been successfully applied to explore specific contact history and selection strength [99, 111, 112]. Since it proves difficult to estimate admixture time from the lengths of archaic segments, similar simulations could help to formally estimate the number of gene flows, the time of multiple Denisova gene flows and changes in selection strength across time, perhaps with added evidence from historical genomes.

3. The detected archaic segments provide opportunities to survey the functional effects of archaic alleles in diverse populations. Many of the GWAS effect alleles with archaic origin listed in Table 5.2 are concentrated in certain populations or geographical regions, making them candidates to explore different environmental challenges and evolutionary trajectory of specific traits. Nevertheless, GWAS studies tend to focus on disease-causing alleles and cannot tag rare alleles in the population. A more comprehensive survey could combine unusually high frequency of the archaic haplotype, deep coalescent genealogy but not consistent with incomplete lineage sorting, and potential association to phenotypes to establish evidence for positive or balancing selection, as reviewed in [173].

4. Additional genetic and fossil evidence of modern and archaic hominins from Middle to Upper Paleolithic Eurasia will be of immense help to decipher the interactions between human groups. So far Denisova fossils have only been found in the same cave in Siberia, whereas genetic evidence reveals higher Denisovan ancestry in East Eurasia and Oceania. Filling in the dearth of hominin fossils is crucial to understand the exact geographical range, genetic variation and even social structure of Neanderthal, Denisova and our modern human ancestors.

References

[1] S. Sankararaman, S. Mallick, M. Dannemann, K. Prufer, J. Kelso, et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357, 2014.

[2] L. Hirschfeld and H. Hirschfeld. Serological differences between the blood of different races. *The Lancet*, 194(5016):675–679, 1919.

[3] R. Ottenberg. A classification of human races based on geographic distribution of the blood groups. *JAMA: The Journal of the American Medical Association*, 84(19):1393, 1925.

[4] J. V. Neel, F. M. Salzano, P. C. Junqueira, F. Keiter, D. Maybury-Lewis, et al. Studies on the Xavante Indians of the Brazilian Mato Grosso. *American Journal of Human Genetics*, 16(1):52–140, 1964.

[5] J. S. Friedlaender. *Patterns of human variation: The demography, genetics, and phenetics of Bougainville Islanders*. 1975.

[6] L. L. Cavalli-Sforza. Population structure and human evolution. *Proceedings of the Royal Society of London. Series B, Biological sciences*, 164(995):362–79, 1966.

[7] B. K. Suarez, J. D. Crouse, and D. H. O'rourke. Genetic variation in North Amerindian populations: The geography of gene frequencies. *American Journal of Physical Anthropology*, 67(3):217–232, 1985.

[8] R. C. Lewontin. The apportionment of human diversity. In *Evolutionary Biology*, pages 381–398. Springer US, New York, NY, 1972.

[9] B. D. H. Latter. Genetic differences within and between populations of the major human subgroups. *The American Naturalist*, 116(2):220–237, 1980.

[10] J. H. Edwards. Population genetics of C4 with the use of complementary DNA probes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 306(1129):405–417, 1984.

[11] J. C. Murray, K. A. Mills, C. M. Demopulos, S. Hornung, and A. G. Motulsky. Linkage disequilibrium and evolutionary relationships of DNA variants (restriction enzyme fragment length polymorphisms) at the serum albumin locus. *Proceedings of the National Academy of Sciences of the United States of America*, 81(11):3486–90, 1984.

[12] R. L. Cann, M. Stoneking, and A. C. Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(6099):31–36, 1987.

[13] A. M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455–457, 1994.

[14] A. Di Rienzo, A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer. Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 91(8):3166–70, 1994.

[15] D. F. Conrad, M. Jakobsson, G. Coop, X. Wen, J. D. Wall, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, 38(11):1251–1260, 2006.

[16] A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[17] M. H. Wolpoff, X. Wu, A. G. Thorne, F. H. Smith, F. Spencer, et al. *The origins of modern humans: A world survey of the fossil evidence*. 1984.

[18] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, A. C. Wilson, et al. African populations and the evolution of human mitochondrial DNA. *Science*, 253(5027):1503–7, 1991.

[19] D. A. Merriwether, A. G. Clark, S. W. Ballinger, T. G. Schurr, H. Soodyall, et al. The structure of human mitochondrial DNA variation. *Journal of Molecular Evolution*, 33(6):543–55, 1991.

[20] A. Di Rienzo and A. C. Wilson. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 88(5):1597–601, 1991.

[21] M. F. Hammer. A recent common ancestry for human Y chromosomes. *Nature*, 378(6555):376–378, 1995.

[22] M. F. Hammer, T. Karafet, A. Rasanayagam, E. T. Wood, T. K. Altheide, et al. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Molecular Biology and Evolution*, 15(4):427–441, 1998.

[23] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

[24] M. E. Steiper. DNA markers of human variation. In Michael P. Muehlenbein, editor, *Human Evolutionary Biology*, pages 238–264. Cambridge University Press, Cambridge, 2010.

[25] H. Kaessmann, F. Heißig, A. von Haeseler, and S. Pääbo. DNA sequence variation in a non-coding region of low recombination onthe human X chromosome. *Nature Genetics*, 22(1):78–81, 1999.

[26] Z. Zhao, L. Jin, Y.-X. Fu, M. Ramsay, T. Jenkins, et al. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11354, 2000.

[27] L.L. Cavalli-Sforza, A.C. Wilson, C.R. Cantor, R.M. Cook-Deegan, and M.-C.and others King. Call for a worldwide survey of human genetic diversity: A vanishing opportunity for the Human Genome Project. *Genomics*, 11(2):490–491, 1991.

[28] H. T. Greely. Human genome diversity: What about the other human genome project? *Nature Reviews Genetics*, 2(3):222–227, 2001.

[29] H. M. Cann, C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, et al. A Human Genome Diversity Cell Line Panel. *Science*, 296(5566):261, 2002.

[30] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, et al. Genetic structure of human populations. *Science*, 298(5602):2381–5, 2002.

[31] S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–7, 2005.

[32] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, 2008.

[33] K. L. Hunley, M. E. Healy, and J. C. Long. The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: Implications for biological race. *American Journal of Physical Anthropology*, 139(1):35–46, 2009.

[34] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–6, 2011.

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

[36] S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–25, 2014.

[37] J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.

[38] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, et al. The International HapMap Project. *Nature*, 426(6968):789–796, 2003.

[39] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

[40] L. Pagani, D. John Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538(7624):238–242, 2016.

[41] HGDP-CEPH Human Genome Diversity Cell Line Panel. http://www.cephb.fr/en/hgdp_panel.php. Accessed: 2018-10-07.

[42] L. A. Zhivotovsky, N. A. Rosenberg, and M. W. Feldman. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *American Journal of Human Genetics*, 72:1171–1186, 2003.

[43] N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, et al. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1(6):e70, 2005.

[44] M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003, 2008.

[45] T. J. Pemberton, M. Jakobsson, D. F. Conrad, G. Coop, J. D. Wall, et al. Using population mixtures to optimize the utility of genomic databases: Linkage disequilibrium and association study design in India. *Annals of Human Genetics*, 72(4):535–546, 2008.

[46] T. J. Pemberton, C. I. Sandefur, M. Jakobsson, and N. A. Rosenberg. Sequence determinants of human microsatellite variability. *BMC Genomics*, 10(1):612, 2009.

[47] M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–6, 2012.

[48] Z. A. Szpiech, J. Xu, T. J. Pemberton, W. Peng, S. Zöllner, et al. Long runs of homozygosity are enriched for deleterious variation. *American Journal of Human Genetics*, 93(1):90–102, 2013.

[49] L. C. Francioli, A. Menelaou, S. L. Pulit, F. van Dijk, P. F. Palamara, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8):818–825, 2014.

[50] A. Ruiz-Linares, K. Adhikari, V. Acuña-Alonzo, M. Quinto-Sanchez, C. Jaramillo, et al. Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genetics*, 10(9):e1004572, 2014.

[51] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, et al. The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314, 2015.

[52] R. G. Harrison and E. L. Larson. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105(S1):795–809, 2014.

[53] J. Mallet, N. Besansky, and M. W. Hahn. How reticulated are species? *BioEssays*, 38(2):140–149, 2016.

[54] J. Mallet. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5):229–237, 2005.

[55] P. R. Grant. Hybridization of Darwin's finches on Isla Daphne Major, Galapagos. *Philosophical Transactions - Royal Society of London, B*, 340(1291):127–139, 1993.

[56] P. F. Smith, A. Konings, and I. Kornfield. Hybrid origin of a cichlid population in Lake Malawi: implications for genetic variation and species diversity. *Molecular Ecology*, 12(9):2497–504, 2003.

[57] T. L. Turner, M. W. Hahn, and S. V. Nuzhdin. Genomic islands of speciation in Anopheles gambiae. *PLoS Biology*, 3(9):e285, 2005.

[58] The Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94–8, 2012.

[59] E. Trinkaus. European early modern humans and the fate of the Neandertals. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7367–72, 2007.

[60] P. D. Evans, N. Mekel-Bobrov, E. J. Vallender, R. R. Hudson, and B. T.and others Lahn. Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 103(48):18178–18183, 2006.

[61] V. Plagnol and J. D. Wall. Possible ancestral structure in human populations. *PLoS Genetics*, 2(7):e105, 2006.

[62] M. Krings, A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, et al. Neandertal DNA sequences and the origin of modern humans. *Cell*, 90(1):19–30, 1997.

[63] I. V. Ovchinnikov, A. Götherström, G. P. Romanova, V. M. Kharitonov, K. Lidén, et al. Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, 404(6777):490–493, 2000.

[64] R. E. Green, A.-S. Malaspinas, J. Krause, A. W. Briggs, P. L.F. Johnson, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–426, 2008.

[65] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, et al. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, 2010.

[66] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–9, 2014.

[67] K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, 358(6363):655–658, 2017.

[68] M. Kuhlwilm, I. Gronau, M. J. Hubisz, C. de Filippo, J. Prado-Martinez, et al. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, 530(7591):429–433, 2016.

[69] M. Hajdinjak, Q. Fu, A. Hübner, M. Petr, F. Mafessoni, et al. Reconstructing the genetic history of late Neanderthals. *Nature*, 555(7698):652–656, 2018.

[70] M. Meyer, J.-L. Arsuaga, C. de Filippo, S. Nagel, A. Aximu-Petri, et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature*, 531(7595):504–507, 2016.

[71] C. Posth, C. Wißing, K. Kitagawa, L. Pagani, L. van Holstein, et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nature Communications*, 8:16046, 2017.

[72] J. Krause, Q. Fu, J. M. Good, B. Viola, M. V. Shunkov, et al. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464(7290):894–897, 2010.

[73] D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, 2010.

[74] P. Skoglund and M. Jakobsson. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences of the United States of America*, 108(45):18301–18306, 2011.

[75] P. Qin and M. Stoneking. Denisovan ancestry in east Eurasian and native American populations. *Molecular Biology and Evolution*, 32(10):2665–2674, 2015.

[76] S. Sankararaman, S. Mallick, N. Patterson, and D. Reich. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology*, 26(9):1241–7, 2016.

[77] V. Slon, F. Mafessoni, B. Vernot, C. de Filippo, S. Grote, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, 561(7721):113–116, 2018.

[78] C. Pinho and J. Hey. Divergence with gene flow: models and data. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):215–230, 2010.

[79] C. Roux, G. Tsagkogeorga, N. Bierne, and N. Galtier. Crossing the species barrier: genomic hotspots of introgression between two highly divergent ciona intestinalis species. *Molecular Biology and Evolution*, 30(7):1574–1587, 2013.

[80] V. C. Sousa, M. Carneiro, N. Ferrand, and J. Hey. Identifying loci under selection against gene flow in isolation-with-migration models. *Genetics*, 194(1):211–233, 2013.

[81] B. K. Rosenzweig, J. B. Pease, N. J. Besansky, and M. W. Hahn. Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, 25(11):2387–2397, 2016.

[82] S Wright. Evolution in Mendelian populations. *Genetics*, 52(1-2):241–95; discussion 201–7, 1931.

[83] M. C. Murray and M. P. Hare. A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, Crassostrea virginica. *Molecular Ecology*, 15(13):4229–4242, 2006.

[84] D. E. Neafsey, B. M. Barker, T. J. Sharpton, J. E. Stajich, D. J Park, et al. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Research*, 20(7):938–46, 2010.

[85] B. Charlesworth. Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, 15(5):538–543, 1998.

[86] T. E. Cruickshank and M. W. Hahn. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13):3133–3157, 2014.

[87] E. Y. Durand, N. Patterson, D. Reich, and M. Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.

[88] S. H. Martin, J. W. Davey, and C. D. Jiggins. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 32(1):244–257, 2015.

[89] A. J. Geneva, C. A. Muirhead, S. B. Kingan, and D. Garrigan. A new method to scan genomes for introgression in a secondary contact model. *PLoS ONE*, 10(4):e0118621, 2015.

[90] B. Vernot and J. M. Akey. Resurrecting surviving Neandeltal linages from modern human genomes. *Science*, 343(February):1017–1021, 2014.

[91] B. Vernot, S. Tucci, J. Kelso, J. G. Schraiber, A. B. Wolf, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*, 352(6282):235–9, 2016.

[92] M. Steinrücken, J. P. Spence, J. A. Kamm, E. Wieczorek, Y. S. Song, et al. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Molecular Ecology*, 27(19):3873–3888, 2018.

[93] S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, et al. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell*, 173(1):53–61.e9, 2018.

[94] A. Durvasula and S. Sankararaman. Recovering signals of ghost archaic admixture in the genomes of present-day Africans. *bioRxiv*, page 285734, 2018.

[95] L. Skov, R. Hui, V. Shchur, A. Hobolth, A. Scally, et al. Detecting archaic introgression using an unadmixed outgroup. *PLOS Genetics*, 14(9):e1007641, 2018.

[96] J. D. Wall, M. A Yang, F. Jay, S. K. Kim, E. Y. Durand, et al. Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics*, 194(1):199–209, 2013.

[97] S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, et al. The date of interbreeding between Neandertals and modern humans. *PLoS Genetics*, 8(10):e1002947, 2012.

[98] A.-S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, et al. A genomic history of Aboriginal Australia. *Nature*, 538(7624):207–214, 2016.

[99] B. Y. Kim and K. E. Lohmueller. Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. *American Journal of Human Genetics*, 96(3):454–461, 2015.

[100] B. Vernot and J. M. Akey. Complex history of admixture between modern humans and Neandertals. *American Journal of Human Genetics*, 96(3):448–53, 2015.

[101] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

[102] I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, 2016.

[103] Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, 524(7564):216–219, 2015.

[104] F. A. Villanea and J. Schraiber. Spectrum of Neandertal introgression across modern-day humans indicates multiple episodes of human-Neandertal interbreeding. *bioRxiv*, page 343087, 2018.

[105] J. M. Burke and M. L. Arnold. Genetics and the fitness of hybrids. *Annual Review of Genetics*, 35(1):31–52, 2001.

[106] N. H. Barton. The role of hybridization in evolution. *Molecular Ecology*, 10(3):551–568, 2008.

[107] S. H. Martin and C. D. Jiggins. Interpreting the genomic landscape of introgression, 2017.

[108] S. Maheshwari and D. A. Barbash. The genetics of hybrid incompatibilities. *Annual Review of Genetics*, 45(1):331–355, 2011.

[109] S. Lamichhaney, J. Berglund, M. S. Almén, K. Maqbool, M. Grabherr, et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(7539):371–375, 2015.

[110] L. C. Norris, B. J. Main, Y. Lee, T. C. Collier, A. Fofana, et al. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proceedings of the National Academy of Sciences of the United States of America*, 112(3):815–20, 2015.

[111] K. Harris and R. Nielsen. The genetic cost of Neanderthal introgression. *Genetics*, 203(2):881–91, 2016.

[112] I. Juric, S. Aeschbacher, and G. Coop. The strength of selection against Neanderthal introgression. *PLOS Genetics*, 12(11):e1006340, 2016.

[113] M. Schumer, C Xu, D. L. Powell, A. Durvasula, L. Skov, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, 360(6389):656–660, 2018.

[114] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7523):445–449, 2014.

[115] Martin Petr, Svante Pääbo, Janet Kelso, and Benjamin Vernot. Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences of the United States of America*, 116(5):1639–1644, 2019.

[116] L. Abi-Rached, M. J. Jobin, S. Kulkarni, A. McWhinnie, K. Dalva, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, 334(6052):89–94, 2011.

[117] F. L. Mendez, J. C. Watkins, and M. F Hammer. A haplotype at STAT2 introgressed from Neanderthals and serves as a candidate of positive selection in Papua New Guinea. *American journal of human genetics*, 91(2):265–74, 2012.

[118] E. Huerta-Sánchez, X. Jin, Asan, Z. Bianba, B. M. Peter, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513):194–7, 2014.

[119] R. M. Gittelman, J. G. Schraiber, B. Vernot, C. Mikacenic, M. M. Wurfel, et al. Archaic hominin admixture facilitated adaptation to out-of-Africa environments. *Current Biology*, 26(24):3375–3382, 2016.

[120] C. N. Simonti, B. Vernot, L. Bastarache, E. Bottinger, D. S. Carrell, et al. The phenotypic legacy of admixture between modern humans and Neandertals. *Science*, 351(6274):737–41, 2016.

[121] M. Dannemann and J. Kelso. The contribution of Neanderthals to phenotypic variation in modern humans. *American Journal of Human Genetics*, 101(4):578–589, 2017.

[122] H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, 2014.

[123] B. S. Pedersen, R. M. Layer, and A. R. Quinlan. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biology*, 17(1):118, 2016.

[124] B. Paten, J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research*, 18(11):1829–43, 2008.

[125] G. McVicker, D. Gordon, C. Davis, and P. Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, 5(5):e1000471, 2009.

[126] F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, et al. The UCSC known genes. *Bioinformatics*, 22(9):1036–1046, 2006.

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

[128] P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448, 2016.

[129] Y. Choi, A. P. Chan, E. Kirkness, A. Telenti, and N. J.and others Schork. Comparison of phasing strategies for whole human genomes. *PLOS Genetics*, 14(4):e1007308, 2018.

[130] D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534):327–332, 2015.

[131] O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, J. Marchini, et al. Haplotype estimation using sequencing reads. *American Journal of Human Genetics*, 93(4):687–96, 2013.

[132] R. Poplin, Valentin R.-R., M. A. DePristo, T. J. Fennell, M. O. Carneiro, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, page 201178, 2018.

[133] O. Delaneau, J. Marchini, G. A. McVean, P. Donnelly, G. Lunter, et al. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, 5:3934, 2014.

[134] B. L. Browning and S. R. Browning. Genotype imputation with millions of reference samples. *American Journal of Human Genetics*, 98(1):116–26, 2016.

[135] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24):3326–3328, 2012.

[136] S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, et al. The genetic structure and history of Africans and African Americans. *Science*, 324(5930):1035–44, 2009.

[137] E. Patin, M. Lopez, R. Grollemund, P. Verdu, C. Harmant, et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, 356(6337):543–546, 2017.

[138] P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, et al. Reconstructing prehistoric African population structure. *Cell*, 171(1):59–71.e21, 2017.

[139] P. Skoglund, H. Malmström, M. Raghavan, J. Storå, P. Hall, et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, 336(6080):466–9, 2012.

[140] W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, 2015.

[141] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, et al. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32(11):1749–1751, 2016.

[142] T. J. Pemberton, D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, et al. Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics*, 91(2):275–292, 2012.

[143] J. K. Pickrell, N. Patterson, P.-R. Loh, M. Lipson, B. Berger, et al. Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 111(7):2632–7, 2014.

[144] P. Skoglund, S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hünemeier, et al. Genetic evidence for two founding populations of the Americas. *Nature*, 525(7567):104, 2015.

[145] C. L. Scheib, H. Li, T. Desai, V. Link, C. Kendall, et al. Ancient human parallel lineages within North America contributed to a coastal expansion. *Science*, 360(6392):1024–1027, 2018.

[146] B. L. Browning and S. R. Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[147] A. H. Bittles and M. L. Black. Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 107 Suppl(Suppl 1):1779–86, 2010.

[148] S. R. Browning and B. L. Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics*, 97(3):404–418, 2015.

[149] J. Kelleher, Y. Wong, P. K. Albers, A. W. Wohns, and G. McVean. Inferring the ancestry of everyone. *bioRxiv*, page 458067, 2018.

[150] L. Speidel, M. Forest, S. Shi, and S. Myers. A method for genome-wide genealogy estimation for thousands of samples. *bioRxiv*, page 550558, 2019.

[151] B.-J. Yoon. Hidden Markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–15, 2009.

[152] J. Kelleher, A. M. Etheridge, and G. McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5):e1004842, 2016.

[153] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.

[154] J. L. Morales and J. Nocedal. Remark on "algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". *ACM Transactions on Mathematical Software*, 38(1):1–4, 2011.

[155] S. Becker. L-BFGS-B, converted from Fortran to C, with Matlab wrapper, 2015.

[156] R. Fraile, E. García-Ortega, and R. Fraile. Fitting an exponential distribution. *Journal of Applied Meteorology*, 44(10):1620–1625, 2005.

[157] L. Skov, A. Bergström, C. Tyler-Smith, Y. Xue, R. Durbin, et al. Detecting introgressed archaic haplotypes in Oceanic population genome sequences, 2016.

[158] G. S. Jacobs, G. Hudjashov, L. Saag, P. Kusuma, C. C. Darusallam, et al. Multiple deeply divergent Denisovan ancestries in Papuans. *Cell*, 177(4):1010–1021.e32, 2019.

[159] H. Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–51, 2014.

[160] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 2017.

[161] E. E. Khrameeva, K. Bozek, L. He, Z. Yan, X. Jiang, et al. Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nature Communications*, 5(1):3584, 2014.

[162] Y. Nédélec, J. Sanz, G. Baharian, Z. A. Szpiech, A. Pacis, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*, 167(3):657–669.e21, 2016.

[163] R. O. Taskent, N. D. Alioglu, E. Fer, H. Melike Donertas, M. Somel, et al. Variation and functional impact of Neanderthal ancestry in western Asia. *Genome Biology and Evolution*, 9(12):3516–3524, 2017.

[164] M. Nei and W. H. Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5269–73, 1979.

[165] A. Bekada, L. R. Arauna, T. Deba, F. Calafell, S. Benhamamouch, et al. Genetic heterogeneity in Algerian human populations. *PLOS ONE*, 10(9):e0138453, 2015.

[166] L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, M. Foll, et al. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10):e1003905, 2013.

[167] A. Bergström, S. J. Oppenheimer, A. J. Mentzer, K. Auckland, K. Robson, et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science*, 357(6356):1160–1163, 2017.

[168] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

[169] E. Paradis. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, 26(3):419–420, 2010.

[170] H. J. Bandelt, P. Forster, and A. Rohl. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1):37–48, 1999.

[171] J. Saillard, P. Forster, N. Lynnerup, H.-J. Bandelt, S. Nørby, et al. mtDNA variation among Greenland Eskimos: The edge of the Beringian expansion. *American Journal of Human Genetics*, 67(3):718, 2000.

[172] A. B. Popejoy and S. M. Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 2016.

[173] F. Racimo, S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371, 2015.

Appendix A

List of samples from the HGDP panel

Table A.1 Information of 929 genomes from the HGDP panel

Sample	Accession No.	Library Type	population	region	sex	coverage
HGDP00001	ERS474507	PCR	Brahui	Central/South Asia	M	34.16
HGDP00003	ERS474508	PCR free	Brahui	Central/South Asia	M	34.64
HGDP00005	ERS474509	PCR	Brahui	Central/South Asia	M	33.74
HGDP00007	ERS474510	PCR free	Brahui	Central/South Asia	M	40.83
HGDP00009	ERS474511	PCR	Brahui	Central/South Asia	M	33.96
HGDP00011	ERS474512	PCR free	Brahui	Central/South Asia	M	35.16
HGDP00013	ERS474513	PCR free	Brahui	Central/South Asia	M	30.34
HGDP00015	ERS474514	PCR free	Brahui	Central/South Asia	M	36.53
HGDP00017	ERS474515	PCR free	Brahui	Central/South Asia	M	36.87
HGDP00019	ERS1042153	PCR free	Brahui	Central/South Asia	M	35.06
HGDP00021	ERS474517	PCR free	Brahui	Central/South Asia	M	26.71
HGDP00023	ERS474518	PCR free	Brahui	Central/South Asia	M	32.84
HGDP00025	ERS474519	PCR free	Brahui	Central/South Asia	M	30.61
HGDP00027	ERS1042152	PCR free	Brahui	Central/South Asia	M	37.92
HGDP00029	ERS474521	PCR free	Brahui	Central/South Asia	M	35.7
HGDP00031	ERS474522	PCR free	Brahui	Central/South Asia	M	33.93
HGDP00033	ERS474523	PCR free	Brahui	Central/South Asia	M	34.24
HGDP00035	ERS474524	PCR free	Brahui	Central/South Asia	M	31.37
HGDP00037	ERS474525	PCR free	Brahui	Central/South Asia	M	34.99
HGDP00039	ERS474526	PCR free	Brahui	Central/South Asia	M	31.23
HGDP00041	ERS474527	PCR free	Brahui	Central/South Asia	M	32.61
HGDP00043	ERS474528	PCR free	Brahui	Central/South Asia	M	33.44
HGDP00045	ERS474529	PCR free	Brahui	Central/South Asia	M	34.94
HGDP00047	ERS474530	PCR free	Brahui	Central/South Asia	M	38.32
HGDP00049	ERS474531	PCR free	Brahui	Central/South Asia	M	40.75

HGDP00052	ERS474532	PCR free	Balochi	Central/South Asia	M	30.39
HGDP00054	ERS474533	PCR free	Balochi	Central/South Asia	M	35.17
HGDP00056	ERS474534	PCR	Balochi	Central/South Asia	M	33.5
HGDP00057	ERS474535	PCR	Balochi	Central/South Asia	M	33.23
HGDP00058	ERS1042182	PCR free	Balochi	Central/South Asia	M	42.72
HGDP00060	ERS474537	PCR free	Balochi	Central/South Asia	M	32.73
HGDP00062	ERS474538	PCR	Balochi	Central/South Asia	M	31.87
HGDP00064	ERS474539	PCR free	Balochi	Central/South Asia	M	31.77
HGDP00066	ERS474540	PCR free	Balochi	Central/South Asia	M	31.43
HGDP00068	ERS474541	PCR free	Balochi	Central/South Asia	M	39.64
HGDP00070	ERS474542	PCR free	Balochi	Central/South Asia	M	34.13
HGDP00072	ERS474543	PCR free	Balochi	Central/South Asia	M	32.84
HGDP00074	ERS474544	PCR free	Balochi	Central/South Asia	M	31.04
HGDP00076	ERS474545	PCR free	Balochi	Central/South Asia	M	39.43
HGDP00078	ERS474546	PCR free	Balochi	Central/South Asia	M	31.34
HGDP00080	ERS474547	PCR free	Balochi	Central/South Asia	M	36.58
HGDP00082	ERS474548	PCR free	Balochi	Central/South Asia	M	39.28
HGDP00086	ERS474549	PCR free	Balochi	Central/South Asia	M	37.7
HGDP00088	ERS474550	PCR free	Balochi	Central/South Asia	M	31.96
HGDP00090	ERS1042137	PCR free	Balochi	Central/South Asia	M	42.69
HGDP00092	ERS474552	PCR free	Balochi	Central/South Asia	M	35.28
HGDP00094	ERS474553	PCR free	Balochi	Central/South Asia	M	33.73
HGDP00096	ERS474554	PCR free	Balochi	Central/South Asia	M	30.02
HGDP00098	ERS474555	PCR free	Balochi	Central/South Asia	M	34.23
HGDP00099	ERS474389	PCR free	Hazara	Central/South Asia	M	34.69
HGDP00100	ERS474390	PCR free	Hazara	Central/South Asia	M	30.88
HGDP00102	ERS474391	PCR free	Hazara	Central/South Asia	M	32.11
HGDP00103	ERS474392	PCR	Hazara	Central/South Asia	M	35.23
HGDP00104	ERS474393	PCR free	Hazara	Central/South Asia	M	28.43
HGDP00105	ERS474394	PCR	Hazara	Central/South Asia	M	35.3
HGDP00106	ERS474395	PCR free	Hazara	Central/South Asia	M	28.03
HGDP00108	ERS474396	PCR free	Hazara	Central/South Asia	M	32.01
HGDP00109	ERS474397	PCR free	Hazara	Central/South Asia	M	31.24
HGDP00110	ERS474398	PCR free	Hazara	Central/South Asia	M	33.07
HGDP00115	ERS474400	PCR free	Hazara	Central/South Asia	M	36.22
HGDP00118	ERS474402	PCR	Hazara	Central/South Asia	M	35.52
HGDP00120	ERS474404	PCR free	Hazara	Central/South Asia	M	31.55
HGDP00121	ERS1063305	PCR free	Hazara	Central/South Asia	M	36.52
HGDP00122	ERS1063301	PCR free	Hazara	Central/South Asia	M	34.39
HGDP00124	ERS1042251	PCR free	Hazara	Central/South Asia	M	64.01
HGDP00125	ERS1042252	PCR free	Hazara	Central/South Asia	M	65.89
HGDP00127	ERS474409	PCR	Hazara	Central/South Asia	M	35.53

HGDP00129	ERS1063300	PCR free	Hazara	Central/South Asia	M	36.23
HGDP00130	ERS474556	PCR	Makrani	Central/South Asia	M	33.91
HGDP00131	ERS474557	PCR free	Makrani	Central/South Asia	M	38.55
HGDP00133	ERS474558	PCR free	Makrani	Central/South Asia	M	37.22
HGDP00134	ERS474559	PCR free	Makrani	Central/South Asia	M	34.49
HGDP00135	ERS474560	PCR	Makrani	Central/South Asia	M	30.56
HGDP00136	ERS474561	PCR	Makrani	Central/South Asia	M	31.94
HGDP00137	ERS474562	PCR free	Makrani	Central/South Asia	M	34.72
HGDP00139	ERS474563	PCR free	Makrani	Central/South Asia	M	34.22
HGDP00140	ERS474564	PCR free	Makrani	Central/South Asia	M	33.61
HGDP00141	ERS474565	PCR free	Makrani	Central/South Asia	M	35.87
HGDP00143	ERS474566	PCR free	Makrani	Central/South Asia	M	33.61
HGDP00144	ERS474567	PCR free	Makrani	Central/South Asia	M	36.23
HGDP00145	ERS474568	PCR free	Makrani	Central/South Asia	M	34.83
HGDP00146	ERS474569	PCR free	Makrani	Central/South Asia	M	34.95
HGDP00148	ERS474570	PCR free	Makrani	Central/South Asia	M	34.21
HGDP00149	ERS474571	PCR free	Makrani	Central/South Asia	M	33.57
HGDP00150	ERS474572	PCR free	Makrani	Central/South Asia	M	33.21
HGDP00151	ERS474573	PCR free	Makrani	Central/South Asia	F	34.75
HGDP00153	ERS474574	PCR free	Makrani	Central/South Asia	F	56.61
HGDP00154	ERS474575	PCR free	Makrani	Central/South Asia	F	34.73
HGDP00155	ERS474576	PCR free	Makrani	Central/South Asia	F	35.76
HGDP00157	ERS1042161	PCR free	Makrani	Central/South Asia	F	40.23
HGDP00158	ERS474578	PCR free	Makrani	Central/South Asia	M	39.32
HGDP00160	ERS1042160	PCR free	Makrani	Central/South Asia	M	42.24
HGDP00161	ERS474580	PCR free	Makrani	Central/South Asia	M	37.54
HGDP00163	ERS474460	PCR	Sindhi	Central/South Asia	M	35.78
HGDP00165	ERS474461	PCR	Sindhi	Central/South Asia	M	34.65
HGDP00167	ERS474462	PCR	Sindhi	Central/South Asia	M	34
HGDP00169	ERS474463	PCR free	Sindhi	Central/South Asia	M	30.18
HGDP00171	ERS474464	PCR free	Sindhi	Central/South Asia	M	35.54
HGDP00173	ERS474465	PCR free	Sindhi	Central/South Asia	M	31.43
HGDP00175	ERS474466	PCR free	Sindhi	Central/South Asia	M	31.59
HGDP00177	ERS474467	PCR free	Sindhi	Central/South Asia	M	30.25
HGDP00179	ERS474468	PCR free	Sindhi	Central/South Asia	M	38.12
HGDP00181	ERS474469	PCR free	Sindhi	Central/South Asia	M	35.16
HGDP00183	ERS474470	PCR free	Sindhi	Central/South Asia	M	37.32
HGDP00185	ERS474471	PCR free	Sindhi	Central/South Asia	M	30.71
HGDP00187	ERS474472	PCR free	Sindhi	Central/South Asia	M	40.17
HGDP00189	ERS474473	PCR free	Sindhi	Central/South Asia	M	29.4
HGDP00191	ERS474474	PCR free	Sindhi	Central/South Asia	M	32.75
HGDP00192	ERS474475	PCR free	Sindhi	Central/South Asia	F	34.08

HGDP00195	ERS1042275	PCR free	Sindhi	Central/South Asia	F	38.5
HGDP00197	ERS474477	PCR free	Sindhi	Central/South Asia	M	33.28
HGDP00199	ERS474478	PCR free	Sindhi	Central/South Asia	M	32.45
HGDP00201	ERS474479	PCR free	Sindhi	Central/South Asia	M	35.9
HGDP00205	ERS474480	PCR free	Sindhi	Central/South Asia	M	32.07
HGDP00206	ERS474481	PCR free	Sindhi	Central/South Asia	F	34.5
HGDP00208	ERS1042274	PCR free	Sindhi	Central/South Asia	M	39.93
HGDP00210	ERS474483	PCR free	Sindhi	Central/South Asia	F	45.19
HGDP00213	ERS474436	PCR	Pathan	Central/South Asia	M	36.14
HGDP00214	ERS474437	PCR	Pathan	Central/South Asia	M	35.74
HGDP00216	ERS1042165	PCR free	Pathan	Central/South Asia	M	40.49
HGDP00218	ERS474439	PCR	Pathan	Central/South Asia	M	35.78
HGDP00222	ERS474440	PCR free	Pathan	Central/South Asia	M	34.98
HGDP00224	ERS474441	PCR free	Pathan	Central/South Asia	M	34.89
HGDP00226	ERS474442	PCR free	Pathan	Central/South Asia	M	35.65
HGDP00228	ERS474443	PCR free	Pathan	Central/South Asia	M	32.99
HGDP00230	ERS474444	PCR free	Pathan	Central/South Asia	M	34.91
HGDP00232	ERS1042166	PCR free	Pathan	Central/South Asia	F	41.28
HGDP00234	ERS474446	PCR free	Pathan	Central/South Asia	M	31.53
HGDP00237	ERS474447	PCR free	Pathan	Central/South Asia	F	35.48
HGDP00239	ERS474448	PCR free	Pathan	Central/South Asia	F	32.09
HGDP00241	ERS474449	PCR free	Pathan	Central/South Asia	M	34.86
HGDP00243	ERS474450	PCR free	Pathan	Central/South Asia	M	34.14
HGDP00244	ERS474451	PCR free	Pathan	Central/South Asia	F	36.29
HGDP00247	ERS474452	PCR free	Pathan	Central/South Asia	F	30.67
HGDP00248	ERS474453	PCR free	Pathan	Central/South Asia	M	32.76
HGDP00251	ERS474454	PCR free	Pathan	Central/South Asia	M	63.97
HGDP00254	ERS474455	PCR free	Pathan	Central/South Asia	M	33.57
HGDP00258	ERS474456	PCR free	Pathan	Central/South Asia	M	29.61
HGDP00259	ERS474457	PCR free	Pathan	Central/South Asia	M	36.34
HGDP00262	ERS474458	PCR free	Pathan	Central/South Asia	M	37.46
HGDP00264	ERS474459	PCR free	Pathan	Central/South Asia	M	37.41
HGDP00274	ERS474485	PCR free	Kalash	Central/South Asia	F	33.74
HGDP00277	ERS474486	PCR free	Kalash	Central/South Asia	M	30.63
HGDP00279	ERS474487	PCR	Kalash	Central/South Asia	M	32.38
HGDP00281	ERS474488	PCR free	Kalash	Central/South Asia	M	31.29
HGDP00285	ERS474489	PCR free	Kalash	Central/South Asia	M	35.36
HGDP00286	ERS1042260	PCR free	Kalash	Central/South Asia	F	33.34
HGDP00288	ERS474491	PCR	Kalash	Central/South Asia	M	31.32
HGDP00290	ERS474492	PCR free	Kalash	Central/South Asia	M	40.24
HGDP00298	ERS474493	PCR free	Kalash	Central/South Asia	F	37.43
HGDP00302	ERS474494	PCR free	Kalash	Central/South Asia	M	34.92

HGDP00304	ERS474495	PCR free	Kalash	Central/South Asia	F	29.98
HGDP00307	ERS474496	PCR free	Kalash	Central/South Asia	M	32.98
HGDP00309	ERS474497	PCR free	Kalash	Central/South Asia	M	31.21
HGDP00311	ERS474498	PCR free	Kalash	Central/South Asia	M	37.41
HGDP00313	ERS474499	PCR free	Kalash	Central/South Asia	M	34.31
HGDP00315	ERS474500	PCR free	Kalash	Central/South Asia	M	40.64
HGDP00319	ERS474501	PCR	Kalash	Central/South Asia	M	33.83
HGDP00323	ERS474502	PCR free	Kalash	Central/South Asia	F	45.79
HGDP00326	ERS474503	PCR free	Kalash	Central/South Asia	M	32.49
HGDP00328	ERS1042259	PCR free	Kalash	Central/South Asia	M	37.38
HGDP00330	ERS474505	PCR free	Kalash	Central/South Asia	M	31.2
HGDP00333	ERS474506	PCR free	Kalash	Central/South Asia	M	32.21
HGDP00338	ERS1042154	PCR free	Burusho	Central/South Asia	F	40.54
HGDP00341	ERS474413	PCR	Burusho	Central/South Asia	M	35.13
HGDP00346	ERS474414	PCR	Burusho	Central/South Asia	M	33.86
HGDP00351	ERS474415	PCR free	Burusho	Central/South Asia	M	30.73
HGDP00356	ERS474416	PCR free	Burusho	Central/South Asia	F	26.69
HGDP00359	ERS474417	PCR	Burusho	Central/South Asia	M	32.96
HGDP00364	ERS474418	PCR free	Burusho	Central/South Asia	M	30.79
HGDP00371	ERS474419	PCR free	Burusho	Central/South Asia	F	32.89
HGDP00372	ERS474420	PCR free	Burusho	Central/South Asia	M	30.68
HGDP00376	ERS474421	PCR free	Burusho	Central/South Asia	M	31.77
HGDP00382	ERS474422	PCR free	Burusho	Central/South Asia	M	28.64
HGDP00388	ERS474423	PCR free	Burusho	Central/South Asia	M	40.82
HGDP00392	ERS474424	PCR free	Burusho	Central/South Asia	M	33.17
HGDP00397	ERS474425	PCR free	Burusho	Central/South Asia	M	35.7
HGDP00402	ERS474426	PCR free	Burusho	Central/South Asia	M	35.35
HGDP00407	ERS474427	PCR free	Burusho	Central/South Asia	M	36.25
HGDP00412	ERS474428	PCR free	Burusho	Central/South Asia	M	35.54
HGDP00417	ERS474429	PCR free	Burusho	Central/South Asia	M	42.26
HGDP00423	ERS474430	PCR free	Burusho	Central/South Asia	M	32.84
HGDP00428	ERS1042242	PCR free	Burusho	Central/South Asia	M	42.19
HGDP00433	ERS474432	PCR free	Burusho	Central/South Asia	M	35.69
HGDP00438	ERS474433	PCR free	Burusho	Central/South Asia	M	35.32
HGDP00444	ERS474434	PCR free	Burusho	Central/South Asia	F	36.96
HGDP00445	ERS474435	PCR free	Burusho	Central/South Asia	M	35.03
HGDP00449	ERS1042179	PCR free	Mbuti	Africa	M	38.82
HGDP00450	ERS474066	PCR	Mbuti	Africa	M	34.52
HGDP00452	ERS474042	PCR	Biaka	Africa	M	32.08
HGDP00453	ERS474043	PCR	Biaka	Africa	M	34.79
HGDP00454	ERS813221	PCR free	Biaka	Africa	M	37.38
HGDP00455	ERS813222	PCR free	Biaka	Africa	M	38.9

HGDP00456	NA	PCR	Mbuti	Africa	M	26.28
HGDP00457	ERS1042238	PCR free	Biaka	Africa	M	43.94
HGDP00458	ERS474047	PCR	Biaka	Africa	M	35.03
HGDP00459	ERS474048	PCR	Biaka	Africa	M	34.87
HGDP00460	ERS474049	PCR free	Biaka	Africa	M	37.92
HGDP00461	ERS1042239	PCR free	Biaka	Africa	M	43.37
HGDP00462	ERS474068	PCR	Mbuti	Africa	M	36.45
HGDP00463	ERS474069	PCR	Mbuti	Africa	M	34.84
HGDP00464	ERS474051	PCR	Biaka	Africa	M	34.64
HGDP00466	ERS474053	PCR	Biaka	Africa	M	33.9
HGDP00467	ERS474070	PCR	Mbuti	Africa	M	38.65
HGDP00469	ERS474054	PCR	Biaka	Africa	M	36.72
HGDP00470	ERS474055	PCR	Biaka	Africa	M	30.77
HGDP00471	ERS474071	PCR	Mbuti	Africa	F	34.73
HGDP00472	ERS474056	PCR	Biaka	Africa	M	36.01
HGDP00473	ERS474057	PCR	Biaka	Africa	M	34.8
HGDP00474	ERS1358102	PCR free	Mbuti	Africa	M	41.18
HGDP00475	ERS474058	PCR	Biaka	Africa	M	34.79
HGDP00476	ERS1042178	PCR free	Mbuti	Africa	F	37.4
HGDP00478	ERS474074	PCR	Mbuti	Africa	M	38.3
HGDP00479	ERS474059	PCR	Biaka	Africa	M	34.35
HGDP00491	ERS474901	PCR	Bougainville	Oceania	M	36.94
HGDP00511	ERS474623	PCR free	French	Europe	M	34.01
HGDP00512	ERS474624	PCR free	French	Europe	M	31.47
HGDP00513	ERS474625	PCR free	French	Europe	F	35.59
HGDP00514	ERS474626	PCR free	French	Europe	F	34.91
HGDP00515	ERS474627	PCR free	French	Europe	M	35.3
HGDP00516	ERS474628	PCR free	French	Europe	F	32.29
HGDP00517	ERS474629	PCR free	French	Europe	F	38.18
HGDP00518	ERS474630	PCR free	French	Europe	M	37.03
HGDP00519	ERS474631	PCR free	French	Europe	M	27.32
HGDP00520	ERS474632	PCR free	French	Europe	F	37.45
HGDP00521	NA	PCR	French	Europe	M	28.63
HGDP00522	ERS474634	PCR free	French	Europe	M	32.87
HGDP00523	ERS474635	PCR free	French	Europe	F	35.25
HGDP00524	ERS474636	PCR free	French	Europe	F	35.34
HGDP00525	ERS474637	PCR free	French	Europe	M	35.43
HGDP00526	ERS1042156	PCR free	French	Europe	F	39.48
HGDP00527	ERS474639	PCR free	French	Europe	F	32.56
HGDP00528	ERS474640	PCR free	French	Europe	M	39.09
HGDP00529	ERS474641	PCR free	French	Europe	F	31.27
HGDP00530	ERS1042155	PCR free	French	Europe	M	43.81

HGDP00531	ERS474643	PCR free	French	Europe	F	35.58
HGDP00533	ERS1358149	PCR	French	Europe	M	44.37
HGDP00534	ERS474645	PCR free	French	Europe	F	40.22
HGDP00535	ERS474646	PCR free	French	Europe	F	34.18
HGDP00536	ERS474647	PCR free	French	Europe	F	38.95
HGDP00537	ERS474648	PCR free	French	Europe	F	35.95
HGDP00538	ERS474649	PCR free	French	Europe	M	32.79
HGDP00539	ERS474650	PCR free	French	Europe	F	36.91
HGDP00540	ERS1358135	PCR free	PapuanSepik	Oceania	M	46.59
HGDP00541	ERS474913	PCR	PapuanSepik	Oceania	M	33.03
HGDP00542	ERS474914	PCR free	PapuanSepik	Oceania	M	37.56
HGDP00543	ERS474915	PCR	PapuanSepik	Oceania	M	34.31
HGDP00544	ERS474916	PCR	PapuanSepik	Oceania	F	33.31
HGDP00545	ERS474917	PCR	PapuanSepik	Oceania	M	32.57
HGDP00546	ERS474918	PCR free	PapuanSepik	Oceania	M	38.04
HGDP00547	ERS474919	PCR	PapuanSepik	Oceania	M	33.65
HGDP00548	ERS1255110	PCR free	PapuanHighlands	Oceania	M	49.02
HGDP00549	ERS474921	PCR free	PapuanHighlands	Oceania	M	37.25
HGDP00550	ERS474922	PCR	PapuanHighlands	Oceania	F	32.05
HGDP00551	ERS474923	PCR free	PapuanHighlands	Oceania	M	35.82
HGDP00552	ERS1255095	PCR free	PapuanHighlands	Oceania	F	42.39
HGDP00553	ERS474925	PCR free	PapuanHighlands	Oceania	M	35.28
HGDP00554	ERS1255102	PCR free	PapuanHighlands	Oceania	F	34.26
HGDP00555	ERS474927	PCR	PapuanHighlands	Oceania	M	33.34
HGDP00556	ERS474928	PCR free	PapuanHighlands	Oceania	M	34.56
HGDP00557	ERS474739	PCR free	Druze	Middle East	F	40.46
HGDP00558	ERS474740	PCR free	Druze	Middle East	F	36.55
HGDP00559	ERS474741	PCR free	Druze	Middle East	F	35.82
HGDP00560	ERS474742	PCR free	Druze	Middle East	F	43.12
HGDP00561	ERS474743	PCR free	Druze	Middle East	F	29.41
HGDP00562	ERS474744	PCR	Druze	Middle East	M	32.53
HGDP00563	ERS474745	PCR free	Druze	Middle East	F	42.33
HGDP00564	ERS474746	PCR free	Druze	Middle East	F	35
HGDP00565	ERS474747	PCR free	Druze	Middle East	F	33.91
HGDP00566	ERS474748	PCR free	Druze	Middle East	F	39.33
HGDP00567	ERS474749	PCR free	Druze	Middle East	F	37.99
HGDP00568	ERS474750	PCR free	Druze	Middle East	F	32.92
HGDP00569	ERS1255088	PCR free	Druze	Middle East	F	48.98
HGDP00571	ERS474752	PCR free	Druze	Middle East	F	34.19
HGDP00572	ERS474753	PCR free	Druze	Middle East	F	35.37
HGDP00573	ERS474754	PCR free	Druze	Middle East	F	37.79
HGDP00574	ERS474755	PCR free	Druze	Middle East	F	40.64

HGDP00575	ERS474756	PCR free	Druze	Middle East	F	36.89
HGDP00576	ERS474757	PCR	Druze	Middle East	M	28.61
HGDP00577	ERS474758	PCR free	Druze	Middle East	F	39.12
HGDP00578	ERS474759	PCR free	Druze	Middle East	F	36.96
HGDP00579	ERS474760	PCR free	Druze	Middle East	F	27.21
HGDP00580	ERS474761	PCR	Druze	Middle East	M	35.2
HGDP00581	ERS474762	PCR free	Druze	Middle East	F	33.88
HGDP00582	ERS474763	PCR free	Druze	Middle East	F	35.9
HGDP00583	ERS474764	PCR free	Druze	Middle East	F	35.2
HGDP00584	ERS474765	PCR free	Druze	Middle East	F	37.31
HGDP00586	ERS474766	PCR free	Druze	Middle East	F	40.03
HGDP00587	ERS474767	PCR free	Druze	Middle East	F	36.34
HGDP00588	ERS474768	PCR free	Druze	Middle East	M	35.53
HGDP00590	ERS474769	PCR free	Druze	Middle East	F	29.72
HGDP00591	ERS474770	PCR free	Druze	Middle East	F	38.81
HGDP00594	ERS474771	PCR free	Druze	Middle East	M	33.47
HGDP00595	ERS474772	PCR free	Druze	Middle East	M	38.17
HGDP00597	ERS1042245	PCR free	Druze	Middle East	M	44.17
HGDP00598	ERS474773	PCR free	Druze	Middle East	M	40.42
HGDP00599	ERS474774	PCR free	Druze	Middle East	M	37.99
HGDP00600	ERS474775	PCR free	Druze	Middle East	M	35.95
HGDP00601	ERS474776	PCR free	Druze	Middle East	F	37.44
HGDP00602	ERS474777	PCR free	Druze	Middle East	M	34.31
HGDP00604	ERS474778	PCR free	Druze	Middle East	M	32.84
HGDP00606	ERS474779	PCR free	Druze	Middle East	F	36.19
HGDP00607	ERS474826	PCR free	Bedouin	Middle East	F	32.16
HGDP00608	ERS474827	PCR	Bedouin	Middle East	M	30.14
HGDP00609	ERS474828	PCR	Bedouin	Middle East	M	35.94
HGDP00610	ERS474829	PCR	Bedouin	Middle East	M	34.68
HGDP00611	ERS474830	PCR free	Bedouin	Middle East	M	43.32
HGDP00612	ERS474831	PCR free	Bedouin	Middle East	F	33.53
HGDP00613	ERS474832	PCR free	Bedouin	Middle East	F	33.4
HGDP00614	ERS474833	PCR free	Bedouin	Middle East	F	35.69
HGDP00615	ERS1063307	PCR free	Bedouin	Middle East	F	27.53
HGDP00616	ERS1042237	PCR free	Bedouin	Middle East	M	43.49
HGDP00618	ERS474836	PCR free	Bedouin	Middle East	M	34.79
HGDP00619	ERS474837	PCR free	Bedouin	Middle East	M	37.08
HGDP00620	ERS474838	PCR free	Bedouin	Middle East	M	38.56
HGDP00621	ERS474839	PCR free	Bedouin	Middle East	M	34.66
HGDP00622	ERS474840	PCR free	Bedouin	Middle East	M	32.3
HGDP00623	ERS474841	PCR free	Bedouin	Middle East	M	35.33
HGDP00624	ERS474842	PCR free	Bedouin	Middle East	M	36.88

HGDP00625	ERS813223	PCR free	Bedouin	Middle East	M	39.19
HGDP00626	ERS474844	PCR free	Bedouin	Middle East	M	36.14
HGDP00627	ERS474845	PCR free	Bedouin	Middle East	M	35.24
HGDP00628	ERS1063302	PCR free	Bedouin	Middle East	M	38.08
HGDP00629	ERS474847	PCR free	Bedouin	Middle East	M	41.68
HGDP00630	ERS474848	PCR free	Bedouin	Middle East	M	47.2
HGDP00631	ERS1063413	PCR free	Bedouin	Middle East	M	35.27
HGDP00632	ERS474850	PCR free	Bedouin	Middle East	F	30.75
HGDP00634	ERS474851	PCR free	Bedouin	Middle East	F	31.61
HGDP00635	ERS474852	PCR free	Bedouin	Middle East	F	32.62
HGDP00636	ERS474853	PCR free	Bedouin	Middle East	F	37.5
HGDP00637	ERS474854	PCR free	Bedouin	Middle East	F	34.25
HGDP00638	ERS474855	PCR free	Bedouin	Middle East	F	32.24
HGDP00639	ERS474856	PCR free	Bedouin	Middle East	M	31.97
HGDP00640	ERS474857	PCR free	Bedouin	Middle East	M	36.15
HGDP00641	ERS474858	PCR free	Bedouin	Middle East	M	33.7
HGDP00642	ERS474859	PCR free	Bedouin	Middle East	M	27.36
HGDP00643	ERS474860	PCR free	Bedouin	Middle East	F	34.57
HGDP00644	ERS474861	PCR free	Bedouin	Middle East	M	31.34
HGDP00645	ERS474862	PCR free	Bedouin	Middle East	M	31.85
HGDP00646	ERS474863	PCR free	Bedouin	Middle East	F	33.44
HGDP00647	ERS474864	PCR free	Bedouin	Middle East	F	35.29
HGDP00648	ERS474865	PCR free	Bedouin	Middle East	M	31.45
HGDP00649	ERS474866	PCR free	Bedouin	Middle East	F	34.77
HGDP00650	ERS1042149	PCR free	Bedouin	Middle East	F	41.04
HGDP00651	ERS474868	PCR free	Bedouin	Middle East	F	31.39
HGDP00653	ERS474869	PCR free	Bedouin	Middle East	F	32.13
HGDP00654	ERS474870	PCR free	Bedouin	Middle East	M	38.41
HGDP00655	ERS474902	PCR	Bougainville	Oceania	M	37.06
HGDP00656	ERS1042151	PCR free	Bougainville	Oceania	F	38.28
HGDP00660	ERS1042150	PCR free	Bougainville	Oceania	F	41.1
HGDP00661	ERS474905	PCR free	Bougainville	Oceania	F	44.29
HGDP00662	ERS474906	PCR	Bougainville	Oceania	M	37.24
HGDP00663	ERS474907	PCR free	Bougainville	Oceania	F	36.14
HGDP00664	ERS474908	PCR free	Bougainville	Oceania	F	37.28
HGDP00665	NA	PCR	Sardinian	Europe	M	26.48
HGDP00666	ERS474712	PCR free	Sardinian	Europe	M	33.83
HGDP00667	ERS474713	PCR free	Sardinian	Europe	F	35.6
HGDP00668	ERS474714	PCR free	Sardinian	Europe	M	30.31
HGDP00669	ERS474715	PCR free	Sardinian	Europe	F	33.6
HGDP00670	ERS474716	PCR free	Sardinian	Europe	M	46.76
HGDP00671	ERS474717	PCR free	Sardinian	Europe	M	38.55

HGDP00672	ERS474718	PCR free	Sardinian	Europe	F	35.37
HGDP00673	ERS474719	PCR free	Sardinian	Europe	F	37.2
HGDP00674	ERS474720	PCR free	Sardinian	Europe	M	37.47
HGDP00675	ERS474780	PCR free	Palestinian	Middle East	M	39.8
HGDP00676	ERS474781	PCR free	Palestinian	Middle East	M	40.14
HGDP00677	ERS474782	PCR free	Palestinian	Middle East	M	34.12
HGDP00678	ERS474783	PCR free	Palestinian	Middle East	M	33.88
HGDP00679	ERS474784	PCR free	Palestinian	Middle East	F	34.87
HGDP00680	ERS474785	PCR free	Palestinian	Middle East	F	32.87
HGDP00682	ERS474786	PCR free	Palestinian	Middle East	F	33.54
HGDP00683	ERS474787	PCR free	Palestinian	Middle East	F	33.29
HGDP00684	ERS474788	PCR free	Palestinian	Middle East	F	35.63
HGDP00685	ERS474789	PCR free	Palestinian	Middle East	F	36.05
HGDP00686	ERS474790	PCR free	Palestinian	Middle East	F	33.15
HGDP00687	ERS474791	PCR free	Palestinian	Middle East	F	32.29
HGDP00688	ERS474792	PCR free	Palestinian	Middle East	F	37.32
HGDP00689	ERS474793	PCR free	Palestinian	Middle East	F	33.04
HGDP00690	ERS1085932	PCR free	Palestinian	Middle East	F	34.69
HGDP00691	ERS474795	PCR free	Palestinian	Middle East	F	33.71
HGDP00692	ERS474796	PCR free	Palestinian	Middle East	F	60.64
HGDP00693	ERS474797	PCR free	Palestinian	Middle East	F	35.13
HGDP00694	ERS474798	PCR free	Palestinian	Middle East	F	33.71
HGDP00696	ERS474799	PCR free	Palestinian	Middle East	F	35.17
HGDP00697	ERS474800	PCR free	Palestinian	Middle East	F	37.19
HGDP00698	ERS474801	PCR free	Palestinian	Middle East	F	36.66
HGDP00699	ERS474802	PCR free	Palestinian	Middle East	F	33.83
HGDP00700	ERS474803	PCR free	Palestinian	Middle East	F	35.71
HGDP00701	ERS474871	PCR free	Bedouin	Middle East	F	31.42
HGDP00702	ERS1042167	PCR free	Colombian	America	F	40.24
HGDP00703	ERS474965	PCR	Colombian	America	M	32.18
HGDP00704	ERS474966	PCR free	Colombian	America	F	38.11
HGDP00706	ERS1042168	PCR free	Colombian	America	F	40.7
HGDP00708	ERS474968	PCR free	Colombian	America	F	36.44
HGDP00710	ERS474969	PCR	Colombian	America	M	35.29
HGDP00711	ERS474315	PCR	Cambodian	East Asia	M	36.07
HGDP00712	ERS474316	PCR free	Cambodian	East Asia	F	28.63
HGDP00713	ERS1042244	PCR free	Cambodian	East Asia	F	38.49
HGDP00714	ERS474318	PCR	Cambodian	East Asia	M	36.34
HGDP00715	ERS474319	PCR	Cambodian	East Asia	M	35.7
HGDP00716	ERS474320	PCR free	Cambodian	East Asia	M	37.01
HGDP00717	ERS1042243	PCR free	Cambodian	East Asia	M	42.66
HGDP00719	ERS474322	PCR free	Cambodian	East Asia	F	32.47

HGDP00721	ERS474324	PCR free	Cambodian	East Asia	F	33.07
HGDP00722	ERS1358098	PCR free	Palestinian	Middle East	M	46.18
HGDP00723	ERS474805	PCR	Palestinian	Middle East	M	31.53
HGDP00724	ERS474806	PCR	Palestinian	Middle East	M	31.25
HGDP00725	ERS1042288	PCR free	Palestinian	Middle East	M	42.31
HGDP00726	ERS474808	PCR	Palestinian	Middle East	M	35.06
HGDP00727	ERS474809	PCR free	Palestinian	Middle East	M	33.3
HGDP00729	ERS1063416	PCR free	Palestinian	Middle East	M	38.93
HGDP00730	ERS474811	PCR free	Palestinian	Middle East	M	36.57
HGDP00731	ERS474812	PCR free	Palestinian	Middle East	M	32.15
HGDP00732	ERS474813	PCR free	Palestinian	Middle East	M	35.79
HGDP00733	ERS474814	PCR free	Palestinian	Middle East	M	35.27
HGDP00734	ERS474815	PCR free	Palestinian	Middle East	M	37.21
HGDP00735	ERS474816	PCR free	Palestinian	Middle East	F	32.99
HGDP00736	ERS474817	PCR free	Palestinian	Middle East	F	37.72
HGDP00737	ERS1042233	PCR free	Palestinian	Middle East	F	39.39
HGDP00738	ERS474819	PCR free	Palestinian	Middle East	F	32.8
HGDP00739	ERS474820	PCR free	Palestinian	Middle East	F	33.7
HGDP00740	ERS474821	PCR free	Palestinian	Middle East	F	36.23
HGDP00741	ERS474822	PCR free	Palestinian	Middle East	F	34.42
HGDP00744	ERS474823	PCR free	Palestinian	Middle East	F	31.24
HGDP00745	ERS474824	PCR free	Palestinian	Middle East	F	36.24
HGDP00746	ERS474825	PCR free	Palestinian	Middle East	F	33.54
HGDP00747	ERS474325	PCR free	Japanese	East Asia	M	28.24
HGDP00748	ERS474326	PCR free	Japanese	East Asia	M	33.35
HGDP00749	ERS1042140	PCR free	Japanese	East Asia	M	83.81
HGDP00750	ERS474328	PCR free	Japanese	East Asia	M	31
HGDP00751	ERS474329	PCR free	Japanese	East Asia	M	29.67
HGDP00752	ERS1063417	PCR free	Japanese	East Asia	M	38.04
HGDP00753	ERS1063418	PCR free	Japanese	East Asia	M	32.86
HGDP00754	ERS474332	PCR free	Japanese	East Asia	F	27.32
HGDP00755	ERS474333	PCR free	Japanese	East Asia	M	34.7
HGDP00756	ERS474334	PCR free	Japanese	East Asia	F	31.68
HGDP00757	ERS474335	PCR free	Japanese	East Asia	M	34.98
HGDP00758	ERS474336	PCR free	Japanese	East Asia	M	29.05
HGDP00759	ERS474337	PCR free	Japanese	East Asia	M	37.13
HGDP00760	ERS474338	PCR free	Japanese	East Asia	F	28.73
HGDP00761	ERS474339	PCR free	Japanese	East Asia	F	29.12
HGDP00762	ERS474340	PCR free	Japanese	East Asia	M	31.61
HGDP00764	ERS474342	PCR free	Japanese	East Asia	M	30.05
HGDP00765	ERS474343	PCR free	Japanese	East Asia	F	34
HGDP00766	ERS474344	PCR free	Japanese	East Asia	M	30.76

HGDP00767	ERS474345	PCR free	Japanese	East Asia	M	30.82
HGDP00769	ERS474347	PCR free	Japanese	East Asia	M	41.03
HGDP00771	ERS474348	PCR free	Japanese	East Asia	F	28.38
HGDP00772	ERS474349	PCR free	Japanese	East Asia	F	28.46
HGDP00773	ERS1042175	PCR free	Japanese	East Asia	F	41.23
HGDP00774	ERS474147	PCR free	Han	East Asia	M	38.08
HGDP00775	ERS1358123	PCR	Han	East Asia	M	36.89
HGDP00776	ERS474149	PCR free	Han	East Asia	F	32.23
HGDP00777	ERS474150	PCR free	Han	East Asia	M	30.97
HGDP00778	NA	PCR	Han	East Asia	M	29.87
HGDP00779	ERS474152	PCR free	Han	East Asia	M	32.56
HGDP00780	ERS474153	PCR free	Han	East Asia	M	34.18
HGDP00781	ERS474154	PCR free	Han	East Asia	F	35.3
HGDP00782	ERS474155	PCR free	Han	East Asia	M	31.78
HGDP00783	ERS1042157	PCR free	Han	East Asia	F	67.17
HGDP00784	ERS474157	PCR free	Han	East Asia	F	32.03
HGDP00785	ERS1042158	PCR free	Han	East Asia	M	42.13
HGDP00786	ERS474159	PCR free	Han	East Asia	M	34.05
HGDP00787	ERS474909	PCR free	Bougainville	Oceania	F	41.41
HGDP00788	ERS474910	PCR	Bougainville	Oceania	M	35.63
HGDP00790	ERS474351	PCR free	Japanese	East Asia	M	37.95
HGDP00791	ERS474352	PCR free	Japanese	East Asia	M	32.05
HGDP00794	ERS474675	PCR free	Orcadian	Europe	F	36.73
HGDP00795	ERS474676	PCR free	Orcadian	Europe	M	34.14
HGDP00796	ERS1042164	PCR free	Orcadian	Europe	F	39.14
HGDP00797	ERS474678	PCR free	Orcadian	Europe	F	32.77
HGDP00798	ERS1042163	PCR free	Orcadian	Europe	M	40.31
HGDP00799	ERS474680	PCR free	Orcadian	Europe	F	36.34
HGDP00800	ERS474681	PCR free	Orcadian	Europe	F	36.79
HGDP00802	ERS474682	PCR free	Orcadian	Europe	F	37.67
HGDP00803	ERS474683	PCR free	Orcadian	Europe	M	33.76
HGDP00804	ERS474684	PCR free	Orcadian	Europe	M	38.93
HGDP00805	ERS474685	PCR free	Orcadian	Europe	F	31.63
HGDP00806	ERS474686	PCR free	Orcadian	Europe	F	33.85
HGDP00807	ERS474687	PCR free	Orcadian	Europe	M	64.75
HGDP00808	ERS474688	PCR free	Orcadian	Europe	M	36.08
HGDP00810	ERS474689	PCR free	Orcadian	Europe	M	37.82
HGDP00811	ERS474160	PCR free	Han	East Asia	F	31.25
HGDP00812	ERS474161	PCR free	Han	East Asia	F	33.36
HGDP00813	ERS474162	PCR free	Han	East Asia	F	31.53
HGDP00814	ERS474163	PCR free	Han	East Asia	F	33.07
HGDP00815	ERS474164	PCR free	Han	East Asia	M	34.29

HGDP00817	ERS474165	PCR free	Han	East Asia	F	34.27
HGDP00818	ERS474166	PCR free	Han	East Asia	F	31.13
HGDP00819	ERS474167	PCR free	Han	East Asia	M	36.19
HGDP00820	ERS474168	PCR free	Han	East Asia	F	29.73
HGDP00821	ERS474169	PCR free	Han	East Asia	M	32.82
HGDP00822	ERS474170	PCR free	Han	East Asia	M	61.96
HGDP00828	ERS474353	PCR free	Japanese	East Asia	M	37.4
HGDP00832	ERS474971	PCR free	Surui	America	F	37.9
HGDP00837	ERS474972	PCR	Surui	America	M	36.74
HGDP00838	ERS474973	PCR free	Surui	America	F	38.29
HGDP00843	ERS474974	PCR	Surui	America	M	37.17
HGDP00845	ERS474975	PCR	Surui	America	M	35.16
HGDP00846	ERS1042171	PCR free	Surui	America	F	40.81
HGDP00849	ERS474977	PCR free	Surui	America	M	35.02
HGDP00852	ERS1042172	PCR free	Surui	America	F	34.96
HGDP00854	ERS474943	PCR free	Maya	America	F	39
HGDP00855	ERS1042268	PCR free	Maya	America	F	36.85
HGDP00856	ERS474945	PCR	Maya	America	M	33.69
HGDP00857	ERS1042269	PCR free	Maya	America	F	44.4
HGDP00858	ERS474947	PCR free	Maya	America	F	35.72
HGDP00859	ERS474948	PCR free	Maya	America	F	42.23
HGDP00860	ERS474949	PCR free	Maya	America	F	40.73
HGDP00861	ERS474950	PCR free	Maya	America	F	38.07
HGDP00862	ERS474951	PCR free	Maya	America	F	36.63
HGDP00863	ERS474952	PCR free	Maya	America	F	37.7
HGDP00864	ERS474953	PCR free	Maya	America	F	34.78
HGDP00865	ERS474954	PCR free	Maya	America	F	38.72
HGDP00868	ERS474955	PCR free	Maya	America	F	34.83
HGDP00869	ERS474956	PCR free	Maya	America	F	38.45
HGDP00870	ERS474957	PCR free	Maya	America	F	39.23
HGDP00871	ERS474958	PCR free	Maya	America	F	36.32
HGDP00872	ERS474959	PCR free	Maya	America	F	37.14
HGDP00873	ERS474960	PCR free	Maya	America	F	41.93
HGDP00875	ERS474961	PCR free	Maya	America	F	33
HGDP00876	ERS474962	PCR free	Maya	America	F	37.2
HGDP00877	ERS474963	PCR	Maya	America	M	35.08
HGDP00879	ERS474598	PCR free	Russian	Europe	M	37.14
HGDP00880	ERS474599	PCR free	Russian	Europe	M	36.55
HGDP00881	ERS474600	PCR free	Russian	Europe	F	36.22
HGDP00882	ERS474601	PCR free	Russian	Europe	M	32.88
HGDP00883	ERS474602	PCR free	Russian	Europe	M	33.92
HGDP00884	ERS474603	PCR free	Russian	Europe	F	31.8

HGDP00885	ERS474604	PCR free	Russian	Europe	F	34.38
HGDP00886	ERS474605	PCR free	Russian	Europe	M	43.35
HGDP00887	ERS1042231	PCR free	Russian	Europe	M	39.72
HGDP00888	ERS474607	PCR free	Russian	Europe	M	38.12
HGDP00889	ERS474608	PCR free	Russian	Europe	F	35.94
HGDP00890	ERS474609	PCR free	Russian	Europe	M	40.33
HGDP00891	ERS474610	PCR free	Russian	Europe	M	37.2
HGDP00892	ERS474611	PCR free	Russian	Europe	M	36.68
HGDP00893	ERS474612	PCR free	Russian	Europe	M	38.54
HGDP00894	ERS474613	PCR free	Russian	Europe	M	37.48
HGDP00895	ERS474614	PCR free	Russian	Europe	M	34.05
HGDP00896	ERS474615	PCR free	Russian	Europe	M	36.38
HGDP00897	ERS474616	PCR free	Russian	Europe	M	38.14
HGDP00898	ERS474617	PCR free	Russian	Europe	F	39.92
HGDP00899	ERS474618	PCR free	Russian	Europe	F	36.32
HGDP00900	ERS474619	PCR free	Russian	Europe	M	41.01
HGDP00901	ERS474620	PCR free	Russian	Europe	F	39.57
HGDP00902	ERS474621	PCR free	Russian	Europe	F	34.83
HGDP00903	ERS1042232	PCR free	Russian	Europe	F	38.22
HGDP00904	ERS474125	PCR	Mandenka	Africa	M	34.12
HGDP00905	ERS474126	PCR	Mandenka	Africa	M	32.72
HGDP00906	ERS813224	PCR free	Mandenka	Africa	M	39.64
HGDP00907	ERS474128	PCR	Mandenka	Africa	M	34.13
HGDP00908	ERS474129	PCR	Mandenka	Africa	M	34.17
HGDP00909	ERS813225	PCR free	Mandenka	Africa	F	42.88
HGDP00910	ERS474131	PCR	Mandenka	Africa	F	33.04
HGDP00911	ERS474132	PCR	Mandenka	Africa	M	33.32
HGDP00912	ERS813226	PCR free	Mandenka	Africa	M	36.23
HGDP00913	ERS813227	PCR free	Mandenka	Africa	M	32.01
HGDP00914	ERS474135	PCR	Mandenka	Africa	F	33.48
HGDP00915	ERS1042267	PCR free	Mandenka	Africa	F	38.23
HGDP00917	ERS474137	PCR free	Mandenka	Africa	F	32.69
HGDP00918	ERS474138	PCR	Mandenka	Africa	F	34
HGDP00920	ERS474103	PCR	Yoruba	Africa	F	32.63
HGDP00924	ERS474104	PCR	Yoruba	Africa	F	32.05
HGDP00925	ERS474105	PCR	Yoruba	Africa	F	32.19
HGDP00926	ERS474106	PCR	Yoruba	Africa	F	30.64
HGDP00927	NA	PCR	Yoruba	Africa	M	34.21
HGDP00928	ERS1042283	PCR free	Yoruba	Africa	F	35.17
HGDP00929	ERS474109	PCR	Yoruba	Africa	M	34.24
HGDP00930	ERS474110	PCR free	Yoruba	Africa	M	35.69
HGDP00931	ERS474111	PCR	Yoruba	Africa	M	36.24

HGDP00932	ERS1042284	PCR free	Yoruba	Africa	M	34.3
HGDP00933	ERS474113	PCR	Yoruba	Africa	F	34.55
HGDP00934	ERS474114	PCR	Yoruba	Africa	F	35.84
HGDP00935	ERS474115	PCR	Yoruba	Africa	F	36.17
HGDP00936	ERS1358129	PCR	Yoruba	Africa	M	40.02
HGDP00937	ERS474117	PCR	Yoruba	Africa	M	34.18
HGDP00938	ERS474118	PCR	Yoruba	Africa	F	27.9
HGDP00939	ERS474119	PCR	Yoruba	Africa	F	34.53
HGDP00940	ERS474120	PCR	Yoruba	Africa	M	36.3
HGDP00941	ERS474121	PCR	Yoruba	Africa	M	37.67
HGDP00942	ERS474122	PCR	Yoruba	Africa	M	35.18
HGDP00943	ERS474123	PCR	Yoruba	Africa	M	36.23
HGDP00944	ERS474124	PCR	Yoruba	Africa	M	35.35
HGDP00945	ERS474354	PCR	Yakut	East Asia	M	36.48
HGDP00946	ERS474355	PCR free	Yakut	East Asia	M	34.62
HGDP00947	ERS474356	PCR	Yakut	East Asia	M	35.99
HGDP00948	ERS474357	PCR	Yakut	East Asia	M	35.36
HGDP00949	ERS474358	PCR free	Yakut	East Asia	M	30.29
HGDP00950	ERS474359	PCR free	Yakut	East Asia	M	27.87
HGDP00951	ERS1255089	PCR free	Yakut	East Asia	M	47
HGDP00952	ERS474361	PCR free	Yakut	East Asia	M	27.9
HGDP00953	ERS474362	PCR free	Yakut	East Asia	M	29.58
HGDP00954	ERS474363	PCR free	Yakut	East Asia	M	34.45
HGDP00955	ERS474364	PCR free	Yakut	East Asia	F	31.16
HGDP00956	ERS1255042	PCR free	Yakut	East Asia	F	39.19
HGDP00957	ERS474366	PCR free	Yakut	East Asia	F	38.75
HGDP00958	ERS474367	PCR free	Yakut	East Asia	M	45.08
HGDP00959	ERS474368	PCR free	Yakut	East Asia	F	28.59
HGDP00960	ERS474369	PCR free	Yakut	East Asia	M	29.43
HGDP00961	ERS474370	PCR free	Yakut	East Asia	M	27.82
HGDP00962	ERS474371	PCR free	Yakut	East Asia	M	31.53
HGDP00963	ERS474372	PCR free	Yakut	East Asia	F	33.05
HGDP00964	ERS474373	PCR free	Yakut	East Asia	M	30.1
HGDP00965	ERS474374	PCR free	Yakut	East Asia	M	31.29
HGDP00966	ERS474375	PCR free	Yakut	East Asia	F	32.2
HGDP00967	ERS474376	PCR free	Yakut	East Asia	F	30.69
HGDP00968	ERS474377	PCR free	Yakut	East Asia	M	24.92
HGDP00969	ERS474378	PCR free	Yakut	East Asia	M	29.22
HGDP00970	ERS474970	PCR free	Colombian	America	F	36.31
HGDP00971	ERS474171	PCR free	Han	East Asia	M	31.51
HGDP00972	ERS474172	PCR free	Han	East Asia	F	30.92
HGDP00973	ERS474173	PCR free	Han	East Asia	M	35.44

HGDP00974	ERS474174	PCR free	Han	East Asia	F	34.33
HGDP00975	ERS474175	PCR free	Han	East Asia	F	34.15
HGDP00976	ERS474176	PCR free	Han	East Asia	F	28.98
HGDP00977	ERS474177	PCR free	Han	East Asia	M	33.26
HGDP00982	ERS1358125	PCR	Mbuti	Africa	M	38.71
HGDP00984	ERS474076	PCR	Mbuti	Africa	M	38.35
HGDP00985	ERS474060	PCR	Biaka	Africa	M	32.83
HGDP00986	ERS474061	PCR	Biaka	Africa	M	32.98
HGDP00987	ERS1042176	PCR free	San	Africa	M	34.15
HGDP00991	ERS1255120	PCR free	San	Africa	M	46.17
HGDP00992	ERS474080	PCR	San	Africa	M	37.17
HGDP00993	ERS474084	PCR	BantuSouthAfrica	Africa	M	34.85
HGDP00994	ERS474085	PCR	BantuSouthAfrica	Africa	M	33.6
HGDP00995	ERS474979	PCR free	Karitiana	America	F	46.32
HGDP00998	NA	PCR	Karitiana	America	M	27.72
HGDP00999	ERS474981	PCR free	Karitiana	America	F	36.02
HGDP01001	ERS474982	PCR free	Karitiana	America	F	32.38
HGDP01009	ERS474985	PCR	Karitiana	America	M	36.97
HGDP01010	ERS474986	PCR free	Karitiana	America	F	39.26
HGDP01012	ERS1042261	PCR free	Karitiana	America	M	44.93
HGDP01013	ERS474988	PCR	Karitiana	America	M	37.61
HGDP01014	ERS474989	PCR free	Karitiana	America	F	36.11
HGDP01015	ERS1358130	PCR	Karitiana	America	M	36.16
HGDP01018	ERS1042262	PCR free	Karitiana	America	F	38.3
HGDP01019	ERS474992	PCR free	Karitiana	America	M	37.46
HGDP01021	ERS474178	PCR free	Han	East Asia	F	32.13
HGDP01023	ERS474179	PCR free	Han	East Asia	F	33.15
HGDP01027	ERS474911	PCR free	Bougainville	Oceania	F	43.23
HGDP01028	ERS1255098	PCR free	BantuSouthAfrica	Africa	M	44.76
HGDP01029	ERS474081	PCR free	San	Africa	M	35.63
HGDP01030	ERS1255115	PCR free	BantuSouthAfrica	Africa	M	43.71
HGDP01031	ERS474088	PCR	BantuSouthAfrica	Africa	M	34.48
HGDP01032	ERS1042177	PCR free	San	Africa	M	37.19
HGDP01033	ERS474089	PCR free	BantuSouthAfrica	Africa	M	35.7
HGDP01034	ERS1042125	PCR free	BantuSouthAfrica	Africa	M	38.94
HGDP01035	ERS1042145	PCR free	BantuSouthAfrica	Africa	M	41.05
HGDP01036	ERS1358127	PCR	San	Africa	M	39.97
HGDP01037	ERS474929	PCR	Pima	America	M	34.17
HGDP01041	ERS475066	PCR free	Pima	America	F	33.11
HGDP01043	ERS474931	PCR	Pima	America	M	29.84
HGDP01044	ERS1042230	PCR free	Pima	America	F	37.38
HGDP01047	ERS1042229	PCR free	Pima	America	M	42.96

HGDP01050	ERS474934	PCR	Pima	America	M	33.09
HGDP01053	ERS474936	PCR free	Pima	America	F	35.62
HGDP01055	ERS474937	PCR free	Pima	America	M	36.41
HGDP01056	ERS474938	PCR free	Pima	America	F	36.24
HGDP01057	ERS474939	PCR free	Pima	America	M	40.14
HGDP01058	ERS474940	PCR free	Pima	America	F	36.91
HGDP01059	ERS474941	PCR free	Pima	America	M	36.83
HGDP01060	ERS474942	PCR free	Pima	America	M	65.05
HGDP01062	ERS474721	PCR free	Sardinian	Europe	F	36.21
HGDP01063	ERS474722	PCR free	Sardinian	Europe	M	35.97
HGDP01064	ERS474723	PCR free	Sardinian	Europe	F	34.95
HGDP01065	ERS474724	PCR free	Sardinian	Europe	F	36.43
HGDP01066	ERS474725	PCR free	Sardinian	Europe	M	35.29
HGDP01067	ERS474726	PCR free	Sardinian	Europe	M	39.24
HGDP01068	ERS474727	PCR free	Sardinian	Europe	F	42.12
HGDP01069	ERS474728	PCR free	Sardinian	Europe	M	37
HGDP01070	ERS474729	PCR free	Sardinian	Europe	F	42.44
HGDP01071	ERS474730	PCR free	Sardinian	Europe	M	49.13
HGDP01072	ERS474731	PCR free	Sardinian	Europe	F	50.13
HGDP01073	ERS474732	PCR free	Sardinian	Europe	M	41.68
HGDP01074	ERS474733	PCR free	Sardinian	Europe	F	36.39
HGDP01075	ERS474734	PCR free	Sardinian	Europe	M	40.4
HGDP01076	ERS1358128	PCR	Sardinian	Europe	M	39.37
HGDP01077	ERS474736	PCR free	Sardinian	Europe	M	37.24
HGDP01078	ERS1042170	PCR free	Sardinian	Europe	F	34.14
HGDP01079	ERS1042169	PCR free	Sardinian	Europe	M	36.52
HGDP01081	ERS474077	PCR free	Mbuti	Africa	M	33.15
HGDP01086	ERS474062	PCR	Biaka	Africa	M	37.94
HGDP01090	ERS474063	PCR	Biaka	Africa	M	33.03
HGDP01094	ERS474064	PCR	Biaka	Africa	M	36.76
HGDP01095	ERS1255065	PCR free	Tujia	East Asia	M	50.04
HGDP01096	ERS474249	PCR	Tujia	East Asia	M	33.33
HGDP01098	ERS1042278	PCR free	Tujia	East Asia	F	38.74
HGDP01099	ERS474252	PCR	Tujia	East Asia	M	33.44
HGDP01100	ERS474253	PCR free	Tujia	East Asia	M	34.83
HGDP01101	ERS474254	PCR free	Tujia	East Asia	M	36.11
HGDP01102	ERS474255	PCR free	Tujia	East Asia	M	33.62
HGDP01103	ERS474256	PCR free	Tujia	East Asia	M	38.19
HGDP01104	ERS474257	PCR free	Tujia	East Asia	M	36.52
HGDP01149	ERS474691	PCR free	BergamoItalian	Europe	M	38.97
HGDP01151	ERS474692	PCR free	BergamoItalian	Europe	M	37.39
HGDP01152	ERS474693	PCR free	BergamoItalian	Europe	M	31.23

HGDP01153	ERS1042183	PCR free	BergamoItalian	Europe	M	68.66
HGDP01155	ERS474695	PCR free	BergamoItalian	Europe	M	36.62
HGDP01156	ERS474696	PCR free	BergamoItalian	Europe	F	35.49
HGDP01157	ERS474697	PCR free	BergamoItalian	Europe	F	39.29
HGDP01161	ERS474703	PCR free	Tuscan	Europe	M	31.44
HGDP01162	ERS474704	PCR free	Tuscan	Europe	M	59.63
HGDP01163	ERS1042280	PCR free	Tuscan	Europe	M	41.61
HGDP01164	ERS474706	PCR free	Tuscan	Europe	M	34.45
HGDP01166	ERS474707	PCR free	Tuscan	Europe	M	39.65
HGDP01167	ERS474708	PCR free	Tuscan	Europe	M	38.7
HGDP01168	ERS1042279	PCR free	Tuscan	Europe	F	39.07
HGDP01169	ERS474710	PCR free	Tuscan	Europe	F	39.04
HGDP01171	ERS474698	PCR free	BergamoItalian	Europe	F	34.75
HGDP01172	ERS1042138	PCR free	BergamoItalian	Europe	F	79.65
HGDP01173	ERS474700	PCR free	BergamoItalian	Europe	M	34.76
HGDP01174	ERS474701	PCR free	BergamoItalian	Europe	M	38.6
HGDP01177	ERS474702	PCR free	BergamoItalian	Europe	F	37.92
HGDP01179	ERS1255048	PCR free	Yi	East Asia	M	42.15
HGDP01180	ERS474229	PCR free	Yi	East Asia	M	32.11
HGDP01181	ERS474230	PCR free	Yi	East Asia	M	31.58
HGDP01182	ERS474231	PCR free	Yi	East Asia	M	34.9
HGDP01183	ERS474232	PCR free	Yi	East Asia	M	33.27
HGDP01184	ERS474233	PCR free	Yi	East Asia	M	27.8
HGDP01185	ERS474234	PCR free	Yi	East Asia	M	33.67
HGDP01186	ERS474235	PCR free	Yi	East Asia	M	30.71
HGDP01187	ERS474236	PCR free	Yi	East Asia	M	32.39
HGDP01188	ERS1255057	PCR free	Yi	East Asia	F	40.01
HGDP01189	ERS474238	PCR	Miao	East Asia	M	32.13
HGDP01190	ERS474239	PCR	Miao	East Asia	M	34.91
HGDP01191	ERS1042124	PCR free	Miao	East Asia	M	35.03
HGDP01192	ERS474241	PCR	Miao	East Asia	M	32.6
HGDP01193	ERS474242	PCR free	Miao	East Asia	M	33.94
HGDP01194	ERS474243	PCR free	Miao	East Asia	M	34.23
HGDP01195	ERS474244	PCR free	Miao	East Asia	M	32.18
HGDP01196	ERS474245	PCR free	Miao	East Asia	F	33.5
HGDP01197	ERS474246	PCR free	Miao	East Asia	F	31.78
HGDP01198	ERS1042162	PCR free	Miao	East Asia	F	36.79
HGDP01199	ERS1042266	PCR free	Mandenka	Africa	M	33.49
HGDP01200	ERS474140	PCR	Mandenka	Africa	M	32.01
HGDP01201	ERS474141	PCR	Mandenka	Africa	F	29.71
HGDP01202	ERS474142	PCR	Mandenka	Africa	M	36.06
HGDP01203	ERS1042272	PCR free	Oroqen	East Asia	M	38.43

HGDP01204	ERS474307	PCR free	Oroqen	East Asia	M	36.43
HGDP01205	ERS474308	PCR free	Oroqen	East Asia	M	31.85
HGDP01206	ERS474309	PCR free	Oroqen	East Asia	M	35.73
HGDP01207	ERS474310	PCR free	Oroqen	East Asia	M	30.28
HGDP01208	ERS474311	PCR free	Oroqen	East Asia	M	30.95
HGDP01209	ERS474312	PCR free	Oroqen	East Asia	F	55.41
HGDP01211	ERS1042273	PCR free	Oroqen	East Asia	F	39.81
HGDP01212	ERS474314	PCR free	Oroqen	East Asia	F	29.13
HGDP01213	ERS474296	PCR	Daur	East Asia	M	33.55
HGDP01214	ERS474297	PCR	Daur	East Asia	M	35.02
HGDP01215	ERS1042139	PCR free	Daur	East Asia	F	45.4
HGDP01216	ERS474299	PCR	Daur	East Asia	M	35.24
HGDP01217	ERS474300	PCR free	Daur	East Asia	M	38.62
HGDP01218	ERS474301	PCR free	Daur	East Asia	M	30.57
HGDP01220	ERS474303	PCR free	Daur	East Asia	M	31.15
HGDP01221	ERS474304	PCR free	Daur	East Asia	M	31.54
HGDP01222	ERS474305	PCR free	Daur	East Asia	F	30.33
HGDP01223	ERS1042271	PCR free	Mongolian	East Asia	F	44.33
HGDP01224	ERS474287	PCR	Mongolian	East Asia	M	29.74
HGDP01225	ERS474288	PCR	Mongolian	East Asia	M	33.37
HGDP01227	ERS474290	PCR free	Mongolian	East Asia	M	32.74
HGDP01228	ERS1042270	PCR free	Mongolian	East Asia	M	39.26
HGDP01229	ERS474292	PCR free	Mongolian	East Asia	M	32.96
HGDP01230	ERS474293	PCR free	Mongolian	East Asia	M	36.14
HGDP01231	ERS474294	PCR free	Mongolian	East Asia	F	30.93
HGDP01232	ERS474295	PCR free	Mongolian	East Asia	F	34.41
HGDP01233	ERS474277	PCR free	Hezhen	East Asia	M	32.33
HGDP01234	ERS474278	PCR free	Hezhen	East Asia	F	37.68
HGDP01236	ERS474279	PCR free	Hezhen	East Asia	M	32.52
HGDP01237	ERS474280	PCR free	Hezhen	East Asia	M	38.05
HGDP01238	ERS474281	PCR free	Hezhen	East Asia	F	31.29
HGDP01239	ERS474282	PCR free	Hezhen	East Asia	M	30.99
HGDP01240	ERS1042253	PCR free	Hezhen	East Asia	M	44.23
HGDP01241	ERS474284	PCR free	Hezhen	East Asia	M	32.61
HGDP01242	ERS1042254	PCR free	Hezhen	East Asia	F	39.89
HGDP01243	ERS474268	PCR	Xibo	East Asia	M	33.72
HGDP01244	ERS474269	PCR	Xibo	East Asia	M	28.33
HGDP01245	ERS474270	PCR	Xibo	East Asia	M	29.68
HGDP01246	ERS1255081	PCR free	Xibo	East Asia	M	51.99
HGDP01247	ERS474272	PCR free	Xibo	East Asia	M	33.75
HGDP01248	ERS474273	PCR free	Xibo	East Asia	M	38.4
HGDP01249	ERS474274	PCR free	Xibo	East Asia	M	32.86

HGDP01250	ERS1042282	PCR free	Xibo	East Asia	M	36.2
HGDP01251	ERS474276	PCR free	Xibo	East Asia	F	32.16
HGDP01253	ERS1042227	PCR free	Mozabite	Middle East	M	37.91
HGDP01254	ERS474873	PCR free	Mozabite	Middle East	F	36.87
HGDP01255	ERS474874	PCR	Mozabite	Middle East	M	35.92
HGDP01256	ERS474875	PCR	Mozabite	Middle East	M	28.87
HGDP01257	ERS474876	PCR	Mozabite	Middle East	M	29.99
HGDP01258	ERS1063420	PCR free	Mozabite	Middle East	M	34.38
HGDP01259	ERS474878	PCR free	Mozabite	Middle East	M	38.39
HGDP01260	ERS474879	PCR free	Mozabite	Middle East	M	31.9
HGDP01261	ERS474880	PCR free	Mozabite	Middle East	M	33.8
HGDP01262	ERS1063421	PCR free	Mozabite	Middle East	M	36.58
HGDP01263	ERS1063422	PCR free	Mozabite	Middle East	M	40.95
HGDP01264	ERS474883	PCR free	Mozabite	Middle East	M	33.63
HGDP01265	ERS474884	PCR free	Mozabite	Middle East	M	33.08
HGDP01266	ERS474885	PCR free	Mozabite	Middle East	M	35.94
HGDP01267	ERS474886	PCR free	Mozabite	Middle East	F	34.73
HGDP01268	ERS474887	PCR free	Mozabite	Middle East	M	32.15
HGDP01269	ERS474888	PCR free	Mozabite	Middle East	M	33.39
HGDP01270	ERS474889	PCR free	Mozabite	Middle East	F	32.32
HGDP01271	ERS474890	PCR free	Mozabite	Middle East	M	44.38
HGDP01272	ERS474891	PCR free	Mozabite	Middle East	M	34.46
HGDP01274	ERS1042228	PCR free	Mozabite	Middle East	F	39.48
HGDP01275	ERS474894	PCR free	Mozabite	Middle East	F	41.83
HGDP01276	ERS474895	PCR free	Mozabite	Middle East	F	42.38
HGDP01277	ERS474896	PCR free	Mozabite	Middle East	F	40.53
HGDP01279	ERS474898	PCR free	Mozabite	Middle East	M	38.06
HGDP01280	ERS1063425	PCR free	Mozabite	Middle East	F	41.29
HGDP01282	ERS474900	PCR free	Mozabite	Middle East	M	41.14
HGDP01283	ERS474143	PCR	Mandenka	Africa	M	35.99
HGDP01284	NA	PCR	Mandenka	Africa	M	26.05
HGDP01285	ERS474145	PCR	Mandenka	Africa	M	36.44
HGDP01286	ERS1358124	PCR	Mandenka	Africa	M	38.29
HGDP01287	ERS474181	PCR free	NorthernHan	East Asia	F	33.24
HGDP01288	ERS474182	PCR free	NorthernHan	East Asia	M	35.59
HGDP01289	ERS474183	PCR free	NorthernHan	East Asia	M	33.05
HGDP01290	ERS474184	PCR free	NorthernHan	East Asia	M	30.89
HGDP01291	ERS474185	PCR free	NorthernHan	East Asia	F	37.96
HGDP01292	ERS474186	PCR free	NorthernHan	East Asia	M	32.13
HGDP01293	ERS474187	PCR free	NorthernHan	East Asia	M	32.43
HGDP01294	ERS474188	PCR free	NorthernHan	East Asia	M	34.89
HGDP01295	ERS474189	PCR free	NorthernHan	East Asia	M	34.52

HGDP01296	ERS474190	PCR free	NorthernHan	East Asia	M	31.96
HGDP01297	ERS1255072	PCR free	Uygur	Central/South Asia	M	50.94
HGDP01298	ERS474380	PCR	Uygur	Central/South Asia	M	33.72
HGDP01299	ERS474381	PCR	Uygur	Central/South Asia	M	34.68
HGDP01300	ERS474382	PCR	Uygur	Central/South Asia	M	35.67
HGDP01301	ERS1063426	PCR free	Uygur	Central/South Asia	M	31.49
HGDP01302	ERS1063427	PCR free	Uygur	Central/South Asia	M	37.1
HGDP01303	ERS474385	PCR free	Uygur	Central/South Asia	M	28.83
HGDP01304	ERS474386	PCR free	Uygur	Central/South Asia	M	31.93
HGDP01305	ERS474387	PCR free	Uygur	Central/South Asia	F	32.03
HGDP01306	ERS1042281	PCR free	Uygur	Central/South Asia	F	38.28
HGDP01307	NA	PCR	Dai	East Asia	M	30.43
HGDP01308	ERS1358148	PCR	Dai	East Asia	M	38.86
HGDP01309	ERS813229	PCR free	Dai	East Asia	M	33.45
HGDP01310	ERS474194	PCR free	Dai	East Asia	M	34.32
HGDP01311	ERS474195	PCR free	Dai	East Asia	M	30.67
HGDP01312	ERS1255071	PCR free	Dai	East Asia	M	40.38
HGDP01313	ERS474197	PCR free	Dai	East Asia	M	32.25
HGDP01314	ERS1358104	PCR free	Dai	East Asia	F	48.91
HGDP01315	ERS1042174	PCR free	Dai	East Asia	F	38.12
HGDP01317	ERS474201	PCR free	Lahu	East Asia	M	31.39
HGDP01318	ERS474202	PCR free	Lahu	East Asia	M	33.05
HGDP01319	ERS474203	PCR free	Lahu	East Asia	M	35.86
HGDP01320	ERS1255097	PCR free	Lahu	East Asia	M	58.02
HGDP01321	ERS474205	PCR free	Lahu	East Asia	M	34.78
HGDP01322	ERS474206	PCR free	Lahu	East Asia	M	36.28
HGDP01323	ERS1042159	PCR free	Lahu	East Asia	F	38.09
HGDP01326	ERS474208	PCR free	Lahu	East Asia	M	34.31
HGDP01327	ERS474218	PCR free	She	East Asia	M	33.86
HGDP01328	ERS474219	PCR free	She	East Asia	M	36
HGDP01329	ERS474220	PCR free	She	East Asia	M	49.63
HGDP01330	ERS474221	PCR free	She	East Asia	M	30.1
HGDP01331	ERS474222	PCR free	She	East Asia	M	31.7
HGDP01332	ERS474223	PCR free	She	East Asia	M	32.36
HGDP01333	ERS1255106	PCR free	She	East Asia	M	48.08
HGDP01334	ERS474225	PCR free	She	East Asia	F	35.87
HGDP01335	ERS1255114	PCR free	She	East Asia	F	48.15
HGDP01336	ERS474227	PCR free	She	East Asia	F	35.91
HGDP01337	ERS474209	PCR free	Naxi	East Asia	M	31.99
HGDP01338	ERS1255103	PCR free	Naxi	East Asia	M	41.23
HGDP01339	ERS813230	PCR free	Naxi	East Asia	M	32.31
HGDP01340	ERS474212	PCR free	Naxi	East Asia	M	31.12

HGDP01341	ERS474213	PCR free	Naxi	East Asia	M	32.33
HGDP01342	ERS474214	PCR free	Naxi	East Asia	M	34.2
HGDP01345	ERS1042180	PCR free	Naxi	East Asia	F	42.13
HGDP01346	ERS474217	PCR free	Naxi	East Asia	F	35.28
HGDP01347	ERS474258	PCR free	Tu	East Asia	M	30.5
HGDP01348	ERS474259	PCR free	Tu	East Asia	M	34.24
HGDP01349	ERS474260	PCR free	Tu	East Asia	M	35.04
HGDP01350	ERS1255122	PCR free	Tu	East Asia	M	51.62
HGDP01351	ERS474262	PCR free	Tu	East Asia	M	32.74
HGDP01352	ERS474263	PCR free	Tu	East Asia	M	34.61
HGDP01353	ERS474264	PCR free	Tu	East Asia	M	37.46
HGDP01354	ERS474265	PCR free	Tu	East Asia	F	36.33
HGDP01355	ERS1042173	PCR free	Tu	East Asia	F	38.06
HGDP01356	ERS474267	PCR free	Tu	East Asia	F	32.86
HGDP01357	ERS474651	PCR free	Basque	Europe	M	34.85
HGDP01358	ERS474652	PCR free	Basque	Europe	M	36.59
HGDP01359	ERS474653	PCR free	Basque	Europe	M	38.71
HGDP01360	ERS474654	PCR free	Basque	Europe	M	38.33
HGDP01361	ERS474655	PCR free	Basque	Europe	M	35.82
HGDP01362	ERS474656	PCR free	Basque	Europe	M	33.32
HGDP01363	ERS474657	PCR free	Basque	Europe	F	34.96
HGDP01364	ERS1042147	PCR free	Basque	Europe	M	41.71
HGDP01365	ERS1042148	PCR free	Basque	Europe	F	37.71
HGDP01366	ERS474660	PCR free	Basque	Europe	F	37.84
HGDP01367	ERS474661	PCR free	Basque	Europe	F	30.11
HGDP01368	ERS474662	PCR free	Basque	Europe	F	38.42
HGDP01369	ERS474663	PCR free	Basque	Europe	F	36.36
HGDP01370	ERS474664	PCR free	Basque	Europe	M	36.29
HGDP01372	ERS474666	PCR free	Basque	Europe	M	37.04
HGDP01373	ERS474667	PCR free	Basque	Europe	F	28.88
HGDP01374	ERS474668	PCR free	Basque	Europe	M	30.17
HGDP01375	ERS474669	PCR free	Basque	Europe	M	37.96
HGDP01376	ERS474670	PCR free	Basque	Europe	M	37.66
HGDP01377	ERS474671	PCR free	Basque	Europe	M	37.21
HGDP01378	ERS474672	PCR free	Basque	Europe	M	34
HGDP01379	ERS474673	PCR free	Basque	Europe	M	34.84
HGDP01380	ERS474674	PCR free	Basque	Europe	F	36.53
HGDP01382	ERS474582	PCR free	Adygei	Europe	F	34.32
HGDP01383	ERS474583	PCR	Adygei	Europe	M	33.08
HGDP01384	ERS474584	PCR free	Adygei	Europe	F	31.49
HGDP01385	ERS474585	PCR	Adygei	Europe	M	34.44
HGDP01386	ERS474586	PCR free	Adygei	Europe	F	34.66

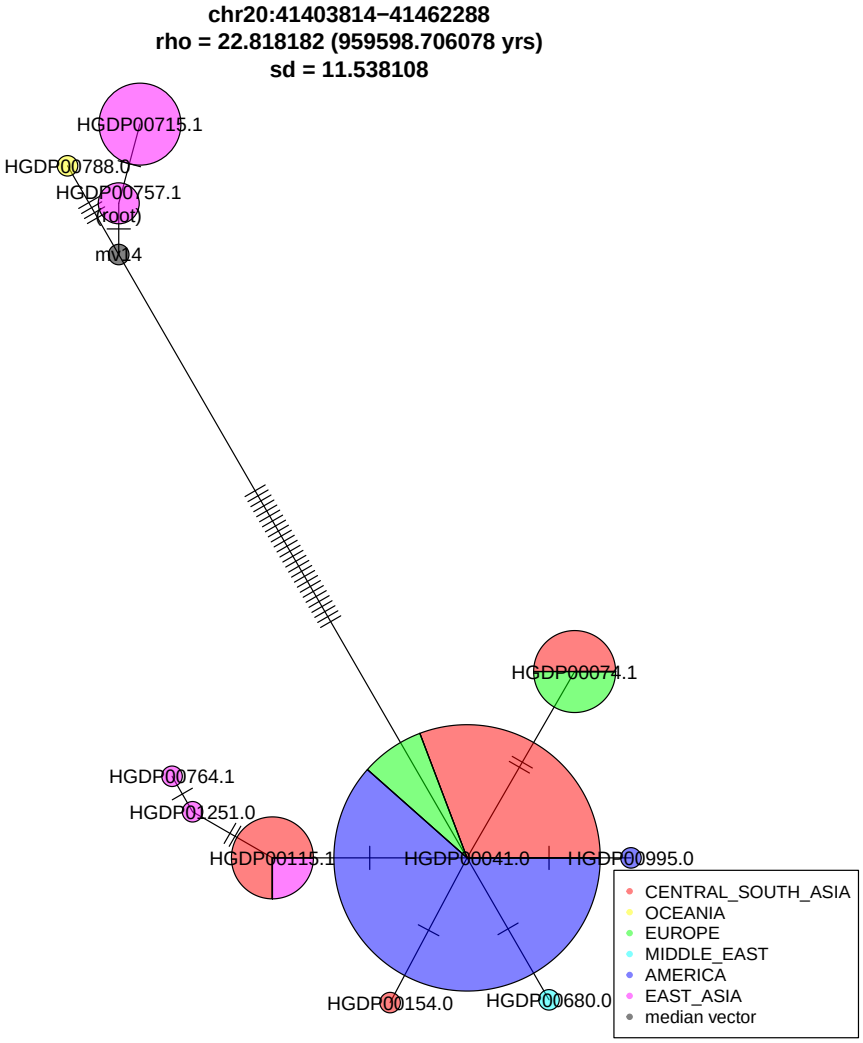
HGDP01387	ERS474587	PCR free	Adygei	Europe	F	35.57
HGDP01388	ERS474588	PCR free	Adygei	Europe	F	28.06
HGDP01396	ERS474589	PCR	Adygei	Europe	M	31.58
HGDP01397	ERS474590	PCR free	Adygei	Europe	M	37.49
HGDP01398	ERS474591	PCR free	Adygei	Europe	F	35.73
HGDP01399	ERS474592	PCR free	Adygei	Europe	F	35
HGDP01400	ERS474593	PCR free	Adygei	Europe	F	34.03
HGDP01401	ERS1042144	PCR free	Adygei	Europe	F	34.83
HGDP01402	ERS1042143	PCR free	Adygei	Europe	M	41.6
HGDP01403	ERS474596	PCR free	Adygei	Europe	M	34.86
HGDP01404	ERS474597	PCR free	Adygei	Europe	M	32.5
HGDP01405	ERS474092	PCR	BantuKenya	Africa	M	34.59
HGDP01406	ERS474093	PCR	BantuKenya	Africa	M	33.21
HGDP01408	ERS474094	PCR	BantuKenya	Africa	M	33.78
HGDP01411	ERS474095	PCR	BantuKenya	Africa	M	34.82
HGDP01412	ERS474096	PCR	BantuKenya	Africa	M	32.85
HGDP01414	ERS1042146	PCR free	BantuKenya	Africa	F	37.01
HGDP01415	ERS474098	PCR	BantuKenya	Africa	M	32.72
HGDP01416	ERS474099	PCR	BantuKenya	Africa	M	33.58
HGDP01417	ERS1255064	PCR free	BantuKenya	Africa	M	48.35
HGDP01418	ERS474101	PCR	BantuKenya	Africa	M	33.8
HGDP01419	ERS474102	PCR	BantuKenya	Africa	M	32.16

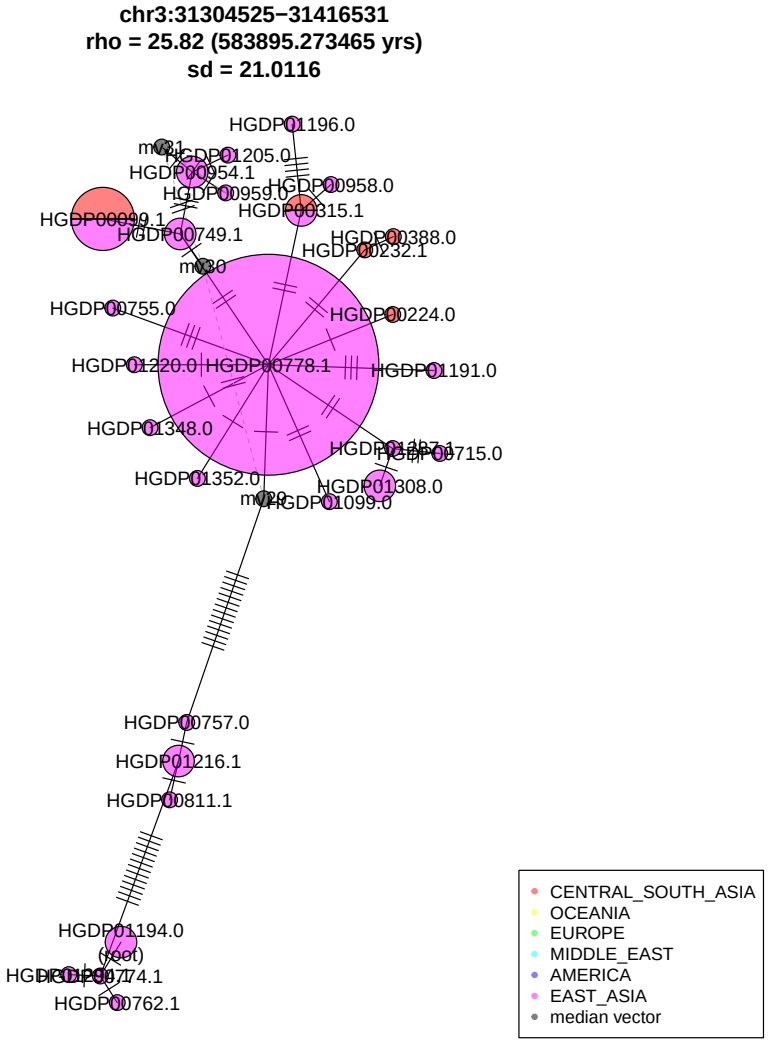
Appendix B

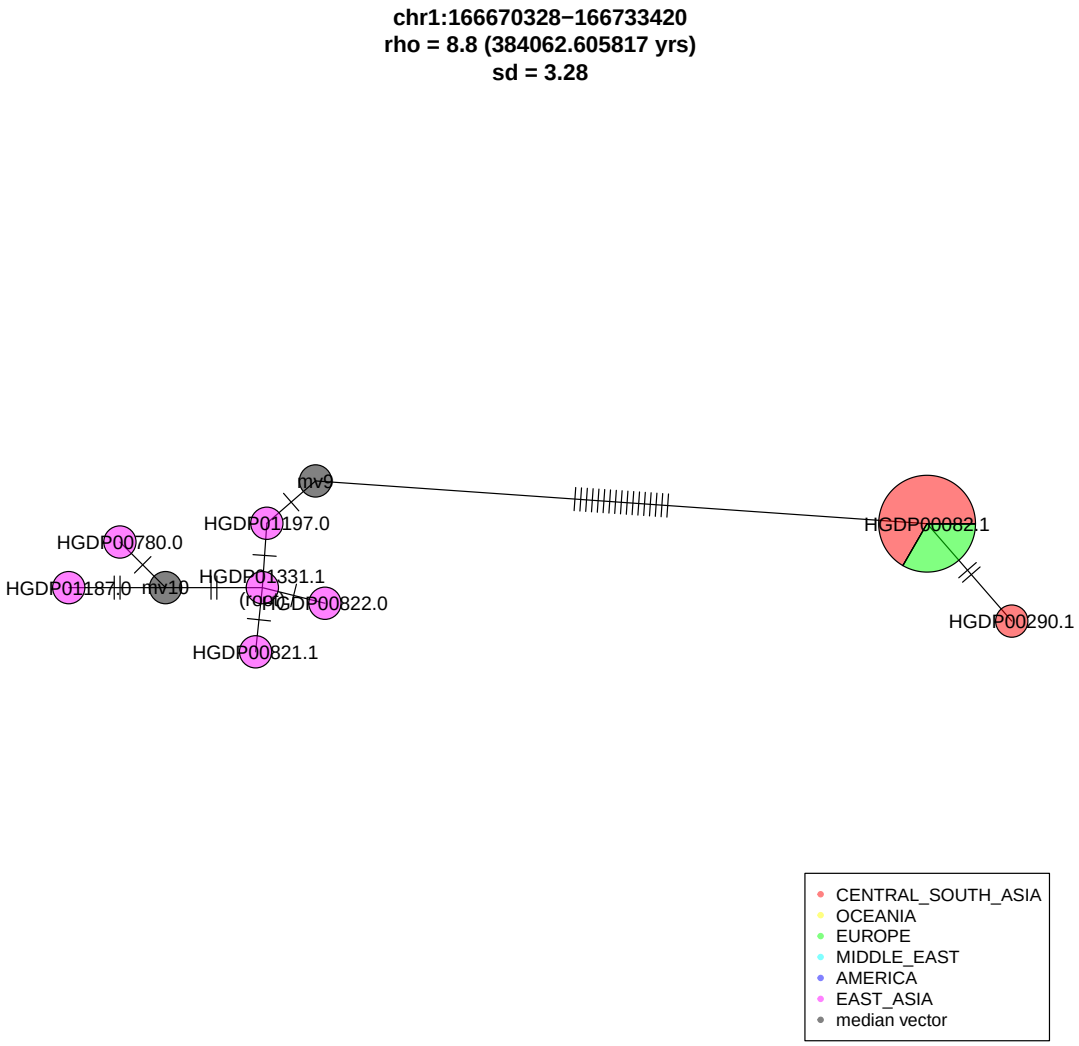
Examples of archaic haplotype networks

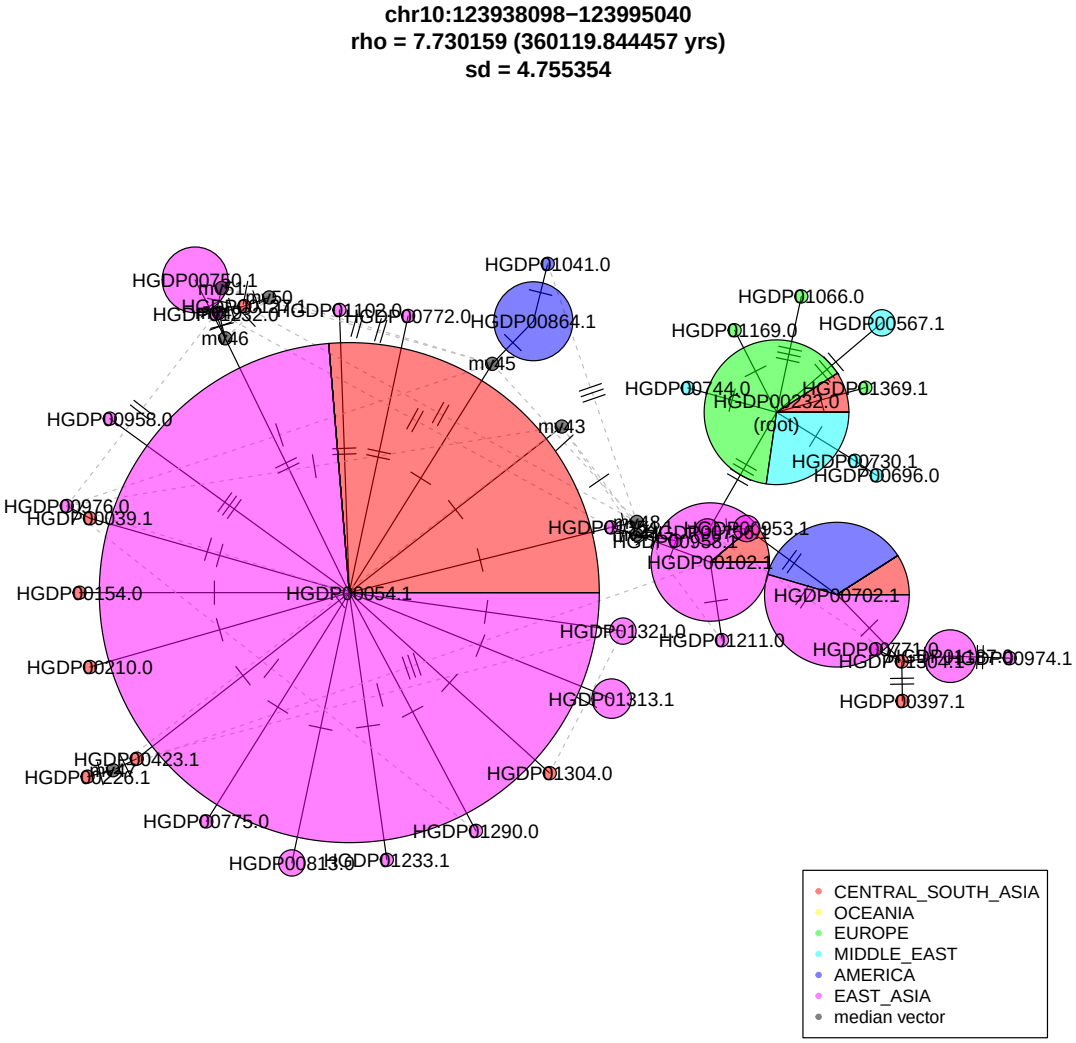
This section contains examples of Neanderthal and Denisova haplotype networks as described in 5.7. In all the plots, each solid circle represents a distinct haplotype, where the radius is proportional to the number of samples carrying that haplotype. The circles are coloured by the geographical origins of the samples and labelled by one of the samples. The number of bars on the edges equals the number of mutations separating the haplotypes. Small grey circles labelled "mv" represents median vectors reconstructed in the median joining algorithm. The dashed lines represent alternative links.

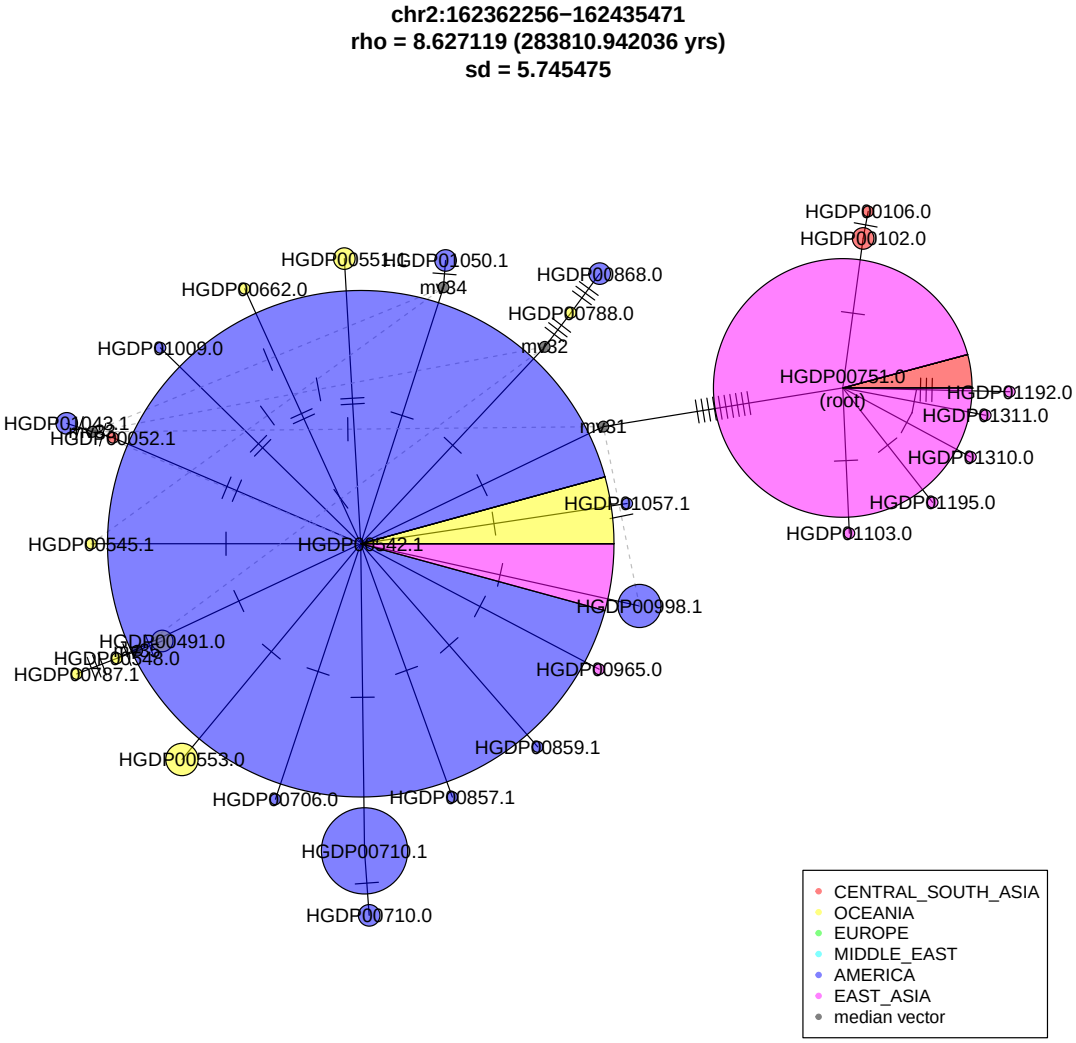
B.1 Examples of Neanderthal haplotype networks older than 100k years

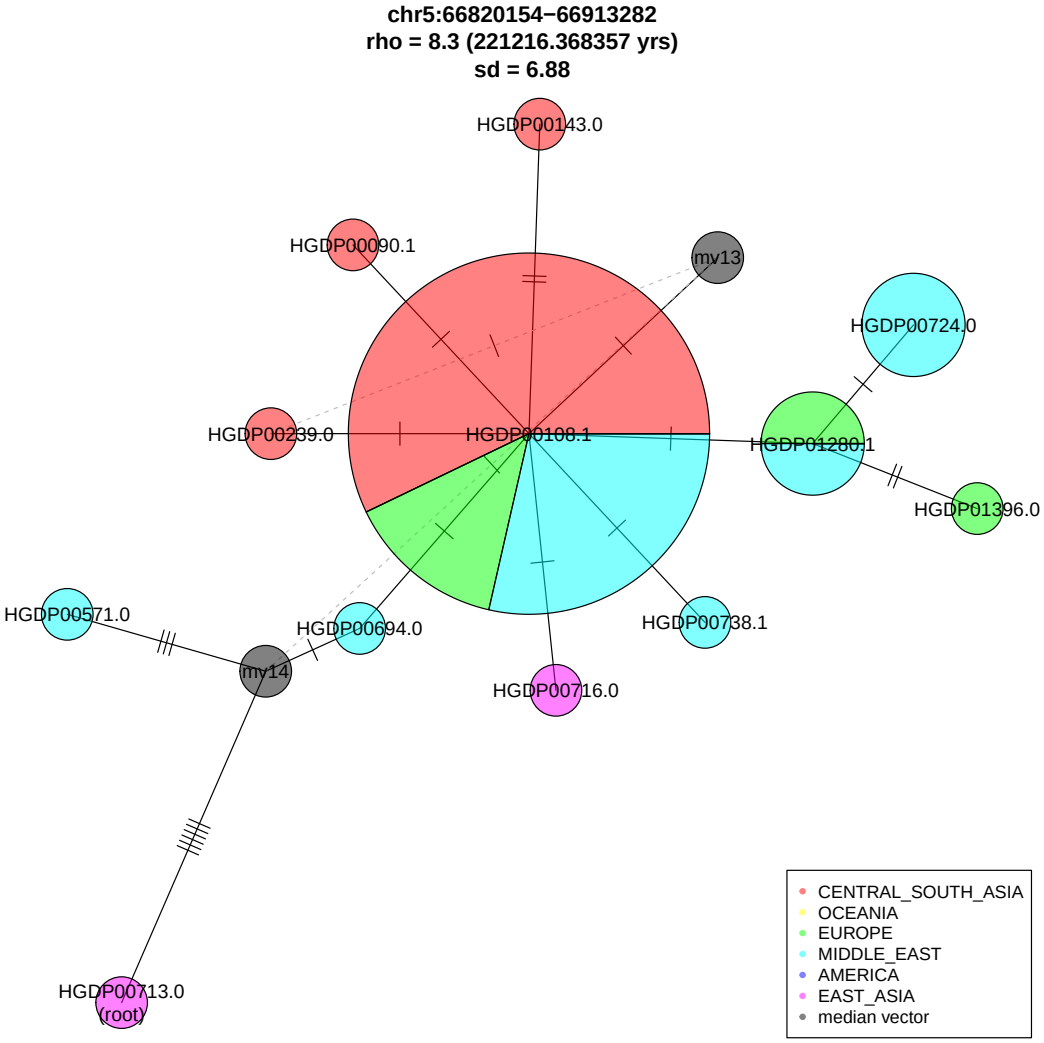


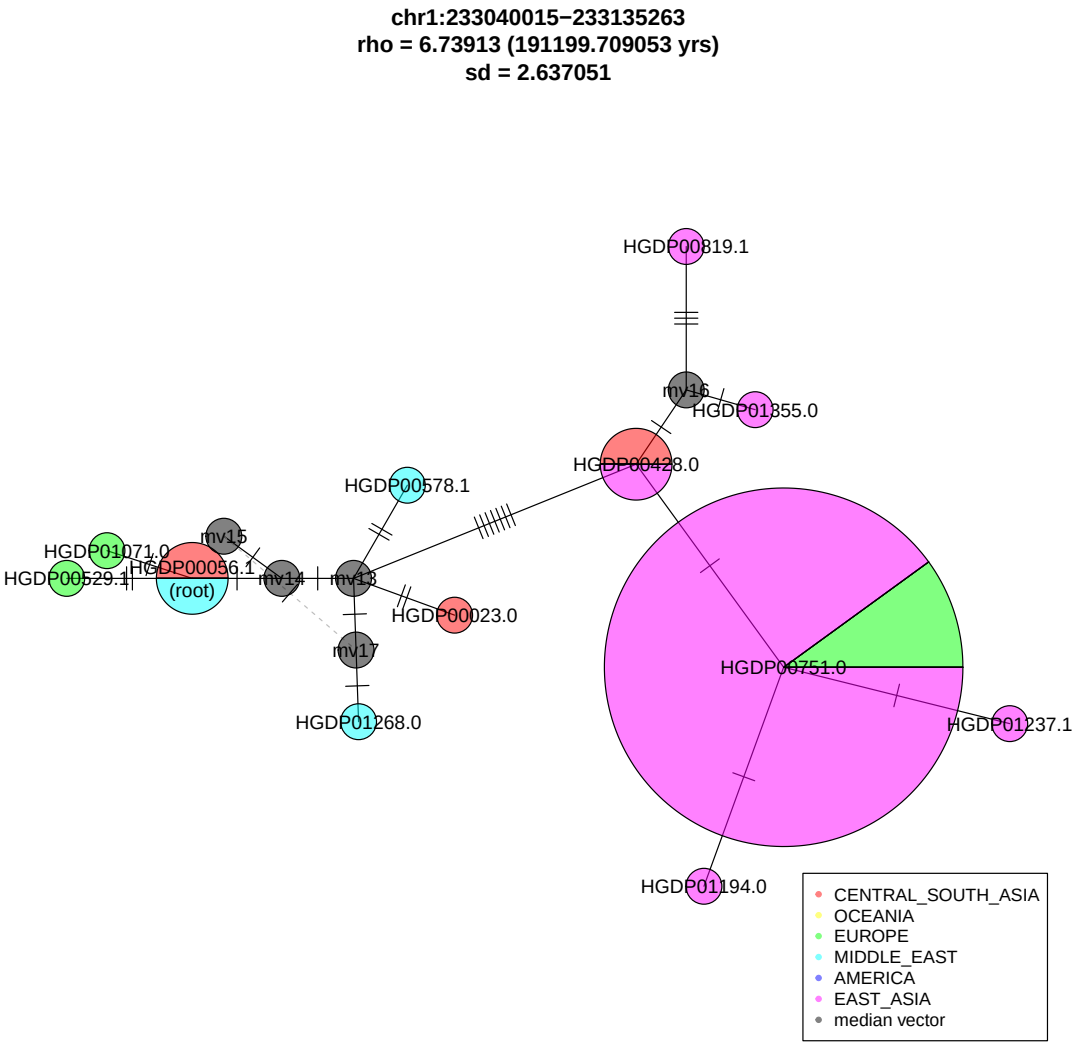

chr3:31304525−31416531
rho = 25.82 (583895.273465 yrs)
sd = 21.0116
HGDP01196.0
HGDP01205.0
HGDP00954.1
HGDP00959.0
HGDP00958.0
HGDP00815.1
HGDP00749.1
HGDP00388.0
HGDP00232.1
HGDP00755.0
HGDP00224.0
HGDP01220.0
HGDP00778.1
HGDP01191.0
HGDP01348.0
HGDP00715.0
HGDP01352.0
HGDP01308.0
HGDP01099.0
HGDP00757.0
HGDP01216.1
HGDP00811.1
HGDP01194.0
(root)
HGDP00762.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

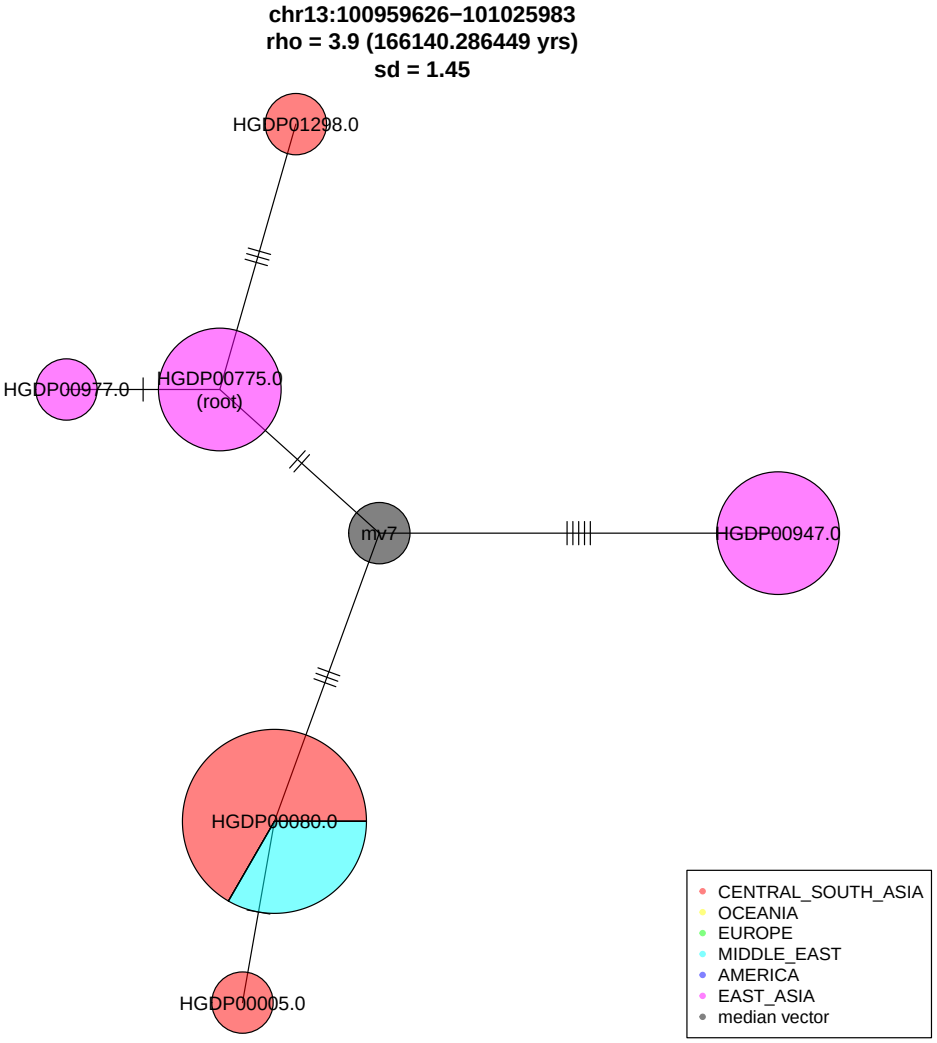

chr1:166670328−166733420
rho = 8.8 (384062.605817 yrs)
sd = 3.28
mv9
HGDP01197.0
HGDP00780.0
HGDP01187.0
mv10
HGDP01331.1
HGDP00822.0
HGDP00821.1
HGDP00082.1
HGDP00290.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

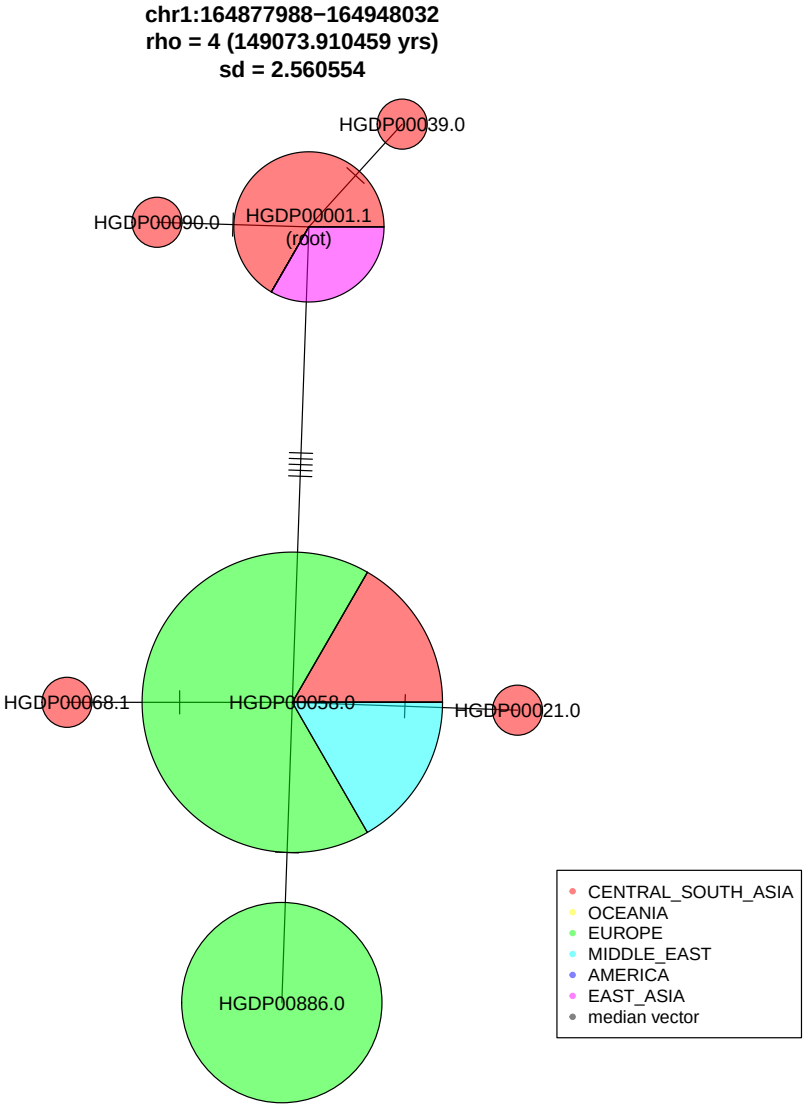

chr10:123938098−123995040
rho = 7.730159 (360119.844457 yrs)
sd = 4.755354
HGDP01041.0
HGDP00864.1
HGDP01066.0
HGDP00567.1
HGDP01169.0
HGDP00744.0
HGDP01369.1
HGDP00232.0
(root)
HGDP00730.1
HGDP00696.0
HGDP00958.0
HGDP00976.0
HGDP00039.1
HGDP00154.0
HGDP00210.0
HGDP00423.1
HGDP00775.0
HGDP01290.0
HGDP01233.1
HGDP00054.1
HGDP01321.0
HGDP01313.1
HGDP01304.0
HGDP00102.1
HGDP01211.0
HGDP00702.1
HGDP00397.1
HGDP00974.1
m043
m045
m046
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

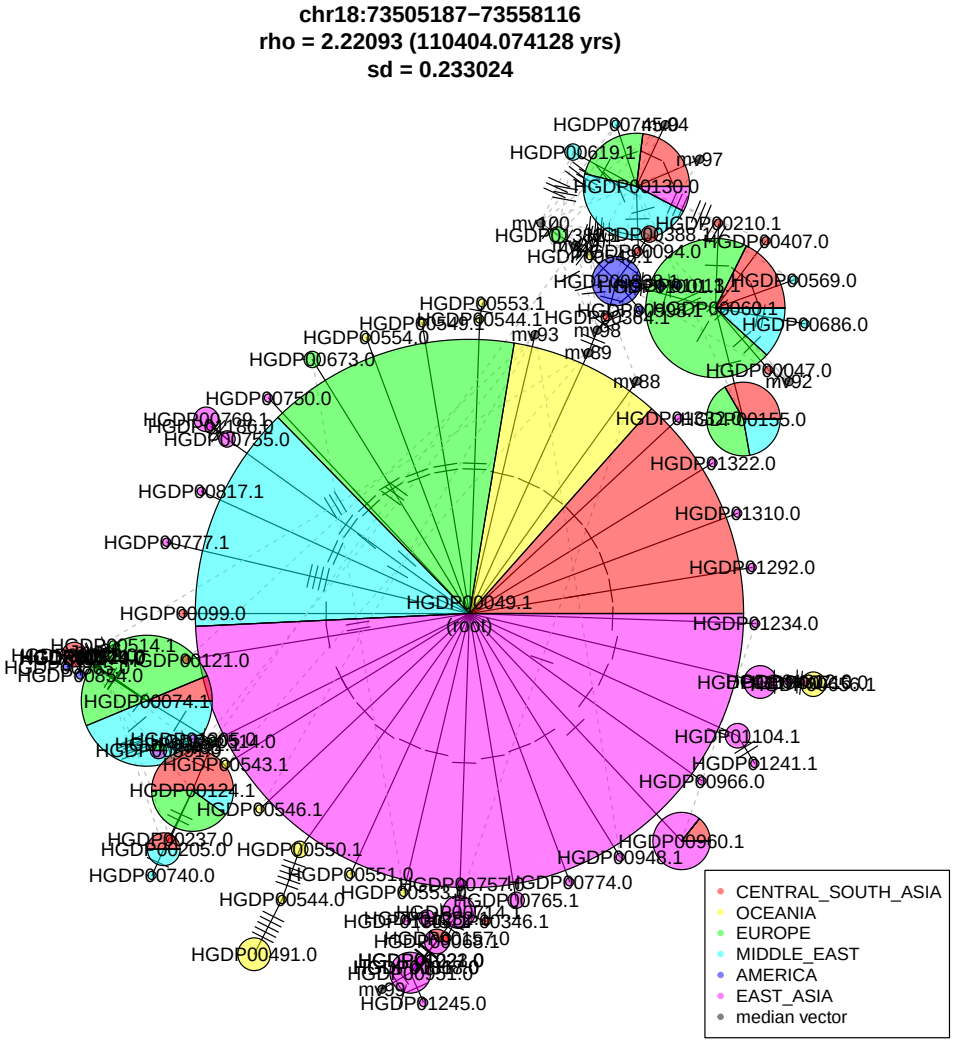

chr2:162362256−162435471
rho = 8.627119 (283810.942036 yrs)
sd = 5.745475
HGDP00106.0
HGDP00102.0
HGDP00751.0
(root)
HGDP01192.0
HGDP01311.0
HGDP01310.0
HGDP01195.0
HGDP01103.0
HGDP00551.1
HGDP01050.1
HGDP00662.0
m284
HGDP00868.0
HGDP00788.0
m282
HGDP01009.0
m281
HGDP00052.1
HGDP01057.1
HGDP00545.1
HGDP00542.1
HGDP00998.1
HGDP00491.0
HGDP00787.1
HGDP00965.0
HGDP00553.0
HGDP00859.1
HGDP00706.0
HGDP00857.1
HGDP00710.1
HGDP00710.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector


chr5:66820154−66913282
rho = 8.3 (221216.368357 yrs)
sd = 6.88
HGDP00143.0
HGDP00090.1
mv13
HGDP00724.0
HGDP00239.0
HGDP00108.1
HGDP01280.1
HGDP01396.0
HGDP00571.0
HGDP00694.0
HGDP00738.1
mv14
HGDP00716.0
HGDP00713.0
(root)
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

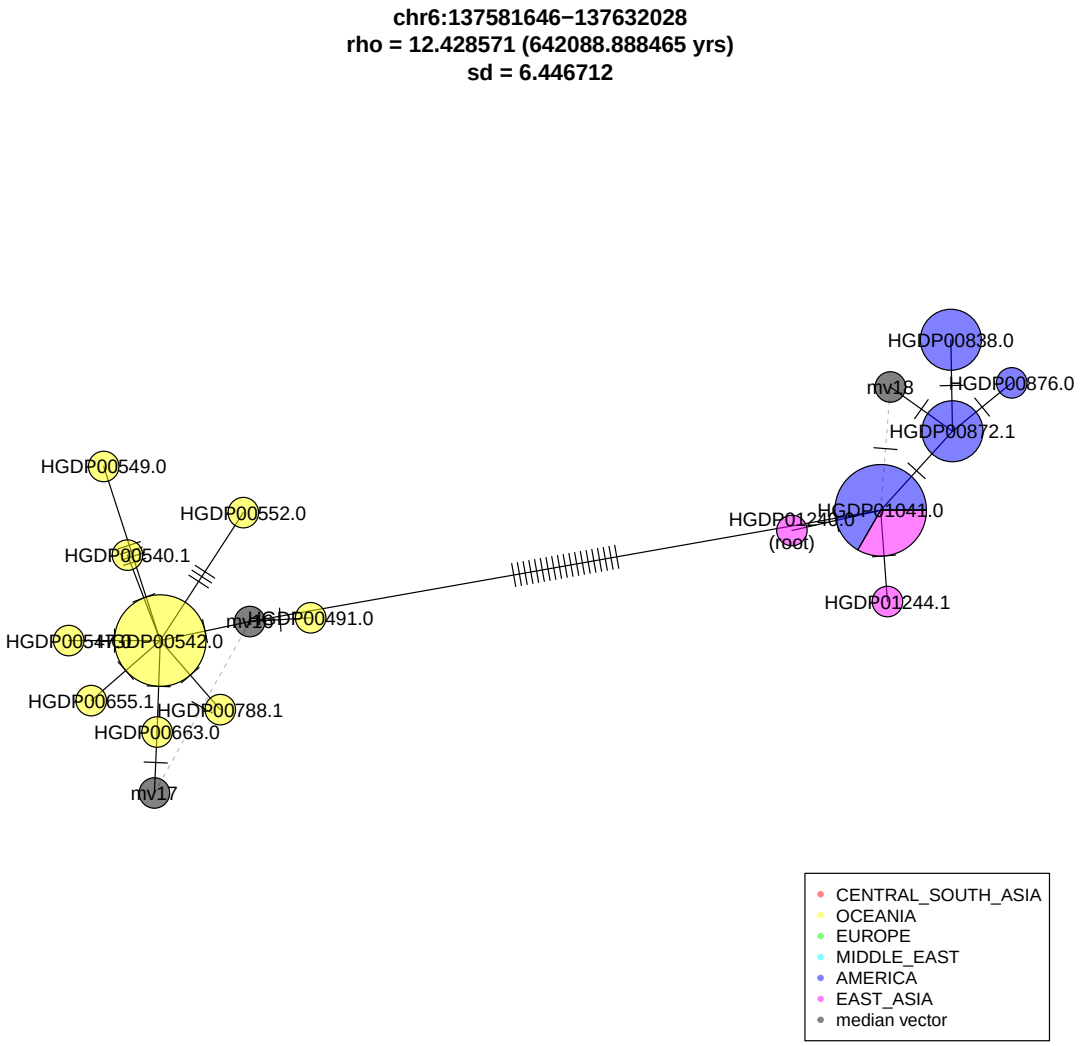

chr1:233040015−233135263
rho = 6.73913 (191199.709053 yrs)
sd = 2.637051
HGDP00819.1
mv16
HGDP01355.0
HGDP00428.0
HGDP00578.1
mv15
HGDP01071.0
HGDP00529.1
HGDP00056.1
(root)
mv14
mv13
HGDP00023.0
mv17
HGDP01268.0
HGDP00751.0
HGDP01237.1
HGDP01194.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

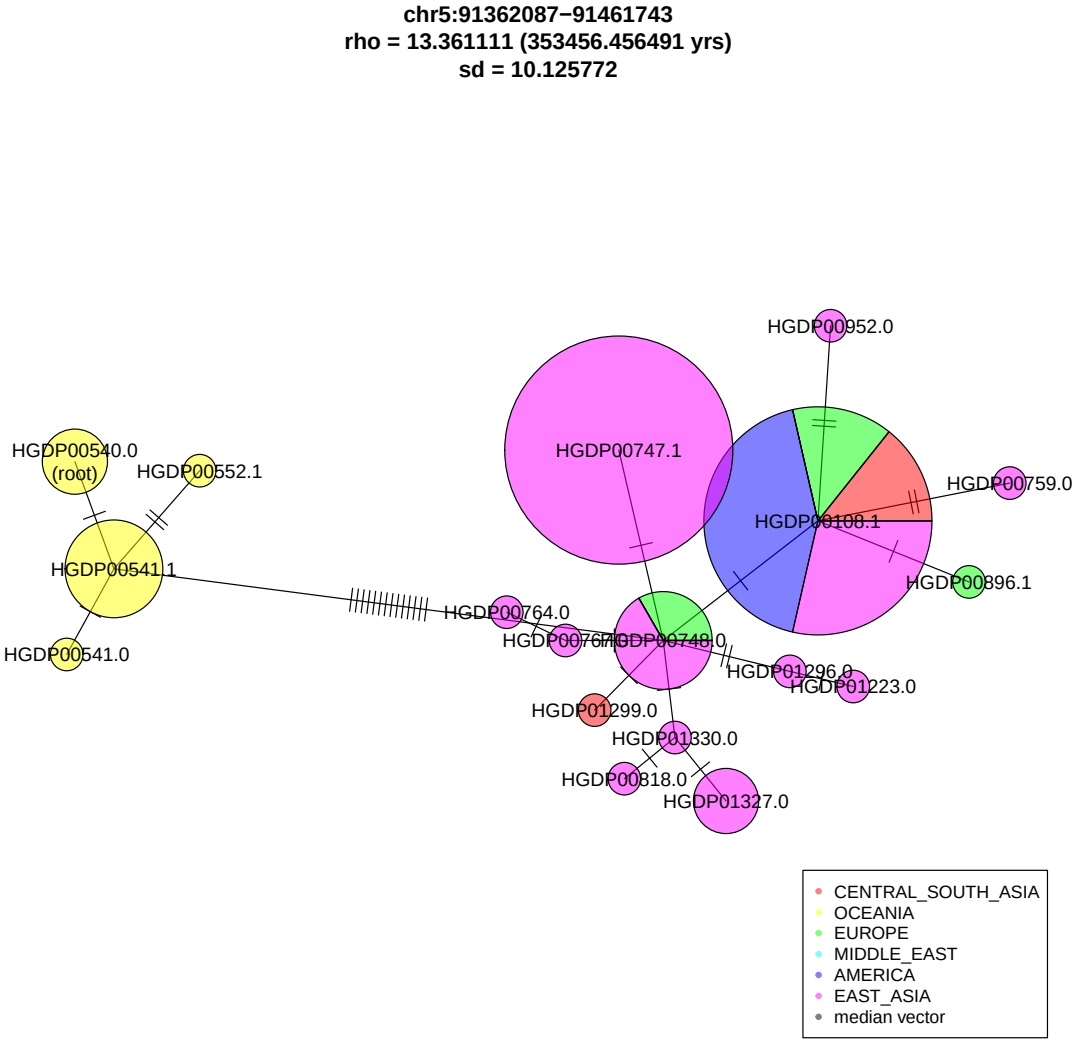

chr13:100959626−101025983
rho = 3.9 (166140.286449 yrs)
sd = 1.45
HGDP01298.0
HGDP00977.0
HGDP00775.0
(root)
mv7
HGDP00947.0
HGDP00080.0
HGDP00005.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

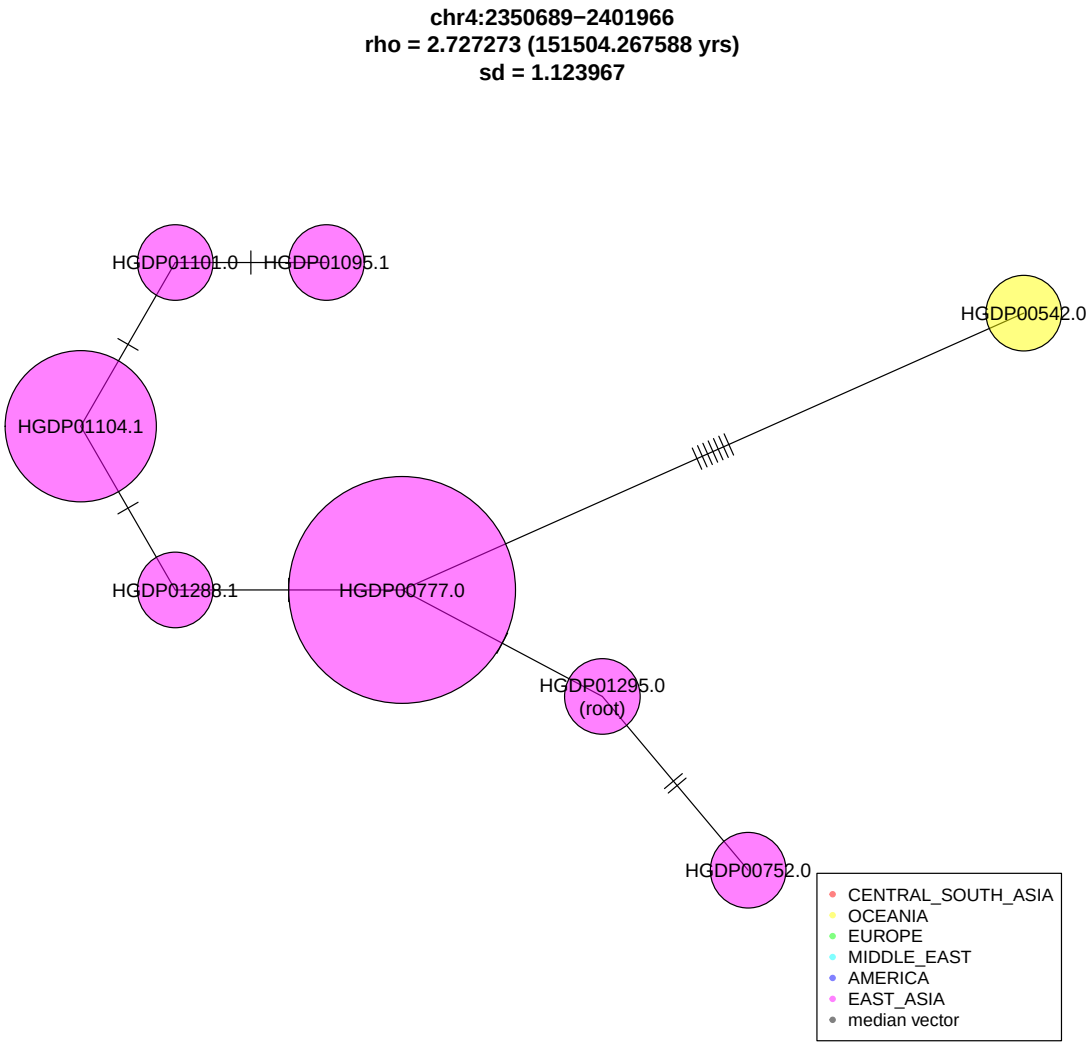

chr1:164877988−164948032
rho = 4 (149073.910459 yrs)
sd = 2.560554
HGDP00039.0
HGDP00090.0
HGDP00001.1
(root)
HGDP00068.1
HGDP00058.0
HGDP00021.0
HGDP00886.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

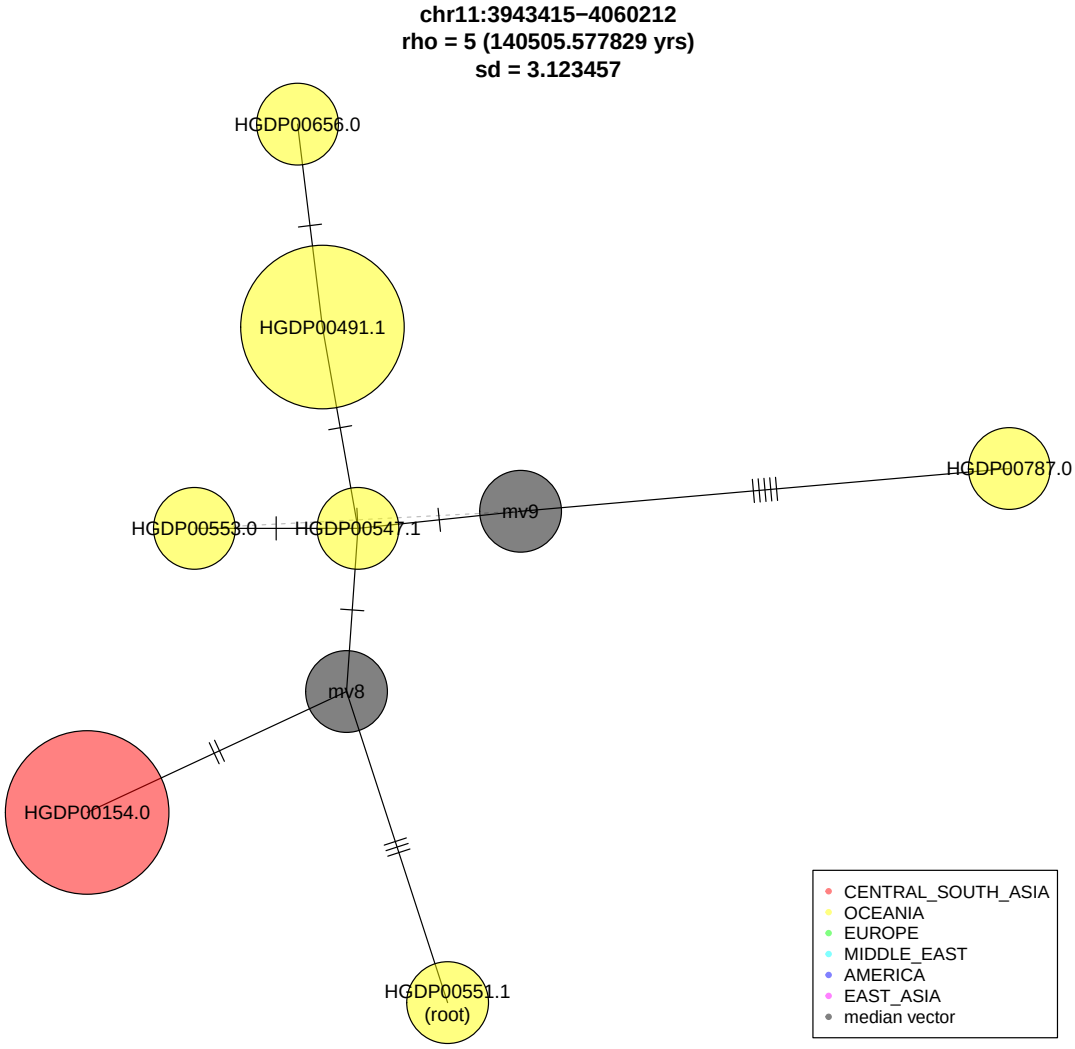

chr18:73505187−73558116
rho = 2.22093 (110404.074128 yrs)
sd = 0.233024
HGDP00619.1
HGDP00130.0
HGDP00210.1
HGDP00407.0
HGDP00569.0
HGDP00686.0
HGDP00047.0
HGDP00155.0
HGDP01322.0
HGDP01310.0
HGDP01292.0
HGDP01234.0
HGDP01104.1
HGDP01241.1
HGDP00966.0
HGDP00960.1
HGDP00948.1
HGDP00774.0
HGDP00765.1
HGDP00553.1
HGDP00544.1
HGDP00554.0
HGDP00673.0
HGDP00750.0
HGDP00735.0
HGDP00817.1
HGDP00777.1
HGDP00099.0
HGDP00049.1
(root)
HGDP00121.0
HGDP00074.1
HGDP00543.1
HGDP00124.1
HGDP00546.1
HGDP00237.0
HGDP00205.0
HGDP00550.1
HGDP00740.0
HGDP00551.0
HGDP00544.0
HGDP00491.0
HGDP01245.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

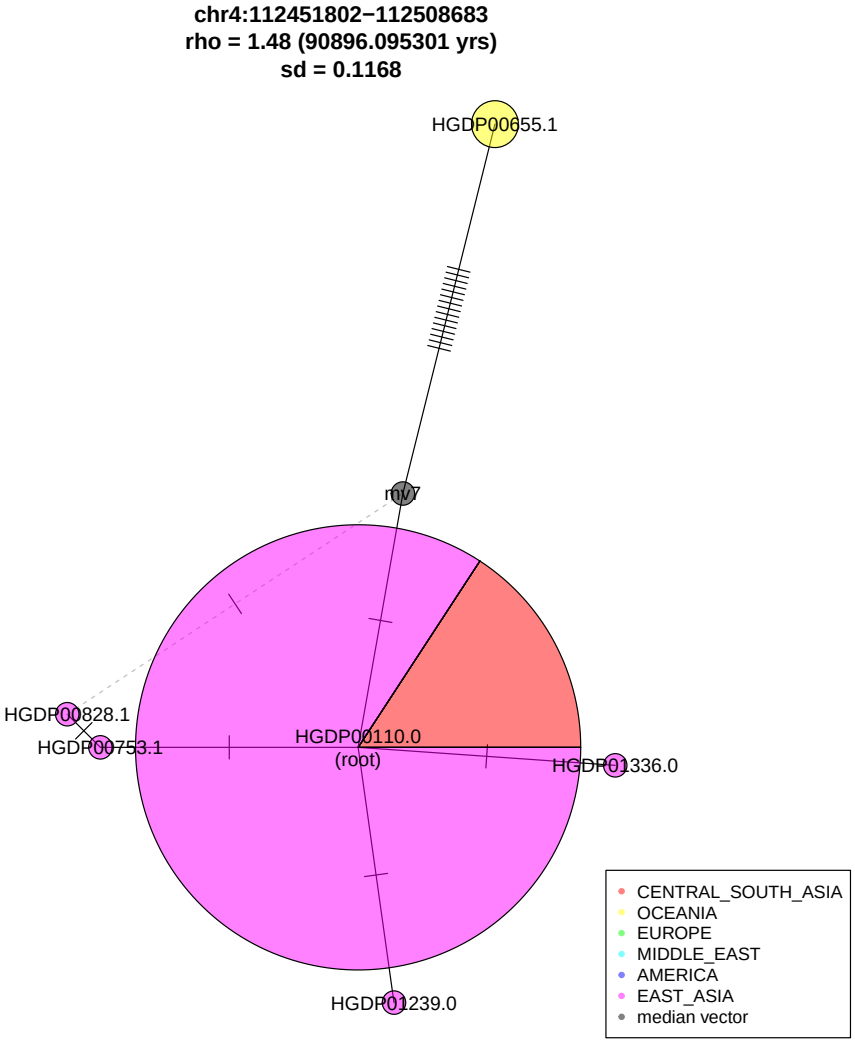
B.2 Examples of Denisova haplotype networks

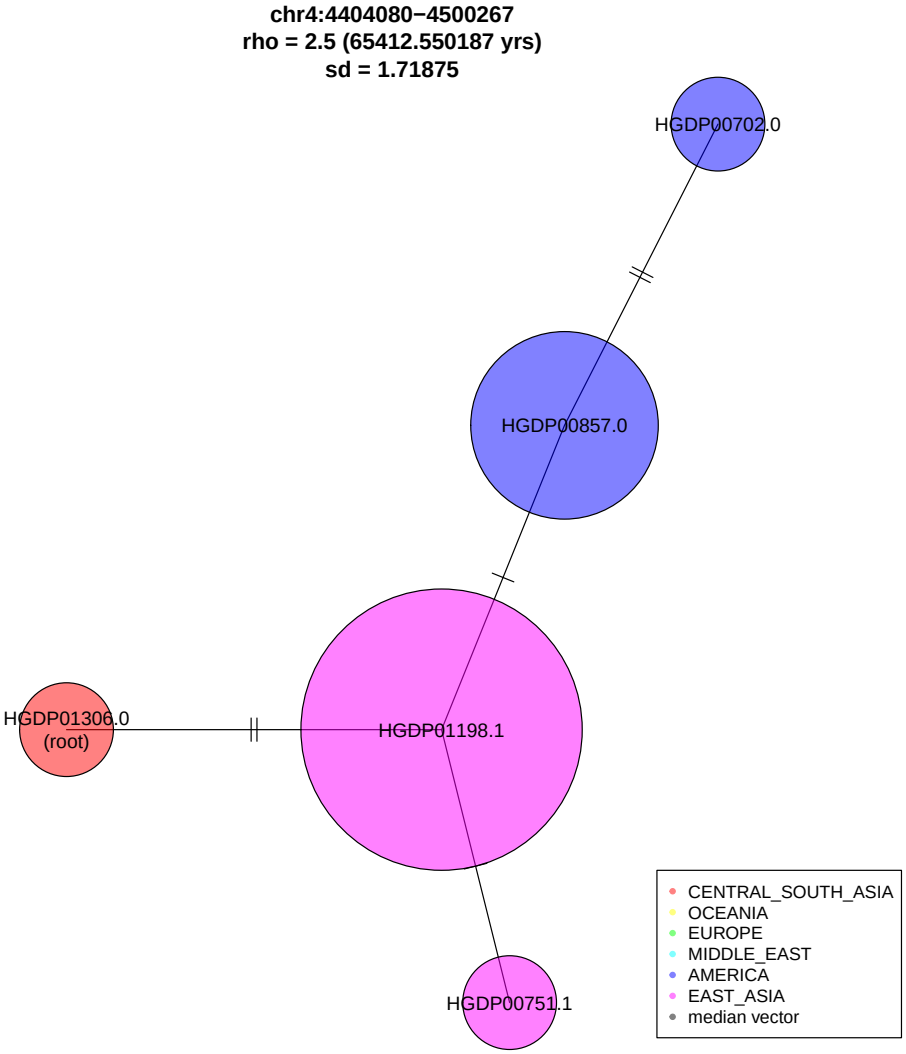


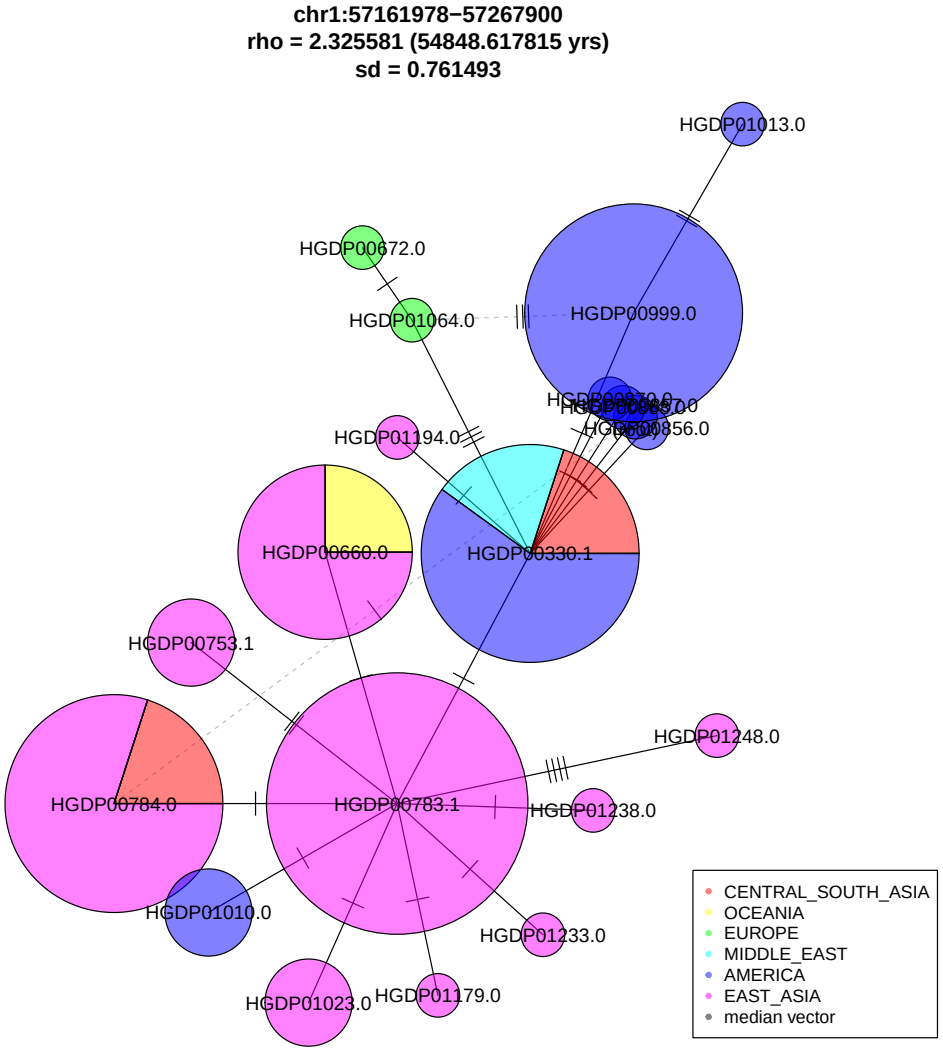

chr5:91362087−91461743
rho = 13.361111 (353456.456491 yrs)
sd = 10.125772
HGDP00952.0
HGDP00747.1
HGDP00540.0
(root)
HGDP00552.1
HGDP00759.0
HGDP00108.1
HGDP00541.1
HGDP00896.1
HGDP00764.0
HGDP00748.0
HGDP01296.0
HGDP01223.0
HGDP00541.0
HGDP01299.0
HGDP01330.0
HGDP00818.0
HGDP01327.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

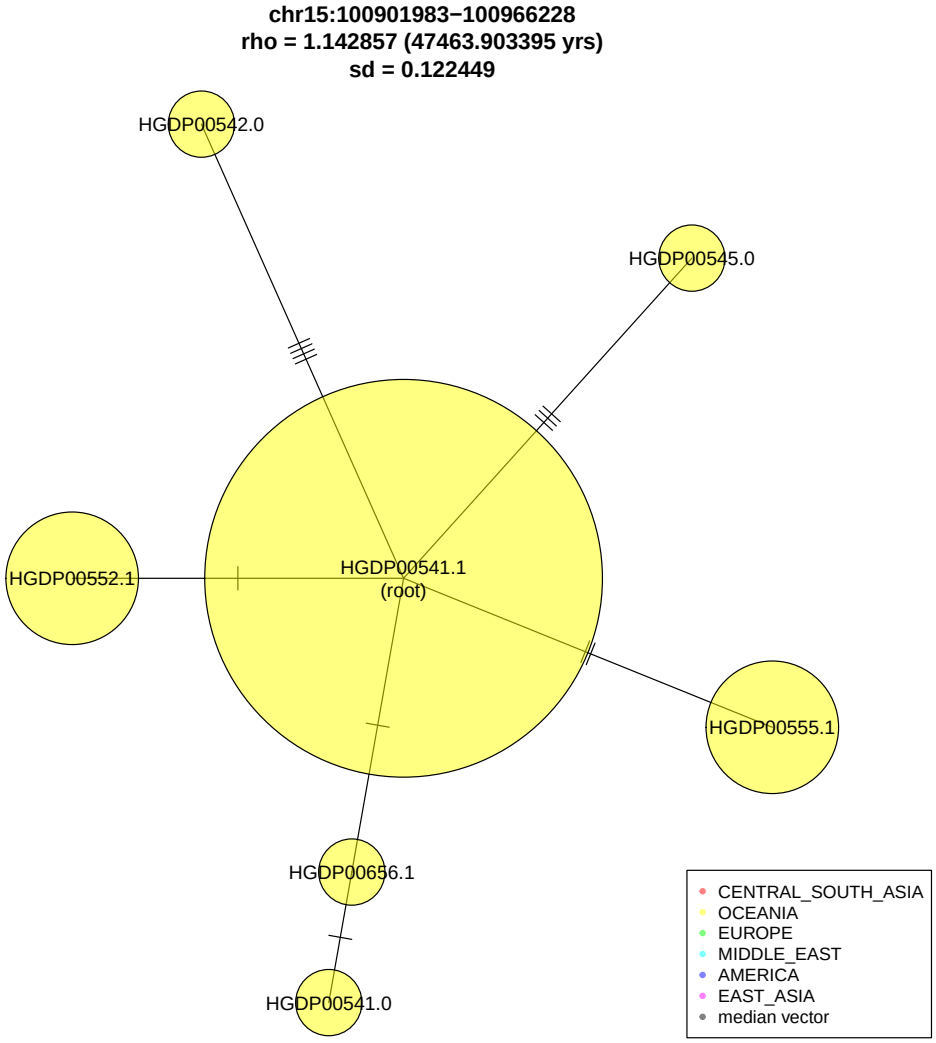

chr4:2350689−2401966
rho = 2.727273 (151504.267588 yrs)
sd = 1.123967
HGDP01101.0
HGDP01095.1
HGDP00542.0
HGDP01104.1
HGDP01288.1
HGDP00777.0
HGDP01295.0
(root)
HGDP00752.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

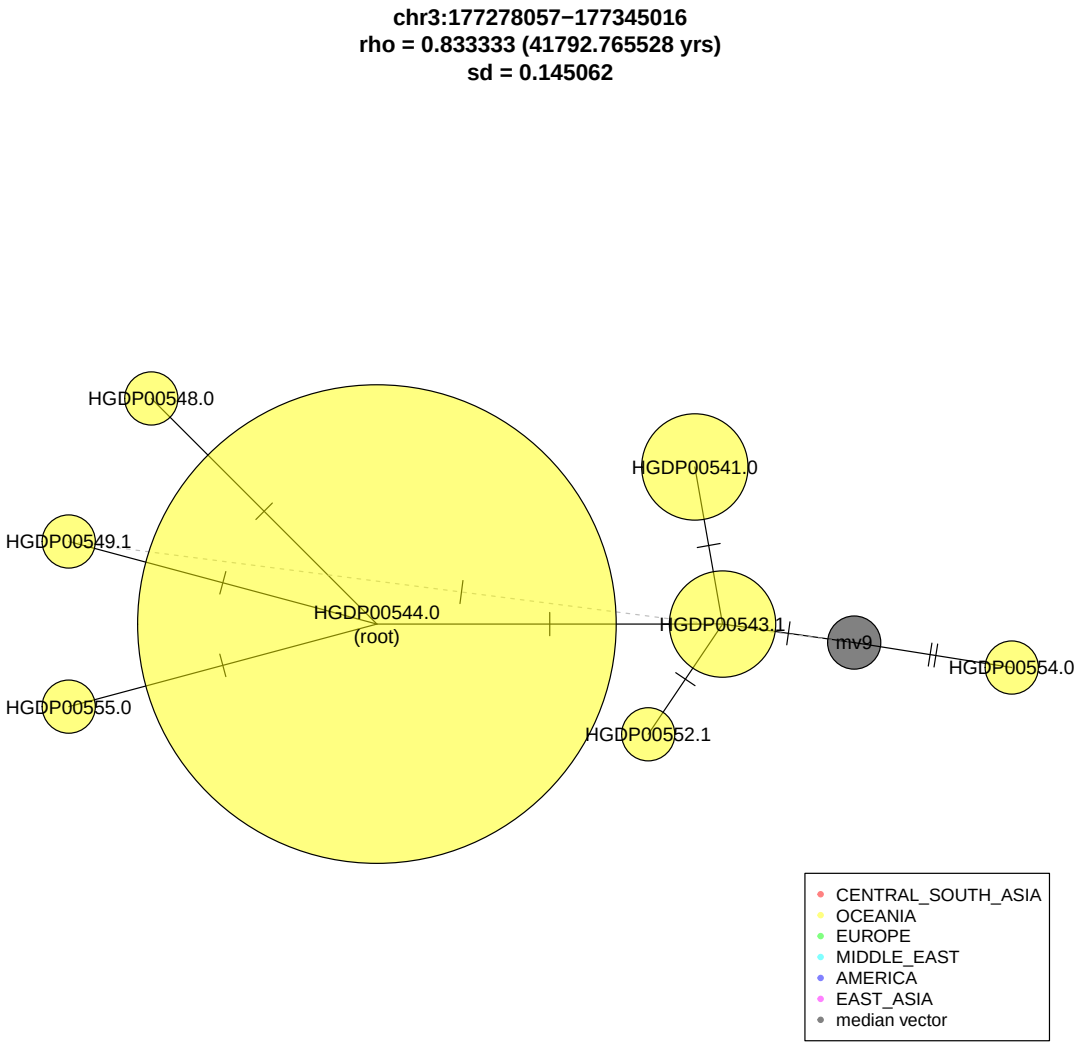

chr11:3943415−4060212
rho = 5 (140505.577829 yrs)
sd = 3.123457
HGDP00656.0
HGDP00491.1
HGDP00787.0
HGDP00553.0
HGDP00547.1
mv9
mv8
HGDP00154.0
HGDP00551.1
(root)
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

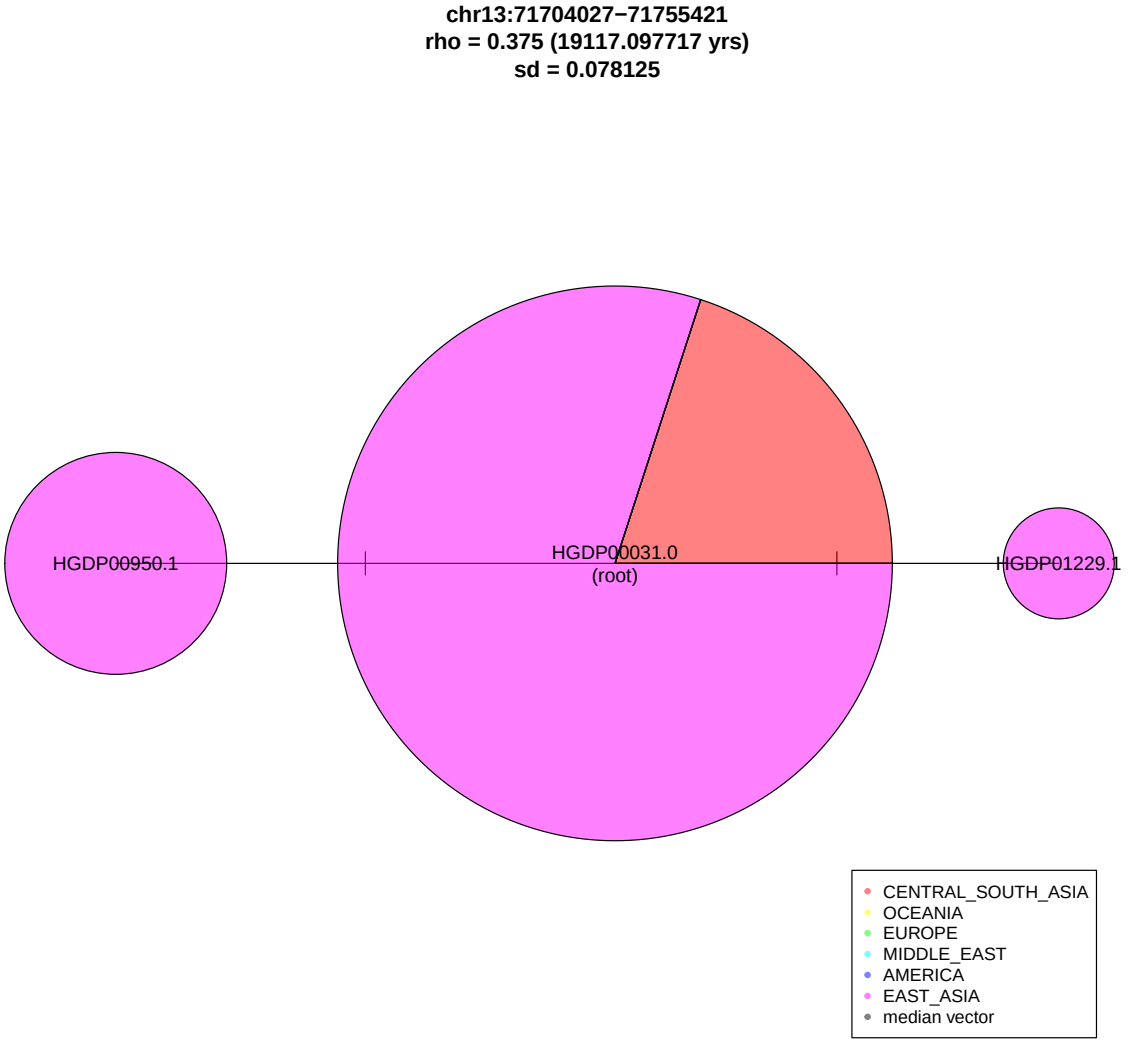

chr4:112451802−112508683
rho = 1.48 (90896.095301 yrs)
sd = 0.1168
HGDP00655.1
mV
HGDP00828.1
HGDP00753.1
HGDP00110.0
(root)
HGDP01336.0
HGDP01239.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector


chr4:4404080−4500267
rho = 2.5 (65412.550187 yrs)
sd = 1.71875
HGDP00702.0
HGDP00857.0
HGDP01306.0
(root)
HGDP01198.1
HGDP00751.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

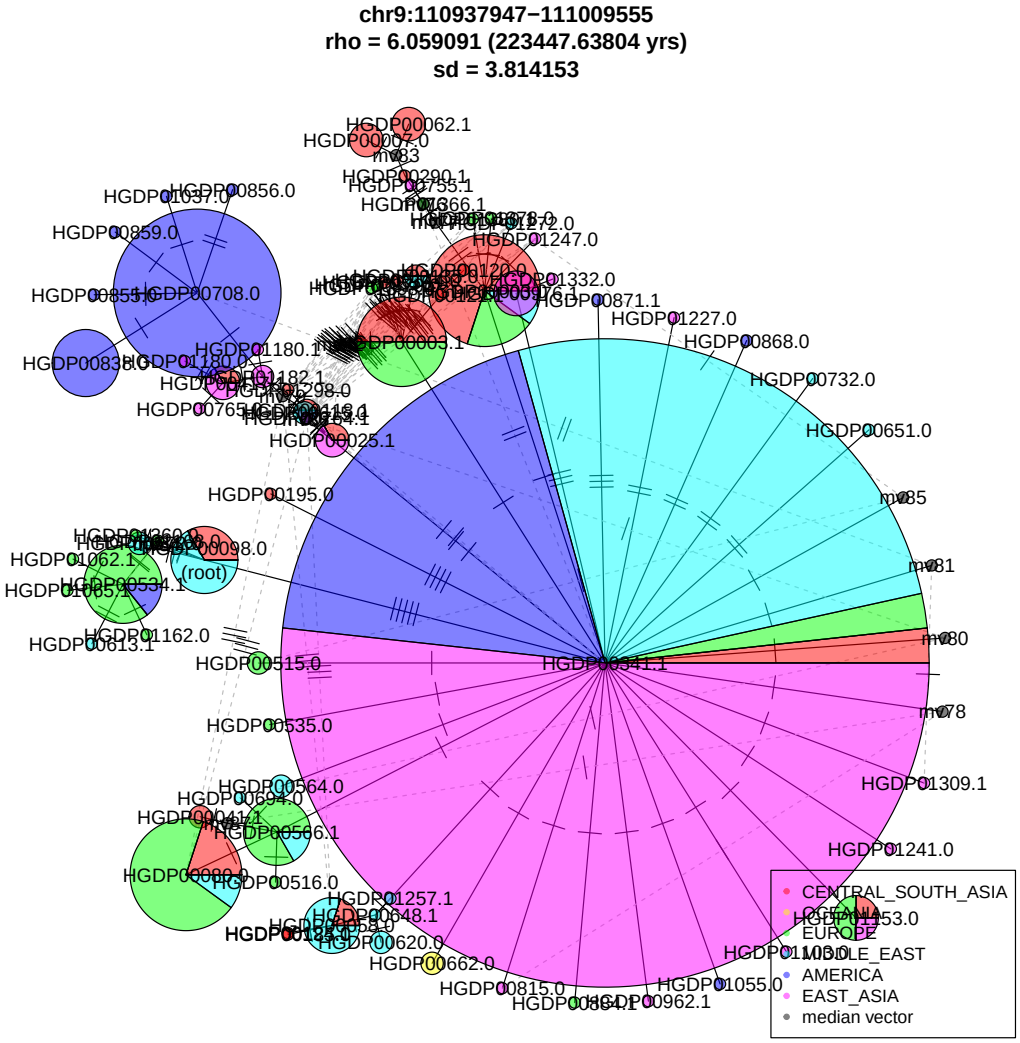

chr1:57161978−57267900
rho = 2.325581 (54848.617815 yrs)
sd = 0.761493
HGDP01013.0
HGDP00672.0
HGDP01064.0
HGDP00999.0
HGDP01194.0
HGDP00660.0
HGDP00330.1
HGDP00753.1
HGDP01248.0
HGDP00784.0
HGDP00783.1
HGDP01238.0
HGDP01010.0
HGDP01233.0
HGDP01023.0
HGDP01179.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

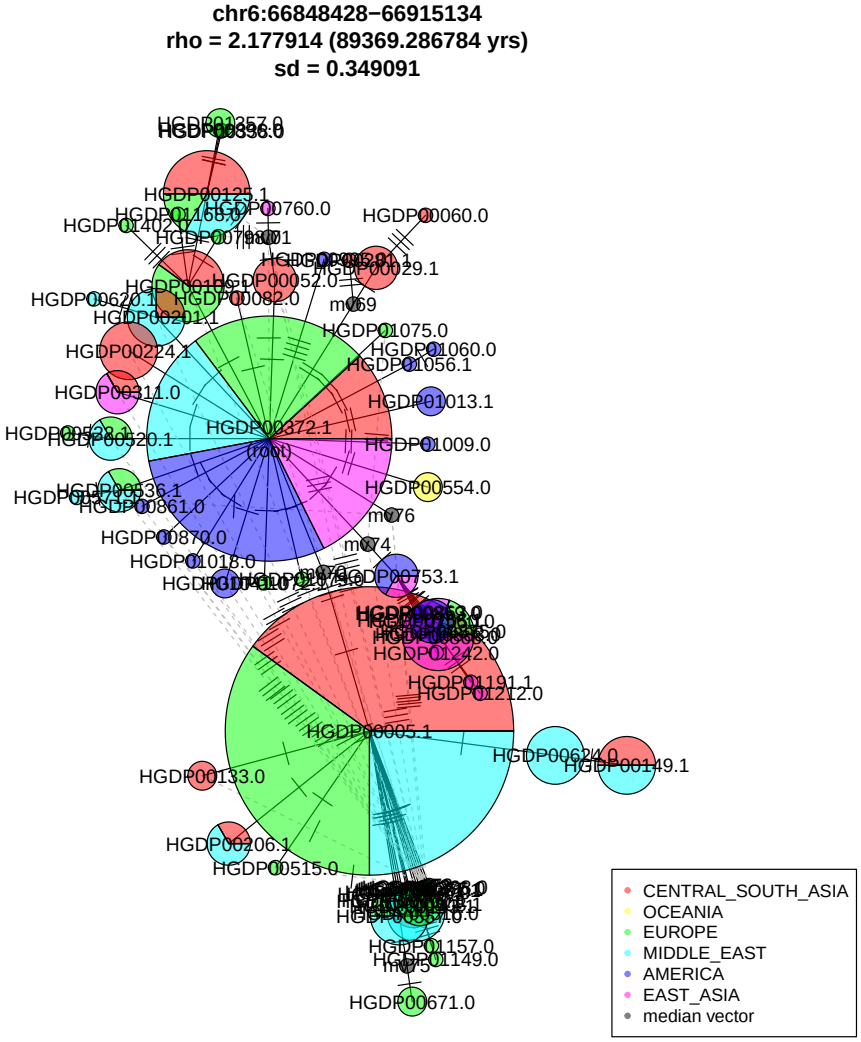

chr15:100901983−100966228
rho = 1.142857 (47463.903395 yrs)
sd = 0.122449
HGDP00542.0
HGDP00545.0
HGDP00552.1
HGDP00541.1
(root)
HGDP00555.1
HGDP00656.1
HGDP00541.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

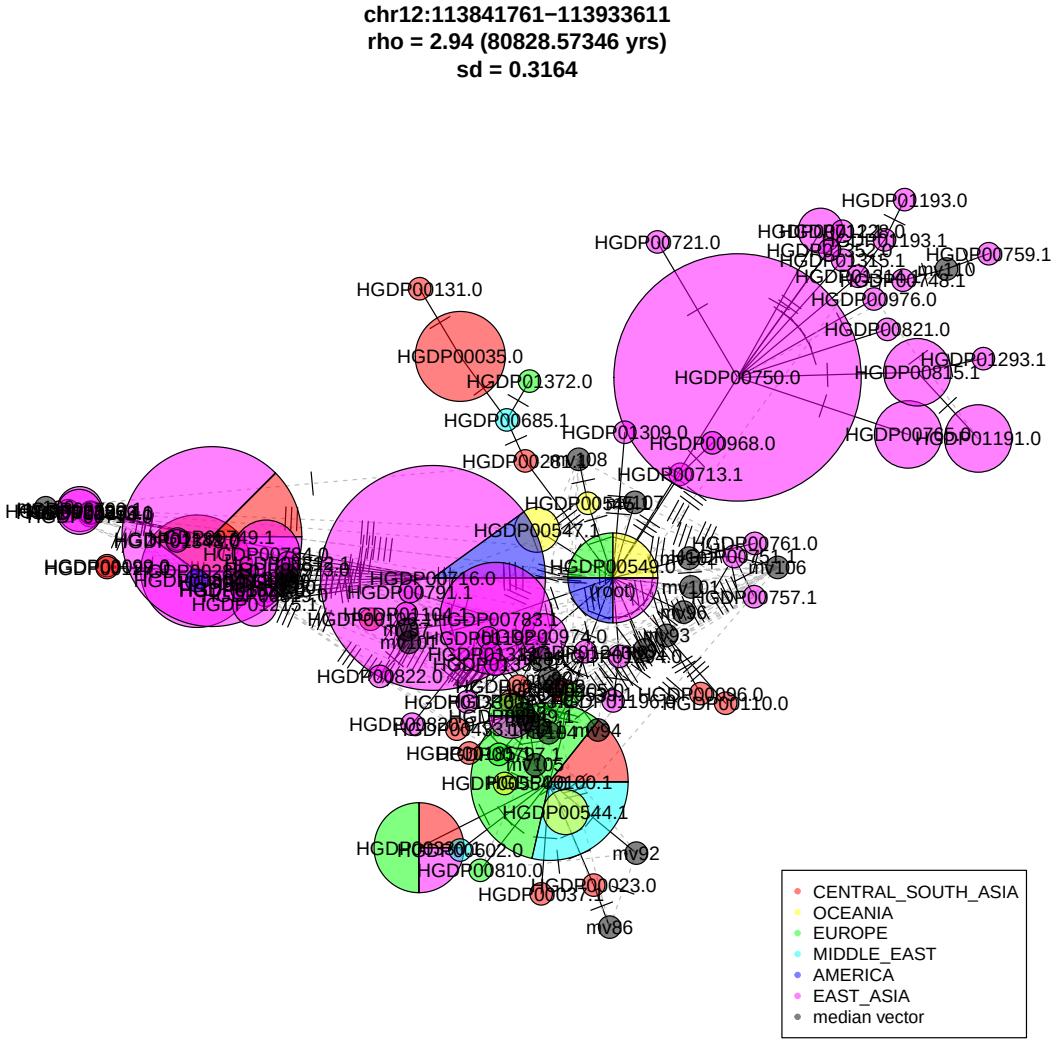

chr3:177278057−177345016
rho = 0.833333 (41792.765528 yrs)
sd = 0.145062
HGDP00548.0
HGDP00549.1
HGDP00555.0
HGDP00544.0
(root)
HGDP00541.0
HGDP00543.1
mv9
HGDP00554.0
HGDP00552.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

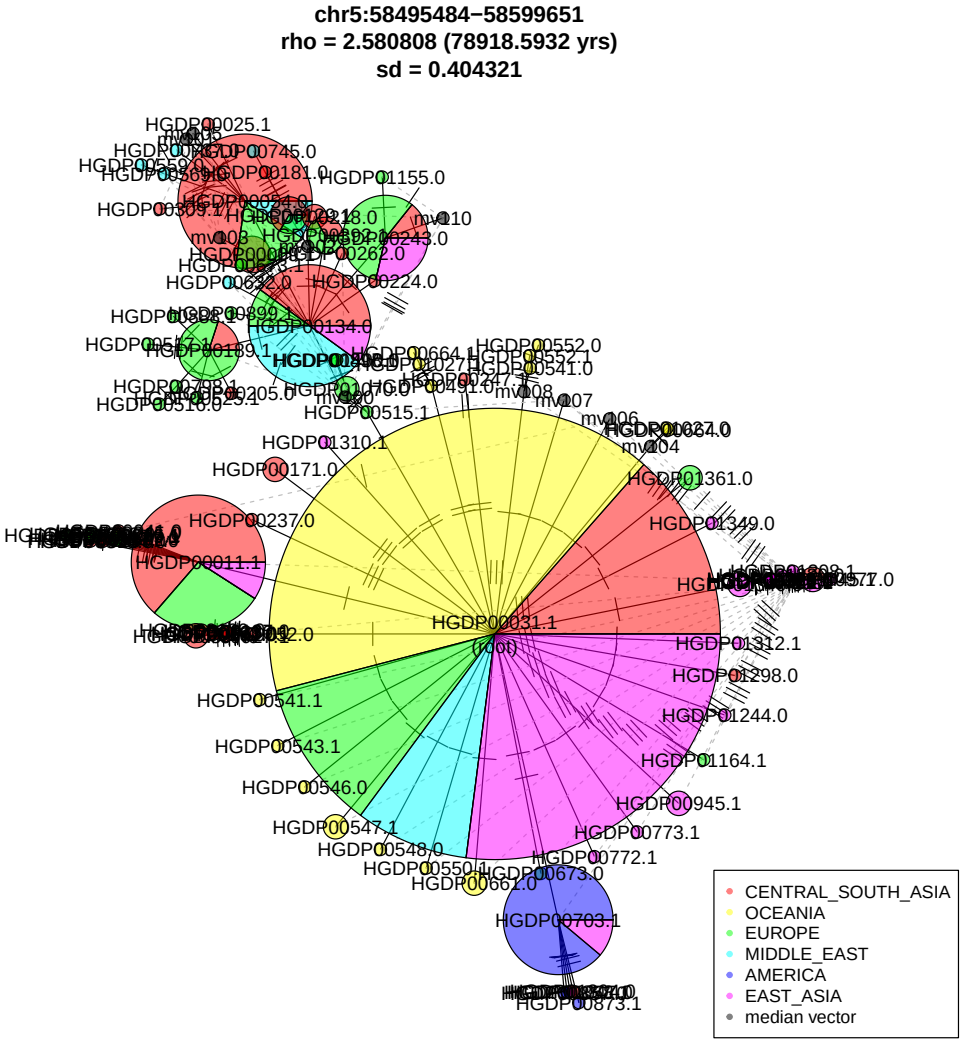

chr13:71704027−71755421
rho = 0.375 (19117.097717 yrs)
sd = 0.078125
HGDP00950.1
HGDP00031.0
(root)
HGDP01229.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

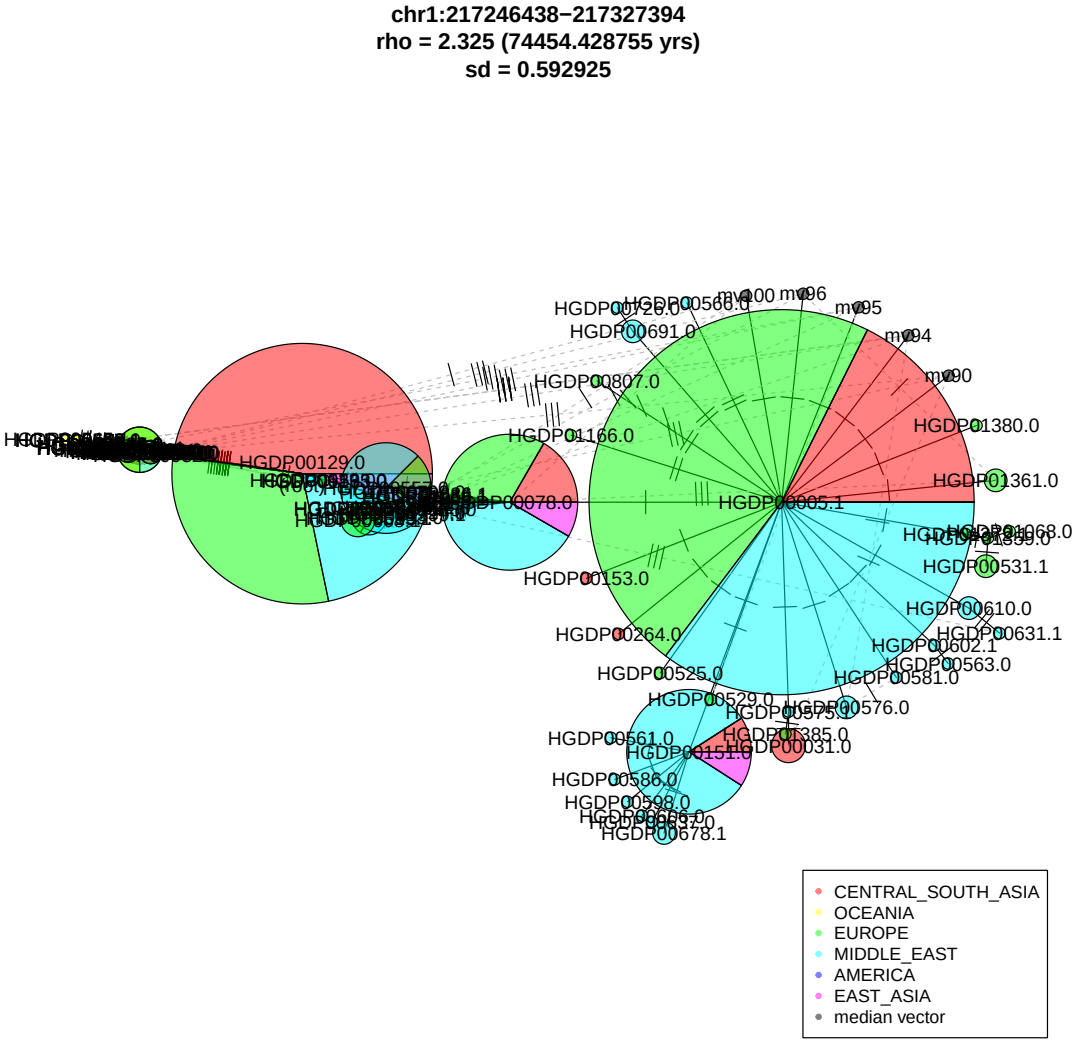
B.3 17 Neanderthal haplotype networks estimated to contain more than 20 founding haplotypes

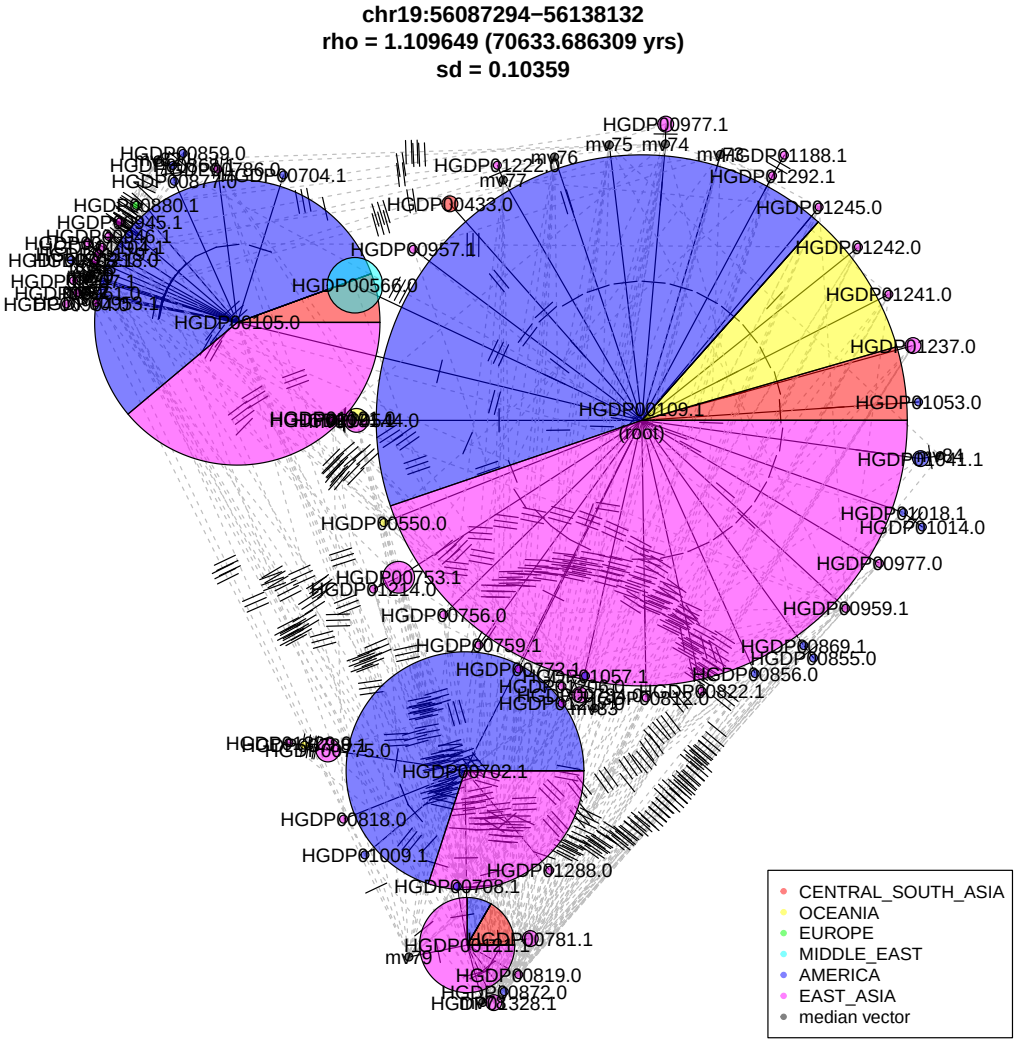


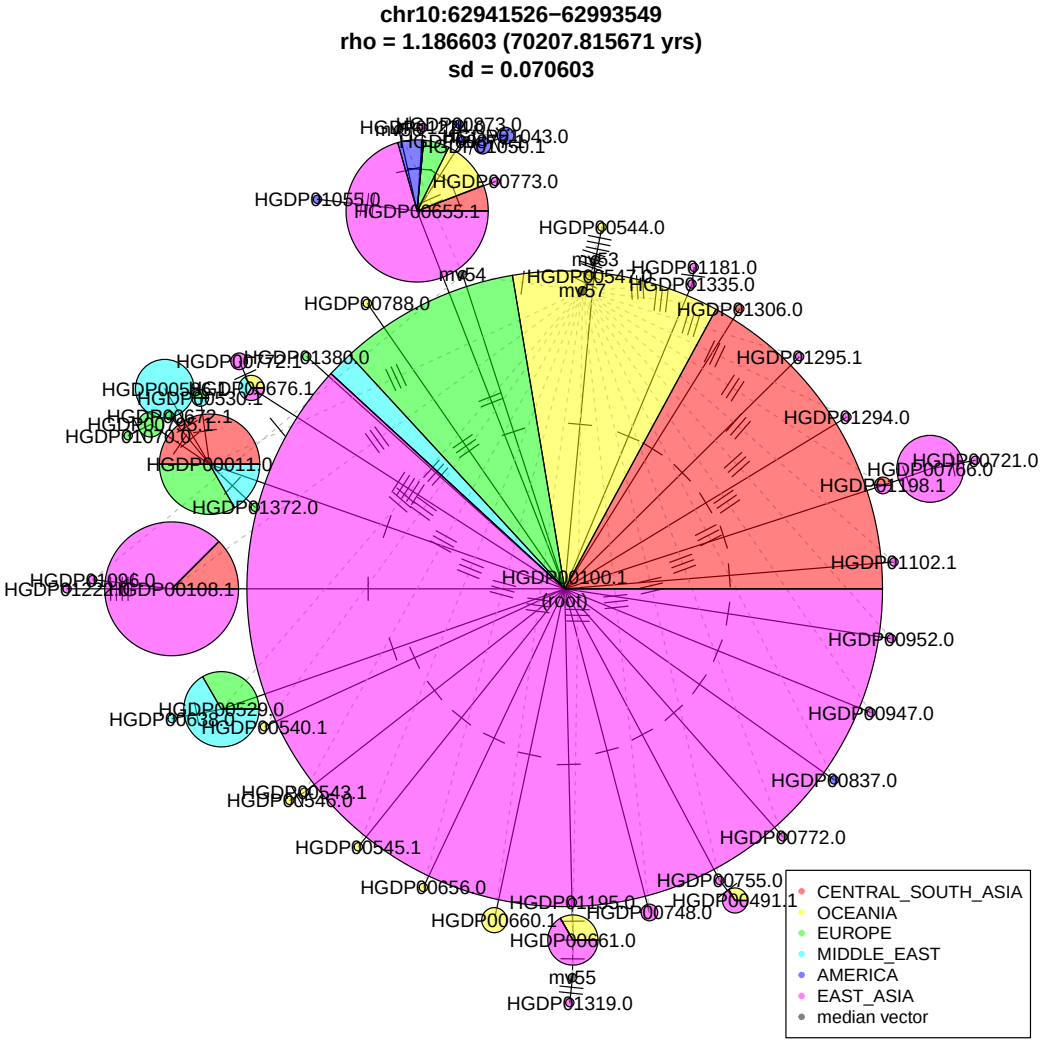

chr6:66848428−66915134
rho = 2.177914 (89369.286784 yrs)
sd = 0.349091
HGDP00125.1
HGDP00760.0
HGDP00060.0
HGDP00029.1
HGDP00052.0
HGDP00620.1
HGDP01082.0
m069
HGDP01075.0
HGDP01060.0
HGDP01056.1
HGDP00224.1
HGDP00311.0
HGDP01013.1
HGDP00372.1
(700)
HGDP01009.0
HGDP00554.0
m076
m074
HGDP00753.1
HGDP01242.0
HGDP01191.1
HGDP01212.0
HGDP00005.1
HGDP00624.0
HGDP00149.1
HGDP00133.0
HGDP00206.1
HGDP00515.0
HGDP01157.0
HGDP01149.0
HGDP00671.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

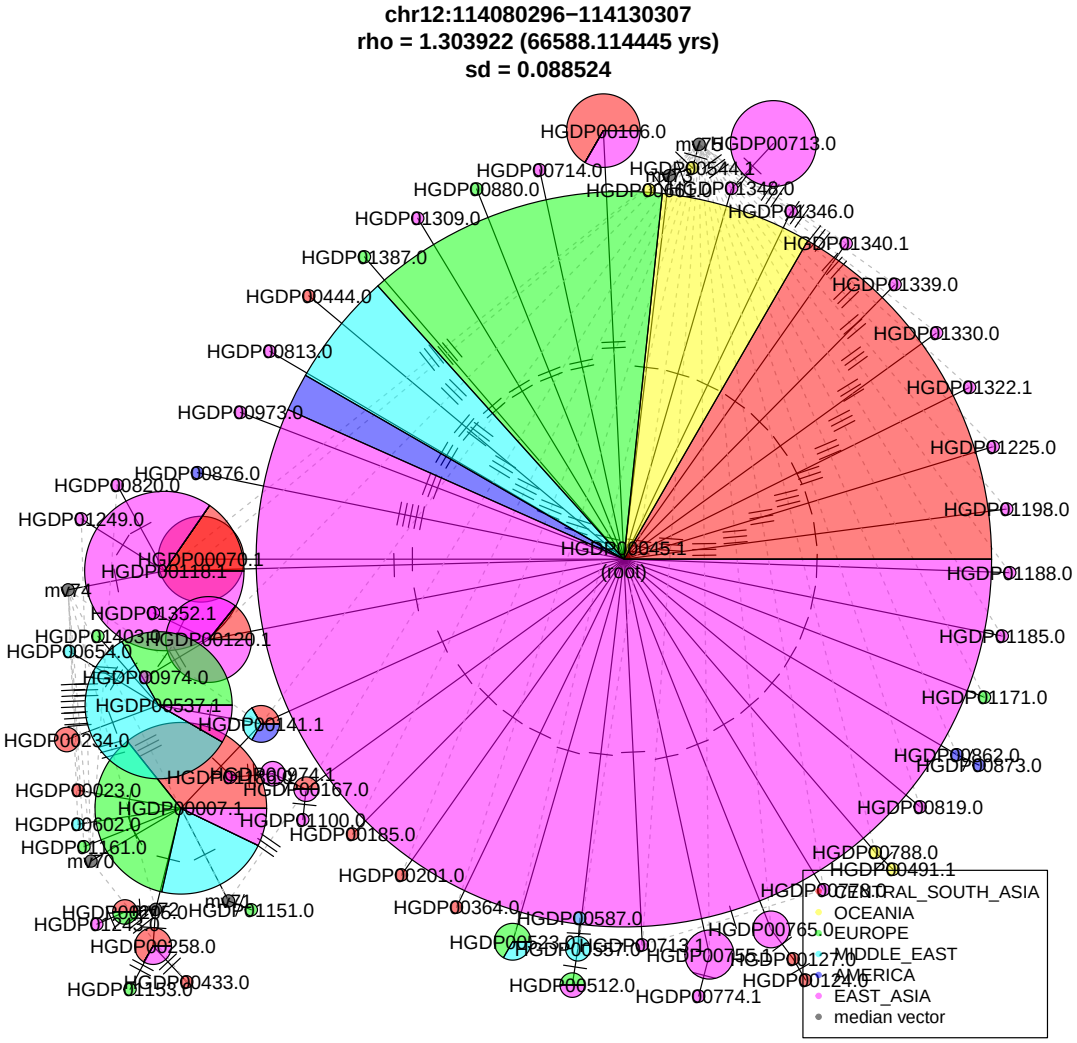

chr12:113841761−113933611
rho = 2.94 (80828.57346 yrs)
sd = 0.3164
HGDP01193.0
HGDP00721.0
HGDP00759.1
HGDP00131.0
HGDP00976.0
HGDP00821.0
HGDP00035.0
HGDP01372.0
HGDP00750.0
HGDP00815.1
HGDP01293.1
HGDP00685.1
HGDP01309.0
HGDP00968.0
HGDP00765.0
HGDP01191.0
HGDP00713.1
HGDP00547.1
HGDP00761.0
HGDP00549.0
HGDP00716.0
HGDP00791.1
HGDP00757.1
HGDP00783.1
HGDP00822.0
HGDP00896.0
HGDP00110.0
HGDP00544.1
HGDP00810.0
HGDP00023.0
mv92
mv86
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

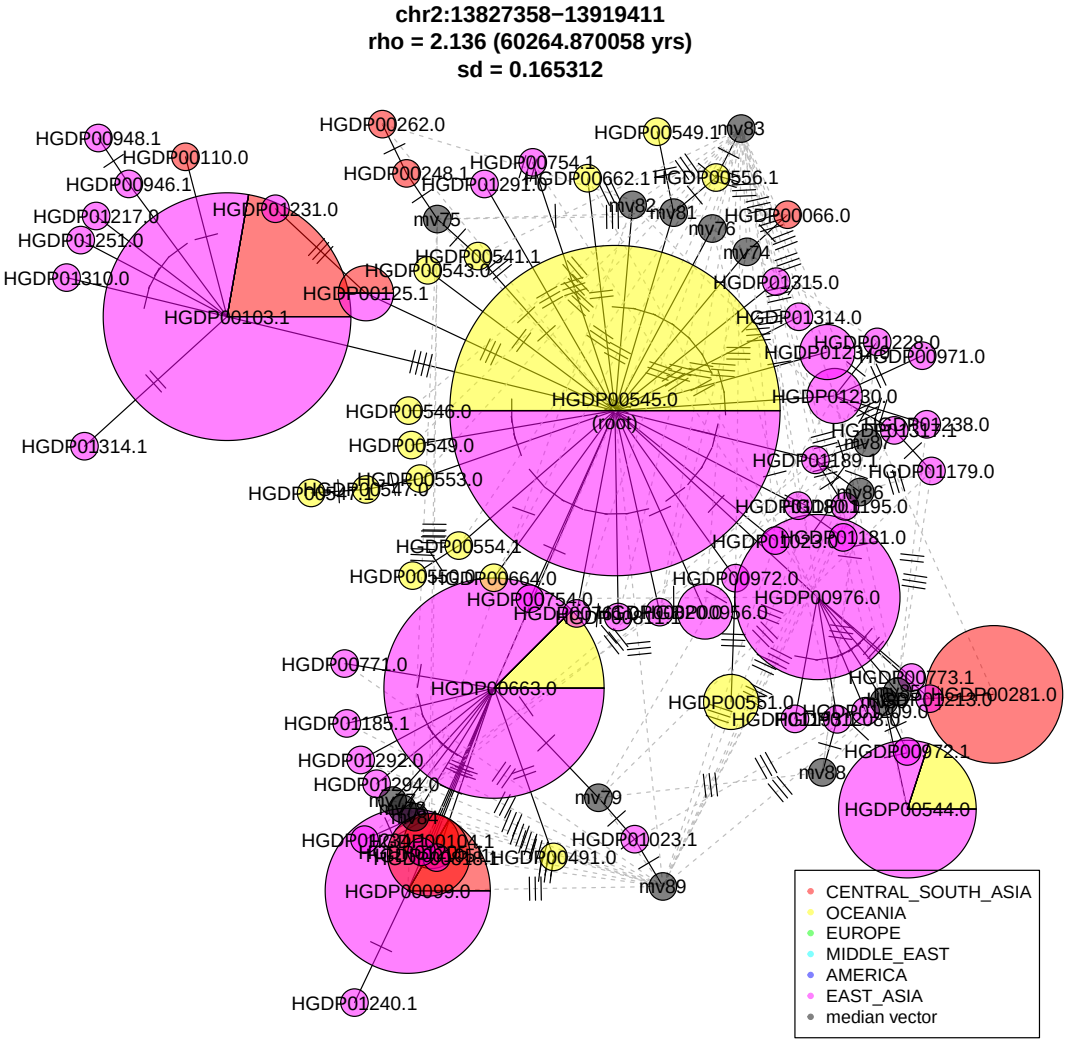

chr5:58495484−58599651
rho = 2.580808 (78918.5932 yrs)
sd = 0.404321
HGDP00031.1
(root)
HGDP00237.0
HGDP00011.1
HGDP01310.1
HGDP00171.0
HGDP01361.0
HGDP01349.0
HGDP01312.1
HGDP01298.0
HGDP01244.0
HGDP01164.1
HGDP00945.1
HGDP00773.1
HGDP00772.1
HGDP00673.0
HGDP00661.0
HGDP00703.1
HGDP00873.1
HGDP00541.1
HGDP00543.1
HGDP00546.0
HGDP00547.1
HGDP00548.0
HGDP00550.1
HGDP00515.1
HGDP00134.0
HGDP00224.0
HGDP01155.0
HGDP00745.0
HGDP00181.0
HGDP00025.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

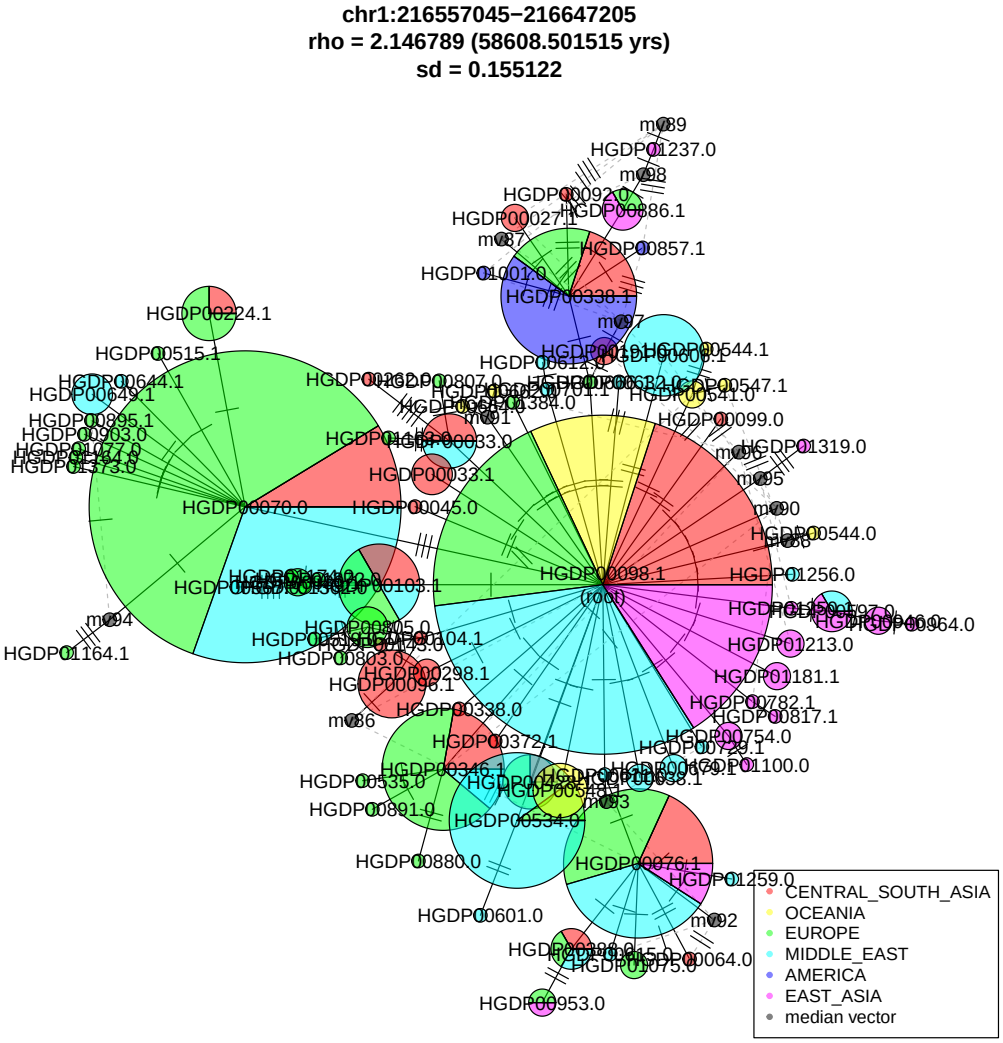

chr1:217246438−217327394
rho = 2.325 (74454.428755 yrs)
sd = 0.592925
HGDP00691.0
HGDP00807.0
HGDP01166.0
HGDP01380.0
HGDP01361.0
HGDP00129.0
HGDP00078.0
HGDP00005.1
HGDP00153.0
HGDP00531.1
HGDP00610.0
HGDP00631.1
HGDP00602.1
HGDP00563.0
HGDP00581.0
HGDP00264.0
HGDP00525.0
HGDP00529.0
HGDP00576.0
HGDP00031.0
HGDP00561.0
HGDP00586.0
HGDP00598.0
HGDP00678.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

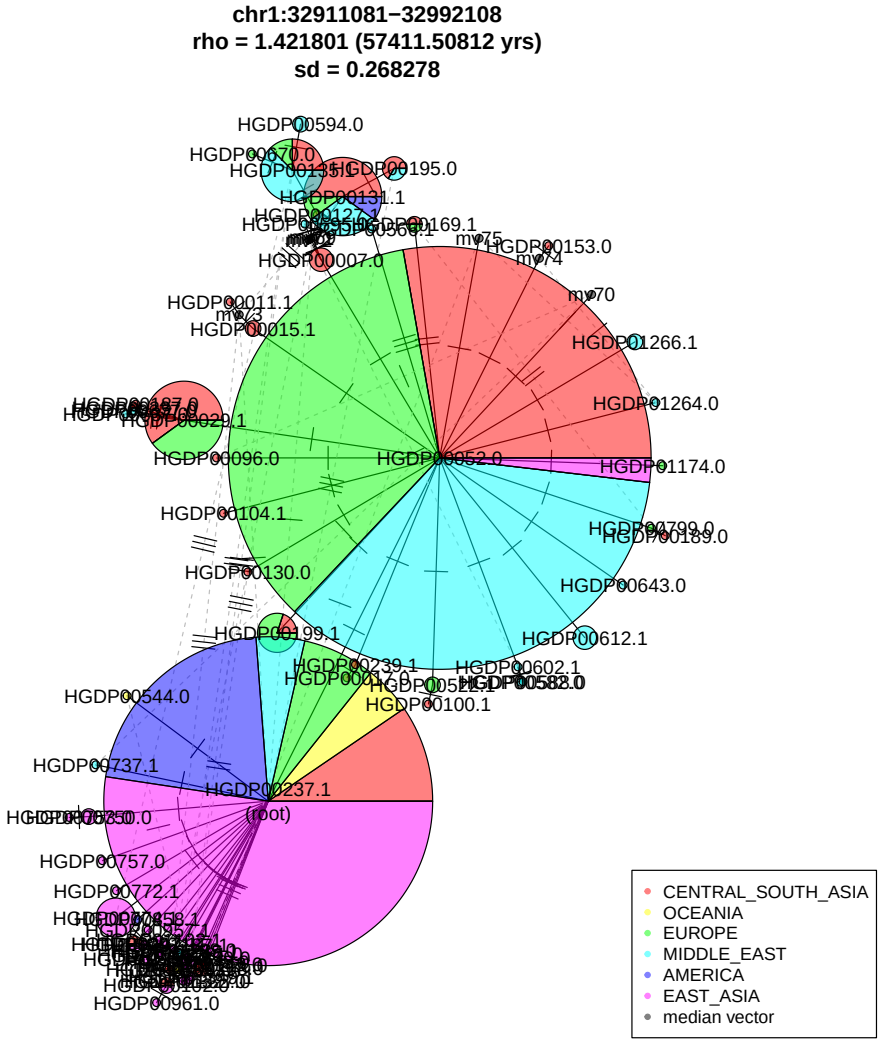

chr19:56087294−56138132
rho = 1.109649 (70633.686309 yrs)
sd = 0.10359
HGDP00977.1
mv75
mv74
HGDP01188.1
HGDP01292.1
HGDP01245.0
HGDP01242.0
HGDP01241.0
HGDP01237.0
HGDP01053.0
HGDP01018.1
HGDP01014.0
HGDP00977.0
HGDP00959.1
HGDP00869.1
HGDP00855.0
HGDP00856.0
HGDP00822.1
HGDP00812.0
HGDP00781.1
HGDP00819.0
HGDP00872.0
HGDP00818.0
HGDP01009.1
HGDP01288.0
HGDP00708.1
HGDP00702.1
HGDP00759.1
HGDP00756.0
HGDP00753.1
HGDP01214.0
HGDP00550.0
HGDP00566.0
HGDP00105.0
HGDP00957.1
HGDP00433.0
HGDP00704.1
HGDP00880.1
HGDP00859.0
mv79
(root)
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

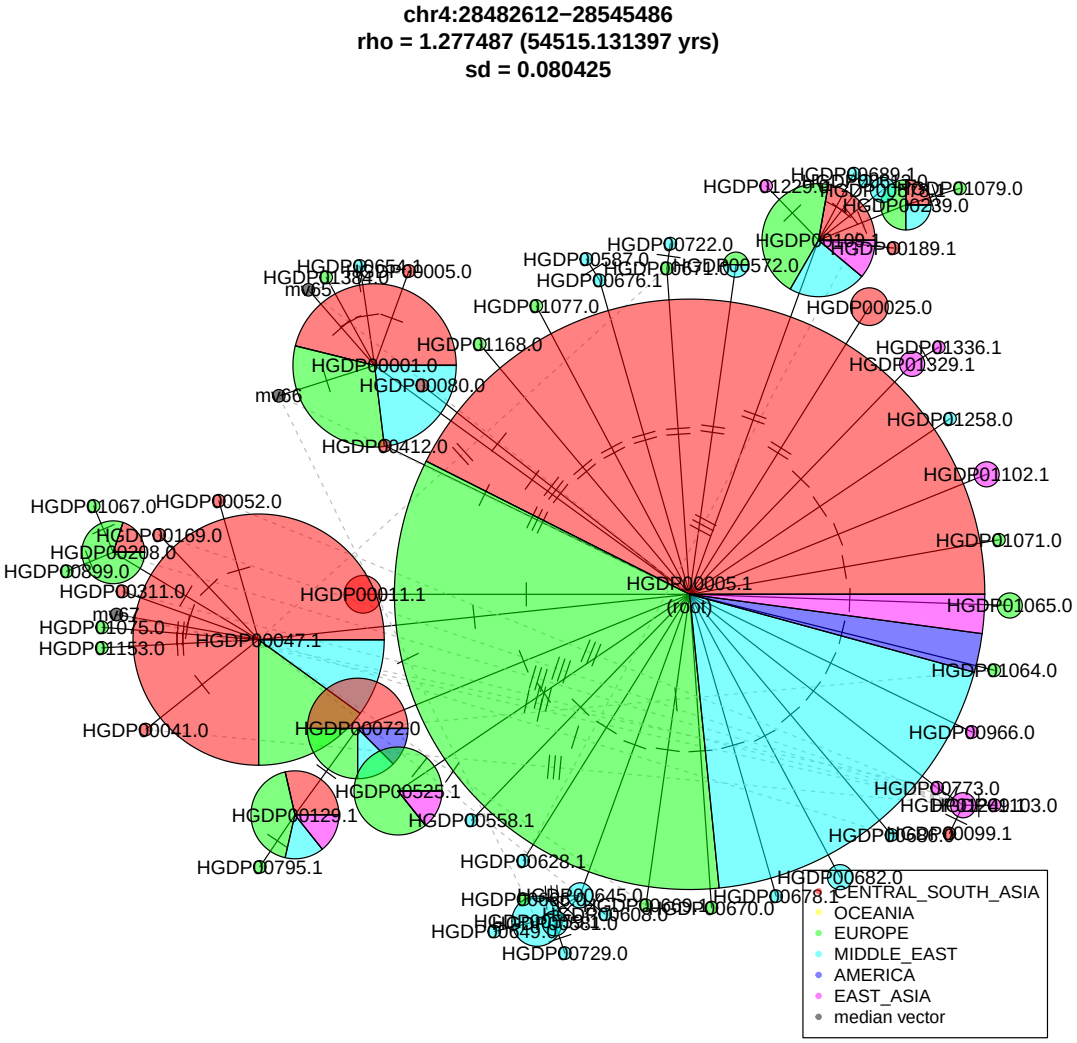

chr10:62941526−62993549
rho = 1.186603 (70207.815671 yrs)
sd = 0.070603
HGDP00773.0
HGDP01055.0
HGDP00655.1
HGDP00544.0
mv53
HGDP01181.0
mv54
HGDP00547.0
HGDP01335.0
mv57
HGDP00788.0
HGDP01306.0
HGDP01380.0
HGDP01295.1
HGDP00676.1
HGDP01294.0
HGDP00011.0
HGDP00721.0
HGDP00766.0
HGDP01198.1
HGDP01372.0
HGDP01102.1
HGDP00108.1
HGDP01100.1
(root)
HGDP00952.0
HGDP00529.0
HGDP00540.1
HGDP00947.0
HGDP00543.1
HGDP00837.0
HGDP00545.1
HGDP00772.0
HGDP00656.0
HGDP00755.0
HGDP01195.0
HGDP00748.0
HGDP00660.1
HGDP00661.0
mv55
HGDP01319.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

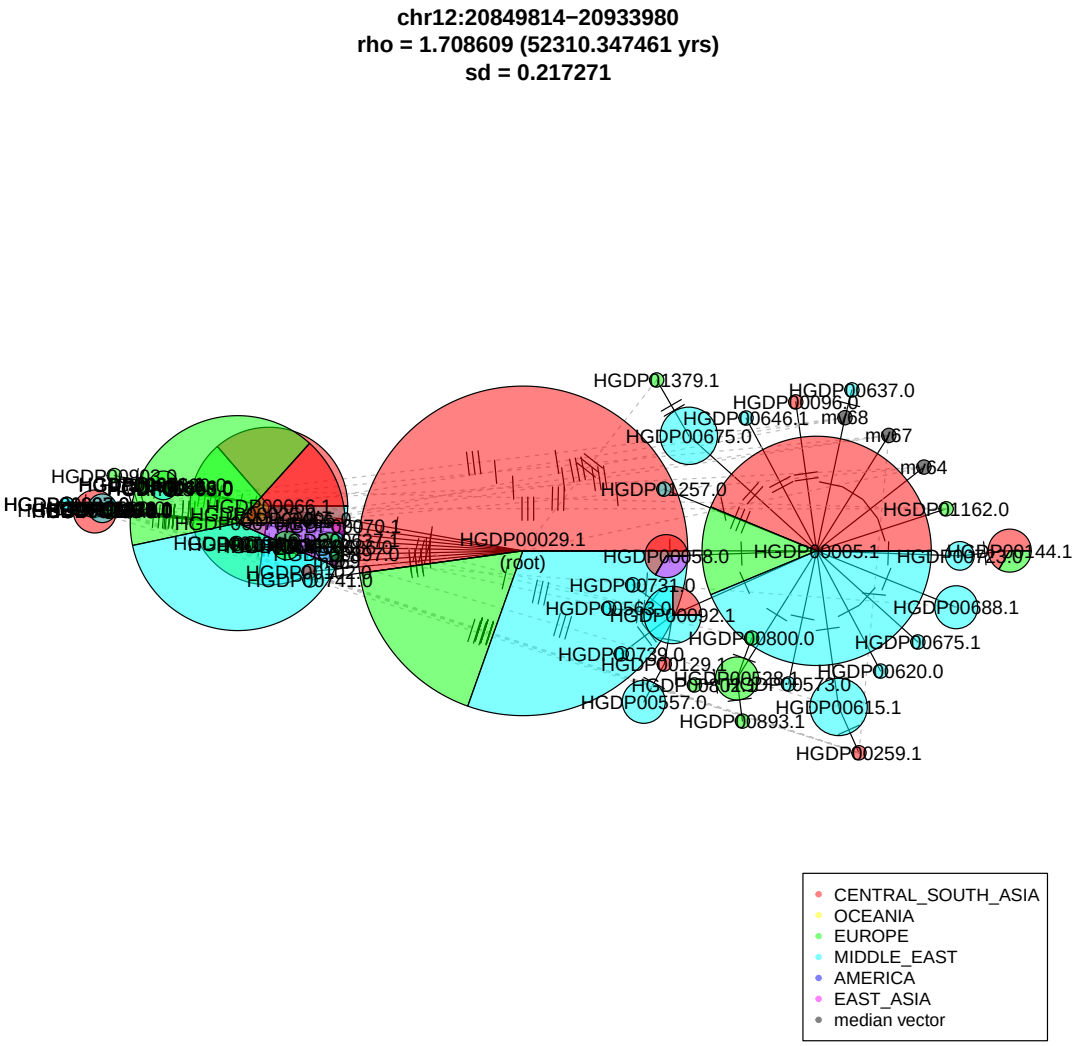

chr12:114080296−114130307
rho = 1.303922 (66588.114445 yrs)
sd = 0.088524
HGDP00166.0
HGDP00713.0
HGDP00714.0
HGDP00880.0
HGDP01309.0
HGDP01387.0
HGDP00444.0
HGDP00813.0
HGDP00973.0
HGDP00876.0
HGDP00820.0
HGDP01249.0
HGDP00070.1
HGDP00118.1
HGDP01252.1
HGDP00120.1
HGDP00651.0
HGDP00974.0
HGDP00537.1
HGDP00234.0
HGDP00141.1
HGDP00023.0
HGDP00007.1
HGDP00602.0
HGDP01161.0
HGDP00167.0
HGDP01100.0
HGDP00185.0
HGDP01151.0
HGDP01243.0
HGDP00258.0
HGDP00433.0
HGDP01133.0
HGDP00201.0
HGDP00364.0
HGDP00587.0
HGDP00512.0
HGDP00774.1
HGDP00765.0
HGDP00124.0
HGDP01348.0
HGDP01346.0
HGDP01340.1
HGDP01339.0
HGDP01330.0
HGDP01322.1
HGDP01225.0
HGDP01198.0
HGDP00045.1
(root)
HGDP01188.0
HGDP01185.0
HGDP01171.0
HGDP00862.0
HGDP00873.0
HGDP00819.0
HGDP00788.0
HGDP00491.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

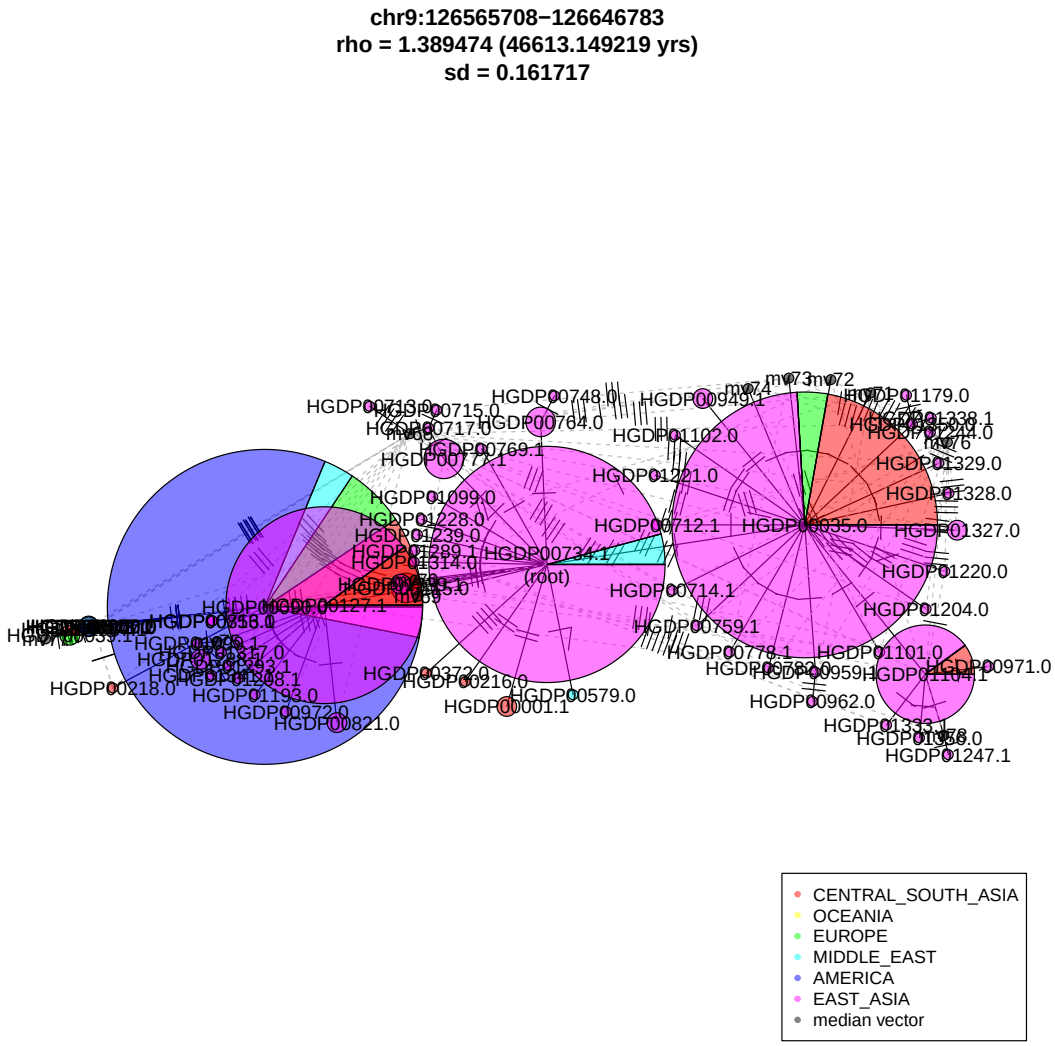

chr2:13827358−13919411
rho = 2.136 (60264.870058 yrs)
sd = 0.165312
HGDP00948.1
HGDP00110.0
HGDP00946.1
HGDP01217.0
HGDP01251.0
HGDP01310.0
HGDP01231.0
HGDP00103.1
HGDP01314.1
HGDP00262.0
HGDP00248.1
HGDP01291.0
HGDP00754.1
HGDP00662.1
HGDP00549.1
mv83
HGDP00556.1
mv82
mv81
HGDP00066.0
mv76
mv74
mv75
HGDP00541.1
HGDP00543.0
HGDP00125.1
HGDP01315.0
HGDP01314.0
HGDP01228.0
HGDP00971.0
HGDP00545.0
(root)
HGDP01230.0
HGDP00546.0
HGDP00549.0
HGDP00553.0
HGDP01179.0
mv86
HGDP01195.0
HGDP01181.0
HGDP00554.1
HGDP00664.0
HGDP00754.0
HGDP00972.0
HGDP00976.0
HGDP00771.0
HGDP00663.0
HGDP00773.1
HGDP00281.0
HGDP00551.0
HGDP01213.0
HGDP01185.1
HGDP01292.0
HGDP01294.0
HGDP00972.1
mv88
mv79
HGDP00544.0
HGDP01023.1
HGDP00491.0
mv89
HGDP00099.0
HGDP01240.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

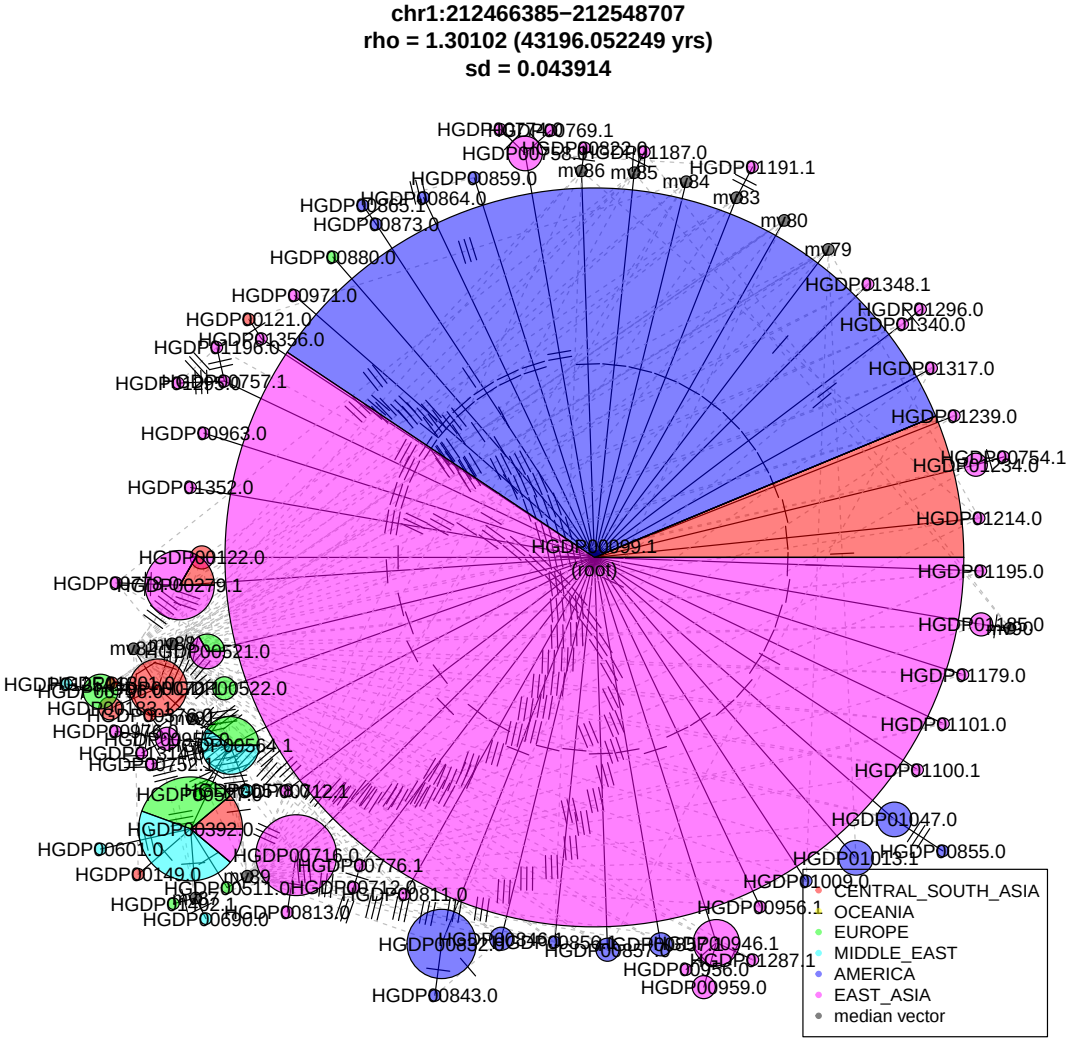

chr1:216557045−216647205
rho = 2.146789 (58608.501515 yrs)
sd = 0.155122
m89
HGDP01237.0
m88
HGDP00886.1
HGDP00027.1
m87
HGDP00857.1
HGDP01001.0
HGDP00338.1
m97
HGDP00224.1
HGDP00515.1
HGDP00644.1
HGDP00649.1
HGDP00895.1
HGDP00903.0
HGDP00544.1
HGDP00547.1
HGDP00541.0
HGDP00099.0
HGDP01319.0
m96
m95
m90
HGDP00544.0
HGDP00033.1
HGDP00070.0
HGDP00045.0
HGDP00098.1
HGDP01256.0
(root)
HGDP00964.0
HGDP01213.0
HGDP01181.1
HGDP00782.1
HGDP00817.1
HGDP00754.0
HGDP00729.1
HGDP01100.0
m94
HGDP01164.1
HGDP00803.0
HGDP00298.1
HGDP00338.0
m86
HGDP00372.1
HGDP00535.0
HGDP00891.0
HGDP00534.0
m93
HGDP00880.0
HGDP00076.1
HGDP01259.0
HGDP00601.0
m92
HGDP01075.0
HGDP00953.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

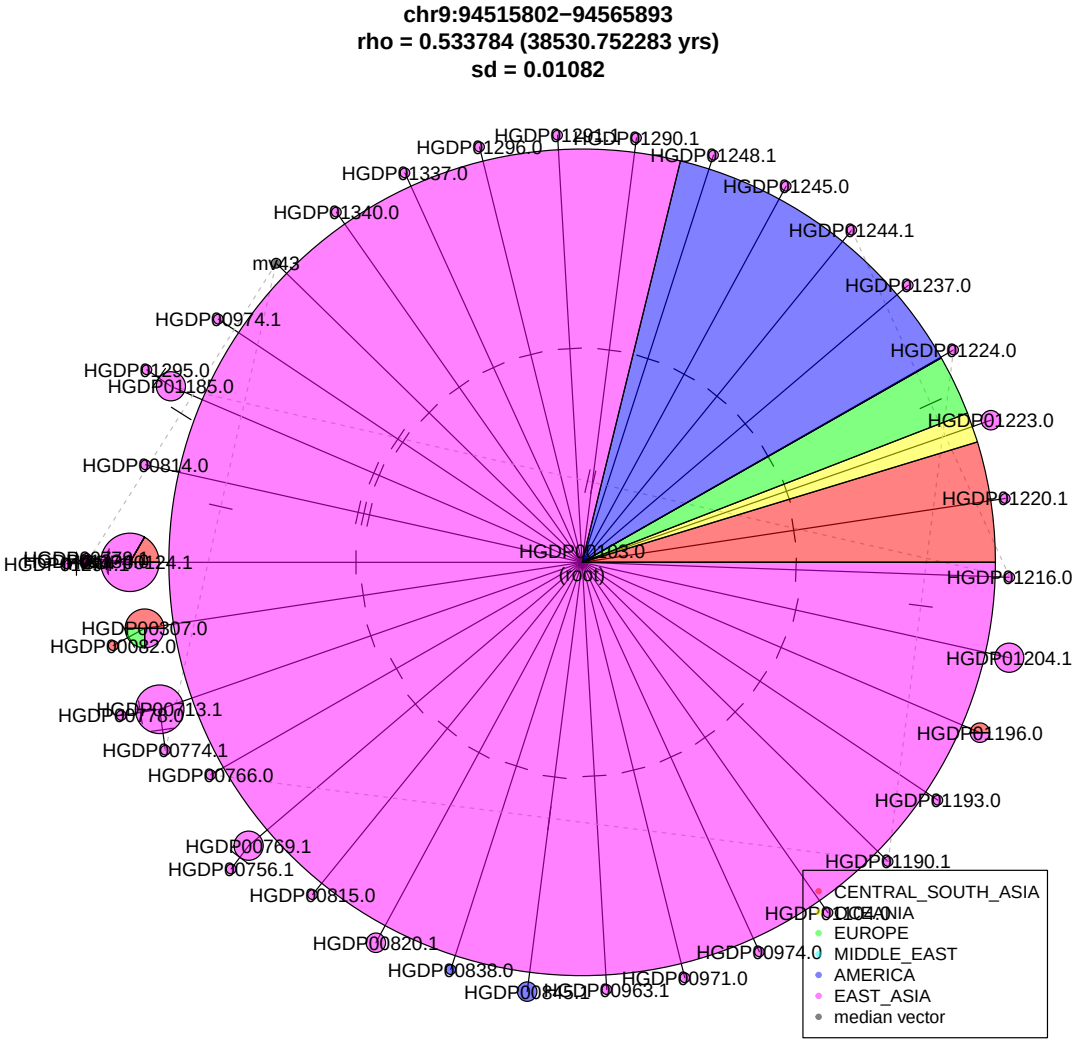

chr1:32911081−32992108
rho = 1.421801 (57411.50812 yrs)
sd = 0.268278
HGDP00594.0
HGDP00670.0
HGDP00195.0
HGDP00131.1
HGDP00153.0
mv75
mv74
mv70
HGDP00007.0
HGDP00011.1
mv73
HGDP00015.1
HGDP01266.1
HGDP01264.0
HGDP00029.1
HGDP00096.0
HGDP00052.0
HGDP01174.0
HGDP00104.1
HGDP00799.0
HGDP00189.0
HGDP00130.0
HGDP00643.0
HGDP00199.1
HGDP00612.1
HGDP00239.1
HGDP00602.1
HGDP00100.1
HGDP00544.0
HGDP00737.1
HGDP00237.1
(root)
HGDP00757.0
HGDP00772.1
HGDP00961.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

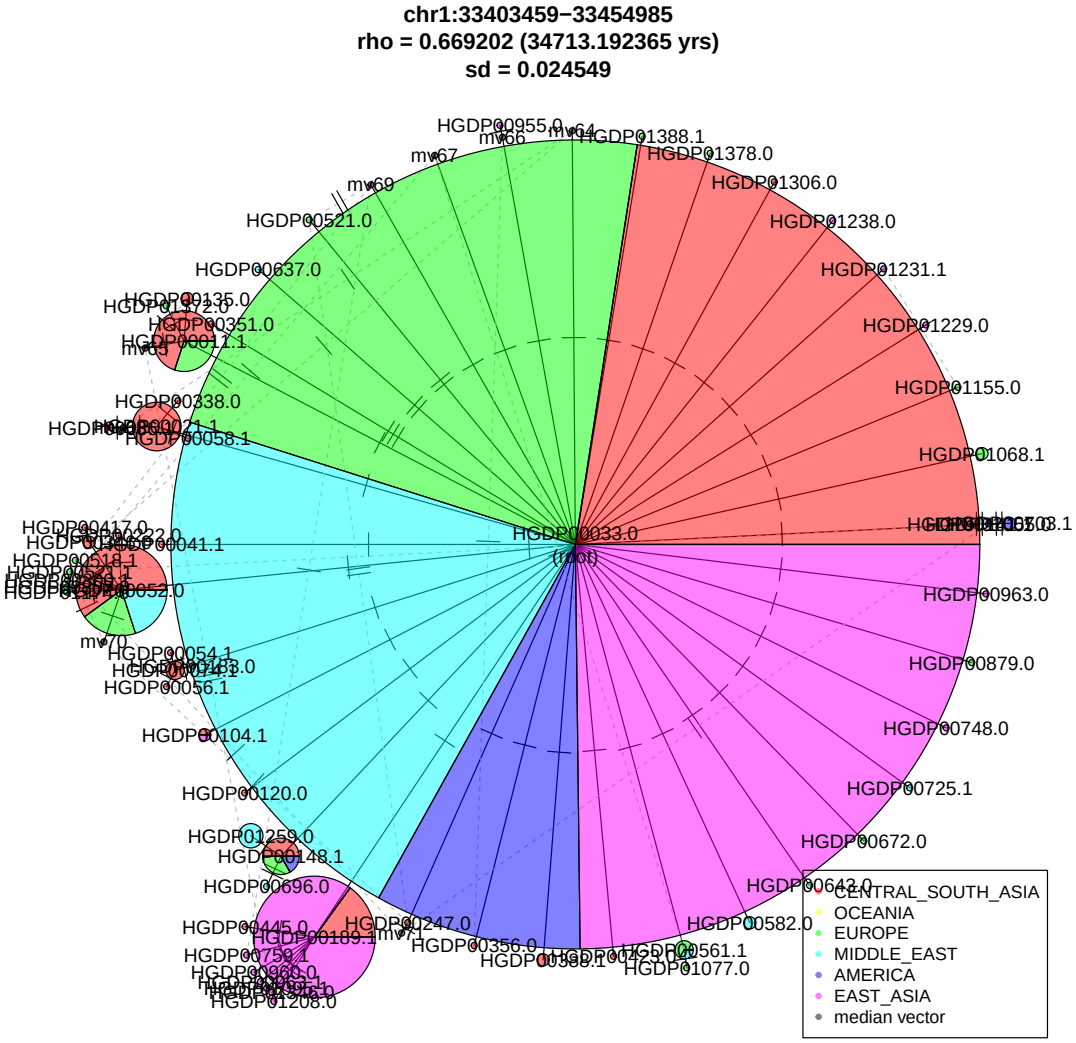

chr4:28482612−28545486
rho = 1.277487 (54515.131397 yrs)
sd = 0.080425
HGDP00005.1
(root)
HGDP00001.0
HGDP00011.1
HGDP00025.0
HGDP00041.0
HGDP00047.1
HGDP00052.0
HGDP00072.0
HGDP00080.0
HGDP00129.1
HGDP00189.1
HGDP00311.0
HGDP00412.0
HGDP00525.1
HGDP00558.1
HGDP00572.0
HGDP00587.0
HGDP00628.1
HGDP00645.0
HGDP00670.0
HGDP00671.0
HGDP00676.1
HGDP00678.1
HGDP00682.0
HGDP00722.0
HGDP00729.0
HGDP00773.0
HGDP00795.1
HGDP00899.0
HGDP00966.0
HGDP01064.0
HGDP01065.0
HGDP01067.0
HGDP01071.0
HGDP01075.0
HGDP01077.0
HGDP01079.0
HGDP01102.1
HGDP01153.0
HGDP01168.0
HGDP01258.0
HGDP01329.1
HGDP01336.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector


chr12:20849814−20933980
rho = 1.708609 (52310.347461 yrs)
sd = 0.217271
HGDP01379.1
HGDP00637.0
HGDP00646.1
HGDP00675.0
HGDP01257.0
HGDP01162.0
HGDP00029.1
(root)
HGDP00058.0
HGDP00731.0
HGDP00688.1
HGDP00800.0
HGDP00675.1
HGDP00620.0
HGDP00557.0
HGDP00615.1
HGDP00893.1
HGDP00259.1
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

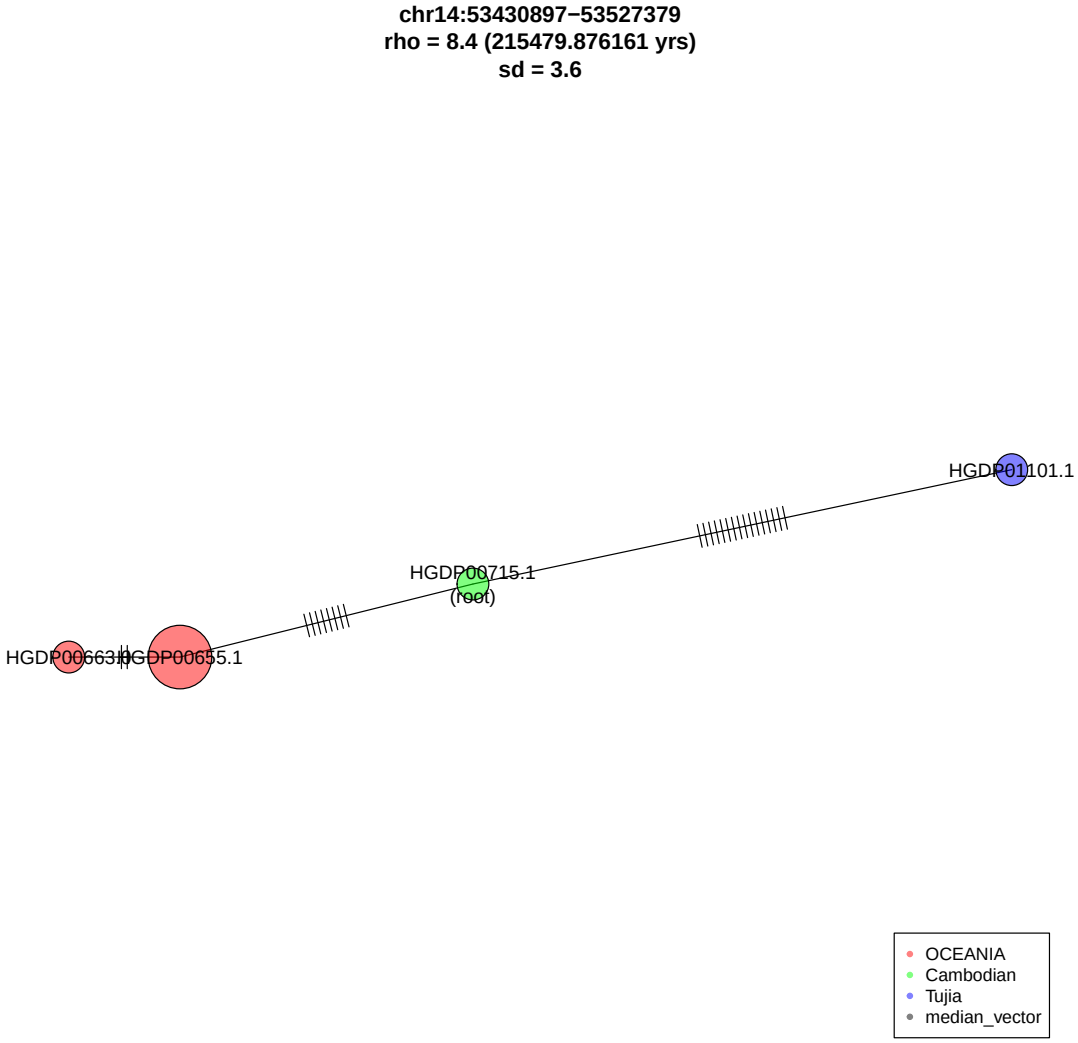

chr9:126565708−126646783
rho = 1.389474 (46613.149219 yrs)
sd = 0.161717
HGDP00748.0
HGDP00949.1
HGDP01102.0
HGDP01221.0
HGDP00764.0
HGDP00769.1
HGDP00771.1
HGDP01099.0
HGDP01228.0
HGDP01239.0
HGDP01299.1
HGDP01314.0
HGDP00734.1
(root)
HGDP00712.1
HGDP00035.0
HGDP01179.0
HGDP01329.0
HGDP01328.0
HGDP01327.0
HGDP01220.0
HGDP01204.0
HGDP00714.1
HGDP00759.1
HGDP00778.1
HGDP01101.0
HGDP00971.0
HGDP00962.0
HGDP01247.1
HGDP00127.1
HGDP00218.0
HGDP01193.0
HGDP00972.0
HGDP00821.0
HGDP00372.0
HGDP00216.0
HGDP00001.1
HGDP00579.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

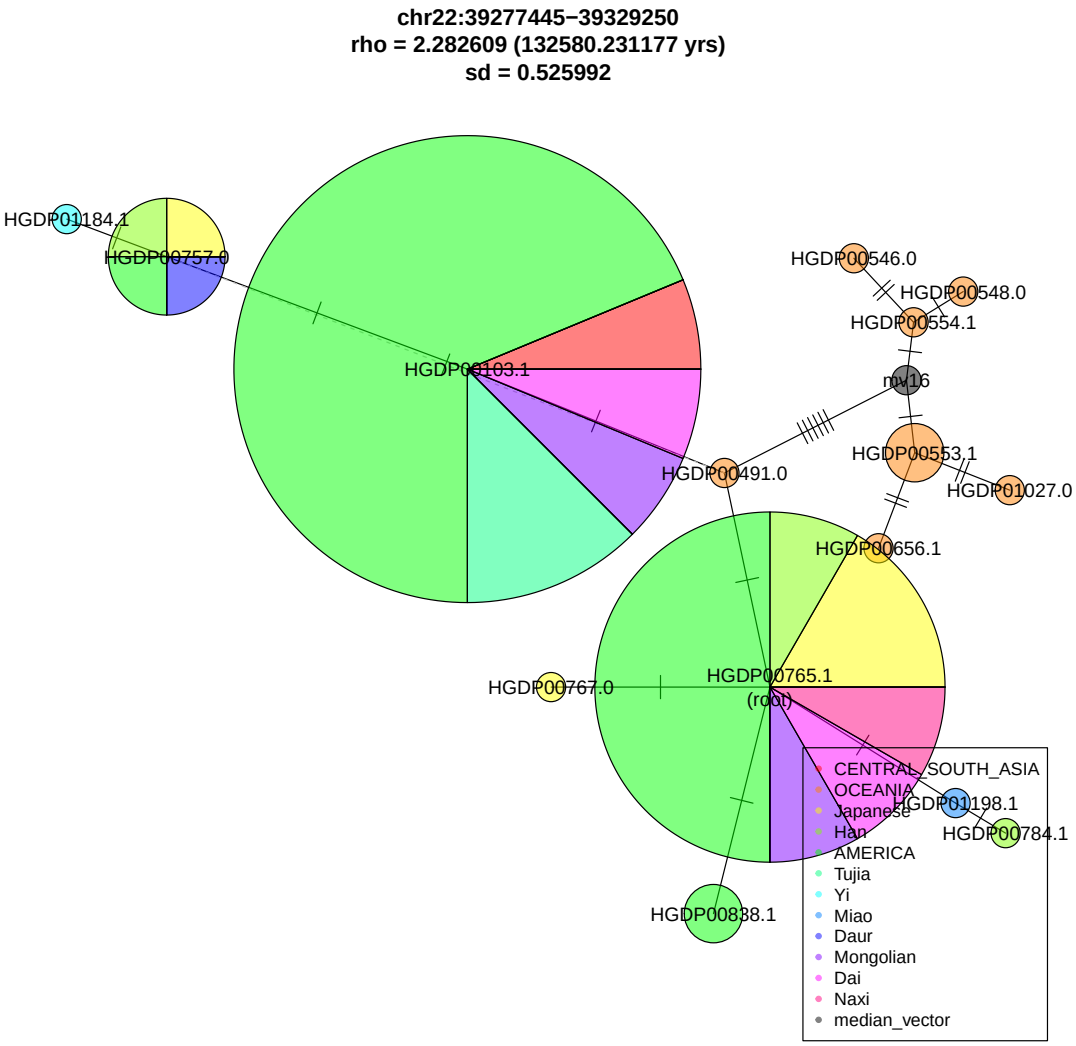

chr1:212466385−212548707
rho = 1.30102 (43196.052249 yrs)
sd = 0.043914
HGDP00769.1
HGDP00882.0
HGDP01187.0
HGDP01191.1
HGDP00859.0
HGDP00864.0
HGDP00873.0
HGDP00880.0
HGDP00971.0
HGDP00121.0
HGDP01356.0
HGDP01196.0
HGDP00757.1
HGDP00963.0
HGDP01352.0
HGDP00122.0
HGDP00279.1
HGDP00521.0
HGDP00522.0
HGDP00564.1
HGDP00712.1
HGDP00392.0
HGDP00601.0
HGDP00149.0
HGDP00716.0
HGDP00776.1
HGDP00712.0
HGDP00811.0
HGDP00813.0
HGDP00690.0
HGDP00843.0
HGDP00946.1
HGDP01287.1
HGDP00956.0
HGDP00959.0
HGDP01348.1
HGDP01296.0
HGDP01340.0
HGDP01317.0
HGDP01239.0
HGDP00754.1
HGDP01234.0
HGDP01214.0
HGDP01195.0
HGDP01179.0
HGDP01101.0
HGDP01100.1
HGDP01047.0
HGDP00855.0
HGDP01009.0
HGDP00099.1
(root)
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector


chr9:94515802−94565893
rho = 0.533784 (38530.752283 yrs)
sd = 0.01082
HGDP01290.1
HGDP01248.1
HGDP01245.0
HGDP01244.1
HGDP01237.0
HGDP01224.0
HGDP01223.0
HGDP01220.1
HGDP01216.0
HGDP01204.1
HGDP01196.0
HGDP01193.0
HGDP01190.1
HGDP00974.0
HGDP00971.0
HGDP00963.1
HGDP00820.1
HGDP00815.0
HGDP00769.1
HGDP00766.0
HGDP00774.1
HGDP00814.0
HGDP01185.0
HGDP01295.0
HGDP00974.1
HGDP01340.0
HGDP01337.0
HGDP01296.0
(root)
CENTRAL_SOUTH_ASIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

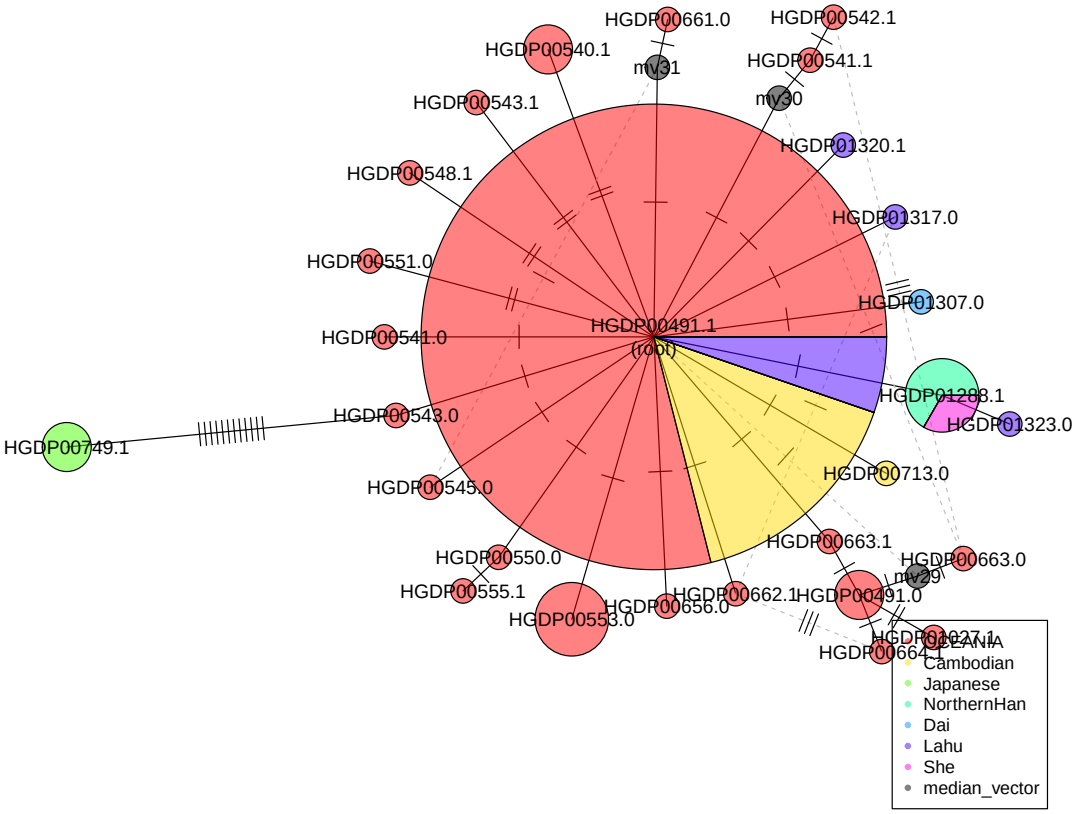

chr1:33403459−33454985
rho = 0.669202 (34713.192365 yrs)
sd = 0.024549
HGDP00955.0
mv64
mv66
mv67
mv69
HGDP01388.1
HGDP01378.0
HGDP01306.0
HGDP01238.0
HGDP01231.1
HGDP01229.0
HGDP01155.0
HGDP01068.1
HGDP00521.0
HGDP00637.0
HGDP00135.0
HGDP00351.0
HGDP00011.1
mv65
HGDP00338.0
HGDP00058.1
HGDP00033.0
(root)
HGDP00963.0
HGDP00879.0
HGDP00748.0
HGDP00725.1
HGDP00672.0
HGDP00643.0
HGDP00582.0
HGDP00561.1
HGDP01077.0
HGDP00356.0
HGDP00247.0
mv71
HGDP00189.1
HGDP00445.0
HGDP00759.1
HGDP00960.0
HGDP01208.0
HGDP00696.0
HGDP00148.1
HGDP01259.0
HGDP00120.0
HGDP00104.1
HGDP00056.1
HGDP00054.1
mv70
HGDP00052.0
HGDP00041.1
HGDP00417.0
CENTRAL_SOUTH_ASIA
OCEANIA
EUROPE
MIDDLE_EAST
AMERICA
EAST_ASIA
median vector

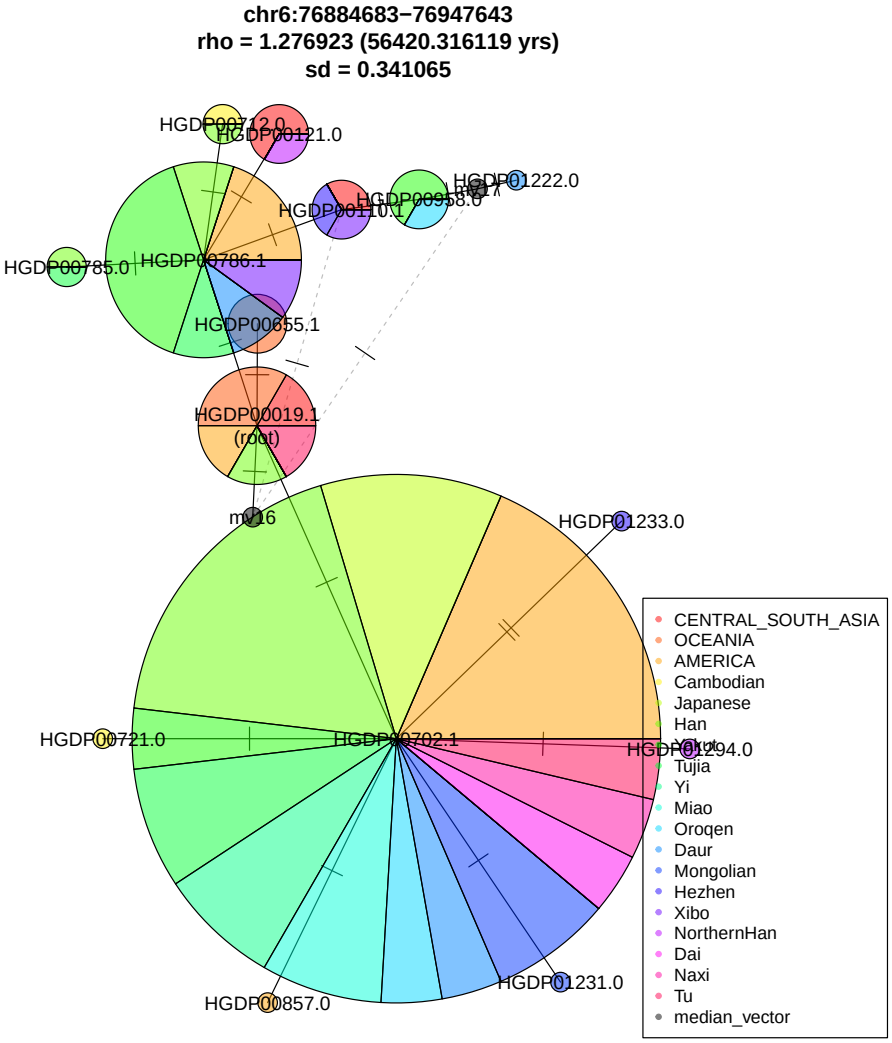
B.4 5 Denisova haplotype networks where East Asia and Oceania are not completely separated

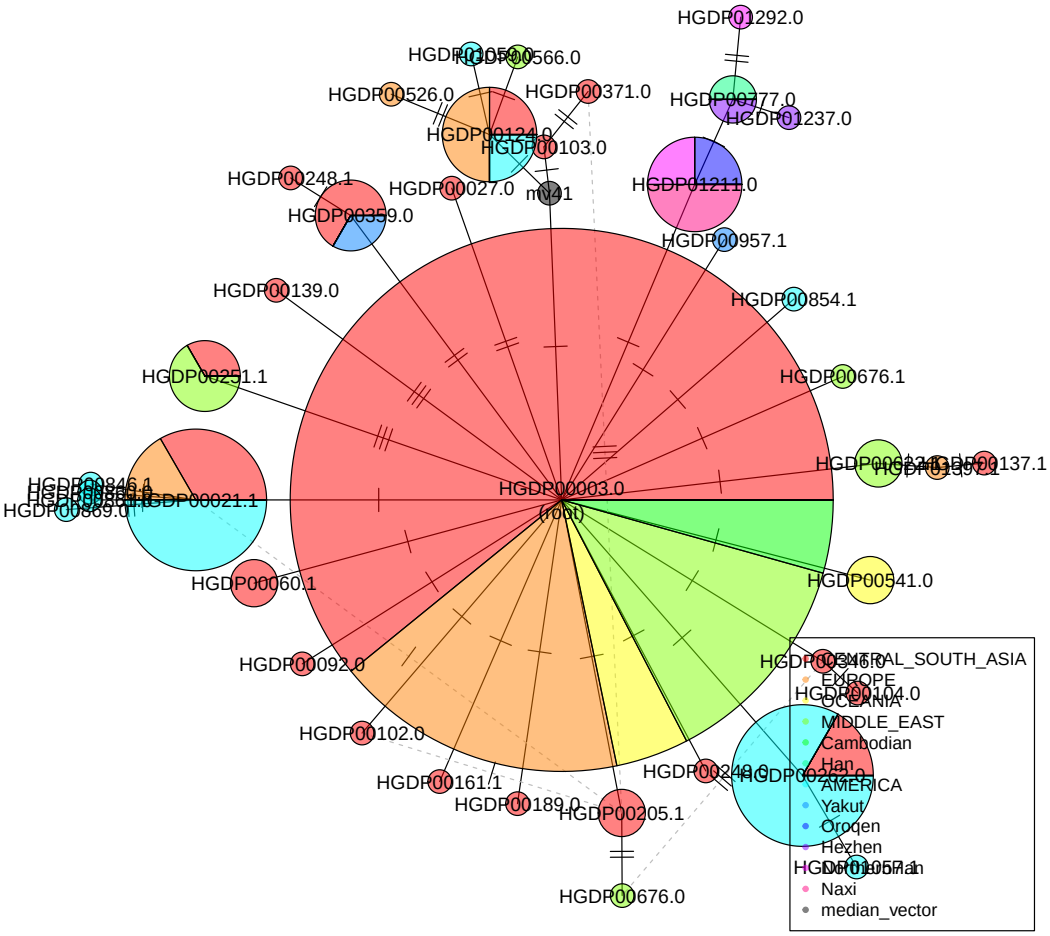



chr22:39277445−39329250
rho = 2.282609 (132580.231177 yrs)
sd = 0.525992
HGDP01184.1
HGDP00757.0
HGDP00103.1
HGDP00546.0
HGDP00548.0
HGDP00554.1
mv16
HGDP00553.1
HGDP00491.0
HGDP01027.0
HGDP00656.1
HGDP00767.0
HGDP00765.1
(root)
HGDP01198.1
HGDP00784.1
HGDP00838.1
CENTRAL_SOUTH_ASIA
OCEANIA
Japanese
Han
AMERICA
Tujia
Yi
Miao
Daur
Mongolian
Dai
Naxi
median_vector

chr2:96945440−97083113
rho = 1.45283 (61189.565713 yrs)
sd = 0.060876




chr6:76884683−76947643
rho = 1.276923 (56420.316119 yrs)
sd = 0.341065
HGDP00712.0
HGDP00121.0
HGDP01222.0
HGDP00958.0
HGDP00110.1
HGDP00785.0
HGDP00786.1
HGDP00855.1
HGDP00019.1
(root)
m916
HGDP01233.0
HGDP00721.0
HGDP00702.1
HGDP01294.0
HGDP01231.0
HGDP00857.0
CENTRAL_SOUTH_ASIA
OCEANIA
AMERICA
Cambodian
Japanese
Han
Tujia
Yi
Miao
Oroqen
Daur
Mongolian
Hezhen
Xibo
NorthernHan
Dai
Naxi
Tu
median_vector


chr15:54655539−54746738
rho = 1.287356 (37791.077763 yrs)
sd = 0.063945
HGDP01292.0
HGDP00566.0
HGDP00526.0
HGDP00371.0
HGDP00377.0
HGDP01237.0
HGDP00124.0
HGDP00103.0
HGDP00248.1
HGDP00027.0
mN41
HGDP01211.0
HGDP00359.0
HGDP00957.1
HGDP00139.0
HGDP00854.1
HGDP00251.1
HGDP00676.1
HGDP00137.1
HGDP00003.0
(root)
HGDP00021.1
HGDP00869.0
HGDP00060.1
HGDP00541.0
HGDP00092.0
HGDP00102.0
HGDP00161.1
HGDP00189.0
HGDP00205.1
HGDP00676.0
CENTRAL_SOUTH_ASIA
EUROPE
OCEANIA
MIDDLE_EAST
Cambodian
Han
AMERICA
Yakut
Oroqen
Hezhen
Naxi
median_vector